Supplementary Materials

A Correlation Heatmap for Mapping Matrix M

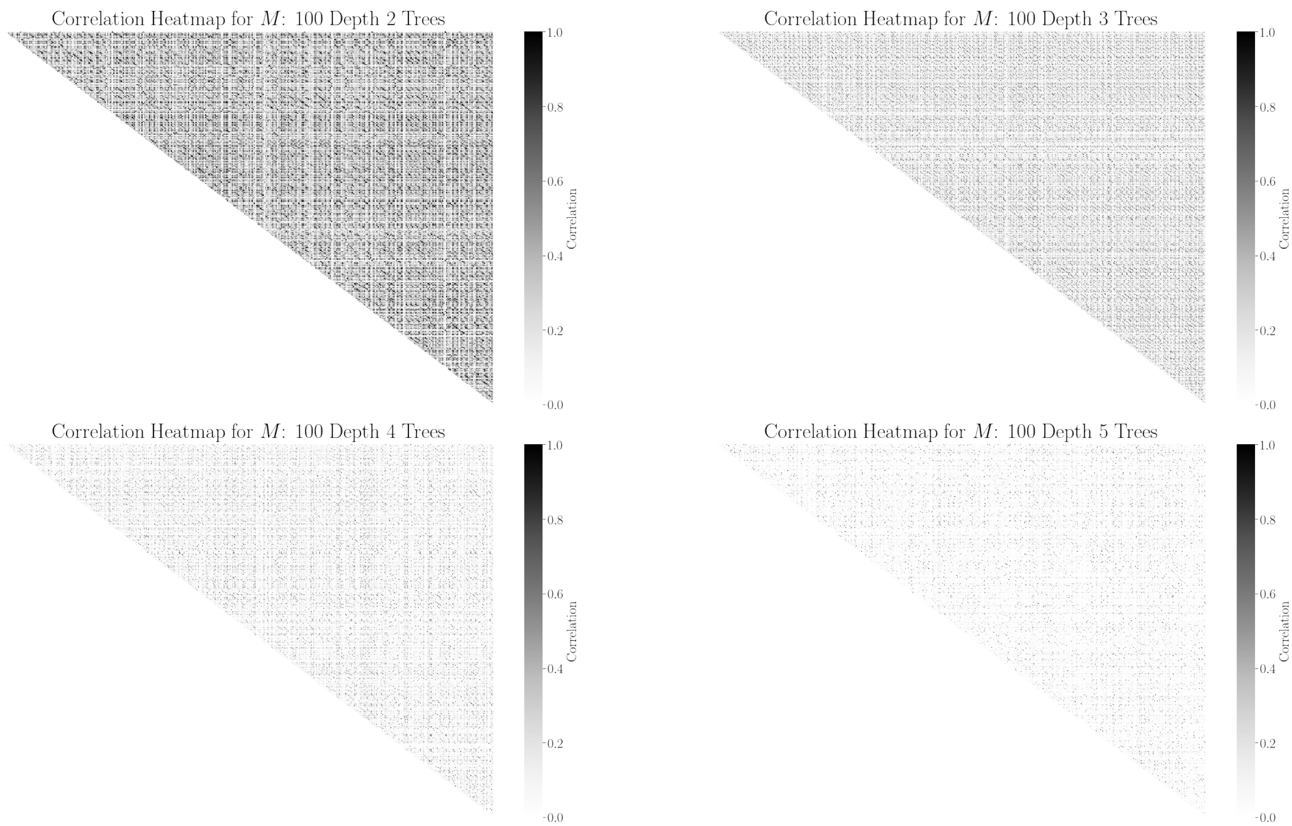

Figure 1: Heatmap showing the absolute value of the correlation coefficients between the columns of mapping matrix M. Shallow tree ensembles contain many similar trees so, as a result, mapping matrix M contains many correlated columns. RuleFit performs poorly when selecting sparse ensembles due to the limitations of LASSO selection [3]. Even for deeper ensembles, a few columns of M are highly correlated.

B MCP Penalty Function Visualization

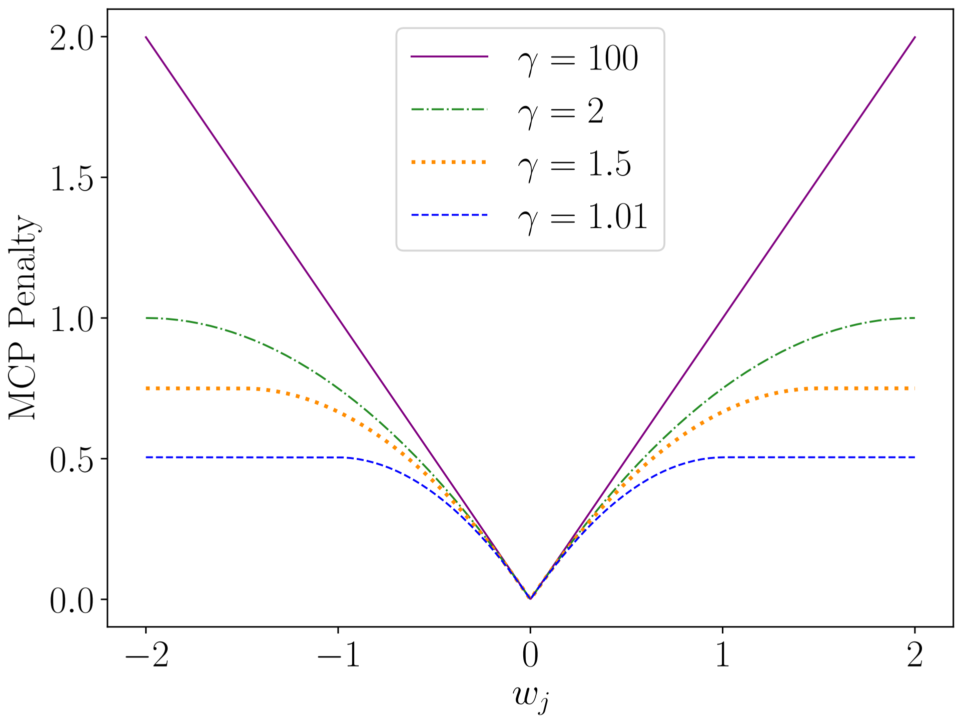

Figure 2: Parameter γ controls the behavior of the MCP penalty, and increasing γ increases shrinkage in w_j. For large values of γ the MCP penalty behaves similar to the ℓ_1-penalty. When $\gamma \to 1^+$, the MCP penalty behaves like the ℓ_0-penalty.

C Block Update Derivations and Proofs

Below we present detailed derivations for expressions referred to in the block update section §4.1 of the main paper.

C.1 Derivation of the proximal block update closed-form minimizer for the ℓ_1-sparsity penalty only.

We want to derive the closed form minimizer for the univariate LASSO problem:

$$\min_{\theta} \quad \frac{1}{2}\left\|\theta - \hat{\theta}\right\|_2^2 + \frac{\lambda_s}{L_t}\left\|\theta\right\|_1.$$

We follow the steps outlined in [1] and compute the optimality conditions. The gradient of the smooth loss function is $\theta - \hat{\theta}$ and the subgradient of the non-smooth ℓ_1-penalty is given elementwise by:

$$\partial\left\|\theta_j\right\|_1 \in \begin{cases} \operatorname{sign}(\theta_j) & \theta_j \neq 0 \\ [-1, 1] & \theta_j = 0. \end{cases} \tag{1}$$

Stationary point θ^* must satisfy the following subgradient condition:

$$0 \in \theta^* - \hat{\theta} + \frac{\lambda_s}{L_t}\partial\left\|\theta^*\right\|_1. \tag{2}$$

We can write these conditions elementwise as:

$$0 \in \begin{cases} \theta_j^* - \hat{\theta}_j + \frac{\lambda_s}{L_t}\operatorname{sign}(\theta_j^*) & \theta_j^* \neq 0 \\ [-\hat{\theta}_j - \frac{\lambda_s}{L_t}, -\hat{\theta}_j + \frac{\lambda_s}{L_t}] & \theta_j^* = 0. \end{cases} \tag{3}$$

This decomposes into three cases.

Case 1: $\theta_j^* < 0$

$$\theta_j^* = \hat{\theta}_j + \frac{\lambda_s}{L_t}$$

Case 2: $\theta_j^* > 0$

$$\theta_j^* = \hat{\theta}_j - \frac{\lambda_s}{L_t}$$

Case 3: $\theta_j^* = 0$

$$0 \in [-\hat{\theta}_j - \frac{\lambda_s}{L_t}, -\hat{\theta}_j + \frac{\lambda_s}{L_t}] \tag{4}$$

$$\implies -\frac{\lambda_s}{L_t} \leq \hat{\theta}_j \leq \frac{\lambda_s}{L_t} \tag{5}$$

From these 3 cases, we see that θ_j^* is defined by:

$$\theta_j^* = \begin{cases} \hat{\theta}_j + \frac{\lambda_s}{L_t} & \hat{\theta}_j \leq -\frac{\lambda_s}{L_t} \\ \hat{\theta}_j - \frac{\lambda_s}{L_t} & \hat{\theta}_j \geq \frac{\lambda_s}{L_t} \\ 0 & -\frac{\lambda_s}{L_t} \leq \hat{\theta}_j \leq \frac{\lambda_s}{L_t}, \end{cases}$$

which is equivalent to $\theta_j^* = S_{\frac{\lambda_s}{L_t}}(\hat{\theta}_j)$, where S_λ is the soft-thresholding operator. This shows that the closed-form minimizer for the univariate LASSO problem is the soft-thresholding operator, as desired.

C.2 Derivation of the proximal block update closed-form minimizer for the MCP-sparsity penalty only.

We want to derive the closed form minimizer for this univariate problem:

$$\min_{\theta} \quad \frac{1}{2}\left\|\theta - \hat{\theta}\right\|_2^2 + \frac{1}{L_t}\sum_{j=1}^{R_t} P_{\gamma_t}(\theta_j, \lambda_s). \tag{6}$$

We can expand the MCP sparsity penalty and rewrite the problem as:

$$\min_{\theta} \quad \frac{1}{2}\left\|\theta - \hat{\theta}\right\|_2^2 + \frac{1}{L_t}\sum_{j:\ |\theta_j| \leq \lambda_s\gamma_t} \left(\lambda_s|\theta_j| - \frac{\theta_j^2}{2\gamma_t}\right) + \frac{1}{2L_t}\sum_{j:\ |\theta_j| > \lambda_s\gamma_t} \gamma_t\lambda_s^2.$$

The subgradient optimality conditions can be expressed elementwise by:

$$0 \in \begin{cases} \left[-\hat{\theta}_j - \frac{\lambda_s}{L_t}, -\hat{\theta}_j + \frac{\lambda_s}{L_t}\right] & |\theta_j^*| = 0 \\ \theta_j^* - \hat{\theta}_j + \frac{\lambda_s}{L_t}\text{sign}(\theta_j^*) - \frac{\theta_j^*}{\gamma_t L_t} & 0 < |\theta_j^*| \leq \lambda_s\gamma_t \\ \theta_j^* - \hat{\theta}_j & |\theta_j^*| > \lambda_s\gamma_t. \end{cases}$$

We can again decompose this into cases:

Case 1: $\theta_j^* = 0$

$$0 \in \left[-\hat{\theta}_j - \frac{\lambda_s}{L_t}, -\hat{\theta}_j + \frac{\lambda_s}{L_t}\right]$$

$$\implies -\frac{\lambda_s}{L_t} \leq \hat{\theta}_j \leq \frac{\lambda_s}{L_t}$$

Case 2: $0 < \theta_j^* \leq \lambda_s \gamma_t$

$$0 = \theta_j^* \left(1 - \frac{1}{\gamma_t} L_t\right) - \hat{\theta}_j + \frac{\lambda_s}{L_t}$$

$$\implies \theta_j^* = \left(\frac{\gamma_t L_t}{\gamma_t L_t - 1}\right)\left(\hat{\theta}_j - \frac{\lambda_s}{L_t}\right)$$

Case 3: $\lambda_s \gamma_t \leq \theta_j^* < 0$

$$\theta_j^* = \left(\frac{\gamma_t L_t}{\gamma_t L_t - 1}\right)\left(\hat{\theta}_j + \frac{\lambda_s}{L_t}\right)$$

Case 4: $\theta_j^* > \lambda_s \gamma_t$

$$\theta_j^* = \hat{\theta}$$

Combining these 4 cases, we get that:

$$\theta_j^* = \begin{cases} \left(\frac{\gamma_t L_t}{\gamma_t L_t - 1}\right) S_{\frac{\lambda_s}{L_t}}(\hat{\theta}_j) & |\hat{\theta}_j| \leq \lambda_s \gamma_t \\ \hat{\theta}_j & |\hat{\theta}_j| > \lambda_s \gamma_t. \end{cases} \tag{7}$$

Let $\gamma = \gamma_t L_t$. With this substitution we have that:

$$\theta_j^* = \begin{cases} (\frac{\gamma}{\gamma - 1}) S_{\frac{\lambda_s}{L_t}}(\hat{\theta}_j) & |\hat{\theta}_j| \leq \frac{\lambda_s \gamma}{L_t} \\ \hat{\theta}_j & |\hat{\theta}_j| > \frac{\lambda_s \gamma}{L_t}, \end{cases} \tag{8}$$

as desired; we denote this the MCP thresholding operator. The parameter γ is the global parameter for the concavity of the MCP penalty; γ_t varies by block and is equal to $\frac{\gamma}{L_t}$.

C.3 Proof of soft-thresholding operator for the fused lasso solution.

We want to show that soft-thresholding the solution for $\theta^*(0, \lambda_f)$ is the optimal solution for the problem when $\lambda_s > 0$. This directly follows the proof for lemma A.1 in [2]. Start with the block update problem for when the fusion and sparsity penalties are nonzero:

$$\min_{\theta} \quad \frac{1}{2} \left\|\theta - \hat{\theta}\right\|_2^2 + \frac{\lambda_s}{L_t} \|\theta\|_1 + \frac{\lambda_f}{L_t} \|D_t \theta\|_1. \tag{9}$$

The subgradient equations are for this objective are [4]:

$$0 \in \theta^* - \hat{\theta} + \frac{\lambda_s}{L_t} \partial \|\theta^*\|_1 + \frac{\lambda_f}{L_t} D_t^\intercal \partial \|D_t \theta^*\|_1, \tag{10}$$

where element wise:

$$\partial \|(D_t \theta)_j\|_1 = \begin{cases} \text{sign}(D_t \theta)_j & (D_t \theta)_j \neq 0 \\ [-1, 1] & (D_t \theta^*)_j = 0. \end{cases}$$

The subgradient equations can be equivalently expressed elementwise by:

$$0 = \theta_j^* - \hat{\theta}_j + \frac{\lambda_s}{L_t}\partial \left\|\theta_j^*\right\|_1 + \underbrace{\frac{\lambda_f}{L_t}\partial \left\|\theta_j^* - \theta_{j-1}^*\right\|_1 - \frac{\lambda_f}{L_t}\partial \left\|\theta_{j+1}^* - \theta_j^*\right\|_1}_{\text{fusion penalty subgradients}}. \tag{11}$$

Assume $\theta^*(0, \lambda_f)$ be the optimal solution to problem 9 with $\lambda_s = 0$. We want to show that $\theta^* = S_{\frac{\lambda_s}{L_t}}(\theta^*(0, \lambda_f))$ satisfies equation 11. One important thing to note is that applying elementwise soft-thresholding does not change the fusion penalty subgradients [2]. Let $\theta_j' = \theta^*(0, \lambda_f)_j$:

$$\frac{\lambda_f}{L_t}\partial \left\|\theta_j^* - \theta_{j-1}^*\right\|_1 - \frac{\lambda_f}{L_t}\partial \left\|\theta_{j+1}^* - \theta_j^*\right\|_1 = \frac{\lambda_f}{L_t}\partial \left\|\theta_j' - \theta_{j-1}'\right\|_1 - \frac{\lambda_f}{L_t}\partial \left\|\theta_{j+1}' - \theta_j'\right\|_1.$$

Equation 11 decomposes into two cases:

Case 1: $|\theta_j'| \geq \frac{\lambda_s}{L_t}$

$$\theta_j^* = S_{\frac{\lambda_s}{L_t}}(\theta_j') = \text{sign}(\theta_j')\left(|\theta_j'| - \frac{\lambda_s}{L_t}\right)_+ = \text{sign}(\theta_j')\left(|\theta_j'| - \frac{\lambda_s}{L_t}\right) = \theta_j' - \text{sign}(\theta_j')\frac{\lambda_s}{L_t}$$

Since $|\theta_j'| \geq \frac{\lambda_s}{L_t} > 0$ we have that:

$$\text{sign}(\theta_j')\frac{\lambda_s}{L_t} = \frac{\lambda_s}{L_t}\partial \left\|\theta_j'\right\|_1.$$

Plugging this in yields:

$$0 = -\hat{\theta}_j + \theta_j' - \frac{\lambda_s}{L_t}\partial \left\|\theta_j'\right\|_1 + \frac{\lambda_s}{L_t}\partial \left\|S_{\frac{\lambda_s}{L_t}}(\theta_j')\right\|_1 + \frac{\lambda_f}{L_t}\partial \left\|\theta_j' - \theta_{j-1}'\right\|_1 - \frac{\lambda_f}{L_t}\partial \left\|\theta_{j+1}' - \theta_j'\right\|_1.$$

Elementwise soft-thresholding does not change signs, and when $S_{\frac{\lambda_s}{L_t}}(\theta_j') = 0$ we can set $\partial \left\|S_{\frac{\lambda_s}{L_t}}(\theta_j')\right\|_1 = \partial \left\|\theta_j'\right\|_1$. Therefore, we have that:

$$0 = -\hat{\theta}_j + \theta_j' - \frac{\lambda_f}{L_t}\partial \left\|\theta_j' - \theta_{j-1}'\right\|_1 - \frac{\lambda_f}{L_t}\partial \left\|\theta_{j+1}' - \theta_j'\right\|_1.$$

This equation is the optimality condition for problem 9 with $\lambda_s = 0$ and it holds since θ' is the optimal solution by assumption.

Case 2: $|\theta_j'| < \frac{\lambda_s}{L_t}$

$$0 = -\hat{\theta}_j + \frac{\lambda_s}{L_t}\partial \left\|S_{\frac{\lambda_s}{L_t}}(\theta_j')\right\|_1 + \frac{\lambda_f}{L_t}\partial \left\|\theta_j' - \theta_{j-1}'\right\|_1 - \frac{\lambda_f}{L_t}\partial \left\|\theta_{j+1}' - \theta_j'\right\|_1$$

We have that $S_{\frac{\lambda_s}{L_t}}(\theta_j') = 0$ since $|\theta_j'| < \frac{\lambda_s}{L_t}$, so:

$$\partial \left\|S_{\frac{\lambda_s}{L_t}}(\theta_j')\right\|_1 \in [-1, 1],$$

and we can chose:

$$\partial \left\|S_{\frac{\lambda_s}{L_t}}(\theta_j')\right\|_1 = \frac{\theta_j' L_t}{\lambda_s}.$$

Plugging this in to the top expression yields:

$$0 = -\hat{\theta}_j + \theta_j' - \frac{\lambda_f}{L_t}\partial \left\|\theta_j' - \theta_{j-1}'\right\|_1 - \frac{\lambda_f}{L_t}\partial \left\|\theta_{j+1}' - \theta_j'\right\|_1,$$

which holds because θ' is the optimal solution by assumption.

With these two cases satisfied, we show that soft-thresholding the optimal solution to the $\lambda_s = 0$ problem, $\theta^*(0, \lambda_f)$ returns the optimal solution to problem 9 with $\lambda_s > 0$.

C.4 Proof of MCP-thresholding operator for the fused lasso solution.

Start with this optimization problem:

$$\min_{\theta} \quad \frac{1}{2}\left\|\theta-\hat{\theta}\right\|_2^2+\frac{1}{L_t}\sum_{j=1}^{R_t}P_{\gamma_t}(\theta_j,\lambda_s)+\frac{\lambda_f}{L_t}\left\|D_t\theta\right\|_1. \tag{12}$$

We want to show that applying the MCP thresholding operator to $\theta^*(0,\lambda_f)$ returns a stationary point for problem 12 with $\lambda_s>0$.

The subgradient equations for the objective can be expressed elementwise by:

$$0=\theta_j^*-\hat{\theta}_j+\frac{1}{L_t}\partial P_{\gamma_t}(\theta_j^*,\lambda_s)+\underbrace{\frac{\lambda_f}{L_t}\partial\left\|\theta_j^*-\theta_{j-1}^*\right\|_1-\frac{\lambda_f}{L_t}\partial\left\|\theta_{j+1}^*-\theta_j^*\right\|_1}_{\text{fusion penalty subgradients}}, \tag{13}$$

where the subgradient of the MCP penalty function is defined by:

$$\partial P_{\gamma_t}(\theta_j,\lambda_s)\in\begin{cases}[-\lambda_s,\lambda_s] & |\theta_j|=0\\ \lambda_s\text{sign}(\theta_j)-\frac{\theta_j}{\gamma_t} & 0<|\theta_j|\le\lambda_s\gamma_t\\ 0 & |\theta_j|>\lambda_s\gamma_t.\end{cases} \tag{14}$$

Recall from §C.2 that the MCP thresholding operator $MCP(\theta_j,\frac{\lambda_s}{L_t},\gamma_t)$ is given by:

$$MCP(\theta_j,\frac{\lambda_s}{L_t},\gamma_t)=\begin{cases}\left(\frac{\gamma_tL_t}{\gamma_tL_t-1}\right)S_{\frac{\lambda_s}{L_t}}(\theta_j) & |\theta_j|\le\lambda_s\gamma_t\\ \theta_j & |\theta_j|>\lambda_s\gamma_t,\end{cases} \tag{15}$$

where S is the soft thresholding operator. Again, assume $\theta^*(0,\lambda_f)=\theta'$ be the optimal solution to problem 12 with $\lambda_s=0$. We want to show that $\theta^*=MCP(\theta',\frac{\lambda_s}{L_t},\gamma_t)$ satisfies the necessary optimality condition (equation 13) and is a stationary point.

Lemma 1. *Applying the MCP thresholding operator elementwise does not change the subgradients of the fusion penalty.*

Proof of Lemma 1. Consider two arbitrary elements θ_1 and θ_2. If $\theta_1 = \theta_2 = 0$, we can set the fusion penalty subgradients before MCP thresholding equal to the fusion penalty subgradients after MCP thresholding, since the $MCP(0) = 0$ so both subgradients can be chosen arbitrarily in $[-1, 1]$.

Assume without loss of generality that $|\theta_1| \geq |\theta_2|$. We need to show that applying the MCP thresholding operation does not change the order of these elements to show that the subgradients are equivalent after MCP thresholding [2]. We can decompose this into 3 cases.

Case 1: $|\theta_1| \geq |\theta_2| > \lambda_s \gamma_t$

The MCP thresholding operator does not change any of the elements for this case so the ordering is preserved.

Case 2: $\lambda_s \gamma_t \geq |\theta_1| \geq |\theta_2|$

For this case, the MCP thresholding operator is the soft thresholding operator scaled by a nonnegative constant. We know from §C.3 and [2] that the soft-thresholding operator preserves ordering between elements. Multiplying both elements by a nonnegative scalar preserves ordering as well.

Case 3: $|\theta_1| > \lambda_s \gamma_t \geq |\theta_2|$

We have that:

$$\left| MCP(\theta_1, \frac{\lambda_s}{L_t}, \gamma_t) \right| = |\theta_1| > \lambda_s \gamma_t.$$

Consider θ_2:

$$\left| MCP(\theta_2, \frac{\lambda_s}{L_t}, \gamma_t) \right| = \left| \left(\frac{\gamma_t L_t}{\gamma_t L_t - 1} \right) \text{sign}(\theta_2) \left(|\theta_2| - \frac{\lambda_s}{L_t} \right)_+ \right|$$

$$\leq \left(\frac{\gamma_t L_t}{\gamma_t L_t - 1} \right) \left(\lambda_s \gamma_t - \frac{\lambda_s}{L_t} \right)_+$$

$$\leq \left(\frac{\gamma_t L_t}{\gamma_t L_t - 1} \right) \left(\lambda_s \left(\frac{\gamma_t L_t - 1}{L_t} \right) \right)_+$$

$$= \gamma_t \lambda_s < |\theta_1|,$$

since we have that $\gamma_t L_t = \gamma$, the MCP concavity hyperparameter, which is defined to be greater than 1. Therefore, we have that:

$$\left| MCP(\theta_2, \frac{\lambda_s}{L_t}, \gamma_t) \right| \leq \left| MCP(\theta_1, \frac{\lambda_s}{L_t}, \gamma_t) \right|,$$

so the ordering is preserved.

We have that:

$$\frac{\lambda_f}{L_t} \partial \left\| \theta_j^* - \theta_{j-1}^* \right\|_1 - \frac{\lambda_f}{L_t} \partial \left\| \theta_{j+1}^* - \theta_j^* \right\|_1 = \frac{\lambda_f}{L_t} \partial \left\| \theta_j' - \theta_{j-1}' \right\|_1 - \frac{\lambda_f}{L_t} \partial \left\| \theta_{j+1}' - \theta_j' \right\|_1,$$

when $\theta^* = MCP(\theta', \frac{\lambda_s}{L_t}, \gamma_t)$. Plug in $\theta^* = MCP(\theta', \frac{\lambda_s}{L_t}, \gamma_t)$ into equation 13, the equation can be decomposed into the following cases.

Case 1: $|\theta_j'| > \lambda_s \gamma_t$

This implies $\theta_j' = \theta_j^*$.

$$0 = \theta'_j - \hat{\theta}_j + \frac{1}{L_t}\partial P_{\gamma_t}(\theta'_j, \lambda_s) + \frac{\lambda_f}{L_t}\partial \left\|\theta'_j - \theta'_{j-1}\right\|_1 - \frac{\lambda_f}{L_t}\partial \left\|\theta'_{j+1} - \theta'_j\right\|_1$$

$$= \theta'_j - \hat{\theta}_j + \frac{\lambda_f}{L_t}\partial \left\|\theta'_j - \theta'_{j-1}\right\|_1 - \frac{\lambda_f}{L_t}\partial \left\|\theta'_{j+1} - \theta'_j\right\|_1$$

This holds because θ' is assumed to be the solution to problem 12 with $\lambda_s = 0$.

Case 2: $\frac{\lambda_s}{L_t} \leq |\theta'_j| \leq \lambda_s\gamma_t$

Recall again we have defined $\gamma = \gamma_t L_t > 1$ in §C.2, so $\frac{\lambda_s}{L_t} \leq \lambda_s\gamma_t$ holds.

$$\theta^* = MCP\left(\theta'_j, \frac{\lambda_s}{L_t}, \gamma_t\right) = \left(\frac{\gamma_t L_t}{\gamma_t L_t - 1}\right) S_{\frac{\lambda_s}{L_t}}(\theta'_j) = \left(\frac{\gamma_t L_t}{\gamma_t L_t - 1}\right)\text{sign}(\theta'_j)\left(|\theta'_j| - \frac{\lambda_s}{L_t}\right)$$

$$= \left(\frac{\gamma_t L_t}{\gamma_t L_t - 1}\right)\theta'_j - \text{sign}(\theta'_j)\frac{\lambda_s}{L_t}\left(\frac{\gamma_t L_t}{\gamma_t L_t - 1}\right)$$

Plugging this in we have that:

$$0 = -\hat{\theta}_j + \left(\frac{\gamma_t L_t}{\gamma_t L_t - 1}\right)\theta'_j - \text{sign}(\theta'_j)\frac{\lambda_s}{L_t}\left(\frac{\gamma_t L_t}{\gamma_t L_t - 1}\right) + \frac{1}{L_t}\partial P_{\gamma_t}\left(MCP\left(\theta'_j, \frac{\lambda_s}{L_t}, \gamma_t\right), \lambda_s\right)$$
$$+ \frac{\lambda_f}{L_t}\partial \left\|\theta'_j - \theta'_{j-1}\right\|_1 - \frac{\lambda_f}{L_t}\partial \left\|\theta'_{j+1} - \theta'_j\right\|_1.$$

Consider the term:

$$\frac{1}{L_t}\partial P_{\gamma_t}\left(MCP\left(\theta'_j, \frac{\lambda_s}{L_t}, \gamma_t\right), \lambda_s\right).$$

Since $|\theta'_j| \leq \lambda_s\gamma_t$ we can follow the steps in the case 3 of the proof for Lemma 1 to obtain:

$$\left|MCP\left(\theta'_j, \frac{\lambda_s}{L_t}, \gamma_t\right)\right| \leq \lambda_s\gamma_t.$$

Therefore:

$$\frac{1}{L_t}\partial P_{\gamma_t}\left(MCP\left(\theta'_j, \frac{\lambda_s}{L_t}, \gamma_t\right), \lambda_s\right) = \frac{1}{L_t}\left(\lambda_s \text{sign}\left(MCP\left(\theta'_j, \frac{\lambda_s}{L_t}, \gamma_t\right)\right) - \frac{MCP\left(\theta'_j, \frac{\lambda_s}{L_t}, \gamma_t\right)}{\gamma_t}\right).$$

The MCP thresholding function does not change sign so:

$$= \frac{1}{L_t}\left(\lambda_s \text{sign}(\theta'_j) - \frac{MCP\left(\theta'_j, \frac{\lambda_s}{L_t}, \gamma_t\right)}{\gamma_t}\right) = \frac{\lambda_s}{L_t}\text{sign}(\theta'_j) - \frac{1}{\gamma_t L_t}\left(\frac{\gamma_t L_t}{\gamma_t L_t - 1}\right)\text{sign}(\theta'_j)\left(|\theta'_j| - \frac{\lambda_s}{L_t}\right)$$

$$= \frac{\lambda_s}{L_t}\text{sign}(\theta'_j) - \left(\frac{1}{\gamma_t L_t - 1}\right)\left(\theta'_j - \frac{\lambda_s}{L_t}\text{sign}(\theta'_j)\right) = \frac{\lambda_s}{L_t}\text{sign}(\theta'_j) - \frac{\theta'_j}{\gamma_t L_t - 1} + \left(\frac{1}{\gamma_t L_t - 1}\right)\frac{\lambda_s}{L_t}\text{sign}(\theta'_j)$$

$$= -\frac{\theta_j'}{\gamma_t L_t - 1} + \frac{\lambda_s}{L_t}\text{sign}(\theta_j')\left(1 + \frac{1}{\gamma_t L_t - 1}\right) = -\left(\frac{1}{\gamma_t L_t - 1}\right)\theta_j' + \frac{\lambda_s}{L_t}\text{sign}(\theta_j')\left(\frac{\gamma_t L_t}{\gamma_t L_t - 1}\right).$$

Substitute this term into the subgradient equations to obtain:

$$0 = -\hat{\theta}_j + \left(\frac{\gamma_t L_t}{\gamma_t L_t - 1}\right)\theta_j' - \text{sign}(\theta_j')\frac{\lambda_s}{L_t}\left(\frac{\gamma_t L_t}{\gamma_t L_t - 1}\right) - \left(\frac{1}{\gamma_t L_t - 1}\right)\theta_j' + \frac{\lambda_s}{L_t}\text{sign}(\theta_j')\left(\frac{\gamma_t L_t}{\gamma_t L_t - 1}\right)$$
$$+\frac{\lambda_f}{L_t}\partial\left\|\theta_j' - \theta_{j-1}'\right\|_1 - \frac{\lambda_f}{L_t}\partial\left\|\theta_{j+1}' - \theta_j'\right\|_1$$

$$= -\hat{\theta}_j + \left(\frac{\gamma_t L_t}{\gamma_t L_t - 1}\right)\theta_j' - \left(\frac{1}{\gamma_t L_t - 1}\right)\theta_j' + \frac{\lambda_f}{L_t}\partial\left\|\theta_j' - \theta_{j-1}'\right\|_1 - \frac{\lambda_f}{L_t}\partial\left\|\theta_{j+1}' - \theta_j'\right\|_1$$

$$= -\hat{\theta}_j + \left(\frac{\gamma_t L_t}{\gamma_t L_t - 1} - \frac{1}{\gamma_t L_t - 1}\right)\theta_j' + \frac{\lambda_f}{L_t}\partial\left\|\theta_j' - \theta_{j-1}'\right\|_1 - \frac{\lambda_f}{L_t}\partial\left\|\theta_{j+1}' - \theta_j'\right\|_1$$

$$= -\hat{\theta}_j + \theta_j' + \frac{\lambda_f}{L_t}\partial\left\|\theta_j' - \theta_{j-1}'\right\|_1 - \frac{\lambda_f}{L_t}\partial\left\|\theta_{j+1}' - \theta_j'\right\|_1.$$

This holds because θ_j' is assumed to be the solution to problem 12 with $\lambda_s = 0$.

Case 3: $|\theta_j'| < \frac{\lambda_s}{L_t} \leq \lambda_s\gamma_t$

We have that:

$$\theta_j^* = MCP\left(\theta_j', \frac{\lambda_s}{L_t}, \gamma_t\right) = 0.$$

Therefore, the subgradient equations are:

$$0 = -\hat{\theta}_j + \frac{1}{L_t}\partial P_{\gamma_t}(0, \lambda_s) + \frac{\lambda_f}{L_t}\partial\left\|\theta_j' - \theta_{j-1}'\right\|_1 - \frac{\lambda_f}{L_t}\partial\left\|\theta_{j+1}' - \theta_j'\right\|_1.$$

We have that:

$$\frac{1}{L_t}\partial P_{\gamma_t}(0, \lambda_s) \in \left[-\frac{\lambda_s}{L_t}, \frac{\lambda_s}{L_t}\right].$$

So we can chose:

$$\frac{1}{L_t}\partial P_{\gamma_t}(0, \lambda_s) = \theta_j',$$

since $|\theta_j'| \leq \frac{\lambda_s}{L_t}$.

Therefore:

$$0 = -\hat{\theta}_j + \theta_j' + \frac{\lambda_f}{L_t}\partial\left\|\theta_j' - \theta_{j-1}'\right\|_1 - \frac{\lambda_f}{L_t}\partial\left\|\theta_{j+1}' - \theta_j'\right\|_1.$$

which holds because θ_j' is assumed to be the solution to problem 9 with $\lambda_s = 0$.

With these three cases satisfied, we show that MCP thresholding the optimal solution to the $\lambda_s = 0$ problem, $\theta^*(0, \lambda_f)$, returns a stationary point to problem 12 with $\lambda_s > 0$.

D Block Selection Derivations

In this section we present derivations on computing the direction vector for our greedy block selection algorithm.

D.1 ℓ_1-penalty only

We want to derive the closed form solution for:

$$d_j = \min_{s \in \lambda_s \partial \|w_j\|_1} |\nabla f_j(w) + s|.$$

We have that:

$$\lambda_s \partial \|w_j\|_1 \in \begin{cases} \lambda_s \text{sign}(w_j) & w_j \neq 0 \\ [-\lambda_s, \lambda_s] & w_j = 0. \end{cases}$$

Decompose this problem into cases.

Case 1: $w_j \neq 0$

$$d_j = \nabla f_j(w) + \lambda_s \text{sign}(w_j)$$

Case 2: $w_j = 0$

$$d_j = \min_{s \in [-\lambda_s, \lambda_s]} |\nabla f_j(w) + s|$$

If $|\nabla f_j(w)| > \lambda_s$:

$$d_j = |\nabla f_j(w)| - \lambda_s \text{sign}(\nabla f_j(w)).$$

If $|\nabla f_j(w)| \leq \lambda_s$:

$$d_j = 0.$$

So combined:

$$d_j = S_{\lambda_s}(\nabla f_j(w)),$$

where S is the soft thresholding operator.

Therefore direction vector d is defined by:

$$d_j = \begin{cases} d_j = \nabla f_j(w) + \lambda_s \text{sign}(w_j) & w_j \neq 0 \\ S_{\lambda_s}(\nabla f_j(w)) & w_j = 0. \end{cases}$$

D.2 MCP-penalty only

We want to derive the closed form solution for:

$$d_j = \min_{s \in \partial P_{\gamma_t}(w, \lambda_s)} |\nabla f_j(w) + s|,$$

where:

$$\partial P_{\gamma_t}(w_j, \lambda_s) \in \begin{cases} [-\lambda_s, \lambda_s] & |w_j| = 0 \\ \lambda_s \text{sign}(w_j) - \frac{w_j}{\gamma_t} & 0 < |w_j| \leq \lambda_s \gamma_t \\ 0 & |w_j| > \lambda_s \gamma_t. \end{cases}$$

Note that $\gamma_t = \frac{\gamma}{L_t}$ where L_t is the block that corresponds to index j.

Decompose this problem into cases.

Case 1: $w_j = 0$

$$d_j = \min_{s \in [-\lambda_s, \lambda_s]} |\nabla f_j(w) + s|$$

$$d_j = S_{\lambda_s}(\nabla f_j(w))$$

Case 2: $0 < |w_j| \leq \lambda_s \gamma_t$

$$d_j = \nabla f_j(w) + \lambda_s \text{sign}(w_j) - \frac{w_j}{\gamma_t}$$

Case 3: $|w_j| > \lambda_s \gamma_t$

$$d_j = \nabla f_j(w)$$

Direction vector d_j is fully defined elementwise by the 3 cases above.

D.3 ℓ_1-penalty with Fusion

We want to derive the closed form solution for:

$$d_j = \min_{s \in \lambda_s \partial \|w_j\|_1 + \partial g_j} |\nabla f_j(w) + s|,$$

where g is the fusion penalty.

Rewrite this as:

$$d_j = \min_{s_1 \in \lambda_s \partial \|w_j\|_1 \wedge \ s_2 \in \partial g_j} |\nabla f_j(w) + s_1 + s_2|. \tag{16}$$

Fusion penalty g is of the form:

$$g(w, \lambda_f) = \lambda_f \sum_{t=1}^{T} \|D_t w_t\|_1,$$

where w_t are the elements of w that correspond to block t and D_t is the fusion matrix of the form:

$$D_t = \begin{bmatrix} -1 & 1 & 0 & 0 & \dots & 0 \\ 0 & -1 & 1 & 0 & \dots & 0 \\ 0 & 0 & -1 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}. \tag{17}$$

One thing to note is that the fusion penalty on penalizes the differences in coefficients within a tree, and not across trees. We can re-express the penalty as:

$$g(w, \lambda_f) = \lambda_f \|Dw\|_1, \tag{18}$$

where $D \in \{-1, 0, 1\}^{(R-1)\times R}$ is a modified fusion matrix, that incorporates spacer rows of all 0's to avoid penalizing the differences in w across trees. Matrix D has the form:

$$D = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix},$$

where the spacer rows are placed between trees.

To compute the subgradient of the fusion penalty, we follow [4] for the derivation of the subgradients for generalized LASSO problems:

$$\partial g(w, \lambda_f) = \lambda_f D^\intercal \partial \left\| Dw \right\|_1 . \tag{19}$$

It is important to note the effect of the spacer rows on the subgradients of the modified fused lasso problem. Consider the case when there is a spacer row at index j'. This affects the subgradient at index j' and $j'+1$. We show why below, using the first 5 rows of D as an example, with a spacer row placed in row 3. The calculations below shows what happens to the subgradient around spacer row 3.

$$Dw = \begin{bmatrix} w_2 - w_1 \\ w_3 - w_2 \\ 0 \\ w_5 - w_4 \\ w_6 - w_5 \\ \vdots \end{bmatrix}$$

$$D^\intercal \partial \left\| Dw \right\|_1 = \begin{bmatrix} \vdots \\ \partial \left\| w_2 - w_1 \right\|_1 - \partial \left\| w_3 - w_2 \right\|_1 \\ \partial \left\| w_3 - w_2 \right\|_1 \\ -\partial \left\| w_5 - w_4 \right\|_1 \\ \partial \left\| w_5 - w_4 \right\|_1 - \partial \left\| w_6 - w_5 \right\|_1 \\ \vdots \end{bmatrix}$$

The matrix above shows indices 2,3,4,5 of the subgradient for the fusion penalty. We say that if j' is a spacer row:

$$\partial g(w_{j'}, \lambda_f) = \lambda_f \partial \left\| w_{j'} - w_{j'-1} \right\|_1 ,$$

and:

$$\partial g(w_{j'+1}, \lambda_f) = -\lambda_f \partial \left\| w_{j'+2} - w_{j'+1} \right\|_1 .$$

.

Finally for thoroughness consider two special cases, when j is the last and first index in w.

When $j = R$, we have that the modified fusion subgradient is:

$$\partial g(w_R, \lambda_f) = \lambda_f \partial \left\| w_R - w_{R-1} \right\|_1 ,$$

so this subgradient is the same as j' spacer row case.

When $j = 1$ the modified fusion subgradient is:

$$\partial g(w_1, \lambda_f) = -\lambda_f \partial \left\| w_2 - w_1 \right\|_1,$$

so this subgradient is the same as $j' + 1$ row following the spacer row case.

We have the following results for the fusion subgradients of our modified fused lasso problem:

Corollary 1. ***Modified Fused Lasso with Spacer Rows Subgradients***

Let δ_1 represent the set of indices for the spacer rows (and the last row) and let δ_2 represent the set of indices for the rows immediately after the spacer rows (and the first row).
If $j \notin \delta_1 \vee \delta_2$:

$$\partial g(w_j, \lambda_f) = \lambda_f \partial \left\| w_j - w_{j-1} \right\|_1 - \lambda_f \partial \left\| w_{j+1} - w_j \right\|_1. \tag{20}$$

If $j \in \delta_1$:

$$\partial g(w_j, \lambda_f) = \lambda_f \partial \left\| w_j - w_{j-1} \right\|_1. \tag{21}$$

If $j \in \delta_2$:

$$\partial g(w_{j+1}, \lambda_f) = -\lambda_f \partial \left\| w_{j+1} - w_j \right\|_1. \tag{22}$$

With this result, we can start to decompose problem 16 by cases. Let:

$$\alpha_j = w_j - w_{j-1},$$

$$\beta_j = w_{j+1} - w_j.$$

Case A: $j \notin \delta_1 \vee \delta_2$

In this case note that the subgradients depend on w_j, α_j and β_j and whether these values are 0. We can further decompose the problem into the following 8 cases.

Case A.1: $w_j = \alpha_j = \beta_j = 0$

$$\begin{aligned} d_j = \min_{s_1, s_2} \quad & |\nabla f_j(w) + s_1 + s_2|. \\ \text{s.t.} \quad & s_1 \in [-\lambda_s, \lambda_s], \\ & s_2 \in [-2\lambda_f, 2\lambda_f] \end{aligned}$$

$$d_j = S_{2\lambda_f + \lambda_s}(\nabla f_j(w))$$

Case A.2: $w_j = 0$, $\alpha_j = 0$, $\beta_j \neq 0$

$$\begin{aligned} d_j = \min_{s_1, s_2'} \quad & |\nabla f_j(w) - \lambda_f \text{sign}(\beta_j) + s_1 + s_2'|. \\ \text{s.t.} \quad & s_1 \in [-\lambda_s, \lambda_s], \\ & s_2' \in [\lambda_f, \lambda_f] \end{aligned}$$

$$d_j = S_{\lambda_f + \lambda_s}(\nabla f_j(w) - \lambda_f \text{sign}(\beta_j))$$

Case A.3: $w_j = 0$, $\alpha_j \neq 0$, $\beta_j = 0$

$$\begin{aligned} d_j = \min_{s_1, s_2'} \quad & |\nabla f_j(w) + \lambda_f \text{sign}(\alpha_j) + s_1 + s_2'|. \\ \text{s.t.} \quad & s_1 \in [-\lambda_s, \lambda_s], \\ & s_2' \in [\lambda_f, \lambda_f] \end{aligned}$$

$$d_j = S_{\lambda_f+\lambda_s}(\nabla f_j(w) + \lambda_f \text{sign}(\alpha_j))$$

Case A.4: $w_j \neq 0$, $\alpha_j = 0$, $\beta_j = 0$

$$\begin{aligned} d_j = \min_{s_2} \quad & |\nabla f_j(w) + \lambda_s \text{sign}(w_j) + s_2|. \\ \text{s.t.} \quad & s_2 \in [2\lambda_f, 2\lambda_f] \end{aligned}$$

$$d_j = S_{2\lambda_f+\lambda_s}(\nabla f_j(w) + \lambda_s \text{sign}(w_j))$$

Case A.5: $w_j \neq 0$, $\alpha_j \neq 0$, $\beta_j = 0$

$$\begin{aligned} d_j = \min_{s_2'} \quad & |\nabla f_j(w) + \lambda_s \text{sign}(w_j) + \lambda_f \text{sign}(\alpha_j) + s_2'|. \\ \text{s.t.} \quad & s_2' \in [\lambda_f, \lambda_f] \end{aligned}$$

$$d_j = S_{\lambda_f}(\nabla f_j(w) + \lambda_s \text{sign}(w_j) + \lambda_f \text{sign}(\alpha_j))$$

Case A.6: $w_j \neq 0$, $\alpha_j = 0$, $\beta_j \neq 0$

$$\begin{aligned} d_j = \min_{s_2'} \quad & |\nabla f_j(w) + \lambda_s \text{sign}(w_j) - \lambda_f \text{sign}(\beta_j) + s_2'|. \\ \text{s.t.} \quad & s_2' \in [\lambda_f, \lambda_f] \end{aligned}$$

$$d_j = S_{\lambda_f}(\nabla f_j(w) + \lambda_s \text{sign}(w_j) - \lambda_f \text{sign}(\beta_j))$$

Case A.7: $w_j = 0$, $\alpha_j \neq 0$, $\beta_j \neq 0$

$$\begin{aligned} d_j = \min_{s_1} \quad & |\nabla f_j(w) + \lambda_f \text{sign}(\alpha_j) - \lambda_f \text{sign}(\beta_j) + s_1|. \\ \text{s.t.} \quad & s_1 \in [\lambda_s, \lambda_s] \end{aligned}$$

$$d_j = S_{\lambda_s}(\nabla f_j(w) + \lambda_f \text{sign}(\alpha_j) - \lambda_f \text{sign}(\beta_j))$$

Case A.8: $w_j \neq 0$, $\alpha_j \neq 0$, $\beta_j \neq 0$

$$d_j = \nabla f_j(w) + \lambda_s \text{sign}(w_j) + \lambda_f \text{sign}(\alpha_j) - \lambda_f \text{sign}(\beta_j)$$

Case B $j \in \delta_1$

In this case note that the subgradients depend on w_j and α_j. We can further decompose the problem in the following 4 cases.

Case B.1 $w_j = \alpha_j = 0$

$$\begin{aligned} d_j = \min_{s_1, s_2} \quad & |\nabla f_j(w) + s_1 + s_2|. \\ \text{s.t.} \quad & s_1 \in [-\lambda_s, \lambda_s], \\ & s_2 \in [-\lambda_f, \lambda_f] \end{aligned}$$

$$d_j = S_{\lambda_f+\lambda_s}(\nabla f_j(w))$$

Case B.2 $w_j = 0$, $\alpha_j \neq 0$

$$\begin{aligned} d_j = \min_{s_1} \quad & |\nabla f_j(w) + s_1 + \lambda_f \text{sign}(\alpha_j)|. \\ \text{s.t.} \quad & s_1 \in [-\lambda_s, \lambda_s] \end{aligned}$$

$$d_j = S_{\lambda_s}(\nabla f_j(w) + \lambda_f \text{sign}(\alpha_j))$$

Case B.3 $w_j \neq 0$, $\alpha_j = 0$

$$\begin{aligned} d_j = \min_{s_2} \quad & |\nabla f_j(w) + s_2 + \lambda_s \text{sign}(w_j)|. \\ \text{s.t.} \quad & s_2 \in [-\lambda_f, \lambda_f] \end{aligned}$$

$$d_j = S_{\lambda_f}(\nabla f_j(w) + \lambda_s \text{sign}(w_j))$$

Case B.4 $w_j \neq 0$, $\alpha_j \neq 0$

$$d_j = \nabla f_j(w) + \lambda_s \text{sign}(w_j) + \lambda_f \text{sign}(\alpha_j)$$

Case C

In this case note that the subgradients depend on w_j and β_j. We can further decompose the problem in the following 4 cases.

Case C.1 $w_j = \beta_j = 0$

$$\begin{aligned} d_j = \min_{s_1, s_2} \quad & |\nabla f_j(w) + s_1 + s_2|. \\ \text{s.t.} \quad & s_1 \in [-\lambda_s, \lambda_s], \\ & s_2 \in [-\lambda_f, \lambda_f] \end{aligned}$$

$$d_j = S_{\lambda_f+\lambda_s}(\nabla f_j(w))$$

Case C.2 $w_j = 0$, $\beta_j \neq 0$

$$\begin{aligned} d_j = \min_{s_1} \quad & |\nabla f_j(w) + s_1 - \lambda_f \text{sign}(\beta_j)|. \\ \text{s.t.} \quad & s_1 \in [-\lambda_s, \lambda_s] \end{aligned}$$

$$d_j = S_{\lambda_s}(\nabla f_j(w) - \lambda_f \text{sign}(\beta_j))$$

Case C.3 $w_j \neq 0$, $\beta_j = 0$

$$\begin{aligned} d_j = \min_{s_2} \quad & |\nabla f_j(w) + s_2 + \lambda_s \text{sign}(w_j)|. \\ \text{s.t.} \quad & s_2 \in [-\lambda_f, \lambda_f] \end{aligned}$$

$$d_j = S_{\lambda_f}(\nabla f_j(w) + \lambda_s \text{sign}(w_j))$$

Case C.4 $w_j \neq 0$, $\beta_j \neq 0$

$$d_j = \nabla f_j(w) + \lambda_s \text{sign}(w_j) - \lambda_f \text{sign}(\beta_j)$$

These cases above give a closed form solution to find elements of direction vector, d_j.

D.4 MCP penalty with Fusion

We want to derive the closed form solution to:

$$d_j = \min_{s_1 \in \partial P_{\gamma_t}(w_j, \lambda_s) \wedge \ s_2 \in \partial g_j} |\nabla f_j(w) + s_1 + s_2|, \tag{23}$$

where t is the block corresponding to index j and $\gamma = \gamma_t L_t$. We have that:

$$\partial P_{\gamma_t}(w_j, \lambda_s) \in \begin{cases} [-\lambda_s, \lambda_s] & |w_j| = 0 \\ \lambda_s \text{sign}(w_j) - \frac{w_j}{\gamma_t} & 0 < |w_j| \leq \lambda_s \gamma_t \\ 0 & |w_j| > \lambda_s \gamma_t. \end{cases}$$

In addition, ∂g_j is defined in the section above.

We decompose problem 23 into cases. Following Corollary 1, let δ_1 represent the set of indices for the spacer rows (indices separating decision trees in the ensemble) and the last row, and let δ_2 represent the set of indices for the rows directly after the spacer rows and the first row. Also again given index j, let $\alpha_j = w_j - w_{j-1}$ and $\beta_j = w_{j+1} - w_j$.

The closed form solutions to find d_j are presented below. We omit the derivations since they follow closely from the previous section.

Case A: Let $j \notin \delta_1 \vee \delta_2$. We have from Corollary 1 that:

$$\partial g(w_j, \lambda_f) = \lambda_f \partial \left\|\alpha_j\right\|_1 - \lambda_f \partial \left\|\beta_j\right\|_1.$$

Case A.1: w_j = α_j = β_j = 0

$$d_j = S_{2\lambda_f + \lambda_s}(\nabla f_j(w))$$

Case A.2: $w_j = 0$, $\alpha_j = 0$, $\beta_j \neq 0$

$$d_j = S_{\lambda_f + \lambda_s}(\nabla f_j(w) - \lambda_f \text{sign}(\beta_j))$$

Case A.3: $w_j = 0$, $\alpha_j \neq 0$, $\beta_j = 0$

$$d_j = S_{\lambda_f + \lambda_s}(\nabla f_j(w) + \lambda_f \text{sign}(\alpha_j))$$

Case A.4: $w_j = 0$, $\alpha_j \neq 0$, $\beta_j \neq 0$

$$d_j = S_{\lambda_s}(\nabla f_j(w) + \lambda_f \text{sign}(\alpha_j) - \lambda_f \text{sign}(\beta_j))$$

Case A.5: $0 < |w_j| \leq \lambda_s \gamma_t$, $\alpha_j = 0$, $\beta_j = 0$

$$d_j = S_{2\lambda_f}\left(\nabla f_j(w) + \lambda_s \text{sign}(w_j) - \frac{w_j}{\gamma_t}\right)$$

Case A.6: $0 < |w_j| \leq \lambda_s \gamma_t$, $\alpha_j = 0$, $\beta_j \neq 0$

$$d_j = S_{\lambda_f}\left(\nabla f_j(w) + \lambda_s \text{sign}(w_j) - \frac{w_j}{\gamma_t} - \lambda_f \text{sign}(\beta_j)\right)$$

Case A.7: $0 < |w_j| \leq \lambda_s \gamma_t$, $\alpha_j \neq 0$, $\beta_j = 0$

$$d_j = S_{\lambda_f}\left(\nabla f_j(w) + \lambda_s \text{sign}(w_j) - \frac{w_j}{\gamma_t} + \lambda_f \text{sign}(\alpha_j)\right)$$

Case A.8: $0 < |w_j| \le \lambda_s \gamma_t$, $\alpha_j \ne 0$, $\beta_j \ne 0$

$$d_j = \nabla f_j(w) + \lambda_s \text{sign}(w_j) - \frac{w_j}{\gamma_t} + \lambda_f \text{sign}(\alpha_j) - \lambda_f \text{sign}(\beta_j)$$

Case A.9: $|w_j| > \lambda_s \gamma_t$, $\alpha_j = 0$, $\beta_j = 0$

$$d_j = S_{2\lambda_f}(\nabla f_j(w))$$

Case A.10: $|w_j| > \lambda_s \gamma_t$, $\alpha_j = 0$, $\beta_j \ne 0$

$$d_j = S_{\lambda_f}(\nabla f_j(w) - \lambda_f \text{sign}(\beta_j))$$

Case A.11: $|w_j| > \lambda_s \gamma_t$, $\alpha_j \ne 0$, $\beta_j = 0$

$$d_j = S_{\lambda_f}(\nabla f_j(w) + \lambda_f \text{sign}(\alpha_j))$$

Case A.12: $|w_j| > \lambda_s \gamma_t$, $\alpha_j \ne 0$, $\beta_j \ne 0$

$$d_j = \nabla f_j(w) + \lambda_f \text{sign}(\alpha_j) - \lambda_f \text{sign}(\beta_j)$$

Case B: $j \in \delta_1$

From Corollary 1 we have that:

$$\partial g(w_j, \lambda_f) = \lambda_f \partial \left\|\alpha_j\right\|_1 .$$

Case B.1: $w_j = \alpha_j = 0$

$$d_j = S_{\lambda_s + \lambda_f}(\nabla f_j(w))$$

Case B.2: $w_j = 0$, $\alpha_j \ne 0$

$$d_j = S_{\lambda_s}(\nabla f_j(w) + \lambda_f \text{sign}(\alpha_j))$$

Case B.3: $0 < |w_j| \le \lambda_s \gamma_t$, $\alpha_j = 0$

$$d_j = S_{\lambda_f}\left(\nabla f_j(w) + \lambda_s \text{sign}(w_j) - \frac{w_j}{\gamma_t}\right)$$

Case B.4: $0 < |w_j| \le \lambda_s \gamma_t$, $\alpha_j \ne 0$

$$d_j = \nabla f_j(w) + \lambda_s \text{sign}(w_j) - \frac{w_j}{\gamma_t} + \lambda_f \text{sign}(\alpha_j)$$

Case B.5: $|w_j| > 0$, $\alpha_j = 0$

$$d_j = S_{\lambda_f}(\nabla f_j(w))$$

Case B.6: $|w_j| > 0$, $\alpha_j \ne 0$

$$d_j = \nabla f_j(w) + \lambda_f \text{sign}(\alpha_j)$$

Case C: $j \in \delta_2$

From Corollary 1 we have that:

$$\partial g(w_j, \lambda_f) = -\lambda_f \partial \left\| \beta_j \right\|_1 .$$

Case C.1: $w_j = \beta_j = 0$

$$d_j = S_{\lambda_s + \lambda_f}(\nabla f_j(w))$$

Case C.2: $w_j = 0$, $\beta_j \neq 0$

$$d_j = S_{\lambda_s}(\nabla f_j(w) - \lambda_f \text{sign}(\beta_j))$$

Case C.3: $0 < |w_j| \leq \lambda_s \gamma_t$, $\beta_j = 0$

$$d_j = S_{\lambda_f}\left(\nabla f_j(w) + \lambda_s \text{sign}(w_j) - \frac{w_j}{\gamma_t}\right)$$

Case C.4: $0 < |w_j| \leq \lambda_s \gamma_t$, $\beta_j \neq 0$

$$d_j = \nabla f_j(w) + \lambda_s \text{sign}(w_j) - \frac{w_j}{\gamma_t} - \lambda_f \text{sign}(\beta_j)$$

Case C.5: $|w_j| > 0$, $\beta_j = 0$

$$d_j = S_{\lambda_f}(\nabla f_j(w))$$

Case C.6: $|w_j| > 0$, $\beta_j \neq 0$

$$d_j = \nabla f_j(w) - \lambda_f \text{sign}(\beta_j)$$

The cases above define the close form solutions to find d_j for all j.

E Additional Timing Experiments

We present the additional results from the timing experiment in §3.4 of the main text with the fusion penalty g included and h as the MCP penalty.

MCP only		
rows	variables	time
1316	2000	13.1 (0.1)
1316	8000	29.6 (0.3)
1316	25000	20.2 (0.2)
4338	2000	130.2 (3.0)
4338	8000	206.5 (2.1)
4338	25000	214.4 (5.3)
10955	2000	177.1 (8.7)
10955	8000	304.2 (1.5)
10955	25000	609.3 (6.3)
ℓ_1 with Fusion		
rows	variables	time
1316	2000	54.1 (0.3)
1316	8000	160.4 (4.0)
1316	25000	232.1 (2.3)
4338	2000	231.6 (2.1)
4338	8000	491.3 (2.0)
4338	25000	755.7 (13.8)
10955	2000	208.1 (1.4)
10955	8000	454.1 (3.5)
10955	25000	1048.4 (37.9)
MCP with Fusion		
rows	variables	time
1316	2000	35.3 (0.6)
1316	8000	196.3 (3.9)
1316	25000	298.5 (10.9)
4338	2000	342.1 (3.4)
4338	8000	422.1 (25.1)
4338	25000	798 (24.1)
10955	2000	235.1 (6.5)
10955	8000	444.2 (4.2)
10955	25000	1096.3 (7.2)

F OpenML Datasets

We present a table of the OpenML datasets used in our experiments.

Dataset Name	**Rows**	**Features**
humandevel	130	2
triazines	186	61
tecator	240	125
autoMpg	398	8
no2	500	8
boston	506	14
stock	950	10
socmob	1156	6
Moneyball	1232	15
balloon	2001	2
space_ga	3107	7
abalone	4177	9
Mercedes_Benz_Greener_Manufacturing	4209	377
mtp	4450	203
wine_quality	6497	12
wind	6574	15
kin8nm	8192	9
cpu_small	8192	13
puma32H	8192	33
bank32nh	8192	33
pol	15000	49
elevators	16599	19
houses	20640	9
house_16H	22784	17
2dplanes	40768	11

Table 1: OpenML Datasets used in the experiments along with metadata.

G Performance Experiment Results

Dataset	FIRE	GLRM	GLRM Debias	SIRUS	Full Model
Mercedes_Benz_Greener_Manufacturing	0.45	1.213	1.035	0.716	0.61
Moneyball	0.125	1.167	1.012	0.21	0.31
abalone	0.542	1.088	1.09	0.656	0.59
autoMpg	0.241	1.043	1.041	0.369	0.25
bank32nh	0.588	0.587	0.665	0.609	0.75
boston	0.247	1.365	1.16	0.292	0.31
cpu_small	0.349	0.583	0.748	0.232	0.12
elevators	0.437	0.989	0.985	0.654	0.53
house_16H	0.566	0.709	0.82	0.677	0.68
houses	0.415	0.473	0.609	0.555	0.53
kin8nm	0.559	0.625	0.705	0.664	0.66
mtp	0.687	1.062	1.024	0.765	0.73
no2	0.555	0.516	0.57	0.561	0.59
pol	0.175	0.152	0.176	0.278	0.61
puma32H	0.31	0.382	0.517	0.521	0.86
socmob	0.316	4.004	2.17	0.472	0.49
space_ga	0.498	0.481	0.53	0.514	0.54
stock	0.097	0.147	0.321	0.272	0.11
tecator	0.116	0.126	0.303	0.113	0.2
us_crime	0.424	1.036	1.014	0.422	0.48
wind	0.331	0.404	0.638	0.389	0.33
wine_quality	0.71	0.724	0.768	0.775	0.75

Table 2: Results of performance experiment comparing FIRE against competing algorithms.

Supplement References

[1] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[2] Jerome Friedman, Trevor Hastie, Holger Höfling, and Robert Tibshirani. Pathwise coordinate optimization. *The annals of applied statistics*, 1(2):302–332, 2007.

[3] Rahul Mazumder, Jerome H Friedman, and Trevor Hastie. Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, 106(495):1125–1138, 2011.

[4] Ryan J Tibshirani and Jonathan Taylor. The solution path of the generalized lasso. *The annals of statistics*, 39(3): 1335–1371, 2011.