From Dual Descent to ADMM

Hailiang ZHAO @ ZJU-CS

http://hliangzhao.me

November 1, 2022

Outline

Outline

Dual Descent

Consider the following linearly constrained convex problem:

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}), \qquad s.t. \quad \mathbf{A}\boldsymbol{x} = \boldsymbol{b}, \tag{1}$$

where $f(\boldsymbol{x})$ is proper, closed, and convex. The corresponding Lagrangian function is

$$L(\boldsymbol{x}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) + \langle \boldsymbol{\lambda}, \mathbf{A}\boldsymbol{x} - \boldsymbol{b} \rangle. \tag{2}$$

The dual function is

$$\begin{aligned} d(\boldsymbol{\lambda}) &= \min_{\boldsymbol{x}} L(\boldsymbol{x}, \boldsymbol{\lambda}) \\ &= -\max_{\boldsymbol{x}} \big(-f(\boldsymbol{x}) - \langle \mathbf{A}^T \boldsymbol{\lambda}, \boldsymbol{x} \rangle \big) - \langle \boldsymbol{\lambda}, \boldsymbol{b} \rangle \\ &= -f^*(-\mathbf{A}^T \boldsymbol{\lambda}) - \langle \boldsymbol{\lambda}, \boldsymbol{b} \rangle, \end{aligned} \tag{3}$$

where $f^*(\cdot)$ is the conjugate function.

Dual Descent

$d(\boldsymbol{\lambda})$ is concave, and its domain is $\mathcal{D} = \{\boldsymbol{\lambda} \mid d(\boldsymbol{\lambda}) > -\infty\}$. The dual problem is

$$\max_{\boldsymbol{\lambda} \in \mathcal{D}} d(\boldsymbol{\lambda}). \tag{4}$$

With the optimal solution of the dual problem $\boldsymbol{\lambda}^*$, we can recover the optimal solution of the primal problem as

$$\boldsymbol{x}^* = \operatorname*{argmin}_{\boldsymbol{x}} L(\boldsymbol{x}, \boldsymbol{\lambda}^*), \tag{5}$$

as the strong duality holds. According to Danskin's Theorem (Theorem B.1) and Proposition B.3 (*see the Preliminaries slide*), we know that $d(\boldsymbol{\lambda})$ is differentiable and $\nabla d(\lambda^k) = \mathbf{A}\boldsymbol{x}^{k+1} - \boldsymbol{b}$, where $\boldsymbol{x}^{k+1}$ is the minimizer of $L(\boldsymbol{x}, \boldsymbol{\lambda}^k)$.

Dual Descent

Thus, we can use the following iterations to solve the primal problem:

$$\boldsymbol{x}^{k+1} = \operatorname*{argmin}_{\boldsymbol{x}} L(\boldsymbol{x}, \boldsymbol{\lambda}^k) \tag{6}$$

$$\boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k + \alpha_k(\mathbf{A}\boldsymbol{x}^{k+1} - \boldsymbol{b}), \tag{7}$$

where α_k is the step size of the gradient ascent method.

The first step is a minimization step in the primal space, while the second step is the update in the dual space. We call this algorithm Dual Descent.

Outline

Augmented Lagrangian Method

The disadvantage of the dual ascent method is that to make the dual function differentiable, we require f to be strictly convex. Otherwise, (7) is a subgradient ascent of the dual function, which converges much slower. Even worse, the subproblem (6) may not have a solution. To address these issues, we can use the augmented Lagrangian method.

Firstly, we introduce the augmented Lagrangian function:

$$L_\beta(\boldsymbol{x}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) + \langle \boldsymbol{\lambda}, \mathbf{A}\boldsymbol{x} - \boldsymbol{b} \rangle + \frac{\beta}{2} \|\mathbf{A}\boldsymbol{x} - \boldsymbol{b}\|^2. \tag{8}$$

β is called the penalty parameter. The associated dual function is

$$d_\beta(\boldsymbol{\lambda}) = \min_{\boldsymbol{x}} L_\beta(\boldsymbol{x}, \boldsymbol{\lambda}). \tag{9}$$

Augmented Lagrangian Method

Because the optimal solution of $\boldsymbol{x}^*$ satisfies that $\mathbf{A}\boldsymbol{x}^* = \boldsymbol{b}$, thus we have $d_\beta(\boldsymbol{\lambda}^*) \leq f(\boldsymbol{x}^*)$. Moreover, for any $\boldsymbol{\lambda}$ we have $d(\boldsymbol{\lambda}) \leq d_\beta(\boldsymbol{\lambda})$. Since $d(\boldsymbol{\lambda}^*) = f(\boldsymbol{x}^*)$, we can conclude that

$$d(\boldsymbol{\lambda}^*) = d_\beta(\boldsymbol{\lambda}^*) = f(\boldsymbol{x}^*). \tag{10}$$

In other words, the augmented term does not change the solution. However, using the augmented Lagrangian function brings great benefits: for $d_\beta(\boldsymbol{\lambda})$ to be differentiable we only require f to be convex, not strictly convex. The result is shown by the following lemma.

Augmented Lagrangian Method

Lemma 1
Let $\mathcal{D}(\boldsymbol{\lambda})$ denote the optimal solution set of $\min_{\boldsymbol{x}} L_\beta(\boldsymbol{x}, \boldsymbol{\lambda})$. Then $\mathbf{A}\boldsymbol{x}$ is invariant over $\mathcal{D}(\boldsymbol{\lambda})$. Moreover, $d_\beta(\boldsymbol{\lambda})$ is differentiable and

$$\nabla d_\beta(\boldsymbol{\lambda}) = \mathbf{A}\boldsymbol{x}(\boldsymbol{\lambda}) - \boldsymbol{b}, \tag{11}$$

where $\boldsymbol{x}(\boldsymbol{\lambda}) \in \mathcal{D}(\boldsymbol{\lambda})$ is any minimizer of $L_\beta(\boldsymbol{x}, \boldsymbol{\lambda})$. We also have that $d_\beta(\boldsymbol{\lambda})$ is $\frac{1}{\beta}$-smooth, i.e.,

$$\|\nabla d_\beta(\boldsymbol{\lambda}) - \nabla d_\beta(\boldsymbol{\lambda}')\| \leq \frac{1}{\beta}\|\boldsymbol{\lambda} - \boldsymbol{\lambda}'\|. \tag{12}$$

Augmented Lagrangian Method

Proof Sketch of Lemma 1

Suppose that $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{D}(\boldsymbol{\lambda})$ and $\mathbf{A}\boldsymbol{x} \neq \mathbf{A}\boldsymbol{x}'$. Then, according to the convexity of $L_\beta(\boldsymbol{x}, \boldsymbol{\lambda})$ we have

$$d_\beta(\boldsymbol{\lambda}) > L_\beta(\bar{\boldsymbol{x}}, \boldsymbol{\lambda}), \tag{13}$$

where $\bar{\boldsymbol{x}} := \frac{\boldsymbol{x}+\boldsymbol{x}'}{2} \in \mathcal{D}(\boldsymbol{\lambda})$. The result contradicts with the definition of $d_\beta(\boldsymbol{\lambda})$. To prove that $d_\beta(\boldsymbol{\lambda})$ is $\frac{1}{\beta}$-smooth, we need to use the fact that $\nabla d_\beta(\boldsymbol{\lambda}) = \mathbf{A}\boldsymbol{x}(\boldsymbol{\lambda}) - \boldsymbol{b}$,

$$\mathbf{0} \in \nabla L_\beta(\boldsymbol{x}, \boldsymbol{\lambda}), \text{where } \boldsymbol{x} = \operatorname*{argmin}_{\boldsymbol{x}} L_\beta(\boldsymbol{x}, \boldsymbol{\lambda}), \tag{14}$$

and the monotonicity of ∂f (*Proposition B.6 in the Preliminaries slide*).

Augmented Lagrangian Method

Applying the dual descent to $d_\beta(\boldsymbol{\lambda})$, we have

$$\boldsymbol{x}^{k+1} = \operatorname*{argmin}_{\boldsymbol{x}} L_\beta(\boldsymbol{x}, \boldsymbol{\lambda}^k) \tag{15}$$

$$\boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k + \beta(\mathbf{A}\boldsymbol{x}^{k+1} - \boldsymbol{b}), \tag{16}$$

We call it the Augmented Lagrangian Method (a.k.a. Method of Multipliers). Note that the step size in (16) is fixed as β.

Augmented Lagrangian Method

The augmented Lagrangian method can also be derived from the dual problem. With (3), the dual problem can be formulated as

$$\min_{\boldsymbol{\lambda}} \quad f^*(-\mathbf{A}^T\boldsymbol{\lambda}) + \langle \boldsymbol{\lambda}, \boldsymbol{b} \rangle. \tag{17}$$

We use the Proximal Point Method to solve it:

$$\boldsymbol{\lambda}^{k+1} = \operatorname*{argmin}_{\boldsymbol{\lambda}} \left(f^*(-\mathbf{A}^T\boldsymbol{\lambda}) + \langle \boldsymbol{\lambda}, \boldsymbol{b} \rangle + \frac{1}{2\beta} \| \boldsymbol{\lambda} - \boldsymbol{\lambda}^k \|^2 \right). \tag{18}$$

The optimality condition is

$$\mathbf{0} \in -\mathbf{A}\partial f^*\big(- \mathbf{A}^T\boldsymbol{\lambda}^{k+1}\big) + \boldsymbol{b} + \frac{1}{\beta}\big(\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^k\big). \tag{19}$$

Augmented Lagrangian Method

(19) means that there exists

$$\boldsymbol{x}^{k+1} \in \partial f^*\big(-\mathbf{A}^T\boldsymbol{\lambda}^{k+1}\big) \tag{20}$$

such that $\mathbf{0} = -\mathbf{A}\boldsymbol{x}^{k+1} + \boldsymbol{b} + \frac{1}{\beta}\big(\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^k\big)$, which leads to (16). On the other hand, according to Proposition B.7.5 and (20), we have

$$-\mathbf{A}^T\boldsymbol{\lambda}^{k+1} \in \partial f(\boldsymbol{x}^{k+1}), \tag{21}$$

which means

$$\begin{aligned}\mathbf{0} &\in \partial f(\boldsymbol{x}^{k+1}) + \mathbf{A}^T\boldsymbol{\lambda}^{k+1} \\ &= \partial f(\boldsymbol{x}^{k+1}) + \mathbf{A}^T\big(\boldsymbol{\lambda}^k + \beta(\mathbf{A}\boldsymbol{x}^{k+1} - \boldsymbol{b})\big).\end{aligned} \tag{22}$$

(22) gives $\boldsymbol{x}^{k+1} = \operatorname{argmin}_{\boldsymbol{x}} L_\beta(\boldsymbol{x}, \boldsymbol{\lambda}^k)$.

Outline

Alternating Direction Method of Multipliers

Consider a special case of problem (1), which has the following separable structure:

$$\min_{\boldsymbol{x},\boldsymbol{y}} \quad f(\boldsymbol{x}) + g(\boldsymbol{y}), \quad s.t. \quad \mathbf{A}\boldsymbol{x} + \mathbf{B}\boldsymbol{y} = \boldsymbol{b}. \tag{23}$$

Introduce the augmented Lagrangian function:

$$\begin{aligned} L_\beta(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) + g(\boldsymbol{y}) + \langle \mathbf{A}\boldsymbol{x} + \mathbf{B}\boldsymbol{y} - \boldsymbol{b}, \boldsymbol{\lambda} \rangle \\ + \frac{\beta}{2} \|\mathbf{A}\boldsymbol{x} + \mathbf{B}\boldsymbol{y} - \boldsymbol{b}\|^2. \end{aligned} \tag{24}$$

When we use the augmented Lagrangian method to solve (23), we need to solve the following subproblem:

$$(\boldsymbol{x}^{k+1}, \boldsymbol{y}^{k+1}) = \operatorname*{argmin}_{\boldsymbol{x},\boldsymbol{y}} L_\beta(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\lambda}). \tag{25}$$

Alternating Direction Method of Multipliers

Sometimes, it is much simpler when we solve (23) for $\boldsymbol{x}$ and $\boldsymbol{y}$ separately, which motivates the ADMM. Different from the augmented Lagrangian method, ADMM updates $\boldsymbol{x}$ and $\boldsymbol{y}$ in an alternating (or called sequential) fashion:

$$\boldsymbol{x}^{k+1} = \underset{\boldsymbol{x}}{\operatorname{argmin}}\, L_\beta(\boldsymbol{x}, \boldsymbol{y}^k, \boldsymbol{\lambda}^k) \tag{26}$$

$$\boldsymbol{y}^{k+1} = \underset{\boldsymbol{y}}{\operatorname{argmin}}\, L_\beta(\boldsymbol{x}^{k+1}, \boldsymbol{y}, \boldsymbol{\lambda}^k) \tag{27}$$

$$\boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k + \beta\big(\mathbf{A}\boldsymbol{x}^{k+1} + \mathbf{B}\boldsymbol{y}^{k+1} - \boldsymbol{b}\big). \tag{28}$$

ADMM is superior to the augmented Lagrangian method when the $\boldsymbol{x}$ and $\boldsymbol{y}$ subproblems can be more efficiently solved.

Outline

References

1. Lin, Zhouchen, Huan Li, and Cong Fang. *Alternating Direction Method of Multipliers for Machine Learning.* Springer Nature, 2022.
2. Boyd, Stephen, Stephen P. Boyd, and Lieven Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.