Nonsmooth optimization: beyond first order methods.

A tutorial focusing on bundle methods

Claudia Sagastizábal

(IMECC-UniCamp, Campinas Brazil, adjunct researcher)

SESO 2018, Paris, May 23 and 25, 2018

Computational NSO: what do we mean?

For the unconstrained problem

$$\min f(x),$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is convex but not differentiable at some points

Algorithms defined according on **how much** information is provided by certain oracle

Computational NSO: what do we mean?

For the unconstrained problem

$$\min f(x),$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is convex but not differentiable at some points,

Algorithms defined according on **how much** information is provided by certain oracle

an **informative oracle**

Computational NSO: what do we mean?

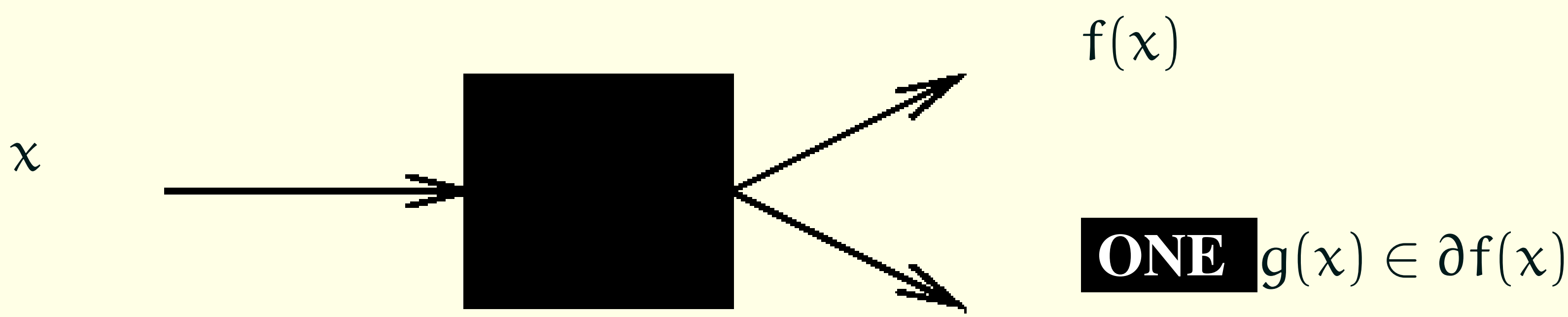
For the unconstrained problem

$$\min f(x),$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is convex but not differentiable at some points

Algorithms defined according on **how much** information is provided by certain oracle

a **“black box”**



Computational NSO: what do we mean?

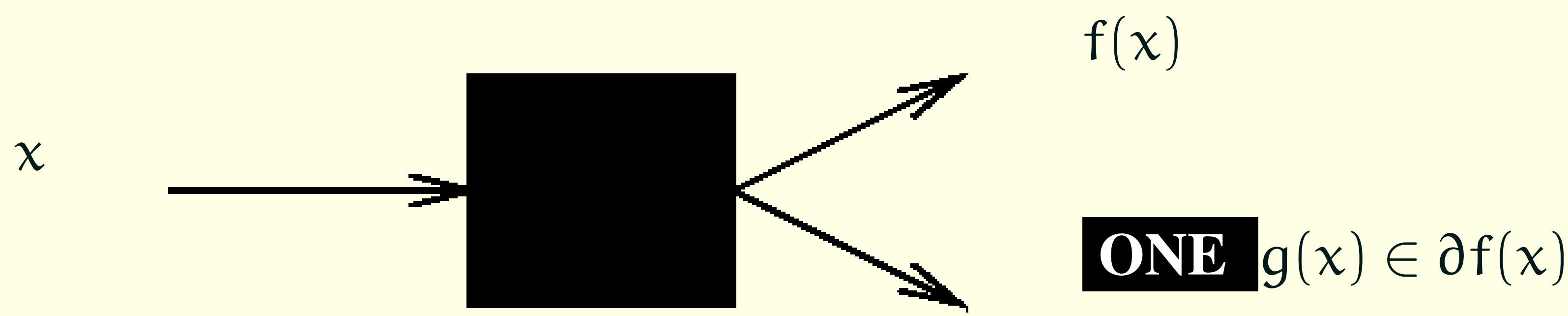
For the unconstrained problem

$$\min f(x),$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is convex but not differentiable at some points,

Algorithms defined according on **how much** information is provided by certain oracle

a **“black box”**



How common are nonsmooth objective functions in optimization?

When does nonsmoothness appear?

* if the **nature** of the problem imposes a nonsmooth model; or
* if **sparsity** of the solution is a concern; or
* in problems difficult to solve,

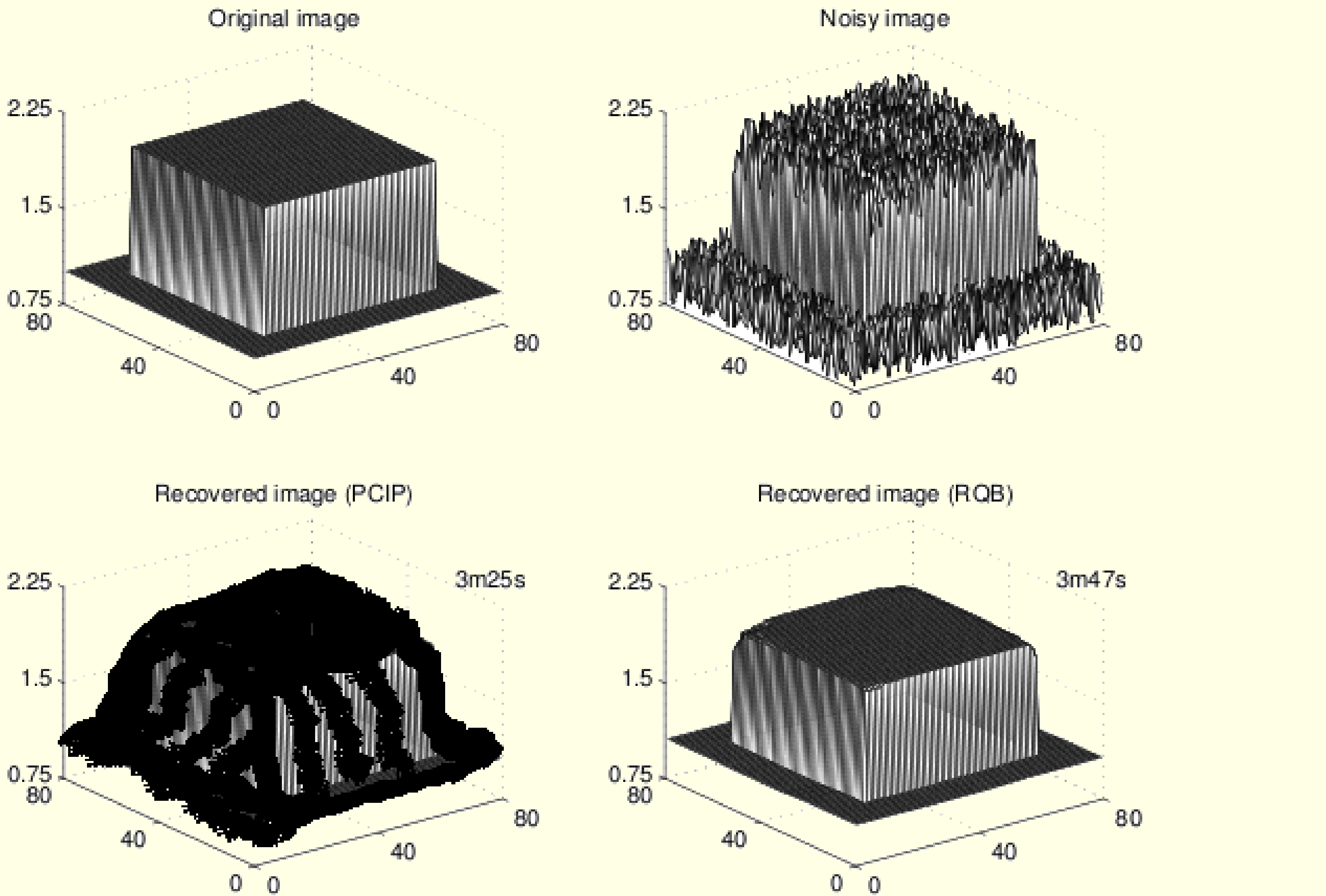
– because they are large scale

– because they are heterogeneous

sometimes the **solution method** induces nonsmoothness

Example of NS model

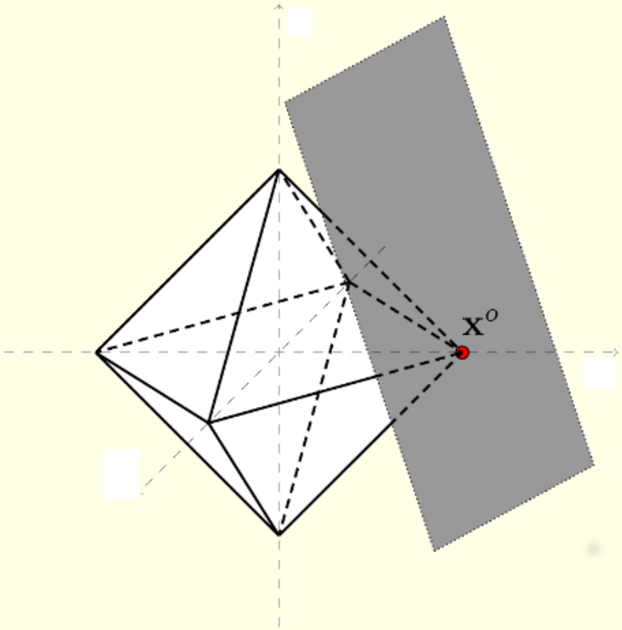
Recovery of **blocky** images (ℓ_1-regularization of TV)



Example of sparse optimization $\min\{\|x\|_1 : Ax = b\}$

Basis pursuit: find least 1-norm point on the affine plane

Tends to return a sparse point (sometimes, the sparsest)

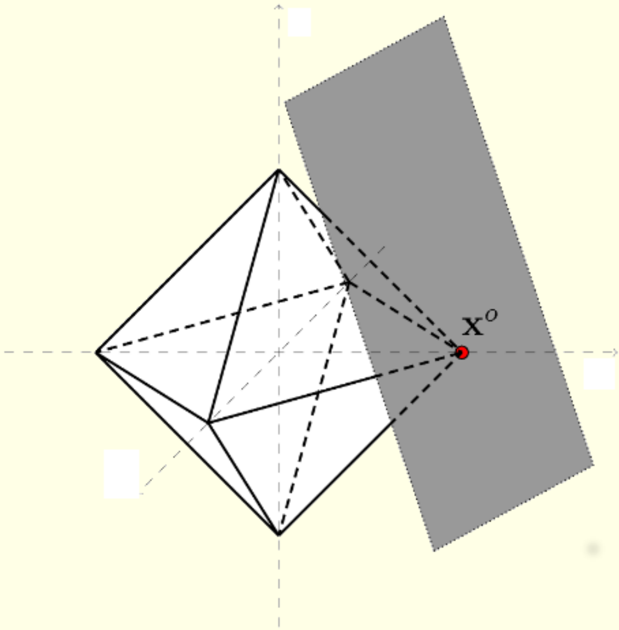


ℓ_1 ball touches the affine plane

Example of sparse optimization $\min\{\|x\|_1 : Ax = b\}$

Basis pursuit: find least 1-norm point on the affine plane

Tends to return a sparse point (sometimes, the sparsest)



ℓ_1 ball touches the affine plane

LASSO denoises basis pursuit

$$\min\left\{\|Ax-b\|_2^2 : \|x\|_1 \le \tau\right\}$$

or

$$\min\left\{\|x\|_1 + \frac{\mu}{2}\|Ax-b\|_2^2\right\}$$

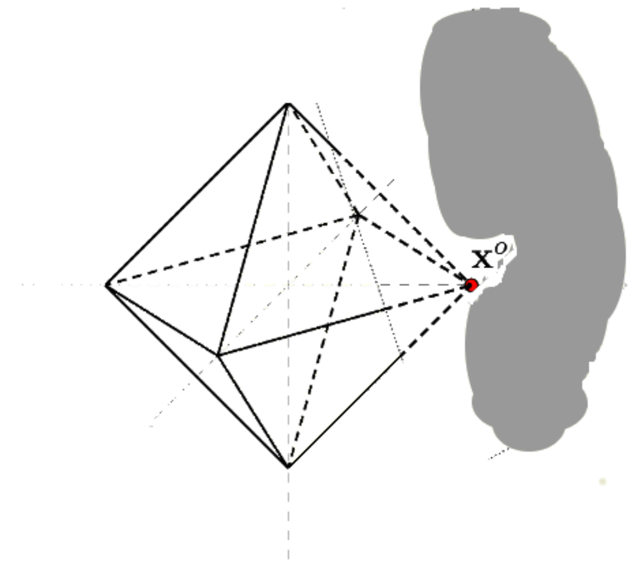
or

$$\min\left\{\|x\|_1 : \|Ax-b\|_2^2 \le \sigma\right\}$$

Example of sparse optimization $\min\{\|x\|_1 : \mathbf{h}(\mathbf{x}) \leq \mathbf{b}\}$

Basis pursuit: find least 1-norm point on a **nonlinear set**

Tends to return a sparse point (sometimes, the sparsest)



ℓ_1 ball touches the set

LASSO denoises basis pursuit

$$\min\left\{\left\|\left(\mathbf{h}(\mathbf{x}) - b\right)^+\right\|_2^2 : \|x\|_1 \leq \tau\right\}$$

or

$$\min\left\{\|x\|_1 + \frac{\mu}{2}\left\|\left(\mathbf{h}(\mathbf{x})x - b\right)^+\right\|_2^2\right\}$$

or

$$\min\left\{\|x\|_1 : \left\|\left(\mathbf{h}(\mathbf{x})x - b\right)^+\right\|_2^2 \leq \sigma\right\}$$

Lagrangian Relaxation Example

Real-life optimization problems

$$
\texttt{(primal)} \quad \begin{cases} \min & \sum_{j \in J} \mathcal{C}^j(p^j) \\ & \text{for } j \in J\text{: } p^j \in \mathcal{P}^j \\ & \sum_{j \in J} g^j(p^j) = \mathrm{Dem} \end{cases}
$$

Lagrangian Relaxation Example

Real-life optimization problems

$$
\texttt{(primal)} \quad \left\{ \begin{array}{ll} \max & \displaystyle\sum_{j \in J} -\mathcal{C}^j(p^j) \\ & \text{for } j \in J\text{: } p^j \in \mathcal{P}^j \\ & \displaystyle\sum_{j \in J} g^j(p^j) = \mathrm{Dem} \end{array} \right. \quad \leftarrow x
$$

Lagrangian Relaxation Example

Real-life optimization problems

$$
\texttt{(primal)} \quad \begin{cases} \max & \sum_{j \in J} -\mathcal{C}^j(p^j) \\ & \text{for } j \in J\text{: } p^j \in \mathcal{P}^j \\ & \sum_{j \in J} g^j(p^j) = \mathrm{Dem} \qquad \leftarrow x \end{cases}
$$

often exhibit separable structure passing to the `(dual)`:

Lagrangian Relaxation Example

Real-life optimization problems

$$
\texttt{(primal)} \quad \begin{cases} \max & \sum_{j \in J} -\mathcal{C}^j(p^j) \\ & \text{for } j \in J\text{: } p^j \in \mathcal{P}^j \\ & \sum_{j \in J} g^j(p^j) = \mathrm{Dem} \quad \leftarrow x \end{cases}
$$

often exhibit separable structure passing to the `(dual)` :

$$
\min_x \quad f(x) := \quad f_0(x) \quad + \sum_{j \in J} \quad f^j(x)
$$

$$
\min_x \quad -\langle x, \mathrm{Dem} \rangle \quad + \sum_{j \in J} \begin{cases} \max & -\mathcal{C}^j(p^j) + \left\langle x, g^j(p^j) \right\rangle \\ & p^j \in \mathcal{P}^j \end{cases}
$$

Lagrangian Relaxation Example

Real-life optimization problems

$$\texttt{(primal)} \quad \begin{cases} \max & \sum_{j\in J} -\mathcal{C}^j(p^j) \\ & \text{for } j \in J\text{: } p^j \in \mathcal{P}^j \\ & \sum_{j\in J} g^j(p^j) = \mathrm{Dem} \end{cases} \quad \leftarrow x$$

often exhibit separable structure passing to the $\texttt{(dual)}$:

$$\min_x \quad f(x) := \quad f_0(x) \quad + \sum_{j\in J} f^j(x)$$

$$\min_x \quad -\langle x, \mathrm{Dem}\rangle \quad + \sum_{j\in J} \begin{cases} \max & -\mathcal{C}^j(p^j) + \left\langle x, g^j(p^j)\right\rangle \\ & p^j \in \mathcal{P}^j \end{cases}$$

Benders Decomposition Example

Similar situation, but now the uncoupling is done on a primal level

$$
\texttt{(primal)} \quad \left\{ \begin{array}{ll} \min & \displaystyle\sum_{j \in J} \mathcal{I}^j(\Delta p^j) + \mathcal{C}^j(p^j) \\ & \text{for } j \in J\text{: } p^j \in \mathcal{P}^j \\ & \Delta p \in D \end{array} \right. \qquad \Longleftrightarrow \quad \mathbf{p^j \leq \bar{p}^j + \Delta p^j}
$$

Benders Decomposition Example

Similar situation, but now the uncoupling is done on a primal level

$$
(\mathtt{primal}) \quad \left\{ \begin{array}{ll} \min & \displaystyle\sum_{j \in J} \mathcal{I}^j(\Delta p^j) + \mathcal{C}^j(p^j) \\ & \text{for } j \in J\text{: } p^j \in \mathcal{P}^j \\ & \Delta p \in D \end{array} \right. \qquad \Longleftrightarrow \quad \mathbf{p^j \leq \bar{p}^j + \Delta p^j}
$$

$$
\left\{ \begin{array}{ll} \min\limits_{\Delta p} & \displaystyle\sum_{j \in J} \mathcal{I}^j(\Delta p^j) + \boldsymbol{\mathcal{V}^j(\Delta p^j)} \\ & \Delta p \in D \end{array} \right. \qquad \mathcal{V}^j(\Delta p^j) := \left\{ \begin{array}{ll} \min & \mathcal{C}^j(p^j) \\ & p^j \leq \bar{p}^j + \Delta p^j \end{array} \right.
$$

$$
\min f(x) := \sum_{j \in J} f^j(\Delta p^j) \quad f^j(\Delta p^j) :=
$$

Benders Decomposition Example

Similar situation, but now the uncoupling is done on a primal level

$$\text{(primal)} \quad \left\{ \begin{array}{ll} \min & \sum_{j \in J} \mathcal{I}^j(\Delta p^j) + \mathcal{C}^j(p^j) \\ & \text{for } j \in J\text{: } p^j \in \mathcal{P}^j \\ & \Delta p \in D \end{array} \right. \qquad \Longleftrightarrow \quad \mathbf{p^j \leq \bar{p}^j + \Delta p^j}$$

$$\left\{ \begin{array}{ll} \min\limits_{\Delta p} & \sum_{j \in J} \mathcal{I}^j(\Delta p^j) + \boldsymbol{\mathcal{V}^j(\Delta p^j)} \\ & \Delta p \in D \end{array} \right. \qquad \mathcal{V}^j(\Delta p^j) := \left\{ \begin{array}{ll} \min & \mathcal{C}^j(p^j) \\ & p^j \leq \bar{p}^j + \Delta p^j \end{array} \right.$$

$$\min f(x) := \sum_{j \in J} f^j(\Delta p^j) \qquad \text{for } f^j(\Delta p^j) := \mathcal{I}^j(\Delta p^j) + \mathcal{V}^j(\Delta p^j)$$

Computing $\partial f(x^k)$: how difficult is it?

1. $f(x) = |x|$, for $n = 1$
2. A linear Lasso function, $f(x) = \|x\|_1 + \frac{\mu}{2}\|Ax - b\|_2^2$
3. A nonlinear Lasso function, $h \in C^1$,
   $f(x) = \|x\|_1 + \frac{\mu}{2}\|\Big(h(x) - b\Big)^+\|_2^2$
4. One of the local subproblems in the Lagrangian example,
   $$f^j(x^k) := \begin{cases} \max & -\mathcal{C}^j(p^j) + \Big\langle x^k, g^j(p^j)\Big\rangle \\ & p^j \in \mathcal{P}^j \end{cases}$$
5. One of the local subproblems in the Benders example,
   $(\mathcal{I}^j(\Delta p^j) + \mathcal{V}^j(\Delta p^j) = f^j(x^{k,j}) = \min\Big\{\mathcal{C}^j(p^j) : p^j \le \bar{p}^j + x^{k,j}\Big\}$

But why would one want **ALL** of $\partial f(x^k)$?

But why would one want ALL of $\partial f(x^k)$?

Indisponsible to calculate the proximal point

$$p = \mathrm{prox}_t^f(x) \iff p = \arg\min f(y) + \frac{1}{2t}\|y - x\|_2^2$$

$$\iff \in \partial f(p) + \frac{1}{t}(p - x)$$

$$\iff$$

But why would one want ALL of $\partial f(x^k)$?

Indisplensible to calculate the proximal point

$$\begin{aligned} p = \operatorname{prox}_t^f(x) \quad &\Longleftrightarrow \quad p = \arg\min f(y) + \frac{1}{2t}\|y - x\|_2^2 \\ &\Longleftrightarrow \quad 0 \in \partial f(p) + \frac{1}{t}(p - x) \end{aligned}$$

But why would one want ALL of $\partial f(x^k)$?

Indisspensible to calculate the proximal point

$$
\begin{aligned}
p = \mathrm{prox}_t^f(x) \quad &\Longleftrightarrow \quad p = \arg\min f(y) + \frac{1}{2t}\|y - x\|_2^2 \\
&\Longleftrightarrow \quad 0 \in \partial f(p) + \frac{1}{t}(p - x) \\
&\Longleftrightarrow \quad \boldsymbol{\frac{1}{t}(x - p) \in \partial f(p)}
\end{aligned}
$$

But why would one want ALL of $\partial f(x^k)$?

Indisprensible to calculate the proximal point

$$\begin{aligned} p = \operatorname{prox}_t^f(x) \quad &\Longleftrightarrow \quad p = \arg\min f(y) + \tfrac{1}{2t}\|y - x\|_2^2 \\ &\Longleftrightarrow \quad 0 \in \partial f(p) + \tfrac{1}{t}(p - x) \\ &\Longleftrightarrow \quad \boxed{\tfrac{\mathbf{1}}{\mathbf{t}}(\mathbf{x} - \mathbf{p}) \in \partial \mathbf{f}(\mathbf{p})} \end{aligned}$$

Without full knowledge of the subdifferential, the **implicit** inclusion cannot be solved!

But why would one want ALL of $\partial f(x^k)$?

Indispensible to calculate the proximal point

$$p = \mathrm{prox}_t^f(x) \iff p = \arg\min f(y) + \frac{1}{2t}\|y - x\|_2^2$$
$$\iff 0 \in \partial f(p) + \frac{1}{t}(p - x)$$
$$\iff \boxed{\frac{1}{t}(\mathbf{x} - \mathbf{p}) \in \partial \mathbf{f}(\mathbf{p})}$$

Without full knowledge of the subdifferential, the **implicit** inclusion cannot be solved!

note: $p \in x - t\partial f(p)$ akin to a subgradient method

Proximal point algorithms (Accel. Nesterov, FISTA, AugLag)

$$x^{k+1} = \mathrm{prox}_{t_k}^{f}(x^k)$$

$$\Longleftrightarrow$$

$$x^{k+1} = \arg\min f(y) + \frac{1}{2t_k}\|y - x^k\|_2^2$$

Proximal point algorithms (Accel. Nesterov, FISTA, AugLag)

$$x^{k+1} = \mathrm{prox}_{t_k}^{f}(x^k)$$

$$\Longleftrightarrow$$

$$x^{k+1} = \arg\min f(y) + \frac{1}{2t_k}\|y - x^k\|_2^2$$

- of interest if computing $\mathrm{prox}_{t_k}^{f}(x^k)$ is much easier than minimizing f
- stepsize $t_k > 0$ impacts on the number of iterations

Proximal point algorithms (Accel. Nesterov, FISTA, AugLag)

$$x^{k+1} = \mathrm{prox}_{t_k}^{f}(x^k)$$

$$\Longleftrightarrow$$

$$x^{k+1} = \arg\min f(y) + \frac{1}{2t_k}\|y - x^k\|_2^2$$

- of interest if **computing $\mathrm{prox}_{t_k}^{f}(x^k)$** is much easier than minimizing f
- stepsize $t_k > 0$ impacts on the number of iterations

Proximal point: calculus rules

- separable sum:

$$f(x,y) = (g(x), h(y)) \Longrightarrow$$
$$\mathrm{prox}_t^f(x) = \Big(\mathrm{prox}_t^g(x), \mathrm{prox}_t^h(y)\Big)$$

- scalar factor ($\alpha \neq 0$) and translation ($v \neq 0$):

$$f(x) = g(\alpha x + v) \Longrightarrow$$
$$\mathrm{prox}_t^f(x) = \frac{1}{\alpha}\Big(\mathrm{prox}_t^{\alpha^2 g}(\alpha x + v) - v\Big)$$

- “perspective" ($\alpha > 0$):

$$f(x) = \alpha g(\tfrac{1}{\alpha}x) \Longrightarrow \mathrm{prox}_t^f(x) = \alpha\,\mathrm{prox}_t^{g/\alpha}(\tfrac{x}{\alpha})$$

Proximal point: special functions

- + linear term ($v \neq 0$):

  $$f(x) = g(x) + \langle v, x \rangle \Longrightarrow \mathrm{prox}_t^f(x) = \mathrm{prox}_t^g(x - v)$$

- + convex quadratic term ($t > 0$):

  $$f(x) = g(x) + \frac{1}{2t}\|x - v\|^2 \Longrightarrow$$

  $$\mathrm{prox}_t^f(x) = \mathrm{prox}_t^{\lambda g}(\lambda x + (1-\lambda)v) \text{ for } \lambda = \frac{t}{t+1}$$

- composition with linear term such that $A^\top A = \frac{1}{\alpha} I$, ($\alpha \neq 0$):

  $$f(x) = g(Ax + v) \Longrightarrow$$

  $$\mathrm{prox}_t^f(x) = (I - \alpha A^\top A)x + \alpha A^\top \left[\mathrm{prox}_t^{g/\alpha}(Ax + v) - v\right]$$

Proximal point algorithm: convergence

If $\arg\min f \neq \emptyset$ then

$$f(x^k) - f(\bar{x}) \leq \frac{\|x^0 - \bar{x}\|^2}{2\sum_{i=1}^{k} t_i}$$

Proximal point algorithm: convergence

If $\operatorname{argmin} f \neq \emptyset$ then

$$f(x^k) - f(\bar{x}) \leq \frac{\|x^0 - \bar{x}\|^2}{2\sum_{i=1}^k t_i}$$

$\Longrightarrow$ convergence if $\sum t_i \to +\infty$

$\Longrightarrow$ rate $1/k$ if $\{t_k\}$ bounded away from zero

Proximal point algorithm: acceleration

$$x^{k+1} = \mathrm{prox}^{f}_{t_k}\left(x^k + \boldsymbol{\theta_{k+1}(\frac{1}{\theta_k} - 1)(x^k - x^{k-1})}\right)$$

for

$$\frac{\theta_{k+1}^2}{t_{k+1}} = (1 - \theta_{k+1})\frac{\theta_k^2}{t_k}$$

Proximal point algorithm: acceleration

$$x^{k+1} = \mathrm{prox}^f_{t_k}\left(x^k + \boldsymbol{\theta_{k+1}\left(\frac{1}{\theta_k} - 1\right)\left(x^k - x^{k-1}\right)}\right)$$

for

$$\frac{\theta^2_{k+1}}{t_{k+1}} = (1 - \theta_{k+1})\frac{\theta^2_k}{t_k}$$

$\Longrightarrow$ convergence if $\sum \sqrt{t_i} \to +\infty$

$\Longrightarrow$ rate $1/k^2$ if $\{t_k\}$ bounded away from zero

What if prox_t^f is not computable?

What if prox_t^f is not computable?

Use bundle methods!

What if prox_t^f is not computable?

Use bundle methods!

When do bundle method prove most useful?

What if prox_t^f is not computable?

Use bundle methods!

When do bundle method prove most useful?

In situations

– when the objective function is not available explicitly

and/or

– when we do not have access to the full subdifferential

and/or

– when calculations need to be done with high precision

Bundling to approximate the prox

WANT: $$p = \operatorname{prox}_t^f(x) \iff p = \arg\min f(y) + \frac{1}{2t}\|y-x\|_2^2$$
$$\iff 0 \in \partial f(p) + \frac{1}{t}(p-x)$$
$$\iff \frac{1}{t}(x-p) \in \partial f(p)$$

Bundling to approximate the prox

WANT: $p = \mathrm{prox}_t^f(x) \iff p = \arg\min f(y) + \frac{1}{2t}\|y - x\|_2^2$

$$\iff 0 \in \partial f(p) + \frac{1}{t}(p - x)$$

$$\iff \frac{1}{t}(x - p) \in \partial f(p)$$

HAVE: $q = \mathrm{prox}_t^{\mathbf{M}}(x) \iff q = \arg\min \mathbf{M}(y) + \frac{1}{2t}\|y - x\|_2^2$

$$\iff 0 \in \partial \mathbf{M}(q) + \frac{1}{t}(q - x)$$

$$\iff \frac{1}{t}(x - q) \in \partial \mathbf{M}(q)$$

Bundling to approximate the prox

WANT: $p = \mathrm{prox}_t^f(x) \iff p = \arg\min f(y) + \frac{1}{2t}\|y - x\|_2^2$

$$\iff 0 \in \partial f(p) + \frac{1}{t}(p - x)$$

$$\iff \frac{1}{t}(x - p) \in \partial f(p)$$

HAVE: $q = \mathrm{prox}_t^{\mathbf{M}}(x) \iff q = \arg\min \mathbf{M}(y) + \frac{1}{2t}\|y - x\|_2^2$

$$\iff 0 \in \partial \mathbf{M}(q) + \frac{1}{t}(q - x)$$

$$\iff \frac{1}{t}(x - q) \in \partial \mathbf{M}(q)$$

$\mathbf{M}$ is a model of f for which we do have full knowledge of the subdifferential: the **implicit** inclusion can be solved!

Bundling to approximate the prox

WANT: $p = \operatorname{prox}_t^f(x) \iff p = \arg\min f(y) + \frac{1}{2t}\|y - x\|_2^2$

$$\iff 0 \in \partial f(p) + \frac{1}{t}(p - x)$$

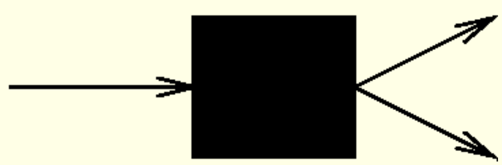
$$\iff \frac{1}{t}(x - p) \in \partial f(p)$$

HAVE: $q = \operatorname{prox}_t^{\mathbf{M}}(x) \iff q = \arg\min \mathbf{M}(y) + \frac{1}{2t}\|y - x\|_2^2$

$$\iff 0 \in \partial \mathbf{M}(q) + \frac{1}{t}(q - x)$$

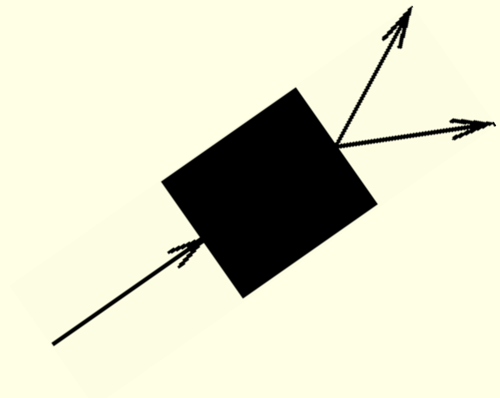
$$\iff \frac{1}{t}(x - q) \in \partial \mathbf{M}(q)$$

M is a model of f for which we do have full knowledge of the subdifferential: the **implicit** inclusion can be solved!



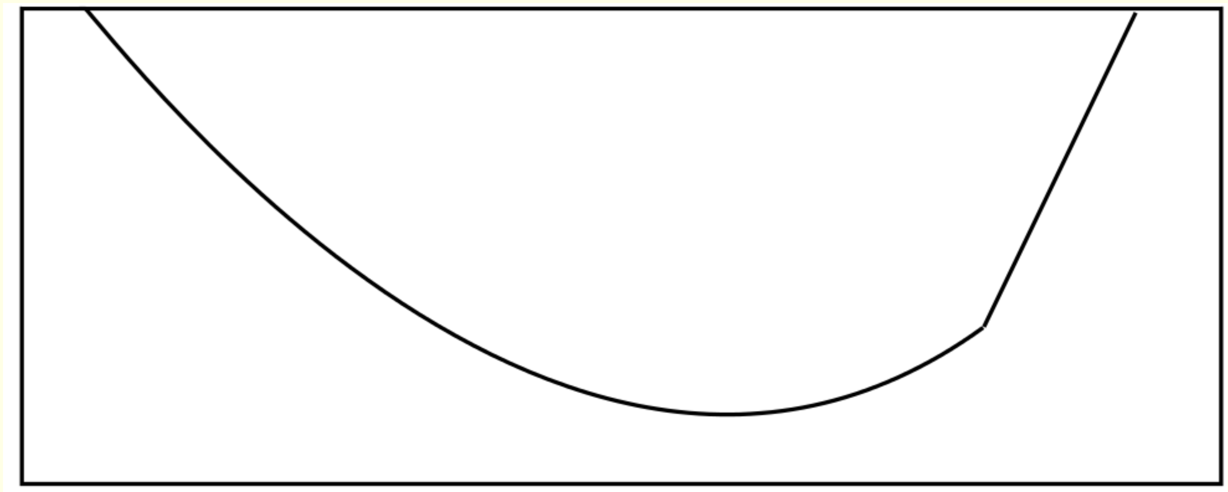
How is the model built?

Model built with the black box



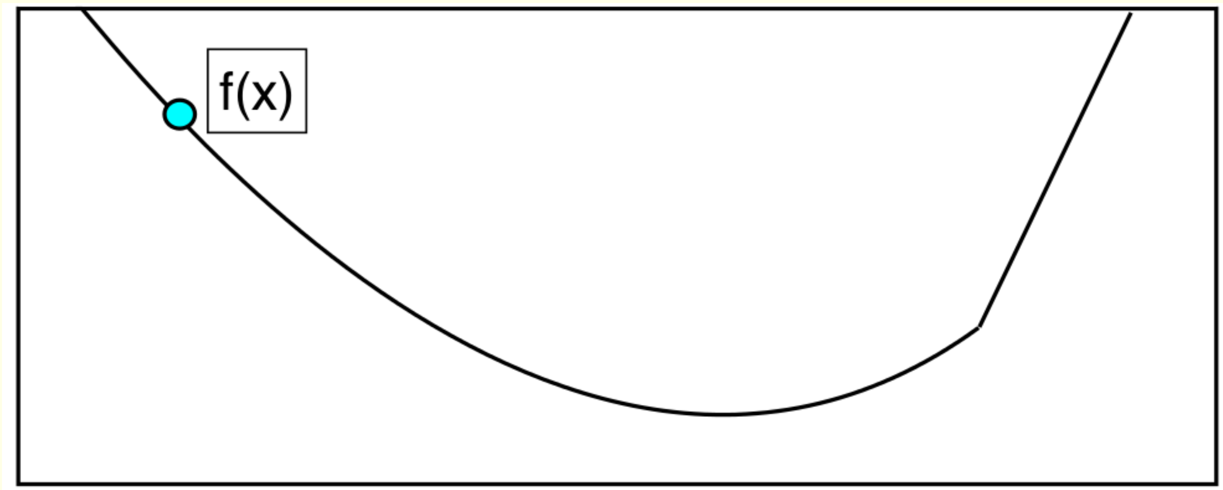
A quick overview of Convex Analysis

An example of a convex nonsmooth function



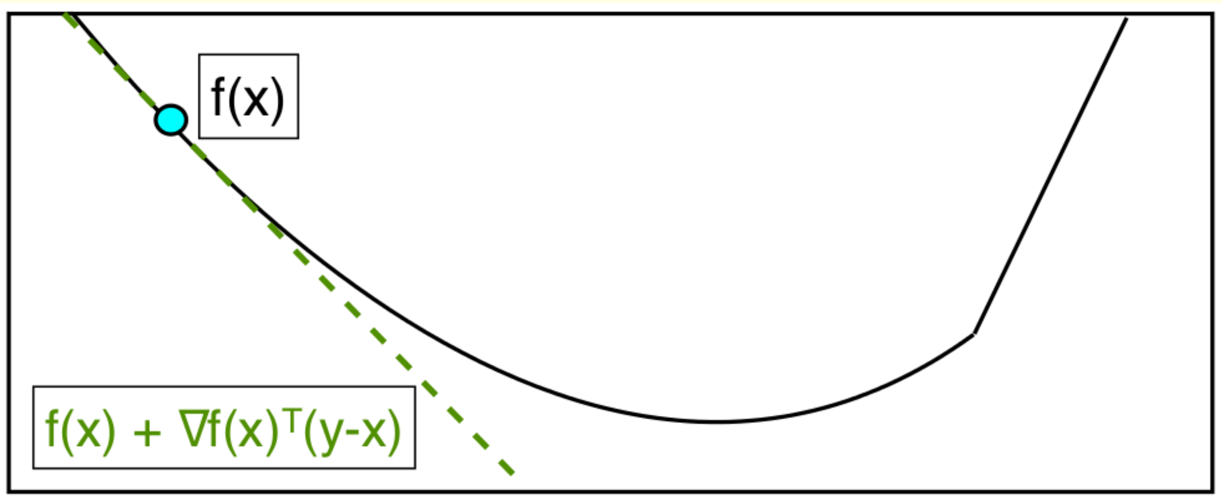
A quick overview of Convex Analysis

An example of a convex nonsmooth function



A quick overview of Convex Analysis

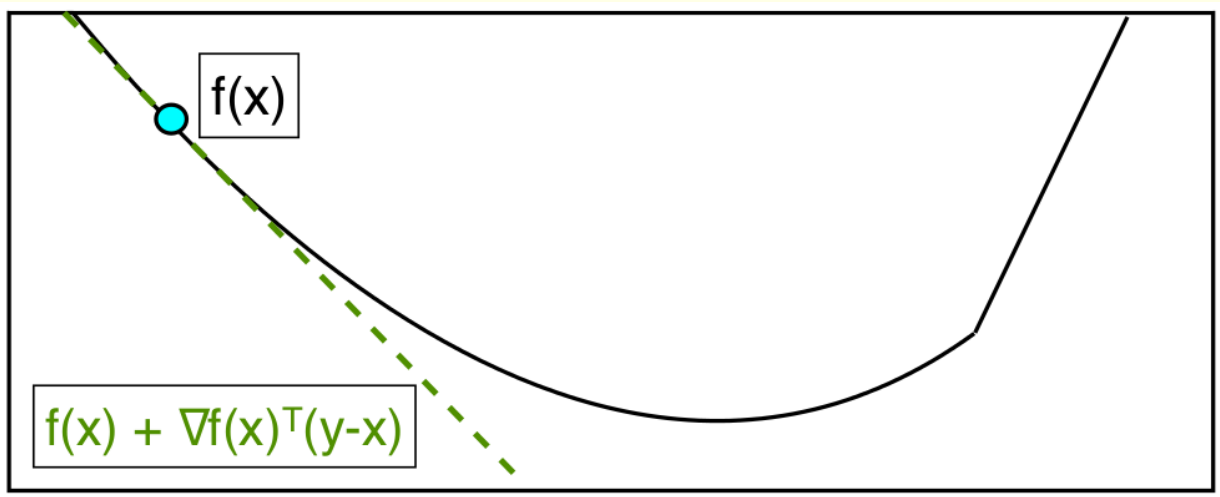
An example of a convex nonsmooth function



$$\{\nabla f(x)\} \quad = \quad \{\text{slope of the linearization supporting } f, \text{ tangent at } x\}$$

A quick overview of Convex Analysis

An example of a convex nonsmooth function



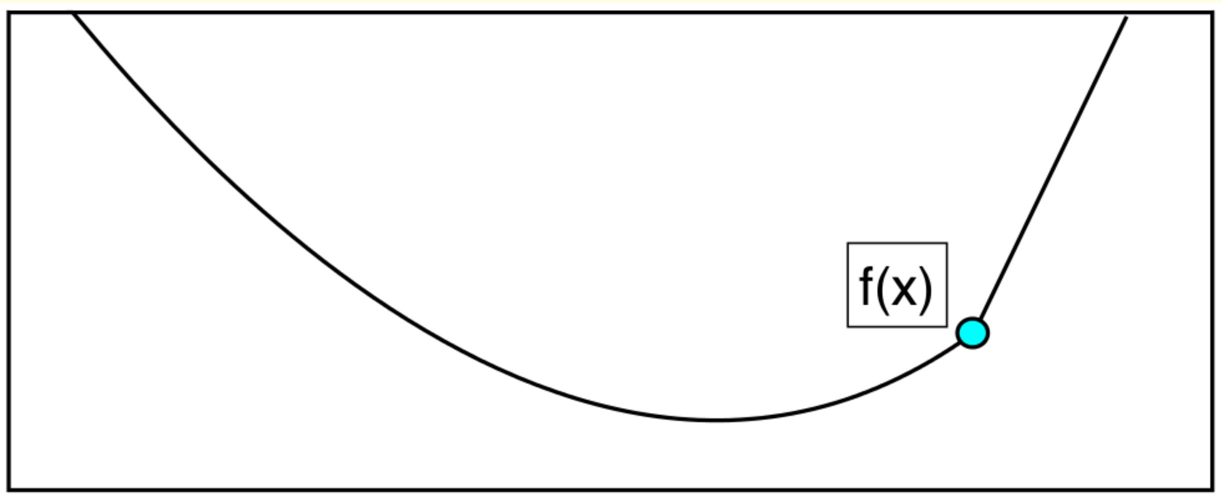
$$\{\nabla f(x)\} \quad = \quad \{\text{slope of the linearization supporting } f\text{, tangent at } x\}$$

By convexity,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \text{ for all } y$$

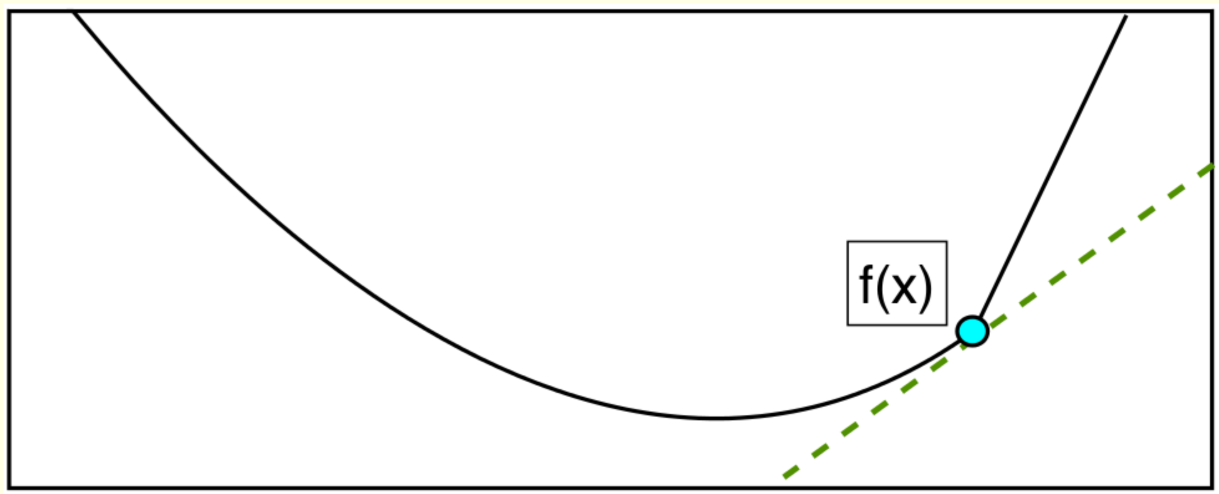
A quick overview of Convex Analysis

An example of a convex nonsmooth function



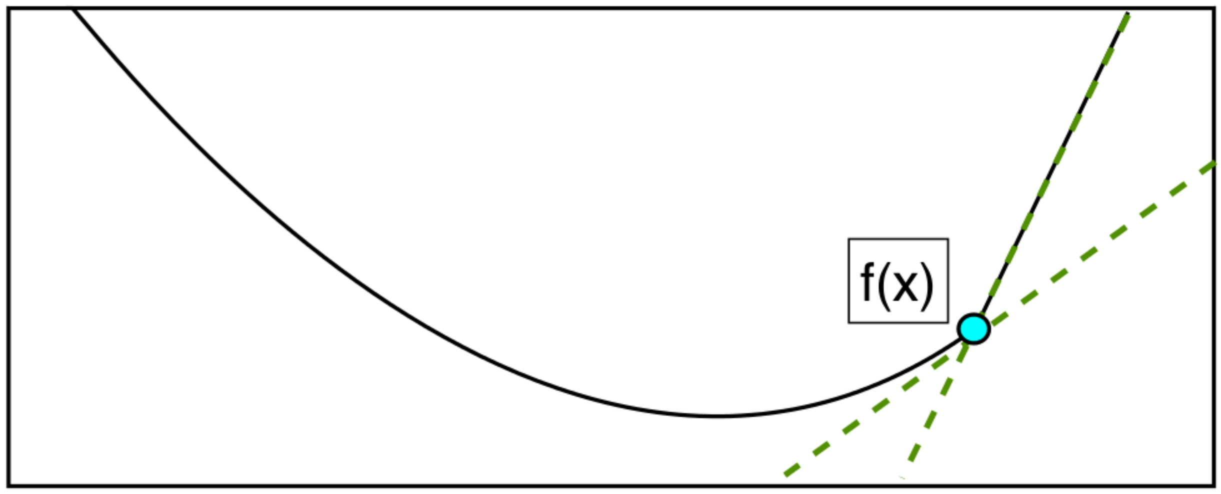
A quick overview of Convex Analysis

An example of a convex nonsmooth function



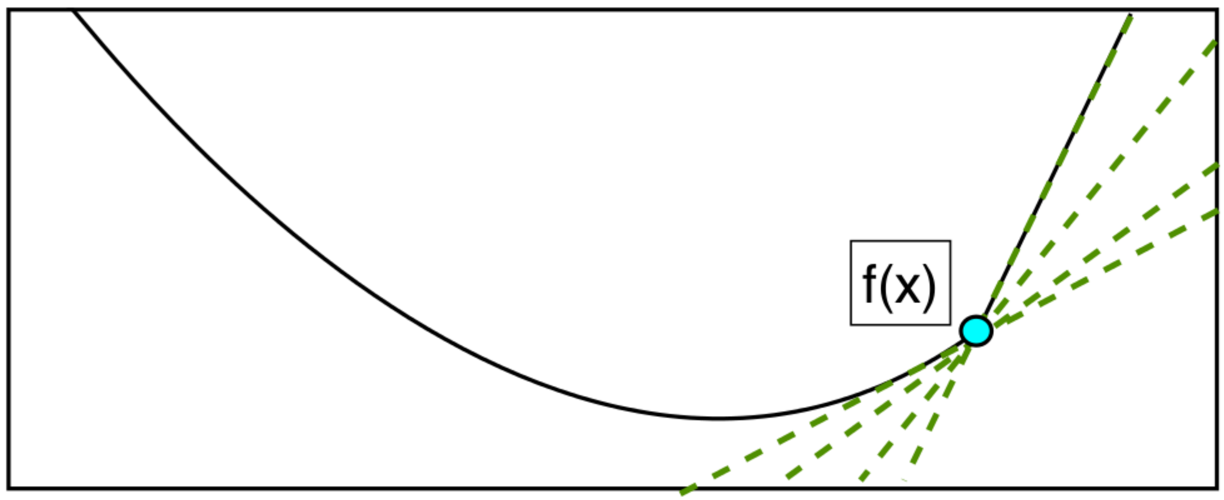
A quick overview of Convex Analysis

An example of a convex nonsmooth function



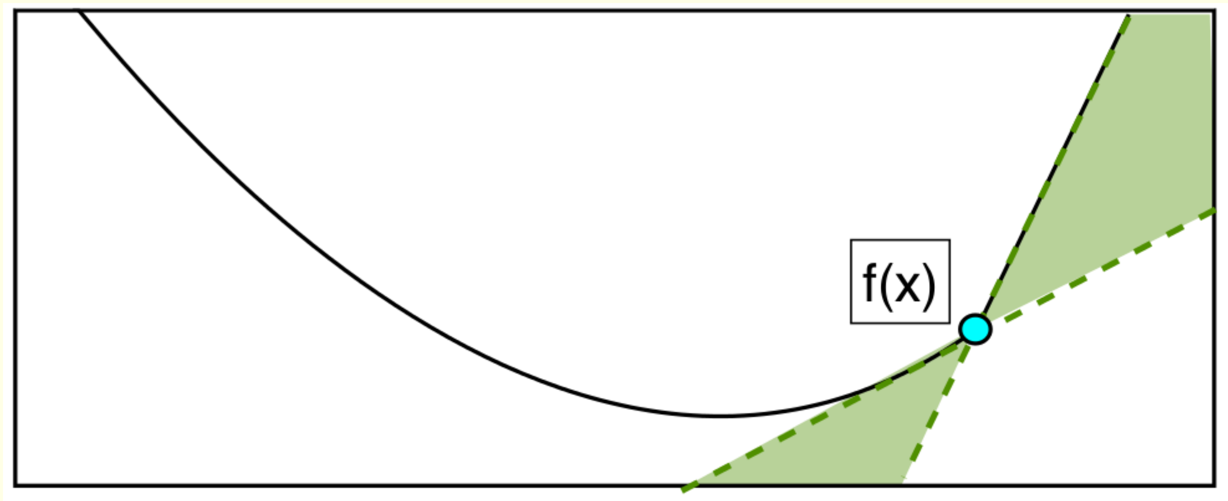
A quick overview of Convex Analysis

An example of a convex nonsmooth function



A quick overview of Convex Analysis

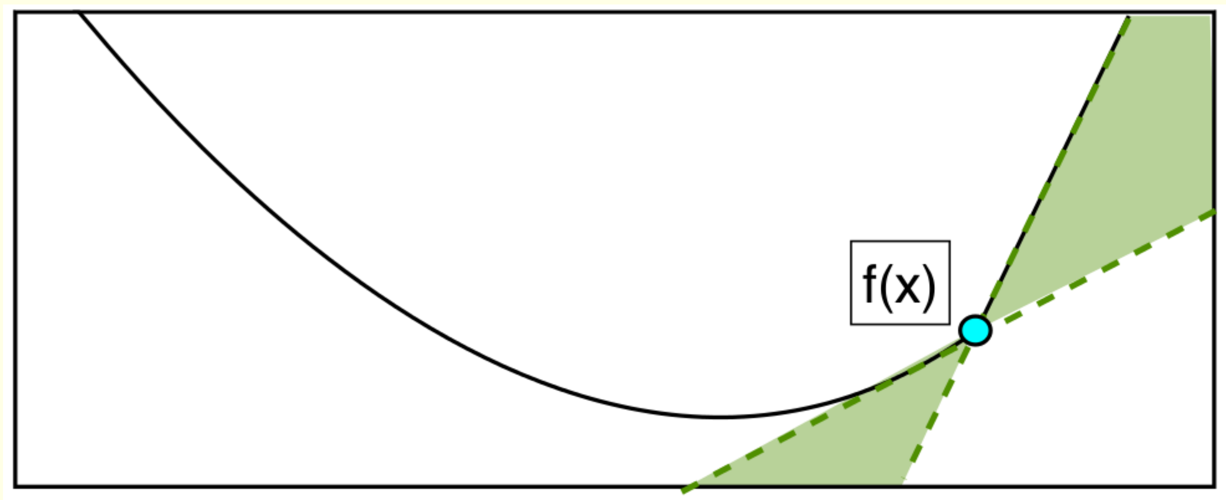
An example of a convex nonsmooth function



$$\partial f(x) \quad = \quad \{g \in \mathbb{R}^n : f(y) \geq f(x) + \langle g, y - x \rangle \text{ for all } y\}$$

A quick overview of Convex Analysis

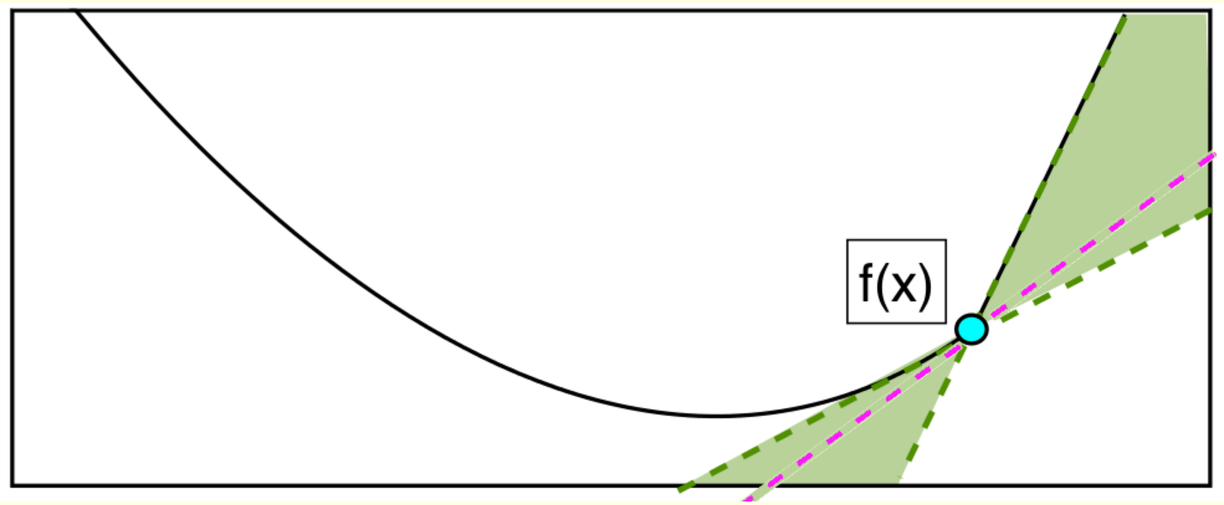
An example of a convex nonsmooth function



$$\begin{aligned}\partial f(x) &= \{g \in \mathbb{R}^n : f(y) \geq f(x) + \langle g, y - x \rangle \text{ for all } y\} \\ &= \{\text{slopes of linearizations supporting } f, \text{ tangent at } x\}\end{aligned}$$

What can be done with the oracle output?

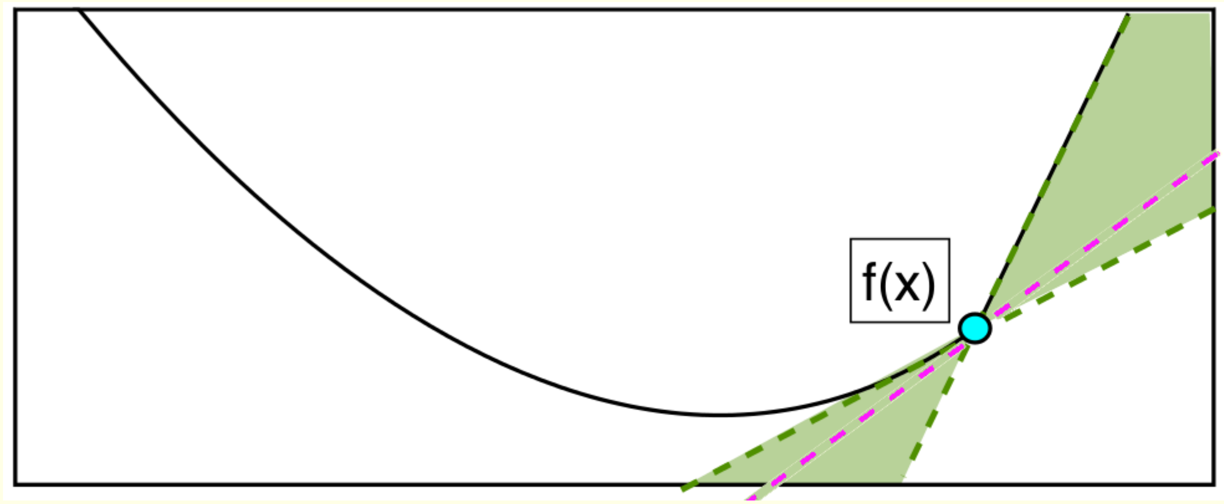
An example of a convex nonsmooth function



$$\begin{aligned} \partial f(x) &= \{g \in \mathbb{R}^n : f(y) \geq f(x) + \langle g, y - x \rangle \text{ for all } y\} \\ &= \{\text{slopes of linearizations supporting } f, \text{ tangent at } x\} \end{aligned}$$

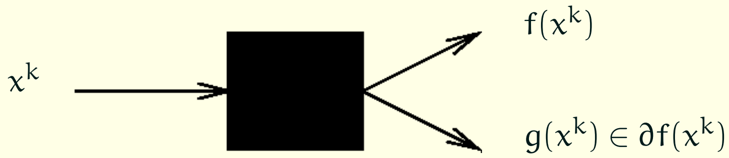
What can be done with the oracle output?

An example of a convex nonsmooth function



$$
\begin{aligned}
\partial f(x) &= \{g \in \mathbb{R}^n : f(y) \geq f(x) + \langle g, y - x\rangle \text{ for all } y\} \\
&= \{\text{slopes of linearizations supporting } f\text{, tangent at } x\}
\end{aligned}
$$

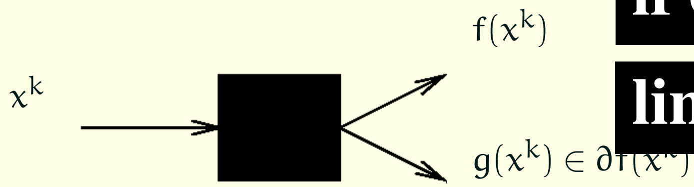
1 oracle call



=

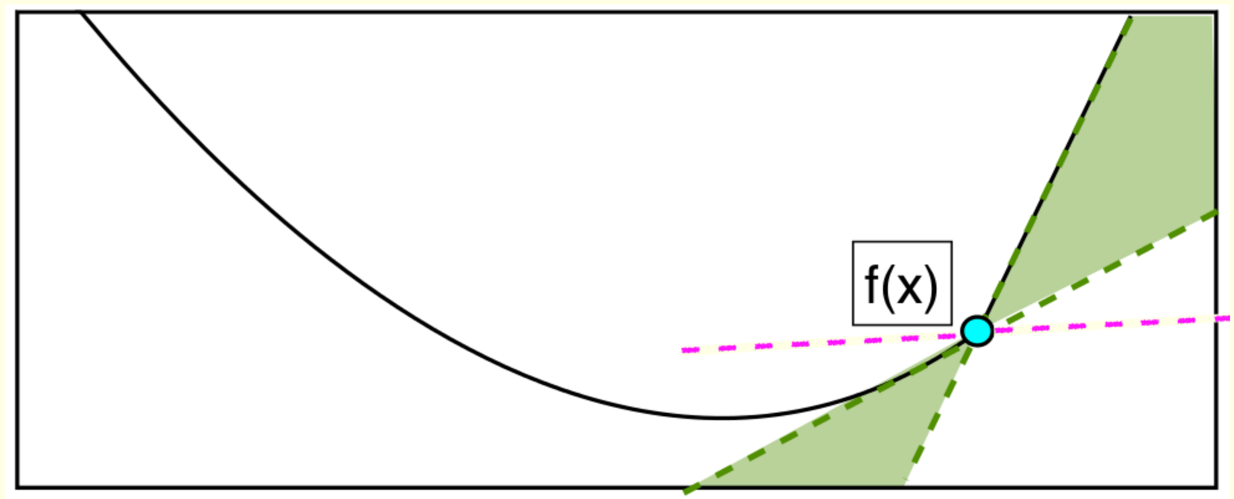
1 linearization

BEWARE:



if oracle output is not accurate,

linearization can be wrong!



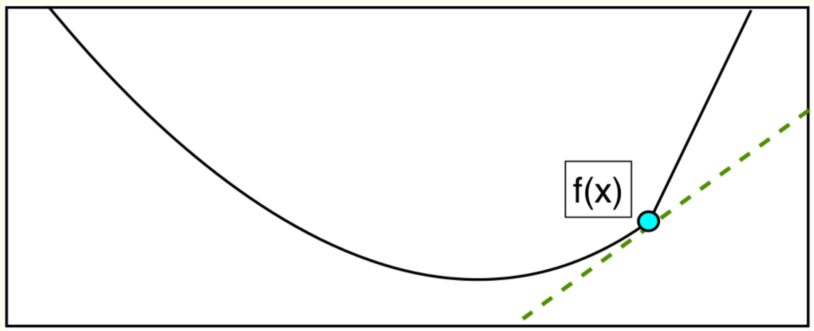
wrong $g(x^k)$ gives bad linearization at x^k

$$\partial f(x) \;=\; \{g \in \mathbb{R}^n : f(y) \geq f(x) + \langle g, y - x\rangle \text{ for all } y\}$$

(similarly if wrong $f(x^k)$, more on this later)

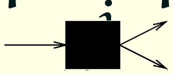
How is the oracle information used?

Putting together linearizations



creates a cutting-plane **model M** for f

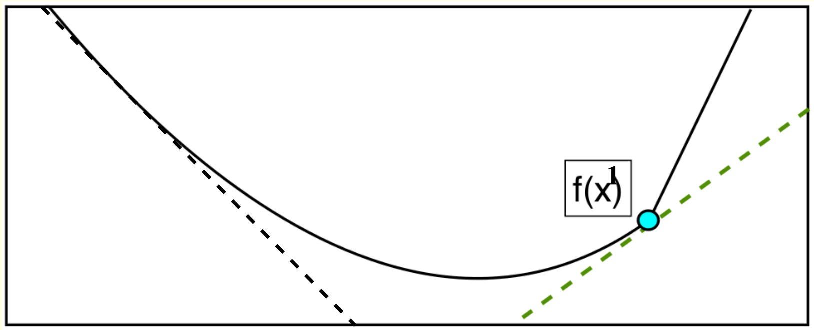
$$f^i = f(x^i)$$



$$g^i = g(x^i) \qquad \Longrightarrow \qquad f^i + \left\langle g^i, x - x^i \right\rangle$$

How is the oracle information used?

Putting together linearizations



creates a cutting-plane **model M** for f

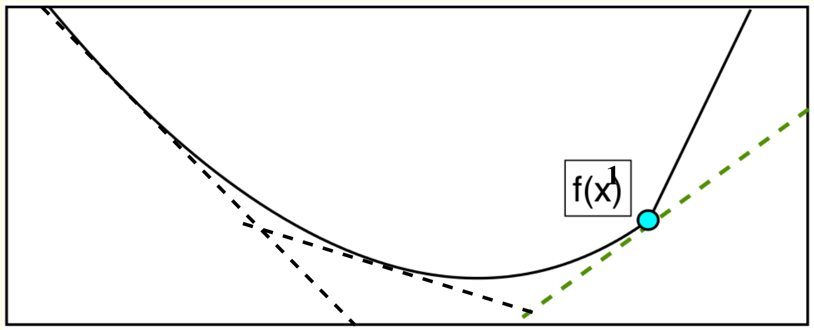
$$f^i = f(x^i)$$

$$g^i = g(x^i)$$

$$\Longrightarrow \mathbf{M}(y) = \max_i \left\{ f^i + \left\langle g^i, x - x^i \right\rangle \right\}$$

How is the oracle information used?

Putting together linearizations



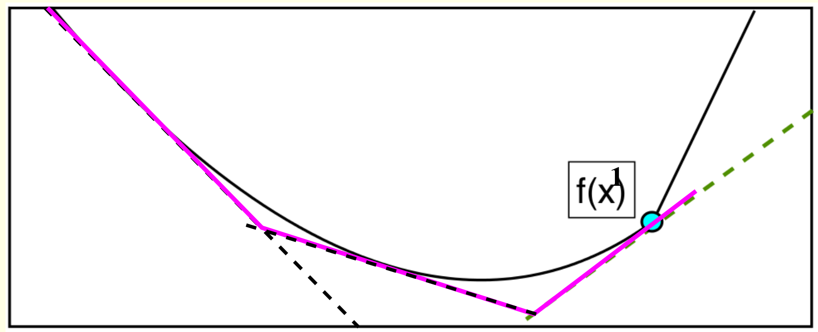
creates a cutting-plane **model M** for f

$$f^i = f(x^i) \qquad \Longrightarrow \mathbf{M}(y) = \max_i \left\{ f^i + \left\langle g^i, x - x^i \right\rangle \right\}$$

$$g^i = g(x^i)$$

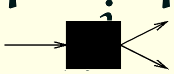
How is the oracle information used?

Putting together linearizations



creates a cutting-plane **model M** for f

$f^i = f(x^i)$



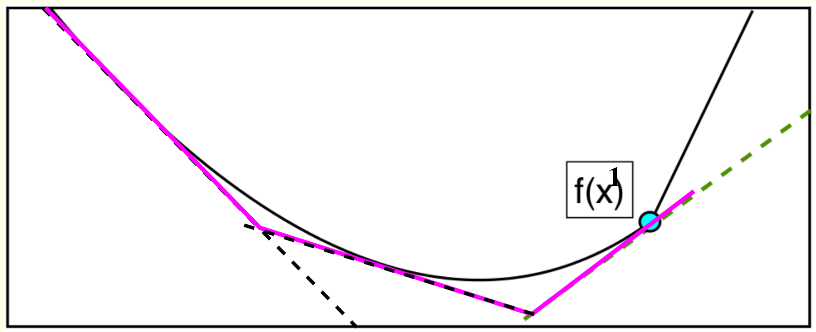
$g^i = g(x^i)$

$$\Longrightarrow \mathbf{M}(y) = \max_i \left\{ f^i + \left\langle g^i, y - x^i \right\rangle \right\}$$

(just one type of model, many others are possible)

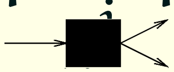
How is the oracle information used?

Putting together linearizations



creates a cutting-plane **model M** for f

$$f^i = f(x^i)$$



$$g^i = g(x^i)$$

$$\Longrightarrow \mathbf{M}_k(y) = \max_{i \le k} \left\{ f^i + \left\langle g^i, y - x^i \right\rangle \right\}$$

(just one type of model, many others are possible)

Infinite bundling yields prox_t^f

$$\text{WANT: } p = \operatorname{prox}_t^f(x) \iff p = \arg\min f(y) + \frac{1}{2t}\|y - x\|_2^2$$

$$\text{HAVE: } q^k = \operatorname{prox}_{t_k}^{M_k}(x) \iff q^k = \arg\min M_k(y) + \frac{1}{2t_k}\|y - x^k\|_2^2$$

$$\iff 0 = G^k + \frac{1}{t_k}(q^k - x^k)$$

for $G^k \in \partial M_k(q^k)$

Theorem [CL93] Suppose the models satisfy

- $M_k(y) \le f(y)$ for all k and y
- $M_{k+1}(y) \ge f(q^k) + \left\langle g(q^k), y - x^k \right\rangle$
- $M_{k+1}(y) \ge M_k(q^k) + \left\langle G^k, y - x^k \right\rangle$

If $0 < t_{\min} \le t_{k+1} \le t_k$, then

$$\lim_{k\to\infty} q^k = p \quad \text{and} \quad \lim_{k\to\infty} M_k(q^k) = f(p)$$

Infinite bundling yields prox_t^f

$$\text{WANT: } p = \operatorname{prox}_t^f(x) \quad \Longleftrightarrow \quad p = \arg\min f(y) + \tfrac{1}{2t}\|y - x\|_2^2$$

$$\text{HAVE: } q^k = \operatorname{prox}_{t_k}^{M_k}(x) \quad \Longleftrightarrow \quad q^k = \arg\min M_k(y) + \tfrac{1}{2t_k}\|y - x^k\|_2^2$$

$$\Longleftrightarrow \quad 0 = G^k + \tfrac{1}{t_k}(q^k - x^k)$$

for $G^k \in \partial M_k(q^k)$

Theorem [CL93] Suppose the models satisfy

- $M_k(y) \leq f(y)$ for all k and y
- $M_{k+1}(y) \geq f(q^k) + \left\langle g(q^k), y - x^k \right\rangle$
- $M_{k+1}(y) \geq M_k(q^k) + \left\langle G^k, y - x^k \right\rangle$

If $0 < t_{\min} \leq t_{k+1} \leq t_k$, then

$$\lim_{k\to\infty} q^k = p \quad \text{and} \quad \lim_{k\to\infty} M_k(q^k) = f(p)$$

Models for the half-and-half function

STRUCTURE	$f(x)$	
none	$\sqrt{x^\top A x} + x^\top B x$	
sum	$f_1(x) + f_2(x)$	$f_1(x) = \sqrt{x^\top A x}$ $f_2(x) = x^\top B x$ f_2 is smooth
compo sition	$(h \circ c)(x)$	$c(x) = (x, x^\top B x) \in \mathbb{R}^{n+1}$ c $h(C) = \sqrt{C_{1:n}^\top A C_{1:n}} + C_{n+1}$ h

Models for the half-and-half function

STRUCTURE	$f(x)$	
none	$\sqrt{x^\top A x} + x^\top B x$	
sum	$f_1(x) + f_2(x)$	$f_1(x) = \sqrt{x^\top A x}$ $f_2(x) = x^\top B x$ f_2 is smooth
compo sition	$(h \circ c)(x)$	$c(x) = (x, x^\top B x) \in \mathbb{R}^{n+1}$ c is smooth $h(C) = \sqrt{C_{1:n}^\top A C_{1:n}} + C_{n+1}$ h is sublinear

Models for the half-and-half function

STRUCTURE	$f(x)$	
none	$\sqrt{x^\top A x} + x^\top B x$	$f^k := f(x^k),\ g^k \in \partial f(x^k)$
sum	$f_1(x) + f_2(x)$	$f_1(x) = \sqrt{x^\top A x}$ $f_2(x) = x^\top B x$
compo sition	$(h \circ c)(x)$	$c(x) = (x, x^\top B x) \in \mathbb{R}^{n+1}$ $h(C) = \sqrt{C_{1:n}^\top A C_{1:n}} + C_{n+1}$

Models for the half-and-half function

STRUCTURE	$f(x)$	
none	$\sqrt{x^\top A x} + x^\top B x$	$f^k := f(x^k),\ g^k \in \partial f(x^k)$
sum	$f_1(x) + f_2(x)$	$f_1(x) = \sqrt{x^\top A x}$ $f_2(x) = x^\top B x$ $f_1^k, g_1^k, f_2^k, \nabla f_2(x^k)$
compo **sition**	$(h \circ c)(x)$	$c(x) = (x, x^\top B x) \in \mathbb{R}^{n+1}$ $h(C) = \sqrt{C_{1:n}^\top A C_{1:n}} + C_{n+1}$

Models for the half-and-half function

STRUCTURE	$f(x)$	
none	$\sqrt{x^\top A x} + x^\top B x$	$f^k := f(x^k),\ g^k \in \partial f(x^k)$
sum	$f_1(x) + f_2(x)$	$f_1(x) = \sqrt{x^\top A x}$ $f_2(x) = x^\top B x$ $f_1^k, g_1^k, f_2^k, \nabla f_2(x^k)$
compo sition	$(h \circ c)(x)$	$c(x) = (x, x^\top B x) \in \mathbb{R}^{n+1}$ $c^k = c(x^k), c'(x^k)$ $h(C) = \sqrt{C_{1:n}^\top A C_{1:n}} + C_{n+1}$ $h^k, g^k \in \partial h(c^k)$

NSO pitfalls

NSO pitfalls

NSO pitfalls

NSO pitfalls

Stopping test in smooth optimization

Algorithms for unconstrained smooth optimization use as optimality certificate Fermat's rule

$$0 = \nabla f(\bar{x})$$

and generate a minimizing sequence:

$$\{x^k\} \to \bar{x} \text{ such that } \nabla f(x^k) \to 0.$$

If $f \in C^1$, then $\nabla f(\bar{x}) = 0$

Stopping test in smooth optimization

Algorithms for unconstrained smooth optimization use as optimality certificate Fermat's rule

$$0 = \nabla f(\bar{x})$$

and generate a minimizing sequence:

$$\{x^k\} \to \bar{x} \text{ such that } \nabla f(x^k) \to 0.$$

If $f \in C^1$, then $\nabla f(\bar{x}) = 0$ things are less straightforward if f is nonsmooth…

What happens with the stopping test in NSO?

Algorithms for unconstrained NSO use as optimality certificate the inclusion

$$0 \in \partial f(\bar{x})$$

- As a set-valued mapping $\partial f(x)$ is osc:

$$\left(x^k, g(x^k) \in \partial f(x^k)\right) : \begin{cases} x^k \to \bar{x} \\ g(x^k) \to \bar{g} \end{cases} \Longrightarrow \bar{g} \in \partial f(\bar{x})$$

What happens with the stopping test in NSO?

Algorithms for unconstrained NSO use as optimality certificate the inclusion

$$0 \in \partial f(\bar{x})$$

- As a set-valued mapping $\partial f(x)$ is osc:

$$\left(x^k, g(x^k) \in \partial f(x^k)\right) : \begin{cases} x^k \to \bar{x} \\ g(x^k) \to \bar{g} \end{cases} \Longrightarrow \bar{g} \in \partial f(\bar{x})$$

- As a set-valued mapping, $\partial f(x)$ is **not** isc:
Given $\bar{g} \in \partial f(\bar{x})$

$$\not\exists \left(x^k, g(x^k) \in \partial f(x^k)\right) : \begin{cases} x^k \to \bar{x} \\ g(x^k) \to \bar{g} \end{cases}$$

What happens with the stopping test in NSO?

Algorithms for unconstrained NSO use as optimality certificate the inclusion

$$0 \in \partial f(\bar{x})$$

- As a set-valued mapping $\partial f(x)$ is osc:

$$\left(x^k, g(x^k) \in \partial f(x^k)\right) : \begin{cases} x^k \to \bar{x} \\ g(x^k) \to \bar{g} \end{cases} \Longrightarrow \bar{g} \in \partial f(\bar{x})$$

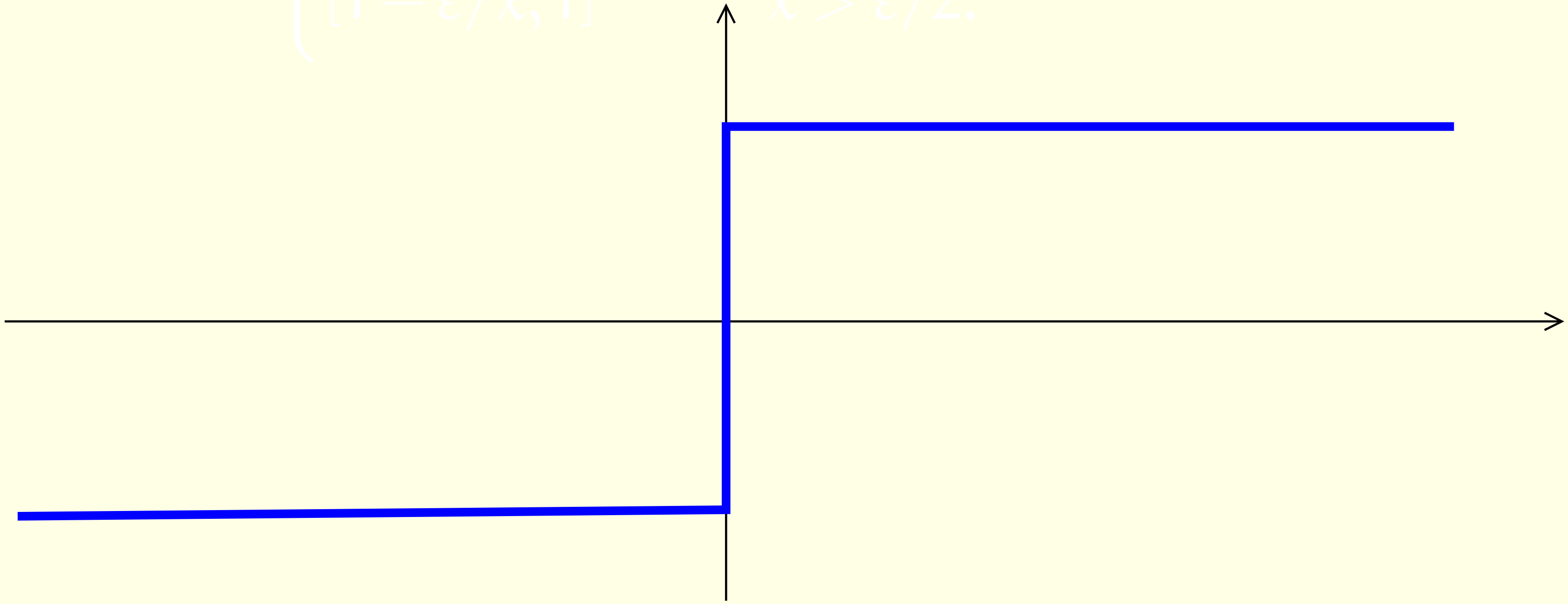
- As a set-valued mapping, $\partial f(x)$ is **not** isc:
  Given $\bar{g} \in \partial f(\bar{x})$

$$\cancel{\exists}\left(x^k, g(x^k) \in \partial f(x^k)\right) : \begin{cases} x^k \to \bar{x} \\ g(x^k) \to \bar{g} \end{cases}$$

The subdifferential

For the absolute value function, $f(x) = |x|$

$$\partial f(x) = \begin{cases} -1 & x < 0 \\ [-1, 1] & x = 0 \\ 1 & x > 0 \end{cases}$$



What happens with the stopping test in NSO?

We need to design a sound stopping test that does not rely on the straightforward extension of Fermat's rule.

What happens with the stopping test in NSO?

We need to design a sound stopping test that does not rely on the straightforward extension of Fermat's rule. We use instead

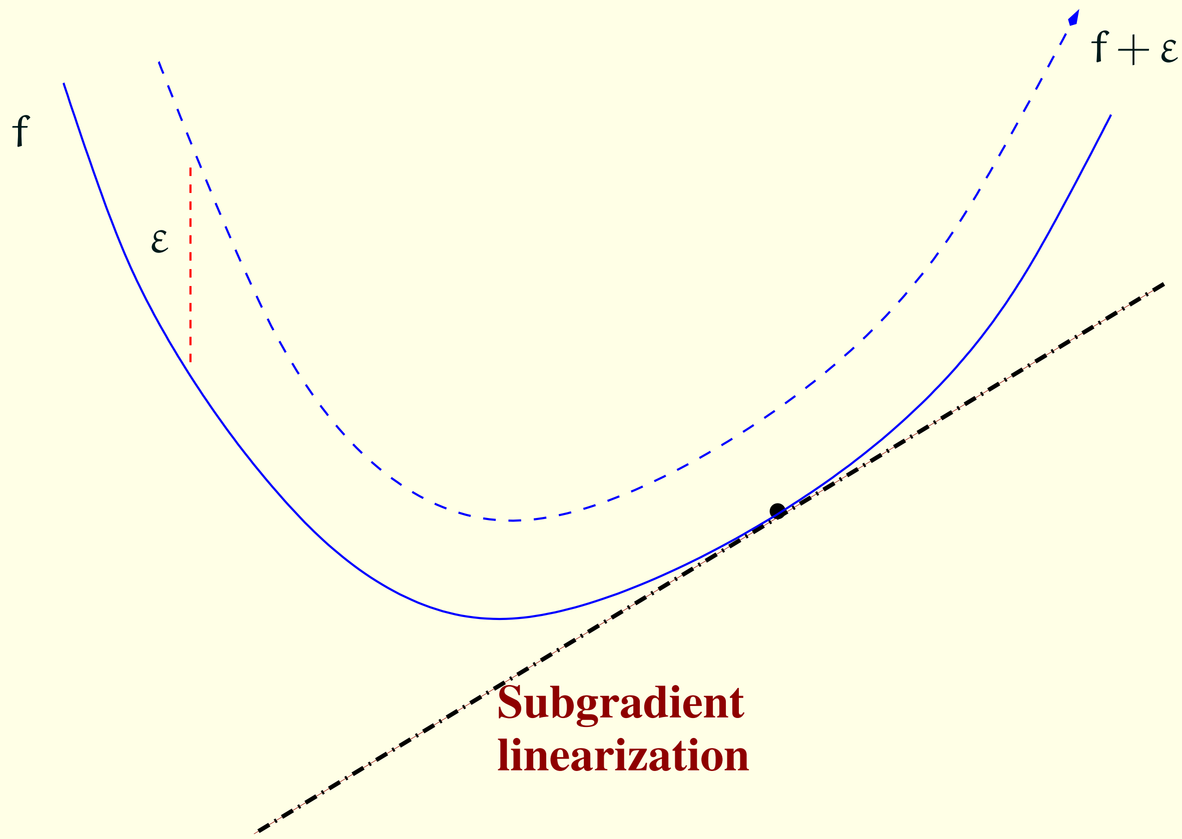
$$\bar{g} \in \partial_{\bar{\varepsilon}} f(\bar{x}) \qquad \text{for } \|\bar{g}\| \text{ and } \bar{\varepsilon} \text{ small}$$

where the ε-subdifferential contains the slopes of linearizations supporting f **up to** ε, tangent at x:

$$\partial_\varepsilon f(x) = \{g \in \mathbb{R}^n : f(y) \geq f(x) + \langle g, y - x\rangle - \varepsilon \text{ for all } y\}$$

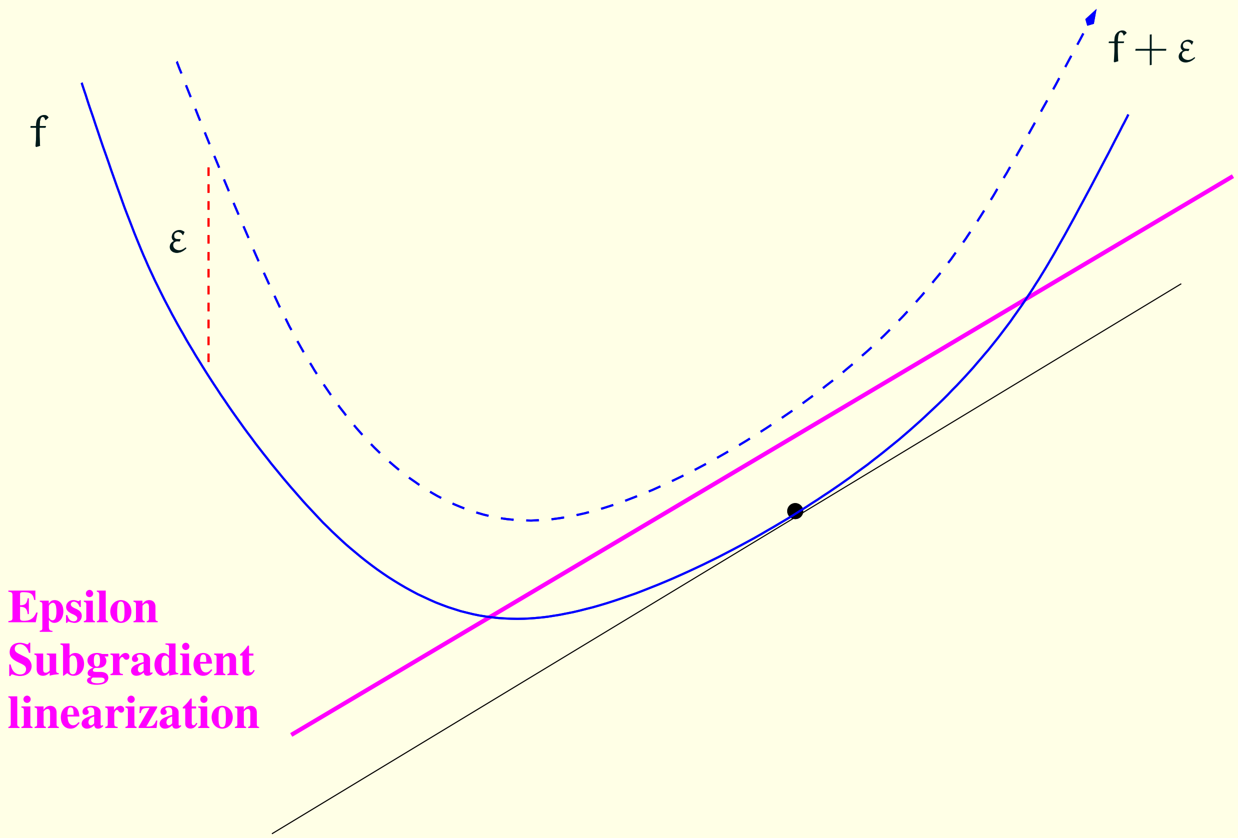
The ε-subdifferential

$$\partial_\varepsilon f(x) = \{g \in \mathbb{R}^n : f(y) \geq f(x) + \langle g, y - x\rangle - \varepsilon \text{ for all } y\}$$



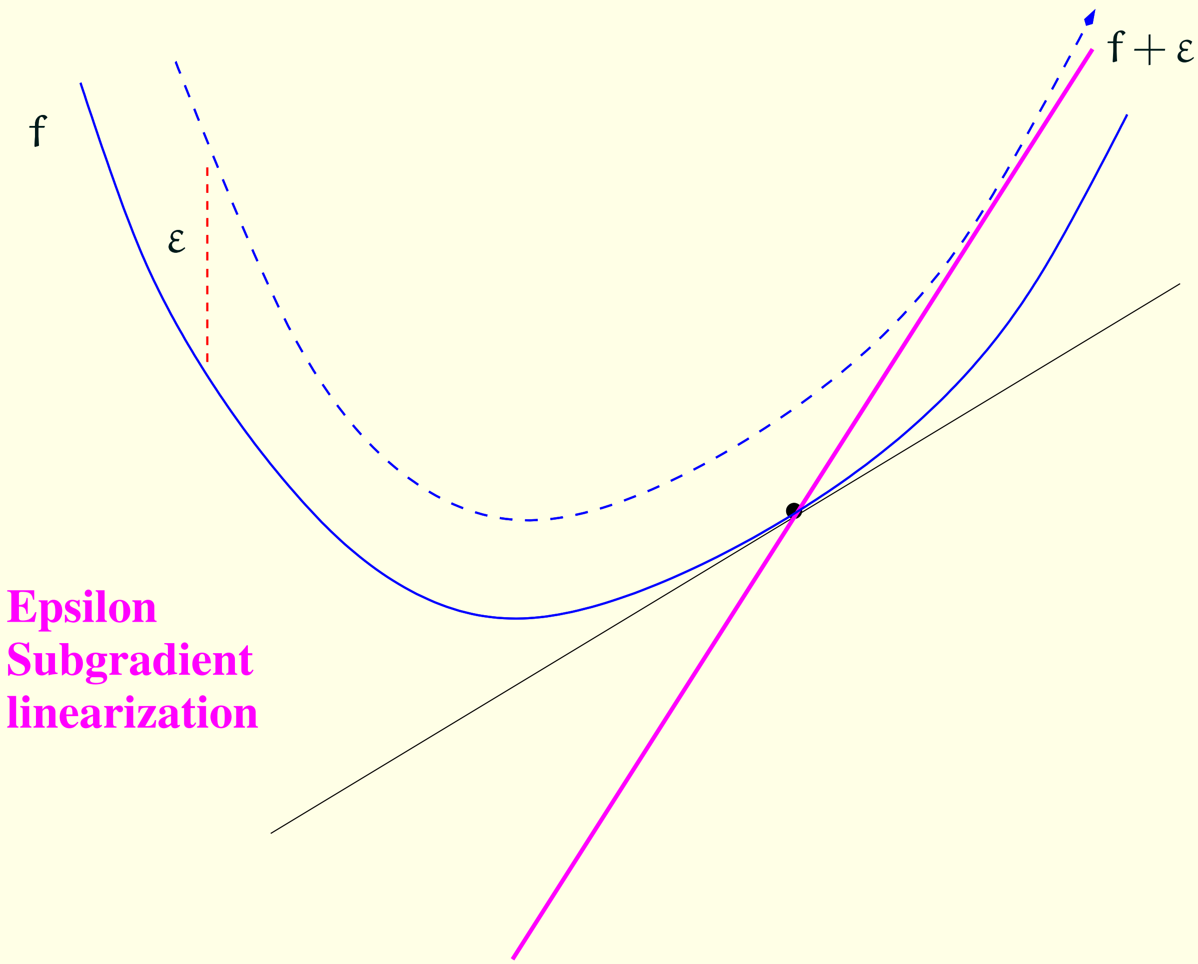
The ε-subdifferential

$$\partial_\varepsilon f(x) = \{g \in \mathbb{R}^n : f(y) \geq f(x) + \langle g, y - x \rangle \boxed{-\varepsilon} \text{ for all } y\}$$



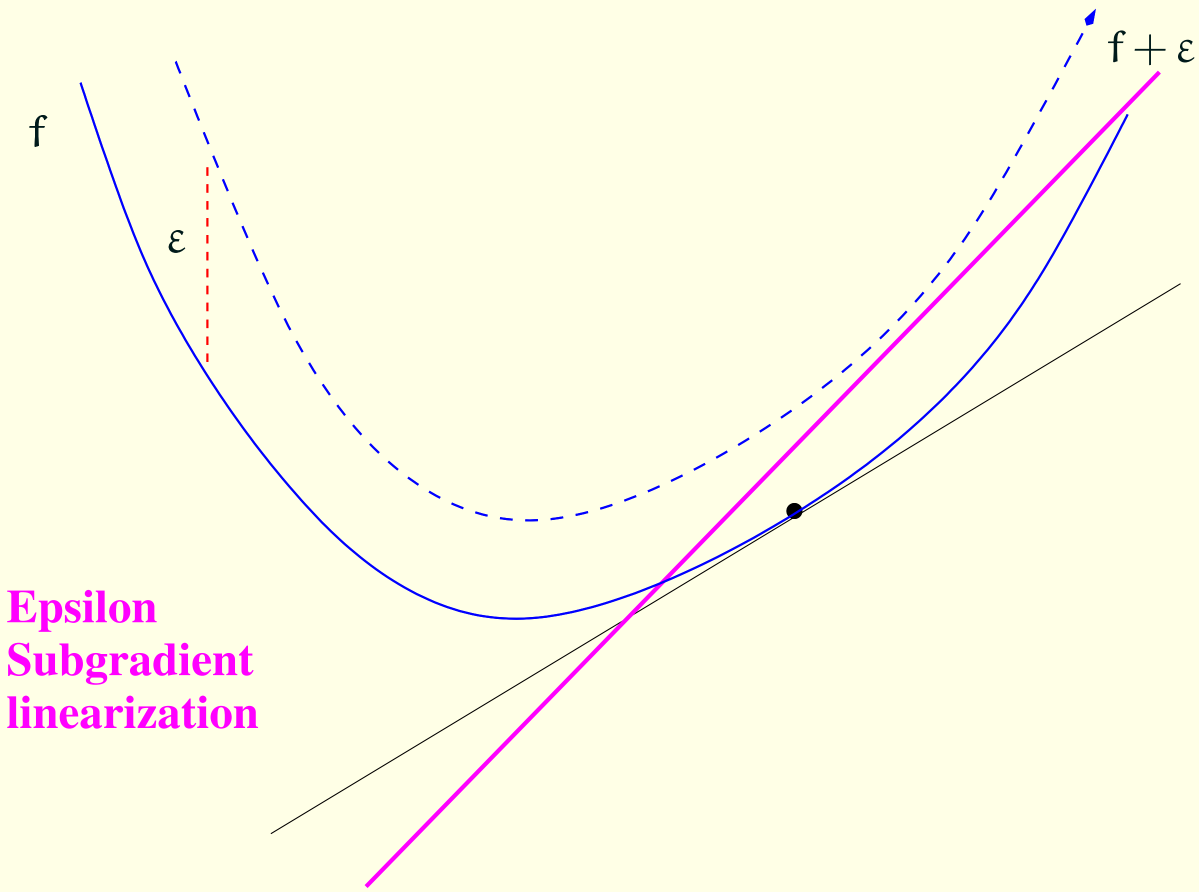
The ε-subdifferential

$$\partial_\varepsilon f(x) = \{g \in \mathbb{R}^n : f(y) \geq f(x) + \langle g, y - x\rangle - \varepsilon \text{ for all } y\}$$



The ε-subdifferential

$$\partial_\varepsilon f(x) = \{g \in \mathbb{R}^n : f(y) \geq f(x) + \langle g, y - x \rangle - \varepsilon \text{ for all } y\}$$

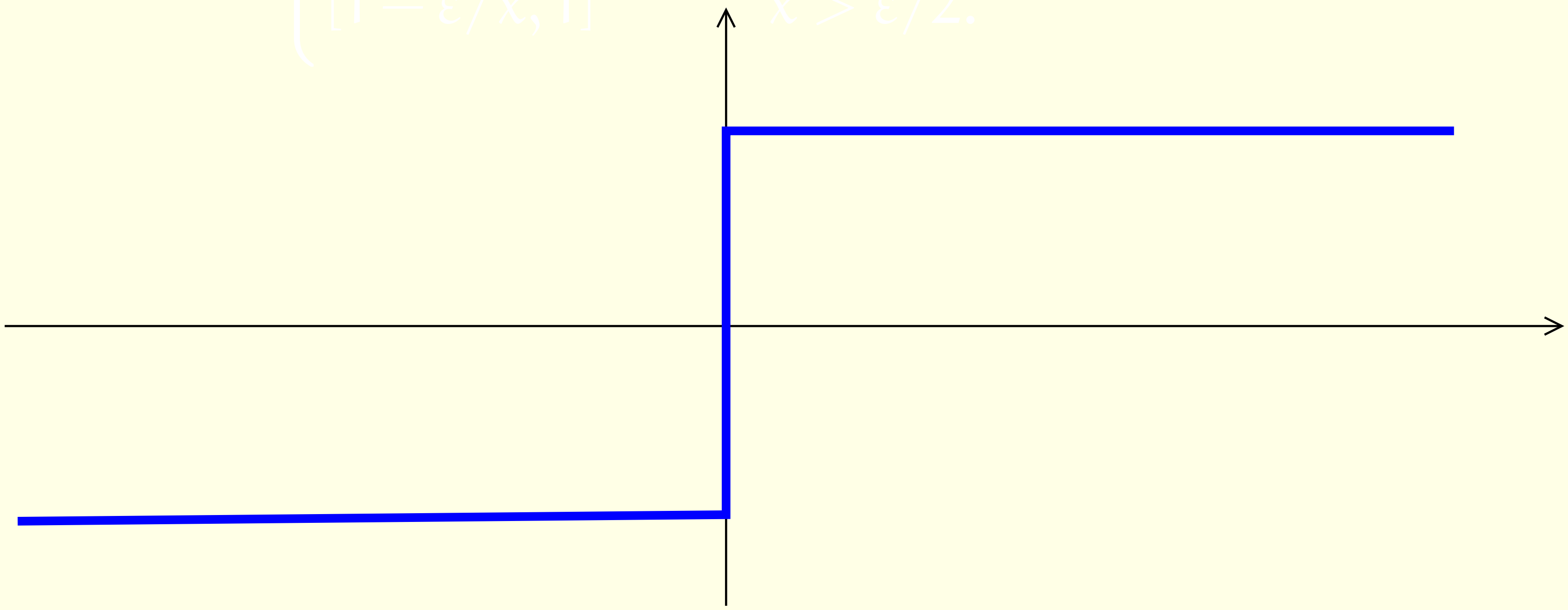


The ε-subdifferential

For the absolute value function, $f(x) = |x|$

$$\partial f(x) = \begin{cases} -1 & x < 0 \\ [-1, 1] & x = 0 \\ 1 & x > 0 \end{cases}$$

$$\partial_\varepsilon f(x) = \begin{cases} [-1, -1 - \varepsilon/x] & x < -\varepsilon/2, \\ [-1, 1] & -\varepsilon/2 \leq x1 \leq \varepsilon1/2, \\ [1 - \varepsilon/x, 1] & x > \varepsilon/2. \end{cases}$$

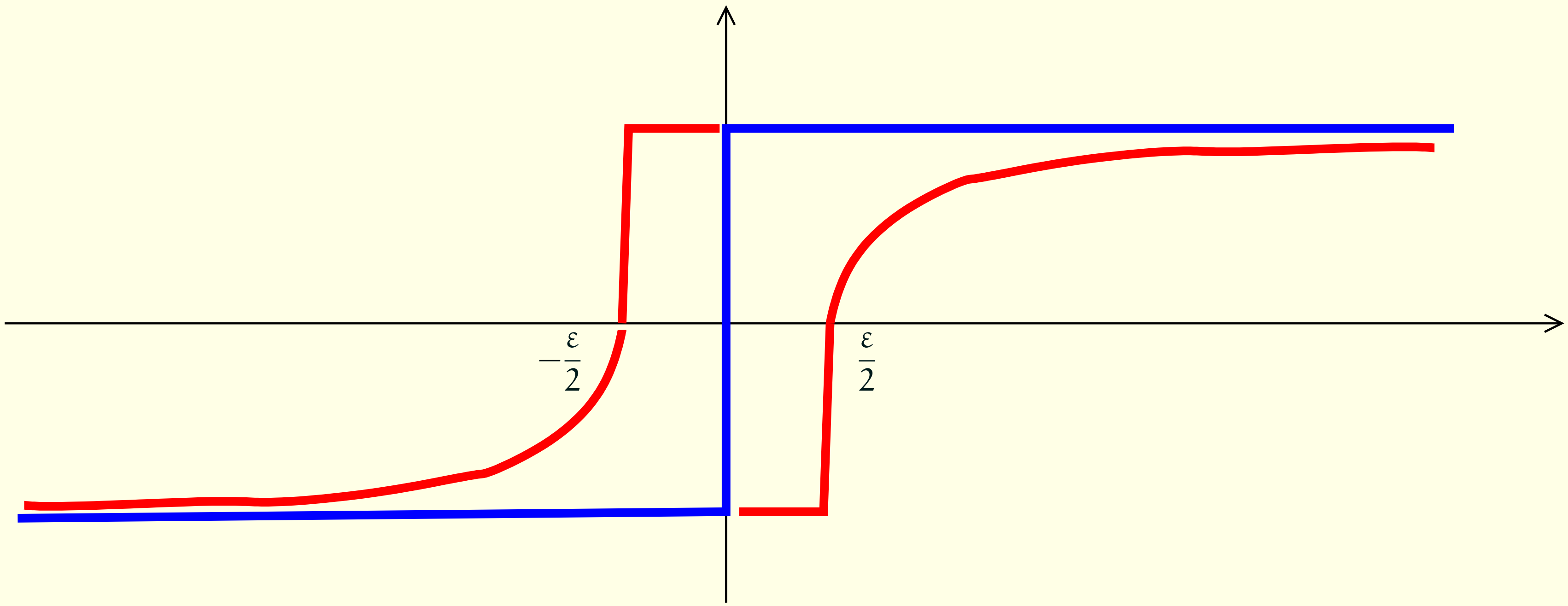


The ε-subdifferential

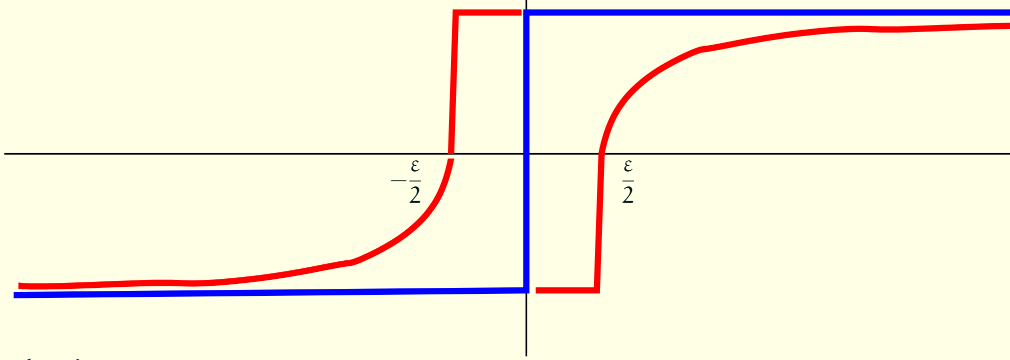
For the absolute value function, $f(x) = |x|$

$$\partial f(x) = \begin{cases} -1 & x < 0 \\ [-1,1] & x = 0 \\ 1 & x > 0 \end{cases}$$

$$\partial_\varepsilon f(x) = \begin{cases} [-1, -1-\varepsilon/x] & \text{if } x < -\varepsilon/2, \\ [-1,1] & \text{if } -\varepsilon/2 \leq x1 \leq \varepsilon1/2, \\ [1-\varepsilon/x, 1] & \text{if } x > \varepsilon/2. \end{cases}$$



The ε-subdifferential



- As a set-valued mapping $\partial_\varepsilon f(x)$ is osc:

$$\left(\varepsilon^k, x^k, G(x^k) \in \partial_{\varepsilon^k} f(x^k)\right) : \begin{cases} \varepsilon^k \longrightarrow \varepsilon \\ x^k \longrightarrow \bar{x} \\ G(x^k) \longrightarrow \bar{g} \end{cases} \Longrightarrow \bar{g} \in \partial_{\bar{\varepsilon}} f(\bar{x})$$

- As a set-valued mapping, $\partial_\varepsilon f(x)$ is isc: Given $\bar{g} \in \partial_{\bar{\varepsilon}} f(\bar{x})$

$$\exists \left(\varepsilon^k, x^k, G(x^k) \in \partial_{\varepsilon^k} f(x^k)\right) : \begin{cases} \varepsilon^k \longrightarrow \bar{\varepsilon} \\ x^k \longrightarrow \bar{x} \\ G(x^k) \longrightarrow \bar{g} \end{cases}$$

The ε-subdifferential and bundle methods

Generate iterates so that for a **subsequence** $\{\hat{x}^k\}$

- As a set-valued mapping $\partial_\varepsilon f(x)$ is osc:

$$\left(\varepsilon^k, \hat{x}^k, G(\hat{x}^k) \in \partial_{\varepsilon^k} f(\hat{x}^k)\right) : \begin{cases} \varepsilon^k \to \bar{\varepsilon} \\ x^k \to \bar{x} \\ G(\hat{x}^k) \to \bar{g} \end{cases} \Longrightarrow \bar{g} \in \partial_{\bar{\varepsilon}} f(\bar{x})$$

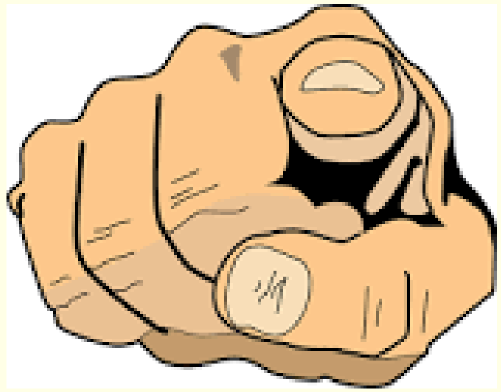
with $\bar{\varepsilon} = 0$ and $\bar{g} = 0$

- As a set-valued mapping, $\partial_\varepsilon f(x)$ is isc:Given $\bar{g} \in \partial_{\bar{\varepsilon}} f(\bar{x})$:

$$\exists\left(\varepsilon^k, \hat{x}^k, G(\hat{x}^k) \in \partial_{\varepsilon^k} f(\hat{x}^k)\right) : \begin{cases} \varepsilon^k \to \bar{\varepsilon} \\ x^k \to \bar{x} \\ G(x^k) \to \bar{g} \end{cases}$$

The ε-subdifferential and bundle methods

You told us



we were going to use subgradient information provided by an oracle or a black box, and now you want to use ε-subgradients!

The transportation formula

How to express subgradients at x^i as ε-subgradients at $\hat{x}^k$?

$$g^i \in \partial f(x^i) \quad \text{if and only if, for all } y \in \mathbb{R}^n$$

$$f(y) \;\geq\; f(x^i) + \left\langle g^i, y - x^i \right\rangle$$

The transportation formula

How to express subgradients at x^i as ε-subgradients at $\hat{x}^k$?

$$g^i \in \partial f(x^i) \quad \text{if and only if, for all } y \in \mathbb{R}^n$$

$$\begin{aligned}
f(y) &\geq f(x^i) + \left\langle g^i, y - x^i \right\rangle \\
&= f(x^i) + \left\langle g^i, y - x \right\rangle \pm f(\hat{x}^k) \\
&= f(\hat{x}^k) + g^{i\top}(y - x) - \left(f(\hat{x}^k) - f(x^i) \right) \\
&= f(\hat{x}^k) + g^{i\top}(y - x \pm \hat{x}^k) - \left(f(\hat{x}^k) - f(x^i) \right) \\
&= f(\hat{x}^k) + g^{i\top}(y - \hat{x}^k) - \left(f(\hat{x}^k) - f(x^i) - \left\langle g^i, \hat{x}^k - x^i \right\rangle \right) \\
&= f(\hat{x}^k) + g^{i\top}(y - \hat{x}^k) - e^i(\hat{x}^k)
\end{aligned}$$

$$\Longrightarrow g^i \in \partial_{e^i(\hat{x}^k)} f(\hat{x}^k)$$

$$e^i(\hat{x}^k) := f(\hat{x}^k) - f(x^i) - \left\langle g^i, \hat{x}^k - x^i \right\rangle \geq 0$$

The transportation formula

How to express subgradients at x^i as ε-subgradients at $\hat{x}^k$?

$$g^i \in \partial f(x^i) \quad \text{if and only if, for all } y \in \mathbb{R}^n$$

$$\begin{aligned}
f(y) &\geq f(x^i) + \left\langle g^i, y - x^i \right\rangle \\
&= f(x^i) + \left\langle g^i, y - x \right\rangle \pm f(\hat{x}^k) \\
&= f(\hat{x}^k) + \left\langle g^i, y - x \right\rangle - \left(f(\hat{x}^k) - f(x^i) \right) \\
&= f(\hat{x}^k) + \left\langle g^i y - x \pm \hat{x}^k \right\rangle - \left(f(\hat{x}^k) - f(x^i) \right) \\
&= f(\hat{x}^k) + \left\langle g^i y - \hat{x}^k \right\rangle - \left(f(\hat{x}^k) - f(x^i) - \left\langle g^i, \hat{x}^k - x^i \right\rangle \right) \\
&= f(\hat{x}^k) + \left\langle g^i, y - \hat{x}^k \right\rangle - e^i(\hat{x}^k)
\end{aligned}$$

$$\Longrightarrow g^i \in \partial_{e^i(\hat{x}^k)} f(\hat{x}^k)$$

$$e^i(\hat{x}^k) := f(\hat{x}^k) - f(x^i) - g^i, \hat{x}^k - x^i) \geq 0$$

The transportation formula

How to express subgradients at x^i as ε-subgradients at $\hat{x}^k$?

$$g^i \in \partial f(x^i) \quad \text{if and only if, for all } y \in \mathbb{R}^n$$

$$\begin{aligned}
f(y) &\geq f(x^i) + \left\langle g^i, y - x^i \right\rangle \\
&= f(x^i) + \left\langle g^i, y - x \right\rangle \pm f(\hat{x}^k) \\
&= f(\hat{x}^k) + \left\langle g^i, y - x \right\rangle - \Big(f(\hat{x}^k) - f(x^i) \Big) \\
&= f(\hat{x}^k) + \left\langle g^i, y - x \pm \hat{x}^k \right\rangle - \Big(f(\hat{x}^k) - f(x^i) \Big) \\
&= f(\hat{x}^k) + \left\langle g^i, y - \hat{x}^k \right\rangle - \Big(f(\hat{x}^k) - f(x^i) - \left\langle g^i, \hat{x}^k - x^i \right\rangle \Big) \\
&= f(\hat{x}^k) + \left\langle g^i, y - \hat{x}^k \right\rangle - e^i(\hat{x}^k)
\end{aligned}$$

$$\Longrightarrow g^i \in \partial_{e^i(\hat{x}^k)} f(\hat{x}^k)$$

$$e^i(\hat{x}^k) := f(\hat{x}^k) - f(x^i) - g^i, \hat{x}^k - x^i) \geq 0$$

The transportation formula

How to express subgradients at x^i as ε-subgradients at $\hat{x}^k$?

$$g^i \in \partial f(x^i) \quad \text{if and only if, for all } y \in \mathbb{R}^n$$

$$\begin{aligned}
f(y) &\geq f(x^i) + \left\langle g^i, y - x^i \right\rangle \\
&= f(x^i) + \left\langle g^i, y - x \right\rangle \pm f(\hat{x}^k) \\
&= f(\hat{x}^k) + \left\langle g^i, y - x \right\rangle - \left(f(\hat{x}^k) - f(x^i) \right) \\
&= f(\hat{x}^k) + \left\langle g^i, y - x \pm \hat{x}^k \right\rangle - \left(f(\hat{x}^k) - f(x^i) \right) \\
&= f(\hat{x}^k) + \left\langle g^i, y - \hat{x}^k \right\rangle - \left(f(\hat{x}^k) - f(x^i) - \left\langle g^i, \hat{x}^k - x^i \right\rangle \right) \\
&= f(\hat{x}^k) + \left\langle g^i, y - \hat{x}^k \right\rangle - e^i(\hat{x}^k)
\end{aligned}$$

$$\Longrightarrow g^i \in \partial_{e^i(\hat{x}^k)} f(\hat{x}^k)$$

$$e^i(\hat{x}^k) := f(\hat{x}^k) - f(x^i) - \langle g^i, \hat{x}^k - x^i \rangle \geq 0$$

The transportation formula

How to express subgradients at x^i as ε-subgradients at $\hat{x}^k$?

$$g^i \in \partial f(x^i) \quad \text{if and only if, for all } y \in \mathbb{R}^n$$

$$\begin{aligned}
f(y) &\geq f(x^i) + \left\langle g^i, y - x^i \right\rangle \\
&= f(x^i) + \left\langle g^i, y - x \right\rangle \pm f(\hat{x}^k) \\
&= f(\hat{x}^k) + \left\langle g^i, y - x \right\rangle - \left(f(\hat{x}^k) - f(x^i) \right) \\
&= f(\hat{x}^k) + \left\langle g^i, y - x \pm \hat{x}^k \right\rangle - \left(f(\hat{x}^k) - f(x^i) \right) \\
&= f(\hat{x}^k) + \left\langle g^i, y - \hat{x}^k \right\rangle - \left(f(\hat{x}^k) - f(x^i) - \left\langle g^i, \hat{x}^k - x^i \right\rangle \right) \\
&= f(\hat{x}^k) + \left\langle g^i, y - \hat{x}^k \right\rangle - e^i(\hat{x}^k)
\end{aligned}$$

$$\Longrightarrow g^i \in \partial_{e^i(\hat{x}^k)} f(\hat{x}^k)$$

$$e^i(\hat{x}^k) := f(\hat{x}^k) - f(x^i) - \langle g^i, \hat{x}^k - x^i \rangle \geq 0$$

The transportation formula

How to express subgradients at x^i as ε-subgradients at $\hat{x}^k$?

$$g^i \in \partial f(x^i) \quad \text{if and only if, for all } y \in \mathbb{R}^n$$

$$\begin{aligned}
f(y) &\geq f(x^i) + \left\langle g^i, y - x^i \right\rangle \\
&= f(x^i) + \left\langle g^i, y - x \right\rangle \pm f(\hat{x}^k) \\
&= f(\hat{x}^k) + \left\langle g^i, y - x \right\rangle - \left(f(\hat{x}^k) - f(x^i) \right) \\
&= f(\hat{x}^k) + \left\langle g^i, y - x \pm \hat{x}^k \right\rangle - \left(f(\hat{x}^k) - f(x^i) \right) \\
&= f(\hat{x}^k) + \left\langle g^i, y - \hat{x}^k \right\rangle - \left(f(\hat{x}^k) - f(x^i) - \left\langle g^i, \hat{x}^k - x^i \right\rangle \right) \\
&= f(\hat{x}^k) + \left\langle g^i, y - \hat{x}^k \right\rangle - e^i(\hat{x}^k)
\end{aligned}$$

$$\Longrightarrow g^i \in \partial_{e^i(\hat{x}^k)} f(\hat{x}^k)$$

$$e^i(\hat{x}^k) := f(\hat{x}^k) - f(x^i) - \left\langle g^i, \hat{x}^k - x^i \right\rangle \geq 0$$

The transportation formula

How to express subgradients at x^i as ε-subgradients at $\hat{x}^k$?

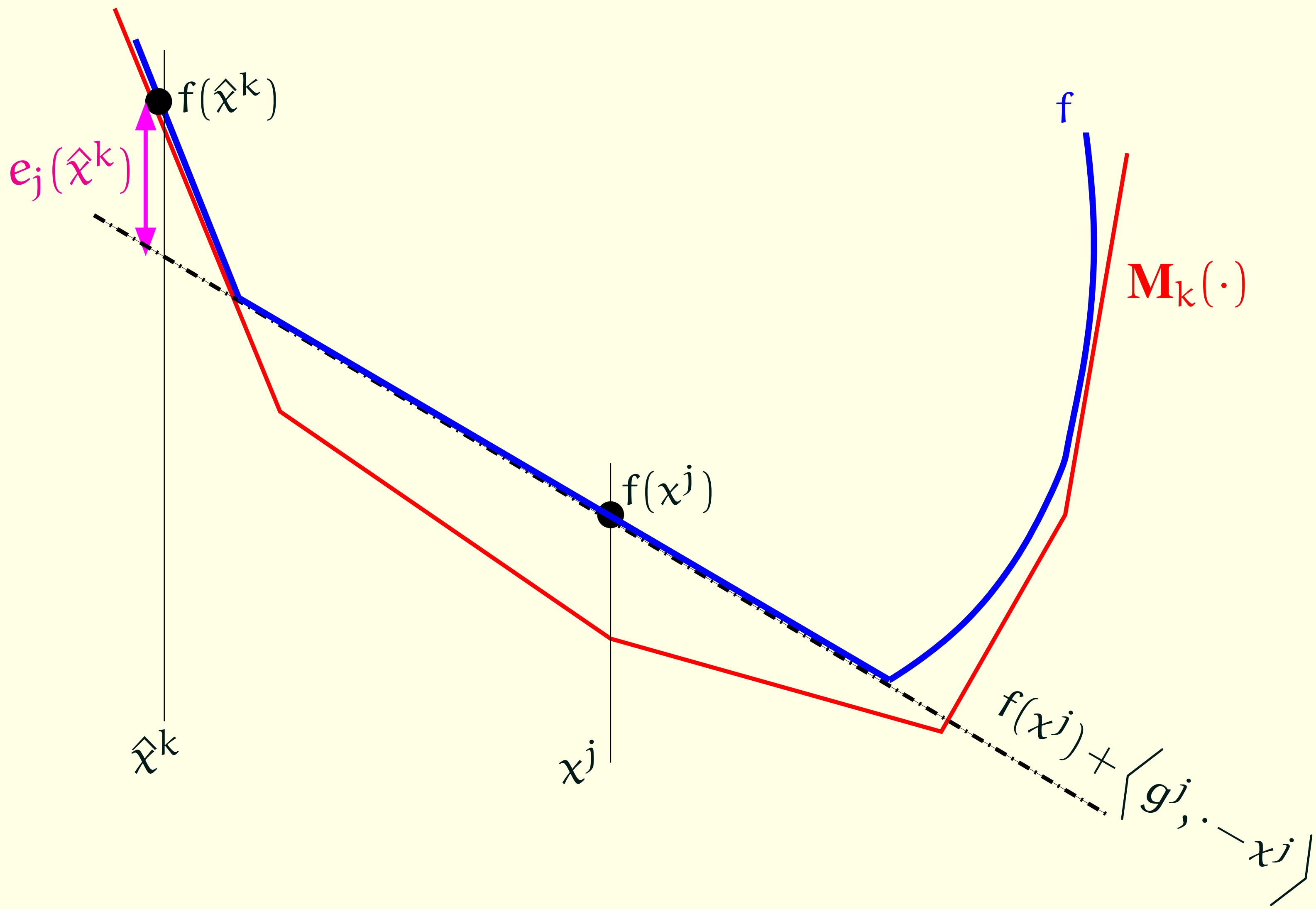
$$g^i \in \partial f(x^i) \quad \text{if and only if, for all } y \in \mathbb{R}^n$$

$$\begin{aligned}
f(y) &\geq f(x^i) + \left\langle g^i, y - x^i \right\rangle \\
&= f(x^i) + \left\langle g^i, y - x \right\rangle \pm f(\hat{x}^k) \\
&= f(\hat{x}^k) + \left\langle g^i, y - x \right\rangle - \left(f(\hat{x}^k) - f(x^i) \right) \\
&= f(\hat{x}^k) + \left\langle g^i, y - x \pm \hat{x}^k \right\rangle - \left(f(\hat{x}^k) - f(x^i) \right) \\
&= f(\hat{x}^k) + \left\langle g^i, y - \hat{x}^k \right\rangle - \left(f(\hat{x}^k) - f(x^i) - \left\langle g^i, \hat{x}^k - x^i \right\rangle \right) \\
&= f(\hat{x}^k) + \left\langle g^i, y - \hat{x}^k \right\rangle - e^i(\hat{x}^k)
\end{aligned}$$

$$\Longrightarrow g^i \in \partial_{e^i(\hat{x}^k)} f(\hat{x}^k)$$

$$e^i(\hat{x}^k) := f(\hat{x}^k) - f(x^i) - \left\langle g^i, \hat{x}^k - x^i \right\rangle \geq 0$$

Linearization errors



The ε-subdifferential and bundle methods

We collect the black-box

$x^i, i = 1, 2, \ldots, k$, so that at iteration k we can define a **bundle** of information, centered at a special iterate $\hat{x}^k \in \{x^i\}$

$$\mathcal{B}^k := \begin{pmatrix} e^i(\hat{x}^k) = f(\hat{x}^k) - f(x^i) - \left\langle g^i, \hat{x}^k - x^i \right\rangle \\ g^i \in \partial_{e^i(\hat{x}^k)} f(\hat{x}^k) \end{pmatrix}$$

The ε-subdifferential and bundle methods

We collect the black-box

$x^i, i = 1, 2, \ldots, k$, so that at iteration k we can define a **bundle** of information, centered at a special iterate $\hat{x}^k \in \{x^i\}$

$$\mathcal{B}^k := \begin{pmatrix} e^i(\hat{x}^k) = f(\hat{x}^k) - f(x^i) - \left\langle g^i, \hat{x}^k - x^i \right\rangle \\ g^i \in \partial_{e^i(\hat{x}^k)} f(\hat{x}^k) \end{pmatrix}$$

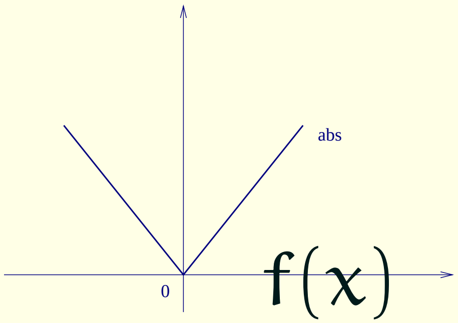
A suitable convex combination

$$\varepsilon^k := \sum_{i \in \mathcal{B}^k} \alpha^i e^i(\hat{x}^k) \text{ and } G^k := \sum_{i \in \mathcal{B}^k} \alpha^i g^i$$

will eventually satisfy the optimality condition!

Why special NSO methods?

Smooth optimization techniques **do not work**



$$\begin{aligned} f(x) &= |x| \\ |\nabla f(x^k)| &= 1,\ \forall x^k \neq 0 \quad \partial f(0) = [-1, 1] \end{aligned}$$

Smooth stopping test **fails**:

$|\nabla f(x^k)| \leq$TOL $\qquad (\leftrightarrow |g(x^k)| \leq$TOL$)$

Why special NSO methods?

Smooth optimization techniques **do not work**

Smooth approximations of derivatives by finite differences **fail**

For $f : \mathbb{R}^3 \to \mathbb{R}$ defined by $f(x) = \max(x_1, x_2, x_3)$

$\partial f(0) = ?$

Forward finite difference $\frac{f(x+\Delta x)-f(x)}{\Delta x}$

Central finite difference $\frac{f(x+\Delta x)-f(x-\Delta)}{2\Delta x}$

Why special NSO methods?

Smooth optimization techniques **do not work**

Smooth approximations of derivatives by finite differences **fail**

For $f : \mathbb{R}^3 \longrightarrow \mathbb{R}$ defined by $f(x) = \max(x_1, x_2, x_3)$

$\partial f(0) = ?$

Forward finite difference $\frac{f(x+\Delta x) - f(x)}{\Delta x} = (\mathbf{1}, \mathbf{1}, \mathbf{1})$

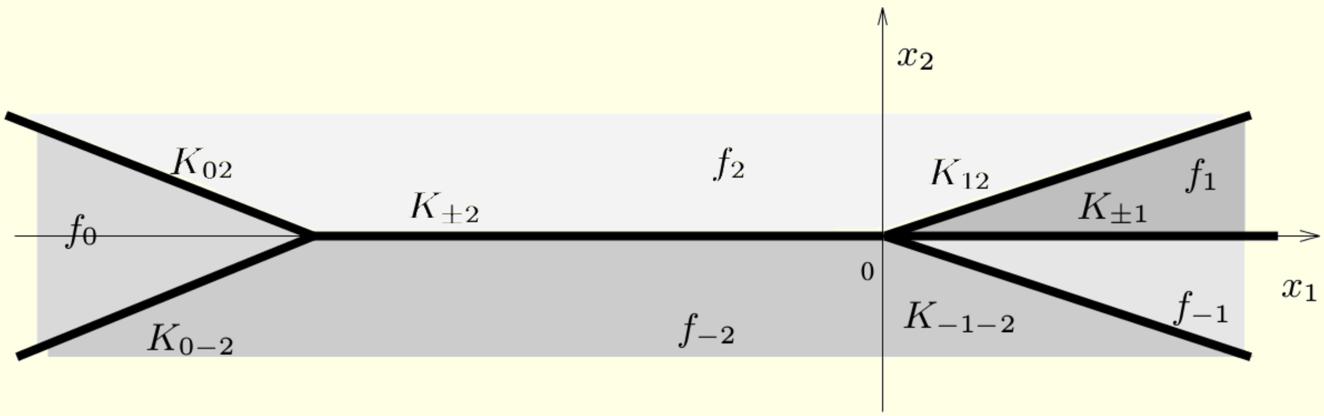
Central finite difference $\frac{f(x+\Delta x) - f(x-\Delta)}{2\Delta x} = (\frac{\mathbf{1}}{\mathbf{2}}, \frac{\mathbf{1}}{\mathbf{2}}, \frac{\mathbf{1}}{\mathbf{2}}$

none of them in the subdifferential!

Why special NSO methods?

Smooth optimization techniques **do not work**

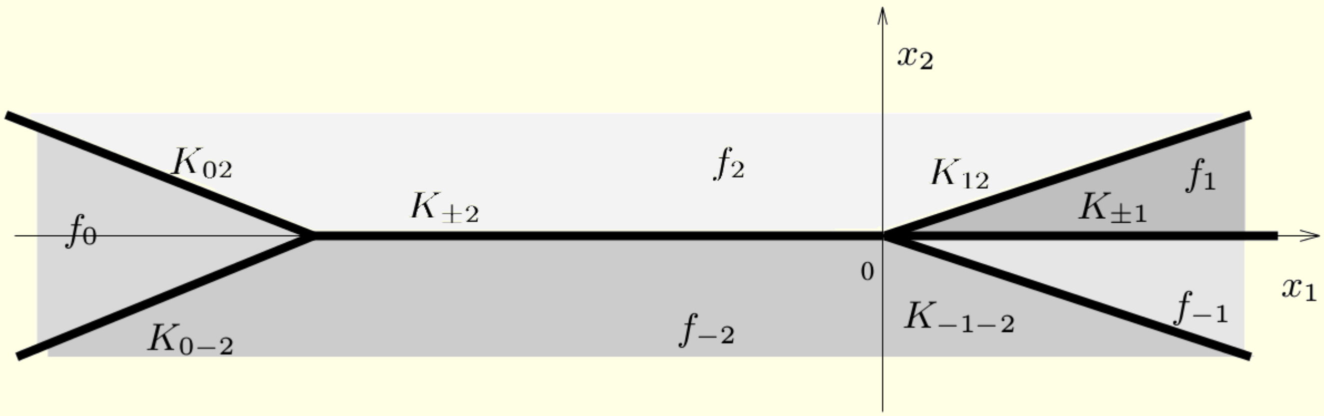
Linesearches get trapped in kinks and **fail**



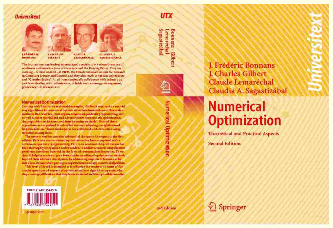
Why special NSO methods?

Smooth optimization techniques **do not work**

Linesearches get trapped in kinks and **fail**



Example 9.1 "Instability of steepest descent"



Why special NSO methods?

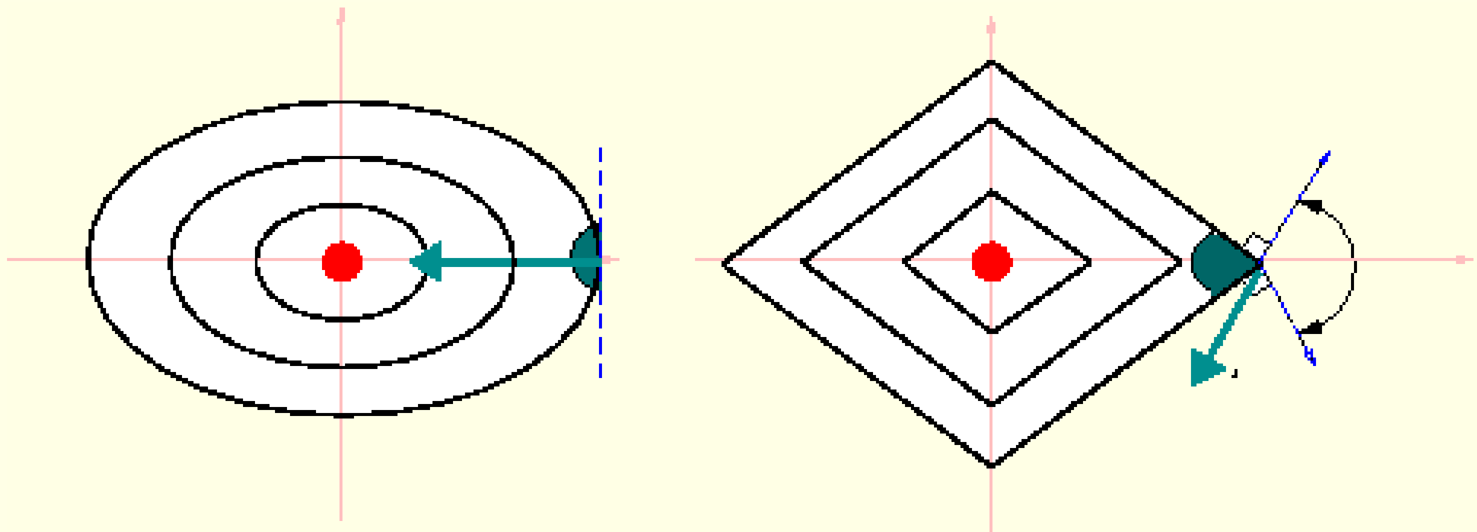
Smooth optimization techniques **do not work**

$-g(x^k)$ may **not** provide descent

Why special NSO methods?

Smooth optimization techniques **do not work**

$-g(x^k)$ may **not** provide descent



Why special NSO methods?

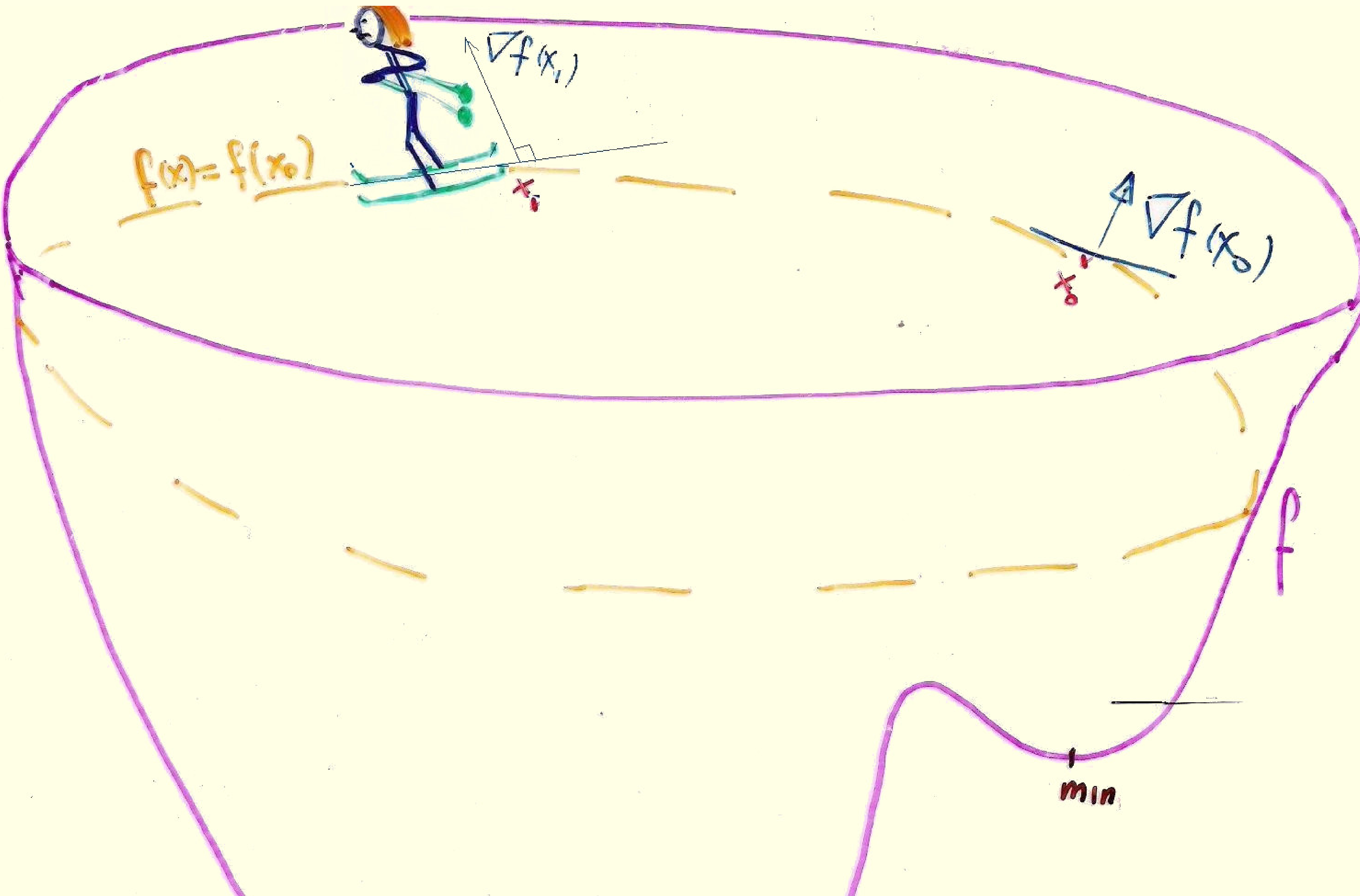
Smooth optimization techniques **do not work**

Smooth stopping test **fails**

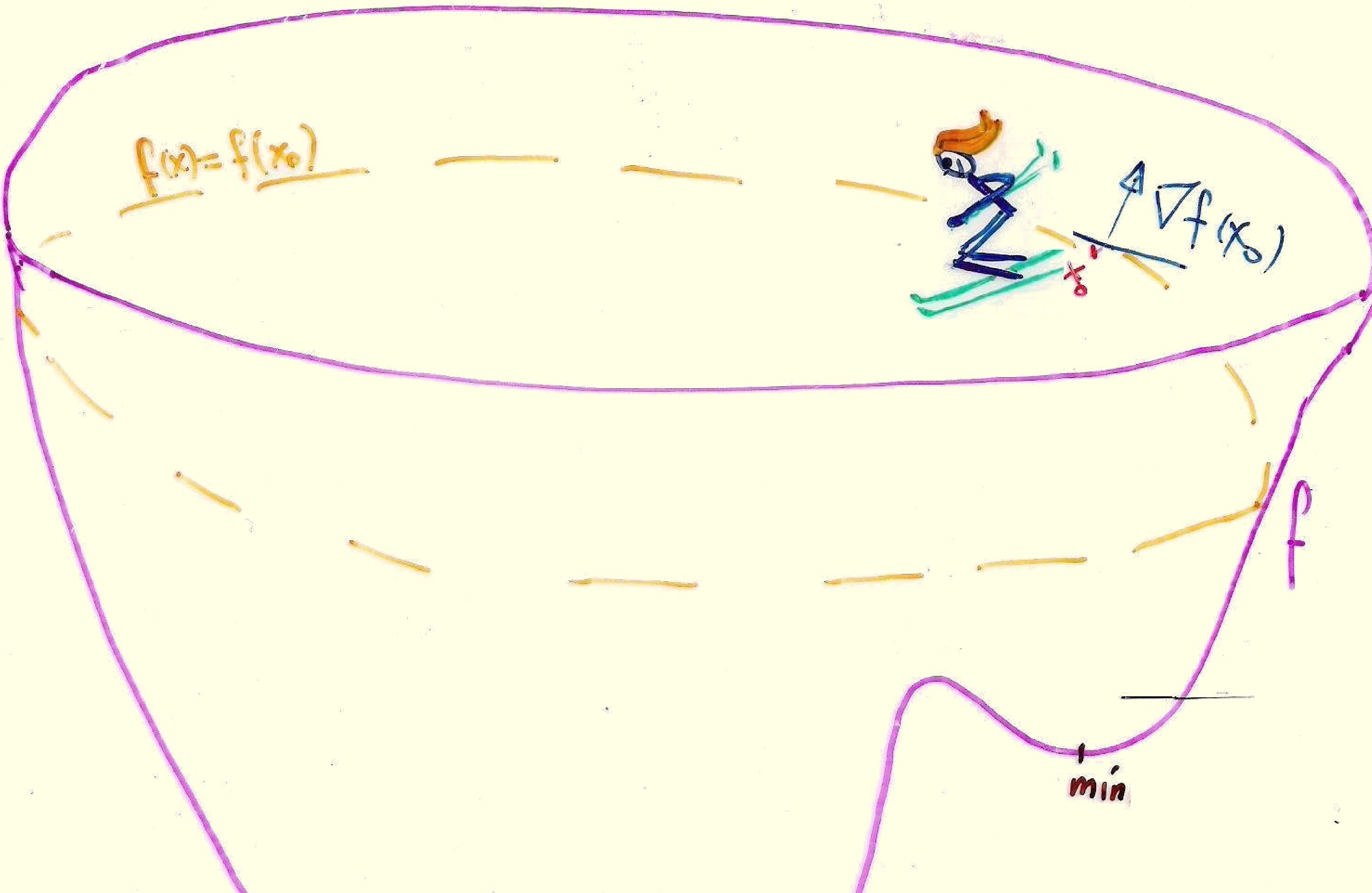
Finite difference approximations **fail**

Linesearches get trapped in kinks and **fail**

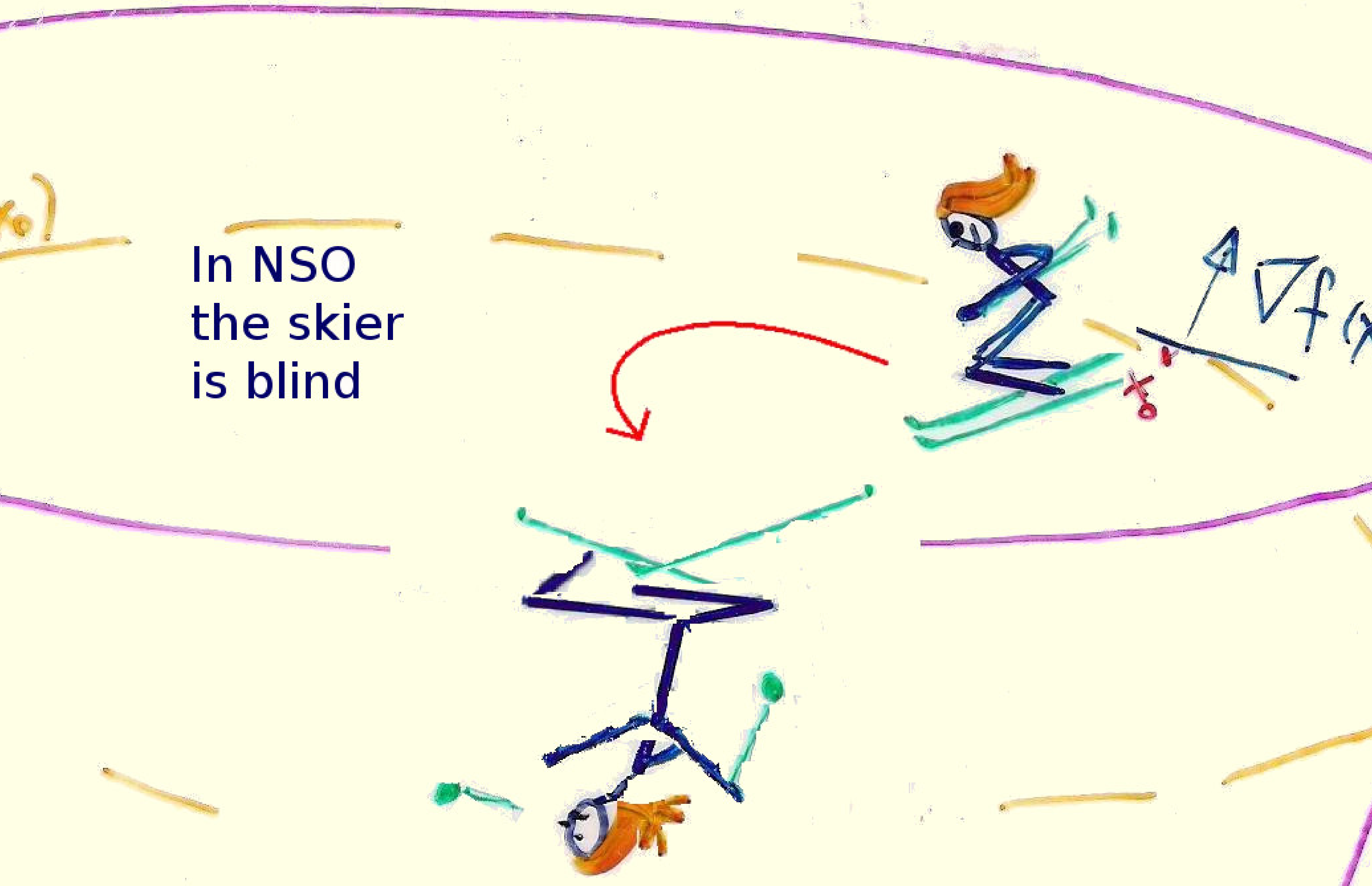
Direction opposite to a subgradient may **increase** the functional values



$\nabla f(x_1)$
$f(x) = f(x_0)$
x_1
$\nabla f(x_0)$
x_0
f
min



f(x) = f(x_0)
∇f(x_0)
x_0
f
min


In NSO
the skier
is blind
x_0

Bundle Methods

$$\text{WANT: } p = \text{prox}_t^f(x) \iff p = \arg\min f(y) + \frac{1}{2t}\|y - x\|_2^2$$

$$\text{HAVE: } q^k = \text{prox}_{t_k}^{M_k}(x) \iff q^k = \arg\min M_k(y) + \frac{1}{2t_k}\|y - x^k\|_2^2$$

$$\iff 0 = G^k + \frac{1}{t_k}(q^k - x^k)$$

$$\text{for } G^k \in \partial M_k(q^k)$$

$$\iff G^k \in \partial_{\varepsilon_k} f(x)$$

$$\text{for } \varepsilon_k = f(x) - M_k(q^k) - t_k\|G^k\|_2^2$$

Bundle Methods

$$\begin{aligned}
\text{WANT: } p = \operatorname{prox}_t^f(x) \quad &\Longleftrightarrow \quad p = \arg\min f(y) + \tfrac{1}{2t}\|y - x\|_2^2 \\
\text{HAVE: } q^k = \operatorname{prox}_{t_k}^{M_k}(x) \quad &\Longleftrightarrow \quad q^k = \arg\min M_k(y) + \tfrac{1}{2t_k}\|y - x^k\|_2^2 \\
&\Longleftrightarrow \quad 0 = G^k + \tfrac{1}{t_k}(q^k - x^k) \\
&\qquad\qquad \text{for } G^k \in \partial M_k(q^k) \\
&\Longleftrightarrow \quad G^k \in \partial_{\varepsilon_k} f(x) \\
&\qquad\qquad \text{for } \varepsilon_k = f(x) - M_k(q^k) - t_k\|G^k\|_2^2
\end{aligned}$$

Two subsequences

- Iterates giving sufficiently good approximal points
- Iterates just helping the optimization process

Bundle Methods

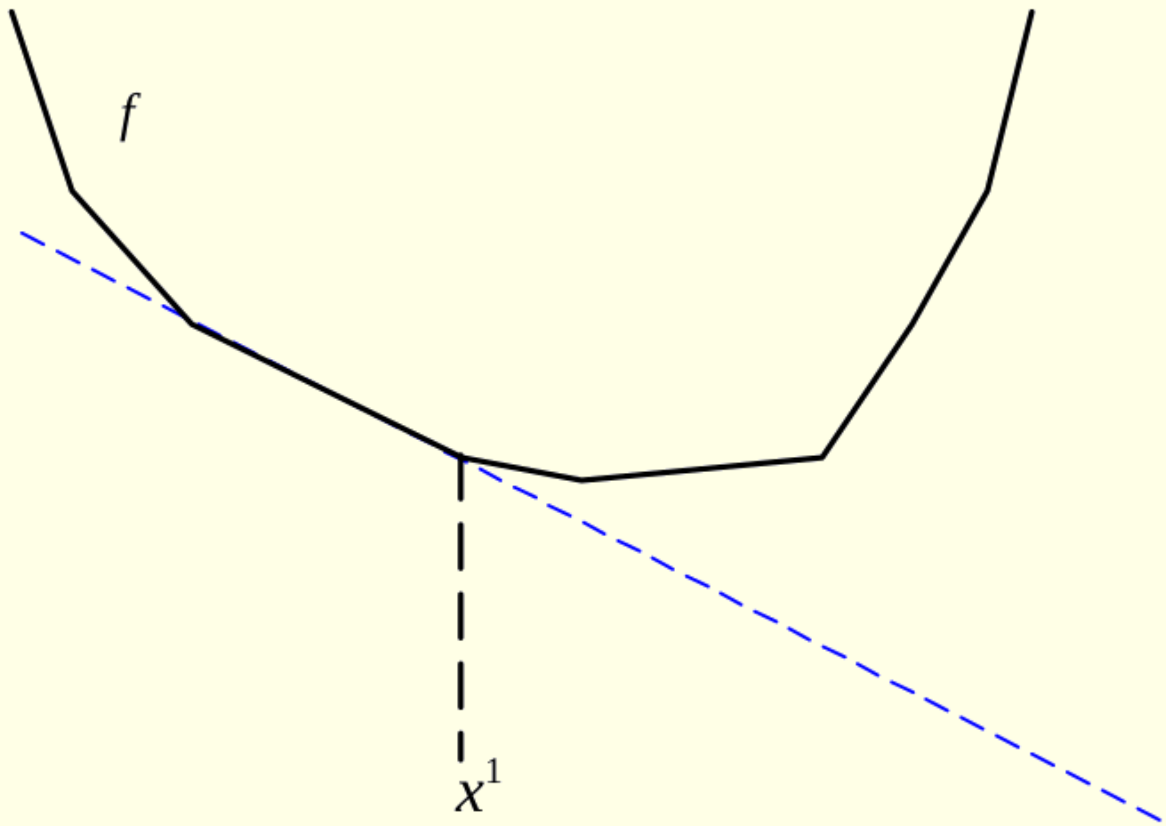
HAVE: $q^k = \mathrm{prox}_{t_k}^{M_k}(x) \;=\; x^k + t_k G^k \qquad\qquad G^k \in \partial_{\varepsilon_k} f(x)$

for $\varepsilon_k = f(x) - M_k(q^k) - t_k \|G^k\|_2^2$

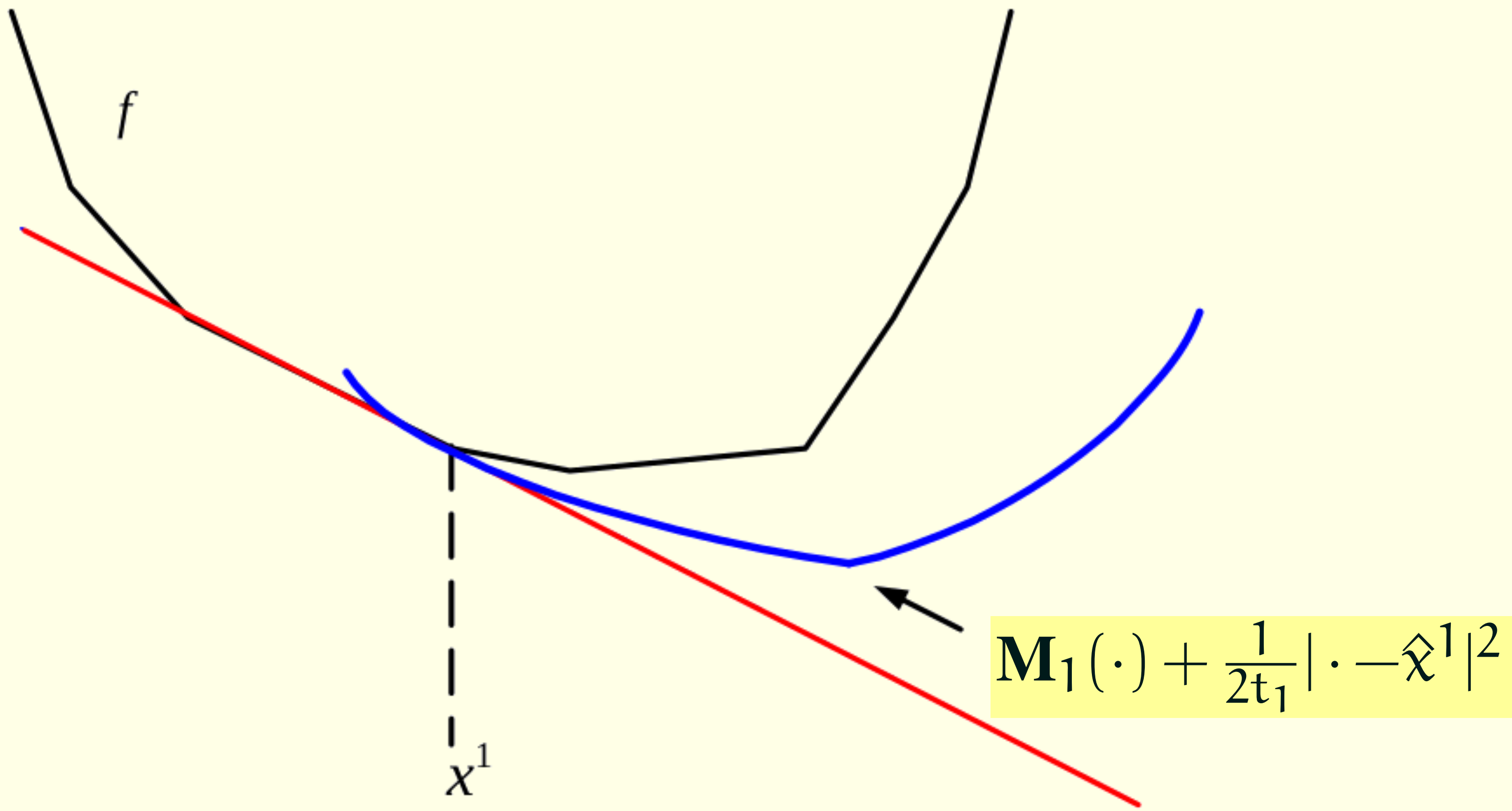
Two subsequences

- Iterates giving sufficiently good approximal points
  moving towards minimum
  in a manner that makes $\delta_k := \varepsilon_k + t_k \|G^k\|_2^2 \to 0$
  (serious)
- Iterates just helping the optimization process
  CL93 eventually applies **(null)**

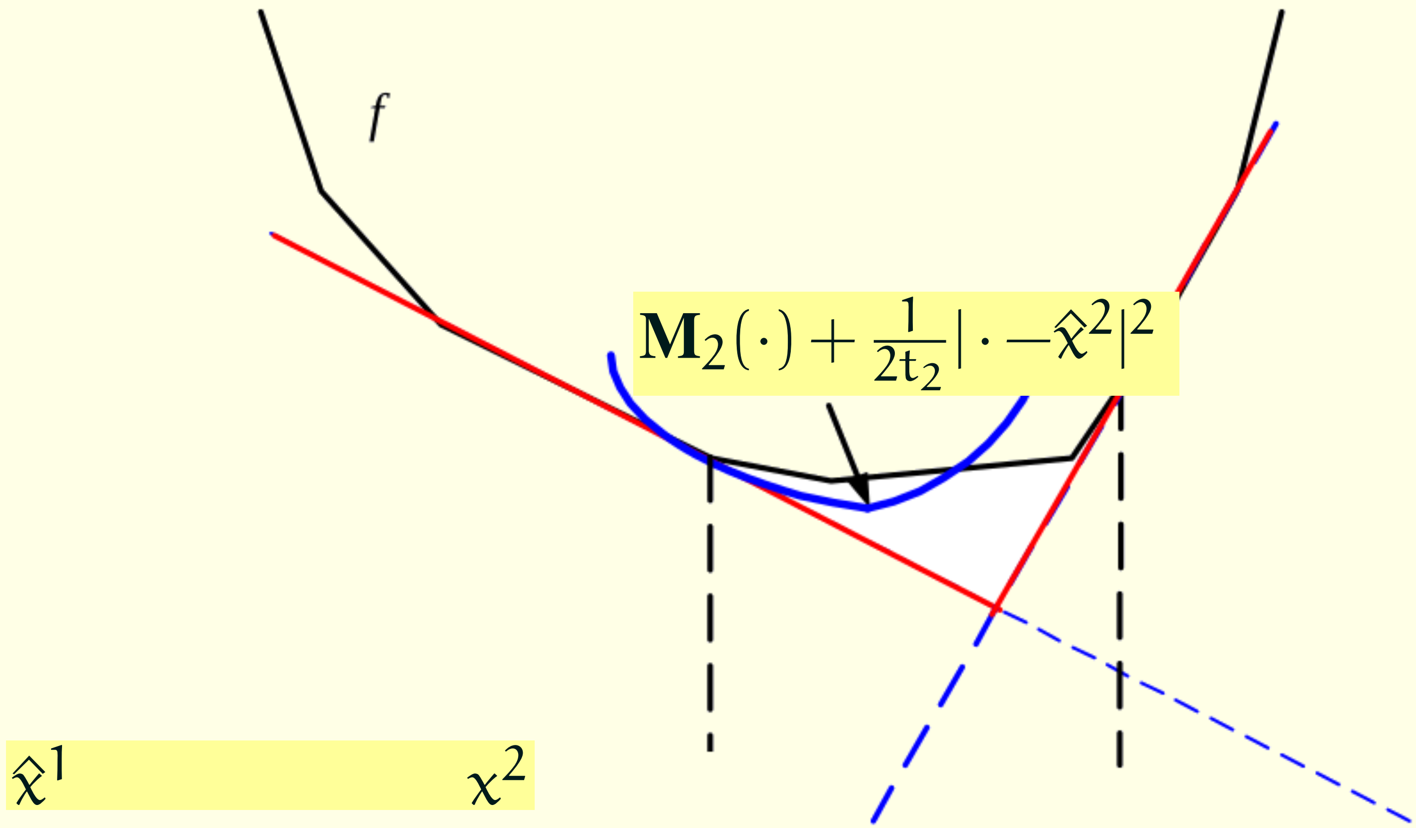
Bundle Methods



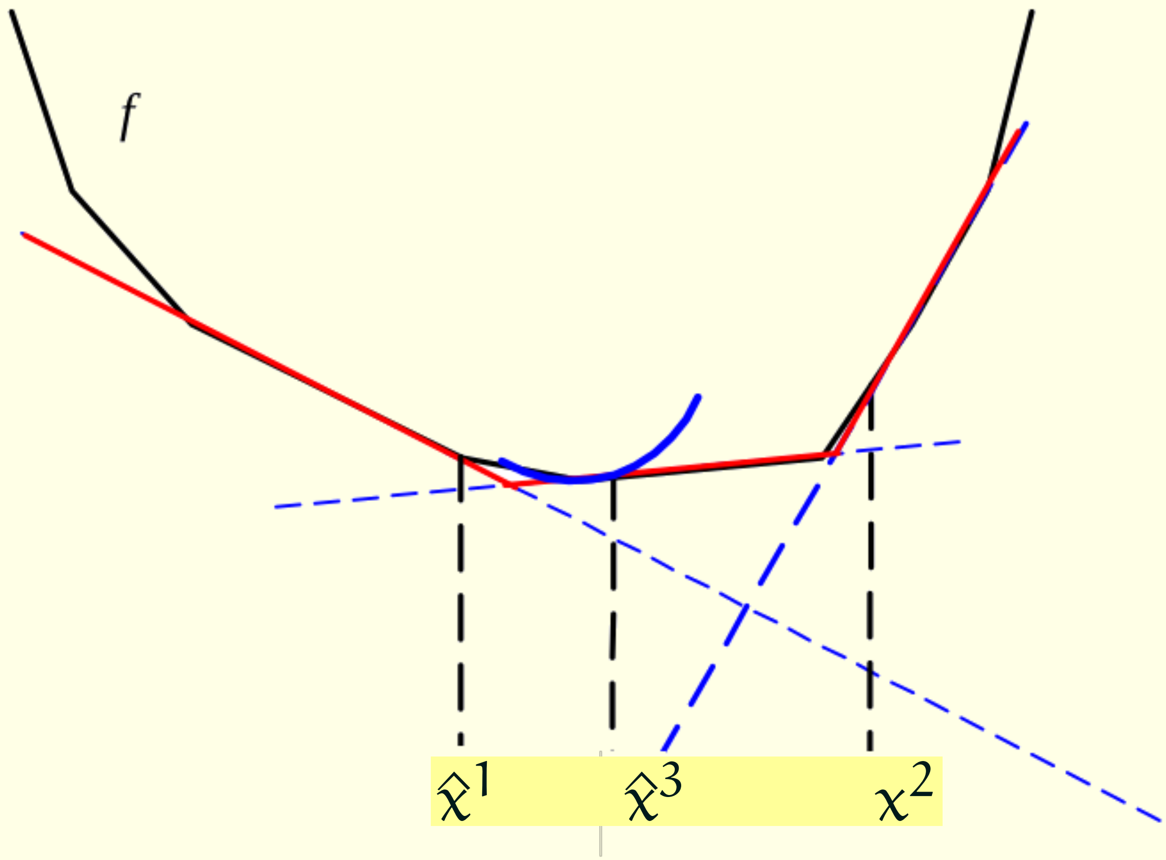
Bundle Methods



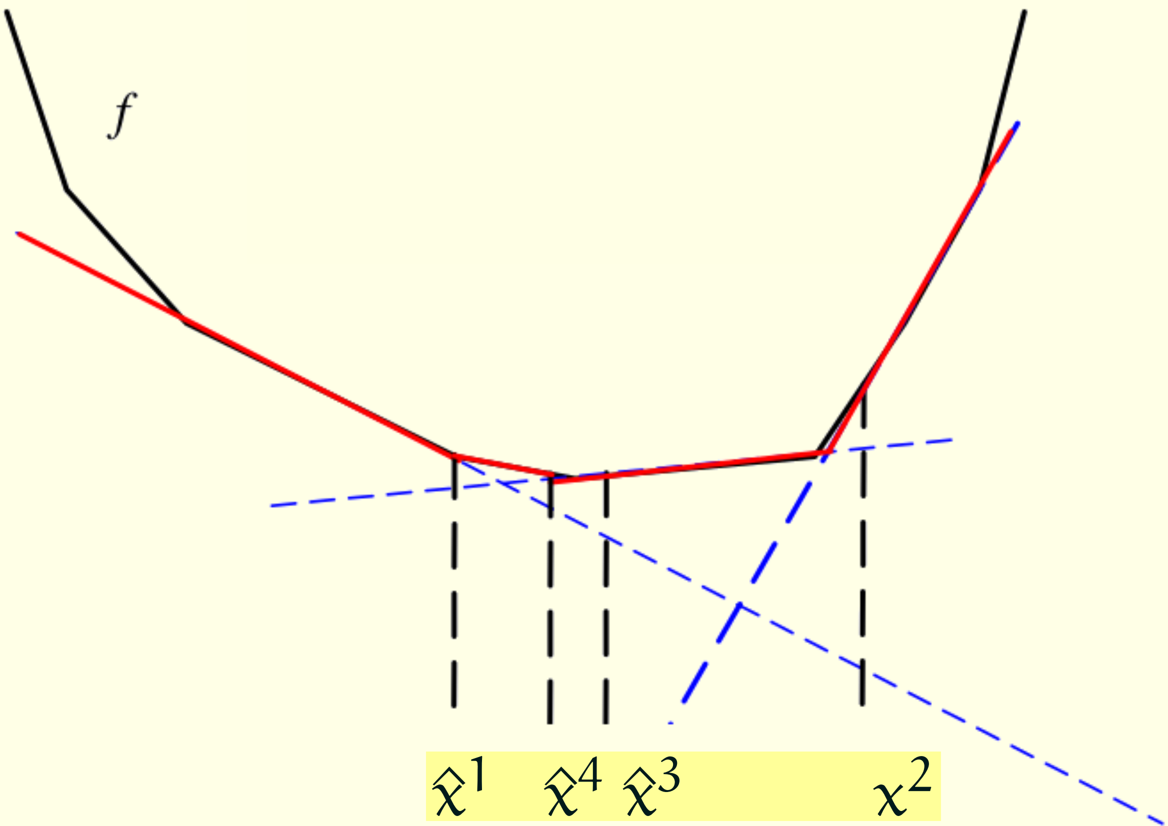
Bundle Methods



Bundle Methods



Bundle Methods



Bundle Methods

0 Choose x^1, set $k = 1$, and let $\hat{x}^1 = x^1$.

1 Compute $x^{k+1} = \arg\min \mathbf{M}_k(x) + \frac{1}{2t_k}|x - \hat{x}^k|^2$

2 If $\delta_{\mathbf{k}} := f(\hat{x}^k) - \mathbf{M}_k(x^{k+1}) \leq \mathtt{tol}$ STOP

3 Call the oracle at x^{k+1}. If

$f(x^{k+1}) \leq f(\hat{x}^k) - m\delta_k$, set $\hat{x}^{k+1} = x^{k+1}$ •

(Serious Step) Otherwise, maintain $\hat{x}^{k+1} = \hat{x}^k$
(Null Step)

4 Define $\mathbf{M}_{k+1}$, t_{k+1}, make $k = k + 1$, and loop to 1.

Bundle Methods: selection mechanism

$$\mathbf{M}_{k+1}(\cdot) = \max\left(\mathbf{M}_k(\cdot), f^k + \left\langle g^k, \cdot - x^k \right\rangle\right),$$

now the choice of the new model is more flexible:

$$x^{k+1} \in \arg\min \mathbf{M}_k(x) + \frac{1}{2t_k}|x - \hat{x}^k|^2$$

with $\mathbf{M}_k(x) = \max_{i \leq k}\{f^i + \left\langle g^i, x - x^i \right\rangle\}$ is equivalent to a QP:

$$\begin{cases} \min_{r \in \mathbb{R}, x \in \mathbb{R}^n} & r + \frac{1}{2t_k}|x - \hat{x}^k|^2 \\ \text{s.t.} & r \geq f^i + \left\langle g^i, x - x^i \right\rangle \text{ for } i \leq k \end{cases}$$

A posteriori, the solution remains the same if …

Bundle Methods: selection mechanism

$$\mathbf{M}_{k+1}(\cdot) = \max\left(\mathbf{M}_k(\cdot), f^k + \left\langle g^k, \cdot - x^k \right\rangle\right),$$

now the choice of the new model is more flexible:

$$x^{k+1} \in \arg\min \mathbf{M}_k(x) + \frac{1}{2t_k}|x - \hat{x}^k|^2$$

with $\mathbf{M}_k(x) = \max_{i \leq k}\{f^i + \left\langle g^i, x - x^i \right\rangle\}$ is equivalent to a QP:

$$\begin{cases} \min_{r \in \mathbb{R}, x \in \mathbb{R}^n} & r + \frac{1}{2t_k}|x - \hat{x}^k|^2 \\ \text{s.t.} & r \geq f^i + \left\langle g^i, x - x^i \right\rangle \text{ for } \mathbf{i \leq k} \end{cases}$$

A posteriori, the solution remains the same if all, or

...

Bundle Methods: selection mechanism

$$\mathbf{M}_{k+1}(\cdot) = \max\left(\mathbf{M}_k(\cdot), f^k + \left\langle g^k, \cdot - x^k \right\rangle\right),$$

now the choice of the new model is more flexible:

$$x^{k+1} \in \arg\min \mathbf{M}_k(x) + \frac{1}{2t_k}|x - \hat{x}^k|^2$$

with $\mathbf{M}_k(x) = \max_{i \leq k}\{f^i + \left\langle g^i, x - x^i \right\rangle\}$ is equivalent to a QP:

$$\begin{cases} \min_{r \in \mathbb{R}, x \in \mathbb{R}^n} & r + \frac{1}{2t_k}|x - \hat{x}^k|^2 \\ \text{s.t.} & r \geq f^i + \left\langle g^i, x - x^i \right\rangle \text{ for } \textbf{active i's} \end{cases}$$

A posteriori, the solution remains the same if all, or active, or . . .

Bundle Methods: selection mechanism

$$\mathbf{M}_{k+1}(\cdot) = \max\left(\mathbf{M}_k(\cdot), f^k + \left\langle g^k, \cdot - x^k \right\rangle\right),$$

now the choice of the new model is more flexible:

$$x^{k+1} \in \arg\min \mathbf{M}_k(x) + \frac{1}{2t_k}|x - \hat{x}^k|^2$$

with $\mathbf{M}_k(x) = \max_{i \le k}\{f^i + \left\langle g^i, x - x^i \right\rangle\}$ is equivalent to a QP:

$$\begin{cases} \min_{r \in \mathbb{R}, x \in \mathbb{R}^n} & r + \frac{1}{2t_k}|x - \hat{x}^k|^2 \\ \text{s.t.} & r \ge \sum_{\mathbf{i}} \bar{\alpha}^{\mathbf{i}}\left(f^i + \left\langle g^i, x - x^i \right\rangle\right) \end{cases}$$

A posteriori, the solution remains the same if all, or active, or the **optimal convex combination**

Bundle Methods: selection mechanism

$$\mathbf{M}_{k+1}(\cdot) = \max\left(\mathbf{M}_k(\cdot), f^k + \left\langle g^k, \cdot - x^k \right\rangle\right),$$

now the choice of the new model is more flexible:

$$x^{k+1} \in \arg\min \mathbf{M}_k(x) + \frac{1}{2t_k}|x - \hat{x}^k|^2$$

with $\mathbf{M}_k(x) = \max_{i \leq k}\{f^i + \left\langle g^i, x - x^i \right\rangle\}$ is equivalent to a QP:

$$\begin{cases} \min_{r \in \mathbb{R}, x \in \mathbb{R}^n} & r + \frac{1}{2t_k}|x - \hat{x}^k|^2 \\ \text{s.t.} & r \geq \sum_i \bar{\alpha}^i\left(f^i + \left\langle g^i, x - x^i \right\rangle\right) \end{cases}$$

A posteriori, the solution remains the same if all, or active, or the optimal convex combination are kept

Bundle Methods: next model options

$$\mathbf{M}_{k+1}(\cdot) = \max\left(\mathbf{M}_k(\cdot), f^k + \left\langle g^k, \cdot - x^k \right\rangle\right)$$

or

$$\mathbf{M}_{k+1}(\cdot) = \max\left(\max_{active}, f^k + \left\langle g^k, \cdot - x^k \right\rangle\right)$$

or

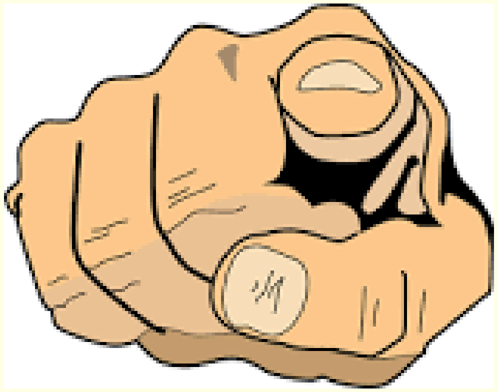
$$\mathbf{M}_{k+1}(\cdot) = \max\left(aggregate, f^k + \left\langle g^k, \cdot - x^k \right\rangle\right)$$

Same QP solution if all, or active, or the optimal convex combination

aggregate=full Bundle Compression: QP with only 2 constraints (but slows down the overall process)

The cutting-plane model

You told us

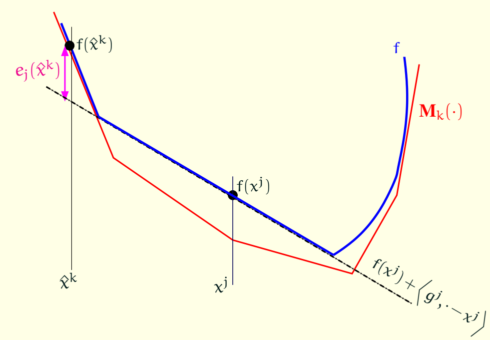


we were going to use a bundle $\mathcal{B}_k$ composed by linearization errors and ε-subgradients at $\hat{x}^k$, but the model uses f^i and $g^i \in \partial f(x^i)$

Rewriting the cutting-plane model

The transportation formula centers the ith linearization in the serious iterate

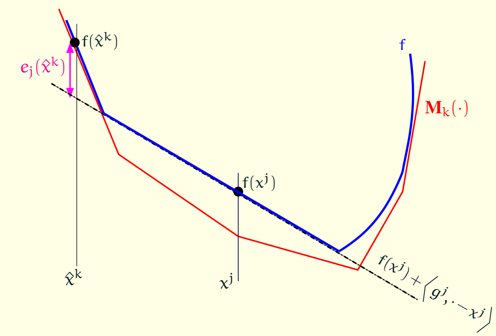
$$\begin{aligned} f(y) &\geq f(x^i) + \left\langle g^i, y - x^i \right\rangle \\ &= f(\hat{x}^k) + \left\langle g^i, y - \hat{x}^k \right\rangle - e^i(\hat{x}^k) \end{aligned}$$



Rewriting the cutting-plane model

The transportation formula centers the ith linearization in the serious iterate

$$\begin{aligned} f(y) &\geq f(x^i) + \left\langle g^i, y - x^i \right\rangle \\ &= f(\hat{x}^k) + \left\langle g^i, y - \hat{x}^k \right\rangle - e^i(\hat{x}^k) \end{aligned}$$



this translates into the model as follows

$$\begin{aligned} \mathbf{M}_k(y) &= \max\left\{ f(x^i) + \left\langle g^i, y - x^i \right\rangle \quad : i \in \mathcal{B}_k \right\} \\ &= \max\left\{ f(\hat{x}^k) + \left\langle g^i, y - \hat{x}^k \right\rangle - e^i(\hat{x}^k) \quad : i \in \mathcal{B}_k \right\} \\ &= f(\hat{x}^k) + \max\left\{ \left\langle g^i, y - \hat{x}^k \right\rangle - e^i(\hat{x}^k) \quad : i \in \mathcal{B}_k \right\} \end{aligned}$$

Bundle Method

0 Choose x^1, set $k = 1$, and let $\hat{x}^1 = x^1$.

1 Compute $x^{k+1} = \arg\min \mathbf{M}_k(x) + \frac{1}{2t_k}|x - \hat{x}^k|^2$

2 If $\delta_k := f(\hat{x}^k) - \mathbf{M}_k(x^{k+1}) \leq \mathtt{tol}$ STOP

3 Call the oracle at x^{k+1}. If

$f(x^{k+1}) \leq f(\hat{x}^k) - m\delta_k$, set $\hat{x}^{k+1} = x^{k+1}$

Otherwise, maintain $\hat{x}^{k+1} = \hat{x}^k$

4 Define $\mathbf{M}_{k+1}$, t_{k+1}, make $k = k + 1$, and loop to 1.

Bundle Method

0 Choose x^1, set $k = 1$, and let $\hat{x}^1 = x^1$.

1 Compute $x^{k+1} = \arg\min \mathbf{M}_k(x) + \frac{1}{2t_k}|x - \hat{x}^k|^2$

2 If $\delta_k := f(\hat{x}^k) - \mathbf{M}_k(x^{k+1}) \leq \mathtt{tol}$ STOP

3 Call the oracle at x^{k+1}. If

$f(x^{k+1}) \leq f(\hat{x}^k) - m\delta_k$, set $\hat{x}^{k+1} = x^{k+1}$

(Serious Step) $\mathbf{k \in K_S}$

Otherwise, maintain $\hat{x}^{k+1} = \hat{x}^k$

(Null Step) $k \in K_N$

4 Define $\mathbf{M}_{k+1}$, t_{k+1}, make $k = k + 1$, and loop to 1.

Bundle Method

When $k \rightarrow \infty$, the algorithm generates two subsequences. Convergence analysis addresses the mutually exclusive situations

- either the SS subsequence is infinite

  $\mathbf{K}_\infty := \{\mathbf{k} \in \mathbf{K}_S\}$

- or there is a last SS, followed by infinitely many null steps

Bundle Method

When $k \to \infty$, the algorithm generates two subsequences. Convergence analysis addresses the mutually exclusive situations

- either the SS subsequence is infinite

  $$\mathbf{K}_\infty := \{\mathbf{k} \in \mathbf{K_S}\}$$

- or there is a last SS, followed by infinitely many null steps

  $$\mathbf{K}_\infty := \{\mathbf{k} \in \mathbf{K_N} : \mathbf{k} \geq \textbf{last SS}\}$$

Bundle Method

When $k \to \infty$, the algorithm generates two subsequences. Convergence analysis addresses the mutually exclusive situations

- either the SS subsequence is infinite

  $\mathbf{K}_\infty := \{\mathbf{k} \in \mathbf{K}_\mathbf{S}\}$ **(limit point minimizes f)**

- or there is a last SS, followed by infinitely many null steps

  $\mathbf{K}_\infty := \{\mathbf{k} \in \mathbf{K}_\mathbf{N} : \mathbf{k} \geq \textbf{last SS}\}$

  (last SS minimizes f and null$\to$ last SS)

Equivalent QPs

1. Given t_k, the stepsize parameter of the proximal bundle method, with QP subproblem given by

$$(PB)_k \quad \min \mathbf{M}_k(x) + \frac{1}{2t_k}|x - \hat{x}^k|^2$$

2. Given Δ_k, the radius parameter of the trust-region bundle method, with QP subproblem given by

$$(TRB)_k \quad \begin{cases} \min & \mathbf{M}_k(x) \\ \text{s.t.} & |x - \hat{x}^k|^2 \leq \Delta_k \end{cases}$$

3. Given ℓ_k, the level parameter of the level bundle method,

with QP subproblem given by

$$(\mathrm{LB})_k \quad \begin{cases} \min & \frac{1}{2}|x-\hat{x}^k|^2 \\ \text{s.t.} & \mathbf{M}_k(x) \le \ell_k \end{cases}$$

Show that

1. given t_k, there exists Δ_k such that if x^{k+1} solves $(\mathrm{PB})_k$, then x^{k+1} solves $(\mathrm{TRB})_k$.
2. given Δ_k, there exists ℓ_k such that if x^{k+1} solves $(\mathrm{TRB})_k$, then x^{k+1} solves $(\mathrm{LB})_k$.
3. given
   ell_k, there exists t_k such that if x^{k+1} solves $(\mathrm{LB})_k$, then x^{k+1} solves $(\mathrm{PB})_k$.

Theorem $\mathbf{K}_\infty := \{\mathbf{k} \in \mathbf{K}_S\}$

Suppose the bundle method loops forever and there are infinitely many serious steps. Either the solution set of $\min f$ is empty and $f(\hat{x}^k) \searrow -\infty$ or the following holds

(i) $\lim\limits_{k\in K_S} \delta_k = 0$ and $\lim\limits_{k\in K_S} \varepsilon_k = 0$.

(ii) If the stepsizes are chosen so that $\sum\limits_{k\in K_S} t_k = +\infty$ then $\{\hat{x}^k\}$ is a minimizing sequence.

(iii) If, in addition, $t_k \le t^{up}$ for all $k \in K_S$, then the subsequence $\{\hat{x}^k\}$ is bounded. In this case, any limit point x^∞ minimizes f and the whole sequence converges to x^∞

Theorem $\mathbf{K}_\infty := \left\{\mathbf{k} \in \mathbf{K_N} \geq \hat{\mathbf{k}}\right\}$

Suppose the bundle method loops forever and there are infinitely many null steps after a last serious one, denoted by $\hat{x}$ and generated at iteration $\hat{k}$. Suppose stepsizes are chosen so that

$$t_{lo} \leq t_{k+1} \leq t_k \quad \text{for all } k \in K_\infty$$

The following holds

1. The sequence $\{x^{k+1}\}$ is bounded
2. $\lim_{k \in K_\infty} \mathbf{M}_k(x^{k+1}) = f(\hat{x})$
3. $\hat{x}$ minimizes f
4. $\lim_{k \in K_\infty} x^{k+1} = \hat{x}$

Model requirements

1. $\mathbf{M}_k \leq f$

2. If k was declared a null step

    a) $\mathbf{M}_{k+1}(x) \geq f^{k+1} + \left\langle g^{k+1}, x - x^{k+1} \right\rangle$

    b) $\mathbf{M}_{k+1}(x) \geq A_k(x) = \mathbf{M}_k(x^{k+1}) + \left\langle G^k, x - x^{k+1} \right\rangle$

Model requirements

1. $\mathbf{M}_k \le f$
2. If k was declared a null step
   a) $\mathbf{M}_{k+1}(x) \ge f^{k+1} + \left\langle g^{k+1}, x - x^{k+1} \right\rangle$
   b) $\mathbf{M}_{k+1}(x) \ge A_k(x) = \mathbf{M}_k(x^{k+1}) + \left\langle G^k, x - x^{k+1} \right\rangle$

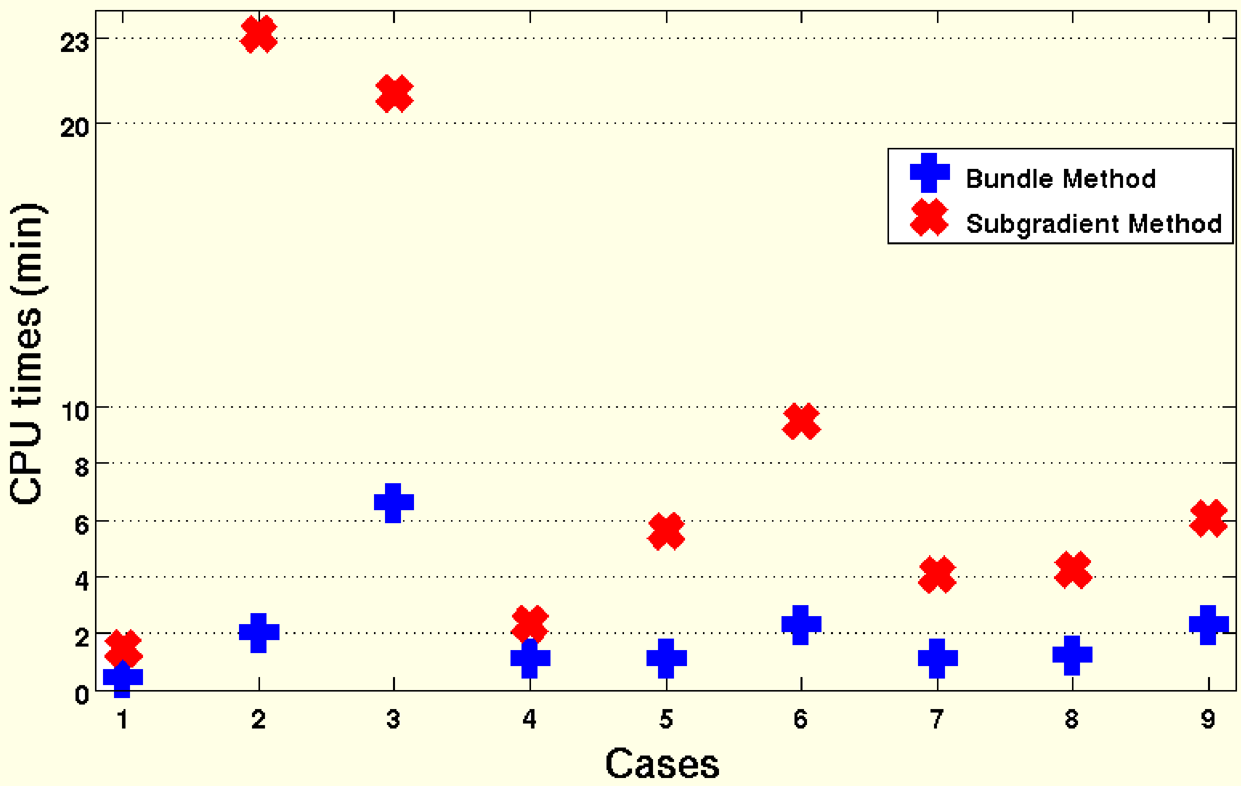
Any model satisfying these conditions

that is used in the QP

maintains the convergence results

Comparing the methods: bundle and SG

Typical performance on a battery of Unit Commitment problems

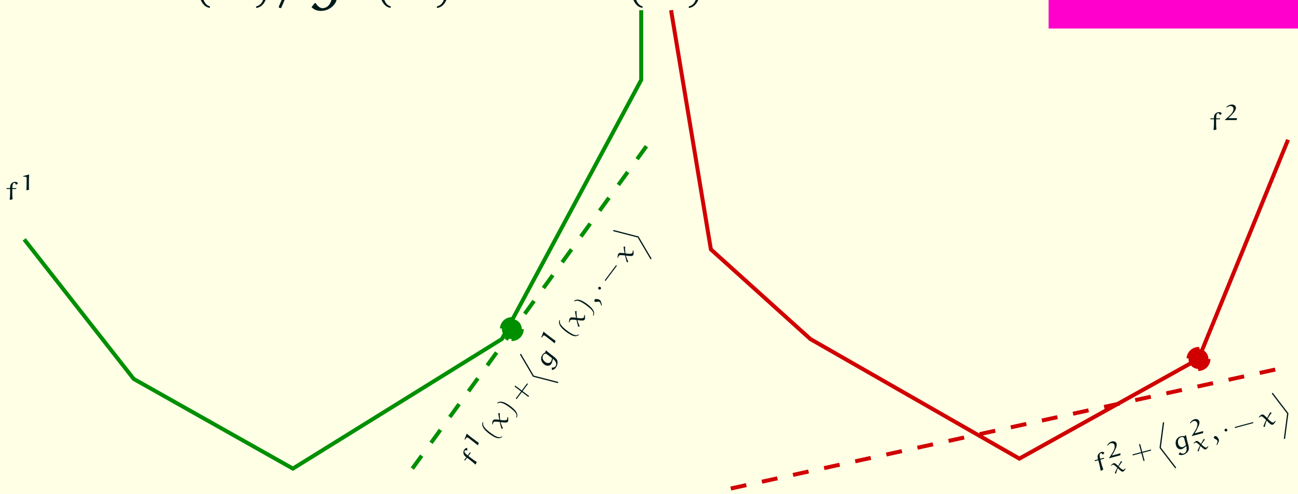


Bundle Methods with on-demand accuracy

the new generation

Oracle types: exact and upper

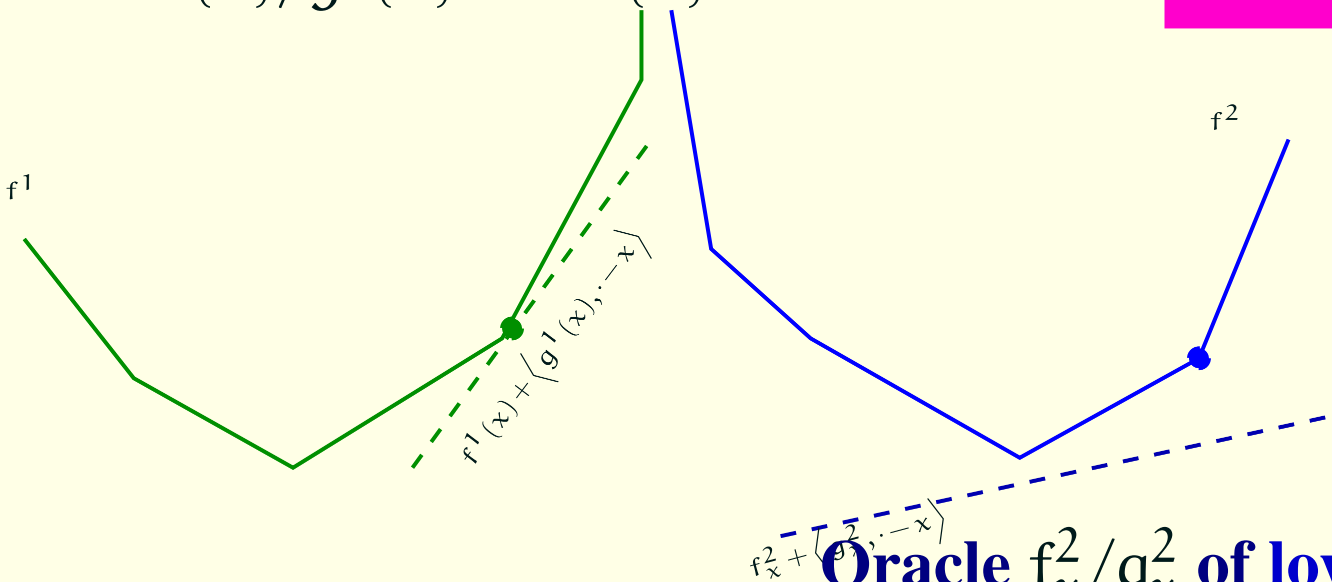
- $f^1(x)/g^1(x) \in \partial f^1(x)$ is **easy**: **exact $f^1(x)/g^1(x)$**
- $f^2(x)/g^2(x) \in \partial f^2(x)$ is **difficult:** **inexact f^2_x/g^2_x**



Oracle f^2_x/g^2_x NOT of lower type

Oracle types: exact and lower

- $f^1(x)/g^1(x) \in \partial f^1(x)$ is **easy**: **exact $f^1(x)/g^1(x)$**
- $f^2(x)/g^2(x) \in \partial f^2(x)$ is **less difficult:** **inexact f^2_x/g^2_x**



Oracle f^2_x/g^2_x of lower type

For the EM problem $f^j(x) = \max\{-\mathcal{C}^j(q^j) + \left\langle x, g^j(q^j)\right\rangle : q^j \in \mathcal{P}^j\}$

By computing f_{x^k} and g_{x^k} satisfying

$$f_{x^k} = f(x^k) - \eta^k \quad \text{and} \quad g_{x^k} \in \partial_{\eta^k} f(x^k)$$

we can build

- A lower oracle
- An asymptotically exact oracle

$$\eta^k \to 0 \quad \text{as} \quad k \to \infty$$

- A partly asymptotically exact oracle

$$\eta^k \to 0 \quad \text{as} \quad K_s \ni k \to \infty$$

- An on-demand accuracy oracle

$$\eta^k \leq \bar{\eta}^k \quad \text{when} \quad f_{x^k} \leq f_{\hat{x}^k} - m\delta_k$$

BM with lower inexact oracles

- $\mathbf{M}_k(x) = \max\{f_{x^i} + \left\langle g_{x^i}, x - x^i \right\rangle : i \in \boldsymbol{\mathcal{B}}_k\}$
- $\delta^k = \varepsilon_k + t_k|G^k|^2$
- SS test: $f_{x^{k+1}} \leq \hat{f}^k - m\delta^k$
- $\hat{f}^k := \max\left\{f_{\hat{x}^k}, \max\left(\mathbf{M}_j(\hat{x}^k), j \geq \hat{k}\right)\right\}$

+ Oracle inaccuracy is locally bounded:
$\forall R \geq 0 \exists \eta(R) \geq 0 : |x| \leq R \Longrightarrow \eta \leq \eta(R)$ convergence as before, up to the accuracy on SS

BM with lower inexact oracles

- $\mathbf{M}_k(x) = \max\{f_{x^i} + \left\langle g_{x^i}, x - x^i \right\rangle : i \in \mathcal{B}_k\}$
- $\delta^k = \varepsilon_k + t_k |G^k|^2$
- SS test: $f_{x^{k+1}} \leq \hat{f}^k - m\delta^k$
- $\hat{f}^k := \max\left\{f_{\hat{x}^k}, \max\left(\mathbf{M}_j(\hat{x}^k), j \geq \hat{k}\right)\right\}$

+ Oracle inaccuracy is locally bounded:
$\forall R \geq 0 \exists \eta(R) \geq 0 : |x| \leq R \Longrightarrow \eta \leq \eta(R)$ convergence as before, up to the accuracy on SS Convex proximal bundle methods in depth: a unified analysis for inexact oracles W. de Oliveira, C. Sagastizábal, C. Lemaréchal MathProg 148, pp 241-277, 2014

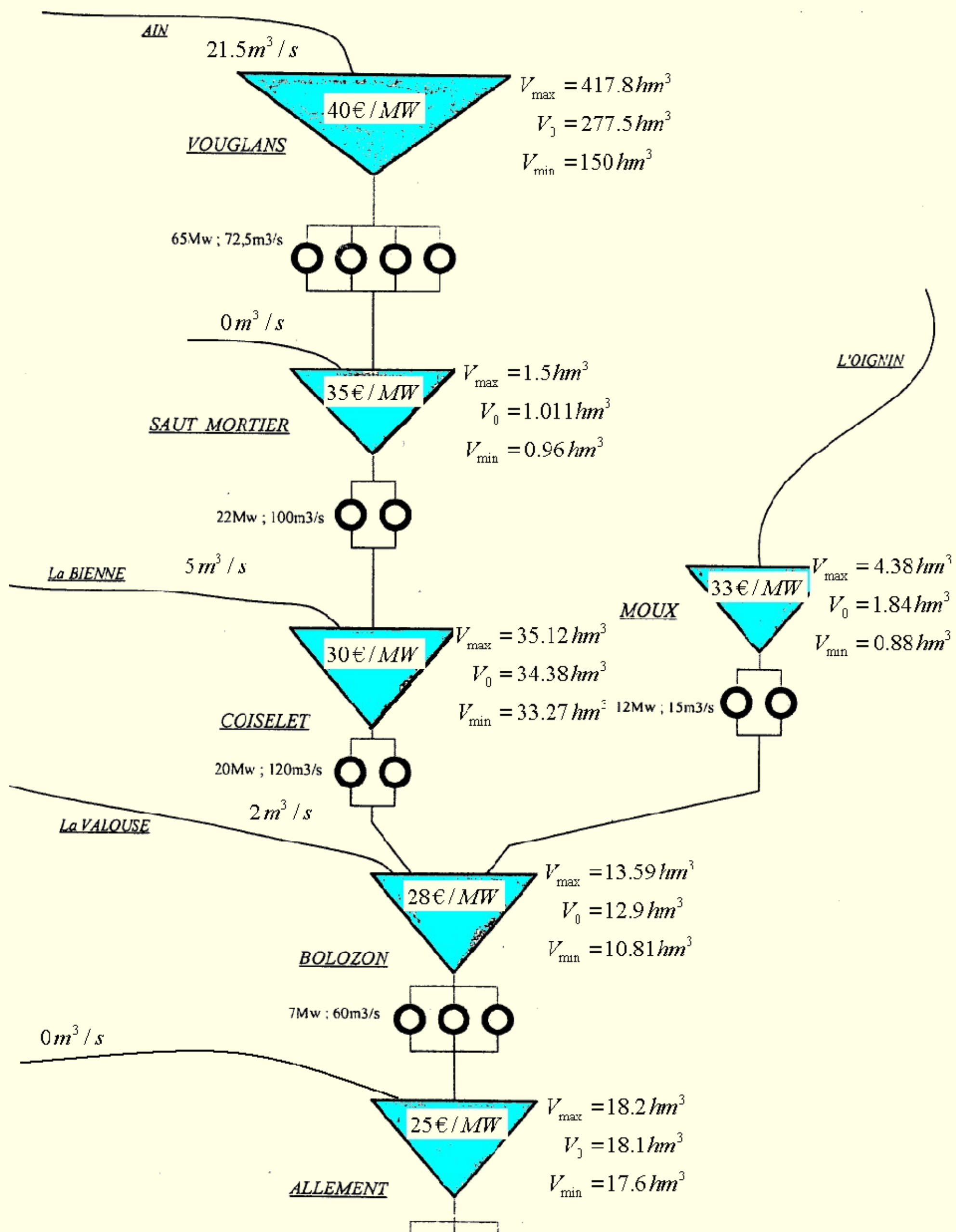
General comments

Bundle methods are

- robust (do not oscillate, as CP methods do)
- reliable (have a stopping test, unlike SG methods)
- can deal with inaccuracy in a reasonable manner

Extending bundle methods

Constrained NSO problems: an example



Optimal management of the hydrovalley

$$\begin{cases} \max & \lambda^\top \mathbb{E}_\eta\Big(\rho(u)\Big) \\ \text{s.t.} & (x,u) \in \mathcal{P} \\ & \mathbb{P}_\eta\Big(Au + a_{\min} \leq M\eta \leq Au + a_{\max}\Big) \geq p \end{cases}$$

Optimal management of the hydrovalley

$$\begin{cases} \max & \lambda^\top \mathbb{E}_\eta\Big(\rho(u)\Big) \\ \text{s.t.} & (x,u) \in \mathcal{P} \\ & \mathbb{P}_\eta\Big(Au + a_{\min} \leq M\eta \leq Au + a_{\max}\Big) \geq p \end{cases}$$

Is this a convex program?

Optimal management of the hydrovalley

$$\begin{cases} \max & \lambda^\top \mathbb{E}_\eta\Big(\rho(u)\Big) \\ \text{s.t.} & (x,u) \in \boldsymbol{\mathcal{P}} \\ & \mathbb{P}_\eta\Big(Au + a_{\min} \le M\eta \le Au + a_{\max}\Big) \ge p \end{cases}$$

Is this a convex program? YES: the function

$$u \mapsto \log\Big(\mathbb{P}_\eta\Big(Au + a_{\min} \le M\eta \le Au + a_{\max}\Big)\Big) \qquad \text{is convex.}$$

We need to solve $\begin{cases} \min & f(u) \\ \text{s.t.} & (x,u) \in \boldsymbol{\mathcal{P}} \\ & c(u) \le 0 \end{cases}$ for linear f and with

$$c(u) := \log\Big(\mathbb{P}_\eta\Big(Au + a_{\min} \le M\eta \le Au + a_{\max}\Big)\Big) - \log p$$

Optimal management of the hydrovalley

$$\begin{cases} \max & \lambda^\top \mathbb{E}_\eta\Big(\rho(u)\Big) \\ \text{s.t.} & (x,u) \in \mathcal{P} \\ & \mathbb{P}_\eta\Big(Au + a_{\min} \le M\eta \le Au + a_{\max}\Big) \ge p \end{cases}$$

Is this a convex program? YES: the function

$$u \mapsto \log\left(\mathbb{P}_\eta\left(Au + a_{\min} \le M\eta \le Au + a_{\max}\right)\right) \qquad \text{is convex.}$$

We need to solve $\begin{cases} \min & f(u) \\ \text{s.t.} & (x,u) \in \mathcal{P} \\ & c(u) \le 0 \end{cases}$ for linear f and with

$$c(u) := \log\left(\mathbb{P}_\eta\left(Au + a_{\min} \le M\eta \le Au + a_{\max}\right)\right) - \log p$$

difficult to compute!

Need to solve the constrained problem

$$(P) \qquad \left\{ \begin{array}{ll} \min & f(u) \\ \text{s.t.} & (x,u) \in \mathcal{P} \\ & c(u) \leq 0 \end{array} \right.$$

for linear f and with inexact evaluation of c and its gradient, via a black box with controllable inaccuracy (bounded by a given tolerance ε, with confidence level 99%, noting that evaluation errors can be positive or negative)

Handling constraints in NSO

For nonsmooth constrained problems

$$\min f(u) \quad \text{s.t.} \quad c(u) \leq 0$$

use the Improvement Function

$$\max_u \{f(u) - f(\hat{u}), c(u)\}$$

(changes with each serious point $\hat{u}$ and supposes exact f/c values available) [SagSol SiOPT, 2005 and KarasRibSagSol MPB, 2009]

Improvement function

Let $(\bar{x}, \bar{u})$ be a solution to (P). The function

$$H_{\bar{u}}(u) := \max_{(x,u)\in\mathcal{P}} \{f(u) - f(\bar{u}), c(u)\}$$

has perfect theoretical properties:

If Slater condition ($\exists\,(x,u) \in \mathcal{P}$ s.t. $c(u) < 0$) holds, then

$$\bar{u} \text{ solves } \min_{(x,u)\in\mathcal{P}} f(u) \quad \text{s.t.} \quad c(u) \le 0 \qquad \textbf{(P)}$$

$$\Updownarrow$$

$$\min_{(x,u)\in\mathcal{P}} H_{\bar{u}}(u) = H_{\bar{u}}(\bar{u}) = 0$$

$$\Updownarrow$$

$$0 \in \partial H(\bar{u}) \text{ for } H(\cdot) := H_{\bar{u}}(\cdot)$$

Improvement function

Let $(\bar{x}, \bar{u})$ be a solution to (P). The function

$$H_{\bar{u}}(u) := \max_{(x,u)\in\mathcal{P}} \{f(u) - f(\bar{u}), c(u)\}$$

has perfect theoretical properties:

Without Slater condition

$$\bar{u} \text{ solves } \min_{(x,u)\in\mathcal{P}} f(u) \quad \text{s.t.} \quad c(u) \leq 0 \qquad \textbf{(P)}$$

$$\Downarrow \quad \textbf{BUT} \not\Uparrow$$

$$\min_{(x,u)\in\mathcal{P}} H_{\bar{u}}(u) = H_{\bar{u}}(\bar{u}) = 0$$

$$\Downarrow \quad \text{and also} \Uparrow$$

$$0 \in \partial H(\bar{u}) \text{ for } H(\cdot) := H_{\bar{u}}(\cdot)$$

Improvement function

Let $(\bar{x}, \bar{u})$ be a solution to (P). The function

$$H_{\bar{u}}(u) := \max_{(x,u)\in\mathcal{P}} \{f(u) - f(\bar{u}), c(u)\}$$

has perfect theoretical properties:

Without Slater condition

$$\bar{u} \text{ solves } \min_{(x,u)\in\mathcal{P}} f(u) \quad \text{s.t.} \quad c(u) \le 0 \qquad \textbf{(P)}$$

$\Uparrow$: **when $c(\bar{u}) \le 0$ $\bar{u}$ solves (P), otherwise it minimizes infeasibility over $\mathcal{P}$**

$$\min_{(x,u)\in\mathcal{P}} H_{\bar{u}}(u) = H_{\bar{u}}(\bar{u}) = 0$$

$$\Updownarrow$$

$$0 \in \partial H(\bar{u}) \text{ for } H(\cdot) := H_{\bar{u}}(\cdot)$$

Handling nonconvex problems

- Nonconvex proximal point mapping [PR96]

$$p_R f(x) := \operatorname{argmin}_{y \in \mathbb{R}^N} \left\{ f(y) + \frac{R}{2}|y - x|^2 \right\}$$

x is the <u>prox-center</u> and $R > R_x$ is the <u>prox-parameter</u>

Theorem If f is convex

- $p_R f$ is well defined **for any** $R > 0$.
- $p_R f$ is single valued and loc. Lip.
- $p = p_R f(x) \iff R(x - p) \in \partial f(p)$
- x^* minimizes $f \iff x^* = p_R f(x^*)$ for any $R > 0$.
- $x_{k+1} = p_R f(x_k)$ converges to a minimizer x^*.

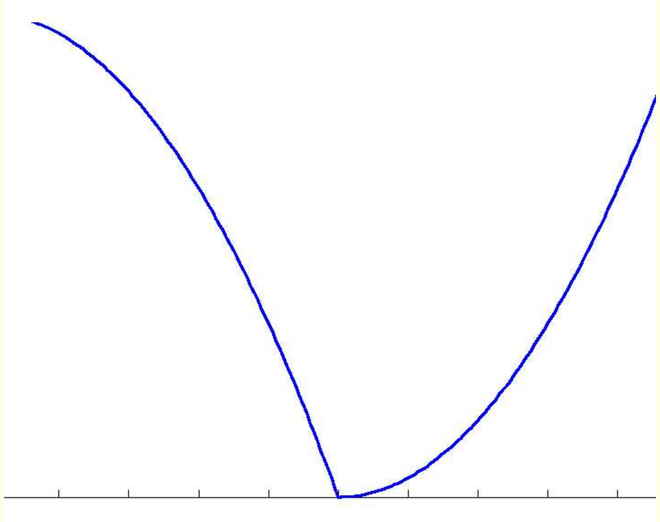
What is f is nonconvex?

Nonconvex difficulties

Proximal Bundle Methods are the most robust and reliable (oracle) methods for convex minimization. Their success relies heavily on convexity. If f is convex:

- $x_{k+1} = p_R f(x_k)$ converges to a minimizer x^*.
- $\check{f}_k$ lies **entirely** below f.

May no longer be true for nonconvex f

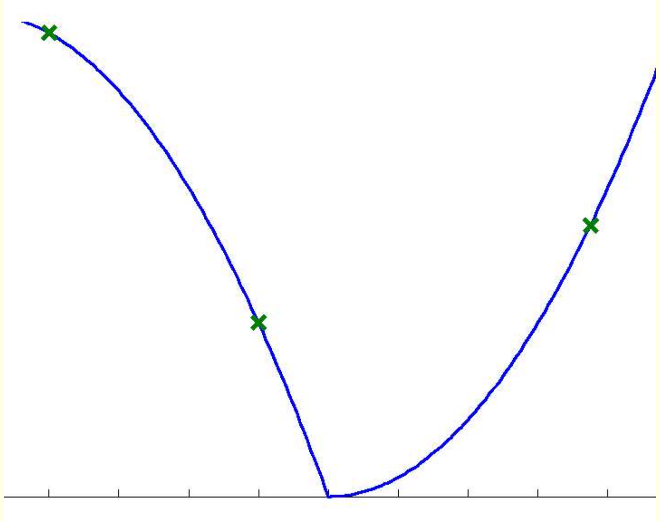


Nonconvex difficulties

Proximal Bundle Methods are the most robust and reliable (oracle) methods for convex minimization. Their success relies heavily on convexity. If f is convex

- $x_{k+1} = p_R f(x_k)$ converges to a minimizer x^*.
- $\check{f}_k$ lies **entirely** below f.

May no longer be true for nonconvex f

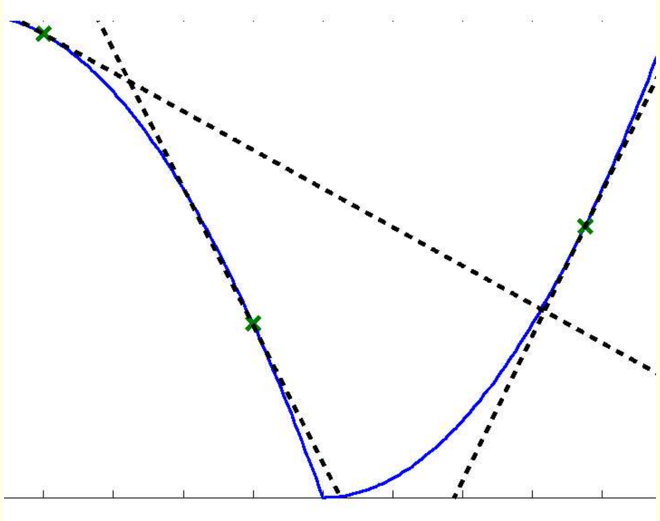


Nonconvex difficulties

Proximal Bundle Methods are the most robust and reliable (oracle) methods for convex minimization. Their success relies heavily on convexity. If f is convex

- $x_{k+1} = p_R f(x_k)$ converges to a minimizer x^*.
- $\check{f}_k$ lies **entirely** below f.

May no longer be true for nonconvex f

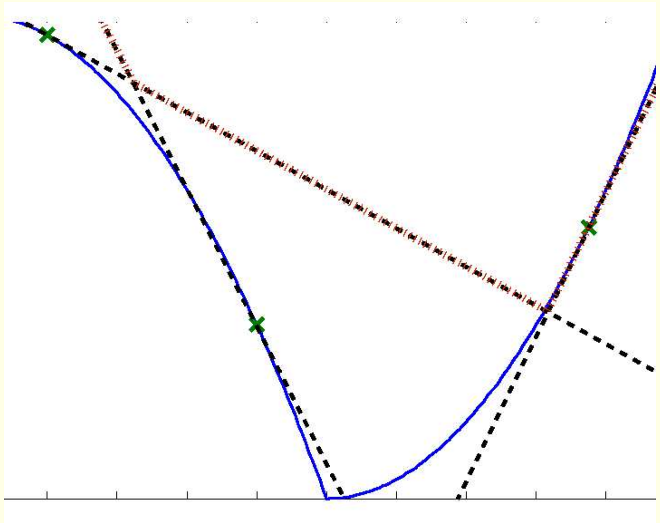


Nonconvex difficulties

Proximal Bundle Methods are the most robust and reliable (oracle) methods for convex minimization. Their success relies heavily on convexity. If f is convex

- $x_{k+1} = p_R f(x_k)$ converges to a minimizer x^*.
- $\check{f}_k$ lies **entirely** below f.

May no longer be true for nonconvex f

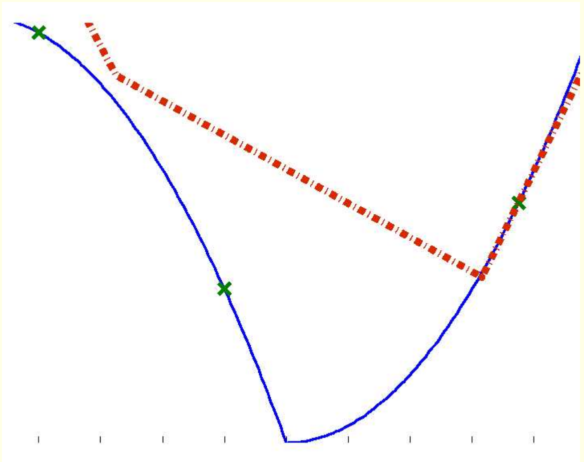


Nonconvex difficulties

Proximal Bundle Methods are the most robust and reliable (oracle) methods for convex minimization. Their success relies heavily on convexity. If f is convex

- $x_{k+1} = p_R f(x_k)$ converges to a minimizer x^*.
- $\check{f}_k$ lies **entirely** below f.

May no longer be true for nonconvex f



How this difficulty has been addressed?

Take each plane in the model: $f_i + \langle g_i, \cdot - y_i \rangle$ and rewrite it, centered at x_k:

$$\begin{array}{lcl} f(x_k) - & \left[f(x_k) - \left(f_i + \langle g_i, x_k - y_i \rangle \right) \right] & + \langle g_i, \cdot - x_k \rangle \\ f(x_k) - & e^f_{k,i} & + \langle g_i, \cdot - x_k \rangle \end{array}$$

$$\Rightarrow \check{f}_k(y) = \max \left\{ f(x_k) - e^f_{k,i} + \langle g_i, y - x_k \rangle \right\}$$

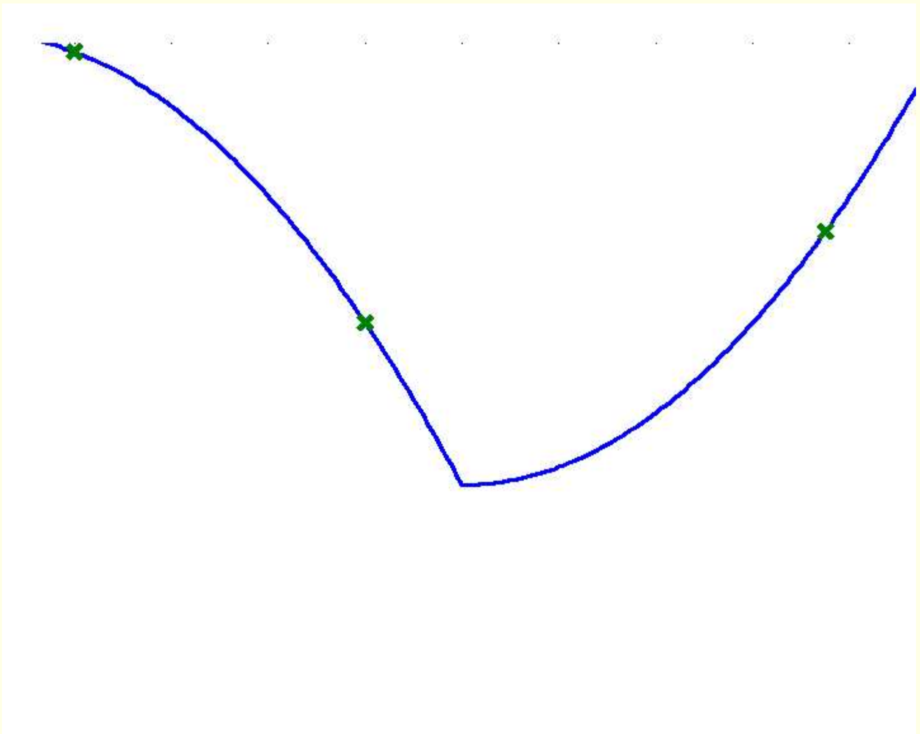
Good: $e^f_{k,i}$ positive $\Rightarrow$ convergence **Good**: If f convex $\Rightarrow e^f_{k,i}$ positive. **BAD**: If f nonconvex, $e^f_{k,i}$ may be **negative**

Nonconvex bundle methods

fix negative linearization errors, replacing $\check{f}_k$ by:

$$\check{f}_k^{\mathbf{FIX}}(y) = \max\left\{ f(x_k) - |\mathbf{e}_{\mathbf{k,i}}^{\mathbf{f}}| + \langle g_i, y - x_k \rangle \right\}$$

[Mif77, Lem80, Kiw85, Luk98]

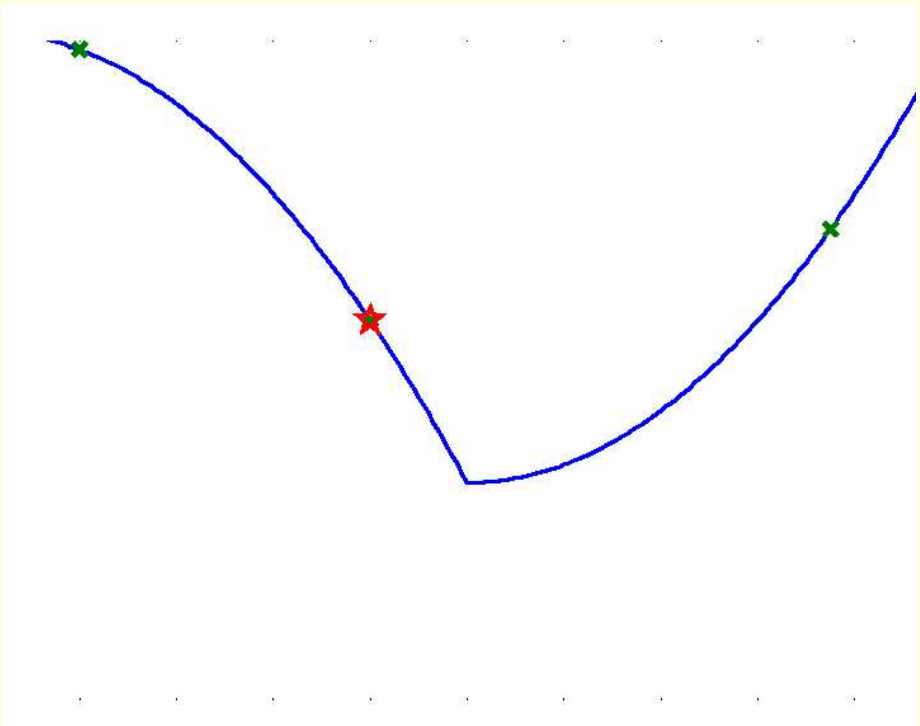


Nonconvex bundle methods

fix negative linearization errors, replacing $\check{f}_k$ by:

$$\check{f}_k^{\mathbf{FIX}}(y) = \max\left\{ f(x_k) - |\mathbf{e^f_{k,i}}| + \langle g_i, y - x_k \rangle \right\}$$

[Mif77, Lem80, Kiw85, Luk98]

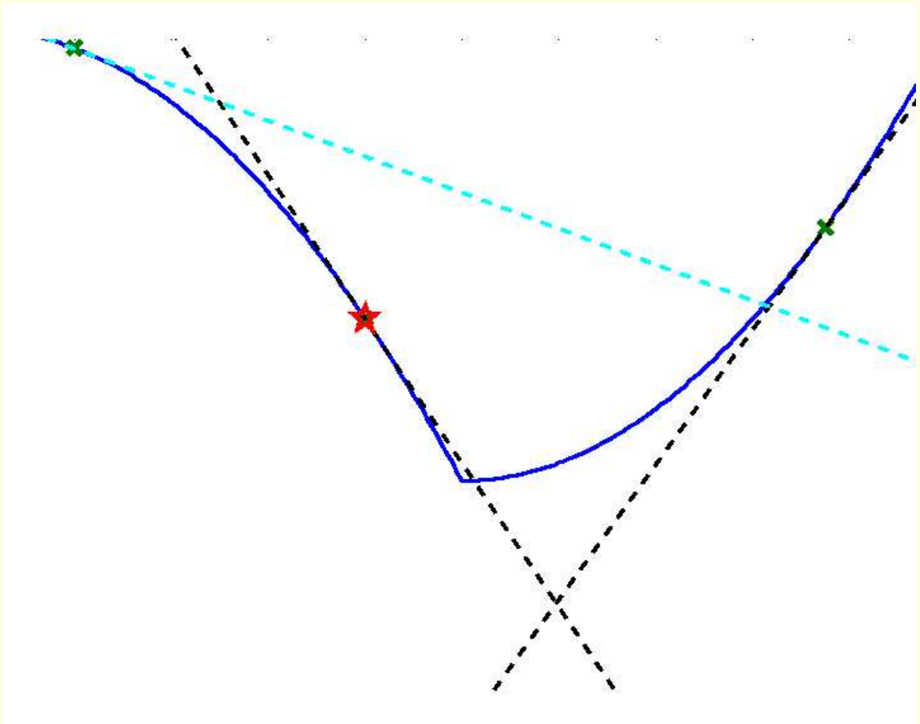


Nonconvex bundle methods

fix negative linearization errors, replacing $\check{f}_k$ by:

$$\check{f}_k^{\mathbf{FIX}}(y) = \max\left\{ f(x_k) - |\mathbf{e}_{\mathbf{k,i}}^{\mathbf{f}}| + \langle g_i, y - x_k\rangle \right\}$$

[Mif77, Lem80, Kiw85, Luk98]

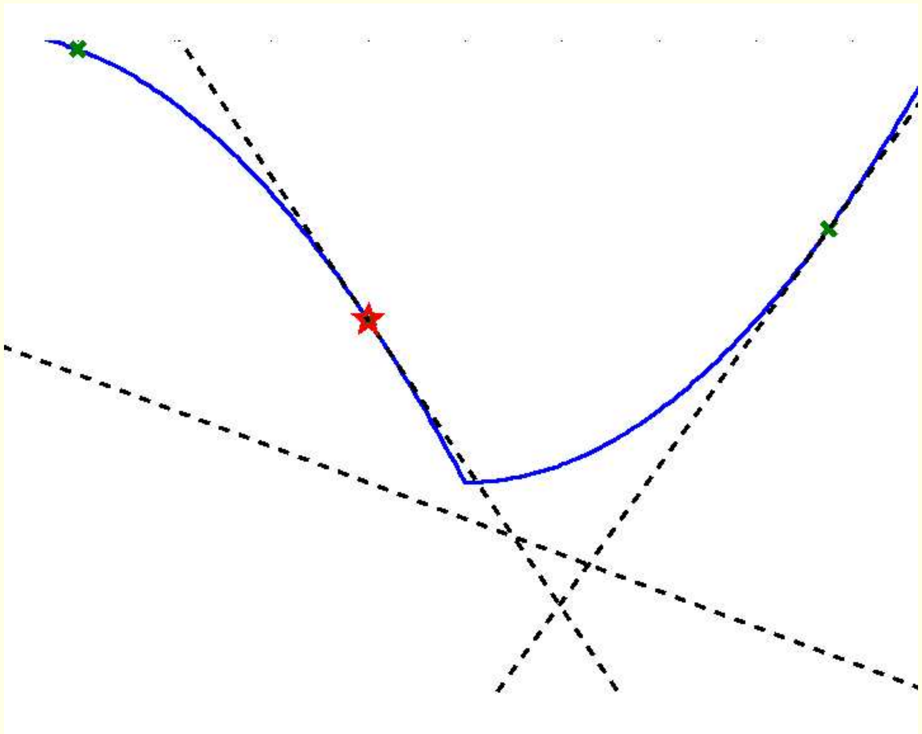


Nonconvex bundle methods

fix negative linearization errors, replacing $\check{f}_k$ by:

$$\check{f}_k^{\mathbf{FIX}}(y) = \max\left\{ f(x_k) - |\mathbf{e}_{\mathbf{k,i}}^{\mathbf{f}}| + \langle g_i, y - x_k \rangle \right\}$$

[Mif77, Lem80, Kiw85, Luk98]

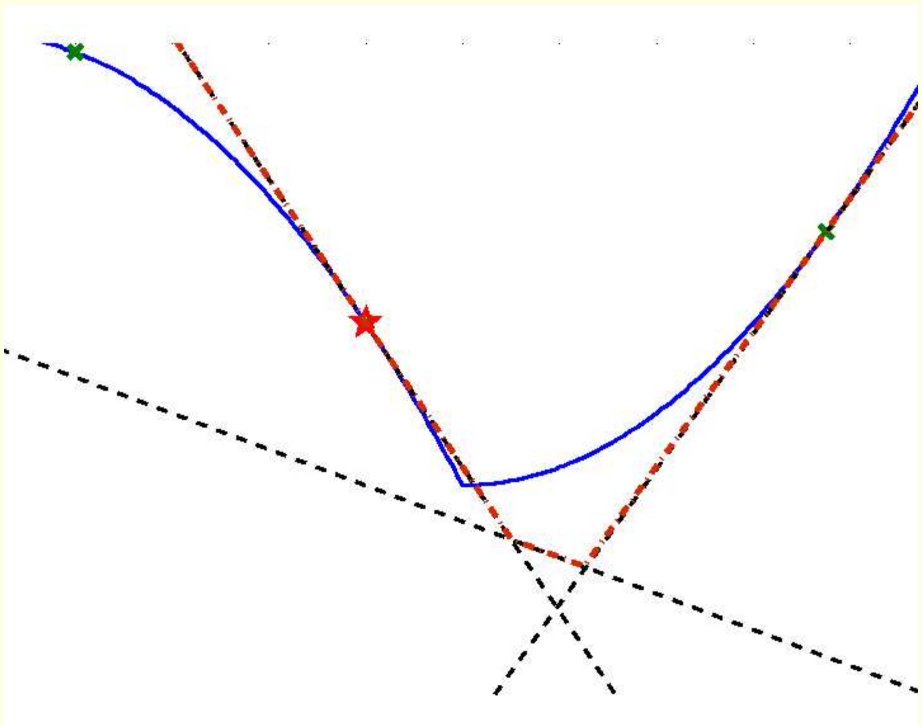


Nonconvex bundle methods

fix negative linearization errors, replacing $\check{f}_k$ by:

$$\check{f}_k^{\mathbf{FIX}}(y) = \max\left\{ f(x_k) - |\mathbf{e}^{\mathbf{f}}_{\mathbf{k,i}}| + \langle g_i, y - x_k \rangle \right\}$$

[Mif77, Lem80, Kiw85, Luk98]

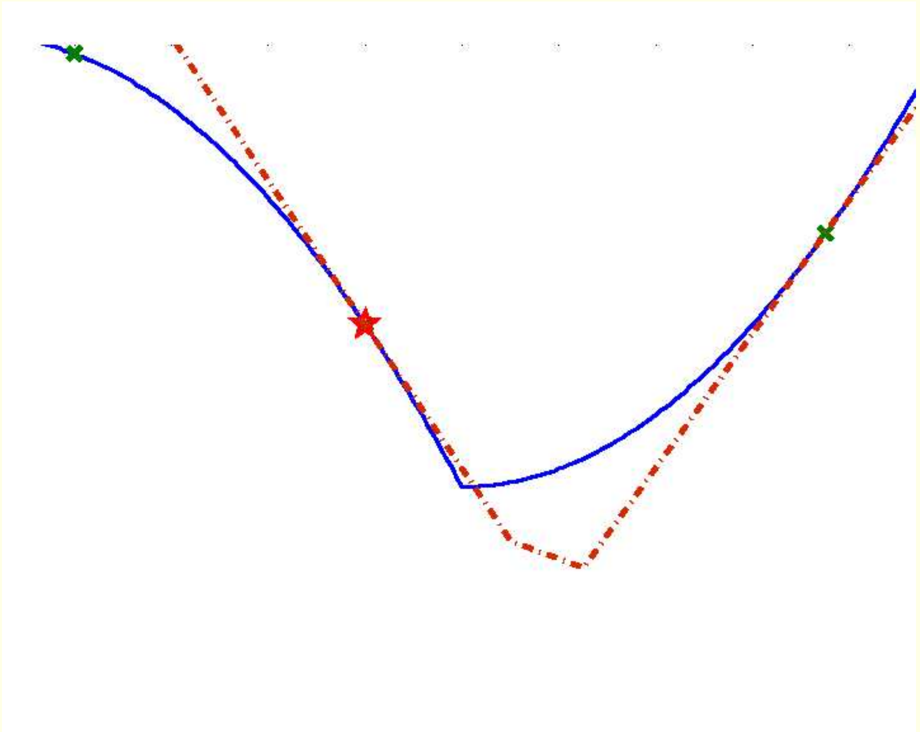


Nonconvex bundle methods

fix negative linearization errors, replacing $\check{f}_k$ by:

$$\check{f}_k^{\mathbf{FIX}}(y) = \max\left\{ f(x_k) - |\mathbf{e}_{\mathbf{k,i}}^{\mathbf{f}}| + \langle g_i, y - x_k\rangle \right\}$$

[Mif77, Lem80, Kiw85, Luk98]



A new method

A different approach (ours) is based on the following trick

Take $\eta, \mu > 0 : R = \eta + \mu$ and note

$$
\begin{aligned}
p_R f(x_k) &= \min_w \Big\{ f(w) + R \tfrac{1}{2}|w - x_k|^2 \Big\} \\
&= \min_w \Big\{ f(w) + (\eta + \mu) \tfrac{1}{2}|w - x_k|^2 \Big\} \\
&= \min_w \Big\{ f(w) + \eta \tfrac{1}{2}|w - x_k|^2 + \mu \tfrac{1}{2}|w - x_k|^2 \Big\} \\
&= \min_w \Big\{ F_\eta(w) + \mu \tfrac{1}{2}|w - x_k|^2 \Big\} \\
&= p_\mu(F_\eta) \ (x_k)
\end{aligned}
$$

$$\Rightarrow p_R f = p_\mu(F_\eta)$$

Redistributed Proximal Bundle Method

At ℓ^{th}-iteration, for $k = k(\ell)$, given R_k, x_k and a bundle
$\mathcal{B} = \{y_i, f_i, g_i\,, i \in I_\ell\}$

0. Split R_k into η_ℓ and μ_ℓ.

1. Model F_{η_ℓ} $\quad \check{F}_{\eta_\ell,\ell}(y) = \max_{i\in\mathcal{B}}\{F_{\eta_\ell\, i} + \langle g_{\eta_\ell\, i}, y - y_i\rangle\}$

2. Minimize **the penalized model**
$y_{\ell+1} = \arg\min\{\check{F}_{\eta_\ell,\ell}(y) + \frac{\mu_\ell}{2}|y - x_k|^2\}$

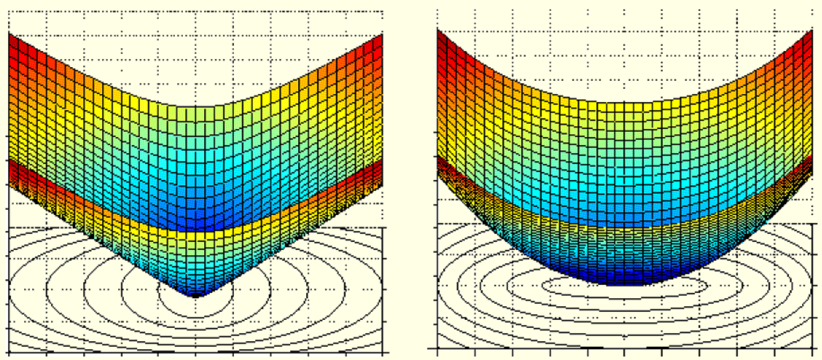
3. Descent test If $y_{\ell+1}$ good: $x_{k+1} \leftarrow y_{\ell+1}$, define R_{k+1} **serious step**

If $y_{\ell+1}$ bad: **null step**

4. Update bundle $\quad \mathcal{B} \leftarrow \mathcal{B} \cup \{y_{\ell+1}, f_{\ell+1}, g_{\ell+1}\}$

$\mathcal{VU}$ quasi-Newton bundle

For $x \in \mathbb{R}^n$, given matrices $A \succeq 0$, $B \succ 0$, $f(x) = \sqrt{x^\top A x} + x^\top B x$ has a unique minimizer at $\bar{x} = 0$. On $\mathcal{N}(A)$ the function is not differentiable, and the first term vanishes: $f|_{\mathcal{N}(A)}$ looks smooth.



$\mathcal{R}(A)$ $\mathcal{N}(A)$

$\mathcal{V}$ $\mathcal{U}$

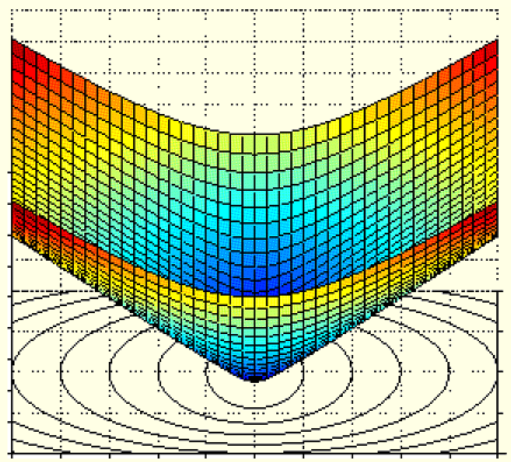
parallel to $\partial f(\bar{x})$ $\mathcal{U}$ perpendicular to $\mathcal{V}$

$\mathcal{V}$ is parallel to $\mathcal{N}(A)$, the “ridge" of nonsmoothness

$\mathcal{VU}$-Algorithm:

(Mifflin&Sagastizábal, MathProg 05) Recall that

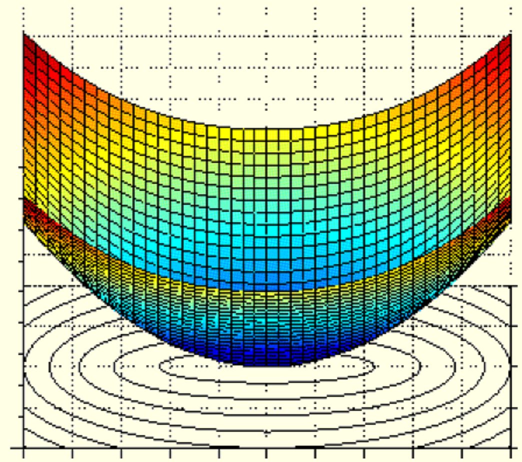
$f|_{\mathcal{V}\|\mathcal{N}(A)}$ is nice: **the key is the two QP-solves**



$\mathcal{R}(A)$

$\mathcal{V}$

2 bundle QPs



$\mathcal{N}(A)$

$\mathcal{U}$

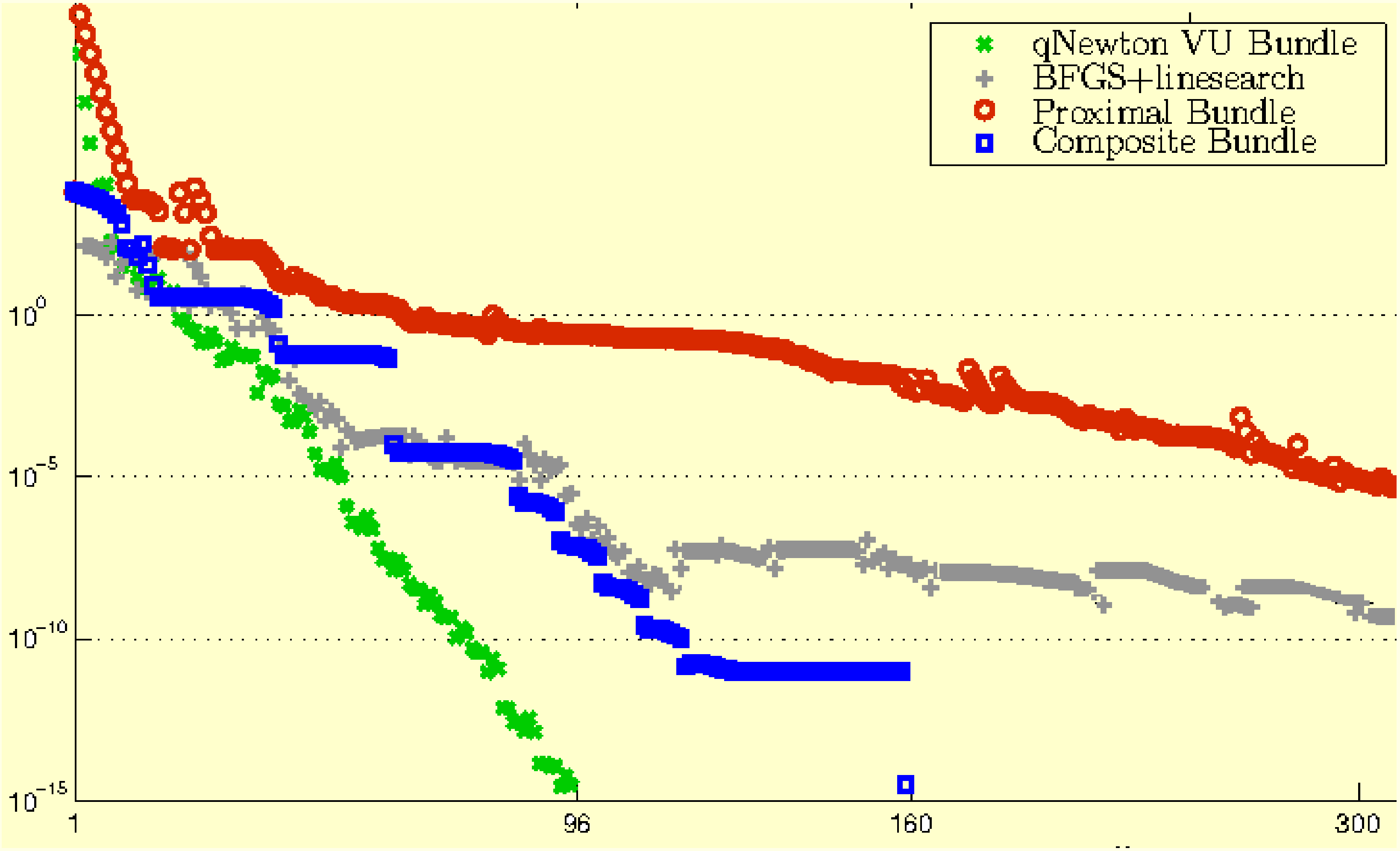
Newton-move

Two successive bundle QPs identify the “ridge” of nonsmoothness

Solve a 2nd QP to create an approximation of $\mathcal{V}$ based on $\partial \check{f}(\hat{p}^k)$

$\mathcal{VU}$-Algorithm:

superlinear "serious" subsequence (Mifflin&Sag, MathProg 05)



To learn more

Bundle methods history

R. MIFFLIN, C. SAGASTIZÁBAL, Documenta Math, 2012. A Science Fiction Story in Nonsmooth Optimization Originating at IIASA
https://www.math.uni-bielefeld.de/documenta/vol-ismp/44_mifflin-robert.pdf

(exact) Bundle books

J.F. BONNANS, J.C. GILBERT, C. LEMARÉCHAL, AND C. SAGASTIZÁBAL, Numerical Optimization: Theoretical and Practical Aspects, Springer, 2nd ed., 2006.

J.B. HIRIART-URRUTY AND C. LEMARÉCHAL, Convex Analysis and Minimization Algorithms II, no. 306 in Grund. der math. Wissenschaften, Springer, 2nd ed., 1996.

Inexact Bundle theory

(next page)

Inexact Bundle theory

M. HINTERMÜLLER, A proximal bundle method based on approximate subgradients, COAp 20 (2001), pp. 245–266.

M. V. SOLODOV, On Approximations with Finite Precision in Bundle Methods for Nonsmooth Optimization. JoTA 119.1 (2003), pp. 151–165

K.C. KIWIEL, A proximal bundle method with approximate subgradient linearizations, SiOpt 16 (2006), pp. 1007–1023.

W. DE OLIVEIRA, C. SAGASTIZÁBAL, AND C. LEMARÉCHAL, Convex proximal bundle methods in depth: a unified analysis for inexact oracles, MathProg 148 (2014), pp. 241–277.

Inexact Bundle variants with applications

G. EMIEL AND C. SAGASTIZÁBAL, Incremental-like bundle methods with application to energy planning, COAp 46 (2010), pp. 305–332.

W. DE OLIVEIRA, C. SAGASTIZÁBAL, AND S. SCHEIMBERG, Inexact bundle methods for two-stage stochastic programming, SiOpt 21 (2011), pp. 517–544.

(next page)

W. van Ackooij and C. Sagastizábal, Constrained bundle methods for upper inexact oracles with application to joint chance constrained energy problems, SiOpt 24 (2014), pp. 733–765.

W. de Oliveira and C. Sagastizábal, Level bundle methods for oracles with on-demand accuracy, OMS 29 (2014), pp. 1180 –1209

W. de Oliveira and C. Sagastizábal, Bundle methods in the xxi century: A birds'-eye view, Pesquisa Operacional 34 (2014), pp. 647 – 670.

W. de Oliveira and M. Solodov, A doubly stabilized bundle method for nonsmooth convex optimization, MathProg 156(1), pp. 126–159, 2016.

Any doubts or questions?

Just e-mail me