H4. SAE Loss Function The total loss is $\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda \mathcal{L}_{\text{sp}}$, where $\lambda > 0$. $\mathcal{L}_{\text{rec}} = \mathbb{E}_{h \sim \mathcal{D}} \left[\|h - \hat{h}\|_2^2 \right]$. $\mathcal{L}_{\text{sp}} = \mathbb{E}_{h \sim \mathcal{D}} \sum_i |z_i|$. Our analysis will focus on the contributions of z_1, z_2 to $\mathcal{L}_{\text{rec}}$ and $\mathcal{L}_{\text{sp}}$.

H5. Definition of δ-Absorption We define a specific parameterization for the encoder and decoder weights associated with f_1 and f_2 by a parameter $\delta \in [0, 1]$:

- $W_{e,1} = f_1 - \delta f_2$
- $W_{e,2} = f_2$
- $W_{d,1} = f_1$
- $W_{d,2} = f_2 + \delta f_1$

$\delta = 0$ represents no absorption, while $\delta = 1$ represents full absorption.

2. Proposition 1: Perfect Reconstruction under δ-Absorption

For any $\delta \in [0, 1]$, and for inputs h consisting only of f_1, f_2 or 0, the reconstruction $\hat{h} = W_{d,1} z_1 + W_{d,2} z_2$ perfectly reconstructs h, i.e., the reconstruction loss component $\mathcal{L}_{rec}^{(1,2)}$ due to these features is 0.

Proof We consider the possible input types based on f_1, f_2:

Case 1: $h = f_1$ (only parent feature f_1 is present). The latent activations are:

$$\begin{aligned} z_1 &= \text{ReLU}(W_{e,1} \cdot h) = \text{ReLU}((f_1 - \delta f_2) \cdot f_1) \\ &= \text{ReLU}(f_1 \cdot f_1 - \delta f_2 \cdot f_1) \\ &= \text{ReLU}(1 - \delta \cdot 0) = 1 \quad \text{(by H1)} \\ z_2 &= \text{ReLU}(W_{e,2} \cdot h) = \text{ReLU}(f_2 \cdot f_1) = \text{ReLU}(0) = 0 \quad \text{(by H1)} \end{aligned}$$

The reconstruction is:

$$\hat{h} = z_1 W_{d,1} + z_2 W_{d,2} = 1 \cdot f_1 + 0 \cdot (f_2 + \delta f_1) = f_1$$

Thus, $\hat{h} = h$.

Case 2: $h = f_1 + f_2$ (both parent f_1 and child f_2 are present). The latent activations are:

$$\begin{aligned} z_1 &= \text{ReLU}(W_{e,1} \cdot h) = \text{ReLU}((f_1 - \delta f_2) \cdot (f_1 + f_2)) \\ &= \text{ReLU}(f_1 \cdot f_1 + f_1 \cdot f_2 - \delta f_2 \cdot f_1 - \delta f_2 \cdot f_2) \\ &= \text{ReLU}(1 + 0 - \delta \cdot 0 - \delta \cdot 1) = \text{ReLU}(1 - \delta) \quad \text{(by H1)} \end{aligned}$$

Since $\delta \in [0, 1]$, $1 - \delta \geq 0$, so $z_1 = 1 - \delta$.

$$\begin{aligned} z_2 &= \text{ReLU}(W_{e,2} \cdot h) = \text{ReLU}(f_2 \cdot (f_1 + f_2)) \\ &= \text{ReLU}(f_2 \cdot f_1 + f_2 \cdot f_2) \\ &= \text{ReLU}(0 + 1) = 1 \quad \text{(by H1)} \end{aligned}$$

The reconstruction is:

$$\begin{aligned} \hat{h} &= z_1 W_{d,1} + z_2 W_{d,2} = (1 - \delta) f_1 + 1 \cdot (f_2 + \delta f_1) \\ &= (1 - \delta) f_1 + f_2 + \delta f_1 = f_1 - \delta f_1 + f_2 + \delta f_1 = f_1 + f_2 \end{aligned}$$

Thus, $\hat{h} = h$.

Case 3: $h = 0$ (neither f_1 nor f_2 is present).

$$\begin{aligned} z_1 &= \text{ReLU}((f_1 - \delta f_2) \cdot 0) = 0 \\ z_2 &= \text{ReLU}(f_2 \cdot 0) = 0 \end{aligned}$$

The reconstruction is:

$$\hat{h} = 0 \cdot f_1 + 0 \cdot (f_2 + \delta f_1) = 0$$

Thus, $\hat{h} = h$.

Case 4: $h = f_2$ **(only child feature** f_2 **is present).** This case is disallowed by assumption H2 ($p_{01} = 0$), as $f_2 \subset f_1$ implies f_1 must be present if f_2 is.

In all permissible cases, $h - \hat{h} = 0$, so $\|h - \hat{h}\|_2^2 = 0$. Therefore, the reconstruction loss component due to f_1, f_2, denoted $\mathcal{L}_{\text{rec}}^{(1,2)}$, is 0 for any $\delta \in [0, 1]$.

3. Proposition 2: Sparsity Loss under δ-Absorption

The expected sparsity loss contribution from latents z_1 and z_2, denoted $\mathcal{L}_{sp}^{(1,2)} = \mathbb{E}_{h\sim\mathcal{D}}[|z_1| + |z_2|]$, is given by:

$$\mathcal{L}_{sp}^{(1,2)} = p_{11}(2 - \delta) + p_{10}$$

Furthermore, its derivative with respect to δ is:

$$\frac{d\mathcal{L}_{sp}^{(1,2)}}{d\delta} = -p_{11}$$

Proof We calculate the sum of absolute latent activations $|z_1| + |z_2|$ for each case from Proposition 1 and weight them by their probabilities (H2):

- If $h = f_1 + f_2$ (probability p_{11}): $z_1 = 1 - \delta$, $z_2 = 1$. Since $\delta \in [0, 1]$, $1 - \delta \geq 0$, so $|z_1| = 1 - \delta$. $|z_2| = 1$. Thus, $|z_1| + |z_2| = (1 - \delta) + 1 = 2 - \delta$.
- If $h = f_1$ (probability p_{10}): $z_1 = 1$, $z_2 = 0$. Thus, $|z_1| + |z_2| = 1 + 0 = 1$.
- If $h = 0$ (probability p_{00}, neither f_1 nor f_2 present): $z_1 = 0$, $z_2 = 0$. Thus, $|z_1| + |z_2| = 0 + 0 = 0$.

The case corresponding to p_{01} does not occur.

The expected sparsity loss from z_1, z_2 is:

$$\begin{aligned} \mathcal{L}_{\text{sp}}^{(1,2)} &= p_{11} \cdot (2 - \delta) + p_{10} \cdot 1 + p_{00} \cdot 0 \\ &= p_{11}(2 - \delta) + p_{10} \end{aligned}$$

Taking the derivative with respect to δ:

$$\frac{d\mathcal{L}_{\text{sp}}^{(1,2)}}{d\delta} = \frac{d}{d\delta}(2p_{11} - \delta p_{11} + p_{10})$$

Since p_{11} and p_{10} are constants with respect to δ:

$$\frac{d\mathcal{L}_{\text{sp}}^{(1,2)}}{d\delta} = -p_{11}$$

□

4. Corollary: Increasing Absorption Decreases Sparsity Loss

If $p_{11} > 0$ (i.e., the child feature f_2 co-occurs with f_1 with non-zero probability), then increasing δ strictly decreases $\mathcal{L}_{sp}^{(1,2)}$.

Proof From Proposition 2, $\frac{d\mathcal{L}_{\text{sp}}^{(1,2)}}{d\delta} = -p_{11}$. If $p_{11} > 0$, then $-p_{11} < 0$. A negative derivative implies that $\mathcal{L}_{\text{sp}}^{(1,2)}$ is a decreasing function of δ for $\delta \in [0, 1]$. The minimum value of $\mathcal{L}_{\text{sp}}^{(1,2)}$ over this interval occurs at $\delta = 1$ (full absorption), yielding $\mathcal{L}_{\text{sp}}^{(1,2)}(\delta = 1) = p_{11}(2 - 1) + p_{10} = p_{11} + p_{10}$. The maximum value occurs at $\delta = 0$ (no absorption), yielding $\mathcal{L}_{\text{sp}}^{(1,2)}(\delta = 0) = p_{11}(2 - 0) + p_{10} = 2p_{11} + p_{10}$. □

5. Conclusion

Given the specified δ-absorption mechanism for an SAE handling two hierarchical features f_1, f_2 (where $f_2 \subset f_1$):

1. Perfect reconstruction of inputs composed of f_1 and f_2 is maintained irrespective of the degree of absorption δ. Thus, $\mathcal{L}_{\text{rec}}^{(1,2)}$ is unaffected by δ.
2. The sparsity loss component $\mathcal{L}_{\text{sp}}^{(1,2)}$ associated with these features is $p_{11}(2-\delta)+p_{10}$.
3. If $p_{11} > 0$, the total loss $\mathcal{L}$ (focusing on the components related to f_1, f_2) decreases as δ increases because $\mathcal{L}_{\text{rec}}^{(1,2)}$ is constant (zero) and $\mathcal{L}_{\text{sp}}^{(1,2)}$ decreases.
4. Therefore, an optimization process like gradient descent, when minimizing the total loss $\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda \mathcal{L}_{\text{sp}}$ (where $\lambda > 0$), will favor increasing δ towards 1, thereby promoting feature absorption for these hierarchically related features, assuming the SAE learns or is constrained to these forms of W_e and W_d.

This formalizes the argument that, under the given conditions and definitions, absorption is a mechanism that can reduce SAE loss by improving sparsity without harming reconstruction for hierarchical features.

A.3 Extended toy model experiments

In this section we explore further variants on absorption in toy models. We use the same setting as our main toy model experiment, with four mutually-orthogonal true features, and train an SAE with four latents. Each true feature $f \in \mathbb{R}^{50}$. Unless otherwise stated, every time feature 1 fires feature 0 must also fire, but feature 0 is allowed to fire on its own as well. This is to simulate hierarchal features such as our example "starts with S" and "short" features, where every time the "short" feature fires we expect "starts with S" must also fire since "short" starts with "S", but "starts with S" can fire on its own as well. Feature 2 and 3 are fully independent. All features fire with magnitude 1.0 and variance 0.0 unless otherwise stated.

Magnitude variance causes partial absorption In our main toy model experiment, each true feature fires with magnitude exactly 1.0. This is not very realistic, though - likely there will be some variance in feature firing magnitudes in real LLMs. We simulate this by adding variance of 0.1 to the firing magnitude of feature 0, so the relative magnitudes of feature 0 and 1 are no longer fixed. We show the plots of cosine similarity between SAE encoder and decoder in Figure 10. Here, we still see the same absorption pattern in the SAE encoder and decoder with the latent 3 encoder containing a negative component of feature 1, and the latent 0 decoder merging features 0 and 1. We show some sample true feature firings and corresponding SAE latent activations in Table 2.

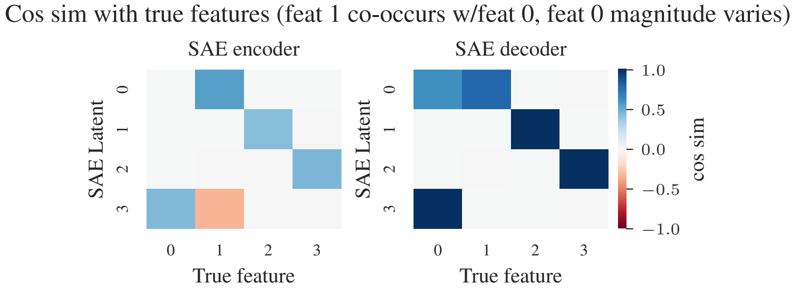

Figure 10: SAE encoder and decoder with true features. The firing magnitude of feature 0 has variance 0.1, while the remaining features fire with variance 0.0. When there is variance in the firing magnitudes of parent and child features, we still see an absorption pattern in the SAE encoder and decoder with the latent 3 encoder containing a negative component of feature 1, and the latent 0 decoder merging features 0 and 1.

We see that now the SAE latent tracking feature 0 still fires when the true values of features 0 and 1 are both 1.0, but very weakly. However, if the magnitude of feature 0 drops down to 0.75, then the feature 0 latent fully turns off.