IE 539: Convex Optimization **KAIST, Fall 2023**
Lecture #20: Dual gradient ascent November 15, 2023
Lecturer: Dabeen Lee

1 Outline

In this lecture, we study

- Dual gradient method,
- Moreau-Yosida smoothing.
- Optimization of the Moreau envelope.

2 Dual gradient 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

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

Then the problem is equivalent to

$$(-1) \quad \times \quad \text{minimize} \quad f^*(-A^\top \mu) + b^\top \mu.$$

As f^* is convex, this dual forumulation is a convex minimization problem. Let us apply the subgradient method to the dual.

2.1 Subgradient method for the dual problem

Given μ_t, let $g_t \in \partial\left(f^*(-A^\top \mu_t) + b^\top \mu_t\right)$. Then the subgradient method applies the following update rule.

$$\mu_{t+1} = \mu_t - \eta_t g_t.$$

Here, what is a subgradient g_t? Note that

$$\underbrace{\partial\left(f^*(-A^\top \mu_t) + b^\top \mu_t\right)}_{\text{subdifferential of } f^*(-A^\top \mu) + b^\top \mu \text{ at } \mu = \mu_t} = -A \underbrace{\partial f^*(-A^\top \mu_t)}_{\text{subdifferential of } f^*(\mu) \text{ at } \mu = -A^\top \mu_t} + b.$$

Hence, $g_t \in \partial\left(f^*(-A^\top \mu_t) + b^\top \mu_t\right)$ if and only if

$$g_t \in -A\partial f^*(-A^\top \mu_t) + b.$$

Therefore,

$$g_t = -Ax_t + b \quad \text{for some } x_t \in \partial f^*(-A^\top \mu_t).$$

Moreover, we have also observed that $x_t \in \partial f^*(-A^\top \mu_t)$ if and only if $-A^\top \mu_t \in \partial f(x_t)$. Here, $-A^\top \mu_t \in \partial f(x_t)$ holds if and only if $0 \in \partial f(x_t) + A^\top \mu_t$ which is equivalent to

$$x_t \in \underset{x}{\operatorname{argmin}}\, f(x) + \mu_t^\top A x.$$

Note that $\mu_t^\top b$ remains constant as x changes, so $x_t \in \operatorname{argmin}_x f(x) + \mu_t^\top A x$ is equivalent to

$$x_t \in \underset{x}{\operatorname{argmin}}\, f(x) + \mu_t^\top (Ax - b).$$

Therefore, the subgradient method applied to the dual problem proceeds with

$$\begin{aligned} &x_t \in \underset{x}{\operatorname{argmin}}\, f(x) + \mu_t^\top (Ax - b), \\ &\mu_{t+1} = \mu_t + \eta_t (A x_t - b). \end{aligned}$$

Here, $f(x) + \mu_t^\top (Ax - b)$ is the Lagrangian function $\mathcal{L}(x, \mu)$ at $\mu = \mu_t$. In words, the subgradient method applied to the dual problem works as follows. At each iteration t with a given dual multiplier μ_t, we find a minimizer of the Lagrangian function $\mathcal{L}(x, \mu_t)$. Then we use the corresponding dual subgradient $Ax_t - b$ to obtain a new multiplier μ_{t+1}.

Algorithm 1 Subgradient method for the dual problem

```
Initialize μ_1.
for t = 1, ..., T − 1 do
    Obtain x_t ∈ argmin_x f(x) + μ_t^⊤(Ax − b),
    Update μ_{t+1} = μ_t + η_t(Ax_t − b) for a step size η_t > 0.
end for
```

At each iteration, we find a minimizer of the Lagrangian function $\mathcal{L}(x, \mu_t)$, which gives rise to an unconstrained optimization problem. Hence, the dual approach is useful when there is a complex system of constraints.

2.2 Smoothness and strong convexity

Another motivation for using dual methods is that the dual objective can become smooth even if the primal objective is not.

Theorem 20.1. *Let $f : \mathbb{R}^d :\to \mathbb{R}$ be closed and α-strongly convex in the ℓ_2 norm. Then f^* is $(1/\alpha)$-smooth in the ℓ_2 norm.*

Proof. Given $y \in \mathbb{R}^d$, we have

$$f^*(y) = \sup_{x \in \operatorname{dom}(f)} \left\{ y^\top x - f(x) \right\}.$$

Note that

$$\begin{aligned} x^* \in \partial f^*(y) \quad &\leftrightarrow \quad y \in \partial f(x^*) \\ &\leftrightarrow \quad 0 \in y - \partial f(x^*) \\ &\leftrightarrow \quad x^* \in \underset{x \in \operatorname{dom}(f)}{\operatorname{argmax}} \left\{ y^\top x - f(x) \right\}. \end{aligned}$$

Since f is strongly convex, there exists a unique maximizer x^* for the supremum. This implies that the subdifferential of f^* contains a unique point, and therefore, f^* is differentiable.

Let $y_1 \in \partial f(x_1)$ and $y_2 \in \partial f(x_2)$. Since f is α-strongly convex, we have

$$f(x_1) \geq f(x_2) + y_2^\top (x_1 - x_2) + \frac{\alpha}{2}\|x_1 - x_2\|_2^2,$$
$$f(x_2) \geq f(x_1) + y_1^\top (x_2 - x_1) + \frac{\alpha}{2}\|x_2 - x_1\|_2^2.$$

Summing up these two inequalities, we obtain

$$(y_1 - y_2)^\top (x_1 - x_2) \geq \alpha \|x_1 - x_2\|_2^2.$$

Hence,

$$\|x_1 - x_2\|_2 \leq \frac{1}{\alpha}\|y_1 - y_2\|_2.$$

As $y_1 \in \partial f(x_1)$ and $y_2 \in \partial f(x_2)$, it follows that $x_1 = \nabla f^*(y_1)$ and $x_2 = \nabla f^*(y_2)$. Therefore,

$$\|\nabla f^*(y_1) - \nabla f^*(y_2)\|_2 \leq \frac{1}{\alpha}\|y_1 - y_2\|_2,$$

which implies that f^* is $(1/\alpha)$-smooth in the ℓ_2 norm. □

Remember that the subgradient method for strongly convex functions guarantees a convergence rate of $O(1/T)$. However, the dual problem of a strongly convex function minimization is a smooth convex function minimization, for which the accelerated gradient method guarantees a convergence rate of $O(1/T^2)$.

Theorem 20.2. *Let $f : \mathbb{R}^d :\to \mathbb{R}$ be a closed convex β-smooth function in the ℓ_2 norm. Then f^* is $(1/\beta)$-strongly convex in the ℓ_2 norm.*

Proof. To show that f^* is $(1/\beta)$-strongly convex in the ℓ_2 norm, we will argue that

$$h(y) = f^*(y) - \frac{1}{2\beta}\|y\|_2^2$$

is convex. Note that

$$\partial h(y) = \partial f^*(y) - \frac{1}{\beta}y.$$

We will use the fact that if ∂h is monotone, then h is convex. In other words, it is sufficient to show that for any $x_1 \in \partial f^*(y_1)$ and $x_2 \in \partial f^*(y_2)$, the following holds.

$$(y_1 - y_2)^\top ((x_1 - (1/\beta)y_1) - (x_2 - (1/\beta)y_2)) \geq 0,$$

which is equivalent to

$$(y_1 - y_2)^\top (x_1 - x_2) \geq \frac{1}{\beta}\|y_1 - y_2\|_2^2.$$

Remember that if f is β-smooth,

$$(\nabla f(x_1) - \nabla f(x_2))^\top (x_1 - x_2) \geq \frac{1}{\beta}\|\nabla f(x_1) - \nabla f(x_2)\|_2^2.$$

Moreover, for any $x_1 \in \partial f^*(y_1)$ and $x_2 \in \partial f^*(y_2)$, we have $y_1 = \nabla f(x_1)$ and $y_2 = \nabla f(x_2)$. Then the above inequality can be rewritten as

$$(y_1 - y_2)^\top (x_1 - x_2) \geq \frac{1}{\beta}\|y_1 - y_2\|_2^2,$$

as required. □

2.3 Dual gradient method for separable problems

We can use dual methods when the objective is separable while there is a system of linking constraints. We consider

$$\begin{array}{rl} \text{minimize} & f_1(x_1) + f_2(x_2) \\ \text{subject to} & A_1x_1 + A_2x_2 = b. \end{array}$$

Let us derive its dual. The Lagrangian dual function is given by

$$\begin{aligned}
&\inf_{x_1,x_2} \left\{ f_1(x_1) + f_2(x_2) + \mu^\top (A_1x_1 + A_2x_2 - b) \right\} \\
&= -b^\top \mu + \inf_{x_1} \left\{ f_1(x_1) + \mu^\top A_1 x_1 \right\} + \inf_{x_2} \left\{ f_2(x_2) + \mu^\top A_2 x_2 \right\} \\
&= -b^\top \mu - \sup_{x_1} \left\{ -f_1(x_1) + (-A_1^\top \mu)^\top x_1 \right\} - \sup_{x_2} \left\{ -f_2(x_2) + (-A_2^\top \mu)^\top x_2 \right\} \\
&= -b^\top \mu - f_1^*(-A_1^\top \mu) - f_2^*(-A_2^\top \mu).
\end{aligned}$$

Therefore, the Lagrangian dual problem is given by

$$\text{maximize} \quad -f_1^*(-A_1^\top \mu) - f_2^*(-A_2^\top \mu) - b^\top \mu.$$

Again, this problem is equivalent to the following convex minimization problem.

$$(-1) \quad \times \quad \text{minimize} \quad f_1^*(-A_1^\top \mu) + f_2^*(-A_2^\top \mu) + b^\top \mu.$$

Given μ_t, let $g_t \in \partial \left(f_1^*(-A_1^\top \mu_t) + f_2^*(-A_2^\top \mu_t) + b^\top \mu_t \right)$. We can argue that

$$\partial \left(f_1^*(-A_1^\top \mu_t) + f_2^*(-A_2^\top \mu_t) + b^\top \mu_t \right) = -A_1 \partial f_1^*(-A_1^\top \mu_t) - A_2 \partial f_2^*(-A_2^\top \mu_t) + b.$$

Note that $x_{1,t} \in \partial f_1^*(-A_1^\top \mu_t)$ if and only if $-A_1^\top \mu_t \in \partial f_1(x_{1,t})$. This is equvialent to

$$x_{1,t} \in \operatorname*{argmin}_{x_1} \left\{ f_1(x_1) + \mu_t^\top A_1 x_1 \right\}.$$

Similarly, $x_{2,t} \in \partial f_2^*(-A_2^\top \mu_t)$ if and only if

$$x_{2,t} \in \operatorname*{argmin}_{x_2} \left\{ f_2(x_2) + \mu_t^\top A_2 x_2 \right\}.$$

Therefore, the subgradient method applied to the dual problem proceeds with the following update rule.

$$\mu_{t+1} = \mu_t + \eta_t (A_1 x_{1,t} + A_2 x_{2,t} - b)$$

where

$$\begin{aligned}
x_{1,t} &\in \operatorname*{argmin}_{x_1} \left\{ f_1(x_1) + \mu_t^\top A_1 x_1 \right\}, \\
x_{2,t} &\in \operatorname*{argmin}_{x_2} \left\{ f_2(x_2) + \mu_t^\top A_2 x_2 \right\}.
\end{aligned}$$

Here, at each iteration, computing the iterates $x_{1,t}$ and $x_{2,t}$ can be done in parallel. For the primal problem, the variables x_1 and x_2 are connected through the constraints $A_1x_1 + A_2x_2 = b$. However, for the dual method, we separate the variables and x_1 and x_2 by the Lagrangian multiplier.

Algorithm 2 Subgradient method for the dual problem of a separable minimization

Initialize μ_1.
for $t = 1, \dots, T-1$ **do**
  Obtain $x_{1,t} \in \operatorname{argmin}_{x_1} \left\{ f_1(x_1) + \mu_t^\top A_1 x_1 \right\}$ and $x_{2,t} \in \operatorname{argmin}_{x_2} \left\{ f_2(x_2) + \mu_t^\top A_2 x_2 \right\}$.
  $\mu_{t+1} = \mu_t + \eta_t (A_1 x_{1,t} + A_2 x_{2,t} - b)$ for a step size $\eta_t > 0$.
end for

3 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

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

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

3.1 Convexity and smoothness

Proposition 20.3. *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 20.4. *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 20.4, we deduce the the Moreau-Yosida smoothing is smooth.

Proposition 20.5. *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 20.4 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. $\square$

Let us consider an example.

Example 20.6. 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 20.1[1]).

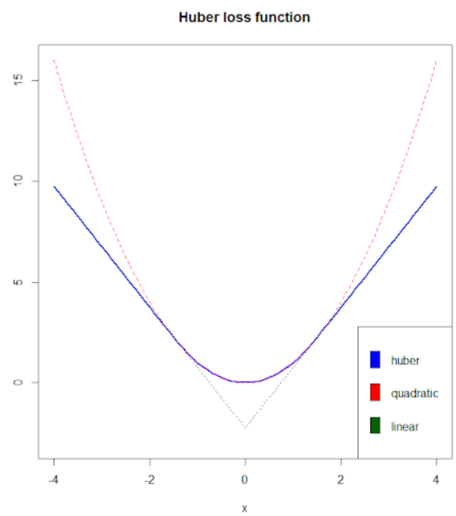

Figure 20.1: Huber loss

3.2 Optimization of the Moreau envelope

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

Proposition 20.7. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex. Then*

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

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

Proof. By Proposition 20.5, 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 20.4 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

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

so

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

as required. □

Proposition 20.8. *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 20.7, 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 20.3. Hence, we can attempt to solve the problem by gradient descent. By Proposition 20.7, 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 20.5. Hence, the gradient descent update rule proceeds with step size η given as follows

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

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.