Supplementary to "Learning Hierarchical Interactions at Scale: A Convex Optimization Approach"

Hussein Hazimeh and Rahul Mazumder

Massachusetts Institute of Technology

A Does the MIP enforce Strong Hierarchy?

We note that there can be a pathological case where the MIP does not satisfy SH. We will briefly discuss why this case corresponds to a zero probability event, when y is drawn from a continuous distribution (this happens for example, if $\epsilon \sim N(0, \sigma^2)$ with $\sigma > 0$). First, we assume that α_1 and α_2 are chosen large enough so that the number of selected variables (as counted by $\sum z_i + \sum_{i<j} z_{ij}$) is less than or equal to the number of samples n. We will also assume that the X_i's and $\tilde{X}_{ij}$'s corresponding to the nonzero z_i's and z_{ij}'s have full rank.

A pathological case can happen when the optimal solution of the MIP satisfies: $z^*_{ij} = 1$, $z^*_i = 1$, $\beta^*_i = 0$, and $\theta^*_{ij} \neq 0$ for some i and j (for example). However, in the latter case, β_i is a free variable, i.e., $|\beta_i| \leq M$ (since $z^*_i = 1$ and we assume M to be sufficiently large). Thus, $\beta^*_i = 0$ is equivalent to saying that a least squares solution on the support defined by the nonzero z^*_i's and z^*_{ij}'s, leads to a coordinate β^*_i which is exactly zero. We know that this is a zero probability event when y is drawn from a continuous distribution (assuming the number of variables is less than or equal to the number of samples and the corresponding columns have full rank).

B Proof of Lemma 1

Suppose the z_i's and z_{ij}'s are relaxed to $[0, 1]$ and fix some feasible solution β, θ. Let us (partially) minimize the objective function with respect to z, while keeping β, θ fixed, to obtain a solution z^*. Then, z^* must satisfy $z^*_i = \max\{\frac{|\beta_i|}{M}, \max_{k,j \in G_i} z_{kj}\}$ for every i (since this choice is the smallest feasible z_i). Moreover, $z^*_{ij} = \frac{|\theta_{ij}|}{M}$ for every $i < j$ (by the same reasoning). Substituting the optimal values z^*_i and z^*_{ij} leads to

$$\min_{\beta, \theta} \quad f(\beta, \theta) + \Omega(\beta, \theta) \quad \text{s.t. } \|\beta, \theta\|_\infty \leq M \tag{B.1}$$

where $\Omega(\beta,\theta) \stackrel{\text{def}}{=} \lambda_1 \sum_{i=1}^{p} \max\{|\beta_i|, \|\theta_{G_i}\|_\infty\} + \lambda_2 \|\theta\|_1$ and $\lambda_1 = \alpha_1/M, \lambda_2 = \alpha_2/M$. Finally, we note that the box constraint in the above formulation can be removed, and the resulting formulation is still a valid relaxation.

C Dual Reformulation of the Proximal Problem

In this section, we present a dual reformulation of the proximal problem, which will facilitate solving the problem. We note that Jenatton et al. (2011) and Mairal et al. (2011) dualize a proximal problem which involves sum of ℓ_∞ norms (their proximal problem thus includes ours as a special case). However, the dual we will present here uses $\Theta(p^2)$ less variables as it exploits the presence of the ℓ_1 norm in the objective. First, we introduce some necessary notation. We define the soft-thresholding operator as follows:

$$\mathcal{S}_\gamma(\tilde{v}) = \begin{cases} 0 & \text{if } |\tilde{v}| \le \gamma \\ (|\tilde{v}| - \gamma)\,\text{sign}(\tilde{v}) & \text{o.w.} \end{cases}$$

We associate every β^i with a dual variable $u^i \in \mathbb{R}$, and every θ_{ij} with two dual variables: $w_j^i \in \mathbb{R}$ and $w_i^j \in \mathbb{R}$. Moreover, we use the notation $w^i \in \mathbb{R}^{p-1}$ to refer to the vector composed of w_j^i for all j such that $j \neq i$.

Theorem C.1. *(Dual formulation) A dual of the proximal problem is:*

$$\max_{u,w} \quad q(u,w) \quad \textit{s.t.} \qquad \|(u^i, w^i)\|_1 \le 1 \quad \forall\, i \in [p] \tag{C.2}$$

where $q(u,w)$ is a continuously differentiable function with a Lipschitz continuous gradient, and

$$\nabla_{u^i} q(u,w) = \lambda_1 \Big(\tilde{\beta}_i - \frac{\lambda_1}{L} u^i \Big)$$

$$\nabla_{w_j^i} q(u,w) = \nabla_{w_i^j} q(u,w) = \lambda_1 \mathcal{S}_{\frac{\lambda_2}{L}} \big(\tilde{\theta}_{ij} - \frac{\lambda_1}{L}(w_j^i + w_i^j) \big).$$

If u^*, w^* is a solution to (C.2)*, then the solution to the proximal problem is:*

$$\beta_i^* = \frac{\nabla_{u^i} q(u^*, w^*)}{\lambda_1} \quad \textit{and} \quad \theta_{ij}^* = \frac{\nabla_{w_j^i} q(u^*, w^*)}{\lambda_1}. \tag{C.3}$$

Proof. Since the ℓ_1 norm is the dual of the l_∞ norm, we have:

$$\max\{|\beta_i|, \|\theta_{G_i}\|_\infty\} = \max_{u^i \in \mathbb{R},\ w^i \in \mathbb{R}^{p-1}} u^i \beta_i + \langle w^i, \theta_{G_i} \rangle \quad \text{s.t.} \|(u^i, w^i)\|_1 \le 1 \tag{C.4}$$

Plugging the above into the proximal problem and switching the order of the min and max (which

is justified by strong duality), we arrive to the dual of the proximal problem:

$$\begin{aligned}\max_{u,w}\min_{\beta,\theta}\frac{L}{2}\left\|\begin{bmatrix}\beta-\tilde{\beta}\\ \theta-\tilde{\theta}\end{bmatrix}\right\|_2^2+\lambda_1\sum_i(u^i\beta_i+\langle w^i,\theta_{G_i}\rangle)+\lambda_2\|\theta\|_1\\ \text{s.t. } \|(u^i,w^i)\|_1\leq 1,\forall i\in[p]\end{aligned}\tag{C.5}$$

Note that each dual variable u^i is a scalar which corresponds to the primal variable β^i. Similarly, the dual vector $w^i\in\mathbb{R}^{p-1}$ corresponds to θ_{G_i}. The term $\sum_i(u^i\beta_i+\langle w^i,\theta_{G_i}\rangle)$ in (C.5) can be written as $\sum_i u^i\beta_i+\sum_{i<j}\theta_{ij}(w^i_j+w^j_i)$, where w^i_j and w^j_i are the components of the vectors w^i and w^j, respectively, corresponding to θ_{ij}. Using this notation, we can rewrite Problem (C.5) as follows:

$$\begin{aligned}\max_{u,w}\min_{\beta,\theta}\sum_i h(\beta_i,u^i;\tilde{\beta}_i)+\sum_{i<j}g(\theta_{ij},w^i_j,w^j_i;\tilde{\theta}_{ij})\\ \text{s.t. } \|(u^i,w^i)\|_1\leq 1,\forall i\in[p]\end{aligned}\tag{C.6}$$

where

$$h(a,b;\tilde{a})\overset{\text{def}}{=}\frac{L}{2}(a-\tilde{a})^2+\lambda_1 ab\quad\text{and}\quad g(a,b,c;\tilde{a})\overset{\text{def}}{=}\frac{L}{2}(a-\tilde{a})^2+\lambda_1 a(b+c)+\lambda_2|a|$$

The optimal solution of the inner minimization in (C.6) is (uniquely) given by:

$$\begin{aligned}\beta_i^*&\overset{\text{def}}{=}\arg\min_{\beta_i}h(\beta_i,u^i;\tilde{\beta}_i)=\tilde{\beta}_i-\frac{\lambda_1}{L}u^i\\ \theta_{ij}^*&\overset{\text{def}}{=}\arg\min_{\theta_{ij}}g(\theta_{ij},w^i_j,w^j_i;\tilde{\theta}_{ij})=\mathcal{S}_{\frac{\lambda_2}{L}}\left(\tilde{\theta}_{ij}-\frac{\lambda_1}{L}(w^i_j+w^j_i)\right)\end{aligned}\tag{C.7}$$

Therefore, the dual problem can equivalently written as:

$$\max_{u,w}\underbrace{\sum_i h(\beta_i^*,u^i;\tilde{\beta}_i)+\sum_{i<j}g(\theta_{ij}^*,w^i_j,w^j_i;\tilde{\theta}_{ij})}_{q(u,w)}\quad\text{s.t. }\|(u^i,w^i)\|_1\leq 1\quad\forall\ i\tag{C.8}$$

Finally, since the solution β^*,θ^* is defined in (C.7) is unique, Danskin's theorem implies that the dual objective function $q(u,w)$ is continuously differentiable and that

$$\nabla_{u^i}q(u,w)=\lambda_1\beta_i^*\quad\text{and}\quad\nabla_{w^i_j}q(u,w)=\nabla_{w^j_i}q(u,w)=\lambda_1\theta_{ij}^*.\tag{C.9}$$

□

In problem (C.2), the separability of the feasible set across the (u^i,w^i)'s and the smoothness of $q(u,w)$ make the problem well-suited for the application of block coordinate ascent (BCA) (Bertsekas (2016); Tseng (2001)), which optimizes with respect to a single block at a time. When

updating a particular block in BCA, we perform inexact maximization by taking a step in the direction of the gradient of the block and projecting the resultant vector onto the feasible set, i.e., the ℓ_1 ball. We present the algorithm more formally below.

Algorithm 3: BCA for Solving (C.2)

- Initialize with u, w and take step size $\alpha_i, i \in [p]$.
- For $i \in [p]$ perform updates (till convergence):

$$\begin{bmatrix} u^i \\ w^i \end{bmatrix} \leftarrow \mathcal{P}_{\|.\|_1 \leq 1}\left(\begin{bmatrix} u^i \\ w^i \end{bmatrix} + \alpha_i \nabla_{u^i, w^i} q(u, w) \right),$$

where for a vector a, $\mathcal{P}_{\|.\|_1 \leq 1}(a)$ denotes projection of a onto the unit ℓ_1-ball.

The Lipschitz parameter of $\nabla_{u^i, w^i} q(u, w)$ is given by $L_i = p\frac{\lambda_1^2}{L}$ (this follows by observing that each component of $\nabla_{u^i, w^i} q(u, w)$ is a piece-wise linear function with a maximal slope of $\frac{\lambda_1^2}{L}$). Thus, by standard results on block coordinate descent (e.g., Beck and Tetruashvili (2013); Bertsekas (2016)), Algorithm 3 with step size $\alpha_i = \frac{1}{L_i}$ converges at a rate of $\mathcal{O}(\frac{1}{t})$ (where t is the iteration counter). We note that BCA has been applied to the dual of structured sparsity problems (e.g., Jenatton et al. (2011); Yan and Bien (2017))—however, the duals considered in the latter works are different.

D Proof of Theorem 1

We prove the theorem using the dual reformulation presented in Theorem (C.1) and the block coordinate ascent (BCA) algorithm presented in Section C. Suppose $\sum_{t \in G_i} \max\{|\tilde{\theta}_t| - \frac{\lambda_2}{L}, 0\} \leq \frac{\lambda_1}{L} - |\tilde{\beta}_i|$ is satisfied for some i. Let u, w be some feasible solution to Problem (C.2) (e.g., solution of all zeros). Now update u, w as follows:

$$u^i = \frac{L}{\lambda_1}\tilde{\beta}_i, \tag{D.10}$$

and for every $t \in G_i$, let j be the index in t different from i and set:

$$w_j^i = \max\left\{ \frac{L}{\lambda_1}|\tilde{\theta}_t| - \frac{\lambda_2}{\lambda_1}, 0 \right\} \operatorname{sign}(\tilde{\theta}_t) \quad \text{and} \quad w_i^j = 0 \tag{D.11}$$

It is easy to check that u, w is still feasible for Problem (C.2) after this update and that $\nabla_{u^i, w^i} q(u, w) = 0$ and $\nabla_{w_i^j} q(u, w) = 0$ for every j. Thus, BCA will never change u^i, w^i, or w_i^j (for any j) in subsequent iterations. Since BCA is guaranteed to converge to an optimal solution, we conclude that the values in (D.10) and (D.11) (which correspond to $\beta_i^*, \theta_{G_i}^* = 0$) are optimal.

For the case when $|\tilde{\theta}_{ij}| \leq \frac{\lambda_2}{L}$, if we set $w_j^i = w_i^j = 0$, then $\nabla_{w_j^i} q(u, w) = \nabla_{w_i^j} q(u, w) = 0$, so BCA will never change w_j^i or w_i^j, which leads to $\theta_{ij}^* = 0$.

E Proof of Theorem 2

By Theorem 1, we have $\beta^*_{\mathcal{V}^c} = 0, \quad \theta^*_{\mathcal{E}^c} = 0$. Plugging the latter into the proximal problem leads to:

$$\begin{bmatrix} \beta^*_{\mathcal{V}} \\ \theta^*_{\mathcal{E}} \end{bmatrix} = \arg\min_{\beta_{\mathcal{V}}, \theta_{\mathcal{E}}} \frac{L}{2} \left\| \begin{bmatrix} \beta_{\mathcal{V}} - \tilde{\beta}_{\mathcal{V}} \\ \theta_{\mathcal{E}} - \tilde{\theta}_{\mathcal{E}} \end{bmatrix} \right\|_2^2 + \Omega(\beta_{\mathcal{V}}, \theta_{\mathcal{E}}) \tag{E.12}$$

where $\Omega(\beta_{\mathcal{V}}, \theta_{\mathcal{E}}) = \lambda_1 \sum_{i \in \mathcal{V}} \max\{|\beta_i|, \|\theta_{\tilde{G}_i}\|_\infty\} + \lambda_2 \|\theta_{\mathcal{E}}\|_1$ and $\tilde{G}_i = G_i \setminus \mathcal{E}^c$. But by the definition of the connected components, the following holds:

$$\Omega(\beta_{\mathcal{V}}, \theta_{\mathcal{E}}) = \sum_{l \in [k]} \Big[\lambda_1 \sum_{i \in \mathcal{V}_l} \max\{|\beta_i|, \|\theta_{\tilde{G}_i}\|_\infty\} + \lambda_2 \|\theta_{\mathcal{E}_l}\|_1 \Big] = \sum_{l \in [k]} \Omega(\beta_{\mathcal{V}_l}, \theta_{\mathcal{E}_l}).$$

The above implies that the proximal problem is separable across the blocks $\beta_{\mathcal{V}_l}, \theta_{\mathcal{E}_l}$, which leads to the result of the theorem.

F Proof of Lemma 2

First, note that the full gradient $\nabla_{\beta,\theta} f(\hat{\beta}, \hat{\theta})$ is sufficient for constructing $\mathcal{G}$ (see steps 2 and 3 of Algorithm 2). The (i, j)'s in $\mathcal{T}^c$ whose $|\tilde{\theta}_{ij}| \leq \lambda_2 / L$ are not needed to construct $\mathcal{G}$ (this follows from the definitions of $\mathcal{V}$ and $\mathcal{E}$). For every $(i, j) \in \mathcal{T}^c$, we have $\hat{\theta}_{ij} = 0$, so the condition $|\tilde{\theta}_{ij}| \leq \lambda_2 / L$ is equivalent to $|\nabla_{\theta_{ij}} f(\hat{\beta}, \hat{\theta})| \leq \lambda_2$ (see the definition of $\tilde{\theta}_{ij}$ in step 2 of Algorithm 2). Thus, the (i, j)'s in $\mathcal{T}^c$ with $|\nabla_{\theta_{ij}} f(\hat{\beta}, \hat{\theta})| \leq \lambda_2$ are not needed to construct $\mathcal{G}$. The latter indices are exactly those in $\mathcal{S}^c$. Thus, the remaining parts of the gradient are: $\nabla_\beta f(\hat{\beta}, \hat{\theta})$, $\nabla_{\theta_{\mathcal{T}}} f(\hat{\beta}, \hat{\theta})$, and $\nabla_{\theta_{\mathcal{S}}} f(\hat{\beta}, \hat{\theta})$—these are sufficient to construct $\mathcal{G}$.

G Proof of Lemma 3

Let $(i, j) \in \mathcal{S}$. By the triangle inequality:

$$|\nabla_{\theta_{ij}} f(\hat{\beta}, \hat{\theta})| \leq |\nabla_{\theta_{ij}} f(\beta^w, \theta^w)| + |\nabla_{\theta_{ij}} f(\hat{\beta}, \hat{\theta}) - \nabla_{\theta_{ij}} f(\beta^w, \theta^w)| \tag{G.13}$$

Writing down the gradients explicitly and using Cauchy-Schwarz, we get:

$$\begin{aligned} |\nabla_{\theta_{ij}} f(\hat{\beta}, \hat{\theta}) - \nabla_{\theta_{ij}} f(\beta^w, \theta^w)| &= |\tilde{X}_{ij}^T (y - X\hat{\beta} - \tilde{X}\hat{\theta}) - \tilde{X}_{ij}^T (y - X\beta^w - \tilde{X}\theta^w)| \\ &\leq \|\tilde{X}_{ij}\|_2 \|(X\hat{\beta} + \tilde{X}\hat{\theta}) - (X\beta^w + \tilde{X}\theta^w)\|_2 \\ &\leq C\|\gamma\|_2 \end{aligned}$$

Plugging the upper bound above into (G.13), we get $|\nabla_{\theta_{ij}} f(\hat{\beta}, \hat{\theta})| \leq |\nabla_{\theta_{ij}} f(\beta^w, \theta^w)| + C\|\gamma\|_2$. Therefore, if $(i,j) \in \mathcal{S}$, i.e., $|\nabla_{\theta_{ij}} f(\hat{\beta}, \hat{\theta})| > \lambda_2$ then $|\nabla_{\theta_{ij}} f(\beta^w, \theta^w)| + C\|\gamma\|_2 > \lambda_2$, implying that $(i,j) \in \hat{\mathcal{S}}$.

H Results of Additional Experiments

H.1 Sizes of Connected Components

For the Riboflavin ($p = 4088$, $n = 71$) and Coepra ($p = 5786$, $n = 89$) datasets (discussed in the paper), we fit a regularization path with 100 solutions using Algorithm 2. In Table H.1, we report the maximum number of edges and vertices encountered across all the connected components and for all the 100 solutions in the path.

Table H.1: Maximum size of the connected components across a regularization path of 100 solutions.

Dataset	$\lambda_2 = 2\lambda_1$		$\lambda_2 = \lambda_1$	
	# Edges	# Vertices	# Edges	# Vertices
Ribo	350	149	1855	693
Coepra	227	86	400	103

H.2 Prediction Error

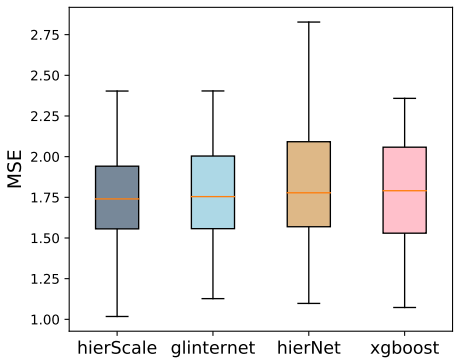

Figure H.1: MSE on the test set for synthetic data (Anti-Hierarchical truth).

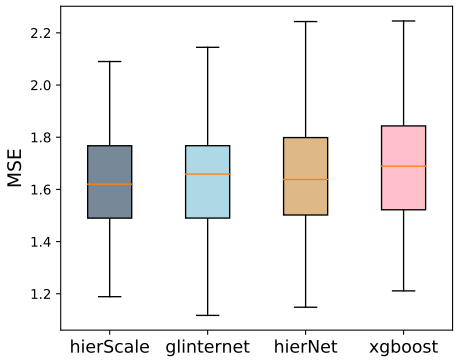

Figure H.2: MSE on the test set for synthetic data (Hierarchical truth).

References

Amir Beck and Luba Tetruashvili. On the convergence of block coordinate descent type methods. *SIAM Journal on Optimization*, 23(4):2037–2060, 2013. doi: 10.1137/120887679. URL `http://dx.doi.org/10.1137/120887679`.

D.P. Bertsekas. *Nonlinear Programming*. Athena scientific optimization and computation series. Athena Scientific, 2016. ISBN 9781886529052. URL `https://books.google.com/books?id=TwOujgEACAAJ`.

Rodolphe Jenatton, Julien Mairal, Guillaume Obozinski, and Francis Bach. Proximal methods for hierarchical sparse coding. *Journal of Machine Learning Research*, 12(Jul):2297–2334, 2011.

Julien Mairal, Rodolphe Jenatton, Guillaume Obozinski, and Francis Bach. Convex and network flow optimization for structured sparsity. *Journal of Machine Learning Research*, 12(Sep):2681–2720, 2011.

P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109(3):475–494, 2001. ISSN 1573-2878. doi: 10.1023/A:1017501703105. URL `http://dx.doi.org/10.1023/A:1017501703105`.

Xiaohan Yan and Jacob Bien. Hierarchical sparse modeling: A choice of two group lasso formulations. *Statistical Science*, 32(4):531–560, 2017.