1 Outline

In this lecture, we study

- the dual gradient method for separable problems,
- the augmented Lagrangian method, and
- the dual of composite minimization problems.

2 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

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

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 equivalent 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

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

Algorithm 1 Subgradient method for the dual problem of a separable minimization

Initialize μ_1.
for $t = 1, \ldots, T-1$ **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

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 $A_1 x_1 + A_2 x_2 = b$. However, for the dual method, we separate the variables and x_1 and x_2 by the Lagrangian multiplier.

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

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

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 2 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 method 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 Moreau-Yosida smoothing of the dual

In the previous section, we saw that the proximal point algorithm applied to the dual is equivalent to the gradient method applied to the dual of

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

From the previous lecture, we learned that the proximal point algorithm is equivalent to the gradient method applied to the Moreau-Yosida smoothing. Does this imply that the dual of the augmented problem $\min\{f(x) + \frac{\eta}{2}\|Ax - b\|_2^2 : Ax = b\}$ is equivalent to the Moreau-Yosida smoothing of the dual of the original problem $\min\{f(x) : Ax = b\}$?

Let us derive the Moreau-Yosida smoothing of the dual of the original problem $\min\{f(x) : Ax = b\}$. Recall that the dual of

$$\min\{f(x) : \ Ax = b\}$$

is equivalent to

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

Then the Moreau-Yosida smoothing of the dual $\min\{h(\mu)\}$ is given by

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

where

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

is the Moreau-Yosida smoothing of $h(\mu)$.

Next we claim that $\min\{h_\eta(\mu)\}$ is equal to the dual of

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

To show this, we take the dual of $\min\{h_\eta(\mu)\}$. Note that

$$\text{minimize } h_\eta(\mu) \quad = \quad \text{minimize } h_\eta(-(-I)^\top \mu) + 0^\top \mu.$$

Therefore, the dual of $\min\{h_\eta(\mu)\}$ is

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

Recall that

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

Note that

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

where the last equality follows from strong duality. Then

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

This implies that the dual problem is equivalent to

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

4 Dual of composite minimization

We consider

$$\text{minimize} \quad f(x) + g(Ax),$$

which is equivalent to

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

Moreover, it can be rewritten as

$$\begin{aligned} \text{minimize} \quad & f(x) + g(y) \\ \text{subject to} \quad & Ax - y = 0. \end{aligned}$$

Then we may apply the dual gradient method developed for separable objective functions.

Algorithm 3 Dual gradient method for composite problems

Initialize μ_1.
for $t = 1, \ldots, T-1$ **do**
  Obtain $x_t \in \operatorname{argmin}_x f(x) + \mu_t^\top Ax$ and $y_t \in \operatorname{argmin}_y g(y) - \mu_t^\top y$.
  Update $\mu_{t+1} = \mu_t + \eta_t(Ax_t - y_t)$ for a step size $\eta_t > 0$.
end for

Basically, at each iteration, we minimize the Lagrangian function at $\mu = \mu_t$:

$$f(x) + g(y) + \mu_t^\top (Ax - y).$$

Instead, the augmented Lagrangian method considers the augmented Lagrangian function given by

$$f(x) + g(y) + \mu_t^\top (Ax - y) + \frac{\eta}{2}\|Ax - y\|_2^2.$$

Here, μ_t changes over iterations while η remains constant.

Algorithm 4 Augmented Lagrangian method for composite problems

Initialize μ_1.
for $t = 1, \ldots, T-1$ **do**
  Obtain $(x_t, y_t) \in \operatorname{argmin}_{(x,y)} f(x) + g(y) + \mu_t^\top (Ax - y) + \frac{\eta}{2}\|Ax - y\|_2^2$,
  Update $\mu_{t+1} = \mu_t + \eta(Ax_t - y_t)$.
end for

As we observed that the augmented Lagrangian method is equivalent to the proximal point algorithm applied to the dual, we will check this for the composite optimization problem as well.

4.1 Proximal gradient applied to the dual

Let us apply the proximal gradient method to the dual of the composite problem. Throughout this subsection, let us assume that f^* is differentiable. Again, the dual is given by

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

The proximal gradient method proceeds with

$$\mu_{t+1} = \operatorname{prox}_{\eta g^*}\left(\mu_t + \eta A \nabla f^*(-A^\top \mu_t)\right)$$

since the gradient of $h(\mu) = f^*(-A^\top \mu)$ is $\nabla h(\mu) = -A\nabla f^*(-A^\top \mu)$. Moreover, $x_t = \nabla f^*(-A^\top \mu_t)$ if and only if $-A^\top \mu_t \in \partial f(x_t)$ which is equivalent to $x_t \in \operatorname{argmin}_x f(x) + \mu_t^\top Ax$. Hence, the update rule is equivalent to

$$\begin{aligned} x_t &\in \operatorname*{argmin}_x f(x) + \mu_t^\top Ax, \\ \mu_{t+1} &= \operatorname{prox}_{\eta g^*}\left(\mu_t + \eta A x_t\right). \end{aligned}$$

Furthermore, by the Moreau decomposition theorem, it follows that

$$\mu_{t+1} = \mu_t + \eta A x_t - \eta \operatorname{prox}_{g/\eta}(\mu_t/\eta + Ax_t).$$

Here, $y_t = \operatorname{prox}_{g/\eta}(\mu_t/\eta + Ax_t)$ if and only if

$$\frac{\mu_t}{\eta} + Ax_t - y_t \in \frac{1}{\eta}\partial g(y_t)$$

which is equivalent to

$$y_t \in \operatorname*{argmin}_y \left\{ g(y) + \mu_t^\top (Ax_t - y) + \frac{\eta}{2}\|Ax_t - y\|_2^2 \right\}.$$

Therefore, the proximal gradient descent applied to the dual is given by the following pseudo-code.

Algorithm 5 Proximal gradient for composite problems

```
Initialize μ_1.
for t = 1, ..., T − 1 do
    Obtain x_t ∈ argmin_x f(x) + μ_t^⊤ Ax,
    Obtain y_t ∈ argmin_y {g(y) + μ_t^⊤ (Ax_t − y) + (η/2)‖Ax_t − y‖_2^2},
    Update μ_{t+1} = μ_t + η(Ax_t − y_t).
end for
```