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

- This work is released under a Creative Commons License with the following terms:

  - **Attribution**
    - The licensor permits others to copy, distribute, display, and perform the work. In return, licensees must give the original authors credit.

  - **Non-Commercial**
    - The licensor permits others to copy, distribute, display, and perform the work. In return, licensees may not use the work for commercial purposes – unless they get the licensor’s permission.

  - **Share Alike**
    - The licensor permits others to distribute derivative works only under a license identical to the one that governs the licensor’s work.

- **Full Text of the License**
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:

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

$f(x)$ is proper and closed.

Gradient descent

Choose a starting point $x^0$ and iterate

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

where $\alpha_k$ is a step-size to be chosen so that $x^k$ converges to $x^*$.

<table>
<thead>
<tr>
<th></th>
<th>$f$ is $L$-smooth &amp; convex</th>
<th>$f$ is $L$-gradient Lipschitz &amp; non-convex</th>
</tr>
</thead>
<tbody>
<tr>
<td>GD</td>
<td>$O(1/k)$ (fast)</td>
<td>$O(1/k)$ (optimal)</td>
</tr>
<tr>
<td>AGD</td>
<td>$O(1/k^2)$ (optimal)</td>
<td>$O(1/k)$ (optimal) [15]</td>
</tr>
</tbody>
</table>
Recall: Gradient descent

Problem (Unconstrained optimization problem)

Consider the following minimization problem:

\[ f^* = \min_{x \in \mathbb{R}^p} f(x) \]

\( f(x) \) is proper and closed.

Gradient descent

Choose a starting point \( x^0 \) and iterate

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

where \( \alpha_k \) is a step-size to be chosen so that \( x^k \) converges to \( x^* \).

<table>
<thead>
<tr>
<th></th>
<th>( f ) is ( L )-smooth &amp; convex</th>
<th>( f ) is ( L )-gradient Lipschitz &amp; non-convex</th>
</tr>
</thead>
<tbody>
<tr>
<td>GD</td>
<td>( O(1/k) ) (fast)</td>
<td>( O(1/k) ) (optimal)</td>
</tr>
<tr>
<td>AGD</td>
<td>( O(1/k^2) ) (optimal)</td>
<td>( O(1/k) ) (optimal) [15]</td>
</tr>
</tbody>
</table>

Why should we study anything else?
Statistical learning with streaming data

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

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

Abstract gradient method

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

Remark:
- This algorithm can not be implemented as the distribution of \((a, b)\) is unknown.
Statistical learning with streaming data

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

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

**Abstract gradient method**

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

**Remark:**

○ This algorithm can not be implemented as the distribution of $(a, b)$ is unknown.

○ In practice, data can arrive in a streaming way.

**A parametric example: Markowitz portfolio optimization**

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

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

Problem (Mathematical formulation)

Consider the following convex minimization problem:

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

- \(\theta\) is a random vector whose probability distribution is supported on set \(\Theta\).
- \(f(\mathbf{x}) := \mathbb{E}[f(\mathbf{x}, \theta)]\) is proper, closed, and convex.
- The solution set \(S^* := \{ \mathbf{x}^* \in \text{dom}(f) : f(\mathbf{x}^*) = f^* \}\) is nonempty.
Stochastic gradient descent (SGD)

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

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

- $G(x^k, \theta_k)$ is an unbiased estimate of the full gradient:

\[ \mathbb{E}[G(x^k, \theta_k)] = \nabla f(x^k). \]
Stochastic gradient descent (SGD)

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

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

- $G(x^k, \theta_k)$ is an unbiased estimate of the full gradient:

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

Remarks:
- The cost of computing $G(x^k, \theta_k)$ is $n$ times cheaper than that of $\nabla f(x^k)$.
- As $G(x^k, \theta_k)$ 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

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

can be rewritten as

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

A stochastic gradient descent (SGD) variant for finite sums

\[
x^{k+1} = x^k - \alpha_k \nabla f_i(x^k) \quad i \text{ is uniformly distributed over } \{1, \ldots, n\}
\]

Remarks:

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

\[
\min_x \left\{ f(x) := \frac{1}{2n} \|Ax - b\|_2^2 : x \in \mathbb{R}^p \right\}
\]

Setup

- **A** := \text{randn}(n, p) - standard Gaussian \( \mathcal{N}(0, I) \), with \( n = 10^4 \), \( p = 10^2 \).
- **x**\(^\dagger\) is 50 sparse with zero mean Gaussian i.i.d. entries, normalized to \( \|x\|^2 = 1 \).
- **b** := **A**\(x\)^\dagger\ + **w**, where **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

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

Then

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

**Observation:** $O(1/\sqrt{k})$ 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

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

Then

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

**Observations:**

- Converge fast (linearly) to a neighborhood around \( x^* \).
- Smaller step-sizes \( \alpha \rightarrow \) converge to a better point, but with a slower rate.
- Zero variance \( (\sigma = 0) \rightarrow \) 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

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

Then

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

where $C$ is a constant independent of $k$.

Observations:

\begin{itemize}
  \item Using the $L$-smooth property,
    $$
    \mathbb{E}[f(x^k) - f(x^*)] \leq L\mathbb{E}[\|x^k - x^*\|^2] \leq \frac{C}{k + 1}.
    $$
  \item The rate is optimal if $\sigma^2 > 0$ with the assumption of strongly-convexity.
\end{itemize}
Example: SGD with different step sizes

![Graph](image)

---

**Setup**

- Synthetic least-squares problem as before.
- We use $\alpha_k = \alpha_0 / (k + k_0)$. 
Example: SGD with different step sizes

\[ \alpha_0 = \frac{1}{(3\mu)} \]
\[ \alpha_0 = \frac{1}{(2\mu)} \]
\[ \alpha_0 = \frac{1}{\mu} \]
\[ \alpha_0 = \frac{2}{\mu} \]

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

\[ f^* := \min_{x \in \mathbb{R}^p} \left\{ f(x) := \frac{1}{n} \sum_{j=1}^{n} f_j(x) \right\}. \]

- \( f \): \( \mu \)-strongly convex with \( L \)-Lipschitz smooth.

<table>
<thead>
<tr>
<th></th>
<th>rate</th>
<th>iteration complexity</th>
<th>cost per iteration</th>
<th>total cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>GD</td>
<td>( \rho^k )</td>
<td>( \log(1/\epsilon) )</td>
<td>( n )</td>
<td>( n \log(1/\epsilon) )</td>
</tr>
<tr>
<td>SGD</td>
<td>( 1/k )</td>
<td>( 1/\epsilon )</td>
<td>( 1 )</td>
<td>( 1/\epsilon )</td>
</tr>
</tbody>
</table>

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:
  \[ \bar{x}^k = \frac{1}{k} \sum_{j=1}^{k} \alpha_j x^j \] (vanilla averaging)
  \[ \bar{x}^k = \frac{\sum_{j=1}^{k} \alpha_j x^j}{\sum_{j=1}^{k} \alpha_j} \] (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 $x^0 \in \mathbb{R}^p$ and $(\alpha_k)_{k \in \mathbb{N}} \in ]0, +\infty[^\mathbb{N}$.

2a. For $k = 0, 1, \ldots$ perform:

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

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

Theorem (Convergence of SGD-A [23])

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

Then,

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

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

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

Observation: ∙ Same convergence rate with vanilla SGD.
Convergence for SGD-A II: strongly convex case

<table>
<thead>
<tr>
<th>Stochastic gradient method with averaging (SGD-A)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Choose $x^0 \in \mathbb{R}^p$ and $(\alpha_k)_{k \in \mathbb{N}} \in ]0, +\infty[\mathbb{N}$.</td>
</tr>
<tr>
<td>2a. For $k = 0, 1, \ldots$ perform: $x^{k+1} = x^k - \alpha_k G(x^k, \theta_k)$.</td>
</tr>
<tr>
<td>2b. $\bar{x}^k = \frac{1}{k} \sum_{j=1}^{k} x^j$.</td>
</tr>
</tbody>
</table>

Theorem (Convergence of SGD-A [26])

Assume

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

Then

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

Observation: Same convergence rate with vanilla SGD.
Example: SGD-A method with different step sizes

\[
\min_x \left\{ f(x) := \frac{1}{2n} \| Ax - b \|_2^2 : x \in \mathbb{R}^p \right\}
\]

Setup

- Synthetic least-squares problem as before
- \( \alpha_k = \alpha_0 / (k + k_0) \).
Example: SGD-A method with different step sizes

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

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

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

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

### Stochastic gradient method with averaging

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

\[ x^k = x^{k-1} - \alpha \left( \langle a_k, x^{k-1} \rangle - b_k \right) a_k. \]

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

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

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

\[ \mathbb{E}f(\bar{x}^{k-1}) - f^* \leq \frac{2}{k} \left( \sigma \sqrt{p} + R \|x^0 - x^*\|_2 \right)^2. \]
Popular SGD Variants

- Mini-batch SGD: For each iteration,
  \[ x^{k+1} = x^k - \alpha_k \frac{1}{b} \sum_{\theta \in \Gamma} G(x^k, \theta). \]
  - \( \alpha_k \): step-size
  - \( b \): mini-batch size
  - \( \Gamma \): a set of random variables \( \theta \) 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:

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

**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 \( \sigma^2 \) is the variance of the gradients and \( C > 0 \) is constant. Then, it holds that

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

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

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

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

---

1We 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:

\[ b = Ax^\dagger + w \]

with \( b \in \mathbb{R}^n \), \( A \in \mathbb{R}^{n \times p} \) are known, \( x^\dagger \) is unknown, and \( w \) 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:

\[ b = Ax^\dagger + w \]

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

- **Standard approach:** Least squares: \( x^*_{LS} \in \arg\min_x \|b - Ax\|_2^2 \)
  - Convex, smooth, and an explicit solution: \( x^*_{LS} = (A^T A)^{-1} A^T b = A^\dagger b \)

- **Alternative approach:** Least absolute value deviation: \( x^* \in \arg\min_x \|b - Ax\|_1 \)
  - 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 $x^\star$.

**Practical performance**

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

$$
\|x^t - x^\dagger\|_2 \leq \|x^t - x^\star\|_2 \quad \text{numerical error} + \|x^\star - x^\dagger\|_2 \quad \text{statistical error}.
$$

**Remarks:**

- *Non-smooth* estimators of $x^\dagger$ 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 $x^b \in \mathbb{R}^p$ and $A \in \mathbb{R}^{n \times p}$. The samples are given by $b = Ax^b + w$, where $w$ denotes the unknown noise. The LS estimator for $x^b$ given $A$ and $b$ is defined as

$$x_{LS}^* = \arg \min_{x \in \mathbb{R}^p} \left\{ \| b - Ax \|_2^2 \right\}.$$

**Remarks:**

- If $A$ has full column rank, $x_{LS}^* = A^\dagger b$ is uniquely defined.
- *When* $n < p$, $A$ cannot have full column rank, and hence $x_{LS}^* \in \left\{ A^\dagger b + h : h \in \text{null} (A) \right\}$.

**Observation:**

- The estimation error $\| x_{LS}^* - x^b \|_2$ can be *arbitrarily large*!
A candidate solution

Continuing the LS example:

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

Observation: Unfortunately, \text{this still fails when } n < p
**Proposition ([16])**

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

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

with probability at least $1 - 2 \exp \left[-\frac{1}{4}(p - n)\epsilon^2\right] - 2 \exp \left[-\frac{1}{4}p\epsilon^2\right]$, for all $\epsilon > 0$ and $x^\# \in \mathbb{R}^p$. 
Summarizing the findings so far

The message so far:
- Even in the absence of noise, we cannot recover $x^\dagger$ from the observations $b = Ax^\dagger$ 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 $x$?
A natural signal model

**Definition (s-sparse vector)**

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

**Sparse representations**

$x^\Psi$: *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!

\[ b \tilde{A} = y \]

\[ b \in \mathbb{R}^n, \tilde{A} \in \mathbb{R}^{n \times p}, \text{ and } n < p \]
Sparse representations strike back!

\[ \begin{bmatrix} b \\ \tilde{A} \end{bmatrix} = \Psi \begin{bmatrix} x^\dagger \end{bmatrix} \]

- \( b \in \mathbb{R}^n, \tilde{A} \in \mathbb{R}^{n \times p}, \) and \( n < p \)
- \( \Psi \in \mathbb{R}^{p \times p}, \) \( x^\dagger \in \mathbb{R}^p, \) and \( \|x^\dagger\|_0 \leq s < n \)
Sparse representations strike back!

\[ \mathbf{b} \in \mathbb{R}^n, \mathbf{A} \in \mathbb{R}^{n \times p}, \text{ and } \mathbf{x}^\dagger \in \mathbb{R}^p, \text{ and } \|\mathbf{x}^\dagger\|_0 \leq s < n < p \]
Sparse representations strike back!

\[ \begin{bmatrix} \mathbf{b} \\ \mathbf{A} \end{bmatrix} = \begin{bmatrix} \mathbf{x}^\dagger \end{bmatrix} \]

\( n \times 1 \quad n \times s \quad s \times 1 \)

Observations:
- The matrix \( \mathbf{A} \) effectively becomes overcomplete.
- We could solve for \( \mathbf{x}^\dagger \) if we knew the location of the non-zero entries of \( \mathbf{x}^\dagger \).
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 \( \mathbf{b} = \mathbf{A} \mathbf{x} + \mathbf{w} \)

**A realistic linear model**

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

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

Then we have

\[
\mathbf{b} = \mathbf{\tilde{A}} \Psi \left( \mathbf{x} + \mathbf{x}_{\text{real}} - \mathbf{x} \right) + \mathbf{\tilde{w}}.
\]

\[
:= \left( \mathbf{\tilde{A}} \Psi \right) \mathbf{x} + \left[ \mathbf{\tilde{w}} + \mathbf{\tilde{A}} \Psi \left( \mathbf{x}_{\text{real}} - \mathbf{x} \right) \right],
\]

equivalently, \( \mathbf{b} = \mathbf{A} \mathbf{x} + \mathbf{w} \).
Peeling the onion

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

### Practical performance

The practical performance at time $t$ is determined by

$$\| x^t - x_{\text{real}} \|_2 \leq \| x^t - x^* \|_2 + \| x^* - x^\natural \|_2 + \| x_{\text{real}} - x^\natural \|_2.$$
Approach 1: Sparse recovery via exhaustive search

Approach 1 for estimating $x^\sharp$ from $b = Ax^\sharp + w$

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

We may search over all $\binom{p}{s}$ subsets $S \subset \{1, \ldots, p\}$ of cardinality $s$, solve the restricted least-squares problem

$$\min_{\mathbf{x}_S} \| \mathbf{b} - A_S \mathbf{x}_S \|_2^2,$$

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 $\ell_1$-norm heuristic

**Heuristic:** The $\ell_1$-ball with radius $c_\infty$ is an “approximation” of the set of sparse vectors $\hat{x} \in \{x : \|x\|_0 \leq s, \|x\|_\infty \leq c_\infty\}$ parameterized by their sparsity $s$ and maximum amplitude $c_\infty$.

$$\hat{x} \in \{x : \|x\|_1 \leq c_\infty\} \text{ with some } c_\infty > 0.$$
Sparse recovery via the Lasso

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

\[ x_{Lasso}^* := \arg \min_{x \in \mathbb{R}^p} \| b - Ax \|_2^2 + \rho \| x \|_1 \]

with some \( \rho \geq 0 \).

- The second term in the objective function is called the *regularizer*.

- The parameter \( \rho \) is called the *regularization parameter*. It is used to trade off the objectives:
  - Minimize \( \| b - Ax \|_2^2 \), so that the solution is consistent with the observations
  - Minimize \( \| x \|_1 \), 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 \( x^\natural \) is \( s \)-sparse and the noise \( w \) is sub-Gaussian (e.g., Gaussian or bounded) with parameter \( \sigma \), then choosing \( \rho = \sqrt{\frac{16\sigma^2 \log p}{n}} \) yields an error of

\[
\| x_\text{Lasso}^* - x^\natural \|_2 \leq \frac{8\sigma}{\kappa(A)} \sqrt{\frac{s \ln p}{n}},
\]

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

Remark:  
- The number of measurements is \( \mathcal{O}(s \ln p) \) – 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?

\[ F^* := \min_{x \in \mathbb{R}^p} f(x) \]  (1)

where \( f \) is proper, closed, convex, but not everywhere differentiable.
Subdifferentials: A generalization of the gradient

**Definition**

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

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

Each element \( v \) of \( \partial f(x) \) is called subgradient of \( f \) at \( x \).

**Lemma**

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

---

**Figure:** (Left) Non-differentiability at point \( y \). (Right) Gradient as a subdifferential with a singleton entry.
(Sub)gradients in convex functions

Example

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

Figure: Subgradients of \( f(x) = |x| \) in \( \mathbb{R} \).
Subdifferentials: Two basic results

Lemma (Necessary and sufficient condition)

\( x^* \in \text{dom}(F) \) is a **globally optimal** solution to (1) iff \( 0 \in \partial F(x^*) \).

**Sketch of the proof.**

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

\[
F(x) - F(x^*) \geq 0^T(x - x^*) = 0,
\]

that is, \( x^* \) is a global solution to (1).

\( \Rightarrow \): If \( x^* \) is a global of (1) then for every \( x \in \text{dom}(F) \), \( F(x) \geq F(x^*) \) and hence

\[
F(x) - F(x^*) \geq 0^T(x - x^*), \forall x \in \mathbb{R}^p,
\]

which leads to \( 0 \in \partial F(x^*) \). \( \square \)

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

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

\[
\partial(f + g) = \partial f + \partial g.
\]
Non-smooth unconstrained convex minimization

Problem (Non-smooth convex minimization)

\[ F^* := \min_{x \in \mathbb{R}^p} f(x) \quad (2) \]

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.

<table>
<thead>
<tr>
<th>Subgradient method</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Choose ( x^0 \in \mathbb{R}^p ) as a starting point.</td>
</tr>
<tr>
<td>2. For ( k = 0, 1, \cdots ), perform:</td>
</tr>
</tbody>
</table>
| \[
\begin{align*}
    x^{k+1} &= x^k - \alpha_k d^k, \\
\end{align*}
\]
| where \( d^k \in \partial f(x^k) \) 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. \( \| g \|_2 \leq G \) for all \( g \in \partial f(x) \) for any \( x \in \mathbb{R}^p \).
2. \( \| x^0 - x^* \|_2 \leq R \)

Let the stepsize be chosen as

\[
\alpha_k = \frac{R}{G \sqrt{k}}
\]

then the iterates generated by the subgradient method satisfy

\[
\min_{0 \leq i \leq k} f(x^i) - f^* \leq \frac{RG}{\sqrt{k}}.
\]

**Remarks**

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

○ An unbiased stochastic subgradient

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

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

<table>
<thead>
<tr>
<th>The classic stochastic subgradient methods (SG)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Choose ( x_1 \in \mathbb{R}^p ) and ( (\gamma_k)_{k \in \mathbb{N}} \in (0, +\infty)^\mathbb{N} ).</td>
</tr>
<tr>
<td>2. For ( k = 1, \ldots ) perform:</td>
</tr>
<tr>
<td>( x_{k+1} = x_k - \gamma_k G(x_k). )</td>
</tr>
</tbody>
</table>

Theorem (Convergence in expectation [27])

Suppose that:

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

Then,

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

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):
  - One on source separation (Math of Data Lecture 6 from 2014)
  - One on convexification of structured sparsity models (research presentation)
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: \( \nabla f(x) \Rightarrow G(x, \theta) \)

<table>
<thead>
<tr>
<th>AdaGrad with ( H_k = \lambda_k I ) [17]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Set ( Q^0 = 0 ).</td>
</tr>
<tr>
<td>2. For ( k = 0, 1, \ldots ), iterate</td>
</tr>
</tbody>
</table>
| \[
| Q^k = Q^{k-1} + \| G(x^k, \theta) \|^2 \\
| H_k = \sqrt{Q^k} I \\
<table>
<thead>
<tr>
<th>x^{k+1} = x_t - \alpha_k H_k^{-1} G(x^k, \theta)</th>
</tr>
</thead>
</table>

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 \( K \) with diameter \( D \). Also consider bounded variance for unbiased gradient estimates, i.e., \( \mathbb{E} \left[ \| G(x, \theta) - \nabla f(x) \|^2 | x \right] \leq \sigma^2 \). Then,

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

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

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

<table>
<thead>
<tr>
<th>AcceleGrad (Accelerated Adaptive Gradient Method)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input</strong> : $x^0 \in \mathcal{K}$, diameter $D$, weights ${\alpha_k}<em>{k \in \mathbb{N}}$, learning rate ${\eta_k}</em>{k \in \mathbb{N}}$</td>
</tr>
<tr>
<td><strong>1.</strong> Set $y^0 = z^0 = x^0$</td>
</tr>
<tr>
<td><strong>2.</strong> For $k = 0, 1, \ldots$, iterate</td>
</tr>
<tr>
<td>$\tau_k := 1/\alpha_k$</td>
</tr>
<tr>
<td>$x^{k+1} = \tau_tz^k + (1 - \tau_k)y^k$, define $g_k := \nabla f(x^{k+1})$</td>
</tr>
<tr>
<td>$z^{k+1} = \Pi\mathcal{K}(z^k - \alpha_k\eta_kg_k)$</td>
</tr>
<tr>
<td>$y^{k+1} = x^{k+1} - \eta_kg_k$</td>
</tr>
<tr>
<td><strong>Output</strong> : $\bar{y}^k \propto \sum_{i=0}^{k-1} \alpha_i y^{i+1}$</td>
</tr>
</tbody>
</table>

**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., $E[\|G(x, \theta) - \nabla f(x)\|^2 | x] \leq \sigma^2$. Then,

$$E[f(\bar{y}^k)] - \min_x f(x) = O\left(\frac{GD\sqrt{\log k}}{\sqrt{k}}\right).$$
**Example: Synthetic least squares**

- $A \in \mathbb{R}^{n \times d}$, where $n = 200$ and $d = 50$.
- Number of epochs: 20.
UniXGrad for stochastic optimization

**UniXGrad**

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

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

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

**Theorem (Convergence rate of UniXGrad)**

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

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

UniXGrad guarantees the following:

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

**Problem**

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

$$Ax = b.$$  

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

### Randomized Kaczmarz algorithm (RKA)

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

### Linear convergence [28]

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

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

- RKA can be seen as a particular case of SGD [21].
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
\[
\mathcal{A} := \{ \pm \mathbf{e}_1, \ldots, \pm \mathbf{e}_p \}, \quad \text{where } \mathbf{e}_i := (\delta_{1,i}, \delta_{2,i}, \ldots, \delta_{p,i}) \text{ for all } i.
\]

**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 \( \psi_k \) denotes the \( k \)th column of \( \Psi \).
**Atomic norms**

- Recall the Lasso problem
  \[ x^*_{\text{Lasso}} := \arg \min_{x \in \mathbb{R}^p} \| b - Ax \|_2^2 + \rho \| x \|_1 \]

**Observations:**
- $\ell_1$-norm is the *atomic norm* associated with the atomic set $A := \{ \pm e_1, \ldots, \pm e_p \}$.
- The norm is closely tied with the convex hull of the set.
- We can extend the same principle for a wide range of regularizers

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

\[ C := \text{conv}(A). \]
**Gauge functions and atomic norms**

**Definition (Gauge function)**

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

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

**Definition (Atomic norm)**

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

$$\|x\|_A := g_{\text{conv}(A)}(x), \quad \forall x \in \mathbb{R}^p,$$

where $\text{conv}(A)$ denotes the convex hull of $A$.

**A generalization of the Lasso**

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

$$x^* = \arg \min_{x \in \mathbb{R}^p} \| b - Ax \|_2^2 + \rho \| x \|_A$$

(4)
*Pop quiz

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

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

\[ x = \begin{bmatrix} -\frac{1}{5} \\ 1 \end{bmatrix} \]
*Pop quiz 2

What is the expression of $\| x \|_A$ for any $x := (x_1, x_2, x_3)^T \in \mathbb{R}^3$?
*Pop quiz 2

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

ANS: $\| x \|_A = |x_1| + \| (x_2, x_3)^T \|_2$. 

![Diagram of a convex set in 3D space with axes labeled $x_1$, $x_2$, and $x_3$, and the convex hull labeled $\text{conv}(A)$]
Application: Multi-knapsack feasibility problem

Problem formulation [19]

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

The answer: We can use the $\ell_\infty$-norm, $\| \cdot \|_\infty$ as $\| \cdot \|_A$. The regularized estimator is given by

$$x^* \in \arg \min_{x \in \mathbb{R}^p} \| b - Ax \|_2^2 + \rho \| x \|_\infty, \rho > 0.$$
Application: Multi-knapsack feasibility problem

Problem formulation [19]

Let \( \mathbf{x}^b \in \mathbb{R}^p \) 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 \( \mathbf{x}^b \) given \( \mathbf{A} \) and \( \mathbf{b} = \mathbf{A} \mathbf{x}^b \)?

The answer: ○ We can use the \( \ell_\infty \)-norm, \( \| \cdot \|_\infty \) as \( \| \cdot \|_\mathcal{A} \). The regularized estimator is given by

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

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

\[
\mathbf{x} = t\mathbf{c} \Rightarrow \forall i, |x_i| = |tc_i| \leq t \\
\Rightarrow g_{\text{conv}(\mathcal{A})}(\mathbf{x}) \geq \max_i |x_i|.
\]

○ 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 \( X^\dagger \in \mathbb{R}^{p \times p} \) with \( \text{rank}(X^\dagger) = r \), and let \( A_1, \ldots, A_n \) be matrices in \( \mathbb{R}^{p \times p} \). How do we estimate \( X^\dagger \) given \( A_1, \ldots, A_n \) and \( b_i = \text{Tr}(A_i X^\dagger) + w_i, i = 1, \ldots, n \), where \( w := (w_1, \ldots, w_n)^T \) denotes unknown noise?

The answer: We can use the nuclear norm, \( \| \cdot \|_* \) as \( \| \cdot \|_A \). The regularized estimator is given by

\[
x^* \in \arg \min_{X \in \mathbb{R}^{p \times p}} \sum_{i=1}^{n} (b_i - \text{Tr}(A_i X))^2 + \rho \| X \|_* , \rho > 0.
\]
Application: Matrix completion

Problem formulation [5, 12]

Let $X^\dagger \in \mathbb{R}^{p \times p}$ with $\text{rank}(X^\dagger) = r$, and let $A_1, \ldots, A_n$ be matrices in $\mathbb{R}^{p \times p}$. How do we estimate $X^\dagger$ given $A_1, \ldots, A_n$ and $b_i = \text{Tr}(A_i X^\dagger) + w_i$, $i = 1, \ldots, n$, where $w := (w_1, \ldots, w_n)^T$ denotes unknown noise?

The answer: We can use the nuclear norm, $\| \cdot \|_*$ as $\| \cdot \|_A$. The regularized estimator is given by

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

The derivation:

Let $X \in \mathbb{R}^{p \times p}$, $C = \sum_i \lambda_i C_i \in \text{conv}(A)$, $\sum_i \lambda_i = 1$, $C_i \in A$, $t > 0$. Then, we have

$$X = t \sum_i \lambda_i C_i \Rightarrow \| X \|_* = t \left\| \sum_i \lambda_i C_i \right\|_* \leq t \sum_i \lambda_i \| C_i \|_* \leq t \Rightarrow g_{\text{conv}(A)}(X) \geq \| X \|_*.$$ 

Let $X \neq 0$, let $X = \sum_i \sigma_i u_i v_i^T$ be its SVD decomposition, where $\sigma_i$'s are its singular values.

Let $t = \| X \|_* = \sum_i |\sigma_i|$, $C_i = u_i v_i^T \in A$, $\forall i$. Then, $X = t \sum_i \lambda_i C_i$, $\lambda_i = \tfrac{|\sigma_i|}{t}$.

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

Mathematics of Data | Prof. Volkan Cevher, volkan.cevher@epfl.ch
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 $A$ and how to choose $\rho$.

**Design of $A$:**
- Sometimes $A$ 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 $\rho$:**
- 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

(Cited on page 27.)

\textit{Advances in neural information processing systems}, 26, 2013. 
(Cited on page 24.)

(Cited on page 75.)

(Cited on pages 14 and 15.)
References II

Exact matrix completion via convex optimization.
(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.
(Cited on page 27.)

On the linear convergence of the stochastic gradient method with constant step-size.
(Cited on page 14.)

The convex geometry of linear inverse problems.
(Cited on page 64.)
The estimation of prediction error: Covariance penalties and cross-validation.  
(Cited on page 76.)

A totally unimodular view of structured sparsity.  
(Cited on page 75.)

SPIDER: near-optimal non-convex optimization via stochastic path-integrated differential estimator.  
(Cited on page 27.)
Quantum tomography via compressed sensing: Error bounds, sample complexity and efficient estimators.  
(Cited on pages 73 and 74.)

*A mathematical introduction to compressive sensing*, volume 1.  
(Cited on page 76.)

Stochastic first-and zeroth-order methods for nonconvex stochastic programming.  
(Cited on page 26.)

Accelerated gradient methods for nonconvex nonlinear and stochastic programming.  
(Cited on pages 4 and 5.)
[16] Rémi Gribonval, Volkan Cevher, and Mike E. Davies.
Compressible distributions for high-dimensional statistics.
(Cited on page 34.)

Online to offline conversions, universality and adaptive minibatch sizes.
(Cited on page 58.)

Online adaptive methods, universality and acceleration.
(Cited on page 59.)

Probability of unique integer solution to a system of linear equations.
(Cited on pages 71 and 72.)
Learning in games from a stochastic approximation viewpoint.
(Cited on page 26.)

Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm.
(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.
(Cited on page 48.)

Robust stochastic approximation approach to stochastic programming.
(Cited on page 20.)
References VII

*Introduction to Optimization.*
(Cited on page 14.)

*Convex Analysis.*
(Cited on page 52.)

[26] Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter.
Pegasos: Primal estimated sub-gradient solver for svm.
(Cited on page 21.)

[27] Ohad Shamir and Tong Zhang.
Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes.
(Cited on pages 13 and 55.)
References VIII
