Nonsmooth Convex Unconstrained Multivariate Optimization

Antonio Frangioni

Department of Computer Science
University of Pisa
https://www.di.unipi.it/~frangio
mailto:frangio@di.unipi.it

Computational Mathematics for Learning and Data Analysis
Master in Computer Science – University of Pisa

A.Y. 2024/25

Outline

Motivation I: Incremental (a.k.a. Stochastic) Gradient in ML

- $I = \{ 1, \ldots, m \}$, $X = [x^i \in \mathbb{R}^h]_{i \in I}$ inputs, $y = [y^i \in \mathbb{R}^k]_{i \in I}$ outputs
- Arbitrarily complex predictor $\pi(x ; w) : \mathbb{R}^h \to \mathbb{R}^k$ parametric on $w \in \mathbb{R}^n$, $\mathcal{L} : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}$ loss function (could be $\mathcal{L}^i$), fitting

  $$\min \{ f(w) = \sum_{i \in I} [f^i(w) = \mathcal{L}(y^i , \pi(x^i ; w))] \; : \; w \in \mathbb{R}^n \}$$
- $\nabla f(w) = \sum_{i \in I} \nabla f^i(w)$: sum of the m gradients of individual f^i
- Linear least squares: $\pi(x ; w) = \langle x , w \rangle$, $\mathcal{L} = (y , z) = (y - z)^2 / 2 \implies$

  $$f^i(w) = (y^i - \langle x^i , w \rangle)^2 / 2 \quad , \quad \nabla f^i(w) = -x^i(y^i - \langle x^i , w \rangle)$$
- Each ∇f^i cheap, but m large $\implies$ computing "the full" ∇f costly already
- Intuition: x^i are i.i.d. $\implies$ ∇f^i are $\implies$ "many of them will cancel out" $\implies$ a small sample is enough to compute a close $\approx$ to the "true" ∇f
- $K \subset I$ "small", $\nabla f^K(w) = \sum_{i \in K} \nabla f^i(w)$ = incremental gradient
- Cheaper but $-\nabla f^K$ not a descent direction, a $\neq$ analysis is needed (but Heavy Ball and ACCG are not descent methods, either)

- How to choose K? What $\# K$ should be?
- Apparently no better way than at random $\equiv$ stochastic gradient
- Iteration with $K = I$ "batch", $\# K < m$ "mini batch" (often $\# K = 1$)
- "Extreme" version: on-line. Observations keep coming (typically fast), have to be used immediately one by one and immediately discarded (no memory)
- Results often given in terms of $\mathbb{E}(\cdot)$ and of the "mean of iterates"
  $\bar{x}^i = (\sum_{k=0}^{i} x^k) / i$ (Cesáro average), $\{ \bar{x}^i \} \to x_*$ if $\{ x^i \}$ does
- With $\# K = 1$, results rather worse than deterministic case, e.g. [1, Th. 6.3]
  $f \in C^1$ and τ-convex $\Longrightarrow$ $\mathbb{E}(f(\bar{x}^i) - f_*) \leq \varepsilon$ for $i \geq O(1 / \varepsilon^2)$
- Things improve as $\# K \nearrow$ [1, p. 334], but iteration cost $\nearrow$ too
- General observation: first-order methods are "quite robust" to errors in ∇f
- Will come in handy presently

- Any ML expert would add a regularizer $\Omega(\,w\,)$ (better in theory & practice)

  $$\min\left\{\ \sum_{i\in I}\mathcal{L}(\,y^i\,,\,\pi(\,x^i\,;\,w\,)\,)+\mu\Omega(\,w\,)\ :\ w\in\mathbb{R}^n\,\right\}$$

  μ hyper-parameter (scalarization of multi-objective), grid search ...

- Standard ridge regularization: $\Omega(\,w\,)=\|\,w\,\|_2^2\,/\,2\ \in C^\infty$, $\nabla\Omega(\,w\,)=w$

- Regularization simplifies model $\implies$ better generalization (if done well)

- Other way to simplify model: decrease n $\equiv$ feature selection

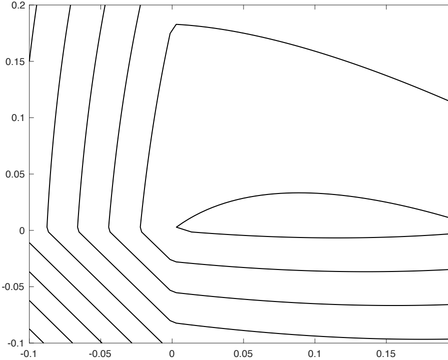
- Can kill two birds with a stone: $\Omega=\|\cdot\|_0$ very nasty function, $\notin C^0$
  (could be written as a Mixed-Integer Nonlinear Problem [2] ...)

- Workable alternative: $\Omega=\|\cdot\|_1$ = Lasso, best convex approximation of $\|\cdot\|_0$

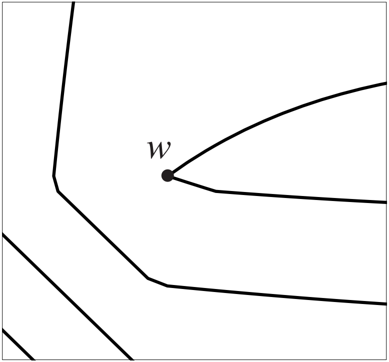
- Increases sparsity in practice, convex, $\in C^0$ but $\notin C^1$

- Is this a real problem? You bet.

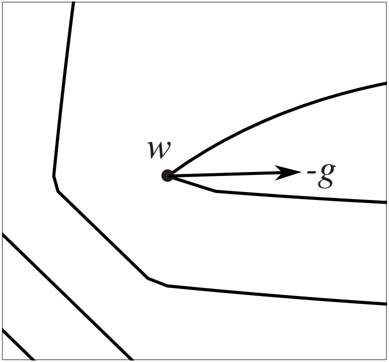
- $x^1 = [\,3\,,\,2\,]$, $y^1 = 2$, $\mu = 10 \implies$
  $$f(\,w_1\,,\,w_2\,) = (\,3w_1 + 2w_2 - 2\,)^2 + 10(\,|\,w_1\,| + |\,w_2\,|\,)$$
- $w_1 = 0$ or $w_2 = 0 \implies S(\,f\,,\,\cdot\,)$ "kinky"
- $[\,|\cdot|\,]'(\,0\,)$ undefined: -1? 1? 0?
- What if I choose arbitrarily?



- $x^1 = [\,3\,,\,2\,]$, $y^1 = 2$, $\mu = 10 \implies$
  $$f(\,w_1\,,\,w_2\,) = (\,3w_1 + 2w_2 - 2\,)^2 + 10(\,|\,w_1\,| + |\,w_2\,|\,)$$
- $w_1 = 0$ or $w_2 = 0 \implies S(\,f\,,\,\cdot\,)$ "kinky"
- $[\,|\cdot|\,]'(\,0\,)$ undefined: -1? 1? 0?
- What if I choose arbitrarily?

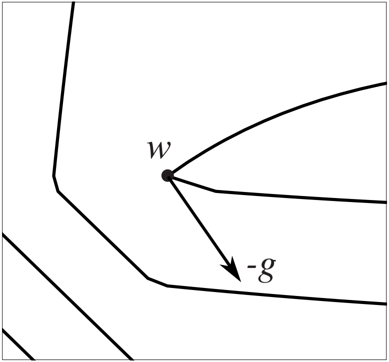


- $x^1 = [\,3\,,\,2\,]$, $y^1 = 2$, $\mu = 10 \implies$
  $$f(\,w_1\,,\,w_2\,) = (\,3w_1 + 2w_2 - 2\,)^2 + 10(\,|\,w_1\,| + |\,w_2\,|\,)$$
- $w_1 = 0$ or $w_2 = 0 \implies S(\,f\,,\,\cdot\,)$ "kinky"
- $[\,|\cdot|\,]'(\,0\,)$ undefined: -1? 1? 0?
- What if I choose arbitrarily?

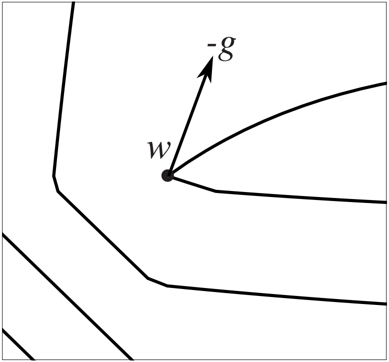


- $\exists\,(-)g \approx \nabla f(\,w\,)$ "pointing inside $S(\,f\,,\,f(\,w\,)\,)$" $\equiv$ descent direction

- $x^1 = [\,3\,,\,2\,]$, $y^1 = 2$, $\mu = 10 \implies$
  $f(\,w_1\,,\,w_2\,) = (\,3w_1 + 2w_2 - 2\,)^2$
  $+ 10(\,|\,w_1\,| + |\,w_2\,|\,)$

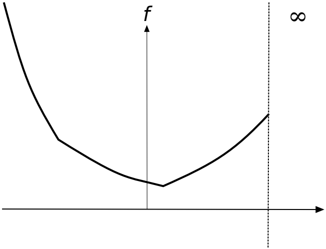


- $w_1 = 0$ or $w_2 = 0 \implies S(\,f\,,\,\cdot\,)$ "kinky"
- $[\,|\cdot|\,]'(\,0\,)$ undefined: -1? 1? 0?
- What if I choose arbitrarily?
- $\exists\,(-)g \approx \nabla f(\,w\,)$ "pointing inside $S(\,f\,,\,f(\,w\,)\,)$" $\equiv$ descent direction
- But many others "point outside"

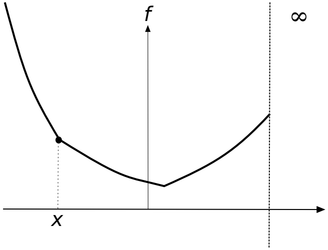


- $x^1 = [\,3\,,\,2\,]$, $y^1 = 2$, $\mu = 10 \implies$
  $f(\,w_1\,,\,w_2\,) = (\,3w_1 + 2w_2 - 2\,)^2$
  $+\,10(\,|\,w_1\,| + |\,w_2\,|\,)$
- $w_1 = 0$ or $w_2 = 0 \implies S(\,f\,,\,\cdot\,)$ "kinky"
- $[\,|\cdot|\,]'(\,0\,)$ undefined: -1? 1? 0?
- What if I choose arbitrarily?
- $\exists\,(-)g \approx \nabla f(\,w\,)$ "pointing inside $S(\,f\,,\,f(\,w\,)\,)$" $\equiv$ descent direction
- But many others "point outside" $\equiv$ no descent direction
- A descent method with $d^i = -g^i \implies \alpha^i = 0 \implies w^{i+1} = w^i$ ↯
- Methods need not be of descent + f is convex, and this can be exploited

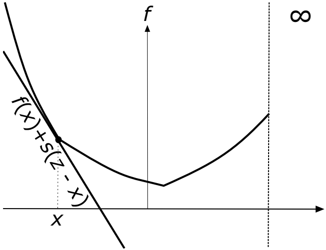
Outline



- s subgradient of f



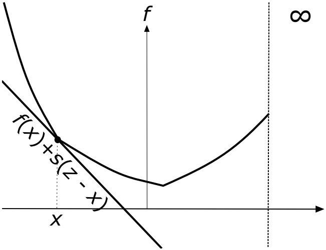
- s subgradient of f at x:



- s subgradient of f at x:

$$f(z) \geq f(x) + \langle s , z - x \rangle \quad \forall z \in \mathbb{R}^n$$

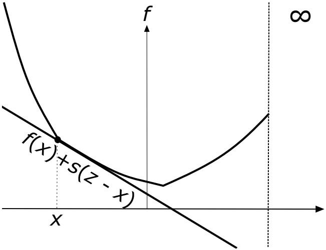
- No lack of first-order information, rather



- s subgradient of f at x:

$$f(z) \geq f(x) + \langle s, z - x \rangle \quad \forall z \in \mathbb{R}^n$$

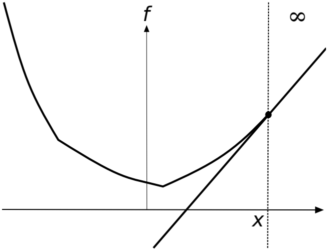
- No lack of first-order information, rather too much of it



- s subgradient of f at x:

$$f(z) \geq f(x) + \langle s , z - x \rangle \quad \forall z \in \mathbb{R}^n$$

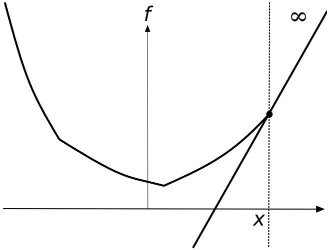
- No lack of first-order information, rather too much of it



- s subgradient of f at x:

$$f(z) \geq f(x) + \langle s, z - x \rangle \quad \forall z \in \mathbb{R}^n$$

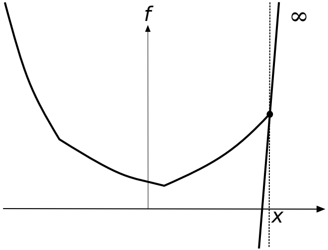
- No lack of first-order information, rather too much of it
- for x "on the border" of $dom(f)$,



- s subgradient of f at x:

$$f(z) \geq f(x) + \langle s, z - x \rangle \quad \forall z \in \mathbb{R}^n$$

- No lack of first-order information, rather too much of it
- for x "on the border" of $dom(f)$,

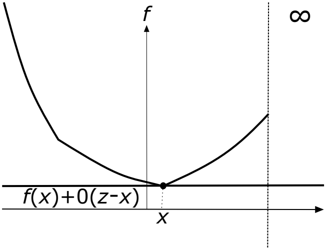


- s subgradient of f at x:

  $$f(z) \geq f(x) + \langle s, z - x \rangle \quad \forall z \in \mathbb{R}^n$$

- No lack of first-order information, rather too much of it

- for x "on the border" of $dom(f)$, $\| s \| \to \infty$



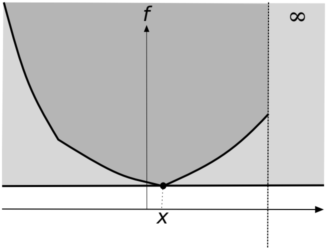
- s subgradient of f at x:

  $$f(z) \geq f(x) + \langle s, z - x \rangle \quad \forall z \in \mathbb{R}^n$$

- No lack of first-order information, rather too much of it

- for x “on the border” of $dom(f)$, $\| s \| \to \infty$

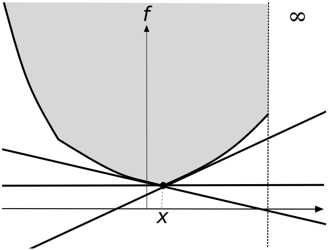
- $s = 0$



- s subgradient of f at x:

$$f(z) \geq f(x) + \langle s, z - x \rangle \quad \forall z \in \mathbb{R}^n$$

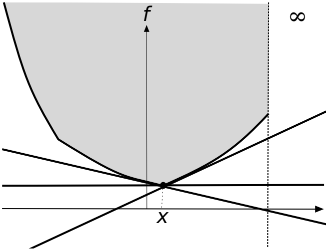
- No lack of first-order information, rather too much of it
- for x "on the border" of $dom(f)$, $\| s \| \to \infty$
- $s = 0 \implies x$ local $\equiv$ global minimum



- s subgradient of f at x:

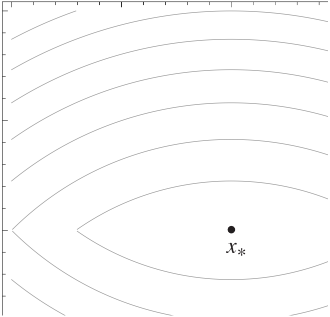
$$f(z) \geq f(x) + \langle s, z - x \rangle \quad \forall z \in \mathbb{R}^n$$

- No lack of first-order information, rather too much of it
- for x "on the border" of $dom(f)$, $\| s \| \to \infty$
- $s = 0 \implies x$ local $\equiv$ global minimum
- However, there can be (∞-ly) many $s \neq 0$ at a local $\equiv$ global minimum

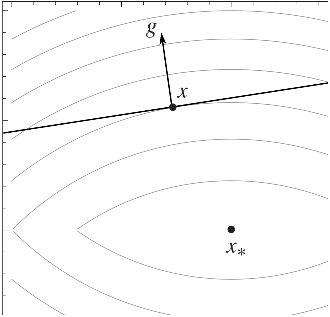


- s subgradient of f at x:
  $$f(z) \geq f(x) + \langle s, z - x \rangle \quad \forall z \in \mathbb{R}^n$$
- No lack of first-order information, rather too much of it
- for x "on the border" of $dom(f)$, $\| s \| \to \infty$
- $s = 0 \implies x$ local $\equiv$ global minimum
- However, there can be (∞-ly) many $s \neq 0$ at a local $\equiv$ global minimum
- $\partial f(x) = \{ s \in \mathbb{R}^n : s$ is a subgradient at $x \} \equiv$ subdifferential (a set)
- $\partial f(x) = \{ \nabla f(x) \} \iff f$ differentiable at x
- $\frac{\partial f}{\partial d}(x) \geq \langle s, d \rangle \; \forall s \in \partial f(x) \implies$

  d is a descent direction $\iff \langle s, d \rangle < 0 \;\; \forall s \in \partial f(x)$
- $s_* = -\mathrm{argmin}\{ \| s \| : s \in \partial f(x) \} =$ steepest descent direction
- x global minimum $\iff 0 \in \partial f(x)$

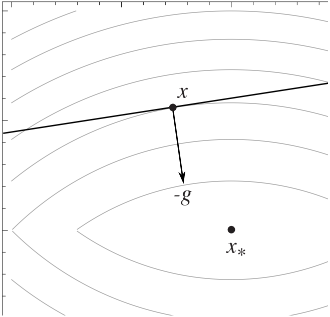


- $f(x_1, x_2) = \max\{x_1^2 + (x_2 - 1)^2, x_1^2 + (x_2 + 1)^2\}$
  convex, nondifferentiable, $x_* = [0, 0]$



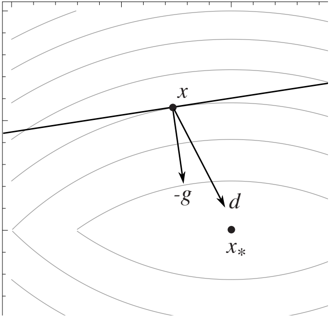
- $f(x_1, x_2) = \max\{x_1^2+(x_2-1)^2, x_1^2+(x_2+1)^2\}$
  convex, nondifferentiable, $x_* = [0, 0]$

- if $\partial f(x) = \{g = \nabla f(x)\}$, $g \perp S(f, f(x))$

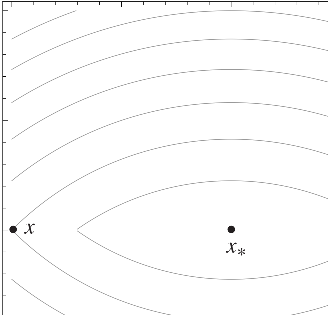


- $f(x_1, x_2) = \max\{x_1^2 + (x_2 - 1)^2, x_1^2 + (x_2 + 1)^2\}$
  convex, nondifferentiable, $x_* = [0, 0]$

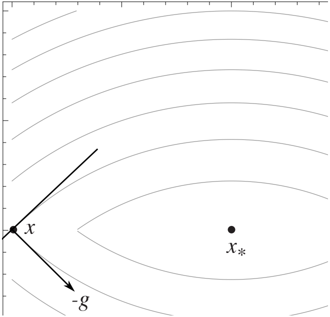
- if $\partial f(x) = \{g = \nabla f(x)\}$, $g \perp S(f, f(x))$
  i.e., $-g$ "points towards x_*"



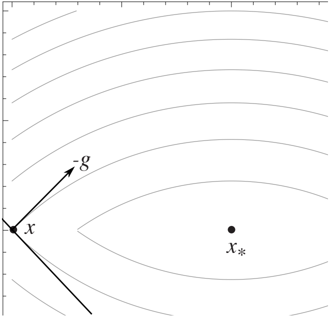
- $f(x_1, x_2) = \max\{ x_1^2 + (x_2 - 1)^2 , x_1^2 + (x_2 + 1)^2 \}$
  convex, nondifferentiable, $x_* = [0, 0]$
- if $\partial f(x) = \{ g = \nabla f(x) \}$, $g \perp S(f, f(x))$
  i.e., $-g$ "points towards x_*"
- d s.t. $\langle g , d \rangle < 0 \equiv$ descent direction



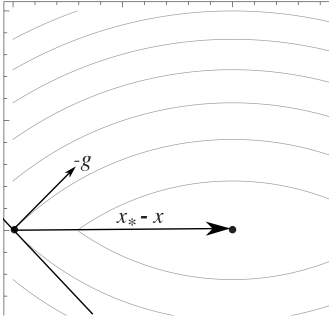
- $f(x_1, x_2) = \max\{ x_1^2 + (x_2 - 1)^2 , x_1^2 + (x_2 + 1)^2 \}$
  convex, nondifferentiable, $x_* = [0, 0]$
- if $\partial f(x) = \{ g = \nabla f(x) \}$, $g \perp S(f, f(x))$
  i.e., $-g$ "points towards x_*"
- d s.t. $\langle g , d \rangle < 0 \equiv$ descent direction
- But if f is nondifferentiable in x



- $f(x_1, x_2) = \max\{ x_1^2 + (x_2 - 1)^2 , x_1^2 + (x_2 + 1)^2 \}$ convex, nondifferentiable, $x_* = [0, 0]$
- if $\partial f(x) = \{ g = \nabla f(x) \}$, $g \perp S(f, f(x))$ i.e., $-g$ "points towards x_*"
- d s.t. $\langle g, d \rangle < 0 \equiv$ descent direction
- But if f is nondifferentiable in x there are many different $(-)g$



- $f(x_1, x_2) = \max\{ x_1^2 + (x_2 - 1)^2 , x_1^2 + (x_2 + 1)^2 \}$ convex, nondifferentiable, $x_* = [0, 0]$
- if $\partial f(x) = \{ g = \nabla f(x) \}$, $g \perp S(f, f(x))$ i.e., $-g$ "points towards x_*"
- d s.t. $\langle g , d \rangle < 0 \;\equiv\;$ descent direction
- But if f is nondifferentiable in x there are many different $(-)g$
- All of them are "$\perp S(f, f(x))$" ($x =$ "kink point")
- Not all of them are descent directions



- $f(x_1, x_2) = \max\{ x_1^2 + (x_2 - 1)^2 , x_1^2 + (x_2 + 1)^2 \}$ convex, nondifferentiable, $x_* = [0, 0]$
- if $\partial f(x) = \{ g = \nabla f(x) \}$, $g \perp S(f, f(x))$ i.e., $-g$ "points towards x_*"
- d s.t. $\langle g , d \rangle < 0 \equiv$ descent direction
- But if f is nondifferentiable in x there are many different $(-)g$
- All of them are "$\perp S(f, f(x))$" ($x =$ "kink point")
- Not all of them are descent directions
- However, any $(-)$ subgradient "points towards x_*":
  $$f(x_*) \geq f(x) + \langle g , x_* - x \rangle \Longrightarrow \langle g , x_* - x \rangle \leq f(x_*) - f(x) \leq 0$$
- Enough for gradient-type approaches (but don'h hold your breath on efficiency)

- $f : \mathbb{R}^n \to \mathbb{R} \implies \partial f(x)$ a compact convex set $\forall\, x$ [6, Def. VI.1.1.4]
- As with ∇f, $\exists$ rules for "computing" ∂f, some look familiar
  - i $\alpha, \beta \in \mathbb{R}_+ \implies \partial[\alpha f + \beta g](x) = \alpha \partial f(x) + \beta \partial g(x)$
  - ii $\partial[f(Ax + b)] = A^T [\partial f](Ax + b)$ (pre-composition with linear function)
  - iii $g : \mathbb{R} \to \mathbb{R}$ increasing $\implies \partial[g(f(x))] = [\partial g](f(x))[\partial f](x)$
    (post-composition with increasing convex function, "chain rule")
  - iv $f(x) = \max\{f_1(x), \ldots, f_m(x)\}$, $I(x) = \{i : f_i(x) = f(x)\} \implies$
    $\partial f(x) = \mathrm{conv}(\cup_{i \in I(x)} \partial f_i(x))$ $\approx$ extends to ∞-ly many [6, §VI.4.4]
  - v $g(x, y) : \mathbb{R}^{n+m} \to \mathbb{R}$, $f(x) = \inf\{g(x, y) : y \in \mathbb{R}^m\} \implies$
    $\partial f(x) = \{s \in \mathbb{R}^n : (s, 0) \in \partial g(x, y)\}$ (partial minimization)
  - vi $f(x) = \inf\{f_1(x_1) + f_2(x_2) : x_1 + x_2 = x\}$ (infimal convolution) $\implies$
    $\partial f(x) = \partial f_1(x_1) \cap \partial f_2(x_1)$ where $x_1 + x_2 = x$ and $f(x) = f_1(x_1) + f_2(x_2)$
- Some more complicated ones (value function, perspective, ... [6, §VI.4.5])

Exercise: prove $\supseteq$ in i. "from prime principles"

Exercise: compute $\partial f(x)$ for $f(x) = |x|$

Outline

(Convex) Nondifferentiable optimization is hard(er)

- Nondifferentiable optimization is orders of magnitude slower [1, Th. 3.13]

- Nondifferentiable optimization is orders of magnitude slower [1, Th. 3.13]

$f \in C^1$	τ-convex	L-smooth	$O(\log(1/\varepsilon))$

- Nondifferentiable optimization is orders of magnitude slower [1, Th. 3.13]

$f \in C^1$	τ-convex	L-smooth	$O(\log(1/\varepsilon))$
$f \notin C^1$	τ-convex	L-Lipschitz	$\Omega(L^2/\varepsilon)$

- Nondifferentiable optimization is orders of magnitude slower [1, Th. 3.13]

$f \in C^1$	τ-convex	L-smooth	$O(\log(1/\varepsilon))$
$f \notin C^1$	τ-convex	L-Lipschitz	$\Omega(L^2/\varepsilon)$
$f \in C^1$	convex	L-smooth	$O(1/\sqrt{\varepsilon})$

(Convex) Nondifferentiable optimization is hard(er)

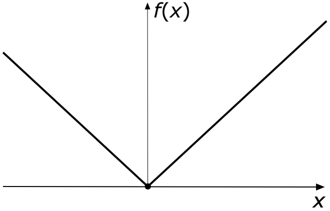
- Nondifferentiable optimization is orders of magnitude slower [1, Th. 3.13]

$f \in C^1$	τ-convex	L-smooth	$O(\log(1/\varepsilon))$
$f \notin C^1$	τ-convex	L-Lipschitz	$\Omega(L^2/\varepsilon)$
$f \in C^1$	convex	L-smooth	$O(1/\sqrt{\varepsilon})$
$f \notin C^1$	convex	L-Lipschitz	$\Omega(L/\varepsilon^2)$

- Nondifferentiable optimization is orders of magnitude slower [1, Th. 3.13]

$f \in C^1$	τ-convex	L-smooth	$O(\log(1/\varepsilon))$
$f \notin C^1$	τ-convex	L-Lipschitz	$\Omega(L^2/\varepsilon)$
$f \in C^1$	convex	L-smooth	$O(1/\sqrt{\varepsilon})$
$f \notin C^1$	convex	L-Lipschitz	$\Omega(L/\varepsilon^2)$

- Furthermore, Fixed Step "cannot work" for $f \notin C^1$

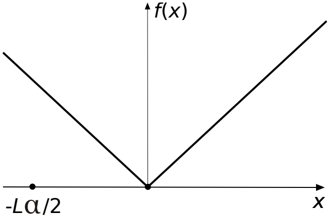


- $f(x) = L|x|$,

- Nondifferentiable optimization is orders of magnitude slower [1, Th. 3.13]

$f \in C^1$	τ-convex	L-smooth	$O(\log(1/\varepsilon))$
$f \notin C^1$	τ-convex	L-Lipschitz	$\Omega(L^2/\varepsilon)$
$f \in C^1$	convex	L-smooth	$O(1/\sqrt{\varepsilon})$
$f \notin C^1$	convex	L-Lipschitz	$\Omega(L/\varepsilon^2)$

- Furthermore, Fixed Step "cannot work" for $f \notin C^1$

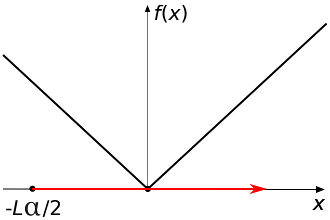


- $f(x) = L|x|$, $x^0 = -\alpha L/2$

- Nondifferentiable optimization is orders of magnitude slower [1, Th. 3.13]

$f \in C^1$	τ-convex	L-smooth	$O(\log(1/\varepsilon))$
$f \notin C^1$	τ-convex	L-Lipschitz	$\Omega(L^2/\varepsilon)$
$f \in C^1$	convex	L-smooth	$O(1/\sqrt{\varepsilon})$
$f \notin C^1$	convex	L-Lipschitz	$\Omega(L/\varepsilon^2)$

- Furthermore, Fixed Step "cannot work" for $f \notin C^1$

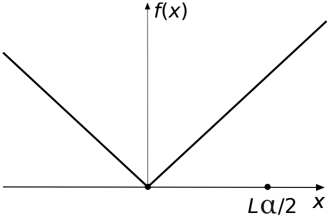


- $f(x) = L|x|$, $x^0 = -\alpha L/2$
- $g^1 = -L$,

- Nondifferentiable optimization is orders of magnitude slower [1, Th. 3.13]

$f \in C^1$	τ-convex	L-smooth	$O(\log(1/\varepsilon))$
$f \notin C^1$	τ-convex	L-Lipschitz	$\Omega(L^2/\varepsilon)$
$f \in C^1$	convex	L-smooth	$O(1/\sqrt{\varepsilon})$
$f \notin C^1$	convex	L-Lipschitz	$\Omega(L/\varepsilon^2)$

- Furthermore, Fixed Step "cannot work" for $f \notin C^1$

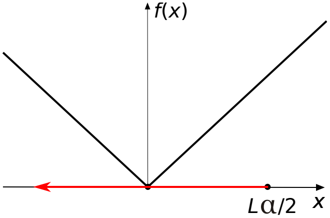


- $f(x) = L|x|$, $x^0 = -\alpha L/2$
- $g^1 = -L$, $x^1 = x^0 - \alpha g^1 = \alpha L/2$

- Nondifferentiable optimization is orders of magnitude slower [1, Th. 3.13]

$f \in C^1$	τ-convex	L-smooth	$O(\log(1/\varepsilon))$
$f \notin C^1$	τ-convex	L-Lipschitz	$\Omega(L^2/\varepsilon)$
$f \in C^1$	convex	L-smooth	$O(1/\sqrt{\varepsilon})$
$f \notin C^1$	convex	L-Lipschitz	$\Omega(L/\varepsilon^2)$

- Furthermore, Fixed Step "cannot work" for $f \notin C^1$

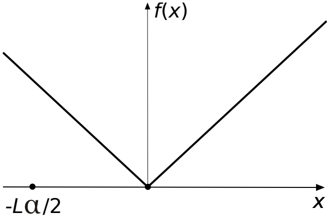


- $f(x) = L|x|$, $x^0 = -\alpha L/2$
- $g^1 = -L$, $x^1 = x^0 - \alpha g^1 = \alpha L/2$
- $g^2 = L$, $x^2 = x^1 - \alpha g^2 =$

- Nondifferentiable optimization is orders of magnitude slower [1, Th. 3.13]

$f \in C^1$	τ-convex	L-smooth	$O(\log(1/\varepsilon))$
$f \notin C^1$	τ-convex	L-Lipschitz	$\Omega(L^2/\varepsilon)$
$f \in C^1$	convex	L-smooth	$O(1/\sqrt{\varepsilon})$
$f \notin C^1$	convex	L-Lipschitz	$\Omega(L/\varepsilon^2)$

- Furthermore, Fixed Step "cannot work" for $f \notin C^1$

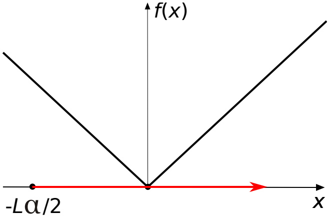


- $f(x) = L|x|$, $x^0 = -\alpha L/2$
- $g^1 = -L$, $x^1 = x^0 - \alpha g^1 = \alpha L/2$
- $g^2 = L$, $x^2 = x^1 - \alpha g^2 = -\alpha L/2 = x^0$

- Nondifferentiable optimization is orders of magnitude slower [1, Th. 3.13]

$f \in C^1$	τ-convex	L-smooth	$O(\log(1/\varepsilon))$
$f \notin C^1$	τ-convex	L-Lipschitz	$\Omega(L^2/\varepsilon)$
$f \in C^1$	convex	L-smooth	$O(1/\sqrt{\varepsilon})$
$f \notin C^1$	convex	L-Lipschitz	$\Omega(L/\varepsilon^2)$

- Furthermore, Fixed Step "cannot work" for $f \notin C^1$

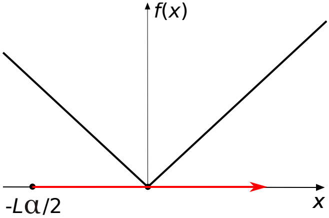


  - $f(x) = L|x|$, $x^0 = -\alpha L/2$
  - $g^1 = -L$, $x^1 = x^0 - \alpha g^1 = \alpha L/2$
  - $g^2 = L$, $x^2 = x^1 - \alpha g^2 = -\alpha L/2 = x^0$
  - $g^3 = -L$,

- Nondifferentiable optimization is orders of magnitude slower [1, Th. 3.13]

$f \in C^1$	τ-convex	L-smooth	$O(\log(1/\varepsilon))$
$f \notin C^1$	τ-convex	L-Lipschitz	$\Omega(L^2/\varepsilon)$
$f \in C^1$	convex	L-smooth	$O(1/\sqrt{\varepsilon})$
$f \notin C^1$	convex	L-Lipschitz	$\Omega(L/\varepsilon^2)$

- Furthermore, Fixed Step "cannot work" for $f \notin C^1$

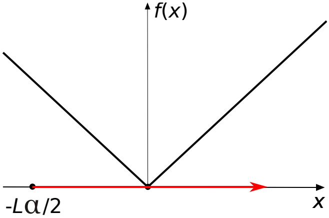


- $f(x) = L|x|$, $x^0 = -\alpha L/2$
- $g^1 = -L$, $x^1 = x^0 - \alpha g^1 = \alpha L/2$
- $g^2 = L$, $x^2 = x^1 - \alpha g^2 = -\alpha L/2 = x^0$
- $g^3 = -L$, $x^1 = x^0 - \alpha g^1 = \dots$

- Nondifferentiable optimization is orders of magnitude slower [1, Th. 3.13]

$f \in C^1$	τ-convex	L-smooth	$O(\log(1/\varepsilon))$
$f \notin C^1$	τ-convex	L-Lipschitz	$\Omega(L^2/\varepsilon)$
$f \in C^1$	convex	L-smooth	$O(1/\sqrt{\varepsilon})$
$f \notin C^1$	convex	L-Lipschitz	$\Omega(L/\varepsilon^2)$

- Furthermore, Fixed Step "cannot work" for $f \notin C^1$



- $f(x) = L|x|$, $x^0 = -\alpha L/2$
- $g^1 = -L$, $x^1 = x^0 - \alpha g^1 = \alpha L/2$
- $g^2 = L$, $x^2 = x^1 - \alpha g^2 = -\alpha L/2 = x^0$
- $g^3 = -L$, $x^1 = x^0 - \alpha g^1 = \ldots$

- $f_{\text{best}} - f_* = L^2\alpha/2$, $O(L)$ for $\alpha = 1/L$, and the algorithm cycles forever

Why (convex) nondifferentiable optimization is hard(er)

- $f \in C^1$, the gradient is unique, $d = -\nabla f(x)$
  - $f(x + \alpha d) < f(x)$ for all (small enough) $\alpha \geq 0$
  - $\| d \|$ is a two-sided proxy of $A(x)$: $\| d \|$ "small" $\Longleftrightarrow$ $f(x)$ "close" to f_*
    $\equiv$ "$\| d \| \leq \varepsilon$" effective stopping criterion
  - can use Fixed Step since $\| x^{i+1} - x^i \| \to 0$ automatically:
    $\| d^i \| \to 0$ even if $\alpha^i \geq \bar{\alpha} > 0$

- $f \notin C^1$, there can be many different subgradients, $d = -[\, g \in \partial f(x)\,]$
  any one of them (can't choose, the oracle does for you)
  - $f(x + \alpha d)$ may be $\geq f(x)$ for all α
  - $\| d \|$ is a one-sided proxy of $A(x)$:
    - $\| d \|$ "small" $\Longrightarrow$ $f(x)$ "close" to f_*
    - $f(x)$ "close" to f_* $\not\Longrightarrow$ $\| d \|$ "small"

    $\equiv$ "$\| d \| \leq \varepsilon$" ineffective stopping criterion (almost never happens)
  - can't use Fixed Step since $\| d \|$ can be "big" even if $x = x_*$:
    to ensure $\| x^{i+1} - x^i \| \to 0$ one has to force $\alpha^i \to 0$ (but not too fast)

Outline

- Any $(-)$ subgradient "points towards x_*"

  $\Longrightarrow$ an appropriate step along $-g$ brings closer to x_*

  $\Longrightarrow$ $x^{i+1} = x^i - \alpha^i g^i$ for makes sense with the right α^i

- Fundamental relationship: $\| x^{i+1} - x_* \|^2 = \| x^i - \alpha^i g^i - x_* \|^2 =$

$$= \| x^i - x_* \|^2 + 2\alpha^i \langle g^i , x_* - x^i \rangle + (\alpha^i)^2 \| g^i \|^2$$

$$\leq \| x^i - x_* \|^2 + \underbrace{2\alpha^i (f_* - f(x_i))}_{[\ < 0\ ]} + \underbrace{(\alpha^i)^2 \| g^i \|^2}_{[\ > 0\ ]}$$

- As $\alpha \searrow 0$ (short step), blue term dominates $\Longrightarrow$ x^{i+1} closer to x^* than x^i

Exercise: check / justify the previous two points

- Any $(-)$ subgradient "points towards x_*"

  $\Longrightarrow$ an appropriate step along $-g$ brings closer to x_*

  $\Longrightarrow$ $x^{i+1} = x^i - \alpha^i g^i$ for makes sense with the right α^i

- Fundamental relationship: $\| x^{i+1} - x_* \|^2 = \| x^i - \alpha^i g^i - x_* \|^2 =$

$$= \| x^i - x_* \|^2 + 2\alpha^i \langle g^i , x_* - x^i \rangle + (\alpha^i)^2 \| g^i \|^2$$

$$\leq \| x^i - x_* \|^2 + \underbrace{2\alpha^i(f_* - f(x_i))}_{[\ < 0\ ]} + \underbrace{(\alpha^i)^2 \| g^i \|^2}_{[\ > 0\ ]}$$

- As $\alpha \searrow 0$ (short step), blue term dominates $\Longrightarrow$ x^{i+1} closer to x^* than x^i

Exercise: check / justify the previous two points

- Short but not too short = "Diminishing–Square Summable" stepsize:

  (DSS) $\quad \sum_{i=1}^{\infty} \alpha^i = \infty \ \wedge \ \sum_{i=1}^{\infty} (\alpha^i)^2 < \infty$

  "$\alpha^i \searrow 0$ but not fast enough that the series converges" ($\alpha^i = 1 / i$)

- DSS just "works": $\forall \varepsilon > 0 \ \exists\, i$ s.t. $f^i - f_* \leq \varepsilon$

- Any $(-)$ subgradient "points towards x_*"

  $\Longrightarrow$ an appropriate step along $-g$ brings closer to x_*

  $\Longrightarrow$ $x^{i+1} = x^i - \alpha^i g^i$ for makes sense with the right α^i

- Fundamental relationship: $\| x^{i+1} - x_* \|^2 = \| x^i - \alpha^i g^i - x_* \|^2 =$

$$= \| x^i - x_* \|^2 + 2\alpha^i \langle g^i , x_* - x^i \rangle + (\alpha^i)^2 \| g^i \|^2$$

$$\leq \| x^i - x_* \|^2 + \underbrace{2\alpha^i (f_* - f(x_i))}_{[\ < 0\ ]} + \underbrace{(\alpha^i)^2 \| g^i \|^2}_{[\ > 0\ ]}$$

- As $\alpha \searrow 0$ (short step), blue term dominates $\Longrightarrow$ x^{i+1} closer to x^* than x^i

Exercise: check / justify the previous two points

- Short but not too short = "Diminishing–Square Summable" stepsize:

  (DSS) $\sum_{i=1}^{\infty} \alpha^i = \infty \ \wedge \ \sum_{i=1}^{\infty} (\alpha^i)^2 < \infty$

  "$\alpha^i \searrow 0$ but not fast enough that the series converges" ($\alpha^i = 1 / i$)

- DSS just "works": $\forall \varepsilon > 0 \ \exists \, i$ s.t. $f^i - f_* \leq \varepsilon$ but not $\forall \, h \geq i$, not monotone

- Incredibly robust result: α^i chosen a priori, $f(x^i)$ not even used (only g^i)

- Need $\| g^i \| \leq L \impliedby f$ L-c
- Can do without, e.g., $\| x^i \| \leq M < \infty$ enough, and bounding strategies $\exists$ [8]
- DSS "works": by contradiction, $f(x^i) - f_* \geq \delta / 2 > 0 \;\; \forall i$
- $\| x^{i+1} - x_* \|^2 \leq \| x^i - x_* \|^2 + 2\alpha^i(f_* - f(x_i)) + (\alpha^i)^2 \| g^i \|^2$

  $\leq \| x^i - x_* \|^2 - \delta\alpha^i + L^2(\alpha^i)^2$ [induction] $\implies$

  $\| x^{k+1} - x_* \|^2 \leq \| x^1 - x_* \|^2 + \left[v^k = -\delta \sum_{i=1}^k \alpha^i + L^2 \sum_{i=1}^k (\alpha^i)^2 \right]$
- $\sum_{i=1}^\infty \alpha^i = \infty$ and $\sum_{i=1}^\infty (\alpha^i)^2 < \infty \implies v^k \to -\infty$ as $k \to \infty \implies$

  $\exists\, k$ s.t. $0 \leq \| x^{k+1} - x_* \|^2 \leq \| x^1 - x_* \|^2 + v^k < 0 \quad \lightning$
- Proves that $\exists\, x^i$ arbitrarily close to x_*, but x^{i+1} could be very far
- α that was "good" at iteration i can be "very bad" at $i+1$
- No control on individual stepsizes, only on "long term average"

- Practical convergence speed of DSS abysmal, cannot use it
- Look again: $\| x^{i+1} - x_* \|^2 \leq \| x^i - x_* \|^2 + 2\alpha^i (f_* - f^i) + (\alpha^i)^2 \| g^i \|^2$
  $\Longrightarrow$ if we knew f_* we could estimate α^i ...

- Practical convergence speed of DSS abysmal, cannot use it
- Look again: $\| x^{i+1} - x_* \|^2 \leq \| x^i - x_* \|^2 + 2\alpha^i (f_* - f^i) + (\alpha^i)^2 \| g^i \|^2$
  $\Longrightarrow$ if we knew f_* we could estimate α^i … let's pretend we do
- Recall: $\phi(\alpha) = a\alpha^2 + b\alpha$, $a > 0 \implies \alpha_* = \text{argmin}\{ \phi(\alpha) \} = -b / 2a$
  $b < 0 \implies \phi(\alpha) < 0 \; \forall \alpha \in (0 , 2\alpha_*)$
- $a = \| g^i \|^2$, $b = 2(f_* - f^i) \implies \alpha^i_* = (f^i - f_*) / \| g^i \|^2$ $[\geq 0]$
- Polyak stepsize (PSS): $\alpha^i \in (0 , 2\alpha^i_*) \implies \| x^{i+1} - x_* \|^2 < \| x^i - x_* \|^2$
- Vastly better in practice as far as it can go = not much:
  $$\min\{ f(x^h) \,:\, h \leq i \} - f_* \leq L \| x^1 - x_* \| / \sqrt{i} \quad \Longrightarrow \quad O(1 / \varepsilon^2)$$
- $\varepsilon = 1\text{e}{-3} \to \varepsilon = 1\text{e}{-4} \implies 100\times$ iterations $\implies \varepsilon < 1\text{e}{-4}$ impractical

- (PSS) $\implies \| x^{i+1} - x_* \| < \| x^i - x_* \| \implies \| x^i - x_* \| < \| x^1 - x_* \| < \infty \; \forall \, i$
  $\implies \| g^i \| \leq L$ [6, Proposition VI.6.2.2] (or just ask f L-c)

- $\alpha^i = \alpha^*_i \implies (f^i - f_*)^2 / \| g^i \|^2 \leq \| x^i - x_* \|^2 - \| x^{i+1} - x_* \|^2$ (**check**)

- $\bar{f}^i = \min\{ f^h \, : \, h \leq i \}$ record value up to iteration i

  $$\implies \frac{(\bar{f}^i - f_*)^2}{L^2} \leq \frac{(f(x^i) - f_*)^2}{\| g^i \|^2} \leq \| x^i - x_* \|^2 - \| x^{i+1} - x_* \|^2$$

- Sum for $i = 1, \ldots, k$: intermediate terms cancel out $\implies$

  $$k \frac{(\bar{f}^k - f_*)^2}{L^2} \leq \| x^1 - x_* \|^2 - \| x^{k+1} - x_* \|^2 \leq \| x^1 - x_* \|^2$$

  $$\implies \quad \bar{f}^k - f_* \leq L \| x^1 - x_* \|^2 / \sqrt{k} \quad \implies \quad O(1 / \varepsilon^2)$$

- "Good news": Polyak would be optimal if we knew f_*, which we don't

- "If you don't know it estimate it, but be ready to revise your estimate"

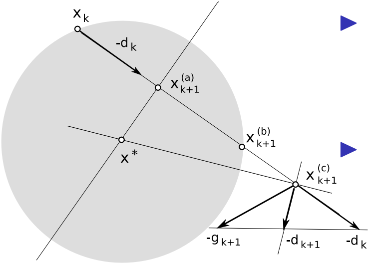
procedure $x = SGPTL\ (f , x , i_{max} , \beta , \delta_0 , R , \rho)$
  $r \leftarrow 0$; $\delta \leftarrow \delta_0$; $f_{ref} \leftarrow \bar{f} \leftarrow f(x)$; $i \leftarrow 1$;
  while($i < i_{max}$) **do**
    $g \in \partial f(x)$; $\alpha \leftarrow \beta(f(x) - (f_{ref} - \delta)) / \| g \|^2$; $x \leftarrow x - \alpha g$;
    if($f(x) \leq f_{ref} - \delta / 2$) **then** { $f_{ref} \leftarrow \bar{f}$; $r \leftarrow 0$; }
      else if($r > R$) **then** { $\delta \leftarrow \delta\rho$; $r \leftarrow 0$; }
        else $r \leftarrow r + \alpha \| g \|$;
    $\bar{f} \leftarrow \min\{ \bar{f} , f(x) \}$; $i \leftarrow i + 1$;

- reference value $f_{ref} -$ threshold δ = target level (ideally) $\approx f_*$
- "Good improvement" $\Longrightarrow f_{ref} \searrow \Longrightarrow$ target level $\searrow$
- "Too many steps without improvement" $\Longrightarrow \delta \searrow \Longrightarrow$ target level $\nearrow$
- (Too) many parameters: $\rho \in (0 , 1)$, $\beta \in (0 , 2)$, $\delta_0 > 0$ (??), $R > 0$ (???)
- $\{ \bar{f}^i \} \rightarrow f_*$ but no reasonable stopping criterion, just "stop after a while"
- Convergences, but slowly: can it be made any better?

- "Want a better direction? Use a better model!"
- There is no second-order information, but deflection is possible:

$$d^i = \gamma^i g^i + (\,1 - \gamma^i\,) d^{i-1} \;,\; x^{i+1} = x^i - \alpha^i d^i \approx \text{"conjugate subgradient"}$$

- If you want theoretical convergence some funny rules are needed



- Stepsize-restricted $\equiv$ deflection-first: Polyak

  $\alpha^i = \beta^i(\,f^i - f_*\,) \,/\, \|\, d^i \,\|^2 \quad \wedge \quad \beta^i \leq \gamma^i$

  "as deflection $\nearrow$, stepsize has to $\searrow$"

- Deflection-restricted $\equiv$ stepsize-first: (DSS) +

  $$\frac{\alpha^{i-1}\|\, d^{i-1} \,\|^2}{(\,f^i - f_*\,) + \alpha^{i-1}\|\, d^{i-1} \,\|^2} \leq \gamma^i$$

  "as $f(\,x^i\,) \to f_*$, deflection $\searrow$"

- In both cases, target level to replace f_* (many ugly parameters)
- $\gamma^i \in \operatorname{argmin}\{\, \|\, \gamma g^i + (\,1 - \gamma\,) d^{i-1} \,\|^2 \,:\, \gamma \in [\,0\,,\,1\,]\,\}$ (closed formula)
- Actually helps in practice, as far as it can go = not much

Outline

- "But the speed of light is a property of the space, master!"
  "OK, so let's just change the space!" [10]

- "But the speed of convergence is a property of the function, master!"
  "OK, so let's just (slightly) change the function!" [10]
- Requires $f(x) = \max\{ x^T Az \,:\, z \in Z \}$ convex (**check**), assumed "easy"
- $\bar{z}$ optimal for $x \implies A\bar{z} \in \partial f(x) \implies f \notin C^1$ (many different $\bar{z}$ can $\exists$)
- Smoothed $f_\mu(x) = \max\{ x^T Az - \mu \| z \|^2 / 2 \,:\, z \in Z \} \in C^1$ (hopefully easy)

Exercise: construct f_μ for $f(x) = |x|$ then plot it to see the "smoothing"

- "But the speed of convergence is a property of the function, master!"
  "OK, so let's just (slightly) change the function!" [10]
- Requires $f(x) = \max\{ x^T Az : z \in Z \}$ convex (**check**), assumed "easy"
- $\bar{z}$ optimal for $x \implies A\bar{z} \in \partial f(x) \implies f \notin C^1$ (many different $\bar{z}$ can $\exists$)
- Smoothed $f_\mu(x) = \max\{ x^T Az - \mu \| z \|^2 / 2 : z \in Z \} \in C^1$ (hopefully easy)

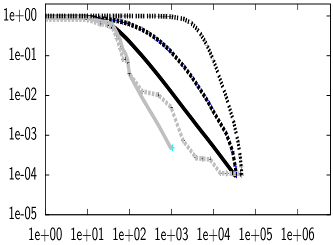
Exercise: construct f_μ for $f(x) = |x|$ then plot it to see the "smoothing"

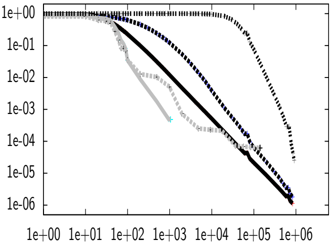
- Choose "small" $\mu = O(\varepsilon)$ + "fast" minimization of $f_\mu \implies$
  "only" $O(1/\varepsilon)$, "much better" than $O(1/\varepsilon^2)$
- Have to pry open the black box and change it, nontrivial (if at all possible)
- In theory parameter-free, but several caveats
- Convergence in practice non that great:
  constructed to optimize worst-case behaviour, gets what is constructed for

- Z convex and compact, "$+\phi(z)$" concave and "$+h(x) \in C^1$" allowed
- $f_\mu \to f$ as $\mu \to 0$ depending on $K = \max\{ \| z \|^2 / 2 \,:\, z \in Z \}$ (assuming $K < \infty$, easy, but computing it is not: convex maximization)
- $f_\mu(x) \leq f(x) \leq f_\mu(x) + \mu K$: as $\mu \searrow 0$, "argmin $\{ f_\mu(x) \} \to x_*$"
- f_μ L-smooth with $L = \| A \|^2 / \mu$ [9, Th. 1] ("less and less Lipschitz" as $\mu \searrow 0$)
- L-smooth is $O(LD / \sqrt{\varepsilon})$: $f(x^i) - f_* \leq 2LD^2 / i^2$ [1, Th. 3.19]
- Choose $\mu = \varepsilon / (2K) \implies L = 2\| A \|^2 K / \varepsilon$ to get
  - $f_\mu(x^i) \leq f(x^i) \leq f_\mu(x^i) + \varepsilon / 2 \implies f_{\mu,*} \leq f_* \leq f_{\mu,*} + \varepsilon / 2$
  - $f(x^i) - f_* \leq \varepsilon \impliedby f_\mu(x^i) + \varepsilon / 2 - f_{\mu,*} \leq \varepsilon \equiv f_\mu(x^i) - f_{\mu,*} \leq \varepsilon / 2$
  - $f_\mu(x^i) - f_{\mu,*} \leq 4\| A \|^2 KD^2 / (\varepsilon i^2) \leq \varepsilon / 2$

    $\equiv \; 4\| A \|^2 KD^2 / i^2 \leq \varepsilon^2 / 2 \; \equiv \; \sqrt{8K}\| A \| D / \varepsilon \leq i$
- Would be parameter-free but have to estimate K to choose μ (not easy)

- How does this work in practice? Consistently slowish
  $\approx$ superlinear in a doubly-logarithmic chart after a long flat leg





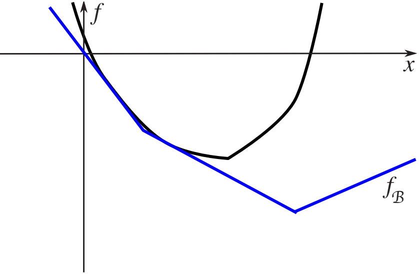
- Subgradients faster but flatline at $\varepsilon \approx$ 1e−4, smoothed does $\varepsilon =$ 1e−6 but it requires 1e+6 iterations to get there
- And with $\varepsilon =$ 1e−6 the flat leg is way longer
- ACCG does steps $1 \,/\, L_\mu = O(\,\mu\,) = O(\,\varepsilon\,)$, far too short at start
- Exploiting information about f_* helps (black solid line)

Exercise: how would you exploit information about f_*? (hint: $\varepsilon \Longrightarrow \varepsilon^i$)

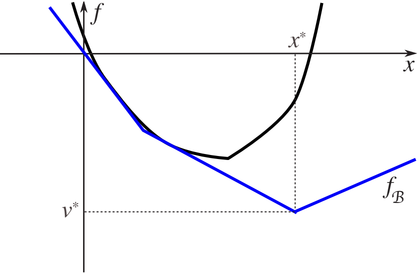
Outline

- "Want a better direction? Use a better model!"
- But $\nexists$ second-order information and first-order one is crap ...

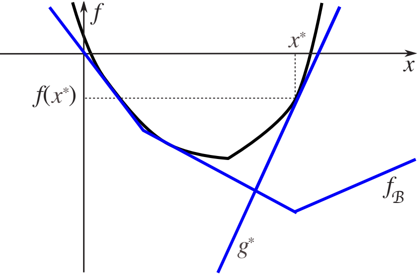
- "Want a better direction? Use a better model!"
- But $\nexists$ second-order information and first-order one is crap ... or is it?
- f convex $\Longrightarrow$ first-order information not so crap: globally valid
- $x \rightsquigarrow$ oracle $\rightsquigarrow f(x)$, $g \in \partial f(x) \Longrightarrow$ first-order model at x
  $l_{x,f(x),g}(z) = f(x) + \langle g , z - x \rangle \leq f(z) \ \forall z \in \mathbb{R}^n$ (not uniquely defined)
- What if I collect it all along the way and use it all?
- $\{ x^i \} \Longrightarrow$ bundle $\mathscr{B}^i = \{ (x^h , f^h = f(x^h) , g^h \in \partial f(x^h)) \}_{h<i}$
- $f^i_{\mathscr{B}}(x) = \max\{ l^h(x) = f^h + \langle g^h , x - x^h \rangle \ : \ (x^h , f^h , g^h) \in \mathscr{B}^i \} \leq f(x) \ \ \forall x$
  Cutting Plane (CP) model of f, "$(1+\varepsilon)$-order" model, convex
- $x^* \in \operatorname{argmin}\{ f_{\mathscr{B}}(x) \}$, $f_{\mathscr{B}}(x^*) \leq f_*$: use x^* as next iterate a-la Newton
- $f_{\mathscr{B}} \notin C^1$ but computing x^* a Linear Program $\Longrightarrow$ "easy" (if $\#\mathscr{B}$ "small")
  $\min\{ f^i_{\mathscr{B}}(x) \} = \min\{ v \ : \ v \geq f^h + \langle g^h , x - x^h \rangle \quad (x^h , f^h , g^h) \in \mathscr{B}^i \}$



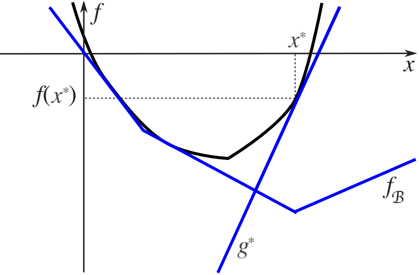
- $v^* = \min\{\, f_{\mathscr{B}}(\, x\,)\,\}$ master problem



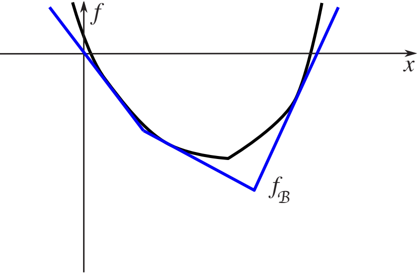
- $v^* = \min\{\, f_{\mathscr{B}}(\, x\,) \,\}$ master problem
- $x^* \in \mathrm{argmin}\{\, f_{\mathscr{B}}(\, x\,) \,\}$, $v^* = f_{\mathscr{B}}(\, x^*\,) \implies$



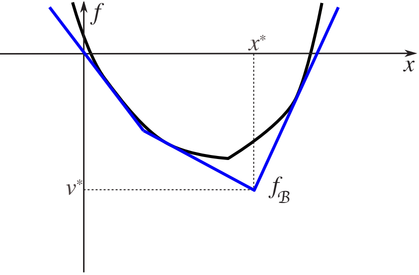
- $v^* = \min\{ f_{\mathscr{B}}(x) \}$ master problem
- $x^* \in \operatorname{argmin}\{ f_{\mathscr{B}}(x) \}$, $v^* = f_{\mathscr{B}}(x^*) \implies$ new $(x^* , f(x^*) , g^* \in \partial f(x^*))$



- $v^* = \min\{\, f_{\mathscr{B}}(\, x\,)\,\}$ master problem
- $x^* \in \text{argmin}\{\, f_{\mathscr{B}}(\, x\,)\,\}$, $v^* = f_{\mathscr{B}}(\, x^*\,) \implies$ new $(\, x^*\,,\, f(\, x^*\,)\,,\, g^* \in \partial f(\, x^*\,)\,)$
- $f(\, x^*\,) \leq v^* \implies x^*$ optimal **(check)**



- $v^* = \min\{ f_{\mathscr{B}}(x) \}$ master problem
- $x^* \in \operatorname{argmin}\{ f_{\mathscr{B}}(x) \}$, $v^* = f_{\mathscr{B}}(x^*) \implies$ new $(x^* , f(x^*), g^* \in \partial f(x^*))$
- $f(x^*) \leq v^* \implies x^*$ optimal **(check)**
- otherwise $\mathscr{B} \leftarrow \mathscr{B} \cup (x^* , f(x^*), g^*)$ $\implies f_{\mathscr{B}}$ becomes a “better” CP model



- $v^* = \min\{ f_{\mathscr{B}}(x) \}$ master problem
- $x^* \in \operatorname{argmin}\{ f_{\mathscr{B}}(x) \}$, $v^* = f_{\mathscr{B}}(x^*) \implies$ new $(x^*, f(x^*), g^* \in \partial f(x^*))$
- $f(x^*) \leq v^* \implies x^*$ optimal (**check**)
- otherwise $\mathscr{B} \leftarrow \mathscr{B} \cup (x^*, f(x^*), g^*)$ $\implies f_{\mathscr{B}}$ becomes a "better" CP model

- $\underline{f}^i = v^{*,i} = f^i_{\mathscr{B}}(x^{*,i}) \leq f_*$ model value, $\underline{f}^i \nearrow$ (**check**)
- $\bar{f}^i = \min\{ f^h : h \leq i \} \geq f_*$ record value up to iteration i, $\bar{f}^i \searrow$
- Under appropriate assumption $\{ \bar{f}^i \} \to f_* \leftarrow \{ \underline{f}^i \}$ [4, Th. 1]
- Practical stopping criterion $\bar{f}^i - \underline{f}^i \leq \varepsilon$, unlike subgradient algorithm; in fact, better than most other approaches so far, even for $f \in C^1$ (thanks convexity)
- But $\#\mathscr{B} \nearrow \infty \implies$ master problem cost per iteration $\nearrow \infty$
- Can be $O((1/\varepsilon)^{n/2})$ [7, Ex. 1.1.2], practical convergence often horrible

- Not surprising: every convex function is the max of its first-order models
  $f(x) = \max\{\, f(z) + \langle g \,,\, x - z \rangle \;:\; z \in \mathbb{R}^n \,,\; g \in \partial f(z) \,\}$

- Even better: can take any one $g \in \partial f(x)$ $\forall\, z \in \mathbb{R}^n$ [7, Th. XI.1.3.8]
  i.e., "fire any oracle for f" in all points of the space

- That is, $f = f_{\mathscr{B}}$ for (uncountably) ∞-ly large $\mathscr{B}$ $\equiv$ ∞-ly many x^i
  while we can only use finitely (in theory countably) many

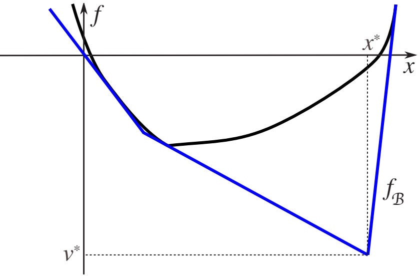
- But we don't need $f(x) = f_{\mathscr{B}}(x)$ $\forall\, x$, only close to x_*

- "Algorithmic proof": assume $x^{*,i} \in \mathcal{B}(x_* \,,\, \varepsilon)$ for any $\varepsilon > 0 \implies$ still works

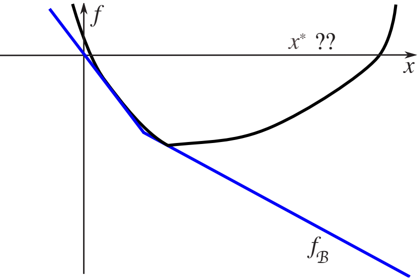
Exercise: prove the statement above

- min in the master problem (hopefully) focuses $\{\, x^i \,\}$ in some $\mathcal{B}(x_* \,,\, \varepsilon)$

- Unfortunately, not efficient at doing so, some help needed



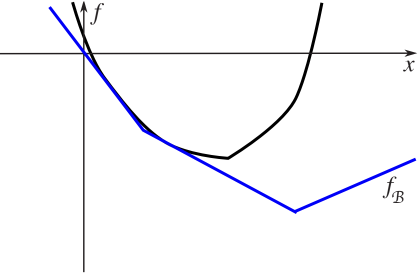
- x^* may be very far from x_*



- x^* may be very far from x_*
  ... up to infinitely far $\Longrightarrow$ assumptions or some dirty trick when $\mathscr{B}$ is small
- Iterates have no locality property: $\| x^{*,i+1} - x^{*,i} \|$ can be very large and does not go "smoothly" to 0
- Forget "fast convergence in the tail"

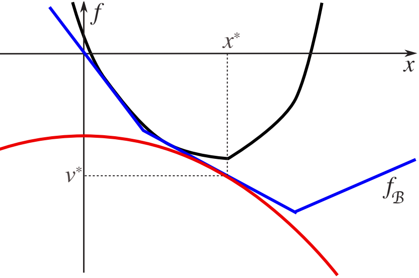
- $\approx$ unavoidable: linear functions have no curvature (really?), you need very many linear functions to make a quadratic one
- Unless f polyhedral and "few facets active in x_*", sometimes happens
- Many iterations $\Longrightarrow \#\mathscr{B} \nearrow \Longrightarrow$ the master problem grows costly
- Pruning $\mathscr{B}$ possible but not easy [4, Ex. 1], no a-priori bound on $\#\mathscr{B}$
- All in all, looks better than subgradient but impractical as it is

- "If something is unstable, then stabilize it" (a.k.a. "regularize")



- $\bar{x}$ stability center ($\approx$ best x^i so far)
- μ stability parameter: "how far from $\bar{x}$ $f_{\mathscr{B}}$ is a good model of f" (??)

- "If something is unstable, then stabilize it" (a.k.a. "regularize")

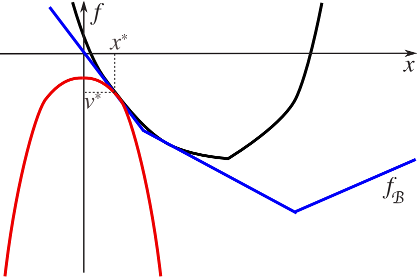


- $\bar{x}$ stability center ($\approx$ best x^i so far)
- μ stability parameter: "how far from $\bar{x}$ $f_{\mathscr{B}}$ is a good model of f" (??)
- Stabilized master problem (not an LP):

  $$\min\{\, f_{\mathscr{B}}(\, x\,) + \mu \|\, x - \bar{x}\,\|^2 \,/\, 2 \,\}$$

- Keeps x^* "close" to $\bar{x}$

- "If something is unstable, then stabilize it" (a.k.a. "regularize")

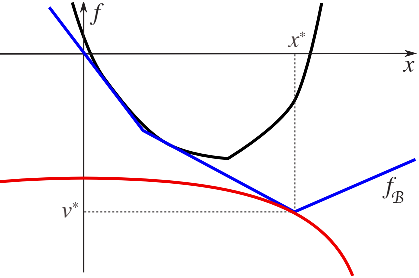


- $\bar{x}$ stability center ($\approx$ best x^i so far)
- μ stability parameter: "how far from $\bar{x}$ $f_{\mathscr{B}}$ is a good model of f" (??)
- Stabilized master problem (not an LP):

  $$\min\{\, f_{\mathscr{B}}(\,x\,) + \mu\|\,x - \bar{x}\,\|^2 / 2\,\}$$

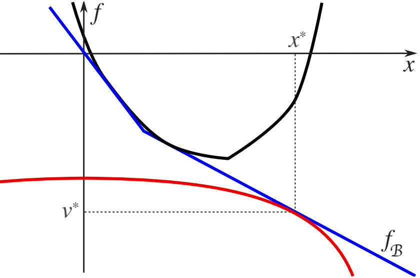
- Keeps x^* "close" to $\bar{x}$
  perhaps too close (μ too large)

- "If something is unstable, then stabilize it" (a.k.a. "regularize")



- $\bar{x}$ stability center ($\approx$ best x^i so far)
- μ stability parameter: "how far from $\bar{x}$ $f_{\mathscr{B}}$ is a good model of f" (??)
- Stabilized master problem (not an LP):
  $$\min\{\, f_{\mathscr{B}}(\, x\,) + \mu \|\, x - \bar{x}\,\|^2 \,/\, 2 \,\}$$
- Keeps x^* "close" to $\bar{x}$
  perhaps too close (μ too large)
- Or not close enough (μ too small) $\approx\equiv$ un-stabilized cutting plane algorithm

- "If something is unstable, then stabilize it" (a.k.a. "regularize")



- $\bar{x}$ stability center ($\approx$ best x^i so far)
- μ stability parameter: "how far from $\bar{x}$ $f_{\mathscr{B}}$ is a good model of f" (??)
- Stabilized master problem (not an LP):
  $$\min\{\, f_{\mathscr{B}}(\, x\,) + \mu \|\, x - \bar{x}\,\|^2 \,/\, 2\,\}$$
- Keeps x^* "close" to $\bar{x}$
  perhaps too close (μ too large)

- Or not close enough (μ too small) $\approx\equiv$ un-stabilized cutting plane algorithm
  except always bounded below $\Longrightarrow$ x^* always well-defined

Exercise: explain why the curious upside-down parabola graphically finds x^*

- Enforces stability $\approx$ trust region ($\nabla^2 f \not\succeq 0$); in fact trust region version $\exists$
- Graft "poorman's Hessian" μI onto $f_{\mathscr{B}}$ $\Longrightarrow$ "poorman's Newton"
- But how to manage $\bar{x}$ and μ?

```
procedure x = PBM ( f , x , m1 , ε , μ )
  B ← { ( x , f( x ) , g ∈ ∂f( x ) ) };
  while ( true ) do
    d* ← argmin { f_B( x + d ) + μ|| d ||^2 / 2 };
    if( μ|| d* || ≤ ε ) then break;
    if( f( x + d* ) − f( x ) ≤ m1[ f_B( x + d* ) − f( x ) ] )
      then { x ← x + d*; possibly μ ↘; } else possibly μ ↗;
    B ← B ∪ { ( x + d* , f( x + d* ) , g ∈ ∂f( x + d* ) ) };
```

- $f(x + d^*) \ll f(x) \implies x \leftarrow x + d^*$ (Armijo-type rule), a Serious Step (SS): this means $f_{\mathscr{B}}$ is "good", $\mu \searrow$ reasonable (try even longer steps)
- x unchanged a Null Step (NS): $f_{\mathscr{B}}$ is "bad", $\mu \nearrow$ reasonable (try shorter steps)
- How to increase / decrease μ? Heuristics $\equiv$ parameters, parameters, ...
- $\{ x^i \} \to x_*$, "optimal" $O(1 / \varepsilon^2)$ complexity: a lot of work but $\approx$ subgradient
- Rather different in practice: it does have "fast convergence in the tail" in practice if $f_{\mathscr{B}}$ succeeds in accruing enough information around x_*
- Can "compress $\mathscr{B}$": master problem cost $\searrow$ but iterations $\nearrow$
- $\#\mathscr{B} \approx 2 \implies$ Bundle $\approx$ subgradient: need "fat" $\mathscr{B}$ for fast convergence

- $0 \in \partial[\, f_{\mathscr{B}}(\, x + \cdot \,) + \mu \| \cdot \|^2 \,](\, d^* \,) \,/\, 2 \implies -\mu d^* \in \partial f_{\mathscr{B}}(\, x + d^* \,) \implies$ $f_{\mathscr{B}}(\, x + d^* \,) - f(\, x \,) \leq -\mu \|\, d^* \,\|^2$ (since $f_{\mathscr{B}}(\, x \,) = f(\, x \,)$) (**check**)

- Need a technical result: $\{\, f^i \,\} \to f^\infty$ (pointwise), $\{\, x^i \,\} \to x \implies$ $\partial f^i(\, x^i \,) \subset \partial f^\infty(\, x \,) + \mathcal{B}(\, 0 \,,\, \varepsilon \,) \; \forall\, \varepsilon$ and large enough i [6, Th. VI.6.2.7]

- Thus: $\{\, x^i \,\} \to x$ and $\{\, \|\, d^{*,i} \,\| \,\} \to 0 \implies 0 \in \partial f(\, x \,)$ (**check**)

- Easy part: ∞ SS made $\implies$ either $f(\, x \,) \to -\infty$ or $\|\, d^* \,\| \to 0$ (**check**) (but $\{\, x^i \,\} \to x$ not obvious, several ways around it)

- Complicated part: $\#$ SS $< \infty \;\equiv\; \infty$ consecutive NS $\implies \|\, d^* \,\| \to 0$ (but at least here $\{\, x^i \,\} \to x$ obvious, finitely happens)

- Intuitively clear: x fixed and $\#\mathscr{B} \nearrow \implies$ "$f_{\mathscr{B}} \to f$ close to x"

- Proof with dual (??) master problem [4] tells which $(\, x^i \,,\, f^i \,,\, g^i \,)$ can be removed from $\mathscr{B}$ and that $\mathscr{B}$ can be "compressed" down to $\#\mathscr{B} = 2$

- $\#\mathscr{B} \searrow \implies$ Bundle $\to$ subgradient: trade-off (iteration $\# \nearrow$ but cost $\searrow$), it often pays to make $\mathscr{B}$ as fat as you can, even with dirty tricks

Outline

- Lack of continuous derivatives is un-good
- No surprise: lack of derivatives is double-plus-un-good
  (although sometimes necessary, e.g., tuning a few hyperparameters)
- Nonsmooth algorithms can be trivial, very robust, and very slow
- Forget high accuracy unless you fight hard:
  either you cheat on the function, or you work with a fat model
- And then the approaches are nontrivial and not-as-robust
- Good news: learning typically does not require high accuracy
- We have have repeatedly seem problems with constraints:
  high time that we move to constrained optimization

[1] S. Bubeck *Convex Optimization: Algorithms and Complexity*, arXiv:1405.4980v2, https://arxiv.org/abs/1405.4980, 2015

[2] M. Cacciola, A. Frangioni, X. Li, A. Lodi *Deep Neural Networks pruning via the Structured Perspective Regularization* arXiv:2206.14056, 2022 https://doi.org/10.48550/arXiv.2206.14056

[3] G. d'Antonio, A. Frangioni *Convergence Analysis of Deflected Conditional Approximate Subgradient Methods* http://pages.di.unipi.it/frangio/abstracts.html#SIOPT08 SIAM Journal on Optimization 20(1), 357–386, 2009.

[4] A. Frangioni *Standard Bundle Methods: Untrusted Models and Duality* http://pages.di.unipi.it/frangio/abstracts.html#NDOB in Numerical Nonsmooth Optimization: State of the Art Algorithms, A.M. Bagirov, M. Gaudioso, N. Karmitsa, M. Mäkelä, S. Taheri (Eds.), 61–116, Springer, 2020

[5] A. Frangioni, B. Gendron, E. Gorgone
Dynamic Smoothness Parameter for Fast Gradient Methods
http://pages.di.unipi.it/frangio/abstracts.html#ORL17a
Optimization Letters 12(1), 43–53, 2018.

[6] J.-B. Hiriart-Urruty, C. Lemaréchal "Convex Analysis and Minimization Algorithms I" Springer-Verlag, 1993

[7] J.-B. Hiriart-Urruty, C. Lemaréchal "Convex Analysis and Minimization Algorithms II" Springer-Verlag, 1993

[8] K.C. Kiwiel *Convergence of Approximate and Incremental Subgradient Methods for Convex Optimization* SIAM Journal on Optimization 14(3), 807–840, 2003

[9] Y. Nesterov *Smooth minimization of non-smooth functions* Mathematical Programming 103, 127–152, 2005.

[10] Wikipedia – Islands of Space
https://en.wikipedia.org/wiki/Islands_of_Space

Outline

- $v \in \partial f(x) \equiv f(z) \geq f(x) + \langle v, z - x \rangle$, and $w \in \partial g(x) \equiv g(z) \geq g(x) + \langle w, z - x \rangle$. Hence, $\alpha f(z) + \beta g(z) \geq$ $\geq \alpha[f(x) + \langle v, z - x \rangle] + \beta[g(x) + \langle v, z - x \rangle] =$ $[\alpha f(x) + \beta g(x)] + \langle \alpha v + \beta w, z - x \rangle = [\alpha f + \beta g](x) + \langle \zeta, z - x \rangle$ for $\zeta = \alpha v + \beta w \implies \zeta \in \partial[\alpha f + \beta g](x)$ **[back]**

- $f(x) = |x| = \max\{ f_1(x) = x, f_2(x) = -x \}$, hence we can use rule iv. $x > 0 \equiv I(x) = \{1\} \equiv \partial f(x) = \{ f_1'(x) \} = \{1\}$. Symmetrically, $x < 0 \equiv I(x) = \{2\} \equiv \partial f(x) = \{ f_2'(x) \} = \{-1\}$. Thus, $f(x)$ is differentiable $\forall x \neq 0$. However, $I(x) = \{1, 2\} \implies \partial f(x) =$ $\mathrm{conv}(f_1'(0) \cup f_2'(0)) = \mathrm{conv}(\{1, -1\}) = [-1, 1]$, hence $f(x)$ is not differentiable in 0 **[back]**

- The crucial relationship comes from expanding $\|\,[\,x^i - x_*\,] - \alpha^i g^i\,\|^2 = \|\,x^i - x_*\,\|^2 - 2\alpha^i\langle\,x^i - x_*\,,\,g^i\,\rangle + (\alpha^i)^2\|\,g^i\,\|^2$, changing sign in the middle term as $+2\alpha^i\langle\,x_* - x^i\,,\,g^i\,\rangle$, and then using the subgradient inequality $g^i \in \partial f(\,x^i\,) \quad\equiv\quad f(\,z\,) \geq f(\,x^i\,) + \langle\,g^i\,,\,z - x^i\,\rangle$ at $z = x_*$, yielding $\langle\,g^i\,,\,x_* - x^i\,\rangle \leq f(\,x_*\,) - f(\,x_i\,)$ $[< 0]$ as already recalled Then, the quadratic function $\phi(\,\alpha\,) = a\alpha^2 + b\alpha$ with $a = \|\,g^i\,\|^2 > 0$ and $b = 2(\,f_* - f(\,x_i\,)\,) < 0$ notoriously has the two roots $\alpha = 0$ and $\alpha = -b\,/\,a > 0$: $\phi(\,\alpha^i\,) < 0$ between the two roots, i.e., $\forall\,\alpha^i \in (\,0\,,\,2(\,f(\,x_i\,) - f_*\,)\,/\,\|\,g^i\,\|^2\,)$, yielding $\|\,x^{i+1} - x_*\,\|^2 < \|\,x^i - x_*\,\|^2$, i.e., the algorithm would succeed in decreasing the distance from x_*. In particular, it is well-known that $\phi(\,\alpha\,)$ has minimum (most negative) value in the middle of the interval, i.e., $\alpha^i = (\,f^i - f_*\,)\,/\,\|\,g^i\,\|^2$ is the step guaranteeing the largest decrease in the distance from x_*. The issue here is clearly that f_* is unknown, and therefore the "optimal" step cannot be computed [**back**]

- In the usual $\| x^{i+1} - x_* \|^2 \leq \| x^i - x_* \|^2 + 2\alpha^i(f_* - f^i) + (\alpha^i)^2 \| g^i \|^2$ plug $\alpha^i = \alpha_i^* = (f^i - f_*) / \| g^i \|^2$ to get $\| x^{i+1} - x_* \|^2 \leq \| x^i - x_* \|^2$ $+2[(f^i - f_*) / \| g^i \|^2](f_* - f^i) + [(f^i - f_*) / \| g^i \|^2]^2 \| g^i \|^2 =$ $= \| x^i - x_* \|^2 - 2(f^i - f_*)^2 / \| g^i \|^2 + (f^i - f_*)^2 / \| g^i \|^2 =$ $= \| x^i - x_* \|^2 - (f^i - f_*)^2 / \| g^i \|^2$ [**back**]

- $f(x)$ is the pointwise maximum of (possibly, ∞-ly many) linear functions $f_z(x)$, one for each $z \in Z$; since each f_z is convex, their maximum also is [**back**]

- The first step is to write $f(x) = | x | = \max\{ x , -x \} =$ $= \max\{ zx : z \in \{ -1 , 1 \} \}$. One then has to realise that "$z \in \{ -1 , 1 \}$" can equivalently be taken as "$z \in [-1 , 1]$": in fact, for (say) $x > 0$ the maximum is still attained in $z = 1$, as for all $z < 1$ one has $zx < x$ (the case $x < 0$ is symmetric). Hence, $f_\mu(x) = \max\{ zx - \mu z^2 / 2 : z \in [-1 , 1] \}$. This is the maximum of the (concave) quadratic non-homogeneous univariate function $\phi(z) = zx - \mu z^2 / 2$ on the interval $[-1 , 1]$, that we know well how to compute: first we write the unconstrained maximum $z_*(x) = x / \mu$, and then we project it on the interval, i.e., the maximum is $\max\{\, 1\, ,\, \min\{\, -1\, ,\, x \,/\, \mu \,\}\,\}$. Plugging this formula into the function gives, after a bit of algebra,

$$f_\mu(\, x\,) = \begin{cases} x^2 \,/\, (2\mu) & \text{if } |\, x\,| \leq \mu \\ |\, x\,| - \mu \,/\, 2 & \text{if } |\, x\,| \geq \mu \end{cases}$$

Hence, $f_\mu(\, x\,)$ "has the same shape" as $f(\, x\,)$ "far from 0" (i.e., for $|\, x\,| \geq \mu$), in that $f_\mu(\, x\,) = f(\, x\,) - \mu \,/\, 2$, whereas $f_\mu(\, x\,)$ is a simple quadratic function that "approximates" the absolute value close to 0; in particular, $f_\mu(\, x\,) = 0$. It is easy to verify that f_μ is continuous ($f_\mu(\, \mu\,) = \mu^2 \,/\, (2\mu) = \mu \,/\, 2 = \mu - \mu \,/\, 2$) and differentiable ($f'_\mu(\, \mu\,) = \mu \,/\, \mu = 1$), as expected since all convex functions are continuous (on the interior of their domain) and the cardinality of the subdifferential is that of the optimal solutions of the max problem, which always has a unique optimal solution. Thus, $f_\mu(\, x\,)$ is indeed a "smoothed version" of $f(\, x\,)$ [**back**]

- The crucial formula is $\mu = \varepsilon / (2K)$; in the standard approach, ε is fixed and so μ is $\Longrightarrow$ if ε is small then so μ is, which makes the algorithm perform very small steps at the beginning (and for a long while) slowing down convergence. A simple idea is to rather take $\mu^i = \max\{ f^i - f_* , \varepsilon \}$ and just run the algorithm with this varying μ. This results in much longer steps at the beginning while the two will tend to behave similarly as $f^i \to f_*$. Of course, requires either having information about f_*, which is unlikely (but not impossible), or some form of target-level approach **[back]**

- $v^* = f_{\mathscr{B}}(x^*_{\mathscr{B}}) \geq f_*$, and $f(x^*) \geq f_*$ by definition, hence $f(x^*) \leq v^* \Longrightarrow f_* \leq f(x^*) \leq v^* \leq f_* \Longrightarrow f_* = f(x^*) \Longrightarrow x^*$ optimal; in fact, it is not possible that $f(x^*) < v^*$, so the check could just be $f(x^*) = v^*$ (save of course for the issue of numerical errors) **[back]**

- Since $\mathscr{B}^{i+1} \supset \mathscr{B}^i$, it is immediate to see that $f^{i+1}_{\mathscr{B}}(x) \geq f^i_{\mathscr{B}}(x) \; \forall \in \mathbb{R}^n$, whence $\underline{f}^{i+1} = f^{i+1}_{\mathscr{B}}(x^{*,i+1}) = v^{*,i+1} = \min\{ f^{i+1}_{\mathscr{B}}(x) \} \geq \min\{ f^i_{\mathscr{B}}(x) \} = v^{*,i} = f^i_{\mathscr{B}}(x^{*,i}) = \underline{f}^i$ **[back]**

- Consider the convex extended-value function $g(x) = f(x) \; \forall\, x \in \mathcal{B}(x_*, \varepsilon)$, while $g(x) = \infty$ otherwise. Also, consider the variant to the Cutting Plane algorithm in which the constraint "$x \in \mathcal{B}(x_*, \varepsilon)$" is added to the master problem (which is still a Linear Problem if the ball is, say, in the ∞-norm, but even with the Euclidean norm it becomes a problem with convex quadratic—in fact, conic as we will see—constraints and therefore still "easy"). The convergence proof of the Cutting Plane algorithm [4, Th. 1] allows for this constraint in the master problem, and in fact it requires it unless $\mathscr{B}^0$ is "large enough" so that the master problem is bounded below; see next slide. So, the algorithm solves $\min\{\, g(x) \,\} = \min\{\, f(x) \,\}$ (the two problems obviously have the same optimal value and an optimal solution in x_*) when all the iterates are forced to remain in an arbitrarily small ball around x^*. Interestingly, this is not only an abstract proof: [4, Table 1] shows that if one would actually be able to force $x^{*,i} \in \mathcal{B}(x_*, \varepsilon)$ then the practical convergence of the Cutting Plane algorithm would typically be faster (dramatically so when ε is small). This is unfortunately impossible since x_* is usually unknown, but the mechanism does suggest the crucial idea behind the practically useful stabilisation approaches [**back**]

- Let $s(x) = \mu \| x - \bar{x} \|^2 / 2$ be the stabilising term, which clearly is a parabola with curvature μ and centred in $\bar{x}$. The optimality condition of the master problem is $0 \in \partial[f_{\mathscr{B}}(\cdot) + s(\cdot)](x^*) \equiv \exists g \in \partial f_{\mathscr{B}}(x^*)$ s.t. $g + \nabla s(x^*) = 0$ $\equiv g = -\nabla s(x^*)$. That is, the derivative of $f_{\mathscr{B}}$ must be the opposite of that of s in x^*. Of course, $-\nabla s(x^*) = \nabla[-s(\cdot)](x^*)$, and $-s$ is the same parabola "upside-down". Hence, x^* is the point where the upside-down parabola and $f_{\mathscr{B}}$ have the same derivative. Geometrically, this can be found by imagining the upside-down parabola shifted by a negative constant, i.e., $-s(x) - M$, so that the value in $\bar{x}$ is $-M$. Then, one starts with a "very large" $M > 0$, so that the upside-down parabola is "pushed down a lot", and gradually decreases M so that it "gradually moves up". By stopping for the first (smallest) value of M such that the graph of $-s(x) - M$ and that of $f_{\mathscr{B}}$ touch, which must exist (for $M = 0$ the two graphs surely touch), one has that either the two derivatives are equal of $f_{\mathscr{B}}$ is differentiable there, or at least there exists a subgradient g of $f_{\mathscr{B}}$ that does the requisite job (see the pictures). Thus, the x coordinate of the point where the two meet is x_* [**back**]

- The first step is due to the fact (already seen in the previous exercise) that $0 \in \partial[\, f_{\mathscr{B}}(\, x + \cdot\,) + \mu\| \cdot \|^2 \,/\, 2\,](\, d^*\,) \;\equiv\; \exists g \in \partial f_{\mathscr{B}}(\, x + d^*\,)$ s.t. $g + \mu d^* = 0$ $\Longrightarrow\; -\mu d^* \in \partial f_{\mathscr{B}}(\, x + d^*\,)$ since $\mu d^* = \nabla[\, \mu\| \cdot \|^2 \,/\, 2\,](\, d^*\,)$. The second step is just the subgradient inequality for $f_{\mathscr{B}}$ evaluated in $x + d^*$: $f(\, x\,) = f_{\mathscr{B}}(\, x\,) \geq$ $\geq f_{\mathscr{B}}(\, x + d^*\,) + \langle\, -\mu d^*\,,\, x - (\, x + d^*\,)\,\rangle$ (properly rearranged) [**back**]

- First note that $\{\, x^i\,\}$ is the sequence of the stability centres, not of the iterates. However, $\{\, x^i\,\} \to x$ and $\{\, \|\, d^{*,i}\,\|\,\} \to 0$ imply that $\{\, x^i + d^{*,i}\,\} \to x$ as well: both the stability centres and the iterates converge to the same point. Note that the stability centres may or may not finitely converge, i.e., after finitely many SS the centre may no longer be changed and only (infinitely many consecutive) NS will be done; yet, this is immaterial for the current result. Now, let $f^i_{\mathscr{B}}$ be the cutting plane model at iteration i, and $f^\infty_{\mathscr{B}}$ be the convex function defined by the set $\mathscr{B}^\infty$ containing all the infinitely many triples $(\, x^i\,,\, f^i\,,\, g^i\,)$: clearly, $\{\, f^i_{\mathscr{B}}\,\} \to f^\infty_{\mathscr{B}}$ pointwise, i.e., however fixed $z \in \mathbb{R}^n$ one has $\lim_{i\to\infty} f^i_{\mathscr{B}}(\, z\,) = \lim_{i\to\infty} \max\{\, l^h(\, z\,) \;:\; h \leq i\,\} = \sup\{\, l^i(\, z\,) \;:\; i \in \mathbb{N}\,\} =$ $f^\infty_{\mathscr{B}}(\, z\,)$. Thus the theorem applies. Also, since $f^i_{\mathscr{B}}(\, z\,) \leq f(\, z\,)$ and $f^\infty_{\mathscr{B}}(\, z\,) = \lim_{i\to\infty} f^i_{\mathscr{B}}(\, z\,)$, then $f^\infty_{\mathscr{B}}(\, z\,) \leq f(\, z\,)$, i.e., $f^\infty_{\mathscr{B}}$ is still a correct lower model of f. Now, $-\mu d^{*,i} \in \partial f^i_{\mathscr{B}}(\, x^i + d^{*,i}\,)$ and (again) $\{\, \|\, d^{*,i} \,\|\,\} \to 0$: thus, however chosen $\varepsilon > 0$ and $\delta > 0$ exists $g \in \partial f^\infty_{\mathscr{B}}(\, x\,)$, v s.t. $\|\, v \,\| \leq \delta$, z s.t. $\|\, z \,\| \leq \varepsilon$ and $z = g + v$ (just wait until i is large enough so that both $\|\, -\mu d^{*,i} \,\| \leq \delta$ and $\partial f^i_{\mathscr{B}}(\, x^i + d^{*,i}\,) \subset \partial f^\infty_{\mathscr{B}}(\, x\,) + \mathcal{B}(\, 0\,,\, \varepsilon\,)$ hold). Hence, $\|\, g \,\| \leq \|\, z - v \,\| \leq \|\, z \,\| + \|\, v \,\| \leq \varepsilon + \delta$: there are elements in $\partial f^\infty_{\mathscr{B}}(\, x\,)$ arbitrarily close to 0. But $f : \mathbb{R}^n \to \mathbb{R}$ is finite-valued and therefore $f^\infty_{\mathscr{B}} \leq f$ is also finite-valued, hence $\partial f^\infty_{\mathscr{B}}(\, x\,)$ is compact and therefore in particular closed: as a consequence, $0 \in \partial f^\infty_{\mathscr{B}}(\, x\,)$, i.e., x is a minimum of $f^\infty_{\mathscr{B}}$. Since $f^\infty_{\mathscr{B}} \leq f$, $f^\infty_{\mathscr{B}}(\, x\,) \leq f_* \leq f(\, x\,)$. Now, $\mathscr{B}^{i+1} = \mathscr{B}^i \cup \{\, (\, x^i + d^{*,i}\,,\, f(\, x^i + d^{*,i}\,)\,,\, g^i\,)\,\} \implies f^{i+1}_{\mathscr{B}}(\, x^i + d^{*,i}\,) = f(\, x^i + d^{*,i}\,)$. Send $i \to \infty$ to yield $f^\infty_{\mathscr{B}}(\, x\,) = f(\, x\,)$: together with $f^\infty_{\mathscr{B}}(\, x\,) \leq f_* \leq f(\, x\,)$ this gives $f^\infty_{\mathscr{B}}(\, x\,) = f(\, x\,) = f_*$

In the proof μ is fixed, but it easily extends to μ^i bounded above by some constant, so that $\{\, \|\, d^{*,i} \,\|\,\} \to 0 \implies \{\, \|\, \mu^i d^{*,i} \,\|\,\} \to 0$ **[back]**

- Direct from $f_{\mathscr{B}}(\, x + d^* \,) - f(\, x \,) \leq -\mu\|\, d^* \,\|^2$ and the SS condition $f(\, x + d^* \,) - f(\, x \,) \leq m_1[\, f_{\mathscr{B}}(\, x + d^* \,) - f(\, x \,)]$: $\|\, d^{i,*} \,\|^2 \geq \varepsilon \; \forall\, i \implies f_{\mathscr{B}}(\, x^i + d^{i,*} \,) - f(\, x^i \,) \leq -\mu\varepsilon \; \forall\, i \implies f(\, x^i + d^{i,*} \,) - f(\, x^i \,) \leq -\mu m_1 \varepsilon$ at each i where a SS is declared ($x^{i+1} = x^i + d^{i,*}$). Thus, if ∞-ly many SS are declared, $f(\, x^i \,) \to -\infty$; conversely, if $f_* > -\infty$ (f is bounded below) this cannot happen, which means that $\{\, \|\, d^{*,i} \,\| \,\} \to 0$ must happen istead
Note that, again, this proof is using a fixed μ, but is easily extended to μ^i bounded away from 0, or even $\mu^i \to 0$ provided that $\sum_{i=1}^{\infty} \mu^i = \infty$ (with the series actually only running on the SS iterations i) [**back**]