A. Proof of Theorem 1

We will first prove a particular case of Theorem 1, the single-output case ($p = 1$).

Proposition 1. *Let $h_{V,W}(x) = V^T\sigma(Wx) : \mathbb{R}^m \to \mathbb{R}$ be a neural network where $V \in \mathbb{R}^{n\times 1}$ and $W \in \mathbb{R}^{n\times m}$. Suppose that that the derivative of the activation is globally bounded between zero and one. Its Lipschitz constant with respect to the ℓ_∞ norm (for the input space) and the ℓ_1-norm (for the output space) is bounded as follows:*

$$L_{V,W} \leq \sum_{i=1}^{n}\sum_{j=1}^{m} |W_{i,j}V_{i,1}| \leq \|V\|_1\|W\|_\infty \tag{16}$$

First, note that because the output space is $\mathbb{R}$, the ℓ_1-norm is just the absolute value of the output. In this case the Lipschitz constant of the single-output function h is equal to the supremum of the ℓ_1-norm of its gradient, over its domain (c.f., Latorre et al. (2020, Theorem 1)).

Proof.

$$\begin{aligned}
L_{V,W} &= \sup_x \|\nabla h_{V,W}(x)\|_1 \\
&= \sup_x \sup_{\|t\|_\infty\leq 1} t^T\nabla h_{V,W}(x) \\
&= \sup_x \sup_{\|t\|_\infty\leq 1} t^T W^T\sigma'(Wx)V \\
&\leq \sup_{0\leq s\leq 1} \sup_{\|t\|_\infty\leq 1} t^T W^T \operatorname{Diag}(s)V \\
&= \sup_{0\leq s\leq 1} \sup_{\|t\|_\infty\leq 1} \sum_{i=1}^{n}\sum_{j=1}^{m} t_i(W^T\operatorname{Diag}(V))_{i,j}s_j \\
&\leq \sum_{i=1}^{n}\sum_{j=1}^{m} \sup_{0\leq s_j\leq 1} \sup_{-1\leq t_i\leq 1} t_i(W^T\operatorname{Diag}(V))_{i,j}s_j \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} |W^T\operatorname{Diag}(V)|_{i,j} = \sum_{i=1}^{n}\sum_{j=1}^{m} |W_{i,j}V_{i,1}|
\end{aligned}$$

This shows the first inequality in (16). We now show the second inequality. Denote the i-th row of the matrix W as w_i:

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{m} |W_{i,j}V_{i,1}| &= \sum_{i=1}^{n} |V_{i,1}| \sum_{j=1}^{m} |W_{i,j}| \\
&= \sum_{i=1}^{n} |V_{i,1}| \|w_i\|_1 \\
&\leq \sum_{i=1}^{n} |V_{i,1}| \max_{j=1,\ldots,m} \|w_j\|_1 \\
&= \sum_{i=1}^{n} |V_{i,1}| \|W\|_\infty \\
&= \|V\|_1\|W\|_\infty
\end{aligned}$$

In the fourth line we have used the fact that the ℓ_∞ operator norm of a matrix is equal to the maximum ℓ_1-norm of the rows.

□

Proof of Theorem 1. We now proceed with the general case where $V \in \mathbb{R}^{n\times p}$, $W \in \mathbb{R}^{n\times m}$ and $h_{V,W}(x) = V^T\sigma(Wx)$.

Proof. Denote the columns of V, in order, as $V_1, \ldots, V_p$. Using Proposition 1 we have

$$\begin{aligned}\|V^T\sigma(Wx) - V^T\sigma(Wy)\|_1 &= \sum_{k=1}^{p} |V_k^T\sigma(Wx) - V_k^T\sigma(Wy)| \\ &\leq \sum_{k=1}^{p}\sum_{i=1}^{n}\sum_{j=1}^{m} |W_{i,j}V_{i,k}|\|x-u\|_\infty \\ &\leq \sum_{k=1}^{p} \|V_k\|_1\|W\|_\infty\|x-y\|_\infty \\ &= \|V^T\|_{\infty,1}\|W\|_\infty\|x-y\|_\infty\end{aligned}$$

where in the fourth line we have used the fact that the (ℓ_∞, ℓ_1) operator norm of a matrix V^T is equal to the sum of the ℓ_1 norm of its rows i.e., the columns of V. This shows that $L_{V,W} \leq \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{p}|W_{i,j}V_{i,k}| \leq \|V^T\|_{\infty,1}\|W\|_\infty$

□

B. Proof of Theorem 2

In this section we prove the theoretical guarantees stated in Theorem 2 of the prox-grad method described by Algorithm 1. The first and second parts of Theorem 2 follow immediately from the results establish by (Bolte et al., 2013). Part two in Theorem 2 states that Algorithm 1 is globally convergent under the celebrated Kurdyka–Lojasiewicz (KL) property (Attouch et al., 2010). The broad classes of semi-algebraic and subanalytic functions, widely used in optimization, satisfy the KL property (see e.g. (Bolte et al., 2013, Section 5)), and in particular, most convex functions encountered in finite dimensional applications satisfy it (see (Bolte et al., 2013, Section 5.1)). We refer the reader to the works (Attouch et al., 2010; 2011; Bolte et al., 2013), in particular to (Bolte et al., 2013, Sections 3.2-3.5) for additional information and results.

For Part three we require the sufficient decrease property stated next.

Lemma 11 (Sufficient decrease property (Bolte et al., 2013, Lemma 2))**.** *Let $\Psi : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function with gradient assumed L_Ψ-Lipschitz continuous, and let $\sigma : \mathbb{R}^n \to (-\infty, \infty]$ be a proper l.s.c function satisfying that $\inf\sigma > -\infty$. Fix any $t \in (0, 1/L_\Psi)$. Then, for any $\mathbf{u} \in \operatorname{dom}\sigma$ and any $\mathbf{u}^+ \in \mathbb{R}^n$ defined by*

$$\mathbf{u}^+ \in \operatorname{prox}_{\sigma t}\left(\mathbf{u} - t\nabla\Psi(\mathbf{u})\right)$$

we have

$$\Psi(\mathbf{u}) + \sigma(\mathbf{u}) - \Psi(\mathbf{u}^+) - \sigma(\mathbf{u}) \geq \frac{1 - tL_\Psi}{2t}\|\mathbf{u}^+ - \mathbf{u}\|^2.$$

Proof of Theorem 2. The first and second parts follow from the results established by (Bolte et al., 2013). We will now prove the third part. By Lemma 11 we have that

$$\mathcal{F}(z^k) - \mathcal{F}(z^{k+1}) = f(z^k) + \lambda g(z^k) - f(z^{k+1}) - \lambda g(z^{k+1})) \geq \frac{1 - L\eta_k}{2\eta_k}\|z^{k+1} - z^k\|^2. \tag{17}$$

Hence $\{f(z^k) + \lambda g(z^k)\}_{k\geq 0}$ is a non-increasing sequence that strictly decreasing unless a critical point is obtained in a finite number of steps. By summing (17) over $k = 0, 1, \ldots, K$ and using the fact that $\{f(z^k) + \lambda g(z^k)\}_{k\geq 0}$ is non-increasing and is bounded below by $\mathcal{F}_*$, we obtain that

$$\begin{aligned}\mathcal{F}(z^0) - \mathcal{F}_* &\geq \sum_{k=0}^{K} \frac{1 - L\eta_k}{2\eta_k}\|z^{k+1} - z^k\|^2 \\ &\geq \frac{1}{2}(c - L)K \min_{k=0,\ldots,K}\|z^{k+1} - z^k\|_2^2.\end{aligned}$$

Consequently,

$$\min_{k=0,\ldots,K}\|z^{k+1} - z^k\|_2 \leq \sqrt{\frac{2(\mathcal{F}(z^0) - \mathcal{F}_*)}{(c - L)K}}.$$

□

C. Single output proximal map computation

This section provides the theoretical background and the required intermediate results to prove Theorem 3.

C.1. Moving to an Equivalent Easier Problem

We are interested in minimizing the nonconvex twice continuously differentiable function

$$\min_{v,w\in\mathbb{R}\times\mathbb{R}^m} \frac{1}{2}(v-x)^2 + \frac{1}{2}\sum_{j=1}^m (w_j - y_j)^2 + \lambda|v|\sum_{j=1}^m |w_j|. \tag{18}$$

The signs of the elements of the decision variables in (18) are determined by the signs of (x, y), and consequently, the problem in (18) is equivalent to problem (19); this is (partly) formulated by Lemma 12.

$$\min_{v,w\in\mathbb{R}_+\times\mathbb{R}_+^m} h_\lambda(v, w; x, y) \equiv \frac{1}{2}(v-|x|)^2 + \frac{1}{2}\sum_{j=1}^m (w_j - |y_j|)^2 + \lambda v\sum_{j=1}^m w_j. \tag{19}$$

Lemma 12. *Let $(v^*, w^*) \in \mathbb{R}_+ \times \mathbb{R}_+^n$ be an optimal solution of problem (19). Then $(\mathit{sign}(x)\cdot v^*, \mathit{sign}(y)\circ w^*)$ is an optimal solution of problem (18).*

Proof. We have that

$$\begin{aligned}
\tilde{h}_\lambda(v, w; x, y) &\equiv \frac{1}{2}(v-x)^2 + \frac{1}{2}\sum_{j=1}^m (w_j - y_j)^2 + \lambda|v|\sum_{j=1}^m |w_j| \\
&= \frac{1}{2}(\mathrm{sign}(x)v - |x|)^2 + \frac{1}{2}\sum_{j=1}^m (\mathrm{sign}(y_j)w_j - |y_j|)^2 + \lambda|v|\sum_{j=1}^m |w_j| \\
&\geq \frac{1}{2}(|v| - |x|)^2 + \frac{1}{2}\sum_{j=1}^m (|w_j| - |y_j|)^2 + \lambda v\sum_{j=1}^m w_j \\
&\geq h_\lambda(v^*, w^*; |x|, |y|),
\end{aligned}$$

where the last inequality follows from the fact that (v^*, w^*) is an optimal solution of (19). Since equality with the lower bound is attained by setting $(v, w) = (\mathrm{sign}(x)\cdot v^*, \mathrm{sign}(y)\circ w^*)$, we conclude that $(\mathrm{sign}(x)\cdot v^*, \mathrm{sign}(y)\circ w^*)$ is an optimal solution of (18). □

To summarize, we have established that, finding an optimal solution to (19) and then changing signs accordingly, yields an optimal solution to (18). We will now focus on obtaining an optimal solution for (18).

C.2. Solving the Prox Problem

First we note that problem (19) is well-posed.

Lemma 13 (Well-posedness of (19))**.** *For any $\lambda \geq 0$ and any $(x, y) \in \mathbb{R}\times\mathbb{R}^m$, the problem (19) has a global optimal solution.*

Proof. The claim follows from the fact that the objective function is coercive, cf. (Beck, 2014, Thm. 2.32). □

In light of Lemma 13, and due to the fact that in (19) we minimize a continuously differentiable function over a closed convex set, the set of optimal solutions of (19) is a nonempty subset of the set of stationary points of (19). These satisfy the following conditions (cf. (Beck, 2014, Ch. 9.1)).

Lemma 14 (Stationarity conditions)**.** *Let $(v^*, w^*) \in \mathbb{R}_+ \times \mathbb{R}_+^m$ be an optimal solution of (19) for a given $(x, y) \in \mathbb{R} \times \mathbb{R}^m$. Then*

$$w_j^* = \max\{0, |y_j| - \lambda v^*\} \quad \textit{for any } j = 1, 2, \ldots, m,$$
$$v^* = \max\left\{0, |x| - \lambda \sum_{j=1}^{m} w_j^*\right\}.$$

Proof. The stationarity (first-order) conditions of (19) (cf. (Beck, 2014, Ch. 9.1)) state that

$$\frac{\partial h_\lambda}{\partial v}(v^*, w^*; x, y) \begin{cases} = 0, & v^* > 0, \\ \geq 0, & v^* = 0, \end{cases} \quad \frac{\partial h_\lambda}{\partial w_j}(v^*, w^*; x, y) \begin{cases} = 0, & w_j^* > 0, \\ \geq 0, & w_j^* = 0, \end{cases}$$

which translates to

$$v^* - |x| + \lambda \sum_{j=1}^{m} w_j^* \begin{cases} = 0, & v^* > 0, \\ \geq 0, & v^* = 0, \end{cases} \quad w_j^* - |y_j| + \lambda v^* \begin{cases} = 0, & w_j^* > 0, \\ \geq 0, & w_j^* = 0, \end{cases}$$

and the required follows. □

The stationarity conditions given in Lemma 14 imply a solution form that we exploit in Algorithm 2; this is described by Corollary 3, where we use the convention that $\sum_{j=1}^{0} a_j \equiv 0$ for any $\{a_j\} \subseteq \mathbb{R}$.

Corollary 3. *Let $(v^*, w^*) \in \mathbb{R}_+ \times \mathbb{R}_+^m$ be an optimal solution of (19) for a given $(x, y) \in \mathbb{R} \times \mathbb{R}^m$.*

1. *The vector w^* satisfies that for any $j, l \in \{1, 2, \ldots, m\}$ it holds that $w_j^* \geq w_l^*$ only if $|y_j| \geq |y_l|$.*
2. *If $v^* = 0$, then $w^* = y$.*
3. *If $v^* > 0$, and $s = |\{j : w_j^* > 0\}|$, then we have that*

$$v^* = \frac{1}{1 - s\lambda^2}\left(|x| - \lambda \sum_{j=1}^{s} |\bar{y}_j|\right), \tag{20}$$

where $\bar{y}$ is the sorted vector of y in descending magnitude order.

Proof. The first part follows trivially from the stationarity conditions on w^* given in Lemma 14. The second part also follows trivially from the problem definition.

From the first part and the conditions in Lemma 14 we have that $\sum_{j=1}^{m} w_j^* = \sum_{j=1}^{s} |\bar{y}_j| - \lambda s v^*$. Plugging the latter to the stationarity condition on v^* (given in Lemma 14) then implies the required. □

In our analysis, we strictly distinguish between the trivial solution $(v^*, w^*) = (0, y)$, and the non-trivial solution in which $v^* > 0$. A practical point-of-view suggests that if $v^* = 0$, then the corresponding succeeding weights should also be zero, even though the optimality conditions imply otherwise. However, to avoid hindering the training process, this observation is considered only in the end of the training.

Recall that our analysis so-far implies that the magnitude order of y determines the order magnitude of w, effectively implying on set of non-zero entries in w (cf. Remark 5). For clarity of indices, and without loss of generality, we assume throughout this section that the vector y is already sorted in decreasing order of magnitude, that is $y \equiv \bar{y}$. We will use, without confusion, both notation to describe the same entity in order to maintain coherence with our procedures and results.

Denote

$$v^{(s)} = \frac{1}{1 - s\lambda^2}\left(|x| - \lambda \sum_{j=1}^{s} |y_j|\right)$$
$$w_j^{(s)} = |y_j| - \lambda v^{(s)} \text{ for } j = 1, 2, \ldots, s, \text{ and } w_j^{(s)} = 0 \text{ otherwise.} \tag{21}$$

Lemma 5 which states the monotonicity property

$$h_\lambda(v^{(s)}, w^{(s)}; x, y) < h_\lambda(v^{(s-1)}, w^{(s-1)}; x, y).$$

is proved next.

Proof of Lemma 5. Recall that $h_\lambda(v, w; x, y) := \frac{1}{2}(v - |x|)^2 + \frac{1}{2}\sum_{j=1}^m (w_j - |y_j|)^2 + \lambda v \sum_{j=1}^m w_j$. By plugging $w^{(s)}$ defined in (21) to h_λ we obtain that

$$\begin{aligned} h_\lambda(v^{(s)}, w^{(s)}; x, y) &= \frac{1}{2}(v^{(s)} - |x|)^2 + \frac{1}{2}\sum_{i=1}^s (|\bar{y}_i| - (|\bar{y}_i| - \lambda v^{(s)}))^2 + \frac{1}{2}\sum_{i=s+1}^m |\bar{y}_i|^2 + \lambda v^{(s)} \sum_{i=1}^s (|\bar{y}_i| - \lambda v^{(s)}) \\ &= \frac{1}{2}(v^{(s)} - |x|)^2 + \frac{\lambda^2}{2} s (v^{(s)})^2 + \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\sum_{i=1}^s |\bar{y}_i|^2 + \lambda v^{(s)} \sum_{i=1}^s |\bar{y}_i| - \lambda^2 s (v^{(s)})^2. \end{aligned}$$

Consequently, plugging $v^{(s)}$, defined in (21), yields

$$\begin{aligned} h_\lambda(v^{(s)}, w^{(s)}; x, y) &= \frac{1}{2}\left(\frac{\lambda^2 s}{1-\lambda^2 s}|x| - \frac{\lambda}{1-\lambda^2 s}\sum_{i=1}^s |\bar{y}_i|\right)^2 - \frac{\lambda^2 s}{2(1-\lambda^2 s)^2}\left(|x| - \lambda \sum_{i=1}^s |\bar{y}_i|\right)^2 \\ &+ \frac{\lambda}{1-\lambda^2 s}\sum_{i=1}^s |\bar{y}_i| \left(|x| - \lambda \sum_{i=1}^s |\bar{y}_i|\right) - \frac{1}{2}\sum_{i=1}^s |\bar{y}_i|^2 + \frac{1}{2}\|y\|_2^2 \\ &= \frac{\lambda^2 s}{2(1-\lambda^2 s)^2} x^2 (\lambda^2 s - 1) + \frac{\lambda^2}{2(1-\lambda^2 s)^2}\left(\sum_{i=1}^s |\bar{y}_i|\right)^2 (1 - \lambda^2 s - 2(1-\lambda^2 s)) \\ &+ |x| \sum_{i=1}^s |\bar{y}_i| \left(-\frac{\lambda^3 s}{(1-\lambda^2 s)^2} + \frac{\lambda^3 s}{(1-\lambda^2 s)^2} + \frac{\lambda}{1-\lambda^2 s}\right) - \frac{1}{2}\sum_{i=1}^s |\bar{y}_i|^2 + \frac{1}{2}\|y\|_2^2 \\ &= \frac{1}{2(1-\lambda^2 s)}\left(-\lambda^2 s x^2 - \left(|x| - \lambda \sum_{i=1}^s |\bar{y}_i|\right)^2 + x^2\right) - \frac{1}{2}\sum_{i=1}^s |\bar{y}_i|^2 + \frac{1}{2}\|y\|_2^2 \\ &= -\frac{1}{2(1-\lambda^2 s)}\left(|x| - \lambda \sum_{i=1}^s |\bar{y}_i|\right)^2 + \frac{1}{2}\|x\|_2^2 - \frac{1}{2}\sum_{i=1}^s |\bar{y}_i|^2 + \frac{1}{2}\|y\|_2^2 \\ &= -\left(1 + \frac{\lambda^2}{1-\lambda^2 s}\right)\frac{1}{2(1-\lambda^2(s-1))}\left(|x| - \lambda \sum_{i=1}^{s-1} |\bar{y}_i| - \lambda |\bar{y}_s|\right)^2 + \frac{1}{2}\|x\|_2^2 - \frac{1}{2}\sum_{i=1}^{s-1} |\bar{y}_i|^2 - \frac{1}{2}|\bar{y}_s|^2 + \frac{1}{2}\|y\|_2^2 \\ &= h_\lambda(v^{(s-1)}, w^{(s-1)}; x, y) - \frac{1}{2(1-\lambda^2 s + \lambda^2)}\left(-2\lambda |\bar{y}_s|\left(|x| - \lambda \sum_{i=1}^{s-1} |\bar{y}_i|\right) + \lambda^2 |\bar{y}_s|^2\right) \\ &- \frac{\lambda^2}{2(1-\lambda^2 s)(1-\lambda^2 s + \lambda^2)}\left(|x| - \lambda \sum_{i=1}^s |\bar{y}_i|\right)^2 - \frac{1}{2}|\bar{y}_s|^2. \end{aligned}$$

Therefore,

$$
\begin{aligned}
&h_\lambda(v^{(s)}, w^{(s)}; x, y) - h_\lambda(v^{(s-1)}, w^{(s-1)}; x, y) \\
&= -\frac{1}{2(1-\lambda^2 s+\lambda^2)}\left(-2\lambda|\bar{y}_s|\left(|x|-\lambda\sum_{i=1}^{s}|\bar{y}_i|\right)-\lambda^2|\bar{y}_s|^2+\frac{\lambda^2}{1-\lambda^2 s}\left(|x|-\lambda\sum_{i=1}^{s}|\bar{y}_i|\right)^2+(1-\lambda^2 s+\lambda^2)|\bar{y}_s|^2\right) \\
&= -\frac{1}{2(1-\lambda^2 s+\lambda^2)}\left((1-\lambda^2 s)|\bar{y}_s|^2-2\lambda|\bar{y}_s|\left(|x|-\lambda\sum_{i=1}^{s}|\bar{y}_i|\right)+\frac{\lambda^2}{1-\lambda^2 s}\left(|x|-\lambda\sum_{i=1}^{s}|\bar{y}_i|\right)^2\right) \\
&= -\frac{1-\lambda^2 s}{2(1-\lambda^2 s+\lambda^2)}\left(|\bar{y}_s|^2-2\lambda|\bar{y}_s|v^{(s)}+\lambda^2(v^{(s)})^2\right) \\
&= -\frac{1-\lambda^2 s}{2(1-\lambda^2 s+\lambda^2)}\left(|\bar{y}_s|-\lambda v^{(s)}\right)^2 \leq 0,
\end{aligned}
$$

meaning that

$$h_\lambda(v^{(s)}, w^{(s)}; x, y) \leq h_\lambda(v^{(s-1)}, w^{(s-1)}; x, y).$$

□

We can now prove our key result formulated in Corollary 2, that states that $(v^{(s^*)}, w^{(s^*)})$ is an optimal solution of (7) for

$$s^* = \max\left\{s \in [\bar{s}] : v^{(s)}, w^{(s)} > 0\right\}, \quad \text{where} \quad \bar{s} = \min(\lfloor\lambda^{-2}\rfloor, m).$$

Proof of Corollary 2. By Lemma 3, $(v^{(s^*)}, w^{(s^*)})$ is a stationary point of (7). Moreover, according to Corollary 1 and Lemma 4, $(v^{(s^*)}, w^{(s^*)})$ belongs to the set of $\bar{s}$ stationary points that are candidates to be optimal solutions of (7). Invoking Lemma 5, we have that

$$h_\lambda(v^{(s^*)}, w^{(s^*)}; x, y) < h_\lambda(v^{(j)}, w^{(j)}; x, y), \quad \forall s^* > j. \tag{22}$$

Hence, $(v^{(j)}, w^{(j)})$ is not an optimal solution for any $j < s^*$.

Let us now consider the complementary case. By Lemma 4, for any $i > \bar{s}$ the pair $(v^{(i)}, w^{(i)})$ does not satisfy the second-order optimality conditions, and therefore is not an optimal solution. On the other hand, by the definition of s^*, for any $\bar{s} > i > s^*$ the pair $(v^{(i)}, w^{(i)})$ is not a feasible solution , and subsequently not a stationary point. To conclude, $h_\lambda(v^{(s^*)}, w^{(s^*)}; x, y) < h_\lambda(v^{(j)}, w^{(j)}; x, y)$ holds for any $j \neq s^*$ such that $(v^{(j)}, w^{(j)})$ is a stationary point, meaning that $(v^{(s^*)}, w^{(s^*)})$ is an optimal solution of (7).

□

Finally, we will show that the problem of finding s^* can be easily solved using binary search. To this end, we show that the feasibility criterion (i.e., $v^{(s)} > 0$ and $w^{(s)} > 0$) satisfies that

$$(v^{(k)}, w^{(k)}) \text{ is feasible } \Rightarrow (v^{(i)}, w^{(i)}) \text{ is feasible } \forall i < k$$

Proof of Lemma 6. Suppose that $(v^{(k)}, w^{(k)})$ is feasible for some $k \in \{2, \ldots, \bar{s}\}$. By induction principle, it is sufficient to show that $(v^{(k-1)}, w^{(k-1)})$ is feasible in order to prove the result.

By (21), we have:

$$(1-k\lambda^2)v^{(k)} = |x| - \lambda\sum_{j=1}^{k}|y_j| = (1-k\lambda^2+\lambda^2)v^{(k-1)} - |y_k|.$$

which implies

$$v^{(k-1)} = \frac{1}{(1-k\lambda^2+\lambda^2)}((1-k\lambda^2)v^{(k)} + |y_k|) \geq 0.$$

For $w^{(k)}$, it is easy to see from (21) that since the vector y is sorted in decreasing order of magnitude, the vector $w^{(k)}$ is also sorted in decreasing order, and thus $w^{(k)}$ is feasible if and only if $w_k^{(k)} > 0$.

$$\begin{aligned}(1-k\lambda^2)w_k^{(k)} &= (1-k\lambda^2)|y_k| - \lambda|x| + \lambda^2\sum_{j=1}^{k}|y_j| \\ &= -\lambda|x| + (1-(k-1)\lambda^2)|y_{k-1}| + \lambda^2\sum_{j=1}^{k-1}|y_j| + \lambda^2|y_k| + (1-k\lambda^2)|y_k| - (1-(k-1)\lambda^2)|y_{k-1}| \\ &= (1-(k-1)\lambda^2)w_{k-1}^{(k-1)} + (1-k\lambda^2+\lambda^2)(|y_k| - |y_{k-1}|),\end{aligned}$$

where the last line uses the identity of the first line for $k-1$. We thus have:

$$w_{k-1}^{(k-1)} = \frac{1}{(1-(k-1)\lambda^2)}(1-k\lambda^2)w_k^{(k)} + |y_{k-1}| - |y_k| \geq 0,$$

since $|y_{k-1}| \geq |y_k|$ and $k \leq \lambda^{-2}$.

Therefore, there exists a value s^* such that $v^{(k)} > 0$ and $w^{(k)} > 0\ \forall k \geq s^*$ and $v^{(k)} \geq 0$ or $w^{(k)} \geq 0\ \forall k > s^*$. This value of s^* can thus efficiently be found using binary search.

□

D. Multi-output proximal map computation

D.1. Solving the prox problem

Returning to the multi-output setting, recall that $h_{V,W}(x) = V^T\sigma(Wx)$ with $V \in \mathbb{R}^{p\times n}, W \in \mathbb{R}^{n\times m}$ and

$$g(V,W) = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{p} W_{ij}V_{ki}.$$

The proximal mapping can then be written as:

$$\begin{aligned}\text{prox}_{\lambda g}(\bar{V},\bar{W}) &= \operatorname*{argmin}_{V,W} \frac{1}{2}\|V-\bar{V}\|_F + \frac{1}{2}\|W-\bar{W}\|_F + \lambda\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{p} W_{ij}V_{ki} \\ &= \operatorname*{argmin}_{V,W} \sum_{i=1}^{n}\left(\frac{1}{2}\sum_{k=1}^{p}(V_{ki}-\bar{V}_{ki})^2 + \sum_{j=1}^{p}(W_{ij}-\bar{W}_{ij})^2 + \sum_{j=1}^{m}\sum_{k=1}^{p} W_{ij}V_{ki}\right).\end{aligned}$$

Noting that the proximal mapping is separable with respect to the columns of W and the rows of V, and using the same sign trick applied in the single-output case, it is enough to solve for any $i = 1,\ldots,n$,

$$\min_{v,w\in\mathbb{R}_+^p\times\mathbb{R}_+^m} h_\lambda(v,w;x,y) \equiv \frac{1}{2}\sum_{k=1}^{p}(v_k-|x_k|)^2 + \frac{1}{2}\sum_{j=1}^{m}(w_j-|y_j|)^2 + \lambda\sum_{j=1}^{m}\sum_{k=1}^{p} v_k w_j, \tag{23}$$

where x denotes the i-th row of V and y the i-th column of W, in order to compute the prox operator.

The stationarity conditions for (23) are stated next; the arguments are the same as in the single-output case.

Lemma 15 (Stationarity conditions)**.** *Let $(v^*,w^*) \in \mathbb{R}_+^p\times\mathbb{R}_+^m$ be an optimal solution of (23) for a given $(x,y)\in\mathbb{R}^p\times\mathbb{R}^m$. Then*

$$w_j^* = \max\left\{0, |y_j| - \lambda\sum_{k=1}^{p} v_k^*\right\} \text{ for any } j = 1,2,\ldots,m,$$

$$v_k^* = \max\left\{0, |x_k| - \lambda\sum_{j=1}^{m} w_j^*\right\} \text{ for any } k = 1,2,\ldots,p.$$

The next lemma restates the result in Lemma 7 which expands on the monotonic relation in magnitude originally established for single-output networks in Corollary 1.

Lemma 16. *Let $(v^*, w^*) \in \mathbb{R}^p_+ \times \mathbb{R}^m_+$ be an optimal solution of (19) for a given $(x, y) \in \mathbb{R}^p \times \mathbb{R}^m$.*

1. *The vector w^* satisfies that for any $j, l \in \{1, 2, \ldots, m\}$ it holds that $w^*_j \geq w^*_l$ only if $|y_j| \geq |y_l|$.*
2. *The vector v^* satisfies that for any $k, l \in \{1, 2, \ldots, p\}$ it holds that $v^*_k \geq v^*_l$ only if $|x_k| \geq |x_l|$.*
3. *Let $\bar{x}, \bar{y}$ be the sorted vector of x and y respectively in descending magnitude order. Let $s_v = |\{k : v^*_k > 0\}|$ and $s_w = |\{j : w^*_j > 0\}|$. If $v^*, w^* \neq 0$, then*

$$v^*_k = |x_k| + \frac{1}{1 - s_v s_w \lambda^2} \left(\lambda^2 s_w \sum_{l=1}^{s_v} |\bar{x}_l| - \lambda \sum_{j=1}^{s_w} |\bar{y}_j| \right), \tag{24}$$

$$w^*_j = |y_j| + \frac{1}{1 - s_v s_w \lambda^2} \left(\lambda^2 s_v \sum_{l=1}^{s_w} |\bar{y}_l| - \lambda \sum_{k=1}^{s_v} |\bar{x}_k| \right). \tag{25}$$

Proof. The two first points are direct applications of the stationary conditions of Lemma 15.

From the conditions in Lemma 15 we have that

$$\begin{aligned}
\sum_{j=1}^{m} w^*_j &= \sum_{j=1}^{s_w} |\bar{y}_j| - \lambda s_w \sum_{k=1}^{p} v^*_k \\
\sum_{k=1}^{p} v^*_k &= \sum_{k=1}^{s_v} |\bar{x}_k| - \lambda s_v \sum_{j=1}^{m} w^*_j \\
&= \sum_{k=1}^{s_v} |\bar{x}_k| - \lambda s_v \sum_{j=1}^{s_w} |\bar{y}_j| + \lambda^2 s_v s_w \sum_{k=1}^{p} v^*_k \\
&= \frac{1}{1 - \lambda^2 s_v s_w} \left(\sum_{k=1}^{s_v} |\bar{x}_k| - \lambda s_v \sum_{j=1}^{s_w} |\bar{y}_j| \right).
\end{aligned}$$

Thus,

$$\begin{aligned}
\sum_{j=1}^{m} w^*_j &= \sum_{j=1}^{s_w} |\bar{y}_j| - \frac{\lambda s_w}{1 - \lambda^2 s_v s_w} \left(\sum_{k=1}^{s_v} |\bar{x}_k| - \lambda s_v \sum_{j=1}^{s_w} |\bar{y}_j| \right) \\
&= \frac{1}{1 - \lambda^2 s_v s_w} \left(-\lambda s_w \sum_{k=1}^{s_v} |\bar{x}_k| + \sum_{j=1}^{s_w} |\bar{y}_j| \right).
\end{aligned}$$

Plugging the latter to the stationarity condition on v^* (given in Lemma 15) then implies the result. □

Finally, we show, as in the single-output case, that second order stationarity condition constraints the ranges of sparsities of v^* and w^*; this relation is given by Lemma 8, and is proved next.

Proof of Lemma 8. Since (v^*, w^*) is an optimal solution of (23) and the objective function in (23) is twice continuously differentiable, (v^*, w^*) satisfies the second order necessary optimality conditions. That is, for any $d \in \mathbb{R}^p \times \mathbb{R}^m$ satisfying that $(v^*, w^*) + d \in \mathbb{R}^p_+ \times \mathbb{R}^m_+$ and $d^T \nabla h_\lambda(v^*, w^*; x, y) = 0$ it holds that

$$d^T \nabla^2 h_\lambda(v^*, w^*; x, y) d = d^T \begin{pmatrix} I_{p\times p} & \Lambda_{p\times m} \\ \Lambda_{m\times p} & I_{m\times m} \end{pmatrix} d \geq 0,$$

where the first row/column corresponds to v and the others correspond to w, I denotes the identity matrix and Λ denotes a matrix completely filled with λ. Similarly as in the single output case, we require that the submatrix of $\nabla^2 h_\lambda(v^*, w^*; x, y)$ containing the rows and columns corresponding to the positive coordinates in (v^*, w^*) is positive semidefinite. Since the minimal eigenvalue of this submatrix equals $1 - \lambda\sqrt{|S_v||S_w|}$, we have that

$$\lambda^{-2} \geq |S_v||S_w|.$$

□

A possible way of solving this proximal problem is thus to exhaustively compute the value of h_λ at each stationary point associated with sparsities $s_v = 1, \ldots, p$, $s_w = 1, \ldots, m$ such that $s_v s_w \leq \lambda^{-2}$. However, trying all possible pairs of sparsities (s_v, s_w) is computationally costly. Similarly as is the single output case, we can exploit some structure of the objective function h_λ in order to reduce the possible candidate pairs of sparsities.

Without loss of generality, we assume hereafter that the vectors x, y are already sorted in decreasing order of magnitude.

Lemma 16 shows that for each pair (s_v, s_w), $s_v = 0, 1, \ldots, p$, $s_w = 0, 1, \ldots, m$, there exists a stationary point $(v^{(s_v,s_w)}, w^{(s_v,s_w)})$ of $h_\lambda(\cdot, \cdot; x, y)$ such that $|\{k : v_k^{(s_v,s_w)} > 0\}| = s_v$, $|\{j : w_j^{(s_v,s_w)} > 0\}| = s_w$, given by

$$\begin{aligned} v_k^{(s_v,s_w)} &= |x_k| + \frac{1}{1 - s_v s_w \lambda^2} \left(\lambda^2 s_w \sum_{l=1}^{s_v} |x_l| - \lambda \sum_{j=1}^{s_w} |y_j| \right) \quad \text{for } k = 1, 2, \ldots, s_v, \text{ and } v_k^{(s_v,s_w)} = 0 \text{ otherwise} \\ w_j^{(s_v,s_w)} &= |y_j| + \frac{1}{1 - s_v s_w \lambda^2} \left(\lambda^2 s_v \sum_{l=1}^{s_w} |y_l| - \lambda \sum_{k=1}^{s_v} |x_k| \right) \quad \text{for } j = 1, 2, \ldots, s_w, \text{ and } w_j^{(s_v,s_w)} = 0 \text{ otherwise.} \end{aligned} \tag{26}$$

We now move to prove the monotonicity property stated in Lemma 9.

Proof of Lemma 9. The proof follows exactly the same lines as in the single output case. We recall the definition of the objective function:

$$h_\lambda(v, w; x, y) \equiv \frac{1}{2} \sum_{k=1}^{p} (v_k - |x_k|)^2 + \frac{1}{2} \sum_{j=1}^{m} (w_j - |y_j|)^2 + \lambda \left(\sum_{k=1}^{p} v_k \right) \left(\sum_{j=1}^{m} w_j \right).$$

Plugging the definitions from equation (26), we have

$$
\begin{aligned}
h_\lambda\left(v^{(s_v,s_w)}, w^{(s_v,s_w)}; x, y\right) &= \frac{s_v}{2}\left(\frac{1}{1-\lambda^2 s_v s_w}\left(\lambda^2 s_w \sum_{k=1}^{s_v}|x_k| - \lambda\sum_{j=1}^{s_w}|y_j|\right)\right)^2 + \frac{1}{2}\sum_{k=s_v+1}^{p} x_k^2 \\
&+ \frac{s_w}{2}\left(\frac{1}{1-\lambda^2 s_v s_w}\left(\lambda^2 s_v \sum_{j=1}^{s_w}|y_j| - \lambda\sum_{k=1}^{s_v}|x_k|\right)\right)^2 + \frac{1}{2}\sum_{j=s_w+1}^{m} y_j^2 \\
&+ \frac{\lambda}{(1-\lambda^2 s_v s_w)^2}\left(\sum_{k=1}^{s_v}|x_k| - \lambda s_v\sum_{j=1}^{s_w}|y_j|\right)\left(-\lambda s_w\sum_{k=1}^{s_v}|x_k| + \sum_{j=1}^{s_w}|y_j|\right) \\
&= \frac{1}{2(1-\lambda^2 s_v s_w)^2}\left(\left(\sum_{k=1}^{s_v}|x_k|\right)^2(\lambda^4 s_v s_w^2 + \lambda^2 s_w - 2\lambda^2 s_w) + \left(\sum_{j=1}^{s_w}|y_j|\right)^2(\lambda^2 s_v + \lambda^4 s_v^2 s_w - 2\lambda^2 s_v)\right. \\
&\left.\left(\sum_{k=1}^{s_v}|x_k|\right)\left(\sum_{j=1}^{s_w}|y_j|\right)(-2\lambda^3 s_v s_w - 2\lambda^3 s_v s_w + 2\lambda + 2\lambda^3 s_v s_w)\right) + \frac{1}{2}\sum_{k=s_v+1}^{p} x_k^2 + \frac{1}{2}\sum_{j=s_w+1}^{m} y_j^2 \\
&= \frac{1}{2(1-\lambda^2 s_v s_w)}\left(-\lambda^2 s_w\left(\sum_{k=1}^{s_v}|x_k|\right)^2 - \lambda^2 s_v\left(\sum_{j=1}^{s_w}|y_j|\right)^2 + 2\lambda\left(\sum_{k=1}^{s_v}|x_k|\right)\left(\sum_{j=1}^{s_w}|y_j|\right)\right) + \frac{1}{2}\sum_{k=s_v+1}^{p} x_k^2 + \frac{1}{2}\sum_{j=s_w+1}^{m} y_j^2
\end{aligned}
\tag{27}
$$

$$
\begin{aligned}
&= \left(1 + \frac{\lambda^2 s_v}{1-\lambda^2 s_v s_w}\right)\frac{1}{2(1-\lambda^2 s_v(s_w-1))}\left(-\lambda^2(s_w-1)\left(\sum_{k=1}^{s_v}|x_k|\right)^2 - \lambda^2\left(\sum_{k=1}^{s_v}|x_k|\right)^2\right. \\
&\left.-\lambda^2 s_v\left(\left(\sum_{j=1}^{s_w-1}|y_j|\right)^2 + 2\lambda|y_{s_w}|\sum_{j=1}^{s_w-1}|y_j| + y_{s_w}^2\right) + 2\lambda\sum_{k=1}^{s_v}|x_k|\left(\sum_{j=1}^{s_w-1}|y_j| + |y_{s_w}|\right)\right) + \frac{1}{2}\sum_{k=s_v+1}^{p} x_k^2 + \frac{1}{2}\sum_{j=s_w-1+1}^{m} y_j^2 - \frac{1}{2}y_{s_w}^2.
\end{aligned}
\tag{28}
$$

By applying equation (27) at $s_v,\ s_w - 1$, we can express the right hand side of equation (28) in terms of $h_\lambda\left(v^{(s_v,s_w-1)}, w^{(s_v,s_w-1)}; x, y\right)$ as:

$$
\begin{aligned}
h_\lambda\left(v^{(s_v,s_w)}, w^{(s_v,s_w)}; x, y\right) &= h_\lambda\left(v^{(s_v,s_w-1)}, w^{(s_v,s_w-1)}; x, y\right) + \frac{1}{2(1-\lambda^2 s_v(s_w-1))}\left(-\lambda^2\left(\sum_{k=1}^{s_v}|x_k|\right)^2\right. \\
&\left.-\lambda^2 s_v|y_{s_w}|\left(2\sum_{j=1}^{s_w-1}|y_j| + |y_{s_w}|\right) + 2\lambda|y_{s_w}|\sum_{k=1}^{s_v}|x_k|\right) + \frac{\lambda^2 s_v}{2(1-\lambda^2 s_v s_w)(1-\lambda^2 s_v(s_w-1))}\left(-\lambda^2 s_w\left(\sum_{k=1}^{s_v}|x_k|\right)^2\right. \\
&\left.-\lambda^2 s_v\left(\sum_{j=1}^{s_w}|y_j|\right)^2 - 2\lambda\left(\sum_{k=1}^{s_v}|x_k|\right)\left(\sum_{j=1}^{s_w}|y_j|\right)\right) - \frac{1}{2}y_{s_w}^2.
\end{aligned}
$$

Therefore:

$$
\begin{aligned}
&h_\lambda\left(v^{(s_v,s_w)}, w^{(s_v,s_w)}; x, y\right) - h_\lambda\left(v^{(s_v,s_w-1)}, w^{(s_v,s_w-1)}; x, y\right) \\
&\quad = -\frac{1}{2(1-\lambda^2 s_v(s_w-1))}\left(-2\lambda|y_{s_w}|\left(\sum_{k=1}^{s_v}|x_k| - \lambda s_v\sum_{j=1}^{s_w}|y_j|\right) - \lambda^2 s_v|y_{s_w}|^2 + \lambda^2\left(\sum_{k=1}^{s_v}|x_k|\right)^2\right. \\
&\quad \left. + \frac{\lambda^2 s_v}{1-\lambda^2 s_v s_w}\left(\lambda^2 s_w\left(\sum_{k=1}^{s_v}|x_k|\right)^2 + \lambda^2 s_v\left(\sum_{j=1}^{s_w}|y_j|\right)^2 - 2\lambda\left(\sum_{k=1}^{s_v}|x_k|\right)\left(\sum_{j=1}^{s_w}|y_j|\right)\right) + (1-\lambda^2 s_v s_w + \lambda^2 s_v)|y_{s_w}|\right) \\
&\quad = -\frac{1}{2(1-\lambda^2 s_v(s_w-1))}\left((1-\lambda^2 s_v s_w)y_{s_w}^2 - 2\lambda|y_{s_w}|(1-\lambda^2 s_v s_w)\sum_{k=1}^{s_v} v_k^{(s_v,s_w)} + \frac{\lambda^2}{1-\lambda^2 s_v s_w}\left(\sum_{k=1}^{s_v}|x_k| - \lambda s_v\sum_{j=1}^{s_w}|y_j|\right)^2\right) \\
&\quad = -\frac{1-\lambda^2 s_v s_w}{2(1-\lambda^2 s_v(s_w-1))}\left(|y_{s_w}| - \lambda\sum_{k=1}^{s_v} v_k^{(s_v,s_w)}\right)^2.
\end{aligned}
$$

The second result is obtain directly by symmetry between v and w. □

In order to derive an efficient algorithm , we will again exploit the monotone property of the feasibility criterion $v^{(s_v,s_w)} > 0$, $w^{(s_v,s_w)} > 0$ restated from Lemma 10:

Lemma 17 (Restatement of Lemma 10)**.** *Let $(k,l) \in [p] \times [m]$ be such that $kl \leq \lambda^{-2}$. Suppose that*

$$v^{(k,l)} \geq 0 \text{ and } w^{(k,l)} \geq 0.$$

Then for any $i = 1, \ldots, k$ and any $j = 1, \ldots, l$, it holds that

$$v^{(i,j)} \geq 0 \text{ and } w^{(i,j)} \geq 0.$$

Proof of Lemma 10. Since the first k entries of $v^{(k,l)}$ are ordered in decreasing order, we have that $v^{(k,l)} \geq 0$ if and only if $v_k^{(k,l)} \geq 0$. Similarly, $w^{(k,l)} \geq 0$ if and only if $w_l^{(k,l)} \geq 0$.

Suppose that $v^{(k,l)} \geq 0$ and $w^{(k,l)} \geq 0$. By induction, in order to prove the result, it is sufficient to prove that $v_{k-1}^{(k-1,l)} \geq 0$, $v_k^{(k,l-1)} \geq 0$, $w_l^{(k-1,l)} \geq 0$ and $w_{l-1}^{(k,l-1)} \geq 0$. We only prove the result for v, as the proof for w is identical.

Using equation (26), we have:

$$
\begin{aligned}
(1-kl\lambda^2)v_k^{(k,l)} &= (1-kl\lambda^2)|x_k| + \lambda^2 l\sum_{i=1}^{k}|x_i| - \lambda\sum_{j=1}^{l}|y_j| \qquad (29)\\
&= (1-kl\lambda^2)|x_k| + (1-(k-1)l\lambda^2)|x_{k-1}| - (1-(k-1)l\lambda^2)|x_{k-1}| + \lambda^2 l\sum_{i=1}^{k-1}|x_i| + \lambda^2 l|x_k| - \lambda\sum_{j=1}^{l}|y_j| \\
&= (1-(k-1)l\lambda^2)v_{k-1}^{(k-1,l)} + (1-(k-1)l\lambda^2)(|x_k| - |x_{k-1}|).
\end{aligned}
$$

Therefore:

$$v_{k-1}^{(k-1,l)} = \frac{1-(k-1)l\lambda^2}{1-kl\lambda^2}v_k^{(k,l)} + |x_{k-1}| - |x_k| \geq 0,$$

since the vector x is ordered in decreasing order of magnitude, and thus $|x_{k-1}| - |x_k| \geq 0$.

Using again equation (29), we have:

$$\begin{aligned}(1-kl\lambda^2)v_k^{(k,l)} &= (1-kl\lambda^2)|x_k| + (1-k(l-1)\lambda^2)|x_k| - (1-k(l-1)\lambda^2)|x_k| \\ &\quad + \lambda^2(l-1)\sum_{i=1}^{k}|x_i| + \lambda^2\sum_{i=1}^{k}|x_i| - \lambda\sum_{j=1}^{l-1}|y_j| - \lambda|y_l| \\ &= (1-k(l-1)\lambda^2)v_k^{(k,l-1)} - k\lambda^2|x_k| + \lambda^2\sum_{i=1}^{k}|x_i| - \lambda|y_l|,\end{aligned}$$

where the last equality follows (again) from equation (29) for $v_k^{(k,l-1)}$. Thus,

$$(1-k(l-1)\lambda^2)v_k^{(k,l-1)} = (1-kl\lambda^2)v_k^{(k,l)} + k\lambda^2|x_k| - \lambda^2\sum_{i=1}^{k}|x_i| + \lambda|y_l|. \tag{30}$$

From the definition of $v_k^{(k,l)}$ (equation (26)), we have that $v_k^{(k,l)} \geq 0$ is equivalent to the condition:

$$|x_k| \geq \frac{\lambda\sum_{j=1}^{l}|y_j| - l\lambda^2\sum_{i=1}^{k}|x_i|}{1-kl\lambda^2}.$$

Plugging this inequality in equation (30), we obtain:

$$\begin{aligned}(1-k(l-1)\lambda^2)v_k^{(k,l-1)} &\geq (1-kl\lambda^2)v_k^{(k,l)} + \frac{k\lambda^2}{1-kl\lambda^2}\left(\lambda\sum_{j=1}^{l}|y_j| - l\lambda^2\sum_{i=1}^{k}|x_i|\right) + \lambda|y_l| - \lambda^2\sum_{i=1}^{k}|x_i| \\ &= (1-kl\lambda^2)v_k^{(k,l)} + \frac{\lambda}{1-kl\lambda^2}\left(k\lambda^2\sum_{j=1}^{l}|y_j| - kl\lambda^3\sum_{i=1}^{k}|x_i| + (1-kl\lambda^2)|y_l| - \lambda(1-kl\lambda^2)\sum_{i=1}^{k}|x_i|\right) \\ &= (1-kl\lambda^2)v_k^{(k,l)} + \frac{\lambda}{1-kl\lambda^2}\left(k\lambda^2\sum_{j=1}^{l}|y_j| + (1-kl\lambda^2)|y_l| - \lambda\sum_{i=1}^{k}|x_i|\right). \end{aligned}\tag{31}$$

From the definition of $w_l^{(k,l)}$ (equation (26)), we have that $w_l^{(k,l)} \geq 0$ is equivalent to the condition:

$$(1-kl\lambda^2)|y_l| + k\lambda^2\sum_{j=1}^{l}|y_j| - \lambda\sum_{i=1}^{k}|x_i| \geq 0. \tag{32}$$

Since the expression of equation (32) is exactly the same as the one inside the parentheses of equation (31), plugging this relation to (30) thus shows that $(1-k(l-1)\lambda^2)v_k^{(k,l-1)} \geq 0$, i.e. $v_k^{(k,l-1)} \geq 0$. □

We now introduce the efficient procedure to compute the maximal feasibility boundary (MFB), and prove that it indeed delivers, as promised, all sparsity pairs in the MFB set.

Lemma 18. *The set S returned by Algorithm 5 contains all, and only, the sparsity pairs that are on the maximal feasibility boundary.*

Proof. First recall that the MFB is defined as all pairs $(s_v, s_w) \in \{0,\dots,p\} \times \{0,\dots,m\}$ satisfying the conditions:

1. $v_{s_v}^{(s_v,s_w)} > 0$ and $w_{s_w}^{(s_v,s_w)} > 0$ and $s_v s_w \leq \lambda^{-2}$,
2. $v_{s_v+1}^{(s_v+1,s_w)} \leq 0$ or $w_{s_w}^{(s_v+1,s_w)} \leq 0$ or $(s_v+1)s_w > \lambda^{-2}$ or $s_v = p$,

Algorithm 5 Finding sparsity pairs on the maximal feasibility boundary

Input: $x \in \mathbb{R}^p$, $y \in \mathbb{R}^m$ ordered in decreasing magnitude order, $\lambda > 0$.

1: $s_v \leftarrow 0$, $s_w \leftarrow m$
2: $S \leftarrow \emptyset$
3: $maximal \leftarrow True$
4: **while** $s_v \leq p$ and $s_w \geq 0$ **do**
5: Compute $v_{s_v}^{(s_v,s_w)}$ and $w_{s_w}^{(s_v,s_w)}$ as shown in equation (26)
6: **if** $v_{s_v}^{(s_v,s_w)} < 0$ or $w_{s_w}^{(s_v,s_w)} < 0$ or $s_v s_w \geq \lambda^{-2}$ **then**
7: **if** $maximal$ **then**
8: $S \leftarrow S \cup \{(s_v - 1, s_w)\}$
9: $maximal \leftarrow False$
10: **end if**
11: $s_w \leftarrow s_w - 1$
12: **else**
13: $s_v \leftarrow s_v + 1$
14: $maximal \leftarrow True$
15: **end if**
16: **end while**
17: **if** $s_v == p + 1$ **then**
18: $S \leftarrow S \cup \{(s_v - 1, s_w)\}$
19: **end if**
20: **return** S

3. $v_{s_v}^{(s_v,s_w+1)} \leq 0$ or $w_{s_w+1}^{(s_v,s_w+1)} \leq 0$ or $s_v(s_w + 1) > \lambda^{-2}$ or $s_w = m$.

Algorithm 5 plays on the properties of *feasibility-infeasibility* of the sparsity levels to build the MFB. We say that a pair of the sparsity levels of v and w (s_v, s_w) is *feasible* if $v_{s_v}^{(s_v,s_w)} \geq 0$, $w_{s_w}^{(s_v,s_w)} \geq 0$ and $s_v s_w < \lambda^{-2}$, and denote this by the property $P(i, j)$, i.e.

$$(i, j) \text{ is feasible } \Leftrightarrow P(i, j).$$

Our claim can be read as: Let $(i, j) \in \{0, \ldots, p\} \times \{0, \ldots, m\}$, then (i, j) is added to S by Algorithm 5 if and only if (i, j) belongs to the MFB, i.e.,

$$(i, j) \in \text{MFB} \Leftrightarrow (i, j) \in S.$$

Obviously, only feasible sparsity pairs belong to the MFB, and it is quite easy to see that only feasible sparsity pairs will belong to an output S of Algorithm 5. Indeed, Algorithm 5 monotonically decrements s_w starting from $s_w = m$ and increments s_v starting from $s_v = 0$. For each value of s_w, it increases s_v while the current pair (s_v, s_w) is feasible (lines $12 - 15$). Once it reaches an infeasible point (i, s_w), and in the case where s_v has been increased at least once for this particular value of s_w, it adds to S the pair encountered just before, i.e., $(i - 1, s_w)$, and then decrements s_w (lines $6 - 11$).

We first prove the $\Rightarrow$ statement. Suppose that some pair (i, j) belongs to the MFB. Let us first leave aside the corner cases, and assume that $i < p$ and $j < m$.

Suppose first that s_w reaches j before s_v reaches i, i.e., $s_v < i$. Since the pair (i, j) is feasible, and due to the monotonicity property of the feasibility condition (Lemma 9), all pairs (k, s_w) with $k \leq i$ must be feasible. Therefore, s_v will be increased until reaching $i + 1$. By definition of the MFB, the pair $(i + 1, j)$ must be infeasible. Since s_v has necessarily been increased at least once for this value of $s_w = j$, and so the pair $(i + 1 - 1, j) = (i, j)$ will be added to S before decrementing s_w.

In the special case where $i = p$, no infeasible point will be found. The loop will thus finish with $s_w = j$ and $s_v = p + 1$. The condition at line 17 will thus hold, and the pair (p, j) will be added to S.

Suppose now that s_v reaches i before s_w reaches j, i.e., $s_w > j$. Since (i, j) is in the MFB, then the pair $(i, j + 1)$ must be infeasible. Thanks to the monotonicity property of the feasibility condition (Lemma 9), all pairs (s_v, k) with $k \geq i$ must also be infeasible. Therefore, s_w will be decreased until reaching $s_w = j$. Then, similarly as in the previous case, since (i, j) is feasible, s_v will be increased, and the pair (i, j) added to S.

We now prove the $\Leftarrow$ statement. We show that if (i, j) is added to S, then it must belong to the MFB, i.e., it satisfies all three properties recalled in the beginning of the proof.

Let us first show that for each pair (s_v, s_w) encountered during the algorithm, the pair $(s_v - 1, s_w)$ is always feasible (or $s_v = 0$). We can show that this property is conserved each time the algorithm either increases s_v or decreases s_w. First note that the pair $(0, m)$ is always feasible. The algorithm will then necessarily first goes to the pair $(1, m)$ and $P(1, m)$ is true. Then suppose that $P(s_v, s_w)$ is true for some pair (s_v, s_w) encountered during the algorithm. Then, if s_v is increases, it means that the pair (s_v, s_w) is feasible. The next encountered pair is then $(s_v + 1, s_w)$ and $P(s_v + 1, s_w)$ is true. On the other hand, suppose that s_w is decreased. The next encountered pair is thus $(s_v, s_w - 1)$. Since $P(s_v, s_w)$ is true, it means that $(s_v - 1, s_w)$ is feasible. By Lemma 9, it implies that $(s_v - 1, s_w - 1)$ is also feasible, and thus $P(s_v, s_w - 1)$ is true. We thus proved that $P(s_v, s_w)$ is true for any pair (s_v, s_w) encountered during the algorithm. Therefore, since any pair added to S is of the form $(s_v - 1, s_w)$ for some pair (s_v, s_w) encountered during the algorithm, then any pair added to S must be feasible.

The second property of the MFB is straightforward to show. Indeed, if $(i - 1, j)$ is added to S, it means that the pair (i, j) is infeasible due to condition on line 6.

Finally, the third property follows from the fact that, when reaching $s_w = j$, s_v must be increased at least once for adding a pair of the form (i, j) to S. Let $s_v^{(j)}$ be the value of s_v when the algorithm reaches $s_w = j$. We necessarily have $s_v^{(j)} \leq i$. This implies that the pair $(s_v^{(j)}, j + 1)$ is infeasible, otherwise s_v would have been increased to a greater value at the previous value $s_w = j + 1$. By Lemma 9, and since $s_v^{(j)} \leq i$ this implies that the pair $(i, j + 1)$ is also infeasible, hence the result.

□

Time complexity of Algorithm 5 At each iteration of the loop, either s_v is incremented by 1 or s_w is decremented by 1. Since s_v starts from 0 and s_w from m, and that the stopping criterion is $s_v > p$ or $s_w < 0$, it follows that the maximal number of iterations inside the loop is $m + p$. At each iteration, we must compute $v_{s_v}^{(s_v, s_w)}$ and $w_{s_w}^{(s_v, s_w)}$, which requires in particular to compute $\sum_{k=1}^{s_v} |x_k|$ and $\sum_{j=1}^{s_w} |y_j|$. However, these cumulative sums can be efficiently computed before the loop in time $\mathcal{O}(m + p)$, so that computing $v_{s_v}^{(s_v, s_w)}$ and $w_{s_w}^{(s_v, s_w)}$ inside the loop can be done in constant time. The overall complexity of this algorithm is thus $\mathcal{O}(m + p)$.

Moreover, we can see that each time we add a pair to S, we must both decrement s_w by 1 (just after adding the element in the algorithm), and increment s_v by 1 (in order for the boolean $maximal$ to become true again). Therefore, there can be at most $\min(m, p)$ pairs in the final set s at the end of the algorithm.

Merging all previous results, we can finally prove Theorem 4.

Proof of Theorem 4. Thanks to the separability argument, it is sufficient to prove that Algorithm 3 returns a solution of problem (11).

Lemma 7 states that given the number of nonzero elements $s_v = |\{k : v_k^* > 0\}|$, $s_w = |\{j : w_j^* > 0\}|$, the optimal solution (v^*, w^*) can be obtained in close form (equations (12), (13)).

Due the monotonicity property of the objective function h_λ (Lemma 9), it follows that the sparsity pair (s_v, s_w) of the optimal solution must lie on the MFB. Indeed, if it does not lie on the MFB, then it means that the candidate solution associated with either the sparsity pair $(s_v + 1, s_w)$ or $(s_v, s_w + 1)$ must be feasible. According to Lemma 9, this pair would then yield a lower value of h_λ, and would then be a better solution.

Algorithm 3 computes the candidate solution associated with all sparsity pair lying on the MFB, and returns the one achieving the lowest value of h_λ. Therefore, the returned solution must necessarily be the optimal solution. □

E. Experimental details and other plots

We consider the following values for the parameters that determine the training loop:

- ▷ batch size: `100`
- ▷ epochs: `20`

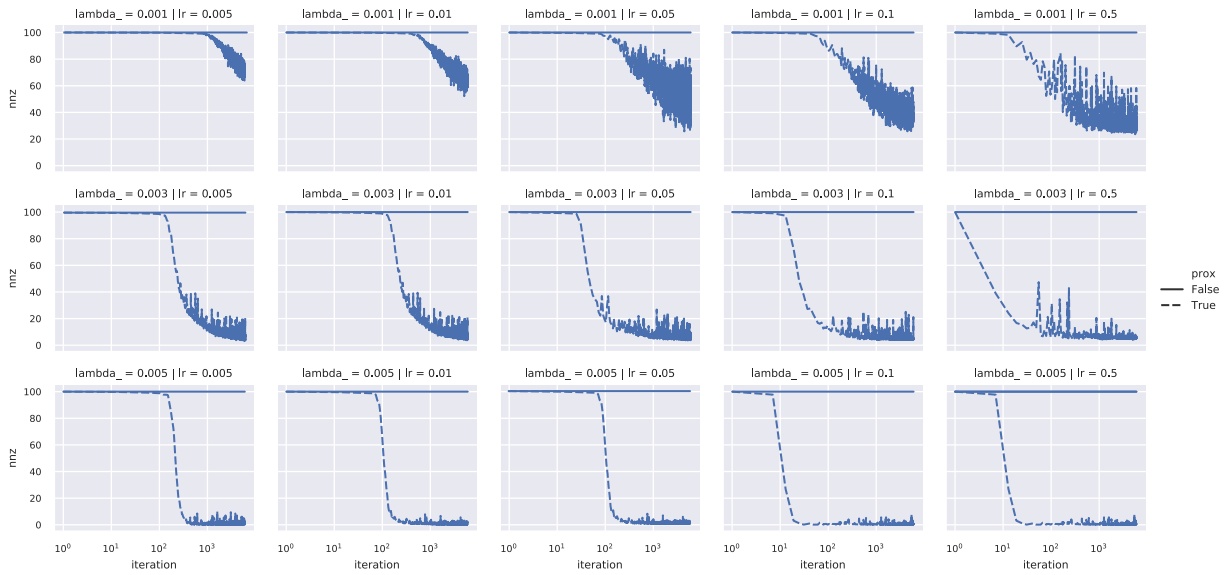

Figure 4. percentage of nonzero weights in the network, as a function of iteration count (path regularization - fmnist dataset).

- ▷ learning rate: `1e-1, 1e-2, 1e-3, 1e-4, 5e-1, 5e-2, 5e-3, 5e-4`
- ▷ dataset: `mnist, fmnist, kmnist`
- ▷ hidden neurons: `200`
- ▷ lambda (λ): `0., 1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 1e0, 1e1, 1e2, 2e-5, 2e-4, 2e-3, 2e-2, 2e-1, 2e0, 2e1, 2e2, 3e-5, 3e-4, 3e-3, 3e-2, 3e-1, 3e0, 3e1, 3e2, 4e-5, 4e-4, 4e-3, 4e-2, 4e-1, 4e0, 4e1, 4e2, 5e-5, 5e-4, 5e-3, 5e-2, 5e-1, 5e0, 5e1, 5e2`

The ℓ_∞-bounded adversarial examples used to evaluate the robustness of the networks were generated using the PGD method described in (Madry et al., 2018) and implemented in the *advertorch* toolbox (`https://github.com/BorealisAI/advertorch`) using the following parameters:

- ▷ epsilon: `0.05, 0.1, 0.15, 0.2, 0.25, 0.3`
- ▷ iterations: 40
- ▷ step size: epsilon / 20
- ▷ random initialization: `True`

E.1. sparsity per iteration

One advantage of the proximal mapping of the 1-path-norm and the ℓ_1-norm is that they can set many weights to exactly zero. This has the effect of providing sparse networks from early iterations. This is in contrast to SGD with a constant stepsize which does not generate sparse iterates. In Figures 4, 5, 6 and 7 we plot the percentage of nonzero weights as a function of the iteration count, for both plain SGD and proximal SGD. We observe that in fact this is the case, and that the sparsity of the ℓ_1 and 1-path-norm regularized network can be controlled with the regularization parameter λ.

E.2. Robustness vs accuracy tradeoff

For all possible values of λ, in Figure 8 we plot the data corresponding to the lerning rate with least error. We plot the value of the error on clean samples and the error on adversarial examples. This allows us to understand the tradeoff between accuracy and robustness that is controlled by the regularization paramter λ.

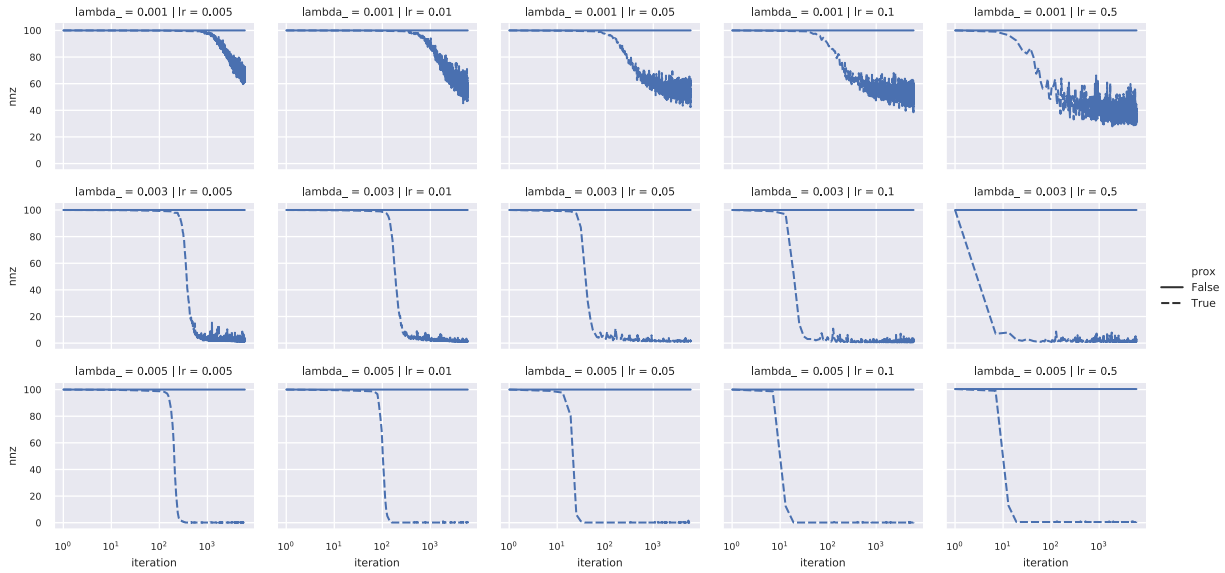

Figure 5. percentage of nonzero weights in the network, as a function of iteration count (path regularization - kmnist dataset).

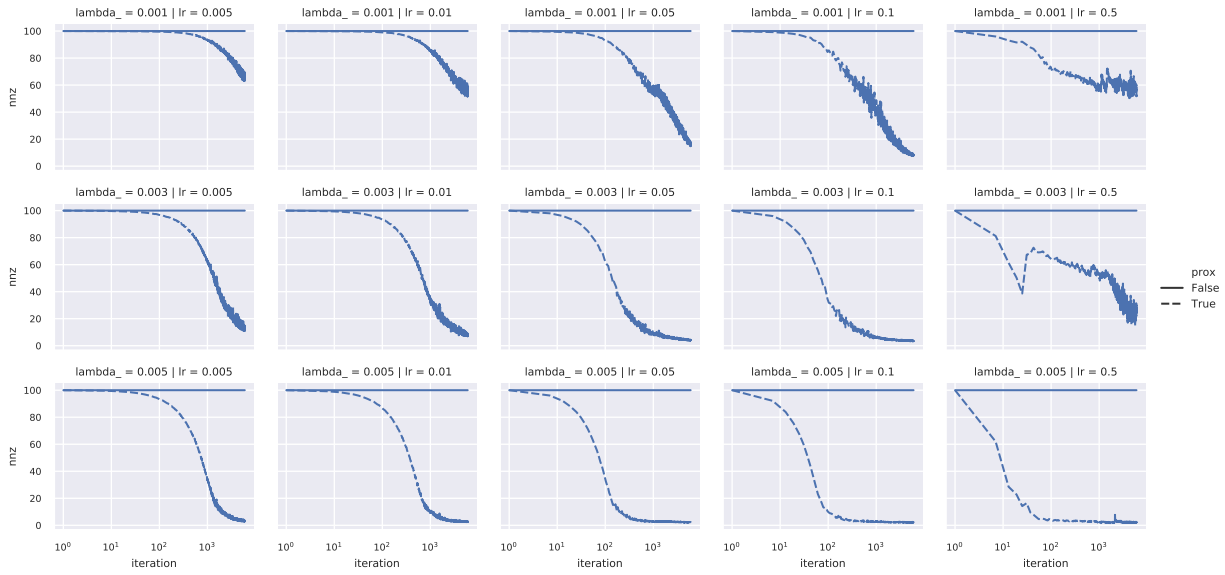

Figure 6. percentage of nonzero weights in the network, as a function of iteration count (ℓ_1 regularization - fmnist dataset).

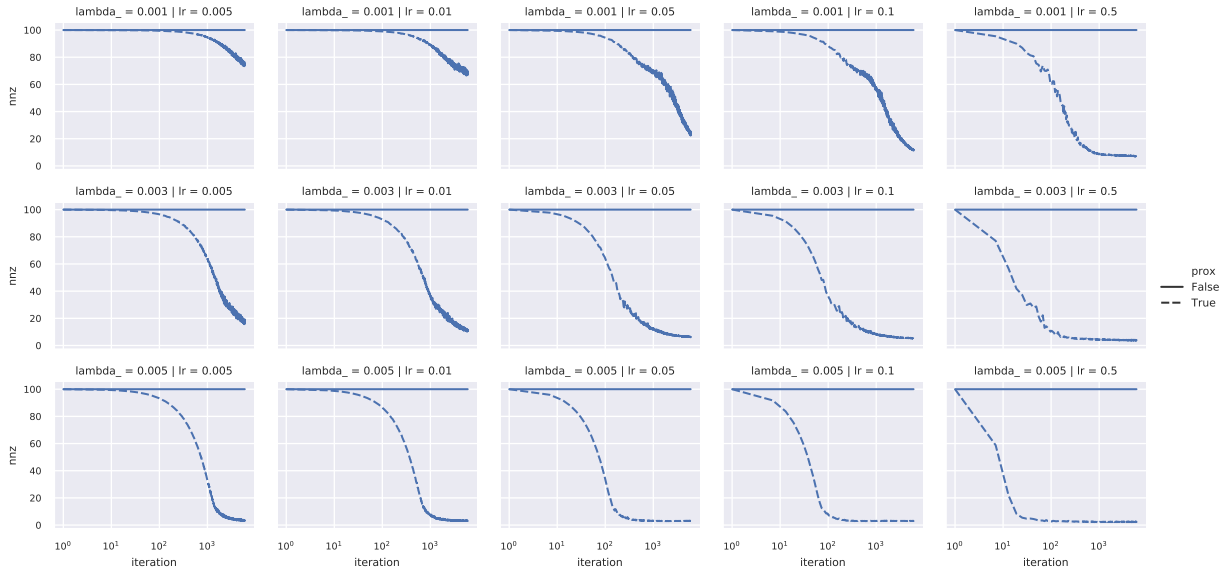

Figure 7. percentage of nonzero weights in the network, as a function of iteration count (ℓ_1 regularization - kmnist dataset).

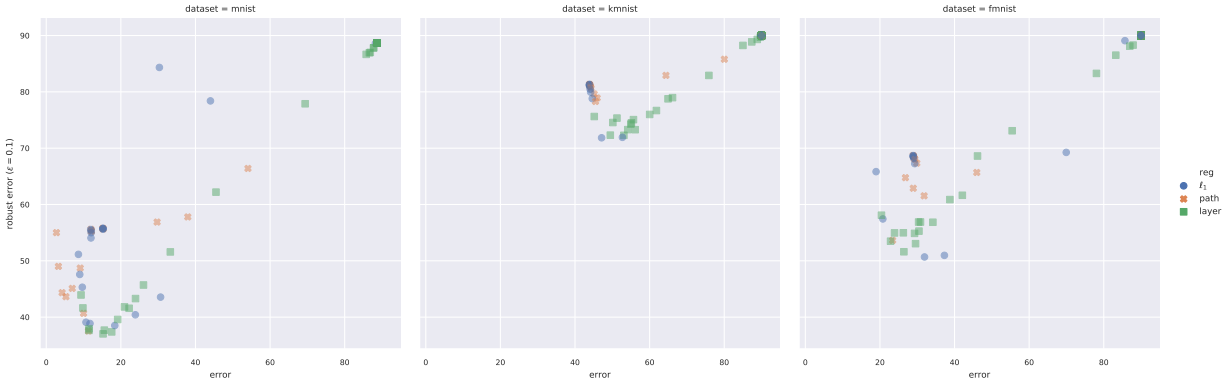

Figure 8. Robustness vs accuracy tradeoff for the different regularizers studied.