Mathematics of Data: From Theory to Computation

Prof. Volkan Cevher
volkan.cevher@epfl.ch

Lecture 9: Composite convex minimization I

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

EE-556 (Fall 2018)

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

- Today
  1. Motivation for non-smooth optimization
  2. Composite convex minimization
  3. Proximal operator and computational complexity
  4. Proximal gradient methods
- Next week
  1. Proximal Newton-type methods
  2. Stochastic proximal methods

Recommended reading material

- A. Beck and M. Tebulle, A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems, SIAM J. Imaging Sciences, 2(1), 183–202, 2009.
- Y. Nesterov, Smooth minimization of non-smooth functions, Math. Program, 103(1), 127–152, 2005.
- Q. Tran-Dinh, A. Kyrillidis and V. Cevher, Composite Self-Concordant Minimization, LIONS-EPFL Tech. Report. `http://arxiv.org/abs/1308.2867`, 2013.
- N. Parikh and S. Boyd, Proximal Algorithms, Foundations and Trends in Optimization, 1(3):123-231, 2014.

Convex minimization: non-smooth and composite

Motivation

Data analytics problems in various disciplines can often be simplified to nonsmooth **composite convex minimization** problems. To this end, this lecture provides **efficient numerical solution methods** for such problems.

Intriguingly, composite minimization problems are far from generic nonsmooth problems and we can exploit individual function structures to obtain numerical solutions nearly as efficiently as if they are smooth problems.

Nonsmooth: When smooth models are deficient

Example: 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: $\hat{\mathbf{x}}_{\text{LS}} \in \arg\min_{\mathbf{x}} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2$

- Convex, smooth, and an *explicit solution*: $\hat{\mathbf{x}}_{\text{LS}} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b} = \mathbf{A}^\dagger\mathbf{b}$

Practical performance of an estimator $\hat{\mathbf{x}}$

Denote the numerical approximation by $\mathbf{x}_\epsilon^\star$. The practical performance is determined by

$$\left\|\mathbf{x}_\epsilon^\star - \mathbf{x}^\natural\right\|_2 \leq \underbrace{\|\mathbf{x}_\epsilon^\star - \hat{\mathbf{x}}\|_2}_{\text{approximation error}} + \underbrace{\left\|\hat{\mathbf{x}} - \mathbf{x}^\natural\right\|_2}_{\text{statistical error}} .$$

Motivation: *non-smooth* estimators of $\mathbf{x}^\natural$ can help *reduce the statistical error*.

Non-smooth unconstrained convex minimization

Alternative approach for Linear Regression: Least absolute value deviation:

$$\hat{\mathbf{x}} \in \arg\min_{\mathbf{x}} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_1 .$$

- The advantage: Improved robustness against outliers (high noise values)
- The bad (!) news: A *non-differentiable* objective function

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, $f \in \mathcal{F}$.

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., $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$.

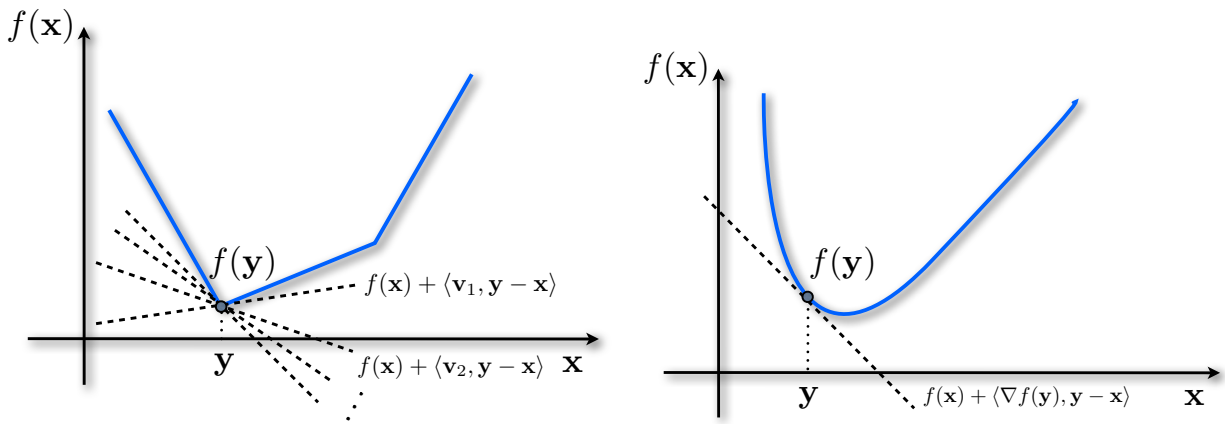


Figure: (**Left**) Non-differentiability at point $\mathbf{y}$. (**Right**) Gradient as a subdifferential with a singleton entry.

(Sub)gradients in convex functions

Example

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

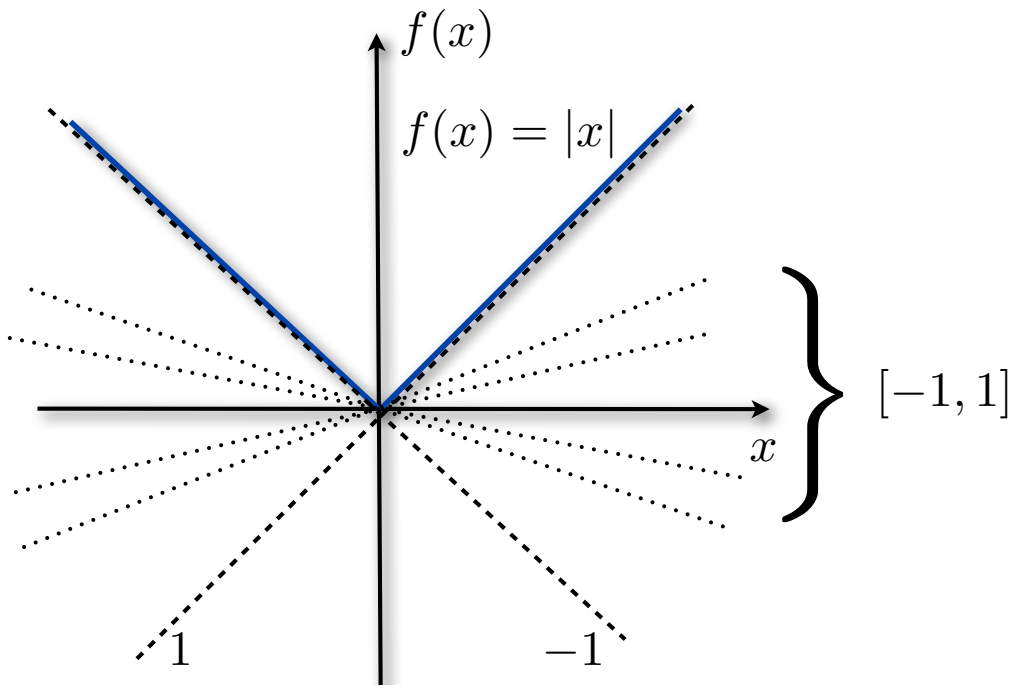


Figure: Subdifferentials of $f(x) = |x|$ in $\mathbb{R}$.

Subdifferentials: Two basic results

Lemma (Necessary and sufficient condition)

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

Sketch of the proof.

- $\Rightarrow$: 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).

- $\Leftarrow$: If $\mathbf{x}^\star$ is a global of (1) then for every $\mathbf{x} \in \mathrm{dom}(F)$, $F(\mathbf{x}) \geq F(\mathbf{x}^\star)$ 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 [8])

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

$$\left\{ \quad \mathbf{x}^{k+1} \quad = \mathbf{x}^k - \alpha_k \mathbf{d}^k, \right. \tag{3}$$

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 $\mathcal{O}(1/\sqrt{k})$ is the slowest we have seen so far!**

Next lecture: Achieving guarantees for (many) non-smooth optimization problems that are just as good as those for smooth ones

Stochastic subgradient methods

- An unbiased stochastic subgradient

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

- Algorithms in Lecture 7 can be extended using unbiased subgradients instead of unbiased gradients.

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 [10])

Suppose that:

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

Then,

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

- **Note:** Rate is $\mathcal{O}(\log k/\sqrt{k})$ instead of $\mathcal{O}(1/\sqrt{k})$ for the deterministic algorithm.

Composite convex minimization

Problem (Unconstrained composite convex minimization)

$$F^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \{F(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x})\} \tag{4}$$

- f *and* g *are both proper, closed, and convex.*
- $dom(F) := dom(f) \cap dom(g) \neq \emptyset$ *and* $-\infty < F^\star < +\infty$.
- *The solution set* $\mathcal{S}^\star := \{\mathbf{x}^\star \in dom(F) : F(\mathbf{x}^\star) = F^\star\}$ *is nonempty.*

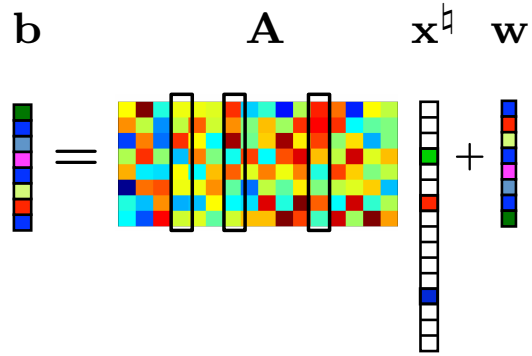
Two remarks

- Nonsmoothness: At least one of the two functions f and g is **nonsmooth**
  - General nonsmooth convex optimization methods (e.g., classical subgradient methods, level, or bundle methods) lack efficiency and numerical robustness.
    - Require $\mathcal{O}(\epsilon^{-2})$ iterations to reach a point $\mathbf{x}^\star_\epsilon$ such that $F(\mathbf{x}^\star_\epsilon) - F^\star \leq \epsilon$. Hence, to reach $\mathbf{x}^\star_{0.01}$ such that $F(\mathbf{x}^\star_{0.01}) - F^\star \leq 0.01$, we need $\mathcal{O}(10^4)$ iterations.
- Generality: it covers a wider range of problems than smooth unconstrained problems, e.g., when handling regularized M-estimation,
  - f is a loss function, a data fidelity, or negative log-likelihood function.
  - g is a regularizer, encouraging structure and/or constraints in the solution.

Example 1: Sparse regression in generalized linear models (GLMs)

Problem (Sparse regression in GLM)

Our goal is to estimate $\mathbf{x}^\natural \in \mathbb{R}^p$ *given* $\{b_i\}_{i=1}^n$ *and* $\{\mathbf{a}_i\}_{i=1}^n$, *knowing that the likelihood function at* y_i *given* $\mathbf{a}_i$ *and* $\mathbf{x}^\natural$ *is given by* $\mathcal{L}(b_i; \langle \mathbf{a}_i, \mathbf{x}^\natural \rangle)$, *and that* $\mathbf{x}^\natural$ *is sparse*.



Optimization formulation

$$\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \underbrace{-\sum_{i=1}^{n} \log \mathcal{L}(b_i; \langle \mathbf{a}_i, \mathbf{x} \rangle)}_{f(\mathbf{x})} + \underbrace{\rho_n \|\mathbf{x}\|_1}_{g(\mathbf{x})} \right\}$$

where $\rho_n > 0$ is a parameter which controls the strength of sparsity regularization.

Theorem (cf. [5] for details)

Under some technical conditions, there exists $\{\rho_i\}_{i=1}^\infty$ *such that with high probability,*

$$\left\| \mathbf{x}^\star - \mathbf{x}^\natural \right\|_2^2 = \mathcal{O}\left(\frac{s \log p}{n} \right), \quad \operatorname{supp} \mathbf{x}^\star = \operatorname{supp} \mathbf{x}^\natural.$$

Recall ML: $\left\| \mathbf{x}_{ML} - \mathbf{x}^\natural \right\|_2^2 = \mathcal{O}\left(p/n\right)$.

Example 2: Image processing

Problem (Imaging denoising/deblurring)

Our goal is to obtain a clean image $\mathbf{x}$ *given "dirty" observations* $\mathbf{b} \in \mathbb{R}^{n \times 1}$ *via* $\mathbf{b} = \mathcal{A}(\mathbf{x}) + \mathbf{w}$, *where* $\mathcal{A}$ *is a linear operator, which, e.g., captures camera blur as well as image subsampling, and* $\mathbf{w}$ *models perturbations, such as Gaussian or Poisson noise.*

Optimization formulation

$$\text{Gaussian}: \quad \min_{\mathbf{x} \in \mathbb{R}^{n \times p}} \left\{ \underbrace{(1/2)\|\mathcal{A}(\mathbf{x}) - \mathbf{b}\|_2^2}_{f(\mathbf{x})} + \underbrace{\rho \|\mathbf{x}\|_{\mathrm{TV}}}_{g(\mathbf{x})} \right\}$$

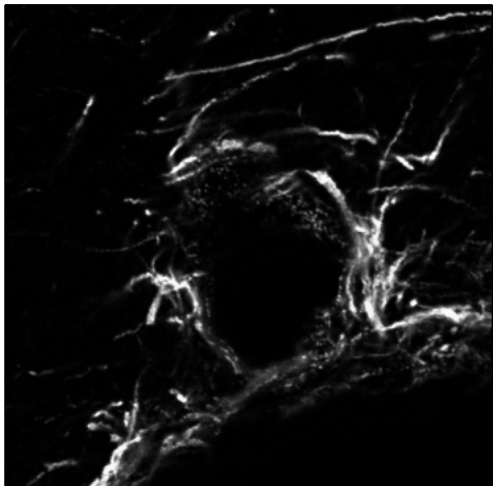
$$\text{Poisson}: \quad \min_{\mathbf{x} \in \mathbb{R}^{n \times p}} \left\{ \underbrace{\frac{1}{n} \sum_{i=1}^{n} \left[\langle \mathbf{a}_i, \mathbf{x} \rangle - b_i \ln \left(\langle \mathbf{a}_i, \mathbf{x} \rangle \right) \right]}_{f(\mathbf{x})} + \underbrace{\rho \|\mathbf{x}\|_{\mathrm{TV}}}_{g(\mathbf{x})} \right\}$$

where $\rho > 0$ is a regularization parameter and $\|\cdot\|_{\mathrm{TV}}$ is the total variation (TV) norm:

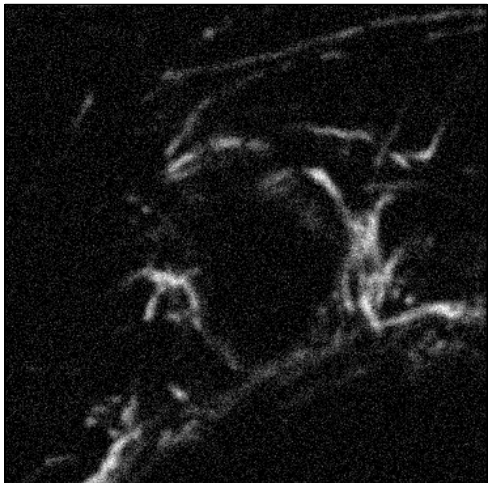
$$\|\mathbf{x}\|_{\mathrm{TV}} := \begin{cases} \sum_{i,j} |\mathbf{x}_{i,j+1} - \mathbf{x}_{i,j}| + |\mathbf{x}_{i+1,j} - \mathbf{x}_{i,j}| & \text{anisotropic case,} \\ \sum_{i,j} \sqrt{|\mathbf{x}_{i,j+1} - \mathbf{x}_{i,j}|^2 + |\mathbf{x}_{i+1,j} - \mathbf{x}_{i,j}|^2} & \text{isotropic case} \end{cases}$$

Example 3: Confocal microscopy with camera blur and Poisson observations

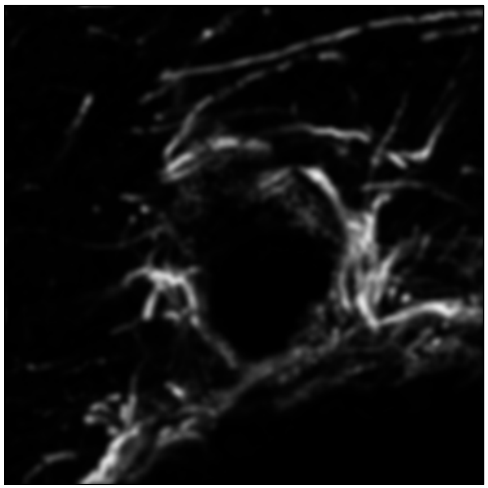
Original image $\mathbf{x}^\natural$



Observed image $\mathbf{b}$



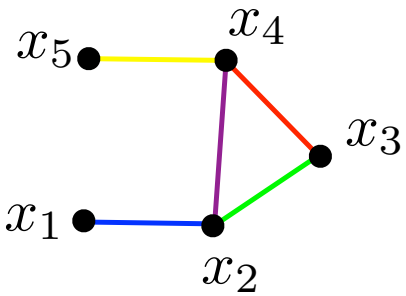
Estimate $\hat{\mathbf{x}}$



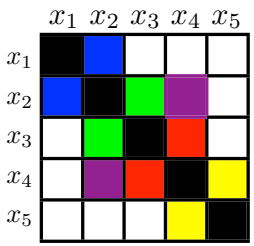
Example 4: Sparse inverse covariance estimation

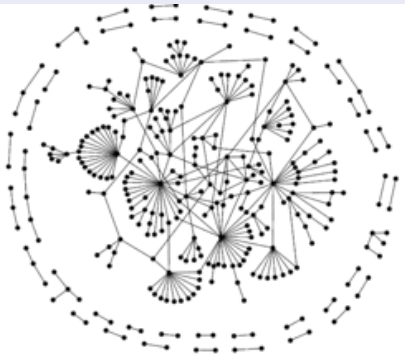
Problem (Graphical model selection)

Given a data set $\mathcal{D} := \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$, where $\mathbf{x}_i$ is a Gaussian random variable. Let Σ be the covariance matrix corresponding to the graphical model of the Gaussian Markov random field. Our goal is to learn a sparse precision matrix Θ (i.e., the inverse covariance matrix Σ^{-1}) that captures the Markov random field structure..



$\Theta =$





Optimization formulation

$$\min_{\Theta \succ 0} \left\{ \underbrace{\operatorname{tr}(\Sigma\Theta) - \log\det(\Theta)}_{f(\mathbf{x})} + \underbrace{\lambda \|\operatorname{vec}(\Theta)\|_1}_{g(\mathbf{x})} \right\} \tag{5}$$

where $\Theta \succ 0$ means that Θ is symmetric and positive definite and $\lambda > 0$ is a regularization parameter and vec is the vectorization operator.

Question: How do we design algorithms for finding a solution $\mathbf{x}^\star$?

Philosophy

- We cannot immediately design algorithms just based on the original formulation

$$F^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \{F(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x})\}. \tag{6}$$

- We need intermediate tools to characterize the solutions $\mathbf{x}^\star$ of this problem
- One key tool is called the **optimality condition**

A short detour: Proximal-point operators

Definition (Proximal operator [9])

Let $g \in \mathcal{F}(\mathbb{R}^p)$ and $\mathbf{x} \in \mathbb{R}^p$. The proximal operator (or prox-operator) of f is defined as:

$$\boxed{\operatorname{prox}_g(\mathbf{x}) \equiv \arg\min_{\mathbf{y}\in\mathbb{R}^p} \left\{ g(\mathbf{y}) + \frac{1}{2}\|\mathbf{y}-\mathbf{x}\|_2^2 \right\}.} \tag{7}$$

A short detour: Proximal-point operators

Definition (Proximal operator [9])

Let $g \in \mathcal{F}(\mathbb{R}^p)$ and $\mathbf{x} \in \mathbb{R}^p$. The proximal operator (or prox-operator) of f is defined as:

$$\boxed{\operatorname{prox}_g(\mathbf{x}) \equiv \arg\min_{\mathbf{y}\in\mathbb{R}^p} \left\{ g(\mathbf{y}) + \frac{1}{2}\|\mathbf{y}-\mathbf{x}\|_2^2 \right\}.} \tag{7}$$

Numerical efficiency: Why do we need proximal operator?

For problem (4):

- For many well-known convex functions g, we can compute $\operatorname{prox}_g(\mathbf{x})$ analytically or very efficiently.
- If $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^p)$, and $\operatorname{prox}_g(\mathbf{x})$ is **cheap** to compute, then solving (4) is as **efficient** as solving $\boxed{\min_{\mathbf{x}\in\mathbb{R}^p} f(\mathbf{x})}$ in terms of complexity.
- If $\operatorname{prox}_f(\mathbf{x})$ and $\operatorname{prox}_g(\mathbf{x})$ are both **cheap** to compute, then *convex splitting* (4) is also efficient.

Tractable prox-operators

Processing non-smooth terms in (4)

- We handle the nonsmooth term g in (4) using the proximal mapping principle.
- Computing proximal operator prox_g of a general convex function g

$$\mathrm{prox}_g(\mathbf{x}) \equiv \arg\min_{\mathbf{y}\in\mathbb{R}^p} \left\{ g(\mathbf{y}) + (1/2)\|\mathbf{y}-\mathbf{x}\|_2^2 \right\}.$$

  can be computationally demanding.
- If we can efficiently compute prox_F, we can use the **proximal-point algorithm** (PPA) [4, 9] to solve (4). Unfortunately, PPA is hardly used in practice to solve (4) since computing $\mathrm{prox}_{\lambda F}(\cdot)$ can be as almost hard as solving (4).

Definition (Tractable proximity)

Given $g \in \mathcal{F}(\mathbb{R}^p)$. We say that g is proximally tractable if prox_g defined by (7) can be computed **efficiently**.

- "**efficiently**" $=$ {closed form solution, low-cost computation, polynomial time}.
- We denote $\mathcal{F}_{\mathrm{prox}}(\mathbb{R}^p)$ the class of proximally tractable convex functions.

Tractable prox-operators

Example

- For separable functions, the prox-operator can be efficient. For instance, $g(\mathbf{x}) := \|\mathbf{x}\|_1 = \sum_{i=1}^n |\mathbf{x}_i|$, we have

$$\boxed{\mathrm{prox}_{\lambda g}(\mathbf{x}) = \mathrm{sign}(\mathbf{x}) \otimes \max\{|\mathbf{x}| - \lambda, 0\}.}$$

- For smooth functions, we can computer the prox-operator via basic algebra. For instance, $g(\mathbf{x}) := (1/2)\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$, one has

$$\boxed{\mathrm{prox}_{\lambda g}(\mathbf{x}) = \Big(\mathbb{I} + \lambda \mathbf{A}^T\mathbf{A}\Big)^{-1}\Big(\mathbf{x} + \lambda \mathbf{A}\mathbf{b}\Big).}$$

- For the indicator functions of simple sets, e.g., $g(\mathbf{x}) := \delta_{\mathcal{X}}(\mathbf{x})$, the prox-operator is the projection operator

$$\boxed{\mathrm{prox}_{\lambda g}(\mathbf{x}) := \pi_{\mathcal{X}}(\mathbf{x})}$$

the projection of $\mathbf{x}$ onto $\mathcal{X}$. For instance, when $\mathcal{X} = \{\mathbf{x} : \|\mathbf{x}\|_1 \leq \lambda\}$, the projection can be obtained efficiently.

Computational efficiency - Example

Proximal operator of quadratic function

The **proximal operator** of a quadratic function $g(\mathbf{x}) := \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$ is defined as

$$\boxed{\operatorname{prox}_{\lambda g}(\mathbf{x}) := \arg\min_{\mathbf{y}\in\mathbb{R}^p} \left\{ \frac{1}{2}\|\mathbf{A}\mathbf{y} - \mathbf{b}\|_2^2 + \frac{1}{2\lambda}\|\mathbf{y} - \mathbf{x}\|_2^2 \right\}.} \tag{8}$$

Computational efficiency - Example

Proximal operator of quadratic function

The **proximal operator** of a quadratic function $g(\mathbf{x}) := \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$ is defined as

$$\boxed{\text{prox}_{\lambda g}(\mathbf{x}) := \arg\min_{\mathbf{y}\in\mathbb{R}^p} \left\{ \frac{1}{2}\|\mathbf{A}\mathbf{y} - \mathbf{b}\|_2^2 + \frac{1}{2\lambda}\|\mathbf{y} - \mathbf{x}\|_2^2 \right\}.} \tag{8}$$

How to compute $\text{prox}_{\lambda g}(\mathbf{x})$?

The optimality condition implies that the solution of (8) should satisfy the following linear system: $\mathbf{A}^T(\mathbf{A}\mathbf{y} - \mathbf{b}) + \lambda^{-1}(\mathbf{y} - \mathbf{x}) = 0$. As a result, we obtain

$$\text{prox}_{\lambda g}(\mathbf{x}) = \left(\mathbb{I} + \lambda \mathbf{A}^T\mathbf{A}\right)^{-1}(\mathbf{x} + \lambda \mathbf{A}\mathbf{b})\,.$$

- When $\mathbf{A}^T\mathbf{A}$ is efficiently diagonalizable (e.g., $\mathbf{U}^T\mathbf{A}^T\mathbf{A}\mathbf{U} := \Lambda$, where $\mathbf{U}$ is a unitary matrix and Λ is a diagonal matrix) then $\text{prox}_{\lambda g}(\mathbf{x})$ can be cheap
  $\text{prox}_{\lambda g}(\mathbf{x}) = \mathbf{U}\left(\mathbb{I} + \lambda\Lambda\right)^{-1}\mathbf{U}^T(\mathbf{x} + \lambda\mathbf{A}\mathbf{b})$.
  - Matrices $\mathbf{A}$ such as TV operator with periodic boundary conditions, index subsampling operators (e.g., as in matrix completion), and circulant matrices (e.g., typical image blur operators) are efficiently diagonalizable with the Fast Fourier transform $\mathbf{U}$.
- If $\mathbf{A}\mathbf{A}^T$ is diagonalizable (e.g., a tight frame $\mathbf{A}$), then we can use the identity
  $$(\mathbb{I} + \lambda\mathbf{A}^T\mathbf{A})^{-1} = \mathbb{I} - \mathbf{A}^T(\lambda^{-1}\mathbb{I} + \mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}.$$

A non-exhaustive list of proximal tractability functions

Name	Function	Proximal operator	Complexity
ℓ_1-norm	$f(\mathbf{x}) := \|\mathbf{x}\|_1$	$\mathrm{prox}_{\lambda f}(\mathbf{x}) = \mathrm{sign}(\mathbf{x}) \otimes [\|\mathbf{x}\| - \lambda]_+$	$\mathcal{O}(p)$
ℓ_2-norm	$f(\mathbf{x}) := \|\mathbf{x}\|_2$	$\mathrm{prox}_{\lambda f}(\mathbf{x}) = [1 - \lambda/\|\mathbf{x}\|_2]_+ \mathbf{x}$	$\mathcal{O}(p)$
Support function	$f(\mathbf{x}) := \max_{\mathbf{y} \in \mathcal{C}} \mathbf{x}^T \mathbf{y}$	$\mathrm{prox}_{\lambda f}(\mathbf{x}) = \mathbf{x} - \lambda \pi_{\mathcal{C}}(\mathbf{x})$	
Box indicator	$f(\mathbf{x}) := \delta_{[\mathbf{a},\mathbf{b}]}(\mathbf{x})$	$\mathrm{prox}_{\lambda f}(\mathbf{x}) = \pi_{[\mathbf{a},\mathbf{b}]}(\mathbf{x})$	$\mathcal{O}(p)$
Positive semidefinite cone indicator	$f(\mathbf{X}) := \delta_{\mathbb{S}^p_+}(\mathbf{X})$	$\mathrm{prox}_{\lambda f}(\mathbf{X}) = \mathbf{U}[\Sigma]_+\mathbf{U}^T$, where $\mathbf{X} = \mathbf{U}\Sigma\mathbf{U}^T$	$\mathcal{O}(p^3)$
Hyperplane indicator	$f(\mathbf{x}) := \delta_{\mathcal{X}}(\mathbf{x})$, $\mathcal{X} := \{\mathbf{x} : \mathbf{a}^T\mathbf{x} = b\}$	$\mathrm{prox}_{\lambda f}(\mathbf{x}) = \pi_{\mathcal{X}}(\mathbf{x}) = \mathbf{x} + \left(\frac{b - \mathbf{a}^T\mathbf{x}}{\|\mathbf{a}\|_2}\right)\mathbf{a}$	$\mathcal{O}(p)$
Simplex indicator	$f(\mathbf{x}) = \delta_{\mathcal{X}}(\mathbf{x}), \mathcal{X} := \{\mathbf{x} : \mathbf{x} \geq 0,\ \mathbf{1}^T\mathbf{x} = 1\}$	$\mathrm{prox}_{\lambda f}(\mathbf{x}) = (\mathbf{x} - \nu\mathbf{1})$ for some $\nu \in \mathbb{R}$, which can be efficiently calculated	$\tilde{\mathcal{O}}(p)$
Convex quadratic	$f(\mathbf{x}) := (1/2)\mathbf{x}^T\mathbf{Q}\mathbf{x} + \mathbf{q}^T\mathbf{x}$	$\mathrm{prox}_{\lambda f}(\mathbf{x}) = (\lambda\mathbb{I} + \mathbf{Q})^{-1}\mathbf{x}$	$\mathcal{O}(p \log p) \to \mathcal{O}(p^3)$
Square ℓ_2-norm	$f(\mathbf{x}) := (1/2)\|\mathbf{x}\|_2^2$	$\mathrm{prox}_{\lambda f}(\mathbf{x}) = (1/(1+\lambda))\mathbf{x}$	$\mathcal{O}(p)$
log-function	$f(\mathbf{x}) := -\log(x)$	$\mathrm{prox}_{\lambda f}(x) = ((x^2 + 4\lambda)^{1/2} + x)/2$	$\mathcal{O}(1)$
log det-function	$f(\mathbf{x}) := -\log\det(\mathbf{X})$	$\mathrm{prox}_{\lambda f}(\mathbf{X})$ is the log-function prox applied to the individual eigenvalues of $\mathbf{X}$	$\mathcal{O}(p^3)$

Here: $[\mathbf{x}]_+ := \max\{0, \mathbf{x}\}$ and $\delta_{\mathcal{X}}$ is the indicator function of the convex set $\mathcal{X}$, sign is the sign function, $\mathbb{S}^p_+$ is the cone of symmetric positive semidefinite matrices.

For more functions, see [1, 7].

A short detour: Basic properties of prox-operator

Property (Basic properties of prox-operator)

1. $\mathrm{prox}_g(\mathbf{x})$ *is well-defined and single-valued (i.e., the prox-operator* (7) *has a unique solution since* $g(\cdot) + (1/2)\|\cdot - \mathbf{x}\|_2^2$ *is strongly convex).*
2. *Optimality condition:*

$$\mathbf{x} \in \mathrm{prox}_g(\mathbf{x}) + \partial g(\mathrm{prox}_g(\mathbf{x})), \ \mathbf{x} \in \mathbb{R}^p.$$

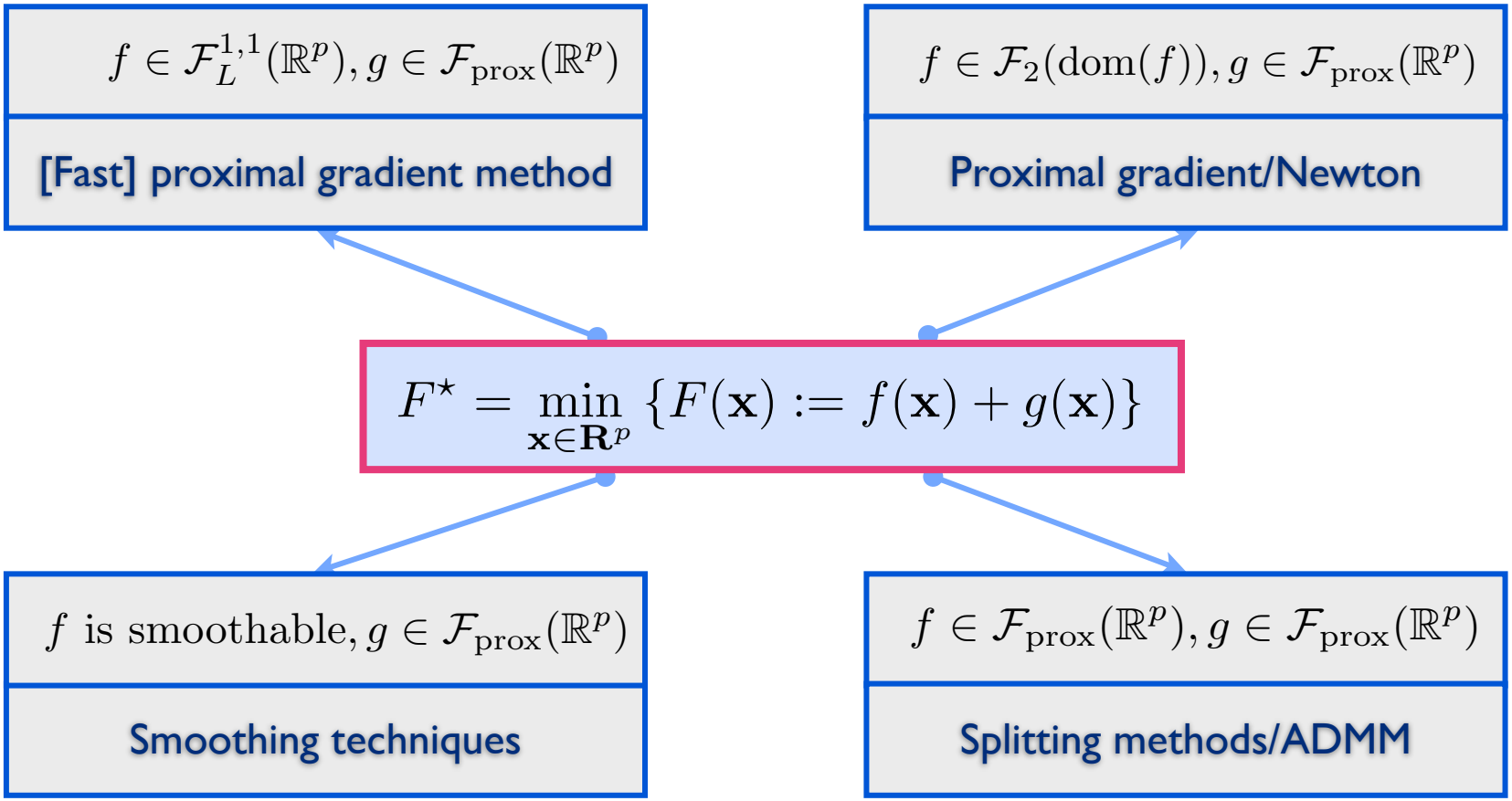
3. $\mathbf{x}^\star$ *is a fixed point of* $\mathrm{prox}_g(\cdot)$*:*

$$0 \in \partial g(\mathbf{x}^\star) \quad \Leftrightarrow \quad \mathbf{x}^\star = \mathrm{prox}_g(\mathbf{x}^\star).$$

4. *Nonexpansiveness:*

$$\|\mathrm{prox}_g(\mathbf{x}) - \mathrm{prox}_g(\tilde{\mathbf{x}})\|_2 \leq \|\mathbf{x} - \tilde{\mathbf{x}}\|_2, \quad \forall \mathbf{x}, \tilde{\mathbf{x}} \in \mathbb{R}^p.$$

Choices of solution methods



- $\mathcal{F}_L^{1,1}$ and $\mathcal{F}_2$ are the class of convex functions with Lipschitz gradient and self-concordant, respectively.
- $\mathcal{F}_{\text{prox}}$ is the class of convex functions with proximity operator (defined in the next slides).
- "smoothable" is defined in the next lectures.

Solution methods

Composite convex minimization

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

Choice of numerical solution methods

- **Solve** (4) = Find $\mathbf{x}^k \in \mathbb{R}^p$ such that

$$\boxed{F(\mathbf{x}^k) - F^\star \leq \varepsilon}$$

for a given tolerance $\varepsilon > 0$.

- **Oracles**: We can use one of the following configurations (oracles):
  1. $\partial f(\cdot)$ and $\partial g(\cdot)$ at any point $\mathbf{x} \in \mathbb{R}^p$.
  2. $\nabla f(\cdot)$ and $\mathrm{prox}_{\lambda g}(\cdot)$ at any point $\mathbf{x} \in \mathbb{R}^p$.
  3. $\mathrm{prox}_{\lambda f}$ and $\mathrm{prox}_{\lambda g}(\cdot)$ at any point $\mathbf{x} \in \mathbb{R}^p$.
  4. $\nabla f(\cdot)$, inverse of $\nabla^2 f(\cdot)$ and $\mathrm{prox}_{\lambda g}(\cdot)$ at any point $\mathbf{x} \in \mathbb{R}^p$.

Using different oracle leads to different types of algorithms

Proximal-gradient method: A quadratic majorization perspective

Definition (Proximal operator [9])

Let $g \in \mathcal{F}(\mathbb{R}^p)$. The proximal operator (or prox-operator) of g is defined as:

$$\operatorname{prox}_g(\mathbf{x}) \equiv \arg\min_{\mathbf{y}\in\mathbb{R}^p} \left\{ g(\mathbf{y}) + \frac{1}{2}\|\mathbf{y}-\mathbf{x}\|_2^2 \right\}.$$

Quadratic upper bound for f

For $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^p)$, we have, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^p$

$$f(\mathbf{x}) \leq f(\mathbf{y}) + \nabla f(\mathbf{y})^T(\mathbf{x}-\mathbf{y}) + \frac{L}{2}\|\mathbf{x}-\mathbf{y}\|_2^2 \coloneqq Q_L(\mathbf{x},\mathbf{y})$$

Quadratic *majorizer* for $f+g$

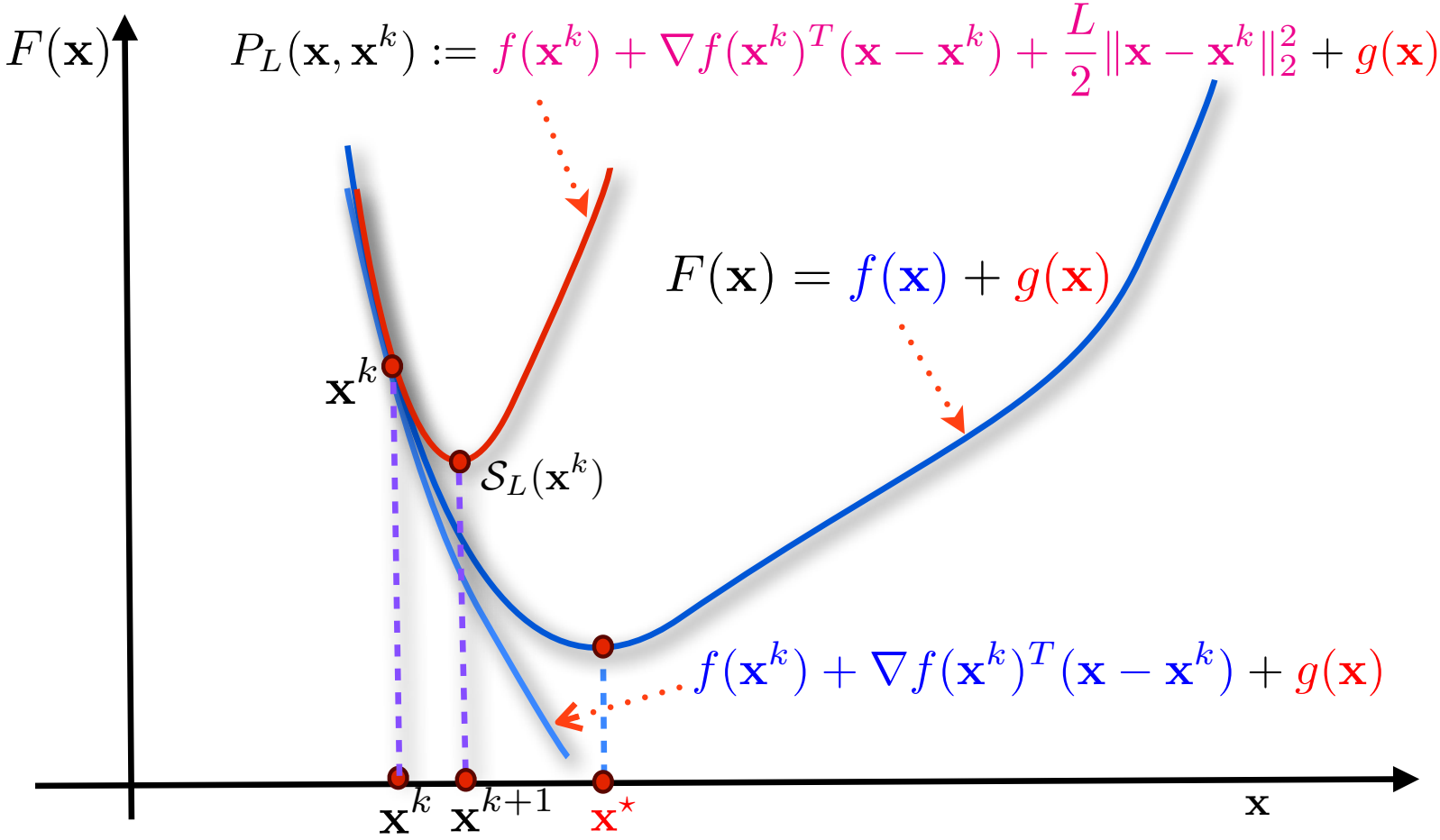
Of course, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^p$,

$$f(\mathbf{x}) \leq Q_L(\mathbf{x},\mathbf{y}) \quad \Rightarrow \quad f(\mathbf{x}) + g(\mathbf{x}) \leq Q_L(\mathbf{x},\mathbf{y}) + g(\mathbf{x}) \coloneqq P_L(\mathbf{x},\mathbf{y})$$

Proximal-gradient from the majorize-minimize perspective

$$\mathbf{x}^{k+1} = \arg\min_{\mathbf{x}} P_L(\mathbf{x},\mathbf{x}^k) = \operatorname{prox}_{g/L}(\mathbf{x} - \nabla f(\mathbf{x}^k)/L)$$

Geometric illustration



Proximal-gradient algorithm

Basic proximal-gradient scheme (ISTA)

1. Choose $\mathbf{x}^0 \in \mathrm{dom}(F)$ arbitrarily as a starting point.

2. For $k = 0, 1, \cdots$, generate a sequence $\{\mathbf{x}^k\}_{k \geq 0}$ as:

$$\mathbf{x}^{k+1} := \mathrm{prox}_{\alpha g}\left(\mathbf{x}^k - \alpha \nabla f(\mathbf{x}^k)\right),$$

where $\alpha := \frac{1}{L}$.

Proximal-gradient algorithm

Basic proximal-gradient scheme (ISTA)

1. Choose $\mathbf{x}^0 \in \mathrm{dom}(F)$ arbitrarily as a starting point.

2. For $k = 0, 1, \cdots$, generate a sequence $\{\mathbf{x}^k\}_{k \geq 0}$ as:

$$\mathbf{x}^{k+1} := \mathrm{prox}_{\alpha g}\left(\mathbf{x}^k - \alpha \nabla f(\mathbf{x}^k)\right),$$

where $\alpha := \frac{1}{L}$.

Theorem (Convergence of ISTA [2])

Let $\{\mathbf{x}^k\}$ be generated by ISTA. Then:

$$F(\mathbf{x}^k) - F^\star \leq \frac{L_f \|\mathbf{x}^0 - \mathbf{x}^\star\|_2^2}{2(k+1)}$$

The worst-case complexity to reach $F(\mathbf{x}^k) - F^\star \leq \varepsilon$ of (ISTA) is $\mathcal{O}\left(\frac{L_f R_0^2}{\varepsilon}\right)$, where $R_0 := \max_{\mathbf{x}^\star \in \mathcal{S}^\star} \|\mathbf{x}^0 - \mathbf{x}^\star\|_2$.

- A line-search procedure can be used to estimate L_k for L based on $(0 < c \leq 1)$:

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) - \frac{c}{2L_k}\|\nabla f(\mathbf{x}^k)\|^2.$$

Fast proximal-gradient algorithm

Fast proximal-gradient scheme (FISTA)

1. Choose $\mathbf{x}^0 \in \mathrm{dom}(F)$ arbitrarily as a starting point.

2. Set $\mathbf{y}^0 := \mathbf{x}^0$ and $t_0 := 1$.

3. For $k = 0, 1, \ldots$, generate two sequences $\{\mathbf{x}^k\}_{k\geq 0}$ and $\{\mathbf{y}^k\}_{k\geq 0}$ as:

$$\begin{cases} \mathbf{x}^{k+1} & := \mathrm{prox}_{\alpha g}\left(\mathbf{y}^k - \alpha \nabla f(\mathbf{y}^k)\right), \\ t_{k+1} & := (1 + \sqrt{4t_k^2 + 1})/2, \\ \mathbf{y}^{k+1} & := \mathbf{x}^{k+1} + \frac{t_k - 1}{t_{k+1}}(\mathbf{x}^{k+1} - \mathbf{x}^k). \end{cases}$$

where $\alpha := L^{-1}$.

From $\mathcal{O}\left(\frac{L_f R_0^2}{\epsilon}\right)$ to $\mathcal{O}\left(R_0\sqrt{\frac{L_f}{\epsilon}}\right)$ iterations at almost no additional cost!.

Complexity per iteration

- **One** gradient $\nabla f(\mathbf{y}^k)$ and **one** prox-operator of g;
- 8 arithmetic operations for t_{k+1} and γ_{k+1};
- 2 more vector additions, and **one** scalar-vector multiplication.

The **cost per iteration** is almost the same as in **gradient scheme** if proximal operator of g is efficient.

Example 1: ℓ_1-regularized least squares

Problem (ℓ_1-regularized least squares)

Given $\mathbf{A} \in \mathbb{R}^{n \times p}$ *and* $\mathbf{b} \in \mathbb{R}^n$, *solve:*

$$F^\star := \min_{\mathbf{x} \in \mathbb{R}^p} \left\{ F(\mathbf{x}) := \frac{1}{2} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \lambda \|\mathbf{x}\|_1 \right\}, \tag{9}$$

where $\lambda > 0$ *is a regularization parameter.*

Complexity per iterations

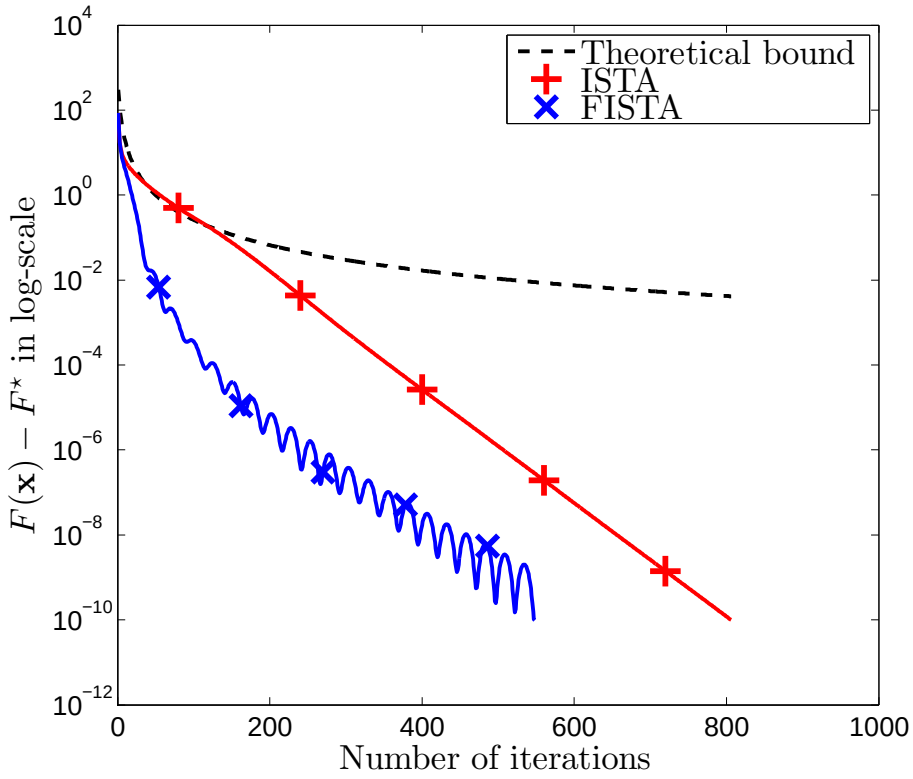
- Evaluating $\nabla f(\mathbf{x}^k) = \mathbf{A}^T(\mathbf{A}\mathbf{x}^k - \mathbf{b})$ requires one $\mathbf{A}\mathbf{x}$ and one $\mathbf{A}^T\mathbf{y}$.
- One soft-thresholding operator $\mathrm{prox}_{\lambda g}(\mathbf{x}) = \mathrm{sign}(\mathbf{x}) \otimes \max\{|\mathbf{x}| - \lambda, 0\}$.
- **Optional**: Evaluating $L = \|\mathbf{A}^T\mathbf{A}\|$ (spectral norm) - via **power iterations** (e.g., 20 iterations, each iteration requires one $\mathbf{A}\mathbf{x}$ and one $\mathbf{A}^T\mathbf{y}$).

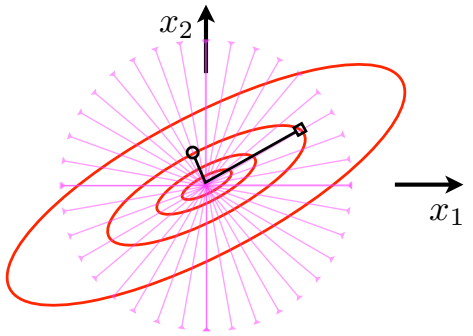
Synthetic data generation

- $\mathbf{A} := \mathrm{randn}(n, p)$ - standard Gaussian $\mathcal{N}(0, \mathbb{I})$.
- $\mathbf{x}^\star$ is a k-sparse vector generated randomly.
- $\mathbf{b} := \mathbf{A}\mathbf{x}^\star + \mathcal{N}(0, 10^{-3})$.

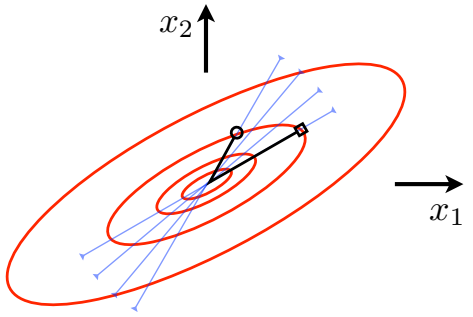
Example 1: Theoretical bounds vs practical performance

- **(Theoretical bounds) FISTA** $:= \frac{2L_f R_0^2}{(k+2)^2}$ and **ISTA** $:= \frac{L_f R_0^2}{2(k+2)}$.





descent directions



restricted descent directions

- ℓ_1-regularized least squares formulation has **restricted strong convexity**. The proximal-gradient method can automatically exploit this structure.

Adaptive Restart

It is possible the preserve $\mathcal{O}(1/k^2)$ convergence guarantee !

One needs to slightly modify the algorithm as below.

Generalized fast proximal-gradient scheme

1. Choose $\mathbf{x}^0 = \mathbf{x}^{-1} \in \mathrm{dom}(F)$ arbitrarily as a starting point.
2. Set $\theta_0 = \theta_{-1} = 1$
3. For $k = 0, 1, \ldots$, generate two sequences $\{\mathbf{x}^k\}_{k\geq 0}$ and $\{\mathbf{y}^k\}_{k\geq 0}$ as:

$$\left\{\begin{array}{l} \mathbf{y}^k := \mathbf{x}^k + \theta_k(\theta_{k-1}^{-1} - 1)(\mathbf{x}^k - \mathbf{x}^{k-1}) \\ \mathbf{x}^{k+1} := \mathrm{prox}_{\lambda g}\left(\mathbf{y}^k - \lambda \nabla f(\mathbf{y}^k)\right), \\ \textbf{if restart test holds} \\ \qquad \theta_{k-1} = \theta_k = 1 \\ \qquad \mathbf{y}^k = \mathbf{x}^k \\ \qquad \mathbf{x}^{k+1} := \mathrm{prox}_{\lambda g}\left(\mathbf{y}^k - \lambda \nabla f(\mathbf{y}^k)\right) \end{array}\right. \tag{10}$$

where $\lambda := L_f^{-1}$.

θ_k is chosen so that it satisfies

$$\theta_{k+1} = \frac{\sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k^2}{2} < \frac{2}{k+3}$$

Adaptive Restart: Guarantee

Theorem (**Global complexity** [3])

The sequence $\{\mathbf{x}^k\}_{k\geq 0}$ *generated by the modified algorithm satisfies*

$$F(\mathbf{x}^k) - F^\star \leq \frac{2L_f}{(k+2)^2}\left(R_0^2 + \sum_{k_i \leq k}\left(\|\mathbf{x}^\star - \mathbf{x}^{k_i}\|_2^2 - \|\mathbf{x}^\star - \mathbf{z}^{k_i}\|_2^2\right)\right) \quad \forall k \geq 0. \quad (11)$$

where $R_0 := \min_{\mathbf{x}^\star \in \mathcal{S}^\star} \|\mathbf{x}^0 - \mathbf{x}^\star\|$, $\mathbf{z}^k = \mathbf{x}^{k-1} + \theta_{k-1}^{-1}(\mathbf{x}^k - \mathbf{x}^{k-1})$ *and* $k_i, i = 1\ldots$ *are the iterations for which the restart test holds.*

Various restarts tests that might coincide with $\|\mathbf{x}^* - \mathbf{x}^{k_i}\|_2^2 \leq \|\mathbf{x}^* - \mathbf{z}^{k_i}\|_2^2$

- Exact non-monotonicity test: $F(\mathbf{x}^{k+1}) - F(\mathbf{x}^k) > 0$
- Non-monotonicity test: $\langle (L_F(\mathbf{y}^{k-1} - \mathbf{x}^k), \mathbf{x}^{k+1} - \frac{1}{2}(\mathbf{x}^k + y^{k-1})\rangle > 0$ (implies exact non-monotonicity and it is advantageous when function evaluations are expensive)
- Gradient-mapping based test: $\langle (L_f(\mathbf{y}^k - \mathbf{x}^{k+1}), \mathbf{x}^{k+1} - \mathbf{x}^k\rangle > 0$

Example 2: Sparse logistic regression

Problem (Sparse logistic regression)

Given $\mathbf{A} \in \mathbb{R}^{n \times p}$ *and* $\mathbf{b} \in \{-1, +1\}^n$, *solve:*

$$F^\star := \min_{\mathbf{x},\beta} \left\{ F(\mathbf{x}) := \frac{1}{n} \sum_{j=1}^{n} \log\left(1 + \exp\left(-\mathbf{b}_j(\mathbf{a}_j^T \mathbf{x} + \beta)\right)\right) + \rho \|\mathbf{x}\|_1 \right\}.$$

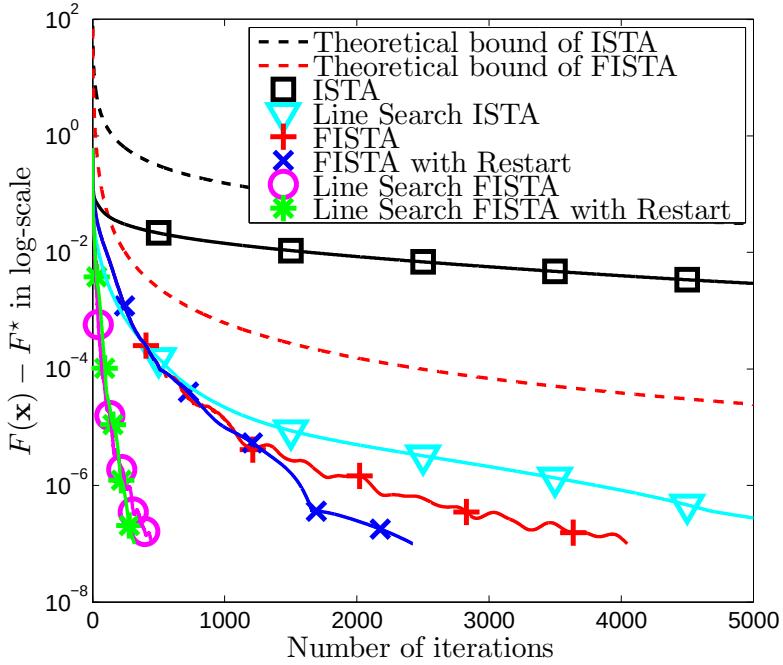
Real data

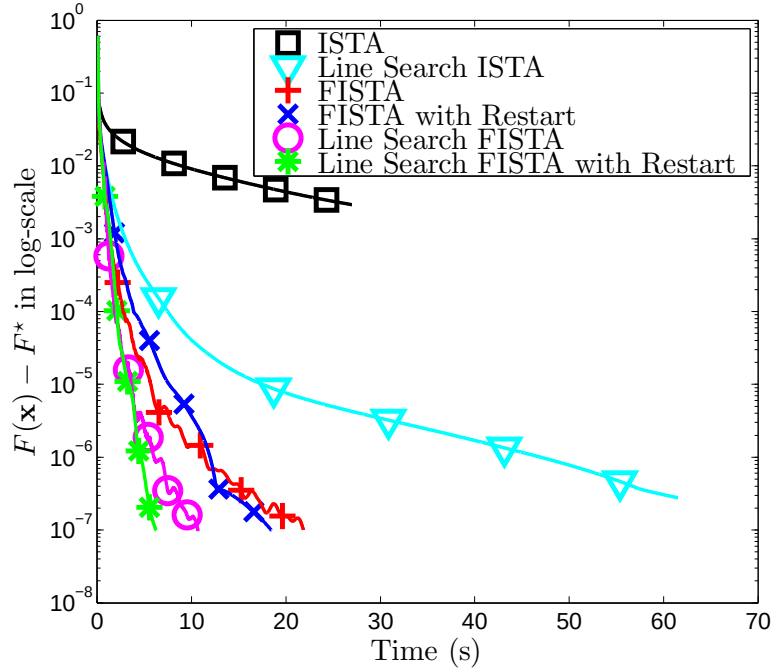
- Real data: `w8a` with $n = 49'749$ data points, $p = 300$ features
- Available at `http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html`.

Parameters

- $\rho = 10^{-4}$.
- Number of iterations 5000, tolerance 10^{-7}.
- Ground truth: Solve problem up to 10^{-9} accuracy by TFOCS to get a high accuracy approximation of $\mathbf{x}^\star$ and $F^\star$.

Example 2: Sparse logistic regression - numerical results





	ISTA	LS-ISTA	FISTA	FISTA-R	LS-FISTA	LS-FISTA-R
Number of iterations	5000	5000	4046	2423	447	317
CPU time (s)	26.975	61.506	21.859	18.444	10.683	6.228
Solution error ($\times 10^{-7}$)	29370	2.774	1.000	0.998	0.961	0.985

Strong convexity case: Algorithms

Proximal-gradient scheme (ISTA$_\mu$)

1. Given $\mathbf{x}^0 \in \mathbb{R}^p$ as a starting point.

2. For $k = 0, 1, \cdots$, generate a sequence $\{\mathbf{x}^k\}_{k\geq 0}$ as:

$$\mathbf{x}^{k+1} := \operatorname{prox}_{\alpha_k g}\Big(\mathbf{x}^k - \alpha_k \nabla f(\mathbf{x}^k)\Big),$$

where $\alpha_k := 2/(L_f + \mu)$ is the optimal step-size.

Fast proximal-gradient scheme (FISTA$_\mu$)

1. Given $\mathbf{x}^0 \in \mathbb{R}^p$ as a starting point. Set $\mathbf{y}^0 := \mathbf{x}^0$.

2. For $k = 0, 1, \cdots$, generate two sequences $\{\mathbf{x}^k\}_{k\geq 0}$ and $\{\mathbf{y}^k\}_{k\geq 0}$ as:

$$\begin{cases} \mathbf{x}^{k+1} := \operatorname{prox}_{\alpha_k g}\Big(\mathbf{y}^k - \alpha_k \nabla f(\mathbf{y}^k)\Big), \\ \mathbf{y}^{k+1} := \mathbf{x}^{k+1} + \left(\frac{\sqrt{c_f}-1}{\sqrt{c_f}+1}\right)(\mathbf{x}^{k+1} - \mathbf{x}^k), \end{cases}$$

where $\alpha_k := L_f^{-1}$ is the optimal step-size.

Strong convexity case: Convergence

Assumption

f *is strongly convex with parameter* $\mu > 0$, *i.e.,* $f \in \mathcal{F}^{1,1}_{L,\mu}(\mathbb{R}^p)$.

Condition number: $c_f := \frac{L_f}{\mu} \geq 0$.

Theorem (**ISTA**$_\mu$ [6])

$$F(\mathbf{x}^k) - F^\star \leq \frac{L_f}{2} \left(\frac{c_f - 1}{c_f + 1} \right)^{2k} \|\mathbf{x}^0 - \mathbf{x}^\star\|_2^2.$$

Convergence rate: **Linear** *with contraction factor:* $\omega := \left(\frac{c_f - 1}{c_f + 1} \right)^2 = \left(\frac{L_f - \mu}{L_f + \mu} \right)^2$.

Theorem (**FISTA**$_\mu$ [6])

$$F(\mathbf{x}^k) - F^\star \leq \frac{L_f + \mu}{2} \left(1 - \sqrt{\frac{\mu}{L_f}} \right)^k \|\mathbf{x}^0 - \mathbf{x}^\star\|_2^2.$$

Convergence rate: **Linear** *with contraction factor:* $\omega_f = \frac{\sqrt{L_f} - \sqrt{\mu}}{\sqrt{L_f}} < \omega$.

A practical issue

Stopping criterion

Fact: If $\mathcal{PG}_{\mathcal{L}}(\mathbf{x}^\star) = 0$, then $\mathbf{x}^\star$ is optimal to (4), where

$$\mathcal{PG}_{\mathcal{L}}(\mathbf{x}) = L\left(\mathbf{x} - \operatorname{prox}_{(1/L)g}\left(\mathbf{x} - (1/L)\nabla f(\mathbf{x})\right)\right).$$

Stopping criterion: (relative solution change)

$$L_k\|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2 \leq \varepsilon \max\{L_0\|\mathbf{x}^1 - \mathbf{x}^0\|_2, 1\},$$

where ε is a given tolerance.

Summary of the worst-case complexities

Software

TFOCS is a good software package to learn about first order methods.

`http://cvxr.com/tfocs/`

Comparison with gradient scheme ($F(\mathbf{x}^k) - F^\star \leq \varepsilon$)

Complexity	Proximal-gradient scheme	Fast proximal-gradient scheme
Complexity [$\mu = 0$]	$\mathcal{O}\left(R_0^2(L_f/\varepsilon)\right)$	$\mathcal{O}\left(R_0\sqrt{L_f/\varepsilon}\right)$
Per iteration	1-gradient, 1-prox, 1-sv, 1-$v+$	1-gradient, 1-prox, 2-sv, 3-$v+$
Complexity [$\mu > 0$]	$\mathcal{O}\left(\kappa\log(\varepsilon^{-1})\right)$	$\mathcal{O}\left(\sqrt{\kappa}\log(\varepsilon^{-1})\right)$
Per iteration	1-gradient, 1-prox, 1-sv, 1-$v+$	1-gradient, 1-prox, 1-sv, 2-$v+$

Here: sv = scalar-vector multiplication, $v+$=vector addition.
$R_0 := \max_{\mathbf{x}^\star \in \mathcal{S}^\star} \|\mathbf{x}^0 - \mathbf{x}^\star\|$ and $\kappa = L_f/\mu_f$ is the condition number.

Summary of the worst-case complexities

Software

TFOCS is a good software package to learn about first order methods.

`http://cvxr.com/tfocs/`

Comparison with gradient scheme ($F(\mathbf{x}^k) - F^\star \leq \varepsilon$)

Complexity	Proximal-gradient scheme	Fast proximal-gradient scheme
Complexity [$\mu = 0$]	$\mathcal{O}\left(R_0^2(L_f/\varepsilon)\right)$	$\mathcal{O}\left(R_0\sqrt{L_f/\varepsilon}\right)$
Per iteration	1-gradient, 1-prox, 1-sv, 1-$v+$	1-gradient, 1-prox, 2-sv, 3-$v+$
Complexity [$\mu > 0$]	$\mathcal{O}\left(\kappa\log(\varepsilon^{-1})\right)$	$\mathcal{O}\left(\sqrt{\kappa}\log(\varepsilon^{-1})\right)$
Per iteration	1-gradient, 1-prox, 1-sv, 1-$v+$	1-gradient, 1-prox, 1-sv, 2-$v+$

Here: sv = scalar-vector multiplication, $v+$=vector addition.
$R_0 := \max_{\mathbf{x}^\star \in \mathcal{S}^\star} \|\mathbf{x}^0 - \mathbf{x}^\star\|$ and $\kappa = L_f/\mu_f$ is the condition number.

Need alternatives when

- f is only self-concordant
- computing $\nabla f(\mathbf{x})$ is much costlier than computing prox_g

⋆Examples

Example (Sparse graphical model selection)

$$\min_{\Theta \succ 0} \left\{ \underbrace{\operatorname{tr}(\Sigma\Theta) - \log\det(\Theta)}_{f(\mathbf{x})} + \underbrace{\rho\|\operatorname{vec}(\Theta)\|_1}_{g(\mathbf{x})} \right\}$$

where $\Theta \succ 0$ means that Θ is symmetric and positive definite, and $\rho > 0$ is a regularization parameter and vec is the vectorization operator.

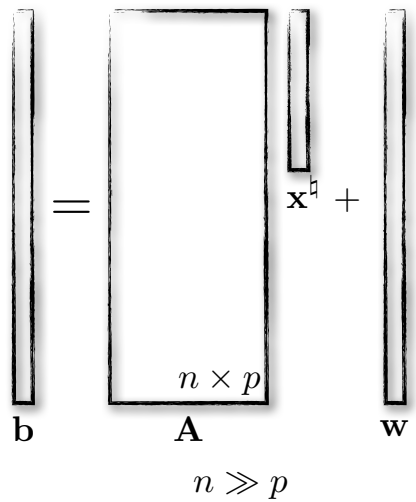
- Computing the gradient is expensive: $\nabla f(\Theta) = \Theta^{-1}$.
- $f \in \mathcal{F}_2$ is self-concordant. However, if $\alpha\mathbf{I} \preceq \Theta \preceq \beta\mathbf{I}$, then $f \in \mathcal{F}_{L,\mu}^{2,1}$ with $L = \sqrt{p}/\alpha^2$ and $\mu = (\beta^2\sqrt{p})^{-1}$.

Example (ℓ_1-regularized Lasso)

$$\min_{\mathbf{x}} \underbrace{\frac{1}{2}\|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2}_{f(\mathbf{x})} + \underbrace{\rho\|\mathbf{x}\|_1}_{g(\mathbf{x})}$$

where $n \gg p$, $\mathbf{A} \in \mathbb{R}^{n\times p}$ is a full column-rank matrix, and $\rho > 0$ is a regularization parameter.

- $f \in \mathcal{F}_{L,\mu}^{2,1}$ and computing the gradient is $\mathcal{O}(n)$.



⋆ Examples (Learned ISTA)

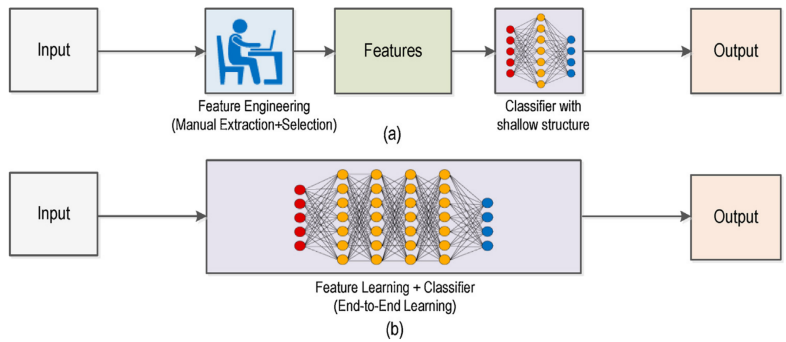


Figure: a) Traditional machine learning approaches b) End-to-end deep learning methods [?]

Example

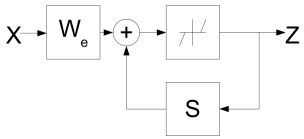
In sparse coding, given a an overcomplete family basis vectors, we would like to find the optimal sparse code vector $Z^* \in \mathbb{R}^m$ to reconstruct the input vector $X \in \mathbb{R}^n$

$$Z^* = \arg\min_Z \frac{1}{2}\|X - W_d Z\|_2^2 + \alpha\|Z\|_1, \tag{12}$$

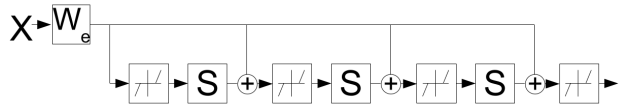
where W_d is $n \times m$ dictionary matrix, whose columns are normalized basis vectors.

- Too slow for real-time applications

⋆ Examples (Learned ISTA)



ISTA method for sparse coding



Learned ISTA - time unfolded version of the ISTA block diagram

Example

$$\mathcal{L}(W) = \frac{1}{P} \sum_{p=0}^{(P-1)} \underbrace{\frac{1}{2} \|Z^{*p} - f_e(W, X_p)\|^2}_{L(W, X_p)} \tag{13}$$

where $(X_1, \ldots, X_P)$ are some samples, and Z^{*p} is the optimal code for sample X^p. Parameters of the encoder are chosen as minimizers of $\mathcal{L}(W)$.

References I

[1] Heinz H Bauschke, Regina Burachik, Patrick L Combettes, Veit Elser, D Russell Luke, and Henry Wolkowicz.
Fixed-point algorithms for inverse problems in science and engineering, volume 49.
Springer Science & Business Media, 2011.

[2] Amir Beck and Marc Teboulle.
A fast iterative shrinkage-thresholding algorithm for linear inverse problems.
SIAM journal on imaging sciences, 2(1):183–202, 2009.

[3] Pontus Giselsson and Stephen Boyd.
Monotonicity and restart in fast gradient methods.
In *Decision and Control (CDC), 2014 IEEE 53rd Annual Conference on*, pages 5058–5063. IEEE, 2014.

[4] Osman Güler.
On the convergence of the proximal point algorithm for convex minimization.
SIAM Journal on Control and Optimization, 29(2):403–419, 1991.

[5] 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.

References II

[6] Yurii Nesterov.
Introductory lectures on convex optimization: A basic course, volume 87.
Springer Science & Business Media, 2013.

[7] Neal Parikh, Stephen Boyd, et al.
Proximal algorithms.
Foundations and Trends® in Optimization, 1(3):127–239, 2014.

[8] R. Tyrrell Rockafellar.
Convex analysis.
Princeton Mathematical Series, No. 28. Princeton University Press, Princeton, N.J., 1970.

[9] R Tyrrell Rockafellar.
Monotone operators and the proximal point algorithm.
SIAM journal on control and optimization, 14(5):877–898, 1976.

[10] Ohad Shamir and Tong Zhang.
Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes.
In *International Conference on Machine Learning*, pages 71–79, 2013.