Recap

KKT (Karush-Kuhn-Tucker) conditions
necessary and sufficient conditions for a solution of the Lasso objective function

$$G_j(\hat{\beta}) = -\mathrm{sign}(\hat{\beta}_j)\lambda \text{ if } \hat{\beta}_j \neq 0$$
$$|G_j(\hat{\beta})| \leq \lambda \text{ if } \hat{\beta}_j = 0$$

where

$$G(\beta) = -2X^T(Y - X\beta)/n$$

(sub-differential must contain the zero element)

sparsity is potentially induced at points of non-differentiability (here the components of β_j)

Coordinate descent algorithms

for optimization, exploiting the KKT conditions

path following algorithms:
compute $\{\hat{\beta}_j(\lambda)\}_{j=1}^{p}$ over all values of $\lambda \in \mathbb{R}^+$

the coefficient paths are typically "non-monotone" in the non-zeros
it may happen that

$$\hat{\beta}_j(\lambda) \neq 0,\ \hat{\beta}_j(\lambda') = 0 \text{ for } \lambda' < \lambda$$

Generalized Linear Models (GLMs)

univariate response Y, covariate $X \in \mathcal{X} \subseteq \mathbb{R}^p$

GLM: $Y_1, \ldots, Y_n$ independent

$$g(\mathbb{E}[Y_i | X_i = x]) = \underbrace{\mu + \sum_{j=1}^{p} \beta_j x^{(j)}}_{= f(x) = f_{\mu,\beta}(x)}$$

μ in the model: one cannot simply center the data

$g(\cdot)$ real-valued, known link function

Lasso: ℓ_1-norm regularized maximum likelihood estimation

$$\hat{\mu}, \hat{\beta} = \text{argmin}_{\mu,\beta}(\underbrace{-\ell(\mu, \beta)}_{\text{neg. log-likelihood}} + \lambda \|\beta\|_1)$$

Group Lasso (Yuan and Lin, 2006)

groups $\mathcal{G}_1, \ldots, \mathcal{G}_q$ which build a partition of $\{1, \ldots, p\}$
write the (high-dimensional) parameter vector as

$$\beta = (\beta_{\mathcal{G}_1}, \beta_{\mathcal{G}_2}, \ldots, \beta_{\mathcal{G}_q})^T$$

goal: an estimator which is "group-sparse", i.e.:
for all $j = 1, \ldots, p$,

$$\text{either } \hat{\beta}_{\mathcal{G}_j} \equiv 0$$
$$\text{or } (\hat{\beta}_{\mathcal{G}_j})_r \neq 0 \ \forall r \in \mathcal{G}_j$$

Group Lasso:

$$\hat{\beta}(\lambda) = \text{argmin}_{\beta} \left(\|Y - X\beta\|_2^2/n + \lambda \sum_{j=1}^{q} m_j \|\beta_{\mathcal{G}_j}\|_2 \right)$$

where typically $m_j = \sqrt{|\mathcal{G}_j|}$

group sparsity because objective function is non-differentiable at $\|\beta_{\mathcal{G}_j}\|_2 = 0 \iff \beta_{\mathcal{G}_j} \equiv 0$ $(j = 1, \ldots, q)$

objective function is non-differentiable at $\|\beta_{\mathcal{G}_j}\|_2$
sub-differential:

$$
\begin{aligned}
& \frac{\partial}{\partial \beta_{\mathcal{G}_j}} \left(\|Y - X\beta\|_2^2/n + \lambda \sum_{j=1}^{q} m_j \|\beta_{\mathcal{G}_j}\|_2 \right) \\
= \quad & G(\beta)_{\mathcal{G}_j} + \lambda m_j E(\beta_{\mathcal{G}_j}) \\
& E(\beta_{\mathcal{G}_j} = \{e \in \mathbb{R}^{|\mathcal{G}_j|};\ e = \frac{\beta_{\mathcal{G}_j}}{\|\beta_{\mathcal{G}_j}\|_2} \text{ if } \beta_{\mathcal{G}_j} \neq 0,\ \|e\|_2 \leq 1 \text{ if } \beta_{\mathcal{G}_j} = 0\}
\end{aligned}
$$

KKT conditions: solution is characterized by

$$0 \in \text{ sub-differential}$$

either $\underbrace{\hat{\beta}_{\mathcal{G}_j} \equiv 0}_{\text{point of non-differentiability}}$ or $(\hat{\beta}_{\mathcal{G}_j})_r \neq 0 \; \forall r \in \mathcal{G}_j$

why the second "or $(\hat{\beta}_{\mathcal{G}_j})_r \neq 0 \; \forall r \in \mathcal{G}_j$?" (when $\|\hat{\beta}_{\mathcal{G}_j}\|_2 \neq 0$)
$\rightsquigarrow$

$$0 = G(\hat{\beta})_{\mathcal{G}_j} + \lambda m_j \frac{\hat{\beta}_{\mathcal{G}_j}}{\|\hat{\beta}_{\mathcal{G}_j}\|_2}$$

suppose $X^T X/n = I$ (orthonormal design) and $\exists r \; (\hat{\beta}_{\mathcal{G}_j})_r = 0$:

$$0 = (-2X^T Y/n)_{\mathcal{G}_j} + 2\hat{\beta}_{\mathcal{G}_j} + \lambda m_j \frac{\hat{\beta}_{\mathcal{G}_j}}{\|\hat{\beta}_{\mathcal{G}_j}\|_2}$$

rth component $\quad 0 = -2((X^T Y/n)_{\mathcal{G}_j})_r + 0 + 0$

but it will not happen that $X^T Y$ is zero (random noise in Y)

Sparse Group Lasso

(Simon, Friedman, Hastie & Tibshirani, 2013)

$$\hat{\beta}(\lambda,\alpha) = \mathrm{argmin}_{\beta} \left(\|Y - X\beta\|_2^2/n + (1-\alpha)\lambda \sum_{j=1}^{q} m_j \|\beta_{\mathcal{G}_j}\|_2 + \alpha\lambda\|\beta\|_1 \right)$$

convex combination of Group Lasso and Lasso penalties
$\leadsto$ may also lead to sparsity within groups for $\alpha > 0$