1 Appendix

1.1 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 $\boldsymbol{B}$:

$$\tilde{L}_T(\boldsymbol{\theta}, \boldsymbol{\lambda}) =_{\boldsymbol{B}} \nabla L_T(\cdot, \boldsymbol{\lambda})|_{\boldsymbol{\theta}} \tag{1}$$

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 $\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}_0)$:

$$_{\boldsymbol{B}}\nabla L_T(\cdot, \boldsymbol{\lambda}_0)|_{\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}_0)} = \tilde{L}_T(\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}_0), \boldsymbol{\lambda}_0) = 0 \tag{2}$$

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

$$\tilde{L}_T(\gamma(\boldsymbol{\lambda}), \boldsymbol{\lambda}) = \nabla_B L_T(\cdot, \boldsymbol{\lambda})|_{\gamma(\boldsymbol{\lambda})} = 0 \tag{3}$$

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 $\boldsymbol{\lambda} \in U$,

$$\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}) = \arg\min_{\boldsymbol{\theta}} L_T(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \arg\min_{\boldsymbol{\theta} \in \Omega^{L_T}(\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}_0), \boldsymbol{\lambda}_0)} L_T(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \gamma(\boldsymbol{\lambda}) \tag{4}$$

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. □

1.2 Gradient Derivations

1.2.1 Un-pooled Sparse Group Lasso

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

$$\begin{aligned} &\min_{\boldsymbol{\lambda} \in \mathbb{R}_+^2} \frac{1}{2n} \left\| \boldsymbol{y}_V - \boldsymbol{X}_V \hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}) \right\|_2^2 \\ \text{s.t. } &\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}) = \arg\min_{\boldsymbol{\theta}} \frac{1}{2n} \|\boldsymbol{y}_T - \boldsymbol{X}_T \boldsymbol{\theta}\|_2^2 + \sum_{m=1}^M \lambda_1^{(m)} \|\boldsymbol{\theta}^{(m)}\|_2 + \lambda_2 \|\boldsymbol{\theta}\|_1 + \frac{1}{2}\epsilon \|\boldsymbol{\theta}\|_2^2 \end{aligned} \tag{5}$$

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

$$\frac{1}{n}\boldsymbol{X}_{T,I(\boldsymbol{\lambda})}^{\top}\boldsymbol{X}_{T,I(\boldsymbol{\lambda})} + \boldsymbol{B}(\boldsymbol{\lambda}) + \epsilon \boldsymbol{I} \tag{6}$$

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

$$\frac{\lambda_1^{(m)}}{||\boldsymbol{\theta}^{(m)}||_2}\left(\boldsymbol{I} - \frac{1}{||\boldsymbol{\theta}^{(m)}||_2^2}\boldsymbol{\theta}^{(m)}\boldsymbol{\theta}^{(m)\top}\right) \tag{7}$$

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

$$\begin{array}{c} \min_{\boldsymbol{\lambda}\in\mathbb{R}_+^2} \frac{1}{2n}\left\|\boldsymbol{y}_V - \boldsymbol{X}_{V,I(\boldsymbol{\lambda})}\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})\right\|_2^2 \\ \text{s.t. } \hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) = \arg\min_{\boldsymbol{\beta}} \frac{1}{2n}\left\|\boldsymbol{y}_T - \boldsymbol{X}_{T,I(\boldsymbol{\lambda})}\boldsymbol{\beta}\right\|_2^2 + \sum_{m=1}^M \lambda_1^{(m)}\|\boldsymbol{\beta}^{(m)}\|_2 + \lambda_2\|\boldsymbol{\beta}\|_1 + \frac{1}{2}\epsilon\|\boldsymbol{\beta}\|_2^2 \end{array} \tag{8}$$

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

$$\begin{aligned} \frac{\partial}{\partial\boldsymbol{\lambda}}\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) &= \left[\frac{\partial}{\partial\boldsymbol{\lambda}_1^{(1)}}\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) \quad \cdots \quad \frac{\partial}{\partial\boldsymbol{\lambda}_1^{(M)}}\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) \quad \frac{\partial}{\partial\boldsymbol{\lambda}_2}\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})\right] \\ &= -\left(\frac{1}{n}\boldsymbol{X}_{T,I(\boldsymbol{\lambda})}^{\top}\boldsymbol{X}_{T,I(\boldsymbol{\lambda})} + \boldsymbol{B}(\boldsymbol{\lambda}) + \epsilon\boldsymbol{I}\right)^{-1}\left[\boldsymbol{C}(\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})) \quad sgn(\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}))\right] \end{aligned} \tag{9}$$

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

$$\begin{bmatrix} 0 \\ \vdots \\ 0 \\ \frac{\hat{\boldsymbol{\beta}}^{(m)}(\boldsymbol{\lambda})}{||\hat{\boldsymbol{\beta}}^{(m)}(\boldsymbol{\lambda})||_2} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{10}$$

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

$$\nabla_{\boldsymbol{\lambda}} L(\boldsymbol{y}_V, \boldsymbol{X}_V\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})) = \frac{1}{n}\left(\boldsymbol{X}_{V,I(\boldsymbol{\lambda})}\frac{\partial}{\partial\boldsymbol{\lambda}}\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})\right)^{\top}(\boldsymbol{y}_V - \boldsymbol{X}_{V,I(\boldsymbol{\lambda})}\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})) \tag{11}$$

1.2.2 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

$$\min_{\boldsymbol{\lambda}\in\mathbb{R}^2_+} \frac{1}{2}\left\|\boldsymbol{y}_V - \boldsymbol{X}_V\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) - (\boldsymbol{I}-\boldsymbol{I}_T)\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})\right\|_2^2$$
$$\text{s.t. } \hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}), \hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}) = \arg\min_{\boldsymbol{\beta},\boldsymbol{\theta}} \frac{1}{2}\|\boldsymbol{y}_T - \boldsymbol{X}_T\boldsymbol{\beta} - \boldsymbol{I}_T\boldsymbol{\theta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1 + \frac{1}{2}\lambda_2\|\boldsymbol{\beta}\|_2^2 + \frac{1}{2}\lambda_3\|\boldsymbol{D}(\boldsymbol{z})\boldsymbol{\theta}\|_2^2 + \frac{1}{2}\epsilon\|\boldsymbol{\theta}\|_2^2 \tag{12}$$

The differentiable space is exactly the same as that given in Section 2.4.5. Also, all three conditions of Theorem 1 are satisfied. Note that the Hessian of the training criterion with respect to the basis in (54) is

$$H = \begin{bmatrix} \boldsymbol{I}_{I(\boldsymbol{\lambda})}^\top \boldsymbol{X}_T^\top \boldsymbol{X}_T \boldsymbol{I}_{I(\boldsymbol{\lambda})} + \lambda_2 \boldsymbol{I} & \boldsymbol{I}_{I(\boldsymbol{\lambda})}^\top \boldsymbol{X}_T^\top \boldsymbol{I}_T \\ \boldsymbol{I}_T^\top \boldsymbol{X}_T \boldsymbol{I}_{I(\boldsymbol{\lambda})} & \boldsymbol{I}_T^\top \boldsymbol{I}_T + \lambda_3 \boldsymbol{D}(\boldsymbol{z})^\top \boldsymbol{D}(\boldsymbol{z}) + \epsilon \boldsymbol{I} \end{bmatrix} \tag{13}$$

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

$$\min_{\boldsymbol{\lambda}\in\mathbb{R}^2_+} \frac{1}{2}\left\|\boldsymbol{y}_V - \boldsymbol{X}_{V,I(\boldsymbol{\lambda})}\hat{\boldsymbol{\eta}}(\boldsymbol{\lambda}) - (\boldsymbol{I}-\boldsymbol{I}_T)\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})\right\|_2^2$$
$$\text{s.t. } \hat{\boldsymbol{\eta}}(\boldsymbol{\lambda}), \hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}) = \arg\min_{\boldsymbol{\eta},\boldsymbol{\theta}} \frac{1}{2}\left\|\boldsymbol{y}_T - \boldsymbol{X}_{T,I(\boldsymbol{\lambda})}\boldsymbol{\eta} - \boldsymbol{I}_T\boldsymbol{\theta}\right\|_2^2 + \lambda_1\|\boldsymbol{\eta}\|_1 + \frac{1}{2}\lambda_2\|\boldsymbol{\eta}\|_2^2 + \frac{1}{2}\lambda_3\|\boldsymbol{D}(\boldsymbol{z})\boldsymbol{\theta}\|_2^2 + \frac{1}{2}\epsilon\|\boldsymbol{\theta}\|_2^2 \tag{14}$$

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

$$\begin{bmatrix} \frac{\partial}{\partial\boldsymbol{\lambda}}\hat{\boldsymbol{\eta}}(\boldsymbol{\lambda}) \\ \frac{\partial}{\partial\boldsymbol{\lambda}}\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}) \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial\lambda_1}\hat{\boldsymbol{\eta}}(\boldsymbol{\lambda}) & \frac{\partial}{\partial\lambda_3}\hat{\boldsymbol{\eta}}(\boldsymbol{\lambda}) & \frac{\partial}{\partial\lambda_3}\hat{\boldsymbol{\eta}}(\boldsymbol{\lambda}) \\ \frac{\partial}{\partial\lambda_1}\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}) & \frac{\partial}{\partial\lambda_2}\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}) & \frac{\partial}{\partial\lambda_3}\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}) \end{bmatrix} = -H^{-1}\begin{bmatrix} sgn(\hat{\boldsymbol{\eta}}(\boldsymbol{\lambda})) & \hat{\boldsymbol{\eta}}(\boldsymbol{\lambda}) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{D}(\boldsymbol{z})^\top\boldsymbol{D}(\boldsymbol{z})\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}) \end{bmatrix} \tag{15}$$

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

$$\nabla_{\boldsymbol{\lambda}} L_V(\boldsymbol{\lambda}) = -\left(\boldsymbol{X}_{V,I(\boldsymbol{\lambda})}\frac{\partial}{\partial\boldsymbol{\lambda}}\hat{\boldsymbol{\eta}}(\boldsymbol{\lambda}) + (\boldsymbol{I}-\boldsymbol{I}_T)\frac{\partial}{\partial\boldsymbol{\lambda}}\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})\right)^\top \left(\boldsymbol{y}_V - \boldsymbol{X}_{V,I(\boldsymbol{\lambda})}\hat{\boldsymbol{\eta}}(\boldsymbol{\lambda}) - (\boldsymbol{I}-\boldsymbol{I}_T)\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})\right) \tag{16}$$

1.3 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 1 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

1.4 Joint Optimization with Accelerated Gradient Descent and Adaptive Restarts

---

Algorithm 2 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

$$
\begin{aligned}
&\min_{\boldsymbol{\lambda}\in\Lambda} L(\boldsymbol{y}_V, f_{\boldsymbol{U}^{(k)}\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})}(\boldsymbol{X}_V)) \\
\text{s.t. } &\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) = \arg\min_{\boldsymbol{\beta}} L(\boldsymbol{y}_T, f_{\boldsymbol{U}^{(k)}\boldsymbol{\beta}}(\boldsymbol{X}_T)) + \sum_{i=1}^{J} \lambda_i P_i(\boldsymbol{U}^{(k)}\boldsymbol{\beta})
\end{aligned}
\tag{17}
$$

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

$$
\frac{\partial}{\partial\boldsymbol{\lambda}}\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) = -\left[{}_{\boldsymbol{U}^{(k)}}\nabla^2\left(L(\boldsymbol{y}_T, f_{\boldsymbol{U}^{(k)}\boldsymbol{\beta}}(\boldsymbol{X}_T)) + \sum_{i=1}^{J}\lambda_i P_i(\boldsymbol{U}^{(k)}\boldsymbol{\beta})\right)\Bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})}\right]^{-1}\left[{}_{\boldsymbol{U}^{(k)}}\nabla P(\boldsymbol{U}^{(k)}\boldsymbol{\beta})|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})}\right]
\tag{18}
$$

    with ${}_{\boldsymbol{U}^{(k)}}\nabla^2$ and ${}_{\boldsymbol{U}^{(k)}}\nabla$ are as defined in (15).
    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

$$
\nabla_{\boldsymbol{\lambda}} L(\boldsymbol{y}_{\boldsymbol{V}}, f_{\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})}(\boldsymbol{X}_{\boldsymbol{V}})) = \left[\boldsymbol{U}^{(k)}\frac{\partial}{\partial\boldsymbol{\lambda}}\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})\right]^{\top}\left[{}_{\boldsymbol{U}^{(k)}}\nabla L(\boldsymbol{y}_{\boldsymbol{V}}, f_{\boldsymbol{U}^{(k)}\boldsymbol{\beta}}(\boldsymbol{X}_{\boldsymbol{V}}))\big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})}\right]
\tag{19}
$$

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

$$
\begin{aligned}
\boldsymbol{\eta} \quad &:= \quad \boldsymbol{\lambda}^{(k)} + \tfrac{k-1}{k+2}\left(\boldsymbol{\lambda}^{(k)} - \boldsymbol{\lambda}^{(k-1)}\right) \\
\boldsymbol{\lambda}^{(k+1)} \quad &:= \quad \boldsymbol{\eta} - t^{(k)}\,\nabla_{\boldsymbol{\lambda}} L\left(\boldsymbol{y}_{\boldsymbol{V}}, f_{\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})}(\boldsymbol{X}_{\boldsymbol{V}})\right)\Big|_{\boldsymbol{\lambda}=\boldsymbol{\eta}}
\end{aligned}
\tag{20}
$$

    if the stopping criteria is reached or

$$
L\left(\boldsymbol{y}_V, f_{\hat{\theta}(\boldsymbol{\lambda}^{(k+1)})}(\boldsymbol{X}_V)\right) > L\left(\boldsymbol{y}_V, f_{\hat{\theta}(\boldsymbol{\lambda}^{(k)})}(\boldsymbol{X}_V)\right),
\tag{21}
$$

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