MDS 6106: Introduction to Optimization

Proximal Gradient Method and Alternating Minimization

Lecture 14

December 24th

Andre Milzarek

SDS / CUHK-SZ

Repetition & Agenda

We considered constrained optimization problems of the form:

$$\min \; f(x) \quad \text{s.t.} \quad x \in X, \tag{1}$$

where $X \subset \mathbb{R}^n$ is a closed, convex, and nonempty set.

Projected Gradient Method:

- Idea: Project the gradient steps $x^k - \lambda_k \nabla f(x^k)$ back onto X to guarantee feasibility.
- The projection is given by $\mathcal{P}_X(x) = \arg\min_{y \in X} \, \frac{1}{2}\|y - x\|^2$.

Algorithmic Components:

- The vector x^* is a stationary point of (1) if and only if

$$x^* - \mathcal{P}_X(x^* - \lambda \nabla f(x^*)) = 0 \quad \text{for any } \lambda > 0.$$

- $d^k := \mathcal{P}_X(x^k - \lambda_k \nabla f(x^k)) - x^k$ is a descent direction for f.

$\rightsquigarrow$ We can apply backtracking and set $x^{k+1} = x^k + \alpha_k d^k$.

Projected Gradient Method

1. Initialization: Choose an initial point $x^0 \in X$ and $\sigma, \gamma \in (0,1)$.

For $k = 0, 1, \ldots$:

2. Select $\lambda_k > 0$ and compute $\nabla f(x^k)$ and the new direction $d^k = \mathcal{P}_X(x^k - \lambda_k \nabla f(x^k)) - x^k$.
3. If $\|d^k\| \leq \lambda_k \varepsilon$, then STOP and x^k is the output.
4. Choose a maximal step size $\alpha_k \in \{1, \sigma, \sigma^2, \ldots\} \subset (0,1]$ that satisfies the Armijo condition

$$f(x^k + \alpha_k d^k) - f(x^k) \leq \gamma \alpha_k \cdot \nabla f(x^k)^\top d^k.$$

5. Set $x^{k+1} = x^k + \alpha_k d^k$.

Agenda & Announcements

Logistics:

- The fifth (smaller) exercise sheet is due on Monday, December 28th, 11:00 pm.
- The deadline for the submission of the project report is Wednesday, December 29th, 12:00 pm.
- The presentations will take place on Thursday, December 30th.

Agenda:

- The proximal gradient method.
- Proximal calculus.
- Alternating direction method of multiplier.

The Proximal Gradient Method

Let us consider the nonsmooth optimization problem:

$$\min_{x} \ \psi(x) = f(x) + \varphi(x) \quad \text{s.t.} \quad x \in \mathbb{R}^n.$$

- $\varphi : \mathbb{R}^n \to \mathbb{R}$ (or $\varphi : \mathbb{R}^n \to (-\infty, \infty]$) is a convex (nonsmooth) function.

Connection to Constrained Problems:

- Let us define the indicator function:

$$\iota_X : \mathbb{R}^n \to (-\infty, +\infty], \quad \iota_X(x) := \begin{cases} 0 & \text{if } x \in X, \\ +\infty & \text{if } x \notin X. \end{cases}$$

  Then, we can write

$$\min_{x \in X} \ f(x) \quad \equiv \quad \min_{x \in \mathbb{R}^n} \ f(x) + \iota_X(x).$$

- Basic Idea: Replace ι_X by a general convex mapping φ.

$\rightsquigarrow$ Transfer techniques and strategies!

Convex Analysis: A Quick Introduction

Let us first assume that φ is differentiable, i.e., we have

$$\varphi(y) - \varphi(x) \geq \nabla\varphi(x)^\top(y - x), \quad \forall\ y \in \mathbb{R}^n.$$

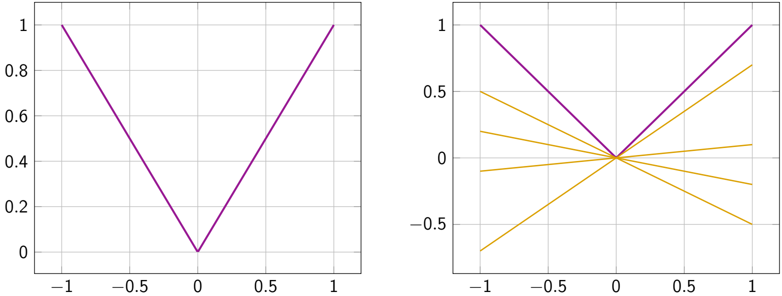
- The tangent $y \mapsto \varphi(x) + \nabla\varphi(x)^\top(y - x)$ supports φ at x from below.
- Generally many such supporting functions might exist!
- The subdifferential of φ is defined as the collection of the subgradients of these supporting functions.

The Convex Subdifferential

The subdifferential of φ at x is the set

$$\partial\varphi(x) := \{g \in \mathbb{R}^n : \varphi(y) - \varphi(x) \geq g^\top(y - x),\ \forall\ y \in \mathbb{R}^n\}.$$

The elements $g \in \partial\varphi(x)$ are called subgradients of φ at x.



(a) Plot of $\varphi(x) = |x|$ (b) Supporting tangents at $x = 0$

Illustration:

- The absolute value $\varphi(x) = |x|$ is differentiable for $x \neq 0$ and we have $\partial\varphi(x) = \{+1\}$ if $x > 0$ and $\partial\varphi(x) = \{-1\}$ if $x < 0$. In the case $x = 0$, we obtain $\partial\varphi(0) = [-1, 1]$.

Chain Rule for Subdifferentials

Let $f : \mathbb{R}^n \to \mathbb{R}$ and $\varphi : \mathbb{R}^m \to \mathbb{R}$ be convex and let $A \in \mathbb{R}^{m \times n}$ be given. Set $\psi(x) := f(x) + \varphi(Ax)$. Then, it holds

$$\partial\psi(x) = \partial f(x) + A^\top \partial\varphi(Ax), \quad \forall\, x \in \mathbb{R}^n.$$

- Next, we present a connection between classical derivatives and subgradients.

Subdifferentiability and Differentiability

Let $\varphi : \mathbb{R}^n \to \mathbb{R}$ be convex and let $x \in \mathbb{R}^n$ be given.

- Suppose that φ is (Fréchet) differentiable at x. Then, we have $\partial\varphi(x) = \{\nabla\varphi(x)\}$.

Example: I

Calculate the subdifferential of the following mapping:

$$\varphi(x) = \|x\|_2.$$

Example: II

Calculate the subdifferential of the following mapping:

$$\varphi(x) = \max\{0, x\}.$$

First-Order Optimality and the Proximity Operator

First-Order Optimality Conditions

Let f be cont. diff. and let $\varphi : \mathbb{R}^n \to \mathbb{R}$ be convex. Suppose that x^* is a local minimizer of $\min_x \, \psi(x)$, then:

$$\nabla f(x^*)^\top (x - x^*) + \varphi(x) - \varphi(x^*) \geq 0 \quad \forall \, x \in \mathbb{R}^n.$$

Remarks:

- If f is convex, then x^* is a global sol. iff the latter cond. holds.
- A point x^* with $\nabla f(x^*)^\top (x - x^*) + \varphi(x) - \varphi(x^*) \geq 0$ for all x is again called stationary point.
- This condition is equivalent to $-\nabla f(x^*) \in \partial \varphi(x^*)$.

$\rightsquigarrow$ We can now generalize the projection $\mathcal{P}_X$ in a similar way!

The Proximity Operator

- For every $x \in \mathbb{R}^n$ and $\lambda > 0$, the optimization problem

$$\min_y \ \varphi(y) + \frac{1}{2\lambda}\|x - y\|^2,$$

has a unique global sol. x^*. This minimizer is called the proximity operator of φ at x and we write $x^* = \mathrm{prox}_{\lambda\varphi}(x)$.

- $\mathrm{prox}_{\lambda\varphi} : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz cont. with constant $L = 1$.
- Let $f : \mathbb{R}^n \to \mathbb{R}$ be C^1. Then, x^* is a stationary point iff

$$F_\lambda(x^*) = x^* - \mathrm{prox}_{\lambda\varphi}(x^* - \lambda\nabla f(x^*)) = 0 \quad \text{for any } \lambda > 0.$$

The proximity operator can be characterized via:

$$p = \mathrm{prox}_{\lambda\varphi}(x) \quad \iff \quad 0 \in \partial\varphi(p) + \frac{1}{\lambda}(p - x).$$

Proximity Operator: Examples

Indicator Functions:

- Let $X \subseteq \mathbb{R}^n$ be a convex, closed, nonempty set. Then, we have:

$$\mathrm{prox}_{\lambda \iota_X}(x) = \mathcal{P}_X(x) \quad \forall\ \lambda > 0.$$

ℓ_1-Norm:

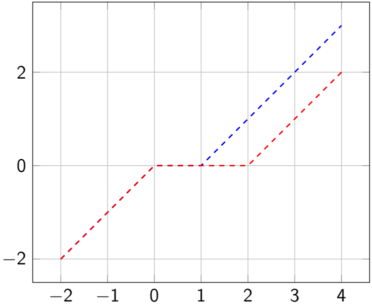
- Set $\varphi(x) = \mu\|x\|_1$. We have:

$$[\mathrm{prox}_{\lambda\varphi}(x)]_i = \mathrm{prox}_{\lambda\mu|\cdot|}(x_i) = \begin{cases} x_i - \lambda\mu & \text{if } x_i > \lambda\mu, \\ 0 & \text{if } x_i \in [-\lambda\mu, \lambda\mu], \\ x_i + \lambda\mu & \text{if } x_i < -\lambda\mu. \end{cases}$$

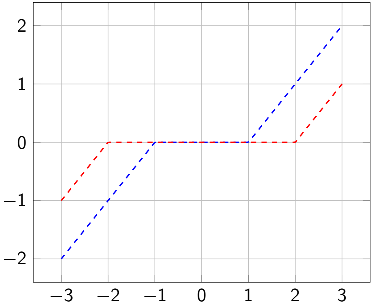
Maximum-Function:

- Set $\varphi(x) = \max\{0, x\}$, $x \in \mathbb{R}$. It holds that:

$$\mathrm{prox}_{\lambda\varphi}(x) = \begin{cases} x - \lambda & \text{if } x > \lambda, \\ 0 & \text{if } x \in [0, \lambda], \\ x & \text{if } x < 0, \end{cases} \quad \text{for all} \quad \lambda > 0.$$



(c) $\mathrm{prox}_{\lambda \max\{0,\cdot\}}(x)$ for $\lambda \in \{1,2\}$



(d) $\mathrm{prox}_{\lambda|\cdot|}$ for $\lambda \in \{1,2\}$

- Plot of the proximity operators $\mathrm{prox}_{\lambda \max\{0,\cdot\}}(x)$ and $\mathrm{prox}_{\lambda|\cdot|}(x)$ for different λ.

ℓ_1-Norm:

- Determine the proximity operator of $\varphi(x) = \mu\|x\|_1$. We have:

ℓ_2-Norm:

- Determine the proximity operator of $\varphi(x) = \mu\|x\|_2$. We have:

The Proximal Gradient Method

Descent Directions for Nonsmooth Problems

Let $x \in \mathbb{R}^n$ and $\lambda > 0$ be given and set $d := -F_\lambda(x)$. Then, we have

$$\Delta := \nabla f(x)^\top d + \varphi(x+d) - \varphi(x) \leq -\frac{1}{\lambda}\|d\|^2.$$

Suppose that x is not a stationary point and choose $\gamma \in (0,1)$. Then, there is $\bar{\alpha} > 0$ such that

$$\psi(x + \alpha d) - \psi(x) \leq \gamma\alpha \cdot \Delta \quad \forall\ \alpha \in [0, \bar{\alpha}].$$

Overall Strategy (As Before):

- $\rightsquigarrow$ Use $d^k = -F_{\lambda_k}(x^k) = \mathrm{prox}_{\lambda_k\varphi}(x^k - \lambda_k \nabla f(x^k)) - x^k$ as a descent direction (with some fixed $\lambda_k > 0$).
- $\rightsquigarrow$ Perform Armijo line-search to find a step size α_k.

Descent Directions for Nonsmooth Problems

Let $x \in \mathbb{R}^n$ and $\lambda > 0$ be given and set $d := -F_\lambda(x)$. Then, we have

$$\Delta := \nabla f(x)^\top d + \varphi(x+d) - \varphi(x) \leq -\frac{1}{\lambda}\|d\|^2.$$

Suppose that x is not a stationary point and choose $\gamma \in (0,1)$. Then, there is $\bar{\alpha} > 0$ such that

$$\psi(x+\alpha d) - \psi(x) \leq \gamma\alpha \cdot \Delta \quad \forall\ \alpha \in [0, \bar{\alpha}].$$

Overall Strategy (As Before):

- $\rightsquigarrow$ Use $d^k = -F_{\lambda_k}(x^k) = \mathrm{prox}_{\lambda_k\varphi}(x^k - \lambda_k \nabla f(x^k)) - x^k$ as a descent direction (with some fixed $\lambda_k > 0$).
- $\rightsquigarrow$ Perform Armijo line-search to find a step size α_k.

The Proximal Gradient Method

Proximal Gradient Method

1. Initialization: Choose an initial point $x^0 \in \mathbb{R}^n$ and $\sigma, \gamma \in (0, 1)$.

For $k = 0, 1, \ldots$:

2. Select $\lambda_k > 0$ and compute $\nabla f(x^k)$ and the new direction $d^k = -F_{\lambda_k}(x^k) = \mathrm{prox}_{\lambda_k \varphi}(x^k - \lambda_k \nabla f(x^k)) - x^k$.
3. If $\|d^k\| \leq \lambda_k \varepsilon$, then STOP and x^k is the output.
4. Choose a maximal step size $\alpha_k \in \{1, \sigma, \sigma^2, \ldots\} \subset (0, 1]$ that satisfies the Armijo condition

$$\psi(x^k + \alpha_k d^k) - \psi(x^k) \leq \gamma \alpha_k \cdot \Delta_k.$$

5. Set $x^{k+1} = x^k + \alpha_k d^k$.

Convergence Guarantees

Let $f : \mathbb{R}^n \to \mathbb{R}$ be C^1 and let $\varphi : \mathbb{R}^n \to \mathbb{R}$ be convex. Let $\{x^k\}_k$ be generated by the PGM and assume that $\{\lambda_k\}_k$ is bounded, i.e.,

$$0 < \underline{\lambda} \leq \lambda_k \leq \overline{\lambda} \quad \forall\ k.$$

Then, we have:

- The function values $\psi(x^k)$, $k \in \mathbb{N}$, decrease and converge to $-\infty$ or some $\psi^* \in \mathbb{R}$.
- Every accumulation point x^* of $\{x^k\}_k$ is a stationary point.

Comments:

- If ∇f is Lipschitz cont. with constant L and $\lambda_k \in (0, \frac{2}{L})$, then we can use:

$$x^{k+1} = \mathrm{prox}_{\lambda_k \varphi}(x^k - \lambda_k \nabla f(x^k)) \tag{2}$$

- If f is also strongly convex, then $\{x^k\}_k$ converges q-linearly to the unique solution x^*.

Convergence Guarantees

Let $f : \mathbb{R}^n \to \mathbb{R}$ be C^1 and let $\varphi : \mathbb{R}^n \to \mathbb{R}$ be convex. Let $\{x^k\}_k$ be generated by the PGM and assume that $\{\lambda_k\}_k$ is bounded, i.e.,

$$0 < \underline{\lambda} \leq \lambda_k \leq \overline{\lambda} \quad \forall\ k.$$

Then, we have:

- The function values $\psi(x^k)$, $k \in \mathbb{N}$, decrease and converge to $-\infty$ or some $\psi^* \in \mathbb{R}$.
- Every accumulation point x^* of $\{x^k\}_k$ is a stationary point.

Comments:

- If ∇f is Lipschitz cont. with constant L and $\lambda_k \in (0, \frac{2}{L})$, then we can use:

$$x^{k+1} = \mathrm{prox}_{\lambda_k \varphi}(x^k - \lambda_k \nabla f(x^k)) \tag{2}$$

- If f is also strongly convex, then $\{x^k\}_k$ converges q-linearly to the unique solution x^*.

The update in (2) can also be written as:

$$\begin{aligned} x^{k+1} &= \mathrm{prox}_{\lambda_k \varphi}(x^k - \lambda_k \nabla f(x^k)) \\ &= \operatorname*{arg\,min}_{y \in \mathbb{R}^n} \ \varphi(y) + \frac{1}{2\lambda_k}\|x^k - \lambda_k \nabla f(x^k) - y\|^2 \\ &= \operatorname*{arg\,min}_{y \in \mathbb{R}^n} \ \varphi(y) + \nabla f(x^k)^\top (y - x^k) + \frac{1}{2\lambda_k}\|y - x^k\|^2. \end{aligned}$$

Hence, the principle idea of PGM can be interpreted as:

- We build a simpler model of $\psi = \varphi + f$ by keeping φ and by using a quadratic approximation

$$f(y) \approx f(x^k) + \nabla f(x^k)^\top (y - x^k) + \frac{1}{2\lambda_k}(y - x^k)^\top (y - x^k)$$

  for the smooth function f.
- The global minimizer of this model is then used to define the next iterate x^{k+1}.

Further Remarks:

- This immediately motivates possible extensions of the form:

$$x^{k+1} = \underset{y \in \mathbb{R}^n}{\arg\min}\ \varphi(y) + \nabla f(x^k)^\top (y - x^k) + \frac{1}{2}(y - x^k)^\top B_k (y - x^k),$$

  where $B_k \in \mathbb{S}^n_{++}$ is a symmetric, positive definite matrix that can either be the Hessian $\nabla^2 f(x^k)$ or a suitable approximation.

In contrast to the simple PGM step such updates typically do not have closed-form expressions.

$\rightsquigarrow$ We need to solve a subproblem to calculate x^{k+1}.

- Methods that are based on this formulation are called proximal Newton methods.

Proximal Calculus

Translation and Scaling

Let $\lambda > 0$ and $x \in \mathbb{R}^n$ be given. We have:

- Define $g(\cdot) := \varphi(\cdot - b)$, $b \in \mathbb{R}^n$. Then, it follows

$$\mathrm{prox}_{\lambda g}(x) = b + \mathrm{prox}_{\lambda\varphi}(x - b).$$

- Define $h(\cdot) := \varphi(\cdot/\beta)$, $\beta \neq 0$. Then, it follows

$$\mathrm{prox}_{\lambda h}(x) = \beta \cdot \mathrm{prox}_{\lambda\varphi/\beta^2}(x/\beta).$$

Separable Functions

Let $\varphi_i : \mathbb{R} \to \mathbb{R}$, $i = 1, \dots, n$, be a family of convex functions and set $\varphi(x) = \sum_{i=1}^n \varphi_i(x_i)$. Then, we have

$$[\mathrm{prox}_{\lambda\varphi}(x)]_i = \mathrm{prox}_{\lambda\varphi_i}(x_i) \quad \forall\ i, \quad \lambda > 0.$$

Composition with a Special Linear Operator

Let $\varphi : \mathbb{R}^m \to \mathbb{R}$ be convex and let $x \in \mathbb{R}^n$, $\lambda > 0$ and $A \in \mathbb{R}^{m \times n}$ be given. Suppose A satisfies $AA^\top = I$.

Setting $g(\cdot) := \varphi(A\cdot)$, it holds that:

$$\mathrm{prox}_{\lambda g}(x) = x - A^\top(Ax - \mathrm{prox}_{\lambda\varphi}(Ax)).$$

Proof: ?

Example:

- Let $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$ be given with $AA^\top = I$. Consider the set $C := \{x : \|Ax - b\| \leq \sigma\}$ and calculate $\mathcal{P}_C$.

Composition with a Special Linear Operator

Let $\varphi : \mathbb{R}^m \to \mathbb{R}$ be convex and let $x \in \mathbb{R}^n$, $\lambda > 0$ and $A \in \mathbb{R}^{m \times n}$ be given. Suppose A satisfies $AA^\top = I$.

Setting $g(\cdot) := \varphi(A\cdot)$, it holds that:

$$\mathrm{prox}_{\lambda g}(x) = x - A^\top(Ax - \mathrm{prox}_{\lambda\varphi}(Ax)).$$

Proof: ?

Example:

- Let $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$ be given with $AA^\top = I$. Consider the set $C := \{x : \|Ax - b\| \le \sigma\}$ and calculate $\mathcal{P}_C$.

The Accelerated Proximal Gradient Method

Let's Accelerate!

It is possible to combine the discussed acceleration techniques and the proximal gradient method under the assumption that f is a convex mapping.

In principle, the acceleration mechanism is identical to the one we already discussed!

- We first perform an extrapolation step

$$y^{k+1} = x^k + \beta_k(x^k - x^{k-1}), \qquad \beta_k > 0$$

  to approximate and extrapolate the next iterate $y^{k+1} \approx x^{k+1}$.

- Afterwards we compute a proximal gradient step based on the predicted information y^{k+1}.

The Accelerated Proximal Gradient Method

Accelerated Proximal Gradient Method

1. Initialization: Choose a point $x^0 \in \mathbb{R}^n$ and set $x^{-1} = x^0$.

For $k = 0, 1, \ldots$:

2. Select an extrapolation parameter β_k and compute the step $y^{k+1} = x^k + \beta_k(x^k - x^{k-1})$.
3. Select $\lambda_k > 0$ and set $x^{k+1} = \mathrm{prox}_{\lambda_k \varphi}(y^{k+1} - \lambda_k \nabla f(y^{k+1}))$.

- For special choices of β_k and $\lambda_k = \bar{\lambda} \in (0, \frac{1}{L}]$, we can show:

$$\psi(x^k) - \psi(x^*) \le \frac{2\|x^0 - x^*\|^2}{\bar{\lambda}(k+1)^2} \quad \forall\ k \in \mathbb{N},$$

where x^* is a solution of the problem $\min_x\ \psi(x)$.

Remarks and Choices of β_k

As in the smooth case, we can utilize the choices $\beta_k = \frac{k-2}{k+1}$ or

$$t_k = \frac{1+\sqrt{1+4t_{k-1}^2}}{2}, \quad \beta_k = \frac{t_{k-1}-1}{t_k}, \quad t_{-1} = t_0 = 1.$$

If the Lipschitz constant L is unknown, then the step size λ_k can be determined by the following line search procedure:

- Choose $\eta \in (0,1)$.
- Set $\lambda_k = \lambda_{k-1}$ and $x^{k+1} = \mathrm{prox}_{\lambda_k\varphi}(y^{k+1} - \lambda_k \nabla f(y^{k+1}))$.
- `while` $f(x^{k+1}) > f(y^{k+1}) + \nabla f(y^{k+1})^\top (x^{k+1} - y^{k+1}) + \frac{\|x^{k+1}-y^{k+1}\|^2}{2\lambda_k}$ `do`:

  Set $\lambda_k = \eta\lambda_k$ and calc. $x^{k+1} = \mathrm{prox}_{\lambda_k\varphi}(y^{k+1} - \lambda_k \nabla f(y^{k+1}))$.

Numerical Experiment: Sparse Reconstruction

We consider the ℓ_1-optimization problem

$$\min_x \ \psi(x) = \frac{1}{2}\|Ax - b\|^2 + \mu\|x\|_1.$$

The data is generated as follows:

- We set $\mathtt{m} = 300$, $\mathtt{n} = 3000$, $\mathtt{s} = 30$ and create an index mask $\mathtt{mask} = \mathtt{randperm(n,s)}$. We generate a sparse signal x^* via

$$x^* = \mathtt{zeros(n,1)}, \quad x^*(\mathtt{mask}) = \mathtt{randn(s,1)}.$$

⇝ x^* has only 30 nonzero randomly chosen components.

- We choose $\mathtt{A} = \mathtt{randn(m,n)}$ and generate the measurement b via $\mathtt{b} = \mathtt{A} \cdot x^* + 0.01 \cdot \mathtt{randn(m,1)}$.

⇝ The goal is to reconstruct the signal x^* from the much smaller measurements b via solving the ℓ_1-problem.

- The Lipschitz constant of ∇f is given by $L = \lambda_{\max}(A^\top A)$.

Numerical Comparison

We consider the accelerated proximal gradient method (APGM) with the following setups:

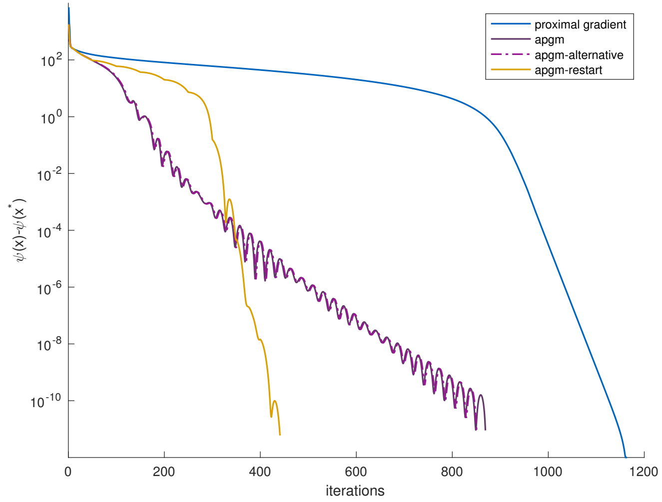
1. We set $\beta_k = (t_{k-1} - 1)t_k^{-1}$, $t_k = 0.5(1 + \sqrt{1 + 4t_{k-1}^2})$, and $\lambda_k \equiv \bar{\lambda} = 1/L$;
2. Same strategy for β_k and λ_k with restart $t_{k-1} = t_k = 1$ after each 50 iterations;
3. Alternative extrapolation strategy: $\beta_k = \frac{k-2}{k+1}$ and $\lambda_k = 1/L$;

We compare APGM with:

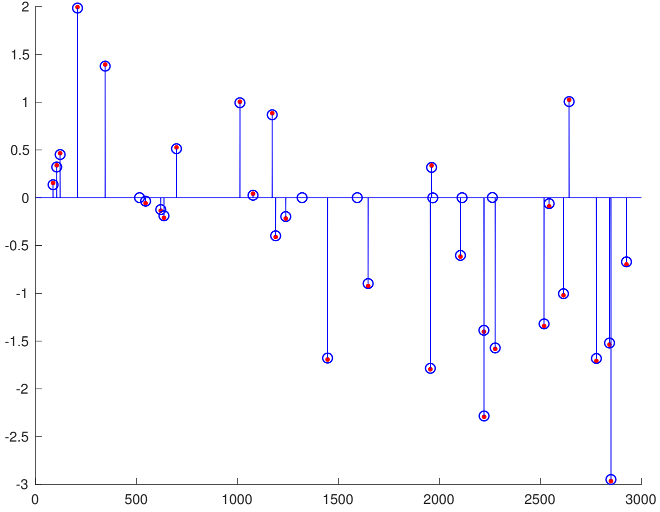
- The basic proximal gradient method with quasi-Armijo line search and $\gamma = 0.1$, $s = 1$, $\sigma = 0.5$.

We use $x^0 = 0$ and $\mu = 5$. The tolerance is set to $\texttt{tol} = 10^{-8}$.

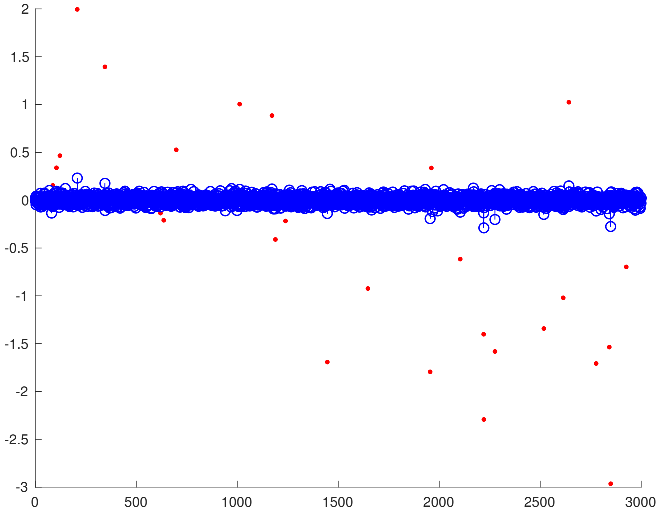
Numerical Results



Comparison: Sparsity Pattern



Comparison: Sparsity Pattern – $\varphi(x) = \frac{\mu}{2}\|x\|^2$



We want to solve: $\min_x \ \psi(x)$

- Identify the structure of the problem. Can ψ be written as

$$\psi = f + \varphi$$

  where f is smooth and φ is convex?

- Is the problem constrained with convex constraints $x \in C$?

Yes $\rightsquigarrow$ Apply Proximal/Projected Gradient Method:

- Expressions for $\mathrm{prox}_{\lambda\varphi}$ or $\mathcal{P}_C$ often exist if φ or C are simple!
- Identify the "simple structure" in φ and C and try to use the proximal calculus.

No $\rightsquigarrow$ ...:

- If $\varphi \equiv 0$ or $C = \mathbb{R}^n$, an unconstrained optimization method can be applied (gradient, Newton, BFGS).

How to Apply the Proximal Gradient Method?

We want to solve: $\min_x \ \psi(x)$

- Identify the structure of the problem. Can ψ be written as

$$\psi = f + \varphi$$

where f is smooth and φ is convex?

- Is the problem constrained with convex constraints $x \in C$?

Yes $\rightsquigarrow$ Apply Proximal/Projected Gradient Method:

- Expressions for $\mathrm{prox}_{\lambda\varphi}$ or $\mathcal{P}_C$ often exist if φ or C are simple!
- Identify the "simple structure" in φ and C and try to use the proximal calculus.

No $\rightsquigarrow$ . . . :

- If $\varphi \equiv 0$ or $C = \mathbb{R}^n$, an unconstrained optimization method can be applied (gradient, Newton, BFGS).

We want to solve: $\min_x \ \psi(x)$

- Identify the structure of the problem. Can ψ be written as

$$\psi = f + \varphi$$

  where f is smooth and φ is convex?
- Is the problem constrained with convex constraints $x \in C$?

Yes $\rightsquigarrow$ Apply Proximal/Projected Gradient Method:

- Expressions for $\mathrm{prox}_{\lambda\varphi}$ or $\mathcal{P}_C$ often exist if φ or C are simple!
- Identify the "simple structure" in φ and C and try to use the proximal calculus.

No $\rightsquigarrow$ . . .:

- If $\varphi \equiv 0$ or $C = \mathbb{R}^n$, an unconstrained optimization method can be applied (gradient, Newton, BFGS).

Bonus: Alternating Direction Method of Multiplier

We now discuss an algorithm that can also be applied to more complicated models.

The starting point for the so-called alternating direction method of multipliers or ADMM are minimization problems of the form:

$$\min_{x \in \mathbb{R}^n} \; f(x) + g(Ax), \tag{3}$$

- A is a given $m \times n$ matrix.
- $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^m \to \mathbb{R}$ are convex functions (or constraint functions: ι_X).

Observation:

$\rightsquigarrow$ Both f and g can be nonsmooth!

We convert this problem to the equivalent constrained problem

$$\min \; f(x) + g(y) \quad x \in \mathbb{R}^n, \quad y \in \mathbb{R}^m, \quad Ax - y = 0.$$

The associated augmented Lagrangian function is given by:

$$L_\sigma(x, y, \lambda) = f(x) + g(y) + \lambda^\top (Ax - y) + \frac{\sigma}{2}\|Ax - y\|^2.$$

The ADMM first minimizes the augmented Lagrangian w.r.t. x, then w.r.t. y, and finally performs a multiplier update:

$$\begin{aligned} x^{k+1} &\in \operatorname*{arg\,min}_{x \in \mathbb{R}^n} \; L_\sigma(x, y^k, \lambda^k) \\ y^{k+1} &\in \operatorname*{arg\,min}_{y \in \mathbb{R}^m} \; L_\sigma(x^{k+1}, y, \lambda^k) \\ \lambda^{k+1} &= \lambda^k + \gamma\sigma(Ax^{k+1} - y^{k+1}), \end{aligned}$$

where $\gamma \in (0, (1+\sqrt{5})/2)$.

The minimization with respect to y can also be written more compactly

$$\begin{aligned} y^{k+1} &= \operatorname*{arg\,min}_{y \in \mathbb{R}^m} \; g(y) + (\lambda^k)^\top (Ax^{k+1} - y) + \frac{\sigma}{2}\|Ax^{k+1} - y\|^2 \\ &= \mathrm{prox}_{g/\sigma}(Ax^{k+1} + \sigma^{-1}\lambda^k). \end{aligned}$$

The penalty parameter σ is typically kept constant in ADMM.

Observations:

- Very simple procedure that performs separate minimization w.r.t. x and y.

$\rightsquigarrow$ Key Modeling Technique: (Re-)Formulate the problem such that the subproblems are easy to solve ($\rightsquigarrow$ proximal calculus).

Application: Support Vector Machines

Example: Sparse Recovery

We consider the ℓ_1-optimization problem

$$\min_x \; \frac{1}{2}\|Ax - b\|^2 + \mu\|x\|_1,$$

where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, and $\mu > 0$ are given.

Application: Support Vector Machines

Given: m data points $a_1, a_2, ..., a_m \in \mathbb{R}^n$ with labels $b_i \in \{-1, 1\}$.

Task: Find a hyperplane $\ell(a) := a^\top x + y$ defined by $(x, y) \in \mathbb{R}^{n+1}$ separating the datapoints such that:

$$b_i = \begin{cases} +1 & \text{if } \ell(a_i) > 0, \\ -1 & \text{if } \ell(a_i) \leq 0. \end{cases}$$

We consider the SVM-model:

$$\min_{x,y} \quad \frac{\lambda}{2}\|x\|^2 + \sum\nolimits_{i=1}^m \max\{0, 1 - b_i(a_i^\top x + y)\}, \quad \lambda > 0.$$

- We want to apply ADMM to solve this problem.

Semi-Proximal Alternating Direction Method of Multiplier

We now analyze a more general version of ADMM.

- We add a quadratic proximity term to the objective function of each subproblem in ADMM.
- Let $S \in \mathbb{S}^n_+$, $T \in \mathbb{S}^m_+$ be given. We set $\|x\|_S^2 = x^\top Sx$ and $\|y\|_T^2 = y^\top Ty$.

We now consider the so-called semi-proximal ADMM:

sp-ADMM

1. Initialization: Choose an initial points $x^0 \in \mathbb{R}^n$, $y^0, \lambda^0 \in \mathbb{R}^m$, and $\sigma > 0$, $\gamma \in (0, (1+\sqrt{5})/2)$.

Perform the following updates:

2. $x^{k+1} \in \arg\min_{x\in\mathbb{R}^n} \; f(x) + \frac{\sigma}{2}\|Ax - y^k + \sigma^{-1}\lambda^k\|^2 + \frac{1}{2}\|x - x^k\|_S^2$.
3. $y^{k+1} \in \arg\min_{y\in\mathbb{R}^m} \; g(y) + \frac{\sigma}{2}\|Ax^{k+1} - y + \sigma^{-1}\lambda^k\|^2 + \frac{1}{2}\|y - y^k\|_T^2$.
4. $\lambda^{k+1} = \lambda^k + \gamma\sigma(Ax^{k+1} - y^{k+1})$.

Important Observation:

- The minimization in step 2 can be considerably simplified by choosing $S = \tau I - \sigma A^\top A$ with $\tau \geq \sigma\lambda_{\max}(A^\top A)$.

Then, we have $S \in \mathbb{S}^+_n$ and the x-step can be simplified as follows:

$$f(x) + \frac{\sigma}{2}\|Ax - y^k + \sigma^{-1}\lambda^k\|^2 + \frac{1}{2}\|x - x^k\|^2_S \ = \ ...$$

The Linearized Semi-Proximal ADMM

Thus, with this choice of S and setting $T = 0$, step 2 and 3 in sp-ADMM can be expressed explicitly via:

$$x^{k+1} = \mathrm{prox}_{f/\tau}(x^k - \frac{\sigma}{\tau} A^\top [Ax^k - y^k + \sigma^{-1}\lambda^k])$$
$$y^{k+1} = \mathrm{prox}_{g/\sigma}(Ax^{k+1} + \sigma^{-1}\lambda^k).$$

Remark:

- This special version of ADMM is called linearized semi-proximal linearized ADMM.

$\rightsquigarrow$ In the updates in step 2 and 3, we actually linearize the original quadratic terms and add a quadratic proximity term.

Merry Christmas!