MATH 301: Advanced Topics in Convex Optimization **Winter 2015**

Lecture 22 — February 27

Lecturer: Emmanuel Candes *Scribe: Qian Yang*

Warning: *These notes may contain factual and/or typographic errors. Some portions of lecture may have been omitted.*

22.1 Overview

In this lecture, we will discuss the Moreau envelope, which is one way to smooth a non-smooth function f, and we will show that the proximal minimization algorithm can be viewed simply as gradient descent on the Moreau envelope. The arguments will proceed as follows:

- First, we define the Moreau-Yosida regularization.
- We use conjugate functions to show that the proximal operator is equivalent to gradient descent on the Moreau envelope f_μ.
- We use strong duality to show that f_μ is itself the conjugate function of the conjugate of f plus a regularization term, and thus it is smooth.
- We will introduce Moreau's decomposition, which can be viewed as a generalization of orthogonal decomposition.
- We conclude that the optimal value of f_μ is also the optimal value of f, and thus the proximal minimization algorithm is a valid method for optimizing non-smooth functions.

22.2 Moreau-Yosida regularization

The *Moreau envelope* or *Moreau-Yosida regularization* is given by

$$f_\mu(x) = \inf_y \left\{ f(y) + \frac{1}{2\mu}\|x-y\|_2^2 \right\}$$

We note that $\operatorname{dom} f_\mu(x) = \mathbb{R}^n$, and that $f_\mu(x)$ is convex.

To see the latter, note that $L(x,y) = f(y) + \frac{1}{2\mu}\|x-y\|_2^2$ is jointly convex in x and y. Then $f_\mu(x) = \inf_y L(x,y)$, which must be convex since its epigraph is the projection of a convex set and thus is itself a convex set.

Example 1 (Huber function)**.** Let $f(x) = |x|$. Then its Moreau envelope is just the familiar Huber function

$$f_\mu(x) = \inf_y \left\{ |y| + \frac{1}{2\mu}(x-y)^2 \right\} = \begin{cases} \frac{1}{2\mu}x^2, & |x| \le \mu, \\ |x| - \frac{\mu}{2}, & |x| > \mu. \end{cases}$$

22.3 Representation via Conjugate Functions

22.3.1 Primal Viewpoint

We can rearrange terms to express $f_\mu(x)$ in the following form:

$$\begin{aligned} f_\mu(x) &= \frac{1}{2\mu}\|x\|^2 - \frac{1}{\mu}\sup_y \left\{ x^T y - \mu f(y) - \frac{1}{2}\|y\|^2 \right\} \\ &= \frac{1}{2\mu}\|x\|^2 - \frac{1}{\mu}\left(\mu f + \frac{1}{2}\|\cdot\|^2 \right)^* (x) \\ \therefore \nabla f_\mu(x) &= \frac{x}{\mu} - \frac{1}{\mu}\operatorname*{argmax}_y \left\{ x^T y - \mu f(y) - \frac{1}{2}\|y\|^2 \right\} \\ &= \frac{1}{\mu}\left(x - \mathbf{prox}_{\mu f}(x) \right) \\ \Rightarrow \mathbf{prox}_{\mu f}(x) &= x - \mu \nabla f_\mu(x) \end{aligned}$$

In the third step, recall the important point from last lecture that the gradient of the conjugate function $f^*(x)$ is equal to the optimal y^* at which $f^*(x) = \sup_{y\in \mathrm{dom}(f)} x^T y - f(y)$ is achieved. In the fourth step, it is easy to derive the standard definition of the proximal operator (see Appendix, Def.2 below) from the given expression.

This derivation gives us an important conclusion: the proximal operator is just performing *gradient descent on a smooth version of* f!

22.3.2 Dual Viewpoint

$$\begin{aligned} f_\mu(x) &= \min_y \left\{ f(y) + \frac{1}{2\mu}\|x - y\|^2 \right\} \\ &= \min_y \left\{ f(y) + \frac{1}{2\mu}\|z\|^2 \right\} \text{ such that } x - y = z \end{aligned}$$

(Note the substitution trick here is a very useful technique.) The Lagrangian and the Lagrange dual function are given by

$$\begin{aligned} \mathcal{L}(y, z, \lambda) &= f(y) + \frac{1}{2\mu}\|z\|^2 + \lambda^T(x - y - z) \\ &= \left[f(y) - \lambda^T y \right] + \left[\frac{1}{2\mu}\|z\|^2 - \lambda^T z \right] + \lambda^T x \\ g(\lambda) &= \inf_{y,z} \mathcal{L}(y, z, \lambda) \\ &= \inf_y \left\{ f(y) - \lambda^T y \right\} - \frac{\mu}{2}\|\lambda\|^2 + \lambda^T x \\ &= -f^*(\lambda) - \frac{\mu}{2}\|\lambda\|^2 + \lambda^T x \end{aligned}$$

By strong duality, we must have that $f_\mu(x)$ is equal to the optimal value of the dual program, and thus

$$\begin{aligned} f_\mu(x) = \sup_\lambda g(\lambda) &= \sup_\lambda \left\{ -f^*(\lambda) - \frac{\mu}{2}\|\lambda\|^2 + \lambda^T x \right\} \\ &= \left(f^* + \frac{\mu}{2}\|\cdot\|^2 \right)^* (x) \end{aligned}$$

Recall from last lecture that the conjugate of a closed, proper, strongly convex function is smooth. Thus, *the Moreau envelope f_μ is smooth* and in particular, its gradient ∇f_μ is Lipschitz with constant at most μ^{-1}.

22.4 Moreau's Decomposition

An important identity is Moreau's decomposition, which states that

$$\mathbf{prox}_f(x) + \mathbf{prox}_{f^*}(x) = x$$

Example 2 (convex cone)**.** Suppose $f(x) = \mathbb{I}_\mathcal{K}(x)$, the indicator function of a convex cone $\mathcal{K}$, defined as $f(x) = 0$ on $\text{dom}(f) = \mathcal{K}$. Then $f^*(x) = \sup_{y\in\mathcal{K}} x^T y$. Consider the polar cone $\mathcal{K}^0 = \{x : x^T y \le 0, \forall y \in \mathcal{K}\}$. Then we see that

$$\begin{aligned} f^*(x) &= \begin{cases} 0 & x \in \mathcal{K}_0 \\ \infty & \text{otherwise} \end{cases} \\ &= \mathbb{I}_{\mathcal{K}_0}(x). \end{aligned}$$

The proximal operator of the indicator function is an Euclidean projection (this is immediate from the definition). Then Moreau's identity in this special case says that

$$x = \Pi_\mathcal{K}(x) + \Pi_{\mathcal{K}_0}(x),$$

where $\Pi_\mathcal{K}(x)$ is the projection of x onto the cone $\mathcal{K}$. In the case where $\mathcal{K}$ is a linear subspace V, we recover the familiar decomposition of x in terms of its projection onto V and onto its orthogonal complement $V^\perp$.

$$x = \Pi_V(x) + \Pi_{V^\perp}(x).$$

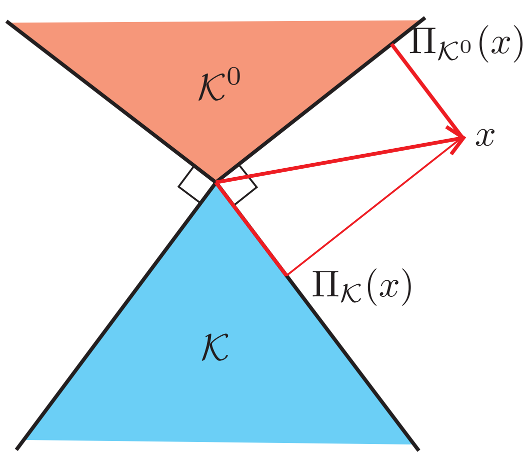

Figure 22.1: An illustration of Moreau's decomposition

We give a simple proof of Moreau's decomposition.

Proof. Let $\mu = 1$. Then from our primal-dual derivations in Section 9.3 above, we have that

$$\begin{aligned} f_1(x) &= \frac{1}{2}\|x\|^2 - \left(f + \frac{1}{2}\|\cdot\|^2\right)^*(x) = \left(f^* + \frac{1}{2}\|\cdot\|^2\right)^*(x) \\ \Rightarrow \quad \frac{1}{2}\|x\|^2 &= \left(f + \frac{1}{2}\|\cdot\|^2\right)^*(x) + \left(f^* + \frac{1}{2}\|\cdot\|^2\right)^*(x) \\ \Rightarrow \quad x &= \mathbf{prox}_f(x) + \mathbf{prox}_{f^*}(x) \end{aligned}$$

where in the last step we took the gradient of both sides.

Corollary. The proximal operator $\mathbf{prox}_f(x)$ is Lipschitz with constant less than 1 (i.e. a contraction) if f is strongly convex.

$$\|\mathbf{prox}_f(x) - \mathbf{prox}_f(y)\| \leq \|x - y\|$$

In fact, whether or not f is strongly convex, we have that

$$\|\mathbf{prox}_f(x) - \mathbf{prox}_f(y)\|^2 \leq (x-y)^T \left(\mathbf{prox}_f(x) - \mathbf{prox}_f(y)\right)$$

This property is called *firm nonexpansiveness.*

22.5 Proximal Minimization Algorithm

Proposition. Consider the usual minimization problem: $\min f(x)$ subject to $x \in C$, where $C \subseteq \mathrm{dom}(f)$, closed, convex, nonempty. Then x^* minimizes $f(x)$ over C iff x^* minimizes $f_\mu(x)$.

Proof.

$$\begin{aligned} \inf_x f_\mu(x) &= \inf_x \inf_y \left\{ f(y) + \frac{1}{2\mu}\|x-y\|^2 \right\} \\ &= \inf_y \inf_x \left\{ f(y) + \frac{1}{2\mu}\|x-y\|^2 \right\} \\ &= \inf_y f(y) \end{aligned}$$

We conclude that $\mathrm{argmin}\, f_\mu(x) = \mathrm{argmin}\, f(y)$.

22.6 Conclusion

Proximal minimization algorithm applied to a nonsmooth function f is equivalent to gradient descent on its smooth Moreau envelope f_μ, with stepsize $\mu = L^{-1}$, where L is the Lipschitz constant of f_μ.

22.7 References

Parikh, N and Boyd, S. Proximal Algorithms. *Foundations and Trends in Optimization*, Vol. 1, No. 3 (2013) 123-231.

22.8 Appendix

For completeness, we repeat here some useful definitions.

Definition 1. The conjugate f^* of a function f is defined as

$$f^*(x) = \sup_{y \in \mathrm{dom}\, f} \left\{ x^T y - f(y) \right\}$$

Definition 2. The proximal operator is defined as

$$\mathbf{prox}_{\mu f}(x) = \operatorname*{argmin}_{y} \left\{ \frac{1}{2\mu} \|x - y\|^2 + f(y) \right\}$$