Proximal gradient (Part II)

Yu-Xiang Wang
CS292F

(Based on Ryan Tibshirani's 10-725)

Last time: proximal gradient descent

Consider the problem

$$\min_x \; g(x) + h(x)$$

with g, h convex, g differentiable, and h "simple" in so much as

$$\mathrm{prox}_t(x) = \underset{z}{\mathrm{argmin}} \; \frac{1}{2t}\|x - z\|_2^2 + h(z)$$

is computable. Proximal gradient descent: let $x^{(0)} \in \mathbb{R}^n$, repeat:

$$x^{(k)} = \mathrm{prox}_{t_k}\big(x^{(k-1)} - t_k \nabla g(x^{(k-1)})\big), \quad k = 1, 2, 3, \ldots$$

Step sizes t_k chosen to be fixed and small, or via backtracking

If ∇g is Lipschitz with constant L, then this has convergence rate $O(1/\epsilon)$. Lastly we can accelerate this, to optimal rate $O(1/\sqrt{\epsilon})$

Last time: proximal gradient descent

In the convergence proof (HW2 Q3), we rewrote update as the following:

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

where G_t is the generalized gradient of f, (Nesterov's Gradient Mapping!)

$$G_t(x) = \frac{x - \mathrm{prox}_t\big(x - t\nabla g(x)\big)}{t}$$

Then we more or less followed the convergence proof of the standard Gradient Descent (Lecture 3).

What is G_t? Is G_t the gradient of some function?

What exactly is the proximal gradient algorithm descent doing?

Outline

Today:

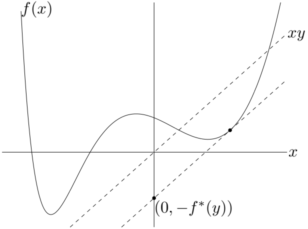
- Fenchel conjugate
- Prox Operator, Moreau Envelope and Smoothing
- Interpreting proximal algorithms

(Fenchel) Conjugate function

Given a function $f : \mathbb{R}^n \to \mathbb{R}$, define its conjugate $f^* : \mathbb{R}^n \to \mathbb{R}$,

$$f^*(y) = \max_x \; y^T x - f(x)$$

Note that f^* is always convex, since it is the pointwise maximum of convex (affine) functions in y (here f need not be convex)



$f^*(y)$: maximum gap between linear function $y^T x$ and $f(x)$

(From B & V page 91)

For differentiable f, conjugation is called the Legendre transform

Examples:

- Simple quadratic: let $f(x) = \frac{1}{2}x^TQx$, where $Q \succ 0$. Then $y^Tx - \frac{1}{2}x^TQx$ is strictly concave in x and is maximized at $x = Q^{-1}y$, so
$$f^*(y) = \frac{1}{2}y^TQ^{-1}y$$
- Indicator function: if $f(x) = I_C(x)$, then its conjugate is
$$f^*(y) = I_C^*(y) = \max_{x \in C} \; y^Tx$$
called the support function of C
- Norm: if $f(x) = \|x\|$, then its conjugate is
$$f^*(y) = I_{\{z \,:\, \|z\|_* \leq 1\}}(y)$$
where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$

Properties:

- Fenchel's inequality: for any x, y,

$$f(x) + f^*(y) \geq x^T y$$

- Conjugate of conjugate f^{**} satisfies $f^{**} \leq f$
- If f is closed and convex, then $f^{**} = f$
- If f is closed and convex, then for any x, y,

$$\begin{aligned} x \in \partial f^*(y) \quad &\Longleftrightarrow \quad y \in \partial f(x) \\ &\Longleftrightarrow \quad f(x) + f^*(y) = x^T y \end{aligned}$$

- If $f(u, v) = f_1(u) + f_2(v)$, then

$$f^*(w, z) = f_1^*(w) + f_2^*(z)$$

Moreau Envelope and Smoothing

We talked about prox operator

$$\mathrm{prox}_{t,f}(x) \in \operatorname*{argmin}_{y} \frac{1}{2t}\|y-x\|^2 + f(y).$$

Note that the output of prox is in the dom_f.
The Moreau envelope of a function f defined as

$$\begin{aligned} M_{t,f}(x) &:= \min_{y} \frac{1}{2t}\|y-x\|^2 + f(y) \\ &= \frac{1}{2t}\|\mathrm{prox}_{t,f}(x) - x\|^2 + f(\mathrm{prox}_{t,f}(x)). \end{aligned}$$

The Moreau envelope outputs the optimal objective value.

These quantities can be defined by for general functions but many of their remarkable properties only apply to convex f.

Example: Huber function

Coming from robust statistics (Huber, 1964, Annals of Statistics).

$$L_\delta(x) = \begin{cases} \frac{1}{2}x^2 & \text{if } |x| \le \delta \\ \delta(|x| - \frac{1}{2}\delta) & \text{otherwise.} \end{cases}$$

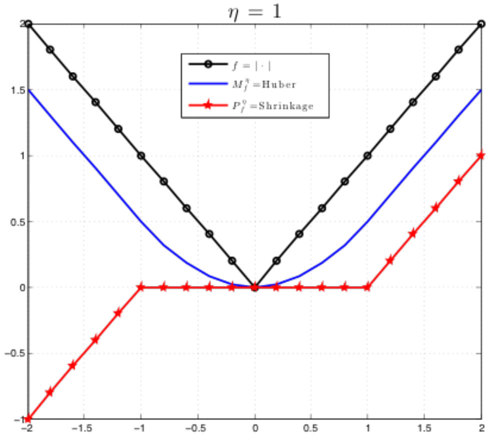
We can rewrite the Huber function as the Moreau Envelope of the absolute value function $|\cdot|$.

$$M_{\delta|\cdot|}(x) = \min_y \frac{1}{2}(x-y)^2 + \delta|y|.$$

Proof.

We know that the argmax is the soft-shresholding operator. Substitute that into the equation. If $|x| > \delta$, the optimal solution $y^* = x - \delta \mathrm{sign}(x)$, and the criterion value is $\frac{1}{2}\delta^2 + \delta|x| - \delta^2$. If $|x| < \delta$, the $y^* = 0$ and $M_{\delta|\cdot|}(x) = \frac{1}{2}x^2$ ☐

Example: Huber function



(Stolen from Yaoliang Yu's wonderful notes. [Click Here].)

Properties of a Moreau Envelope and Prox Operator

1. (Yoshida-Moreau Smoothing) $M_{t,f}(x)$ of any convex function is $1/t$-smooth. (Need duality to write down a clean proof.)
2. (Preservation of optimal criterion.) $\min_x f(x) = \min_x M_f(x)$.
3. (Preservation of optimal solution.) x minimizes f *if and only if* x minimizes $M_{t,f}(x)$ for all $t > 0$ (even for nonconvex functions).
4. (Gradient of a Moreau-Envelope) $\nabla M_{t,f}(x) = \frac{x - \mathrm{prox}_{t,f}(x)}{t}$
5. (Fixed Point Iteration) x^* minimizes f *if and only if* $x^* = \mathrm{prox}_{t,f}(x^*)$.

More properties of a Moreau Envelope and Prox Operator

1. (Moreau Decomposition) $x = \mathrm{prox}_f(x) + \mathrm{prox}_{f^*}(x)$
    - You can think of it as a generalization of the orthogonal projection decomposition to a subspace S

$$x = \Pi_S(x) + \Pi_{S^\perp}(x).$$

    - Combine with the gradients, you have: $\nabla M_f(x) = \mathrm{prox}_{f^*}(x)$.
2. (Proximal average) Let $f_1, ..., f_m$ be closed proper convex functions, there exists a convex function g, such that

$$\frac{1}{m}\sum_{i=1}^{m} \mathrm{prox}_f = \mathrm{prox}_g.$$

3. (Non-Expansiveness) prox_f is a non-expansion, namely, for all x, y

$$\|\mathrm{prox}_f(x) - \mathrm{prox}_f(y)\|^2 \leq \langle x - y, \mathrm{prox}_f(x) - \mathrm{prox}_f(y)\rangle.$$

Operator-theoretic view of a prox operator

∂f maps a point $x \in \mathrm{dom} f$ to the set $\partial f(x)$.
$(I + t\partial f)^{-1}$ is called the resolvent of an operator ∂f.

Theorem: Consider convex function f (so that the subgradient exists in the rel-int)

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

Proof: Recall the definition:

$\mathrm{prox}_f(x) = \mathrm{argmin}_y \frac{1}{2}\|y - x\|^2 + f(y)$.
By the first order optimality condition x^* obeys that

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

if an only if

$$x^* = (\mathrm{I} + \partial f)^{-1} x.$$

Proximal Point Algorithm (aka Proximal Minimization)

To minimize a convex function f. Iterate:

$$x^{k+1} = \mathrm{prox}_{tf}(x^k).$$

1. This is a fixed point iteration (note that prox is a non-expansion).

$$x^{k+1} = (\mathrm{I} + t\partial f)^{-1} x^k.$$

2. Also, this is a gradient descent on the Moreau Envelope.

$$x^{k+1} = x_k - (\mathrm{I} - (\mathrm{I} + t\partial f)^{-1}) x_k = x_k - t\nabla M_f(x_k).$$

Question: Is the learning rate appropriate for the GD to converge?

Proximal Gradient Algorithm

For minimizing a composition objective $f + h$

$$x^{k+1} = \mathrm{prox}_{th}(x^k - t\nabla f(x^k)).$$

1. As a fixed point iteration:

$$x^{k+1} = (I + t\partial h)^{-1}(I - t\nabla f)x_k$$

2. As a Smoothed Majorization-Minimization objective

$$x^{k+1} = \operatorname*{argmin}_y f(x^k) + \langle \nabla f(x^k), y - x^k \rangle + \frac{1}{2t}\|y - x_k\|^2 + h(y)$$

3. The generalized gradient is the gradient of a Moreau-Envelope of $f_{Linearized} + h$ at x^k.

Summary of Proximal Algorithms

1. Proximal point algorithm is to minimize the smoothed version of a nonsmooth objective using gradient descent.
2. Proximal gradient is to combine the idea of local quadratic approximation (with Majorization-Minimization) with the Moreau-Yoshida smoothing.
3. We can express things in operator-theoretic form as fixed point iterations.
4. If the fixed point iterations are conducted using a contraction map, then we have linear convergence.

References and further reading

- Parikh, N. and Boyd, S. (2014). “Proximal algorithms”. Foundations and Trends® in Optimization, 1(3), 127-239.
- Yaoliang Yu (2015). “Proximity Operator”. `https://cs.uwaterloo.ca/~y328yu/mynotes/po.pdf`.
- Fenchel, W. (1949). “On conjugate convex functions”. Canadian Journal of Mathematics, 1(1), 73-77.
- Rockafellar, R. T. (1976). “Monotone operators and the proximal point algorithm. SIAM journal on control and optimization”, 14(5), 877-898.
- Vandenberghe’s Lecture Notes for ECE 236C “Proximal Operator”. `http://www.seas.ucla.edu/~vandenbe/236C/lectures/proxop.pdf`