Notes on Approximate Leave-One-Out for Elastic Net

Linyun He Wanchao Qin Peng Xu Yuze Zhou

August 28, 2018

1 ALO for Elastic Net, Approximation in the Primal Domain

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

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

2 ALO for Elastic Net, Approximation in the Dual Domain

The original problem for elastic net is to solve for $\hat{\boldsymbol{\beta}}$ such that:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left(\frac{1}{2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2^2 \right) \tag{5}$$

By adding the Lagrangian, we get the formulation of L:

$$L = \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{z}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2^2 + u^\top (\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta}). \tag{6}$$

The original problem is solving the primal of the Lagrangian such that $p^* = \min_{\boldsymbol{\beta},\boldsymbol{z}} \max_{\boldsymbol{u}} L$ and the dual formulation $d^* = \max_{\boldsymbol{u}} \min_{\boldsymbol{\beta},\boldsymbol{z}} L$, to minimize over $\boldsymbol{z}$:

$$\frac{\partial L}{\partial \boldsymbol{z}} = \boldsymbol{z} - \boldsymbol{y} + \boldsymbol{u} = \boldsymbol{0} \implies \boldsymbol{y} = \boldsymbol{u} + \boldsymbol{z}.$$

Since $\boldsymbol{\beta}$ is penalized element-wisely, we can minimize over $\boldsymbol{\beta}$ by minimizing over each $\boldsymbol{\beta}_i$, that is, we have to minimize $\lambda_1 |\beta_i| + \lambda_2 \beta_i^2 - \boldsymbol{u}^\top \boldsymbol{X}_i \boldsymbol{\beta}$ for each dimension of $\boldsymbol{\beta}$, where $\boldsymbol{X}_i$ denotes the ith column of $\boldsymbol{X}$, therefore:

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

By taking all the above to the Lagrangian, we obtain the dual problem d^* as:

$$d^* = \min_{\boldsymbol{u}} \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{u}\|_2^2 + \sum_{j : |\boldsymbol{X}_j^\top \boldsymbol{u}| > \lambda_1} \frac{(\lambda_1 - |\boldsymbol{u}^\top \boldsymbol{X}_i|)^2}{4\lambda_2}. \tag{7}$$

The minimizer $\hat{\boldsymbol{u}}$ could also be obtained from the dual problem through a proximal approach:

$$\hat{\boldsymbol{u}} = \mathbf{prox}_R(y), \qquad R(\boldsymbol{u}) = \sum_{j : |\boldsymbol{X}_j^\top \boldsymbol{u}| > \lambda_1} \frac{(\lambda_1 - |\boldsymbol{u}^\top \boldsymbol{X}_i|)^2}{4\lambda_2}.$$

By replacing the full data problem $\boldsymbol{y}$ with $\boldsymbol{y}_\alpha = \boldsymbol{y} + (y_i^{\setminus i} - y_i) e_i$, where $y_i^{\setminus i}$ is the true LOO estimator and e_i is the i-th standard vector, and let $\boldsymbol{u}^{\setminus i} = \mathbf{prox}_R(\boldsymbol{y}_\alpha)$, we have:

$$\begin{aligned} 0 &= e_i^\top \boldsymbol{u}^{\setminus i} \\ &= e_i^\top \, \mathbf{prox}_R(\boldsymbol{y}_\alpha) \\ &\approx e_i^\top [\mathbf{prox}_R(\boldsymbol{y}) + \boldsymbol{J}_R(\boldsymbol{y})(\boldsymbol{y}_\alpha - \boldsymbol{y})] \\ &\approx \hat{u}_i + \boldsymbol{J}_{ii}(y_i^{\setminus i} - y_i). \end{aligned}$$

Here $\boldsymbol{J}_R(\boldsymbol{y})$ denotes the Jacobian matrix of the proximal operator at $\boldsymbol{y}$, thus the ALO estimator $\tilde{y}_i$ is obtained as

$$\tilde{y}_i = y_i - \frac{\hat{u}_i}{\boldsymbol{J}_{ii}}. \tag{8}$$

The Jacobian could locally be obtained as:

$$\boldsymbol{J}_R(\boldsymbol{y}) = (\boldsymbol{I} + \nabla^2 R(\mathbf{prox}_R(\boldsymbol{y})))^{-1} = (\boldsymbol{I} + \nabla^2 R(\hat{\boldsymbol{u}}))^{-1} = \left(\boldsymbol{I} + \frac{1}{2\lambda_2}\boldsymbol{X}_E\boldsymbol{X}_E^\top\right)^{-1} \tag{9}$$

for $E = \{j : |\boldsymbol{X}_j^\top \boldsymbol{u}| > \lambda_1\}$.

3 ALO for Elastic Net, Approximation with Proximal Formulation

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

$$\mathbf{prox}_R(\boldsymbol{z}) = \gamma \operatorname{sgn}(\boldsymbol{z}) \odot (|\boldsymbol{z}| - \lambda \mathbf{1}_p)_+, \qquad \gamma = \frac{1}{1 + (1-\alpha)\lambda}. \tag{10}$$

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

$$\boldsymbol{H} = \gamma \boldsymbol{X}_{\cdot,E}\left[\gamma \boldsymbol{X}_{\cdot,E}^\top \boldsymbol{X}_{\cdot,E} + (1-\gamma)\,\boldsymbol{I}_{E,E}\right]^{-1}\boldsymbol{X}_{\cdot,E}^\top. \tag{11}$$

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

$$\boldsymbol{J}_{E,E} = [\boldsymbol{J}(\boldsymbol{u})]_{E,E} = \begin{bmatrix}1 & 0 & 0 & \dots & 0 \\ 0 & (1-\alpha)\lambda & 0 & \dots & 0 \\ 0 & 0 & (1-\alpha)\lambda & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & (1-\alpha)\lambda\end{bmatrix}^{-1}. \tag{12}$$

The ALO formula is then immediate by Thm. 5.1.

4 ALO for LASSO, with Intercept through Generalized LASSO

For the generalized LASSO:

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\| + \lambda\|\boldsymbol{D}\boldsymbol{\beta}\|_1, \tag{13}$$

the dual problem can be derived as:

$$\min_{\boldsymbol{u}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{\theta}\|_2^2, \qquad \boldsymbol{\theta} \in \{\boldsymbol{X}^\top\boldsymbol{\theta} = \boldsymbol{D}^\top\boldsymbol{u}, \|\boldsymbol{u}\|_\infty \le \lambda\}. \tag{14}$$

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

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

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:

$$\boldsymbol{y}^{\setminus i} = \boldsymbol{y}_i - \frac{\hat{\boldsymbol{u}}_i}{\boldsymbol{J}_{ii}}. \tag{15}$$

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 $\boldsymbol{D} = [\mathbf{0}, \boldsymbol{I}]$. 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 $\tilde{\boldsymbol{X}}_E = [\mathbf{1}, \boldsymbol{X}_E]$, then the Jacobian is given as $\boldsymbol{I} - \tilde{\boldsymbol{X}}_E(\tilde{\boldsymbol{X}}_E^\top \tilde{\boldsymbol{X}}_E)\tilde{\boldsymbol{X}}_E^\top$.

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

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

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

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:

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

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

$$\begin{aligned} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\boldsymbol{\beta}} \end{bmatrix} &= \arg\min \frac{1}{2} \left\| \boldsymbol{y}^* - \beta_0 \begin{bmatrix} \mathbf{1}_n \\ \mathbf{0}_p \end{bmatrix} - \boldsymbol{X}^* \boldsymbol{\beta} \right\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 \\ &= \arg\min \frac{1}{2} \left\| \begin{bmatrix} \boldsymbol{y} \\ \mathbf{0}_p \end{bmatrix} - \begin{bmatrix} \mathbf{1}_n & \boldsymbol{X} \\ \mathbf{0}_p & \sqrt{\lambda_2} \boldsymbol{I}_p \end{bmatrix} \begin{bmatrix} \beta_0 \\ \boldsymbol{\beta} \end{bmatrix} \right\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1, \end{aligned} \tag{16}$$

which is a special case of the general LASSO.

6 Generalized Loss function

6.1 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

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

Adding in intercept terms is then straightforward.

6.2 ALO for Logistic Regression, Approximation in the Dual Domain

6.2.1 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.

6.2.2 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.

6.3 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

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

6.4 ALO for Poisson Regression, Approximation in the Dual Domain

6.4.1 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.

6.4.2 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.

7 Usage of ALO formulae with `glmnet` package

The `glmnet` package scales the elastic net loss function by a factor of $1/n$. Furthermore, for linear problems `glmnet` implicitly "standardizes y to have unit variance before computing its λ sequence (and then unstandardizes the resulting coefficients)" (*cf.* [Glmnet Vignette]). So to get comparable results, it is necessary to rescale $\boldsymbol{y}$ by the MLE $\hat{\sigma}_y$ before fitting the model. More precisely, `glmnet` is in fact optimizing the following problem:

$$\min_{\boldsymbol{\beta}} \frac{1}{2n} \sum_{j=1}^{n} \left(\frac{\boldsymbol{x}_j^\top \boldsymbol{\beta}}{\hat{\sigma}_y} - \frac{y_j}{\hat{\sigma}_y} \right)^2 + \frac{\lambda}{\hat{\sigma}_y} \alpha \|\boldsymbol{\beta}\|_1 + \frac{\lambda}{\hat{\sigma}_y^2} \frac{1-\alpha}{2} \|\boldsymbol{\beta}\|_2^2. \tag{17}$$

We thus have

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

Hence, for the linear elastic net problem, the primal ALO is:

$$\tilde{y}_j^{\setminus i} = \hat{y}_j + \frac{\boldsymbol{H}_{ii}(\hat{y}_j - y_j)}{n\hat{\sigma}_y - \boldsymbol{H}_{ii}}, \qquad \boldsymbol{H} = \boldsymbol{X}_{\cdot,A} \left[\frac{1}{n\hat{\sigma}_y} \boldsymbol{X}_{\cdot,A}^\top \boldsymbol{X}_{\cdot,A} + \frac{(1-\alpha)\,\lambda}{\hat{\sigma}_y^2} \boldsymbol{I}_{A,A} \right]^{-1} \boldsymbol{X}_{\cdot,A}^\top.$$

We get further complication when option `standardization = T` is given, in which case `glmnet` first standardize the data $\boldsymbol{X}$ using $\hat{\sigma}_{\boldsymbol{X}}$ (assuming $\boldsymbol{X}$ is standardized):

- If `intercept = F`, compute $\boldsymbol{X}^* = \text{diag}[\hat{\sigma}_y \hat{\sigma}_{\boldsymbol{X}}]^{-1} \boldsymbol{X}$.
- If `intercept = T`, compute $\boldsymbol{X}^* = \text{diag}[\hat{\sigma}_y \hat{\sigma}_{\boldsymbol{X}}]^{-1} (\boldsymbol{X} - \bar{\boldsymbol{X}} \mathbf{1}^\top)$.

Afterwards, the the coefficients are returned de-standardized:

- Let $(\beta_0, \boldsymbol{\beta})$ be the original intercept and coefficients, compute $\boldsymbol{\beta}^* = \hat{\sigma}_y \, \text{diag}[\hat{\sigma}_{\boldsymbol{X}}]^{-1} \boldsymbol{\beta}$ then $\beta_0^* = \beta_0 - \bar{\boldsymbol{X}} \boldsymbol{\beta}^*$.
- Return pair $(\beta_0^*, \boldsymbol{\beta}^*)$.

For logistics and Poisson regression the standardization procedure is basically the same, except `glmnet` no longer standardize by $\hat{\sigma}_y$, which make sense since $\boldsymbol{y}$ is now either categorical or count data.

8 Benchmark

n	p	k	Average ALO	Average 5-fold CV	Relative	n full fit
300	100	60	0.016	0.053	3.313	1.2
500	800	500	0.251	0.533	2.124	36.5
1000	1200	800	0.489	1.200	2.454	211.0
2500	2000	1200	2.267	3.623	1.598	1577.5
5000	2500	2000	5.017	8.097	1.614	7740.0
10000	10000	2500	27.236	36.520	1.341	62530.0

Table 1: Average elapsed time (in seconds) comparison, 10 runs, $\alpha = 0.5$.