B.9. Restatement of Theorem 3.1

Here, we state Theorem 3.1 when parameters v irrelevant to the symmetry is explicitly shown. There is no essential difference from the original version, and the proof is the same.

Theorem B.3. *Let u, w, and v be weight vectors of arbitrary dimensions. Let $\ell(u,w,v,x)$ satisfy $\ell(u,w,v,x) = \ell(\lambda u, w/\lambda, v, x)$ for arbitrary x and any $\lambda \in \mathbb{R}_+$. Then,*

$$\frac{d}{dt}(||u||^2 - ||w||^2) = -T(u^T C_1 u - w^T C_2 w), \tag{108}$$

where $C_1 = \mathbb{E}[A^T A] - \mathbb{E}[A^T]\mathbb{E}[A]$, $C_2 = \mathbb{E}[AA^T] - \mathbb{E}[A]\mathbb{E}[A^T]$ and $A_{ki} = \partial\tilde{\ell}/\partial(u_i w_k)$ with $\tilde{\ell}(u_i w_k, v, x) \equiv \ell(u_i, w_k, v, x)$.[7]

B.10. Derivation of Eq. (7)

We here prove inequality (7). At stationarity, $d(\|u\|^2 - \|w\|^2)/dt = 0$, indicating

$$\lambda_{1M}\|u\|^2 - \lambda_{2m}\|w\|^2 \geq 0,\ \lambda_{1m}\|u\|^2 - \lambda_{2M}\|w\|^2 \leq 0. \tag{109}$$

The first inequality in Eq. (109) gives the solution

$$\frac{\|u\|^2}{\|w\|^2} \geq \frac{\lambda_{2m}}{\lambda_{1M}}. \tag{110}$$

The second inequality in Eq. (109) gives the solution

$$\frac{\|u\|^2}{\|w\|^2} \leq \frac{\lambda_{2M}}{\lambda_{1m}}. \tag{111}$$

Combining these two results, we obtain

$$\frac{\lambda_{2m}}{\lambda_{1M}} \leq \frac{\|u\|^2}{\|w\|^2} \leq \frac{\lambda_{2M}}{\lambda_{1m}}, \tag{112}$$

which is Eq. (7).

B.11. Additional Experiment

As an example application of our theory, we perform an experiment with the simplest version of the transformer, a two-layer single-head self-attention network without MLP in between. Here, the input dimension is 50×6 such that for each data point X, elements of $X_{:,1:5}$ are i.i.d. from $\mathcal{N}(0,1)$, and the target

$$X_{1:49,6} = X_{1:49,1:5}w^* + \epsilon, \tag{113}$$

where $w^* \in \mathbb{R}^5$ is a ground truth vector, generated also from $\mathcal{N}(0, I_5)$, and ϵ is an i.i.d. noise for each data point. The tasks are the simplest type of in-context learning, where the first 49 vectors serve as demonstrations of feature-target pairs, and the last row of X is the feature that the model needs to predict, whose label is $X_{:,1:5}w^*$. Following this data generation procedure, we train in the online setting, and the training proceeds with SGD with a learning rate of 4×10^{-3} with a batch size of 200 and without weight decay.

For this problem, 5 independent rescaling symmetries exist between the query and key matrices of each of the two self-attention layers:

$$\ell(W_K W_Q) = \ell(\sum_i^5 W_K^i (W_Q^i)^T) \tag{114}$$

where W_K and W_Q are matrices in $\mathbb{R}^{6\times 6}$ and W^i denotes the i-th column of W_K and W_Q. Therefore, for every $i \in [6]$, there is a rescaling symmetry between W_K^i and $(W_Q^i)^T$. According to our theory, there are 6 quantities that will be balanced at the end of training.

See Figure 6. We plot the first (denoted as g_K^i) and second quantity (g_Q^i) of the right-hand side of Eq. (35) for each W_K^i and $(W_Q^i)^T$. We see that as the training proceeds, the two quantities become closer and ultimately overlap within an acceptable range of uncertainty.

[7] Our result also holds if we consider the effect of a finite step-size by using the modified loss (See Appendix B.7).

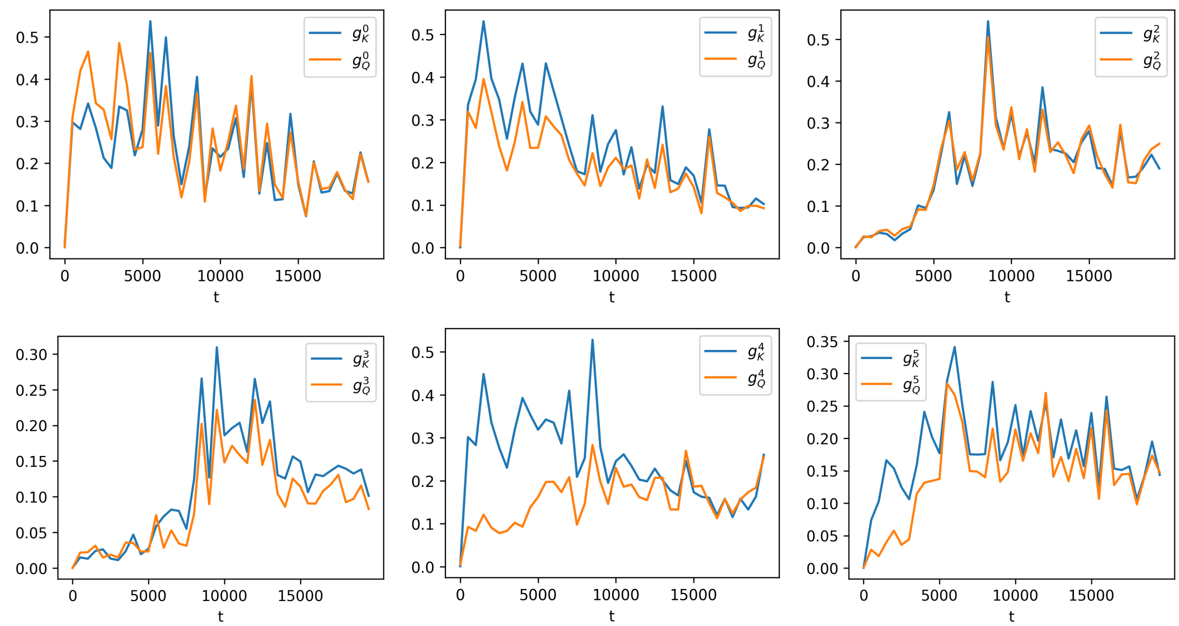

Figure 6: Evolution of the two quantities in Eq. (35) during training. The difference between the two quantities becomes smaller and smaller during training and overlaps well at the end of training.