

1 Statement

The goal of this analysis is to understand the shared representations model.

We have words in a vocabulary:

$$\{\mathbf{x}_1, \dots, \mathbf{x}_V\} \subseteq \mathbb{R}^D,$$

and a set of L languages, indexed by $[L] = \{1, \dots, L\}$.

We train a gated deep linear network (GDLN) to “translate” words from one language into another. The gated deep linear network is represented by a collection of L input and output weight matrices of dimensions $D \times H$ and $H \times D$, respectively.

$$\hat{y}(\mathbf{x}) = \sum_{l' \in [L]} g_{l'}(\mathbf{x}) \mathbf{W}_{l'} \left(\sum_{l \in [L]} g_l(\mathbf{x}) \mathbf{W}_l \mathbf{x} \right) = \sum_{l, l' \in [L]} g_l(\mathbf{x}) g_{l'}(\mathbf{x}) \mathbf{W}_{l'} \mathbf{W}_l \mathbf{x}.$$

Let us consider an alternative way to write the gating mechanism:

$$\sum_{l \in [L]} g_l(\mathbf{x}) \mathbf{W}_l \mathbf{x} = [\mathbf{g}(\mathbf{x}) \otimes I_D]^\top \mathbf{W} \mathbf{x},$$

where

$$\mathbf{g}(\mathbf{x}) = \begin{pmatrix} g_1(\mathbf{x}) \\ \vdots \\ g_L(\mathbf{x}) \end{pmatrix} \quad \text{and} \quad \mathbf{W} = \begin{pmatrix} \mathbf{W}_1 \\ \vdots \\ \mathbf{W}_L \end{pmatrix}.$$

2 Thoughts

1. Is the loss for Gaussian data the same as the loss for one-hots? I think it may be.
2. For Gaussian data, lowering the initialization scale lowers the generalization threshold. Does this imply that as the initialization scale goes to zero, the generalization threshold goes to zero? To investigate this, consider $T = I_L$ and take $\sigma \rightarrow 0$. Can we describe the dynamics of the model in this case?