Proximal Operators and First Order Methods

Robert Baraldi

Firedrake 2020

February 11^{th}, 2020

Problem Class - Decomposable Functions

Consider (possibly) nonconvex composite problems of the form

$$\min_{x} \phi(x) + \varphi(x) \quad (1)$$

- $x \in \mathbb{R}^n$ are the decision variables
- φ may be nonsmooth, is convex typically
- ϕ is a 'nice' function (smooth, convex)

Applications and Utility

- Nonsmooth (and nonconvex) functions are useful but difficult
  - How do you minimize things without gradients or Hessians?
  - Optimization community has focused on using first order methods for nonsmooth functions
- Why do difficult functions arise?
  - Implementation of nonsmooth/nonconvex regularizers and constraints
    - Promote simplicity in ill-posed or high-dimensional setting - TV regularization
    - "Classic" optimization examples: sparse regression, matrix completion, phase retrieval
- Separable nonsmooth/nonconvex optimization is much easier than the general case
  - Special function structures can be exploited [4].

Approach/Goals

- Can you exploit function structure to find minima?
  - Tools: Infimal Convolution, Proximal Gradient Descent, various accelerations (FISTA)
- Purpose: Bridge the gap between the optimization and PDE communities.
- Numerical Example: Obstacle problem

Definition (Infimal Convolution)

Let $f, g : \mathcal{H} \to]-\infty, +\infty]$. The infimal convolution or *epi-sum* of f and g is

$$f \Box g : \mathcal{H} \to [-\infty, +\infty] : x \mapsto \inf_{y \in \mathcal{H}} (f(y) + g(x - y)) \qquad (2)$$

and it is *exact at a point* $x \in \mathcal{H}$ if $(f \Box g)(x) = \min_{y \in \mathcal{H}} f(y) + g(x - y)$, i.e.

$$(\exists y \in \mathcal{H}) \ (f \Box g)(x) = f(y) + g(x - y) \in]-\infty, +\infty]. \qquad (3)$$

Infimal Convolution: Discussion

- Backbone of many convex (and sometimes nonconvex optimization techniques)
- Has many useful properties (i.e. $f \Box g = g \Box f$, $\mathrm{dom}(f \Box g) = \mathrm{dom}(f) + \mathrm{dom}(g)$ given f, g have affine minorants, etc...)
- Useful for smoothing functions, looking at dual functions.

Now with Norms

Proposition (Inf. Conv. of p-Norms)

For $f \in \Gamma_0(\mathcal{H})$, let $\gamma \in \mathbb{R}_{++}$ and $p \in]1, +\infty]$. Then

$$f \square \left(\frac{1}{\gamma p} \| \cdot \|^p \right) : \mathcal{H} \to]-\infty, +\infty] : x \mapsto \inf_{y \in \mathcal{H}} \left(f(y) + \frac{1}{\gamma p} \|x - y\|^p \right) \tag{4}$$

is convex, real-valued, continuous, and exact. Moreover, for every $x \in \mathcal{H}$, the infimum is uniquely attained.

- Leads into β-Pasch-Hausdorff envelopes, with useful properties of Lipschitz functions that we will skip for now.

The Most Important Norm: $p = 2$, Moreau-Yoshida envelope/regularization [1]

Definition (Moreau Envelope)

Let $f : \mathcal{H} \mapsto]-\infty, +\infty]$ and let $\gamma \in \mathbb{R}_{++}$ the Moreau-Envelope of f of parameter γ is

$$^{\gamma}f = f \square \left(\frac{1}{2\gamma} \| \cdot \|^2 \right). \tag{5}$$

Definition (Proximal Operator/Mapping)

Let $f \in \Gamma_0(\mathcal{H}),\ x \in \mathcal{H}$. Then $\mathrm{prox}_{\gamma f}(x)$ is the unique point in $\mathcal{H}$ that satisfies

$$\mathrm{prox}_{\gamma f}(x) = \arg\min_{y} {}^{\gamma}f(x) = f(\mathrm{prox}_{\gamma f}(x)) + \frac{1}{2\gamma} \|x - \mathrm{prox}_{\gamma f}(x)\|^2. \tag{6}$$

Proposition (Firm Nonexpansivity)

Let $f \in \Gamma_0(\mathcal{H})$. Then prox_f and $I - \mathrm{prox}_f$ are firmly nonexpansive.

Proposition (Differentiability and Lipschitz)

Let $f \in \Gamma_0(\mathcal{H})$ and $\gamma \in \mathbb{R}_{++}$. Then ${}^{\gamma}f : \mathcal{H} \to \mathbb{R}$ is Freéchet Differentiable on $\mathcal{H}$, and its gradient

$$\nabla({}^{\gamma}f) = \gamma^{-1}(I - \mathrm{prox}_{\gamma f}) \tag{7}$$

is γ^{-1}-Lipschitz continuous.

- We can compute the proximal operator analytically for many functions[?]:
  - ℓ_1-norm: soft-thresholding
  - Indicator Functions: if C is a nonempty closed convex subset of $\mathcal{H}$, then $\mathrm{prox}_{\delta_C} = \mathrm{proj}_C$.

Properties of the Prox

- ${}^{\gamma}f$ of convex f is $1/\gamma$ smooth.
- Preserves optimal criterion: $\min_x {}^{\gamma}f = \min_x f(x)$
- Preserves optimal solution: x minimizes f iff x minimizes ${}^{\gamma}f$ for all $\gamma > 0$ (even for nonconvex)
- Fixed point iteration: x^* minimizes f iff $x^* = \mathrm{prox}_{\gamma f}(x^*)$

Subgradients

Definition (Subgradient)

A vector g is a subgradient of convex f at $x \in \text{dom}(f)$ if $\forall z \in \text{dom}(f)$,

$$f(z) \geq f(x) + g^T(z - x)$$

or more generally for nonconvex f

$$f(z) \geq f(x) + g^T(z - x) + o(\|z - x\|).$$

and $\partial f(x)$ is the set of all g for which the above holds.

- General first order optimality:

$$0 \in \partial f(x) \Leftrightarrow x \in \arg\min_x f(x)$$

Operator View

- First order optimality conditions of ${}^{\gamma}f$:

$$0 \in (x^* - x) + \partial f(x^*) \Leftrightarrow x \in x^* + \partial f(x^*) = (I + \partial f)(x^*)$$

- $\mathrm{prox}_{\gamma f}(x) = (I + \nu \partial f)^{-1}(x).$

Prox as Backwards Euler

- Gradient flow: $x'(t) = -\nabla f(x),\ x(0) = x_0$.
- First order numerical method for tracing path from x_0 to x^* with finite difference (backwards)

$$\begin{aligned} \gamma^{-1}(x(t) - x(t-\gamma)) &\approx -\nabla f(x(t)) \\ x^{k+1} &= x^k - \gamma \nabla f(x^{k+1}) \end{aligned}$$

- We can get the same thing with proximal operator:

$$x^{k+1} = \arg\min_x\ f(x) + \frac{1}{2\gamma}\|x - x^k\|^2$$

$$\Downarrow \quad \text{differentiate w.r.t. } x^{k+1}$$

$$0 = \nabla f(x^{k+1}) + \gamma^{-1}(x^{k+1} - x^k)$$

How is this used?

- Smooths difficult regularizers
- Separate the composite function into distinct entities
- Generally: subgradients have nicer properties
- Naively: nonsmooth derivatives are subgradients, use the subgradient method

Subgradient Algorithm [3]

Input: x^0
Initialize: $k = 0$.
while not converged **do**
    $x^k \leftarrow x^{k-1} + t_k g^{k-1}$ for $g^{k-1} \in \partial f(x^{k-1})$
end while
Output: x

- Not necessarily a descent method
- Step sizes t_k are pre-specified as fixed or diminishing; not obvious.
- Objective function error level of $O(1/\sqrt{k})$ after k iterations even for Lipschitz, convex functions.

Taking Advantage of Problem Structure

- Recall our problem:

$$\min_x f(x) := \phi(x) + \varphi(x)$$

where we know that *some* pure gradient info exists - $\nabla\phi$.

- With gradient descent, we'd minimize a 1st order approximation of ϕ around x:

$$x^+ = \arg\min_z \underbrace{\phi(x) + \nabla\phi^T(z - x) + \frac{1}{2\nu}\|z - x\|^2}_{\tilde{\phi}_\nu(z)}$$

with $\nabla^2\phi(x) \approx \nu^{-1} I$.

Proximal-Gradient Derivation [3]

- Since f is not differentiable - approximate ϕ but leave φ:

$$
\begin{aligned}
x^+ &= \arg\min_z \ \tilde{\phi}_\nu(z) + \varphi(z) \\
&= \arg\min_z \phi(x) + \nabla\phi^T(z-u) + \frac{1}{2\nu}\|z-x\|_2^2 + \varphi(z) \\
&= \arg\min_z \underbrace{\phi(x) - \nu\|\phi(x)\|^2}_{\text{adds nothing}} \dots \\
&\dots + \underbrace{\nabla\phi^T(z-x) + \frac{1}{2\nu}\|z-x\|_2^2 + \nu\|\phi(x)\|^2}_{\text{complete the square}} + \varphi(z) \\
&= \arg\min_z \ \frac{1}{2\nu}\|z-(x-\nu\nabla\phi(x))\|_2^2 + \varphi(z).
\end{aligned}
$$

- $x^{k+1} = (I + \nu\partial\varphi)^{-1}(I - \nu\nabla\phi)x_k.$

Proximal Gradient Algorithm

Input: x^0
Initialize: $k = 0$.
while not converged **do**
    $x^k \leftarrow \mathrm{prox}_{\nu_k \varphi}(x^{k-1} + \nu_k \nabla \phi(x^{k-1}))$
end while
Output: x

Connections

- Moreau Envelope at the gradient update:

$$\arg\min_z \underbrace{\frac{1}{2\nu}\|z - (x - \nu\nabla\phi(x))\|_2^2}_{\text{stay close to gradient update of } \phi} + \underbrace{\varphi(z)}_{\text{minimize } \varphi}$$

- For $G_\nu(x) = \nu^{-1}(x - \mathrm{prox}_{\nu\varphi}(x - \nu\nabla\phi(x)))$ (Generalized gradient update - also gradient of Moreau Envelope)

$$x^k = x^{k-1} - \nu_k G_{\nu_k}(x^{k-1})$$

- Only need gradients of ϕ, hopefully closed-form prox evaluation

More Connections & Extensions

- Can combine with backtracking linesearch to choose ν_k
- Convergence rate of $O(1/k)$.
- 'Generalized Gradient Descent':
  - $\varphi = 0$: gradient descent
  - $\varphi = \delta_C$: projected gradient descent
  - $\phi = 0$: proximal point algorithm
- Current work - inexact prox evaluation
- Accelerate with momentum weights - FISTA [2]

Acceleration - FISTA

Input: $x^0, \ x^{-1}, \ t^0$
Initialize: $k = 0$.
while not converged **do**
  $x^k \leftarrow \mathrm{prox}_{\nu_k \varphi}(y + \nu_k \nabla \phi(y))$
  $t^k \leftarrow \frac{1}{2}(1 + \sqrt{1 + 4(t^{k-1})^2})$
  $y \leftarrow x^k + \frac{t^{k-1}-1}{t^k}(x^k - x^{k-1})$
end while
Output: x

FISTA

- Utilizes "momentum weights" in t_k
- Iterations are proximal gradient steps at extrapolated points y
- x^k are feasible, y are possibly outside the domain of φ
- Convergence $O(1/k^2)$

Conclusions

- A variety of fast, first order methods exist for nonsmooth problems - complete with analysis in finite dimensions
- More communication between PDE and optimization communities
  - Extensions into Sobolev spaces?
  - Implementation in UFL languages?

References I

Convex Analysis and Monotone Operator Theory in Hilbert Spaces.
Number 3. Springer Science + Business Media, Berlin, 2011.

Amir Beck.
First-Order Methods in Optimization.
SIAM-Society for Industrial and Applied Mathematics, Philadelphia, USA, 2017.

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.

Neal Parikh and Stephen Boyd.
Proximal algorithms.
Foundations and Trends in Optimization, 1(3):123–231, 2014.

References II

Ewout van den Berg and Michael P. Friedlander.
Probing the pareto frontier for basis pursuit solutions.
SIAM J. Sci. Comput., 31(2):890–912, November 2008.