# 1 Dual Formulation of Elastic Net

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

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

The augmented Lagrangian for the problem is:

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

By taking the derivatives with respect to $z$ and $\beta$, we could obtain:

$$0 = \tfrac{\partial L}{\partial z} = z - y + u$$
$$0 = \tfrac{\partial L}{\partial \beta} = -X^\tau u - D^\tau v$$

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

$$\min_{\omega_i} \lambda_1|\omega_i| + \lambda_2\omega_i^2 - v^\tau X_i\omega_i$$
$$= \begin{cases} 0 & if \quad |v^\tau X_i| \leq \lambda_1 \\ -\dfrac{(\lambda_1 - |v^\tau X_i|)^2}{4\lambda_2} & if \quad |v^\tau X_i| > \lambda_1 \end{cases}$$

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

$$d^* = \min_u \tfrac{1}{2}||y - u||_2^2 + \sum_{j:|X_j^\tau u| > \lambda_1} \tfrac{(\lambda_1 - |u^\tau X_i|)^2}{4\lambda_2}$$
$$subject \quad to \quad \mathbf{1}^\tau u = 0$$

First denote $f(u) = \sum_{j:|X_j^\tau u| > \lambda_1} \frac{(\lambda_1 - |u^\tau X_i|)^2}{4\lambda_2}$, clearly $f(u)$ is of quadratic form, $f(u) = \frac{1}{2}u^\tau Au + a^\tau u + b$, where $b$ is a constant and does not matter in the optimization of the dual problem, $A$ and $a$ are:

$$A = \tfrac{1}{2\lambda_2} X_E X_E^\tau$$
$$E := \{i : |X_i^\tau u| > \lambda\}$$

$$a = \tfrac{\lambda_1}{2\lambda_2}\Big(\sum_{i:X_i^\tau < -\lambda_1} X_i - \sum_{i:X_i^\tau > \lambda_1} X_i\Big)$$

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

$$\hat{u} = \mathbf{prox}_{\tilde{f}}(y)$$
$$\tilde{f} = \mathbf{I}(\mathbf{1}^\tau u = 0) f(u) + \mathbf{I}(\mathbf{1}^\tau u \neq 0)\infty$$

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

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

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

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

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

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

where $\nabla(\hat{\lambda})^\tau$ denotes the gradient of $\lambda$ as a function of $y$.

By taking $u = (I+A)^{-1}(y-a-\lambda\mathbf{1})$ back to the Lagrangian, the dual problem will become a second-order equation of $\lambda$:

$$d^* = \max_\lambda \frac{1}{2}||y - (I+A)^{-1}(y-a-\lambda\mathbf{1})||_2^2 + \frac{1}{2}(y-a-\lambda\mathbf{1})^\tau (I+A)^{-1} A (I+A)^{-1}(y-a-\lambda\mathbf{1}) + a^\tau (I+A)^{-1}(y-a-\lambda\mathbf{1}) + \lambda * \mathbf{1}^\tau (I+A)^{-1}(y-a-\lambda\mathbf{1})$$

More specifically, the second-order term is:

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

and the first-order term is:

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

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

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

and the gradient

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

By taking the gradient back to $J = (I+A)^{-1} - (I+A)^{-1}\mathbf{1}\nabla(\hat{\lambda})^\tau$, we could obtain the Jacobian.