A Appendix

A.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:

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

A.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 $\boldsymbol{B}$:

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

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

$${}_{\boldsymbol{B}}\nabla L_T(\cdot, \boldsymbol{\lambda}_0)|_{\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}_0)} = \tilde{L}_T(\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}_0), \boldsymbol{\lambda}_0) = 0 \tag{3}$$

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

$$\tilde{L}_T(\gamma(\boldsymbol{\lambda}), \boldsymbol{\lambda}) = \nabla_B L_T(\cdot, \boldsymbol{\lambda})|_{\gamma(\boldsymbol{\lambda})} = 0 \tag{4}$$

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

$$\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}) = \arg\min_{\boldsymbol{\theta}} L_T(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \arg\min_{\boldsymbol{\theta}\in\Omega^{L_T}(\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}_0), \boldsymbol{\lambda}_0)} L_T(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \gamma(\boldsymbol{\lambda}) \tag{5}$$

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. □

A.3 Regression Examples

A.3.1 Elastic Net

We show that the joint optimization problem for the Elastic Net satisfies all three conditions in Theorem 1:

> Condition 1: The elastic net solution paths are piecewise linear (Zou & Hastie 2003), which means that the nonzero indices of the elastic net estimates stay locally constant for almost every $\boldsymbol{\lambda}$. Therefore, $S_{\boldsymbol{\lambda}}$ as defined in Section 2.4.1 is a local optimality space for $L_T(\cdot, \boldsymbol{\lambda})$. ✓

> Condition 2: We only need to establish that the ℓ_1 penalty is twice-continuously differentiable in the directions of $S_{\boldsymbol{\lambda}}$ since the quadratic loss function and the ridge penalty are both smooth. The absolute value function is twice-continuously differentiable everywhere except at zero. Hence the training criterion is smooth when restricted to $S_{\boldsymbol{\lambda}}$. ✓

> Condition 3: The Hessian matrix of $L_T(\cdot, \boldsymbol{\lambda})$ with respect to $\boldsymbol{I}_{I(\boldsymbol{\lambda})}$ is $\boldsymbol{I}_{I(\boldsymbol{\lambda})}^\top \boldsymbol{X}_T^\top \boldsymbol{X}_T \boldsymbol{I}_{I(\boldsymbol{\lambda})} + \lambda_2 \boldsymbol{I}$. The first summand is positive semi-definite. As long as $\lambda_2 > 0$, the contribution of the identity matrix ensures the Hessian is positive definite. ✓

A.3.2 Additive Models with Sparsity and Smoothness Penalties

Let

$$\boldsymbol{U} = \begin{bmatrix} \boldsymbol{U}^{(i_1)} & \ldots & \boldsymbol{U}^{(i_{|J(\boldsymbol{\lambda})|})} \end{bmatrix} \tag{6}$$

where $i_\ell \in J(\boldsymbol{\lambda})$. Then the Hessian matrix in this problem is

$$\boldsymbol{H}(\boldsymbol{\lambda}) = \boldsymbol{U}^\top \boldsymbol{I}_T^\top \boldsymbol{I}_T \boldsymbol{U} + \lambda_0 \, diag\left(\frac{1}{||\boldsymbol{U}^{(i)} \hat{\boldsymbol{\beta}}^{(i)}(\boldsymbol{\lambda})||_2} \left(\boldsymbol{I} - \frac{\hat{\boldsymbol{\beta}}^{(i)}(\boldsymbol{\lambda})^\top \hat{\boldsymbol{\beta}}^{(i)}(\boldsymbol{\lambda})}{||\boldsymbol{U}^{(i)} \hat{\boldsymbol{\beta}}^{(i)}(\boldsymbol{\lambda})||_2} \right) \right) + \epsilon \boldsymbol{I} \tag{7}$$

Now we check that all three conditions are satisfied.

> Condition 1: It seems likely that the space spanned by $S_{\boldsymbol{\lambda}}$ is a local optimality space, though we are unable to formally prove this fact. The training criterion for this problem is composed of generalized lasso penalties and a group lasso penalties. For the generalized lasso, Tibshirani et al. (2011) proved that the solution path is smooth almost everywhere. For the group lasso, there is empirical evidence that the active set is locally constant almost everywhere with respect to the penalty parameter (Yuan & Lin 2006), but this has not been formally proven. Vaiter et al. (2012) showed that the active set is locally constant with respect to the response; we suspect similar techniques could be used to prove our hypothesis.

> Condition 2: We only need to establish that the generalized lasso and group lasso penalties are twice-continuously differentiable in the directions of $S_{\boldsymbol{\lambda}}$ since the rest of the training criterion is smooth. $\|\boldsymbol{D}\boldsymbol{\theta}\|_1$ is not differentiable at the points where $\boldsymbol{D}\boldsymbol{\theta}$

has zero elements. We must therefore restrict the derivatives to be taken in directions such that the zero elements of $\boldsymbol{D\theta}$ remain constant. The ℓ_2 norm is twice-continuously differentiable everywhere except at the zero vector. Hence the training criterion is smooth when restricted to the differentiable space $S_{\boldsymbol{\lambda}}$ specified in Section 2.4.2. ✓

Condition 3: The Hessian matrix in (7) is the sum of positive semi-definite matrices. As long as $\epsilon > 0$, the contribution of the last summand $\epsilon \boldsymbol{I}$ will make the Hessian matrix positive-definite. ✓

In the gradient calculation for this problem, the matrix $\boldsymbol{C}(\boldsymbol{\beta}(\boldsymbol{\lambda}))$ in (29) has columns $i = 1, ..., p$

$$\boldsymbol{C}_i(\boldsymbol{\beta}(\boldsymbol{\lambda})) = \begin{cases} \begin{bmatrix} \mathbf{0} \\ \boldsymbol{U}^{(i)\top} \boldsymbol{D}^{(2)\top}_{\boldsymbol{x}_i} sgn(\boldsymbol{D}^{(2)}_{\boldsymbol{x}_i} \boldsymbol{U}^{(i)} \hat{\boldsymbol{\beta}}^{(i)}) \\ \mathbf{0} \end{bmatrix} & \text{for } i \in J(\boldsymbol{\lambda}) \\ \mathbf{0} & \text{for } i \notin J(\boldsymbol{\lambda}) \end{cases} \tag{8}$$

A.3.3 Un-pooled Sparse Group Lasso

The Hessian in this problem is

$$\boldsymbol{H}(\boldsymbol{\lambda}) = \frac{1}{n} \boldsymbol{X}^{\top}_{T,I(\boldsymbol{\lambda})} \boldsymbol{X}_{T,I(\boldsymbol{\lambda})} + diag\left(\frac{\lambda_m}{||\boldsymbol{\theta}^{(m)}||_2} \left(\boldsymbol{I} - \frac{\boldsymbol{\theta}^{(m)} \boldsymbol{\theta}^{(m)\top}}{||\boldsymbol{\theta}^{(m)}||_2^2} \right) \right) + \epsilon \boldsymbol{I} \tag{9}$$

The logic for checking all three conditions in Theorem 1 is similar to the other examples:

Condition 1: We hypothesize that the differentiable space $S_{\boldsymbol{\lambda}}$ is also a local optimality space, though we have not formally proven this fact. We suspect this is true for the same reasons discussed in Section A.3.2.

Condition 2: The ℓ_1 and ℓ_2 penalties are twice-differentiable when restricted to $S_{\boldsymbol{\lambda}}$ for the same reasons discussed in Section A.3.2. ✓

Condition 3: The Hessian matrix in (9) is the sum of positive semi-definite matrices. It is positive definite for any $\epsilon > 0$ due to the contribution from the last summand $\epsilon \boldsymbol{I}$. ✓

In the gradient calculations for this problem, the matrix $\boldsymbol{C}(\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}))$ in (33) has columns $m = 1, 2..., M$

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

where $\mathbf{0}$ are the appropriate dimensions.

A.3.4 Low-rank Matrix Completion

We first derive a differentiable space of the training criterion (35) with respect to $\boldsymbol{\Gamma}$. Suppose $\boldsymbol{\Gamma} = \boldsymbol{U} diag(\boldsymbol{\sigma}) \boldsymbol{V}^\top$ is the singular value decomposition of $\boldsymbol{\Gamma}$ where the $\boldsymbol{\sigma}$ is a vector of the singular values in non-increasing order. The subdifferential of the nuclear norm is (Parikh et al. 2014)

$$\partial \|\boldsymbol{\Gamma}\|_* = \left\{ \boldsymbol{U} \text{diag}(\boldsymbol{\mu}) \boldsymbol{V}^\top \middle| \mu_i \in \begin{cases} [-1, 1] & \text{if } \sigma_i = 0 \\ \text{sign}(\sigma_i) & \text{if } \sigma_i \neq 0 \end{cases} \right\} \tag{11}$$

The subdifferential of the nuclear norm at $\hat{\boldsymbol{\Gamma}}(\boldsymbol{\lambda})$ reduces to a gradient if we restrict the derivative to be taken in the directions

$$S_{\boldsymbol{\lambda},\boldsymbol{\Gamma}} = \left\{ \boldsymbol{G} \in \mathbb{R}^{N \times N} \middle| \text{range}(\boldsymbol{G}) \subseteq \text{range}(\hat{\boldsymbol{\Gamma}}(\boldsymbol{\lambda})) \right\} \tag{12}$$

$$= \text{span}\left(\left\{ \boldsymbol{B}^{(ij)} = \boldsymbol{u}_j \boldsymbol{e}_i^\top \middle| i, j \in \{1, ..., N\}, \sigma_j \neq 0 \right\} \right) \tag{13}$$

where $\boldsymbol{u}_i$ is the ith column of $\hat{\boldsymbol{U}}(\boldsymbol{\lambda})$ and $\boldsymbol{e}_i \in \mathbb{R}^N$ is the ith standard basis vector. Note that the matrices in (13) form an orthonormal basis of $S_{\boldsymbol{\lambda},\boldsymbol{\Gamma}}$. By re-parameterizing the matrix $\boldsymbol{\Gamma}$ with its basis representation $\sum_{i=1}^N \sum_{j:\sigma_j \neq 0} b_{ij} \boldsymbol{B}^{(ij)}$, the derivative of the nuclear norm with respect to $\boldsymbol{B}^{(ij)}$ reduces to

$$_{\boldsymbol{B}^{(ij)}} \nabla \left\| \sum_{i=1}^N \sum_{j:\sigma_j \neq 0} b_{ij} \boldsymbol{B}^{(ij)} \right\|_* = \text{vec}\left(\boldsymbol{B}^{(ij)} \right)^\top \text{vec}\left(\boldsymbol{U} \text{sign}(\boldsymbol{\sigma}) \boldsymbol{V} \right) \tag{14}$$

where vec$(\cdot)$ denotes the vectorization of a matrix.

Now we derive the gradient of the validation loss with respect to the penalty parameters. One method is to follow Algorithm 2, which gives us expressions involving the partial derivatives of $\hat{\boldsymbol{\eta}}(\boldsymbol{\lambda})$,$\hat{\boldsymbol{\gamma}}(\boldsymbol{\lambda})$, $\hat{\boldsymbol{b}}(\boldsymbol{\lambda})$, $\hat{\boldsymbol{U}}(\boldsymbol{\lambda})$, and $\hat{\boldsymbol{V}}(\boldsymbol{\lambda})$. To solve for their partial derivatives, we would need to know how the basis representation of $\boldsymbol{\Gamma} = \sum_{i=1}^N \sum_{j:\sigma_j \neq 0} b_{ij} \boldsymbol{B}^{(ij)}$ maps to the singular value decomposition of $\boldsymbol{\Gamma} = \boldsymbol{U} \text{diag}(\boldsymbol{\sigma}) \boldsymbol{V}^\top$. Unfortunately, this mapping is complicated.

We can avoid this issue altogether by considering a different representation of the differentiable space. It is easy to show that $S_{\boldsymbol{\lambda},\boldsymbol{\Gamma}}$ is the set of directions that preserve the rank of $\hat{\boldsymbol{\Gamma}}(\boldsymbol{\lambda})$

$$S_{\boldsymbol{\lambda},\boldsymbol{\Gamma}} = \left\{ \boldsymbol{G} \middle| \text{rank}\left(\hat{\boldsymbol{\Gamma}}(\boldsymbol{\lambda}) + \delta \boldsymbol{G} \right) \leq \text{rank}\left(\hat{\boldsymbol{\Gamma}}(\boldsymbol{\lambda}) \right), \delta \in [0, 1] \right\} \tag{15}$$

Therefore, we can express $\boldsymbol{\Gamma}$ in the training criterion using a singular value decomposition where the dimensions of the components are restricted: $\boldsymbol{U} \in \mathbb{R}^{\text{rank}(\hat{\boldsymbol{\Gamma}}(\boldsymbol{\lambda})) \times N}$ $\boldsymbol{\sigma} \in \mathbb{R}^{\text{rank}(\hat{\boldsymbol{\Gamma}}(\boldsymbol{\lambda}))}$ and $\boldsymbol{V} \in \mathbb{R}^{N \times \text{rank}(\hat{\boldsymbol{\Gamma}}(\boldsymbol{\lambda}))}$. This re-parameterization of $\boldsymbol{\Gamma}$ gives us an equivalent, smooth training

criterion at $\boldsymbol{\lambda}$:

$$\begin{aligned}\underset{\boldsymbol{\eta},\boldsymbol{\gamma},\boldsymbol{U},\boldsymbol{\sigma},\boldsymbol{V}}{\arg\min}\ & \frac{1}{2|T|}\left\|\boldsymbol{M}-\boldsymbol{X}_{I_r(\boldsymbol{\lambda})}\boldsymbol{\eta}\mathbf{1}^\top-(\boldsymbol{Z}_{I_c(\boldsymbol{\lambda})}\boldsymbol{\gamma}\mathbf{1}^\top)^\top-\boldsymbol{U}diag(\boldsymbol{\sigma})\boldsymbol{V}^\top\right\|_T^2+\lambda_0\left\|\boldsymbol{U}diag(\boldsymbol{\sigma})\boldsymbol{V}^\top\right\|_* \\ & +\sum_{g=1}^{G}\lambda_g\|\boldsymbol{\eta}^{(g)}\|_2+\sum_{g=1}^{G}\lambda_{G+g}\|\boldsymbol{\gamma}^{(g)}\|_2+\frac{1}{2}\epsilon\left(\|\boldsymbol{\eta}\|_2^2+\|\boldsymbol{\gamma}\|_2^2+\left\|\boldsymbol{U}diag(\boldsymbol{\sigma})\boldsymbol{V}^\top\right\|_F^2\right)\end{aligned} \tag{16}$$

$$\text{s.t. } \boldsymbol{V}^\top\boldsymbol{V}=\boldsymbol{I} \text{ and } \boldsymbol{U}^\top\boldsymbol{U}=\boldsymbol{I} \tag{17}$$

We can take the subgradient of the training criterion with respect to $\boldsymbol{\Gamma}=\boldsymbol{U}diag(\boldsymbol{\sigma})\boldsymbol{V}^\top$ and multiply it by its left singular vectors to get the following gradient optimality condition:

$$\begin{aligned}\mathbf{0}=&-\frac{1}{|T|}\hat{\boldsymbol{U}}(\boldsymbol{\lambda})^\top\left(\boldsymbol{M}-\boldsymbol{X}_{I_r(\boldsymbol{\lambda})}\hat{\boldsymbol{\eta}}(\boldsymbol{\lambda})\mathbf{1}^\top-(\boldsymbol{Z}_{I_c(\boldsymbol{\lambda})}\hat{\boldsymbol{\gamma}}(\boldsymbol{\lambda})\mathbf{1}^\top)^\top\right)+\frac{1}{|T|}\text{diag}(\hat{\boldsymbol{\sigma}}(\boldsymbol{\lambda}))\hat{\boldsymbol{V}}^\top(\boldsymbol{\lambda}) \\ &+\lambda_0\text{diag}(\text{sign}(\hat{\boldsymbol{\sigma}}(\boldsymbol{\lambda})))\hat{\boldsymbol{V}}(\boldsymbol{\lambda})^\top+\epsilon\text{diag}(\hat{\boldsymbol{\sigma}}(\boldsymbol{\lambda}))\hat{\boldsymbol{V}}(\boldsymbol{\lambda})^\top\end{aligned} \tag{18}$$

Similarly, multiply the result by its right singular vectors to get

$$\begin{aligned}\mathbf{0}=&-\frac{1}{|T|}\left(\boldsymbol{M}-\boldsymbol{X}_{I_r(\boldsymbol{\lambda})}\hat{\boldsymbol{\eta}}(\boldsymbol{\lambda})\mathbf{1}^\top-(\boldsymbol{Z}_{I_c(\boldsymbol{\lambda})}\hat{\boldsymbol{\gamma}}(\boldsymbol{\lambda})\mathbf{1}^\top)^\top\right)\hat{\boldsymbol{V}}(\boldsymbol{\lambda})+\frac{1}{|T|}\hat{\boldsymbol{U}}(\boldsymbol{\lambda})\text{diag}(\hat{\boldsymbol{\sigma}}(\boldsymbol{\lambda})) \\ &+\lambda_0\hat{\boldsymbol{U}}(\boldsymbol{\lambda})\text{diag}(\text{sign}(\hat{\boldsymbol{\sigma}}(\boldsymbol{\lambda})))+\epsilon\hat{\boldsymbol{U}}(\boldsymbol{\lambda})\text{diag}(\hat{\boldsymbol{\sigma}}(\boldsymbol{\lambda}))\end{aligned} \tag{19}$$

The gradient optimality conditions with respect to $\boldsymbol{\eta}$ and $\boldsymbol{\gamma}$ do not require any special manipulation and follows Algorithm 2 exactly. Finally, we implicitly differentiate the gradient optimality conditions, as well as the conditions in (17), with respect to $\boldsymbol{\lambda}$. We can easily solve the resulting system of linear equations to get the partial derivatives of $\hat{\boldsymbol{\eta}}(\boldsymbol{\lambda}), \hat{\boldsymbol{\gamma}}(\boldsymbol{\lambda}), \hat{\boldsymbol{U}}(\boldsymbol{\lambda}), \hat{\boldsymbol{\sigma}}(\boldsymbol{\lambda})$, and $\hat{\boldsymbol{V}}(\boldsymbol{\lambda})$.

We now show that the conditions in Theorem 1 are satisfied.

Condition 1: We hypothesize that the differentiable space $S_{\boldsymbol{\lambda}}$ defined in (37) is also a local optimality space $\boldsymbol{\lambda}$. For the group lasso penalties, we use the same reasons mentioned in A.3.2 to justify this hypothesis. For the nuclear norm penalty, it has been observed empirically that small perturbations in the penalty parameter result in matrices with same rank (Mazumder et al. 2010). This supports our belief that $S_{\boldsymbol{\lambda},\boldsymbol{\Gamma}}$ is a local optimality space with respect to $\boldsymbol{\Gamma}$ at $\boldsymbol{\lambda}$.

Condition 2: The only non-smooth components of the training criterion are the group lasso and nuclear norm penalties. The group lasso penalty is twice-differentiable when restricted to the differentiable space, using the same reasoning in Section A.3.2. From (14), we see that the nuclear norm $\|\boldsymbol{\Gamma}\|_*$ is also twice-differentiable with respect to $\boldsymbol{\Gamma}$ when restricted to $S_{\boldsymbol{\lambda},\boldsymbol{\Gamma}}$. ✓

Condition 3: As shown in this section, there is an orthonormal basis $\boldsymbol{B}$ of the differentiable space. Therefore the Hessian matrix of the training criterion with respect to $\boldsymbol{B}$ exists. Since the training criterion is the sum of convex functions with ridge penalties on all the variables, the Hessian of the training criterion is positive definite for any $\epsilon > 0$. ✓

A.4 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 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

A.5 Sensitivity to initialization points

Since the results of gradient descent and Nelder-Mead depend on their initialization points, we ran a simulation to see how sensitive the methods were to where they were initialized and how many initializations were used.

We tested a smaller version of the joint optimization problem in Section 2.4.2. Here we use 60 training, 30 validation, and 30 test observations and $p = 15$ covariates. The response was generated from (40). We initialized $\boldsymbol{\lambda}$ by considering all possible combinations of $(\lambda_0, \lambda_1 \mathbf{1})$ where $\lambda_0, \lambda_1 \in \{10^i : i \in \{-2, -1, 0, 1\}\}$.

In Figure A.5 (left), we plot the validation error as the number of initializations increases. The validation errors from both methods plateau quickly. Gradient descent manages to find penalty parameters with lower validation error than Nelder-Mead. Figure A.5 (right) presents the distribution of validation errors resulting from the random initializations. On average, gradient descent finds penalty parameters with lower validation error compared to Nelder-Mead. The plots show that the methods are indeed sensitive to their initialization points. For example, one could run a very coarse grid search on the two-parameter version of the joint optimization problem and use the best one or two penalty parameter values.

A.6 Additional simulation results

The simulation results in Section 3 show that joint optimization problems with many penalty parameters can produce better models than those with only two penalty parameters. One may wonder if this difference is due to the method used to tune the penalty parameters. Here we present results from tuning the two-penalty-parameter joint optimization problems from Sections 3.2, 3.3, and 3.4 using gradient descent, Nelder-Mead, and Spearmint. As shown in Table 7, the performance of these methods are very similar to grid search. Regardless of the method used to tune the two-penalty parameter joint optimization, the resulting models all have higher validation and test error compared to the models from the joint optimization problem with many penalty parameters tuned by gradient descent.

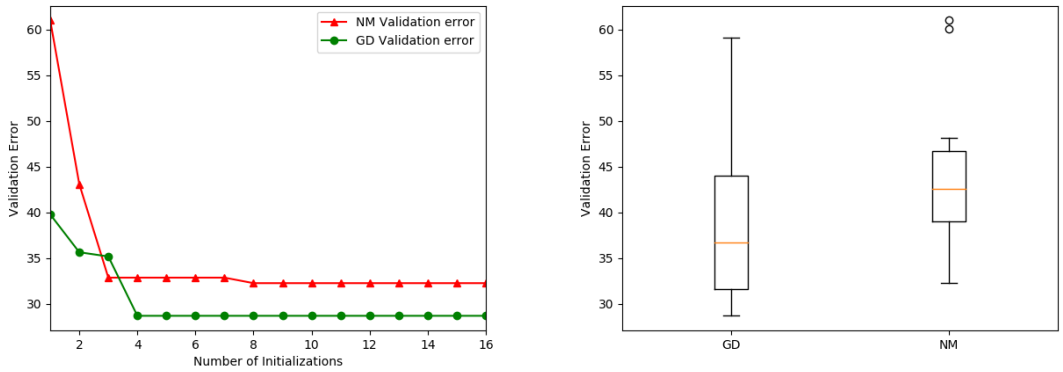
Table 7: Two-parameter joint optimization problems for the examples in Section 3. Standard errors are given in parentheses. We abbreviated the methods as follows: Gradient Descent = GD, Nelder-Mead = NM, Spearmint = SP, Grid Search = GS

Sparse additive models			
	Validation Error	Test Error	# Solves
GD	23.87 (0.97)	26.10 (0.86)	13.07
NM	28.86 (1.04)	29.97 (0.96)	100
SP	29.18 (1.07)	30.09 (1.08)	100
GS	28.71 (0.97)	29.42 (0.96)	100

Sparse Group Lasso			
n=90, p=600, M=30			
	Validation Err	Test Err	# Solves
GD	46.82 (2.21)	49.33 (1.36)	21.43
NM	46.37 (2.24)	48.95 (1.35)	100
SP	45.70 (2.32)	49.35 (1.56)	100
GS	47.23 (2.26)	50.01 (1.40)	100
n=90, p=900, M=60			
	Validation Error	Test Error	# Solves
GD	45.71 (2.26)	50.31 (1.93)	20.77
NM	44.95 (2.24)	50.18 (1.82)	100
SP	49.59 (2.27)	56.54 (2.14)	100
GS	45.70 (2.27)	51.34 (1.86)	100
n=90, p=1200, M=100			
	Validation Error	Test Error	# Solves
GD	50.46 (2.30)	57.02 (1.94)	19.80
NM	49.92 (2.33)	55.46 (1.89)	100
SP	49.70 (2.26)	56.51 (2.16)	100
GS	50.00 (2.16)	57.14 (2.18)	100

Low-rank Matrix Completion			
	Validation Err	Test Err	Num Solves
GD	0.70 (0.04)	0.71 (0.04)	8.03 (0.79)
NM	0.71 (0.04)	0.71 (0.04)	100
SP	0.73 (0.04)	0.74 (0.04)	100
GS	0.71 (0.04)	0.72 (0.04)	100

Figure 1: Error of additive models tuned by Gradient Descent vs. Nelder-Mead. Left: Validation error of models after as the number of initialization points increases. Right: The distribution of validation errors. (Gradient Descent = GD, Nelder-Mead = NM)



References

Mazumder, R., Hastie, T. & Tibshirani, R. (2010), 'Spectral regularization algorithms for learning large incomplete matrices', *Journal of machine learning research* **11**(Aug), 2287–2322.

Parikh, N., Boyd, S. et al. (2014), 'Proximal algorithms', *Foundations and Trends® in Optimization* **1**(3), 127–239.

Tibshirani, R. J., Taylor, J. et al. (2011), 'The solution path of the generalized lasso', *The Annals of Statistics* **39**(3), 1335–1371.

Vaiter, S., Deledalle, C., Peyré, G., Fadili, J. & Dossal, C. (2012), 'The degrees of freedom of the group lasso', *arXiv preprint arXiv:1205.1481* .

Yuan, M. & Lin, Y. (2006), 'Model selection and estimation in regression with grouped variables', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**(1), 49–67.

Zou, H. & Hastie, T. (2003), 'Regression shrinkage and selection via the elastic net, with applications to microarrays', *Journal of the Royal Statistical Society: Series B. v67* pp. 301–320.