1 Appendix

1.1 Gradient Descent

Algorithm 1 (WEIRDLY NUMBERED) Updated Algorithm 1

Initialize $\boldsymbol{\lambda}^{(0)}$.

for each iteration $k = 0, 1, ...$ until stopping criteria is reached **do**

Solve for $\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}^{(k)}) = \arg\min_{\boldsymbol{\theta}\in\Theta} L_T(\boldsymbol{\theta}, \boldsymbol{\lambda}^{(k)})$.

Calculate the derivative of the model parameters with respect to the regularization parameters

$$\frac{\partial}{\partial \boldsymbol{\lambda}}\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}) = -\left[\left[\nabla_\theta^2\left(L\left(\boldsymbol{y}_T, f_\theta(\boldsymbol{X}_T)\right) + \sum_{i=1}^{J}\lambda_i P_i(\boldsymbol{\theta})\right)\right]^{-1}\nabla_{\boldsymbol{\theta}} P(\boldsymbol{\theta})\right]\Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}^{(k)})} \tag{1}$$

Calculate the gradient

$$\nabla_{\boldsymbol{\lambda}} L\left(\boldsymbol{y}_{\boldsymbol{V}}, f_{\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})}(\boldsymbol{X}_{\boldsymbol{V}})\right)\Big|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^{(k)}} = \left[\frac{\partial}{\partial \boldsymbol{\theta}} L(\boldsymbol{y}_V, f_{\boldsymbol{\theta}}(\boldsymbol{X}_{\boldsymbol{V}}))\Big|_{\theta=\hat{\theta}(\boldsymbol{\lambda}^{(k)})}\right]^\top \frac{\partial}{\partial \boldsymbol{\lambda}}\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})\Big|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^{(k)}} \tag{2}$$

Perform gradient step with step size $t^{(k)}$

$$\boldsymbol{\lambda}^{(k+1)} := \boldsymbol{\lambda}^{(k)} - t^{(k)}\,\nabla_{\boldsymbol{\lambda}} L\left(\boldsymbol{y}_V, f_{\hat{\theta}(\boldsymbol{\lambda})}(\boldsymbol{X}_V)\right)\Big|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^{(k)}} \tag{3}$$

end for

1.1.1 K-fold Cross Validation

We can perform joint optimization for K-fold cross validation by reformulating the problem. Let $(\boldsymbol{y}, \boldsymbol{X})$ be the full data set. We denote the kth fold as $(\boldsymbol{y}_k, \boldsymbol{X}_k)$ and its complement as $(\boldsymbol{y}_{-k}, \boldsymbol{X}_{-k})$. Then the objective of this joint optimization problem is the average validation cost across all K folds:

$$\begin{gathered}\arg\min_{\boldsymbol{\lambda}\in\Lambda} \tfrac{1}{K}\textstyle\sum_{k=1}^{K} L(\boldsymbol{y}_k, f_{\hat{\boldsymbol{\theta}}^{(k)}(\boldsymbol{\lambda})}(\boldsymbol{X}_k)) \\ \text{s.t. } \hat{\boldsymbol{\theta}}^{(k)}(\boldsymbol{\lambda}) = \arg\min_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{y}_{-k}, f_{\boldsymbol{\theta}}(\boldsymbol{X}_{-k})) + \sum_{i=1}^{J}\lambda_i P_i(\boldsymbol{\theta}) \text{ for } k = 1, ..., K\end{gathered} \tag{4}$$

1.2 Proof of Theorem 1

Proof. We will show that for a given $\boldsymbol{\lambda}_0$ that satisfies the given conditions, the validation loss is continuously differentiable within some neighborhood of $\boldsymbol{\lambda}_0$. It then follows that if the theorem conditions hold true for almost every $\boldsymbol{\lambda}$, then the validation loss is continuously differentiable with respect to $\boldsymbol{\lambda}$ at almost every $\boldsymbol{\lambda}$.

Suppose the theorem conditions are satisfied at $\boldsymbol{\lambda}_0$. Let $\boldsymbol{B}'$ be an orthonormal set of basis vectors that span the differentiable space $\Omega^{L_T}(\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}_0), \boldsymbol{\lambda}_0)$ with the subset of vectors $\boldsymbol{B}$ that span the model parameter space.

Let $\tilde{L}_T(\boldsymbol{\theta}, \boldsymbol{\lambda})$ be the gradient of $L_T(\cdot, \boldsymbol{\lambda})$ at $\boldsymbol{\theta}$ with respect to the basis $\boldsymbol{B}$:

$$\tilde{L}_T(\boldsymbol{\theta}, \boldsymbol{\lambda}) =_{\boldsymbol{B}} \nabla L_T(\cdot, \boldsymbol{\lambda})|_{\boldsymbol{\theta}} \tag{5}$$

Since $\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}_0)$ is the minimizer of the training loss, the gradient of $L_T(\cdot, \boldsymbol{\lambda}_0)$ with respect to the basis $\boldsymbol{B}$ must be zero at $\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}_0)$:

$${}_{\boldsymbol{B}}\nabla L_T(\cdot, \boldsymbol{\lambda}_0)|_{\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}_0)} = \tilde{L}_T(\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}_0), \boldsymbol{\lambda}_0) = 0 \tag{6}$$

From our assumptions, we know that there exists a neighborhood W containing $\boldsymbol{\lambda}_0$ such that $\tilde{L}_T$ is continuously differentiable along directions in the differentiable space $\Omega^{L_T}(\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}_0), \boldsymbol{\lambda}_0)$. Also, the Jacobian matrix $D\tilde{L}_T(\cdot, \boldsymbol{\lambda}_0)|_{\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}_0)}$ with respect to basis $\boldsymbol{B}$ is nonsingular. Therefore, by the implicit function theorem, there exist open sets $U \subseteq W$ containing $\boldsymbol{\lambda}_0$ and V containing $\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}_0)$ and a continuously differentiable function $\gamma : U \to V$ such that for every $\boldsymbol{\lambda} \in U$, we have that

$$\tilde{L}_T(\gamma(\boldsymbol{\lambda}), \boldsymbol{\lambda}) = \nabla_B L_T(\cdot, \boldsymbol{\lambda})|_{\gamma(\boldsymbol{\lambda})} = 0 \tag{7}$$

That is, we know that $\gamma(\boldsymbol{\lambda})$ is a continuously differentiable function that minimizes $L_T(\cdot, \boldsymbol{\lambda})$ in the differentiable space $\Omega^{L_T}(\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}_0), \boldsymbol{\lambda}_0)$. Since we assumed that the differentiable space is a local optimality space of $L_T(\cdot, \boldsymbol{\lambda})$ in the neighborhood W, then for every $\boldsymbol{\lambda} \in U$,

$$\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}) = \arg\min_{\boldsymbol{\theta}} L_T(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \arg\min_{\boldsymbol{\theta} \in \Omega^{L_T}(\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}_0), \boldsymbol{\lambda}_0)} L_T(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \gamma(\boldsymbol{\lambda}) \tag{8}$$

Therefore, we have shown that if $\boldsymbol{\lambda}_0$ satisfies the assumptions given in the theorem, the fitted model parameters $\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})$ is a continuously differentiable function within a neighborhood of $\boldsymbol{\lambda}_0$. We can then apply the chain rule to get the gradient of the validation loss. □

1.3 Examples - Checking Theorem 1 Conditions

1.3.1 Elastic Net

Next, we show that the joint optimization problem satisfies all three conditions in Theorem 1:

Condition 1: The nonzero indices of the elastic net estimates stay locally constant for almost every $\boldsymbol{\lambda}$. Therefore, $S_{\boldsymbol{\lambda}}$ is a local optimality space for $L_T(\cdot, \boldsymbol{\lambda})$ ✓

Condition 2: The ℓ_1 penalty is smooth when restricted to $S_{\boldsymbol{\lambda}}$. ✓

Condition 3: The Hessian matrix of $L_T(\cdot, \boldsymbol{\lambda})$ with respect to the columns of $\boldsymbol{I}_{I(\boldsymbol{\lambda})}$ is $\boldsymbol{I}_{I(\boldsymbol{\lambda})}^\top \boldsymbol{X}_T^\top \boldsymbol{X}_T \boldsymbol{I}_{I(\boldsymbol{\lambda})} + \lambda_2 \boldsymbol{I}$. This is positive definite if $\lambda_2 > 0$. ✓

1.3.2 Sparse Group Lasso

Since the reasoning for the first two conditions is exactly the same as in the Elastic Net (Section 1.3.1), we just give the calculations for the third condition.

Condition 3: The Hessian matrix of $L_T(\cdot, \boldsymbol{\lambda})$ with respect to the columns of $\boldsymbol{I}_{I(\boldsymbol{\lambda})}$ is

$$\frac{1}{n}\boldsymbol{I}_{I(\boldsymbol{\lambda})}^{\top}\boldsymbol{X}_T^{\top}\boldsymbol{X}_T\boldsymbol{I}_{I(\boldsymbol{\lambda})} + \lambda_1\boldsymbol{B}(\boldsymbol{\lambda}) + \epsilon\boldsymbol{I}_p \tag{9}$$

where $\boldsymbol{B}(\boldsymbol{\lambda})$ is a block diagonal matrix with components

$$\left\|\tilde{\boldsymbol{\theta}}^{(m)}\boldsymbol{\lambda})\right\|_2^{-1}\left(\boldsymbol{I} - \frac{\tilde{\boldsymbol{\theta}}^{(m)}(\boldsymbol{\lambda})\tilde{\boldsymbol{\theta}}^{(m)}(\boldsymbol{\lambda})^{\top}}{\|\tilde{\boldsymbol{\theta}}^{(m)}(\boldsymbol{\lambda})\|_2^2}\right) \tag{10}$$

for $m = 1, ..., M$ from top left to bottom right. The Hessian is positive definite for any fixed $\epsilon > 0$. ✓

1.3.3 Generalized Lasso (condition check)

The first two conditions in Theorem 1 are satisfied by similar reasoning to that discussed in Section 1.3.1. For the third condition, we need to check that the Hessian matrix is invertible.

Condition 3: The Hessian matrix of $L_T(\cdot, \boldsymbol{\lambda})$ with respect to $\boldsymbol{U}_\lambda$ is

$$\boldsymbol{U}_\lambda^{\top}\boldsymbol{X}_T^{\top}\boldsymbol{X}_T\boldsymbol{U}_\lambda + \epsilon\boldsymbol{U}_\lambda \tag{11}$$

This is positive definite for any fixed $\epsilon > 0$. ✓

1.3.4 Generalized Lasso (the whole thing)

The generalized lasso (**?**) penalizes the ℓ_1 norm of the coefficients $\boldsymbol{\theta}$ weighted by some matrix $\boldsymbol{D}$. Depending on the choice of $\boldsymbol{D}$, the generalized lasso induces different structural constraints on the regression coefficients. Special cases include the fused lasso, trend filtering, and wavelet smoothing (**?**), (**?**), (**?**).

To tune the regularization parameter λ, we formulate the generalized lasso as a joint optimization problem:

$$\begin{gathered}\min_{\lambda\in\mathbb{R}_+} \tfrac{1}{2}\|\boldsymbol{y}_V - \boldsymbol{X}_V\hat{\boldsymbol{\theta}}(\lambda)\|^2 \\ \text{s.t. } \hat{\boldsymbol{\theta}}(\lambda) = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \tfrac{1}{2}\|\boldsymbol{y}_T - \boldsymbol{X}_T\boldsymbol{\theta}\|^2 + \lambda\|\boldsymbol{D}\boldsymbol{\theta}\|_1 + \tfrac{1}{2}\epsilon\|\boldsymbol{\theta}\|_2^2\end{gathered} \tag{12}$$

Let $I(\lambda)$ denote the indices of the zero elements of $\boldsymbol{D}\hat{\boldsymbol{\theta}}(\lambda)$:

$$I(\lambda) = \left\{i|(\boldsymbol{D}\hat{\boldsymbol{\theta}}(\lambda))_i = 0 \text{ for } i = 1, ..., p\right\} \tag{13}$$

Since $\|\boldsymbol{D}\boldsymbol{\theta}\|_1$ is differentiable in θ only along directions where the current zero elements of $\boldsymbol{D}\boldsymbol{\theta}$ remain zero, the differentiable space S_λ is the null space of $\boldsymbol{I}_{I(\lambda)}^{\top}\boldsymbol{D}$, denoted $\mathcal{N}(\boldsymbol{I}_{I(\lambda)}^{\top}\boldsymbol{D})$. Let $\boldsymbol{U}_\lambda$ be an orthonormal basis for $\mathcal{N}(\boldsymbol{I}_{I(\lambda)}^{\top}\boldsymbol{D})$.

Now we show the gradient calculations. Following Algorithm 2, we first define the locally equivalent joint optimization problem:

$$\begin{aligned} &\min_{\lambda\in\mathbb{R}_+} \tfrac{1}{2}\|\boldsymbol{y}_V - \boldsymbol{X}_V\boldsymbol{U}_\lambda\hat{\boldsymbol{\beta}}(\lambda)\|^2 \\ \text{s.t. } &\hat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta}} \tfrac{1}{2}\|\boldsymbol{y}_T - \boldsymbol{X}_T\boldsymbol{U}_\lambda\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{D}\boldsymbol{U}_\lambda\boldsymbol{\beta}\|_1 + \tfrac{1}{2}\epsilon\|\boldsymbol{U}_\lambda\boldsymbol{\beta}\|_2^2 \end{aligned} \tag{14}$$

Implicit differentiation with respect to λ of the gradient condition for $\hat{\boldsymbol{\beta}}(\lambda)$ gives us

$$\frac{\partial}{\partial\lambda}\hat{\boldsymbol{\beta}}(\lambda) = -(\boldsymbol{U}_\lambda^\top\boldsymbol{X}_T^\top\boldsymbol{X}_T\boldsymbol{U}_\lambda + \epsilon\boldsymbol{U}_\lambda)^{-1}\boldsymbol{U}_\lambda^\top\boldsymbol{D}^\top sgn(\boldsymbol{D}\boldsymbol{U}_\lambda\hat{\boldsymbol{\beta}}(\lambda)) \tag{15}$$

The chain rule then gives the gradient of the validation loss with respect to λ:

$$\nabla_\lambda L(\boldsymbol{y}_{\boldsymbol{V}}, f_{\hat{\boldsymbol{\theta}}(\lambda)}(\boldsymbol{X}_{\boldsymbol{V}})) = -\left(\boldsymbol{X}_V\boldsymbol{U}_\lambda\frac{\partial}{\partial\lambda}\hat{\boldsymbol{\beta}}(\lambda)\right)^\top\left(\boldsymbol{y}_V - \boldsymbol{X}_V\boldsymbol{U}_\lambda\hat{\boldsymbol{\beta}}(\lambda)\right) \tag{16}$$

1.3.5 Ugly Math from Additive Models

To formalize our optimization problem we give a bit of notation. Let $\boldsymbol{Z}_T \in \mathbb{R}^{|T|\times q}$, $\boldsymbol{Z}_V \in \mathbb{R}^{|V|\times q}$ be the linear covariates from the training and validation sets. Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ be the design matrix for the nonlinear covariates where the first $|T|$ observations are from the training set and the last $|V|$ observations are from the validation set. Let $\boldsymbol{I}_T$ and $\boldsymbol{I}_V$ be matrices such that $X_T = I_T X$ and $X_V = I_V X$. We combine the validation and training nonlinear covariates into one design matrix since we use the training data to fit estimates for f_i at the validation points.

The joint optimization problem is

$$\begin{aligned} &\min_{\boldsymbol{\lambda}\in\mathbb{R}_+^{p+1}} \tfrac{1}{2}\left\|\boldsymbol{y}_V - \boldsymbol{Z}_V\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) - \boldsymbol{I}_V\textstyle\sum_{i=1}^p\hat{\boldsymbol{\theta}}^{(i)}(\boldsymbol{\lambda})\right\|_2^2 \\ \text{s.t. } &\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}), \hat{\boldsymbol{\theta}}^{(i)}(\boldsymbol{\lambda}) = \arg\min_{\boldsymbol{\beta},\boldsymbol{\theta}} \tfrac{1}{2}\left\|\boldsymbol{y}_T - \boldsymbol{Z}_T\boldsymbol{\beta} - \boldsymbol{I}_T\textstyle\sum_{i=1}^p\boldsymbol{\theta}^{(i)}\right\|_2^2 + \lambda_0\|\boldsymbol{\beta}\|_1 \\ &\quad + \tfrac{1}{2}\textstyle\sum_{i=1}^p\lambda_i\left\|\boldsymbol{D}_{\boldsymbol{x}_i}^{(2)}\boldsymbol{\theta}^{(i)}\right\|_2^2 + \tfrac{1}{2}\epsilon\left(\|\boldsymbol{\beta}\|_2^2 + \textstyle\sum_{i=1}^p\|\boldsymbol{\theta}^{(i)}\|_2^2\right) \end{aligned} \tag{17}$$

The matrix $\boldsymbol{D}_{\boldsymbol{x}_i}^{(2)}$ gives the second-order differences between the nonparametric estimates of f_i for unevenly-spaced inputs. Construction of $\boldsymbol{D}_{\boldsymbol{x}_i}^{(2)}$ is given in the Appendix. Note that we have presented an extremely general form on the training criterion, where there are separate λ_i for $\left\|\boldsymbol{D}_{\boldsymbol{x}_i}^{(2)}\boldsymbol{\theta}^{(i)}\right\|_2^2$. In practice, one would pool λ_i since tuning more than 2 penalty parameters is difficult.

In this example, the lasso is the only penalty which is not everywhere differentiable. Let the nonzero indices of $\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$ be denoted $I(\boldsymbol{\lambda}) = \{i|\hat{\beta}_i(\boldsymbol{\lambda}) \neq 0 \text{ for } i = 1, ..., p\}$. The differentiable space is then $S_{\boldsymbol{\lambda}} = \boldsymbol{C}(\boldsymbol{I}_{I(\boldsymbol{\lambda})}) \oplus \mathbb{R}^{n\times p}$.

We now calculate the gradient of the validation loss. Given $I(\boldsymbol{\lambda})$, the nonzero set of $\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$, the locally equivalent joint optimization problem as

$$\begin{aligned} &\min_{\boldsymbol{\lambda}\in\mathbb{R}_+^2} \tfrac{1}{2}\left\|\boldsymbol{y}_V - \boldsymbol{X}_{V,I(\boldsymbol{\lambda})}\hat{\boldsymbol{\eta}}(\boldsymbol{\lambda}) - \boldsymbol{I}_V\textstyle\sum_{i=1}^p\hat{\boldsymbol{\theta}}^{(i)}(\boldsymbol{\lambda})\right\|_2^2 \\ \text{s.t. } &\hat{\boldsymbol{\eta}}(\boldsymbol{\lambda}), \hat{\boldsymbol{\theta}}^{(i)}(\boldsymbol{\lambda}) = \arg\min_{\boldsymbol{\eta},\boldsymbol{\theta}} \tfrac{1}{2}\left\|\boldsymbol{y}_T - \boldsymbol{X}_{T,I(\boldsymbol{\lambda})}\boldsymbol{\eta} - \boldsymbol{I}_T\textstyle\sum_{i=1}^p\boldsymbol{\theta}^{(i)}\right\|_2^2 \\ &\quad + \lambda_0\|\boldsymbol{\eta}\|_1 + \tfrac{1}{2}\textstyle\sum_{i=1}^p\lambda_i\|\boldsymbol{D}(\boldsymbol{z})\boldsymbol{\theta}^{(i)}\|_2^2 + \tfrac{1}{2}\epsilon\left(\|\boldsymbol{\eta}\|_2^2 + \textstyle\sum_{i=1}^p\|\boldsymbol{\theta}^{(i)}\|_2^2\right) \end{aligned} \tag{18}$$

We follow the same steps as before to calculate the gradient. The gradient is then

$$\nabla_{\lambda_j} L_V(\boldsymbol{\lambda}) = -\left(\boldsymbol{X}_{V,I(\boldsymbol{\lambda})} \frac{\partial}{\partial \lambda_j} \hat{\boldsymbol{\eta}}(\boldsymbol{\lambda}) + \boldsymbol{I}_V \sum_{i=1}^{p} \frac{\partial}{\partial \lambda_j} \hat{\boldsymbol{\theta}}^{(i)}(\boldsymbol{\lambda})\right)^\top \left(\boldsymbol{y}_V - \boldsymbol{X}_{V,I(\boldsymbol{\lambda})} \hat{\boldsymbol{\eta}}(\boldsymbol{\lambda}) - \boldsymbol{I}_V \sum_{i=1}^{p} \hat{\boldsymbol{\theta}}^{(i)}(\boldsymbol{\lambda})\right)$$

where

$$\begin{bmatrix} \frac{\partial}{\partial \boldsymbol{\lambda}} \hat{\boldsymbol{\eta}}(\boldsymbol{\lambda}) \\ \frac{\partial}{\partial \boldsymbol{\lambda}} \hat{\boldsymbol{\theta}}^{(1)}(\boldsymbol{\lambda}) \\ \dots \\ \frac{\partial}{\partial \boldsymbol{\lambda}} \hat{\boldsymbol{\theta}}^{(p)}(\boldsymbol{\lambda}) \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial \boldsymbol{\lambda}_0} \hat{\boldsymbol{\eta}}(\boldsymbol{\lambda}) & \dots & \frac{\partial}{\partial \boldsymbol{\lambda}_p} \hat{\boldsymbol{\eta}}(\boldsymbol{\lambda}) \\ \frac{\partial}{\partial \boldsymbol{\lambda}_0} \hat{\boldsymbol{\theta}}^{(1)}(\boldsymbol{\lambda}) & \dots & \frac{\partial}{\partial \boldsymbol{\lambda}_p} \hat{\boldsymbol{\theta}}^{(1)}(\boldsymbol{\lambda}) \\ \dots & \dots & \dots \\ \frac{\partial}{\partial \boldsymbol{\lambda}_0} \hat{\boldsymbol{\theta}}^{(p)}(\boldsymbol{\lambda}) & \dots & \frac{\partial}{\partial \boldsymbol{\lambda}_p} \hat{\boldsymbol{\theta}}^{(p)}(\boldsymbol{\lambda}) \end{bmatrix} = H^{-1} \begin{bmatrix} sgn\left(\hat{\boldsymbol{\eta}}(\boldsymbol{\lambda})\right) & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & D_{x_1}^T D_{x_1} \hat{\boldsymbol{\theta}}^{(1)}(\boldsymbol{\lambda}) & \dots & \mathbf{0} \\ \dots & \dots & \dots & \dots \\ \mathbf{0} & \mathbf{0} & \dots & D_{x_p}^T D_{x_p} \hat{\boldsymbol{\theta}}^{(p)}(\boldsymbol{\lambda}) \end{bmatrix} \tag{19}$$

The hideous matrix H is given in the appendix.

The second-order difference matrix construction: Let the n observations be ordered according to the ith covariate such that $x_{ik_1} \leq x_{ik_2} \leq \dots \leq x_{ik_n}$. $\boldsymbol{D}_{\boldsymbol{x}_i}^{(1)} \in \mathbb{R}^{n \times n}$ is the corresponding first-order difference matrix; so row $j = 1, \dots, n-1$ of $\boldsymbol{D}_{\boldsymbol{x}_i}^{(1)}$ has a -1 in position k_j, 1 in position k_{j+1}, and 0 elsewhere and row n is all zeros. Then $\boldsymbol{D}_{\boldsymbol{x}_i}^{(2)}$ is

$$\boldsymbol{D}_{\boldsymbol{x}_i}^{(1)} \cdot \text{diag}\left(\frac{1}{x_{ik_2} - x_{ik_1}}, \frac{1}{x_{ik_3} - x_{ik_2}}, \dots, \frac{1}{x_{ik_n} - x_{ik_{n-1}}}, 0\right) \cdot \boldsymbol{D}_{\boldsymbol{x}_i}^{(1)} \tag{20}$$

1.3.6 Additive Partially Linear Models

By the same reasoning in Section 1.3.1, the first two conditions of Theorem 1 are satisfied. We now check for the third condition.

Condition 3: The Hessian matrix of $L_T(\cdot, \boldsymbol{\lambda})$ with respect to the basis

$$\begin{bmatrix} \boldsymbol{I}_{I(\boldsymbol{\lambda})} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_n \end{bmatrix} \tag{21}$$

is

$$H = \begin{bmatrix} \boldsymbol{I}_{I(\boldsymbol{\lambda})}^\top \boldsymbol{X}_T^\top \boldsymbol{X}_T \boldsymbol{I}_{I(\boldsymbol{\lambda})} + \epsilon \boldsymbol{I} & \boldsymbol{I}_{I(\boldsymbol{\lambda})}^\top \boldsymbol{X}_T^\top \boldsymbol{I}_T \\ \boldsymbol{I}_T^\top \boldsymbol{X}_T \boldsymbol{I}_{I(\boldsymbol{\lambda})} & \boldsymbol{I}_T^\top \boldsymbol{I}_T + \lambda_2 \boldsymbol{D}(\boldsymbol{z})^\top \boldsymbol{D}(\boldsymbol{z}) + \epsilon \boldsymbol{I} \end{bmatrix} \tag{22}$$

The Hessian matrix is invertible for any $\lambda_2 > 0$ and any fixed $\epsilon > 0$.

1.3.7 Additive Models with Heterogeneous Smoothness

The problem is composed solely of quadratic functions, so the validation loss $L(\boldsymbol{y}_v, \boldsymbol{I}_V \sum_{i=1}^{p} \hat{\boldsymbol{\theta}}^{(i)}(\boldsymbol{\lambda}))$ differentiable everywhere with respect to $\boldsymbol{\lambda}$. Choosing $\epsilon > 0$ will satisfy the positive definite condition.

1.4 Gradient Derivations

1.4.1 Un-pooled Sparse Group Lasso

The joint optimization formulation of the un-pooled sparse group lasso is

$$\begin{array}{c} \min_{\boldsymbol{\lambda}\in\mathbb{R}^2_+} \frac{1}{2n}\left\|\boldsymbol{y}_V - \boldsymbol{X}_V\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})\right\|_2^2 \\ \text{s.t. } \hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}) = \arg\min_{\boldsymbol{\theta}} \frac{1}{2n}\|\boldsymbol{y}_T - \boldsymbol{X}_T\boldsymbol{\theta}\|_2^2 + \sum_{m=1}^M \lambda_1^{(m)}\|\boldsymbol{\theta}^{(m)}\|_2 + \lambda_2\|\boldsymbol{\theta}\|_1 + \frac{1}{2}\epsilon\|\boldsymbol{\theta}\|_2^2 \end{array} \tag{23}$$

Let $I(\boldsymbol{\lambda}) = \{i|\hat{\theta}_i(\boldsymbol{\lambda}) \neq 0 \text{ for } i = 1, ..., p\}$. With similar reasoning in Section 2.4.3, the differentiable space for this problem is $span(\boldsymbol{I}_{I(\boldsymbol{\lambda})})$. All three conditions of Theorem 1 are satisfied. We note that the Hessian in this problem is

$$\frac{1}{n}\boldsymbol{X}_{T,I(\boldsymbol{\lambda})}^\top\boldsymbol{X}_{T,I(\boldsymbol{\lambda})} + \boldsymbol{B}(\boldsymbol{\lambda}) + \epsilon\boldsymbol{I} \tag{24}$$

where $\boldsymbol{B}(\boldsymbol{\lambda})$ is the block diagonal matrix with components $m = 1, 2, ..., M$

$$\frac{\lambda_1^{(m)}}{||\boldsymbol{\theta}^{(m)}||_2}\left(\boldsymbol{I} - \frac{1}{||\boldsymbol{\theta}^{(m)}||_2^2}\boldsymbol{\theta}^{(m)}\boldsymbol{\theta}^{(m)\top}\right) \tag{25}$$

from top left to bottom right. This is positive definite for any $\epsilon > 0$.

To find the gradient, the locally equivalent joint optimization with a smooth training criterion is

$$\begin{array}{c} \min_{\boldsymbol{\lambda}\in\mathbb{R}^2_+} \frac{1}{2n}\left\|\boldsymbol{y}_V - \boldsymbol{X}_{V,I(\boldsymbol{\lambda})}\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})\right\|_2^2 \\ \text{s.t. } \hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) = \arg\min_{\boldsymbol{\beta}} \frac{1}{2n}\left\|\boldsymbol{y}_T - \boldsymbol{X}_{T,I(\boldsymbol{\lambda})}\boldsymbol{\beta}\right\|_2^2 + \sum_{m=1}^M \lambda_1^{(m)}\|\boldsymbol{\beta}^{(m)}\|_2 + \lambda_2\|\boldsymbol{\beta}\|_1 + \frac{1}{2}\epsilon\|\boldsymbol{\beta}\|_2^2 \end{array} \tag{26}$$

Implicit differentiation of the gradient condition with respect to the regularization parameters gives us

$$\begin{array}{rcl} \frac{\partial}{\partial\boldsymbol{\lambda}}\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) & = & \left[\frac{\partial}{\partial\boldsymbol{\lambda}_1^{(1)}}\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) \quad \cdots \quad \frac{\partial}{\partial\boldsymbol{\lambda}_1^{(M)}}\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) \quad \frac{\partial}{\partial\boldsymbol{\lambda}_2}\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})\right] \\ & = & -\left(\frac{1}{n}\boldsymbol{X}_{T,I(\boldsymbol{\lambda})}^\top\boldsymbol{X}_{T,I(\boldsymbol{\lambda})} + \boldsymbol{B}(\boldsymbol{\lambda}) + \epsilon\boldsymbol{I}\right)^{-1}\left[\boldsymbol{C}(\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})) \quad sgn(\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}))\right] \end{array} \tag{27}$$

where $\boldsymbol{C}(\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}))$ has columns $m = 1, 2..., M$

$$\begin{bmatrix} 0 \\ \vdots \\ 0 \\ \frac{\hat{\boldsymbol{\beta}}^{(m)}(\boldsymbol{\lambda})}{||\hat{\boldsymbol{\beta}}^{(m)}(\boldsymbol{\lambda})||_2} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{28}$$

By the chain rule, we get that the gradient of the validation error is

$$\nabla_{\boldsymbol{\lambda}} L(\boldsymbol{y}_V, \boldsymbol{X}_V\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})) = \frac{1}{n}\left(\boldsymbol{X}_{V,I(\boldsymbol{\lambda})}\frac{\partial}{\partial\boldsymbol{\lambda}}\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})\right)^\top(\boldsymbol{y}_V - \boldsymbol{X}_{V,I(\boldsymbol{\lambda})}\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})) \tag{29}$$

1.4.2 Additive Partially Linear Model with three penalties

The joint optimization formulation of the additive partially linear model with the elastic net penalty for the linear model $\boldsymbol{\beta}$ and the H-P filter for the nonparametric estimates $\boldsymbol{\theta}$ is

$$\begin{aligned} &\min_{\boldsymbol{\lambda}\in\mathbb{R}^2_+} \tfrac{1}{2}\left\|\boldsymbol{y}_V - \boldsymbol{X}_V\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) - (\boldsymbol{I}-\boldsymbol{I}_T)\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})\right\|_2^2 \\ \text{s.t. } &\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}),\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}) = \arg\min_{\boldsymbol{\beta},\boldsymbol{\theta}} \tfrac{1}{2}\left\|\boldsymbol{y}_T - \boldsymbol{X}_T\boldsymbol{\beta} - \boldsymbol{I}_T\boldsymbol{\theta}\right\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1 + \tfrac{1}{2}\lambda_2\|\boldsymbol{\beta}\|_2^2 + \tfrac{1}{2}\lambda_3\|\boldsymbol{D}(\boldsymbol{z})\boldsymbol{\theta}\|_2^2 + \tfrac{1}{2}\epsilon\|\boldsymbol{\theta}\|_2^2 \end{aligned} \tag{30}$$

The differentiable space is exactly the same as that given in Section **??**. Also, all three conditions of Theorem 1 are satisfied. Note that the Hessian of the training criterion with respect to the basis in (18) is

$$H = \begin{bmatrix} \boldsymbol{I}_{I(\boldsymbol{\lambda})}^\top \boldsymbol{X}_T^\top \boldsymbol{X}_T \boldsymbol{I}_{I(\boldsymbol{\lambda})} + \lambda_2 \boldsymbol{I} & \boldsymbol{I}_{I(\boldsymbol{\lambda})}^\top \boldsymbol{X}_T^\top \boldsymbol{I}_T \\ \boldsymbol{I}_T^\top \boldsymbol{X}_T \boldsymbol{I}_{I(\boldsymbol{\lambda})} & \boldsymbol{I}_T^\top \boldsymbol{I}_T + \lambda_3 \boldsymbol{D}(\boldsymbol{z})^\top \boldsymbol{D}(\boldsymbol{z}) + \epsilon \boldsymbol{I} \end{bmatrix} \tag{31}$$

To find the gradient, we first consider the locally equivalent joint optimization problem with a smooth training criterion:

$$\begin{aligned} &\min_{\boldsymbol{\lambda}\in\mathbb{R}^2_+} \tfrac{1}{2}\left\|\boldsymbol{y}_V - \boldsymbol{X}_{V,I(\boldsymbol{\lambda})}\hat{\boldsymbol{\eta}}(\boldsymbol{\lambda}) - (\boldsymbol{I}-\boldsymbol{I}_T)\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})\right\|_2^2 \\ \text{s.t. } &\hat{\boldsymbol{\eta}}(\boldsymbol{\lambda}),\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}) = \arg\min_{\boldsymbol{\eta},\boldsymbol{\theta}} \tfrac{1}{2}\left\|\boldsymbol{y}_T - \boldsymbol{X}_{T,I(\boldsymbol{\lambda})}\boldsymbol{\eta} - \boldsymbol{I}_T\boldsymbol{\theta}\right\|_2^2 + \lambda_1\|\boldsymbol{\eta}\|_1 + \tfrac{1}{2}\lambda_2\|\boldsymbol{\eta}\|_2^2 + \tfrac{1}{2}\lambda_3\|\boldsymbol{D}(\boldsymbol{z})\boldsymbol{\theta}\|_2^2 + \tfrac{1}{2}\epsilon\|\boldsymbol{\theta}\|_2^2 \end{aligned} \tag{32}$$

After implicit differentiation of the gradient condition with respect to the regularization parameters, we get that

$$\begin{bmatrix} \frac{\partial}{\partial\boldsymbol{\lambda}}\hat{\boldsymbol{\eta}}(\boldsymbol{\lambda}) \\ \frac{\partial}{\partial\boldsymbol{\lambda}}\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}) \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial\lambda_1}\hat{\boldsymbol{\eta}}(\boldsymbol{\lambda}) & \frac{\partial}{\partial\lambda_3}\hat{\boldsymbol{\eta}}(\boldsymbol{\lambda}) & \frac{\partial}{\partial\lambda_3}\hat{\boldsymbol{\eta}}(\boldsymbol{\lambda}) \\ \frac{\partial}{\partial\lambda_1}\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}) & \frac{\partial}{\partial\lambda_2}\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}) & \frac{\partial}{\partial\lambda_3}\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}) \end{bmatrix} = -H^{-1}\begin{bmatrix} sgn(\hat{\boldsymbol{\eta}}(\boldsymbol{\lambda})) & \hat{\boldsymbol{\eta}}(\boldsymbol{\lambda}) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{D}(\boldsymbol{z})^\top\boldsymbol{D}(\boldsymbol{z})\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}) \end{bmatrix} \tag{33}$$

We then apply the chain rule to get the gradient direction of the validation loss with respect to $\boldsymbol{\lambda}$

$$\nabla_{\boldsymbol{\lambda}} L_V(\boldsymbol{\lambda}) = -\left(\boldsymbol{X}_{V,I(\boldsymbol{\lambda})}\frac{\partial}{\partial\boldsymbol{\lambda}}\hat{\boldsymbol{\eta}}(\boldsymbol{\lambda}) + (\boldsymbol{I}-\boldsymbol{I}_T)\frac{\partial}{\partial\boldsymbol{\lambda}}\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})\right)^\top\left(\boldsymbol{y}_V - \boldsymbol{X}_{V,I(\boldsymbol{\lambda})}\hat{\boldsymbol{\eta}}(\boldsymbol{\lambda}) - (\boldsymbol{I}-\boldsymbol{I}_T)\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})\right) \tag{34}$$

1.5 Backtracking Line Search

Let the criterion function be $L:\mathbb{R}^n\to\mathbb{R}$. Suppose that the descent algorithm is currently at point x with descent direction Δx. Backtracking line search uses a heuristic for finding a step size $t\in(0,1]$ such that the value of the criterion is minimized. The method depends on constants $\alpha\in(0,0.5)$ and $\beta\in(0,1)$.

Algorithm 2 Backtracking Line Search

Initialize $t = 1$.
while $L(\boldsymbol{x} + t\boldsymbol{\Delta x}) > L(\boldsymbol{x}) + \alpha t \nabla L(\boldsymbol{x})^T \boldsymbol{\Delta x}$ **do**
  Update $t := \beta t$
end while

1.6 Joint Optimization with Accelerated Gradient Descent and Adaptive Restarts

Algorithm 3 Joint Optimization with Accelerated Gradient Descent and Adaptive Restarts

Initialize $\boldsymbol{\lambda}^{(0)}$.
while stopping criteria is not reached **do**
  for each iteration $k = 0, 1, ...$ **do**
    Solve for $\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}^{(k)}) = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} L_T(\boldsymbol{\theta}, \boldsymbol{\lambda}^{(k)})$.
    Construct matrix $\boldsymbol{U}^{(k)}$, an orthonormal basis of $\Omega^{L_T(\cdot,\boldsymbol{\lambda})}\left(\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}^{(k)})\right)$.
    Define the locally equivalent joint optimization problem

$$\begin{aligned} &\min_{\boldsymbol{\lambda}\in\Lambda} L(\boldsymbol{y}_V, f_{\boldsymbol{U}^{(k)}\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})}(\boldsymbol{X}_V)) \\ \text{s.t. } &\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) = \arg\min_{\boldsymbol{\beta}} L(\boldsymbol{y}_T, f_{\boldsymbol{U}^{(k)}\boldsymbol{\beta}}(\boldsymbol{X}_T)) + \sum_{i=1}^{J} \lambda_i P_i(\boldsymbol{U}^{(k)}\boldsymbol{\beta}) \end{aligned} \tag{35}$$

    Calculate $\frac{\partial}{\partial\boldsymbol{\lambda}}\hat{\beta}(\boldsymbol{\lambda})|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^{(k)}}$ where

$$\frac{\partial}{\partial\boldsymbol{\lambda}}\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) = -\left[{}_{\boldsymbol{U}^{(k)}}\nabla^2\left(L(\boldsymbol{y}_T, f_{\boldsymbol{U}^{(k)}\boldsymbol{\beta}}(\boldsymbol{X}_T)) + \sum_{i=1}^{J}\lambda_i P_i(\boldsymbol{U}^{(k)}\boldsymbol{\beta})\right)\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})}\right]^{-1}\left[{}_{\boldsymbol{U}^{(k)}}\nabla P(\boldsymbol{U}^{(k)}\boldsymbol{\beta})|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})}\right] \tag{36}$$

    with ${}_{\boldsymbol{U}^{(k)}}\nabla^2$ and ${}_{\boldsymbol{U}^{(k)}}\nabla$ are as defined in (12).
    Calculate the gradient $\nabla_{\boldsymbol{\lambda}} L(\boldsymbol{y}_{\boldsymbol{V}}, f_{\hat{\theta}(\boldsymbol{\lambda})}(\boldsymbol{X}_{\boldsymbol{V}}))|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^{(k)}}$ where

$$\nabla_{\boldsymbol{\lambda}} L(\boldsymbol{y}_{\boldsymbol{V}}, f_{\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})}(\boldsymbol{X}_{\boldsymbol{V}})) = \left[\boldsymbol{U}^{(k)}\frac{\partial}{\partial\boldsymbol{\lambda}}\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})\right]^{\top}\left[{}_{\boldsymbol{U}^{(k)}}\nabla L(\boldsymbol{y}_{\boldsymbol{V}}, f_{\boldsymbol{U}^{(k)}\boldsymbol{\beta}}(\boldsymbol{X}_{\boldsymbol{V}}))\big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})}\right] \tag{37}$$

    Perform Neterov's update with step size $t^{(k)}$:

$$\begin{aligned} \boldsymbol{\eta} &:= \boldsymbol{\lambda}^{(k)} + \tfrac{k-1}{k+2}\left(\boldsymbol{\lambda}^{(k)} - \boldsymbol{\lambda}^{(k-1)}\right) \\ \boldsymbol{\lambda}^{(k+1)} &:= \boldsymbol{\eta} - t^{(k)}\,\nabla_{\boldsymbol{\lambda}} L\left(\boldsymbol{y}_{\boldsymbol{V}}, f_{\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})}(\boldsymbol{X}_{\boldsymbol{V}})\right)\Big|_{\boldsymbol{\lambda}=\boldsymbol{\eta}} \end{aligned} \tag{38}$$

    if the stopping criteria is reached or

$$L\left(\boldsymbol{y}_V, f_{\hat{\theta}(\boldsymbol{\lambda}^{(k+1)})}(\boldsymbol{X}_V)\right) > L\left(\boldsymbol{y}_V, f_{\hat{\theta}(\boldsymbol{\lambda}^{(k)})}(\boldsymbol{X}_V)\right), \tag{39}$$

    then
      set $\boldsymbol{\lambda}^{(0)} := \boldsymbol{\lambda}^{(k)}$ and break
    end if
  end for
end while
return $\boldsymbol{\lambda}^{(0)}$ and $\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}^{(0)})$