THE ALTERNATING DIRECTION METHOD OF MULTIPLIERS

Overview

ADMM is a highly versatile and efficient technique for minimizing the sum of two convex functions subject to equality constraints. It turns out that many machine learning optimization problems have this form.

Motivation

Consider the optimization problem

$$\min f(x)$$
$$\text{s.t. } Ax = b$$

where $x \in \mathbb{R}^p$, $A \in \mathbb{R}^{r \times p}$, $b \in \mathbb{R}^r$, and $f: \mathbb{R}^p \longrightarrow \mathbb{R}$ is convex. The Lagrangian is

$$L(x, \lambda) = f(x) + \lambda^T (Ax - b)$$

where $\lambda \in \mathbb{R}^r$. If strong duality holds, recall that we can recover a primal optimal point x^* from a dual optimal point λ^*. Indeed, x^* solves

$$\begin{aligned} \min_x \; & f(x) \\ \text{s.t. } & Ax = b \end{aligned} \quad = \min_x \max_\lambda L(x, \lambda)$$

$$= \max_\lambda \min_x L(x, \lambda)$$

$$= \min_x L(x, \lambda^*)$$

To solve the dual, the following iterative strategy, called <u>dual ascent</u>, can be used:

Initialize λ^0

Iterate

- $x^{k+1} = \arg\min_x L(x, \lambda^k)$ — approximates gradient of dual
- $\lambda^{k+1} = \lambda^k + \alpha^k (Ax^k - b)$ — α^k: stepsize

The λ update moves the entries of λ in the right

direction.

This procedure has desirable convergence properties under certain assumptions. A related method, called the method of multipliers, is similar but works with the so-called augmented Lagrangian:

$$L_\rho(x, \lambda) = f(x) + \lambda^T (Ax - b) + \frac{\rho}{2} \|Ax - b\|_2^2$$

where $\rho > 0$. The algorithm is

Initialize λ^0

Iterate

- $x^{k+1} = \arg\min_x \; L_\rho(x, \lambda^k)$
- $\lambda^{k+1} = \lambda^k + \rho(Ax^{k+1} - b)$

↖ Note: step size equals regularization parameter

The quadratic penalty essentially encourages x^{k+1} to approximately satisfy the constraint. The algorithm works under more relaxed assumptions when compared to dual ascent.

ADMM

ADMM applies to problems of the form

$$\min \quad f(x) + g(y)$$

$$\text{s.t.} \quad Ax + By = c$$

where $x \in \mathbb{R}^p$, $y \in \mathbb{R}^q$, $A \in \mathbb{R}^{r \times p}$, $B \in \mathbb{R}^{r \times q}$, $c \in \mathbb{R}^r$,

$f: \mathbb{R}^p \to \mathbb{R} \cup \{\infty\}$ and $g: \mathbb{R}^q \to \mathbb{R} \cup \{\infty\}$.

f and g are assumed to be convex. Define the augmented Lagrangian

$$L_\rho(x, y, \lambda) = f(x) + g(y) + \lambda^T (Ax + By - c) + \frac{\rho}{2} \|Ax + By - c\|_2^2$$

ADMM is the following algorithm.

Initialize y^0, λ^0

Iterate

- $x^{k+1} = \arg\min_x \; L_\rho(x, y^k, \lambda^k)$
- $y^{k+1} = \arg\min_y \; L_\rho(x^{k+1}, y, \lambda^k)$
- $\lambda^{k+1} = \lambda^k + \rho(Ax^{k+1} + By^{k+1} - c)$

$$\cdot \lambda^{k+1} = \lambda^k + \rho(Ax^{k+1} + By^{k+1} - c)$$

The following convergence result holds.

A convex function $f: \mathbb{R}^p \to \mathbb{R} \cup \{\infty\}$ is said to be proper if

$$\text{epi } f := \{(x,t) \in \mathbb{R}^p \times \mathbb{R} \mid f(x) \leq t\}$$

is nonempty, and closed if epi f is a closed set.

Theorem Assume f and g are closed and proper convex functions, and L_0 (L_ρ with $\rho = 0$) has at least one saddle point. Then

$$\|Ax^k + By^k - c\|_2 \longrightarrow 0$$

and

$$f(x^k) + g(y^k)$$

converges to the optimal objective function value.

Note that the result holds for any $\rho > 0$. Choice of ρ will impact the rate of convergence empirically.

There are [illegible] stopping criteria. One is

There are many possible stopping criteria. One is

$$\max\left\{\frac{\|x^{k+1}-x^k\|_2}{\|x^k\|_2}, \frac{\|y^{k+1}-y^k\|_2}{\|y^k\|_2}\right\} \leq \varepsilon$$

Proximal Operators

We can express ADMM as follows. For the x update, observe

$$x^{k+1} = \arg\min_x \left(f(x) + g(y^k) + (\lambda^k)^T(Ax+By^k-c) + \frac{\rho}{2}\|Ax+By^k-c\|_2^2 \right)$$

$$= \arg\min_x \left(f(x) + \underbrace{(\lambda^k)^T Ax + \frac{\rho}{2}\|Ax+By^k-c\|_2^2}_{\text{complete the square: these differ by a constant}} \right)$$

$$= \arg\min_x \left(f(x) + \frac{\rho}{2}\|Ax+By^k-c+\frac{1}{\rho}\lambda^k\|_2^2 \right)$$

Similarly for the y update. Thus, if we define the scaled Lagrange multiplier

$$u^k = \frac{1}{\rho}\lambda^k$$

ADMM can be expressed

ADMM can be expressed

Initialize y^0, u^0

Iterate:

- $x^{k+1} = \arg\min_x \left(f(x) + \frac{\rho}{2} \| Ax + By^k - c + u^k \|_2^2 \right)$
- $y^{k+1} = \arg\min_y \left(g(y) + \frac{\rho}{2} \| Ax^{k+1} + By - c + u^k \|_2^2 \right)$
- $u^{k+1} = u^k + Ax^{k+1} + By^{k+1} - c$

Because of this form of ADMM, an important concept is that of a proximity operator, which we now define.

For any proper closed convex f, define the proximity operator $\text{prox}_f : \mathbb{R}^p \to \mathbb{R}^p$ by

$$\text{prox}_f(v) = \arg\min_x \left(f(x) + \frac{1}{2} \| x - v \|_2^2 \right).$$

strictly convex, so minimizer is unique

We will see that the proximity operator of many functions

of interest can be computed efficiently.

Example

Consider

$$f(x) = \frac{1}{2} x^T R x + s^T x + t$$

where R is PSD. Then

$$f(x) + \frac{\rho}{2} \|x - v\|_2^2 = \frac{1}{2} x^T R x + \frac{\rho}{2} x^T x$$

$$+ s^T x - \rho v^T x + \text{constant}$$

$$= \frac{1}{2} x^T (R + \rho I) x - (\rho v - s)^T x + \text{constant}$$

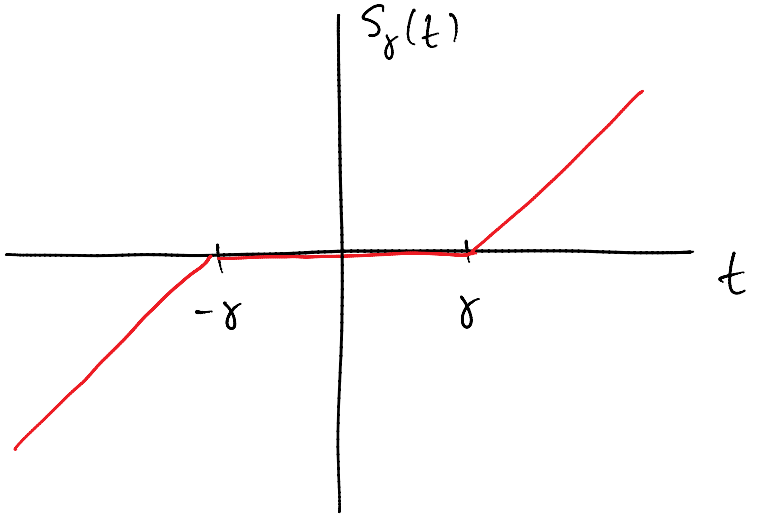
and therefore

$$\text{prox}_{f/\rho}(v) = (R + \rho I)^{-1} (\rho v - s).$$

Example

Consider $g(t) = \lambda |t|$, $\lambda > 0$. Define the soft-thresholding operator

$$S_\gamma(t) = \begin{cases} t - \gamma & \text{if } t \geq \gamma \\ 0 & \text{if } -\gamma < t < \gamma \\ t + \gamma & \text{if } t \leq -\gamma \end{cases}$$



With a little algebra it can be shown that

$$\text{prox}_{g/\rho}(t) = S_{\lambda/\rho}(t).$$

Generalizing to multiple dimensions, now consider $g(y) = \lambda \|y\|_1$. We can write

$$g(y) + \frac{\rho}{2}\|x-v\|^2 = \sum_{j=1}^{p}\left[\lambda|y^{(j)}| + \frac{\rho}{2}(y^{(j)}-v^{(j)})^2\right]$$

and solve for each $y^{(j)}$ independently. Thus

$$\text{prox}_{g/\rho}(v) = \left(S_{\lambda/\rho}(v^{(j)})\right)_{j=1}^{p} \in \mathbb{R}^p.$$

For convenience, for $v \in \mathbb{R}^p$ we will write $S_\gamma(v)$ to denote the vector that results from applying the soft-thresholding operator to v element-wise.

Solving the Lasso

Consider the problem

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

The lasso corresponds to

$$A = \sqrt{\frac{2}{n}}\begin{bmatrix} \tilde{x}_1^T \\ \vdots \\ \tilde{x}_n^T \end{bmatrix}, \quad b = \sqrt{\frac{2}{n}}\begin{bmatrix} \tilde{y}_1 \\ \vdots \\ \tilde{y}_n \end{bmatrix}$$

where $\tilde{x}_i = x_i - \bar{x}$, $\tilde{y}_i = y_i - \bar{y}$.

This is equivalent to

$$\begin{aligned} \min_{w,z} \quad & f(w) + g(z) \\ \text{s.t.} \quad & w - z = 0 \end{aligned}$$

this idea is known as "variable splitting"

where $f(x) = \frac{1}{2}\|Ax-b\|_2^2$ and $g(y) = \lambda\|y\|_1$

Since $f(x) = \frac{1}{2}x^TA^TAx + b^TAx + \text{constant}$, and applying the previous examples, ADMM becomes

$$x^{k+1} = \text{prox}_{f/\rho}(y^k - u^k)$$

$$= (A^TA + \rho I)^{-1}(A^Tb + \rho(y^k - u^k))$$

$$y^{k+1} = \text{prox}_{g/\rho}(x^{k+1} + u^k)$$

$$= S_{\lambda/\rho}(x^{k+1} + u^k)$$

$$u^{k+1} = u^k + x^{k+1} - y^{k+1}.$$

The soft-thresholding step can be implemented very efficiently. The x update can be performed efficiently if the dimension p of x is not too large. If p is large but n is small, the matrix inversion lemma can be employed to convert to a smaller matrix inverse. Numerical methods such as conjugate gradient can be used if an exact update is not possible. In some applications, A has special structure that can be exploited, such as circulant structure.

Final Thoughts

There are many optimization problems in ML that can be solved efficiently via ADMM. It is a very

flexible and powerful framework because many proximal functions of interest can be computed efficiently. ADMM is also easily parallelizable for distributed implementation.

For more on ADMM, see Boyd et al., "Distributed Optimization and Machine Learning via the Alternating Direction Method of Multipliers," Foundations and Trends in Machine Learning, 2010.