Exclusive Group Lasso

Coleman Zhang

August 24, 2020

1 Introduction

Assume design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, response $\mathbf{y} \in \mathbb{R}^n$. Let us consider penalized linear regression, where we are minimizing the objective:

$$\underset{\beta}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \boldsymbol{\Omega}(\boldsymbol{\beta}) \tag{1}$$

where $\lambda \in \mathbb{R}_{++}$, and $\boldsymbol{\Omega}(\cdot) : \mathbb{R}^p \to \mathbb{R}_+$ defined as

$$\boldsymbol{\Omega}(\boldsymbol{\beta}) = \frac{1}{2}\sum_{g \in \mathcal{G}} \omega_g \|\boldsymbol{\beta}_g\|_1^2 = \frac{1}{2}\sum_{g \in \mathcal{G}} \omega_g \left(\sum_{j \in g} |\beta_j|\right)^2 \tag{2}$$

is the exclusive group lasso norm. The group weights should be a positive vector, i.e.,

$$\omega_g \geq 0 \ \forall g \in \mathcal{G}$$

A convention introduced in the group lasso is to pick the group weight to be proportional to the group size: e.g. $\omega_g := n_g$. In addition, we limit group assignments to be mutually exclusive, and collectively exhaustive, i.e.,

$$g_k \cap g_l = \varnothing \ \forall k \neq l, \ \bigcup_k g_k = \{1, \ldots, p\} = [p]$$

In this case one could prove that $\boldsymbol{\Omega}(\cdot)$ is a norm. In addition, our objective is convex, but not differentiable at $\beta_i = 0$ for some i.

Definition 1. *(Vector Norm) Let $\|\cdot\| : \mathbb{C}^m \to \mathbb{R}_+$. Then $\|\cdot\|$ is norm if $\forall\ x, y \in \mathbb{C}^m$ and $\forall\ \alpha \in \mathbb{C}$*

- $x \neq 0 \implies \|x\| > 0$ *(positive definite);*
- $\|\alpha x\| = |\alpha|\|x\|$ *(homogeneous);*
- $\|x + y\| \leq \|x\| + \|y\|$ *(triangle inequality).*

2 Literature

This work relies on extensive results from convex analysis and optimization theory. For a review, see Boyd & Vandenberghe [BV04, V20], Beck [B17], and Tibshirani's lecture notes [T18]. In particular, the ability to derive the dual program using the KKT conditions, taking the Fenchel conjugate of functions, and computing the dual norm is especially important.

Kong et. al [KF14, KL16] provided an iterative re-weighted least square algorithm. Sun et. al [SC20] built on this work and provided a bisection algorithm that solves the lasso problem at each iteration. Campbell & Allen [CA15] provided a coordinate descent method and extensive theoretical analysis including a result for the dual norm of $\mathbf{\Omega}(\cdot)$.

However, none of these work provides an algorithm that adopts both the state-of-the-art coordinate descent, and convergence checks based on strong duality. Xiang et. al [XW14] and Fercoq et. al [FG15], for example, provided screening rules to safely set coefficients to 0 using duality of the lasso. Ndiaye et. al [NF16,NF17, WY14] generalizes the "duality gap safe" screening rule first to the sparse group lasso [SF13], and finally to a general sparsity-enforcing penalty. We follow the methodologies of the previously mentioned work for the following reasons. First, (block) coordinate descent achieves the state-of-the-art performance in terms of speed of convergence. Second, we could check if the algorithm converges and early stop by computing the duality gap.

3 Primal, Dual, & Optimality

Let $P_\lambda(\boldsymbol{\beta})$ be the primal objective,

$$\hat{\boldsymbol{\beta}}^{(\lambda)} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}}\ P_\lambda(\boldsymbol{\beta}) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}}\ \frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\mathbf{\Omega}(\boldsymbol{\beta})$$

This is equivalent to,

$$\begin{aligned} &\text{(P)} \quad \min_{\boldsymbol{\beta}\in\mathbb{R}^p,\mathbf{z}\in\mathbb{R}^n} \frac{1}{2}\|\mathbf{y}-\mathbf{z}\|_2^2 + \frac{\lambda}{2}\sum_{g\in\mathcal{G}}\omega_g\|\boldsymbol{\beta}_g\|_1^2 \\ &\qquad\quad \text{s.t. } \mathbf{z} = \mathbf{X}\boldsymbol{\beta} \end{aligned} \tag{3}$$

The Lagrangian dual of this problem is,

$$\begin{aligned} &\min_{\boldsymbol{\beta}\in\mathbb{R}^p,\mathbf{z}\in\mathbb{R}^n} \frac{1}{2}\|\mathbf{y}-\mathbf{z}\|_2^2 + \frac{\lambda}{2}\sum_{g\in\mathcal{G}}\omega_g\|\boldsymbol{\beta}_g\|_1^2 + \mathbf{u}^T(\mathbf{z}-\mathbf{X}\boldsymbol{\beta}) \\ &= \min_{\mathbf{z}\in\mathbb{R}^n} \frac{1}{2}\|\mathbf{y}-\mathbf{z}\|_2^2 + \mathbf{u}^T\mathbf{z} + \min_{\boldsymbol{\beta}\in\mathbb{R}^p} \lambda\sum_{g\in\mathcal{G}}\frac{1}{2}\omega_g\|\boldsymbol{\beta}_g\|_1^2 - \mathbf{u}^T\mathbf{X}\boldsymbol{\beta} \\ &= \frac{1}{2}\|\mathbf{y}\|_2^2 - \frac{1}{2}\|\mathbf{y}-\mathbf{u}\|_2^2 - \lambda\max_{\boldsymbol{\beta}\in\mathbb{R}^p}\left(\frac{\mathbf{X}^T\mathbf{u}}{\lambda}\right)^T\boldsymbol{\beta} - \mathbf{\Omega}(\boldsymbol{\beta}) \\ &= \frac{1}{2}\|\mathbf{y}\|_2^2 - \frac{1}{2}\|\mathbf{y}-\mathbf{u}\|_2^2 - \lambda\mathbf{\Omega}^*\left(\frac{\mathbf{X}^T\mathbf{u}}{\lambda}\right) \end{aligned} \tag{4}$$

where $\mathbf{\Omega}^*$ is the Fenchel conjugate of $\mathbf{\Omega}$,

Definition 2. *(Conjugate functions) Let $f:\mathbb{R}^n\to\mathbb{R}$. The function $f^*:\mathbb{R}^n\to\mathbb{R}$ defined as*

$$f^*(\boldsymbol{y}) = \sup_{\boldsymbol{x}\in dom f}\left(\boldsymbol{y}^T\boldsymbol{x} - f(\boldsymbol{x})\right)$$

is called the conjugate of function f.

Property. *The conjugate function has the following calculus rules:*

- *(Scaling) Let $f : \mathbb{E} \to (-\infty, \infty]$ and let $\alpha \in \mathbb{R}_{++}$. The conjugate of $g(\mathbf{x}) = \alpha f(\mathbf{x})$ is given by*

$$g^*(\mathbf{y}) = \alpha f^*(\frac{\mathbf{y}}{\alpha})$$

- *(Separable functions) Let $g : \mathbb{E}_1 \times \mathbb{E}_2 \times \cdots \times \mathbb{E}_p \to (-\infty, \infty]$ be given by $g(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p) = \sum_{i=1}^{p} f_i(\mathbf{x}_i)$, where $f_i : \mathbb{E}_i \to (-\infty, \infty]$ is a proper function for any $i = 1, 2, \ldots, p$. Then*

$$g^*(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_p) = \sum_{i=1}^{p} f_i^*(\mathbf{y}_i)$$

Definition 3. *(Dual norm) Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$. The associated dual norm, denoted $\|\cdot\|_*$, is defined as*

$$\|\boldsymbol{z}\|_* = \sup_{\|\boldsymbol{x}\| \le 1} \boldsymbol{z}^T \boldsymbol{x}$$

From the definition we have the following Hölder's inequality,

$$\mathbf{z}^T \mathbf{x} \le \|\mathbf{x}\| \|\mathbf{z}\|_*$$

Example. *(Norm squared).* Now consider the function $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2$, where $\|\cdot\|$ is a norm, with dual norm $\|\cdot\|_*$. From Hölder's inequality,

$$\mathbf{y}^T \mathbf{x} - \frac{1}{2}\|\mathbf{x}\|^2 \le \|\mathbf{x}\| \|\mathbf{y}\|_* - \frac{1}{2}\|\mathbf{x}\|^2$$

for all $\mathbf{x}$. The righthand side is a quadratic function of $\|\mathbf{x}\|$, which has maximum $\frac{1}{2}\|\mathbf{y}\|_*^2$. Therefore for all $\mathbf{x}$, we have

$$\mathbf{y}^T \mathbf{x} - \frac{1}{2}\|\mathbf{x}\|^2 \le \frac{1}{2}\|\mathbf{y}\|_*^2$$

which shows that $f^*(\mathbf{y}) \le \frac{1}{2}\|\mathbf{y}\|_*^2$.

To show the other inequality, pick $\mathbf{x}$ to be such that $\mathbf{y}^T \mathbf{x} = \|\mathbf{y}\|_* \|\mathbf{x}\|$, scaled so that $\|\mathbf{x}\| = \|\mathbf{y}\|_*$. Then, from definition and by construction,

$$f^*(\mathbf{y}) \ge \mathbf{y}^T \mathbf{x} - \frac{1}{2}\|\mathbf{x}\|^2 = \frac{1}{2}\|\mathbf{y}\|_*^2.$$

In conclusion, it must be that $f^*(\mathbf{y}) = \frac{1}{2}\|\mathbf{y}\|_*^2$

Proposition 1. *Let*

$$\boldsymbol{\Omega}(\boldsymbol{\beta}) = \frac{1}{2}\sum_{g \in \mathcal{G}} \omega_g \|\boldsymbol{\beta}_g\|_1^2 = \frac{1}{2}\sum_{g \in \mathcal{G}} \omega_g \left(\sum_{j \in g} |\beta_j|\right)^2$$

Then the conjugate of $\boldsymbol{\Omega}$ is

$$\boldsymbol{\Omega}^*(\boldsymbol{\xi}) = \frac{1}{2}\sum_{g \in \mathcal{G}} \frac{1}{\omega_g} \|\boldsymbol{\xi}_g\|_\infty^2 = \frac{1}{2}\sum_{g \in \mathcal{G}} \frac{1}{\omega_g} \left(\max_{j \in g} |\xi_j|\right)^2$$

Proof. First by observation $\mathbf{\Omega}$ is group separable. Define $\mathbf{\Omega}(\boldsymbol{\beta}) = \sum_{g\in\mathcal{G}} \mathbf{\Omega}_g(\boldsymbol{\beta}_g) = \sum_{g\in\mathcal{G}} \frac{1}{2}\omega_g \|\boldsymbol{\beta}_g\|_1^2$, then $\mathbf{\Omega}^*(\boldsymbol{\xi}) = \sum_{g\in\mathcal{G}} \mathbf{\Omega}_g^*(\boldsymbol{\xi}_g)$. Define $f(\boldsymbol{\beta}_g) = \frac{1}{2}\|\boldsymbol{\beta}_g\|_1^2$, then $\mathbf{\Omega}_g(\boldsymbol{\beta}_g) = \omega_g f(\boldsymbol{\beta}_g)$. By the previous example and the calculus rule for scaling, we have $\mathbf{\Omega}_g^*(\boldsymbol{\xi}_g) = \frac{1}{2}\omega_g \|\frac{\boldsymbol{\xi}_g}{\omega_g}\|_\infty^2 = \frac{1}{2}\frac{1}{\omega_g}\|\boldsymbol{\xi}_g\|_\infty^2$. Hence $\mathbf{\Omega}^*(\boldsymbol{\xi}) = \sum_{g\in\mathcal{G}} \mathbf{\Omega}_g^*(\boldsymbol{\xi}_g) = \frac{1}{2}\sum_{g\in\mathcal{G}} \frac{1}{\omega_g}\|\boldsymbol{\xi}_g\|_\infty^2$. □

From the above illustration, the dual formulation is given by,

$$\text{(D)} \ \max_{\mathbf{u}\in\mathbb{R}^n} \frac{1}{2}\|\mathbf{y}\|_2^2 - \frac{1}{2}\|\mathbf{y}-\mathbf{u}\|_2^2 - \frac{1}{2\lambda}\sum_{g\in\mathcal{G}} \frac{1}{\omega_g}\|\mathbf{X}_g^T\mathbf{u}\|_\infty^2 =: D_\lambda(\boldsymbol{u}) \tag{5}$$

By Slater's condition, strong duality applies. If $\boldsymbol{\beta}^*$ and $\boldsymbol{u}^*$ are both primal and dual optimal, then there is no duality gap, i.e., $P_\lambda(\boldsymbol{\beta}^*) = D_\lambda(\boldsymbol{u}^*)$.

4 Subgradient, KKT conditions

We establish the following optimality condition based on subgradient:

Proposition 2. *(Fermat's Rule) (see (Bauschke and Combettes, 2011, Proposition 26.1) for a more general result) For any convex function $f:\mathbb{R}^d \to \mathbb{R}$,*

$$x^* \in \underset{x}{argmin}\ f(x) \implies 0 \in \partial f(x^*)$$

From the sufficient conditions for optimality, if $\boldsymbol{\beta}^*$ and $\boldsymbol{u}^*$ satisfy KKT conditions (Fermat's Rule + primal dual feasible), then they are primal and dual optimal, which implies zero duality gap. In the following section, we derive the subgradient equations for the exclusive lasso problem.

Recall,

$$\text{(P)} \ \min_{\beta} \frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \frac{\lambda}{2}\sum_{g\in\mathcal{G}}\omega_g\left(\sum_{j\in g}|\beta_j|\right)^2$$

The subgradient equation for a particular β_i satisfies:

$$\mathbf{X}_i^T(\mathbf{r}_{-i} - \mathbf{X}_i\beta_i) = \lambda\omega_g\left(\sum_{k\in g\setminus i}|\beta_k| + |\beta_i|\right)\partial|\beta_i| \tag{6}$$

where

$$\boldsymbol{r}_{-i} = \boldsymbol{y} - \sum_{j\neq i}\boldsymbol{X}_j\beta_j$$

is the partial residual, and subgradient of ℓ_1 norm

$$\partial|\beta_i| = \begin{cases} -1 & \text{if } \beta_i < 0 \\ [-1,1] & \text{if } \beta_i = 0 \\ 1 & \text{if } \beta_i > 0 \end{cases} \tag{7}$$

For the lasso problem, the subgradient equations are crucial for deriving the soft-threshold operator used in coordinate-wise descent. For a brief review, please see Tibshirani's lecture notes. Here it is possible to derive a closed form solution. Define the soft-threshold operator (at level $\tau \geq 0$) for $\mathbf{x} \in \mathbb{R}^d$ as $[\mathcal{S}_\tau(\mathbf{x})]_j = \text{sign}(x_j)(|x_j| - \tau)_+$,

$$\begin{aligned}
\tilde{\tau} &= \frac{\lambda\omega_g \sum_{k\in g\setminus i} |\beta_k|}{\lambda\omega_g + \|\mathbf{X}_i\|_2^2} \\
\tilde{z} &= \frac{\mathbf{X}_i^T \mathbf{r}_{-i}}{\lambda\omega_g + \|\mathbf{X}_i\|_2^2} \\
\hat{\beta}_i &= \mathcal{S}_{\tilde{\tau}}(\tilde{z})
\end{aligned} \tag{8}$$

5 Stopping criterion

We stop according the following criterion,

$$P_\lambda(\hat{\boldsymbol{\beta}}) - D_\lambda(\boldsymbol{r}) < \epsilon$$

which reflects the tolerance on the duality gap.

6 Bonus: Sparse Exclusive Group Lasso

Consider the penalty

$$\begin{aligned}
\boldsymbol{\Omega}\left(\boldsymbol{\beta}\right) &= \boldsymbol{\Omega}_1^{1-\alpha}\left(\boldsymbol{\beta}\right) + \boldsymbol{\Omega}_2^{\alpha}\left(\boldsymbol{\beta}\right) \\
&= \frac{1}{2}(1-\alpha)\sum_{g\in\mathcal{G}} \omega_g \|\boldsymbol{\beta}_g\|_1^2 + \alpha\|\boldsymbol{\beta}\|_1
\end{aligned} \tag{9}$$

where $\alpha \in [0, 1]$.

Theorem 1. *(conjugate of sum). Let $h_1 : \mathbb{E} \to (-\infty, \infty]$ be a proper convex function and $h_2 : \mathbb{E} \to \mathbb{R}$ be a real-valued convex function. Then*

$$(h_1 + h_2)^* = h_1^* \Box h_2^*$$

where $\Box$ denotes the infimal convolution operator, defined as

$$(f_1 \Box f_2)(\mathbf{x}) = \inf_{\mathbf{u}\in\mathbb{E}} f_1(\mathbf{u}) + f_2(\mathbf{x} - \mathbf{u})$$

We know that,

$$\begin{aligned}
(\boldsymbol{\Omega}_1^{1-\alpha})^*\left(\boldsymbol{\xi}\right) &= \frac{1}{2}\sum_{g\in\mathcal{G}} \frac{1}{(1-\alpha)\omega_g}\|\boldsymbol{\xi}_g\|_\infty^2 \\
(\boldsymbol{\Omega}_2^{\alpha})^*\left(\boldsymbol{\xi}\right) &= \mathbf{I}_{\mathcal{B}_\infty}(\boldsymbol{\xi}/\alpha) = \sum_{g\in\mathcal{G}} \mathbf{I}_{\mathcal{B}_\infty}(\boldsymbol{\xi}_g/\alpha)
\end{aligned} \tag{10}$$

By definition we have,

$$
\begin{aligned}
\boldsymbol{\Omega}^*(\boldsymbol{\xi}) &= ((\boldsymbol{\Omega}_1^{1-\alpha})^* \Box (\boldsymbol{\Omega}_2^{\alpha})^*)(\boldsymbol{\xi}) \\
&= \inf_{\boldsymbol{\eta}} (\boldsymbol{\Omega}_1^{1-\alpha})^*(\boldsymbol{\xi} - \boldsymbol{\eta}) + (\boldsymbol{\Omega}_2^{\alpha})^*(\boldsymbol{\eta}) \\
&= \sum_{g \in \mathcal{G}} \inf_{\boldsymbol{\eta}_g} \frac{1}{2} \frac{1}{(1-\alpha)\omega_g} \|\boldsymbol{\xi}_g - \boldsymbol{\eta}_g\|_\infty^2 + \mathbf{I}_{\mathcal{B}_\infty}(\boldsymbol{\eta}_g/\alpha) \\
&= \sum_{g \in \mathcal{G}} \inf_{\|\boldsymbol{\eta}_g\|_\infty \le \alpha} \frac{1}{2} \frac{1}{(1-\alpha)\omega_g} \|\boldsymbol{\xi}_g - \boldsymbol{\eta}_g\|_\infty^2
\end{aligned} \tag{11}
$$

which is equivalent to solving

$$
\begin{aligned}
\nu_g^* = &\min_{\boldsymbol{\eta}_g} \|\boldsymbol{\xi}_g - \boldsymbol{\eta}_g\|_\infty \\
&\text{s.t. } \|\boldsymbol{\eta}_g\|_\infty \le \alpha
\end{aligned} \tag{12}
$$

We can see that $\boldsymbol{\eta}_g^*(\boldsymbol{\xi}_g)$ is indeed the projection of $\boldsymbol{\xi}$ on $\alpha\mathcal{B}_\infty$, which admits closed form solution:

$$
\left[\boldsymbol{\eta}_g^*(\boldsymbol{\xi}_g)\right]_i = \left[\mathbf{P}_{\alpha\mathcal{B}_\infty}(\boldsymbol{\xi}_g)\right]_i = \begin{cases} \alpha & \text{if } \left[\boldsymbol{\xi}_g\right]_i > \alpha \\ \left[\boldsymbol{\xi}_g\right]_i & \text{if } |\left[\boldsymbol{\xi}_g\right]_i| \le \alpha \\ -\alpha & \text{if } \left[\boldsymbol{\xi}_g\right]_i < -\alpha \end{cases} \tag{13}
$$

Hence (12) can be solved as

$$
\begin{aligned}
\nu_g^* &= \|\mathcal{S}_\alpha\left(\boldsymbol{\xi}_g\right)\|_\infty \\
\boldsymbol{\Omega}^*(\boldsymbol{\xi}) &= \sum_{g \in \mathcal{G}} \frac{1}{2} \frac{1}{(1-\alpha)\omega_g} \|\mathcal{S}_\alpha\left(\boldsymbol{\xi}_g\right)\|_\infty^2
\end{aligned} \tag{14}
$$

The soft-threshold operator for the sparse exclusive group lasso problem becomes:

$$
\begin{aligned}
\tilde{\tau} &= \frac{\lambda(1-\alpha)\omega_g \sum_{k \in g \setminus i} |\beta_k| + \lambda\alpha}{\lambda(1-\alpha)\omega_g + \|\mathbf{X}_i\|_2^2} \\
\tilde{z} &= \frac{\mathbf{X}_i^T \mathbf{r}_{-i}}{\lambda(1-\alpha)\omega_g + \|\mathbf{X}_i\|_2^2} \\
\hat{\beta}_i &= \mathcal{S}_{\tilde{\tau}}(\tilde{z})
\end{aligned} \tag{15}
$$

Design choice: not choosing

$$
\|\boldsymbol{\beta}\|_{1,2} = \sqrt{\sum_{g \in \mathcal{G}} \omega_g \|\boldsymbol{\beta}_g\|_1^2}
$$

with dual norm [B17]

$$
\|\boldsymbol{\beta}\|_{1,2}^D = \sqrt{\sum_{g \in \mathcal{G}} \frac{1}{\omega_g} \|\boldsymbol{\beta}_g\|_\infty^2}
$$

since the penalty is not coordinate-wise separable, so it might be problematic to apply coordinate descent algorithm.

Question: What is $\lambda_{\max}$? How to determine the dual scaling constant to find the dual feasible point?

6.1 Group Level & Feature Level Screening

For λ large enough, $\mathbf{0} \in \partial P_\lambda(\boldsymbol{\beta})$. Using first-order conditions, we can determine $\lambda_{\max}$. For a particular lambda, we can determine when a particular $\beta_i = 0$ or $\boldsymbol{\beta}_g = 0$.

Feature level screening:

$$\mathbf{X}_j^T \left(\mathbf{r}_{-j} - \mathbf{X}_j \beta_j\right) = \lambda\alpha\partial|\beta_j| + \lambda(1-\alpha)\omega_g\|\boldsymbol{\beta}_g\|_1\partial|\beta_j|$$

$$\forall j \in g,\ |\mathbf{X}_j^T \mathbf{r}_{-j}| < \lambda\left(\alpha + (1-\alpha)\omega_g \sum_{k \in g\setminus j} |\beta_k|\right) \implies \hat{\beta}_j = 0.$$

Group level screening:

$$\mathbf{X}_g^T \left(\mathbf{r}_{-g} - \mathbf{X}_g \boldsymbol{\beta}_g\right) = \lambda\alpha\partial\|\boldsymbol{\beta}_g\|_1 + \lambda(1-\alpha)\omega_g\|\boldsymbol{\beta}_g\|_1\partial\|\boldsymbol{\beta}_g\|_1$$

$$\forall g \in \mathcal{G},\ \|\mathbf{X}_g^T \mathbf{r}_{-g}\|_\infty < \lambda\alpha \implies \hat{\boldsymbol{\beta}}_g = 0.$$

$$\lambda_{\max} = \max_{g \in \mathcal{G}} \|\mathbf{X}_g^T \mathbf{r}_{-g}\|_\infty / \alpha$$

$$\forall \lambda > \lambda_{\max},\ \hat{\boldsymbol{\beta}} = \mathbf{0}$$

References

[BV04] S. Boyd and L. Vandenberghe "Convex Optimization," Cambridge University Press, 2004.

[B17] A. Beck "First-Order Methods in Optimization," Society for Industrial and Applied Mathematics, 2017.

[T18] R. Tibshirani "Convex Optimization Fall 2018," `https://www.stat.cmu.edu/~ryantibs/convexopt-F18`, 2018.

[V20] L. Vandenberghe "ECE236C - Optimization Methods for Large-Scale Systems," `http://www.seas.ucla.edu/~vandenbe/ee236c.html`, 2020.

[KF14] D. Kong and R. Fujimaki and J. Liu and F. Nie and C. Chris "Exclusive Feature Learning on Arbitrary Structures via $\ell_{1,2}$ -norm," Advances in Neural Information Processing Systems 27, 2014.

[KL16] D. Kong and J. Liu and B. Liu and X. Bao "Uncorrelated Group LASSO," AAAI, 2016.

[SC20] Y. Sun and B. Chain and S. Kaski and J. Shawe-Taylor "Correlated Feature Selection with Extended Exclusive Group Lasso," 2020.

[CA15] F. Campbell and G. Allen "Within Group Variable Selection through the Exclusive Lasso," Electronic Journal of Statistics, 2015.

[FG15] O. Fercoq and A. Gramfort and J. Salmon "Mind the duality gap: safer rules for the Lasso," Proceedings of Machine Learning Research, 2015.

[NF16] E. Ndiaye and O. Fercoq and A. Gramfort and J. Salmon "GAP Safe Screening Rules for Sparse-Group Lasso," Advances in Neural Information Processing Systems 29, 2016.

[NF17] E. Ndiaye and O. Fercoq and A. Gramfort and J. Salmon "Gap Safe Screening Rules for Sparsity Enforcing Penalties," Journal of Machine Learning Research, 2017.

[WY14] J. Wang and J. Ye "Two-Layer Feature Reduction for Sparse-Group Lasso via Decomposition of Convex Sets," Journal of Machine Learning Research, 2014.

[XW14] Z. Xiang and Y. Wang and P. Ramadge "Screening Tests for Lasso Problems," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014.

[SF13] N. Simon and J. Friedman and T. Hastie and R. Tibshirani "A sparse-group lasso," Journal of Computational and Graphical Statistics, 2013.