A. Supplementary Material

A.1. S annealing schedule

Acceleration factor α, annealing start iteration i_{min}, annealing end iteration i_{max}, current iteration i_{cur}, $d_{target} = \frac{1}{\alpha}$.

$$d_{cur} = d_{target} + (1 - d_{target}) \left(1 - \frac{i_{cur} - i_{min}}{i_{max} - i_{min}}\right)$$
$$S_{cur} = d_{cur} \cdot D$$

A.2. Proof of Constraint Projection

Proof for Eq. 6 and Eq. 7 is taken and adapted from [38]. Transforming updated parameters $\tilde{\boldsymbol{\theta}} \in \mathbb{R}^D$ into $\boldsymbol{\theta}$, which fulfills the sparsification constraint can be described as a least-squares convex problem:

$$\arg\min_{\boldsymbol{\theta}} \frac{1}{2}||\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}||^2 \; s.t. \sum_{i=1}^{D} \theta_i = \mathbf{1}^\top \boldsymbol{\theta} \leq S \text{and} 0 \leq \theta_i \leq 1\,. \tag{9}$$

This can be solved by the Lagrangian multiplier method:

$$\begin{aligned} \mathfrak{L}(\boldsymbol{\theta}, \lambda) &= \frac{1}{2}||\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}||^2 + \lambda(\mathbf{1}^\top \boldsymbol{\theta} - S) && (10)\\ &= \frac{1}{2}||\boldsymbol{\theta} - (\tilde{\boldsymbol{\theta}} - \lambda \mathbf{1})||^2 + \lambda(\mathbf{1}^\top \tilde{\boldsymbol{\theta}} - S) - \frac{n}{2}\lambda^2\,, && (11) \end{aligned}$$

where $\lambda \geq 0$ and $0 \leq \theta_i \leq 1$. Minimizing w.r.t. $\boldsymbol{\theta}$ results in

$$\bar{\boldsymbol{\theta}} = \mathbf{1}_{\tilde{s} - \lambda \mathbf{1} \geq 1} + (\tilde{s} - \lambda \mathbf{1})_{1 > \tilde{s} - \lambda \mathbf{1} > 0}\,. \tag{12}$$

Thus, for $\lambda \geq 0$

$$\begin{aligned} g(\lambda) &= \mathfrak{L}(\bar{\boldsymbol{\theta}}, \lambda) \\ &= \frac{1}{2}||[\tilde{s} - \lambda \mathbf{1}]_- + [\tilde{\boldsymbol{\theta}} - (\lambda + 1)\mathbf{1}]_+||^2 \\ &+ \lambda(\mathbf{1}^\top \tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) - \frac{D}{2}\lambda^2 \\ &= \frac{1}{2}||[\tilde{s} - \lambda \mathbf{1}]_-|| + \frac{1}{2}||[\tilde{\boldsymbol{\theta}} - (\lambda + 1)\mathbf{1}]_+||^2 \\ &+ \lambda(\mathbf{1}^\top \tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) - \frac{D}{2}\lambda^2 \end{aligned} \tag{13}$$

and

$$\begin{aligned} g'(\lambda) &= \mathbf{1}^\top [\lambda \mathbf{1} - \tilde{\boldsymbol{\theta}}]_+ + \mathbf{1}^\top [(\lambda + 1)\mathbf{1} - \boldsymbol{\theta}]_- \\ &+ (\mathbf{1}^\top \tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) - D\lambda \\ &= \mathbf{1}^\top \min(\mathbf{1}, \max(\mathbf{0}, \tilde{\boldsymbol{\theta}} - \lambda \mathbf{1})) - S \\ &= [\textstyle\sum_{i=1}^{D} \min(1, \max(0, \tilde{\theta}_i - \lambda))] - S\,. \end{aligned} \tag{14}$$

With $g'(\lambda)$ being a monotone function, λ_1^* a solution for $g'(\lambda) = 0$ can be obtained by e.g. a convex solver or a bisection method. The maximum of $g(\lambda)$ with $\lambda \geq 0$ is at $\lambda_2^* = \max(0, \lambda_1^*)$. Eventually,

$$\begin{aligned} \boldsymbol{\theta}^* &= \mathbf{1}_{\tilde{s} - \lambda_2^* \mathbf{1} \geq 1} + (\tilde{s} - \lambda_2^* \mathbf{1})_{1 > \tilde{s} - \lambda_2^* \mathbf{1} > 0} && (15)\\ &= \min(\mathbf{1}, \max(\mathbf{0}, \tilde{\boldsymbol{\theta}} - \lambda_2^* \mathbf{1}) && (16)\\ &= \min(\mathbf{1}, \max(\mathbf{0}, \tilde{\boldsymbol{\theta}} - \max(0, \lambda_1^*)\mathbf{1})\,. && (17) \end{aligned}$$