Supplementary Material

In this document we give proofs for propositions (1) and (2) in the main paper. We use a slightly different notation for simplicity. We give a constructive proof for Proposition (2) that inherently implies Proposition (1). In the following section we give the necessary definitions and define the proximal operator for ℓ_∞^T-norm followed by proof in the next section.

1 Definitions

Let us consider a tree-structured set of groups of variables $\mathcal{G}$, which are subsets of $\{1, \dots, p\}$. The tree-structure definition follows [1], where two groups g and g' are either disjoint or one is included in the other.

Definition 1 (Tree-structured set of groups).
A set of groups $\mathcal{G} \triangleq \{g\}_{g\in\mathcal{G}}$ is said to be tree-structured in $\{1,\dots,p\}$, if $\bigcup_{g\in\mathcal{G}} g = \{1,\dots,p\}$ and if for all $g, h \in \mathcal{G}$, $g \cap h = \emptyset$, or $g \subseteq h$, or $h \subseteq g$. We also define for each group g,

- *the set of variables $root(g) \subseteq g$ is such that $i \in root(g)$ is not in g' for all group $g' \subseteq g$;*
- *the set of groups $children(g)$ is the set of groups g' such that $g' \subseteq g$.*

We are now interested in the following optimization problem

$$\min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{u}-\mathbf{w}\|_2^2 + \lambda \sum_{g\in\mathcal{G}} \|\mathbf{w}_g\|_\infty. \tag{1}$$

Following [1], it can be solved by Algorithm 1 where Π_λ is the Euclidean projection on the ℓ_1-ball of radius λ.

Lemma 1 (Equivalent Views of the ℓ_∞-proximal Operator).
Let us consider the proximal operator $Prox_\lambda^g$:

$$Prox_\lambda^g : \mathbf{u} \mapsto \arg\min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{u}-\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}_g\|_\infty.$$

Then,

$$[Prox_\lambda^g(\mathbf{u})]_g = \mathbf{u}_g - \Pi_\lambda(\mathbf{u}_g), \tag{2}$$

Algorithm 1 Computation of the Proximal Operator.

Inputs: $\mathbf{u} \in \mathbb{R}^p$ and an ordered tree-structured set of groups $\mathcal{G}$ with root g_0.
Initialization: $\mathbf{w} \leftarrow \mathbf{u}$;
Call `recursiveProx`(g_0);
Return $\mathbf{w}$.

Procedure `recursiveProx`(g)

for $h \in \text{child}(g)$ **do**
  Call `recursiveProx`(h);
end for
$\mathbf{w}_g \leftarrow \mathbf{w}_g - \Pi_\lambda(\mathbf{w}_g)$.

and there exists $\tau \geq 0$ *such that for all* $j \in g$,

$$[Prox_\lambda^g(\mathbf{u})]_j = \text{sign}(\mathbf{u}_j)\min(|\mathbf{u}_j|, \tau) \quad \textit{and} \tag{3}$$

$$\left\{ \|\Pi_\lambda(\mathbf{u}_g)\|_1 = \sum_{j \in g} \max(|\mathbf{u}_j| - \tau, 0) = \lambda \quad \textit{or} \quad \tau = 0 \right\}. \tag{4}$$

Proof. The proof of Eq. (2) can be found in [1]. The proof of Eq. (4) consists of noticing that the projection on the ℓ_1-ball is obtained by a soft-thresholding operator [1]. In other words, there exists $\tau \geq 0$ such that $[\Pi_\lambda(\mathbf{u})]_j = \text{sign}(\mathbf{u}_j)\max(|\mathbf{u}_j| - \tau, 0)$ for all j in g. We notice that by definition of the Euclidean projection, either $\|\Pi_\lambda(\mathbf{u}_g)\|_1 < \lambda$ and $\Pi_\lambda(\mathbf{u}_g) = \mathbf{u}_g$ (meaning $\tau = 0$), or $\|\Pi_\lambda(\mathbf{u}_g)\|_1 = \lambda$. This yields (4). □

By using the definition of Prox_λ^g, we see that Algorithm 1 in fact performs a composition of proximal operators. Suppose that the groups in $\mathcal{G} = \{g_1, \ldots, g_k\}$ are ordered according to depth-first search order, we have

$$\text{Prox}_{\lambda\Omega} = \text{Prox}^{g_k} \circ \ldots \circ \text{Prox}^{g_1},$$

where Ω is the tree-structured penalty $\Omega(\mathbf{w}) = \sum_{g \in \mathcal{G}} \|\mathbf{w}_g\|_\infty$, and $\circ$ is a composition operator.

We now have the following (Proposition 2 of main paper) to compose proximal step over constant value non-branching paths or nested groups. We prove this by showing that in consecutive projections the τ in 3 can only be smaller than the previous one forcing the values along a non-branching path to be equal.

Lemma 2 (**Composition Lemma Along Nested Groups).**
Assume that for all groups g *in* $\mathcal{G}$, *root*(g) *is a singleton* $\{r(g)\}$. *Consider a particular group* g *with a single child* g', *such that* $\mathbf{u}_{r(g)} = \mathbf{u}_{r(g')}$. *Then,*

$$\left(Prox_\lambda^g \circ Prox_\lambda^{g'}\right)(\mathbf{u}) = Prox_{2\lambda}^g(\mathbf{u}).$$

Proof. Without loss of generality, let us assume that all the entries of $\mathbf{u}$ are non-negative. Indeed, it is sufficient to store beforehand the signs of that vector, compute the proximal operator of the vector with nonnegative entries, and assign the stored signs to the result [1]. We also have

$$\left[\left(\text{Prox}_\lambda^g \circ \text{Prox}_\lambda^{g'}\right)(\mathbf{u})\right]_j = [\text{Prox}_{2\lambda}^g(\mathbf{u})]_j = \mathbf{u}_j \quad \text{for all} \quad j \notin g,$$

since all the proximal operators only affect the variables in g and g'. Let us now define $\mathbf{v} \triangleq \text{Prox}_\lambda^{g'}(\mathbf{u})$, $\mathbf{w}^\star \triangleq \text{Prox}_\lambda^g(\mathbf{v})$

Consider τ' defined in Lemma 1, such that $\mathbf{v}_{g'} = \min(\mathbf{u}_{g'}, \tau')$, and τ such that $\mathbf{w}_g^\star = \min(\mathbf{v}_g, \tau)$.

First step: $\tau \leq \tau'$:
Let us proceed by contradiction and assume that $\tau' < \tau$. Then, we have $\mathbf{v}_{g'} \leq \tau$ and thus, Eq. (4) applied to the group g gives us that $\mathbf{u}_{r(g)} - \tau = \mathbf{v}_{r(g)} - \tau = \lambda$ since $\tau \neq 0$ and $g = g' \cup \{r(g)\}$. Note also that $\mathbf{u}_{r(g')} - \tau' \leq \|\Pi_\lambda(\mathbf{u}_{g'})\|_1 \leq \lambda$ according to Eq. (4) applied to the group g'. Since $\mathbf{u}_{r(g')} = \mathbf{u}_{r(g)}$, we have $\mathbf{u}_{r(g')} - \tau' \leq \mathbf{u}_{r(g)} - \tau$, and $\tau \leq \tau'$, which is a contradiction.
End of the proof:
By using Eq. (4), and using the fact that $\tau \leq \tau'$, we now have a closed form solution for $\mathbf{w}_g^\star$:

$$\mathbf{w}_g^\star = \min(\mathbf{u}_g, \tau).$$

We now consider two cases

- if $\tau = 0$, we have $\mathbf{w}_g^\star = 0$, and thus $\mathbf{v}_g = \Pi_\lambda(\mathbf{v}_g)$, meaning that $\|\mathbf{v}_g\|_1 \leq \lambda$. Thus, $\|\mathbf{u}_g\|_1 = \|\mathbf{v}_g\|_1 + \|\mathbf{u}_{g'} - \mathbf{v}_{g'}\|_1 \leq \lambda + \|\Pi_\lambda(\mathbf{u}_{g'})\|_1 \leq 2\lambda$. Thus, $[\text{Prox}_{2\lambda}^g(\mathbf{u})]_g = 0 = \mathbf{w}_g^\star$;

- if $\tau > 0$, we define the quantity $\mathbf{z}_g = \mathbf{u}_g - \mathbf{w}_g^\star = \max(\mathbf{u}_g - \tau, 0)$, which has the form of an orthogonal projection of $\mathbf{u}_g$ onto the ℓ_1-ball of some radius λ' (see [1]). It remains to compute $\|\mathbf{z}_g\|_1$ to know the radius of λ'. We have

  $$\|\mathbf{z}_g\|_1 = \|\mathbf{u}_g - \mathbf{w}_g^\star\|_1 = \|\mathbf{u}_g - \mathbf{v}_g + \mathbf{v}_g - \mathbf{w}_g^\star\|_1 = \|\mathbf{u}_{g'} - \mathbf{v}_{g'}\|_1 + \|\mathbf{v}_g - \mathbf{w}_g^\star\|_1 = 2\lambda,$$

  where we apply again Eq. (4). Thus, $\mathbf{z}_g = \Pi_{2\lambda}(\mathbf{u}_g)$ and $\mathbf{w}_g^\star = \text{Prox}_{2\lambda}^g(\mathbf{u})]_g$ by using Eq. (2).

□

This proof can be put together for paths with more than two nested groups to inductively construct single-step proximal projections for longer paths.

It is easy to see from this definition 4 that all entries with the same value $u_j = \delta \forall j$ will continue to share a value after applying the proximal operator $\min(\delta, \tau)$. We see from 2 that all entries at nested groups will be projected to the same value. This in fact turns out to be a single projection with the λ scaled appropriately. These two put together we have the property that constant value non-branching paths are preserved.

References

[1] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for hierarchical sparse coding. *Journal of Machine Learning Research*, 12:2297–2334, 2011.