1 Statement

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

We have words in a vocabulary:

$$\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_V\} \subseteq \mathbb{R}^D,$$

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

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.

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

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

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

where

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