# A. Experimental details

## A.1. CIFAR-10

The CIFAR-10 datasest (Krizhevsky, 2009) was fetched using the TensorFlow datasets[5].

In all of the CIFAR-10 experiments, the data was preprocessed by subtracting mean and dividing by a standard deviation for each pixel and data point separately (equivalent to using `LayerNorm` as the first layer). We inflated all of the standard deviations by $10^{-15}$ to avoid division by zero.

All the classification tasks were converted into regression tasks by encoding the targets as $C$–dimensional vectors, where $C$ is the number of classes, with the entry corresponding to the correct label set to $\frac{C-1}{C}$ and all other entries to $-\frac{1}{C}$. This enabled us to perform closed form NNGP and NTK inference using the Gaussian likelihood/MSE loss.

### A.1.1. HYPERPARAMETER SEARCH

The hyperparameter search was on a fixed architecture with 8x Convolution + ReLU, Attention, Flatten, and a Dense readout layer. We used $1.7562$ and $0.1841$ respectively for the weight and bias variances as in (Novak et al., 2019, appendix G.1) except for the attention output variance $\sigma_O^2$ which was set to one. The convolutional layers were used with the `SAME` padding, stride one, and filter size $3 \times 3$. For attention kernels with positional encodings, the reported $\rho$ parameter (Equation (12)) is actually $\rho/(\sigma_Q^2 \sigma_K^2)$ so that the relative scale of the contribution of $R$ remains the same with changing $\sigma_Q^2 \sigma_K^2$.

There were two stages of the hyperparameter search, first to identify the most promising candidates (Table 5), and second to refine the parameters of these candidate kernels (Table 6). The second stage also included the *residual attention kernel* (Equation (14)); the $\alpha$ in the second table should thus be interpreted as the one stated in Equation (14) (cf. Appendix D). The best hyperparameters used in Figure 4 and Table 2 can be found in a bold typeset in Table 6.

All computation was done in 32-bit precision, and run on up to 8 NVIDIA V100 GPUs with 16Gb of RAM each.

*Table 5.* Hyperparameter values for the first stage of search. VALUE POSITIONAL ENCODING stands for whether the positional encodings should be added to all $Q$, $K$, and $V$ (TRUE), or only to the inputs of $Q$ and $K$ (FALSE; see Section 4.2.2). ENCODINGS COVARIANCE represent whether positional encodings should be added (0 for no), and if so, what should their initialisation covariance be ($I$ for identity, and $R$ for the covariance defined in Equation (12)). $\zeta =$ softmax was only used when QUERY/KEY SCALING was $d^{-1}$ (see Section 4). $\varphi, \rho, \alpha$ were skipped when VALUE POSITIONAL ENCODING was FALSE, and $\varphi$ was only varied when ENCODINGS COVARIANCE was $R$.

| HYPERPARAMETER | VALUES |
|---|---|
| QUERY/KEY SCALING | $\{d^{-1/2}, d^{-1}\}$ |
| $\zeta$ (SECTION 4.1) | {SOFTMAX, IDENTITY} |
| VALUE POSITIONAL ENCODING | {TRUE, FALSE} |
| ENCODINGS COVARIANCE | $\{0, I, R\}$ |
| $\varphi$ (EQUATION (12)) | $\{1, 5\}$ |
| $\rho$ (EQUATION (12)) | $\{1\}$ |
| $\alpha$ (EQUATION (9)) | $\{0.5, 0.8\}$ |
| $\sigma_Q \cdot \sigma_K$ | $\{0.1, 1.0\}$ |

*Table 6.* Hyperparameter values for the first stage of search. See Table 5 for description of the individual hyperparameters. The parameters that achieved the best *NNGP* validation accuracy and were selected for the subsequent experiments are in a bold typeset.

| HYPERPARAMETER | STRUCT | RESIDUAL |
|---|---|---|
| QUERY/KEY SCALING | $\{\boldsymbol{d^{-1}}\}$ | $\{\boldsymbol{d^{-1}}\}$ |
| $\zeta$ (SECTION 4.1) | {**SOFTMAX**} | {**IDENTITY**} |
| VALUE POSITIONAL ENCODING | {TRUE, **FALSE**} | – |
| ENCODINGS COVARIANCE | $\{\boldsymbol{R}\}$ | $\{\boldsymbol{R}\}$ |
| $\varphi$ (EQUATION (12)) | $\{1, \mathbf{5}, 10\}$ | $\{1, \mathbf{5}, 10\}$ |
| $\alpha$ (EQUATIONS (9) AND (14)) | $\{\mathbf{0.4}, 0.5, 0.65, 0.8, 0.9\}$ | $\{0.4, 0.5, \mathbf{0.65}, 0.8, 0.9\}$ |
| $\rho$ (EQUATION (12)) | $\{0.5, 1, \mathbf{1.5}\}$ | $\{\mathbf{0.5}, 1, 1.5\}$ |
| $\sigma_Q \cdot \sigma_K$ | $\{0.001, \mathbf{0.1}, 1.0\}$ | – |

[5] https://www.tensorflow.org/datasets/catalog/cifar10

#### A.1.2. DETAILS FOR FIGURE 2

The downsampling was performed using `skimage.transform.resize` with parameters `mode="reflect"` and `anti_aliasing=True`, using downsampled height and width of size 8 as mentioned.

Both the convergence and accuracy plots are for the $d^{-1/2}$ vanilla NNGP kernel with $\zeta =$ softmax. The intractable softmax integral of the limiting covariance function was estimated using MC integration with 2048 samples.

We used $1.7562$ and $0.1841$ respectively for the weight and bias variances as in (Novak et al., 2019, appendix G.1) for all the convolutional and dense layers, $1.7562$ for the $\sigma_K^2, \sigma_Q^2$ and $\sigma_V^2$, and $\sigma_O^2 = 1$. The convolutional layer used `VALID` paddingstride one, and filter size $3 \times 3$.

As in (Novak et al., 2019), The reported distance between kernel matrices is the logarithm of

$$\frac{\|\hat{\mathcal{K}} - \mathcal{K}\|_F^2}{\|\mathcal{K}\|_F^2} \,, \tag{15}$$

where $\hat{\mathcal{K}}$ and $\mathcal{K}$ are respectively the empirical and the predicted theoretical covariance matrices for the training set.

All computation was done in 32-bit precision, and run on up to 8 NVIDIA V100 GPUs with 16Gb of RAM each.

#### A.1.3. DETAILS FOR FIGURE 3

We used a 45K/5K train/validation split of the usual 50K CIFAR-10 training set and reported the validation set accuracy after training for 1000 epochs with batch size 64 and the Adam optimiser.

The attention layers used the usual $d^{-1/2}$ scaling of the query/key inner products, and the convolutional layers used the `SAME` padding, stride one, and filter size $3 \times 3$. We used $2.0$ and $10^{-2}$ respectively for the weight and bias variances except in the attention where $\sigma_Q^2 = \sigma_K^2 = \sigma_V^2 = 2$ but $\sigma_O^2 = 1$. Further, we used the `append` type positional encodings (Section 4.2) with the same embedding dimension as N_CHANNELS (Table 7), thus doubling the embedding dimension of the attention layer inputs.

All computation was done in 32-bit precision, and run on a single NVIDIA V100 GPU with 16Gb of RAM each.

*Table 7.* Hyperparameter values for which results are reported in Figure 3. N_CHANNELS is the number of channels used in the convolutional layers. The same number was used for $d_n^{\ell,G}$ and output dimension in the attention layer, but $d_n^{\ell,H} = d_n^{\ell,V} = \lfloor \text{N\_CHANNELS} \rfloor$ to reduce the memory footprint. The learning rate was fixed throughout the training, relying only on Adam to adapt step size. Each configuration was run with three random seeds and each of the corresponding results was included in the appropriate column in Figure 3.

| HYPERPARAMETER | VALUES |
|---|---|
| $\zeta$ (ATTENTION) | {RELU, ABS, SOFTMAX} |
| LAYERNORM | {NONE, PER_HEAD , AT_OUTPUT} |
| N_CHANNELS | $\{32, 192\}$ |
| LEARNING RATE | $\{10^{-3}, 10^{-2}\}$ |

#### A.1.4. DETAILS FOR FIGURE 4

We used $1.7562$ and $0.1841$ respectively for the weight and bias variances as in (Novak et al., 2019, appendix G.1) except for the attention output variance $\sigma_O^2$ which was set to one. The convolutional layers were used with the `SAME` padding, stride one, and filter size $3 \times 3$. For the vanilla attention kernels, we report the best performance over $\sigma_Q \sigma_K = \{10^{-3}, 10^{-1}, 1, 2, 10\}$ at each depth. The `Struct` and `Residual` were used with the best hyperparameters found during hyperparameter search as reported in Appendix A.1.1.

All computation was done in 32-bit precision, and run on up to 8 NVIDIA V100 GPUs with 16Gb of RAM each.

#### A.1.5. DETAILS FOR TABLE 2

The best set-up from Appendix A.1.1 was used (including the best hyperparameters as stated in Table 6).

All computation was done in 64-bit precision, and run on up to 8 NVIDIA V100 GPUs with 16Gb of RAM each.

## A.2. IMDb

### A.2.1. General settings for Table 3 and Table 4.

The IMDb reviews dataset (Maas et al., 2011) was fetched using TensorFlow datasets[6].

All sentences were truncated or padded to 1000 tokens using the default settings of `tf.keras.preprocessing.text.Tokenizer`[7]. No words were removed from the embedding model dictionary. Tokens were embedded using GloVe embeddings (Pennington et al., 2014) with no other pre-processing. Binary targets were mapped to $\{-0.5, 0.5\}$ values. Diagonal regularizers for inference were selected based on validation performance among the values of $10^{-7}, 10^{-6}, \ldots, 1$ multiplied by the mean trace of the kernel.

When applicable, all models used ReLU nonlinearities, `Struct` (Structured positional encoding, $d^{-1}$ scaling, Table 1) kernel with $\zeta$ being the row-wise softmax function (Equation (18)), decaying positional embeddings used only for the attention keys and queries, with $\varphi = 2.5$ (Equation (12)), $\alpha = 0.75$, and $\rho = 1$ (Equation (9)). These parameters were selected based on preliminary experiments with CIFAR-10, and fine-tuning on IMDb specifically is an interesting avenue for future research.

All preliminary and validation experiments were carried out in 32-bit precision, while test evaluation (reported in the Table 3 and Table 4) were done in 64-bit precision. All experiments were run on machines with up to 8 NVIDIA V100 GPUs with 16Gb of RAM each.

### A.2.2. Details for Table 3

Words were embedded using GloVe 840B.300d embeddings.

The embedding model was selected on a small-scale experiment (4000 train and 4000 validation sets) among GloVe 6B 50-, 100-, 200-, and 300-dimensional variants, as well as GloVe 840B.300d, and 1024-dimensional ELMO (Peters et al., 2018) embeddings (using TensorFlow Hub[8]). In this preliminary experiment, GloVe 840B.300d, GloVe6B.300d, and ELMO.1024d performed similarly, and GloVe 840B.300d was chosen for the full dataset experiment.

The validation experiment was run on the 25K training set partitioned into a 15K and 10K training and validation sets, with the best models then evaluated on the 25K training and 25K test sets.[9]

All layers used weight and bias variances 2 and $0.01$ respectively, expect for attention outputs and values variances which were set to 1, and the top linear readout layer with weight variance 1 and no bias.

Three classes of models were considered:

1. `GAP-only`, doing only global average pooling over inputs followed by the linear readout.
2. `GAP-FCN`, in which GAP was followed by 0, 1, or 2 fully connected layers.
3. `Struct`, allowing the same models as `GAP-FCN`, except for necessarily having an attention layer before GAP.

Each class could also have an optional `LayerNorm` layer following GAP. The best model from each class was then evaluated on the test set.

### A.2.3. Details for Table 4

All convolutional layers used the total window (context) size of 9 tokens, stride 1, and `SAME` (zero) padding.

Experiments were run on a 3200/1600/1600 train/validation/test splits. Four classes of models were considered:

1. `GAP-only`, identical to the one in Appendix A.2.2.

[6] https://www.tensorflow.org/datasets/catalog/imdb_reviews
[7] https://www.tensorflow.org/api_docs/python/tf/keras/preprocessing/text/Tokenizer
[8] https://tfhub.dev/google/elmo/3
[9] Precisely, subsets of sizes 14880/9920 and 24960/24960 were used to make the dataset be divisible by 8 (the number of GPUs) times 20 (the batch size), which is a technical limitation of the Neural Tangents (Novak et al., 2020) library.

2. `GAP-FCN`, also identical to the one in Appendix A.2.2.
3. `CNN-GAP`, allowing the same models as in `GAP-FCN`, but having GAP preceeded by 0, 1, 2, 4, or 8 CNN layers.
4. `Struct`, allowing the same models as in `CNN-GAP`, but having 1 or 2 attention layers (each optionally followed by `LayerNorm` over channels) before GAP. If the model also had CNN layers, attention and CNN layers were interleaved, attention layers being located closer to GAP (for example, a model with 8 CNN layers and 2 attention layers would have 7 CNN layers followed by attention, CNN, attention, GAP).

All models were allowed to have either ReLU or Erf nonlinearity, with weight and bias variances set to 2 and 0.01 for ReLU, and 1.7562 and 0.1841 for Erf, with the same values used by attention keys and queries layers, but having variance 1 for values and output layers. The readout linear layer had weight variance 1 and no bias.

*Table 8.* Best validation accuracy for various finite attention architectures. The reported numbers are an average over three random seeds.

| | SOFTMAX | RELU | IDENTITY |
|---|---|---|---|
| NONE | 64.10 | 69.68 | 70.46 |
| PER_HEAD | 68.96 | 77.40 | 75.28 |
| AT_OUTPUT | 71.70 | 79.00 | 79.56 |

# B. Proofs

**Assumptions:** We assume the input set $\mathcal{X} \subset \mathbb{R}^{\mathbb{N} \times d^0}$ is *countable*, and the usual *Borel product $\sigma$-algebra* on any of the involved countable real spaces (inputs, weights, outputs of intermediary layers). We also assume that the nonlinearities $\phi$ and $\zeta$ are *continuous* and (entrywise) *polynomially bounded*, i.e., $|\phi(z)| \leq \sum_{t=0}^{m} c_t |z|^t$ for some $m \in \mathbb{N}$ and $c_0, \ldots, c_m \in \mathbb{R}_+$ independent of $z$,[10] and $|\zeta(G)_{ai}| \leq \sum_{t=0}^{m} c_0 |G_{ai}|^t$ for some $m \in \mathbb{N}$ and $c_0, \ldots, c_M \in \mathbb{R}_+$ independent of $G$. For the NTK proofs, we further assume that $\nabla\phi$ and $\nabla\zeta$ are continuous bounded almost everywhere, where for ReLU, Leaky ReLU, or similar, we set $\nabla\phi(0) \coloneqq \lim_{z \to 0^-} \nabla\phi(z)$ which for ReLU/Leaky ReLU is equal to zero.

As Matthews et al. (2018), we will need to use the 'infinite width, finite fan-out' construction of the sequence of NNs. In particular, we will assume that for any attention layer $\ell \in [L+1]$ and $n \in \mathbb{N}$, the output is computed as defined in Equation (3), but we will add a countably infinite number of additional heads which do not affect the output of the $n^{\text{th}}$ network, but are used by wider networks, i.e., each head $h > d_n^{\ell,H}$ is only used to compute the outputs by networks with index $m \in \mathbb{N}$ such that $d_m^{\ell,H} \geq h$. Similar construction can be used for fully connected, convolutional, and other types of layers as demonstrated in (Matthews et al., 2018; Garriga-Alonso et al., 2019). Since the outputs remain unchanged, a proof of convergence of the 'infinite width, finite fan-out networks' implies convergence of the standard finite width networks, and thus the construction should be viewed only as an analytical tool which will allow us to treat all the random variables

$$\{f_{n,ij}^{\ell}(x), f_{n,ij}^{\ell h}(x) \colon n, h, i, j \in \mathbb{N}, \ell \in [L+1], x \in \mathcal{X}\}\,,$$

as defined on the same probability space, and thus allows us to make claims about convergence in probability and similar.

Finally, we will be using the *NTK parametrisation* (Jacot et al., 2018) within the NTK convergence proofs, i.e., we implicitly treat each weight $W_{ij} \sim \mathcal{N}(0, \sigma^2/d)$, i.i.d., as $W = \frac{\sigma}{\sqrt{d}} \widetilde{W}$ where only $\widetilde{W}$ is trainable. This parametrisation ensures that not only the forward but also the backward pass are properly normalised; under certain conditions, proofs for NTK parametrisation can be extended to standard parametrisation (Lee et al., 2019).

**Notation:** For random variables $X, (X_n)_{n \geq 1}$, $X_n \rightsquigarrow X$ denotes convergence in distribution, and $X_n \xrightarrow{P} X$ convergence in probability. For vectors $x, y \in \mathbb{R}^m$, $\langle x, y \rangle = \sum_{j=1}^{m} x_j y_j$ denotes the usual inner product, and for matrices $A, B \in \mathbb{R}^{m \times m}$, $\langle A, B \rangle = \langle \text{vect}(A), \text{vect}(B) \rangle = \sum_{i,j=1}^{m} A_{ij} B_{ij}$ denotes the Frobenius inner product. For any $A \in \mathbb{R}^{m \times k}$, we will use

[10] This is a relaxation of the original 'linear envelope' condition $|\phi(z)| \leq c + m|z|$ for some $c, m \in \mathbb{R}_+$, used in (Matthews et al., 2018; Garriga-Alonso et al., 2019) and stated in Theorem 3. We decided to keep the reference to the linear envelope condition in the main text since it is general enough to guarantee convergence for all bounded (e.g., softmax, tanh) and ReLU like (e.g., ReLU, Leaky ReLU, SeLU) nonlinearities, and matches the existing literature with which the readers may already be familiar. Nevertheless, all the presented proofs are valid for the polynomially bounded nonlinearities, similarly to (Yang, 2019b).

$A_{i \cdot} \in \mathbb{R}^{1 \times k}$ and $A_{\cdot j} \in \mathbb{R}^{m \times 1}$ to respectively denote $i^{\text{th}}$ row and $j^{\text{th}}$ column of the matrix. Later on, we will be working with finite subsets $\mathcal{L} \subset \mathcal{X} \times \mathbb{N}$ for which we define the coordinate projections

$$\mathcal{L}_{\mathcal{X}} := \{x \in \mathcal{X} \colon \exists i \in \mathbb{N} \text{ s.t. } (x, i) \in \mathcal{L}\}\,, \qquad \mathcal{L}_{\mathbb{N}} := \{i \in \mathbb{N} \colon \exists x \in \mathcal{X} \text{ s.t. } (x, i) \in \mathcal{L}\}\,.$$

Since Yang (2019b) provides convergence for attention architectures under the $d^{-1}$ only in the NNGP regime, we use $\tau \in \{1, \frac{1}{2}\}$ to refer to the different $d^{-\tau}$ within the NTK proofs. As explained in Section 3.2, the $\tau = 1$ limit is not very interesting when $W^Q$ and $W^K$ are initialised independently with zero mean, and thus we will be assuming $W^Q = W^K$ a.s. whenever $\tau = 1$. Finally, we use $\sigma_{OV} := \sigma_O \sigma_V$, $\sigma_{QK} := \sigma_Q \sigma_K$, $\lesssim$ as 'less then up to a universal constant', $\operatorname{poly}(x_1, \ldots, x_M)$ for a polynomial in $x_1, \ldots, x_m \in \mathbb{R}$, and the shorthand

$$\widetilde{G}^{\ell h}_{n,ai}(x) := \zeta(G^{\ell h}_{n,ai}(x))\,. \tag{16}$$

**Proof technique:** The now common way of establishing convergence of various deep NNs architectures is to inductively prove that whenever a preceding layer's outputs converge in distribution to a GP, the outputs of the subsequent layer converge to a GP too under the same assumptions on the nonlinearities and initialisation (e.g., Matthews et al., 2018; Lee et al., 2018; Novak et al., 2019; Garriga-Alonso et al., 2019; Yang, 2019a;b). We prove this induction step for NNGP under the $d^{-1/2}$ scaling in Theorem 3 (recall that the equivalent result under the $d^{-1}$ is already known due to Yang (2019b)), and for NTK in Theorem 18. As in (Matthews et al., 2018), our technique is based on exchangeability (Lemma 5), and we repeatedly make use of Theorem 29 which says that if a sequence of real valued random variables $(X_n)_{n \geq 1}$ converges in distribution to $X$, and the $(X_n)_{n \geq 1}$ are uniformly integrable (Definition 6 below), then $X$ is integrable and $\mathbb{E}[X_n] \to \mathbb{E}[X]$.

**Lemma 5** (Exchangeability). *For any $n \in \mathbb{N}$, the outputs of an attention layer $f^{\ell}_{n,ai}(x)$ are exchangeable along the $i$ index. Furthermore, each of $f^{\ell h}_n(x), G^{\ell h}_n(x), Q^{\ell h}_n(x), K^{\ell h}_n(x), V^{\ell h}_n(x)$ is exchangeable over the $h$ index, and for a fixed $h$, each of $f^{\ell h}_{n,ai}(x), Q^{\ell h}_{n,ai}(x), K^{\ell h}_{n,ai}(x), V^{\ell h}_{n,ai}(x)$ is exchangeable over the $i$ index.*

**Definition 6** (Uniform integrability). *A collection of real valued random variables $\mathcal{C}$ is called uniformly integrable if for any $\varepsilon > 0$ there exists $c_\varepsilon \geq 0$ s.t. $\mathbb{E}\,|X| \mathbb{1}_{|X| \geq c_\varepsilon} \leq \varepsilon$ for all $X \in \mathcal{C}$ simultaneously.*

*Proof of Lemma 5.* Recall that by the de Finnetti's theorem, it is sufficient to exhibit a set of random variables conditioned on which the set of random variables becomes i.i.d. This is is trivial for the columns of $f^{\ell}_{n,ai}(x)$ as we can simply condition on $\{f^{\ell h}_{n,ai}(x) \colon h \in [d^{\ell,H}_n]\}$. The remainder of the claims can be obtained by observing that

$$f^{\ell h}_n(x) = \zeta\left( \frac{1}{\sqrt{d^{\ell,G}_n}} g^{\ell-1}_n(x) W^{\ell h,Q}_n (g^{\ell-1}_n(x) W^{\ell h,K}_n)^\top \right) g^{\ell-1}_n(x) W^{\ell h,V}_n\,,$$

and thus if we condition on $g^{\ell-1}_n(x)$, the variables associated with individual heads are i.i.d. ∎

### B.1. $d^{-1/2}$ NNGP convergence proof

**Theorem 3** ($d^{-1/2}$ limit). *Let $\ell \in \{2, \ldots, L+1\}$, and $\phi$ be such that $|\phi(x)| \leq c + m|x|$ for some $c, m \in \mathbb{R}_+$. Assume $f^{\ell-1}_n$ converges in distribution to $f^{\ell-1} \sim \mathcal{GP}(0, \kappa^{\ell-1})$, such that $f^{\ell-1}_{\cdot j}$ and $f^{\ell-1}_{\cdot k}$ are independent for any $j \neq k$, the variables $\{f^{\ell-1}_{n,\cdot j} \colon j \in \mathbb{N}\}$ are exchangeable over $j$.*

*Then as $\min\{n, d^{\ell,H}_n, d^{\ell,G}_n\} \to \infty$:*

**(I)** $G^{\ell}_n = \{G^{\ell h}_n(x) \colon x \in \mathcal{X}, h \in \mathbb{N}\}$ *converges in distribution to $G^{\ell} \sim \mathcal{GP}(0, \kappa^{\ell,G})$ with*

$$\mathbb{E}[G^{\ell h}_{ai}(x) G^{\ell h'}_{bj}(x')] = \delta_{h=h'} \tilde{\kappa}^{\ell}_{ab}(x, x')\, \tilde{\kappa}^{\ell}_{ij}(x, x')\,.$$

**(II)** $f^{\ell}_n$ *converges in distribution to $f^{\ell} \sim \mathcal{GP}(0, \kappa^{\ell})$ with $f^{\ell}_{\cdot k}$ and $f^{\ell}_{\cdot l}$ independent for any $k \neq l$, and*

$$\begin{aligned} \kappa^{\ell}_{ab}(x, x') &= \mathbb{E}[f^{\ell}_{a1}(x) f^{\ell}_{b1}(x')] \\ &= \sum_{i,j=1}^{d^s} \tilde{\kappa}^{\ell}_{ij}(x, x')\, \mathbb{E}[\zeta(G^{\ell 1}(x))_{ai} \zeta(G^{\ell 1}(x'))_{bj}]\,. \end{aligned} \tag{7}$$

*Proof.* Since we have assumed that the input set $\mathcal{X}$ is countable, we can use Lemma 27 to see that all that we need to do to prove Theorem 3 is to show that every finite dimensional marginal of $f_n^\ell$ converges to the corresponding Gaussian limit. Because the finite coordinate projections are continuous by definition of the product topology, the continuous mapping theorem (Dudley, 2002, theorem 9.3.7) tells us it is sufficient to prove convergence of the finite dimensional marginals of

$$\{f_{n,\cdot j}^\ell(x) \colon x \in \mathcal{X}, j \in \mathbb{N}\}\,, \tag{17}$$

as any finite dimensional marginal of $f_n^\ell$ can be obtained by a finite coordinate projection.

Focusing on an arbitrary finite marginal $\mathcal{L} \subset \mathcal{X} \times \mathbb{N}$, we follow Matthews et al. and use the Cramér-Wold device (Billingsley, 1986, p. 383) to reduce the problem to that of establishing convergence of

$$\mathcal{T}_n \coloneqq \sum_{(x,i)\in\mathcal{L}} \langle \alpha^{x,i}, f_{n,\cdot i}^\ell(x) \rangle\,, \tag{18}$$

for any choice of $\{\alpha^{(x,i)} \in \mathbb{R}^{d^s} : (x,i) \in \mathcal{L}\} \subset \mathbb{R}^{d^s \times \mathcal{L}}$. We can rewrite $\mathcal{T}_n$ as

$$\begin{aligned}
\mathcal{T}_n &= \sum_{x,i\in\mathcal{L}} \langle \alpha^{x,i}, f_{n,\cdot i}^\ell(x) \rangle = \sum_{x,i\in\mathcal{L}} \langle \alpha^{x,i}, \big[f_n^{\ell 1}(x), \ldots, f_n^{\ell d_n^{\ell,H}}(x)\big] W_{n,\cdot i}^{\ell,O} \rangle \\
&= \frac{1}{\sqrt{d_n^{\ell,H}}} \sum_{h=1}^{d_n^{\ell,H}} \sum_{(x,i)} \big\langle \alpha^{x,i}, \sqrt{d_n^{\ell,H}} f_n^{\ell h}(x) W_{n,\cdot i}^{\ell h,O} \big\rangle \eqqcolon \frac{1}{\sqrt{d_n^{\ell,H}}} \sum_{h=1}^{d_n^{\ell,H}} \gamma_{n,h}\,,
\end{aligned}$$

where we have defined $W_{n,\cdot i}^{\ell h,O} \coloneqq [W_{n,(hd_n^\ell+1)i}^{\ell,O}, \ldots, W_{n,(hd_n^\ell+d_n^\ell-1)i}^{\ell,O}] \subset \mathbb{R}^{d_n^\ell}$.

We are now prepared to apply lemma 10 from (Matthews et al., 2018) which we restate (with minor modifications) here.

**Lemma 7** (Adaptation of theorem 2 from (Blum et al., 1958))**.** *For each $n \in \mathbb{N}$, let $\{X_{n,i} \colon i = 1, 2, \ldots\}$ be an infinitely exchangeable sequence with $\mathbb{E}\, X_{n,1} = 0$ and $\mathbb{E}\, X_{n,1}^2 = \sigma_n^2$, such that $\lim_{n\to\infty} \sigma_n^2 = \sigma_*^2$ for some $\sigma_*^2 \geq 0$. Let*

$$S_n \coloneqq \frac{1}{\sqrt{d_n}} \sum_{i=1}^{d_n} X_{n,i}\,, \tag{19}$$

*for some sequence $(d_n)_{n\geq 1} \subset \mathbb{N}$ s.t. $\lim_{n\to\infty} d_n = \infty$. Assume:*

*(a)* $\mathbb{E}\, X_{n,1} X_{n,2} = 0$

*(b)* $\lim_{n\to\infty} \mathbb{E}\, X_{n,1}^2 X_{n,2}^2 = \sigma_*^4$

*(c)* $\mathbb{E}\, |X_{n,1}|^3 = \mathrm{o}(\sqrt{d_n})$

*Then $S_n \rightsquigarrow Z$, where $Z = 0$ (a.s.) if $\sigma_*^2 = 0$, and $Z \sim \mathcal{N}(0, \sigma_*^2)$ otherwise.*

Substituting $S_n = \mathcal{T}_n$ and $X_{n,h} = \gamma_{n,h}$, convergence of $\mathcal{T}_n$ follows from Lemma 7:

- Exchangeability requirement is satisfied by Lemma 8.
- Zero mean and covariance follow from Lemma 9.
- Convergence of variance is established in Lemma 10.
- Convergence of $\mathbb{E}[\varphi_{n,1}^2 \varphi_{n,2}^2]$ and $\mathbb{E}\, |\gamma_{n,h}|^3 = \mathrm{o}(\sqrt{d_n^\ell})$ are implied by Lemma 11.

Combining the above with Lemmas 12 and 33 concludes the proof. □

**Lemma 8** (Infinite exchangeability)**.** *$\gamma_{n,h}$ are exchangeable over the index $h$.*

*Proof.* Recall that by the de Finetti's theorem, it is sufficient to exhibit a set of random variables conditioned on which the $\{\gamma_{n,h} \colon h \in \mathbb{N}\}$ are i.i.d. From Section 2, we have

$$f_n^{\ell h}(x) = \zeta\bigg(\frac{1}{\sqrt{d_n^{\ell,G}}} g_n^{\ell-1}(x) W_n^{\ell h,Q} (g_n^{\ell-1}(x) W_n^{\ell h,K})^\top\bigg) g_n^{\ell-1}(x) W_n^{\ell h,V} .$$

Hence if we condition on $\{g_{n,\cdot j}^{\ell-1}(x) \colon j \in [d_n^{\ell-1}], x \in \mathcal{L}_\mathcal{X}\}$, where $\mathcal{L}_\mathcal{X} \coloneqq \{x \in \mathcal{X} \colon \exists i \in \mathbb{N} \text{ s.t. } (x,i) \in \mathcal{L}\}$, it is easy to see that $\{\gamma_{n,h}\}_{h \geq 1}$ are exchangeable. □

**Lemma 9** (Zero mean and covariance)**.** $\mathbb{E}\,\gamma_{n,1} = \mathbb{E}\,\gamma_{n,1}\gamma_{n,2} = 0$.

*Proof.* Using $\mathbb{E}\, W_{n,\cdot i}^{\ell 1,O} = 0$, $\mathbb{E}\,\gamma_{n,1} = 0$ if $|\mathbb{E}\, f_{n,ij}^{\ell 1}(x)| < \infty$ for all $(i,j) \in [d^s] \times \mathbb{N}$. Substituting $\mathbb{E}\, f_{n,ij}^{\ell 1}(x) = \mathbb{E}\,\zeta(G_n^{\ell 1})_{i\cdot} g_n^{\ell-1}(x) W_{n,\cdot j}^{\ell h,V} = 0$ as long as $|\mathbb{E}\,\zeta(G_n^{\ell 1})_{ia} g_{n,bk}^{\ell-1}(x)| < \infty$ for any $a, b, k \in [d^s]$. This can be obtained by combining Hölder's inequality with Lemma 32. An analogous argument applies for $\mathbb{E}\,\gamma_{n,1}\gamma_{n,2} = 0$ since $\mathbb{E}[W_{n,\cdot i}^{\ell 1,O} (W_{n,\cdot i}^{\ell 2,O})^\top] = 0$ by assumption. □

**Lemma 10** (Convergence of variance)**.** $\lim_{n\to\mathbb{N}} \mathbb{E}\,\gamma_{n,1}^2 = \sigma_*^2$.

*Proof.* Observe that $\mathbb{E}\,\gamma_{n,1}^2$ can be written as

$$\frac{\sigma_O^2}{d_n^\ell} \sum_{(x,i),(x',j)} (\alpha^{x,i})^\top\, \mathbb{E}\Big[f_n^{\ell 1}(x) \varepsilon_{\cdot i} (\varepsilon_{\cdot j})^\top f_n^{\ell 1}(x')^\top\Big] \alpha^{x',j} = \frac{\sigma_O^2}{d_n^\ell} \sum_{(x,i)} (\alpha^{x,i})^\top\, \mathbb{E}\Big[f_n^{\ell 1}(x) f_n^{\ell 1}(x)^\top\Big] \alpha^{x,i} ,$$

and thus it will be sufficient to show that $\mathbb{E}[f_n^{\ell 1}(x) f_n^{\ell 1}(x)^\top]/d_n^\ell$ converges to the mean of the weak distributional limit.

$$\begin{aligned}\frac{1}{d_n^\ell}\,\mathbb{E}\Big[f_n^{\ell 1}(x) f_n^{\ell 1}(x)^\top\Big] &= \frac{1}{d_n^\ell}\,\mathbb{E}\Big[\zeta(G_n^{\ell 1}(x)) g_n^{\ell-1}(x) W_n^{\ell 1,V} (W_n^{\ell 1,V})^\top g_n^{\ell-1}(x)^\top \zeta(G_n^{\ell 1}(x))^\top\Big] \\ &= \sigma_V^2\,\mathbb{E}\bigg[\zeta(G_n^{\ell 1}(x)) \frac{g_n^{\ell-1}(x) g_n^{\ell-1}(x)^\top}{d_n^{\ell-1}} \zeta(G_n^{\ell 1}(x))^\top\bigg] ,\end{aligned}$$

suggests the desired result could be obtained by application of Theorem 29 which requires that the integrands converge in distribution to the relevant limit, and that their collection is uniformly integrable. Combination of the continuous mapping theorem and Lemmas 12, 30 and 33 yields convergence in distribution; application of the Hölder's inequality, the polynomial bound on $\zeta$, and Lemma 32 yields uniform integrability by Lemma 28, concluding the proof. □

**Lemma 11.** *For any $h, h' \in \mathbb{N}$, $\mathbb{E}[\gamma_{n,h}^2 \gamma_{n,h'}^2]$ to the mean of the weak limit of $\{\gamma_{n,h}^2 \gamma_{n,h'}^2\}_{n \geq 1}$'s distributions.*

*Proof of Lemma 11.* Defining $\tilde{f}_{n,\cdot i}^h(x) \coloneqq \sqrt{d_n^{\ell,H}} f_n^{\ell h}(x) W_{n,\cdot i}^{\ell h,O}$, we observe $\mathbb{E}[\gamma_{n,h}^2 \gamma_{n,h'}^2]$ equals

$$\sum_{\substack{(x_1,i_1)\\(x_2,i_2)}} \sum_{\substack{(x_1',j_1)\\(x_2',j_2)}} (\alpha^{x_1,i_1})^\top\, \mathbb{E}\Big[\tilde{f}_{n,\cdot i_1}^h(x_1) \tilde{f}_{n,\cdot i_2}^h(x_2) \alpha^{x_2,i_2} (\alpha^{x_1',j_1})^\top \tilde{f}_{n,\cdot j_1}^{h'}(x_1') \tilde{f}_{n,\cdot j_2}^{h'}(x_2')\Big] \alpha^{x_2',j_2} ,$$

which means that the expectation can be evaluated as a weighted sum of terms of the form

$$\mathbb{E}\big[\tilde{f}_{n,ai}^h(s) \tilde{f}_{n,bj}^h(t) \tilde{f}_{n,ck}^{h'}(u) \tilde{f}_{n,dl}^{h'}(v)\big] ,$$

where $a, b, c, d \in [d^s]$, $i, j, k, l \in \mathcal{L}_\mathcal{I}$, and $s, t, u, v \in \mathcal{L}_\mathcal{X}$. We therefore only need to show convergence of these expectations. Substituting:

$$\begin{aligned}&\mathbb{E}\big[\tilde{f}_{n,ai}^h(s) \tilde{f}_{n,bj}^h(t) \tilde{f}_{n,ck}^{h'}(u) \tilde{f}_{n,dl}^{h'}(v)\big] \\ &\qquad = \bigg(\frac{\sigma_O^2}{d_n^\ell}\bigg)^2\, \mathbb{E}\Big[f_{n,a\cdot}^{\ell h}(s) \varepsilon_{\cdot i}^h (\varepsilon_{\cdot j}^h)^\top f_{n,b\cdot}^{\ell h}(t)^\top f_{n,c\cdot}^{\ell h'}(u) \varepsilon_{\cdot k}^{h'} (\varepsilon_{\cdot l}^{h'})^\top f_{n,d\cdot}^{\ell h'}(v)^\top\Big]\end{aligned}$$

$$= \left(\frac{\sigma_O^2}{d_n^\ell}\right)^2 \mathbb{E}\left[f_{n,a\cdot}^{\ell h}(s) f_{n,b\cdot}^{\ell h}(t)^\top f_{n,c\cdot}^{\ell h'}(u) f_{n,d\cdot}^{\ell h'}(v)^\top\right] \delta_{i=j}\delta_{k=l} \,,$$

where $\varepsilon_i^h$ are i.i.d. standard normal random variables, and re-purposing the $i, j$ indices, we have

$$\frac{1}{(d_n^\ell)^2} \mathbb{E}\left[f_{n,a\cdot}^{\ell h}(s) f_{n,b\cdot}^{\ell h}(t)^\top f_{n,c\cdot}^{\ell h'}(u) f_{n,d\cdot}^{\ell h'}(v)^\top\right] = \frac{1}{(d_n^\ell)^2} \sum_{i,j=1}^{d_n^\ell} \mathbb{E}\left[f_{n,ai}^{\ell h}(s) f_{n,bi}^{\ell h}(t) f_{n,cj}^{\ell h'}(u) f_{n,dj}^{\ell h'}(v)\right]$$
$$= \frac{1}{d_n^\ell} \mathbb{E}\left[f_{n,a1}^{\ell h}(s) f_{n,b1}^{\ell h}(t) f_{n,c1}^{\ell h'}(u) f_{n,d1}^{\ell h'}(v)\right] + \frac{d_n^\ell - 1}{d_n^\ell} \mathbb{E}\left[f_{n,a1}^{\ell h}(s) f_{n,b1}^{\ell h}(t) f_{n,c2}^{\ell h'}(u) f_{n,d2}^{\ell h'}(v)\right].$$

Note that we can bound the integrands by a universal constant (Lemma 32), and thus we can focus only on the latter term on the r.h.s. We can thus turn to

$$\mathbb{E}\left[f_{n,a1}^{\ell h}(s) f_{n,b1}^{\ell h}(t) f_{n,c2}^{\ell h'}(u) f_{n,d2}^{\ell h'}(v)\right]$$
$$= \sigma_V^4 \mathbb{E}\left[\zeta(G_n^{\ell h}(s))_{a\cdot} \frac{g_n^{\ell-1}(s) g_n^{\ell-1}(t)^\top}{d_n^{\ell-1}} \zeta(G_n^{\ell h}(s))_{b\cdot}^\top \zeta(G_n^{\ell h'}(u))_{c\cdot} \frac{g_n^{\ell-1}(u) g_n^{\ell-1}(v)^\top}{d_n^{\ell-1}} \zeta(G_n^{\ell h'}(v))_{d\cdot}^\top\right].$$

Observe that by Lemma 33,

$$\left(\frac{g_n^{\ell-1}(s) g_n^{\ell-1}(t)^\top}{d_n^{\ell-1}} \,,\, \frac{g_n^{\ell-1}(u) g_n^{\ell-1}(v)^\top}{d_n^{\ell-1}}\right) \xrightarrow{P} (\tilde{\kappa}^\ell\,(s,t)\,,\, \tilde{\kappa}^\ell\,(u,v))\,,$$

and by Lemma 12 and the continuous mapping theorem

$$\zeta(G_n^{\ell h}(s))_{a\cdot} \zeta(G_n^{\ell h}(s))_{b\cdot} \zeta(G_n^{\ell h'}(u))_{c\cdot} \zeta(G_n^{\ell h'}(v))_{d\cdot} \,,$$

converges in distribution. By Lemma 30, this means that the integrand converges in distribution. Finally, to obtain the convergence of the expectation, we apply Theorem 29 where the required uniform integrability can be obtained by applying Hölder's inequality and Lemma 32. □

### B.1.1. Convergence of $G_n^{\ell h}$

**Lemma 12.** *Let the assumptions of Theorem 3 hold. Then* $G_n^\ell \coloneqq \{G_n^{\ell h}(x) \colon x \in \mathcal{X}, h \in \mathbb{N}\}$ *converges in distribution to a centred GP with covariance as described in Theorem 3.*

*Proof.* Using Lemma 27 and the Cramér Wold device (Billingsley, 1986, p. 383), we can again restrict our attention to one dimensional projections of finite dimensional marginals of $G_n^\ell$

$$\mathcal{T}_n^G \coloneqq \sum_{(x,h)\in\mathcal{L}} \langle \beta^{x,h}, G_n^{\ell h}(x)\rangle_F = \sum_{(x,h)\in\mathcal{L}} \Big\langle \beta^{x,h}, \frac{1}{\sqrt{d_n^\ell}} \sum_{j=1}^{d_n^\ell} Q_{n,\cdot j}^{\ell h}(x) (K_{n,\cdot j}^{\ell h}(x))^\top \Big\rangle_F$$
$$= \frac{1}{\sqrt{d_n^\ell}} \sum_{j=1}^{d_n^\ell} \underbrace{\sum_{(x,h)\in\mathcal{L}} \Big\langle \beta^{x,h}, Q_{n,\cdot j}^{\ell h}(x) (K_{n,\cdot j}^{\ell h}(x))^\top \Big\rangle_F}_{=:\varphi_{n,j}} \,,$$

The above formula suggests the desired result follows from Lemma 7:

- Exchangeability requirement is satisfied by Lemma 13.
- Zero mean and covariance follow from Lemma 14.
- Convergence of variance is established in Corollary 15.

- Convergence of $\mathbb{E}[\varphi_{n,1}^2 \varphi_{n,2}^2]$ is proved in Lemma 16.
- $\mathrm{o}(d_n^\ell)$ growth of the third absolute moments is implied by Lemma 17. □

**Lemma 13.** *Under the assumptions of Theorem 3, $\varphi_{n,j}$ are exchangeable over the $j$ index.*

*Proof.* Observe

$$Q_{n,\cdot j}^{\ell h}(x)(K_{n,\cdot j}^{\ell h}(x))^\top = g_n^{\ell-1}(x) W_{n,\cdot i}^{\ell h,Q} (W_{n,\cdot i}^{\ell h,K})^\top g_n^{\ell-1}(x)^\top ,$$

which means that the individual terms $\varphi_{n,j}$ are i.i.d. if we condition on $\{g_n^{\ell-1}(x) \colon x \in \mathcal{L}_\mathcal{X}\}$. Application of de Finetti's theorem concludes the proof. □

**Lemma 14.** *Under the assumptions of Theorem 3, $\mathbb{E}[\varphi_{n,1}] = \mathbb{E}[\varphi_{n,1}\varphi_{n,2}] = 0$.*

*Proof.* For $\mathbb{E}[\varphi_{n,1}] = 0$, note that for any $h \in \mathbb{N}$, $\mathbb{E}[\varphi_{n,1}]$ can be expressed as a sum over terms

$$\langle \beta^{x,h}, \mathbb{E}[g_n^{\ell-1}(x) W_{n,\cdot j}^{\ell h,Q} (W_{n,\cdot j}^{\ell h,K})^\top g_n^{\ell-1}(x)^\top] \rangle_F = \langle \beta^{x,h}, 0 \rangle_F = 0 ,$$

as long as $\mathbb{E}[g_{n,\cdot 1}^{\ell-1}(x) g_{n,\cdot 1}^{\ell-1}(x)^\top]$ is entry-wise finite for any $(x, n) \in \mathcal{L}_\mathcal{X} \times \mathbb{N}$ which can be obtained by Lemma 32. For $\mathbb{E}[\varphi_{n,1}\varphi_{n,2}] = 0$, we have to evaluate a weighted sum of terms of the form

$$(\beta^{x,h})^\top \, \mathbb{E}\Big[ g_n^{\ell-1}(x) W_{n,\cdot 1}^{\ell h,Q} (W_{n,\cdot 1}^{\ell h,K})^\top g_n^{\ell-1}(x)^\top g_n^{\ell-1}(x') W_{n,\cdot 2}^{\ell h',Q} (W_{n,\cdot 2}^{\ell h',K})^\top g_n^{\ell-1}(x')^\top \Big] \beta^{x',h'} ,$$

which are all equal to zero as long as

$$\mathbb{E}\Bigg[ \frac{g_n^{\ell-1}(x) g_n^{\ell-1}(x)^\top}{d_n^{\ell-1}} \frac{g_n^{\ell-1}(x') g_n^{\ell-1}(x')^\top}{d_n^{\ell-1}} \Bigg] ,$$

is entry-wise finite. Since the integrand converges in probability to $\tilde{\kappa}^\ell\,(x,x)\tilde{\kappa}^\ell\,(x',x')$ by Lemma 33, an argument analogous to the one made above for the $\mathbb{E}[\varphi_{n,1}] = 0$ concludes the proof. □

**Corollary 15.** *Under the assumptions of Theorem 3, $\lim_{n\to\infty} \mathbb{E}[\varphi_{n,1}^2] = \sigma_*^2$.*

*Proof.* The second half the proof of Lemma 14 establishes $\mathbb{E}[\varphi_{n,i}\varphi_{n,j}]$ converges for any $i, j$. □

**Lemma 16.** *Under the assumptions of Theorem 3, $\lim_{n\to\infty} \mathbb{E}[\varphi_{n,1}^2 \varphi_{n,2}^2] = \sigma_*^4$.*

*Proof.* Defining $R_{n,j}^h(x) \coloneqq Q_{n,\cdot j}^{\ell h}(x)(K_{n,\cdot j}^{\ell h}(x))^\top$, we can rewrite $\mathbb{E}\big[\varphi_{n,1}^2 \varphi_{n,2}^2\big]$ as

$$\sum_{(x,h),(x',h')} (\beta^{x,h})^\top \, \mathbb{E}\Big[ R_{n,1}^h(x) R_{n,1}^h(x)^\top \beta^{x,h} (\beta^{x',h'})^\top R_{n,2}^{h'}(x') R_{n,2}^{h'}(x')^\top \Big] \beta^{x',h'} ,$$

where we have w.l.o.g. assumed all matrices have been flattened as $\langle A, B \rangle_F = \operatorname{vect}(A)^\top \operatorname{vect}(B)$. The above could be further rewritten as a weighted sum of terms which take the following form:

$$\begin{aligned} &\mathbb{E}\Big[ Q_{n,a_1 1}^{\ell h}(x) K_{n,b_1 1}^{\ell h}(x) Q_{n,a_2 1}^{\ell h}(x) K_{n,b_2 1}^{\ell h}(x) Q_{n,a_3 2}^{\ell h'}(x') K_{n,b_3 2}^{\ell h'}(x') Q_{n,a_4 2}^{\ell h'}(x') K_{n,b_4 2}^{\ell h'}(x') \Big] \\ &\propto \mathbb{E}\Bigg[ \frac{g_{n,a_1\cdot}^{\ell-1}(x) g_{n,a_2\cdot}^{\ell-1}(x)^\top}{d_n^{\ell-1}} \frac{g_{n,b_1\cdot}^{\ell-1}(x) g_{n,b_2\cdot}^{\ell-1}(x)^\top}{d_n^{\ell-1}} \frac{g_{n,a_3\cdot}^{\ell-1}(x') g_{n,a_4\cdot}^{\ell-1}(x')^\top}{d_n^{\ell-1}} \frac{g_{n,b_3\cdot}^{\ell-1}(x') g_{n,b_4\cdot}^{\ell-1}(x')^\top}{d_n^{\ell-1}} \Bigg] . \end{aligned}$$

Thanks to Lemma 33 and the continuous mapping theorem, we know that the integrand converges in probability to

$$\sigma_Q^4 \sigma_K^4 \tilde{\kappa}_{a_1 a_2}^\ell\,(x,x) \tilde{\kappa}_{b_1 b_2}^\ell\,(x,x) \tilde{\kappa}_{a_3 a_4}^\ell\,(x',x') \tilde{\kappa}_{b_3 b_4}^\ell\,(x',x') ,$$

and thus we can use Theorem 29 to obtain that the above expactation converges as long as the sequence of integrands is uniformly integrable. Noting that we can upper bound by $\max_{c \in [d^s} \max_{z \in \mathcal{L}_\mathcal{X}} \mathbb{E}\, |g_{n,c1}^{\ell-1}(z)|^8$ by Hölder's inequality and exchangeability, uniform integrability can be obtained by Lemma 28. □

**Lemma 17.** *Under the assumptions of Theorem 3,* $\mathbb{E}\,|\varphi_{n,1}|^3 = \mathrm{o}(\sqrt{d_n^\ell})$.

*Proof.* Using Hölder's inequality, it is sufficient to show $\limsup_n \mathbb{E}\,|\varphi_{n,1}|^4 < \infty$. Setting $R_{n,j}^h(x) := Q_{n,\cdot j}^{\ell h}(x)(K_{n,\cdot j}^{\ell h}(x))^\top$

$$\mathbb{E}\,|\varphi_{n,1}|^4 = \sum_{(x,h),(x',h')} (\beta^{x,h})^\top\, \mathbb{E}\Big[R_{n,1}^h(x) R_{n,1}^h(x)^\top \beta^{x,h} (\beta^{x',h'})^\top R_{n,1}^{h'}(x') R_{n,1}^{h'}(x')^\top\Big]\beta^{x',h'}\,,$$

analogously to the proof of Lemma 16. Substituting for the individual terms and using Hölder's inequality, we we can see that each of the terms in the above sum can be itself decomposed into a sum over $(d_n^{\ell-1})^8$ terms that are up to a constant upper bounded by

$$\max_{a\in[d^s]}\max_{z\in\{x,x'\}} \mathbb{E}\,|g_{n,a1}^{\ell-1}(z)|^8\,,$$

which means we can conclude this proof by bounding this quantity by a constant independent of $n$ by Lemma 32. □

### B.2. NTK convergence proof

We need to prove convergence of the attention NTK at initialisation, i.e., for any $a, b \in [d^s]$, $i, j \in \mathbb{N}$, and $x, x' \in \mathcal{X}$

$$\frac{\partial f_{n,ai}^\ell(x)}{\partial \theta_n^{\leq \ell}} \frac{\partial f_{n,bj}^\ell(x')}{\partial \theta_n^{\leq \ell}}^\top \xrightarrow{P} \delta_{i=j}\Theta_{ab}^\ell(x,x')\,, \tag{20}$$

where $\theta_n^{\leq\ell}$ is the collection of trainable parameters in the first $\ell$ layers, as $n\to\infty$. We will further use $\theta_n^\ell$ to refer to the trainable parameters of the $\ell^{\text{th}}$ layer; e.g., for the attention layer $\theta_n^\ell = \{\widetilde{W}_n^\ell\} \cup \bigcup_{h=1}^{d_n^{\ell,H}} \{\widetilde{W}_n^{\ell h,Q}, \widetilde{W}_n^{\ell h,K}, \widetilde{W}_n^{\ell h,V}\}$.

Note that

$$\frac{\partial f_{n,ai}^\ell(x)}{\partial \theta_n^{\leq \ell}} \frac{\partial f_{n,bj}^\ell(x')}{\partial \theta_n^{\leq \ell}}^\top = \underbrace{\frac{\partial f_{n,ai}^\ell(x)}{\partial \theta_n^{\ell}} \frac{\partial f_{n,bj}^\ell(x')}{\partial \theta_n^{\ell}}^\top}_{\text{direct}} + \underbrace{\frac{\partial f_{n,ai}^\ell(x)}{\partial g_n^{\ell-1}(x)} \frac{\partial g_n^{\ell-1}(x)}{\partial \theta_n^{<\ell}} \frac{\partial g_n^{\ell-1}(x')}{\partial \theta_n^{<\ell}}^\top \frac{\partial f_{n,bj}^\ell(x')}{\partial g_n^{\ell-1}(x')}^\top}_{\text{indirect}}\,, \tag{21}$$

where the *direct* part corresponds to the contribution due to gradient w.r.t. the parameters of the $\ell^{\text{th}}$ layer itself, and the *indirect* part is due to effect of the $\ell^{\text{th}}$ layer on the contribution due to the parameters of preceding layers. The next two sections show convergence of each of these terms to a constant in probability, implying the desired result:

**Theorem 18** (NTK convergence)**.** *Under the assumptions of Theorem 3 (including those stated at the beginning of Appendix B), for any* $a, b \in [d^s]$, *and* $x, x' \in \mathcal{X}$

$$\frac{\partial f_{n,ai}^\ell(x)}{\partial \theta_n^{\leq \ell}} \frac{\partial f_{n,bj}^\ell(x')}{\partial \theta_n^{\leq \ell}}^\top \xrightarrow{P} \delta_{i=j}\Theta_{ab}^\ell(x,x')\,,$$

*where*

$$\begin{aligned}\Theta_{ab}^\ell(x,x') =& 2\kappa_{ab}^\ell(x,x') + \sigma_{OV}^2 \sum_{a',b'}^{d^s} \widetilde{\Theta}_{a'b'}^\ell(x,x')\,\mathbb{E}[\widetilde{G}_{aa'}^{\ell 1}(x)\widetilde{G}_{bb'}^{\ell 1}(x')] + \\ & \delta_{\tau=\frac{1}{2}}\sigma_{OV}^2\sigma_{QK}^2 (2\tilde{\kappa}_{ab}^\ell(x,x') + \widetilde{\Theta}_{ab}^\ell(x,x')) \sum_{\substack{c_1,c_2\\ d_1,d_2}}^{d^s} \tilde{\kappa}_{c_1c_2}^\ell(x,x')\tilde{\kappa}_{d_1d_2}^\ell(x,x')\,\mathbb{E}\left[\frac{\partial \widetilde{G}_{ac_1}^{\ell 1}(x)}{\partial G_{ad_1}^{\ell 1}(x)} \frac{\partial \widetilde{G}_{bc_2}^{\ell 1}(x')}{\partial G_{bd_2}^{\ell 1}(x')}\right] + \\ & \delta_{\tau=\frac{1}{2}}\sigma_{OV}^2\sigma_{QK}^2 \tilde{\kappa}_{ab}^\ell(x,x') \sum_{\substack{c_1,c_2\\ d_1,d_2}}^{\ell} \tilde{\kappa}_{c_1c_2}^\ell(x,x')\widetilde{\Theta}_{d_1d_2}^\ell(x,x')\,\mathbb{E}\left[\frac{\partial \widetilde{G}_{ac_1}^{\ell 1}(x)}{\partial G_{ad_1}^{\ell 1}(x)} \frac{\partial \widetilde{G}_{bc_2}^{\ell 1}(x')}{\partial G_{bd_2}^{\ell 1}(x')}\right]\,.\end{aligned} \tag{22}$$

Theorem 18 will be proven in the following two subsections.

### B.2.1. Direct contribution

The direct contribution of an attention layer can be expanded as

$$\frac{\partial f^{\ell}_{n,ai}(x)}{\partial \theta^{\ell}_n}\frac{\partial f^{\ell}_{n,bj}(x')}{\partial \theta^{\ell}_n}^{\top} = \frac{\partial f^{\ell}_{n,ai}(x)}{\partial \widetilde{W}^{\ell,O}_n}\frac{\partial f^{\ell}_{n,bj}(x')}{\partial \widetilde{W}^{\ell,O}_n}^{\top} + \sum_{h=1}^{d^{\ell,H}_n} \frac{\partial f^{\ell}_{n,ai}(x)}{\partial \widetilde{W}^{\ell h,V}_n}\frac{\partial f^{\ell}_{n,bj}(x')}{\partial \widetilde{W}^{\ell h,V}_n}^{\top} + \frac{\partial f^{\ell}_{n,ai}(x)}{\partial \widetilde{W}^{\ell h,Q}_n}\frac{\partial f^{\ell}_{n,bj}(x')}{\partial \widetilde{W}^{\ell h,Q}_n}^{\top} + \frac{\partial f^{\ell}_{n,ai}(x)}{\partial \widetilde{W}^{\ell h,K}_n}\frac{\partial f^{\ell}_{n,bj}(x')}{\partial \widetilde{W}^{\ell h,K}_n}^{\top} .$$

We prove convergence of each of these terms next.

**Lemma 19.** $\frac{\partial f^{\ell}_{n,ai}(x)}{\partial \widetilde{W}^{\ell,O}_n}\frac{\partial f^{\ell}_{n,bj}(x')}{\partial \widetilde{W}^{\ell,O}_n}^{\top} \xrightarrow{P} \delta_{i=j}\kappa^{\ell}_{ab}(x,x')$.

*Proof of Lemma 19.* Observe

$$\begin{aligned}\frac{\partial f^{\ell}_{n,ai}(x)}{\partial \widetilde{W}^{\ell,O}_n}\frac{\partial f^{\ell}_{n,bj}(x')}{\partial \widetilde{W}^{\ell,O}_n}^{\top} &= \delta_{i=j}\sum_{h=1}^{d^{\ell,H}_n}\sum_{k=1}^{d^{\ell,V}_n}\frac{\sigma_O^2}{d^{\ell,H}_n d^{\ell,V}_n} f^{\ell h}_{n,ak}(x) f^{\ell h}_{n,bk}(x') \\ &= \delta_{i=j}\frac{\sigma_O^2}{d^{\ell,H}_n}\sum_{h=1}^{d^{\ell,H}_n}\sum_{c_1,c_2=1}^{d^s}\widetilde{G}^{\ell h}_{n,ac_1}(x)\widetilde{G}^{\ell h}_{n,bc_2}(x')\frac{\langle V^{\ell h}_{n,c_1\cdot}(x), V^{\ell h}_{n,c_2\cdot}(x')\rangle}{d^{\ell,V}_n} .\end{aligned}$$

Since $d^s$ is fixed, we can focus on an arbitrary pair $c_1, c_2 \in [d^s]$. Notice that by the continuous mapping theorem and Lemmas 30 and 33, the individual summands converge in distribution

$$\widetilde{G}^{\ell h}_{n,ac_1}(x)\widetilde{G}^{\ell h}_{n,bc_2}(x')\frac{\langle V^{\ell h}_{n,c_1\cdot}(x), V^{\ell h}_{n,c_2\cdot}(x')\rangle}{d^{\ell,V}_n} \rightsquigarrow \sigma_V^2\widetilde{G}^{\ell h}_{ac_1}(x)\widetilde{G}^{\ell h}_{bc_2}(x')\tilde{\kappa}^{\ell}_{c_1c_2}(x,x') ,$$

where $\widetilde{G}^{\ell h}$ follows the $\zeta_{\#}$ pushforward of the GP distribution of $G^{\ell}$ described in Theorem 3 if $\tau = \frac{1}{2}$, or $\widetilde{G}^{\ell}(x) = \zeta(\sigma_Q\sigma_K\tilde{\kappa}^{\ell}(x,x))$ a.s. if $\tau = 1$ (Yang, 2019b, appendix A). The desired result could thus be established by application of Lemma 31, averaging over the $h$ index, if its assumptions hold.

Starting with the exchangeability assumption, note that if we condition on $g^{\ell-1}_n(x), g^{\ell-1}_n(x')$, the individual terms are i.i.d. because the parameters of individual heads are i.i.d. Since $\{\widetilde{G}^{\ell h}_{ac_1}(x)\widetilde{G}^{\ell h}_{bc_2}(x')\}_{h\geq 1}$ are also i.i.d. (see Theorem 3 for $\tau = \frac{1}{2}$, and constancy under $\tau = 1$), it is also clear that the $\mathbb{E}[X_{*,1}X_{*,2}] = (\mathbb{E}[X_{*,1}])^2$ is satisfied. All that remains is to show $\limsup_{n\to\infty}\mathbb{E}|X_{n,1}|^{2+\varepsilon} < \infty$, and where we will use $\varepsilon = 2$ for convenience. By Hölder's inequality

$$\begin{aligned}&\mathbb{E}\left\{\left[\widetilde{G}^{\ell 1}_{n,ac_1}(x)\widetilde{G}^{\ell 1}_{n,bc_2}(x') \quad \frac{\langle V^{\ell 1}_{n,c_1\cdot}(x), V^{\ell 1}_{n,c_2\cdot}(x')\rangle}{d^{\ell,V}_n}\right]^4\right\} \\ &\qquad\lesssim \operatorname{poly}\left(\max_{c,c'\in[d^s], z\in\{x,x'\}}\mathbb{E}|\widetilde{G}^{\ell 1}_{n,cc'}(z)|^{16}, \max_{c\in[d^s], z\in\{x,x'\}}\mathbb{E}|g^{\ell-1}_{n,c1}(z)|^{16}\right) ,\end{aligned}$$

where we used the assumed exchangeability of $g^{\ell-1}_n(z)$ over its columns. Application of Lemma 32 implies that the above can be bounded by a constant independent of $n$, implying all assumptions of Lemma 31 are satisfied. □

**Lemma 20.** $\sum_{h=1}^{d^{\ell,H}_n}\frac{\partial f^{\ell}_{n,ai}(x)}{\partial \widetilde{W}^{\ell h,V}_n}\frac{\partial f^{\ell}_{n,bj}(x')}{\partial \widetilde{W}^{\ell h,V}_n}^{\top} \xrightarrow{P} \delta_{i=j}\kappa^{\ell}_{ab}(x,x')$.

*Proof of Lemma 20.* Note that

$$\sum_{h=1}^{d^{\ell,H}_n}\frac{\partial f^{\ell}_{n,ai}(x)}{\partial \widetilde{W}^{\ell h,V}_n}\frac{\partial f^{\ell}_{n,bj}(x')}{\partial \widetilde{W}^{\ell h,V}_n}^{\top} = \frac{\sigma_O^2}{d^{\ell,H}_n d^{\ell,V}_n}\sum_{h=1}^{d^{\ell,H}_n}\sum_{k=1}^{d^{\ell,V}_n}\widetilde{W}^{\ell h,O}_{n,ki}\widetilde{W}^{\ell h,O}_{n,kj}\frac{\sigma_V^2}{d^{\ell-1}_n}\left\langle \widetilde{G}^{\ell h}_{n,a\cdot}(x)g^{\ell-1}(x), \widetilde{G}^{\ell h}_{n,b\cdot}(x')g^{\ell-1}(x')\right\rangle$$

$$= \frac{\sigma_O^2 \sigma_V^2}{d_n^{\ell,H} d_n^{\ell,V}} \sum_{h,k} \sum_{c_1,c_2=1}^{d^s} \widetilde{W}_{n,ki}^{\ell h,O} \widetilde{W}_{n,kj}^{\ell h,O} \widetilde{G}_{n,ac_1}^{\ell h}(x) \widetilde{G}_{n,bc_2}^{\ell h}(x') \frac{\langle g_{c_1 \cdot}^{\ell-1}(x), g_{c_2 \cdot}^{\ell-1}(x') \rangle}{d_n^{\ell-1}} .$$

Since $d^s$ is fixed, we can focus on an arbitrary $c_1, c_2 \in [d^s]$. Notice that by the assumed independence of the entries of $\widetilde{W}_n^{\ell h,O}$, the continuous mapping theorem and Lemmas 30 and 33, the individual summands converge in distribution

$$\widetilde{W}_{n,ki}^{\ell h,O} \widetilde{W}_{n,kj}^{\ell h,O} \widetilde{G}_{n,ac_1}^{\ell h}(x) \widetilde{G}_{n,bc_2}^{\ell h}(x') \frac{\langle g_{c_1 \cdot}^{\ell-1}(x), g_{c_2 \cdot}^{\ell-1}(x') \rangle}{d_n^{\ell-1}} \rightsquigarrow \widetilde{W}_{n,ki}^{\ell h,O} \widetilde{W}_{n,kj}^{\ell h,O} \widetilde{G}_{ac_1}^{\ell h}(x) \widetilde{G}_{bc_2}^{\ell h}(x') \tilde{\kappa}_{c_1 c_2}^{\ell}(x, x') ,$$

with the distribution of $\widetilde{G}^{\ell h}(x)$ as in the proof of Lemma 19. The desired result can thus again be obtained by applying Lemma 31, averaging over $h$ and $k$, if its assumptions hold. As $\mathbb{E}[\widetilde{W}_{n,ki}^{\ell h,O} \widetilde{W}_{n,kj}^{\ell h,O}] = \delta_{i=j}$ and $\mathbb{E}\,|\widetilde{W}_{n,ki}^{\ell h,O}|^t < \infty$ for any $t \geq 1$, the same argument as in Lemma 19 applies. □

**Lemma 21.** $\sum_{h=1}^{d_n^{\ell,H}} \frac{\partial f_{n,ai}^{\ell}(x)}{\partial \widetilde{W}_n^{\ell h,Q}} \frac{\partial f_{n,bj}^{\ell}(x')}{\partial \widetilde{W}_n^{\ell h,Q}}^{\top} + \frac{\partial f_{n,ai}^{\ell}(x)}{\partial \widetilde{W}_n^{\ell h,K}} \frac{\partial f_{n,bj}^{\ell}(x')}{\partial \widetilde{W}_n^{\ell h,K}}^{\top}$ *converges in probability to*

$$\delta_{i=j} \delta_{\tau=\frac{1}{2}} 2\sigma_{OV}^2 \sigma_{QK}^2 \tilde{\kappa}_{ab}^{\ell}(x, x') \sum_{\substack{c_1,c_2\\ d_1,d_2}}^{d^s} \tilde{\kappa}_{c_1 c_2}^{\ell}(x, x') \tilde{\kappa}_{d_1 d_2}^{\ell}(x, x') \, \mathbb{E}\left[ \frac{\partial \widetilde{G}_{ac_1}^{\ell 1}(x)}{\partial G_{ad_1}^{\ell 1}(x)} \frac{\partial \widetilde{G}_{bc_2}^{\ell 1}(x')}{\partial G_{bd_2}^{\ell 1}(x')} \right] .$$

*Proof of Lemma 21.* By symmetry, it is sufficient to prove convergence for the gradients w.r.t. $\widetilde{W}_n^{\ell h,K}$. Observe

$$\sum_{h=1}^{d_n^{\ell,H}} \frac{\partial f_{n,ai}^{\ell}(x)}{\partial \widetilde{W}_n^{\ell h,K}} \frac{\partial f_{n,bj}^{\ell}(x')}{\partial \widetilde{W}_n^{\ell h,K}}^{\top}$$

$$= \frac{\sigma_O^2}{d_n^{\ell,H} d_n^{\ell,V}} \sum_{h=1}^{d_n^{\ell,H}} \sum_{k_1,k_2=1}^{d_n^{\ell,V}} \sum_{\substack{c_1,c_2\\ d_1,d_2}}^{d^s} \widetilde{W}_{n,k_1 i}^{\ell h,O} \widetilde{W}_{n,k_2 j}^{\ell h,O} V_{n,c_1 k_1}^{\ell h}(x) V_{n,c_2 k_2}^{\ell h}(x') \frac{\partial \widetilde{G}_{n,ac_1}^{\ell h}(x)}{\partial G_{n,ad_1}^{\ell h}(x)} \frac{\partial \widetilde{G}_{n,bc_2}^{\ell h}(x')}{\partial G_{n,bd_2}^{\ell h}(x')} \frac{\partial G_{n,ad_1}^{\ell h}(x)}{\partial \widetilde{W}_n^{\ell h,K}} \frac{\partial G_{n,bd_2}^{\ell h}(x')}{\partial \widetilde{W}_n^{\ell h,K}}^{\top} .$$

Since $d^s$ is fixed, we can focus on arbitrary $c_1, c_2, d_1, d_2 \in [d^s]$. Rewriting the r.h.s. above for one such choice, we obtain

$$\frac{\sigma_O^2 \sigma_K^2}{d_n^{\ell,H} d_n^{\ell,V}} \sum_{h,k_1,k_2} \widetilde{W}_{n,k_1 i}^{\ell h,O} \widetilde{W}_{n,k_2 j}^{\ell h,O} V_{n,c_1 k_1}^{\ell h}(x) V_{n,c_2 k_2}^{\ell h}(x') \frac{\partial \widetilde{G}_{n,ac_1}^{\ell h}(x)}{\partial G_{n,ad_1}^{\ell h}(x)} \frac{\partial \widetilde{G}_{n,bc_2}^{\ell h}(x')}{\partial G_{n,bd_2}^{\ell h}(x')} \frac{\langle Q_{n,a\cdot}^{\ell h}(x), Q_{n,b\cdot}^{\ell h}(x') \rangle}{(d_n^{\ell,G})^{2\tau}} \frac{\langle g_{d_1 \cdot}^{\ell-1}(x), g_{d_2 \cdot}^{\ell-1}(x') \rangle}{d_n^{\ell-1}} .$$

Noting that $\langle g_{d_1 \cdot}^{\ell-1}(x), g_{d_2 \cdot}^{\ell-1}(x') \rangle / d_n^{\ell-1}$ only depends on the spatial dimension indices $d_1$ and $d_2$, we can use Lemma 33 to establish it converges in probability to $\tilde{\kappa}_{d_1 d_2}^{\ell}(x, x')$, implying that we only need to prove that the rest of the terms in the above sum also converges in probability. Let

$$\bar{S}_n = \frac{1}{d_n^{\ell,H} d_n^{\ell,V}} \sum_{h,k_1,k_2} \widetilde{W}_{n,k_1 i}^{\ell h,O} \widetilde{W}_{n,k_2 j}^{\ell h,O} V_{n,c_1 k_1}^{\ell h}(x) V_{n,c_2 k_2}^{\ell h}(x') \frac{\partial \widetilde{G}_{n,ac_1}^{\ell h}(x)}{\partial G_{n,ad_1}^{\ell h}(x)} \frac{\partial \widetilde{G}_{n,bc_2}^{\ell h}(x')}{\partial G_{n,bd_2}^{\ell h}(x')} \frac{\langle Q_{n,a\cdot}^{\ell h}(x), Q_{n,b\cdot}^{\ell h}(x') \rangle}{(d_n^{\ell,G})^{2\tau}} ,$$

and note that $\mathbb{E}[\bar{S}_n] = \delta_{i=j} \, \mathbb{E}\left[ \frac{\langle Q_{n,a\cdot}^{\ell 1}(x), Q_{n,b\cdot}^{\ell 1}(x') \rangle}{(d_n^{\ell,G})^{2\tau}} \frac{\langle g_{n,c_1\cdot}^{\ell-1}(x), g_{n,c_2\cdot}^{\ell-1}(x') \rangle}{d_n^{\ell-1}} \frac{\partial \widetilde{G}_{n,ac_1}^{\ell 1}(x)}{\partial G_{n,ad_1}^{\ell 1}(x)} \frac{\partial \widetilde{G}_{n,bc_2}^{\ell 1}(x')}{\partial G_{n,bd_2}^{\ell 1}(x')} \right]$ by exchangeability. This suggests that the required result could be obtained using the Chebyshev's inequality

$$\mathbb{P}(|\bar{S}_n - \mathbb{E}\,\bar{S}_n| \geq \delta) \leq \frac{\mathbb{E}[\bar{S}_n^2] - \{\mathbb{E}[\bar{S}_n]\}^2}{\delta^2} ,$$

if $\mathbb{E}[\bar{S}_n]$ converges to the desired limit. To establish this convergence, observe

$$\delta_{i=j} \frac{\langle Q_{n,a\cdot}^{\ell 1}(x), Q_{n,b\cdot}^{\ell 1}(x') \rangle}{(d_n^{\ell,G})^{2\tau}} \frac{\langle g_{n,c_1\cdot}^{\ell-1}(x), g_{n,c_2\cdot}^{\ell-1}(x') \rangle}{d_n^{\ell-1}} \frac{\partial \widetilde{G}_{n,ac_1}^{\ell 1}(x)}{\partial G_{n,ad_1}^{\ell 1}(x)} \frac{\partial \widetilde{G}_{n,bc_2}^{\ell 1}(x')}{\partial G_{n,bd_2}^{\ell 1}(x')}$$

$$\rightsquigarrow \delta_{i=j} \delta_{\tau=\frac{1}{2}} \sigma_Q^2 \tilde{\kappa}_{ab}^{\ell}(x, x') \tilde{\kappa}_{c_1 c_2}^{\ell}(x, x') \frac{\partial \widetilde{G}_{ac_1}^{\ell 1}(x)}{\partial G_{ad_1}^{\ell 1}(x)} \frac{\partial \widetilde{G}_{bc_2}^{\ell 1}(x')}{\partial G_{bd_2}^{\ell 1}(x')} , \tag{23}$$

since the first two terms converge in probability (Lemma 33), and the last converges in distribution by Theorem 3 and the continuous mapping theorem, implying that the product of all three thus converges in distribution by Lemma 30. Convergence of $\mathbb{E}[\bar{S}_n]$ could thus be obtained by establishing uniform integrability of the $(\bar{S}_n)_{n\geq 1}$ sequence (Theorem 29).

By Lemma 28, uniform integrability of $\bar{S}_n$ can be established by showing $\mathbb{E}[\bar{S}_n^2] \to \{\mathbb{E}[\bar{S}_*]\}^2$ which would also imply $\bar{S}_n \xrightarrow{P} \mathbb{E}[\bar{S}_*]$ by the above Chebyshev's inequality. For the rest of this proof, we drop the $x, x'$ from our equations for brevity; this allows us to write

$$\mathbb{E}[\bar{S}_n^2] = \frac{1}{(d_n^{\ell,H} d_n^{\ell,V})^2} \sum_{\substack{h_1,h_2\\ k_1,k_2,k_3,k_4}} \mathbb{E}\left[\prod_{t=0}^{1} \widetilde{W}^{\ell h_{t+1},O}_{n,k_{2t+1}i} \widetilde{W}^{\ell h_{t+1},O}_{n,k_{2t+2}j} V^{\ell h_{t+1}}_{n,c_1 k_{2t+1}} V^{\ell h_{t+1}}_{n,c_2 k_{2t+2}} \frac{\langle Q^{\ell h_{t+1}}_{n,a\cdot}, Q^{\ell h_{t+1}}_{n,b\cdot}\rangle}{(d_n^{\ell,G})^{2\tau}} \frac{\partial \widetilde{G}^{\ell h_{t+1}}_{n,ac_1}}{\partial G^{\ell h_{t+1}}_{n,ad_1}} \frac{\partial \widetilde{G}^{\ell h_{t+1}}_{n,bc_2}}{\partial G^{\ell h_{t+1}}_{n,bd_2}}\right].$$

From above, we can restrict our attention to groups of terms that include at least $\mathcal{O}((d_n^{\ell,H} d_n^{\ell,V})^2)$ of the summands as long as the expectation of the square of each term can be bounded by a constant independent of the $h, k$ and $n$ indices.

$$\begin{aligned}\mathbb{E}&\left\{\left[\prod_{t=0}^{1} \widetilde{W}^{\ell h_{t+1},O}_{n,k_{2t+1}i} \widetilde{W}^{\ell h_{t+1},O}_{n,k_{2t+2}j} V^{\ell h_{t+1}}_{n,c_1 k_{2t+1}} V^{\ell h_{t+1}}_{n,c_2 k_{2t+2}} \frac{\langle Q^{\ell h_{t+1}}_{n,a\cdot}, Q^{\ell h_{t+1}}_{n,b\cdot}\rangle}{(d_n^{\ell,G})^{2\tau}} \frac{\partial \widetilde{G}^{\ell h_{t+1}}_{n,ac_1}}{\partial G^{\ell h_{t+1}}_{n,ad_1}} \frac{\partial \widetilde{G}^{\ell h_{t+1}}_{n,bc_2}}{\partial G^{\ell h_{t+1}}_{n,bd_2}}\right]^2\right\}\\
&\lesssim \mathbb{E}\left\{\left[\prod_{t=0}^{1} V^{\ell h_{t+1}}_{n,c_1 k_{2t+1}} V^{\ell h_{t+1}}_{n,c_2 k_{2t+2}} \frac{\langle Q^{\ell h_{t+1}}_{n,a\cdot}, Q^{\ell h_{t+1}}_{n,b\cdot}\rangle}{(d_n^{\ell,G})^{2\tau}}\right]^4\right\} \lesssim \operatorname{poly}\left(\max_{c\in[d^s], z\in\{x,x'\}} \mathbb{E}\,|g^{\ell-1}_{n,c1}(z)|^{16}\right), \qquad (24)\end{aligned}$$

by Hölder's inequality and exchangeability. Application of Lemma 32 allows us to bound the above r.h.s. by a constant independent of $h, k$ and $n$ as desired.

We can thus only focus on the terms for which $h_1 \neq h_2$. Among these, the only ones with non-zero expectation are those where $i = j$, $k_1 = k_2$, and $k_3 = k_4$, contributing to $\mathbb{E}[\bar{S}_n^2]$ by

$$\delta_{i=j} \frac{\sigma_O^2 \sigma_V^2}{(d_n^{\ell,H} d_n^{\ell,V})^2} \sum_{\substack{h_1 h_2\\ k_1,k_2}} \mathbb{E}\left[\left(\frac{\langle g^{\ell-1}_{n,c_1\cdot}, g^{\ell-1}_{n,c_2\cdot}\rangle}{d_n^{\ell-1}}\right)^2 \frac{\langle Q^{\ell 1}_{n,a\cdot}, Q^{\ell 1}_{n,b\cdot}\rangle}{(d_n^{\ell,G})^{2\tau}} \frac{\langle Q^{\ell 2}_{n,a\cdot}, Q^{\ell 2}_{n,b\cdot}\rangle}{(d_n^{\ell,G})^{2\tau}} \frac{\partial \widetilde{G}^{\ell 1}_{n,ac_1}}{\partial G^{\ell 1}_{n,ad_1}} \frac{\partial \widetilde{G}^{\ell 1}_{n,bc_2}}{\partial G^{\ell 1}_{n,bd_2}} \frac{\partial \widetilde{G}^{\ell 2}_{n,ac_1}}{\partial G^{\ell 2}_{n,ad_1}} \frac{\partial \widetilde{G}^{\ell 2}_{n,bc_2}}{\partial G^{\ell 2}_{n,bd_2}}\right] \qquad (25)$$

by exchangeability. Noting that the sum cancels out with the $d_n^{\ell,H} d_n^{\ell,V}$ terms, we see that the limit of $\mathbb{E}[\bar{S}_n^2]$ will be identical to that of Equation (25). Applying Lemmas 30 and 33, Theorem 3 (resp. the result by Yang (2019b, appendix A) if $\tau = 1$), and the continuous mapping theorem

$$\begin{aligned}\delta_{i=j} &\left(\frac{\langle g^{\ell-1}_{n,c_1\cdot}, g^{\ell-1}_{n,c_2\cdot}\rangle}{d_n^{\ell-1}}\right)^2 \frac{\langle Q^{\ell 1}_{n,a\cdot}, Q^{\ell 1}_{n,b\cdot}\rangle}{(d_n^{\ell,G})^{2\tau}} \frac{\langle Q^{\ell 2}_{n,a\cdot}, Q^{\ell 2}_{n,b\cdot}\rangle}{(d_n^{\ell,G})^{2\tau}} \frac{\partial \widetilde{G}^{\ell 1}_{n,ac_1}}{\partial G^{\ell 1}_{n,ad_1}} \frac{\partial \widetilde{G}^{\ell 1}_{n,bc_2}}{\partial G^{\ell 1}_{n,bd_2}} \frac{\partial \widetilde{G}^{\ell 2}_{n,ac_1}}{\partial G^{\ell 2}_{n,ad_1}} \frac{\partial \widetilde{G}^{\ell 2}_{n,bc_2}}{\partial G^{\ell 2}_{n,bd_2}}\\
&\rightsquigarrow \delta_{i=j} \delta_{\tau=\frac{1}{2}} \sigma_Q^4 [\tilde{\kappa}^{\ell}_{ab}(x,x')]^2 [\tilde{\kappa}^{\ell}_{c_1 c_2}(x,x')]^2 \frac{\partial \widetilde{G}^{\ell 1}_{ac_1}}{\partial G^{\ell 1}_{ad_1}} \frac{\partial \widetilde{G}^{\ell 1}_{bc_2}}{\partial G^{\ell 1}_{bd_2}} \frac{\partial \widetilde{G}^{\ell 2}_{ac_1}}{\partial G^{\ell 2}_{ad_1}} \frac{\partial \widetilde{G}^{\ell 2}_{bc_2}}{\partial G^{\ell 2}_{bd_2}}, \qquad (26)\end{aligned}$$

where $\frac{\partial \widetilde{G}^{\ell h}}{\partial G^{\ell h}}$ follows the $(\nabla\zeta)_{\#}$ pushforward of the GP distribution of $G^{\ell}$ described in Theorem 3 if $\tau = \frac{1}{2}$, and is a.s. constant if $\tau = 1$ as the limit $\widetilde{G}^{\ell h}$ is a.s. constant (Yang, 2019b, appendix A), both by the assumed continuity of $\nabla\zeta$.

Finally, because Equation (24) establishes uniform integrability, and $\frac{\partial \widetilde{G}^{\ell 1}}{\partial G^{\ell 1}}$ is independent of $\frac{\partial \widetilde{G}^{\ell 2}}{\partial G^{\ell 2}}$ by Theorem 3, we can combine Equations (23) and (26) with Theorem 29 to conclude that both $\mathbb{E}[\bar{S}_n^2]$ and $\{\mathbb{E}[\bar{S}_n]\}^2$ converge to the same limit. □

### B.2.2. Indirect contribution

The indirect contribution of an attention layer can be expanded as

$$\frac{\partial f^{\ell}_{n,ai}(x)}{\partial g^{\ell-1}_n(x)} \frac{\partial g^{\ell-1}_n(x)}{\partial \theta^{<\ell}_n} \frac{\partial g^{\ell-1}_n(x')}{\partial \theta^{<\ell}_n}^{\top} \frac{\partial f^{\ell}_{n,bj}(x')}{\partial g^{\ell-1}_n(x')}^{\top} = \sum_{a',b'=1}^{d^s} \sum_{i',j'=1}^{d_n^{\ell-1}} \widehat{\Theta}^{\ell}_{a'i',b'j'}(x,x') \frac{\partial f^{\ell}_{n,ai}(x)}{\partial g^{\ell-1}_{n,a'i'}(x)} \frac{\partial f^{\ell}_{n,bj}(x')}{\partial g^{\ell-1}_{n,b'j'}(x')},$$

where[11]

$$\widehat{\Theta}^{\ell}_{a'i',b'j'}(x, x') := \left\langle \frac{\partial g^{\ell-1}_{n,a'i'}(x)}{\partial \theta^{\leq \ell}_{n}}, \frac{\partial g^{\ell-1}_{n,b'j'}(x')}{\partial \theta^{\leq \ell}_{n}} \right\rangle , \tag{27}$$

which we know converges a.s., and thus also in probability, to $\delta_{i'=j'}\widetilde{\Theta}^{\ell}_{a'b'}(x, x')$ for architectures without attention layers (Yang, 2019b). Expanding the indirect contribution further

$$\begin{aligned} &\sum_{a',b'}\sum_{i',j'} \widehat{\Theta}^{\ell}_{a'i',b'j'}(x, x') \frac{\partial f^{\ell}_{n,ai}(x)}{\partial g^{\ell-1}_{n,a'i'}(x)} \frac{\partial f^{\ell}_{n,bj}(x')}{\partial g^{\ell-1}_{n,b'j'}(x')} \\ &= \sum_{a',b'}\sum_{i',j'} \widehat{\Theta}^{\ell}_{a'i',b'j'}(x, x') \frac{\sigma^2_O}{d^{\ell,H}_n d^{\ell,V}_n} \sum_{h_1,h_2=1}^{d^{\ell,H}_n} \sum_{k_1,k_2=1}^{d^{\ell,V}_n} \widetilde{W}^{\ell h_1,O}_{n,k_1 i} \widetilde{W}^{\ell h_2,O}_{n,k_2 j} \frac{\partial \widetilde{G}^{\ell h_1}_{n,a\cdot}(x) V^{\ell h_1}_{n,\cdot k_1}(x)}{\partial g^{\ell-1}_{n,a'i'}(x)} \frac{\partial \widetilde{G}^{\ell h_1}_{n,b\cdot}(x) V^{\ell h_2}_{n,\cdot k_2}(x')}{\partial g^{\ell-1}_{n,b'j'}(x')} . \end{aligned}$$

In the rest of this section, we drop the $x, x'$ from most of our equations so as to reduce the number of multi-line expressions. Continuing with the inner sum from above we obtain

$$\begin{aligned} &\frac{\sigma^2_O}{d^{\ell,H}_n d^{\ell,V}_n} \sum_{\substack{h_1,h_2\\k_1,k_2}} \widetilde{W}^{\ell h_1,O}_{n,k_1 i} \widetilde{W}^{\ell h_2,O}_{n,k_2 j} \frac{\partial \widetilde{G}^{\ell h_1}_{n,a\cdot} V^{\ell h_1}_{n,\cdot k_1}}{\partial g^{\ell-1}_{n,a'i'}} \frac{\partial \widetilde{G}^{\ell h_1}_{n,b\cdot} V^{\ell h_2}_{n,\cdot k_2}}{\partial g^{\ell-1}_{n,b'j'}} \\ &= \frac{\sigma^2_O}{d^{\ell,H}_n d^{\ell,V}_n} \sum_{\substack{h_1,h_2\\k_1,k_2}} \widetilde{W}^{\ell h_1,O}_{n,k_1 i} \widetilde{W}^{\ell h_2,O}_{n,k_2 j} \left( \sum_{c_1=1}^{d^s} \widetilde{G}^{\ell h_1}_{n,ac_1} \frac{\partial V^{\ell h_1}_{n,c_1 k_1}}{\partial g^{\ell-1}_{n,a'i'}} + V^{\ell h_1}_{n,c_1 k_1} \sum_{d_1=1}^{d^s} \frac{\partial \widetilde{G}^{\ell h_1}_{n,ac_1}}{\partial G^{\ell h_1}_{n,ad_1}} \frac{\partial G^{\ell h_1}_{n,ad_1}}{\partial g^{\ell-1}_{n,a'i'}} \right) \cdot \\ &\qquad \left( \sum_{c_2=1}^{d^s} \widetilde{G}^{\ell h_2}_{n,bc_2} \frac{\partial V^{\ell h_2}_{n,c_2 k_2}}{\partial g^{\ell-1}_{n,b'j'}} + V^{\ell h_2}_{n,c_2 k_2} \sum_{d_2=1}^{d^s} \frac{\partial \widetilde{G}^{\ell h_2}_{n,bc_2}}{\partial G^{\ell h_2}_{n,bd_2}} \frac{\partial G^{\ell h_2}_{n,bd_2}}{\partial g^{\ell-1}_{n,b'j'}} \right) , \end{aligned}$$

which gives us four sums after multiplying out the terms inside the parenthesis, for each of which we prove convergence separately. Since the spatial dimension $d^s$ does not change with $n$, we will restrict our attention to an arbitrary fixed choice of $a', b', c_1, c_2, d_1, d_2 \in [d^s]$ throughout.

**Lemma 22.** $\frac{\sigma^2_O}{d^{\ell,H}_n d^{\ell,V}_n} \sum_{i',j'} \sum_{\substack{h_1,h_2\\k_1,k_2}} \widehat{\Theta}^{\ell}_{a'i',b'j'} \widetilde{W}^{\ell h_1,O}_{n,k_1 i} \widetilde{W}^{\ell h_2,O}_{n,k_2 j} \widetilde{G}^{\ell h_1}_{n,ac_1} \widetilde{G}^{\ell h_2}_{n,bc_2} \frac{\partial V^{\ell h_1}_{n,c_1 k_1}}{\partial g^{\ell-1}_{n,a'i'}} \frac{\partial V^{\ell h_2}_{n,c_2 k_2}}{\partial g^{\ell-1}_{n,b'j'}}$ *converges in probability to*

$$\delta_{i=j} \delta_{\substack{c_1=a'\\c_2=b'}} \sigma^2_{OV} \widetilde{\Theta}^{\ell}_{a'b'}(x, x') \, \mathbb{E}[\widetilde{G}^{\ell 1}_{ac_1}(x) \widetilde{G}^{\ell 1}_{bc_2}(x')] ,$$

*Proof of Lemma 22.* Note that

$$\begin{aligned} &\frac{1}{d^{\ell,H}_n d^{\ell,V}_n} \sum_{i',j'} \sum_{\substack{h_1,h_2\\k_1,k_2}} \widehat{\Theta}^{\ell}_{a'i',b'j'} \widetilde{W}^{\ell h_1,O}_{n,k_1 i} \widetilde{W}^{\ell h_2,O}_{n,k_2 j} \widetilde{G}^{\ell h_1}_{n,ac_1} \widetilde{G}^{\ell h_2}_{n,bc_2} \frac{\partial V^{\ell h_1}_{n,c_1 k_1}}{\partial g^{\ell-1}_{n,a'i'}} \frac{\partial V^{\ell h_2}_{n,c_2 k_2}}{\partial g^{\ell-1}_{n,b'j'}} \\ &= \delta_{\substack{c_1=a'\\c_2=b'}} \frac{\sigma^2_V}{d^{\ell,H}_n d^{\ell,V}_n d^{\ell-1}_n} \sum_{i',j'} \sum_{\substack{h_1,h_2\\k_1,k_2}} \widehat{\Theta}^{\ell}_{a'i',b'j'} \widetilde{W}^{\ell h_1,O}_{n,k_1 i} \widetilde{W}^{\ell h_2,O}_{n,k_2 j} \widetilde{W}^{\ell h_1,V}_{n,i'k_1} \widetilde{W}^{\ell h_2,V}_{n,j'k_2} \widetilde{G}^{\ell h_1}_{n,ac_1} \widetilde{G}^{\ell h_2}_{n,bc_2} := \delta_{\substack{c_1=a'\\c_2=b'}} \sigma^2_V \bar{S}_n . \end{aligned}$$

Further, $\mathbb{E}[\bar{S}_n] = \mathbb{E}[\widehat{\Theta}^{\ell}_{a'1,b'1} \widetilde{G}^{\ell 1}_{n,ac_1} \widetilde{G}^{\ell 1}_{n,bc_2}]$ by exchangeability. As in the proof of Lemma 21, the desired result could thus be obtained by an application of Chebyshev's inequality, $\mathbb{P}(|\bar{S}_n - \mathbb{E}\,\bar{S}_n| \geq \delta) \leq \delta^{-2}[\mathbb{E}[\bar{S}_n^2] - \{\mathbb{E}[\bar{S}_n]\}^2]$, if $\mathbb{E}[\bar{S}_n]$ converges to the desired limit and $|\mathbb{E}[\bar{S}_n^2] - \{\mathbb{E}[\bar{S}_n]\}^2| \to 0$ as $n \to \infty$.

To establish convergence of the mean, first note that $\widehat{\Theta}^{\ell}_{a'1b'1} \widetilde{G}^{\ell 1}_{n,ac_1} \widetilde{G}^{\ell 1}_{n,bc_2} \rightsquigarrow \widetilde{\Theta}^{\ell}_{a'b'} \widetilde{G}^{\ell 1}_{ac_1} \widetilde{G}^{\ell 1}_{bc_2}$ by Theorem 3, the continuous mapping theorem, $\widehat{\Theta}^{\ell}_{a'1b'1} \xrightarrow{P} \widetilde{\Theta}^{\ell}_{a'b'}$ (Yang, 2019b), and Lemma 30. Inspecting Theorem 29 and Lemma 28, we see it is

[11] $\widehat{\Theta}$ should technically also be subscripted with $n$ as all other variables dependent on the $\theta^{\leq \ell}_n$; we make an exception here and omit this from our notation as the number of subscripts of $\widehat{\Theta}$ is already high.

sufficient to show $\mathbb{E}[\bar{S}_n^2] \to \{\mathbb{E}[\bar{S}_*]\}^2$ to establish both convergence of the mean, and $\bar{S}_n \xrightarrow{P} \mathbb{E}[\bar{S}_*]$. We thus turn to $\mathbb{E}[\bar{S}_n^2]$

$$\frac{1}{(d_n^{\ell,H} d_n^{\ell,V} d_n^{\ell-1})^2} \sum_{\substack{i_1',j_1' \\ i_2',j_2'}} \sum_{\substack{h_1,h_2,h_3,h_4 \\ k_1,k_2,k_3,k_4}} \mathbb{E}\left[ \prod_{t=0}^{1} \widehat{\Theta}^{\ell}_{a'i'_{t+1},b'j'_{t+1}} \widetilde{W}^{\ell h_{2t+1},O}_{n,k_{2t+1}i} \widetilde{W}^{\ell h_{2t+2},O}_{n,k_{2t+2}j} \widetilde{W}^{\ell h_{2t+1},V}_{n,i'_{t+1}k_{2t+1}} \widetilde{W}^{\ell h_{2t+2},V}_{n,j'_{t+1}k_{2t+2}} \widetilde{G}^{\ell h_{2t+1}}_{n,ac_1} \widetilde{G}^{\ell h_{2t+2}}_{n,bc_2} \right].$$

We can thus restrict our attention to groups of terms that include at least $\mathcal{O}((d_n^{\ell,H} d_n^{\ell,V} d_n^{\ell-1})^2)$ of the summands, as long as the expectation of the square of each term can be bounded by a constant independent of the $h, k, i', j'$ and $n$ indices. Observe that

$$\mathbb{E}\left\{ \left[ \prod_{t=0}^{1} \widehat{\Theta}^{\ell}_{a'i'_{t+1},b'j'_{t+1}} \widetilde{W}^{\ell h_{2t+1},O}_{n,k_{2t+1}i} \widetilde{W}^{\ell h_{2t+2},O}_{n,k_{2t+2}j} \widetilde{W}^{\ell h_{2t+1},V}_{n,i'_{t+1}k_1} \widetilde{W}^{\ell h_{2t+2},V}_{n,j'_{t+1}k_2} \widetilde{G}^{\ell h_{2t+1}}_{n,ac_1} \widetilde{G}^{\ell h_{2t+2}}_{n,bc_2} \right]^2 \right\}$$

$$\lesssim \text{poly}\left( \max_{\substack{a',b' \in [d^s], i',j' \in \{1,2\} \\ z,z' \in \{x,x'\}}} \mathbb{E}[\widehat{\Theta}^{\ell}_{a'i',\,b'j'}(z,z')^4]\,,\, \max_{\substack{c,c' \in [d^s] \\ z \in \{x,x'\}}} \mathbb{E}[\widetilde{G}^{\ell 1}_{n,c,c'}(z)^8] \right), \tag{28}$$

and thus we can obtain the desired bound by applying Lemma 32. We thus shift our attention to the terms that are not $\mathrm{o}((d_n^{\ell,H} d_n^{\ell,V} d_n^{\ell-1})^2)$ of which there are three types: (i) $i=j$, $(h_1,k_1,i_1') = (h_2,k_2,j_2')$, and $(h_3,k_3,i_3') = (h_4,k_4,j_4')$; (ii) $i=j$, $(h_1,k_1,i_1') = (h_4,k_4,j_4')$, and $(h_2,k_2,i_2') = (h_3,k_3,j_3')$; (iii) $(h_1,k_1,i_1') = (h_3,k_3,i_3')$, and $(h_2,k_2,i_2') = (h_4,k_4,i_4')$. Hence the limit of $\mathbb{E}[\bar{S}_n^2]$ will up to a constant coincide with that of

$$\mathbb{E}\left[ \left( \widehat{\Theta}^{\ell}_{a'1,b'2} \widetilde{G}^{\ell 1}_{n,aa'} \widetilde{G}^{\ell 2}_{n,bb'} \right)^2 \right] + \delta_{i=j}\, \mathbb{E}\left[ \left( \widehat{\Theta}^{\ell}_{a'1,b'1} \widehat{\Theta}^{\ell}_{a'2,b'2} + \widehat{\Theta}^{\ell}_{a'1,b'2} \widehat{\Theta}^{\ell}_{a'2,b'1} \right) \widetilde{G}^{\ell 1}_{n,aa'} \widetilde{G}^{\ell 1}_{n,bb'} \widetilde{G}^{\ell 2}_{n,aa'} \widetilde{G}^{\ell 2}_{n,bb'} \right],$$

by exchangeability. Noticing $\widetilde{G}^{\ell}_n$ converges in distribution by Theorem 3 and the continuous mapping theorem, and the $\widehat{\Theta}^{\ell}$ converges in this distribution (Yang, 2019b), both integrands converge in distribution by the continuous mapping theorem and Lemma 30. Since the $\widetilde{G}^{\ell h}_n$ corresponding to different heads are independent in the limit (Theorem 3), and the limit of $\widehat{\Theta}^{\ell}_{a'i',b'j'}$ is non-zero only if $i'=j'$ (Yang, 2019b), application of Theorem 29 combined with the bound from Equation (28) concludes the proof. □

**Lemma 23.** $\frac{\sigma_O^2}{d_n^{\ell,H} d_n^{\ell,V}} \sum_{i',j'} \sum_{\substack{h_1,h_2 \\ k_1,k_2}} \widehat{\Theta}^{\ell}_{a'i',b'j'} \widetilde{W}^{\ell h_1,O}_{n,k_1 i} \widetilde{W}^{\ell h_2,O}_{n,k_2 j} V^{\ell h_1}_{n,c_1 k_1} V^{\ell h_2}_{n,c_2 k_2} \frac{\partial \widetilde{G}^{\ell h_1}_{n,ac_1}}{\partial G^{\ell h_1}_{n,ad_1}} \frac{\partial \widetilde{G}^{\ell h_2}_{n,bc_2}}{\partial G^{\ell h_2}_{n,bd_2}} \frac{\partial G^{\ell h_1}_{n,ad_1}}{\partial g^{\ell-1}_{n,a'i'}} \frac{\partial G^{\ell h_2}_{n,bd_2}}{\partial g^{\ell-1}_{n,b'j'}}$ *converges in probability to*

$$\delta_{\substack{i=j \\ \tau=\frac{1}{2}}}\, \sigma_{OV}^2 \sigma_{QK}^2 \widetilde{\Theta}^{\ell}_{a'b'}(x,x') \tilde{\kappa}^{\ell}_{c_1 c_2}(x,x') \left( \delta_{\substack{d_1=a' \\ d_2=b'}} \tilde{\kappa}^{\ell}_{ab}(x,x') + \delta_{\substack{a'=a \\ b'=b}} \tilde{\kappa}^{\ell}_{d_1 d_2}(x,x') \right) \mathbb{E}\left[ \frac{\partial \widetilde{G}^{\ell 1}_{ac_1}(x)}{\partial G^{\ell 1}_{ad_1}(x)} \frac{\partial \widetilde{G}^{\ell 1}_{bc_2}(x')}{\partial G^{\ell 1}_{bd_2}(x')} \right].$$

*Proof of Lemma 23.* To make the notation more succinct, we define

$$\frac{\partial G^{\ell h_1}_{n,ad_1}}{\partial g^{\ell-1}_{n,a'i'}} = \frac{1}{(d_n^{\ell,G})^{\tau} \sqrt{d_n^{\ell-1}}} \sum_{u_1=1}^{d_n^{\ell,G}} \underbrace{\delta_{d_1=a'} \sigma_K Q^{h_1}_{au_1} \widetilde{W}^{\ell h_1,K}_{n,i'u_1} + \delta_{a'=a} \sigma_Q K^{\ell h_1}_{n,d_1 u_1} \widetilde{W}^{\ell h_1,Q}_{n,i'u_1}}_{:=\Gamma^{h_1}_{i'u_1}}. \tag{29}$$

which leads us to

$$\bar{S}_n = \frac{\sigma_O^2}{d_n^{\ell,H} d_n^{\ell,V} d_n^{\ell-1} (d_n^{\ell,G})^{2\tau}} \sum_{\substack{i',j' \\ h_1,h_2}} \sum_{\substack{k_1,k_2 \\ u_1,u_2}} \widehat{\Theta}^{\ell}_{a'i',b'j'} \widetilde{W}^{\ell h_1,O}_{n,k_1 i} \widetilde{W}^{\ell h_2,O}_{n,k_2 j} V^{\ell h_1}_{n,c_1 k_1} V^{\ell h_2}_{n,c_2 k_2} \frac{\partial \widetilde{G}^{\ell h_1}_{n,ac_1}}{\partial G^{\ell h_1}_{n,ad_1}} \frac{\partial \widetilde{G}^{\ell h_2}_{n,bc_2}}{\partial G^{\ell h_2}_{n,bd_2}} \Gamma^{h_1}_{n,i'u_1} \Gamma^{h_2}_{n,j'u_2}.$$

Unlike in the proof of Lemma 22, the mean

$$\mathbb{E}[\bar{S}_n] = \delta_{i=j} \frac{\sigma_{OV}^2}{d_n^{\ell-1} (d_n^{\ell,G})^{2\tau}} \sum_{i',j'} \sum_{u_1,u_2} \mathbb{E}\left[ \widehat{\Theta}^{\ell}_{a'i',b'j'} \frac{\langle g^{\ell-1}_{n,c_1 \cdot}, g^{\ell-1}_{n,c_2 \cdot} \rangle}{d_n^{\ell-1}} \frac{\partial \widetilde{G}^{\ell 1}_{n,ac_1}}{\partial G^{\ell 1}_{n,ad_1}} \frac{\partial \widetilde{G}^{\ell 1}_{n,bc_2}}{\partial G^{\ell 1}_{n,bd_2}} \Gamma^1_{n,i'u_1} \Gamma^1_{n,j'u_2} \right], \tag{30}$$

only eliminates some of the sums. This issue can be resolved with the help of Lemma 24.

**Lemma 24.** *The random variable*

$$\bar{S}_n^{h_1 h_2} := \frac{1}{d_n^{\ell-1}(d_n^{\ell,G})^{2\tau}} \sum_{i',j'} \sum_{u_1,u_2} \widehat{\Theta}^{\ell}_{a'i',b'j'} \Gamma^{h_1}_{n,i'u_1} \Gamma^{h_2}_{n,j'u_2} \,,$$

*converges in probability to*

$$\delta_{\tau=\frac{1}{2}} \delta_{h_1=h_2} \sigma_{QK}^2 \widetilde{\Theta}^{\ell}_{a'b'}(x,x') \left( \delta_{\substack{d_1=a' \\ d_2=b'}} \tilde{\kappa}^{\ell}_{ab}(x,x') + \delta_{\substack{a'=a \\ b'=b}} \tilde{\kappa}^{\ell}_{d_1 d_2}(x,x') \right) .$$

*Proof.* Notice

$$\begin{aligned} \mathbb{E}[\bar{S}_n^{h_1 h_2}] =& \delta_{h_1=h_2} \frac{\sigma_{QK}^2}{(d_n^{\ell,G})^{2\tau-1}} \mathbb{E}\left[ \widehat{\Theta}^{\ell}_{a'1,b'1} \left( \delta_{\substack{d_1=a' \\ d_2=b'}} \frac{\langle g^{\ell}_{n,a\cdot}, g^{\ell}_{n,b\cdot} \rangle}{d_n^{\ell-1}} + \delta_{\substack{a'=a \\ b'=b}} \frac{\langle g^{\ell}_{n,d_1\cdot}, g^{\ell}_{n,d_2\cdot} \rangle}{d_n^{\ell-1}} \right) \right] + \\ & \delta_{h_1=h_2} \frac{\sigma_{QK}^2}{(d_n^{\ell,G})^{2\tau-1}} \mathbb{E}\left[ \widehat{\Theta}^{\ell}_{a'1,b'1} \left( \delta_{\substack{d_1=a' \\ b'=b}} \frac{g^{\ell-1}_{n,a1} g^{\ell-1}_{n,d_2 1}}{d_n^{\ell-1}} + \delta_{\substack{a'=a \\ d_2=b'}} \frac{g^{\ell-1}_{n,d_1 1} g^{\ell-1}_{n,b1}}{d_n^{\ell-1}} \right) \right] , \end{aligned} \tag{31}$$

and thus we can combine the fact that $\widehat{\Theta}^{\ell}_{a'i',b'j'} \xrightarrow{P} \delta_{i'=j'} \widetilde{\Theta}^{\ell}_{a'b'}$ (Yang, 2019b) with Lemmas 32 and 33, the continuous mapping theorem, and Theorem 29 to obtain

$$\mathbb{E}[\bar{S}_n^{h_1 h_2}] \to \delta_{\tau=\frac{1}{2}} \delta_{h_2=h_2} \sigma_{QK}^2 \widetilde{\Theta}^{\ell}_{a'b'}(x,x') \left( \delta_{\substack{d_1=a' \\ d_2=b'}} \tilde{\kappa}^{\ell}_{ab}(x,x') + \delta_{\substack{a'=a \\ b'=b}} \tilde{\kappa}^{\ell}_{d_1 d_2}(x,x') \right) ,$$

as $n \to \infty$. To obtain the convergence of $\bar{S}_n^{h_1 h_2}$ to in probability, it is thus sufficient to show $|\mathbb{E}[(\bar{S}_n^{h_1 h_2})^2] - \{\mathbb{E}[\bar{S}_n^{h_1 h_2}]\}^2|$ converges to zero as $n \to \infty$. Substituting

$$\mathbb{E}[(\bar{S}_n^{h_1 h_2})^2] = \frac{1}{(d_n^{\ell-1}(d_n^{\ell,G})^{2\tau})^2} \sum_{\substack{i_1',j_1' \\ i_2',j_2'}} \sum_{\substack{u_1,u_2 \\ u_3,u_4}} \mathbb{E}\left[ \widehat{\Theta}^{\ell}_{a'i_1',b'j_1'} \widehat{\Theta}^{\ell}_{a'i_2',b'j_2'} \Gamma^1_{n,i_1'u_1} \Gamma^1_{n,j_1'u_2} \Gamma^1_{n,i_2'u_3} \Gamma^1_{n,j_2'u_4} \right] ,$$

we can once again restrict our attention to groups of terms that include at least $\mathcal{O}((d_n^{\ell-1}(d_n^{\ell,G})^{2\tau})^2)$ of the summands as long as each term can be bounded by a constant independent of the $i', j'$ and $n$ indices. This bound can be again obtained by a repeated application of Hölder's inequality, followed by Lemma 32. We can thus shift our attention to the terms for which either of the following holds: (i) $(i_1', u_1) = (j_1', u_2)$ and $(i_2', u_3) = (j_2', u_4)$; or (ii) $(i_1', u_1) = (i_2', u_3)$ and $(j_1', u_2) = (j_2', u_4)$; or $(i_1', u_1) = (j_2', u_4)$ and $(j_1', u_2) = (i_2', u_3)$.

As in Equation (31), we can use the above established boundedness to see that the contribution from any terms that involve the cross terms like $Q^{\ell 1}_{a1} K^{\ell 1}_{n,d_1 1} \widetilde{W}^{\ell 1,K}_{n,11} \widetilde{W}^{\ell 1,Q}_{n11}$, and terms with either of $i_1' \neq j_2'$ and $i_2' \neq j_2'$ (the limit of $\widehat{\Theta}_{a'i',b',j'}$ is zero if $i' \neq j'$), vanish. With some algebraic manipulation analogous to that in Equation (31), we thus obtain

$$\mathbb{E}[(\bar{S}_n^{h_1 h_2})^2] \to \delta_{\tau=\frac{1}{2}} \delta_{h_1=h_2} \left[ \sigma_{QK}^2 \widetilde{\Theta}^{\ell}_{a'b'}(x,x') \left( \delta_{\substack{d_1=a' \\ d_2=b'}} \tilde{\kappa}^{\ell}_{ab}(x,x') + \delta_{\substack{a'=a \\ b'=b}} \tilde{\kappa}^{\ell}_{d_1 d_2}(x,x') \right) \right]^2 ,$$

as desired. Application of Cheybshev's inequality concludes the proof. □

With $\bar{S}_n^{h_1 h_2}$ defined as in Lemma 24, we can revisit Equation (30)

$$\mathbb{E}[\bar{S}_n] = \delta_{i=j} \sigma_{OV}^2 \, \mathbb{E}\left[ \bar{S}_n^{11} \frac{\langle g^{\ell-1}_{n,c_1\cdot}, g^{\ell-1}_{n,c_2\cdot} \rangle}{d_n^{\ell-1}} \frac{\partial \widetilde{G}^{\ell 1}_{n,ac_1}}{\partial G^{\ell 1}_{n,ad_1}} \frac{\partial \widetilde{G}^{\ell 1}_{n,bc_2}}{\partial G^{\ell 1}_{n,bd_2}} \right] .$$

Note that the first two terms converge in probability to constant by Lemmas 24 and 33 and the continuous mapping theorem,

$$\frac{\partial \widetilde{G}^{\ell 1}_{n,ac_1}}{\partial G^{\ell 1}_{n,ad_1}} \frac{\partial \widetilde{G}^{\ell 1}_{n,bc_2}}{\partial G^{\ell 1}_{n,bd_2}} \rightsquigarrow \frac{\partial \widetilde{G}^{\ell 1}_{ac_1}}{\partial G^{\ell 1}_{ad_1}} \frac{\partial \widetilde{G}^{\ell 1}_{bc_2}}{\partial G^{\ell 1}_{bd_2}} \,,$$

if $\tau = \frac{1}{2}$ by Theorem 3, and in probability to a constant if $\tau = 1$ (Yang, 2019b, appendix A), both using the assumed continuity of $\nabla \zeta$. Since $\nabla \zeta$ is also assumed to be bounded, we can combine Hölder's inequality with Lemma 32 to establish uniform integrability (see the proof of Lemma 24 for the bound on $\bar{S}^{11}_n$) via Lemma 28, and with that convergence of $\mathbb{E}[\bar{S}_n]$ by Lemma 30 and Theorem 29, yielding

$$\mathbb{E}[\bar{S}_n] \to \delta_{\substack{i=j \\ \tau=\frac{1}{2}}} \sigma^2_{OV} \sigma^2_{QK} \widetilde{\Theta}^{\ell}_{a'b'}(x,x') \tilde{\kappa}^{\ell}_{c_1 c_2}(x,x') \left( \delta_{\substack{d_1=a' \\ d_2=b'}} \tilde{\kappa}^{\ell}_{ab}(x,x') + \delta_{\substack{a'=a \\ b'=b}} \tilde{\kappa}^{\ell}_{d_1 d_2}(x,x') \right) \mathbb{E}\left[ \frac{\partial \widetilde{G}^{\ell 1}_{ac_1}(x)}{\partial G^{\ell 1}_{ad_1}(x)} \frac{\partial \widetilde{G}^{\ell 1}_{bc_2}(x')}{\partial G^{\ell 1}_{bd_2}(x')} \right] .$$

Convergence of $\bar{S}_n$ to the same constant can be obtained via Chebyshev's inequality by proving $|\mathbb{E}[\bar{S}^2_n] - \{\mathbb{E}[\bar{S}_n]\}^2| \to 0$. Using the notation from Lemma 24, the second moment of $\bar{S}_n$ can be written as

$$\mathbb{E}[\bar{S}^2_n] = \frac{\sigma^4_O}{(d^{\ell,H}_n d^{\ell,V}_n)^2} \sum_{\substack{h_1,h_2,h_3,h_4 \\ k_1,k_2,k_3,k_4}} \mathbb{E}\left[ \prod_{t=0}^{1} \bar{S}^{h_{2t+1} h_{2t+2}}_n \widetilde{W}^{\ell h_{2t+1},O}_{n,k_{2t+1} i} \widetilde{W}^{\ell h_{2t+2},O}_{n,k_{2t+2} j} V^{\ell h_{2t+1}}_{n,c_1 k_{2t+1}} V^{\ell h_{2t+2}}_{n,c_2 k_{2t+2}} \frac{\partial \widetilde{G}^{\ell h_{2t+1}}_{n,ac_1}}{\partial G^{\ell h_{2t+1}}_{n,ad_1}} \frac{\partial \widetilde{G}^{\ell h_{2t+2}}_{n,bc_2}}{\partial G^{\ell h_{2t+2}}_{n,bd_2}} \right] .$$

Because $\nabla \zeta$ is bounded by assumption, we can again use Hölder's inequality together with Lemma 32 to bound each of the summands by a constant independent of the $h, k$ and $n$ indices. This means we can restrict our attention only to groups of terms that include at least $\mathcal{O}((d^{\ell,H}_n d^{\ell,V}_n)^2)$ of the summands. These fall into one of the following three categories: (i) $i = j$, $(h_1,k_1) = (h_2,k_2)$, and $(h_3,k_3) = (h_4,k_4)$; (i) $(h_1,k_1) = (h_3,k_3)$, and $(h_2,k_4) = (h_4,k_4)$; and (iii) $i = j$, $(h_1,k_1) = (h_4,k_4)$, and $(h_2,k_2) = (h_3,k_3)$. Using exchangeability, we thus obtain

$$\begin{aligned} \mathbb{E}[\bar{S}^2_n] =& \sigma^4_{OV} \, \mathbb{E}\left[ \bar{S}^{12}_n \bar{S}^{12}_n \frac{\langle g^{\ell-1}_{n,c_1\cdot}, g^{\ell-1}_{n,c_1\cdot} \rangle}{d^{\ell-1}_n} \frac{\langle g^{\ell-1}_{n,c_2\cdot}, g^{\ell-1}_{n,c_2\cdot} \rangle}{d^{\ell-1}_n} \frac{\partial \widetilde{G}^{\ell 1}_{n,ac_1}}{\partial G^{\ell 1}_{n,ad_1}} \frac{\partial \widetilde{G}^{\ell 1}_{n,bc_2}}{\partial G^{\ell 1}_{n,bd_2}} \frac{\partial \widetilde{G}^{\ell 2}_{n,ac_1}}{\partial G^{\ell 2}_{n,ad_1}} \frac{\partial \widetilde{G}^{\ell 2}_{n,bc_2}}{\partial G^{\ell 2}_{n,bd_2}} \right] + \\ & \delta_{i=j} \sigma^4_{OV} \, \mathbb{E}\left[ (\bar{S}^{11}_n \bar{S}^{22}_n + \bar{S}^{12}_n \bar{S}^{21}_n) \left( \frac{\langle g^{\ell-1}_{n,c_1\cdot}, g^{\ell-1}_{n,c_2\cdot} \rangle}{d^{\ell-1}_n} \right)^2 \frac{\partial \widetilde{G}^{\ell 1}_{n,ac_1}}{\partial G^{\ell 1}_{n,ad_1}} \frac{\partial \widetilde{G}^{\ell 1}_{n,bc_2}}{\partial G^{\ell 1}_{n,bd_2}} \frac{\partial \widetilde{G}^{\ell 2}_{n,ac_1}}{\partial G^{\ell 2}_{n,ad_1}} \frac{\partial \widetilde{G}^{\ell 2}_{n,bc_2}}{\partial G^{\ell 2}_{n,bd_2}} \right] + \\ & \mathrm{o}((d^{\ell,H}_n d^{\ell,V}_n)^2) \, . \end{aligned}$$

Note by the assumed continuity of $\nabla \zeta$, Theorem 3, Lemmas 24 and 33, the continuous mapping theorem, and Lemma 30, both the integrands converge in distribution, which, combined with the above derived bound and Theorem 29, implies

$$\mathbb{E}[\bar{S}^2_n] \to \delta_{\substack{i=j \\ \tau=\frac{1}{2}}} \left[ \sigma^2_{OV} \sigma^2_{QK} \widetilde{\Theta}^{\ell}_{a'b'}(x,x') \tilde{\kappa}^{\ell}_{c_1 c_2}(x,x') \left( \delta_{\substack{d_1=a' \\ d_2=b'}} \tilde{\kappa}^{\ell}_{ab}(x,x') + \delta_{\substack{a'=a \\ b'=b}} \tilde{\kappa}^{\ell}_{d_1 d_2}(x,x') \right) \mathbb{E}\left[ \frac{\partial \widetilde{G}^{\ell 1}_{ac_1}(x)}{\partial G^{\ell 1}_{ad_1}(x)} \frac{\partial \widetilde{G}^{\ell 1}_{bc_2}(x')}{\partial G^{\ell 1}_{bd_2}(x')} \right] \right]^2$$

where we have used the fact that $\bar{S}^{h_1 h_2}_n$ converges in probability to zero whenever $h_1 \neq h_2$ (Lemma 24), and the asymptotic indepedence of $\widetilde{G}^{\ell 1}_n$ and $\widetilde{G}^{\ell 2}_n$ (Theorem 3 if $\tau = \frac{1}{2}$, resp. (Yang, 2019b, appendix A) if $\tau = 1$). □

**Lemma 25.** $\frac{\sigma^2_O}{d^{\ell,H}_n d^{\ell,V}_n} \sum_{i',j'} \sum_{\substack{h_1,h_2 \\ k_1,k_2}} \widehat{\Theta}^{\ell}_{a'i',b'j'} \widetilde{W}^{\ell h_1,O}_{n,k_1 i} \widetilde{W}^{\ell h_2,O}_{n,k_2 j} \widetilde{G}^{\ell h_1}_{n,ac_1} V^{\ell h_2}_{n,c_2 k_2} \frac{\partial V^{\ell h_1}_{n,c_1 k_1}}{\partial g^{\ell-1}_{n,a'i'}} \frac{\partial \widetilde{G}^{\ell h_2}_{n,bc_2}}{\partial G^{\ell h_2}_{n,bd_2}} \frac{\partial G^{\ell h_2}_{n,bd_2}}{\partial g^{\ell-1}_{n,b'j'}} \xrightarrow{P} 0 \,.$

*Proof of Lemma 25.* Observing that $\frac{\partial V^{\ell h_1}_{n,c_1 k_1}}{\partial g^{\ell-1}_{n,a'i'}} = \delta_{c_1=a'} \sigma_V \frac{\widetilde{W}^{\ell h_1,V}_{n,i'k_1}}{\sqrt{d^{\ell-1}_n}}$ and setting

$$\bar{S}_n = \frac{\sigma_V}{d^{\ell,H}_n d^{\ell,V}_n \sqrt{d^{\ell-1}_n}} \sum_{i',j'} \sum_{\substack{h_1,h_2 \\ k_1,k_2}} \widehat{\Theta}^{\ell}_{a'i',b'j'} \widetilde{W}^{\ell h_1,O}_{n,k_1 i} \widetilde{W}^{\ell h_2,O}_{n,k_2 j} \widetilde{W}^{\ell h_1,V}_{n,i'k_1} V^{\ell h_2}_{n,c_2 k_2} \widetilde{G}^{\ell h_1}_{n,ac_1} \frac{\partial \widetilde{G}^{\ell h_2}_{n,bc_2}}{\partial G^{\ell h_2}_{n,bd_2}} \frac{\partial G^{\ell h_2}_{n,bd_2}}{\partial g^{\ell-1}_{n,b'j'}} \,,$$

we immediately see that

$$\mathbb{E}[\bar{S}_n] = \delta_{i=j} \frac{\sigma_V^2}{d_n^{\ell-1}} \sum_{i',j'} \mathbb{E}\left[ \widehat{\Theta}^{\ell}_{a'i',b'j'} g^{\ell-1}_{n,c_2 i'} \widetilde{G}^{\ell 1}_{n,ac_1} \frac{\partial \widetilde{G}^{\ell 1}_{n,bc_2}}{\partial G^{\ell 1}_{n,bd_2}} \frac{\partial G^{\ell 1}_{n,bd_2}}{\partial g^{\ell-1}_{n,b'j'}} \right] .$$

Analogously to the proof of Lemma 23, we define

$$\bar{S}^h_n = \frac{1}{d_n^{\ell-1}} \sum_{i',j'} \widehat{\Theta}^{\ell}_{a'i',b'j'} g^{\ell-1}_{n,c_2 i'} \frac{\partial G^{\ell h}_{n,bd_2}}{\partial g^{\ell-1}_{n,b'j'}} \tag{32}$$

$$= \frac{1}{(d_n^{\ell-1})^{3/2} (d_n^{\ell,G})^{\tau}} \sum_{i',j'} \sum_{u_1=1}^{d_n^{\ell,G}} \widehat{\Theta}^{\ell}_{a'i',b'j'} g^{\ell-1}_{n,c_2 i'} \Big( \underbrace{\delta_{d_2=b'} \sigma_K Q^{\ell h}_{n,bu_1} \widetilde{W}^{\ell h,K}_{n,j'u_1} + \delta_{b'=b} \sigma_Q K^{\ell h}_{n,d_2 u_1} \widetilde{W}^{\ell h,Q}_{n,j'u_1}}_{=:\Gamma^h_{n,j'u_1}} \Big) , \tag{33}$$

and make use of an auxiliary lemma.

**Lemma 26.** $\bar{S}^h_n \xrightarrow{P} 0$.

*Proof.* Observe $\mathbb{E}[\bar{S}^h_n] = 0$ if $\tau = \frac{1}{2}$ (independence of key and query weights), and

$$\mathbb{E}[\bar{S}^h_n] = \frac{\sigma_{QK}}{(d_n^{\ell-1})^2} \sum_{i',j'} \mathbb{E}\left[ \widehat{\Theta}^{\ell}_{a'i',b'j'} g^{\ell-1}_{n,c_2 i'} \left( \delta_{d_2=b'} g^{\ell-1}_{n,bj'} + \delta_{b'=b} g^{\ell-1}_{n,d_2 j'} \right) \right] ,$$

if $\tau = 1$ (key and query weights are equal a.s.). Since each of the summands can be bounded by a constant indpendent of the $i', j'$ and $n$ indices by Lemma 32, we can restrict our focus to the terms for which $i' \neq j'$, yielding

$$\mathbb{E}[\bar{S}^h_n] = \sigma_{QK} \frac{d_n^{\ell-1}(d_n^{\ell-1} - 1)}{(d_n^{\ell-1})^2} \mathbb{E}\left[ \widehat{\Theta}^{\ell}_{a'1,b'2} g^{\ell-1}_{n,c_2 1} \left( \delta_{d_2=b'} g^{\ell-1}_{n,bj'} + \delta_{b'=b} g^{\ell-1}_{n,d_2 j'} \right) \right] + \mathrm{o}((d_n^{\ell-1})^2) ,$$

Since $\widehat{\Theta}^{\ell}_{a'1,b'2} \xrightarrow{P} 0$ (Yang, 2019b), and the $g^{\ell-1}_{n,c_2 i'} g^{\ell-1}_{n,bj'}$ products converge in distribution by continuity of the assumed $\phi$ and the continuous mapping theorem, the integrand converges to zero in distribution by Lemma 30. Using Lemmas 28 and 32 and Theorem 29, we again establish $\mathbb{E}[\bar{S}^h_n] \to 0$.

To obtain convergence in probability, observe

$$\mathbb{E}[(\bar{S}^h_n)^2] = \frac{1}{(d_n^{\ell-1})^3 (d_n^{\ell,G})^{2\tau}} \sum_{\substack{i'_1,j'_1 \\ i'_2,j'_2}} \sum_{u_1,u_2} \mathbb{E}\left[ \widehat{\Theta}^{\ell}_{a'i'_1,b'j'_1} \widehat{\Theta}^{\ell}_{a'i'_2,b'j'_2} g^{\ell-1}_{n,c_2 i'_1} g^{\ell-1}_{n,c_2 i'_2} \Gamma^h_{n,j'_1 u_1} \Gamma^h_{n,j'_2 u_2} \right] ,$$

and note that we can again bound each of the summands using Hölder's inequality and Lemma 32 as in to Lemma 24. We can thus restrict our attention to groups of terms that include at least $\mathcal{O}((d_n^{\ell-1})^3 (d_n^{\ell,G})^{2\tau})$ of the summands. If $\tau = 1$, we can thus focus on $u_1 \neq u_2$, in which case integrating $\Gamma^h_{n,j'_1 u_1} \Gamma^h_{n,j'_2 u_2}$ over key and query weights will yield an additional $d_n^{\ell-1}$ factor, for example

$$\mathbb{E}[Q^{\ell h}_{bu_1} \widetilde{W}^{\ell h,K}_{n,j'_1 u_1} Q^{\ell h}_{bu_2} \widetilde{W}^{\ell h,K}_{n,j'_2 u_2}] = \frac{\sigma_Q^2}{d_n^{\ell}} g_{n,b'j'_1} g_{n,b'j'_2} ,$$

using the equality of key and query weights. Since $\widehat{\Theta}^{\ell}_{a'i',b'j'}$ converges in probability to zero whenever $i' \neq j'$ (Yang, 2019b), and there are only $(d_n^{\ell-1})^2$ terms for which $i'_1 = j'_1$ and $i'_2 = j'_2$, we can use the continuous mapping theorem, Lemma 30, and Theorem 29 to establish that $\mathbb{E}[(\bar{S}^h_n)^2] \to 0$. If $\tau = \frac{1}{2}$, all terms for which $u_1 \neq 0$ will have zero expectation (independence of key and query weights), and thus analogous argument to the one for $\tau = 1$. ☐

With Lemma 26 at hand, we can simplify

$$\mathbb{E}[\bar{S}_n] = \delta_{i=j} \sigma_V^2 \, \mathbb{E}\left[ \bar{S}^1_n \widetilde{G}^{\ell 1}_{n,ac_1} \frac{\partial \widetilde{G}^{\ell 1}_{n,bc_2}}{\partial G^{\ell 1}_{n,bd_2}} \right] ,$$

and use the assumed continuity of $\nabla \zeta$ together with the continuous mapping theorem and Lemma 30 to establish that the integrand converges in distribution to zero. Since $\nabla \zeta$ is bounded by assumption, we can used Hölder's inequality and Lemma 32 to establish uniform integrability via Lemma 28 (see the proof of Lemma 26 for the bound on $\bar{S}_n^1$). We thus have $\mathbb{E}[\bar{S}_n] \to 0$ by Theorem 29.

To establish $\bar{S}_n \xrightarrow{P} 0$, it is sufficient to show $\mathbb{E}[(\bar{S}_n)^2] \to 0$ and apply Chebyshev's inequality.[12] We have

$$\mathbb{E}[(\bar{S}_n)^2] = \frac{\sigma_V^2}{(d_n^{\ell,H} d_n^{\ell,V} d_n^{\ell-1})^2} \sum_{\substack{i_1', j_1' \\ i_2', j_2'}} \sum_{\substack{h_1, h_2, h_3, h_4 \\ k_1, k_2, k_3, k_4}} \mathbb{E}\left[ \prod_{t=0}^{1} \widehat{\Theta}_{a' i_{t+1}', b' j_{t+1}'}^{\ell} \widetilde{W}_{n, k_{2t+1} i}^{\ell h_{2t+1}, O} \widetilde{W}_{n, k_{2t+2} j}^{\ell h_{2t+2}, O} \right.$$
$$\left. \widetilde{W}_{n, i_{t+1}' k_{2t+1}}^{\ell h_{2t+1}, V} V_{n, c_2 k_{2t+2}}^{\ell h_{2t+2}} \widetilde{G}_{n, a c_1}^{\ell h_{2t+1}} \frac{\partial \widetilde{G}_{n, b c_2}^{\ell h_{2t+2}}}{\partial G_{n, b d_2}^{\ell h_{2t+2}}} \sqrt{d_n^{\ell-1}} \frac{\partial G_{n, b d_2}^{\ell h_{2t+2}}}{\partial g_{n, b' j_{t+1}'}^{\ell-1}} \right] ,$$

where notice we are multiplying $\frac{\partial G_{n,bd_2}^{\ell h_{2t+2}}}{\partial g_{n,b' j_{t+1}'}^{\ell-1}}$ by $\sqrt{d_n^{\ell-1}}$ as this term scales as $(d_n^{\ell-1})^{-1/2}$ (see Equation (32)). Since $\nabla \zeta$ is bounded by assumption, we can use the Hölder's inequality to bound each of the summands by

$$\operatorname{poly}\left( \max_{\substack{a', b' \in [d^s], i', j' \in \{1,2\} \\ z, z' \in \{x, x'\}}} \mathbb{E}[\widehat{\Theta}_{a' i', b' j'}^{\ell}(z, z')^4] , \max_{\substack{c, c' \in [d^s] \\ z \in \{x, x'\}}} \mathbb{E}[\widetilde{G}_{n, cc'}^{\ell 1}(z)^8] , \max_{\substack{c \in [d^s] \\ z \in \{x, x'\}}} \mathbb{E}\, | g_{c1}^{n, \ell-1}(z) |^{16} \right) ,$$

which will be bounded by a constant independent of the $i', j', h, k$ and $n$ by Lemma 32. By Lemma 28, we can thus restrict our attention to the terms that are not $\mathrm{o}((d_n^{\ell,H} d_n^{\ell,V} d_n^{\ell-1})^2)$, which fall into one of the following three categories: (i) $i = j$, $(h_1, k_1) = (h_2, k_2)$, and $(h_3, k_3) = (h_4, k_4)$; (ii) $(h_1, k_1, i_1') = (h_3, k_3, i_2')$, and $(h_2, k_2, j_1') = (h_4, k_4, j_2')$; (iii) $i = j$, $(h_1, k_1) = (h_3, k_3)$, and $(h_2, k_2) = (h_4, k_4)$. Using exchangeability, we thus obtain

$$\mathbb{E}[(\bar{S}_n)^2] = \sigma_V^4 \, \mathbb{E}\left[ \frac{\langle g_{n, c_2 \cdot}^{\ell-1}, g_{n, c_2 \cdot}^{\ell-1} \rangle}{d_n^{\ell-1}} \left( \widehat{\Theta}_{a'1, b'2}^{\ell} \widetilde{G}_{n, a c_1}^{\ell 1} \frac{\partial \widetilde{G}_{n, b c_2}^{\ell 1}}{\partial G_{n, b d_2}^{\ell 1}} \sqrt{d_n^{\ell-1}} \frac{\partial G_{n, b d_2}^{\ell 2}}{\partial g_{n, b' 2}^{\ell-1}} \right)^2 \right] + \tag{34}$$

$$\delta_{i=j} 2 \sigma_V^2 \, \mathbb{E}\left[ \bar{S}_n^1 \bar{S}_n^2 \widetilde{G}_{n, a c_1}^{\ell 1} \widetilde{G}_{n, a c_1}^{\ell 2} \frac{\partial \widetilde{G}_{n, b c_2}^{\ell 1}}{\partial G_{n, b d_2}^{\ell 1}} \frac{\partial \widetilde{G}_{n, b c_2}^{\ell 2}}{\partial G_{n, b d_2}^{\ell 2}} \right] + \mathrm{o}((d_n^{\ell,H} d_n^{\ell,V} d_n^{\ell-1})^2) , \tag{35}$$

where we have used $\mathbb{E}[\widetilde{W}_{n, i_1' 1}^{\ell 1, V} V_{n, c_2 1}^{\ell 1} \widetilde{W}_{n, i_2' 2}^{\ell 2, V} V_{n, c_2 2}^{\ell 2}] = \frac{\sigma_V^2}{d_n^{\ell-1}} g_{n, c_2 i_1'}^{\ell-1} g_{n, c_2 i_2'}^{\ell-1}$ and the definition of $\bar{S}_n^h$ from Equation (32). We prove convergence of both of these expectations to zero separately.

Starting with the second expectation in Equation (34), we can use the assumed continuity of $\nabla \zeta$, Theorem 3, Equation (32), the continuous mapping theorem, and Lemma 30 to establish that the integrand converges in distribution to zero. Because $\nabla \zeta$ is bounded by assumption, we can combine Hölder's inequality and Lemma 32 to establish uniform integrability via Lemma 28 (see the proof of Lemma 26 for the bound on $\bar{S}_n^h$), and thus convergence of the expectation to zero by Theorem 29.

For the first expectation in Equation (34), note that the absolute value of the expectation can be upper bounded by

$$\mathbb{E}\left[ \left| \frac{\langle g_{n, c_2 \cdot}^{\ell-1}, g_{n, c_2 \cdot}^{\ell-1} \rangle}{d_n^{\ell-1}} \right| (\widehat{\Theta}_{a'1, b'2}^{\ell})^2 \left( \left( \widetilde{G}_{n, a c_1}^{\ell 1} \frac{\partial \widetilde{G}_{n, b c_2}^{\ell 1}}{\partial G_{n, b d_2}^{\ell 1}} \right)^2 + d_n^{\ell-1} \left( \frac{\partial G_{n, b d_2}^{\ell 2}}{\partial g_{n, b' 2}^{\ell-1}} \right)^2 \right) \right] ,$$

where, when multiplied out, we can use that $\widehat{\Theta}_{a'1, b'2}^{\ell} \xrightarrow{P} 0$ (Yang, 2019b), and an argument analogous to the one above—using Lemma 33 and the continuous mapping theorem to obtain convergence in probability for the inner product—to establish convergence to zero. Finally, for the second term, observe

$$\sqrt{d_n^{\ell-1}} \frac{\partial G_{n, b d_2}^{\ell 2}}{\partial g_{n, b' 2}^{\ell-1}} = \frac{1}{(d_n^{\ell,G})^{\tau}} \sum_{u_1=1}^{d_n^{\ell,G}} \delta_{d_2 = b'} \sigma_K Q_{n, b u_1}^{\ell h} \widetilde{W}_{n, j' u_1}^{\ell h, K} + \delta_{b' = b} \sigma_Q K_{n, d_2 u_1}^{\ell h} \widetilde{W}_{n, j' u_1}^{\ell h, Q} ,$$

[12] We will be using the explicit parenthesis here to distinguish between $\bar{S}_n^h$ with $h = 2$, and $(\bar{S}_n)^2$.

which converges in probability to a constant if $\tau = 1$ by Lemma 31 (using (Yang, 2019b) to establish convergence in distribution of the keys and queries, and Lemma 32 for the moment bound). If $\tau = \frac{1}{2}$, the sum will converge in distribution (Yang, 2019b) and since the rest of the term in the expectation converge in probability, their product converges in distribution by Lemma 30. One can then again combine Hölder's inequality, Lemmas 28 and 32 and Theorem 29 to obtain convergence of the expectation to zero. Hence $\mathbb{E}[(\bar{S}_n)^2] \to 0$, implying $\bar{S}_n \xrightarrow{P} 0$ as desired. □

### B.3. Expressivity of $d^{-1}$ and $d^{-1/2}$ induced attention kernels

**Proposition 2.** *There is no set of attention coefficients $\{\bar{\zeta}_{ai}^x \in \mathbb{R}\colon a, i \in [d^s], x \in \mathcal{X}\}$ such that for* all *positive semidefinite kernels $\tilde{\kappa}$ simultaneously*

$$\sum_{i,j=1}^{d^s} \tilde{\kappa}_{ij}(x, x')\bar{\zeta}_{ai}^x \bar{\zeta}_{bj}^{x'} = \sum_{i=1}^{d_f} \tilde{\kappa}_{N_a(i)N_b(i)}(x, x')\frac{1}{d_f},$$

*where $d_f$ is the dimension of the (flattened) convolutional filter, $N_a, N_b \subset [d^s]$ are the ordered subsets of pixels which are used to compute the new values of pixels a and b, respectively, and $N_a(i), N_b(i)$ are the $i^{th}$ pixels in $N_a, N_b$.*

*Proof of Proposition 2.* Consider $\kappa_{aa}^{\text{CNN}}(x, x) = \sum_{i=1}^{d_f} \tilde{\kappa}_{N_a(i)N_a(i)}(x, x)\frac{1}{d_f}$, and the corresponding attention kernel $\kappa_{aa}^{\text{ATTN}}(x, x) = \sum_{i,j=1}^{d^s} \tilde{\kappa}_{ij}(x, x)\bar{\zeta}_{ai}^x\bar{\zeta}_{aj}^x$. Note that $\kappa_{aa}^{\text{CNN}}(x, x)$ is just sum of terms on a subset of the diagonal of $\tilde{\kappa}(x, x)$. Hence it must be that $\bar{\zeta}_{ai}^x = \pm(d_f)^{-1/2}$ since we require that the same set of coefficients $\{\bar{\zeta}_{ai}^x\colon i \in [d^s]\}$ works all kernels $\tilde{\kappa}$ simultaneously, and thus for any $\tilde{\kappa}_{aa}(x, x)$ including all diagonal matrices with non-negative entries. Therefore $\bar{\zeta}_{ai}^x\bar{\zeta}_{aj}^x = \pm(d_f)^{-1}$ for all $i, j$, making signs the only degree of freedom.[13] We conclude by noting that we can make $\kappa_{aa}^{\text{ATTN}}(x, x) \neq \kappa_{aa}^{\text{CNN}}(x, x)$ by choosing $\tilde{\kappa}_{aa}(x, x)$ diagonal except for one pair of off-diagonal entries. □

**Proposition 4.** *Under the $d^{-1/2}$ scaling, there exists a distribution over $G$ such that for any $x, x'$ and $a, b, i, j$*

$$\mathbb{E}[\zeta(G(x))_{ai}\zeta(G(x'))_{bj}] = \begin{cases} \frac{1}{d_f}, & \exists k \in [d] \text{ s.t. } i = N_a(k), j = N_b(k), \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

*Proof of Proposition 4.* We provide a simple construction here, and expand on more realistic ones after the proof.

Consider $\Omega = [0, 1)$ with the usual Borel $\sigma$-algebra $\mathcal{B}$ and the Lebesgue measure $\lambda$. Let $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$ be the extended real axis and $\overline{\mathcal{B}}$ be the $\sigma$-algebra generated by the interval topology on $\overline{\mathbb{R}}$. Now construct the random variables $G_{ai}\colon \Omega \to \overline{\mathbb{R}}$ such that $G_{ai} = -\infty$ a.s. if $i \notin N_a$, and $G_{ai} = \infty \cdot \mathbb{1}_{A_{ai}}$ a.s. where and $A_{ai} := \left[\frac{i_{(a)}-1}{d_f}, \frac{i_{(a)}}{d}\right)$, with $i_{(a)}$ being the position of $i$ in the ordered set $N_a$, and $\infty \cdot 0$ is to be interpreted as 0. □

For a more realistic construction consider the usual $G(x) = d^{-1/2}Q(x)K(x)^\top$ but now additionally multiply each row of $Q(x$ by a corresponding scalar random variable $c_a^Q\colon \Omega \to \overline{\mathbb{R}}$, similarly each row of $K(x)$ by $c_a^K\colon \Omega \to \overline{\mathbb{R}}$. Then $G_{ab}(x) = d^{-1/2}c_a^Q c_b^K \langle Q_{a\cdot}(x), K_{b\cdot}(x)\rangle$ and thus one can achieve the desired result by setting up the joint distribution of $\{c_1^Q, \ldots, c_{d^s}^Q, c_1^K, \ldots, c_{d^s}^K\}$ in analogy to that in the above proof.

### B.4. Auxiliary results

**Lemma 27** (Billingsley, 1986, p. 19)**.** *Let $X, (X_n)_{n\geq 1}$ be random variables taking values in $(\mathbb{R}^{\mathbb{N}}, \mathcal{B}^N)$, $\mathcal{B}^{\mathbb{N}}$ the usual Borel $\sigma$-algebra. Then $X_n \rightsquigarrow X$ if and only if for each finite $J \subset \mathbb{N}$ and the corresponding projection $\Gamma^J\colon \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^J$, $\Gamma^J(X_n) \rightsquigarrow \Gamma^J(X)$ as $n \to \infty$.*

**Lemma 28** (Billingsley, 1986, p. 31)**.** *A sequence of real valued random variables $(X_n)_{n\geq 1}$ is uniformly integrable if*

$$\sup_n \mathbb{E}\,|X_n|^{1+\varepsilon} < \infty\,.$$

[13] As a side note, this degree of freedom disappears when $\bar{\zeta}$ is a limit of the softmax variables (non-negativity).

**Theorem 29** (Billingsley, 1986, theorem 3.5)**.** *If $(X_n)_{n\geq 1}$ are uniformly integrable and $X_n \rightsquigarrow X$, then $X$ is integrable and*

$$\mathbb{E}[X_n] \to \mathbb{E}[X] \, .$$

**Lemma 30** (Slutsky's lemmas)**.** *Let $X, (X_n)_{n\geq 1}$ and $(Y_n)_{n\geq 1}$ be real valued random variables defined on the same probability space, and assume $X_n \rightsquigarrow X$ and $Y_n \xrightarrow{P} c$ for some $c \in \mathbb{R}$. Then*

$$X_n Y_n \rightsquigarrow cX \, , \qquad\qquad X_n + Y_n \rightsquigarrow X + c \, . \tag{36}$$

**Lemma 31** (Weak LLN for exchangeable triangular arrays)**.** *Let $X_n \coloneqq \{X_{n,i} \colon i = 1, 2, \dots\}$ be an infinitely exchangeable sequence of random variables on $\mathbb{R}^{\mathbb{N}}$ s.t. $\limsup_{n\to\infty} \mathbb{E}\,|X_{n,1}|^{2+\varepsilon} < \infty$ for some $\varepsilon > 0$, and define $\bar{S}_n \coloneqq \frac{1}{d_n}\sum_{i=1}^{d_n} X_{n,i}$ , for some sequence $(d_n)_{n\geq 1}$ s.t. $\lim_{n\to\infty} d_n = \infty$. Assuming all $X_n$ are defined on the same space, if $X_n$ converges in distribution to some infinitely exchangeable $X_* = \{X_{*,i} \colon i = 1, 2, \dots\}$ s.t. $\mathbb{E}[X_{*,1}X_{*,2}] = (\mathbb{E}[X_{*,1}])^2$, then as $n \to \infty$, $\mathbb{E}[\bar{S}_n] \to \mathbb{E}[X_{*,1}]$, $\mathbb{E}[\bar{S}_n^2] \to (\mathbb{E}[X_{*,1}])^2$, and*

$$\bar{S}_n \xrightarrow{P} \mathbb{E}[X_{*,1}] \, .$$

*Proof of Lemma 31.* By exchangeability $\mathbb{E}[\bar{S}_n] = \mathbb{E}[X_{n,1}]$, and thus by Lemma 28 and Theorem 29, $\mathbb{E}[\bar{S}_n] \to \mathbb{E}[X_{\star,1}]$. Similarly,

$$\mathbb{E}[\bar{S}_n^2] = \frac{1}{d_n}\,\mathbb{E}[X_{n,1}^2] + \frac{d_n(d_n-1)}{d_n^2}\,\mathbb{E}[X_{n,1}X_{n,2}] \, ,$$

and thus by the continuous mapping theorem and again by Lemma 28 and Theorem 29, $\mathbb{E}[\bar{S}_n^2] \to (\mathbb{E}[X_{\star,1}])^2$ as $n \to \infty$. Finally, the convergence in probability follows by Chebyshev's inequality

$$\mathbb{P}\left\{\left|\bar{S}_n - \mathbb{E}\,\bar{S}_n\right| \geq \delta\right\} \leq \frac{\mathbb{E}[\bar{S}_n^2] - (\mathbb{E}[\bar{S}_n])^2}{\delta^2} \, . \qquad \square$$

**Lemma 32** (Moment propagation)**.** *Under the assumptions of Theorem 3, for any $x, x' \in \mathcal{X}$, $\ell \in [L+1]$, and $t \geq 1$*

$$\sup_{\substack{c\in[d^s]\\ i\in\mathbb{N}}} \sup_n \mathbb{E}\,|g_{n,ci}^{\ell-1}(x)|^t < \infty \, ,$$

$$\sup_{\substack{c\in[d^s]\\ i\in\mathbb{N}}} \sup_n \mathbb{E}\,|f_{n,ci}^{\ell}(x)|^t < \infty \, ,$$

$$\sup_{\substack{c\in[d^s]\\ h,i\in\mathbb{N}}} \sup_n \mathbb{E}\,|f_{n,ci}^{\ell h}(x)|^t < \infty \, ,$$

$$\sup_{\substack{c,c'\in[d^s]\\ h\in\mathbb{N}}} \sup_n \mathbb{E}\,|\widetilde{G}_{n,cc'}^{\ell h}(x)|^t < \infty \, ,$$

$$\sup_{\substack{a,b\in[d^s]\\ i,j\in\mathbb{N}}} \sup_n \mathbb{E}\,|\widehat{\Theta}_{ai,\,bj}^{\ell}(x, x')|^t < \infty \, .$$

*Proof of Lemma 32.* Beginning with $\mathbb{E}\,|g_{n,ci}^{\ell-1}(x)|^t$ and $\mathbb{E}\,|f_{n,ci}^{\ell}(x)|^t$, we see that this condition holds if none of the layers $1, \dots, \ell-1$ uses attention by the assumed polynomial boundedness of $\phi$ as a corollary of (Matthews et al., 2018, lemma 20) for dense, and (Garriga-Alonso et al., 2019, pages 14 and 15) for convolutional networks.[14] To extend to the case where one or more of the preceding layers include attention, we see that it is sufficient to focus on bound for $f^\ell$ in the first attention layer (i.e., with the lowest $\ell$ among the attention layer), as the bound for the following $g^\ell$ can be obtained from the assumed polynomial boundedness of $\phi$ and exchangeability, and the bound on the following layers by a simple inductive argument.

Thus, focusing on $\mathbb{E}\,|f_{n,ci}^{\ell}(x)|^t = \mathbb{E}\,|f_{n,c1}^{\ell}(x)|^t$ (exchangeability), we see that proving the bound for an arbitrary fixed $c \in [d^s]$ will be sufficient as $d^s$ is finite. Substituting

$$\mathbb{E}\,|f_{n,c1}^{\ell}(x)|^t = \mathbb{E}\left\{\mathbb{E}\left[\left|\sum_{h=1}^{d_n^{\ell,H}}\sum_{k=1}^{d_n^{\ell,V}} f_{n,ck}^{\ell h}(x) W_{n,k1}^{\ell h,O}\right|^t \;\middle|\; f_{n,c\cdot}^{\ell 1}(x)\,, \dots, f_{n,c\cdot}^{\ell d_n^{\ell,H}}(x)\right]\right\} \lesssim \mathbb{E}\left|\frac{\sigma_O^2}{d_n^{\ell,H} d_n^{\ell,V}} \sum_{h,k} f_{n,ck}^{\ell h}(x)^2\right|^{\frac{t}{2}} ,$$

[14]Note that the bound on $\mathbb{E}\,|g_{n,ci}^{0}(x)|^t = |x_{ci}|^t$ is trivial, which then leads to a bound on $\mathbb{E}\,|f_{n,ci}^{\ell}(x)|^t$ as the individual columns will be i.i.d. Gaussian for any $n$.

where we have used that if $\varepsilon \sim \mathcal{N}(0, I)$ and $v \in \mathbb{R}^d$ is a fixed vector, then $\langle v, \varepsilon \rangle$ is in distribution equal to $\|v\|_2 \varepsilon'$ where $\varepsilon' \sim \mathcal{N}(0, 1)$ by standard Gaussian identities, and the fact that $\mathbb{E}\, |\varepsilon'|^t < \infty$. Using Hölder's inequality if necessary, we can assume $t$ is even, and with that multiply out the r.h.s. above, leading to

$$\mathbb{E} \left| \frac{1}{d_n^{\ell,H} d_n^{\ell,V}} \sum_{h,k} f_{n,ck}^{\ell h}(x)^2 \right|^{\frac{t}{2}} \lesssim \mathbb{E}\, |f_{n,c1}^{\ell 1}(x)|^t \,,$$

by exchangeability. It will thus be sufficient to establish $\sup_{c \in [d^s], h, i \in \mathbb{N}} \sup_n \mathbb{E}\, |f_{n,ci}^{\ell h}(x)|^t < \infty$ for any $t \geq 1$. Observe

$$\mathbb{E}\, |f_{n,ci}^{\ell h}(x)|^t = \mathbb{E} \left\{ \mathbb{E} \left[ \left| \sum_{j=1}^{d_n^{\ell-1}} \widetilde{G}_{n,c\cdot}^{\ell h}(x) g_{n,\cdot j}^{\ell-1}(x) W_{n,\cdot i}^{\ell h,V} \right|^t \,\middle|\, \widetilde{G}_n^{\ell h}(x) \,, g_n^{\ell-1}(x) \right] \right\} \lesssim \mathbb{E} \left| \frac{\sigma_V^2}{d_n^{\ell-1}} \sum_j \left( \widetilde{G}_{n,c\cdot}^{\ell h}(x) g_{n,\cdot j}^{\ell-1}(x) \right)^2 \right|^{\frac{t}{2}} ,$$

meaning we can combine an argument analogous to the one above with Hölder's inequality and exchangeability to obtain

$$\mathbb{E} \left| \frac{1}{d_n^{\ell-1}} \sum_j \left( \widetilde{G}_{n,c\cdot}^{\ell h}(x) g_{n,\cdot j}^{\ell-1}(x) \right)^2 \right|^{\frac{t}{2}} \lesssim \operatorname{poly}\left( \max_{c' \in [d^s]} \mathbb{E}\, |\widetilde{G}_{n,cc'}^{\ell h}(x)|^{2t} \,, \max_{c' \in [d^s]} \sup_{j \in \mathbb{N}} \mathbb{E}\, |g_{n,c' j}^{\ell-1}(x)|^{2t} \right) .$$

As shown at the beginning of this proof, we can bound $\mathbb{E}\, |g_{n,c'j}^{\ell-1}(x)|^{4t}$ by a constant independent of $c', j$ and $n$, and thus it only remains to show that $\max_{c,c' \in [d^s]} \sup_{h \in \mathbb{N}} \sup_n \mathbb{E}\, |\widetilde{G}_{n,cc'}^{\ell h}(x)|^t < \infty$ in order to bound the $\mathbb{E}\, |f_{n,c1}^{\ell}(x)|^t$. Using the assumed entrywise polynomial boundedness of $\zeta$, we can see it will be sufficient to establish $\max_{c,c' \in [d^s]} \sup_{h \in \mathbb{N}} \sup_n \mathbb{E}\, |G_{n,cc'}^{\ell h}(x)|^t < \infty$. We do this separately for $\tau = 1$ and $\tau = \frac{1}{2}$.

Starting with the former, we can again replicate the argument from above, yielding

$$\mathbb{E}\, |G_{n,cc'}^{\ell h}(x)|^t = \mathbb{E} \left| \frac{1}{d_n^{\ell,G}} \sum_{k=1}^{d_n^{\ell,G}} Q_{n,ck}^{\ell h}(x) K_{n,c'k}^{\ell h}(x) \right|^t \lesssim \mathbb{E}\, |Q_{n,c1}^{\ell 1}(x)|^{2t} \lesssim \mathbb{E}\, |g_{n,c1}^{\ell-1}(x)|^{4t} \,,$$

by exchangeability, Hölder's inequality, and the assumed $W_n^{\ell h,Q} = W_n^{\ell h,K}$ a.s. under $\tau = 1$ (Section 3.2). For the $\tau = \frac{1}{2}$ case, we start by w.l.o.g. assuming we need bound for $t \in \mathbb{N}$ even

$$\mathbb{E}\, |G_{n,cc'}^{\ell h}(x)|^t = \mathbb{E} \left| \frac{1}{\sqrt{d_n^{\ell,G}}} \sum_{k=1}^{d_n^{\ell,G}} Q_{n,ck}^{\ell h}(x) K_{n,c'k}^{\ell h}(x) \right|^t = \left( \frac{1}{\sqrt{d_n^{\ell,G}}} \right)^t \sum_{k_1,\ldots,k_t} \mathbb{E} \left[ \prod_{s=1}^t Q_{n,ck_s}^{\ell h}(x) K_{n,c'k_s}^{\ell h}(x) \right] ,$$

and noting that because $W_n^{\ell h,Q}$ and $W_n^{\ell h,K}$ are assumed independent under $\tau = \frac{1}{2}$, there will be at most $\mathcal{O}((G)^{t/2})$ terms with non-zero expectation, meaning that we can again apply exchangeability and the distributional equivalence between $Q_n^{\ell h}(x)$ and $K_n^{\ell h}(x)$ to obtain

$$\mathbb{E}\, |G_{n,cc'}^{\ell h}(x)|^t \lesssim \mathbb{E}\, |Q_{n,c1}^{\ell 1}(x)|^{2t} \lesssim \mathbb{E}\, |g_{n,c1}^{\ell-1}(x)|^{4t} \,.$$

The convergence of $\max_{a,b \in [d^s]} \sup_{i,j \in \mathbb{N}} \sup_n \mathbb{E}\, |\widehat{\Theta}_{ai,\,bj}^{\ell}(x, x')|^t < \infty$ for non-attention layers can be obtained by combining exchangeability between the two groups of $\widehat{\Theta}_{ai,\,bj}^{\ell}$ variables with $i \neq j$ (resp. $i = j$) indices, and the results in (Yang, 2019a) which show that expectations of polynomially bounded functions converge (this is essentially due to the assumed polynomial boundedness of $\phi$ and $\zeta$ and their (weak) derivatives, the fact that the pre-nonlinearities in the first layer are Gaussian for all of the considered architectures by linearity of Gaussian variables, and the standard combination of Lemma 28 and Theorem 29—see the proofs of theorems 4.3 and 5.1 in (Yang, 2019a)). This can then be used to prove the bound for the first attention layer by inspecting the proofs in Appendix B.2 and noting that $\sup_n \mathbb{E}\, |\widehat{\Theta}_{ai,\,bj}^{\ell}(x, x')|^t$ can be always bounded by a polynomial in suprema over quantities from previous layers that we already know are uniformly bounded. The proof for subsequent attention layers can thus proceed by induction. □

**Lemma 33** (Convergence of inner products)**.** *Under the assumptions of Theorem 3, the following holds for any* $a, b \in [d^s]$, $x, x' \in \mathcal{X}$, $\ell \in [L + 1]$, *and* $h \in [d_n^{\ell,H}]$

$$\mathbb{E} \left[ \frac{\langle g_{n,a\cdot}^{\ell-1}(x), g_{n,b\cdot}^{\ell-1}(x') \rangle}{d_n^{\ell-1}} \right] \overset{n \to \infty}{\to} \tilde{\kappa}_{ab}^{\ell}(x, x') \,, \qquad \frac{\langle g_{n,a\cdot}^{\ell-1}(x), g_{n,b\cdot}^{\ell-1}(x') \rangle}{d_n^{\ell-1}} \xrightarrow{P} \tilde{\kappa}_{ab}^{\ell}(x, x') \,, \tag{37}$$

$$\mathbb{E}\left[\frac{\langle Q^{\ell h}_{n,a\cdot}(x), Q^{\ell h}_{n,b\cdot}(x')\rangle}{d^{\ell,G}_n}\right] \overset{n\to\infty}{\to} \sigma_Q^2 \tilde{\kappa}^{\ell}_{ab}(x,x')\,, \qquad \frac{\langle Q^{\ell h}_{n,a\cdot}(x), Q^{\ell h}_{n,b\cdot}(x')\rangle}{d^{\ell,G}_n} \xrightarrow{P} \sigma_Q^2 \tilde{\kappa}^{\ell}_{ab}(x,x')\,, \tag{38}$$

$$\mathbb{E}\left[\frac{\langle K^{\ell h}_{n,a\cdot}(x), K^{\ell h}_{n,b\cdot}(x')\rangle}{d^{\ell,G}_n}\right] \overset{n\to\infty}{\to} \sigma_K^2 \tilde{\kappa}^{\ell}_{ab}(x,x')\,, \qquad \frac{\langle K^{\ell h}_{n,a\cdot}(x), K^{\ell h}_{n,b\cdot}(x')\rangle}{d^{\ell,G}_n} \xrightarrow{P} \sigma_K^2 \tilde{\kappa}^{\ell}_{ab}(x,x')\,, \tag{39}$$

$$\mathbb{E}\left[\frac{\langle Q^{\ell h}_{n,a\cdot}(x), K^{\ell h}_{n,b\cdot}(x')\rangle}{d^{\ell,G}_n}\right] \overset{n\to\infty}{\to} \delta_{\tau=1}\sigma_{QK} \tilde{\kappa}^{\ell}_{ab}(x,x')\,, \qquad \frac{\langle Q^{\ell h}_{n,a\cdot}(x), K^{\ell h}_{n,b\cdot}(x')\rangle}{d^{\ell,G}_n} \xrightarrow{P} \delta_{\tau=1}\sigma_{QK} \tilde{\kappa}^{\ell}_{ab}(x,x')\,. \tag{40}$$

*Proof of Lemma 33.* Notice that all the statements involving $Q^{\ell h}$ or $K^{\ell h}$ are of the form

$$\frac{g^{\ell-1}_{n,a\cdot}(x) W^{\ell h}_n (W^{\ell h}_n)^\top g^{\ell-1}_{n,b\cdot}(x')^\top}{d^{\ell,G}_n}\,,$$

i.e., with the same weight matrix multiplying the layer inputs $g^{\ell-1}_n(x)$ (recall that under $\tau = 1$, we assumed $W^{\ell h,Q}_n = W^{\ell h,K}_n$ a.s.). Taking the expectation of the above term we obtain a term proportional up to $\sigma_Q$ or $\sigma_K$ to

$$\mathbb{E}\left[\frac{\langle g^{\ell-1}_{n,a\cdot}(x), g^{\ell-1}_{n,b\cdot}(x')\rangle}{d^{\ell-1}_n}\right] = \mathbb{E}\left[g^{\ell-1}_{n,a1}(x), g^{\ell-1}_{n,b1}(x')\right]\,,$$

by the assumed exchangeability of $g^{\ell-1}_n$. Since the integrand converges in distribution by the assumed continuity of $\phi$ and the continuous mapping theorem (Dudley, 2002, theorem 9.3.7), we can combine Lemma 32 with Lemma 28 to obtain uniform integrability and thus $\mathbb{E}[g^{\ell-1}_{n,a1}(x), g^{\ell-1}_{n,b1}(x')] \to \tilde{\kappa}^{\ell}_{ab}(x,x')$ by Theorem 29, proving the convergence of expectations.

The obtain the convergence in probability, it is sufficient to show that

$$\mathbb{E}\left|\frac{g^{\ell-1}_{n,a\cdot}(x) W^{\ell h}_n (W^{\ell h}_n)^\top g^{\ell-1}_{n,b\cdot}(x')^\top}{d^{\ell,G}_n}\right|^2 = \sum_{\substack{i_1,j_1\\ i_2 j_2}}^{d^{\ell-1}_n} \mathbb{E}\left[g^{\ell-1}_{n,ai_1}(x) g^{\ell-1}_{n,bj_1}(x) g^{\ell-1}_{n,ai_2}(x) g^{\ell-1}_{n,bj_2}(x) W^{\ell h}_{n,i_1\cdot}(W^{\ell h}_{n,j_1\cdot})^\top W^{\ell h}_{n,i_2\cdot}(W^{\ell h}_{n,j_2\cdot})^\top\right]\,,$$

converges to the square of the mean as by Chebyshev's inequality: $\mathbb{P}(|X_n - \mathbb{E}\,X| \geq \delta) \leq \delta^{-2}(\mathbb{E}[X_n^2] - \{\mathbb{E}[X_n]\}^2)$. Since we can bound each of the summands using Hölder's inequality combined with Lemma 32, the limit of the above expectation will up to a constant coincide with that of

$$\frac{1}{(d^{\ell-1}_n)^2}\sum_{i,j}^{d^{\ell-1}_n} \mathbb{E}\left[g^{\ell-1}_{n,ai}(x) g^{\ell-1}_{n,bi}(x) g^{\ell-1}_{n,aj}(x) g^{\ell-1}_{n,bj}(x)\right] = \mathbb{E}\left[g^{\ell-1}_{n,a1}(x) g^{\ell-1}_{n,b1}(x) g^{\ell-1}_{n,a2}(x) g^{\ell-1}_{n,b2}(x)\right] + \mathrm{o}((d^{\ell-1}_n)^2)\,,$$

where the equality is by the assumed exchangeability. Since the individual columns of $g^{\ell-1}_n$ are asymptotically independent by assumption, we can use an argument analogous to that we made for the $\mathbb{E}[g^{\ell-1}_{n,a1}(x) g^{\ell-1}_{n,b1}(x)]$ above to obtain the $(\tilde{\kappa}^{\ell}_{ab}(x,x'))^2$ limit. Noting that the l.h.s. above is equal to $\mathbb{E}\big[\big(\langle g^{\ell-1}_{n,a\cdot}(x), g^{\ell-1}_{n,b\cdot}(x')\rangle / d^{\ell-1}_n\big)^2\big]$ concludes the proof. □

## C. Positional encodings

As in the proofs for attention without positional encodings, we assume the 'infinite width, finite fan-out' construction of the sequence of NNs. In particular, we will assume that for any $n \in \mathbb{N}$, there is a countably infinite set of random variables $\{E^{\ell}_{n,\cdot i} : i \in \mathbb{N}\} = \{E^{\ell}_{\cdot i} : i \in \mathbb{N}\}$, where $E^{\ell}_{\cdot i} \sim \mathcal{N}(0, R)$ i.i.d. over the $i$ index, but only a finite number $d^{\ell,E}_n \in \mathbb{N}$ is `add`-ed,

$$\tilde{g}^{\ell-1}_n(x) = \sqrt{\alpha} g^{\ell-1}_n(x) + \sqrt{1-\alpha} E^{\ell}_n\,,$$

or `append`-ed

$$\tilde{g}^{\ell-1}_n(x) = [g^{\ell-1}_n(x), E^{\ell}_n]\,,$$

to each of the layer inputs $g^{\ell-1}_n(x)$. In the `append` case, we further assume $\alpha = \lim_{n\to\infty} d^{\ell-1}_n / (d^{\ell,E}_n + d^{\ell-1}_n)$.

### C.1. NNGP limit

Note that all Theorem 3 relies on is that the layer's inputs $\{g_n^{\ell-1}(x) \colon x \in \mathcal{X}\}$ converge in distribution to some $g^{\ell-1}(x)$ with mutually independent columns, and on the fact that the elementwise absolute moments of $g_{n,ai}^{\ell-1}(x)$ are bounded uniformly in $a, i$ and $n$. Let us thus replace $g_n^{\ell-1}(x)$ by $\tilde{g}_n^{\ell-1}(x)$, and see whether the proofs still apply.

**Exchangeability:** The proofs of exchangeability in Lemmas 5, 8 and 13 are all based on conditioning on $g_n^{\ell-1}(x)$ for some fixed finite subset of the inputs $x$, and then showing that the random variables are conditionally i.i.d. for any given $n \in \mathbb{N}$. If positional encodings are used, the variables will be again i.i.d. if we add $E_n^{\ell}$ into the conditioning set.

**Convergence in distribution:** To establish $\{\tilde{g}_n^{\ell-1}(x) \colon x \in \mathcal{X}\}$ converges in distribution in the `add` case, we can use a simple argument based on the Cramér-Wold device and pointwise convergence of the characteristic function, which implies convergence in distribution by Lévy's continuity theorem (using the fact that $E_{n,\cdot i}^{\ell}$ are assumed to be i.i.d. $\mathcal{N}(0, R)$ and the distribution of a particular $E_{n,\cdot i}^{\ell}$ does not change with $n$). An alternative approach has to be taken for the `append` case where the weak limit of the layer input's distribution may not be well defined; closer inspection of Appendix B.1 reveals that all the proofs depend on the convergence of the layer inputs only through Lemmas 32 and 33, which we discuss next.

**Convergence of inner products and boundedness of moments:** The proof of each statement of Lemma 32 relies on $\{g_n^{\ell-1}(x) \colon x \in \mathcal{X}\}$ only through the bound $\max_{c \in [d^s]} \sup_{i \in \mathbb{N}} \sup_n \mathbb{E}\, |g_{n,ci}^{\ell-1}(x)|^t < \infty$ which is essentially established using the assumed polynomial bound on $|\phi|$ and the Gaussianity of the weights at initialisation. All we need to extend Lemma 32 to the case where positional encodings are used is to establish $\max_{c \in [d^s]} \sup_{i \in \mathbb{N}} \sup_n \mathbb{E}\, |\tilde{g}_{n,ci}^{\ell-1}(x)|^t < \infty$. This can be done by observing $\mathbb{E}\, |\tilde{g}_{n,ci}^{\ell-1}(x)|^t \leq \max\{\mathbb{E}\, |g_{n,ci}^{\ell-1}(x)|^t, \mathbb{E}\, |E_{n,c1}^{\ell}|^t\} < \infty$ by the assumption $E_{n,\cdot i}^{\ell} \sim S(0, R)$ i.i.d. over the $i$ index for any $n \in \mathbb{N}$.

Similarly, the proof of Lemma 33 can be modified by observing that

$$\mathbb{E}\left[\frac{\langle \tilde{g}_{n,a\cdot}^{\ell-1}(x), \tilde{g}_{n,b\cdot}^{\ell-1}(x')\rangle}{d_n^{\ell,E} + d_n^{\ell-1}}\right] = \frac{d_n^{\ell-1}}{d_n^{\ell,E} + d_n^{\ell-1}}\, \mathbb{E}\left[g_{n,a1}^{\ell-1}(x), g_{n,b1}^{\ell-1}(x')\right] + \frac{d_n^{\ell,E}}{d_n^{\ell,E} + d_n^{\ell-1}} \underbrace{\mathbb{E}\left[E_{n,a1}^{\ell} E_{n,b1}^{\ell}\right]}_{=R_{ab}},$$

in the `append` case by the independence of $E_n^{\ell}$ and exchangeability. Using the Gaussianity of positional encodings and $\alpha = \lim_{n \to \infty} d_n^{\ell-1}/(d_n^{\ell,E} + d_n^{\ell-1})$, an analogous argument to that made in Lemma 33 can be used to establish convergence of the r.h.s. to $\mathcal{I} \circ \tilde{\kappa}_{ab}^{\ell}(x, x') = \alpha \tilde{\kappa}_{ab}^{\ell}(x, x') + (1 - \alpha) R_{ab}$ in both probability and expectation. For the `add` case,

$$\mathbb{E}\left[\frac{\langle \tilde{g}_{n,a\cdot}^{\ell-1}(x), \tilde{g}_{n,b\cdot}^{\ell-1}(x')\rangle}{d_n^{\ell-1}}\right] = \alpha\, \mathbb{E}\left[g_{n,a1}^{\ell-1}(x), g_{n,b1}^{\ell-1}(x')\right] + (1 - \alpha) \underbrace{\mathbb{E}\left[E_{n,a1}^{\ell} E_{n,b1}^{\ell}\right]}_{=R_{ab}},$$

again by the independence of $E_n^{\ell}$ and exchangeability, and thus a similar argument to the one above applies.

Putting all of the above together, addition of positional encodings does not prevent GP behaviour in the infinite width limit; the only modification of the results in Appendix B.1 is thus replacement of any $\tilde{\kappa}_{ab}^{\ell}(x, x')$ in the expression for the limiting covariance of $f^{\ell}$ and $G^{\ell}$ by $\mathcal{I} \circ \tilde{\kappa}_{ab}^{\ell}(x, x')$.

### C.2. NTK limit

There are two sets of changes to the NTK limit. First, the gradients w.r.t. $g_n^{\ell-1}(x)$ in the *indirect* part will now be multiplied by $\sqrt{\alpha}$ in the `add` case, and by $[d_n^{\ell-1}/(d_n^{\ell} + d_n^{\ell,E})]^{1/2}$—to ensure convergence of corresponding inner products—in the `append` case, and all the terms of the form

$$\mathbb{E}\left|\frac{\langle \tilde{g}_{n,a\cdot}^{\ell-1}(x), \tilde{g}_{n,b\cdot}^{\ell-1}(x')\rangle}{d_n^{\ell,E} + d_n^{\ell-1}}\right|^k,$$

for some $k > 0$ in the *direct* part will converge to the $k^{\text{th}}$ power of the $\mathcal{I} \circ \tilde{\kappa}_{ab}^{\ell}(x, x') = \alpha \tilde{\kappa}_{ab}^{\ell}(x, x') + (1 - \alpha) R_{ab}$ kernel as discussed. Since we have shown that Lemmas 32 and 33 hold mutatis mutandis in the previous section, the rest of the proofs in the direct part can be modified in the obvious way, replacing $\tilde{\kappa}_{ab}^{\ell}(x, x')$ by $\mathcal{I} \circ \tilde{\kappa}_{ab}^{\ell}(x, x')$ as necessary.

Second, there will be a new contribution to the *direct* part due to the gradient w.r.t. the trainable $E_n^\ell$. Since $E_n^\ell$ is `add`-ed (resp. `append`-ed) to the layer input $g_n^{\ell-1}(x)$, this contribution will be quite similar to the *indirect* contribution, however with $\widehat{\Theta}^{\ell}_{a'i',b'j'}(x,x')$ (Equation (27)) replaced by $(1-\alpha)\delta_{a'=b'}\delta_{i'=j'}$. Inspecting Lemmas 22, 23 and 25, this will lead to two changes. Firstly, since $\mathbb{E}\,|(1-\alpha)\delta_{a'=b'}\delta_{i'=j'}|^t < \infty$, all bounds involving $\mathbb{E}\,|\widehat{\Theta}_{a'i',b'j'}|^t$ can be trivially reduced. Secondly, as shown in the previous section, all appearances of the $\tilde{\kappa}^{\ell}_{ab}(x,x')$ are to be replaced $\mathcal{I}\circ\tilde{\kappa}^{\ell}_{ab}(x,x')$, including those involved indirectly through the modified asymptotic distribution of $G_n^\ell$. The rest of the proofs is affected by the introduction of positional encodings only through Lemmas 32 and 33 which, as mentioned, do hold in a modified form. Substituting $(1-\alpha)\delta_{a'=b'}\delta_{i'=j'}$ for $\widehat{\Theta}^{\ell}_{a'i',b'j'}(x,x')$ in Lemmas 22 and 23, we thus conclude that the new contribution to the NTK due to the gradient w.r.t. $E_n^\ell$ is

$$
\begin{aligned}
&(1-\alpha)\sigma_{OV}^2\sum_{c=1}^{d^s}\mathbb{E}[\widetilde{G}^{\ell 1}_{ac}(x)\widetilde{G}^{\ell 1}_{bc}(x')]+\\
&\delta_{\tau=\frac{1}{2}}(1-\alpha)\sigma_{OV}^2\sigma_{QK}^2\sum_{\substack{c_1,c_2\\d_1,d_2}}^{d^s}\mathcal{I}\circ\tilde{\kappa}^{\ell}_{c_1c_2}(x,x')\left(\delta_{d_1=d_2}\mathcal{I}\circ\tilde{\kappa}^{\ell}_{ab}(x,x')+\delta_{a=b}\mathcal{I}\circ\tilde{\kappa}^{\ell}_{d_1d_2}(x,x')\right)\mathbb{E}\left[\frac{\partial\widetilde{G}^{\ell 1}_{ac_1}(x)}{\partial G^{\ell 1}_{ad_1}(x)}\frac{\partial\widetilde{G}^{\ell 1}_{bc_2}(x')}{\partial G^{\ell 1}_{bd_2}(x')}\right].
\end{aligned}
$$

## D. Residual attention

Observe that by (Garriga-Alonso et al., 2019; Yang, 2019b), the covariance induced by the skip connection, $f_n^\ell(x) = \sqrt{\alpha}g_n^{\ell-1}(x)+\sqrt{1-\alpha}\tilde{f}_n^\ell(x)$, in the infinite width limit is equal to

$$
\begin{aligned}
\mathbb{E}[f^\ell_{a1}(x)f^\ell_{b1}(x')] &= \alpha\,\mathbb{E}[g^{\ell-1}_{a1}(x)g^{\ell-1}_{b1}(x')]+(1-\alpha)\,\mathbb{E}[\tilde{f}^\ell_{a1}(x)\tilde{f}^\ell_{a1}(x')]\\
&= \alpha\tilde{\kappa}^{\ell}_{ab}(x,x')+(1-\alpha)\,\mathbb{E}[\tilde{f}^\ell_{a1}(x)\tilde{f}^\ell_{a1}(x')]\,.
\end{aligned}
$$

To obtain the $\alpha\tilde{\kappa}^{\ell}_{ab}(x,x')+(1-\alpha)R_{a\cdot}\tilde{\kappa}^{\ell}(x,x')R_{b\cdot}^\top$ from Equation (14), it is thus sufficient to choose $\tilde{f}_n^\ell(x)$ to be the output of attention layer under the $d^{-1}$ scaling with structured positional encodings (covariance $R$), identity function for $\zeta$ and the interpolation parameter *for the attention layer* set to zero, resulting in the $R_{a\cdot}\tilde{\kappa}^{\ell}(x,x')R_{b\cdot}^\top$ (see Table 1).