

Alternating Direction Method of Multipliers

Lasso regression invokes the following optimization problem

$$\text{minimize} \quad \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1$$

Rewrite this into ADMM form

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\gamma\|_1 \\ &\text{subject to} \quad \beta - \gamma = 0 \end{aligned}$$

Thus, the augmented Lagrangian is

$$L_\rho(\beta, \gamma, u) = \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\gamma\|_1 + u^T(\beta - \gamma) + \frac{\rho}{2} \|\beta - \gamma\|_2^2$$

Which means that the Lasso ADMM consists of the following iterations:

$$\begin{aligned} \beta^{k+1} &:= \arg \min_{\beta} L_\rho(\beta, \gamma^k, u^k) \\ \gamma^{k+1} &:= \arg \min_{\gamma} L_\rho(\beta^{k+1}, \gamma, u^k) \\ u^{k+1} &:= u^k + \rho(\beta^{k+1} - \gamma^{k+1}) \end{aligned}$$

To obtain the argmin's, we need to take the gradient of the objective function and set it equal to zero.

$$\begin{aligned} \nabla_{\beta} L_\rho(\beta, \gamma^k, u^k) &= \nabla_{\beta} \left\{ \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\gamma^k\|_1 + (u^k)^T(\beta - \gamma^k) + \frac{\rho}{2} \|\beta - \gamma^k\|_2^2 \right\} \\ &= X^T(X\beta - y) + u^k + \rho(\beta - \gamma^k) \stackrel{\text{set}}{=} 0 \\ \Rightarrow \beta^{k+1} &= (X^T X + \rho I)^{-1} (X^T y + \rho \gamma^k - u^k) \end{aligned}$$

And

$$\begin{aligned} \gamma^{k+1} &= \arg \min_{\gamma} L_\rho(\beta^{k+1}, \gamma, u^k) \\ &= \arg \min_{\gamma} \left\{ \frac{1}{2} \|y - X\beta^{k+1}\|_2^2 + \lambda \|\gamma\|_1 + (u^k)^T(\beta^{k+1} - \gamma) + \frac{\rho}{2} \|\beta^{k+1} - \gamma\|_2^2 \right\} \\ &= \arg \min_{\gamma} \left\{ \lambda \|\gamma\|_1 - (u^k)^T \gamma + \frac{\rho}{2} \|\beta^{k+1} - \gamma\|_2^2 \right\} \\ &= \arg \min_{\gamma} \left\{ \lambda \|\gamma\|_1 - (u^k)^T \gamma + \frac{\rho}{2} \gamma^T \gamma - \rho(\beta^{k+1})^T \gamma \right\} \\ &= \arg \min_{\gamma} \left\{ \lambda \|\gamma\|_1 - (u^k + \rho\beta^{k+1})^T \gamma + \frac{\rho}{2} \gamma^T \gamma \right\} \\ &= \arg \min_{\gamma} \left\{ \frac{\lambda}{\rho} \|\gamma\|_1 - \left(\frac{1}{\rho} u^k + \beta^{k+1}\right)^T \gamma + \frac{1}{2} \gamma^T \gamma \right\} \\ &= \arg \min_{\gamma} \left\{ \frac{\lambda}{\rho} \|\gamma\|_1 + \frac{1}{2} \left\| \gamma - \left(\frac{1}{\rho} u^k + \beta^{k+1}\right) \right\|_2^2 \right\} \\ &= S_{\lambda/\rho} \left(\frac{1}{\rho} u^k + \beta^{k+1} \right) \quad (\text{see the proof of exercise06}) \end{aligned}$$

Let $v_k = u_k / \rho$, we have

$$\gamma^{k+1} = S_{\lambda/\rho}(\beta^{k+1} + v^k) = \underset{\tau=1}{prox} \frac{\lambda}{\rho} \|\beta^{k+1} + v^k\|$$

We can also use scaled augmented Lagrangian to obtain the same solutions as in the paper (http://stanford.edu/~boyd/papers/pdf/admm_distr_stats.pdf). Finally, we have the following updates:

$$\begin{aligned}\beta^{k+1} &:= (X^T X + \rho I)^{-1} [X^T y + \rho(\gamma^k - v^k)] \\ \gamma^{k+1} &:= S_{\lambda/\rho}(\beta^{k+1} + v^k) \\ v^{k+1} &:= v^k + \beta^{k+1} - \gamma^{k+1}\end{aligned}$$

Lastly, the stopping rules are

$$\begin{aligned}\|r^k\|_2 &\leq \varepsilon^{pri}, & r^{k+1} &= \beta^{k+1} - \gamma^{k+1} & \text{for } \varepsilon^{pri} > 0 \\ \|s^k\|_2 &\leq \varepsilon^{dual}, & s^{k+1} &= -\rho(\gamma^{k+1} - \gamma^k) & \text{for } \varepsilon^{dual} > 0\end{aligned}$$

Where ε^{pri} and ε^{dual} are feasibility tolerances for primal and dual feasibility conditions. These tolerances can be chosen using an absolute and relative criterion, such as

$$\begin{aligned}\varepsilon^{pri} &= \sqrt{p}\varepsilon^{abs} + \varepsilon^{rel} \max\{\|\beta^k\|_2, \|\gamma^k\|_2\} \\ \varepsilon^{dual} &= \sqrt{n}\varepsilon^{abs} + \varepsilon^{rel} \|v\|_2\end{aligned} \quad \text{where } \varepsilon^{abs} > 0 \quad \varepsilon^{rel} > 0$$

Below are the ADMM lasso objective function plot vs. number of iterations with varying penalty parameters.

