COORDINATE DESCENT FOR SDA

$$\min_x h(x) := f(x) + \lambda\Omega(x)$$

f is smooth, Ω is a regularization function that may be nonsmooth.

Alg. 2 update step:

$$z_{i_k}^k \leftarrow \operatorname*{argmin}_\chi (\chi - x_{i_k}^k)^T[\nabla f(x^k)]_{i_k} + \frac{1}{2\alpha_k}\|\chi - x_{i_k}^k\|_2^2 + \lambda\Omega_i(x) \quad \text{for some } \alpha_k > 0$$
$$x^{k+1} \leftarrow x^k + (z_{i_k}^k - x_{i_k}^k)e_{i_k}$$

For our problem:
$f(\beta) = \|Y\theta^{t+1} - X\beta\|^2 + \gamma\|\beta\|^2$ with $\Omega = I$ where $\|\| = \|\|_2$
$\Omega(\beta) = \|\beta\|_1 \Rightarrow \Omega_i(\beta_i) = |\beta_i|$

Assume we already have β^t. We need to find β^{t+1}. Then the update step for our problem is

$$z_{i_t}^t \leftarrow \operatorname*{argmin}_\chi \chi[\nabla f(\beta^t)]_{i_t} - \beta_{i_t}^t[\nabla f(\beta^t)]_{i_t} + \frac{1}{2\alpha_t}\|\chi - \beta_{i_t}^t\|_2^2 + \lambda|\chi|$$
$$\beta^{t+1} \leftarrow \beta^t + (z_{i_t}^t - \beta_{i_t}^t)e_{i_t}$$

That is, we want to minimize

$$\chi[\nabla f(\beta^t)]_{i_t} - \beta_{i_t}^t[\nabla f(\beta^t)]_{i_t} + \frac{1}{2\alpha_t}|\chi - \beta_{i_t}^t||\chi - \beta_{i_t}^t| + \lambda|\chi|$$
$$\Rightarrow \chi[\nabla f(\beta^t)]_{i_t} - \beta_{i_t}^t[\nabla f(\beta^t)]_{i_t} + \frac{1}{2\alpha_t}(\chi^2 - 2\chi\beta_{i_t}^t + (\beta_{i_t}^t)^2) + \lambda|\chi|.$$

Note that $\chi[\nabla f(\beta^t)]_{i_t} - \beta_{i_t}^t[\nabla f(\beta^t)]_{i_t} + \frac{1}{2\alpha_t}(\chi^2 - 2\chi\beta_{i_t}^t + (\beta_{i_t}^t)^2)$ is a quadratic with respect to χ. So the minimum is found where the gradient with respect to χ is 0. But since we have the $\lambda|\chi|$ term, we must also take the subgradient of the absolute value which will give us an interval that includes 0.

$$\Rightarrow 0 \in [\nabla f(\beta^t)]_{i_t} + \frac{1}{\alpha_t}\chi - \frac{1}{\alpha_t}\beta_{i_t}^t + \partial\lambda|\chi|$$

Say χ^* is the optimal solution. We have three cases.

$$\begin{aligned}
\chi^* > 0 &\Rightarrow [\nabla f(\beta^t)]_{i_t} + \frac{1}{\alpha_t}\chi^* - \frac{1}{\alpha_t}\beta^t_{i_t} + \lambda = 0 \\
&\Rightarrow \chi^* = \beta^t_{i_t} - \alpha_t[\nabla f(\beta^t)]_{i_t} - \alpha_t\lambda \\
\chi^* < 0 &\Rightarrow [\nabla f(\beta^t)]_{i_t} + \frac{1}{\alpha_t}\chi^* - \frac{1}{\alpha_t}\beta^t_{i_t} - \lambda = 0 \\
&\Rightarrow \chi^* = \beta^t_{i_t} - \alpha_t[\nabla f(\beta^t)]_{i_t} + \alpha_t\lambda \\
\chi^* = 0 &\Rightarrow 0 \in [\nabla f(\beta^t)]_{i_t} - \frac{1}{\alpha_t}(0) - \frac{1}{\alpha_t}\beta^t_{i_t} + \lambda[-1,1] \\
&\Rightarrow -\lambda + [\nabla f(\beta^t)]_{i_t} - \frac{1}{\alpha_t}\beta^t_{i_t} \le 0 \le \lambda + [\nabla f(\beta^t)]_{i_t} - \frac{1}{\alpha_t}\beta^t_{i_t} \\
&\Rightarrow -[\nabla f(\beta^t)]_{i_t} + \frac{1}{\alpha_t}\beta^t_{i_t} \in [-\lambda, \lambda] \\
&\Rightarrow |\beta^t_{i_t} - \alpha_t[\nabla f(\beta^t)]_{i_t}| \le \alpha_t\lambda
\end{aligned}$$

Say $D = \beta^t_{i_t} - \alpha_t[\nabla f(\beta^t)]_{i_t}$. Then the optimal solution for our update step is

$$z^t_{i_t} = \begin{cases} 0 & |D| \le \alpha_t\lambda \\ D - \alpha_t\lambda & D > 0 \text{ and } |D| > \alpha_t\lambda \\ D + \alpha_t\lambda & D < 0 \text{ and } |D| > \alpha_t\lambda \end{cases}$$

which is the soft-thresholding operator $S(D, \alpha_t\lambda)$.

Now to calculate β^{t+1} we need $z^t_{i_t} - \beta^t_{i_t}$.

$$z^t_{i_t} - \beta^t_{i_t} = \begin{cases} 0 - \beta^t_{i_t} & |D| \le \alpha_t\lambda \\ D - \alpha_t\lambda - \beta^t_{i_t} & D > 0 \text{ and } |D| > \alpha_t\lambda \\ D + \alpha_t\lambda - \beta^t_{i_t} & D < 0 \text{ and } |D| > \alpha_t\lambda \end{cases}$$

$$D - \alpha_t\lambda - \beta^t_{i_t} = \beta^t_{i_t} - \alpha_t[\nabla f(\beta^t)]_{i_t} - \alpha_t\lambda - \beta^t_{i_t} = -\alpha_t[\nabla f(\beta^t)]_{i_t} - \alpha_t\lambda$$

$$D + \alpha_t\lambda - \beta^t_{i_t} = \beta^t_{i_t} - \alpha_t[\nabla f(\beta^t)]_{i_t} + \alpha_t\lambda - \beta^t_{i_t} = -\alpha_t[\nabla f(\beta^t)]_{i_t} + \alpha_t\lambda$$

$$\Rightarrow z^t_{i_t} - \beta^t_{i_t} = \begin{cases} -\beta^t_{i_t} & |D| \le \alpha_t\lambda \\ -\alpha_t[\nabla f(\beta^t)]_{i_t} - \alpha_t\lambda & D > 0 \text{ and } |D| > \alpha_t\lambda \\ -\alpha_t[\nabla f(\beta^t)]_{i_t} + \alpha_t\lambda & D < 0 \text{ and } |D| > \alpha_t\lambda \end{cases}$$

To calculate $[\nabla f(\beta^t)]_{i_t}$ we see that

$$\begin{aligned}
f(\beta^t) &= (Y\theta^{t+1} - X\beta^t)^T(Y\theta^{t+1} - X\beta^t) + \gamma(\beta^t)^T(\beta^t) \\
&= (\theta^{t+1})^T Y^T Y\theta^{t+1} - (\theta^{t+1})^T Y^T X\beta^t - (\beta^t)^T X^T Y\theta^{t+1} + (\beta^t)^T X^T X\beta^t + \gamma(\beta^t)^T(\beta^t) \\
&= (\theta^{t+1})^T Y^T Y\theta^{t+1} - 2(\beta^t)^T X^T Y\theta^{t+1} + (\beta^t)^T X^T X\beta^t + \gamma(\beta^t)^T(\beta^t)
\end{aligned}$$

$$[\nabla f(\beta^t)]_{i_t} = [-2X^T Y\theta^{t+1} + 2X^T X\beta^t + 2\gamma\beta^t]_{i_t}$$

To find the i_t component of $[\nabla f(\beta^t)]_{i_t}$ we have

$[2\gamma\beta^t]_{i_t} = 2\gamma\beta^t_{i_t}$

$[2X^TX\beta^t]_{i_t} = 2(X_{i_t})^TX\beta^t$ where X_{i_t} denotes the i_t column of X

$[-2X^TY\theta^{t+1}]_{i_t} = -2(X_{i_t})^TY\theta^{t+1}$

$\Rightarrow [\nabla f(\beta^t)]_{i_t} = -2(X_{i_t})^TY\theta^{t+1} + 2(X_{i_t})^TX\beta^t + 2\gamma\beta^t_{i_t}$

Plugging this into $-\alpha_t[\nabla f(\beta^t)]_{i_t} \pm \alpha_t\lambda$ and D we get

$$-\alpha_t[\nabla f(\beta^t)]_{i_t} \pm \alpha_t\lambda = 2\alpha_t(X_{i_t})^TY\theta^{t+1} - 2\alpha_t(X_{i_t})^TX\beta^t - 2\alpha_t\gamma\beta^t_{i_t} \pm \alpha_t\lambda$$
$$D = \beta^t_{i_t} - \alpha_t[\nabla f(\beta^t)]_{i_t} = \beta^t_{i_t} + 2\alpha_t(X_{i_t})^TY\theta^{t+1} - 2\alpha_t(X_{i_t})^TX\beta^t - 2\alpha_t\gamma\beta^t_{i_t}$$

Recall that after we calculate $z^t_{i_t}$ we calculate β^{t+1} by $\beta^{t+1} \leftarrow \beta^t + (z^t_{i_t} - \beta^t_{i_t})e_{i_t}$. So only the i_t component of β^{t+1} is update. Therefore, the coordinate update for β^{t+1} is

$$\begin{aligned}
\beta^{t+1}_{i_t} &= \beta^t_{i_t} + z^t_{i_t} - \beta^t_{i_t} \\
&= \begin{cases} \beta^t_{i_t} - \beta^t_{i_t} & |D| \le \alpha_t\lambda \\ \beta^t_{i_t} + 2\alpha_t(X_{i_t})^TY\theta^{t+1} - 2\alpha_t(X_{i_t})^TX\beta^t - 2\alpha_t\gamma\beta^t_{i_t} - \alpha_t\lambda & D > 0 \text{ and } |D| > \alpha_t\lambda \\ \beta^t_{i_t} + 2\alpha_t(X_{i_t})^TY\theta^{t+1} - 2\alpha_t(X_{i_t})^TX\beta^t - 2\alpha_t\gamma\beta^t_{i_t} + \alpha_t\lambda & D < 0 \text{ and } |D| > \alpha_t\lambda \end{cases} \\
&= \begin{cases} 0 & |D| \le \alpha_t\lambda \\ (1-2\alpha_t\gamma)\beta^t_{i_t} + 2\alpha_t(X_{i_t})^TY\theta^{t+1} - 2\alpha_t(X_{i_t})^TX\beta^t - \alpha_t\lambda & D > 0 \text{ and } |D| > \alpha_t\lambda \\ (1-2\alpha_t\gamma)\beta^t_{i_t} + 2\alpha_t(X_{i_t})^TY\theta^{t+1} - 2\alpha_t(X_{i_t})^TX\beta^t + \alpha_t\lambda & D < 0 \text{ and } |D| > \alpha_t\lambda \end{cases} \\
&= S(D, \alpha_t\lambda)
\end{aligned}$$

where $D = (1-2\alpha_t\gamma)\beta^t_{i_t} + 2\alpha_t(X_{i_t})^TY\theta^{t+1} - 2\alpha_t(X_{i_t})^TX\beta^t$