Notes on Approximate Leave-One-Out for Elastic Net

Linyun He Wanchao Qin Peng Xu Yuze Zhou

August 29, 2018

1 Approximation in the Primal Domain

1.1 ALO for Linear Regression

Recall the objective function for the elastic net problem:

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\sum_{j=1}^{n}(\boldsymbol{x}_j^\top\boldsymbol{\beta}-y_j)^2+\lambda\left(\alpha\|\boldsymbol{\beta}\|_1+\frac{1-\alpha}{2}\|\boldsymbol{\beta}\|_2^2\right). \tag{1}$$

Let $A=\{i:\beta_i\notin K, i=1,\dots,p\}$ be the active set, we have

$$\dot{\ell}(\boldsymbol{x}_j^\top\boldsymbol{\beta};\,y_j)=\boldsymbol{x}_j^\top\boldsymbol{\beta}-y_j,\qquad \ddot{\ell}(\boldsymbol{x}_j^\top\boldsymbol{\beta};\,y_j)=1,\qquad \nabla^2 R(\hat{\boldsymbol{\beta}}_A)=(1-\alpha)\lambda\boldsymbol{I}_{A,A}.$$

Thus, Eqn. 31 reduces to

$$\boldsymbol{H}=\boldsymbol{X}_{\cdot,A}\left[\boldsymbol{X}_{\cdot,A}^\top\boldsymbol{X}_{\cdot,A}+(1-\alpha)\,\lambda\boldsymbol{I}_{A,A}\right]^{-1}\boldsymbol{X}_{\cdot,A}^\top. \tag{2}$$

By augmenting $\boldsymbol{X}$ with an extra column of 1s, adding the intercept back to the model is straightforward, as Eqn. 31 now becomes

$$\boldsymbol{H}=[\mathbf{1}_n,\boldsymbol{X}_{\cdot,A}]\left\{[\mathbf{1}_n,\boldsymbol{X}_{\cdot,A}]^\top\boldsymbol{D}\,[\mathbf{1}_n,\boldsymbol{X}_{\cdot,A}]+\nabla^2 R\left(\hat{\beta}_0,\hat{\boldsymbol{\beta}}_A\right)\right\}^{-1}[\mathbf{1}_n,\boldsymbol{X}_{\cdot,A}]^\top, \tag{3}$$

where

$$\boldsymbol{D}=\operatorname{diag}\left[\ddot{\ell}\left(\hat{\beta}_0+\boldsymbol{x}_j^\top\hat{\boldsymbol{\beta}};y_j\right)\right]_{j\in A}=\boldsymbol{I}_{A,A},\qquad \nabla^2 R\left(\hat{\beta}_0,\hat{\boldsymbol{\beta}}_A\right)=\begin{bmatrix}0&0&\dots&0\\0&(1-\alpha)\lambda&\dots&0\\\vdots&\vdots&\ddots&\vdots\\0&0&\dots&(1-\alpha)\lambda\end{bmatrix}.$$

Then, the ALO can be computed as

$$\begin{bmatrix}1&\boldsymbol{x}_i^\top\end{bmatrix}\begin{bmatrix}\tilde{\beta}_0^{\setminus i}\\\tilde{\boldsymbol{\beta}}^{\setminus i}\end{bmatrix}=(\hat{\beta}_0+\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}})+\frac{\boldsymbol{H}_{ii}}{1-\boldsymbol{H}_{ii}\ddot{\ell}\left(\hat{\beta}_0+\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}};y_i\right)}\dot{\ell}\left(\hat{\beta}_0+\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}};y_i\right) \tag{4}$$

1.2 ALO for Logistic Regression, Approximation in the Primal Domain

For binomial logistic regression, the primal problem is:

$$\min_{\boldsymbol{\beta}} \sum_{j=1}^{n} \left[\ln\left(1 + e^{\boldsymbol{x}_j^\top \boldsymbol{\beta}}\right) - y_j \boldsymbol{x}_j^\top \boldsymbol{\beta} \right] + \lambda \left(\alpha \|\boldsymbol{\beta}\|_1 + \frac{1-\alpha}{2} \|\boldsymbol{\beta}\|_2^2 \right).$$

Let A be the active set, we have

$$\dot{\ell}(\boldsymbol{x}_j^\top \boldsymbol{\beta};\, y_j) = \frac{e^{\boldsymbol{x}_j^\top \boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_j^\top \boldsymbol{\beta}}} - y_j, \qquad \ddot{\ell}(\boldsymbol{x}_j^\top \boldsymbol{\beta};\, y_j) = \frac{e^{\boldsymbol{x}_j^\top \boldsymbol{\beta}}}{\left(1 + e^{\boldsymbol{x}_j^\top \boldsymbol{\beta}}\right)^2}, \qquad \nabla^2 R(\hat{\boldsymbol{\beta}}_A) = (1-\alpha)\lambda \boldsymbol{I}_{A,A}.$$

Therefore, the ALO formula is

$$\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}^{\setminus j} \approx \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}} + \frac{\boldsymbol{H}_{ii} \left(1 + e^{\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}}\right) \left[e^{\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}} - y_j \left(1 + e^{\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}}\right) \right]}{\left(1 + e^{\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}}\right)^2 - \boldsymbol{H}_{ii} e^{\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}}},$$

where

$$\boldsymbol{H} = \boldsymbol{X}_{\cdot,A} \left[\boldsymbol{X}_{\cdot,A}^\top \operatorname{diag}\left(\frac{e^{\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}}}{1 + 2e^{\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}} + e^{2\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}}} \right) \boldsymbol{X}_{\cdot,A} + (1-\alpha)\lambda \boldsymbol{I}_{A,A} \right]^{-1} \boldsymbol{X}_{\cdot,A}^\top.$$

1.3 ALO for Poisson Regression, Approximation in the Primal Domain

For Poisson regression, the primal problem is:

$$\min_{\boldsymbol{\beta}} \sum_{j=1}^{n} \left(e^{\boldsymbol{x}_j^\top \boldsymbol{\beta}} - y_j \boldsymbol{x}_j^\top \boldsymbol{\beta} \right) + \lambda \left(\alpha \|\boldsymbol{\beta}\|_1 + \frac{1-\alpha}{2} \|\boldsymbol{\beta}\|_2^2 \right).$$

Let A be the active set, we have

$$\dot{\ell}(\boldsymbol{x}_j^\top \boldsymbol{\beta};\, y_j) = e^{\boldsymbol{x}_j^\top \boldsymbol{\beta}} - y_j, \qquad \ddot{\ell}(\boldsymbol{x}_j^\top \boldsymbol{\beta};\, y_j) = e^{\boldsymbol{x}_j^\top \boldsymbol{\beta}}, \qquad \nabla^2 R(\hat{\boldsymbol{\beta}}_A) = (1-\alpha)\lambda \boldsymbol{I}_{A,A}.$$

Therefore, the ALO formula is

$$\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}^{\setminus j} \approx \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}} + \frac{\boldsymbol{H}_{ii} \left(e^{\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}} - y_j \right)}{1 - \boldsymbol{H}_{ii} e^{\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}}},$$

where

$$\boldsymbol{H} = \boldsymbol{X}_{\cdot,A} \left[\boldsymbol{X}_{\cdot,A}^\top \operatorname{diag}\left(e^{\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}} \right) \boldsymbol{X}_{\cdot,A} + (1-\alpha)\lambda \boldsymbol{I}_{A,A} \right]^{-1} \boldsymbol{X}_{\cdot,A}^\top.$$

2 Approximation in the Dual Domain

2.1 ALO for Linear Regression

First, to write the optimization problem of elastic net in a matrix form, and denote $\boldsymbol{D} = [\boldsymbol{0}, \boldsymbol{I}]$, the optimization problem becomes:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{D}\boldsymbol{\beta}\|_1 + \lambda_2\|\boldsymbol{D}\boldsymbol{\beta}\|_2^2$$

The augmented Lagrangian for the problem is:

$$L = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{z}\|_2^2 + \lambda_1\|\boldsymbol{\omega}\|_1 + \lambda_2\|\boldsymbol{\omega}\|_2^2 + \boldsymbol{u}^\top(\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta}) + \boldsymbol{v}^\top(\boldsymbol{\omega} - \boldsymbol{D}\boldsymbol{\beta})$$

By taking the derivatives with respect to z and β, we could obtain:

$$0 = \frac{\partial L}{\partial \boldsymbol{z}} = \boldsymbol{z} - \boldsymbol{y} + \boldsymbol{u}, \qquad 0 = \frac{\partial L}{\partial \boldsymbol{\beta}} = -\boldsymbol{X}^\top\boldsymbol{u} - \boldsymbol{D}^\top\boldsymbol{v}.$$

Since the first column of $\boldsymbol{X}$ is filled with ones and the first column of D is filled with zeros, the first row of $-\boldsymbol{X}^\top\boldsymbol{u} - \boldsymbol{D}^\top\boldsymbol{v} = 0$ gives $\boldsymbol{1}^\top\boldsymbol{u} = 0$ and due to $\boldsymbol{D} = [\boldsymbol{0}, \boldsymbol{I}]$, the rest rows give that $-\boldsymbol{X}_j^\top\boldsymbol{u} = \boldsymbol{v}_j$. Moreover, since the rest dimensions of ω is penalized element-wisely in the augmented Lagrangian, we can minimize over ω by minimizing over each $\omega_i, \quad i \geq 2$, that is, we have to minimize $\lambda_1|\omega_i| + \lambda_2\omega_i^2 - u^\top X_i\omega_i$ for each dimension of ω, where X_i denotes the ith column of X, therefore:

$$\min_{\omega_i}\left(\lambda_1|\omega_i| + \lambda_2\omega_i^2 - \boldsymbol{u}^\top\boldsymbol{X}_i\omega_i\right) = \begin{cases} 0 & |\boldsymbol{u}^\top\boldsymbol{X}_i| \leq \lambda_1, \\ -\dfrac{(\lambda_1 - |\boldsymbol{u}^\top\boldsymbol{X}_i|)^2}{4\lambda_2} & |\boldsymbol{u}^\top\boldsymbol{X}_i| > \lambda_1. \end{cases}$$

By taking all the above back to the Lagrangian, we obtain the dual problem as:

$$d^* = \min_{\boldsymbol{u}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{u}\|_2^2 + \sum_{j:|\boldsymbol{X}_j^\top\boldsymbol{u}|>\lambda_1} \frac{(\lambda_1 - |\boldsymbol{u}^\top\boldsymbol{X}_i|)^2}{4\lambda_2}, \qquad \text{subject to } \boldsymbol{1}^\top\boldsymbol{u} = 0.$$

Denote

$$f(\boldsymbol{u}) = \sum_{j:|\boldsymbol{X}_j^\top\boldsymbol{u}|>\lambda_1} \frac{(\lambda_1 - |\boldsymbol{u}^\top\boldsymbol{X}_i|)^2}{4\lambda_2},$$

clearly $f(u)$ is of quadratic form:

$$f(\boldsymbol{u}) = \frac{1}{2}\boldsymbol{u}^\top\boldsymbol{A}\boldsymbol{u} + \boldsymbol{a}^\top\boldsymbol{u} + b,$$

where b is a constant and does not matter in the optimization of the dual problem, $\boldsymbol{A}$ and $\boldsymbol{a}$ are:

$$\boldsymbol{A} = \frac{1}{2\lambda_2}\boldsymbol{X}_{\cdot,E}\boldsymbol{X}_{\cdot,E}^\top, \qquad \boldsymbol{a} = \frac{\lambda_1}{2\lambda_2}\Big(\sum_{i:\boldsymbol{X}_i^\top\leq\lambda_1} \boldsymbol{X}_i - \sum_{i:\boldsymbol{X}_i^\top>\lambda_1} \boldsymbol{X}_i\Big)$$

where $E \coloneqq \{i : |\boldsymbol{X}_i^\top u| > \lambda\}$.

The dual problem could also be written in a proximal form:

$$\hat{\boldsymbol{u}} = \text{prox}_{\tilde{f}}(\boldsymbol{y}), \qquad \tilde{f} = \boldsymbol{I}(\mathbf{1}^\top \boldsymbol{u} = 0) f(\boldsymbol{u}) + \boldsymbol{I}(\mathbf{1}^\top \boldsymbol{u} \neq 0)\infty.$$

After transforming $f(\boldsymbol{u})$ into a quadratic form, we could write the Lagrangian for the dual problem back again:

$$L = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{u}\|_2^2 + \frac{1}{2}\boldsymbol{u}^\top \boldsymbol{A}\boldsymbol{u} + \boldsymbol{a}^\top \boldsymbol{u} + b + \lambda \mathbf{1}^\top \boldsymbol{u}.$$

By taking the derivative with respect to u, we could obtain:

$$\frac{\partial L}{\partial \boldsymbol{u}} = \boldsymbol{u} - \boldsymbol{y} + \boldsymbol{A}\boldsymbol{u} + \boldsymbol{a} + \lambda \mathbf{1} = 0$$

By shifting the terms, $\boldsymbol{u}$ could be written as a formula of $\boldsymbol{y}$ and λ: $\boldsymbol{u} = (\boldsymbol{I} + \boldsymbol{A})^{-1}(\boldsymbol{y} - \boldsymbol{a} - \lambda \mathbf{1})$, by taking the derivative with respect to $\boldsymbol{y}$ at both sides, we could obtain the Jacobian matrix $\boldsymbol{J}$ of the proximal operator $\text{prox}(\tilde{f})$ at $\boldsymbol{y}$ as:

$$\boldsymbol{J} = (\boldsymbol{I} + \boldsymbol{A})^{-1} - (\boldsymbol{I} + \boldsymbol{A})^{-1}\mathbf{1}\nabla(\hat{\lambda})^\top$$

where $\nabla(\hat{\lambda})^\top$ denotes the gradient of λ as a function of $\boldsymbol{y}$. By taking $\boldsymbol{u} = (\boldsymbol{I} + \boldsymbol{A})^{-1}(\boldsymbol{y} - \boldsymbol{a} - \lambda \mathbf{1})$ back to the Lagrangian, the dual problem will become a second-order equation of λ:

$$\begin{aligned} d^* = \max_\lambda & \frac{1}{2}\|\boldsymbol{y} - (\boldsymbol{I} + \boldsymbol{A})^{-1}(\boldsymbol{y} - \boldsymbol{a} - \lambda \mathbf{1})\|_2^2 \\ & + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{a} - \lambda \mathbf{1})^\top (\boldsymbol{I} + \boldsymbol{A})^{-1}\boldsymbol{A}(\boldsymbol{I} + \boldsymbol{A})^{-1}(\boldsymbol{y} - \boldsymbol{a} - \lambda \mathbf{1}) + \boldsymbol{a}^\top (\boldsymbol{I} + \boldsymbol{A})^{-1}(\boldsymbol{y} - \boldsymbol{a} - \lambda \mathbf{1}) \\ & + \lambda \mathbf{1}^\top (\boldsymbol{I} + \boldsymbol{A})^{-1}(\boldsymbol{y} - \boldsymbol{a} - \lambda \mathbf{1}) \end{aligned}$$

More specifically, the second-order term is:

$$\frac{1}{2}\mathbf{1}^\top (\boldsymbol{I} + \boldsymbol{A})^{-2}\mathbf{1} + \frac{1}{2}\mathbf{1}^\top (\boldsymbol{I} + \boldsymbol{A})^{-1}\boldsymbol{A}(\boldsymbol{I} + \boldsymbol{A})^{-1}\mathbf{1} - \mathbf{1}^\top (\boldsymbol{I} + \boldsymbol{A})^{-1}\mathbf{1}$$

and the first-order term is:

$$2\mathbf{1}^\top (\boldsymbol{I} + \boldsymbol{A})^{-1}(\boldsymbol{y} - \boldsymbol{a}) - \mathbf{1}^\top (\boldsymbol{I} + \boldsymbol{A})^{-2}(\boldsymbol{y} - \boldsymbol{a}) - \mathbf{1}^\top (\boldsymbol{I} + \boldsymbol{A})^{-1}\boldsymbol{A}(\boldsymbol{I} + \boldsymbol{A})^{-1}(\boldsymbol{y} - \boldsymbol{a})$$

Thus by solving the second-order equation, we could obtain

$$\hat{\lambda} = \frac{2\mathbf{1}^\top (\boldsymbol{I} + \boldsymbol{A})^{-1}(\boldsymbol{y} - \boldsymbol{a}) - \mathbf{1}^\top (\boldsymbol{I} + \boldsymbol{A})^{-2}(\boldsymbol{y} - \boldsymbol{a}) - \mathbf{1}^\top (\boldsymbol{I} + \boldsymbol{A})^{-1}\boldsymbol{A}(\boldsymbol{I} + \boldsymbol{A})^{-1}(\boldsymbol{y} - \boldsymbol{a})}{\mathbf{1}^\top (\boldsymbol{I} + \boldsymbol{A})^{-2}\mathbf{1} + \mathbf{1}^\top (\boldsymbol{I} + \boldsymbol{A})^{-1}\boldsymbol{A}(\boldsymbol{I} + \boldsymbol{A})^{-1}\mathbf{1} - 2\mathbf{1}^\top (\boldsymbol{I} + \boldsymbol{A})^{-1}\mathbf{1}}$$

and the gradient

$$\nabla(\hat{\lambda}) = \frac{2(\boldsymbol{I} + \boldsymbol{A})^{-1}\mathbf{1} - (\boldsymbol{I} + \boldsymbol{A})^{-2}\mathbf{1} - (\boldsymbol{I} + \boldsymbol{A})^{-1}\boldsymbol{A}(\boldsymbol{I} + \boldsymbol{A})^{-1}\mathbf{1}}{\mathbf{1}^\top (\boldsymbol{I} + \boldsymbol{A})^{-2}\mathbf{1} + \mathbf{1}^\top (\boldsymbol{I} + \boldsymbol{A})^{-1}\boldsymbol{A}(\boldsymbol{I} + \boldsymbol{A})^{-1}\mathbf{1} - 2\mathbf{1}^\top (\boldsymbol{I} + \boldsymbol{A})^{-1}\mathbf{1}}$$

By taking the gradient back to

$$\boldsymbol{J} = (\boldsymbol{I} + \boldsymbol{A})^{-1} - (\boldsymbol{I} + \boldsymbol{A})^{-1}\mathbf{1}\nabla(\hat{\lambda})^\top = (\boldsymbol{I} + \boldsymbol{A})^{-1} - \frac{(\boldsymbol{I} + \boldsymbol{A})^{-1}\mathbf{1}\mathbf{1}^\top (\boldsymbol{I} + \boldsymbol{A})^{-1}}{\mathbf{1}^\top (\boldsymbol{I} + \boldsymbol{A})^{-1}\mathbf{1}},$$

we could obtain the Jacobian.

2.2 Proof of the Equivalence between the primal and the dual solutions

First recall the ALO formula from the dual approach

$$\boldsymbol{y}^{/i} = \boldsymbol{y}_i - \frac{\boldsymbol{u}_i}{\boldsymbol{J}_{ii}} = \frac{\boldsymbol{J}_{ii} - 1}{\boldsymbol{J}_{ii}}\boldsymbol{y}_i + \frac{1}{\boldsymbol{J}_{ii}}\boldsymbol{x}_i\hat{\boldsymbol{\beta}}$$

and the primal formula from the primal approach

$$\boldsymbol{y}^{/i} = \boldsymbol{x}_i\hat{\beta} + \frac{\boldsymbol{H}_{ii}}{1 - \boldsymbol{H}_{ii}}(\boldsymbol{x}_i\hat{\boldsymbol{\beta}} - \boldsymbol{y}_i) = -\frac{\boldsymbol{H}_{ii}}{1 - \boldsymbol{H}_{ii}}\boldsymbol{y}_i + \frac{1}{1 - \boldsymbol{H}_{ii}}\boldsymbol{x}_i\hat{\boldsymbol{\beta}}.$$

In this section, we're going to show that $\boldsymbol{H} + \boldsymbol{J} = \boldsymbol{I}$, thus giving $\boldsymbol{H}_{ii} + \boldsymbol{J}_{ii} = 1$, and that the solutions given by both the primal and the dual approach are equivalent.

First, using matrix inverse lemma, we could calculate the inverse of $(\boldsymbol{I} + \boldsymbol{A}) = (\boldsymbol{I} + \frac{1}{2\lambda_2}\boldsymbol{X}_{\cdot,E}\boldsymbol{X}_{\cdot,E}^\top)$ as:

$$(\boldsymbol{I} + \boldsymbol{A})^{-1} = (\boldsymbol{I} + \frac{1}{2\lambda_2}\boldsymbol{X}_{\cdot,E}\boldsymbol{X}_{\cdot,E}^\top)^{-1} = \boldsymbol{I} - \boldsymbol{X}_{\cdot,E}(2\lambda_2\boldsymbol{I} + \boldsymbol{X}_{\cdot,E}^\top\boldsymbol{X}_{\cdot,E})^{-1}\boldsymbol{X}_{\cdot,E}^\top,$$

therefore the matrix $\boldsymbol{J}$ is:

$$\begin{aligned}
\boldsymbol{J} &= (\boldsymbol{I} + \boldsymbol{A})^{-1} - \frac{(\boldsymbol{I} + \boldsymbol{A})^{-1}\mathbf{1}\mathbf{1}^\top(\boldsymbol{I} + \boldsymbol{A})^{-1}}{\mathbf{1}^\top(\boldsymbol{I} + \boldsymbol{A})^{-1}\mathbf{1}} \\
&= \boldsymbol{I} - \boldsymbol{X}_{\cdot,E}(2\lambda_2\boldsymbol{I} + \boldsymbol{X}_{\cdot,E}^\top\boldsymbol{X}_{\cdot,E})^{-1}\boldsymbol{X}_{\cdot,E}^\top \\
&\quad - \frac{(\mathbf{1} - \boldsymbol{X}_{\cdot,E}(2\lambda_2\boldsymbol{I} + \boldsymbol{X}_{\cdot,E}^\top\boldsymbol{X}_{\cdot,E})^{-1}\boldsymbol{X}_{\cdot,E}^\top\mathbf{1})(\mathbf{1}^\top - \mathbf{1}^\top\boldsymbol{X}_{\cdot,E}(2\lambda_2\boldsymbol{I} + \boldsymbol{X}_{\cdot,E}^\top\boldsymbol{X}_{\cdot,E})^{-1}\boldsymbol{X}_{\cdot,E}^\top)}{\mathbf{1}^\top(\boldsymbol{I} - \boldsymbol{X}_{\cdot,E}(2\lambda_2\boldsymbol{I} + \boldsymbol{X}_{\cdot,E}^\top\boldsymbol{X}_{\cdot,E})^{-1}\boldsymbol{X}_{\cdot,E}^\top)\mathbf{1}}
\end{aligned}$$

Now recall that

$$\boldsymbol{H} = [\mathbf{1}, \boldsymbol{X}_{\cdot,E}]([\mathbf{1}, \boldsymbol{X}_{\cdot,E}]^\top[\mathbf{1}, \boldsymbol{X}_{\cdot,E}] + \mathrm{diag}[0, 2\lambda_2, \cdots, 2\lambda_2])[\mathbf{1}, \boldsymbol{X}_{\cdot,E}]^\top,$$

by adopting block inverse, we could derive $\boldsymbol{H}$ as:

$$\begin{aligned}
\boldsymbol{H} &= [1, \boldsymbol{X}_{\cdot,E}]\begin{pmatrix} n & \mathbf{1}^\top\boldsymbol{X}_{\cdot,E} \\ \boldsymbol{X}_{\cdot,E}^\top\mathbf{1} & \boldsymbol{X}_{\cdot,E}^\top\boldsymbol{X}_{\cdot,E} + 2\lambda_2\boldsymbol{I} \end{pmatrix}[1, \boldsymbol{X}_{\cdot,E}]^\top \\
&= [1, \boldsymbol{X}_{\cdot,E}]\begin{pmatrix} \frac{1}{\mathbf{1}^\top(\boldsymbol{I} - \boldsymbol{X}_{\cdot,E}(2\lambda_2\boldsymbol{I} + \boldsymbol{X}_{\cdot,E}^\top\boldsymbol{X}_{\cdot,E})^{-1}\boldsymbol{X}_{\cdot,E}^\top)\mathbf{1}} & \frac{-\mathbf{1}^\top\boldsymbol{X}_{\cdot,E}(2\lambda_2\boldsymbol{I} + \boldsymbol{X}_{\cdot,E}^\top\boldsymbol{X}_{\cdot,E})^{-1}}{\mathbf{1}^\top(\boldsymbol{I} - \boldsymbol{X}_{\cdot,E}(2\lambda_2\boldsymbol{I} + \boldsymbol{X}_{\cdot,E}^\top\boldsymbol{X}_{\cdot,E})^{-1}\boldsymbol{X}_{\cdot,E}^\top)\mathbf{1}} \\ \frac{-(2\lambda_2\boldsymbol{I} + \boldsymbol{X}_{\cdot,E}^\top\boldsymbol{X}_{\cdot,E})^{-1}\boldsymbol{X}_{\cdot,E}^\top\mathbf{1}}{\mathbf{1}^\top(\boldsymbol{I} - \boldsymbol{X}_{\cdot,E}(2\lambda_2\boldsymbol{I} + \boldsymbol{X}_{\cdot,E}^\top\boldsymbol{X}_{\cdot,E})^{-1}\boldsymbol{X}_{\cdot,E}^\top)\mathbf{1}} & (2\lambda_2\boldsymbol{I} + \boldsymbol{X}_{\cdot,E}^\top\boldsymbol{X}_{\cdot,E})^{-1} + \frac{(2\lambda_2\boldsymbol{I} + \boldsymbol{X}_{\cdot,E}^\top\boldsymbol{X}_{\cdot,E})^{-1}\boldsymbol{X}_{\cdot,E}^\top\mathbf{1}\mathbf{1}^\top\boldsymbol{X}_{\cdot,E}(2\lambda_2\boldsymbol{I} + \boldsymbol{X}_{\cdot,E}^\top\boldsymbol{X}_{\cdot,E})^{-1}}{\mathbf{1}^\top(\boldsymbol{I} - \boldsymbol{X}_{\cdot,E}(2\lambda_2\boldsymbol{I} + \boldsymbol{X}_{\cdot,E}^\top\boldsymbol{X}_{\cdot,E})^{-1}\boldsymbol{X}_{\cdot,E}^\top)\mathbf{1}} \end{pmatrix} \\
&\quad \times [1, \boldsymbol{X}_{\cdot,E}]^\top \\
&= \boldsymbol{X}_{\cdot,E}(2\lambda_2\boldsymbol{I} + \boldsymbol{X}_{\cdot,E}^\top\boldsymbol{X}_{\cdot,E})^{-1}\boldsymbol{X}_{\cdot,E}^\top \\
&\quad + \frac{(\mathbf{1} - \boldsymbol{X}_{\cdot,E}(2\lambda_2\boldsymbol{I} + \boldsymbol{X}_{\cdot,E}^\top\boldsymbol{X}_{\cdot,E})^{-1}\boldsymbol{X}_{\cdot,E}^\top\mathbf{1})(\mathbf{1}^\top - \mathbf{1}^\top\boldsymbol{X}_{\cdot,E}(2\lambda_2\boldsymbol{I} + \boldsymbol{X}_{\cdot,E}^\top\boldsymbol{X}_{\cdot,E})^{-1}\boldsymbol{X}_{\cdot,E}^\top)}{\mathbf{1}^\top(\boldsymbol{I} - \boldsymbol{X}_{\cdot,E}(2\lambda_2\boldsymbol{I} + \boldsymbol{X}_{\cdot,E}^\top\boldsymbol{X}_{\cdot,E})^{-1}\boldsymbol{X}_{\cdot,E}^\top)\mathbf{1}} \\
&= \boldsymbol{I} - \boldsymbol{J}
\end{aligned}$$

Now that we have showed that $\boldsymbol{H} + \boldsymbol{J} = \boldsymbol{I}$, we could also conclude that $\boldsymbol{J}_{ii} + \boldsymbol{H}_{ii} = 1$ and that the ALO solutions given by both the primal and dual approaches are equivalent.

2.3 ALO for Logistic Regression with Lasso penalty

First, let's rewrite the optimization problem with the loss functions separated for each observation, therefore the loss function goes:

$$-\sum_{i=1}^{n}\left[y_i \boldsymbol{x}_i^\top \boldsymbol{\beta} + \ln\left(1 + e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}\right)\right] + \lambda_1 \|\boldsymbol{\beta}\|_1,$$

where the individual loss function is $\ell(x_i^\top \boldsymbol{\beta}; y_i) = y_i x_i^\top \boldsymbol{\beta} + \ln(1 + e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}})$ and the regularizer is $R(\boldsymbol{\beta}) = \lambda_1 \|\boldsymbol{\beta}\|_1$, from which we could derive the dual optimal and the conjugate functions:

$$\hat{\boldsymbol{\theta}} = y - \frac{e^{\boldsymbol{X}\hat{\boldsymbol{\beta}}}}{1 + e^{\boldsymbol{X}\hat{\boldsymbol{\beta}}}}, \qquad \ell^*(-\theta_i; y_i) = (y_i - \theta_i)\ln\frac{y_i - \theta_i}{1 - (y_i - \theta_i)} - \ln\frac{1}{1 - (y_i - \theta_i)}, \qquad R^*(\boldsymbol{\beta}) = \begin{cases} 0 & \|\boldsymbol{\beta}\|_\infty \le \lambda_1, \\ \infty & \text{otherwise.} \end{cases}$$

From the results of the conjugate functions above, we could also obtain the derivatives of the loss functions and the Jacobian of the regularizer:

$$\dot{\ell}^*(-\theta_i; y_i) = \ln\frac{y_i - \theta_i}{1 - (y_i - \theta_i)}, \qquad \ddot{\ell}^*(-\theta_i; y_i) = \frac{1}{(y_i - \theta_i)[1 - (y_i - \theta_i)]}$$

Recall Eqn. 20 from the main paper, the quadratic surrogate of the dual problem is

$$\min_{\boldsymbol{u}} \frac{1}{2}\sum_{i=1}^{n}\left(u_i - \frac{\hat{\theta}_i \ddot{\ell}^*(-\hat{\theta}_i; y_i) + \hat{y}_i}{\sqrt{\ddot{\ell}^*(-\hat{\theta}_i; y_i)}}\right)^2 + R^*(X^\top K u),$$

where $\boldsymbol{K} = \operatorname{diag}\sqrt{\ddot{\ell}^*(-\hat{\theta}_i; y_i)}$. Therefore the Jacobian at $\boldsymbol{y}_u = \hat{\theta}_i \ddot{\ell}^*(-\hat{\theta}_i; y_i) + \hat{\theta}_i / \sqrt{\ddot{\ell}^*(-\hat{\theta}_i; y_i)}$ could locally be treated as the projection onto the orthogonal complement of the polyhedron $\{\|\boldsymbol{X}^\top K \boldsymbol{u}\|_\infty \le \lambda_1\}$, thus $\boldsymbol{J} = \boldsymbol{I} - \boldsymbol{X}_{u,E}(\boldsymbol{X}_{u,E}^\top \boldsymbol{X}_{u,E})^{-1}\boldsymbol{X}_{u,E}$, where $\boldsymbol{X}_{u,E}$ are the columns of $\boldsymbol{X}_u = \boldsymbol{X}^\top \boldsymbol{K}$, such that the columns in the set $E = \{i : |\boldsymbol{X}_i^\top \boldsymbol{\theta}| = \lambda_1\}$ are selected. Take everything to Eqn. 22, $y^{/i} = \boldsymbol{K}_{ii}(y_{u,i} - \boldsymbol{K}_{ii}\hat{\theta}_i / \boldsymbol{J}_{ii})$, we could obtain the ALO for the i-th observation.

2.4 ALO for Logistic Regression with Elastic Net Penalty

The optimization problem for logistic regression with elastic net penalty is:

$$-\sum_{i=1}^{n}\left[y_i x_i^\top \boldsymbol{\beta} + \ln\left(1 + e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}\right)\right] + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2^2.$$

The optimization problem is the same except the regularizer is changed, therefore the only thing different is the conjugate function of the regularizer, R^* and the corresponding Jacobian, here $R(\boldsymbol{\beta}) = \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2^2$:

$$R^*(\boldsymbol{\beta}) = \sum_{|u_i| > \lambda_1} \frac{(\lambda_1 - |u_i|)^2}{4\lambda_2}.$$

The corresponding Jacobian is $\boldsymbol{J} = (\boldsymbol{I} + \boldsymbol{X}_{u,E}\boldsymbol{X}_{u,E}^\top / 2\lambda_2)$, where $\boldsymbol{X}_{u,E}$ are the columns of $\boldsymbol{X}_u = \boldsymbol{X}^\top \boldsymbol{K}$, such that the columns in the set $E = \{i : |\boldsymbol{X}_i^\top \boldsymbol{\theta}| = \lambda_1\}$ are selected. Take everything to Eqn. 22, $y^{/i} = \boldsymbol{K}_{ii}(y_{u,i} - \boldsymbol{K}_{ii}\hat{\theta}_i / \boldsymbol{J}_{ii})$, we could obtain the ALO for the i-th observation.

2.5 ALO for Poisson Regression with Lasso penalty

The optimization function for Poisson regression with lasso penalty is:

$$\sum_{i=1}^{n} \left[-y_i \boldsymbol{x}_i^\top \boldsymbol{\beta} + e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}} + \ln(y_i!) \right] + \lambda_1 \|\boldsymbol{\beta}\|_1$$

The regularizer is the same as the logistic regression with the lasso penalty case, thus the Jacobian will also be the same, therefore we only have to focus on the loss function $\ell(x_i^\top \boldsymbol{\beta}; y_i) = -y_i x_i^\top \boldsymbol{\beta} + e^{x_i^\top \boldsymbol{\beta}} + \ln(y_i!)$. The optimal solution for the dual problem $\hat{\theta} = y - e^{\boldsymbol{X}\hat{\boldsymbol{\beta}}}$ and the conjugate of the loss function is $\ell^*(-\theta_i; y_i) = (y_i - \theta_i)\ln(y_i - \theta_i) - (y_i - \theta_i)$, the corresponding derivatives are therefore:

$$\dot{\ell}^*(-\theta_i; y_i) = \ln(y_i - \theta_i), \qquad \ddot{\ell}^*(-\theta_i; y_i) = \frac{1}{y_i - \theta_i}.$$

By plugging everything into Eqn. 22, we obtain the ALO for Poisson regression with the lasso penalty.

2.6 ALO for Poisson Regression with Elastic Net Penalty

The loss function for Poisson regression with elastic net penalty is the same as that of Poisson regression with the lasso penalty and the regularizer of it is the same as that of logistic regression with elastic net penalty. Thus by plugging everything into Eqn. 22, we could obtain the ALO for Poisson regression with elastic net penalty.

3 Approximation with Proximal Formulation

3.1 ALO for Linear Regression

For the elastic net problem, the proximal mapping is known to be

$$\mathbf{prox}_R(\boldsymbol{z}) = \gamma \operatorname{sgn}(\boldsymbol{z}) \odot (|\boldsymbol{z}| - \lambda \mathbf{1}_p)_+, \qquad \gamma = \frac{1}{1 + (1-\alpha)\lambda}. \tag{5}$$

Let E be the active set, if $z_i \in E$, then

$$\frac{\partial}{\partial z_i} \gamma \operatorname{sgn}(z_i)(|z_i| - \lambda)_+ = \gamma.$$

Plug in $\boldsymbol{z} = \hat{\boldsymbol{\beta}} - \sum_{j=1}^{n} \dot{\ell}(\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}};\, y_j)\boldsymbol{x}_j$, Eqn. 46 thus reduce to

$$\boldsymbol{H} = \gamma \boldsymbol{X}_{\cdot,E} \left[\gamma \boldsymbol{X}_{\cdot,E}^\top \boldsymbol{X}_{\cdot,E} + (1-\gamma)\, \boldsymbol{I}_{E,E} \right]^{-1} \boldsymbol{X}_{\cdot,E}^\top. \tag{6}$$

Bringing back the intercept term is straightforward as well. Noted that

$$\begin{bmatrix} \hat{\boldsymbol{\beta}}_0^{\setminus i} \\ \hat{\boldsymbol{\beta}}^{\setminus i} \end{bmatrix} = \mathbf{prox}_R(\boldsymbol{z}), \qquad \boldsymbol{z} = \begin{bmatrix} \hat{\boldsymbol{\beta}}_0^{\setminus i} \\ \hat{\boldsymbol{\beta}}^{\setminus i} \end{bmatrix} - \sum_{j \neq i} \begin{bmatrix} 1 \\ \boldsymbol{x}_j \end{bmatrix} \dot{\ell}\left(\hat{\boldsymbol{\beta}}_0^{\setminus i} + \boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}^{\setminus i};\, y_j \right).$$

Hence, from the first-order condition $\sum_{j\neq i} \dot{\ell}\left(\hat{\beta}_0^{\setminus i} + \boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}^{\setminus i}; y_j\right) = 0$, we can derive that

$$\boldsymbol{J}_{E,E} = [\boldsymbol{J}(\boldsymbol{u})]_{E,E} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & (1-\alpha)\lambda & 0 & \dots & 0 \\ 0 & 0 & (1-\alpha)\lambda & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & (1-\alpha)\lambda \end{bmatrix}^{-1}. \tag{7}$$

The ALO formula is then immediate by Thm. 5.1.

3.2 ALO for LASSO, with Intercept through Generalized LASSO

For the generalized LASSO:

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\| + \lambda\|\boldsymbol{D}\boldsymbol{\beta}\|_1, \tag{8}$$

the dual problem can be derived as:

$$\min_{\boldsymbol{u}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{\theta}\|_2^2, \qquad \boldsymbol{\theta} \in \{\boldsymbol{X}^\top\boldsymbol{\theta} = \boldsymbol{D}^\top\boldsymbol{u}, \|\boldsymbol{u}\|_\infty \leq \lambda\}. \tag{9}$$

The dual problem could be written in a proximal approach, such that:

$$\hat{\boldsymbol{u}} = \mathbf{prox}_R(\boldsymbol{y}), \qquad R(\boldsymbol{u}) = \begin{cases} 0 & \boldsymbol{\theta} \in \{\boldsymbol{X}^\top\boldsymbol{\theta} = \boldsymbol{D}^\top\boldsymbol{u}, \|\boldsymbol{u}\|_\infty \leq \lambda\}, \\ \infty & \text{otherwise.} \end{cases}$$

Denote $\boldsymbol{J}$ as the Jacobian of the proximal operator at the full data problem $\boldsymbol{y}$, then the ALO estimator could be obtained as:

$$\boldsymbol{y}^{\setminus i} = \boldsymbol{y}_i - \frac{\hat{\boldsymbol{u}}_i}{\boldsymbol{J}_{ii}}. \tag{10}$$

For the case of LASSO with an intercept, we could expand the $\boldsymbol{X}$ with a column of ones in the first column, expand $\boldsymbol{\beta}$ with another dimension and choose $\boldsymbol{D} = [\boldsymbol{0}, \boldsymbol{I}]$. Let $E \coloneqq \{j : |\boldsymbol{X}_j^\top\boldsymbol{\theta}| = \lambda\}$ denote the active set. The Jacobian is locally given as the projection onto the orthogonal complement of the span of $\boldsymbol{X}_E$ and the vector of ones. Further denote $\tilde{\boldsymbol{X}}_E = [\boldsymbol{1}, \boldsymbol{X}_E]$, then the Jacobian is given as $\boldsymbol{I} - \tilde{\boldsymbol{X}}_E(\tilde{\boldsymbol{X}}_E^\top\tilde{\boldsymbol{X}}_E)\tilde{\boldsymbol{X}}_E^\top$.

3.3 ALO for Elastic Net, without Penalty on Intercept through Generalized LASSO

Without penalty on intercept, the elastic net problem can be written as:

$$\begin{aligned} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\boldsymbol{\beta}} \end{bmatrix} &= \arg\min \frac{1}{2}\left\|\boldsymbol{y} - \beta_0 - \boldsymbol{X}\boldsymbol{\beta}\right\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_2^2 \\ &= \arg\min \frac{1}{2}\begin{bmatrix} \beta_0 \\ \boldsymbol{\beta} \end{bmatrix}^\top \left(\begin{bmatrix} 1 & \boldsymbol{X} \end{bmatrix}^\top \begin{bmatrix} 1 & \boldsymbol{X} \end{bmatrix} + \lambda_2 \operatorname{diag}(0; \boldsymbol{1}_p)\right)\begin{bmatrix} \beta_0 \\ \boldsymbol{\beta} \end{bmatrix} - \boldsymbol{y}^\top \begin{bmatrix} 1 & \boldsymbol{X} \end{bmatrix}\begin{bmatrix} \beta_0 \\ \boldsymbol{\beta} \end{bmatrix} + \lambda_1\|\boldsymbol{\beta}\|_1 \end{aligned}$$

where we assume that the size of $\boldsymbol{X}$ is $n \times p$. In the mean time, note the LASSO problem (also without penalty on intercept) is:

$$\begin{aligned}\begin{bmatrix}\hat{\beta}_0\\ \hat{\boldsymbol{\beta}}\end{bmatrix} &= \arg\min \frac{1}{2}\left\|\boldsymbol{y} - \beta_0 - \boldsymbol{X}\boldsymbol{\beta}\right\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1\\ &= \arg\min \frac{1}{2}\begin{bmatrix}\beta_0\\ \boldsymbol{\beta}\end{bmatrix}^\top \begin{bmatrix}\mathbf{1} & \boldsymbol{X}\end{bmatrix}^\top \begin{bmatrix}\mathbf{1} & \boldsymbol{X}\end{bmatrix}\begin{bmatrix}\beta_0\\ \boldsymbol{\beta}\end{bmatrix} - \boldsymbol{y}^\top \begin{bmatrix}\mathbf{1} & \boldsymbol{X}\end{bmatrix}\begin{bmatrix}\beta_0\\ \boldsymbol{\beta}\end{bmatrix} + \lambda_1\|\boldsymbol{\beta}\|_1\end{aligned}$$

Thus we can add some "observations" to the data and let

$$\boldsymbol{y}^* = \begin{bmatrix}\boldsymbol{y}\\ \mathbf{0}_p\end{bmatrix}, \qquad \boldsymbol{X}^* = \begin{bmatrix}\boldsymbol{X}\\ \sqrt{\lambda_2}\boldsymbol{I}_p\end{bmatrix},$$

then the elastic net becomes

$$\begin{aligned}\begin{bmatrix}\hat{\beta}_0\\ \hat{\boldsymbol{\beta}}\end{bmatrix} &= \arg\min \frac{1}{2}\left\|\boldsymbol{y}^* - \beta_0\begin{bmatrix}\mathbf{1}_n\\ \mathbf{0}_p\end{bmatrix} - \boldsymbol{X}^*\boldsymbol{\beta}\right\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1\\ &= \arg\min \frac{1}{2}\left\|\begin{bmatrix}\boldsymbol{y}\\ \mathbf{0}_p\end{bmatrix} - \begin{bmatrix}\mathbf{1}_n & \boldsymbol{X}\\ \mathbf{0}_p & \sqrt{\lambda_2}\boldsymbol{I}_p\end{bmatrix}\begin{bmatrix}\beta_0\\ \boldsymbol{\beta}\end{bmatrix}\right\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1,\end{aligned} \tag{11}$$

which is a special case of the general LASSO.

4 Derivation for Multinomial ALO

4.1 Loss Function

Assume we have K classes, n observations $\boldsymbol{X}$, and $\boldsymbol{R}^p$ parameters $\boldsymbol{\beta}_k$ for each class. Define leave-i-out variables as

$$\boldsymbol{y}^{\setminus i}_{(n-1)K\times 1} = \{y_{jk}\}_{j\neq i}^{k=1,\dots,K} = \begin{bmatrix}\begin{bmatrix}y_{11}\\ y_{12}\\ \vdots\\ y_{1K}\end{bmatrix}\\ \begin{bmatrix}y_{21}\\ y_{22}\\ \vdots\\ y_{2K}\end{bmatrix}\\ \vdots\\ \begin{bmatrix}y_{n1}\\ y_{n2}\\ \vdots\\ y_{nK}\end{bmatrix}\end{bmatrix}, \quad \boldsymbol{X}^{\setminus i}_{(n-1)K\times pK} = \begin{bmatrix}\begin{bmatrix}\boldsymbol{X}_1^\top & 0 & \cdots & 0\\ 0 & \boldsymbol{X}_1^\top & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \boldsymbol{X}_1^\top\end{bmatrix}\\ \begin{bmatrix}\boldsymbol{X}_2^\top & 0 & \cdots & 0\\ 0 & \boldsymbol{X}_2^\top & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \boldsymbol{X}_2^\top\end{bmatrix}\\ \vdots\\ \begin{bmatrix}\boldsymbol{X}_n^\top & 0 & \cdots & 0\\ 0 & \boldsymbol{X}_n^\top & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \boldsymbol{X}_n^\top\end{bmatrix}\end{bmatrix}, \quad \boldsymbol{\mathcal{B}} = \begin{bmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\\ \vdots\\ \boldsymbol{\beta}_K\end{bmatrix}$$

The loss function would be

$$\ell(\boldsymbol{\mathcal{B}}) = -\left\{\sum_{j\neq i}\left[\sum_{k=1}^K y_{jk}\boldsymbol{X}_j^\top\boldsymbol{\beta}_k - \log\left(\sum_{k=1}^K e^{\boldsymbol{X}_j^\top\boldsymbol{\beta}_k}\right)\right]\right\} = \sum_{j\neq i}\log\left(\sum_{k=1}^K e^{\boldsymbol{X}_j^\top\boldsymbol{\beta}_k}\right) - \sum_{j\neq i}^n\sum_{k=1}^K y_{jk}\boldsymbol{X}_j^\top\boldsymbol{\beta}_k$$

By taking the first order derivative, we get

$$\frac{\partial \ell(\boldsymbol{\mathcal{B}})}{\partial \boldsymbol{\mathcal{B}}} = \begin{bmatrix} \frac{\partial \ell(\boldsymbol{\mathcal{B}})}{\partial \boldsymbol{\beta}_1} \\ \vdots \\ \frac{\partial \ell(\boldsymbol{\mathcal{B}})}{\partial \boldsymbol{\beta}_K} \end{bmatrix} = \begin{bmatrix} \sum_{j \neq i} \frac{\exp\left(\boldsymbol{X}_j^\top \boldsymbol{\beta}_1\right)}{\sum_{k=1}^K \exp\left(\boldsymbol{X}_j^\top \boldsymbol{\beta}_k\right)} \boldsymbol{X}_j - \sum_{j \neq i} y_{j1} \boldsymbol{X}_j \\ \vdots \\ \sum_{j \neq i} \frac{\exp\left(\boldsymbol{X}_j^\top \boldsymbol{\beta}_K\right)}{\sum_{k=1}^K \exp\left(\boldsymbol{X}_j^\top \boldsymbol{\beta}_k\right)} \boldsymbol{X}_j - \sum_{j \neq i} y_{jK} \boldsymbol{X}_j \end{bmatrix} = \boldsymbol{\mathcal{X}}^{/iT} \begin{bmatrix} \begin{bmatrix} \frac{\exp\left(\boldsymbol{X}_1^\top \boldsymbol{\beta}_1\right)}{\sum_{k=1}^K \exp\left(\boldsymbol{X}_1^\top \boldsymbol{\beta}_k\right)} \\ \vdots \\ \frac{\exp\left(\boldsymbol{X}_1^\top \boldsymbol{\beta}_K\right)}{\sum_{k=1}^K \exp\left(\boldsymbol{X}_1^\top \boldsymbol{\beta}_k\right)} \end{bmatrix} \\ \begin{bmatrix} \frac{\exp\left(\boldsymbol{X}_2^\top \boldsymbol{\beta}_1\right)}{\sum_{k=1}^K \exp\left(\boldsymbol{X}_2^\top \boldsymbol{\beta}_k\right)} \\ \vdots \\ \frac{\exp\left(\boldsymbol{X}_2^\top \boldsymbol{\beta}_K\right)}{\sum_{k=1}^K \exp\left(\boldsymbol{X}_2^\top \boldsymbol{\beta}_k\right)} \end{bmatrix} \\ \vdots \\ \begin{bmatrix} \frac{\exp\left(\boldsymbol{X}_n^\top \boldsymbol{\beta}_1\right)}{\sum_{k=1}^K \exp\left(\boldsymbol{X}_n^\top \boldsymbol{\beta}_k\right)} \\ \vdots \\ \frac{\exp\left(\boldsymbol{X}_n^\top \boldsymbol{\beta}_K\right)}{\sum_{k=1}^K \exp\left(\boldsymbol{X}_n^\top \boldsymbol{\beta}_k\right)} \end{bmatrix} \end{bmatrix}_{(n-1)K \times 1} - \boldsymbol{\mathcal{X}}^{/iT} \boldsymbol{y}^{/i}$$

$$= \boldsymbol{\mathcal{X}}^{/iT} \left[\boldsymbol{\mathcal{A}}^{/i}(\boldsymbol{\beta}) - \boldsymbol{y}^{/i} \right] = \boldsymbol{\mathcal{X}}^{/iT} \left(\begin{bmatrix} \boldsymbol{A}_1(\boldsymbol{\beta}) \\ \boldsymbol{A}_2(\boldsymbol{\beta}) \\ \vdots \\ \boldsymbol{A}_n(\boldsymbol{\beta}) \end{bmatrix} - \boldsymbol{y}^{/i} \right)$$

Similarly, we can get

$$\frac{\partial^2 \ell(\boldsymbol{\mathcal{B}})}{\partial \boldsymbol{\mathcal{B}} \partial \boldsymbol{\mathcal{B}}^\top} = \boldsymbol{\mathcal{X}}^{/iT} \frac{\partial \boldsymbol{\mathcal{A}}^{/i}(\boldsymbol{\mathcal{B}})}{\partial \boldsymbol{\mathcal{B}}^\top} = \boldsymbol{\mathcal{X}}^{/iT} \begin{bmatrix} \frac{\partial \boldsymbol{A}_1(\boldsymbol{\mathcal{B}})}{\partial \boldsymbol{\mathcal{B}}^\top} \\ \frac{\partial \boldsymbol{A}_2(\boldsymbol{\mathcal{B}})}{\partial \boldsymbol{\mathcal{B}}^\top} \\ \vdots \\ \frac{\partial \boldsymbol{A}_n(\boldsymbol{\mathcal{B}})}{\partial \boldsymbol{\mathcal{B}}^\top} \end{bmatrix}$$

where,

$$\frac{\partial \boldsymbol{A}_j(\boldsymbol{\mathcal{B}})}{\partial \boldsymbol{\mathcal{B}}^\top} = \begin{bmatrix} \frac{\exp\left(\boldsymbol{X}_j^\top \boldsymbol{\beta}_1\right)}{\sum_{k=1}^K \exp\left(\boldsymbol{X}_j^\top \boldsymbol{\beta}_k\right)} & 0 & \cdots & 0 \\ 0 & \frac{\exp\left(\boldsymbol{X}_j^\top \boldsymbol{\beta}_2\right)}{\sum_{k=1}^K \exp\left(\boldsymbol{X}_j^\top \boldsymbol{\beta}_k\right)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{\exp\left(\boldsymbol{X}_j^\top \boldsymbol{\beta}_K\right)}{\sum_{k=1}^K \exp\left(\boldsymbol{X}_j^\top \boldsymbol{\beta}_k\right)} \end{bmatrix}_{K\times K} \begin{bmatrix} \boldsymbol{X}_j^\top & 0 & \cdots & 0 \\ 0 & \boldsymbol{X}_j^\top & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{X}_j^\top \end{bmatrix}_{K\times pK}$$

$$- \begin{bmatrix} \frac{\exp\left(\boldsymbol{X}_j^\top \boldsymbol{\beta}_1\right)}{\sum_{k=1}^K \exp\left(\boldsymbol{X}_j^\top \boldsymbol{\beta}_k\right)} \\ \frac{\exp\left(\boldsymbol{X}_j^\top \boldsymbol{\beta}_2\right)}{\sum_{k=1}^K \exp\left(\boldsymbol{X}_j^\top \boldsymbol{\beta}_k\right)} \\ \vdots \\ \frac{\exp\left(\boldsymbol{X}_j^\top \boldsymbol{\beta}_K\right)}{\sum_{k=1}^K \exp\left(\boldsymbol{X}_j^\top \boldsymbol{\beta}_k\right)} \end{bmatrix}_{K\times 1} \begin{bmatrix} \frac{\exp\left(\boldsymbol{X}_j^\top \boldsymbol{\beta}_1\right)}{\sum_{k=1}^K \exp\left(\boldsymbol{X}_j^\top \boldsymbol{\beta}_k\right)} & \frac{\exp\left(\boldsymbol{X}_j^\top \boldsymbol{\beta}_2\right)}{\sum_{k=1}^K \exp\left(\boldsymbol{x}_j^\top \boldsymbol{\beta}_k\right)} & \cdots & \frac{\exp\left(\boldsymbol{x}_j^\top \boldsymbol{\beta}_K\right)}{\sum_{k=1}^K \exp\left(\boldsymbol{x}_j^\top \boldsymbol{\beta}_k\right)} \end{bmatrix}_{1\times K} \begin{bmatrix} \boldsymbol{x}_j^\top & 0 & \cdots & 0 \\ 0 & \boldsymbol{x}_j^\top & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{x}_j^\top \end{bmatrix}_{K\times pK}$$

$$= \left[\operatorname{diag}\left(\boldsymbol{A}_j(\boldsymbol{\mathcal{B}})\right) - \boldsymbol{A}_j(\boldsymbol{\mathcal{B}})\boldsymbol{A}_j(\boldsymbol{\mathcal{B}})^\top\right]_{K\times K} \begin{bmatrix} \boldsymbol{x}_j^\top & 0 & \cdots & 0 \\ 0 & \boldsymbol{x}_j^\top & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{x}_j^\top \end{bmatrix}_{K\times pK}$$

So, we have

$$
\begin{aligned}
\frac{\partial^2 \ell(\boldsymbol{\mathcal{B}})}{\partial \boldsymbol{\mathcal{B}} \partial \boldsymbol{\mathcal{B}}^\top} =& \boldsymbol{\mathcal{X}}^{/iT} \frac{\partial \boldsymbol{\mathcal{A}}^{/i}(\boldsymbol{\mathcal{B}})}{\partial \boldsymbol{\mathcal{B}}^\top} = \boldsymbol{\mathcal{X}}^{/iT} \begin{bmatrix} \frac{\partial \boldsymbol{A}_1(\boldsymbol{\mathcal{B}})}{\partial \boldsymbol{\mathcal{B}}^\top} \\ \frac{\partial \boldsymbol{A}_2(\boldsymbol{\mathcal{B}})}{\partial \boldsymbol{\mathcal{B}}^\top} \\ \vdots \\ \frac{\partial \boldsymbol{A}_n(\boldsymbol{\mathcal{B}})}{\partial \boldsymbol{\mathcal{B}}^\top} \end{bmatrix} \\
=& \boldsymbol{\mathcal{X}}^{/iT} \begin{bmatrix} \left[\operatorname{diag}(\boldsymbol{A}_1(\boldsymbol{\mathcal{B}})) - \boldsymbol{A}_1(\boldsymbol{\mathcal{B}})\boldsymbol{A}_1(\boldsymbol{\mathcal{B}})^\top\right]_{K\times K} \begin{bmatrix} \boldsymbol{x}_1^\top & 0 & \cdots & 0 \\ 0 & \boldsymbol{x}_1^\top & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{x}_1^\top \end{bmatrix}_{K\times pK} \\ \left[\operatorname{diag}(\boldsymbol{A}_2(\boldsymbol{\mathcal{B}})) - \boldsymbol{A}_2(\boldsymbol{\mathcal{B}})\boldsymbol{A}_2(\boldsymbol{\mathcal{B}})^\top\right]_{K\times K} \begin{bmatrix} \boldsymbol{x}_2^\top & 0 & \cdots & 0 \\ 0 & \boldsymbol{x}_2^\top & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{x}_2^\top \end{bmatrix}_{K\times pK} \\ \vdots \\ \left[\operatorname{diag}(\boldsymbol{A}_n(\boldsymbol{\mathcal{B}})) - \boldsymbol{A}_n(\boldsymbol{\mathcal{B}})\boldsymbol{A}_n(\boldsymbol{\mathcal{B}})^\top\right]_{K\times K} \begin{bmatrix} \boldsymbol{x}_n^\top & 0 & \cdots & 0 \\ 0 & \boldsymbol{x}_n^\top & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{x}_n^\top \end{bmatrix}_{K\times pK} \end{bmatrix}_{(n-1)K\times pK} \\
=& \boldsymbol{\mathcal{X}}^{/iT} \boldsymbol{\mathcal{D}}^{/i}_{(n-1)K\times(n-1)K}(\boldsymbol{\mathcal{B}}) \boldsymbol{\mathcal{X}}^{/i}
\end{aligned}
$$

where,

$$
\boldsymbol{\mathcal{D}}^{/i}(\boldsymbol{\mathcal{B}}) = \begin{bmatrix} \left[\operatorname{diag}(\boldsymbol{A}_1(\boldsymbol{\mathcal{B}})) - \boldsymbol{A}_1(\boldsymbol{\mathcal{B}})\boldsymbol{A}_1(\boldsymbol{\mathcal{B}})^\top\right] & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \left[\operatorname{diag}(\boldsymbol{A}_n(\boldsymbol{\mathcal{B}})) - \boldsymbol{A}_n(\boldsymbol{\mathcal{B}})\boldsymbol{A}_n(\boldsymbol{\mathcal{B}})^\top\right] \end{bmatrix}
$$

4.2 Newton's Method

With Newton's method, we have the one step update as

$$
\begin{aligned}
\tilde{\boldsymbol{\mathcal{B}}}^{/i} =& \hat{\boldsymbol{\mathcal{B}}} - \left[\boldsymbol{\mathcal{X}}^{/iT} \boldsymbol{\mathcal{D}}^{/i}(\boldsymbol{\mathcal{B}}) \boldsymbol{\mathcal{X}}^{/i} + \nabla^2 R(\boldsymbol{\mathcal{B}})\right]^{-1} \left[\boldsymbol{\mathcal{X}}^{/iT} \left(\boldsymbol{\mathcal{A}}^{/iT}(\boldsymbol{\mathcal{B}}) - \boldsymbol{y}^{/i}\right) + \nabla R(\boldsymbol{\mathcal{B}})\right] \\
=& \hat{\boldsymbol{\mathcal{B}}} + \left[\boldsymbol{\mathcal{X}}^\top \boldsymbol{\mathcal{D}}(\boldsymbol{\mathcal{B}}) \boldsymbol{\mathcal{X}} + \nabla^2 R(\boldsymbol{\mathcal{B}}) - \boldsymbol{X}_i^\top \left[\operatorname{diag}(\boldsymbol{A}_i(\boldsymbol{\mathcal{B}})) - \boldsymbol{A}_i(\boldsymbol{\mathcal{B}})\boldsymbol{A}_i(\boldsymbol{\mathcal{B}})^\top\right] \boldsymbol{X}_i\right]^{-1} \boldsymbol{X}_i^\top \left(\boldsymbol{A}_i(\boldsymbol{\mathcal{B}}) - \boldsymbol{y}_i\right)
\end{aligned}
$$

where,

$$
\boldsymbol{X}_i = \begin{bmatrix} \boldsymbol{x}_i^\top & 0 & \cdots & 0 \\ 0 & \boldsymbol{x}_i^\top & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{x}_i^\top \end{bmatrix}_{K\times pK}, \quad \boldsymbol{y}_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iK} \end{bmatrix}
$$

Defining $\mathcal{K}(\mathcal{B}) = \mathcal{X}^\top \mathcal{D}(\mathcal{B})\mathcal{X} + \nabla^2 R(\mathcal{B})$, with matrix inversion lemma, we can get

$$\begin{aligned}\tilde{\mathcal{B}}^{/i} =& \hat{\mathcal{B}} + \left[\mathcal{K}(\mathcal{B}) - \boldsymbol{X}_i^\top \left[\operatorname{diag}\left(\boldsymbol{A}_i(\mathcal{B})\right) - \boldsymbol{A}_i(\mathcal{B})\boldsymbol{A}_i(\mathcal{B})^\top\right] \boldsymbol{X}_i\right]^{-1} \boldsymbol{X}_i^\top \left(\boldsymbol{A}_i(\mathcal{B}) - \boldsymbol{y}_i\right) \\ =& \mathcal{B} + \mathcal{K}(\mathcal{B})^{-1} \boldsymbol{X}_i^\top \left(\boldsymbol{A}_i(\mathcal{B}) - \boldsymbol{y}_i\right) \\ & - \mathcal{K}(\mathcal{B})^{-1} \boldsymbol{X}_i^\top \left\{-\left[\operatorname{diag}\left(\boldsymbol{A}_i(\mathcal{B})\right) - \boldsymbol{A}_i(\mathcal{B})\boldsymbol{A}_i(\mathcal{B})^\top\right]^{-1} + \boldsymbol{X}_i \mathcal{K}(\mathcal{B})^{-1} \boldsymbol{X}_i^\top\right\}^{-1} \boldsymbol{X}_i \mathcal{K}(\mathcal{B})^{-1} \boldsymbol{X}_i^\top \left(\boldsymbol{A}_i(\mathcal{B}) - \boldsymbol{y}_i\right)\end{aligned}$$

4.3 Approximate Leave-i-Out Prediction

Given the approximate leave-i-out estimation, we can do the approximate leave-i-out prediciton as

$$\begin{aligned}\boldsymbol{y}_i^{/i} =& \begin{bmatrix} y_{i1}^{/i} \\ \vdots \\ y_{iK}^{/i} \end{bmatrix} = \boldsymbol{X}_i \tilde{\mathcal{B}}^{/i} \\ =& \boldsymbol{X}_i \mathcal{B} + \boldsymbol{X}_i \mathcal{K}(\mathcal{B})^{-1} \boldsymbol{X}_i^\top \left(\boldsymbol{A}_i(\mathcal{B}) - \boldsymbol{y}_i\right) \\ & - \boldsymbol{X}_i \mathcal{K}(\mathcal{B})^{-1} \boldsymbol{X}_i^\top \left\{-\left[\operatorname{diag}\left(\boldsymbol{A}_i(\mathcal{B})\right) - \boldsymbol{A}_i(\mathcal{B})\boldsymbol{A}_i(\mathcal{B})^\top\right]^{-1} + \boldsymbol{X}_i \mathcal{K}(\mathcal{B})^{-1} \boldsymbol{X}_i^\top\right\}^{-1} \boldsymbol{X}_i \mathcal{K}(\mathcal{B})^{-1} \boldsymbol{X}_i^\top \left(\boldsymbol{A}_i(\mathcal{B}) - \boldsymbol{y}_i\right)\end{aligned}$$