DSCC/CSC 435 & ECE 412 Optimization for Machine Learning — Lecture 10

Augmented Lagrangian

Lecturer: Jiaming Liang — *October 24, 2023*

1 Augmented Lagrangian

In this section, we are interested in solving optimization problems with simple inequality contraints. We begin our discussion on solving constrained optimization using augmented Lagrangian method by first presenting the dual ascent method. Intuitively, dual ascent is the counterpart of (primal) gradient descent in the dual space.

1.1 Dual ascent

Consider the problem

$$\begin{aligned} &\min_{x\in\mathbb{R}^n} f(x) \\ &\text{s.t. } Ax = b \end{aligned}$$

where $A : \mathbb{R}^m \to \mathbb{R}^n$, $b \in \mathbb{R}^m$, and f is a closed and convex function. We define the Lagrangian as

$$L(x, y) = f(x) + y^\top (Ax - b),$$

and the dual function is

$$\begin{aligned} d(y) &= \inf_x L(x, y) \\ &= \inf_x \left\{ f(x) + y^\top Ax \right\} - y^\top b \\ &= -\sup_x \left\{ (-A^\top y)^\top x - f(x) \right\} - y^\top b \\ &= -f^*(-A^\top y) - y^\top b, \end{aligned} \tag{1}$$

where f^* denotes the conjugate of f. Thus, the dual problem is

$$\max_{y\in\mathbb{R}^n} d(y).$$

An iteration of the dual ascent method reads as

$$y_{k+1} = y_k + \alpha_k d'(y_k),$$

where (in view of (1))

$$d'(y_k) \in A\partial f^*(-A^\top y_k) - b.$$

It follows from Corollary 1 of Lecture 5 and (1) that

$$d'(y_k) = Ax_k - b$$

where

$$x_k \in \operatorname{Argmin}_{x\in\mathbb{R}^n}\{f(x) + y_k^\top(Ax - b)\}.$$

Summarizing the steps above, we have the dual ascent method.

Algorithm 1 Dual ascent method

Input: Initial point $y_0 \in \mathbb{R}^n$
for $k \geq 0$ **do**
  Step 1. Compute $x_{k+1} = \operatorname{argmin}_{x\in\mathbb{R}^n} L(x, y_k)$.
  Step 2. Choose $\alpha_k > 0$ and set $y_{k+1} = y_k + \alpha_k(Ax_{k+1} - b)$.
end for

1.2 Method of Multipliers

The method of multipliers shares the same idea as dual ascent, but we augment the Lagrangian to make the primal update more robust. It is clear that

$$\begin{aligned} &\min_{x\in\mathbb{R}^n} f(x) \\ &\text{s.t. } Ax = b \end{aligned}$$

and

$$\begin{aligned} &\min_{x\in\mathbb{R}^n} f(x) + \frac{\rho}{2}\|Ax - b\|^2 \\ &\text{s.t. } Ax = b \end{aligned}$$

have exactly the same set of solutions for all $\rho \geq 0$. The Lagrangian for the second program is

$$L_\rho(x, y) = f(x) + \frac{\rho}{2}\|Ax - b\|^2 + y^\top(Ax - b).$$

This is called the *augmented Lagrangian* of the original problem. Let us apply the proximal point method to the dual problem

$$y_{k+1} = \operatorname{argmax}_{y\in\mathbb{R}^m}\left\{d(y) - \frac{1}{2\rho}\|y - y_k\|^2\right\},$$

i.e.,

$$y_{k+1} = \operatorname{argmin}_{y\in\mathbb{R}^m} \left\{ -d(y) + \frac{1}{2\rho}\|y - y_k\|^2 \right\}.$$

The optimality condition is

$$0 \in -d'(y_{k+1}) + \frac{1}{\rho}(y_{k+1} - y_k),$$

and hence (as in the derivation of dual ascent)

$$y_{k+1} = y_k + \rho(Ax_{k+1} - b)$$

where

$$\begin{aligned} x_{k+1} &\in \operatorname{Argmin}_{x\in\mathbb{R}^n} \{f(x) + y_{k+1}^\top (Ax - b)\} \\ &= \operatorname{Argmin}_{x\in\mathbb{R}^n} \{f(x) + \langle y_{k+1}, Ax\rangle\} \\ &= \operatorname{Argmin}_{x\in\mathbb{R}^n} \{f(x) + \langle y_k + \rho(Ax_{k+1} - b), Ax\rangle\}. \end{aligned}$$

Moreover, on the one hand, the optimality condition of the above program is

$$0 \in A^\top [y_k + \rho(Ax_{k+1} - b)] + \partial f(x_{k+1}),$$

which, on the other hand, is also the optimality condition of

$$x_{k+1} \in \operatorname{Argmin}_{x\in\mathbb{R}^n} \left\{ L_\rho(x, y_k) = f(x) + y_k^\top (Ax - b) + \frac{\rho}{2}\|Ax - b\|^2 \right\}.$$

Summarizing the discussion above, we have the method of multipliers.

Algorithm 2 Method of multipliers/Augmented Lagrangian method

Input: Initial point $y_0 \in \mathbb{R}^n$
for $k \geq 0$ **do**
  Step 1. Compute $x_{k+1} = \operatorname{argmin}_{x\in\mathbb{R}^n} L_\rho(x, y_k)$.
  Step 2. Set $y_{k+1} = y_k + \rho(Ax_{k+1} - b)$.
end for

It is worth noting that the method of multipliers is the proximal point method applied to the dual problem.

2 Alternating Direction Method of Multipliers

Consider the problem

$$\begin{aligned} &\min f_1(x) + f_2(Ax) \\ &\text{s.t. } x \in X, \quad Ax \in Z, \end{aligned}$$

we reformulate the problem as

$$\begin{aligned}
&\min f_1(x) + f_2(z)\\
&\text{s.t. } x \in X, \quad z \in Z, \quad Ax = z.
\end{aligned}$$

The augmented Lagarangian is

$$L_\rho(x, z, y) = f_1(x) + f_2(z) + \langle y, Ax - z\rangle + \frac{\rho}{2}\|Ax - z\|^2.$$

The alternating direction method of multipliers (ADMM) to solve the above augmented Lagarangian has three updates

$$\begin{aligned}
x_{k+1} &\in \operatorname{Argmin}_{x\in\mathbb{R}^n} \{L_\rho(x, z_k, y_k)\}\\
z_{k+1} &\in \operatorname{Argmin}_{z\in\mathbb{R}^m} \{L_\rho(x_{k+1}, z, y_k)\}\\
y_{k+1} &= y_k + \rho(Ax_{k+1} - z_{k+1}).
\end{aligned}$$

2.1 ADMM as an instance of IPP

In this subsection, we show that ADMM is an instance of the IPP framework. Define

$$\begin{aligned}
&\tilde{y}_{k+1} = y_k + \rho(Ax_{k+1} - z_k), \quad (\tilde{x}_{k+1}, \tilde{z}_{k+1}) = (x_{k+1}, z_{k+1}),\\
&\tilde{s}_{k+1} = (\tilde{x}_{k+1}, \tilde{z}_{k+1}, \tilde{y}_{k+1}), \quad s_{k+1} = (x_{k+1}, z_{k+1}, y_{k+1}).
\end{aligned}$$

Then, the three updates of ADMM can be rewritten as

$$\begin{aligned}
\partial f_1(\tilde{x}_{k+1}) + A^\top \tilde{y}_{k+1} &\ni 0\\
\partial f_2(\tilde{z}_{k+1}) - \tilde{y}_{k+1} + \rho(z_{k+1} - z_k) &\ni 0\\
-A\tilde{x}_{k+1} + \tilde{z}_{k+1} + \frac{y_{k+1} - y_k}{\rho} &= 0,
\end{aligned}$$

which are equivalent to

$$T^{\varepsilon_{k+1}}(\tilde{s}_{k+1}) \ni \frac{\nabla w(s_k) - \nabla w(s_{k+1})}{\lambda_{k+1}}$$

with $\lambda_{k+1} = 1$, $\varepsilon_{k+1} = 0$,

$$T(s) = T(x, z, y) := \begin{bmatrix} 0 & 0 & A^\top \\ 0 & 0 & -I \\ -A & -I & 0 \end{bmatrix} \begin{bmatrix} x \\ z \\ y \end{bmatrix} + \begin{bmatrix} \partial f_1(x) \\ \partial f_2(z) \\ 0 \end{bmatrix}$$

and

$$w(s) = w(x, z, y) = \frac{\rho}{2}\|z\|^2 + \frac{1}{2\rho}\|y\|^2.$$

Recall that the Bregman divergence for w is

$$D_w(s', s) = w(s') - \ell_w(s'; s).$$

We have shown the inclusion in the IPP framework. It remains to prove the inequality

$$D_w(\tilde{s}_{k+1}, s_{k+1}) + \lambda_{k+1}\varepsilon_{k+1} \le \sigma D_w(\tilde{s}_{k+1}, s_k).$$

By the definition, we have

$$\begin{aligned} D_w(\tilde{s}_{k+1}, s_{k+1}) &= \frac{\rho}{2}\|z_{k+1} - \tilde{z}_{k+1}\|^2 + \frac{1}{2\rho}\|y_{k+1} - \tilde{y}_{k+1}\|^2 \\ &= \frac{1}{2\rho}\|y_{k+1} - \tilde{y}_{k+1}\|^2 = \frac{\rho}{2}\|z_{k+1} - z_k\|^2 \end{aligned}$$

and

$$\begin{aligned} D_w(\tilde{s}_{k+1}, s_k) &= \frac{\rho}{2}\|z_k - \tilde{z}_{k+1}\|^2 + \frac{1}{2\rho}\|y_k - \tilde{y}_{k+1}\|^2 \\ &= \frac{\rho}{2}\|z_k - z_{k+1}\|^2 + \frac{1}{2\rho}\|y_k - \tilde{y}_{k+1}\|^2. \end{aligned}$$

So the IPP inequality holds with $\lambda_{k+1} = 1$, $\varepsilon_{k+1} = 0$, and $\sigma = 1$.

2.2 Examples of ADMM

Least absolute deviations

Consider the problem

$$\begin{aligned} &\min \|Ax - b\|_1 \\ &\text{s.t. } x \in \mathbb{R}^n \end{aligned}$$

where A is an $m \times n$ matrix of rank n and $b \in \mathbb{R}^m$ is a given vector. We reformulate the problem as

$$\begin{aligned} &\min f_1(x) + f_2(z) \\ &\text{s.t. } Ax - b = z, \end{aligned}$$

where $f_1(x) \equiv 0$ and $f_2(z) = \|z\|_1$. The augmented Lagrangian is given by

$$L_\rho(x, z, y) = \|z\|_1 + \langle y, Ax - z - b\rangle + \frac{\rho}{2}\|Ax - z - b\|^2.$$

The ADMM iteration takes the form

$$\begin{aligned} x_{k+1} &= (A^\top A)^{-1}A^\top \left(z_k + b - \frac{y_k}{\rho}\right) \\ z_{k+1} &\in \operatorname{Argmin}_{z\in\mathbb{R}^m} \left\{\|z\|_1 - \langle y_k, z\rangle + \frac{\rho}{2}\|Ax_{k+1} - z - b\|^2\right\} \\ y_{k+1} &= y_k + \rho(Ax_{k+1} - z_{k+1} - b). \end{aligned}$$

Setting $\bar{y}_k = y_k/\rho$, the iteration can be written in the notationally simpler form

$$\begin{aligned}
x_{k+1} &= (A^\top A)^{-1}A^\top\left(z_k + b - \bar{y}_k\right) \\
z_{k+1} &\in \operatorname{Argmin}_{z\in\mathbb{R}^m}\left\{\|z\|_1 + \frac{\rho}{2}\|Ax_{k+1} - z - b + \bar{y}_k\|^2\right\} \\
\bar{y}_{k+1} &= \bar{y}_k + Ax_{k+1} - z_{k+1} - b.
\end{aligned}$$

The minimization over z is expressed in terms of soft-shrinkage as

$$z_{k+1} = \mathcal{T}_{1/\rho}(Ax_{k+1} - b + \bar{y}_k)$$

where $\mathcal{T}_\lambda(y)$ is the soft-thresholding operator introduced in Lecture 6. Recall that it can be implemented component-wisely

$$\operatorname{prox}_{\lambda|\cdot|}(y) = \mathcal{T}_\lambda(y) = [|y| - \lambda]_+ \operatorname{sgn}(y) = \begin{cases} y - \lambda, & y \geq \lambda \\ 0, & |y| < \lambda \\ y + \lambda, & y \leq -\lambda \end{cases}$$

Lasso

Consider the problem

$$\min_x \frac{1}{2}\|Ax - b\|_2^2 + \tau\|x\|_1.$$

Taking

$$f_1(x) = \frac{1}{2}\|Ax - b\|_2^2, \quad f_2(z) = \tau\|z\|_1,$$

we can rewrite the problem as

$$\begin{aligned}
&\min f_1(x) + f_2(z) \\
&\text{s.t. } x - z = 0.
\end{aligned}$$

The augmented Lagrangian is given by

$$L_\rho(x, z, y) = \frac{1}{2}\|Ax - b\|_2^2 + \tau\|z\|_1 + \langle y, x - z\rangle + \frac{\rho}{2}\|x - z\|_2^2.$$

The ADMM iteration takes the form

$$\begin{aligned}
x_{k+1} &= (A^\top A + \rho I)^{-1}\left(A^\top b + \rho z_k - y_k\right) \\
z_{k+1} &\in \operatorname{Argmin}_{z\in\mathbb{R}^m}\left\{\tau\|z\|_1 - \langle y_k, z\rangle + \frac{\rho}{2}\|x_{k+1} - z\|^2\right\} \\
y_{k+1} &= y_k + \rho(x_{k+1} - z_{k+1}).
\end{aligned}$$

Setting $\bar{y}_k = y_k/\rho$, the iteration can be written in the notationally simpler form

$$\begin{aligned}
x_{k+1} &= (A^\top A + \rho I)^{-1} \left(A^\top b + \rho(z_k - \bar{y}_k) \right) \\
z_{k+1} &\in \operatorname{Argmin}_{z\in\mathbb{R}^m} \left\{ \tau \|z\|_1 + \frac{\rho}{2} \|x_{k+1} - z + \bar{y}_k\|^2 \right\} \\
\bar{y}_{k+1} &= \bar{y}_k + x_{k+1} - z_{k+1}.
\end{aligned}$$

The minimization over z is expressed in terms of soft-shrinkage as

$$z_{k+1} = \mathcal{T}_{\tau/\rho}(x_{k+1} + \bar{y}_k)$$

where $\mathcal{T}_\lambda(y)$ is the soft-thresholding operator.

2.3 ADMM applied to separable problems

Separable problem of the form

$$\begin{aligned}
&\min \sum_{i=1}^m f_i(x_i) \\
&\text{s.t. } \sum_{i=1}^m A_i x_i = b, \quad x_i \in X_i,\ i = 1, \ldots, m,
\end{aligned}$$

where $f_i : \mathbb{R}^{n_i} \mapsto \mathbb{R}$ are convex and X_i are closed convex sets. Since the primary attractive feature of ADMM is that it decouples the augmented Lagrangian optimization calculations, it is natural to consider its application to this problem.

An idea that readily comes to mind is to form the augmented Lagrangian

$$L_\rho(x_1, \ldots, x_m, y) = \sum_{i=1}^m f_i(x_i) + \langle y, \sum_{i=1}^m A_i x_i - b\rangle + \frac{\rho}{2} \left\| \sum_{i=1}^m A_i x_i - b \right\|^2,$$

and use an ADMM-like iteration, whereby we minimize L_ρ sequentially w.r.t. $x_1, \ldots, x_m$, i.e.,

$$x_i^{k+1} \in \operatorname{Argmin}_{x_i\in X_i} L_\rho(x_1^{k+1}, \ldots, x_{i-1}^{k+1}, x_i, x_{i+1}^k, \ldots, x_m^k, y^k), \quad i = 1, \ldots, m,$$

and follow these minimization with the multiplier iteration

$$y^{k+1} = y^k + \rho \left(\sum_{i=1}^m A_i x_i^{k+1} - b \right).$$

When $m = 1$, this is the method of multipliers. When $m = 2$, this is ADMM. On the other hand, when $m \geq 3$, this is a special case of the ADMM that we have discussed and the method is not convergent in general.

In what follows, we develop an ADMM by formulating the separable problem as a two-block minimization problem. By introducing additional variables $z_1, \ldots, z_m$, we rewrite the problem as

$$\begin{aligned} &\min \sum_{i=1}^m f_i(x_i) \\ &\text{s.t. } A_i x_i = z_i,, \quad x_i \in X_i, \ i = 1, \ldots, m, \\ &\qquad \sum_{i=1}^m z_i = b. \end{aligned}$$

We denote $x = (x_1, \ldots, x_m)$ and $z = (z_1, \ldots, z_m)$. We view $X = X_1 \times \ldots \times X_m$ as a constraint set for x and view

$$Z = \left\{ z : \sum_{i=1}^m z_i = b \right\}$$

as a constraint set for z. We introduce a multiplier y_i for each of the equality constraints $A_i x_i = z_i$. The augmented Lagrangian has the separable form

$$L_\rho(x, z, y) = \sum_{i=1}^m \left(f_i(x_i) + \langle y_i, A_i x_i - z_i \rangle + \frac{\rho}{2} \|A_i x_i - z_i\|^2 \right),$$

and the ADMM is given by

$$\begin{aligned} x_i^{k+1} &\in \operatorname{Argmin}_{x_i \in X_i} \left\{ f_i(x_i) + \langle y_i^k, A_i x_i - z_i^k \rangle + \frac{\rho}{2} \|A_i x_i - z_i^k\|^2 \right\} \\ z_{k+1} &\in \operatorname{Argmin}_{\sum_{i=1}^m z_i = b} \left\{ \sum_{i=1}^m \langle y_i^k, A_i x_i^{k+1} - z_i \rangle + \frac{\rho}{2} \|A_i x_i^{k+1} - z_i\|^2 \right\} \\ y_i^{k+1} &= y_i^k + \rho(A_i x_i^{k+1} - z_i^{k+1}). \end{aligned}$$

We will show how to simplify the algorithm. Introducing a multiplier λ^{k+1} for the constraint $\sum_{i=1}^m z_i = b$, we have the Lagrangian corresponding to the z-minimization

$$\sum_{i=1}^m \left(\langle y_i^k, A_i x_i^{k+1} - z_i \rangle + \frac{\rho}{2} \|A_i x_i^{k+1} - z_i\|^2 + \langle \lambda^{k+1}, z_i \rangle \right) - \langle \lambda^{k+1}, b \rangle.$$

Settinf its gradient w.r.t. z_i to zero, we see z_i^{k+1} is given by

$$z_i^{k+1} = A_i x_i^{k+1} + \frac{y_i^k - \lambda^{k+1}}{\rho}.$$

A key observation is that

$$\lambda^{k+1} = y_i^k + \rho(A_i x_i^{k+1} - z_i^{k+1}), \quad i = 1, \ldots, m.$$

Hence,

$$y_i^{k+1} = \lambda^{k+1}, \quad i = 1, \ldots, m.$$

Given z^k and λ^k (which is equal to y_i^k for every i), we have

$$x_i^{k+1} \in \operatorname{Argmin}_{x_i \in X_i} \left\{ f_i(x_i) + \langle \lambda^k, A_i x_i - z_i^k \rangle + \frac{\rho}{2} \| A_i x_i - z_i^k \|^2 \right\} \tag{2}$$

and

$$z_i^{k+1} = A_i x_i^{k+1} + \frac{\lambda^k - \lambda^{k+1}}{\rho}. \tag{3}$$

Note

$$\sum_{i=1}^{m} \left(A_i x_i^{k+1} + \frac{\lambda^k - \lambda^{k+1}}{\rho} \right) = \sum_{i=1}^{m} z_i^{k+1} = b,$$

so

$$\lambda^{k+1} = \lambda^k + \frac{\rho}{m} \left(\sum_{i=1}^{m} A_i x_i^{k+1} - b \right). \tag{4}$$

In summary, given (x^k, z^k, λ^k), the iteration to obtain $(x^{k+1}, z^{k+1}, \lambda^{k+1})$ by applying (2), (3), and (4).