1 Outline

In this lecture, we study

- Moreau-Yosida smoothing.
- Proimxal point algorithm applied to the smoothed problem.
- Augmented Lagrangian method.

2 Moreau-Yosida smoothing

GIven a function $f : \mathbb{R}^d \to \mathbb{R}$, the Moreau-Yosida smoothing of f is defined as

$$f_\eta(x) := \inf_u \left\{ f(u) + \frac{1}{2\eta}\|u - x\|_2^2 \right\}$$

for some $\eta > 0$. This is also referred to as the Moreau envelope. Note that

$$f_\eta(x) = f\left(\text{prox}_{\eta f}(x)\right) + \frac{1}{2\eta}\left\|\text{prox}_{\eta f}(x) - x\right\|_2^2.$$

Why do we care about this? There are several nice properties of the Moreau-Yosida smoothing.

2.1 Convexity and smoothness

Proposition 22.1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex. Then f_η is convex.*

Proof. Let

$$g(x, u) = f(u) + \frac{1}{2\eta}\|u - x\|_2^2.$$

Then g is convex in x, and it is convex in u. Moerover, $f_\eta(x)$ is a partial minimization of $g(x, u)$ obtained after minimizing out the variables u. Therefore, f_η is convex. □

Proposition 22.2. *The Fenchel conjugate of f_η is given by*

$$f_\eta^*(y) = f^*(y) + \frac{\eta}{2}\|y\|_2^2.$$

Proof. Note that

$$f_\eta(x) = \inf_{u+v=x} \left\{ f(u) + \frac{1}{2\eta}\|v\|_2^2 \right\}.$$

Hence, f_η is the infimal convolution of f and $\|\cdot\|_2^2/(2\eta)$. This implies that

$$f_\eta^*(y) = f^*(y) + \left(\frac{1}{2\eta}\|\cdot\|_2^2\right)^*(y).$$

Note that
$$\left(\frac{1}{2\eta}\|\cdot\|_2^2\right)^*(y) = \sup_v \left\{ y^\top v - \frac{1}{2\eta}\|v\|_2^2 \right\} = \frac{\eta}{2}\|y\|_2^2$$
where the last equality is deduced from the optimality condition. □

As a direct consequence of Proposition 22.2, we deduce the the Moreau-Yosida smoothing is smooth.

Proposition 22.3. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex. Then its Moreau envelope f_η is $(1/\eta)$-smooth in the ℓ_2 norm.*

Proof. First, as f is convex, f_η is convex. Since f_η is convex, it is continuous on $\mathbb{R}^d$. As $\mathbb{R}^d$ is closed, f_η is a closed function. It follows from Proposition 22.2 that the Fenchel conjugate f_η^* of f_η is η-strongly convex in the ℓ_2 norm. Then the Fenchel conjugate f_η^{**} of f_η^* is $(1/\eta)$-smooth in the ℓ_2 norm. Lastly, as f_η is closed and convex, $f_\eta^{**} = f_\eta$. Therefore, f_η is also $(1/\eta)$-smooth in the ℓ_2 norm. □

Let us consider an example.

Example 22.4. Let $f(x) = \|x\|_1$. Then
$$f_\eta(x) = \sum_{i=1}^d \frac{1}{\eta} L_\eta(x_i)$$
where
$$L_\eta(c) = \begin{cases} \eta|c| - \eta^2/2, & \text{if } |c| \geq \eta, \\ |c|^2/2, & \text{if } |c| \leq \eta. \end{cases}$$
Here, L_η is called the Huber loss (see Figure 22.1[1]).

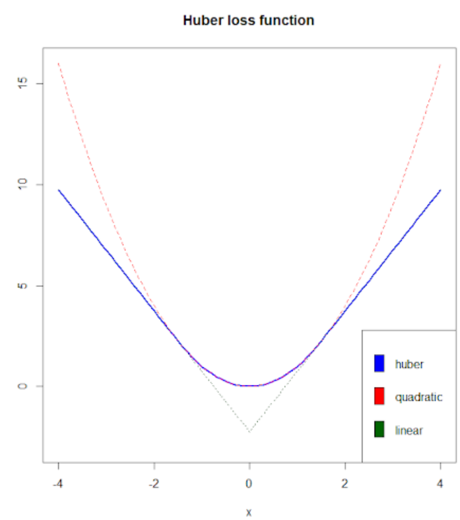

Figure 22.1: Huber loss

[1] Image taken from http://yetanothermathprogrammingconsultant.blogspot.com/2021/09/huber-regression-different-formulations.html

2.2 Optimization of the Moreau envelope

Moreover, we can compute the gradient of the Moreau-Yosida smoothing.

Proposition 22.5. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex. Then*

$$\nabla f_\eta(x) = \operatorname{prox}_{f^*/\eta}\left(\frac{x}{\eta}\right) = \frac{1}{\eta}(x - \operatorname{prox}_{\eta f}(x)).$$

Proof. By Proposition 22.3, f_η is smooth and thus differentiable. Moreover, as f_η is convex and closed, it follows that $y = \nabla f_\eta(x)$ if and only if $x \in \partial f_\eta^*(y)$. Note that Proposition 22.2 implies that

$$\partial f_\eta^*(y) = \partial f^*(y) + \eta y^*.$$

Hence, $x \in \partial f_\eta^*(y)$ if and only if $x - \eta y^* \in \partial f^*(y)$ which is equivalent to

$$\frac{1}{\eta}x - y^* \in \frac{1}{\eta}\partial f^*(y).$$

Furthermore, this is equivalent to

$$\operatorname{prox}_{f^*/\eta}\left(\frac{x}{\eta}\right) = y^*.$$

By the Moreau decomposition theorem, we have

$$x = \operatorname{prox}_{\eta f}(x) + \eta \operatorname{prox}_{f^*/\eta}(x/\eta),$$

so

$$\frac{1}{\eta}(x - \operatorname{prox}_{\eta f}(x)) = \operatorname{prox}_{f^*/\eta}\left(\frac{x}{\eta}\right),$$

as required. □

Proposition 22.6. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be closed. Then a minimizer of the Moreau-Yosida smoothing f_η is a minimizer of f.*

Proof. By Proposition 22.5, it follows that

$$\nabla f_\eta(x) = \frac{1}{\eta}(x - \operatorname{prox}_{\eta f}(x)).$$

Then, by the optimality condition, x^* is a minimizer of f_η if and only if

$$0 = \nabla f_\eta(x^*) = \frac{1}{\eta}(x^* - \operatorname{prox}_{\eta f}(x*))$$

which is equivalent to

$$x^* = \operatorname{prox}_{\eta f}(x^*).$$

Note that $x^* = \operatorname{prox}_{\eta f}(x^*)$ holds if and only if

$$0 = x^* - x^* \in \eta\partial f(x^*).$$

Therefore, $x^* = \operatorname{prox}_{\eta f}(x^*)$ if and only if x^* is a minimizer of f. □

Therefore, the problem

$$\text{minimize} \quad f(x)$$

is equivalent to solving

$$\text{minimize} \quad f_\eta(x) = \inf_u \left\{ f(u) + \frac{1}{2\eta}\|u - x\|_2^2 \right\}.$$

We know that f_η is convex by Proposition 22.1. Hence, we can attempt to solve the problem by gradient descent. By Proposition 22.5, the gradient of f_η is given by

$$\nabla f_\eta(x) = \frac{1}{\eta}(x - \mathrm{prox}_{\eta f}(x)).$$

Moreover, f_η is $(1/\eta)$-smooth by Proposition 22.3. Hence, the gradient descent update rule proceeds with step size η given as follows

$$x_{t+1} = x_t - \eta \nabla f_\eta(x_t) = \mathrm{prox}_{\eta f}(x_t).$$

This is precisely the update rule of the proximal point algorithm! This implies that the proximal point algorithm is equivalent to gradient descent applied to the smoothed objective.

3 Augmented Lagrangian method

We consider

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & Ax = b. \end{aligned}$$

We observed that its dual is given by

$$\text{maximize} \quad -f^*(-A^\top \mu) - b^\top \mu,$$

which is equivalent to

$$\text{minimize} \quad f^*(-A^\top \mu) + b^\top \mu,$$

Remember that the dual subgradient method solves the dual problem. In this section, we derive and study another algorithm that solves the dual formulation.

3.1 Proximal point algorithm applied to the dual

The proximal point algorithm proceeds with the following update rule.

$$\mu_{t+1} = \operatorname*{argmin}_\mu \left\{ f^*(-A^\top \mu) + b^\top \mu + \frac{1}{2\eta}\|\mu - \mu_t\|_2^2 \right\}.$$

By the optimality condition,

$$0 \in -A\partial f^*(-A^\top \mu_{t+1}) + b + \frac{1}{\eta}(\mu_{t+1} - \mu_t).$$

Hence,

$$\mu_{t+1} = \mu_t + \eta(Ax_t - b) \quad \text{where } x_t \in \partial f^*(-A^\top \mu_{t+1}).$$

Note that $x_t \in \partial f^*(-A^\top \mu_{t+1})$ holds if and only if $-A^\top \mu_{t+1} \in \partial f(x_t)$, which is equivalent to

$$\begin{aligned} 0 \in \partial f(x_t) + A^\top \mu_{t+1} \quad &\leftrightarrow \quad 0 \in \partial f(x_t) + A^\top(\mu_t + \eta(Ax_t - b)) \\ &\leftrightarrow \quad 0 \in \partial f(x_t) + A^\top \mu_t + \eta A^\top (Ax_t - b) \\ &\leftrightarrow \quad x_t \in \operatorname*{argmin}_x \left\{ f(x) + \mu_t^\top (Ax - b) + \frac{\eta}{2}\|Ax - b\|_2^2 \right\} \end{aligned}$$

Hence, the proximal point algorithm for the dual problem works with the following update rule.

$$\begin{aligned} &x_t \in \operatorname*{argmin}_x \left\{ f(x) + \mu_t^\top (Ax - b) + \frac{\eta}{2}\|Ax - b\|_2^2 \right\} \\ &\mu_{t+1} = \mu_t + \eta(Ax_t - b). \end{aligned}$$

This is precisely, the augmented Lagrangian method (ALM).

Algorithm 1 Augmented Lagrangian method

Initialize μ_1.
for $t = 1, \ldots, T$ **do**
  Find $x_t \in \operatorname{argmin}_x \left\{ f(x) + \mu_t^\top (Ax - b) + \frac{\eta}{2}\|Ax - b\|_2^2 \right\}$.
  Update $\mu_{t+1} = \mu_t + \eta(Ax_t - b)$.
end for

Notice that the augmented Lagrangian method is the dual gradient ascent applied to the following equivalent formulation of the primal problem.

$$\begin{aligned} \text{minimize} \quad & f(x) + \frac{\eta}{2}\|Ax - b\|_2^2 \\ \text{subject to} \quad & Ax = b. \end{aligned}$$

Note that the objective is strongly convex, which implies that the dual objective becomes smooth.

3.2 Gradient ascent to the smoothed dual

The proximal point algorithm on the dual is given by

$$\mu_{t+1} = \operatorname*{argmin}_\mu \left\{ f^*(-A^\top \mu) + b^\top \mu + \frac{1}{2\eta}\|\mu - \mu_t\|_2^2 \right\} = \operatorname{prox}_{\eta h}(\mu_t)$$

where

$$h(\mu) = f^*(-A^\top \mu) + b^\top \mu.$$

Remember that the proximal point algorithm is equivalent to gradient descent on the smoothed objective. The Moreau-Yosida smoothing of h is given by

$$h_\eta(\mu) = \inf_\gamma \left\{ f^*(-A^\top \gamma) + b^\top \gamma + \frac{1}{2\eta}\|\gamma - \mu\|_2^2 \right\}.$$

Note that

$$\text{minimize } h_\eta(\mu) \quad = \quad -\text{ maximize } -h_\eta(\mu)$$

is the dual of

$$\begin{aligned} \text{minimize} \quad & h_\eta^*(y) \\ \text{subject to} \quad & -y = 0 \end{aligned}$$

Here, what is h_η^*? By Proposition 22.2, we have

$$h_\eta^*(y) = h^*(y) + \frac{\eta}{2}\|y\|_2^2$$

Note that

$$\begin{aligned} h^*(y) &= \sup_\mu \left\{ y^\top \mu - f^*(-A^\top \mu) - b^\top \mu \right\} \\ &= \sup_\mu \left\{ (y-b)^\top \mu - f^*(-A^\top \mu) \right\} \\ &= \inf_x \left\{ f(x) : \ -Ax = y - b \right\}. \end{aligned}$$

Then

$$h_\eta^*(y) = \inf_x \left\{ f(x) : \ y = b - Ax \right\} + \frac{\eta}{2}\|y\|_2^2.$$

This implies that the dual problem is equivalent to

$$\begin{aligned} \text{minimize} \quad & f(x) + \frac{\eta}{2}\|b - Ax\|_2^2 \\ \text{subject to} \quad & Ax = b. \end{aligned}$$