EECS 559 Optimization Methods for SIPML

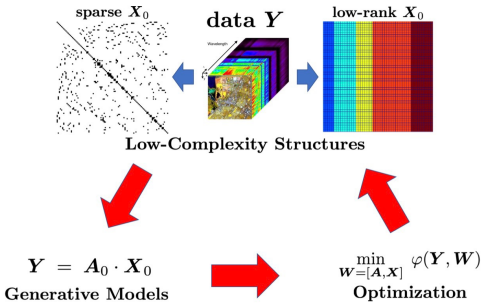
Lecture 6 – Subgradient Methods

Instructor: Prof. Qing Qu (qingqu@umich.edu)

Lecture Agenda

- **Motivating Examples**
- Subgradient and Subdifferential Revisit
- Subgradient Methods

Recover Low-Dimensional Structures

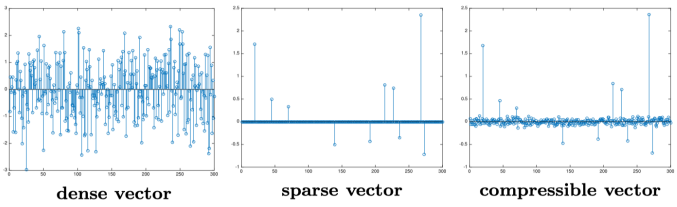


Convex Nonsmooth Problems

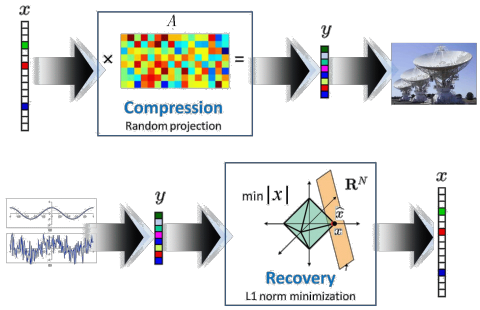
$$\min_{\boldsymbol{x}} \; f(\boldsymbol{x}), \quad \text{s.t.} \quad \boldsymbol{x} \in \mathcal{C}$$

- Here, the function $f(\boldsymbol{x}) : \mathbb{R}^n \mapsto \mathbb{R}$ is *convex*, and the set $\mathcal{C}$ is a *convex set* or just $\mathcal{C} = \mathbb{R}^n$;
- The function $f(\boldsymbol{x}) : \mathbb{R}^n \mapsto \mathbb{R}$ is *nonsmooth* (i.e., the gradient is not well-defined everywhere).

Example I: Sparse Recovery



Example I: Sparse Recovery



Example I: Sparse Recovery

$$\boldsymbol{y}_i = \boldsymbol{y}_{i\text{clean}} + \boldsymbol{z}_i = \underbrace{\boldsymbol{A}}_{\text{patch dictionary}} \cdot \underbrace{\boldsymbol{x}_i}_{\text{sparse coefficient}} + \boldsymbol{z}_i$$

noisy image | denoised image | dictionary for image patches

Example I: Sparse Recovery

Problem: given measurement $\boldsymbol{y} \in \mathbb{R}^m$ and sensing matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ $(m \ll n)$, recover the **sparsest** $\boldsymbol{x} \in \mathbb{R}^n$:

$$\min_{\boldsymbol{x}} \; \|\boldsymbol{x}\|_0, \quad \text{s.t.} \quad \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}.$$

where we have

$$\|\boldsymbol{x}\|_0 := \#\{\, i \mid x(i) \neq 0 \,\} = \sum_i \mathbb{1}_{x(i)\neq 0};$$

Example I: Sparse Recovery

Convex Relaxation via *basis pursuit* (BP)

$$\min_{\boldsymbol{x}} \|\boldsymbol{x}\|_1, \quad \text{s.t.} \quad \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x},$$

Here, $\|\boldsymbol{x}\|_1 = \sum_{k=1}^{m} |x_k|$ denotes the ℓ^1-norm of $\boldsymbol{x}$.

Example I: Sparse Recovery

Basis pursuit denoising (BPDN)

$$\min_{\boldsymbol{x}} \|\boldsymbol{x}\|_1, \quad \text{s.t.} \quad \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2 \le \delta,$$

which is equivalent to

$$\boxed{\min_{\boldsymbol{x}\in\mathbb{R}^n} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda \cdot \|\boldsymbol{x}\|_1,}$$

for some properly chosen $\lambda > 0$. The problem is termed as *Lasso* (least absolute shrinkage and selection operator).

Example I: Sparse Recovery

Basis pursuit denoising (BPDN)

$$\min_{\boldsymbol{x}} \|\boldsymbol{x}\|_1, \quad \text{s.t.} \quad \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2 \le \delta,$$

which is equivalent to

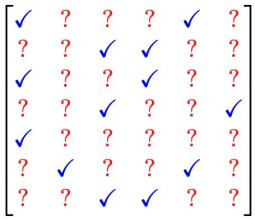
$$\boxed{\min_{\boldsymbol{x}\in\mathbb{R}^n} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda \cdot \|\boldsymbol{x}\|_1,}$$

for some properly chosen $\lambda > 0$. The ℓ_1-norm introduces **nonsmoothness** into the optimization problems.

Example II: Matrix Completion

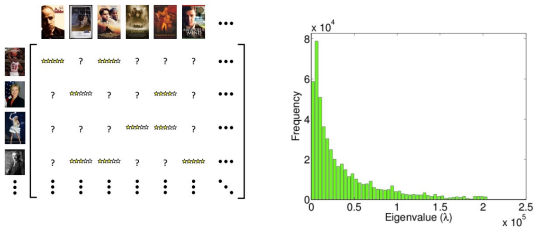
- **Netflix Challenge**: Netflix provides highly incomplete ratings from 0.5 million users for $\approx 17,770$ movies
- How to predict unseen user ratings for movies?

Example II: Matrix Completion



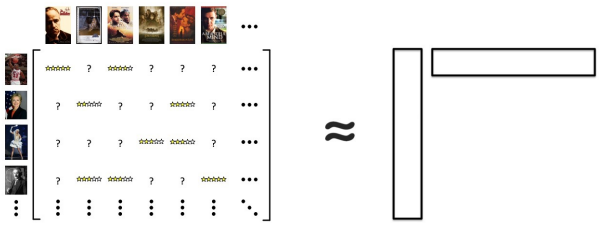
More unknowns than observations (under-determined system)

Example II: Matrix Completion



A few factors explain most of the data: **low-rank** approximation

Example II: Matrix Completion



How to exploit **low-rank structures** in predictions?

Example II: Matrix Completion

Given observed entries $M_{i,j}$ with $(i,j) \in \Omega$, complete the matrix via rank minimization

$$\min_{\boldsymbol{X}} \operatorname{rank}(\boldsymbol{X}), \quad \text{s.t.} \quad X_{i,j} = M_{i,j}, \quad (i,j) \in \Omega,$$

or, equivalently, we can rewrite it as

$$\min_{\boldsymbol{X}} \operatorname{rank}(\boldsymbol{X}), \quad \text{s.t.} \quad \mathcal{P}_\Omega(\boldsymbol{X}) = \mathcal{P}_\Omega(\boldsymbol{M}).$$

Example II: Matrix Completion

$$\min_{\boldsymbol{X}} \operatorname{rank}(\boldsymbol{X}), \quad \text{s.t.} \quad \mathcal{P}_\Omega(\boldsymbol{X}) = \mathcal{P}_\Omega(\boldsymbol{M}),$$

where $\mathcal{P}_\Omega(\cdot)$ is the orthogonal projection onto the subspace of matrices support on Ω.

$$\mathcal{P}_\Omega(\boldsymbol{X}) := \begin{cases} X_{ij}, & \text{if } (i,j) \in \Omega, \\ 0, & \text{if } (i,j) \notin \Omega. \end{cases}$$

Example II: Matrix Completion

$$\min_{\boldsymbol{X}} \operatorname{rank}(\boldsymbol{X}), \quad \text{s.t.} \quad \mathcal{P}_\Omega(\boldsymbol{X}) = \mathcal{P}_\Omega(\boldsymbol{M}),$$

- Similar to ℓ_0-norm minimization, the rank minimization is also **NP-hard**
- Remedy: convex relaxation via *nuclear norm*:

$$\operatorname{rank}(\boldsymbol{X}) = \|\boldsymbol{\sigma}(\boldsymbol{X})\|_0, \quad \|\boldsymbol{X}\|_* = \|\boldsymbol{\sigma}(\boldsymbol{X})\|_1,$$

where $\|\boldsymbol{X}\|_* = \sum_{i=1}^n \sigma_i(\boldsymbol{X})$ is a *convex surrogate* of $\operatorname{rank}(\boldsymbol{X})$

More General: Low-rank Matrix Recovery

$$\min_{\boldsymbol{X}} \operatorname{rank}(\boldsymbol{X}), \quad \text{s.t.} \quad \mathcal{P}_\Omega(\boldsymbol{X}) = \mathcal{P}_\Omega(\boldsymbol{M}),$$

The matrix completion problem can be extended to a more general form

$$\min_{\boldsymbol{X}} \|\boldsymbol{X}\|_*, \quad \text{s.t.} \quad \boldsymbol{y} = \mathcal{A}(\boldsymbol{X}).$$

with given linear sensing matrices $\boldsymbol{A}_i (1 \le i \le m)$

$$\boldsymbol{y} = \mathcal{A}(\boldsymbol{M}) := \begin{bmatrix} \operatorname{tr}\left(\boldsymbol{A}_1^\top \boldsymbol{M}\right) \\ \vdots \\ \operatorname{tr}\left(\boldsymbol{A}_m^\top \boldsymbol{M}\right) \end{bmatrix}$$

More General: Low-rank Matrix Recovery

Low-rank matrix recovery under noise

$$\min_{\boldsymbol{X}} \|\boldsymbol{X}\|_*, \quad \text{s.t.} \quad \|\boldsymbol{y} - \mathcal{A}(\boldsymbol{X})\|_2 \le \delta,$$

is *equivalent* to

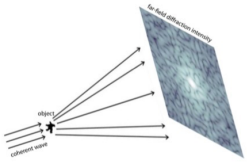
$$\min_{\boldsymbol{x} \in \mathbb{R}^n} \frac{1}{2} \|\boldsymbol{y} - \mathcal{A}(\boldsymbol{X})\|_2^2 + \lambda \cdot \|\boldsymbol{X}\|_*,$$

for some properly chosen $\lambda > 0$.

- Because $\|\cdot\|_*$ is **nonsmooth**, both problems are nonsmooth.

Example III: Phase Retrieval

Electric field $x(t_1, t_2) \rightarrow$ Fourier transform $\widehat{x}(f_1, f_2)$



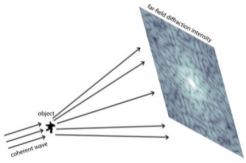
Coherent Diffraction Imaging[1]

Applications: X-ray crystallography, **diffraction imaging** (left), optics, astronomical imaging, and microscopy

Phase retrieval: given intensity $\boldsymbol{y} = |\mathcal{F}(\boldsymbol{x})|$, recover $\boldsymbol{x}$.

Example III: Phase Retrieval

Electric field $x(t_1, t_2) \rightarrow$ Fourier transform $\widehat{x}(f_1, f_2)$

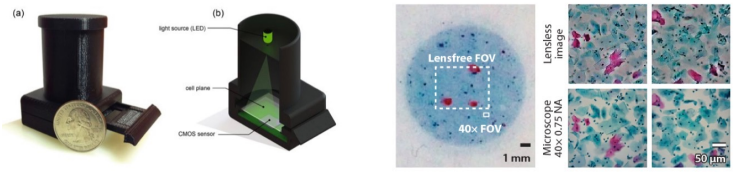


Coherent Diffraction Imaging[1]

Applications: X-ray crystallography, **diffraction imaging** (left), optics, astronomical imaging, and microscopy

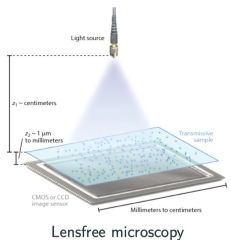
Generalized phase retrieval: given intensity $\boldsymbol{y} = |\boldsymbol{A}\boldsymbol{x}|$ recover $\boldsymbol{x}$, where $\boldsymbol{A}$ can be a general sensing matrix.

Example III: Phase Retrieval

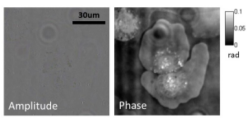


Lensfree microscopy imaging

Example III: Phase Retrieval

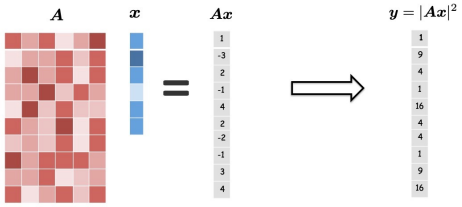


Lensfree microscopy



Phase contrast microscopy: phase is important[1].

Example III: Phase Retrieval



Solve for $\boldsymbol{x} \in \mathbb{R}^n$ in m quadratic equations

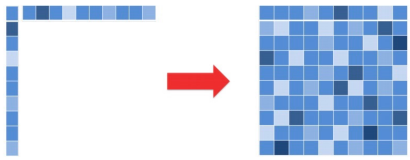
$$y_k = \left|\boldsymbol{a}_k^\top \boldsymbol{x}\right|^2, \quad k = 1, \cdots, m,$$

$$\text{or} \quad \boldsymbol{y} = |\boldsymbol{A}\boldsymbol{x}|^2, \quad \text{where } |\boldsymbol{x}|^2 := \left[|x_1|^2, \cdots, |x_m|^2\right]^\top.$$

Example III: Phase Retrieval

Lifting: introduce $\boldsymbol{X} = \boldsymbol{x}\boldsymbol{x}^*$ to linearize the problem

$$y_k = |\boldsymbol{a}_k^* \boldsymbol{x}|^2 = \boldsymbol{a}_k^* \underbrace{(\boldsymbol{x}\boldsymbol{x}^*)}_{\boldsymbol{X}} \boldsymbol{a}_k \quad \Longrightarrow \quad y_k = \langle \boldsymbol{a}_k \boldsymbol{a}_k^*, \boldsymbol{X} \rangle$$



Example III: Phase Retrieval

$$\text{find } \boldsymbol{X} \succeq \boldsymbol{0}, \quad \text{s.t.} \quad y_k = \langle \boldsymbol{a}_k \boldsymbol{a}_k^*, \boldsymbol{X} \rangle, \quad k = 1, \cdots, m$$
$$\operatorname{rank}(\boldsymbol{X}) = 1$$

- **Convex relaxation:**

$$\min_{\boldsymbol{X}} \|\boldsymbol{X}\|_*, \quad \text{s.t.} \quad \boldsymbol{y} = \mathcal{A}(\boldsymbol{X}), \ \boldsymbol{X} \succeq \boldsymbol{0}.$$

- **Stable version:**

$$\min_{\boldsymbol{X} \succeq \boldsymbol{0}} \frac{1}{2} \|\boldsymbol{y} - \mathcal{A}(\boldsymbol{X})\|_2^2 + \lambda \cdot \|\boldsymbol{X}\|_*.$$

Example III: Sparse Phase Retrieval

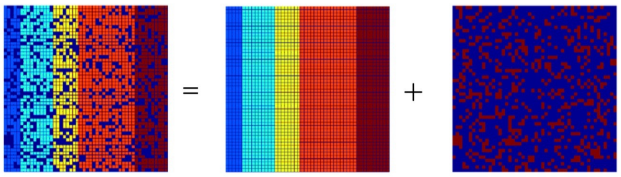
- **Convex relaxation:**

$$\min_{\boldsymbol{X}} \|\boldsymbol{X}\|_* + \lambda \|\boldsymbol{X}\|_1, \quad \text{s.t.} \quad \boldsymbol{y} = \mathcal{A}(\boldsymbol{X}), \ \boldsymbol{X} \succeq \boldsymbol{0}.$$

- **Stable version:**

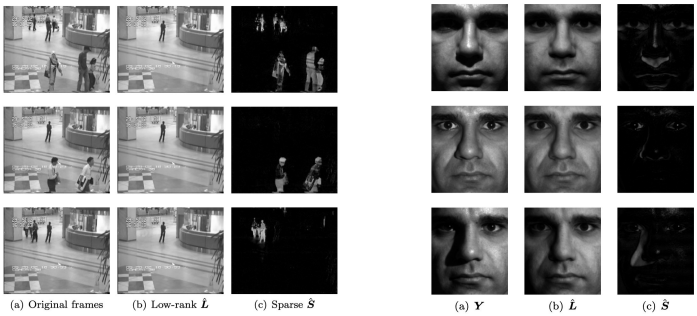
$$\min_{\boldsymbol{X} \succeq \boldsymbol{0}} \|\boldsymbol{X}\|_* + \lambda \|\boldsymbol{X}\|_1 + \frac{\mu}{2} \|\boldsymbol{y} - \mathcal{A}(\boldsymbol{X})\|_2^2.$$

Example IV: Robust PCA



observation $\boldsymbol{Y}$ low-rank $\boldsymbol{L}$ sparse $\boldsymbol{S}$

Example IV: Robust PCA



(a) Original frames (b) Low-rank $\hat{\boldsymbol{L}}$ (c) Sparse $\hat{\boldsymbol{S}}$

(a) $\boldsymbol{Y}$ (b) $\hat{\boldsymbol{L}}$ (c) $\hat{\boldsymbol{S}}$

Example IV: Robust PCA

- **Convex relaxation:**

$$\min_{\boldsymbol{L},\boldsymbol{S}} \|\boldsymbol{L}\|_* + \lambda \|\boldsymbol{S}\|_1, \quad \text{s.t.} \quad \boldsymbol{Y} = \boldsymbol{L} + \boldsymbol{S}.$$

- **Stable version:**

$$\min_{\boldsymbol{L},\boldsymbol{S}} \|\boldsymbol{L}\|_* + \lambda \|\boldsymbol{S}\|_1 + \frac{\mu}{2} \|\boldsymbol{Y} - \boldsymbol{L} - \boldsymbol{S}\|_F^2.$$

Further Readings

- *High-Dimensional Data Analysis with Low-Dimensional Models: Principles, Computation, and Applications*. John Wright, Yi Ma. (**Chapter 3-5**)

Lecture Agenda

- Motivating Examples
- **Subgradient and Subdifferential Revisit**
- Subgradient Methods

(Projected) Subgradient Methods

$$\min_{\boldsymbol{x}} \underbrace{f(\boldsymbol{x})}_{\text{nonsmooth}}, \quad \text{s.t.} \quad \boldsymbol{x} \in \mathcal{C}.$$

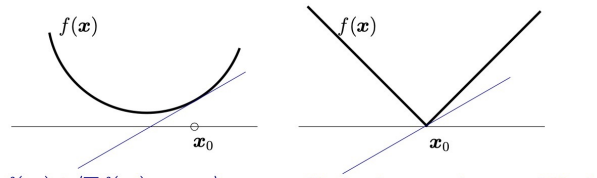
Practically, a natural choice is "subgradient-based methods"

$$\boldsymbol{x}_{k+1} = \mathcal{P}_{\mathcal{C}}\left(\boldsymbol{x}_k - \tau_k \cdot \boldsymbol{g}_k\right),$$

where $\boldsymbol{g}_k \in \partial f(\boldsymbol{x}_k)$ is any subgradient of $f(\cdot)$ at $\boldsymbol{x}_k$.

- **Caution:** unlike gradient descent method, a subgradient direction might *not* be a descent direction on f.

Basics: Subdifferential & Subgradient



Basics: Subgradient & Subdifferential

- **Subgradient.** Let $f: \mathbb{R}^n \mapsto \mathbb{R}$ be *convex. A subgradient* of f at $\boldsymbol{x}_0$ is any $\boldsymbol{u}$ satisfying

$$\boxed{f(\boldsymbol{x}) \geq f(\boldsymbol{x}_0) + \langle \boldsymbol{u}, \boldsymbol{x} - \boldsymbol{x}_0 \rangle, \quad \forall\, \boldsymbol{x}.}$$

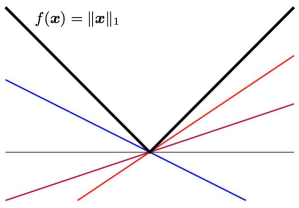
- **Subdifferential.** It is the *set of all subgradients* of f at $\boldsymbol{x}_0$

$$\partial f(\boldsymbol{x}_0) := \{\boldsymbol{u} \mid f(\boldsymbol{x}) \geq f(\boldsymbol{x}_0) + \langle \boldsymbol{u}, \boldsymbol{x} - \boldsymbol{x}_0 \rangle, \ \forall \boldsymbol{x} \in \mathbb{R}^n\}.$$

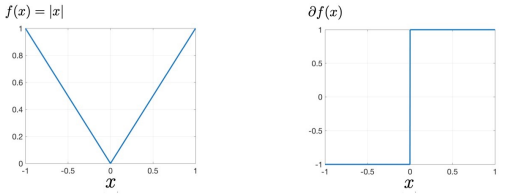
If f is differentiable at $\boldsymbol{x}_0$, then $\partial f(\boldsymbol{x}) = \{\nabla f(\boldsymbol{x}_0)\}$.

Example: $f(x) = |x|$

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



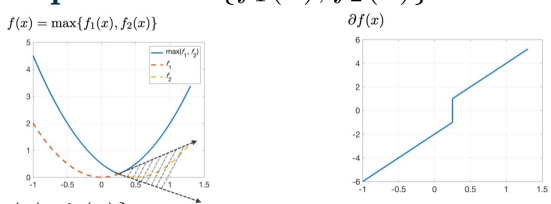
Example: $f(x) = |x|$



$f(x) = |x|$

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

Example: $\max\{f_1(x), f_2(x)\}$



$\max\{f_1(x), f_2(x)\}$, where f_1 and f_2 are differentiable

$$\partial f(x) = \begin{cases} \{f_1'(x)\}, & \text{if } f_1(x) > f_2(x) \\ [f_1'(x), f_2'(x)], & \text{if } f_1(x) = f_2(x) \\ \{f_2'(x)\}, & \text{if } f_1(x) < f_2(x) \end{cases}$$

A Subgradient of Norms at 0

Theorem. Let $f(\boldsymbol{x}) = \|\boldsymbol{x}\|$ for any norm $\|\cdot\|$, then for any $\boldsymbol{g}$ obeying $\|\boldsymbol{g}\|_\star \leq 1$, we have

$$\boxed{\boldsymbol{g} \in \partial f(\boldsymbol{0}),}$$

where $\|\cdot\|_\star$ is the dual norm of $\|\cdot\|$ with

$$\|\boldsymbol{x}\|_\star := \sup_{\boldsymbol{z}:\|\boldsymbol{z}\| \leq 1} \langle \boldsymbol{z}, \boldsymbol{x} \rangle$$

Here, $\|\cdot\|_\star$ is not unclear norm.

A Subgradient of Norms at 0

Proof. To show this, it suffices to show that for $f(\boldsymbol{z}) = \|\boldsymbol{z}\|$,

$$\forall \boldsymbol{z} \in \mathbb{R}^n \quad f(\boldsymbol{z}) \geq f(\boldsymbol{0}) + \langle \boldsymbol{g}, \boldsymbol{z} - \boldsymbol{0} \rangle, \quad \|\boldsymbol{g}\|_\star \leq 1$$

$\Longleftrightarrow$ $\|\boldsymbol{z}\| \geq \langle \boldsymbol{g}, \boldsymbol{z} \rangle \quad \|\boldsymbol{g}\|_\star \leq 1$

To prove the above, notice that

$$\|\boldsymbol{g}\|_\star = \sup_{\|\boldsymbol{z}\| \leq 1} \langle \boldsymbol{g}, \boldsymbol{z} \rangle = \sup_{\boldsymbol{z}} \left\langle \boldsymbol{g}, \frac{\boldsymbol{z}}{\|\boldsymbol{z}\|} \right\rangle \geq \left\langle \boldsymbol{g}, \frac{\boldsymbol{z}}{\|\boldsymbol{z}\|} \right\rangle, \ \forall\, \boldsymbol{z}.$$

$\Longrightarrow$ $\langle \boldsymbol{g}, \boldsymbol{z} \rangle \leq \|\boldsymbol{g}\|_\star \|\boldsymbol{z}\| \leq \|\boldsymbol{z}\|$

Basic Rules for Subdifferential

- **Scaling:** $\partial(\alpha f) = \alpha \partial f$ (for $\alpha > 0$)
- **Summation:** $\partial(f_1 + f_2) = \partial f_1 + \partial f_2$

Example: the ℓ_1-Norm

$$f(\boldsymbol{x}) \ = \ \|\boldsymbol{x}\|_1 \ = \ \sum_{i=1}^{n} \underbrace{|x_i|}_{=: f_i(\boldsymbol{x})}$$

Since

$$\partial f_i(\boldsymbol{x}) \ = \ \begin{cases} \operatorname{sign}(x_i)\boldsymbol{e}_i, & \text{if } x_i \neq 0, \\ [-1,1] \cdot \boldsymbol{e}_i & \text{if } x_i = 0. \end{cases}$$

We have

$$\partial f(\boldsymbol{x}) \ = \ \sum_{i:\, x_i \neq 0} \operatorname{sign}(x_i)\boldsymbol{e}_i + \sum_{j: x_j = 0} [-1,1] \cdot \boldsymbol{e}_j$$

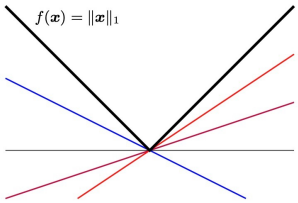
In particular, $\sum_{i:\, x_i \neq 0} \operatorname{sign}(x_i)\boldsymbol{e}_i \in \partial f(\boldsymbol{x})$.

Example: $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_1$, $\boldsymbol{x} \in \mathbb{R}^n$

$$\partial f(\boldsymbol{x}) \ = \ \mathcal{J}_1 \times \cdots \times \mathcal{J}_n,$$

where for $k = 1, 2, \cdots, n$,

$$\mathcal{J}_k \ = \ \begin{cases} \{1\} & x_k > 0, \\ [-1,1] & x_k = 0, \\ \{-1\} & x_k < 0. \end{cases}$$



Basic Rules for Subdifferential

- **Scaling:** $\partial(\alpha f) = \alpha \partial f$ (for $\alpha > 0$)
- **Summation:** $\partial(f_1 + f_2) = \partial f_1 + \partial f_2$
- **Affine transformation:** if $h(\boldsymbol{x}) = f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})$, then

$$\partial h(\boldsymbol{x}) \ = \ \boldsymbol{A}^\top \partial f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})$$

Example: $h(\boldsymbol{x}) = \|\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}\|_1$

Let $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_1$ and $\boldsymbol{A} = [\boldsymbol{a}_1, \cdots, \boldsymbol{a}_m]^\top$, we have

$$\boldsymbol{g} \ = \ \sum_{i: \boldsymbol{a}_i^\top \boldsymbol{x} + b_i \neq 0} \operatorname{sign}(\boldsymbol{a}_i^\top \boldsymbol{x} + b_i) \cdot \boldsymbol{e}_i \ \in \ \partial f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})$$

$$\Longrightarrow \quad \boldsymbol{A}^\top \boldsymbol{g} \ = \ \sum_{i: \boldsymbol{a}_i^\top \boldsymbol{x} + b_i \neq 0} \operatorname{sign}(\boldsymbol{a}_i^\top \boldsymbol{x} + b_i) \cdot \boldsymbol{a}_i \ \in \ \partial h(\boldsymbol{x}).$$

Basic Rules for Subdifferential

- **Chain rule:** suppose $f : \mathbb{R}^n \mapsto \mathbb{R}$ is convex, and $g : \mathbb{R} \mapsto \mathbb{R}$ is differentiable, nondecreasing, and convex. Let $h = f \circ g$,

$$\partial h(\boldsymbol{x}) \ = \ g'(f(\boldsymbol{x})) \cdot \partial f(\boldsymbol{x})$$

- **Composition:** suppose $f(\boldsymbol{x}) = h(f_1(\boldsymbol{x}), \cdots, f_n(\boldsymbol{x}))$, where f_i's are convex, and h is differentiable, nondecreasing, and convex. Let $\boldsymbol{q} = \nabla h(\boldsymbol{y})\,|_{\boldsymbol{y} = [f_1(\boldsymbol{x}), \cdots, f_n(\boldsymbol{x})]}$ and $\boldsymbol{g}_i \in \partial f_i(\boldsymbol{x})$, then

$$\sum_{i=1}^{n} q_i \boldsymbol{g}_i \ \in \ \partial f(\boldsymbol{x}).$$

Basic Rules for Subdifferential

- **Pointwise maximum:** if $f(\boldsymbol{x}) = \max_{1 \leq i \leq k} f_i(\boldsymbol{x})$, then

$$\partial f(\boldsymbol{x}) \ = \ \underbrace{\operatorname{conv}\left\{\cup \left\{\partial f_i(\boldsymbol{x}) \ \middle|\ f_i(\boldsymbol{x}) = f(\boldsymbol{x})\right\}\right\}}_{\text{convex hull of subdifferentials of all active functions}}$$

- **Pointwise supremum:** if $f(\boldsymbol{x}) = \sup_{\alpha \in \mathcal{F}} f_\alpha(\boldsymbol{x})$, then

$$\partial f(\boldsymbol{x}) \ = \ \operatorname{closure}\left(\operatorname{conv}\left\{\cup_{\alpha \in \mathcal{F}} \left\{\partial f_\alpha(\boldsymbol{x}) \ \middle|\ f_\alpha(\boldsymbol{x}) = f(\boldsymbol{x})\right\}\right\}\right)$$

Example: Piece-wise Linear Functions

$$f(\boldsymbol{x}) \;=\; \max_{1\le i\le m} \left\{\boldsymbol{a}_i^\top \boldsymbol{x} + b_i\right\}$$

At any given $\boldsymbol{x}_0$, pick any $\boldsymbol{a}_j$ such that

$$\boldsymbol{a}_j^\top \boldsymbol{x}_0 + b_j = \max\nolimits_{1\le i\le m} \left\{\boldsymbol{a}_i^\top \boldsymbol{x}_0 + b_i\right\}$$

then

$$\boldsymbol{a}_j \;\in\; \partial f(\boldsymbol{x}_0).$$

Example: the ℓ_∞-Norm

$$f(\boldsymbol{x}) \;=\; \|\boldsymbol{x}\|_\infty \;=\; \max_{1\le i\le n} |x_i|\,.$$

At any given $\boldsymbol{x}'$, pick any x'_j such that

$$|x'_j| = \max\nolimits_{1\le i\le m} |x'_i|$$

then

$$\operatorname{sign}(x'_j)\cdot \boldsymbol{e}_j \;\in\; \partial f(\boldsymbol{x}')$$

Example: the Maximum Eigenvalue

$$f(\boldsymbol{x}) \;=\; \lambda_{\max}\left(\sum_{i=1}^n x_i \boldsymbol{A}_i\right)$$

where $\boldsymbol{A}_1, \cdots, \boldsymbol{A}_n$ are real symmetric matrices. Rewrite

$$f(\boldsymbol{x}) \;=\; \sup_{\|\boldsymbol{v}\|=1} \boldsymbol{v}^\top \left(\sum_{i=1}^n x_i \boldsymbol{A}_i\right) \boldsymbol{v}.$$

For any given $\boldsymbol{x}$, taking $\boldsymbol{v}_0$ as the leading eigenvector of $\sum_i x_i \boldsymbol{A}_i$

$$\begin{bmatrix}\boldsymbol{v}_0^\top \boldsymbol{A}_1 \boldsymbol{v}_0 & \cdots & \boldsymbol{v}_0^\top \boldsymbol{A}_n \boldsymbol{v}_0\end{bmatrix}^\top \;\in\; \partial f(\boldsymbol{x})$$

Example: the Nuclear Norm

Let $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ with SVD $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ and

$$f(\boldsymbol{X}) \;=\; \|\boldsymbol{X}\|_* \;:=\; \sum_{i=1}^{\min\{m,n\}} \sigma_i(\boldsymbol{X})$$

Then we have

$$\boldsymbol{U}\boldsymbol{V}^\top \;\in\; \partial f(\boldsymbol{X}).$$

More generally, we can show that

$$\partial \|\boldsymbol{X}\|_* \;=\; \left\{\boldsymbol{U}\boldsymbol{V}^\top + \boldsymbol{W} \;\middle|\; \|\boldsymbol{W}\| \le 1, \boldsymbol{W} \perp \operatorname{col}(\boldsymbol{U}),\; \boldsymbol{W} \perp \operatorname{row}(\boldsymbol{V})\right\}.$$

Example: the Nuclear Norm

Proof. First, notice the fact that

$$\|\boldsymbol{X}\|_* \;=\; \sup_{\|\boldsymbol{Z}\|\le 1} \langle \boldsymbol{X}, \boldsymbol{Z}\rangle$$

Therefore, we have

$$\|\boldsymbol{X}\|_* = \sup_{\|\boldsymbol{X}\|\le 1} \langle \boldsymbol{Z}, \boldsymbol{U}\boldsymbol{\Sigma}_X\boldsymbol{V}^\top\rangle = \sup_{\|\boldsymbol{X}\|\le 1} \langle \boldsymbol{U}^\top\boldsymbol{Z}\boldsymbol{V}, \boldsymbol{\Sigma}_X\rangle$$

where the supreme is achieved when $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{V}^\top$, so that $\|\boldsymbol{X}\|_* = \operatorname{tr}(\boldsymbol{\Sigma}_X)$

Thus, let $f_Z(\boldsymbol{X}) = \langle \boldsymbol{Z}, \boldsymbol{X}\rangle$, we have $\|\boldsymbol{X}\|_* \;=\; \sup_{\|\boldsymbol{Z}\|\le 1} f_Z(\boldsymbol{X})$. By the property of subgradient, we have

$$\nabla_{\boldsymbol{X}} f_Z(\boldsymbol{X})\,|_{\boldsymbol{Y}=\boldsymbol{U}\boldsymbol{V}^\top} \in \partial \|\boldsymbol{X}\|_* \;\Longrightarrow\; \boldsymbol{U}\boldsymbol{V}^\top \in \partial \|\boldsymbol{X}\|_*$$

Optimality Condition for Convex Nonsmooth Problem

Theorem. For any convex function f,

$$f(\boldsymbol{x}_\star) \;=\; \min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) \quad\Longleftrightarrow\quad \boldsymbol{0} \in \partial f(\boldsymbol{x}_\star)$$

that $\boldsymbol{x}_\star$ is a minimizer *iff* $\boldsymbol{0}$ is a subgradient of f at $\boldsymbol{x}_\star$.

Reason: for any $\boldsymbol{y}$, by the definition of subgradient

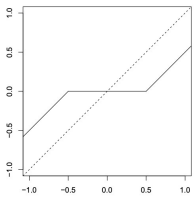
$$f(\boldsymbol{y}) \;\ge\; f(\boldsymbol{x}_\star) + \boldsymbol{0}^\top(\boldsymbol{y} - \boldsymbol{x}_\star)$$

Soft-thresholding Operator

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \; f(\boldsymbol{x}) \;=\; \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{z}\|_2^2 + \lambda \|\boldsymbol{x}\|_1$$

Claim. An optimal solution $\boldsymbol{x}_\star = \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x})$, is the *soft-thresholding operator* $\boldsymbol{x}_\star = \mathrm{S}_\lambda(\boldsymbol{z})$ with

$$[\mathrm{S}_\lambda(\boldsymbol{z})]_i \;=\; \begin{cases} z_i - \lambda & \text{if } z_i > \lambda, \\ 0 & \text{if } |z_i| \le \lambda, \\ z_i + \lambda & \text{if } z_i \le -\lambda. \end{cases}$$



In summary, $\mathrm{S}_\lambda(\boldsymbol{z}) \;=\; \operatorname{sign}(\boldsymbol{z}) \odot (|\boldsymbol{z}| - \lambda)_+$.

Soft-thresholding Operator

Proof. Notice that $\partial f(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{z}) + \lambda \partial \|\boldsymbol{x}\|_1$, we find the optimal solution $\boldsymbol{x}_\star$ by solving $\boldsymbol{0} \in \partial f(\boldsymbol{x}_\star)$

Given that $[\partial f(\boldsymbol{x}_\star)]_i = [\boldsymbol{x}_\star - \boldsymbol{z} + \lambda \partial \|\boldsymbol{x}_\star\|_1]_i$

- If $z_i > \lambda$, $0 \in x_{\star i} - z_i + \lambda\,[\partial \|\boldsymbol{x}_\star\|_1]_i$ $\Longrightarrow$ $x_{\star i} = z_i - \lambda$
- If $z_i < -\lambda$, $0 \in x_{\star i} - z_i + \lambda\,[\partial \|\boldsymbol{x}_\star\|_1]_i$ $\Longrightarrow$ $x_{\star i} = z_i + \lambda$
- If $|z_i| \le \lambda$, $0 \in x_{\star i} - z_i + \lambda\,[\partial \|\boldsymbol{x}_\star\|_1]_i$ $\Longrightarrow$ $x_{\star i} = 0$

Singular Value Thresholding

Theorem. The unique solution $\boldsymbol{X}_\star$ to the program

$$\min_{\boldsymbol{X}} \frac{1}{2} \|\boldsymbol{X} - \boldsymbol{Z}\|_F^2 + \lambda \|\boldsymbol{X}\|_*,$$

is given by

$$\boldsymbol{X}_\star \;=\; \boldsymbol{U} \mathrm{S}_\lambda [\boldsymbol{\Sigma}] \boldsymbol{V}^\top,$$

where $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ and $\mathrm{S}_\lambda(\cdot)$ is the soft-thresholding operator.

Lecture Agenda

- Motivating Examples
- Subgradient and Subdifferential Revisit
- **Subgradient Methods**

(Projected) Subgradient Methods

$$\min_{\boldsymbol{x}} \underbrace{f(\boldsymbol{x})}_{\text{nonsmooth}}, \quad \text{s.t.} \quad \boldsymbol{x} \in \mathcal{C}.$$

Practically, a natural choice is "subgradient-based methods"

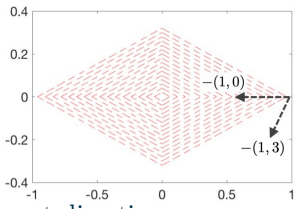
$$\boldsymbol{x}_{k+1} \;=\; \mathcal{P}_{\mathcal{C}} \left(\boldsymbol{x}_k - \tau_k \cdot \boldsymbol{g}_k \right),$$

where $\boldsymbol{g}_k \in \partial f(\boldsymbol{x}_k)$ is any subgradient of $f(\cdot)$ at $\boldsymbol{x}_k$.

- **Caution:** unlike gradient descent method, a subgradient direction might *not* be a descent direction on f.

Negative Subgradient Not a Descent Direction

Example: $f(\boldsymbol{x}) = |x_1| + 3\,|x_2|$



at $\boldsymbol{x} = (1, 0)$:

- $\boldsymbol{g}_1 = (1, 0) \in \partial f(\boldsymbol{x})$, and $-\boldsymbol{g}_1$ is a descent direction;
- $\boldsymbol{g}_2 = (1, 3) \in \partial f(\boldsymbol{x})$, but $-\boldsymbol{g}_2$ is *not* a descent direction!

Reason: lack of continuity. One can change directions significantly without violating the validity of subgradients.

Subgradient Methods

- Since $f(\boldsymbol{x}_k)$ is not necessarily monotone, we keep track of the best point

$$f_{\text{best},k} \;:=\; \min_{1 \le i \le k} f(\boldsymbol{x}_i)$$

- At each step k, we set

$$f_{\text{best},k} \;=\; \min \left\{ f_{\text{best},k-1}, f(\boldsymbol{x}_k) \right\}.$$

Convergence of Subgradient Methods

$$\min_{\boldsymbol{x}} \underbrace{f(\boldsymbol{x})}_{\text{nonsmooth}}, \quad \text{s.t.} \quad \boldsymbol{x} \in \mathbb{R}^n.$$

Lemma. Suppose $f : \mathbb{R}^n \mapsto \mathbb{R}$ is L-Lipschitz, then the iterate

$$\boldsymbol{x}_{k+1} \;=\; \boldsymbol{x}_k - \tau_k \cdot \boldsymbol{g}_k,$$

satisfies

$$f_{\text{best},k} - f_\star \;\le\; \frac{R^2 + L^2 \sum_{i=1}^k \tau_i^2}{2 \sum_{i=1}^k \tau_i},$$

where $R \ge \|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|_2$.

 Subgradient Methods, Notes for EE364b, Stanford University. Stephen Boyd. (Section 3, page 6)

Convergence of Subgradient Methods

Proof. Let us denote $f_k = f(\boldsymbol{x}_k)$ and $f_\star = f(\boldsymbol{x}_\star)$

$$\begin{aligned} 0 \le \|\boldsymbol{x}_{k+1} - \boldsymbol{x}_\star\|_2^2 &= \|\boldsymbol{x}_k - \boldsymbol{x}_\star - \tau_k \boldsymbol{g}_k\|_2^2 \\ &= \|\boldsymbol{x}_k - \boldsymbol{x}_\star\|_2^2 - 2 \langle \tau_k \boldsymbol{g}_k, \boldsymbol{x}_k - \boldsymbol{x}_\star \rangle + \tau_k^2 \|\boldsymbol{g}_k\|_2^2 \\ &\le \|\boldsymbol{x}_k - \boldsymbol{x}_\star\|_2^2 - 2\tau_k (f_k - f_\star) + \tau_k^2 \|\boldsymbol{g}_k\|_2^2 \end{aligned}$$

which has been derived by the def. of subgradient

$$f_\star \ge f_k + g_k^\top (\boldsymbol{x}_\star - \boldsymbol{x}_k) \quad \Rightarrow \quad f_\star - f_k \ge g_k^\top (\boldsymbol{x}_\star - \boldsymbol{x}_k)$$

Therefore, recursively, we have

$$0 \le \|\boldsymbol{x}_{k+1} - \boldsymbol{x}_\star\|_2^2 \le \|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|_2^2 - 2 \sum_{i=0}^k \tau_i (f_i - f_\star) + \sum_{i=0}^k \tau_i^2 \|\boldsymbol{g}_i\|_2^2$$

Convergence of Subgradient Methods

- On one hand:

$$2 \sum_{i=0}^k \tau_i (f_i - f_\star) \le R^2 + \sum_{i=0}^k \tau_i^2 \|\boldsymbol{g}_i\|_2^2 \le R^2 + L^2 \sum_{i=0}^K \tau_i^2$$

- On the other hand:

$$\sum_{i=0}^k \tau_i (f_i - f_\star) \ge \left(\sum_{i=0}^k \tau_i \right) \min_{0 \le i \le k} (f_i - f_\star)$$

Combining the above together, we obtain

$$f_{best,k} - f_\star = \min_{0 \le i \le k} (f_i - f_\star) \le \frac{1}{2 \sum_{i=0}^k \tau_i} \left(R^2 + L^2 \sum_{i=0}^k \tau_i^2 \right)$$

Step Size Rules

$$f_{\text{best},k} - f_\star \;\leq\; \frac{R^2 + L^2 \sum_{i=1}^k \tau_i^2}{2 \sum_{i=1}^k \tau_i}$$

- **Constant step size** $\tau_k \equiv \tau$

$$\lim_{k\to\infty} f_{\text{best},k} - f_\star \;\leq\; \frac{L^2 \tau}{2}.$$

- **Diminishing step size** with $\sum_k \tau_k^2 < +\infty$ and $\sum_k \tau_k \to +\infty$

$$\lim_{k\to\infty} f_{\text{best},k} - f_\star \;=\; 0.$$

Optimal Step Size for Subgradient Method?

Theorem. Suppose $f : \mathbb{R}^n \mapsto \mathbb{R}$ is L-Lipschitz, then choose a constant step size

$$\tau_i = R/(L\sqrt{k}),\; i = 1, \cdots, k$$

gives

$$f_{\text{best},k} - f_\star \;\leq\; \frac{RL}{\sqrt{k}}$$

where $R \geq \|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|_2$.

Optimal Step Size for Subgradient Method?

Proof. We already know that

$$f_{best,k} - f_\star \leq h(\tau) := \frac{R^2 + L^2 \sum_{i=0}^k \tau_i^2}{2 \sum_{i=0}^k \tau_i}$$

and we can show that $h(\tau)$ is *convex* and *symmetric* in $\tau = \{\tau_1, \cdots, \tau_k\}$

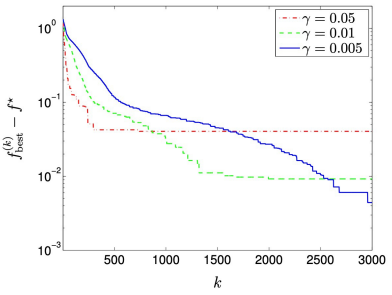
The optimal solution for $\min_\tau \; h(\tau)$ is $\tau_1 = \cdots = \tau_k = \dfrac{R}{L\sqrt{k}}$

Plugging the solution in, we obtain

$$f_{best,k} - f_\star \leq h^*(\tau_*) = \frac{RL}{\sqrt{k}}$$

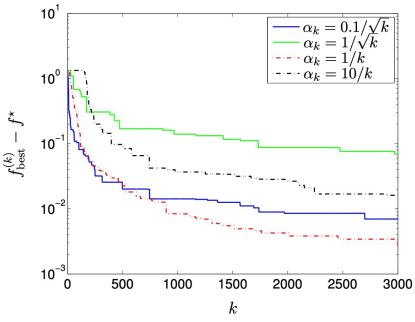
Experiments: Subgradient Methods

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \; f(\boldsymbol{x}) \;=\; \|\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}\|_\infty$$



Experiments: Subgradient Methods

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \; f(\boldsymbol{x}) \;=\; \|\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}\|_\infty$$



Optimal Step Size with Known $f_\star$

Theorem. (Polyak's Stepsize Rule)
Suppose $f : \mathbb{R}^n \mapsto \mathbb{R}$ is L-Lipschitz, and the optimal is $f_\star$ known. Choose stepsizes that satisfy

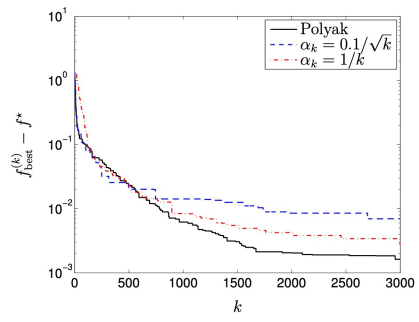
$$\tau_k = \frac{f_k - f_\star}{\|\boldsymbol{g}_k\|_2^2}.$$

Then, we have

$$f_{\text{best},k} - f_\star \;\leq\; \frac{RL}{\sqrt{k}}.$$

Experiments: Subgradient Methods

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \; f(\boldsymbol{x}) \;=\; \|\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}\|_\infty$$



Strongly Convex Functions

Theorem. Let f be μ-strongly convex and L-Lipschitz continuous over $\mathbb{R}^n$. Choose the stepsize as $\tau_k \equiv \tau = \frac{2}{\mu(k+1)}$, then

$$f_{\text{best},k} - f_\star \;\leq\; \frac{2L^2}{\mu}\frac{1}{k+1}.$$

Summary: Subgradient Methods

	Stepsize Rule	Convergence Rate	Iteration Complexity
Convex & Lipschitz Problems	$\tau_k \asymp \frac{1}{\sqrt{k}}$	$O\left(\frac{1}{\sqrt{k}}\right)$	$O\left(\frac{1}{\varepsilon^2}\right)$
Strongly Convex & Lipschitz Problems	$\tau_k \asymp \frac{1}{k}$	$O\left(\frac{1}{k}\right)$	$O\left(\frac{1}{\varepsilon}\right)$

Similar results can be shown for the projected version.

Further Readings

- *Subgradient Methods, Notes for EE364b, Stanford University.* Stephen Boyd. https://stanford.edu/class/ee364b/lectures/subgrad_method_notes.pdf