Aspects of Convex, Nonconvex, and Geometric Optimization

(Lecture 1)

Suvrit Sra

Massachusetts Institute of Technology

Hausdorff Institute for Mathematics (HIM)
Trimester: Mathematics of Signal Processing
January 2016

Outline

- Convex analysis, optimality
- First-order methods
- Proximal methods, operator splitting
- Stochastic optimization, incremental methods
- Nonconvex models, algorithms
- Geometric optimization

Convex analysis

Convex sets (vector space)

Def. Set $C \subset \mathbb{R}^n$ called **convex**, if for any $x, y \in C$, the line-segment $\theta x + (1-\theta)y$, where $\theta \in [0,1]$, also lies in C.

Observations

- Linear: if restrictions on θ_1, θ_2 are dropped
- Conic: if restriction $\theta_1 + \theta_2 = 1$ is dropped
- Convex: $\theta_1 x + \theta_2 y \in C$, where $\theta_1, \theta_2 \geq 0$ and $\theta_1 + \theta_2 = 1$.

Theorem (Intersection).
Let C_1, C_2 be convex sets. Then, $C_1 \cap C_2$ is also convex.

Convex sets (vector space)

♡ Let $x_1, x_2, \ldots, x_m \in \mathbb{R}^n$. Their **convex hull** is

$$\text{co}(x_1, \ldots, x_m) := \left\{ \sum_i \theta_i x_i \mid \theta_i \geq 0, \sum_i \theta_i = 1 \right\}.$$

♡ Let $A \in \mathbb{R}^{m \times n}$, and $b \in \mathbb{R}^m$. The set $\{x \mid Ax = b\}$ is convex (it is an *affine space* over subspace of solutions of $Ax = 0$).

♡ *halfspace* $\{x \mid a^T x \leq b\}$.

♡ *polyhedron* $\{x \mid Ax \leq b, Cx = d\}$.

♡ *ellipsoid* $\{x \mid (x - x_0)^T A(x - x_0) \leq 1\}$, ($A$: semidefinite)

♡ *convex cone* $x \in \mathcal{K} \implies \alpha x \in \mathcal{K}$ for $\alpha \geq 0$ (and $\mathcal{K}$ convex)

Challenge 1

Let $A, B \in \mathbb{R}^{n \times n}$ be symmetric. Prove that

$$R(A, B) := \left\{ (x^T A x, x^T B x) \mid x^T x = 1 \right\}$$

is a compact convex set for $n \geq 3$.

At the heart of S-lemma in control / optimization.

Convex functions

Def. Function $f : I \to \mathbb{R}$ on interval I called **midpoint convex** if

$$f\left(\frac{x+y}{2}\right) \leq \frac{f(x)+f(y)}{2}, \qquad \text{whenever } x, y \in I.$$

Read: f of AM is less than or equal to AM of f.

Convex functions

Def. Function $f : I \to \mathbb{R}$ on interval I called **midpoint convex** if

$$f\left(\frac{x+y}{2}\right) \leq \frac{f(x)+f(y)}{2}, \qquad \text{whenever } x, y \in I.$$

Read: f of AM is less than or equal to AM of f.

Think: What is we use other means, e.g., GM-AM, GM-GM?

Convex functions

Def. Function $f : I \to \mathbb{R}$ on interval I called **midpoint convex** if

$$f\left(\frac{x+y}{2}\right) \leq \frac{f(x)+f(y)}{2}, \qquad \text{whenever } x, y \in I.$$

Read: f of AM is less than or equal to AM of f.
Think: What is we use other means, e.g., GM-AM, GM-GM?

Theorem (J.L.W.V. Jensen). Let $f : I \to \mathbb{R}$ be continuous. Then, f is convex *if and only if* it is midpoint convex.

- Extends to $f : \mathcal{X} \subseteq \mathbb{R}^n \to \mathbb{R}$; useful for proving convexity.

Recognizing convex functions

- If f is continuous and midpoint convex, then it is convex.
- If f is differentiable, then f is convex *if and only if* $\operatorname{dom} f$ is convex and $f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle$ for all $x, y \in \operatorname{dom} f$.
- If f is twice differentiable, then f is convex *if and only if* $\operatorname{dom} f$ is convex and $\nabla^2 f(x) \succeq 0$ at every $x \in \operatorname{dom} f$.
- By showing f to be a pointwise max of convex functions
- By showing $f : \operatorname{dom}(f) \to \mathbb{R}$ is convex *if and only if* its restriction to **any** line that intersects $\operatorname{dom}(f)$ is convex. That is, for any $x \in \operatorname{dom}(f)$ and any v, the function $g(t) = f(x + tv)$ is convex (on its domain $\{t \mid x + tv \in \operatorname{dom}(f)\}$).
- Exercises (Ch. 3) in Boyd & Vandenberghe
- Several more ways . . .

Convex functions – Indicator

Let $\mathbb{1}_{\mathcal{X}}$ be the *indicator function* for $\mathcal{X}$ defined as:

$$\mathbb{1}_{\mathcal{X}}(x) := \begin{cases} 0 & \text{if } x \in \mathcal{X}, \\ \infty & \text{otherwise.} \end{cases}$$

Note: $\mathbb{1}_{\mathcal{X}}(x)$ is convex **if and only if** $\mathcal{X}$ is convex.

Convex functions – distance

Example Let $\mathcal{X}$ be a convex set. Let $x \in \mathbb{R}^n$ be some point. The distance of x to the set $\mathcal{X}$ is defined as

$$\text{dist}(x, \mathcal{X}) := \inf_{y \in \mathcal{X}} \quad \|x - y\|.$$

Note: because $\|x - y\|$ is jointly convex in (x, y), the function $\text{dist}(x, \mathcal{Y})$ is a convex function of x.

Convex functions – norms

Example (ℓ_p-norm): Let $p \geq 1$. $\|x\|_p = \left(\sum_i |x_i|^p\right)^{1/p}$

Example (Frobenius-norm): Let $A \in \mathbb{R}^{m \times n}$. $\|A\|_{\mathrm{F}} := \sqrt{\sum_{ij} |a_{ij}|^2}$

Example Let A be any matrix. Then, the operator norm of A is

$$\|A\| := \sup_{\|x\|_2 \neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \sigma_{\max}(A).$$

Example Operator $q \to p$ norm (typically NP-Hard to compute!)

$$\|A\|_{q \to p} := \sup_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_q}$$

Fenchel conjugate

Def. The **Fenchel conjugate** of a function f is

$$f^*(z) := \sup_{x \in \operatorname{dom} f} \quad x^T z - f(x).$$

Fenchel conjugate

Def. The **Fenchel conjugate** of a function f is

$$f^*(z) := \sup_{x \in \operatorname{dom} f} \quad x^T z - f(x).$$

Note: f^* is pointwise (over x) sup of linear functions of z. Hence, it is always convex (even if f is not convex).

Example $+\infty$ and $-\infty$ conjugate to each other.

Think: What about other notions of conjugates?

Fenchel conjugate

Def. The **Fenchel conjugate** of a function f is

$$f^*(z) := \sup_{x \in \operatorname{dom} f} \quad x^T z - f(x).$$

Note: f^* is pointwise (over x) sup of linear functions of z. Hence, it is always convex (even if f is not convex).

Example $+\infty$ and $-\infty$ conjugate to each other.

Think: What about other notions of conjugates?
[Artstein-Avidan, Milman (2007).] Up to linear terms, Fenchel conjugate only possible transform under basic axioms.

Challenge 2

Consider the following functions on strictly positive variables:

$$
\begin{aligned}
h_1(x) &:= \frac{1}{x} \\
h_2(x,y) &:= \frac{1}{x} + \frac{1}{y} - \frac{1}{x+y} \\
h_3(x,y,z) &:= \frac{1}{x} + \frac{1}{y} + \frac{1}{z} - \frac{1}{x+y} - \frac{1}{y+z} - \frac{1}{x+z} + \frac{1}{x+y+z}
\end{aligned}
$$

- ♡ Prove that h_1, h_2, h_3, and in general h_n are convex!
- ♡ Prove that in fact each $1/h_n$ is concave

$\nabla^2 h_n(x) \succeq 0$ is not recommended 

Arose in studying *expected random broadcast time* in an unreliable star network (I. Affleck, 1994):

$$
h_n(x) = E_n[T(x)] := \sum_{\sigma \in S_n} \left(\prod_{i=1}^{n} \frac{x_{\sigma(i)}}{\sum_{j=i}^{n} x_{\sigma(j)}} \right) \left(\sum_{i=1}^{n} \frac{1}{\sum_{j=i}^{n} x_{\sigma(j)}} \right)
$$

Challenge 3 – Kummer function

(D. Karp, 2009). Let $c > a > 0$, let $x \in \mathbb{R}$. Prove that

$$\mu \mapsto -\log {}_1F_1(a+\mu, c+\mu, x) = -\log \left[\sum_{k \geq 0} \frac{(a+\mu)_k}{(c+\mu)_k} \frac{x^k}{k!} \right],$$

is convex on $(0, \infty)$. (Here $(x)_k := x(x-1)\cdots(x-k+1)$).

Open problem.

(Arose while we were studying "Directional Statistics" using the Multivariate Watson Distribution on $\mathbb{RP}^{n-1}$)

Challenge 4 – DPP entropy

- Determinantal Point Process (DPP): Essentially, probability measure over subsets of a ground set Y, wlog $Y = \{1, \ldots, n\}$
- Suppose random subset A is drawn from 2^Y as per DPP
- Probability that any $S \in 2^Y$ is covered by A is

$$P(S \subseteq A) = \det(K_S), \quad S \subseteq [n],$$

  $0 \preceq K \preceq I$ the (marginal) DPP kernel; $K_S = K[S, S]$
- Sometimes, we may write $P(S) \propto \det(L_S)$ for $L = K(I - K)^{-1}$

Challenge 4 – DPP entropy

- Determinantal Point Process (DPP): Essentially, probability measure over subsets of a ground set Y, wlog $Y = \{1, \ldots, n\}$
- Suppose random subset A is drawn from 2^Y as per DPP
- Probability that any $S \in 2^Y$ is covered by A is

$$P(S \subseteq A) = \det(K_S), \quad S \subseteq [n],$$

  $0 \preceq K \preceq I$ the (marginal) DPP kernel; $K_S = K[S, S]$
- Sometimes, we may write $P(S) \propto \det(L_S)$ for $L = K(I - K)^{-1}$

(R. Lyons, 2003). On $\{0 \preceq K \prec I\}$, the negentropy

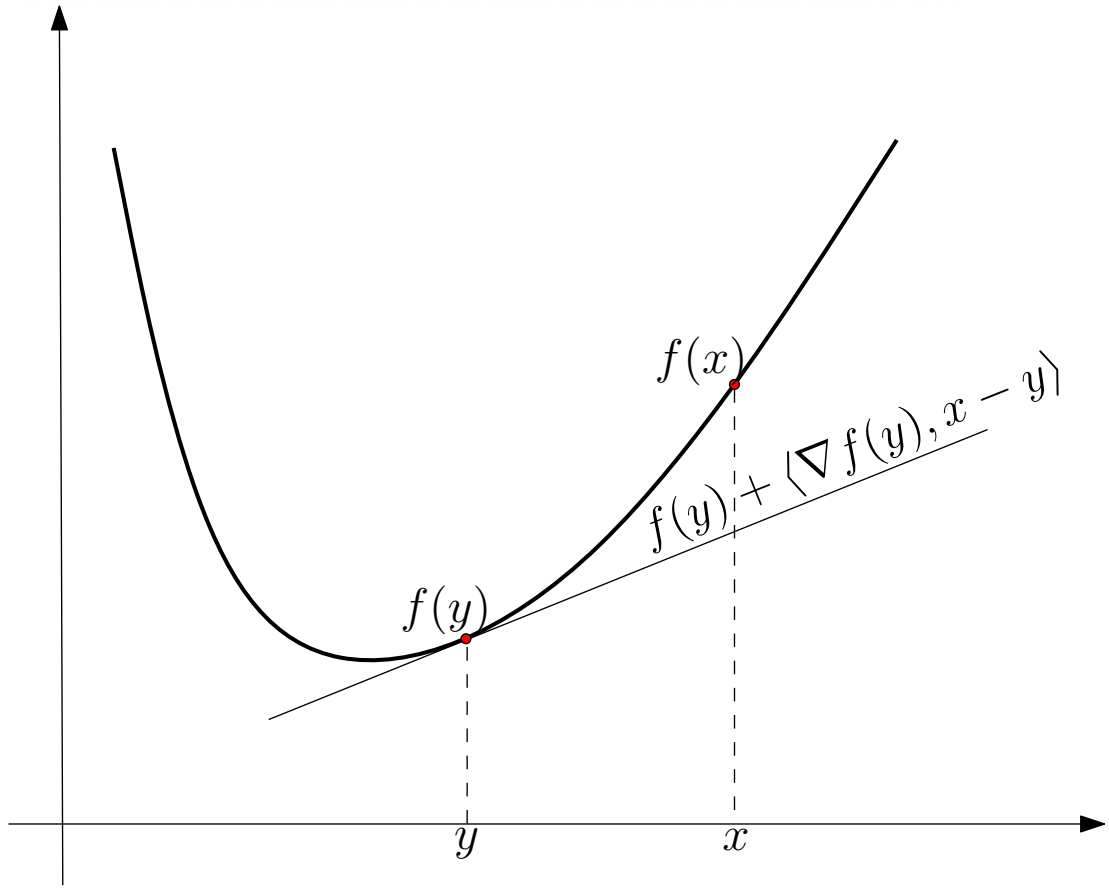
$$H(K) := \sum\nolimits_{S \subseteq [n]} P(S) \log P(S)$$

is convex, where $P(S) = \det(I - K) \det([K(I - K)^{-1}]_S)$.

Open problem.

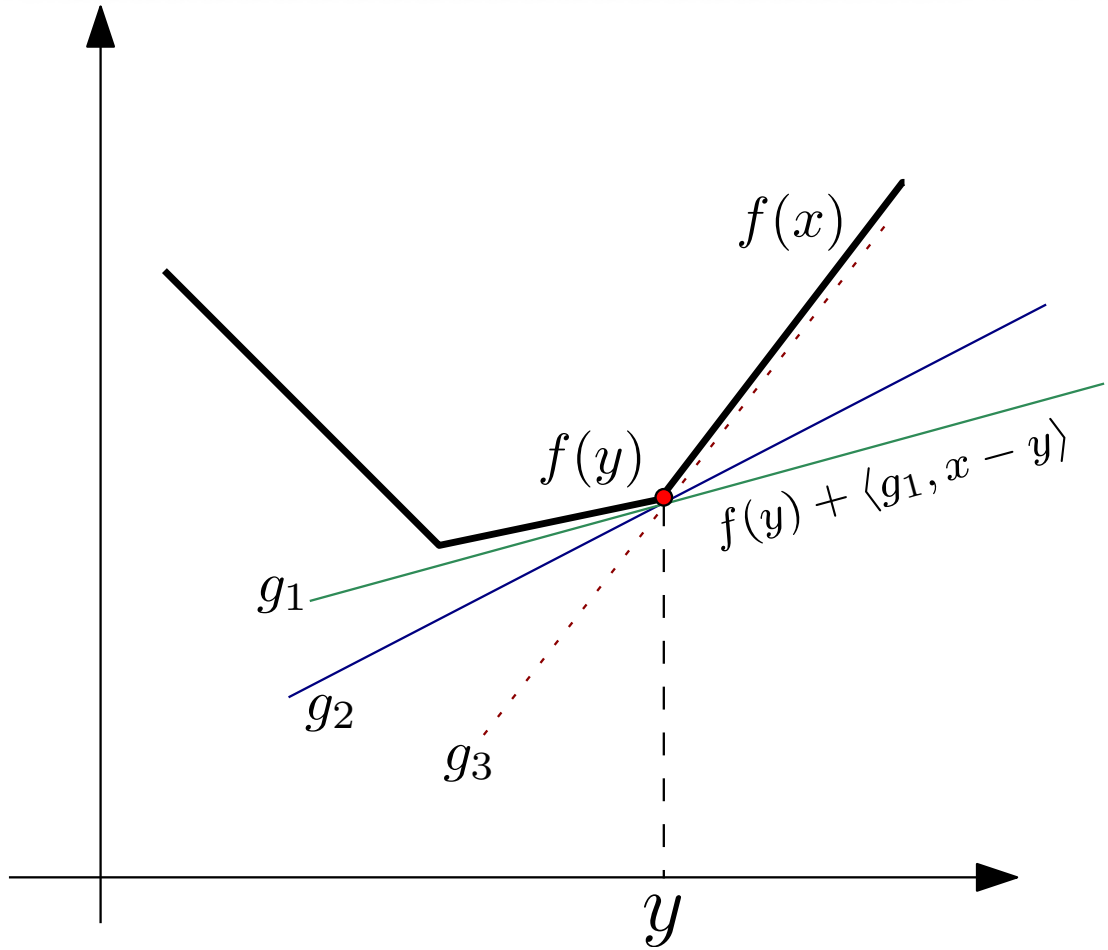
Convex Optimization

Subgradients: global underestimators



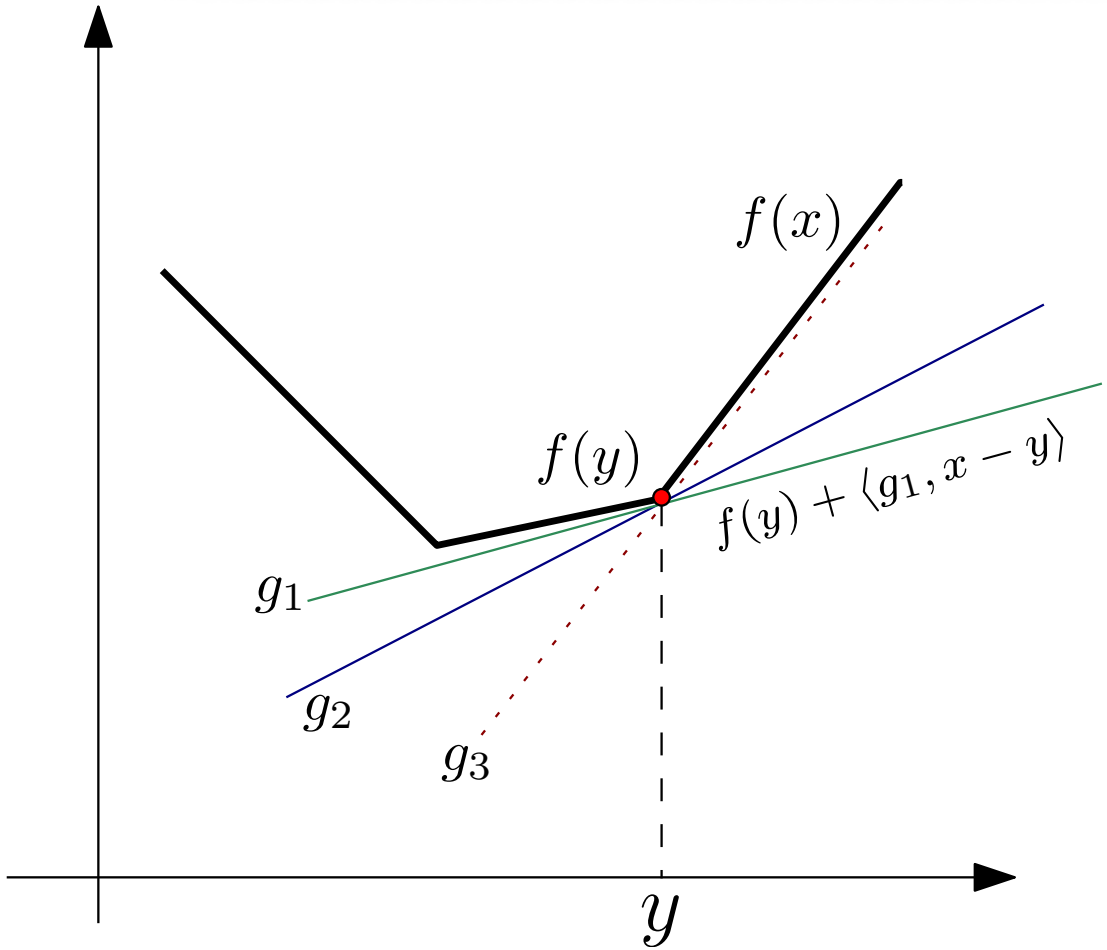
$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle$$

Subgradients: global underestimators



$$f(x) \geq f(y) + \langle g, x - y \rangle$$

Subgradients: global underestimators



g_1, g_2, g_3 are subgradients at y

Subgradients – basic facts

- f is convex, differentiable: $\nabla f(y)$ the **unique** subgradient at y
- A vector g is a subgradient at a point y if and only if $f(y) + \langle g, x - y \rangle$ is **globally** smaller than $f(x)$.
- Usually, **one** subgradient costs approx. as much as $f(x)$

Subgradients – basic facts

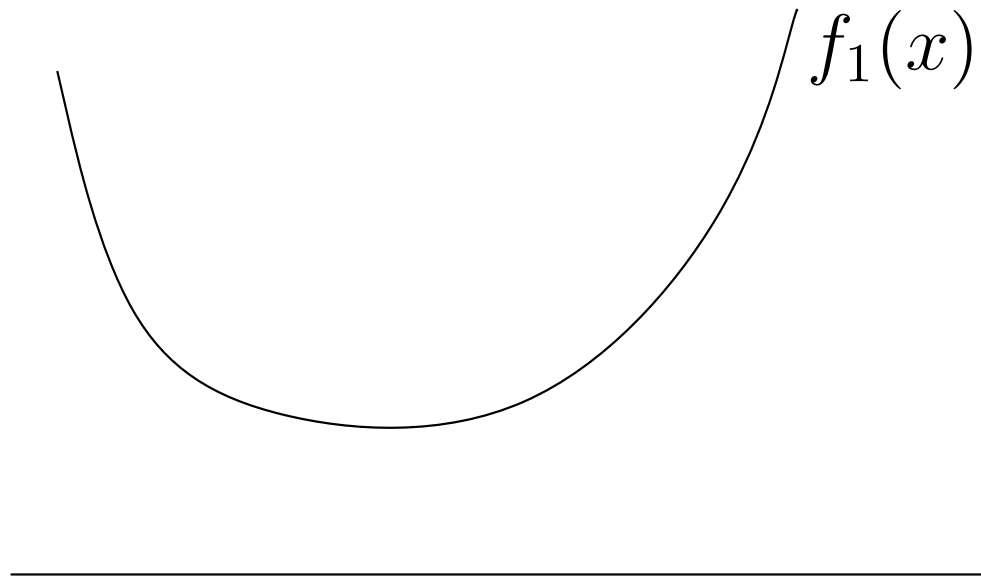
- f is convex, differentiable: $\nabla f(y)$ the **unique** subgradient at y
- A vector g is a subgradient at a point y if and only if $f(y) + \langle g, x - y \rangle$ is **globally** smaller than $f(x)$.
- Usually, **one** subgradient costs approx. as much as $f(x)$
- Determining all subgradients at a given point — difficult.
- Subgradient calculus—major achievement in convex analysis
- Fenchel-Young inequality: $f(x) + f^*(s) \geq \langle s, x \rangle$

Subgradients – example

$f(x) := \max(f_1(x), f_2(x))$; both f_1, f_2 convex, differentiable

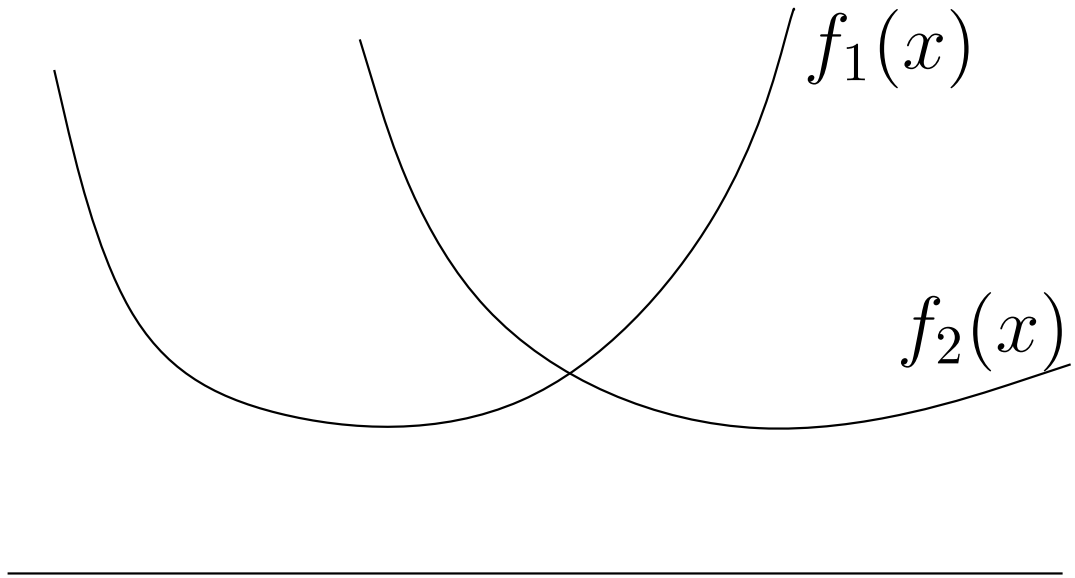
Subgradients – example

$f(x) := \max(f_1(x), f_2(x))$; both f_1, f_2 convex, differentiable



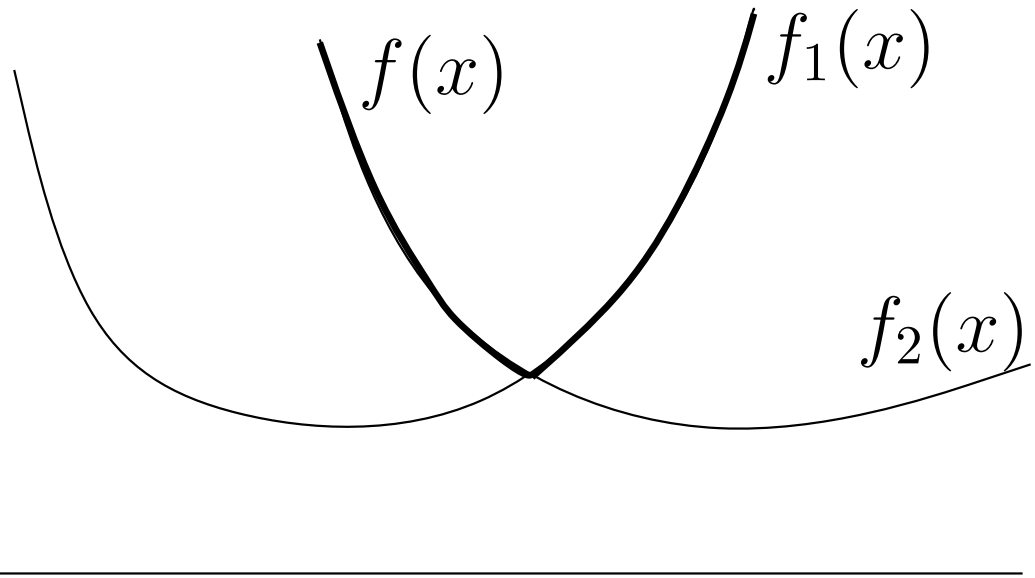
Subgradients – example

$f(x) := \max(f_1(x), f_2(x))$; both f_1, f_2 convex, differentiable



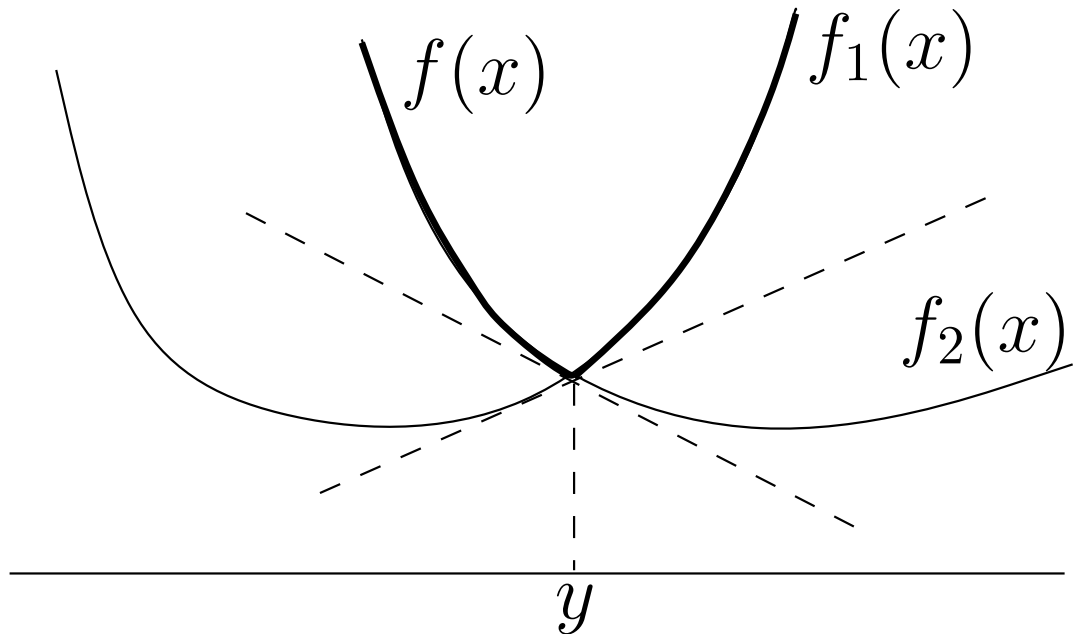
Subgradients – example

$f(x) := \max(f_1(x), f_2(x))$; both f_1, f_2 convex, differentiable



Subgradients – example

$f(x) := \max(f_1(x), f_2(x))$; both f_1, f_2 convex, differentiable



- $f_1(x) > f_2(x)$: unique subgradient of f is $f_1'(x)$
- $f_1(x) < f_2(x)$: unique subgradient of f is $f_2'(x)$
- $f_1(y) = f_2(y)$: subgradients, the segment $[f_1'(y), f_2'(y)]$ (imagine all supporting lines turning about point y)

Subdifferential

Def. The set of all subgradients at y denoted by $\partial f(y)$. This set is called **subdifferential** of f at y

Subdifferential

Def. The set of all subgradients at y denoted by $\partial f(y)$. This set is called **subdifferential** of f at y

If f is convex, $\partial f(x)$ is nice:

- ♣ If $x \in$ relative interior of dom f, then $\partial f(x)$ nonempty

Subdifferential

Def. The set of all subgradients at y denoted by $\partial f(y)$. This set is called **subdifferential** of f at y

If f is convex, $\partial f(x)$ is nice:

- If $x \in$ relative interior of dom f, then $\partial f(x)$ nonempty
- If f differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$

Subdifferential

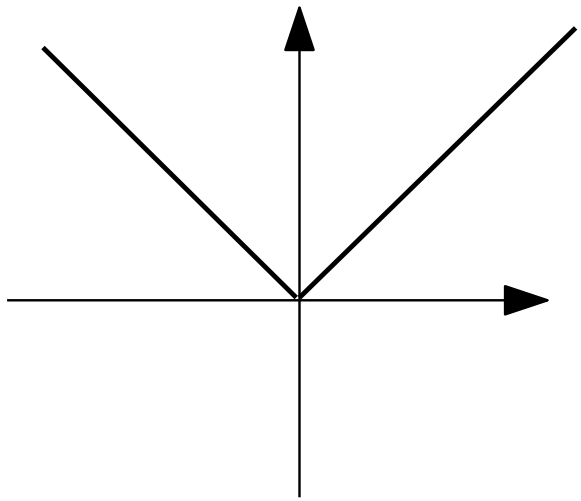
Def. The set of all subgradients at y denoted by $\partial f(y)$. This set is called **subdifferential** of f at y

If f is convex, $\partial f(x)$ is nice:

- ♣ If $x \in$ relative interior of dom f, then $\partial f(x)$ nonempty
- ♣ If f differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$
- ♣ If $\partial f(x) = \{g\}$, then f is differentiable and $g = \nabla f(x)$

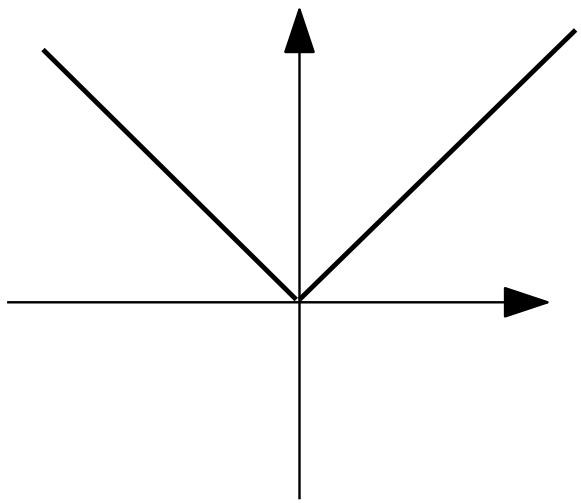
Subdifferential – example

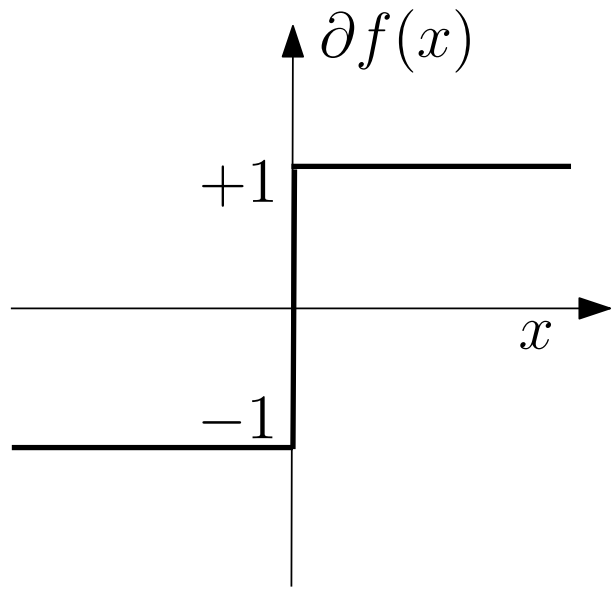
$$f(x) = |x|$$



Subdifferential – example

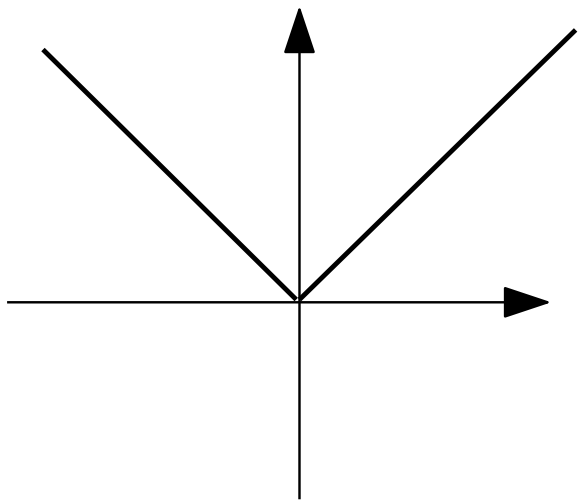
$$f(x) = |x|$$

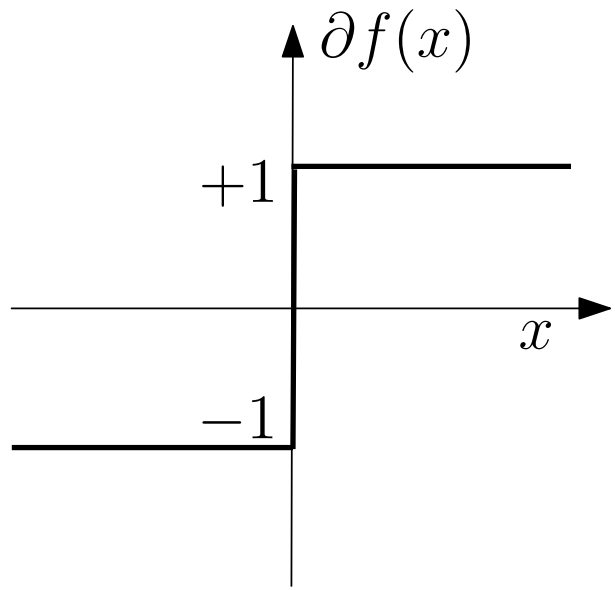




Subdifferential – example

$$f(x) = |x|$$





$$\partial |x| = \begin{cases} -1 & x < 0, \\ +1 & x > 0, \\ [-1, 1] & x = 0. \end{cases}$$

More examples

Example $f(x) = \|x\|_2$. Then,

$$\partial f(x) := \begin{cases} x/\|x\|_2 & x \neq 0, \\ \{z \mid \|z\|_2 \leq 1\} & x = 0. \end{cases}$$

More examples

Example $f(x) = \|x\|_2$. Then,

$$\partial f(x) := \begin{cases} x/\|x\|_2 & x \neq 0, \\ \{z \mid \|z\|_2 \leq 1\} & x = 0. \end{cases}$$

Proof.

$$\|z\|_2 \quad \geq \quad \|x\|_2 + \langle g, z - x \rangle$$

More examples

Example $f(x) = \|x\|_2$. Then,

$$\partial f(x) := \begin{cases} x/\|x\|_2 & x \neq 0, \\ \{z \mid \|z\|_2 \leq 1\} & x = 0. \end{cases}$$

Proof.

$$\begin{aligned} \|z\|_2 &\geq \|x\|_2 + \langle g, z - x \rangle \\ \|z\|_2 &\geq \langle g, z \rangle \end{aligned}$$

More examples

Example $f(x) = \|x\|_2$. Then,

$$\partial f(x) := \begin{cases} x/\|x\|_2 & x \neq 0, \\ \{z \mid \|z\|_2 \leq 1\} & x = 0. \end{cases}$$

Proof.

$$\begin{aligned} \|z\|_2 &\geq \|x\|_2 + \langle g, z - x \rangle \\ \|z\|_2 &\geq \langle g, z \rangle \\ &\implies \|g\|_2 \leq 1. \end{aligned}$$

Example

Example A convex function need not be subdifferentiable everywhere. Let

$$f(x) := \begin{cases} -(1 - \|x\|_2^2)^{1/2} & \text{if } \|x\|_2 \leq 1, \\ +\infty & \text{otherwise.} \end{cases}$$

f diff. for all x with $\|x\|_2 < 1$, but $\partial f(x) = \emptyset$ whenever $\|x\|_2 \geq 1$.

Subdifferential calculus

- If f is differentiable, $\partial f(x) = \{\nabla f(x)\}$
- **Scaling** $\alpha > 0$, $\partial(\alpha f)(x) = \alpha \partial f(x) = \{\alpha g \mid g \in \partial f(x)\}$
- **Addition***: $\partial(f + k)(x) = \partial f(x) + \partial k(x)$ (set addition)
- **Chain rule***: Let $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $f : \mathbb{R}^m \to \mathbb{R}$, and $h : \mathbb{R}^n \to \mathbb{R}$ be given by $h(x) = f(Ax + b)$. Then,

$$\partial h(x) = A^T \partial f(Ax + b).$$

- **Chain rule***: $h(x) = f \circ k$, where $k : X \to Y$ is diff.

$$\partial h(x) = \partial f(k(x)) \circ Dk(x) = [Dk(x)]^T \partial f(k(x))$$

- **Max function***: If $f(x) := \max_{1 \le i \le m} f_i(x)$, then

$$\partial f(x) = \operatorname{conv} \bigcup \{\partial f_i(x) \mid f_i(x) = f(x)\},$$

  convex hull over subdifferentials of "active" functions at x
- **Conjugation:** $z \in \partial f(x)$ if and only if $x \in \partial f^*(z)$

* — can fail to hold without precise assumptions.

Example

It can happen that $\partial(f_1 + f_2) \neq \partial f_1 + \partial f_2$

Example

It can happen that $\partial(f_1 + f_2) \neq \partial f_1 + \partial f_2$

Example Define f_1 and f_2 by

$$f_1(x) := \begin{cases} -2\sqrt{x} & \text{if } x \geq 0, \\ +\infty & \text{if } x < 0, \end{cases} \quad \text{and} \quad f_2(x) := \begin{cases} +\infty & \text{if } x > 0, \\ -2\sqrt{-x} & \text{if } x \leq 0. \end{cases}$$

Then, $f = \max\{f_1, f_2\} = \mathbb{1}_{\{0\}}$, whereby $\partial f(0) = \mathbb{R}$
But $\partial f_1(0) = \partial f_2(0) = \emptyset$.

Example

It can happen that $\partial(f_1 + f_2) \neq \partial f_1 + \partial f_2$

Example Define f_1 and f_2 by

$$f_1(x) := \begin{cases} -2\sqrt{x} & \text{if } x \geq 0, \\ +\infty & \text{if } x < 0, \end{cases} \quad \text{and} \quad f_2(x) := \begin{cases} +\infty & \text{if } x > 0, \\ -2\sqrt{-x} & \text{if } x \leq 0. \end{cases}$$

Then, $f = \max\{f_1, f_2\} = \mathbb{1}_{\{0\}}$, whereby $\partial f(0) = \mathbb{R}$
But $\partial f_1(0) = \partial f_2(0) = \emptyset$.

However, $\partial f_1(x) + \partial f_2(x) \subset \partial(f_1 + f_2)(x)$ always holds.

Optimality – constrained

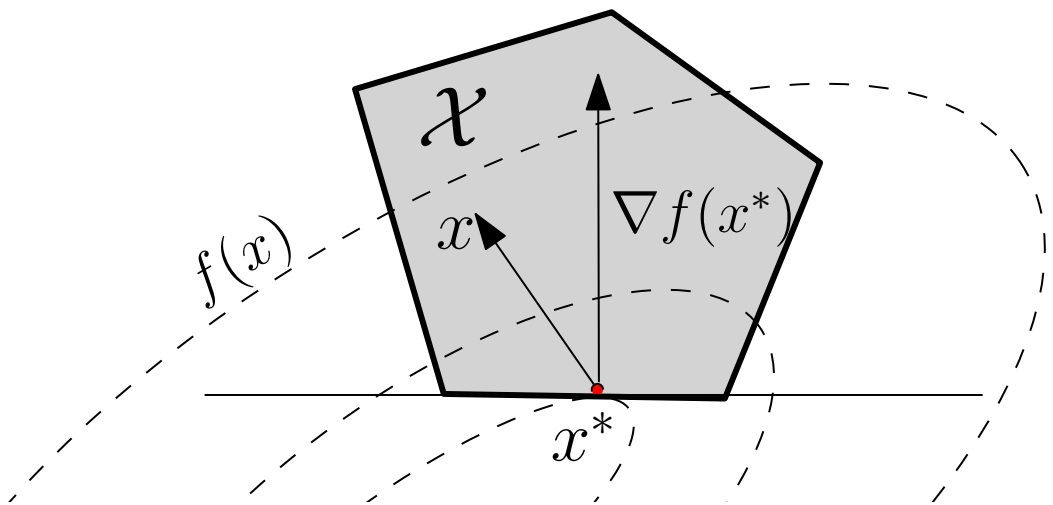
♠ For every $x, y \in \operatorname{dom} f$, we have $f(y) \geq f(x) + \langle \nabla f(x),\ y - x \rangle$.

Optimality – constrained

- For every $x, y \in \operatorname{dom} f$, we have $f(y) \geq f(x) + \langle \nabla f(x),\ y - x \rangle$.
- Thus, x^* is optimal **if** and only if

$$\langle \nabla f(x^*),\ y - x^* \rangle \geq 0, \qquad \text{for all } \ y \in \mathcal{X}.$$

- If $\mathcal{X} = \mathbb{R}^n$, this reduces to $\nabla f(x^*) = 0$



- If $\nabla f(x^*) \neq 0$, it defines supporting hyperplane to $\mathcal{X}$ at x^*

Optimality – nonsmooth

Theorem (Fermat's rule): Let $f : \mathbb{R}^n \to (-\infty, +\infty]$. Then,

$$\operatorname{argmin} f = \operatorname{zer}(\partial f) := \left\{ x \in \mathbb{R}^n \mid 0 \in \partial f(x) \right\}.$$

Optimality – nonsmooth

Theorem (Fermat's rule): Let $f : \mathbb{R}^n \to (-\infty, +\infty]$. Then,

$$\operatorname{argmin} f = \operatorname{zer}(\partial f) := \left\{ x \in \mathbb{R}^n \mid 0 \in \partial f(x) \right\}.$$

Proof: $x \in \operatorname{argmin} f$ implies that $f(x) \leq f(y)$ for all $y \in \mathbb{R}^n$.

Optimality – nonsmooth

Theorem (Fermat's rule): Let $f : \mathbb{R}^n \to (-\infty, +\infty]$. Then,

$$\operatorname{argmin} f = \operatorname{zer}(\partial f) := \left\{ x \in \mathbb{R}^n \mid 0 \in \partial f(x) \right\}.$$

Proof: $x \in \operatorname{argmin} f$ implies that $f(x) \leq f(y)$ for all $y \in \mathbb{R}^n$.
Equivalently, $f(y) \geq f(x) + \langle 0,\, y - x \rangle \quad \forall y$,

Optimality – nonsmooth

Theorem (Fermat's rule): Let $f : \mathbb{R}^n \to (-\infty, +\infty]$. Then,

$$\operatorname{argmin} f = \operatorname{zer}(\partial f) := \left\{ x \in \mathbb{R}^n \mid 0 \in \partial f(x) \right\}.$$

Proof: $x \in \operatorname{argmin} f$ implies that $f(x) \leq f(y)$ for all $y \in \mathbb{R}^n$.
Equivalently, $f(y) \geq f(x) + \langle 0,\, y - x \rangle \quad \forall y, \leftrightarrow 0 \in \partial f(x)$.

Optimality – nonsmooth

Theorem (Fermat's rule): Let $f : \mathbb{R}^n \to (-\infty, +\infty]$. Then,

$$\operatorname{argmin} f = \operatorname{zer}(\partial f) := \left\{ x \in \mathbb{R}^n \mid 0 \in \partial f(x) \right\}.$$

Proof: $x \in \operatorname{argmin} f$ implies that $f(x) \leq f(y)$ for all $y \in \mathbb{R}^n$.
Equivalently, $f(y) \geq f(x) + \langle 0,\, y - x \rangle \quad \forall y, \leftrightarrow 0 \in \partial f(x)$.

Nonsmooth optimality

$$\begin{array}{lll} \min & f(x) & \text{s.t. } x \in \mathcal{X} \\ \min & f(x) + \mathbb{1}_{\mathcal{X}}(x). & \end{array}$$

Optimality – nonsmooth

- Minimizing x must satisfy: $0 \in \partial(f + \mathbb{1}_{\mathcal{X}})(x)$

Optimality – nonsmooth

- Minimizing x must satisfy: $0 \in \partial(f + \mathbb{1}_{\mathcal{X}})(x)$
- (**CQ**) Assuming $\operatorname{ri}(\operatorname{dom} f) \cap \operatorname{ri}(\mathcal{X}) \neq \emptyset$, $0 \in \partial f(x) + \partial \mathbb{1}_{\mathcal{X}}(x)$

Optimality – nonsmooth

- Minimizing x must satisfy: $0 \in \partial(f + \mathbb{1}_{\mathcal{X}})(x)$
- (**CQ**) Assuming $\operatorname{ri}(\operatorname{dom} f) \cap \operatorname{ri}(\mathcal{X}) \neq \emptyset$, $0 \in \partial f(x) + \partial \mathbb{1}_{\mathcal{X}}(x)$
- Recall, $g \in \partial \mathbb{1}_{\mathcal{X}}(x)$ iff $\mathbb{1}_{\mathcal{X}}(y) \geq \mathbb{1}_{\mathcal{X}}(x) + \langle g, y - x \rangle$ for all y.

Optimality – nonsmooth

- Minimizing x must satisfy: $0 \in \partial(f + \mathbb{1}_{\mathcal{X}})(x)$
- (**CQ**) Assuming $\operatorname{ri}(\operatorname{dom} f) \cap \operatorname{ri}(\mathcal{X}) \neq \emptyset$, $0 \in \partial f(x) + \partial \mathbb{1}_{\mathcal{X}}(x)$
- Recall, $g \in \partial \mathbb{1}_{\mathcal{X}}(x)$ iff $\mathbb{1}_{\mathcal{X}}(y) \geq \mathbb{1}_{\mathcal{X}}(x) + \langle g,\, y - x \rangle$ for all y.
- So $g \in \partial \mathbb{1}_{\mathcal{X}}(x)$ means $x \in \mathcal{X}$ and $0 \geq \langle g,\, y - x \rangle \; \forall y \in \mathcal{X}$.

Optimality – nonsmooth

- Minimizing x must satisfy: $0 \in \partial(f + \mathbb{1}_{\mathcal{X}})(x)$
- (**CQ**) Assuming $\operatorname{ri}(\operatorname{dom} f) \cap \operatorname{ri}(\mathcal{X}) \neq \emptyset$, $0 \in \partial f(x) + \partial \mathbb{1}_{\mathcal{X}}(x)$
- Recall, $g \in \partial \mathbb{1}_{\mathcal{X}}(x)$ iff $\mathbb{1}_{\mathcal{X}}(y) \geq \mathbb{1}_{\mathcal{X}}(x) + \langle g, y - x \rangle$ for all y.
- So $g \in \partial \mathbb{1}_{\mathcal{X}}(x)$ means $x \in \mathcal{X}$ and $0 \geq \langle g, y - x \rangle \; \forall y \in \mathcal{X}$.
- **Normal cone:**

$$\mathcal{N}_{\mathcal{X}}(x) := \left\{ g \in \mathbb{R}^n \mid 0 \geq \langle g, y - x \rangle \quad \forall y \in \mathcal{X} \right\}$$

Optimality – nonsmooth

- Minimizing x must satisfy: $0 \in \partial(f + \mathbb{1}_{\mathcal{X}})(x)$
- (**CQ**) Assuming $\operatorname{ri}(\operatorname{dom} f) \cap \operatorname{ri}(\mathcal{X}) \neq \emptyset$, $0 \in \partial f(x) + \partial \mathbb{1}_{\mathcal{X}}(x)$
- Recall, $g \in \partial \mathbb{1}_{\mathcal{X}}(x)$ iff $\mathbb{1}_{\mathcal{X}}(y) \geq \mathbb{1}_{\mathcal{X}}(x) + \langle g,\, y - x \rangle$ for all y.
- So $g \in \partial \mathbb{1}_{\mathcal{X}}(x)$ means $x \in \mathcal{X}$ and $0 \geq \langle g,\, y - x \rangle \; \forall y \in \mathcal{X}$.
- **Normal cone:**

$$\mathcal{N}_{\mathcal{X}}(x) := \left\{ g \in \mathbb{R}^n \mid 0 \geq \langle g,\, y - x \rangle \quad \forall y \in \mathcal{X} \right\}$$

Application. $\min f(x) \quad \text{s.t. } x \in \mathcal{X}$:

♢ If f is diff., we get $0 \in \nabla f(x^*) + \mathcal{N}_{\mathcal{X}}(x^*)$

Optimality – nonsmooth

- Minimizing x must satisfy: $0 \in \partial(f + \mathbb{1}_{\mathcal{X}})(x)$
- (**CQ**) Assuming $\operatorname{ri}(\operatorname{dom} f) \cap \operatorname{ri}(\mathcal{X}) \neq \emptyset$, $0 \in \partial f(x) + \partial \mathbb{1}_{\mathcal{X}}(x)$
- Recall, $g \in \partial \mathbb{1}_{\mathcal{X}}(x)$ iff $\mathbb{1}_{\mathcal{X}}(y) \geq \mathbb{1}_{\mathcal{X}}(x) + \langle g,\, y - x \rangle$ for all y.
- So $g \in \partial \mathbb{1}_{\mathcal{X}}(x)$ means $x \in \mathcal{X}$ and $0 \geq \langle g,\, y - x \rangle \; \forall y \in \mathcal{X}$.
- **Normal cone:**

$$\mathcal{N}_{\mathcal{X}}(x) := \left\{ g \in \mathbb{R}^n \mid 0 \geq \langle g,\, y - x \rangle \quad \forall y \in \mathcal{X} \right\}$$

Application. $\min f(x) \quad \text{s.t. } x \in \mathcal{X}$:

♢ If f is diff., we get $0 \in \nabla f(x^*) + \mathcal{N}_{\mathcal{X}}(x^*)$

♢ $-\nabla f(x^*) \in \mathcal{N}_{\mathcal{X}}(x^*) \iff \langle \nabla f(x^*),\, y - x^* \rangle \geq 0$ for all $y \in \mathcal{X}$.

Problems

Regularized optimization

$$\inf_{x \in \mathcal{X}} \quad f(x) + r(Ax) \quad \text{s.t.} \;\; Ax \in \mathcal{Y}.$$

Regularized optimization

$$\inf_{x \in \mathcal{X}} \quad f(x) + r(Ax) \quad \text{s.t.} \;\; Ax \in \mathcal{Y}.$$

Dual problem

$$\inf_{u \in \mathcal{Y}} \quad f^*(-A^T u) + r^*(u).$$

Regularized optimization

$$\inf_{x \in \mathcal{X}} \quad f(x) + r(Ax) \quad \text{s.t.} \;\; Ax \in \mathcal{Y}.$$

Dual problem

$$\inf_{u \in \mathcal{Y}} \quad f^*(-A^T u) + r^*(u).$$

- Introduce new variable $z = Ax$

$$\inf_{x \in \mathcal{X}, z \in \mathcal{Y}} \quad f(x) + r(z), \qquad \text{s.t.} \quad z = Ax.$$

Regularized optimization

$$\inf_{x \in \mathcal{X}} \quad f(x) + r(Ax) \quad \text{s.t.} \;\; Ax \in \mathcal{Y}.$$

Dual problem

$$\inf_{u \in \mathcal{Y}} \quad f^*(-A^T u) + r^*(u).$$

- Introduce new variable $z = Ax$

$$\inf_{x \in \mathcal{X}, z \in \mathcal{Y}} \quad f(x) + r(z), \qquad \text{s.t.} \quad z = Ax.$$

- The (partial)-Lagrangian is

$$L(x, z; u) := f(x) + r(z) + u^T(Ax - z), \quad x \in \mathcal{X}, z \in \mathcal{Y};$$

Regularized optimization

$$\inf_{x \in \mathcal{X}} \quad f(x) + r(Ax) \quad \text{s.t.} \;\; Ax \in \mathcal{Y}.$$

Dual problem

$$\inf_{u \in \mathcal{Y}} \quad f^*(-A^T u) + r^*(u).$$

- Introduce new variable $z = Ax$

$$\inf_{x \in \mathcal{X}, z \in \mathcal{Y}} \quad f(x) + r(z), \qquad \text{s.t.} \quad z = Ax.$$

- The (partial)-Lagrangian is

$$L(x, z; u) := f(x) + r(z) + u^T(Ax - z), \quad x \in \mathcal{X}, z \in \mathcal{Y};$$

- Associated dual function

$$g(u) := \inf_{x \in \mathcal{X}, z \in \mathcal{Y}} L(x, z; u).$$

Regularized optimization

$$\inf_{x \in \mathcal{X}} \quad f(x) + r(Ax) \quad \text{s.t.} \;\; Ax \in \mathcal{Y}.$$

Dual problem

$$\inf_{y \in \mathcal{Y}} \quad f^*(-A^T y) + r^*(y).$$

The infimum above can be rearranged as follows

$$g(y) \;\; = \;\; \inf_{x \in \mathcal{X}} \; f(x) + y^T A x + \inf_{z \in \mathcal{Y}} \; r(z) - y^T z$$

Regularized optimization

$$\inf_{x \in \mathcal{X}} \quad f(x) + r(Ax) \quad \text{s.t.} \;\; Ax \in \mathcal{Y}.$$

Dual problem

$$\inf_{y \in \mathcal{Y}} \quad f^*(-A^T y) + r^*(y).$$

The infimum above can be rearranged as follows

$$\begin{aligned} g(y) &= \inf_{x \in \mathcal{X}} \; f(x) + y^T A x + \inf_{z \in \mathcal{Y}} \; r(z) - y^T z \\ &= -\sup_{x \in \mathcal{X}} \left\{ -x^T A^T y - f(x) \right\} - \sup_{z \in \mathcal{Y}} \left\{ z^T y - r(z) \right\} \end{aligned}$$

Regularized optimization

$$\inf_{x \in \mathcal{X}} \quad f(x) + r(Ax) \quad \text{s.t.} \;\; Ax \in \mathcal{Y}.$$

Dual problem

$$\inf_{y \in \mathcal{Y}} \quad f^*(-A^T y) + r^*(y).$$

The infimum above can be rearranged as follows

$$\begin{aligned} g(y) &= \inf_{x \in \mathcal{X}} \; f(x) + y^T A x + \inf_{z \in \mathcal{Y}} \; r(z) - y^T z \\ &= -\sup_{x \in \mathcal{X}} \left\{ -x^T A^T y - f(x) \right\} - \sup_{z \in \mathcal{Y}} \left\{ z^T y - r(z) \right\} \\ &= -f^*(-A^T y) - r^*(y) \quad \text{s.t. } y \in \mathcal{Y}. \end{aligned}$$

Regularized optimization

$$\inf_{x \in \mathcal{X}} \quad f(x) + r(Ax) \quad \text{s.t.} \;\; Ax \in \mathcal{Y}.$$

Dual problem

$$\inf_{y \in \mathcal{Y}} \quad f^*(-A^T y) + r^*(y).$$

The infimum above can be rearranged as follows

$$\begin{aligned} g(y) &= \inf_{x \in \mathcal{X}} \; f(x) + y^T A x + \inf_{z \in \mathcal{Y}} \; r(z) - y^T z \\ &= -\sup_{x \in \mathcal{X}} \left\{ -x^T A^T y - f(x) \right\} - \sup_{z \in \mathcal{Y}} \left\{ z^T y - r(z) \right\} \\ &= -f^*(-A^T y) - r^*(y) \quad \text{s.t. } y \in \mathcal{Y}. \end{aligned}$$

Dual problem computes $\sup_{u \in \mathcal{Y}} g(u)$; so equivalently,

$$\inf_{y \in \mathcal{Y}} \quad f^*(-A^T y) + r^*(y).$$

Regularized optimization

Strong duality

$$\inf_x \{f(x) + r(Ax)\} = \sup_y \left\{-f^*(-A^T y) + r^*(y)\right\}$$

if either of the following conditions holds:

Regularized optimization

Strong duality

$$\inf_x \{f(x) + r(Ax)\} = \sup_y \left\{-f^*(-A^T y) + r^*(y)\right\}$$

if either of the following conditions holds:

1. $\exists x \in \operatorname{ri}(\operatorname{dom} f)$ such that $Ax \in \operatorname{ri}(\operatorname{dom} r)$
2. $\exists y \in \operatorname{ri}(\operatorname{dom} r^*)$ such that $A^T y \in \operatorname{ri}(\operatorname{dom} f^*)$

Regularized optimization

Strong duality

$$\inf_x \{f(x) + r(Ax)\} = \sup_y \left\{-f^*(-A^T y) + r^*(y)\right\}$$

if either of the following conditions holds:

1. $\exists x \in \operatorname{ri}(\operatorname{dom} f)$ such that $Ax \in \operatorname{ri}(\operatorname{dom} r)$
2. $\exists y \in \operatorname{ri}(\operatorname{dom} r^*)$ such that $A^T y \in \operatorname{ri}(\operatorname{dom} f^*)$

- Condition 1 ensures 'inf' attained at some x
- Condition 2 ensures 'sup' attained at some y

Example: norm regularized problems

$$\min \quad f(x) + \|Ax\|$$

Example: norm regularized problems

$$\min \quad f(x) + \|Ax\|$$

Dual problem

$$\min_y \quad f^*(-A^T y) \quad \text{s.t. } \|y\|_* \leq 1.$$

Example: norm regularized problems

$$\min \quad f(x) + \|Ax\|$$

Dual problem

$$\min_y \quad f^*(-A^T y) \quad \text{s.t. } \|y\|_* \leq 1.$$

Say $\|\bar{y}\|_* < 1$, such that $A^T \bar{y} \in \text{ri}(\text{dom } f^*)$, then we have strong duality (e.g., for instance $0 \in \text{ri}(\text{dom } f^*)$)

Example: Lasso-like problem

$$p^* := \min_x \quad \|Ax - b\|_2 + \lambda \|x\|_1.$$

Example: Lasso-like problem

$$p^* := \min_x \quad \|Ax - b\|_2 + \lambda \|x\|_1.$$

$$\|x\|_1 = \max \left\{ x^T v \mid \|v\|_\infty \leq 1 \right\}$$

$$\|x\|_2 = \max \left\{ x^T u \mid \|u\|_2 \leq 1 \right\}.$$

Example: Lasso-like problem

$$p^* := \min_x \quad \|Ax - b\|_2 + \lambda \|x\|_1.$$

$$\|x\|_1 = \max \left\{ x^T v \mid \|v\|_\infty \leq 1 \right\}$$

$$\|x\|_2 = \max \left\{ x^T u \mid \|u\|_2 \leq 1 \right\}.$$

Saddle-point formulation

$$p^* \quad = \quad \min_x \max_{u,v} \left\{ u^T(b - Ax) + v^T x \mid \|u\|_2 \leq 1, \quad \|v\|_\infty \leq \lambda \right\}$$

Example: Lasso-like problem

$$p^* := \min_x \quad \|Ax - b\|_2 + \lambda \|x\|_1.$$

$$\|x\|_1 = \max \left\{ x^T v \mid \|v\|_\infty \leq 1 \right\}$$

$$\|x\|_2 = \max \left\{ x^T u \mid \|u\|_2 \leq 1 \right\}.$$

Saddle-point formulation

$$\begin{aligned} p^* &= \min_x \max_{u,v} \left\{ u^T(b - Ax) + v^T x \mid \|u\|_2 \leq 1, \quad \|v\|_\infty \leq \lambda \right\} \\ &= \max_{u,v} \min_x \left\{ u^T(b - Ax) + x^T v \mid \|u\|_2 \leq 1, \quad \|v\|_\infty \leq \lambda \right\} \end{aligned}$$

Example: Lasso-like problem

$$p^* := \min_x \quad \|Ax - b\|_2 + \lambda \|x\|_1.$$

$$\|x\|_1 = \max \left\{ x^T v \mid \|v\|_\infty \leq 1 \right\}$$

$$\|x\|_2 = \max \left\{ x^T u \mid \|u\|_2 \leq 1 \right\}.$$

Saddle-point formulation

$$\begin{aligned} p^* &= \min_x \max_{u,v} \left\{ u^T(b - Ax) + v^T x \mid \|u\|_2 \leq 1, \ \ \|v\|_\infty \leq \lambda \right\} \\ &= \max_{u,v} \min_x \left\{ u^T(b - Ax) + x^T v \mid \|u\|_2 \leq 1, \ \ \|v\|_\infty \leq \lambda \right\} \\ &= \max_{u,v} u^T b \qquad A^T u = v, \ \|u\|_2 \leq 1, \ \ \|v\|_\infty \leq \lambda \end{aligned}$$

Example: Lasso-like problem

$$p^* := \min_x \quad \|Ax - b\|_2 + \lambda \|x\|_1.$$

$$\|x\|_1 = \max \left\{ x^T v \mid \|v\|_\infty \leq 1 \right\}$$

$$\|x\|_2 = \max \left\{ x^T u \mid \|u\|_2 \leq 1 \right\}.$$

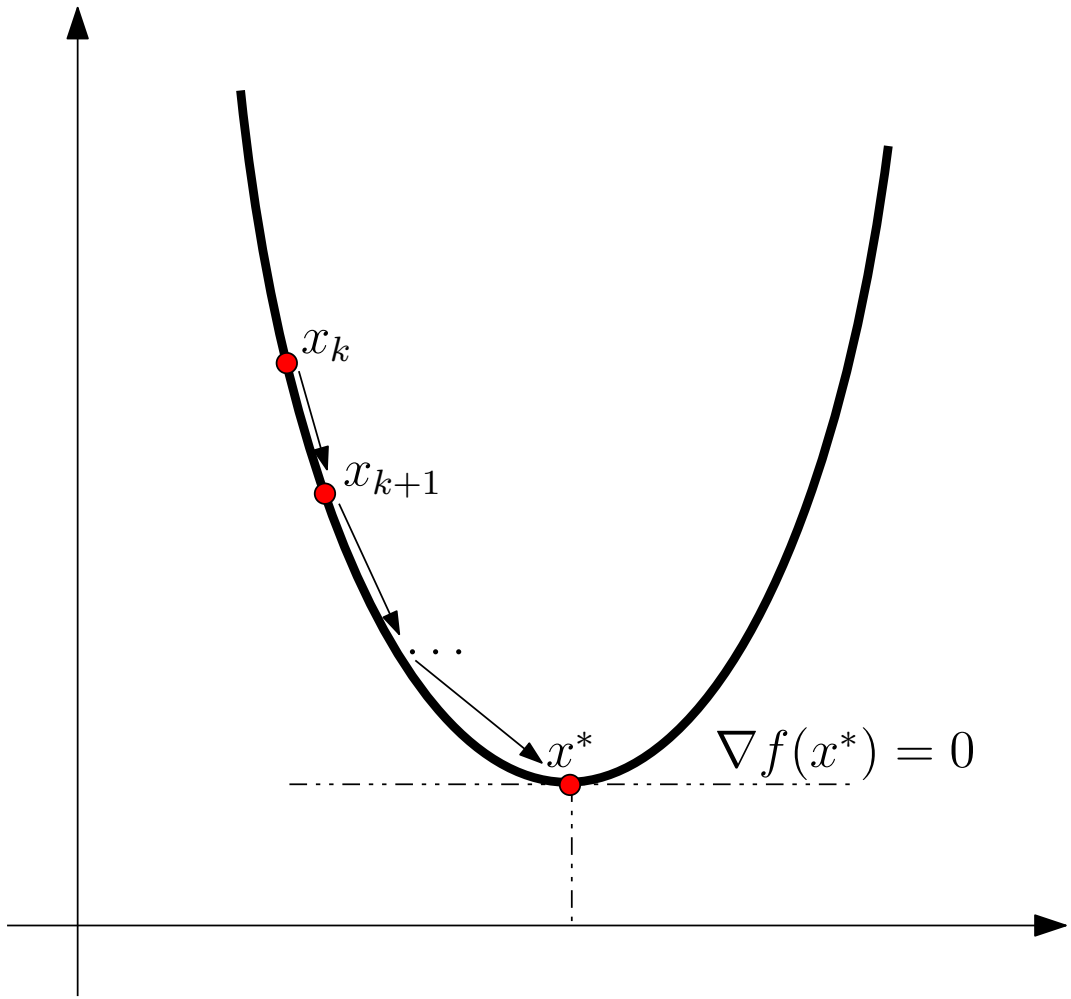
Saddle-point formulation

$$\begin{aligned}
p^* &= \min_x \max_{u,v} \left\{ u^T(b - Ax) + v^T x \mid \|u\|_2 \leq 1, \ \ \|v\|_\infty \leq \lambda \right\} \\
&= \max_{u,v} \min_x \left\{ u^T(b - Ax) + x^T v \mid \|u\|_2 \leq 1, \ \ \|v\|_\infty \leq \lambda \right\} \\
&= \max_{u,v} u^T b \qquad A^T u = v, \ \|u\|_2 \leq 1, \ \ \|v\|_\infty \leq \lambda \\
&= \max_u u^T b \qquad \|u\|_2 \leq 1, \quad \|A^T v\|_\infty \leq \lambda.
\end{aligned}$$

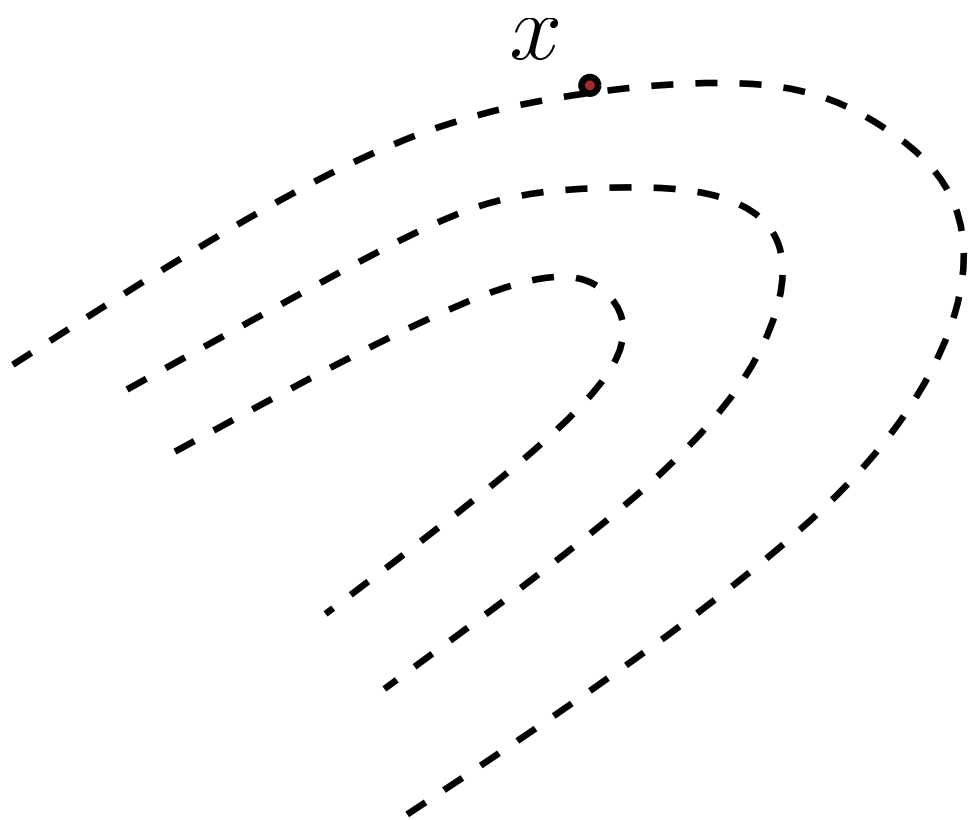
Algorithms

Descent methods

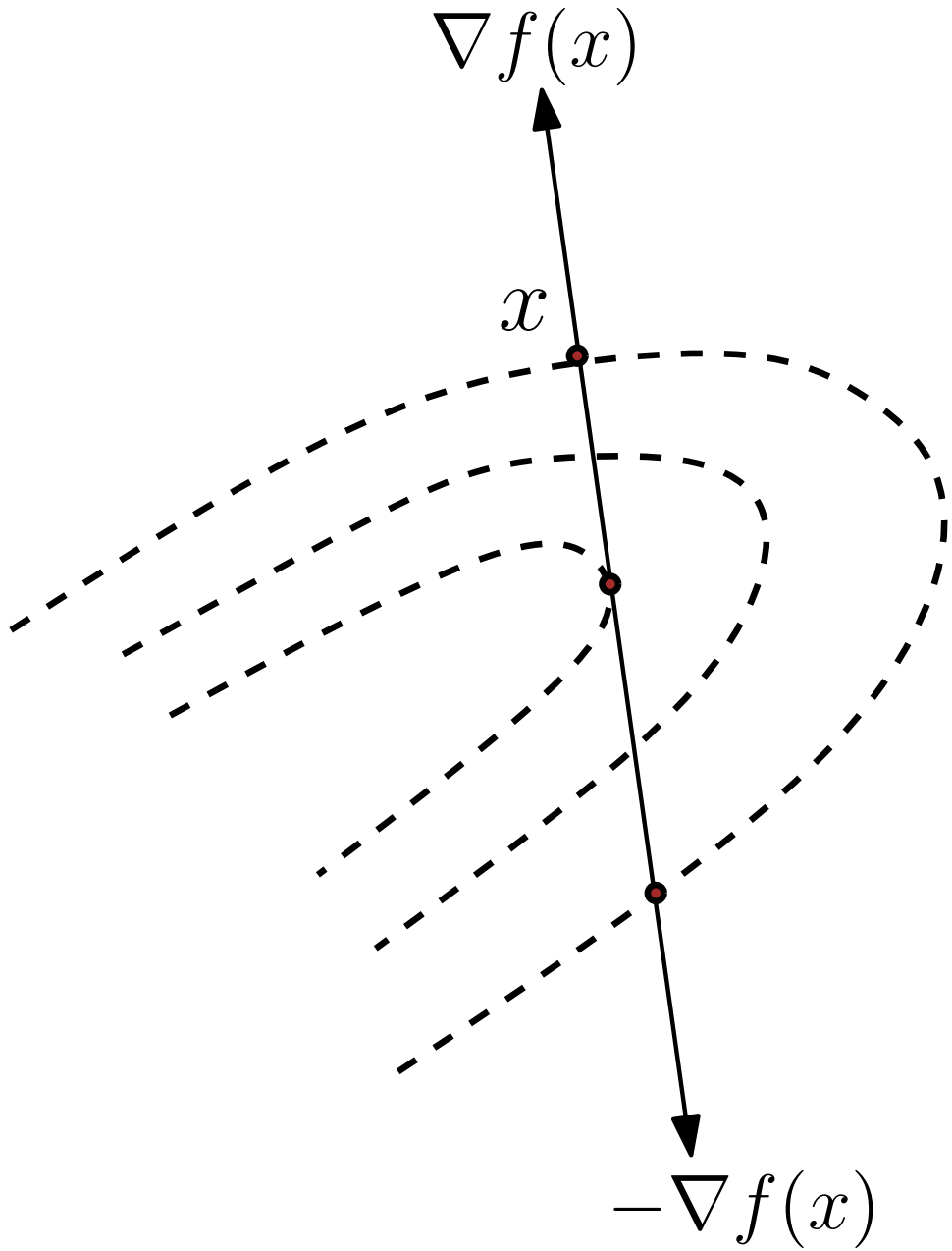
$$\min_x \quad f(x)$$



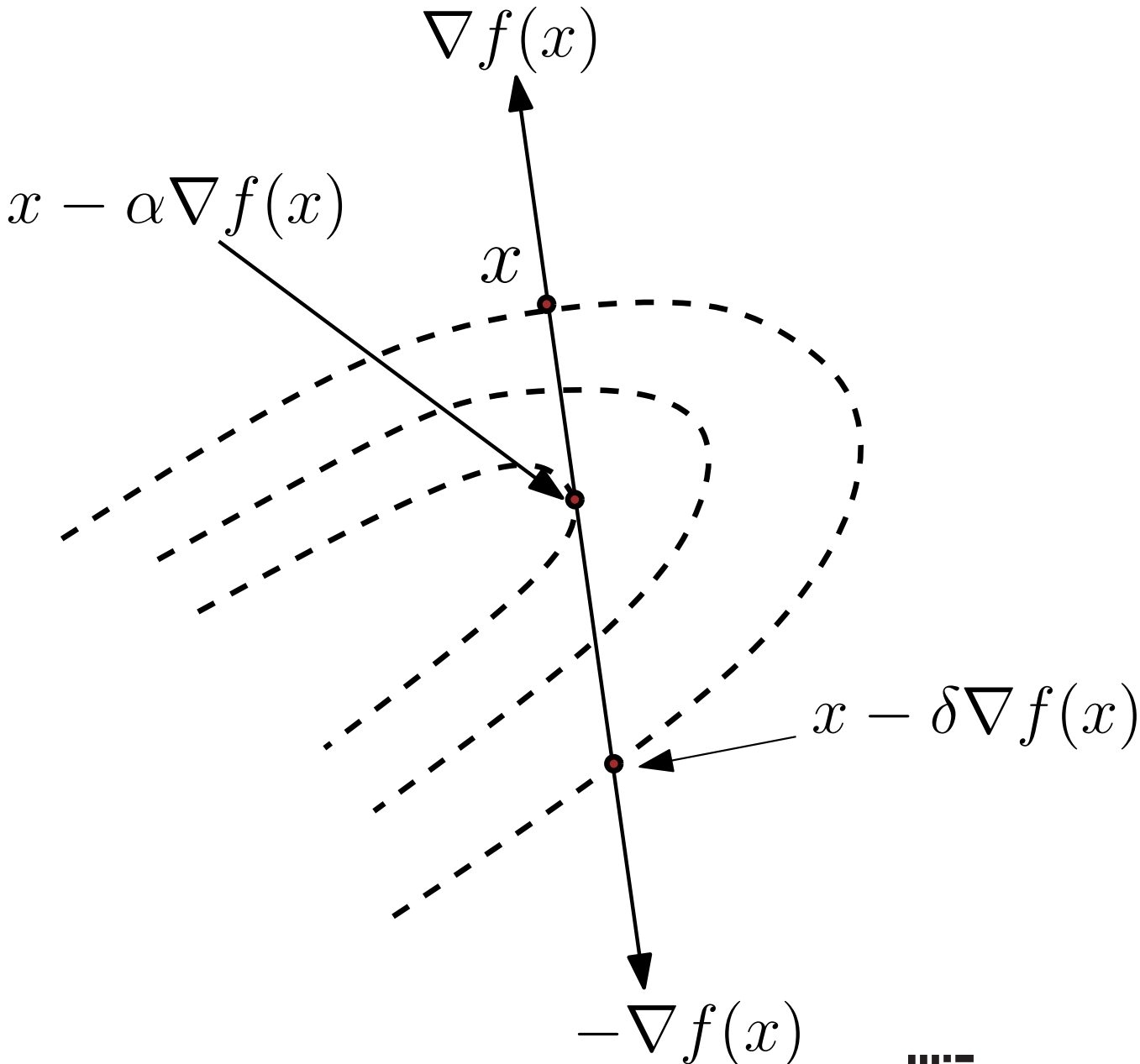
Descent methods



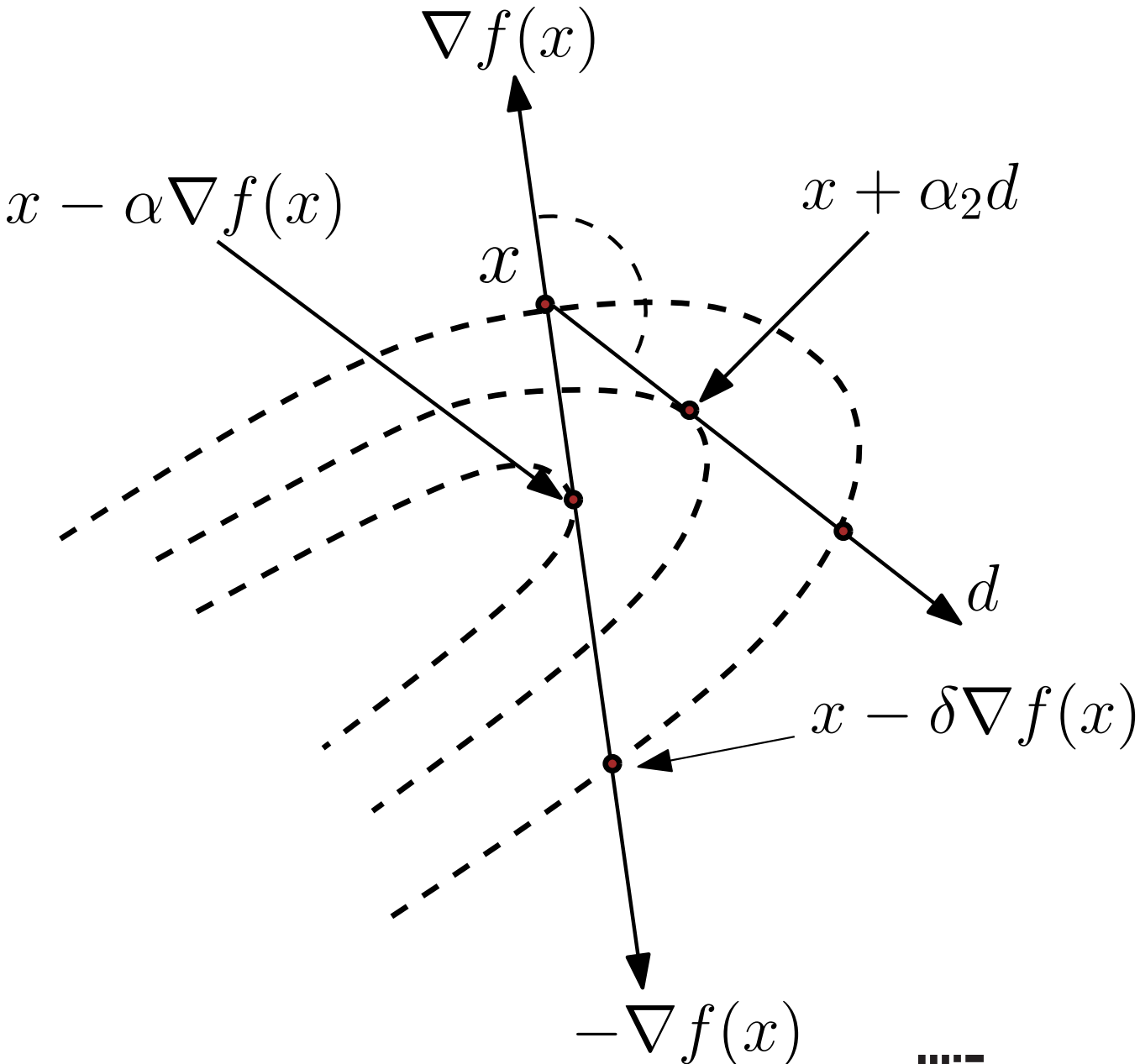
Descent methods



Descent methods



Descent methods



Algorithm

1. Start with some guess x^0;
2. For each $k = 0, 1, \ldots$
   - $x^{k+1} \leftarrow x^k + \alpha_k d^k$
   - Check when to stop (e.g., if $\nabla f(x^{k+1}) = 0$)

Gradient methods

$$x^{k+1} = x^k + \alpha_k d^k, \quad k = 0, 1, \ldots$$

- **stepsize** $\alpha_k \geq 0$, usually ensures $f(x^{k+1}) < f(x^k)$

Gradient methods

$$x^{k+1} = x^k + \alpha_k d^k, \quad k = 0, 1, \ldots$$

- **stepsize** $\alpha_k \geq 0$, usually ensures $f(x^{k+1}) < f(x^k)$
- **Descent direction** d^k satisfies

$$\langle \nabla f(x^k),\, d^k \rangle < 0$$

Numerous ways to select α_k and d^k

Usually methods **seek monotonic descent**

$$f(x^{k+1}) < f(x^k)$$

Giving up on this monotonicity led to breakthroughs!

Gradient methods – direction

$$x^{k+1} = x^k + \alpha_k d^k, \quad k = 0, 1, \ldots$$

- Different choices of direction d^k
  - **Scaled gradient:** $d^k = -D^k \nabla f(x^k)$, $D^k \succ 0$
  - **Newton's method:** $(D^k = [\nabla^2 f(x^k)]^{-1})$
  - **Quasi-Newton:** $D^k \approx [\nabla^2 f(x^k)]^{-1}$
  - **Steepest descent:** $D^k = \boldsymbol{I}$
  - **Diagonally scaled:** D^k diagonal with $D^k_{ii} \approx \left(\frac{\partial^2 f(x^k)}{(\partial x_i)^2} \right)^{-1}$
  - **Discretized Newton:** $D^k = [H(x^k)]^{-1}$, H via finite-diff.
  - ...

Gradient methods – stepsize

- **Exact:** $\alpha_k := \underset{\alpha \geq 0}{\operatorname{argmin}}\, f(x^k + \alpha d^k)$
- **Limited min:** $\alpha_k = \underset{0 \leq \alpha \leq s}{\operatorname{argmin}}\, f(x^k + \alpha d^k)$
- **Armijo-rule**. Given **fixed** scalars, s, β, σ with $0 < \beta < 1$ and $0 < \sigma < 1$ (chosen experimentally). Set $\alpha_k = \beta^{m_k} s$ where we **try** $\beta^m s$ for $m = 0, 1, \ldots$ until **sufficient descent**

$$f(x^k) - f(x + \beta^m s d^k) \geq -\sigma \beta^m s \langle \nabla f(x^k),\, d^k \rangle$$

  If $\langle \nabla f(x^k),\, d^k \rangle < 0$, stepsize guaranteed to exist
  Usually, σ small $\in [10^{-5}, 0.1]$, while β from $1/2$ to $1/10$ depending on how confident we are about initial stepsize s.
- **Constant:** $\alpha_k = 1/L$ (for suitable value of L)
- **Diminishing:** $\alpha_k \to 0$ but $\sum_k \alpha_k = \infty$.

Gradient-descent

Assumption: Lipschitz continuous gradient; denoted $f \in C^1_L$

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2$$

Gradient-descent

Assumption: Lipschitz continuous gradient; denoted $f \in C^1_L$

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2$$

- ♣ Gradient vectors of closeby points are close to each other
- ♣ Objective function has "bounded curvature"
- ♣ Speed at which gradient varies is bounded

Gradient-descent

Assumption: Lipschitz continuous gradient; denoted $f \in C_L^1$

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2$$

- Gradient vectors of closeby points are close to each other
- Objective function has "bounded curvature"
- Speed at which gradient varies is bounded

Lemma (Descent). Let $f \in C_L^1$. Then,

$$f(y) \leq f(x) + \langle \nabla f(x),\, y - x\rangle + \tfrac{L}{2}\|y - x\|_2^2$$

Theorem Let $f \in C_L^1$ and $\{x^k\}$ be sequence generated as above, with $\alpha_k = 1/L$. Then, $f(x^{k+1}) - f(x^*) = O(1/k)$.

Descent lemma

Proof. Since $f \in C^1_L$, by Taylor's theorem, for the vector $z_t = y + t(x - y)$ we have

$$f(x) = f(y) + \int_0^1 \langle \nabla f(z_t),\, x - y \rangle dt.$$

Descent lemma

Proof. Since $f \in C_L^1$, by Taylor's theorem, for the vector $z_t = y + t(x - y)$ we have

$$f(x) = f(y) + \int_0^1 \langle \nabla f(z_t),\, x - y \rangle dt.$$

Add and subtract $\langle \nabla f(y),\, x - y \rangle$ on rhs we have

$$\begin{aligned}
f(x) - f(y) - \langle \nabla f(y),\, x - y \rangle &= \int_0^1 \langle \nabla f(z_t) - \nabla f(y),\, x - y \rangle dt \\
|f(x) - f(y) - \langle \nabla f(y),\, x - y \rangle| &\leq \left| \int_0^1 \langle \nabla f(z_t) - \nabla f(y),\, x - y \rangle dt \right| \\
&\leq \int_0^1 |\langle \nabla f(z_t) - \nabla f(y),\, x - y \rangle| dt \\
&\leq \int_0^1 \|\nabla f(z_t) - \nabla f(y)\| \, \|x - y\| dt \\
&\leq L \int_0^1 t \|x - y\|^2 dt \\
&= \tfrac{L}{2} \|x - y\|^2.
\end{aligned}$$

Bounds $f(x)$ above and below with quadratic functions

Descent lemma – corollaries

Cor. 1 If $f \in C_L^1$, and $0 < \alpha_k < 2/L$, then $f(x^{k+1}) < f(x^k)$

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) + \langle \nabla f(x^k),\, x^{k+1} - x^k \rangle + \tfrac{L}{2}\|x^{k+1} - x^k\|_2 \\ &= f(x^k) - \alpha_k \|\nabla f(x^k)\|_2^2 + \tfrac{\alpha_k^2 L}{2}\|\nabla f(x^k)\|_2^2 \\ &= f(x^k) - \alpha_k(1 - \tfrac{\alpha_k}{2}L)\|\nabla f(x^k)\|_2^2 \end{aligned}$$

Thus, if $\alpha_k < 2/L$ we have descent. Minimize over α_k to get best bound: this yields $\alpha_k = 1/L$

Descent lemma – corollaries

Cor. 1 If $f \in C_L^1$, and $0 < \alpha_k < 2/L$, then $f(x^{k+1}) < f(x^k)$

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) + \langle \nabla f(x^k),\, x^{k+1} - x^k \rangle + \tfrac{L}{2}\|x^{k+1} - x^k\|_2 \\ &= f(x^k) - \alpha_k \|\nabla f(x^k)\|_2^2 + \tfrac{\alpha_k^2 L}{2}\|\nabla f(x^k)\|_2^2 \\ &= f(x^k) - \alpha_k(1 - \tfrac{\alpha_k}{2}L)\|\nabla f(x^k)\|_2^2 \end{aligned}$$

Thus, if $\alpha_k < 2/L$ we have descent. Minimize over α_k to get best bound: this yields $\alpha_k = 1/L$

Cor. 2 If $f \in C_L^1$, then

$$\langle f(x) - f(y),\, x - y \rangle \geq \tfrac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2$$

Linear convergence

Assumption: Strong convexity; denote $f \in S^1_{L,\mu}$

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \tfrac{\mu}{2}\|x - y\|_2^2$$

- Setting $\alpha_k = 2/(\mu + L)$ yields linear rate ($\mu > 0$)

Strongly convex – linear rate

Theorem. If $f \in S^1_{L,\mu}$, $0 < \alpha < 2/(L+\mu)$, then the gradient method generates a sequence $\{x^k\}$ that satisfies

$$\|x^k - x^*\|_2^2 \leq \left(1 - \frac{2\alpha\mu L}{\mu + L}\right)^k \|x^0 - x^*\|_2.$$

Moreover, if $\alpha = 2/(L+\mu)$ then

$$f(x^k) - f^* \leq \frac{L}{2}\left(\frac{\kappa - 1}{\kappa + 1}\right)^{2k} \|x^0 - x^*\|_2^2,$$

where $\kappa = L/\mu$ is the **condition number**.

Convergence – proof sketch

Thm 2. Suppose $f \in S^1_{L,\mu}$. Then, for any $x, y \in \mathbb{R}^n$

$$\langle \nabla f(x) - \nabla f(y),\, x - y \rangle \geq \frac{\mu L}{\mu + L} \|x - y\|_2^2 + \frac{1}{\mu + L} \|\nabla f(x) - \nabla f(y)\|_2^2$$

Proof. Recall descent lemma implies

$$\langle f(x) - f(y),\, x - y \rangle \geq \tfrac{1}{L} \|\nabla f(x) - \nabla f(y)\|^2$$

Apply this result to $\phi(x) = f(x) - \frac{\mu}{2}\|x\|^2$ with $L - \mu$.
$\nabla \phi(x) = \nabla f(x) - \mu x$

$$\langle \nabla \phi(x) - \nabla \phi(y),\, x - y \rangle \geq \frac{1}{L - \mu} \|\nabla \phi(x) - \nabla \phi(y)\|_2$$

Convergence – proof sketch

- Let $r_k = \|x^k - x^*\|_2$, and consider

Convergence – proof sketch

- Let $r_k = \|x^k - x^*\|_2$, and consider

$$r_{k+1}^2 \quad = \quad \|x^k - x^* - \alpha \nabla f(x^k)\|_2^2$$

Convergence – proof sketch

- Let $r_k = \|x^k - x^*\|_2$, and consider

$$\begin{aligned} r_{k+1}^2 &= \|x^k - x^* - \alpha \nabla f(x^k)\|_2^2 \\ &= r_k^2 - 2\alpha \langle \nabla f(x^k),\, x^k - x^* \rangle + \alpha^2 \|\nabla f(x^k)\|_2^2 \end{aligned}$$

Convergence – proof sketch

- Let $r_k = \|x^k - x^*\|_2$, and consider

$$\begin{aligned} r_{k+1}^2 &= \|x^k - x^* - \alpha \nabla f(x^k)\|_2^2 \\ &= r_k^2 - 2\alpha \langle \nabla f(x^k),\, x^k - x^* \rangle + \alpha^2 \|\nabla f(x^k)\|_2^2 \\ &\leq \left(1 - \frac{2\alpha\mu L}{\mu + L}\right) r_k^2 + \alpha \left(\alpha - \frac{2}{\mu + L}\right) \|\nabla f(x^k)\|_2^2 \end{aligned}$$

where we used Thm. 2 with $\nabla f(x^*) = 0$ for last inequality.

Convergence – proof sketch

- Let $r_k = \|x^k - x^*\|_2$, and consider

$$\begin{aligned} r_{k+1}^2 &= \|x^k - x^* - \alpha \nabla f(x^k)\|_2^2 \\ &= r_k^2 - 2\alpha \langle \nabla f(x^k),\, x^k - x^* \rangle + \alpha^2 \|\nabla f(x^k)\|_2^2 \\ &\leq \left(1 - \frac{2\alpha\mu L}{\mu + L}\right) r_k^2 + \alpha \left(\alpha - \frac{2}{\mu + L}\right) \|\nabla f(x^k)\|_2^2 \end{aligned}$$

  where we used Thm. 2 with $\nabla f(x^*) = 0$ for last inequality.

To finish: Use Lipschitz gradient and $\nabla f(x^*) = 0$ on last term to ultimately obtain

$$r_{k+1}^2 \leq \gamma r_k^2,$$

where $\gamma \leq 1 - \frac{2\alpha\mu L}{\mu + L}$.

Gradient methods – lower bounds

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k)$$

Theorem Lower bound I (Nesterov) For any $x^0 \in \mathbb{R}^n$, and $1 \leq k \leq \frac{1}{2}(n-1)$, there is a smooth f, s.t.

$$f(x^k) - f(x^*) \geq \frac{3L\|x^0 - x^*\|_2^2}{32(k+1)^2}$$

Gradient methods – lower bounds

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k)$$

Theorem Lower bound I (Nesterov) For any $x^0 \in \mathbb{R}^n$, and $1 \leq k \leq \frac{1}{2}(n-1)$, there is a smooth f, s.t.

$$f(x^k) - f(x^*) \geq \frac{3L\|x^0 - x^*\|_2^2}{32(k+1)^2}$$

Theorem Lower bound II (Nesterov). For class of smooth, strongly convex, i.e., $S^{\infty}_{L,\mu}$ ($\mu > 0$, $\kappa > 1$)

$$f(x^k) - f(x^*) \geq \frac{\mu}{2}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2k} \|x^0 - x^*\|_2^2.$$