Constrained Problem and Algorithm II

Jong-June Jeon

October 10, 2023

Department of Statistics, University of Seoul

Applications of constrained optimization

Example 1 (Linear regression with constraints of positive coefficients)

An average response of a variable y is determined by x_1 and x_2. Denote the ith observation of y and (x_1, x_2) by y_i and (x_{i1}, x_{i2}). When the positive constraint of a regression coefficient is required, a linearly contained optimization can be applied.

- (Model) $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$, where $\beta_2 \geq 0$
- (Optimization problem)

$$\begin{aligned} \min \quad & \frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2})^2 \\ \text{subject to} \quad & \beta_2 \geq 0 \end{aligned}$$

(Continue with the example)

Let $Y = (y_1, \cdots, y_n)^\top$ and $\tilde{X}$ be the $n \times 2$ data table and $1 \in \mathbb{R}^n$ be the one-column vector. Let $X = (1, \tilde{X}) \in \mathbb{R}^{n\times 3}$, $\beta = (\beta_0, \beta_1, \beta_2)$, and $G = (0, 0, -1)$. Then, the objective function is written by

$$\begin{aligned}\frac{1}{2n}\|Y - X\beta\|^2 &= \frac{1}{2n}(Y - X\beta)^\top(Y - X\beta) \\ &= \frac{1}{2}\beta^\top\left(\frac{X^\top X}{n}\right)\beta - \left(\frac{X^\top Y}{n}\right)^\top \beta + \frac{1}{2n}Y^\top Y,\end{aligned}$$

and the constraint is written by $G\beta \leq 0$.

(Continue with the example)

Thus, in the QP

- $P = X^\top X/n$
- $q = -X^\top Y/n$ and $r = Y^\top Y/n$
- $G = (0, 0, -1) \in \mathbb{R}^{1\times 3}$ and $h = 0 \in \mathbb{R}$
- $A = 0$ and $b = 0$

Example 2 (Logistic linear regression with constraints of positive coefficients)

Modify the example 1 by letting $y \in \{0, 1\}$. The optimization problem for obtaining the MLE is given by

$$\begin{aligned} \min \quad & \frac{1}{n} \sum_{i=1}^{n} (-y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}) + \log(1 + \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}))) \\ \text{subject to} \quad & \beta_2 \geq 0. \end{aligned}$$

Write $L(\beta) = \frac{1}{n} \sum_{i=1}^{n} (-y_i x_i^\top \beta + \log(1 + \exp(x_i^\top \beta)))$, where $x_i = (1, x_{i1}, x_{i2})^\top \in \mathbb{R}^3$ and $\beta = (\beta_0, \beta_1, \beta_2)^\top \in \mathbb{R}^3$.

(Continue with the example)

The quadratic approximation of $L(\beta)$ at $\beta^{(t)}$ is given by

$$
\begin{aligned}
f(\beta;\beta^{(t)}) &= L(\beta^{(t)}) + \nabla L(\beta^{(t)})^\top(\beta-\beta^{(t)}) + \frac{1}{2}(\beta-\beta^{(t)})^\top\nabla^2 L(\beta^{(t)})(\beta-\beta^{(t)}) \\
&= \frac{1}{2}\beta^\top\nabla^2 L(\beta^{(t)})\beta + \left(\nabla L(\beta^{(t)}) - \nabla^2 L(\beta^{(t)})\beta^{(t)}\right)^\top \beta \\
&\quad + \frac{1}{2}\beta^{(t)\top}\nabla^2 L(\beta^{(t)})\beta^{(t)} - \nabla L(\beta^{(t)})^\top\beta^{(t)} + \frac{1}{2}\beta^{(t)\top}\nabla^2 L(\beta^{(t)})\beta^{(t)}
\end{aligned}
$$

(Continue with the example)

Thus, in the QP

- $P = \nabla^2 L(\beta^{(t)})$
- $q = \nabla L(\beta^{(t)}) - \nabla^2 L(\beta^{(t)})\beta^{(t)}$
- $G = (0, 0, -1)$ and $h = 0$

With the P, q, G and h, we can solve $\min f(\beta; \beta^{(t)})$ with the constraint $G\beta \leq 0$.

(Continue with the example)

Computation of $\nabla L(\beta^{(t)})$ and $\nabla^2 L(\beta^{(t)})$: let $\hat{p}(x_i) = 1/(1 - \exp(x_i^\top \hat{\beta}^{(t)}))$.

$$
\begin{aligned}
\nabla L(\beta^{(t)}) &= \frac{1}{n}\sum_{i=1}^{n}(\hat{p}(x_i) - y_i)x_i \in \mathbb{R}^3 \\
\nabla^2 L(\beta^{(t)}) &= \frac{1}{n}\sum_{i=1}^{n}\hat{p}(x_i)(1 - \hat{p}(x_i))x_i x_i^\top \in \mathbb{R}^{3\times 3}
\end{aligned}
$$

Thus, the P and the q in the QP are computed.

(Continue with the example)

(algorithm)

1. Set an initial $\beta^{(0)}$ and $t = 0$
2. $\beta^{(t+1)} \leftarrow \text{argmin} f(\beta; \beta^{(t)})$ with $G\beta \leq 0$.
3. check the convergence of $\beta^{(t+1)}$. If $\beta^{(t+1)}$ converges, stop the algorithm. Otherwise, $t \leftarrow t + 1$ and go to the step 2.

Example 3 (Linear regression with ordered positive coefficients)

- (Model) $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$, where $0 \le \beta_1 \le \beta_2$
- (Optimization problem)

$$
\begin{aligned}
\min \quad & \frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2})^2 \\
\text{subject to} \quad & -\beta_1 \le 0 \\
& \beta_1 - \beta_2 \le 0
\end{aligned}
$$

There are two constraints given by $G\beta \le 0$, where

$$
G = \begin{pmatrix} 0 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix}.
$$

Example 4 (Linear regression with l_1-penalty)

- (Model) $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$.
- (Optimization problem)

$$\min \qquad \frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2})^2 + \lambda(|\beta_1| + |\beta_2|),$$

where $\lambda \geq 0$ is a tuning parameter.

Note that the minimizer of β depends on the section of λ. It is known as the LASSO estimator.

(Continue with the example)

Note that this example just shows an application of solving the regression problem with l_1-penalty. More efficient algorithms have been developed.

Let $\beta_j^+ = \max(\beta_j, 0)$ and $\beta_j^- = \max(-\beta_j, 0)$. Then, $\beta_j = \beta_j^+ - \beta_j^-$, $|\beta_j| = \beta_j^+ + \beta_j^-$ and $\beta_j x_{ij} = \beta_j^+ x_{ij} + \beta_j^-(-x_{ij})$.

Let $\beta = (\beta_0, \beta_1^+, \beta_1^-, \beta_2^+, \beta_2^-)^\top$ and $x_i = (1, x_{i1}, -x_{i1}, x_{i2}, -x_{i2})^\top$ and $d = (0, 1, 1, 1, 1)^\top$. The objective function is written by

$$\frac{1}{2n}\|Y - X\beta\|^2 + \lambda d^\top \beta.$$

(Continue with the example)

The constraints are $\beta_j^+, \beta_j^- \geq 0$. Thus, $G\beta \leq 0$, where

$$G = \begin{pmatrix} 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 \end{pmatrix}.$$

In QP, to avoid the singular problem ($\det(X^\top X) = 0$), the term of $\eta\|\beta\|^2$ with a small $\eta > 0$ is added in the objective function.

Example 5 (Fused lasso [Tibshirani et al., 2005])

- (Model) $y_i = \beta_0 + \mu_i + \epsilon_i$, where $\epsilon_i \sim (0, \sigma^2)$.
- (Optimization problem)

$$\min \quad \frac{1}{2n}\sum_{i=1}^{n}(y_i - \mu_0 - \mu_i)^2 + \lambda_1 \sum_{i=1}^{n}|\mu_i| + \lambda_2 \sum_{i=1}^{n-1}|\mu_{i+1} - \mu_i|,$$

where $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ are the tuning parameters.

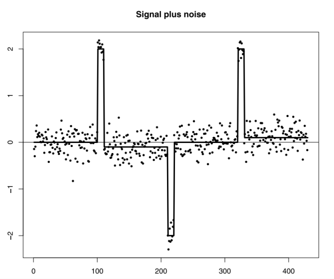

Figure 1: l_1 fused lasso estimator [RINALDO, 2009]

(Continue with the example)

Generalized lasso [Tibshirani and Taylor, 2011] solves the problem:

$$\min \frac{1}{2n}\|Y - X\beta\|^2 + \lambda\|D\beta\|_1,$$

where $\beta \in \mathbb{R}^p$ and $D \in \mathbb{R}^{r\times p}$. Let $X = (1, I) \in \mathbb{R}^{n\times(n+1)}$ and $D = [D_1^\top D_2^\top]^\top$, where

$$D_1 = \begin{pmatrix} 0 & 0_n^\top \\ 0_n & I \end{pmatrix} \in \mathbb{R}^{(n+1)\times(n+1)} \text{ and } D_2 = \begin{pmatrix} 0 & 1 & -1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & -1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & -1 \end{pmatrix} \in \mathbb{R}^{(n-1)\times(n+1)},$$

then the fused lasso estimator is computed by the generalized lasso algorithm.

Example 6 (Linear regression with a strong heredity)

- (Model) $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$, where $x_1, x_2 \in \{0, 1\}$. β_3 is nonzero only when $\beta_1 \neq 0$ and $\beta_2 \neq 0$. (model restriction: the interaction effect is significant only when both main effects are significant)
- (Optimization problem)

$$\begin{aligned} \min \quad & \frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \beta_3 x_{i1} x_{i2})^2 \\ \text{subject to} \quad & |\beta_1| + |\beta_2| + |\beta_3| \leq C \\ & |\beta_3| \leq |\beta_1| \text{ and } |\beta_3| \leq |\beta_2|. \end{aligned}$$

where $C \geq 0$ is a tuning parameter.

Example 7 (Non-crossing composite quantile regression)

- (Model) Denote the cdf of $y|x$ by $F(\cdot|x)$. For $0 < \tau_1 < \cdots < \tau_K < 1$,

$$F^{-1}(\tau_k|x) = \beta_{k0} + \beta_{k1}x_1 + \cdots + \beta_{kp}x_p, \text{ for } k = 1, \cdots, K.$$

  The $F^{-1}(\tau_k|x)$ is the conditional τ_k-qunatile function. We simply denote the quantile regression function $\tilde{x}^\top \beta_k$, where $\tilde{x} = (1, x^\top)^\top \in \mathbb{R}^{p+1}$.

- (Optimization problem)

$$\min \frac{1}{nK} \sum_{k=1}^{K} \sum_{i=1}^{n} \rho_{\tau_k}(y_i - \tilde{x}_i^\top \boldsymbol{\beta}_k),$$

  where $\rho_\tau(z) = \tau \max(z, 0) + (1 - \tau) \max(-z, 0)$.

(Continue with the example)

Crossing problem: Let $\hat{\boldsymbol{\beta}}_k$ be the τ_k-quantile regression coefficients. For $\tau_k < \tau_{k+1}$

$$\tilde{x}^\top \hat{\boldsymbol{\beta}}_k > \tilde{x}^\top \hat{\boldsymbol{\beta}}_{k+1}$$

for some $\tilde{x}$ in the domain of predictors. [Bondell et al., 2010] proposed a reduced version of inequality constraints to prevent the crossing problem.

Let $\boldsymbol{\delta}_1 = \boldsymbol{\beta}_1$ and $\boldsymbol{\delta}_j = \boldsymbol{\beta}_j - \boldsymbol{\beta}_{j-1}$ for $j = 2, \cdots, K$. Since $\boldsymbol{\beta}_k = \sum_{j=1}^{k} \boldsymbol{\delta}_j$

$$
\begin{aligned}
F^{-1}(\tau_k|x) &= \sum_{j=1}^{k} \delta_{j0} + (\sum_{j=1}^{k} \delta_{j1})x_1 + \cdots + (\sum_{j=1}^{k} \delta_{jp})x_p \\
&= \tilde{x}^\top \left(\sum_{j=1}^{k} \boldsymbol{\delta}_j \right)
\end{aligned}
$$

Theorem 8 (non-crossing constraints [Bondell et al., 2010])

Assume that $\tilde{x} \in [0,1]^{p+1}$. *If* $\delta_{k0} - \sum_{j=1}^{p} \max(-\delta_{kj}, 0) \geq 0$ *for* $k = 2, \cdots, K$, *then*

$$\tilde{x}^\top \left(\sum_{j=1}^{k} \boldsymbol{\delta}_j \right) \leq \tilde{x}^\top \left(\sum_{j=1}^{k+1} \boldsymbol{\delta}_j \right) \text{ for all } \tilde{x} \in [0,1]^{p+1} \text{ and } k = 1, \cdots K-1.$$

(proof) See [Bondell et al., 2010] or [Moon et al., 2021]

(Optimization problem for estimating non-crossing quantile regression)

$$\begin{aligned} \min \quad & \frac{1}{nK}\sum_{k=1}^{K}\sum_{i=1}^{n}\rho_{\tau_k}(y_i - \tilde{x}_i^\top \boldsymbol{\delta}_k) \\ \text{subject to} \quad & \delta_{k0} - \sum_{j=1}^{p}\max(-\delta_{kj}, 0) \geq 0 \text{ for } k = 2, \cdots, K \end{aligned}$$

Because the feature vectors in the neural network satisfy the bounded condition of $\tilde{x}$ by using the sigmoid activation function, the non-crossing composite quantile linear regression model easily is extended to the neural network model [Moon et al., 2021].

Example 9 (Monotone regression)

- (Model) $f : \mathbb{R} \mapsto \mathbb{R}$, nondecreasing function.
  Let k_j for $j = 1, \cdots, m$ be knot point and $B_j(z) = \max(z - k_j, 0)$. (The knot points are pre-determined)

$$f(x) = \gamma_0 + \sum_{j=1}^{m} \gamma_j B_j(x), \text{ where } \sum_{j=1}^{k} \gamma_j \geq 0 \text{ for } k = 1, \cdots, m.$$

- (Optimization problem)

$$\begin{aligned} \min \quad & \frac{1}{2n} \sum_{i=1}^{n} (y_i - \gamma_0 - \sum_{j=1}^{m} \gamma_j B_j(x_i))^2 \\ \text{subject to} \quad & \sum_{j=1}^{k} \gamma_j \geq 0 \text{ for } k = 1, \cdots, m. \end{aligned}$$

Appendix

The lasso regression estimator is given by minimizing

$$l_\lambda(\beta) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - x_i^\top \beta)^2 + \lambda\|\beta\|_1. \tag{1}$$

(Here, the intercept is not considered in the model.)

$l_\lambda(\beta)$ is a convex function of β.

Appendix

Coordinatewise algorithm Let $l(\beta_1, \cdots, \beta_p)$ be (strictly) convex function on $\mathbb{R}^p$. If the convex function is differentiable , the following coordinate algorithm gives a minimizer.

(1) Let $k = 0$ and set an initial estimator $(\beta_1^{(k)}, \cdots, \beta_p^{(k)})$

(2) For $j = 1, \cdots, p$

- minimize $l(\beta_1^{(k+1)}, \cdots, \beta_{j-1}^{(k+1)}, \beta_j, \beta_{j+1}^{(k)}, \cdots, \beta_p^{(k)})$ with respect to β_j and let the minimizer be $\beta_j^{(k+1)}$

(3) $k \to k + 1$ and repeat (2) until the solutions converges.

When the nondifferentiable function is separable, the coordinate algorithm gives the minimizer for (1) [Tseng, 2001]. This algorithm is known as "shooting algorithm" [Fu, 1998] and is elaborated by [Friedman et al., 2010].

Appendix

First, consider the minimizer of the following function.

$$\min_{x \in \mathbb{R}} ax^2 + bx + \lambda|x|$$

for $a > 0$ and $\lambda \geq 0$. Let $f_\lambda(x) = ax^2 + bx + \lambda|x|$. Compute the minimizer of $f_\lambda(x)$.

$$\text{argmin}_x f_\lambda(x) = \begin{cases} -\frac{b}{2a} + \text{sign}(b)\frac{\lambda}{2a}, & \text{if } |b| > \lambda \\ 0, & \text{if } |b| \leq \lambda \end{cases}$$

Appendix

First consider the case of $b < 0$ and $-b/2a - \lambda/2a \geq 0$

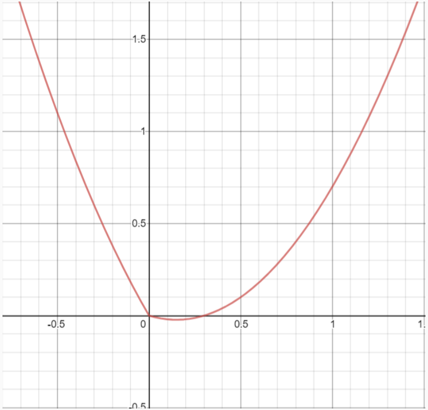

Figure 2: Illustration of $ax^2 + bx + \lambda|x|$

Appendix

Consider a one-dimensional objective function

$$l_\lambda(\beta_1^{(k+1)}, \cdots, \beta_{j-1}^{(k+1)}, \beta_j, \beta_{j+1}^{(k)}, \cdots, \beta_p^{(k)}).$$

Let $\tilde{r}_i^{-j} = y_i - \mathbf{x}_i'(\beta_1^{(k+1)}, \cdots, \beta_{j-1}^{(k+1)}, 0, \beta_{j+1}^{(k)}, \cdots, \beta_p^{(k)})$, then the above objective function is simply written by

$$\begin{aligned}
& l_\lambda(\beta_1^{(k+1)}, \cdots, \beta_{j-1}^{(k+1)}, \beta_j, \beta_{j+1}^{(k)}, \cdots, \beta_p^{(k)}) \\
= \; & \frac{1}{2n}\sum_{i=1}^{n}(\tilde{r}_i^{-j} - x_{ij}\beta_j)^2 + \lambda|\beta_j| + const \\
= \; & \underbrace{\frac{1}{2n}(\sum_{i=1}^{n} x_{ij}^2)}_{a}\,\beta_j^2 + \underbrace{(-\frac{1}{n}\sum_{i=1}^{n}\tilde{r}_i^{-j}x_{ij})}_{b}\,\beta_j + \lambda|\beta_j| + const'
\end{aligned}$$

Appendix

Then, we can apply the minimization algorithm of

$$\min_{x \in \mathbb{R}} ax^2 + bx + \lambda|x|$$

to the lasso problem sequentially.

Bondell, H. D., Reich, B. J., and Wang, H. (2010).
Noncrossing quantile regression curve estimation.
Biometrika, 97(4):825–838.

Friedman, J., Hastie, T., and Tibshirani, R. (2010).
Regularization paths for generalized linear models via coordinate descent.
Journal of statistical software, 33(1):1.

Fu, W. J. (1998).
Penalized regressions: the bridge versus the lasso.
Journal of computational and graphical statistics, 7(3):397–416.

Moon, S. J., Jeon, J.-J., Lee, J. S. H., and Kim, Y. (2021).
Learning multiple quantiles with neural networks.
Journal of Computational and Graphical Statistics, 30(4):1238–1248.

RINALDO, A. (2009).
Properties and refinements of the fused lasso.
Annals of statistics, 37(5):2922–2952.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005).
Sparsity and smoothness via the fused lasso.
Journal of the Royal Statistical Society Series B: Statistical Methodology, 67(1):91–108.

Tibshirani, R. J. and Taylor, J. (2011).
The solution path of the generalized lasso.
The Annals of Statistics, 39(3):1335.

Tseng, P. (2001).
Convergence of a block coordinate descent method for nondifferentiable minimization.
Journal of optimization theory and applications, 109(3):475–494.