STA5001: High Dimensional Statistics

Gong Wenwu 12031299

June 4, 2021

1. Gradient Descent is MM Algorithm

Proof:

- (a.) Majorization Step: We can make *Taylor Expansion* of convex function $f(\mathbf{x})$ around point $\mathbf{x}_{i-1}$:

  $$f(\mathbf{x}) = f(\mathbf{x}_{i-1}) + f^{'}(\mathbf{x}_{i-1})^T(\mathbf{x} - \mathbf{x}_{i-1}) + f^{''}(\mathbf{x}_{i-1})/2\| \mathbf{x} - \mathbf{x}_{i-1} \|^2 + Remainder.$$

  Since $f^{''}(\mathbf{x}) \preceq L1_p,\ L \geq \frac{1}{\delta}$, we can say that $g(\mathbf{x}) - f(\mathbf{x}) = \frac{1}{2} \| \mathbf{x} - \mathbf{x}_{i-1} \|^2 \ (L1_p - f^{''}(\mathbf{x})) \geq 0$ for all $\mathbf{x}$ and $f(\mathbf{x}_{i-1}) = g(\mathbf{x}_{i-1})$. $\square$

- (b.) Minimization Step: We can differentiate $g(\mathbf{x}_i)$ and set $g^{'}(\mathbf{x}_i) = 0$ to find out the minimizer.

  $$g^{'}(\mathbf{x}_i) = f^{'}(\mathbf{x}_{i-1}) + \frac{1}{\delta}(\mathbf{x}_i - \mathbf{x}_{i-1}\ ) = 0 \Rightarrow \mathbf{x}_i = \mathbf{x}_{i-1} - \delta f^{'}(\mathbf{x}_{i-1})$$

  which shows the sol. is a iteration process if $f^{'}(\mathbf{x})$ existed. By the Majorization and Minimization Steps, we assure that it is a MM-algorithm. $\square$

- (c.) Suppose that $\mathbf{x}_{i-1} < \mathbf{x}^* < \mathbf{x}_i$, then $f^{'}(\mathbf{x}_{i-1}) < 0 < f^{'}(\mathbf{x}_i)$ holds. We can expand $f(x)$ around $\mathbf{x}_{i-1}, \mathbf{x}_i$, for fixed $\mathbf{x}^*$, (a.) has shown that $\mathbf{x}_{i-1} - f(\mathbf{x}^*) \leq \frac{1}{2\delta} \parallel \mathbf{x}^* - \mathbf{x}_{i-1} \parallel^2$ and $\mathbf{x}_i - f(\mathbf{x}^*) \leq \frac{1}{2\delta} \parallel \mathbf{x}^i - \mathbf{x}_* \parallel^2$. Then $\mathbf{x}_{i-1} - f(\mathbf{x}^i) \leq \frac{1}{2\delta}\{\parallel \mathbf{x}^* - \mathbf{x}_i \parallel^2 + \parallel \mathbf{x}^{i-1} - \mathbf{x}_* \parallel^2\}$ holds. On the other hands, the convex function assure that $f(\mathbf{x}^i) \leq f(\mathbf{x}^*) + f(\mathbf{x}_{i-1}) - f(\mathbf{x}_i)$, so we have proved that $f(\mathbf{x}^i) \leq f(\mathbf{x}^*) + \frac{1}{2\delta}\{\parallel \mathbf{x}^* - \mathbf{x}_i \parallel^2 + \parallel \mathbf{x}_{i-1} - \mathbf{x}_* \parallel^2\}$. $\square$

- (d.) We firstly do the one-step gradient descent k times, then the result can be attained by summing the result given by (c.). $\square$

2. Sparse Group Lasso

The sparse group Lasso, defining as group lasso with an additional ℓ_1-penalty, leads to the convex program

$$\underset{\left\{\theta_j \in \mathbb{R}^{p_j}\right\}_{j=1}^{J}}{\operatorname{minimize}} \left\{ \frac{1}{2} \left\| \mathbf{y} - \sum_{j=1}^{J} \mathbf{Z}_j \theta_j \right\|_2^2 + \lambda \sum_{j=1}^{J} \left[(1-\alpha) \left\| \theta_j \right\|_2 + \alpha \left\| \theta_j \right\|_1 \right] \right\},$$

with $\alpha \in [0,1]$. Same as the elastic-net, the parameter α creates a bridge between the group lasso ($\alpha = 0$) and the lasso ($\alpha = 1$). So the optimal solution must satisfy the condition:

$$-\mathbf{Z}_j^T \left(\mathbf{y} - \sum_{\ell=1}^{J} \mathbf{Z}_\ell \widehat{\theta}_\ell \right) + \lambda(1-\alpha) \cdot \widehat{s}_j + \lambda\alpha \widehat{t}_j = 0, \text{ for } j = 1, \cdots, J,$$

where $\widehat{s}_j \in \mathbb{R}^{p_j}$ belongs to the subdifferential of the Euclidean norm at $\widehat{\theta}_j$, and $\widehat{t}_j \in \mathbb{R}^{p_j}$ belongs to the subdifferential of the ℓ_1 -norm at $\widehat{\theta}_j$. Further, we have each $\widehat{t}_{jk} \in \operatorname{sign}(\theta_{jk})$ as with the lasso.
Proof:

We can solve these equations via block-wise coordinate descent (Since the problem is convex, and the penalty is block separable, it is guaranteed to converge to an optimal solution). Define r_j as the partial residual in the j^{th} coordinate, it can be seen that $\widehat{\theta}_j = 0$ if and only if the equation

$$\mathbf{Z}_j^T \boldsymbol{r}_j = \lambda(1-\alpha)\widehat{s}_j + \lambda\alpha\widehat{t}_j$$

has a solution with $\left\|\widehat{s}_j\right\|_2 \leq 1$ and $\widehat{t}_{jk} \in [-1,1]$ for $k = 1, \ldots, p_j$.

Now, I will check this condition by solving $\min_{t: t_k \in [-1,1]} J(t)$ where

$$J(t) = \frac{1}{\lambda(1-\alpha)} \left\| \mathbf{Z}_j^T \mathbf{r}_j - \lambda\alpha \cdot t \right\|_2 = \|s\|_2$$

.

For the subdifferential of the Euclidean norm $\|\theta\|_2 = \sqrt{\sum_{j=1}^{p} \theta^2{}_j}$ evaluated at $\widehat{\theta}_j$, we know that: if $\widehat{\theta}_j \neq 0$, then we have $\widehat{s}_j = \widehat{\theta}_j / \left\|\widehat{\theta}_j\right\|_2$; whereas when $\widehat{\theta}_j = 0$, then $\widehat{s}_j$ is any vector with $\left\|\widehat{s}_j\right\|_2 \leq 1$. By the chain rule, we can know that

$$\frac{J(t)}{dt} = -\frac{\lambda\alpha}{\lambda(1-\alpha)} \frac{\mathbf{Z}_j^T \mathbf{r}_j - \lambda\alpha \cdot t}{\left\| \mathbf{Z}_j^T \mathbf{r}_j - \lambda\alpha \cdot t \right\|_2 = \|s\|_2}$$

so, if $\mathbf{Z}_j^T \mathbf{r}_j > \lambda\alpha$, $J'(t) < 0$, i.e., $\arg\min_{t: t_k \in [-1,1]} J(t) = 1$; if $\mathbf{Z}_j^T \mathbf{r}_j < \lambda\alpha$, $J'(t) > 0$, i.e., $\arg\min_{t: t_k \in [-1,1]} J(t) = -1$; if $\|\mathbf{Z}_j^T \mathbf{r}_j\| \leq \lambda\alpha$, $\min_{t: t_k \in [-1,1]} J(t) = 0$. When $\widehat{\theta}_j = 0$, the derivative of $J(t)$ equals to $sgn(\mathbf{Z}_j^T \mathbf{r}_j)(\mathbf{Z}_j^T \mathbf{r}_j - \lambda\alpha)_+ \leq \lambda(1-\alpha)$. So We can find that $\widehat{\theta}_j = 0$ if and only if $\left\| \mathcal{S}_{\lambda\alpha}\left(\mathbf{Z}_j^T r_j\right) \right\|_2 \leq \lambda(1-\alpha)$, where $\mathcal{S}_{\lambda\alpha}(\cdot)$ is the soft-thresholding operator applied here componentwise to its vector argument $\mathbf{Z}_j^T \boldsymbol{r}_j$.

Notice the similarity with the conditions for the group lasso, except here we use the soft-thresholded gradient $\mathcal{S}_{\lambda\alpha}\left(\mathbf{Z}_j^T r_j\right)$. So, if $\mathbf{Z}_j^T\mathbf{Z}_j = \mathbf{I}$ ($\mathbf{Z}_j$ is orthonormal), we have the closed form sol. of Sparse Group Lasso:

$$\widehat{\theta}_j = \left(1 - \frac{\lambda(1-\alpha)}{\left\|\mathcal{S}_{\lambda\alpha}\left(\mathbf{Z}_j^T r_j\right)\right\|_2}\right)_+ \mathcal{S}_{\lambda\alpha}\left(\mathbf{Z}_j^T r_j\right).$$

where $(t)_+ := \max\{0, t\}$ is the positive part function. $\square$

3. Generalized Linear Model: Logistic Regression

Suppose that $\left(\mathbf{X}_i, Y_i\right), i = 1, \cdots, n$, is an independent random sample from a generalized linear model with link $g(\cdot)$, and the conditional distribution of response given the covariates is

$$f\left(Y_i \mid \mathbf{X}_i, \theta_i, \phi\right) = \exp\left[\left\{Y_i\theta_i - b\left(\theta_i\right)\right\}/a_i(\phi) + c\left(Y_i, \phi\right)\right]$$

Denote by $\mu_i = \mu\left(\mathbf{X}_i\right) = \mathrm{E}\left(Y \mid \mathbf{X}_i\right)$. Then

$$\theta_i = \left(b'\right)^{-1}\left(\mu_i\right) = h\left(\mathbf{X}_i^T\boldsymbol{\beta}\right).$$

The likelihood function of $\boldsymbol{\beta}$ and ϕ is

$$\ell_n(\boldsymbol{\beta}, \phi) = \sum_{i=1}^n\left[Y_i h\left(\mathbf{X}_i^T\boldsymbol{\beta}\right) - b\left\{h\left(\mathbf{X}_i^T\boldsymbol{\beta}\right)\right\}\right]/a_i(\phi) + \sum_{i=1}^n c\left(Y_i, \phi\right).$$

Consider the case of logistic regression (Bernoulli, logit link),

$$\pi(\mathbf{x}) = \frac{\exp\left(\mathbf{x}^T\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{x}^T\boldsymbol{\beta}\right)}, \qquad 1-\pi(\mathbf{x}) = \frac{1}{1+\exp\left(\mathbf{x}^T\boldsymbol{\beta}\right)}, \quad \mathbf{x}^T\boldsymbol{\beta} = \beta_0 + \sum_{j=1}^p \beta_j\mathbf{X}_i^{(j)}.$$

- (a.) The negative log-likelihood equals

$$\sum_{i=1}^n\{y_i log(\pi(\mathbf{x})) + (1-y_i)log(1-\pi(\mathbf{x}))\} = \sum_{i=1}^n\left\{-y_i\left(\mathbf{x}^T\boldsymbol{\beta}\right) + \log\left(1+\exp\left(\mathbf{x}^T\boldsymbol{\beta}\right)\right)\right\}.$$

- (b.) We define the loss function of logit regression as

$$\rho(x, y) = -yf + \log(1+\exp(f)), \qquad f = \mathbf{X}^T\boldsymbol{\beta},$$

  and holds for $y = 0 \; or \; 1$. When $y = 0$, $\rho(f, 0) = \log(1+\exp(f))$; whereas $\rho(f, 1) = -f + \log(1+\exp(f)) = \log(\exp(f)(1+\exp(-f))) = \log(1+\exp(-f))$. So we have

$$\begin{aligned}&\rho(f, y) = \log(1+\exp(-(2y-1)f)) = \log(1+\exp(-\widetilde{y}f))\\ &\widetilde{y} = 2y-1 \in \{-1, 1\}\end{aligned}$$

4. Elastic Net Sol.

Proof:

- (a.) For the penalized least squares, the Elastic Net estimator is defined as

$$\arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_2 \|\boldsymbol{\beta}\|^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 \right\},$$

where $p_{\lambda_1,\lambda_2}(t) = \lambda_1|t| + \lambda_2 t^2$ is called the Elastic Net penalty. We can rewrite the Elastic Net form as

$$p_{\lambda,\alpha}(t) = \lambda J(t) = \lambda \left[(1-\alpha)t^2 + \alpha|t| \right] \ ,$$

with $\lambda = \lambda_1 + \lambda_2$ and $\alpha = \frac{\lambda_1}{\lambda_1+\lambda_2}$. For the equivalent form of LASSO, we only need to augment (X, Y) with $(\tilde{X}, \tilde{Y})$ such that $\boldsymbol{\beta} \left\{ \frac{1}{n} \|\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\|^2 + \lambda(1-\alpha)\|\boldsymbol{\beta}\|^2 \right\} = \boldsymbol{\beta} \left\{ \frac{1}{n} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 \right\}$. So, define ($\tilde{X} = [X^T, \sqrt{(1-\alpha)\lambda}1]^T \in \Re^{n+p,p}, \tilde{Y} = [Y^T, 0] \in \Re^{n+p}$). Since the intersection equals to zero, we can infer a equivalent LASSO form. □

- (b.) Just as the proof of LASSO, we suppose that we have two solutions $\hat{\beta}^1$ and $\hat{\beta}^2$ with $X\hat{\beta}^1 \neq X\hat{\beta}^2$. For any $0 < \gamma < 1$, we know that $\alpha\hat{\beta}^1 + (1-\alpha)\hat{\beta}^2$ is also a solution for Elastic-net convex minimization problem. Set the common optimal value of Elastic-net solutions is c^*, then

$$\frac{1}{2n} \| \, Y - X(\gamma\hat{\beta}^1 + (1-\gamma)\hat{\beta}^2) \, \|_2{}^2 + \lambda_1 \, \| \, \gamma\hat{\beta}^1 + (1-\gamma)\hat{\beta}^2 \, \|_1 + \lambda_2 \, \| \, \gamma\hat{\beta}^1 + (1-\gamma)\hat{\beta}^2 \, \|_2$$
$$< \alpha c^* + (1-\gamma)c^* = c^*.$$

where the strict inequality is due to the strict convexity of the function $f(x) = \| \, y - x \, \|_2{}^2$ along with the convexity of the l_2 and l_1-norm. This means that$\gamma\hat{\beta}^1 + (1-\gamma)\hat{\beta}^2$ attains a lower criterion value than c^*, this contradicts our assumption $\hat{\beta}^1$ and $\hat{\beta}^2$ are Elastic-net solutions. □

5. Elastic Net Penalty

Proof: Define

$$f(t) = \left\{ \sum_{\ell=1}^{K} \left[\frac{1}{2}(1-\alpha)\left(\beta_{j\ell} - t\right)^2 + \alpha \left|\beta_{j\ell} - t\right| \right] \right\}.$$

For $\alpha = 0$, $f'(t) = \sum_{\ell=1}^{K} (\beta_{j\ell} - t)$ and $f''(t) > 0$, the unique $c_j(0) = \widehat{\beta_j}$; For $\alpha = 1$, $f'(t) = |\beta_{j\ell} - t|$ and $f''(t) > 0$, the unique $c_j(1) = \widetilde{\beta_j}$. So the lower and upper bound have been proven.

6. Squared Hinge Loss Function

Proof:

- (a.) Clearly, this maximum function is continuous, and we only need to verify $\lim_{t \longrightarrow 1^+} \frac{\Phi_{sqh}(t)-0}{t-1} = \lim_{t \longrightarrow 1^-} \frac{\Phi_{sqh}(t)-0}{t-1} = 0$(existed).
- (b.) Define

  $$g(f) = \mathbb{E}_Y \left[\phi_{sqh}(Yf(x))\right] = p(x)(1-f(x))_+^2 + (1-p(x))(1+f(x))_+^2,\ where\, p(x)\, is\, known.$$

  If $f(x) \geq 1$, $\arg\min g(f) = 1 = 2p(x) - 1,\ p(x) = 1$; If $f(x) \leq -1$, $\arg\min g(f) = -1 = 2p(x) - 1,\ p(x) = 0$; If $-1 < f(x) < 1$, $\arg\min g(f) = 1 = 2p(x) - 1$.
- (c.) If $f(x) \geq 1$, $\arg\min g(f) = 1 = sgn(p(x) - 1/2),\ p(x) > 1/2$; If $f(x) \leq -1$, $\arg\min g(f) = -1 = sgn(p(x) - 1/2),\ p(x) < 1/2$; If $-1 < f(x) < 1$, $\arg\min g(f) = sgn(p(x) - 1/2)$.

7. Algorithm: Unconstrained Gradient Descent

Proof:

- (a.) Let

  $$\nabla f(\beta) = \beta^T \mathbf{Q} - b^T = 0$$

  then the sol. form of β^* is $(\mathbf{Q}, b)$ and the second derivative of $f(\beta) = \mathbf{Q} \succ 0$ ensure that β^* is unique.
- (b.) $\beta^{t+1} = \beta^t - s\nabla f(\beta^t) = \beta^t - s(\mathbf{Q}\beta^t - b)$, for $t = 0, 1, \cdots,$
- (c.) $\lim_{t \longrightarrow +\infty} \frac{\beta^{t+1}-\beta^*}{\beta^t-\beta^*} = 1 - 2sQ := c,\ c \in (0,1)$, then gradient descent converges for any fixed stepsize $s \in (0, c)$.

8. Algorithm: Proximal Gradient Descent

For the objective functions f, we can decompose it as $f = g + h$ where g is convex and differentiable, h is convex but nondifferentiable. Then, make a local approximation to f by linearizing the differentiable component g, but leaving the nondifferentiable component fixed. This leads to the generalized gradient update, defined by

$$\beta_{gg}^{t+1} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \left\langle \nabla g\left(\beta^t\right), \beta - \beta^t \right\rangle + \frac{1}{2s^t} \left\| \beta - \beta^t \right\|_2^2 + h(\beta) \right\},\ g\left(\beta^t\right)\ \textit{is constant}.$$

This update can be viewed as the proximal gradient descent. In order to make this connection explicit, we define the proximal map of a convex function h, a type of generalized projection operator:

$$\operatorname{prox}_h(z) := \arg\min_{\theta \in \mathbb{R}^p} \left\{ \frac{1}{2} \|z - \theta\|_2^2 + h(\theta) \right\}.$$

Then we can infer that $\operatorname{prox}_{sh}(z) = \arg\min_{\theta \in \mathbb{R}^p} \left\{ \frac{1}{2s} \|z - \theta\|_2^2 + h(\theta) \right\}$.

Proof: Generalized gradient update can be viewed as the proximal gradient descent, i.e., $\beta_{gg}^{t+1} = \beta_{pg}^{t+1}$.

The proximal-gradient descent update step defines as

$$\beta_{pg}^{t+1} = \operatorname{prox}_{s^t h}\left(\beta^t - s^t \nabla g\left(\beta^t\right)\right)$$

and this updates will be computationally efficient as long as the proximal map is relatively easy to compute.

By the relationship of $\operatorname{prox}_{sh}(z)$, we have

$$\begin{aligned}
\beta_{pg}^{t+1} &= \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2s^t} \|\beta - \beta^t + s^t \nabla g\left(\beta^t\right) \|_2^2 + h(\beta) \right\} \\
&= \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2s^t} \left\| \beta - \beta^t \right\|_2^2 + \left\langle \nabla g\left(\beta^t\right), \beta - \beta^t \right\rangle + \frac{s^t}{2} \{\nabla g\left(\beta^t\right)\}^2 + h(\beta) \right\} \\
&= \beta_{gg}^{t+1},\ \textit{since}\ \frac{s^t}{2} \{\nabla g\left(\beta^t\right)\}^2\ \textit{is constant when given}\ \beta^t.
\end{aligned}$$

□

9. Algorithm: ADMM for Group LASSO

The augmented Lagrangian is

$$\mathcal{L}_\eta(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{u}) = \frac{1}{2n} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p} \|\boldsymbol{\theta}_j\|_2 + \boldsymbol{u}^T(\boldsymbol{\theta} - \boldsymbol{\beta}) + \frac{\eta}{2} \|\boldsymbol{\theta} - \boldsymbol{\beta}\|_2^2,\ \alpha = 0.$$

where η can be a fixed positive constant set by the user, e.g. $\eta = 1$. The term $\boldsymbol{u}^T(\boldsymbol{\theta} - \boldsymbol{\beta})$ is the Lagrange multiplier and the term $\frac{\eta}{2}\|\boldsymbol{\theta} - \boldsymbol{\beta}\|_2^2$ is its augmentation. The choice of η can affect the convergence speed. ADMM is an iterative procedure. Let $\left(\boldsymbol{\beta}^k, \boldsymbol{\theta}^k, \boldsymbol{u}^k\right)$ denote the k-th iteration of the ADMM algorithm for $k = 0, 1, 2, \ldots$. Then the algorithm proceeds as follows:

$$\begin{aligned}
\boldsymbol{\beta}^{k+1} &= \operatorname{argmin}_{\boldsymbol{\beta}} \mathcal{L}_\eta \left(\boldsymbol{\beta}, \boldsymbol{\theta}^k, \boldsymbol{u}^k\right) \\
\boldsymbol{\theta}^{k+1} &= \operatorname{argmin}_{\boldsymbol{\theta}} \mathcal{L}_\eta \left(\boldsymbol{\beta}^{k+1}, \boldsymbol{\theta}^k, \boldsymbol{u}^k\right), \\
\boldsymbol{u}^{k+1} &= \boldsymbol{u}^k - \left(\boldsymbol{\theta}^{k+1} - \boldsymbol{\beta}^{k+1}\right)
\end{aligned}$$

It is easy to see that $\boldsymbol{\beta}^{k+1}$ has a close form expression and $\boldsymbol{\theta}^{k+1}$ is obtained by solving p group L_2 penalized problems.

$$\theta_j = \left(1 - \frac{\lambda}{\left\|\mathcal{S}_\lambda\left(\mathbf{Z}_j^T r_j\right)\right\|_2}\right)_+ \mathcal{S}_{\lambda\alpha}\left(\mathbf{Z}_j^T r_j\right).$$

where $(t)_+ := \max\{0, t\}$ is the positive part function. More specifically, we have

$$\begin{aligned}
\boldsymbol{\beta}^{k+1} &= \left(\mathbf{X}^T\mathbf{X}/n + \eta\mathbf{I}\right)^{-1}\left(\mathbf{X}^T\mathbf{Y}/n + \eta\boldsymbol{\theta}^k - \eta\boldsymbol{\theta}^k\right) \\
\theta_j^{k+1} &= \operatorname{sgn}\left(\beta_j^{k+1} + \theta_j^k\right)\left(\left|\beta_j^{k+1} + \theta_j^k\right| - \lambda/\eta\right), j = 1, \ldots, p
\end{aligned}$$