Lecturer: Dabeen Lee

1 Outline

In this lecture, we study

- Moreau-Yosida smoothing,

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

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

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

2.1 Convexity and smoothness

Proposition 15.1. *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 15.2. *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 15.2, we deduce the the Moreau-Yosida smoothing is smooth.

Proposition 15.3. *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 15.2 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. □

2.2 Optimization of the Moreau envelope

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

Proposition 15.4. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex. Then*

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

Proof. By Proposition 15.3, 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 15.2 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

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

By the Moreau decomposition theorem, we have

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

so

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

as required. □

Proposition 15.5. *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 15.4, it follows that

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

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

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

which is equivalent to

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

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

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

Therefore, $x^* = \mathrm{prox}_{\eta f}(x^*)$ 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 15.1. Hence, we can attempt to solve the problem by gradient descent. By Proposition 15.4, 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 15.3. 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 Lagrangian Duality

We consider problems of the following structure.

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & g_i(x) \le 0 \quad \text{for } i = 1, \ldots, m \\ & h_j(x) = 0 \quad \text{for } j = 1, \ldots, \ell. \end{aligned} \tag{15.1}$$

We consider the most general setting for which we do not impose the condition that the objective and constraint functions are convex. We may define vector-valued functions $g : \mathbb{R}^d \to \mathbb{R}^m$ and $h : \mathbb{R}^d \to \mathbb{R}^\ell$ such that

- $g(x) = (g_1(x), \ldots, g_m(x))^\top$,
- $h(x) = (h_1(x), \ldots, h_\ell(x))^\top$.

Then (15.1) can be written as

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & g(x) \le 0 \\ & h(x) = 0. \end{aligned} \tag{15.2}$$

3.1 Lagrangian Dual Problem

The *Lagrangian function* of (15.1) is given by

$$\begin{aligned}\mathcal{L}(x,\lambda,\mu) &= f(x) + \lambda^\top g(x) + \mu^\top h(x)\\ &= f(x) + \sum_{i=1}^{m}\lambda_i g_i(x) + \sum_{j=1}^{\ell}\mu_j h_j(x).\end{aligned}$$

When the objective function f is convex, constraint functions $g_1,\ldots,g_m$ are convex, constraint functions $h_1,\ldots,h_\ell$ are affine, and the multiplier $\lambda \geq 0$, the Lagrangian function is convex in x for any fixed λ and μ. Moreover, the Lagrangian function is affine in λ and μ for any fixed x.

The *Lagrangian dual function* of (15.1) is

$$q(\lambda,\mu) = \inf_x \mathcal{L}(x,\lambda,\mu) = \inf_x \left\{ f(x) + \lambda^\top g(x) + \mu^\top h(x)\right\}.$$

Notice that the Lagrangian dual function is concave in (λ,μ), regardless of f, $g_1,\ldots,g_m$, and $h_1,\ldots,h_\ell$. This is because $\mathcal{L}(x,\lambda,\mu)$ is affine in λ and μ for any fixed x, and $q(\lambda,\mu)$ is a point-wise minimum of affine functions.

Proposition 15.6. *Let x be a feasible solution to (15.1), and $\lambda \geq 0$. Then*

$$f(x) \geq q(\lambda,\mu).$$

Proof. Since x is feasible, $g_i(x) \leq 0$ for $i=1,\ldots,m$ and $h_j(x) = 0$ for $j=1,\ldots,\ell$. Then for any $\lambda \geq 0$, we have

$$\sum_{i=1}^{m}\lambda_i g_i(x) + \sum_{j=1}^{\ell}\mu_j h_j(x) \leq 0.$$

This implies that

$$f(x) \geq \mathcal{L}(x,\lambda,\mu).$$

Note that

$$q(\lambda,\mu) = \inf_x \mathcal{L}(x,\lambda,\mu) \leq \mathcal{L}(x,\lambda,\mu).$$

Therefore, $f(x) \geq q(\lambda,\mu)$. □

By Proposition 15.6, if (15.1) is unbounded below, the Lagrangian dual function $q(\lambda,\mu) = -\infty$ for any $\lambda \geq 0$.

With the Lagrangian dual function, we can provide a lower bound on the problem (15.1). The *Lagrangian dual problem* is defined as

$$\begin{aligned}\text{maximize}\quad & q(\lambda,\mu)\\ \text{subject to}\quad & \lambda \geq 0.\end{aligned} \tag{15.3}$$

We often call (15.1) as *primal* and (15.3) as the *associated (Lagrangian) dual*. The following result states that the optimal value of the primal is lower bounded by the optimal value of the dual.

Theorem 15.7 (Weak duality)**.** *Consider the problem (15.1) and the associated Lagrangian dual problem (15.3). Then the following statement holds.*

$$\min_{x\in C} f(x) \geq \max_{\lambda\geq 0} q(\lambda,\mu)$$

where $C = \{x : g_i(x) \leq 0 \text{ for } i=1,\ldots,m,\ h_j(x) = 0 \text{ for } j=1,\ldots,\ell\}$.

Proof. By proposition 15.6, we know that $f(x) \geq q(\lambda, \mu)$ for any $x \in C$ and $\lambda \geq 0$. Then taking the minimum of $f(x)$ over $x \in C$, it follows that $\min_{x \in C} f(x) \geq q(\lambda, \mu)$. Then taking the maximum of $q(\lambda, \mu)$ over $\lambda \geq 0$, we obtain the desired inequality. $\square$

Theorem 15.7 holds regardless of whether the objective and constraint functions are convex or not. Then our next question is whether the equality holds. To answer this, we define the notion of *Slater's condition*.

Definition 15.8 (Slater's condition)**.** Suppose that $g_1, \ldots, g_k$ are affine and $g_{k+1}, \ldots, g_m$ are convex functions that are not affine. Then we say that the problem (15.1) satisfies Slater's condition if there exists a solution $\bar{x}$ such that

$$g_i(\bar{x}) \leq 0 \text{ for } i = 1, \ldots, k, \quad g_i(\bar{x}) < 0 \text{ for } i = k+1, \ldots, m, \quad h_j(\bar{x}) = 0 \text{ for } j = 1, \ldots, \ell.$$

If we assume that the objective f is convex and the constraint functions satisfy Slater's condition, then the inequality given in Theorem 15.7 holds with equality.

Theorem 15.9 (Strong duality)**.** *Consider the primal problem (15.1) and the associated Lagrangian dual problem (15.3). Assume that the objective function f and the constraint functions $g_1, \ldots, g_m$ are convex, and $h_1, \ldots, h_\ell$ are affine. If the primal problem (15.1) has a finite optimal value and Slater's condition, given in Definition 15.8, is satisfied, then there exist $\lambda^* \geq 0$ and μ^* such that*

$$\min_{x \in C} f(x) = q(\lambda^*, \mu^*) = \max_{\lambda \geq 0} q(\lambda, \mu)$$

where $C = \{x : g_i(x) \leq 0 \text{ for } i = 1, \ldots, m,\ h_j(x) = 0 \text{ for } j = 1, \ldots, \ell\}$.

3.2 Examples

Consider the following linear program in standard form.

$$\begin{aligned} \text{minimize} \quad & c^\top x \\ \text{subject to} \quad & Ax = b, \\ & x \geq 0. \end{aligned} \tag{15.4}$$

Then the Lagrangian dual function is given by

$$\begin{aligned} q(\lambda, \mu) &= \inf_x \mathcal{L}(x, \lambda, \mu) \\ &= \inf_x \left\{ c^\top x - \lambda^\top x + \mu^\top (Ax - b) \right\} \\ &= -b^\top \mu + \inf_x \left\{ (c - \lambda + A^\top \mu)^\top x \right\}. \end{aligned}$$

Note that $\inf_x \left\{ (c - \lambda + A^\top \mu)^\top x \right\} = -\infty$ unless $c - \lambda + A^\top \mu = 0$. Hence, to maximize the Lagrangian dual function $q(\lambda, \mu)$, it is sufficient to consider (λ, μ) satisfying $c - \lambda + A^\top \mu = 0$. Therefore, the associated Lagrangian dual problem is equivalent to

$$\begin{aligned} \text{maximize} \quad & -b^\top \mu \\ \text{subject to} \quad & c - \lambda + A^\top \mu = 0, \\ & \lambda \geq 0. \end{aligned} \tag{15.5}$$

In fact, we can eliminate the variable from the constraints $c + A^\top \mu \geq \lambda$ and $\lambda \geq 0$, and they can be equivalently written as $c + A^\top \mu \geq 0$. Moreover, the variables μ are unrestricted, so we can replace μ by $-\mu$. Then (15.5) is equivalent to

$$\begin{aligned} \text{maximize} \quad & b^\top \mu \\ \text{subject to} \quad & A^\top \mu \leq c, \end{aligned} \tag{15.6}$$

which is the dual linear program for (15.4).

Next we consider the following quadratic program.

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2} x^\top Q x + p^\top x \\ \text{subject to} \quad & Ax = b \end{aligned} \tag{15.7}$$

where Q is positive definite and thus is invertible. The corresponding Lagrangian function is given by

$$\begin{aligned} \mathcal{L}(x, \mu) &= \frac{1}{2} x^\top Q x + p^\top x + \mu^\top (Ax - b) \\ &= -b^\top \mu + \left(\frac{1}{2} x^\top Q x + (p + A^\top \mu)^\top x \right). \end{aligned}$$

Then

$$\nabla_x \mathcal{L}(x, \mu) = Qx + (p + A^\top \mu),$$

and therefore, $\nabla_x \mathcal{L}(x, \mu) = 0$ if and only if $x = -Q^{-1}(p + A^\top \mu)$. This in turn implies that the Lagrangian dual function is given by

$$\begin{aligned} q(\mu) &= \inf_x \mathcal{L}(x, \mu) \\ &= \mathcal{L}\left(-Q^{-1}(p + A^\top \mu), \mu \right) \\ &= -b^\top \mu - \frac{1}{2}(p + A^\top \mu)^\top Q^{-1} (p + A^\top \mu) \\ &= -\frac{1}{2} \mu^\top A Q^{-1} A^\top \mu - (b + A Q^{-1} p)^\top \mu - \frac{1}{2} p^\top p. \end{aligned}$$

Hence, the Lagrangian dual problem is

$$\max_\mu \left\{ -\frac{1}{2} \mu^\top A Q^{-1} A^\top \mu - (b + A Q^{-1} p)^\top \mu \right\}.$$

3.3 Lagrangian dual for conic programming

Consider the following conic programming problem

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & g_i(x) \leq_{K_i} 0 \quad \text{for } i = 1, \dots, m \\ & h_j(x) = 0 \quad \text{for } j = 1, \dots, \ell \end{aligned} \tag{15.8}$$

where $K_1, \dots, K_m$ are proper cones. Remember that $g_i(x) \leq_{K_i} 0$ means $-g_i(x) \in K_i$. Moreover, recall that the dual cone of a cone K is given by

$$K^* = \{ y : \ y^\top x \geq 0 \ \forall x \in K \}.$$

As we picked a nonnegative multiplier $\lambda \geq 0$ to define the Lagrangian function, we pick a multiplier λ from the dual cone K^*. The Lagrangian function of (15.8) is given by

$$\mathcal{L}(x, \lambda, \mu) = f(x) + \sum_{i=1}^{m} \lambda_i^\top g_i(x) + \sum_{j=1}^{\ell} \mu_j h_j(x)$$

where $\lambda_i \in K_i^*$ is now a vector from the dual cone of K_i for each i. Then the Lagrangian dual function is similarly defined as $q(\lambda, \mu) = \inf_x \mathcal{L}(x, \lambda, \mu)$. The Lagrangian dual problem is given by

$$\begin{aligned} \text{maximize} \quad & q(\lambda, \mu) \\ \text{subject to} \quad & \lambda_i \geq_{K_i^*} 0 \quad \text{for } i = 1, \dots, m. \end{aligned} \tag{15.9}$$

As an example, we consider the following semidefinite program.

$$\begin{aligned} \text{minimize} \quad & c^\top x \\ \text{subject to} \quad & \sum_{i=1}^{d} x_i A_i \geq_{S_+^m} B \end{aligned} \tag{15.10}$$

where S_+^m denotes the PSD cone containing all $m \times m$ PSD matrices. We learned that the PSD cone is self-dual, so the dual of S_+^m is itself. Let $Y \in S_+^m$, and consider the associated Lagrangian dual function.

$$q(Y) = \inf_x \mathcal{L}(x, Y) = \inf_x \left\{ c^\top x - \sum_{i=1}^{d} x_i \mathrm{tr}(Y^\top A_i) + \mathrm{tr}(Y^\top B) \right\}.$$

Note that the Lagrangian dual function $q(Y)$ has a finite value if and only if $c_i = \mathrm{tr}(Y^\top A_i)$ for every $i \in [d]$. Then the Lagrangian dual problem is given by

$$\begin{aligned} \text{maximize} \quad & \mathrm{tr}(Y^\top B) \\ \text{subject to} \quad & \mathrm{tr}(Y^\top A_i) = c_i \quad \text{for } i = 1, \dots, d. \\ & Y \in S_+^m \end{aligned} \tag{15.11}$$