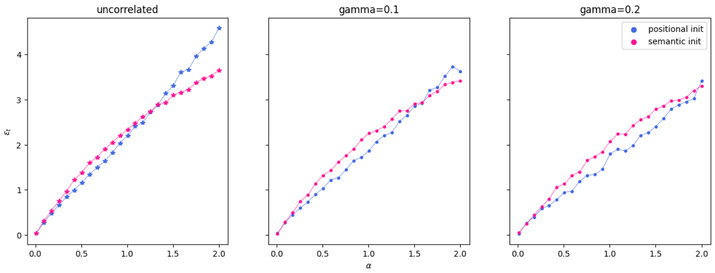

Figure 1: **Uncorrelated vs. correlated inputs** ($\omega = 0.3, \sigma = 0.5, \lambda = 0.001$)**.** In the main text we sample all datapoints $\mathbf{x} \in \mathbb{R}^{d\times L=2}$ such that the columns are independent. We compare this setting with a correlated data structure with a hidden latent: We sample three vectors $u, v, w \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$, and set the first column of x to be $(\sigma u + \gamma v)/\sqrt{\sigma + \gamma}$ and the second one $(\sigma w + \gamma v)/\sqrt{\sigma + \gamma}$. Experiments are repeated 5 times per data point with $d = 1000$.

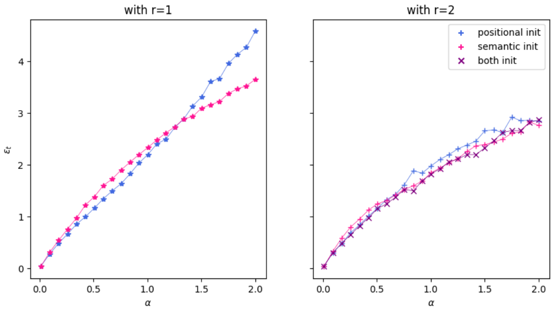

Figure 2: **Rank 1 vs. rank 2 student** ($\omega = 0.3, \sigma = 0.5, \lambda = 0.001$)**.** We compare different initializations of the higher-rank student. Positional is both columns of the student matrix $\hat{Q}$ are initialized using the positional strategy. We do the same for the semantic strategy. The 'both' initialization initializes one column using the positional strategy and one using the semantic strategy. Experiments are repeated 5 times per data point with $d = 1000$.

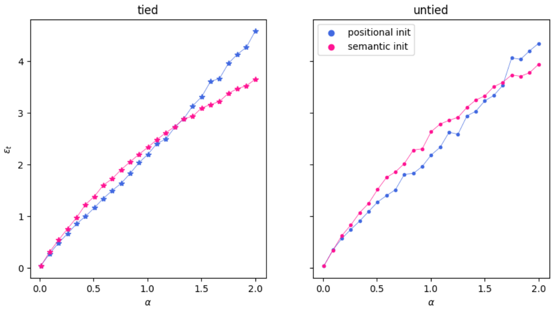

Figure 3: **Tied vs. independent weight Q, K** ($\omega = 0.3, \sigma = 0.5, \lambda = 0.001$)**.** We compare the student setting from the paper, where the query and key matrices are bound to each other with the setting where we set them independently. We initialize with them both being either close to the positional or close to the semantic initialization. The phase transition for the semantic minimum dominating moves to the left, i.e. more samples are now needed. Experiments are repeated 5 times per data point with $d = 1000$.

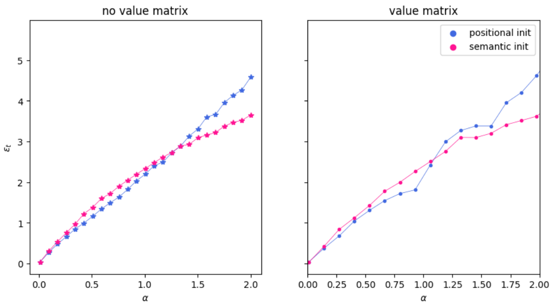

Figure 4: **No value matrix vs. value matrix.** ($\omega = 0.3, \sigma = 0.5, \lambda = 0.001$)**.** We compare the setting from the main text with adding a value matrix, i.e. a trainable parameter $V \in \mathbf{R}^{d\times d}$. This is applied to every embedding before they are averaged over using the attention matrix. We ran the experiment 5 times for each α with $d = 500$, and rescaled the training error to compare to the experiments in the main, where $d = 1000$.