AMATH 840: Advanced Numerical Methods for Computational and Data Science

Giang Tran
Department of Applied Mathematics, University of Waterloo

Winter 2024

Part 1: Sparse Optimization and Compressive Sensing

1.3: ℓ_1-Optimization Algorithms

Winter 2024

Outline

Some Popular Algorithms for Compressive Sensing - Part 2

- Basis pursuit:

$$\min_{\mathbf{z}\in\mathbb{C}^n} \|\mathbf{z}\|_1 \quad s.t. \quad A\mathbf{z} = \mathbf{y}. \tag{BP}$$

- Basis pursuit denoising:

$$\min_{\mathbf{z}\in\mathbb{C}^n} \|\mathbf{z}\|_1 \quad s.t. \quad \|A\mathbf{z} - \mathbf{y}\|_2 \leq \eta, \tag{BP_η}$$

  or

$$\min_{\mathbf{z}\in\mathbb{C}^n} \frac{1}{2}\|A\mathbf{z} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{z}\|_1. \tag{QP_λ}$$

- Lasso:

$$\min_{\mathbf{z}\in\mathbb{C}^n} \frac{1}{2}\|A\mathbf{z} - \mathbf{y}\|_2^2 \quad s.t. \quad \|\mathbf{z}\|_1 \leq \tau. \tag{LS_τ}$$

References:

- Chapter 3 from "A Mathematical Introduction to Compressive Sensing", by S. Foucart and H. Rauhut.
- ECE236C, Optimization Methods for Large-Scale Systems, by Boyd and Vandenberghe

Relations between ℓ_1-Optimization Models

Proposition 3.2. (Relations between (BP_η), (QP_λ), and (LS_τ).

1. If $\mathbf{z}_{qp}$ is a minimizer of (QP_λ) with $\lambda > 0$, then there exists $\sigma = \sigma_{\mathbf{z}_{qp}} \geq 0$ such that $\mathbf{z}_{qp}$ is a minimizer of (BP_η).
2. If $\mathbf{z}_{bp}$ is a unique minimizer of (BP_η) with $\sigma \geq 0$, then there exists $\tau = \tau_{\mathbf{z}_{bp}} \geq 0$ such that $\mathbf{z}_{bp}$ is a unique minimizer of (LS_τ).
3. If $\mathbf{z}_{ls}$ is a minimizer of (LS_τ) with $\tau > 0$, then there exists $\lambda = \lambda_{\mathbf{z}_{ls}} \geq 0$ such that $\mathbf{z}_{ls}$ is a minimizer of (QP_λ).

Proof Sketch.

- $(QP_\lambda \Rightarrow BP_\eta)$. Set $\sigma := \|Az_{qp} - y\|_2$.
- $(BP_\eta \Rightarrow LS_\tau)$. Set $\tau := \|z_{bp}\|_1$.
- $(LS_\tau \Rightarrow QP_\lambda)$. See Theorem B.28 from "A Mathematical Introduction to Compressive Sensing", by S. Foucart and H. Rauhut.

Relations between ℓ_1-Optimization Models (cont'd)

With suitable values of η, λ, τ, the solutions of $BP_\eta, QP_\lambda, LS_\tau$ coincide.

- If A is orthogonal, a suggestion is $\lambda = \eta\sqrt{2\log(n)}$. [1]
- In general, the relations among η, λ, τ cannot be known a priori. [2]
- If λ is large enough, the solution of QP_λ problem is $z_\lambda = 0$.

Theorem (BP vs QP_λ.)
Assume that $Aw = y$ has a solution. For each $\lambda > 0$, let z_λ be a minimizer of (QP_λ). If the (BP) problem has a unique solution $z^\#$, then

$$\lim_{\lambda \to 0^+} z_\lambda = z^\#.$$

[1] *Atomic Decomposition by Basis Pursuit*, by Chen, Donoho, and Saunders, SIAM Review, 2001.

[2] *Probing the Pareto frontier for basis pursuit solutions*, by E. van den Berg and M. P. Friedlander, SIAM J. on Scientific Computing, 2008.

Outline

Relations between ℓ_1-Optimization Models

Minimizing the Sum of Two Convex Functions

- Proximal Operator
- Minimizing the Sum of Two Convex Functions

Some Popular ℓ_1-Optimization Algorithms

- FISTA: A fast iterative shrinkage-thresholding algorithm
- Nesterov's Second Method
- spgl1 and Other Available Packages
- Alternating Direction Method of Multipliers (ADMM)

Proximal Operator

Definition

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a closed, proper, convex function, which means that its epigraph

$$\text{epi}\, f = \{(x, t) \in \mathbb{R}^n \times \mathbb{R} : f(x) \leq t\}$$

is a nonempty closed convex set.

- The proximal operator $\text{prox}_f : \mathbb{R}^n \to \mathbb{R}^n$ is defined as follows:

$$\text{prox}_f(\mathbf{x}_0) := \underset{\mathbf{x}}{\text{argmin}} \left\{ f(\mathbf{x}) + \frac{1}{2}\|\mathbf{x} - \mathbf{x}_0\|_2^2 \right\}, \quad \text{where} \quad \mathbf{x}_0 \in \mathbb{R}^n.$$

- The proximal operator of the scaled function λf, where $\lambda > 0$, is also called the proximal operator of f with parameter λ.

Subgradient characterization: $u = \text{prox}_f(x) \Leftrightarrow x - u \in \partial f(u)$, where

$$\partial f(\mathbf{x}) := \{\mathbf{z} : \mathbf{z}^T(\mathbf{y} - \mathbf{x}) \leq f(\mathbf{y}) - f(\mathbf{x}),\ \forall \mathbf{y} \in \text{dom}(f)\}.$$

[3]

[3] *Proximal Algorithms*, by Parikh and Boyd, Foundations and Trends in Optimization 2013.

Examples of Proximal Operators

- Example 1: The proximal operator of the ℓ_1 function is

$$\begin{aligned} \text{prox}_{\lambda\|\cdot\|_1}(\mathbf{x}_0) &:= \operatorname*{argmin}_{\mathbf{x}} \left\{ \lambda\|\mathbf{x}\|_1 + \frac{1}{2}\|\mathbf{x} - \mathbf{x}_0\|_2^2 \right\} \\ &= \text{sign}(\mathbf{x}_0)\max(|\mathbf{x}_0| - \lambda, 0). \quad \text{(element-wise)} \end{aligned}$$

  It is called a soft-thresholding operator.

- Example 2: The proximal operator of the ℓ_0 function is

$$\text{prox}_{\lambda\|\cdot\|_0}(\mathbf{x}_0) := \operatorname*{argmin}_{\mathbf{x}} \left\{ \lambda\|\mathbf{x}\|_0 + \frac{1}{2}\|\mathbf{x} - \mathbf{x}_0\|_2^2 \right\} = \text{(discuss in class)}.$$

  It is called a hard-thresholding operator.

- Example 3: The proximal operator of the indicator function of closed convex set C is

$$\text{prox}_{i_C}(\mathbf{x}_0) = \operatorname*{argmin}_{\mathbf{u}\in C} \|\mathbf{u} - \mathbf{x}_0\|_2^2 = P_C(\mathbf{x}_0) \quad \text{(projection on } C\text{)}.$$

Examples of Projection on a Closed Convex Set

- For $C = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{a}^T\mathbf{x} = \mathbf{b}\}$ with $\mathbf{a} \neq 0$, then

$$P_C(\mathbf{x}) = \mathbf{x} + \frac{\mathbf{b} - \mathbf{a}^T\mathbf{x}}{\|\mathbf{a}\|_2^2}\mathbf{a}.$$

- For $C = \{\mathbf{x} \in \mathbb{R}^n \mid A\mathbf{x} = \mathbf{b}\}$ (with $A \in \mathbb{R}^{p\times n}$ and $\text{rank}(A) = p \ll n$):

$$P_C(\mathbf{x}) = \mathbf{x} + A^T(AA^T)^{-1}(\mathbf{b} - A\mathbf{x}).$$

- For $C = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{a}^T\mathbf{x} \leq \mathbf{b}\}$ with $\mathbf{a} \neq 0$, then

$$P_C(\mathbf{x}) = \begin{cases} \mathbf{x} + \frac{\mathbf{b} - \mathbf{a}^T\mathbf{x}}{\|\mathbf{a}\|_2^2}\mathbf{a} & \text{if } \mathbf{a}^T\mathbf{x} > \mathbf{b}, \\ \mathbf{x} & \text{if } \mathbf{a}^T\mathbf{x} \leq \mathbf{b}. \end{cases}$$

4

[4] http://www.seas.ucla.edu/~vandenbe/236C/lectures/proxop.pdf

Examples of Projection on a Closed Convex Set (cont'd)

- For $C = \{\mathbf{x} \in \mathbb{R}^n \mid \ell \preceq x \preceq \mathbf{u}\}$, then

$$P_C(\mathbf{x}) = \begin{cases} \ell_k & \text{when} \quad x_k \leq \ell_k, \\ x_k & \text{when} \quad \ell_k \leq x_k \leq u_k, \\ u_k & \text{when} \quad x_k \geq u_k. \end{cases}$$

- For $C = \mathbb{R}^n_+$, then $P_C(\mathbf{x}) = ReLU(\mathbf{x}) \in C$.

- For $C = \{\mathbf{x} \in \mathbb{R}^n \mid 1^T\mathbf{x} = 1, \mathbf{x} \succeq 0\}$, then $P_C(\mathbf{x}) = (\mathbf{x} - \lambda\mathbf{1})_+$, where λ is the solution of the equation

$$\mathbf{1}^T(\mathbf{x} - \lambda\mathbf{1})_+ = \sum_{i=1}^{n} \max\{0, x_k - \lambda\} = 1.$$

5

[5] http://www.seas.ucla.edu/~vandenbe/236C/lectures/proxop.pdf

Minimizing the Sum of Two Convex Functions

- Consider the following nonsmooth convex optimization problem:

$$\min\{F(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^n\}$$

  where

  - $f : \mathbb{R}^n \to \mathbb{R}$ is closed proper convex, continuously differentiable with Lipschitz continous gradient L_f:

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \leq L_f \|\mathbf{x} - \mathbf{y}\|_2.$$

  - $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is closed proper convex, continuous function which is possibly nonsmooth, with inexpensive proximal operator $\text{prox}_g(\cdot)$.
  - The optimization problem is solvable, i.e., $\text{argmin}\, F \neq \emptyset$.

- In the remaining slides, we assume the functions f and g have those properties, unless stated otherwise.

Minimizing the Sum of Two Convex Functions (cont'd)

- Example: The ℓ_1- regularization problem (QP_λ), where

$$f(\mathbf{z}) = \frac{1}{2}\|A\mathbf{z} - \mathbf{y}\|^2, \quad g(\mathbf{z}) = \lambda\|\mathbf{z}\|_1, \quad \text{and } L_f = \lambda_{\max}(A^T A).$$

Outline

FISTA: A fast iterative shrinkage-thresholding algorithm

- FISTA is an iterative shrinkage-thresholding algorithm (ISTA) with complexity result of $\mathcal{O}(1/k^2)$ (see Theorem 4.4 in [6]) to solve

$$\min\{F(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^n\}.$$

FISTA with constant stepsize - $\min_{\mathbf{x}\in\mathbb{R}^n}(f(\mathbf{x}) + g(\mathbf{x}))$

Input: $L = L_f$, a Lipschitz constant of ∇f, and final step K.

Step 0. $\mathbf{y}_1 = \mathbf{x}_0 \in \mathbb{R}^n$, $t_1 = 1$.

Step k. $(k \geq 1)$ Compute

$$\begin{aligned}
\mathbf{x}_k &= \underset{\mathbf{x}}{\operatorname{argmin}} \left\{ g(\mathbf{x}) + \frac{L}{2} \left\| \mathbf{x} - \left(\mathbf{y}_k - \frac{1}{L} \nabla f(\mathbf{y}_k) \right) \right\|^2 \right\} \\
&= \operatorname{prox}_{(1/L)g}\left(\mathbf{y}_k - \frac{1}{L}\nabla f(\mathbf{y}_k)\right),
\end{aligned}$$

$$t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2},$$

$$\mathbf{y}_{k+1} = \mathbf{x}_k + \frac{t_k - 1}{t_{k+1}}(\mathbf{x}_k - \mathbf{x}_{k-1}).$$

Output: $\mathbf{x}_K$

[6] *A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems*, by Beck & Teboulle, SIAM J. Imaging Sciences, 2009.

FISTA (cont'd)

FISTA - **Another Version** - $\min_{\mathbf{x}\in\mathbb{R}^n}(f(\mathbf{x}) + g(\mathbf{x}))$

Input: $L = L_f$, a Lipschitz constant of ∇f, and final step K.

Step 0. Choose any $\mathbf{x}_1 = \mathbf{x}_0 \in \mathbb{R}^n$.

Step k. $(k \geq 1)$ Compute

$$\mathbf{y} = \mathbf{x}_k + \frac{k-1}{k+2}(\mathbf{x}_k - \mathbf{x}_{k-1})$$

$$\mathbf{x}_{k+1} = \text{prox}_{t_{k+1}g}\Big(\mathbf{y} - t_{k+1}\nabla f(\mathbf{y})\Big),$$

where step size $t_k = \dfrac{1}{L}, \forall k$ or is determined by line search.

Output: $\mathbf{x}_K$

[7]

[7] http://www.seas.ucla.edu/~vandenbe/236C/lectures/fista.pdf, by Boyd & Vandenberghe

FISTA: Examples

Using FISTA to solve

$$\min_{\mathbf{x}\in\mathbb{R}^n}\left(\frac{1}{2}\|A\mathbf{x}-\mathbf{b}\|_2^2+\lambda\|\mathbf{x}\|_1\right).$$

In this case, $f(\mathbf{x})=\frac{1}{2}\|A\mathbf{x}-\mathbf{b}\|_2^2$, $L_f=\lambda_{\max}(A^TA)$, and $g(\mathbf{x})=\lambda\|\mathbf{x}\|_1$.

FISTA to Solve - $\min_{\mathbf{x}\in\mathbb{R}^n}\left(\frac{1}{2}\|A\mathbf{x}-\mathbf{b}\|_2^2+\lambda\|\mathbf{x}\|_1\right)$

Input: Final step K.

Step 0. Choose any $\mathbf{x}_1=\mathbf{x}_0\in\mathbb{R}^n$.

Step k. $(k\geq 1)$ Compute

$$\mathbf{y}=\mathbf{x}_k+\frac{k-1}{k+2}(\mathbf{x}_k-\mathbf{x}_{k-1})$$

$$\hat{\mathbf{y}}=\mathbf{y}-\frac{1}{L}A^T(A\mathbf{y}-\mathbf{b})$$

$$\mathbf{x}_{k+1}=\text{prox}_{(1/L)g}(\hat{\mathbf{y}})=\text{sign}(\hat{\mathbf{y}})\max(|\hat{\mathbf{y}}|-\frac{\lambda}{L},0),$$

where $L=\lambda_{\max}(A^TA)$.

Output: $\mathbf{x}_K$

Nesterov's Second Method

- Nesterov's second method is a gradient projection method with $(1/k^2)$ convergence rate.

Nesterov's Second Method
Input: L = L(f), a Lipschitz constant of ∇f, and final step K.
Step 0. Choose any $\mathbf{x}_0 = \mathbf{z}_0 \in \mathbb{R}^n$.
Step k. $(k \geq 1)$ Compute

$$\mathbf{y} = (1-\theta_k)\mathbf{x}_{k-1} + \theta_k \mathbf{z}_{k-1}$$

$$\mathbf{z}_k = \mathrm{prox}_{(t_k/\theta_k)g}\left(\mathbf{z}_{k-1} - \frac{t_k}{\theta_k}\nabla f(\mathbf{y})\right)$$

$$\mathbf{x}_k = (1-\theta_k)\mathbf{x}_{k-1} + \theta_k \mathbf{z}_k,$$

where $\theta_k = \frac{2}{k+1}$ and $t_k = \frac{1}{L}$, or one of the line search methods.
Output: $\mathbf{x}_K$

8 9

[8] http://www.seas.ucla.edu/~vandenbe/236C/lectures/fista.pdf by Boyd & Vandenberghe
[9] *On Accelerated Proximal Gradient Methods for Convex-Concave Optimization*, by Tseng, 2008.

spgl1 and Other Available Packages

- spgl1[10]. Matlab and Python codes can be downloaded from `https://friedlander.io/spgl1/install`
- Python packages: scikit-learn package.
  - Link: `https://scikit-learn.org/stable/modules/linear_model.html`
  - Solve the (QP_λ) by coordinate descent method [11].

[10] *SPGL1: A solver for large-scale sparse reconstruction*, by Den Berg and Friedlander, 2007.
[11] *Regularization Path For Generalized linear Models by Coordinate Descent*, by Friedman, Hastie and Tibshirani.

Alternating Direction Method of Multipliers (ADMM)

Here we assume that f and g are convex, closed, proper and L_0 has a saddle point.

- Consider the following optimization problem:

$$\begin{aligned} \text{minimize} \quad & f(\mathbf{x}) + g(\mathbf{z}) \\ \text{subject to} \quad & A\mathbf{x} + B\mathbf{z} = \mathbf{c}. \end{aligned}$$

- The corresponding augmented Lagrangian is

$$L_\rho(\mathbf{x}, \mathbf{z}, \mathbf{y}) = f(\mathbf{x}) + g(\mathbf{z}) + \mathbf{y}^T(A\mathbf{x} + B\mathbf{z} - \mathbf{c}) + \frac{\rho}{2}\|A\mathbf{x} + B\mathbf{z} - \mathbf{c}\|_2^2$$

- ADMM algorithm:

$$\mathbf{x}_{k+1} := \underset{x}{\operatorname{argmin}}\, L_\rho(\mathbf{x}, \mathbf{z}_k, \mathbf{y}_k) \qquad \text{(x-minimization)}$$

$$\mathbf{z}_{k+1} := \underset{z}{\operatorname{argmin}}\, L_\rho(\mathbf{x}_{k+1}, \mathbf{z}, \mathbf{y}_k) \qquad \text{(z-minimization)}$$

$$\mathbf{y}_{k+1} := y_k + \rho(A\mathbf{x}_{k+1} + B\mathbf{z}_{k+1} - \mathbf{c}) \qquad \text{(dual update)}$$

[12] [13] [14]

[12] ADMM is proposed by Gabay, Mercier, Glowinski, Marrocco in 1976.

[13] *The Split Bregman Method for L1-Regularized Problems*, by Goldstein and Osher, SIAM J. Imaging Sciences, 2009.

[14] https://web.stanford.edu/class/ee364b/lectures/admm_slides.pdf

ADMM and Related Algorithms

- Under the stated assumptions, ADMM converges in the sense that
  - Iterates approach feasibility: $A\mathbf{x}_k + B\mathbf{z}_k - \mathbf{c} \to 0$
  - Objective approaches optimal value: $f(\mathbf{x}_k) + g(\mathbf{z}_k) \to p_*$
- Related algorithms:
  - operator splitting methods
  - proximal point algorithm
  - Bregman iterative methods

ADMM: Examples

Example 1: Consider ADMM for

$$\min f(\mathbf{x}) \quad \text{subject to} \quad \mathbf{x} \in \mathbb{C}.$$

Answer:

- The ADMM form with $g(\mathbf{z}) = I_C(\mathbf{z})$, the indicator function of set C:

$$\begin{aligned} \text{minimize} \quad & f(\mathbf{x}) + g(\mathbf{z}) \\ \text{subject to} \quad & \mathbf{x} - \mathbf{z} = 0. \end{aligned}$$

- ADMM algorithm (discuss in class)

ADMM: Examples

Example 2: Consider ADMM for

$$\min \frac{1}{2}\|A\mathbf{x} - \mathbf{b}\|_2^2 + \lambda\|\mathbf{x}\|_1.$$

Answer:

- The ADMM form with $g(\mathbf{z}) = \lambda\|\mathbf{z}\|_1$:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|A\mathbf{x} - \mathbf{b}\|_2^2 + \lambda\|\mathbf{z}\|_1 \\ \text{subject to} \quad & \mathbf{x} - \mathbf{z} = 0. \end{aligned}$$

- ADMM algorithm (discuss in class)

ADMM: Examples

Example 3: Consider ADMM for

$$\min \frac{1}{2}\|A\mathbf{x} - \mathbf{b}\|_2^2 + \lambda\|C\mathbf{x} - \mathbf{d}\|_1.$$

Answer:

- The ADMM form with $g(\mathbf{z}) = \lambda\|\mathbf{z} - \mathbf{d}\|_1$:

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}\|A\mathbf{x} - \mathbf{b}\|_2^2 + \lambda\|\mathbf{z} - \mathbf{d}\|_1 \\ &\text{subject to} \quad C\mathbf{x} - \mathbf{z} = 0. \end{aligned}$$

- ADMM algorithm (discuss in class)

ADMM: Examples

Example 4: Given a 2D noisy image f, consider ADMM for the TV denoising model:

$$\min_u \frac{\mu}{2}\|u - f\|_2^2 + \|\nabla_x u\|_1 + \|\nabla_y u\|_1.$$

Answer:

- The ADMM form:

$$\begin{aligned} \min_{u, d_x, d_y} \quad & \frac{\mu}{2}\|u - f\|_2^2 + \|d_x\|_1 + \|d_y\|_1 \\ \text{subject to} \quad & d_x - \nabla_x u = 0 \quad \text{and} \quad d_y - \nabla_y u = 0. \end{aligned}$$

- ADMM algorithm (discuss in class)

Primal-Dual Algorithm - TO BE EDITED

Given $A \in \mathbb{C}^{m \times N}$, the functions $f : \mathbb{C}^m \to (-\infty, \infty]$ and $g : \mathbb{C}^N \to (-\infty, \infty]$ are extended real-valued lower semicontinuous convex functions. Consider:

$$\min_{\mathbf{x} \in \mathbb{C}^N} f(A\mathbf{x}) + g(\mathbf{x})$$

ℓ_1-Algorithms for (QP_λ) Problem (cont'd)

Remarks:

- Global rate of convergence $\mathcal{O}(1/k^2)$ can be achieved, for example, with FISTA and Nesterov's 2nd method. [15]
- The speed of some algorithms for ℓ_1-minimization problems does not depend on the sparsity level s, such as the primal-dual algorithm $\rightarrow$ Use ℓ_1-minimization solvers for mildly large s.
- Debiasing technique: Suppose z_{sol} is the num. soln. of the (QP_λ) problem. Let $S := \mathrm{supp}(z_{final})$ and solve

$$\min\{\|Az - y\|_2^2 : \mathrm{supp}(z) \subset S\}.$$

[15] http://www.seas.ucla.edu/~vandenbe/236C/lectures/fista.pdf