CS292A Convex Optimization: Gradient Methods and Online Learning **Spring 2019**

Lecture 6 Proximal gradient (Part II): April 25

Lecturer: Yu-Xiang Wang *Scribes: Kaiqi Zhang*

6.1 Fenchel conjugate

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

$$f^*(y) = \max_x y^T x - f(x) \tag{6.1}$$

Note that f^* is always convex, since it is the pointwise maximum of function convex (affine) functions in y.

It has the following properties:

- Fenchel's inequality: for any x, y
$$f(x) + f^*(y) \geq x^T y \tag{6.2}$$
- Conjugate of conjugate f^{**} satisfies $f^{**} \leq f$.
- If f is closed and convex, then for any x, y,
$$x \in \partial f^*(y) \Longleftrightarrow y \in \partial f(x) \Longleftrightarrow f(x) + f^*(y) = x^T y \tag{6.3}$$
- If $f(u, v) = f_1(u) + f_2(v)$, then $f^*(w, z) = f_1^*(w) + f_2^*(z)$

Examples:

$f(x)$	$f^*(x)$
$\frac{1}{2}x^T Qx(Q \succ 0)$	$\frac{1}{2}y^T Q^{-1} y$
$I_C(x)$ (indicator function)	$\max_{x \in C} y^T x$ (support function)
$\|x\|$	$I_{\{z:\|z\|_* \leq 1\}}(y)$

6.2 Moreau Envelope and Smoothing

$$\begin{aligned} M_{t,f}(x) &:= \min_y \tfrac{1}{2t}\|y - x\|^2 + f(y) \\ &= \tfrac{1}{2t}\|\mathrm{prox}_{t,f}(x) - x\|^2 + f(\mathrm{prox}_{t,f}(x)) \end{aligned} \tag{6.4}$$

Example: Huber function is

$$L_\delta(x) = \begin{cases} \frac{1}{2}x^2 & \text{if } |x| \leq \delta \\ \delta(|x| - \frac{1}{2}\delta) & \text{otherwise} \end{cases} \tag{6.5}$$

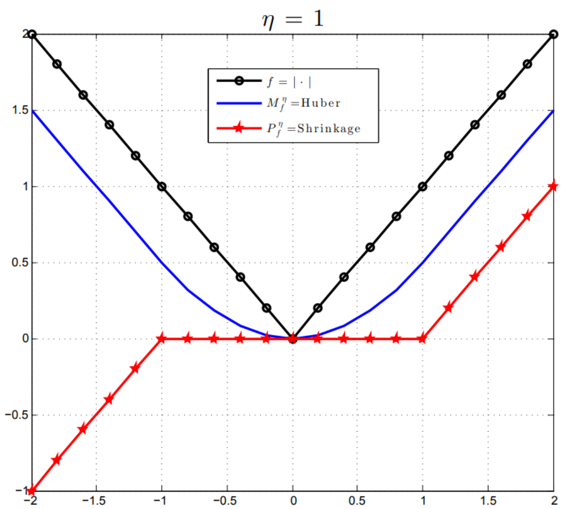

Figure 6.1: Huber envelope of absolute value function

is the Moreau Envelope of the absolute value function

$$M_{\delta|\cdot|}(x) = \min_y \frac{1}{2}(x-y)^2 + \delta|y| \tag{6.6}$$

Huber envelope and prox operators has the following properties:

- (Yoshida-Moreau Smoothing) $M_{t,f}(x)$ of any convex function is $1/t$-smooth.
- (Preservation of optimal criterion.) $\min_x f(x) = \min_x M_f(x)$.
- (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).
- (Gradient of a Moreau-Envelope) $\nabla M_{t,f}(x) = \frac{x - \text{prox}_{t,f}(x)}{t}$.
- (Fixed Point Iteration) x^* minimizes f if and only if $x^* = \text{prox}_{t,f}(x^*)$.
- (Moreau Decomposition) $x = prox_f(x) + prox_{f^*}(x)$. This a generalization of the orthogonal projection decomposition to a subspace S. $x = \Pi_S(x) + \Pi_{S^\perp}(x)$. Combine with the gradients, we have $\nabla M_f(x) = \text{prox}_{f^*}(x)$.
- (Proximal average) Let $f_1, \dots, f_m$ be closed proper convex functions, there exists a convex function g, such that
$$\frac{1}{m}\sum_{i=1}^m \text{prox}_f = \text{prox}_g \tag{6.7}$$
- (Non-Expansiveness) prox_f is a non-expansion, namely, for all x, y,

$$\|\text{prox}_f(x) - \text{prox}_f(y)\|^2 \le \langle x - y, \text{prox}_f(x) - \text{prox}_f(y)\rangle \tag{6.8}$$

6.3 Operator-theoretic view of a prox operator

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

Theorem 6.1 *Consider convex function f,*

$$prox_{t,f}(x) = (I + t\partial f)^{-1}(x). \tag{6.9}$$

Proof: Recall the definition:

$$\text{prox}_f(x) = \arg\min_y \frac{1}{2}\|y - x\|^2 + f(y). \tag{6.10}$$

By the first order optimality condition x^* obeys that

$$0 \in (x^* - x) + \partial f(x^*) \Longleftrightarrow x \in x^* + \partial f(x^*) = (I + \partial f)(x^*) \tag{6.11}$$

if an only if

$$x^* = (I + \partial f)^{-1} x. \tag{6.12}$$

■

6.4 Proximal Point Algorithm (aka Proximal Minimization)

To minimize a convex function f . Iterate:

$$x^{k+1} = \text{prox}_{tf}(x^k). \tag{6.13}$$

- This is a fixed point iteration (note that prox is a non-expansion) $x^{k+1} = (I + t\partial f)^{-1} x^k$.
- Also, this is a gradient descent on the Moreau Envelope. $x^{k+1} = x^k - (I - (I + t\partial f)^{-1})x_k = x_k - t\nabla M_f(x_k)$.

6.5 Proximal Gradient Algorithm

For minimizing a composition objective $f + h$

$$x^{k+1} = \text{prox}_{th}(x^k - t\nabla f(x^k)). \tag{6.14}$$

- It can be taken as a fixed point iteration:

$$x_{k+1} = (I + t\partial h)^{-1}(I - t\nabla f)x^k \tag{6.15}$$

- Or, it can be taken as a Smoothed Majorization-Minimization objective

$$x^{k+1} = \arg\min_y f(x^k) + \langle \nabla f(x^k), y - x^k \rangle + \frac{1}{2t}\|y - x_k\|^2 + h(y) \tag{6.16}$$

Proof:

$$\begin{aligned} x^* \text{ is optimal } &\Longleftrightarrow 0 \in \nabla f(x^*) + \partial h(x^*) \\ &\Longleftrightarrow 0 \in \nabla f(x^*) - x^* + x^* + \partial h(x^*) \\ &\Longleftrightarrow x^* - \nabla f(x^*) \in x^* + \partial h(x^*) \\ &\Longleftrightarrow x^* - \nabla f(x^*) \in (I + \partial h)(x^*) \\ &\Longleftrightarrow x^* = (I + \partial h)^{-1}(I - \nabla f)(x^*) \end{aligned} \tag{6.17}$$

■

- The generalized gradient is the gradient of a Moreau-Envelope of $f_{\text{Linearized}} + h$ at x_k .

We now delve right into the proof.

Lemma 6.2 *This is the first lemma of the lecture.*

Proof: The proof is by induction on .... For fun, we throw in a figure.

Figure 6.1: A Fun Figure

This is the end of the proof, which is marked with a little box. ■

6.5.1 A few items of note

Here is an itemized list:

- this is the first item;
- this is the second item.

Here is an enumerated list:

1. this is the first item;
2. this is the second item.

Here is an exercise:

Exercise: Show that P $\neq$ NP.

Here is how to define things in the proper mathematical style. Let f_k be the $AND - OR$ function, defined by

$$f_k(x_1, x_2, \ldots, x_{2^k}) = \begin{cases} x_1 & \text{if } k = 0; \\ AND(f_{k-1}(x_1, \ldots, x_{2^{k-1}}), f_{k-1}(x_{2^{k-1}+1}, \ldots, x_{2^k})) & \text{if } k \text{ is even;} \\ OR(f_{k-1}(x_1, \ldots, x_{2^{k-1}}), f_{k-1}(x_{2^{k-1}+1}, \ldots, x_{2^k})) & \text{otherwise.} \end{cases}$$

Theorem 6.3 *This is the first theorem.*

Proof: This is the proof of the first theorem. We show how to write pseudo-code now.

Consider a comparison between x and y:

if x or y or both are in S **then**
    answer accordingly
else
    Make the element with the larger score (say x) win the comparison
    if $F(x) + F(y) < \frac{n}{t-1}$ **then**
        $F(x) \leftarrow F(x) + F(y)$
        $F(y) \leftarrow 0$
    else
        $S \leftarrow S \cup \{x\}$
        $r \leftarrow r + 1$
    endif
endif

This concludes the proof. ■

6.6 Next topic

Here is a citation, just for fun [CW87].

References

[CW87] D. Coppersmith and S. Winograd, "Matrix multiplication via arithmetic progressions," *Proceedings of the 19th ACM Symposium on Theory of Computing*, 1987, pp. 1–6.