DSO Screening Exam: 2017 In-Class Exam

Gregory Faletto

Exercise 1 (**Probability/Analysis**). (a)

$$\dot{Z}(t) = \frac{\partial}{\partial t}\sum_{i=1}^{k}\gamma_i(t)X_i = \sum_{i=1}^{k}\frac{\partial\gamma_i(t)}{\partial t}X_i$$

$$\implies \mathbb{E}\left(\dot{Z}(t)\right) = \sum_{i=1}^{k}\mathbb{E}\left(\frac{\partial\gamma_i(t)}{\partial t}X_i\right) = 0$$

$$\implies \text{Cov}(Z(t),\dot{Z}(t)) = \mathbb{E}\big[(Z(t)-\mathbb{E}[Z(t)])(\dot{Z}(t)-\mathbb{E}[\dot{Z}(t)])\big] = \mathbb{E}\big[Z(t)\dot{Z}(t)\big]$$

$$= \mathbb{E}\left[\left(\sum_{i=1}^{k}\gamma_i(t)X_i\right)\left(\sum_{i=1}^{k}\frac{\partial\gamma_i(t)}{\partial t}X_i\right)\right] = \sum_{i=1}^{k}\mathbb{E}\left(\gamma_i(t)\frac{\partial\gamma_i(t)}{\partial t}X_i^2\right) + 0 = \sum_{i=1}^{k}\gamma_i(t)\frac{\partial\gamma_i(t)}{\partial t}$$

$$\vdots$$

$$\mathbb{E}\left[\sum_i \gamma_i\frac{\mathrm{d}\gamma_i}{\mathrm{d}t}\right] = 0$$

because

(b)

$$\ddot{Z}(t) = \sum_{i=1}^{k}\frac{\partial^2\gamma_i(t)}{\partial t^2}X_i$$

Exercise 2 (**Mathematical statistics; hypothesis test. so maybe don't worry about?**). (a)

(b)

Exercise 3 (**Probability; Wen says we should do this**). (a) Let $u_j = \exp(\theta_j)$. In the $n = 1$ case we have

$$I(m,\lambda) = \int_{\mathbb{R}}\frac{\exp(\lambda\theta_1 m_1)}{(1+\exp(\theta_1))^{\lambda}}d\theta_1 = \int_{\mathbb{R}}\frac{\exp(\theta_1)^{\lambda m_1}}{(1+\exp(\theta_1))^{\lambda}}d\theta_1$$

Let $u_1 = 1+\exp(\theta_1) \implies \partial u_1 = \exp(\theta_1)\partial\theta_1$.

$$= \int_{\mathbb{R}}\frac{(u_1-1)^{\lambda m_1-1}}{u_1^{\lambda}}du_1$$

$$= \ldots = \frac{\Gamma(\lambda(1-m_1))\Gamma(\lambda m_1)}{\Gamma(\lambda)} = \frac{\Gamma(\lambda m_0)\Gamma(\lambda m_1)}{\Gamma(\lambda)}$$

(b)

Exercise 4 (**Convex Optimization**). **Like theorem 2 in convex optimization lecture notes 3; probably don't need to worry about since not covered in Boyd. Don't prioritize.**

Exercise 5 (High-dimensional statistics). (a) If $p > n$, then even if $\boldsymbol{X}$ is full rank, it has a nullspace with nonzero entries. That is, the columns of $\boldsymbol{X}$ must be linearly dependent, so there are infinitely many non-zero vectors $\boldsymbol{v}$ such that $\boldsymbol{X}\boldsymbol{v} = \boldsymbol{0}$). Therefore for any solution $\hat{\boldsymbol{\beta}}$ satisfying

$$\hat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$$

and any vector $\boldsymbol{v}$ in the nullspace of $\boldsymbol{X}$, $\hat{\boldsymbol{\beta}} + \boldsymbol{v}$ is also a solution since

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 = \|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|_2^2 = \|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{0}\|_2^2 = \|\boldsymbol{y} - \boldsymbol{X}(\hat{\boldsymbol{\beta}} + \boldsymbol{v})\|_2^2$$

$$\iff \boldsymbol{\beta} + \boldsymbol{v} \in \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2.$$

(b) If we assume that $\|\boldsymbol{\beta}_0\|_0 = s \leq n$, then it may be that there is only one $\boldsymbol{v}$ in the nullspace of $\boldsymbol{X}$ such that $\|\hat{\boldsymbol{\beta}}\|_0 = s$. Intuitively, this makes sense because as long as the features with zero coefficients in $\boldsymbol{\beta}_0$ are sufficiently uncorrelated with the features with nonzero coefficients and the true coefficients are not too small, it is very unlikely that it is possible to construct another s-sparse vector with equally low empirical risk by replacing one of the true features with only one of the other features (or some combination thereof). (We would also hope that $\text{rank}(\boldsymbol{X}) > s$. Otherwise, it could either be the case that features in the true model are linearly dependent, in which case the true model is not identifiable, or some of the noise features are linearly dependent with some of the true features, which could also preclude identifiability.)

(c) **Solution in Section 1.9.1 of Linear Regression notes.**

We will follow the analysis from Zhao and Yu [2006]. The lasso problem is convex but not necessarily strictly convex if $p > n$. That is, there is some flat region, so the minimizer may not be unique. Consider the KKT conditions from convex optimization:

$$g(\beta) = \arg\min \left\{ \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\} = \arg\min \{f_1(\beta) + f_2(\beta)\}$$

Then $\hat{\beta}$ is a lasso solution if and only if 0 is in the subdifferential of $g(\hat{\beta})$. Note that

$$\partial g(\hat{\beta}) = \nabla f_1 + \partial f_2 = \frac{1}{n} X^T (X\beta - \boldsymbol{y}) + \lambda \begin{bmatrix} \vdots \\ \partial |\beta_j| \\ \vdots \end{bmatrix} = \frac{1}{n} X^T (X\beta - \boldsymbol{y}) + \lambda \begin{bmatrix} \vdots \\ \begin{cases} \text{sgn}(\beta_j) & \beta_j \neq 0 \\ [-1, 1] & \beta_j = 0 \end{cases} \\ \vdots \end{bmatrix}$$

Now assume $\text{supp}(\hat{\beta}) = \text{supp}(\beta_0)$ (that is, assume lasso recovers the correct support). Suppose the first s features are nonzero and consider one of them (so we know that we should have $\hat{\beta}_j \neq 0$):

$$0 \in \partial g(\hat{\beta}) \implies 0 \in \partial_j g(\hat{\beta}) = \left[\frac{1}{n} X^T (X\beta - \boldsymbol{y}) \right]_j + \lambda \, \text{sgn}(\hat{\beta}_j)$$

Therefore

$$\frac{1}{n} X_A^T (\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{y}) + \lambda \, \text{sgn}(\hat{\beta}_j) = 0 \tag{1}$$

where X_A is a submatrix of $\boldsymbol{X}$ containing the columns corresponding to the features in the true support, is our first condition. Next, consider what happens for $j > s$ (features not in the true support). We have

$$0 \in \partial g(\hat{\beta}) \implies 0 \in \partial_j g(\hat{\beta}) = \left[\frac{1}{n} X^T (X\beta - \boldsymbol{y})\right] + \lambda[-1, 1]$$

$$\implies \left\| \frac{1}{n} X_{A^C}^T (\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{y}) \right\|_\infty \leq \lambda \tag{2}$$

where X_{A^C} is a submatrix of $\boldsymbol{X}$ containing the columns corresponding to the features not in the true support, is our boundary condition. Recall the true model

$$y = X\beta_0 + \boldsymbol{\varepsilon}$$

and consider the case $X = \begin{bmatrix} \boldsymbol{X}_1 & \boldsymbol{X}_2 \end{bmatrix}$ where $\boldsymbol{X}_1$ are the features in the true model and $\boldsymbol{X}_2$ are noise features; that is, $\beta_0 = \begin{bmatrix} \boldsymbol{\beta}_1 \\ 0 \end{bmatrix}$. Then we are assuming

$$\hat{\beta}_{\text{lasso}} = \begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ 0 \end{bmatrix}.$$

We have from (1)

$$0 = \frac{1}{n} \boldsymbol{X}_1^T (\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{y}) + \lambda \operatorname{sgn}(\hat{\boldsymbol{\beta}}_1) = \frac{1}{n} \boldsymbol{X}_1^T (\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 - \boldsymbol{X}_1\boldsymbol{\beta}_1 - \boldsymbol{\varepsilon}) + \lambda \operatorname{sgn}(\hat{\boldsymbol{\beta}}_1)$$

$$\iff \frac{1}{n} \boldsymbol{X}_1^T \boldsymbol{X}_1 (\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) = \frac{1}{n} \boldsymbol{X}_1^T \boldsymbol{\varepsilon} - \lambda \operatorname{sgn}(\hat{\boldsymbol{\beta}}_1)$$

Let's assume that $\operatorname{sgn}(\hat{\beta}) = \operatorname{sgn}(\beta_0)$ (sign consistency).

$$\iff \frac{1}{n} \boldsymbol{X}_1^T \boldsymbol{X}_1 (\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) = \frac{1}{n} \boldsymbol{X}_1^T \boldsymbol{\varepsilon} - \lambda \operatorname{sgn}(\boldsymbol{\beta}_1)$$

which is linear in $\hat{\beta}$. Solving, we have

$$\iff \hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1 = (\boldsymbol{X}_1^T \boldsymbol{X}_1)^{-1} \big(\boldsymbol{X}_1^T \boldsymbol{\varepsilon} - n\lambda \operatorname{sgn}(\boldsymbol{\beta}_1) \big) \iff \hat{\boldsymbol{\beta}}_1 = \boldsymbol{\beta}_1 + (n^{-1} \boldsymbol{X}_1^T \boldsymbol{X}_1)^{-1} \big(n^{-1} \boldsymbol{X}_1^T \boldsymbol{\varepsilon} - \lambda \operatorname{sgn}(\boldsymbol{\beta}_1) \big) \tag{3}$$

Looking at the second (boundary) condition (2), we have

$$\left\| \frac{1}{n} \boldsymbol{X}_2^T (\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{y}) \right\|_\infty \leq \lambda. \tag{4}$$

Consider that

$$\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{y} = \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 - \boldsymbol{X}_1\boldsymbol{\beta}_1 - \boldsymbol{\varepsilon} = \boldsymbol{X}_1(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) - \boldsymbol{\varepsilon}$$

Substituting in the result from (3) yields

$$\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{y} = \boldsymbol{X}_1 \big[(n^{-1} \boldsymbol{X}_1^T \boldsymbol{X}_1)^{-1} \big(n^{-1} \boldsymbol{X}_1^T \boldsymbol{\varepsilon} - \lambda \operatorname{sgn}(\boldsymbol{\beta}_1) \big) \big] - \boldsymbol{\varepsilon}$$

which when we plug into (4) yields

$$\left\|\frac{1}{n}\boldsymbol{X}_2^T\left[\boldsymbol{X}_1(n^{-1}\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\big(n^{-1}\boldsymbol{X}_1^T\boldsymbol{\varepsilon}-\lambda\,\mathrm{sgn}(\boldsymbol{\beta}_1)\big)-\boldsymbol{\varepsilon}\right]\right\|_\infty \leq \lambda.$$

$$\iff \left\|\frac{1}{n}\boldsymbol{X}_2^T\boldsymbol{X}_1(n^{-1}\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\big(n^{-1}\boldsymbol{X}_1^T\boldsymbol{\varepsilon}-\lambda\,\mathrm{sgn}(\boldsymbol{\beta}_1)\big)-\frac{1}{n}\boldsymbol{X}_2^T\boldsymbol{\varepsilon}\right\|_\infty \leq \lambda.$$

Using the Triangle Inequality, we have

$$\begin{aligned}
&\left\|\frac{1}{n}\boldsymbol{X}_2^T\boldsymbol{X}_1(n^{-1}\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\big(n^{-1}\boldsymbol{X}_1^T\boldsymbol{\varepsilon}-\lambda\,\mathrm{sgn}(\boldsymbol{\beta}_1)\big)-\frac{1}{n}\boldsymbol{X}_2^T\boldsymbol{\varepsilon}\right\|_\infty \\
&\leq \left\|\frac{1}{n}\boldsymbol{X}_2^T\boldsymbol{X}_1(n^{-1}\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\big(n^{-1}\boldsymbol{X}_1^T\boldsymbol{\varepsilon}-\lambda\,\mathrm{sgn}(\boldsymbol{\beta}_1)\big)\right\|_\infty + \left\|\frac{1}{n}\boldsymbol{X}_2^T\boldsymbol{\varepsilon}\right\|_\infty \\
&\leq \left\|\frac{1}{n}\boldsymbol{X}_2^T\boldsymbol{X}_1(n^{-1}\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\right\|_\infty \cdot \left\|n^{-1}\boldsymbol{X}_1^T\boldsymbol{\varepsilon}-\lambda\,\mathrm{sgn}(\boldsymbol{\beta}_1)\right\|_\infty + \left\|\frac{1}{n}\boldsymbol{X}_2^T\boldsymbol{\varepsilon}\right\|_\infty \qquad (5)
\end{aligned}$$

$$\vdots$$

(d) **Solution in Section 1.9.1 of Linear Regression notes.** Theorem 4 from Zhao and Yu [2006] states that if $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$ for some $\sigma^2 > 0$ and some other regularity conditions hold, then the Strong Irrepresentable Condition implies the lasso has sign consistency (which implies model selection consistency) with high probability for λ satisfying

$$\lambda(n) \propto n^{(1+c_4)/2}$$

where c_4 is a constant between 0 and 1 and smaller than $c_2 - c_1$, where c_1 is a constant governing how large the true model can be with $0 \leq c_1$ and c_2 is a constant governing the minimum size of the coefficients in the true model with $c_1 < c_2 \leq 1$.

Recall that the Strong Irrepresentable Condition is the following: if as in part 5(c) $\boldsymbol{X}_1 \in \mathbb{R}^{n\times q}$ is a matrix containing only the q columns of $\boldsymbol{X}$ corresponding to features in the true model and $\boldsymbol{X}_2 \in \mathbb{R}^{n\times p-q}$ contains only the $p-q$ columns of $\boldsymbol{X}$ corresponding to irrelevant features, the Strong Irrepresentable Condition is

$$\left\|\frac{1}{n}\boldsymbol{X}_2^T\boldsymbol{X}_1\left(\frac{1}{n}\boldsymbol{X}_1^T\boldsymbol{X}_1\right)^{-1}\mathrm{sgn}\left(\boldsymbol{\beta}_{(1)}^n\right)\right\|_\infty \leq \boldsymbol{1}-\boldsymbol{\eta}$$

$$\iff \left\|\left(\boldsymbol{X}_1^T\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1^T\boldsymbol{X}_2\right\|_\infty \leq \boldsymbol{1}-\boldsymbol{\eta}$$

for some $\boldsymbol{\eta} \in \mathbb{R}^n$ with $\boldsymbol{\eta} \succeq \boldsymbol{0}$, where $\boldsymbol{\beta}_{(1)}^n$ are the coefficients of $\boldsymbol{X}_1$ in the true model. That is, a regression of the variables in $\boldsymbol{X}_2$ against the variables in $\boldsymbol{X}_1$ may not have any coefficients that are larger in absolute value than $1-\eta$ for some $\eta > 0$. This implies that the features in $\boldsymbol{X}_2$ are not too strongly correlated with the features in $\boldsymbol{X}_1$ (the features of $\boldsymbol{X}_2$ are "irrepresentable" by the features in $\boldsymbol{X}_1$).

(Roughly speaking, the required regularity conditions are that the number of features in the design matrix p_n is not excessively large, the eigenvalues of $\boldsymbol{X}_1$ are not too small, the coefficients of the relevant features in the true model are not too small, and the model is sufficiently sparse. The conditions of this theorem allow p_n to grow asymptotically with n; the requirements for lower bounded eigenvalues and coefficients on the true coefficients are consequences of allowing p_n to grow with n.)

References

P. Zhao and B. Yu. On Model Selection Consistency of Lasso. *Journal of Machine Learning Research*, 7: 2541–2563, 2006.