Math 273a: Optimization
Proximal Operator and Proximal-Point Algorithm

Wotao Yin
Department of Mathematics, UCLA

online discussions on piazza.com
Outline

- Concept
- Definition
- Examples
- Optimization and operator-theoretic properties
- Algorithm
- Interpretations
Why proximal method?

- **Newton's method:**
  - for $C^2$-smooth, unconstrained problems
  - allow modest size

- **Gradient method:**
  - for $C^1$-smooth, unconstrained problems
  - give large size and sometimes distributed implementations

- **Proximal method:**
  - for smooth and non-smooth, constrained and unconstrained
  - but for structured problems
  - gives large size and distributed implementations
Why proximal method?

- **Newton’s method**
  - uses **low-level** (explicit) operation: \( x^{k+1} = x^k - \lambda H^{-1}(x^k) \nabla f(x^k) \).

- **Gradient method**
  - uses **low-level** (explicit) operation: \( x^{k+1} = x^k - \lambda \nabla f(x^k) \).

- **Proximal methods**
  - uses **high-level** (implicit) operation: \( x^{k+1} = \text{prox}_{\lambda f}(x^k) \).
  - \( \text{prox}_{\lambda f} \) is an **optimization** problem
  - only simple for structured \( f \), but there are many of them.
Proximal operator

Assumptions and Notation

1. $f : \mathbb{R}^n \rightarrow \mathbb{R}^n \cup \{\infty\}$ is a closed, proper, convex function

2. $f$ is proper if $\text{dom} f \neq \emptyset$

3. $f$ is closed if its graph is a closed set.

   $\Rightarrow \text{prox}_\lambda f$ is well-defined and unique for $\lambda > 0$

4. the raised $^*$ (e.g. $x^*$) is a global minimizer of some function.

5. $\text{dom} f$ is the domain of $f$, which is where $f(x)$ is finite.
The proximal operator $\text{prox}_f : \mathbb{R}^n \to \mathbb{R}^n$ of a function $f$ is defined by:

$$\text{prox}_f(v) = \arg \min_{x \in \mathbb{R}^n} \left( f(x) + \frac{1}{2} \| x - v \|^2 \right)$$

The scaled proximal operator $\text{prox}_{\lambda f} : \mathbb{R}^n \to \mathbb{R}^n$ is defined by:

$$\text{prox}_{\lambda f}(v) = \arg \min_{x \in \mathbb{R}^n} \left( f(x) + \frac{1}{2\lambda} \| x - v \|^2 \right)$$
Special case: Projection

- Consider a closed convex set $C \neq \emptyset$
- Let $\iota_C$ be the **indicator function** of $C$: $\iota_C(x) = 0$ if $x \in C$; $\infty$ otherwise.

$$\text{prox}_{\iota_C}(x) = \arg \min_y \left( \iota_C(y) + \frac{1}{2} \| y - x \|^2 \right) = \arg \min_{y \in C} \frac{1}{2} \| y - x \|^2 =: P_C(x)$$

- By generalizing $\iota_C$ to $f$, we generalize $P_C$ to $\text{prox}_f$. 
Proximal “step size”

\[ \text{prox}_{\lambda f}(v) = \arg \min_{x \in \mathbb{R}^n} \left( f(x) + \frac{1}{2\lambda} \| x - v \|^2 \right) \]

- \( \lambda > 0 \) is the “step size”:
  - \( \lambda \uparrow \infty \implies \text{prox}_{\lambda f}(v) \to \arg \min_{x \in \mathbb{R}^n} f(x) \)
    (In case there are multiple solutions, pick the one closest to \( v \))
  - \( \lambda \downarrow 0 \implies \text{prox}_{\lambda f}(v) \to P_{\text{dom} f}(v) \)

\[ P_{\text{dom} f}(v) = \arg \min_{x \in \mathbb{R}^n} \left\{ \frac{1}{2} \| v - x \|^2 : f(x) \text{ is finite} \right\} \]

- Given \( v \), (\( \text{prox}_{\lambda f}(v) - v \)) is not linear in \( \lambda \), so \( \lambda \) has the function like a step size is not a step size
**Examples**

- **Linear function:** Let \( a \in \mathbb{R}^n, \ b \in \mathbb{R} \) and

  \[
  f(x) = a^T x + b.
  \]

  Then,

  \[
  \text{prox}_{\lambda f}(v) = \arg \min_{x \in \mathbb{R}^n} \left( (a^T x + b) + \frac{1}{2\lambda} \|x - v\|^2 \right)
  \]

  has first-order optimality conditions:

  \[
  a + \frac{1}{\lambda} (\text{prox}_{\lambda f}(v) - v) = 0 \iff \text{prox}_{\lambda f}(v) = v - \lambda a
  \]

- **Application:** proximal operator of linear approximation of \( f \)

  - let \( f^{(1)}(x) = f(x^0) + \langle \nabla f(x^0), x - x^0 \rangle \)

  - then, \( \text{prox}_{\lambda f^{(1)}}(x^0) = x^0 - \lambda \nabla f(x^0) \) is a gradient step with size \( \lambda \)
Examples

- **Quadratic function** Let $A \in \mathbb{S}_+^n$ be a symmetric positive semi-definite matrix, $b \in \mathbb{R}^n$, and

\[
f(x) = \frac{1}{2} x^T A x - b^T x + c.
\]

The proximal operator

\[
\text{prox}_{\lambda f}(v) = \arg \min_{x \in \mathbb{R}^n} \left( f(x) + \frac{1}{2\lambda} \|x - v\|^2 \right)
\]

has first order optimality conditions:

\[
(Av^* - b) + \frac{1}{\lambda} (v^* - v) = 0 \iff v^* = (\lambda A + I)^{-1} (\lambda b + v)
\]

\[
\iff v^* = (\lambda A + I)^{-1} (\lambda b + \lambda Av + v - \lambda Av)
\]

\[
\iff v^* = v + (A + \frac{1}{\lambda} I)^{-1} (b - Av)
\]

It gives a **iterative refinement method** for least squares problems.
\[
\text{prox}_\lambda f(v) = v + (A + \frac{1}{\lambda} I)^{-1}(b - Av)
\]

- **Application:** *proximal operator of quadratic approximation of \( f \)

  - let \( f^{(2)}(x) = f(x^0) + \langle \nabla f(x^0), x - x^0 \rangle + \frac{1}{2}(x - x^0)^T \nabla^2 f(x^0)(x - x^0) \)

  \[
  = \frac{1}{2} x^T Ax - b^T x + c
  \]

  where

  - \( A = \nabla^2 f(x^0) \)
  - \( b = (\nabla^2 f(x^0))^T x^0 - \nabla f(x) \)

  - by letting \( v = x^0 \), we get

  \[
  \text{prox}_\lambda f^{(2)}(x^0) = x^0 - (\nabla^2 f(x^0) + \frac{1}{\lambda} I)^{-1}\nabla f(x^0)
  \]

- **modified-Hessian Newton** update, **Levenberg-Marquardt** update
Examples

- **$\ell_1$-norm**: $x \in \mathbb{R}^n$, let $f(x) = \|x\|_1$, then

  $$\text{prox}_{\lambda f} = \text{sign}(x) \cdot \max(|x| - \lambda, 0) = x - P_{[-\lambda, \lambda]^n} x.$$  

  The operator is often written as $\text{shrink}(x, \lambda)$

- **$\ell_2$-norm**: let $f(x) = \|x\|_2$, then

  $$\text{prox}_{\lambda f} = \text{shrink}_{\| . \|}(x, \lambda) = \max(\|x\| - \lambda, 0) \frac{x}{\|x\|} = x - P_{B(0, \lambda)} x,$$

  where we let $0/0 = 0$ if $x = 0$.

More examples:

- **$\ell_{\infty}$-norm**.

- **$\ell_{2,1}$-norm**.

- Unitary-invariant matrix norms: **Frobenius-norm, nuclear-norm, maximal singular value**.
Properties

Proposition (separable sums)

Suppose that $f(x, y) = \phi(x) + \psi(y)$ is a block separable function

$$\text{prox}_{\lambda f}(v, w) = (\text{prox}_{\lambda \phi}(v), \text{prox}_{\lambda \psi}(w))$$

- Note: we have observed this with $f(x) = \sum_{i=1}^{n} |x_i|$. 
Properties

**Theorem (minimizer = fixed point)**

Let $\lambda > 0$. Point $x^* \in \mathbb{R}^n$ is a minimizer of $f$ if, and only if, $\text{prox}_{\lambda f}(x^*) = x^*$. 


Proximal-point algorithm (PPA)

- **Iteration:**
  \[ x^{k+1} = \text{prox}_{\lambda f}(x^k) \]

- **Convergence:**
  - For strongly convex \( f \), \( \text{prox}_{\lambda f} \) is a contraction, i.e., \( \exists C_0 < 1: \)
    \[ \| \text{prox}_{\lambda f}(x) - \text{prox}_{\lambda f}(y) \| \leq C_0 \| x - y \|. \]

  Let \( x = x^k \) and \( y = x^* \). We get
  \[ \| x^{k+1} - x^* \| = \| \text{prox}_{\lambda f}(x^k) - \text{prox}_{\lambda f}(x^*) \| \leq C_0 \| x^k - x^* \|. \]

  Iterate this, then
  \[ \| x^{k+1} - x^* \| \leq C_0^{k+1} \| x^k - x^* \|. \]

  \( x^k \) converges \( x^* \) linearly.
For general convex $f$, $\text{prox}_{\lambda f}$ is **firmly nonexpansive**, i.e.,

$$\|\text{prox}_{\lambda f}(x) - \text{prox}_{\lambda f}(y)\|^2 \leq \|x - y\|^2 - \|(x - \text{prox}_{\lambda f}(x)) - (y - \text{prox}_{\lambda f}(y))\|^2.$$ 

From this inequality, one can show:

- $x^k \rightarrow x^*$ weakly; still true if computation error is summable
- fixed-point residual $\|\text{prox}_{\lambda f}(x^k) - x^k\|^2 = o(1/k^2)$
- objective function $f(x^k) - f(x^*) = o(1/k)$
Proximal operator and resolvent

Definition
Given a mapping $T$, $(I + \lambda T)^{-1}$ is called the resolvent of $T$.

Proposition
Suppose that $f$ has subdifferential $\partial f$. We have

$$\text{prox}_{\lambda f} = (I + \lambda \partial f)^{-1}.$$ 

In addition, they are single valued.
Interpretation: implicit (sub)gradient

\[ x^{k+1} = \text{prox}_{\lambda f}(x^k) \quad \iff \quad x^{k+1} = (I + \lambda \partial f)^{-1}(x^k) \]
\[ \iff \quad x^k \in (I + \lambda \partial f)x^{k+1} \]
\[ \iff \quad x^k \in x^{k+1} + \lambda \partial f(x^{k+1}) \]
\[ \iff \quad x^{k+1} = x^k - \lambda \tilde{\nabla} f(x^{k+1}) \]

- **Notation:** \( \tilde{\nabla} f \) is the subgradient mapping uniquely defined by \( \text{prox}_{\lambda f} \)
- Let \( y^{k+1} = \tilde{\nabla} f(x^{k+1}) \in \partial f(x^{k+1}) \). Plugging formula of \( x^{k+1} \), we get
  \[ y^{k+1} \in \partial f(x^k - \lambda y^{k+1}) \].
- Given \( x^k \) and \( \lambda \), compute \( \text{prox}_{\lambda f}(x^k) \iff \text{solve } x^{k+1} \) (primal approach)
  \[ \iff \text{solve } y^{k+1} \] (dual approach)
Summary: proximal operator

- Conceptually simple, easy to understand and derive, a standard tool for nonsmooth and/or constrained optimization

- Work for any $\lambda > 0$, more stable than gradient descent

- Gives a fixed-point optimality condition and a converging algorithm

- “Sits at a high level of abstraction”

- Interpretations: general projection, implicit gradient, backward Euler

- Closed-formed or quick solutions for many basic functions

Next few lectures:

- Prox operation is applied when it is easily evaluated, so often a step in other algorithms, especially operator splitting algorithms

- Under separable structures, it is amenable to parallel and distributed algorithms with interesting applications in machine learning, signal processing, compressed sensing, large-scale modern convex problems