InnerPiSSA Equations

Adapter

$$y = xW_{\text{res}} + xVR(\alpha)(S + \alpha \cdot \Delta S)U^T$$

where:
- $R(\alpha)$ is the Cayley rotation that reverses with sign: $R(-\alpha) = R(\alpha)^T$
- ΔS is learnable singular value scaling
- α is the steering coefficient

Loss Function (Compact)

$$\mathcal{L} = \underbrace{-\text{proj}(\Delta h_\pi, v)}_{\text{separation}} + \underbrace{\sum_t \mathbb{1}[\Delta\ell_t > \tau] \cdot \Delta\ell_t}_{\text{coherence constraint}} + \underbrace{\text{hinge}(\delta_{-1}, \delta_0, \delta_{+1})}_{\text{monotonic constraint}}$$

where:
- $\Delta h_\pi = h_{\text{cho}} - h_{\text{rej}}$ (hidden state difference)
- v is the frozen PCA preference direction
- $\Delta\ell_t = \ell_t^\pi - \ell_t^{\text{ref}}$ is per-token NLL degradation
- $\delta_\alpha = \log p(\text{cho}) - \log p(\text{rej})$ at coefficient α

Loss Function (Expanded)

$$\mathcal{L} = \underbrace{-\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \frac{\left(h^i_{\text{cho}} - h^i_{\text{rej}}\right)^T v}{\|v\|_2}}_{\text{separation along preference direction}}$$

$$\underbrace{+\sum_{t=1}^{T} \mathbb{1}\left[\ell_t^\pi - \ell_t^{\text{ref}} > \tau\right] \cdot \left(\ell_t^\pi - \ell_t^{\text{ref}}\right)}_{\text{coherence constraint}}$$

$$\underbrace{+\max(0, \delta_{-1} - \delta_0) + \max(0, \delta_0 - \delta_{+1})}_{\text{monotonic constraint}}$$

where:
- $h_{\text{cho}}, h_{\text{rej}}$ are hidden states from contrastive prompt prefixes
- $v = \text{PCA}\left(h_{\text{ref,cho}} - h_{\text{ref,rej}}\right)$ is frozen reference direction
- $\ell_t = -\log p(x_t \mid x_{<t})$ is per-token NLL
- $\delta_\alpha = \log p(\text{cho}) - \log p(\text{rej})$ at coefficient α
- Loss computed jointly for $\alpha \in \{-1, +1\}$ to enforce bidirectionality

Monotonic Constraint Detail

The monotonic constraint ensures:

$$\delta_{-1} < \delta_0 < \delta_{+1}$$

This prevents saddle points where both steering directions degrade performance.