Mathematics of Data: From Theory to Computation

Prof. Volkan Cevher
volkan.cevher@epfl.ch

Lecture 6: From stochastic gradient descent to non-smooth optimization

Laboratory for Information and Inference Systems (LIONS)
École Polytechnique Fédérale de Lausanne (EPFL)

EE-556 (Fall 2022)

License Information for Mathematics of Data Slides

Outline

- Stochastic optimization
- Deficiency of smooth models
- Sparsity and compressive sensing
- Non-smooth minimization via Subgradient descent
- *Atomic norms

Recall: Gradient descent

Problem (Unconstrained optimization problem)

Consider the following minimization problem:

$$f^\star = \min_{\mathbf{x}\in\mathbb{R}^p} f(\mathbf{x})$$

$f(\mathbf{x})$ *is proper and closed.*

Gradient descent

Choose a starting point $\mathbf{x}^0$ and iterate

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

where α_k is a step-size to be chosen so that $\mathbf{x}^k$ converges to $\mathbf{x}^\star$.

	f **is L-smooth & convex**	f **is L-gradient Lipschitz & non-convex**
GD	$O(1/k)$ (fast)	$O(1/k)$ (optimal)
AGD	$O(1/k^2)$ (optimal)	$O(1/k)$ (optimal) [15]

Recall: Gradient descent

Problem (Unconstrained optimization problem)

Consider the following minimization problem:

$$f^\star = \min_{\mathbf{x}\in\mathbb{R}^p} f(\mathbf{x})$$

$f(\mathbf{x})$ *is proper and closed.*

Gradient descent

Choose a starting point $\mathbf{x}^0$ and iterate

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

where α_k is a step-size to be chosen so that $\mathbf{x}^k$ converges to $\mathbf{x}^\star$.

	f is L-smooth & convex	f is L-gradient Lipschitz & non-convex
GD	$O(1/k)$ (fast)	$O(1/k)$ (optimal)
AGD	$O(1/k^2)$ (optimal)	$O(1/k)$ (optimal) [15]

Why should we study anything else?

Statistical learning with streaming data

- Recall that statistical learning seeks to find a $h^\star \in \mathcal{H}$ that minimizes the *expected* risk,

$$h^\star \in \arg\min_{h\in\mathcal{H}} \left\{ R(h) := \mathbb{E}_{(\mathbf{a},b)}\left[L(h(\mathbf{a}),b)\right]\right\}.$$

Abstract gradient method

$$h^{k+1} = h^k - \alpha_k \nabla R(h^k) = h^k - \alpha_k \mathbb{E}_{(\mathbf{a},b)}[\nabla L(h^k(\mathbf{a}),b)].$$

Remark:

- This algorithm can not be implemented as the distribution of $(\mathbf{a}, b)$ is unknown.

Statistical learning with streaming data

◦ Recall that statistical learning seeks to find a $h^\star \in \mathcal{H}$ that minimizes the *expected* risk,

$$h^\star \in \arg\min_{h \in \mathcal{H}} \left\{ R(h) := \mathbb{E}_{(\mathbf{a},b)} \left[L(h(\mathbf{a}), b) \right] \right\}.$$

Abstract gradient method

$$h^{k+1} = h^k - \alpha_k \nabla R(h^k) = h^k - \alpha_k \mathbb{E}_{(\mathbf{a},b)}[\nabla L(h^k(\mathbf{a}), b)].$$

Remark: ◦ This algorithm can not be implemented as the distribution of $(\mathbf{a}, b)$ is unknown.

◦ In practice, data can arrive in a *streaming* way.

A parametric example: Markowitz portfolio optimization

$$\mathbf{x}^\star := \min_{\mathbf{x} \in \mathcal{X}} \left\{ \mathbb{E}\left[|b - \langle \mathbf{x}, \mathbf{a} \rangle|^2 \right] \right\}$$

- $h_{\mathbf{x}}(\cdot) = \langle \mathbf{x}, \cdot \rangle$
- $b \in \mathbb{R}$ is the desired return & $\mathbf{a} \in \mathbb{R}^p$ are the stock returns
- $\mathcal{X}$ is intersection of the standard simplex and the constraint: $\langle \mathbf{x}, \mathbb{E}[\mathbf{a}] \rangle \geq \rho$.

Stochastic programming

Problem (**Mathematical formulation**)

Consider the following convex minimization problem:

$$f^\star = \min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) := \mathbb{E}[f(\mathbf{x},\theta)] \right\}$$

- θ *is a random vector whose probability distribution is supported on set* Θ.
- $f(\mathbf{x}) := \mathbb{E}[f(\mathbf{x},\theta)]$ *is proper, closed, and convex.*
- *The solution set* $\mathcal{S}^\star := \{\mathbf{x}^\star \in \mathrm{dom}\,(f) : f(\mathbf{x}^\star) = f^\star\}$ *is nonempty.*

Stochastic gradient descent (SGD)

Stochastic gradient descent (SGD)

1. Choose $\mathbf{x}^0 \in \mathbb{R}^p$ and $(\alpha_k)_{k\in\mathbb{N}} \in]0, +\infty[^{\mathbb{N}}$.
2. For $k = 0, 1, \ldots$ perform:

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha_k G(\mathbf{x}^k, \theta_k).$$

○ $G(\mathbf{x}^k, \theta_k)$ is an unbiased estimate of the full gradient:

$$\mathbb{E}[G(\mathbf{x}^k, \theta_k)] = \nabla f(\mathbf{x}^k).$$

Stochastic gradient descent (SGD)

Stochastic gradient descent (SGD)

1. Choose $\mathbf{x}^0 \in \mathbb{R}^p$ and $(\alpha_k)_{k\in\mathbb{N}} \in]0, +\infty[^{\mathbb{N}}$.
2. For $k = 0, 1, \ldots$ perform:

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha_k G(\mathbf{x}^k, \theta_k).$$

◦ $G(\mathbf{x}^k, \theta_k)$ is an unbiased estimate of the full gradient:

$$\mathbb{E}[G(\mathbf{x}^k, \theta_k)] = \nabla f(\mathbf{x}^k).$$

Remarks:

◦ The cost of computing $G(\mathbf{x}^k, \theta_k)$ is n times cheaper than that of $\nabla f(\mathbf{x}^k)$.

◦ As $G(\mathbf{x}^k, \theta_k)$ is an unbiased estimate of the full gradient, SGD would perform well.

◦ We assume $\{\theta_k\}$ are jointly independent.

◦ SGD is not a monotonic descent method.

Example: Convex optimization with finite sums

Convex optimization with finite sums

The problem

$$\arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) := \frac{1}{n}\sum_{j=1}^{n} f_j(\mathbf{x}) \right\},$$

can be rewritten as

$$\arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) := \mathbb{E}_i[f_i(\mathbf{x})] \right\}, \qquad i \text{ is uniformly distributed over } \{1, 2, \cdots, n\}.$$

A stochastic gradient descent (SGD) variant for finite sums

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha_k \nabla f_i(\mathbf{x}^k) \qquad i \text{ is uniformly distributed over} \{1, ..., n\}$$

Remarks:

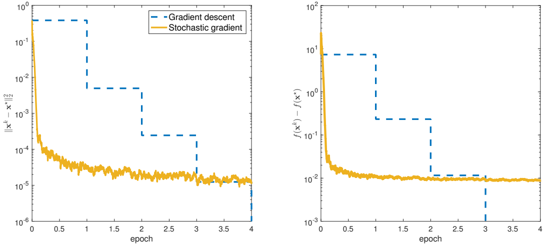
- Note: $\mathbb{E}_i[\nabla f_i(\mathbf{x}^k)] = \sum_{j=1}^{n} \nabla f_j(\mathbf{x}^k)/n = \nabla f(\mathbf{x}^k)$.
- The computational cost of SGD per iteration is p.

Synthetic least-squares problem

$$\min_{\mathbf{x}} \left\{ f(\mathbf{x}) := \frac{1}{2n}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 : \mathbf{x} \in \mathbb{R}^p \right\}$$

Setup

- $\mathbf{A} := \mathrm{randn}(n, p)$ - standard Gaussian $\mathcal{N}(0, \mathbb{I})$, with $n = 10^4$, $p = 10^2$.
- $\mathbf{x}^\natural$ is 50 sparse with zero mean Gaussian i.i.d. entries, normalized to $\|\mathbf{x}^\natural\|_2 = 1$.
- $\mathbf{b} := \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$, where $\mathbf{w}$ is Gaussian white noise with variance 1.



∘ 1 epoch $= 1$ pass over the full gradient

Convergence of SGD when the objective is not strongly convex

Theorem (decaying step-size [27])

Assume

- $\mathbb{E}[\|\mathbf{x}^k - \mathbf{x}^\star\|^2] \leq D^2$ *for all* k,
- $\mathbb{E}[\|G(\mathbf{x}^k, \theta_k)\|^2] \leq M^2$ *(bounded gradient),*
- $\alpha_k = \alpha_0 / \sqrt{k}$.

Then

$$\mathbb{E}[f(\mathbf{x}^k) - f(\mathbf{x}^\star)] \leq \left(\frac{D^2}{\alpha_0} + \alpha_0 M^2\right) \frac{2 + \log k}{\sqrt{k}}.$$

Observation: ∘ $\mathcal{O}(1/\sqrt{k})$ rate is optimal for SGD if we do not consider the strong convexity.

Convergence of SGD for strongly convex problems I

Theorem (strongly convex objective, fixed step-size [4])

Assume

- *f is μ-strongly convex and L-smooth,*
- $\mathbb{E}[\|G(\mathbf{x}^k, \theta_k)\|^2]_2 \leq \sigma^2 + M\|\nabla f(\mathbf{x}^k)\|_2^2$ *(bounded variance),*
- $\alpha_k = \alpha \leq \frac{1}{LM}$.

Then

$$\mathbb{E}[f(\mathbf{x}^k) - f(\mathbf{x}^\star)] \leq \frac{\alpha L \sigma^2}{2\mu} + (1-\mu\alpha)^{k-1}\left(f(\mathbf{x}^1) - f^\star\right).$$

Observations:

- ◦ Converge fast (linearly) to a neighborhood around $\mathbf{x}^\star$.
- ◦ Smaller step-sizes $\alpha \implies$ converge to a better point, but with a slower rate.
- ◦ Zero variance ($\sigma = 0$) $\implies$ linear convergence.
- ◦ This is also known as the relative noise model [24] or the strong growth condition [7].
- ◦ The growth condition is in fact a necessary and sufficient condition for linear convergence [7].
- ◦ The theory applies to the Kaczmarz algorithm (see advanced material).

Convergence of SGD for strongly convex problems II

Theorem (strongly convex objective, decaying step-size [4])

Assume

- *f is μ-strongly convex and L-smooth,*
- *$\mathbb{E}[\|G(\mathbf{x}^k, \theta_k)\|^2]_2 \leq \sigma^2 + M\|\nabla f(\mathbf{x}^k)\|_2^2$ (bounded variance),*
- *$\alpha_k = \frac{c}{k_0+k}$ with some appropriate constants c and k_0.*

Then

$$\mathbb{E}[\|\mathbf{x}^k - \mathbf{x}^\star\|^2] \leq \frac{C}{k+1},$$

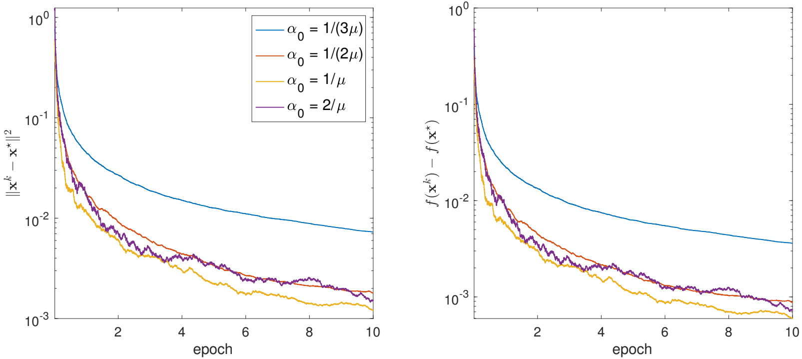
where C is a constant independent of k.

Observations: ◦ Using the L-smooth property,

$$\mathbb{E}[f(\mathbf{x}^k) - f(\mathbf{x}^\star)] \leq L\mathbb{E}[\|\mathbf{x}^k - \mathbf{x}^\star\|^2] \leq \frac{C}{k+1}.$$

◦ The rate is optimal if $\sigma^2 > 0$ with the assumption of strongly-convexity.

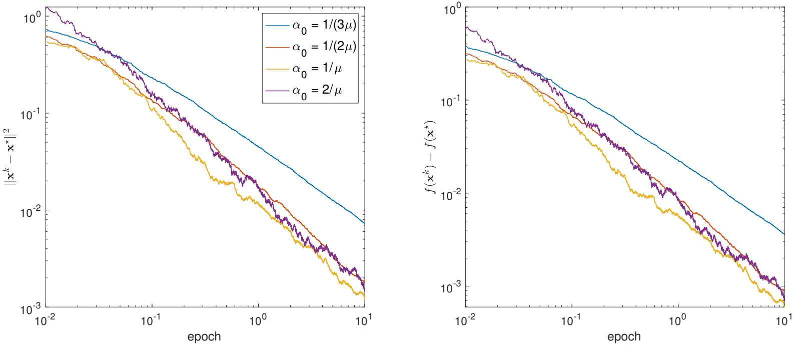
Example: SGD with different step sizes



Setup

- Synthetic least-squares problem as before.
- We use $\alpha_k = \alpha_0/(k + k_0)$.

Example: SGD with different step sizes



Setup

- Synthetic least-squares problem as before.
- We use $\alpha_k = \alpha_0/(k + k_0)$.

Observation: ◦ $\alpha_0 = 1/\mu$ is the best choice.

Comparison with GD

$$\boxed{f^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) := \frac{1}{n}\sum_{j=1}^{n} f_j(\mathbf{x}) \right\}.}$$

∘ f: μ-strongly convex with L-Lipschitz smooth.

	rate	iteration complexity	cost per iteration	total cost
GD	ρ^k	$\log(1/\epsilon)$	n	$n\log(1/\epsilon)$
SGD	$1/k$	$1/\epsilon$	1	$1/\epsilon$

Remark: ∘ SGD is more favorable when n is large — large-scale optimization problems

Motivation for SGD with Averaging

◦ SGD iterates tend to oscillate around global minimizers

◦ Averaging iterates can reduce the oscillation effect

◦ Two types of averaging:

$$\bar{\mathbf{x}}^k = \frac{1}{k}\sum_{j=1}^{k} \alpha_j \mathbf{x}^j \quad \text{(vanilla averaging)}$$

$$\bar{\mathbf{x}}^k = \frac{\sum_{j=1}^{k} \alpha_j \mathbf{x}^j}{\sum_{j=1}^{k} \alpha_j} \quad \text{(weighted averaging)}$$

Remark: ◦ Do not confuse the averaging above with the ones used in Federated Learning.

Convergence for SGD-A I: non-strongly convex case

Stochastic gradient method with averaging (SGD-A)

1. Choose $\mathbf{x}^0 \in \mathbb{R}^p$ and $(\alpha_k)_{k\in\mathbb{N}} \in]0, +\infty[^{\mathbb{N}}$.
2a. For $k = 0, 1, \ldots$ perform:

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha_k G(\mathbf{x}^k, \theta_k).$$

2b. $\bar{\mathbf{x}}^k = (\sum_{j=0}^k \alpha_j)^{-1} \sum_{j=0}^k \alpha_j \mathbf{x}^j$.

Theorem (Convergence of SGD-A [23])

Let $D = \|\mathbf{x}^0 - \mathbf{x}^\star\|$ *and* $\mathbb{E}[\|G(\mathbf{x}^k, \theta_k)\|^2] \leq M^2$.
Then,

$$\mathbb{E}[f(\bar{\mathbf{x}}^{k+1}) - f(\mathbf{x}^\star)] \leq \frac{D^2 + M^2 \sum_{j=0}^k \alpha_j^2}{2\sum_{j=0}^k \alpha_j}.$$

In addition, choosing $\alpha_k = D/(M\sqrt{k+1})$, *we get,*

$$\mathbb{E}[f(\bar{\mathbf{x}}^k) - f(\mathbf{x}^\star)] \leq \frac{MD(2 + \log k)}{\sqrt{k}}.$$

Observation: ◦ Same convergence rate with vanilla SGD.

Convergence for SGD-A II: strongly convex case

Stochastic gradient method with averaging (SGD-A)

1. Choose $\mathbf{x}^0 \in \mathbb{R}^p$ and $(\alpha_k)_{k\in\mathbb{N}} \in]0, +\infty[^{\mathbb{N}}$.
2a. For $k = 0, 1, \ldots$ perform:

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha_k G(\mathbf{x}^k, \theta_k).$$

2b. $\bar{\mathbf{x}}^k = \frac{1}{k}\sum_{j=1}^{k} \mathbf{x}^j$.

Theorem (Convergence of SGD-A [26])

Assume

- *f is μ-strongly convex,*
- $\mathbb{E}[\|G(\mathbf{x}^k, \theta_k)\|^2] \leq M^2$,
- $\alpha_k = \alpha_0/k$ *for some* $\alpha_0 \geq 1/\mu$.

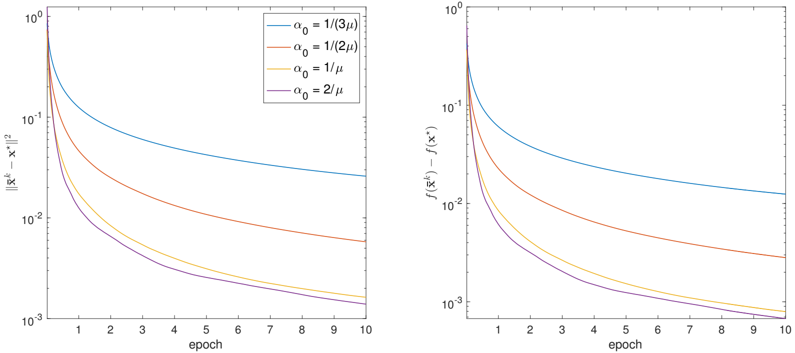
Then

$$\mathbb{E}[f(\bar{\mathbf{x}}^k) - f(\mathbf{x}^\star)] \leq \frac{\alpha_0 M^2(1 + \log k)}{2k}.$$

Observation: ○ Same convergence rate with vanilla SGD.

Example: SGD-A method with different step sizes

$$\min_{\mathbf{x}} \left\{ f(\mathbf{x}) := \frac{1}{2n} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 : \mathbf{x} \in \mathbb{R}^p \right\}$$

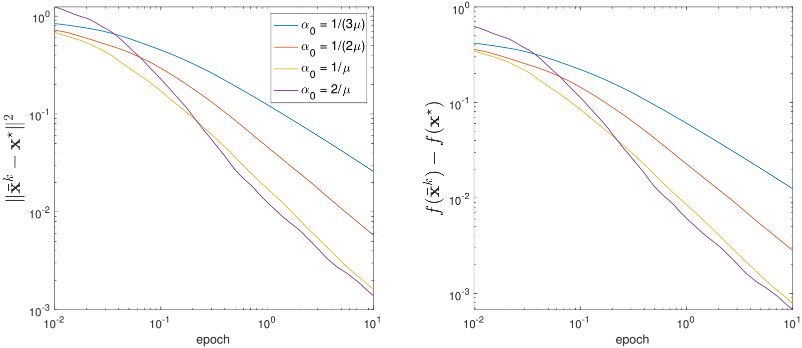


Setup

- ◦ Synthetic least-squares problem as before
- ◦ $\alpha_k = \alpha_0/(k + k_0)$.

Example: SGD-A method with different step sizes

$$\min_{\mathbf{x}} \left\{ f(\mathbf{x}) := \frac{1}{2n} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 : \mathbf{x} \in \mathbb{R}^p \right\}$$



Setup

- Synthetic least-squares problem as before
- $\alpha_k = \alpha_0/(k + k_0)$.

Observations:

- SGD-A is more stable than SGD.
- $\alpha_0 = 2/\mu$ is the best choice.

Least mean squares algorithm

Least-square regression problem

Solve

$$\mathbf{x}^\star \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) := \frac{1}{2}\mathbb{E}_{(\mathbf{a},b)}(\langle \mathbf{a}, \mathbf{x}\rangle - b)^2 \right\},$$

given i.i.d. samples $\{(\mathbf{a}_j, b_j)\}_{j=1}^n$ (particularly in a streaming way).

Stochastic gradient method with averaging

1. Choose $\mathbf{x}^0 \in \mathbb{R}^p$ and $\alpha > 0$.
2a. For $k = 1, \ldots, n$ perform:

$$\mathbf{x}^k = \mathbf{x}^{k-1} - \alpha \left(\langle \mathbf{a}_k, \mathbf{x}^{k-1}\rangle - b_k \right) \mathbf{a}_k.$$

2b. $\bar{\mathbf{x}}^k = \frac{1}{k+1}\sum_{j=0}^{k} \mathbf{x}^j$.

$O(1/k)$ convergence rate, without strongly convexity [2]

Let $\|\mathbf{a}_j\|_2 \leq R$ and $|\langle \mathbf{a}_j, \mathbf{x}^\star\rangle - b_j| \leq \sigma$ a.s.. Pick $\alpha = 1/(4R^2)$. Then, the average sequence $\bar{\mathbf{x}}^{k-1}$ satisfies the following

$$\mathbb{E}f(\bar{\mathbf{x}}^{k-1}) - f^* \leq \frac{2}{k}\left(\sigma\sqrt{p} + R\|\mathbf{x}^0 - \mathbf{x}^\star\|_2\right)^2.$$

Popular SGD Variants

- Mini-batch SGD: For each iteration,

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha_k \frac{1}{b} \sum_{\theta \in \Gamma} G(\mathbf{x}^k, \theta).$$

  - α_k: step-size
  - b : mini-batch size
  - Γ : a set of random variables θ of size b

- Accelerated SGD (Nesterov accelerated technique)

- SGD with Momentum

- Adaptive stochastic methods: AdaGrad...

SGD - Non-convex stochastic optimization

- SGD and several variants are also well-studied for non-convex problems [20].
- Sometimes, there are gaps between SGD's practical performance and theoretical understanding (more later!).
- Recall SGD update rule:

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha_k G(\mathbf{x}^k, \theta)$$

Theorem (A well-known result for SGD & Non-convex problems [14])

Let f be a non-convex and L-smooth function. Set $\alpha_k = \min\left\{\frac{1}{L}, \frac{C}{\sigma\sqrt{T}}\right\}$, $\forall k = 1, ..., T$, where σ^2 is the variance of the gradients and $C > 0$ is constant. Then, it holds that

$$\mathbb{E}[\|\nabla f(\mathbf{x}^R)\|^2] = O\left(\frac{\sigma}{\sqrt{T}}\right),$$

where $\mathbb{P}(R = k) = \frac{2\alpha_k - L\alpha_k^2}{\sum_{k=1}^{T}(2\alpha_k - L\alpha_k^2)}$.

Lower bounds in non-convex optimization

Assumptions on f	Additional assumptions	Sample complexity
L-smooth	Deterministic Oracle $f(\mathbf{x}^0) - \inf_{\mathbf{x}} f(\mathbf{x}) \leq \Delta$	$\Omega(\Delta L \epsilon^{-2})$[6]
L_1-smooth L_2-Lipschitz Hessian	Deterministic Oracle $f(\mathbf{x}^0) - \inf_{\mathbf{x}} f(\mathbf{x}) \leq \Delta$	$\Omega(\Delta L_1^{3/7} L_2^{2/7} \epsilon^{-12/7})$[6]
L-smooth	$\mathbb{E}[G(\mathbf{x},\theta)] = \nabla f(x)$ $\mathbb{E}[\Vert G(\mathbf{x},\theta) - \nabla f(\mathbf{x})\Vert^2] \leq \sigma^2$ $f(\mathbf{x}^0) - \inf_{\mathbf{x}} f(\mathbf{x}) \leq \Delta$	$\Omega(\Delta L \sigma^2 \epsilon^{-4})$[1]
$G(\mathbf{x},\theta)$ has averaged L-Lipschitz gradient $\implies$ L-smooth	$\mathbb{E}[G(\mathbf{x},\theta)] = \nabla f(x)$ $\mathbb{E}[\Vert G(\mathbf{x},\theta) - \nabla f(\mathbf{x})\Vert^2] \leq \sigma^2$ $f(\mathbf{x}^0) - \inf_{\mathbf{x}} f(\mathbf{x}) \leq \Delta$	$\Omega(\Delta L \sigma \epsilon^{-3} + \sigma^2 \epsilon^{-2})$[1]
$f(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^n f_i(\mathbf{x})$ $f_i(\mathbf{x})$ has averaged L-Lipschitz gradient $\implies$ L-smooth	Access to $\nabla f_i(\mathbf{x})$ $f(\mathbf{x}^0) - \inf_{\mathbf{x}} f(\mathbf{x}) \leq \Delta$ $n \leq O(\epsilon^{-4})$[1]	$\Omega(\Delta L \sqrt{n} \epsilon^{-2})$[11]

- Measure of stationarity: $\Vert\nabla f(\mathbf{x})\Vert \leq \epsilon$ or $\mathbb{E}[\Vert\nabla f(\mathbf{x})\Vert \leq \epsilon$
- Sample complexity: $\#$ of total oracle calls (deterministic or stochastic gradients)
- Averaged L-Lipschitz gradient: $\mathbb{E}\left[\Vert\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\Vert^2\right] \leq L^2 \Vert\mathbf{x} - \mathbf{y}\Vert^2$
- $G(\mathbf{x},\theta)$ denotes a stochastic gradient estimate for f at $\mathbf{x}$ with randomness governed by θ.

[1] We have $n \leq O(\epsilon^{-4})$ in order to match the respective *upper bound* of $O(n + \sqrt{n}\epsilon^{-2})$ achieved by [11]

Non-smooth minimization: A simple example

What if we simultaneously want $f_1(x), f_2(x), \ldots, f_k(x)$ to be small?

A natural approach in some cases: Minimize $f(x) = \max\{f_1(x), \ldots, f_k(x)\}$

- *The good news*: If each $f_i(x)$ is convex, then $f(x)$ is convex
- *The bad (!) news*: Even if each $f_i(x)$ is smooth, $f(x)$ may be non-smooth
  - e.g., $f(x) = \max\{x, x^2\}$

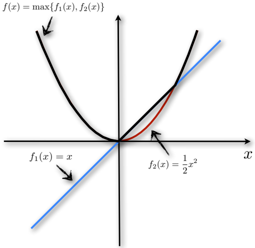


Figure: Maximum of two smooth convex functions.

A statistical learning motivation for non-smooth optimization

Linear Regression

Consider the classical linear regression problem:

$$\mathbf{b} = \mathbf{A}\mathbf{x}^{\natural} + \mathbf{w}$$

with $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{n \times p}$ are known, $\mathbf{x}^{\natural}$ is unknown, and $\mathbf{w}$ is noise. Assume *for now* that $n \geq p$ (more later).

A statistical learning motivation for non-smooth optimization

Linear Regression

Consider the classical linear regression problem:

$$\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$$

with $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{n \times p}$ are known, $\mathbf{x}^\natural$ is unknown, and $\mathbf{w}$ is noise. Assume *for now* that $n \geq p$ (more later).

- *Standard approach:* Least squares: $\mathbf{x}^\star_{\text{LS}} \in \arg\min_{\mathbf{x}} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2$
  - Convex, smooth, and an *explicit solution*: $\mathbf{x}^\star_{\text{LS}} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b} = \mathbf{A}^\dagger\mathbf{b}$
- *Alternative approach:* Least absolute value deviation: $\mathbf{x}^\star \in \arg\min_{\mathbf{x}} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_1$
  - The advantage: Improved robustness against outliers (i.e., less sensitive to high noise values)
  - The bad (!) news: A *non-differentiable* objective function

Our main motivating example this lecture: The case $n \ll p$

Deficiency of smooth models

Recall the practical performance of an estimator $\mathbf{x}^\star$.

Practical performance

Denote the numerical approximation at time t by $\mathbf{x}^t$. The practical performance is determined by

$$\|\mathbf{x}^t - \mathbf{x}^\natural\|_2 \leq \underbrace{\|\mathbf{x}^t - \mathbf{x}^\star\|_2}_{\text{numerical error}} + \underbrace{\|\mathbf{x}^\star - \mathbf{x}^\natural\|_2}_{\text{statistical error}}.$$

Remarks:

- *Non-smooth* estimators of $\mathbf{x}^\natural$ can help *reduce the statistical error*.
- This improvement *may* require higher computational costs.

Example: Least-squares estimation in the linear model

◦ Recall the linear model and the LS estimator.

LS estimation in the linear model

Let $\mathbf{x}^\natural \in \mathbb{R}^p$ and $\mathbf{A} \in \mathbb{R}^{n \times p}$. The samples are given by $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$, where $\mathbf{w}$ denotes the unknown noise. The LS estimator for $\mathbf{x}^\natural$ given $\mathbf{A}$ and $\mathbf{b}$ is defined as

$$\mathbf{x}^\star_{\mathsf{LS}} \in \arg\min_{\mathbf{x} \in \mathbb{R}^p} \left\{ \| \mathbf{b} - \mathbf{A}\mathbf{x} \|_2^2 \right\}.$$

Remarks:

◦ If $\mathbf{A}$ has full column rank, $\mathbf{x}^\star_{\mathsf{LS}} = \mathbf{A}^\dagger \mathbf{b}$ is uniquely defined.

◦ *When $n < p$*, $\mathbf{A}$ cannot have full column rank, and hence $\mathbf{x}^\star_{\mathsf{LS}} \in \left\{ \mathbf{A}^\dagger \mathbf{b} + \mathbf{h} : \mathbf{h} \in \mathrm{null}\,(\mathbf{A}) \right\}$.

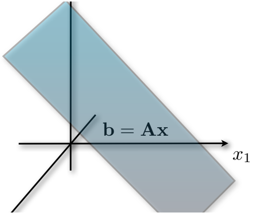
Observation:

◦ The estimation error $\| \mathbf{x}^\star_{\mathsf{LS}} - \mathbf{x}^\natural \|_2$ can be *arbitrarily large*!

A candidate solution

Continuing the LS example:

- There exist infinitely many $\mathbf{x}$'s such that $\mathbf{b} = \mathbf{A}\mathbf{x}$
- Suppose that $\mathbf{w} = 0$ (i.e. no noise). Let us just choose the one $\hat{\mathbf{x}}_{\text{candidate}}$ with the smallest norm $\|\mathbf{x}\|_2$.



Observation: ○ Unfortunately, *this still fails when $n < p$*

A candidate solution contd.

Proposition ([16])

Suppose that $\mathbf{A} \in \mathbb{R}^{n \times p}$ *is a matrix of i.i.d. standard Gaussian random variables, and* $\mathbf{w} = \mathbf{0}$*. We have*

$$(1-\epsilon)\left(1 - \frac{n}{p}\right) \| \mathbf{x}^\natural \|_2^2 \leq \| \hat{\mathbf{x}}_{\mathrm{candidate}} - \mathbf{x}^\natural \|_2^2 \leq (1-\epsilon)^{-1}\left(1 - \frac{n}{p}\right) \| \mathbf{x}^\natural \|_2^2$$

with probability at least $1 - 2\exp\left[-(1/4)(p-n)\epsilon^2\right] - 2\exp\left[-(1/4)p\epsilon^2\right]$*, for all* $\epsilon > 0$ *and* $\mathbf{x}^\natural \in \mathbb{R}^p$*.*

Summarizing the findings so far

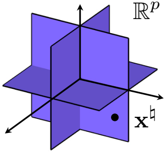
The message so far:

- Even in the absence of noise, we cannot recover $\mathbf{x}^\natural$ from the observations $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural$ unless $n \geq p$
- But in applications, p might be thousands, millions, billions...
- **Can we get away with $n \ll p$ under some further assumptions on $\mathbf{x}$?**

A natural signal model

Definition (s-sparse vector)

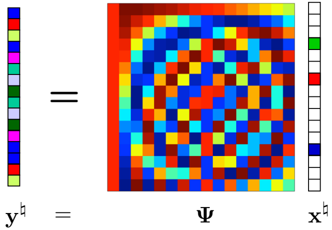
A vector $\mathbf{x} \in \mathbb{R}^p$ is s-sparse if it has at most s non-zero entries.

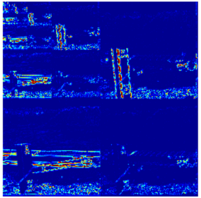


Sparse representations

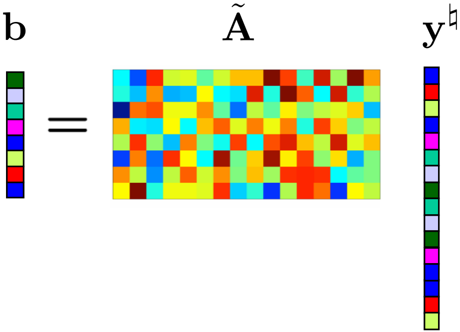
$\mathbf{x}^\natural$: *sparse* transform coefficients

- Basis representations $\Psi \in \mathbb{R}^{p \times p}$
  - *Wavelets*, DCT, ...
- Frame representations $\Psi \in \mathbb{R}^{m \times p}$, $m > p$
  - Gabor, curvelets, shearlets, ...
- Other *dictionary* representations...



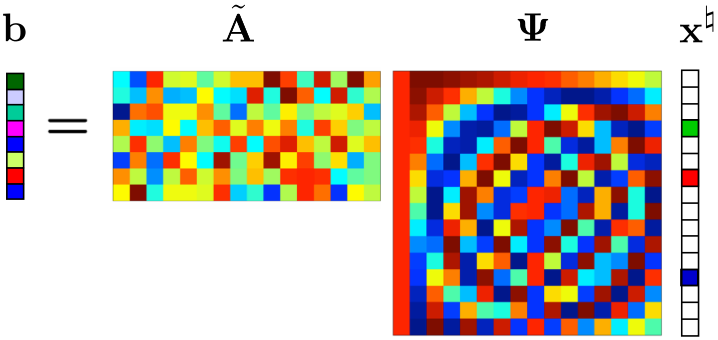


Sparse representations strike back!



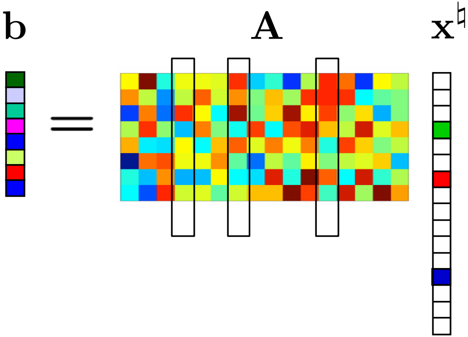
- $\mathbf{b} \in \mathbb{R}^n$, $\tilde{\mathbf{A}} \in \mathbb{R}^{n \times p}$, and $n < p$

Sparse representations strike back!



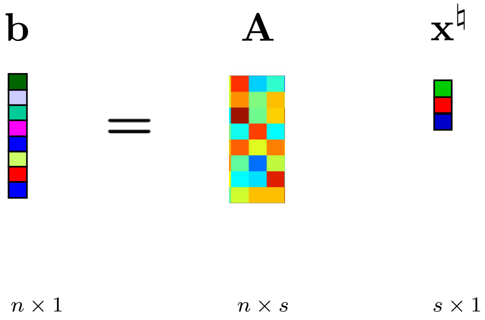
- $\mathbf{b} \in \mathbb{R}^n$, $\tilde{\mathbf{A}} \in \mathbb{R}^{n \times p}$, and $n < p$
- $\boldsymbol{\Psi} \in \mathbb{R}^{p \times p}$, $\mathbf{x}^\natural \in \mathbb{R}^p$, and $\|\mathbf{x}^\natural\|_0 \leq s < n$

Sparse representations strike back!



- $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{n \times p}$, and $\mathbf{x}^\natural \in \mathbb{R}^p$, and $\|\mathbf{x}^\natural\|_0 \leq s < n < p$

Sparse representations strike back!



Observations:

- The matrix $\mathbf{A}$ effectively becomes *overcomplete*.
- We could solve for $\mathbf{x}^\natural$ if we knew *the location of the non-zero entries of* $\mathbf{x}^\natural$.

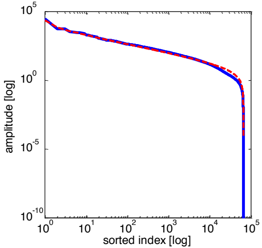
Compressible signals

∘ Real signals may not be exactly sparse, but approximately sparse, or *compressible*.

Definition (Compressible signals)

Roughly speaking, a vector $\mathbf{x} := (x_1, \ldots, x_p)^T \in \mathbb{R}^p$ is compressible if the number of its significant components (i.e., entries larger than some $\epsilon > 0$: $|\{k : |x_k| \geq \epsilon, 1 \leq k \leq p\}|$) is small.

- **Cameraman**@MIT.



- **Solid curve**: Sorted wavelet coefficients of the cameraman image.
- **Dashed curve**: Expected order statistics of generalized Pareto distribution with shape parameter 1.67.

A different tale of the linear model $\mathrm{b} = \mathrm{Ax} + \mathrm{w}$

A *realistic* linear model

Let $\mathbf{b} := \tilde{\mathbf{A}}\mathbf{y}^\natural + \tilde{\mathbf{w}} \in \mathbb{R}^n$.

- Let $\mathbf{y}^\natural := \Psi\mathbf{x}_{\text{real}} \in \mathbb{R}^m$ that admits a *compressible* representation $\mathbf{x}_{\text{real}}$.
- Let $\mathbf{x}_{\text{real}} \in \mathbb{R}^p$ that is *compressible* and let $\mathbf{x}^\natural$ be its *best s-term approximation*.
- Let $\tilde{\mathbf{w}} \in \mathbb{R}^n$ denote the possibly nonzero *noise* term.
- Assume that $\Psi \in \mathbb{R}^{m\times p}$ and $\tilde{\mathbf{A}} \in \mathbb{R}^{n\times m}$ are known.

Then we have

$$\begin{aligned}\mathbf{b} &= \tilde{\mathbf{A}}\Psi\left(\mathbf{x}^\natural + \mathbf{x}_{\text{real}} - \mathbf{x}^\natural\right) + \tilde{\mathbf{w}}.\\ &:= \underbrace{\left(\tilde{\mathbf{A}}\Psi\right)}_{\mathbf{A}}\mathbf{x}^\natural + \underbrace{\left[\tilde{\mathbf{w}} + \tilde{\mathbf{A}}\Psi\left(\mathbf{x}_{\text{real}} - \mathbf{x}^\natural\right)\right]}_{\mathbf{w}},\end{aligned}$$

equivalently, $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$.

Peeling the onion

- The *realistic* linear model uncovers yet another level of difficulty

Practical performance

The practical performance at time t is determined by

$$\|\mathbf{x}^t - \mathbf{x}_{\text{real}}\|_2 \leq \underbrace{\|\mathbf{x}^t - \mathbf{x}^\star\|_2}_{\text{numerical error}} + \underbrace{\|\mathbf{x}^\star - \mathbf{x}^\natural\|_2}_{\text{statistical error}} + \underbrace{\|\mathbf{x}_{\text{real}} - \mathbf{x}^\natural\|_2}_{\text{model error}}.$$

Approach 1: Sparse recovery via exhaustive search

Approach 1 for estimating $\mathbf{x}^\natural$ from $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$

We may search over all $\binom{p}{s}$ subsets $S \subset \{1, \ldots, p\}$ of cardinality s, solve the restricted least-squares problem $\min_{\mathbf{x}_S} \|\mathbf{b} - \mathbf{A}_S \mathbf{x}_S\|_2^2$, and return the resulting $\mathbf{x}$ corresponding to the smallest error, putting zeros in the entries of $\mathbf{x}$ outside S.

◦ Stable and robust recovery of any s-sparse signal is possible using just $n = 2s$ measurements.

Approach 1: Sparse recovery via exhaustive search

Approach 1 for estimating $\mathbf{x}^\natural$ from $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$

We may search over all $\binom{p}{s}$ subsets $S \subset \{1, \ldots, p\}$ of cardinality s, solve the restricted least-squares problem $\min_{\mathbf{x}_S} \|\mathbf{b} - \mathbf{A}_S \mathbf{x}_S\|_2^2$, and return the resulting $\mathbf{x}$ corresponding to the smallest error, putting zeros in the entries of $\mathbf{x}$ outside S.

◦ Stable and robust recovery of any s-sparse signal is possible using just $n = 2s$ measurements.

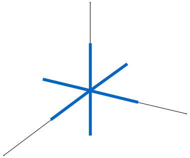
Issues

- $\binom{p}{s}$ is a huge number - too many to search!
- s is not known in practice

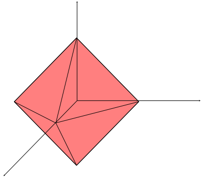
The ℓ_1-norm heuristic

Heuristic: The *ℓ_1-ball with radius c_∞* is an "approximation" of the set of sparse vectors $\hat{\mathbf{x}} \in \{\mathbf{x} : \|\mathbf{x}\|_0 \leq s, \|\mathbf{x}\|_\infty \leq c_\infty\}$ parameterized by their sparsity s and maximum amplitude c_∞.

$$\hat{\mathbf{x}} \in \{\mathbf{x} : \|\mathbf{x}\|_1 \leq c_\infty\} \quad \text{with some } c_\infty > 0.$$



The set
$\left\{\mathbf{x} : \|\mathbf{x}\|_0 \leq 1, \|\mathbf{x}\|_\infty \leq 1, \mathbf{x} \in \mathbb{R}^3\right\}$



The unit ℓ_1-norm ball
$\left\{\mathbf{x} : \|\mathbf{x}\|_1 \leq 1, \mathbf{x} \in \mathbb{R}^3\right\}$

Remark: ◦ This heuristic leads to the so-called *Lasso* optimization problem.

Sparse recovery via the Lasso

Definition (Least absolute shrinkage and selection operator (Lasso))

$$\mathbf{x}^\star_{Lasso} := \arg\min_{\mathbf{x}\in\mathbb{R}^p} \|\,\mathbf{b} - \mathbf{A}\mathbf{x}\,\|_2^2 + \rho\|\,\mathbf{x}\,\|_1$$

with some $\rho \geq 0$.

- The second term in the objective function is called the *regularizer*.
- The parameter ρ is called the *regularization parameter*. It is used to trade off the objectives:
  - Minimize $\|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2$, so that the solution is consistent with the observations
  - Minimize $\|\mathbf{x}\|_1$, so that the solution has the desired sparsity structure

Remark: ◦ The Lasso has a *convex* but *non-smooth* objective function

Performance of the Lasso

Theorem (Existence of a stable solution in polynomial time [22])

This Lasso convex formulation is a second order cone program, which can be solved in polynomial time in terms of the inputs n and p. Surprisingly, if the signal $\mathbf{x}^\natural$ is s-sparse and the noise $\mathbf{w}$ is sub-Gaussian (e.g., Gaussian or bounded) with parameter σ, then choosing $\rho = \sqrt{\frac{16\sigma^2 \log p}{n}}$ yields an error of

$$\| \mathbf{x}^\star_{Lasso} - \mathbf{x}^\natural \|_2 \leq \frac{8\sigma}{\kappa(\mathbf{A})} \sqrt{\frac{s \ln p}{n}},$$

with probability at least $1 - c_1 \exp(-c_2 n \rho^2)$, where c_1 and c_2 are absolute constants, and $\kappa(\mathbf{A}) > 0$ encodes the difficulty of the problem.

Remark:

- The number of measurements is $\mathcal{O}\left(s \ln p\right)$ – this may be *much* smaller than p!

Non-smooth unconstrained convex minimization

Problem (**Mathematical formulation**)

How can we find an optimal solution to the following optimization problem?

$$\boxed{F^\star := \min_{\mathbf{x}\in\mathbb{R}^p} f(\mathbf{x})} \tag{1}$$

where f is proper, closed, convex, but not everywhere differentiable.

Subdifferentials: A generalization of the gradient

Definition

Let $f : \mathcal{Q} \to \mathbb{R} \cup \{+\infty\}$ be a convex function. The subdifferential of f at a point $\mathbf{x} \in \mathcal{Q}$ is defined by the set:

$$\partial f(\mathbf{x}) = \{\mathbf{v} \in \mathbb{R}^p \ : \ f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{v}, \ \mathbf{y} - \mathbf{x} \rangle \text{ for all } \mathbf{y} \in \mathcal{Q}\}.$$

Each element $\mathbf{v}$ of $\partial f(\mathbf{x})$ is called *subgradient* of f at $\mathbf{x}$.

Lemma

Let $f : \mathcal{Q} \to \mathbb{R} \cup \{+\infty\}$ be a differentiable convex function. Then, the subdifferential of f at a point $\mathbf{x} \in \mathcal{Q}$ contains only the gradient, i.e., $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$.

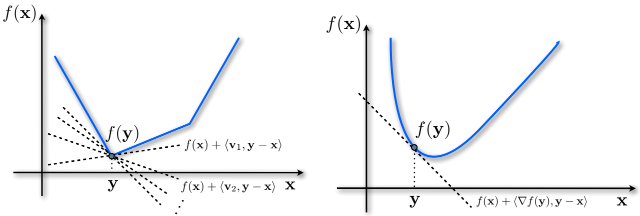


Figure: (**Left**) Non-differentiability at point $\mathbf{y}$. (**Right**) Gradient as a subdifferential with a singleton entry.

(Sub)gradients in convex functions

Example

$f(x) = |x| \quad \longrightarrow \quad \partial|x| = \{\text{sgn}(x)\}\,,$ if $x \neq 0,$ but $[-1,1],$ if $x = 0.$

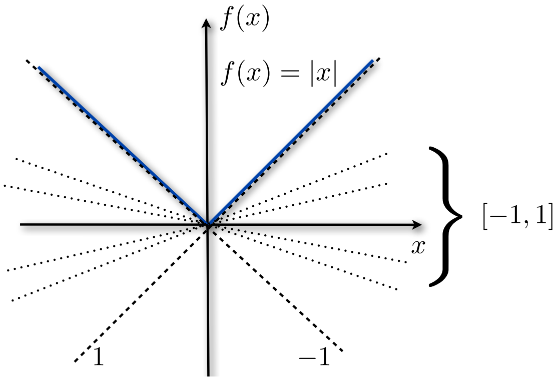


Figure: Subgradients of $f(x) = |x|$ in $\mathbb{R}$.

Subdifferentials: Two basic results

Lemma (Necessary and sufficient condition)

$$\mathbf{x}^\star \in \operatorname{dom}(F) \textit{ is a } \textbf{\textit{globally optimal}} \textit{ solution to } (1) \quad \textbf{\textit{iff}} \quad 0 \in \partial F(\mathbf{x}^\star).$$

Sketch of the proof.

- $\Leftarrow$: For any $\mathbf{x} \in \mathbb{R}^p$, by definition of $\partial F(\mathbf{x}^\star)$:

$$F(\mathbf{x}) - F(\mathbf{x}^\star) \geq 0^T(\mathbf{x} - \mathbf{x}^\star) = 0,$$

that is, $\mathbf{x}^\star$ is a global solution to (1).

- $\Rightarrow$: If $\mathbf{x}^\star$ is a global of (1) then for every $\mathbf{x} \in \operatorname{dom}(F)$, $F(\mathbf{x}) \geq F(\mathbf{x}^\star)$ and hence

$$F(\mathbf{x}) - F(\mathbf{x}^\star) \geq 0^T(\mathbf{x} - \mathbf{x}^\star), \forall \mathbf{x} \in \mathbb{R}^p,$$

which leads to $0 \in \partial F(\mathbf{x}^\star)$. □

Theorem (Moreau-Rockafellar's theorem [25])

Let ∂f and ∂g be the subdifferential of f and g, respectively. If $f, g \in \mathcal{F}(\mathbb{R}^p)$ and $\operatorname{dom}(f) \cap \operatorname{dom}(g) \neq \emptyset$, then:

$$\boxed{\partial(f+g) = \partial f + \partial g.}$$

Non-smooth unconstrained convex minimization

Problem (**Non-smooth convex minimization**)

$$F^\star := \min_{\mathbf{x}\in\mathbb{R}^p} f(\mathbf{x}) \tag{2}$$

Subgradient method

The subgradient method relies on the fact that even though f is non-smooth, we can still compute its **subgradients**, informing of the local descent directions.

Subgradient method

1. Choose $\mathbf{x}^0 \in \mathbb{R}^p$ as a starting point.
2. For $k = 0, 1, \cdots,$ perform:

$$\begin{cases} \mathbf{x}^{k+1} &= \mathbf{x}^k - \alpha_k \mathbf{d}^k, \end{cases} \tag{3}$$

where $\mathbf{d}^k \in \partial f(\mathbf{x}^k)$ and $\alpha_k \in (0, 1]$ is a given step size.

Convergence of the subgradient method

Theorem

Assume that the following conditions are satisfied:

1. $\|\mathbf{g}\|_2 \leq G$ *for all* $\mathbf{g} \in \partial f(\mathbf{x})$ *for any* $\mathbf{x} \in \mathbb{R}^p$.
2. $\|\mathbf{x}^0 - \mathbf{x}^\star\|_2 \leq R$

Let the stepsize be chosen as

$$\alpha_k = \frac{R}{G\sqrt{k}}$$

then the iterates generated by the subgradient method satisfy

$$\min_{0 \leq i \leq k} f(\mathbf{x}^i) - f^\star \leq \frac{RG}{\sqrt{k}}.$$

Remarks

- Condition (1) holds, for example, when f is G-Lipschitz.
- **The convergence rate of $\mathcal{O}\left(1/\sqrt{k}\right)$ is the slowest we have seen so far!**

Stochastic subgradient methods

∘ An unbiased stochastic subgradient

$$\mathbb{E}[G(\mathbf{x})|\mathbf{x}] \in \partial f(\mathbf{x}).$$

∘ Stochastic gradient methods using unbiased subgradients instead of unbiased gradients work

The classic stochastic subgradient methods (SG)
1. Choose $\mathbf{x}_1 \in \mathbb{R}^p$ and $(\gamma_k)_{k\in\mathbb{N}} \in (0, +\infty)^{\mathbb{N}}$. **2.** For $k = 1, \ldots$ perform: $$\mathbf{x}_{k+1} = \mathbf{x}_k - \gamma_k G(\mathbf{x}_k).$$

Theorem (Convergence in expectation [27])

Suppose that:

1. $\mathbb{E}[\|G(\mathbf{x}^k)\|^2] \leq M^2$,
2. $\gamma_k = \gamma_0/\sqrt{k}$.

Then,

$$\mathbb{E}[f(\mathbf{x}^k) - f(\mathbf{x}^\star)] \leq \left(\frac{D^2}{\gamma_0} + \gamma_0 M^2\right)\frac{2 + \log k}{\sqrt{k}}.$$

Remark: ∘ The rate is $\mathcal{O}(\log k/\sqrt{k})$ instead of $\mathcal{O}(1/\sqrt{k})$ for the deterministic algorithm.

Wrap up!

- Three supplementary lectures to take a look once the course is over!
  - One on compressive sensing (Math of Data Lecture 4 from 2014):
    https://archive-wp.epfl.ch/lions/wp-content/uploads/2019/01/lecture-4-2014.pdf
  - One on source separation (Math of Data Lecture 6 from 2014)
    https://archive-wp.epfl.ch/lions/wp-content/uploads/2019/01/lecture-6-2014.pdf
  - One on convexification of structured sparsity models (research presentation)
    https://www.epfl.ch/labs/lions/wp-content/uploads/2019/01/volkan-TU-view-web.pdf

⋆Adaptive methods for stochastic optimization

Remark

- Adaptive methods have extensive applications in stochastic optimization.
- We will see **another nature** of adaptive methods in this lecture.
- Mild additional assumption: **bounded variance** of gradient estimates.

*AdaGrad for stochastic optimization

- Only modification: $\nabla f(\mathbf{x}) \Rightarrow G(\mathbf{x}, \theta)$

AdaGrad with $\mathbf{H}_k = \lambda_k \mathbf{I}$ [17]

1. Set $Q^0 = 0$.
2. For $k = 0, 1, \ldots,$ iterate

$$\begin{cases} Q^k &= Q^{k-1} + \|G(\mathbf{x}^k, \theta)\|^2 \\ \mathbf{H}_k &= \sqrt{Q^k}\mathbf{I} \\ \mathbf{x}^{k+1} &= \mathbf{x}_t - \alpha_k \mathbf{H}_k^{-1} G(\mathbf{x}^k, \theta) \end{cases}$$

Theorem (Convergence rate: stochastic, convex optimization [17])

Assume f is convex and L-smooth, such that minimizer of f lies in a convex, compact set $\mathcal{K}$ with diameter D. Also consider bounded variance for unbiased gradient estimates, i.e., $\mathbb{E}\left[\|G(\mathbf{x}, \theta) - \nabla f(\mathbf{x})\|^2 | \mathbf{x}\right] \leq \sigma^2$. Then,

$$\mathbb{E}[f(\mathbf{x}^k)] - \min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) = O\left(\frac{\sigma D}{\sqrt{k}}\right)$$

- AdaGrad is **adaptive** also in the sense that it adapts to nature of the oracle.

*AcceleGrad for stochastic optimization

○ Similar to AdaGrad, replace $\nabla f(\mathbf{x}) \Rightarrow G(\mathbf{x}, \theta)$

AcceleGrad (Accelerated Adaptive Gradient Method)

Input : $\mathbf{x}^0 \in \mathcal{K}$, diameter D, weights $\{\alpha_k\}_{k\in\mathbb{N}}$, learning rate $\{\eta_k\}_{k\in\mathbb{N}}$

1. Set $\mathbf{y}^0 = \mathbf{z}^0 = \mathbf{x}^0$

2. For $k = 0, 1, \ldots,$ iterate

$$\begin{cases} \tau_k & := 1/\alpha_k \\ \mathbf{x}^{k+1} & = \tau_t \mathbf{z}^k + (1-\tau_k)\mathbf{y}^k, \text{define } \mathbf{g}_k := \nabla f(\mathbf{x}^{k+1}) \\ \mathbf{z}^{k+1} & = \Pi_{\mathcal{K}}(\mathbf{z}^k - \alpha_k \eta_k \mathbf{g}_k) \\ \mathbf{y}^{k+1} & = \mathbf{x}^{k+1} - \eta_k \mathbf{g}_k \end{cases}$$

Output : $\overline{\mathbf{y}}^k \propto \sum_{i=0}^{k-1} \alpha_i \mathbf{y}^{i+1}$

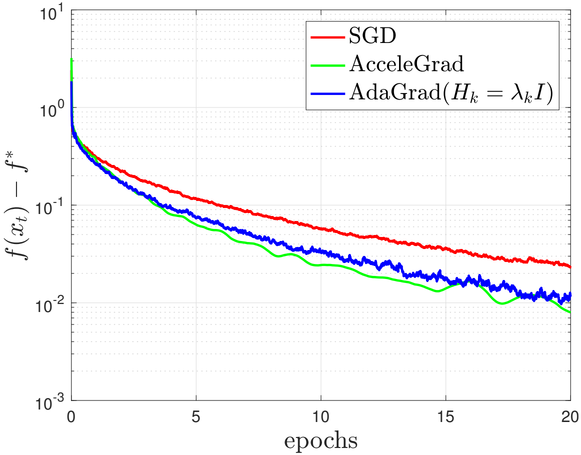
Theorem (Convergence rate [18])

Assume f is convex and G-Lipschitz and that minimizer of f lies in a convex, compact set $\mathcal{K}$ with diameter D. Also consider bounded variance for unbiased gradient estimates, i.e., $\mathbb{E}\left[\|G(\mathbf{x},\theta) - \nabla f(\mathbf{x})\|^2 | \mathbf{x}\right] \leq \sigma^2$. *Then,*

$$\mathbb{E}[f(\overline{\mathbf{y}}^k)] - \min_{\mathbf{x}} f(\mathbf{x}) = O\left(\frac{GD\sqrt{\log k}}{\sqrt{k}}\right).$$

*Example: Synthetic least squares

- $\mathbf{A} \in \mathbb{R}^{n\times d}$, where $n = 200$ and $d = 50$.
- Number of epochs: 20.
- Algorithms: SGD, AdaGrad & AcceleGrad.



$\star$UniXGrad for stochastic optimization

UniXGrad

1. Set $\mathbf{x}^0 = \mathbf{z}^0 = \mathbf{x}^0$
2. For $k = 0, 1, \ldots,$ iterate

$$\begin{cases} \mathbf{x}^{k+1/2} &= \Pi_{\mathcal{X}}\left(\mathbf{x}^k - \alpha_k \eta_k \nabla f(\tilde{\mathbf{x}}^k)\right) \\ \mathbf{x}^{k+1} &= \Pi_{\mathcal{X}}\left(\mathbf{x}^k - \alpha_k \eta_k \nabla f(\bar{\mathbf{x}}^{k+1/2})\right) \end{cases}$$

- $\Pi_{\mathcal{X}}(\mathbf{x})$ is Euclidean projection onto $\mathcal{X}$ and $\alpha_k = k$
- $\tilde{\mathbf{x}}^k = \frac{\alpha_k \mathbf{x}^k + \sum_{i=1}^{k-1} \alpha_i \mathbf{x}^{i+1/2}}{\sum_{i=1}^{k} \alpha_i}$, $\quad \bar{\mathbf{x}}^{k+1/2} = \frac{\sum_{i=1}^{k} \alpha_i \mathbf{x}^{i+1/2}}{\sum_{i=1}^{k} \alpha_i}$
- $\eta_k = \frac{2D}{\sqrt{1 + \sum_{i=1}^{k} (\alpha_k)^2 \|\nabla f(\bar{\mathbf{x}}^{k+1/2}) - \nabla f(\tilde{\mathbf{x}}^k)\|^2}}$

Theorem (Convergence rate of UniXGrad)

Let the sequence $\{\mathbf{x}^{k+1/2}\}$ be generated by UniXGrad. Under the assumptions

- *f is convex and L-smooth,*
- *Constraint set $\mathcal{X}$ has bounded diameter, i.e., $D = \max_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \|\mathbf{x} - \mathbf{y}\|$,*
- $\mathbb{E}[\tilde{\nabla} f(\mathbf{x}) | \mathbf{x}] = \nabla f(\mathbf{x})$ *and* $\mathbb{E}[\|\tilde{\nabla} f(\mathbf{x}) - \nabla f(\mathbf{x})\|^2 | \mathbf{x}] \leq \sigma^2$

UniXGrad guarantees the following:

$$f(\bar{\mathbf{x}}^{k+1/2}) - \min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) \leq O\left(\frac{LD^2}{k^2} + \frac{\sigma D}{\sqrt{k}}\right).$$

*Randomized Kaczmarz algorithm

Problem

Given a full-column-rank matrix $\mathbf{A} \in \mathbb{R}^{n \times p}$ and $b \in \mathbb{R}^n$, solve the linear system

$$\mathbf{A}\mathbf{x} = \mathbf{b}.$$

Notations: $\mathbf{b} := (b_1, \ldots, b_n)^T$ and $\mathbf{a}_j^T$ is the j-th row of $\mathbf{A}$.

Randomized Kaczmarz algorithm (RKA)

1. Choose $\mathbf{x}^0 \in \mathbb{R}^p$.
2. For $k = 0, 1, \ldots$ perform:
   - 2a. Pick $j_k \in \{1, \cdots, n\}$ randomly with $\Pr(j_k = i) = \|\mathbf{a}_i\|_2^2 / \|\mathbf{A}\|_F^2$
   - 2b. $\mathbf{x}^{k+1} = \mathbf{x}^k - \left(\langle \mathbf{a}_{j_k}, \mathbf{x}^k \rangle - b_{j_k}\right) \mathbf{a}_{j_k} / \|\mathbf{a}_{j_k}\|_2^2$.

Linear convergence [28]

Let $\mathbf{x}^\star$ be the solution of $\mathbf{A}\mathbf{x} = \mathbf{b}$ and $\kappa = \|\mathbf{A}\|_F \|\mathbf{A}^{-1}\|$. Then

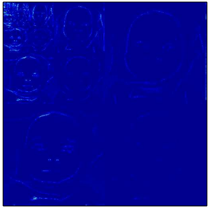
$$\mathbb{E}\|\mathbf{x}^k - \mathbf{x}^\star\|_2^2 \leq (1 - \kappa^{-2})^k \|\mathbf{x}^0 - \mathbf{x}^*\|_2^2$$

◦ RKA can be seen as a particular case of SGD [21].

*Other models with simplicity

p

pixels

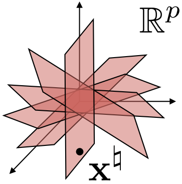


Information level:

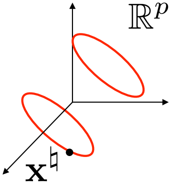
$s \ll p$

large wavelet coefficients

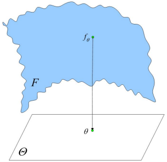
(blue = 0)



sparse signals



low-rank matrices



nonlinear models

There are many models extending far beyond sparsity, coming with other non-smooth regularizers.

*Generalization via simple representations

Definition (Atomic sets & atoms [8])

An *atomic set* $\mathcal{A}$ is a set of vectors in $\mathbb{R}^p$. An *atom* is an element in an atomic set.

Terminology (Simple representation [8])

A parameter $\mathbf{x}^\natural \in \mathbb{R}^p$ admits a simple representation with respect to an atomic set $\mathcal{A} \subseteq \mathbb{R}^p$, if it can be represented as a non-negative combination of few atoms, i.e., $\mathbf{x}^\natural = \sum_{i=1}^k c_i \mathbf{a}_i, \quad \mathbf{a}_i \in \mathcal{A},\ c_i \geq 0.$

Example (Sparse parameter)

Let $\mathbf{x}^\natural$ be s-sparse. Then $\mathbf{x}^\natural$ can be represented as the non-negative combination of s elements in $\mathcal{A}$, with $\mathcal{A} := \{\pm\mathbf{e}_1, \ldots, \pm\mathbf{e}_p\}$, where $\mathbf{e}_i := (\delta_{1,i}, \delta_{2,i}, \ldots, \delta_{p,i})$ for all i.

Example (Sparse parameter with a dictionary)

Let $\Psi \in \mathbb{R}^{m \times p}$, and let $\mathbf{y}^\natural := \Psi\mathbf{x}^\natural$ for some s-sparse $\mathbf{x}^\natural$. Then $\mathbf{y}^\natural$ can be represented as the non-negative combination of s elements in $\mathcal{A}$, with $\mathcal{A} := \{\pm\psi_1, \ldots, \pm\psi_p\}$, where ψ_k denotes the kth column of Ψ.

*Atomic norms

◦ Recall the Lasso problem

$$\mathbf{x}^\star_{\text{Lasso}} := \arg\min_{\mathbf{x}\in\mathbb{R}^p} \|\,\mathbf{b} - \mathbf{A}\mathbf{x}\,\|_2^2 + \rho\|\,\mathbf{x}\,\|_1$$

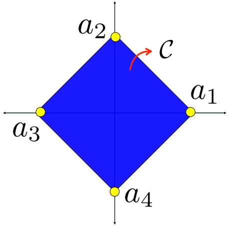
Observations: ◦ ℓ_1-norm is the *atomic norm* associated with the atomic set $\mathcal{A} := \{\pm\mathbf{e}_1, \ldots, \pm\mathbf{e}_p\}$.

◦ The norm is closely tied with the convex hull of the set.

◦ We can extend the same principle for a wide range of regularizers

$$\mathcal{A} := \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \end{bmatrix} \right\}.$$

$$\mathcal{C} := \operatorname{conv}(\mathcal{A}).$$



*Gauge functions and atomic norms

Definition (Gauge function)

Let $\mathcal{C}$ be a convex set in $\mathbb{R}^p$, the **gauge function** associated with $\mathcal{C}$ is given by

$$g_{\mathcal{C}}(\mathbf{x}) := \inf\{t > 0 : \mathbf{x} = t\mathbf{c} \text{ for some } \mathbf{c} \in \mathcal{C}\}.$$

Definition (Atomic norm)

Let $\mathcal{A}$ be a symmetric *atomic set* in $\mathbb{R}^p$ such that if $\mathbf{a} \in \mathcal{A}$ then $-\mathbf{a} \in \mathcal{A}$ for all $\mathbf{a} \in \mathcal{A}$. Then, the **atomic norm** associated with a symmetric atomic set $\mathcal{A}$ is given by

$$\|\mathbf{x}\|_{\mathcal{A}} := g_{\mathrm{conv}(\mathcal{A})}(\mathbf{x}), \quad \forall \mathbf{x} \in \mathbb{R}^p,$$

where $\mathrm{conv}(\mathcal{A})$ denotes the *convex hull* of $\mathcal{A}$.

A generalization of the Lasso

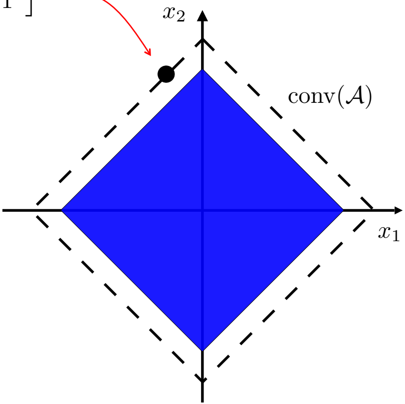
Given an atomic set $\mathcal{A}$, solve the following regularized least-squares problem:

$$\mathbf{x}^\star = \arg\min_{\mathbf{x}\in\mathbb{R}^p} \|\,\mathbf{b} - \mathbf{A}\mathbf{x}\,\|_2^2 + \rho\|\,\mathbf{x}\,\|_{\mathcal{A}} \tag{4}$$

*Pop quiz

Let $\mathcal{A} := \left\{(1,0)^T, (0,1)^T, (-1,0)^T, (0,-1)^T\right\}$, and let $\mathbf{x} := (-\frac{1}{5}, 1)^T$. What is $\|\mathbf{x}\|_{\mathcal{A}}$?

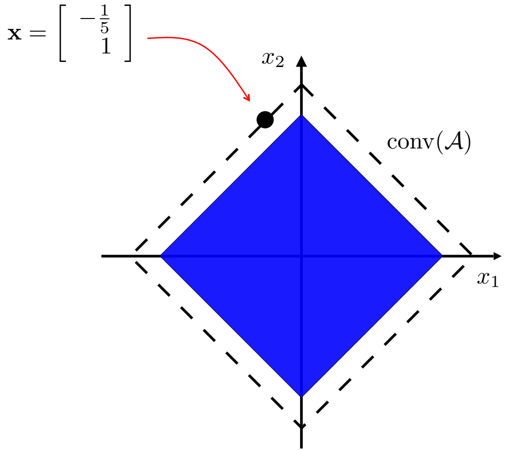
$$\mathbf{x} = \begin{bmatrix} -\frac{1}{5} \\ 1 \end{bmatrix}$$



*Pop quiz

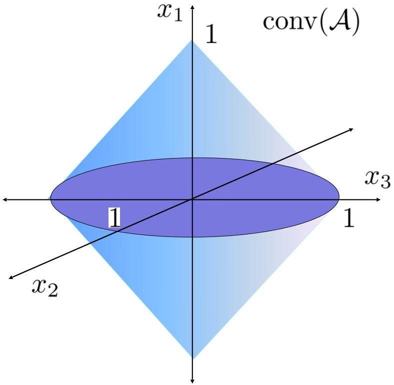
Let $\mathcal{A} := \left\{(1,0)^T, (0,1)^T, (-1,0)^T, (0,-1)^T\right\}$, and let $\mathbf{x} := (-\frac{1}{5}, 1)^T$. What is $\|\,\mathbf{x}\,\|_{\mathcal{A}}$?

ANS: $\|\,\mathbf{x}\,\|_{\mathcal{A}} = \frac{6}{5}$.



*Pop quiz 2

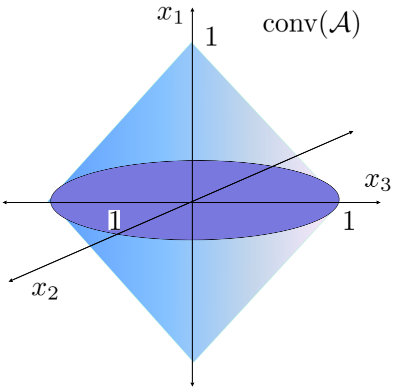
What is the expression of $\|\mathbf{x}\|_{\mathcal{A}}$ for any $\mathbf{x} := (x_1, x_2, x_3)^T \in \mathbb{R}^3$?



*Pop quiz 2

What is the expression of $\| \mathbf{x} \|_{\mathcal{A}}$ for any $\mathbf{x} := (x_1, x_2, x_3)^T \in \mathbb{R}^3$?

ANS: $\| \mathbf{x} \|_{\mathcal{A}} = |x_1| + \| (x_2, x_3)^T \|_2$.



*Application: Multi-knapsack feasibility problem

Problem formulation [19]

Let $\mathbf{x}^\natural \in \mathbb{R}^p$ which is a convex combination of k vectors in $\mathcal{A} := \{-1,+1\}^p$, and let $\mathbf{A} \in \mathbb{R}^{n\times p}$. How can we recover $\mathbf{x}^\natural$ given $\mathbf{A}$ and $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural$?

The answer:

- We can use the *ℓ_∞-norm*, $\|\cdot\|_\infty$ as $\|\cdot\|_{\mathcal{A}}$. The regularized estimator is given by

$$\mathbf{x}^\star \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 + \rho\|\mathbf{x}\|_\infty, \rho > 0.$$

*Application: Multi-knapsack feasibility problem

Problem formulation [19]

Let $\mathbf{x}^\natural \in \mathbb{R}^p$ which is a convex combination of k vectors in $\mathcal{A} := \{-1, +1\}^p$, and let $\mathbf{A} \in \mathbb{R}^{n \times p}$. How can we recover $\mathbf{x}^\natural$ given $\mathbf{A}$ and $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural$?

The answer: ∘ We can use the *ℓ_∞-norm*, $\|\cdot\|_\infty$ as $\|\cdot\|_{\mathcal{A}}$. The regularized estimator is given by

$$\mathbf{x}^\star \in \arg\min_{\mathbf{x} \in \mathbb{R}^p} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 + \rho\|\mathbf{x}\|_\infty, \rho > 0.$$

The derivation: ∘ In this case, we have $\text{conv}(\mathcal{A}) = [-1, 1]^p$ and

$$g_{\text{conv}(\mathcal{A})}(\mathbf{x}) = \inf\left\{t > 0 : \mathbf{x} = t\mathbf{c} \text{ for some } \mathbf{c} \text{ such that } |c_i| \leq 1\ \forall i\right\}.$$

∘ We also have, $\forall \mathbf{x} \in \mathbb{R}^p, \mathbf{c} \in \text{conv}(\mathcal{A}), t > 0$,

$$\begin{aligned} \mathbf{x} = t\mathbf{c} &\Rightarrow \forall i, |x_i| = |tc_i| \leq t \\ &\Rightarrow g_{\text{conv}(\mathcal{A})}(\mathbf{x}) \geq \max_i |x_i|. \end{aligned}$$

∘ Let $\mathbf{x} \neq 0$, let $j \in \arg\max_i |x_i|$ and choose $t = \max_i |x_i|$, $c_i = x_i/t \in [-1, 1]^p$.

∘ Then, $\mathbf{x} = t\mathbf{c}$, and so $g_{\text{conv}(\mathcal{A})}(\mathbf{x}) \leq \max_i |x_i|$.

*Application: Matrix completion

Problem formulation [5, 12]

Let $\mathbf{X}^\natural \in \mathbb{R}^{p\times p}$ with $\mathrm{rank}(\mathbf{X}^\natural) = r$, and let $\mathbf{A}_1, \ldots, \mathbf{A}_n$ be matrices in $\mathbb{R}^{p\times p}$. How do we estimate $\mathbf{X}^\natural$ given $\mathbf{A}_1, \ldots, \mathbf{A}_n$ and $b_i = \mathrm{Tr}\left(\mathbf{A}_i\mathbf{X}^\natural\right) + w_i$, $i = 1, \ldots, n$, where $\mathbf{w} := (w_1, \ldots, w_n)^T$ denotes unknown noise?

The answer: ○ We can use the *nuclear norm*, $\|\cdot\|_*$ as $\|\cdot\|_{\mathcal{A}}$. The regularized estimator is given by

$$\mathbf{x}^\star \in \arg\min_{\mathbf{X}\in\mathbb{R}^{p\times p}} \sum_{i=1}^{n} (b_i - \mathrm{Tr}\,(\mathbf{A}_i\mathbf{X}))^2 + \rho\|\,\mathbf{X}\,\|_*, \rho > 0.$$

*Application: Matrix completion

Problem formulation [5, 12]

Let $\mathbf{X}^\natural \in \mathbb{R}^{p\times p}$ with $\mathrm{rank}(\mathbf{X}^\natural) = r$, and let $\mathbf{A}_1, \ldots, \mathbf{A}_n$ be matrices in $\mathbb{R}^{p\times p}$. How do we estimate $\mathbf{X}^\natural$ given $\mathbf{A}_1, \ldots, \mathbf{A}_n$ and $b_i = \mathrm{Tr}\left(\mathbf{A}_i\mathbf{X}^\natural\right) + w_i$, $i = 1, \ldots, n$, where $\mathbf{w} := (w_1, \ldots, w_n)^T$ denotes unknown noise?

The answer: ◦ We can use the *nuclear norm*, $\|\cdot\|_*$ as $\|\cdot\|_{\mathcal{A}}$. The regularized estimator is given by

$$\mathbf{x}^\star \in \arg\min_{\mathbf{X}\in\mathbb{R}^{p\times p}} \sum_{i=1}^{n} (b_i - \mathrm{Tr}\,(\mathbf{A}_i\mathbf{X}))^2 + \rho\|\,\mathbf{X}\,\|_*, \rho > 0.$$

The derivation: ◦ Let us use the following atomic set $\mathcal{A} = \left\{\mathbf{X} : \mathrm{rank}\ (\mathbf{X}) = 1, \|\,\mathbf{X}\,\|_F = 1, \mathbf{X} \in \mathbb{R}^{p\times p}\right\}$.

◦ Let $\forall\mathbf{X} \in \mathbb{R}^{p\times p}, \mathbf{C} = \sum_i \lambda_i\mathbf{C}_i \in \mathrm{conv}(\mathcal{A}), \sum_i \lambda_i = 1, \mathbf{C}_i \in \mathcal{A}, t > 0$. Then, we have

$$\mathbf{X} = t\sum_i \lambda_i\mathbf{C}_i \Rightarrow \|\mathbf{X}\|_* = t\left\|\sum_i \lambda_i\mathbf{C}_i\right\|_* \leq t\sum_i \lambda_i\,\|\mathbf{C}_i\|_* \leq t \Rightarrow g_{\mathrm{conv}(\mathcal{A})}(\mathbf{X}) \geq \|\mathbf{X}\|_*\,.$$

◦ Let $\mathbf{X} \neq 0$, let $\mathbf{X} = \sum_i \sigma_i\mathbf{u}_i\mathbf{v}_i^t$ be its SVD decomposition, where σ_i's are its singular values.

◦ Let $t = \|\mathbf{X}\|_* = \sum_i |\sigma_i|$, $\mathbf{C}_i = \mathbf{u}_i\mathbf{v}_i^T \in \mathcal{A}$, $\forall i$. Then, $\mathbf{X} = t\sum_i \lambda_i\mathbf{C}_i$, $\lambda_i = \frac{|\sigma_i|}{t}$.

◦ Since t is feasible and $\sum_i \lambda_i = 1$, it follows that $g_{\mathrm{conv}(\mathcal{A})}(\mathbf{X}) \leq \|\mathbf{X}\|_*$.

*Structured Sparsity

There exist many more structures that we have not covered here, each of which is handled using different non-smooth regularizers. Some examples [3, 10]:

- **Group Sparsity:** Many signals are not only sparse, but the non-zero entries tend to cluster according to known patterns.
- **Tree Sparsity:** When natural images are transformed to the Wavelet domain, their significant entries form a *rooted connected tree*.

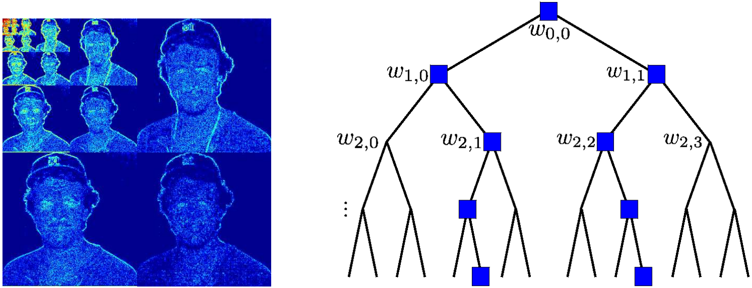


Figure: (**Left panel**) Natural image in the Wavelet domain. (**Right panel**) Rooted connected tree containing the significant coefficients.

*Selection of the Parameters

In all of these problems, there remain the issues of *how to design* $\mathbf{A}$ and *how to choose* ρ.

Design of $\mathbf{A}$:

- Sometimes $\mathbf{A}$ is given "by nature", whereas sometimes it can be designed
- For the latter case, i.i.d. Gaussian designs provide good theoretical guarantees, whereas in practice we must resort to structured matrices permitting more efficient storage and computation
- See [13] for an extensive study in the context of compressive sensing

Selection of ρ:

- Theoretical bounds provide some insight, but usually the direct use of the theoretical choice does not suffice
- In practice, a common approach is *cross-validation* [9], which involves searching for a parameter that performs well on a set of known training signals
- Other approaches include *covariance penalty* [9] and *upper bound heuristic* [29]

References I

[1] Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth.
Lower bounds for non-convex stochastic optimization.
arXiv preprint arXiv:1912.02365, 2019.
(Cited on page 27.)

[2] Francis Bach and Eric Moulines.
Non-strongly-convex smooth stochastic approximation with convergence rate o (1/n).
Advances in neural information processing systems, 26, 2013.
(Cited on page 24.)

[3] R.G. Baraniuk, V. Cevher, M.F. Duarte, and C. Hegde.
Model-based compressive sensing.
Information Theory, IEEE Transactions on, 56(4):1982–2001, 2010.
(Cited on page 75.)

[4] Léon Bottou, Frank E Curtis, and Jorge Nocedal.
Optimization methods for large-scale machine learning.
Siam Review, 60(2):223–311, 2018.
(Cited on pages 14 and 15.)

References II

[5] Emmanuel Candès and Benjamin Recht.
Exact matrix completion via convex optimization.
Found. Comp. Math., 9:717–772, 2009.
(Cited on pages 73 and 74.)

[6] Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford.
Lower bounds for finding stationary points II: first-order methods.
Math. Program., 185(1-2):315–355, 2021.
(Cited on page 27.)

[7] Volkan Cevher and Bang Cong Vu.
On the linear convergence of the stochastic gradient method with constant step-size.
arXiv:1712.01906 [math], June 2018.
(Cited on page 14.)

[8] Venkat Chandrasekaran, Benjamin Recht, Pablo A. Parrilo, and Alan S. Willsky.
The convex geometry of linear inverse problems.
Found. Comp. Math., 12:805–849, 2012.
(Cited on page 64.)

References III

[9] Bradley Efron.
The estimation of prediction error: Covariance penalities and cross-validation.
J. Amer. Math. Soc., 99(467):619–632, September 2004.
(Cited on page 76.)

[10] Marwa El Halabi and Volkan Cevher.
A totally unimodular view of structured sparsity.
preprint, 2014.
arXiv:1411.1990v1 [cs.LG].
(Cited on page 75.)

[11] Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang.
SPIDER: near-optimal non-convex optimization via stochastic path-integrated differential estimator.
In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 687–697, 2018.
(Cited on page 27.)

References IV

[12] Steven T. Flammia, David Gross, Yi-Kai Liu, and Jens Eisert.
Quantum tomography via compressed sensing: Error bounds, sample complexity and efficient estimators.
New J. Phys., 14, 2012.
(Cited on pages 73 and 74.)

[13] Simon Foucart and Holger Rauhut.
A mathematical introduction to compressive sensing, volume 1.
Birkhäuser Basel, 2013.
(Cited on page 76.)

[14] Saeed Ghadimi and Guanghui Lan.
Stochastic first-and zeroth-order methods for nonconvex stochastic programming.
SIAM Journal on Optimization, 23(4):2341–2368, 2013.
(Cited on page 26.)

[15] Saeed Ghadimi and Guanghui Lan.
Accelerated gradient methods for nonconvex nonlinear and stochastic programming.
Math. Program., 156(1–2):59–99, March 2016.
(Cited on pages 4 and 5.)

References V

[16] Rémi Gribonval, Volkan Cevher, and Mike E. Davies.
Compressible distributions for high-dimensional statistics.
IEEE Trans. Inf. Theory, 58(8):5016–5034, 2012.
(Cited on page 34.)

[17] Kfir Levy.
Online to offline conversions, universality and adaptive minibatch sizes.
In *Advances in Neural Information Processing Systems*, pages 1613–1622, 2017.
(Cited on page 58.)

[18] Kfir Levy, Alp Yurtsever, and Volkan Cevher.
Online adaptive methods, universality and acceleration.
In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018.
(Cited on page 59.)

[19] O. L. Mangasarian and Benjamin Recht.
Probability of unique integer solution to a system of linear equations.
Eur. J. Oper. Res., 214:27–30, 2011.
(Cited on pages 71 and 72.)

References VI

[20] Panayotis Mertikopoulos, Ya-Ping Hsieh, and Volkan Cevher.
Learning in games from a stochastic approximation viewpoint.
arXiv preprint arXiv:2206.03922, 2022.
(Cited on page 26.)

[21] Deanna Needell, Rachel Ward, and Nati Srebro.
Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm.
Advances in neural information processing systems, 27, 2014.
(Cited on page 62.)

[22] Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu.
A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers.
Stat. Sci., 27(4):538–557, 2012.
(Cited on page 48.)

[23] Arkadi Semen Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro.
Robust stochastic approximation approach to stochastic programming.
SIAM Journal on Optimization, 19(4):1574–1609, 2009.
(Cited on page 20.)

References VII

[24] Boris T. Polyak.
Introduction to Optimization.
Optimization Softw., Inc., New York, 1987.
(Cited on page 14.)

[25] R. Tyrrell Rockafellar.
Convex Analysis.
Princeton Univ. Press, Princeton, NJ, 1970.
(Cited on page 52.)

[26] Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter.
Pegasos: Primal estimated sub-gradient solver for svm.
Mathematical programming, 127(1):3–30, 2011.
(Cited on page 21.)

[27] Ohad Shamir and Tong Zhang.
Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes.
In *ICML '13: Proceedings of the 30th International Conference on Machine Learning*, 2013.
(Cited on pages 13 and 55.)

References VIII

[28] Thomas Strohmer and Roman Vershynin.
Comments on the randomized kaczmarz method.
J. Fourier Anal. and Apps., 15(4):437–440, 2009.
(Cited on page 62.)

[29] Christos Thrampoulidis, Samet Oymak, and Babak Hassibi.
Simple error bounds for regularized noisy linear inverse problems.
2014.
arXiv:1401.6578v1 [math.OC].
(Cited on page 76.)