

Appendices of the NeurIPS 2025 Paper: **“How Data Mixing Shapes In-Context Learning:** **Asymptotic Equivalence for Transformers with MLPs”**

Deferred proofs

In the following appendices, we provide deferred proofs. We first consider the case of zero-mean inputs, i.e., $\mu_{x,s} = \mathbf{0}$ for all s , and address the extension to non-zero mean inputs, i.e., $\mu_{x,s} \neq \mathbf{0}$, in the final section. This organization streamlines the initial presentation while preserving generality.

Appendix A introduces a reparameterization of the attention mechanism that simplifies the form of the attention outputs. Appendix B analyzes the distribution of $F\text{vec}(\mathbf{H}_Z)$, a key quantity in our analysis, and proves Lemma 4.9. Appendix C presents a decomposition of the gradient matrix corresponding to a single gradient step on the first layer (6), establishing Lemma 4.11. Appendix D builds on prior results to establish a conditional Gaussian equivalence. Appendix E contains the proof of our main theoretical result (Theorem 4.12). Finally, Appendix F considers an extension of the analysis to the setting with non-zero mean inputs.

Notation

We adopt standard notation following [18]. For a random vector \mathbf{v} , its covariance is denoted by $\text{Cov}(\mathbf{v})$. The spectral norm and Frobenius norm of a matrix \mathbf{A} are denoted by $\|\mathbf{A}\|$ and $\|\mathbf{A}\|_F$, respectively, while $\text{Tr}(\mathbf{A})$ denotes the trace. Matrix entries and slices are indicated by $A_{i,j}$, $\mathbf{A}_{:,j}$, and $\mathbf{A}_{i,:}$, where i and j denote the row and column indices, respectively. We write $f(\cdot) \asymp g(\cdot)$ to denote that f and g are of the same asymptotic order with respect to the diverging dimensions. We use standard asymptotic notation: $f(d) = \mathcal{O}(g(d))$ if $f(d)/g(d) \rightarrow C < \infty$ for some $C > 0$, and $f(d) = o(g(d))$ if $f(d)/g(d) \rightarrow 0$ as $d \rightarrow \infty$. Furthermore, we define $\tilde{\mathcal{O}}(f(\cdot))$ as shorthand for $\mathcal{O}(f(\cdot) \text{polylog } d)$ to suppress polylogarithmic factors for brevity. Element-wise multiplication is denoted by \odot , and the Kronecker product by \otimes . Conditional random variables are written as $X \mid C$, representing the distribution of X given condition C . Finally, we use $x \rightarrow \mathcal{N}(0, 1)$ to denote that the random variable x converges in distribution to the standard normal almost surely.

A Reparameterization of the linear attention

We begin by reparameterizing the linear attention mechanism in accordance with prior work [47, 2, 29, 23, 32, 27] to simplify subsequent analysis. Specifically, we consider the following form of linear attention:

$$\mathbf{A} := \mathbf{Z} + \frac{1}{\ell} \mathbf{V} \mathbf{Z} (\mathbf{K} \mathbf{Z})^T (\mathbf{Q} \mathbf{Z}), \quad (\text{S1})$$

where \mathbf{K} , \mathbf{Q} , and \mathbf{V} denote the key, query, and value matrices, respectively. Due to the structure of the embedding matrix introduced in (2), the output of linear attention, denoted \hat{y}_{linear} , corresponds to the $(d+1, \ell+1)$ -th entry of \mathbf{A} , i.e.,

$$\hat{y}_{\text{linear}} := A_{d+1, \ell+1} = \frac{1}{\ell} \mathbf{V}_{d+1, :} (\mathbf{Z} \mathbf{Z}^T) (\mathbf{K}^T \mathbf{Q}) \mathbf{Z}_{:, \ell+1}. \quad (\text{S2})$$

This formulation highlights that the parameters relevant to \hat{y}_{linear} are $\mathbf{V}_{d+1, :}$ and $\mathbf{K}^T \mathbf{Q}$. To isolate these, we express the matrices as follows:

$$\mathbf{V} = \begin{bmatrix} * & * \\ \mathbf{v}_{21} & v_{22} \end{bmatrix}, \quad \text{and} \quad \mathbf{M} := \mathbf{K}^T \mathbf{Q} = \begin{bmatrix} \mathbf{M}_{11} & * \\ \mathbf{m}_{21}^T & * \end{bmatrix}, \quad (\text{S3})$$

where $*$ denotes components not involved in predicting $y_{\ell+1}$, while the submatrices $\mathbf{M}_{11} \in \mathbb{R}^{d \times d}$, $\mathbf{v}_{21}, \mathbf{m}_{21} \in \mathbb{R}^d$, and $v_{22} \in \mathbb{R}$ capture the relevant contributions. Using these components, the prediction \hat{y}_{linear} can be rewritten in terms of the input features and previous outputs as:

$$\frac{1}{\ell} \left\langle \mathbf{x}_{\ell+1}, v_{22} \mathbf{M}_{11}^T \sum_{i \leq \ell} y_i \mathbf{x}_i + v_{22} \mathbf{m}_{21} \sum_{i \leq \ell} y_i^2 + \mathbf{M}_{11}^T \sum_{i \leq \ell+1} \mathbf{x}_i \mathbf{x}_i^T \mathbf{v}_{21} + \mathbf{m}_{21} \sum_{i \leq \ell} y_i \mathbf{x}_i^T \mathbf{v}_{21} \right\rangle, \quad (\text{S4})$$

where $\langle \cdot, \cdot \rangle$ is the standard inner product and we omit the source index s for brevity.

In most of the theoretical literature [47, 2, 29, 23, 32], the parameters \mathbf{v}_{21} and \mathbf{m}_{21} are commonly set to $\mathbf{0}$, since the leading term

$$\frac{1}{\ell} \left\langle \mathbf{x}_{\ell+1}, v_{22} \mathbf{M}_{11}^T \sum_{i \leq \ell} y_i \mathbf{x}_i \right\rangle$$

in (S4) alone is often sufficient to accurately predict $y_{\ell+1}$. More recently, [27] relaxed this assumption by allowing \mathbf{m}_{21} to be trainable while keeping $\mathbf{v}_{21} = \mathbf{0}$, as this modification retains analytical tractability. Following this approach, we adopt a setting in which $\mathbf{v}_{21} = \mathbf{0}$ and the remaining parameters are trainable.

Under this formulation, the output of the linear attention mechanism can be compactly expressed as:

$$\hat{y}_{\text{linear}} = \text{vec}(\mathbf{\Gamma})^T \text{vec}(\mathbf{H}_{\mathbf{Z}}), \quad (\text{S5})$$

where the trainable parameters are consolidated into the matrix $\mathbf{\Gamma}$, and the processed version of the embedding matrix \mathbf{Z} is denoted by $\mathbf{H}_{\mathbf{Z}}$. These are defined as:

$$\mathbf{\Gamma} := v_{22} \begin{bmatrix} \mathbf{M}_{11}^T & \mathbf{m}_{21} \end{bmatrix}, \quad \text{and} \quad \mathbf{H}_{\mathbf{Z}} := \mathbf{x}_{\ell+1} \begin{bmatrix} \frac{1}{\ell} \sum_{i \leq \ell} y_i \mathbf{x}_i^T & \frac{1}{\ell} \sum_{i \leq \ell} y_i^2 \end{bmatrix}. \quad (\text{S6})$$

The advantage of this reformulation lies in the fact that the linear attention output is now a linear function of both the trainable parameters $\mathbf{\Gamma}$ and the data-derived matrix $\mathbf{H}_{\mathbf{Z}}$. As a result, the matrix $\mathbf{\Gamma}$ can be efficiently optimized via ridge regression.

Similarly, in the case of a Transformer architecture equipped with a nonlinear MLP block, the above formulation enables us to express the prediction as:

$$\hat{y}_{\text{nonlinear}} := \frac{1}{\sqrt{k}} \mathbf{w}^T \sigma(\mathbf{F} \text{vec}(\mathbf{H}_{\mathbf{Z}})), \quad (\text{S7})$$

where $\mathbf{F} \in \mathbb{R}^{k \times d(d+1)}$ serves a role analogous to that of $\text{vec}(\mathbf{\Gamma})$, with each row of \mathbf{F} representing the weights associated with one of the k nonlinear neurons. The function σ denotes a nonlinear activation function, allowing the model to capture complex input-output relationships, and $\mathbf{w} \in \mathbb{R}^k$ is a trainable vector that linearly combines the outputs of these nonlinear units. Overall, this formulation captures the structure of a Transformer with linear attention followed by a nonlinear (two-layer) MLP block.

B Asymptotic distribution of $\mathbf{F} \text{vec}(\mathbf{H}_{\mathbf{Z}})$

In this section, we analyze the asymptotic distribution of the term $\mathbf{F} \text{vec}(\mathbf{H}_{\mathbf{Z}})$, which plays a central role in the Transformer with a nonlinear MLP layer. After training the first-layer parameters as described in (6), the prediction of the model (5) takes the form

$$\frac{1}{\sqrt{k}} \mathbf{w}^T \sigma(\hat{\mathbf{F}} \text{vec}(\mathbf{H}_{\mathbf{Z}})) = \frac{1}{\sqrt{k}} \mathbf{w}^T \sigma(\mathbf{F} \text{vec}(\mathbf{H}_{\mathbf{Z}}) + \eta \mathbf{G} \text{vec}(\mathbf{H}_{\mathbf{Z}})),$$

where \mathbf{G} denotes the gradient matrix defined in (7). Since $\mathbf{F} \text{vec}(\mathbf{H}_{\mathbf{Z}})$ is the primary input to the nonlinear activation function σ , understanding its distribution is a natural and informative starting point.

Let \mathbf{f}_i denote the i -th row of the matrix \mathbf{F} , and define $t := \text{Tr}(\text{Cov}(\text{vec}(\mathbf{H}_{\mathbf{Z}})))$. For the purpose of this analysis, we initially assume that $\mathbf{H}_{\mathbf{Z}}$ is fixed. Under Assumption 4.6, each row \mathbf{f}_i is drawn independently from a Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I}/t)$. Consequently, we have the projection

$$\mathbf{f}_i^T \text{vec}(\mathbf{H}_{\mathbf{Z}}) \sim \mathcal{N}(0, \|\text{vec}(\mathbf{H}_{\mathbf{Z}})\|_2^2/t).$$

Thus, the task of proving Lemma 4.9 reduces to showing that $\|\text{vec}(\mathbf{H}_{\mathbf{Z}})\|_2^2$ concentrates around 1, which would imply that $\mathbf{f}_i^T \text{vec}(\mathbf{H}_{\mathbf{Z}}) \rightarrow \mathcal{N}(0, 1)$ almost surely.

To establish this concentration result, we begin with the following lemma, which characterizes $\mathbf{H}_{\mathbf{Z}}$ as an outer product of a Gaussian vector and a sub-exponential vector, conditional on the data source index s and the task vector $\boldsymbol{\xi}|s$. The lemma also provides concentration bounds for the norms of these constituent vectors. For brevity, we omit explicit conditioning in some expressions where the context makes it clear.

Lemma B.1. Let $\mathbf{b} := \left[\frac{1}{\ell} \sum_{i \leq \ell} y_i \mathbf{x}_i \right]$, which implies that $\mathbf{H}_Z = \mathbf{x}_{\ell+1} \mathbf{b}^T$. Suppose that the data source index s and the task vector $\boldsymbol{\xi}|s$ are given. Then, conditioned on s , the vector $\mathbf{x}_{\ell+1}|s$ is Gaussian by definition, and the conditional vector $\mathbf{b}(\boldsymbol{\xi}, s)$ exhibits sub-exponential tails. Furthermore, conditioned on $(s, \boldsymbol{\xi}|s)$, the following concentration bounds hold with high probability:

$$\left| \|\mathbf{x}_{\ell+1}\|^2 - \text{Tr}(\text{Cov}(\mathbf{x}_{\ell+1})) \right| / \text{Tr}(\text{Cov}(\mathbf{x}_{\ell+1})) = o(1), \quad (\text{S8})$$

$$\left| \|\mathbf{b}\|^2 - \text{Tr}(\text{Cov}(\mathbf{b})) \right| / \text{Tr}(\text{Cov}(\mathbf{b})) = o(1). \quad (\text{S9})$$

Proof. Given s and $\boldsymbol{\xi}|s$, we have $\mathbf{x}_i|s \sim \mathcal{N}(\boldsymbol{\mu}_{x,s}, \boldsymbol{\Sigma}_{x,s})$, and $y_i := \phi_s \left(\frac{(\boldsymbol{\xi}|s)^T (\mathbf{x}_i|s)}{\|\boldsymbol{\xi}|s\|_2 \|\boldsymbol{\Sigma}_{x,s}\|_2^{1/2}} \right) + \epsilon_i|s$, where ϕ_s is Lipschitz by Assumption 4.7 and $\epsilon_i|s$ is Gaussian by definition.

We can express \mathbf{b} as: $\mathbf{b} = \frac{1}{\ell} \sum_{i \leq \ell} \mathbf{b}_i$, where $\mathbf{b}_i := \begin{bmatrix} y_i \mathbf{x}_i \\ y_i^2 \end{bmatrix}$. Since ϕ_s is Lipschitz and \mathbf{x}_i is sub-Gaussian, the composition y_i is sub-Gaussian with a constant norm due to Gaussian concentration of Lipschitz functions [40, Theorem 5.2.2].

By Assumption 4.2, we have $\|\boldsymbol{\Sigma}_{x,s}\|_2^2 = \mathcal{O}(d)$ and $\text{Tr}(\boldsymbol{\Sigma}_{x,s}) \asymp d$, so \mathbf{x}_i is sub-Gaussian with norm $\mathcal{O}(d^{1/4})$. Then, since the product of sub-Gaussian random variables is sub-exponential [40, Lemma 2.7.7], the term $y_i \mathbf{x}_i$ is sub-exponential with norm $\mathcal{O}(d^{1/4})$, and y_i^2 is sub-exponential with constant norm. Thus, \mathbf{b}_i is a sub-exponential vector with norm $\mathcal{O}(d^{1/4})$.

Applying the (vector version of) Bernstein's inequality [40, Corollary 2.8.3] to the vector \mathbf{b} (an average over ℓ samples of a sub-exponential) implies that the deviation $(\mathbf{b} - \mathbb{E}[\mathbf{b}])$ has sub-exponential tails with norm $\mathcal{O}(d^{1/4}/\ell)$. Under Assumption 4.1, $\ell/d \in \mathbb{R}$, and hence the norm becomes $\mathcal{O}(d^{-3/4})$.

Finally, applying the Hanson-Wright inequality [40, Theorem 6.2.1] yields the bound in (S8). For (S9), we rely on an extension of the Hanson-Wright inequality to sub-exponential vectors [19]. \square

Now that we have bounds on the norms of the vectors used to construct \mathbf{H}_Z , we can analyze the concentration of the norm of $\text{vec}(\mathbf{H}_Z)$ in the following corollary, which builds on Lemma B.1.

Corollary B.2. With high probability, the following concentration result holds:

$$\frac{\left| \|\text{vec}(\mathbf{H}_Z) \mid (s, \boldsymbol{\xi}|s)\|^2 - \text{Tr}(\text{Cov}(\text{vec}(\mathbf{H}_Z) \mid (s, \boldsymbol{\xi}|s))) \right|}{\text{Tr}(\text{Cov}(\text{vec}(\mathbf{H}_Z) \mid (s, \boldsymbol{\xi}|s)))} = \left| \frac{\|\text{vec}(\mathbf{H}_Z) \mid (s, \boldsymbol{\xi}|s)\|^2}{t} - 1 \right| = o(1),$$

where $t := \text{Tr}(\text{Cov}(\text{vec}(\mathbf{H}_Z)))$.

Proof. We first note that:

$$\text{vec}(\mathbf{H}_Z) = \text{vec}(\mathbf{x}_{\ell+1} \mathbf{b}^T) = \mathbf{b} \otimes \mathbf{x}_{\ell+1}.$$

Using this, we compute the covariance:

$$\text{Cov}(\text{vec}(\mathbf{H}_Z)) = \mathbb{E}[(\mathbf{b} \otimes \mathbf{x}_{\ell+1})(\mathbf{b} \otimes \mathbf{x}_{\ell+1})^T] \quad (\text{S10})$$

$$= \mathbb{E}[(\mathbf{b} \otimes \mathbf{x}_{\ell+1})(\mathbf{b}^T \otimes \mathbf{x}_{\ell+1}^T)] \quad (\text{S11})$$

$$= \mathbb{E}[(\mathbf{b} \mathbf{b}^T) \otimes (\mathbf{x}_{\ell+1} \mathbf{x}_{\ell+1}^T)] \quad (\text{S12})$$

$$= \text{Cov}(\mathbf{b}) \otimes \text{Cov}(\mathbf{x}_{\ell+1}), \quad (\text{S13})$$

where we used properties of the Kronecker product (transpose and mixed-product), the independence of \mathbf{b} and $\mathbf{x}_{\ell+1}$, and the linearity of expectation.

Similarly, we have:

$$\|\text{vec}(\mathbf{H}_Z)\|^2 = \|\mathbf{b} \otimes \mathbf{x}_{\ell+1}\|^2 = \|\mathbf{b}\|^2 \cdot \|\mathbf{x}_{\ell+1}\|^2, \quad (\text{S14})$$

and

$$\text{Tr}(\text{Cov}(\text{vec}(\mathbf{H}_Z))) = \text{Tr}(\text{Cov}(\mathbf{b})) \cdot \text{Tr}(\text{Cov}(\mathbf{x}_{\ell+1})). \quad (\text{S15})$$

Define the relative errors:

$$\delta_x := \frac{|\|\mathbf{x}_{\ell+1}\|^2 - \text{Tr}(\text{Cov}(\mathbf{x}_{\ell+1}))|}{\text{Tr}(\text{Cov}(\mathbf{x}_{\ell+1}))}, \quad \text{and} \quad \delta_v := \frac{|\|\mathbf{b}\|^2 - \text{Tr}(\text{Cov}(\mathbf{b}))|}{\text{Tr}(\text{Cov}(\mathbf{b}))}. \quad (\text{S16})$$

Then, conditioned on $(s, \boldsymbol{\xi}|s)$, we get:

$$\frac{|\|\text{vec}(\mathbf{H}_{\mathbf{Z}})\|^2 - \text{Tr}(\text{Cov}(\text{vec}(\mathbf{H}_{\mathbf{Z}})))|}{\text{Tr}(\text{Cov}(\text{vec}(\mathbf{H}_{\mathbf{Z}})))} = \frac{|\|\mathbf{x}_{\ell+1}\|^2 \|\mathbf{b}\|^2 - \text{Tr}(\text{Cov}(\mathbf{x}_{\ell+1}))\text{Tr}(\text{Cov}(\mathbf{b}))|}{\text{Tr}(\text{Cov}(\mathbf{x}_{\ell+1}))\text{Tr}(\text{Cov}(\mathbf{b}))} \quad (\text{S17})$$

$$\leq \delta_x + \delta_v + \delta_x \delta_v \quad (\text{S18})$$

$$= o(1), \quad (\text{S19})$$

where the final step follows from Lemma B.1, which showed that both δ_x and δ_v are $o(1)$. \square

So far, our analysis has focused on the case where $\mathbf{H}_{\mathbf{Z}}$ is conditioned on the data source s and the task vector $\boldsymbol{\xi}|s$. However, by Assumption 4.4, we have

$$t = \text{Tr}(\text{Cov}(\text{vec}(\mathbf{H}_{\mathbf{Z}}) \mid s = i)) = \text{Tr}(\text{Cov}(\text{vec}(\mathbf{H}_{\mathbf{Z}}) \mid s = j))$$

for any data source indices i, j . Furthermore, Corollary B.2 showed that $\|\text{vec}(\mathbf{H}_{\mathbf{Z}})\|^2/t$ concentrates around 1 for all s and $\boldsymbol{\xi}|s$. This leads us to the following unconditioned concentration result.

Corollary B.3. *Without conditioning on the data source or task vector, the following holds with high probability:*

$$\left| \frac{\|\text{vec}(\mathbf{H}_{\mathbf{Z}})\|^2}{t} - 1 \right| \leq \max_{s, \boldsymbol{\xi}|s} \left| \frac{\|\text{vec}(\mathbf{H}_{\mathbf{Z}}) \mid (s, \boldsymbol{\xi}|s)\|^2}{t} - 1 \right| = o(1), \quad (\text{S20})$$

as a consequence of Lemma B.1 and Corollary B.2.

Proof. We apply a union bound over the (countable) set of data sources and integrate over the distribution of $\boldsymbol{\xi}|s$. Since the conditional deviation is uniformly bounded by $o(1)$, the same bound extends to the unconditioned case. \square

Remark B.4. (Implications of task vector randomness) Corollary B.3 implies that the variability introduced by the task vector $\boldsymbol{\xi}|s$ does not significantly affect the norm of the attention output $\text{vec}(\mathbf{H}_{\mathbf{Z}})$, with deviations bounded by $o(1)$. This concentration result justifies a simplification in our analysis: we may treat $\boldsymbol{\xi}|s$ as effectively fixed and analyze the system under a worst-case $\boldsymbol{\xi}|s$, rather than integrating over its full distribution. This reduces analytical complexity while preserving rigorous control over variation across tasks. In effect, the randomness of task vectors can be absorbed into a uniform bound, allowing our results to hold uniformly over task diversity.

This completes our analysis of the concentration of $\|\text{vec}(\mathbf{H}_{\mathbf{Z}})\|^2/t$ around 1, and thereby concludes the proof of Lemma 4.9.

C Decomposition of the gradient matrix

Having characterized the distribution of $\mathbf{F}\text{vec}(\mathbf{H}_{\mathbf{Z}})$, we now turn our attention to the effect of a single gradient step during training, as defined in Equation (6). Specifically, we analyze the structure of the gradient matrix \mathbf{G} , following the approach of [4, 13].

Recall the definition of the gradient matrix (7):

$$\mathbf{G} := \frac{1}{n} \left(\frac{1}{\sqrt{k}} \left(\mathbf{w}\tilde{\mathbf{y}}^T - \frac{1}{\sqrt{k}} \mathbf{w}\mathbf{w}^T \sigma(\mathbf{F}\tilde{\mathbf{H}}^T) \right) \odot \sigma'(\mathbf{F}\tilde{\mathbf{H}}^T) \right) \tilde{\mathbf{H}}. \quad (\text{S21})$$

A complication arises from the presence of the derivative of the activation function σ , which appears as a multiplicative factor. To simplify this expression, we apply an orthogonal decomposition to the derivative:

$$\sigma'(z) = \alpha + \sigma'_{\perp}(z),$$

where $\alpha := \mathbb{E}_{z \sim \mathcal{N}(0,1)}[\sigma'(z)]$ is the average slope, and $\sigma'_\perp(z) := \sigma'(z) - \alpha$ captures the zero-mean deviation. The decomposition is justified by the fact that the random variable entering σ —namely, $\mathbf{f}_i^T \tilde{\mathbf{h}}_j$ —converges in distribution to $\mathcal{N}(0, 1)$ almost surely for all i, j , by Lemma 4.9.

Using this decomposition, the gradient matrix \mathbf{G} can be rewritten as:

$$\mathbf{G} = \underbrace{\mathbf{u}\mathbf{v}^T}_{\text{spike}} + \underbrace{\frac{1}{n\sqrt{k}} \left(\mathbf{w}\tilde{\mathbf{y}}^T \odot \sigma'_\perp(\mathbf{F}\tilde{\mathbf{H}}^T) \right) \tilde{\mathbf{H}} - \frac{1}{nk} \left(\mathbf{w}\mathbf{w}^T \sigma(\mathbf{F}\tilde{\mathbf{H}}^T) \odot \sigma'(\mathbf{F}\tilde{\mathbf{H}}^T) \right) \tilde{\mathbf{H}}}_{\Delta}, \quad (\text{S22})$$

where we define $\mathbf{u} := \alpha\mathbf{w}$ and $\mathbf{v} := \tilde{\mathbf{H}}^T \tilde{\mathbf{y}} / (n\sqrt{k})$.

This decomposition separates the gradient into two components: A *spike term* $\mathbf{u}\mathbf{v}^T$, which is expected to dominate, and a *residual term* Δ , which we aim to show is negligible.

Our next goal is to establish that $\|\mathbf{u}\mathbf{v}^T\| \gg \|\Delta\|$, thereby confirming that the spike term governs the spectral structure of \mathbf{G} .

A new technical challenge arises in this context: unlike prior work [4, 13], where $\tilde{\mathbf{H}}$ is composed of i.i.d. Gaussian vectors, in our setting each row of $\tilde{\mathbf{H}}$ is a realization of $\text{vec}(\mathbf{H}_Z)$. As shown in Lemma B.1 (Appendix B), \mathbf{H}_Z is the outer product of a Gaussian vector and a sub-exponential vector, making $\text{vec}(\mathbf{H}_Z)$ a heavy-tailed [41] random vector. This deviates from the sub-Gaussian assumptions in prior analyses and introduces additional complexity in bounding the norms of $\tilde{\mathbf{H}}$.

Nevertheless, in the following lemma, we characterize a spectral norm bound for $\tilde{\mathbf{H}}$ that accommodates this heavy-tailed structure.

Lemma C.1. *Under our assumptions, the spectral norm of the matrix $\tilde{\mathbf{H}}$ satisfies the following bound with high probability:*

$$\|\tilde{\mathbf{H}}\| / \|\text{Cov}(\text{vec}(\mathbf{H}_Z))\|^{1/2} = \tilde{\mathcal{O}}(d). \quad (\text{S23})$$

Proof. Corollaries B.2 and B.3 imply that

$$|\|\text{vec}(\mathbf{H}_Z)\|^2 - \text{Tr}(\text{Cov}(\text{vec}(\mathbf{H}_Z)))| = \mathcal{O}(d^2)$$

with high probability, given that $\text{Tr}(\text{Cov}(\text{vec}(\mathbf{H}_Z))) \asymp d^2$ by Assumption 4.4. It follows that

$$\|\text{vec}(\mathbf{H}_Z)\| = \mathcal{O}(d)$$

holds with high probability. We now apply a result from [39, Theorem 5.44] concerning the spectral norm of a random matrix with independent heavy-tailed rows. This yields the stated bound on $\|\tilde{\mathbf{H}}\|$. Note that the assumption $n/d^2 \in \mathbb{R}$ (from Assumption 4.1) ensures the validity of this application. \square

We now establish high-probability norm bounds for other key random quantities involved in the analysis. Specifically, the following lemma bounds the norms of \mathbf{w} , \mathbf{F} , and $\tilde{\mathbf{y}}$.

Lemma C.2. *Under our assumptions, the following bounds hold with high probability:*

- (i) $\|\mathbf{w}\| = \tilde{\mathcal{O}}(1)$ and $\|\mathbf{w}\|_\infty = \tilde{\mathcal{O}}(k^{-1/2})$,
- (ii) $\|\mathbf{F}\| = \tilde{\mathcal{O}}(1)$,
- (iii) $\|\tilde{\mathbf{y}}\| = \tilde{\mathcal{O}}(n^{1/2})$ and $\|\tilde{\mathbf{y}}\|_\infty = \tilde{\mathcal{O}}(1)$,

where $k, n = \mathcal{O}(d^2)$ as specified by Assumption 4.1.

Proof. (i) Follows from standard sub-Gaussian norm bounds, specifically [40, Proposition 2.5.2 and Theorem 3.1.1], in conjunction with Assumption 4.6.

(ii) The spectral norm bound on \mathbf{F} is a direct consequence of the concentration of the spectral norm for sub-Gaussian random matrices; see [40, Theorem 4.4.5].

(iii) The bounds for $\tilde{\mathbf{y}}$ follow from the Gaussian concentration of Lipschitz functions [40, Theorem 5.2.2], noting that each element of $\tilde{\mathbf{y}}$ is defined via a Lipschitz transformation of sub-Gaussian variables. \square

Next, we derive high-probability norm bounds for key quantities appearing in the residual term Δ of the decomposed gradient expression in (S22). Specifically, we consider the terms $\sigma'(\mathbf{F}\tilde{\mathbf{H}}^T)$, $\sigma'_\perp(\mathbf{F}\tilde{\mathbf{H}}^T)$, and $\mathbf{w}^T\sigma(\mathbf{F}\tilde{\mathbf{H}}^T)$.

Lemma C.3. *Under our assumptions, the following bounds hold with high probability:*

- (i) $\|\sigma'_\perp(\mathbf{F}\tilde{\mathbf{H}}^T)\| / \|\tilde{\mathbf{H}}\| = \tilde{\mathcal{O}}(k^{1/2})$,
- (ii) $\|\sigma'(\mathbf{F}\tilde{\mathbf{H}}^T)\| = \tilde{\mathcal{O}}(k)$,
- (iii) $\|\mathbf{w}^T\sigma(\mathbf{F}\tilde{\mathbf{H}}^T)\|_\infty / \|\tilde{\mathbf{H}}\| = \tilde{\mathcal{O}}(1)$,

where $k = \mathcal{O}(d^2)$, as specified by Assumption 4.1.

Proof. (i) From Lemma 4.9, the elements of $\mathbf{F}\tilde{\mathbf{H}}^T$ converge in distribution to $\mathcal{N}(0, 1)$ under the asymptotic regime specified by Assumption 4.1. Furthermore, the columns of $\mathbf{F}\tilde{\mathbf{H}}^T$ are independent by construction, although their covariance is not isotropic due to the structure imposed by $\text{Cov}(\text{vec}(\mathbf{H}_Z))$. The function $\sigma'_\perp(\cdot)$, being the centered component of the derivative, remains bounded and Lipschitz (by Assumption 4.8). Thus, the rows of $\sigma'_\perp(\mathbf{F}\tilde{\mathbf{H}}^T)^T$ are independent anisotropic sub-Gaussian vectors. Using Gaussian concentration of Lipschitz functions [40, Theorem 5.2.2] and spectral norm bounds for matrices with independent sub-Gaussian rows [39, Theorem 5.39 and Eq. (5.26)], we obtain the desired result.

(ii) We bound the full derivative term as

$$\|\sigma'(\mathbf{F}\tilde{\mathbf{H}}^T)\| \leq \|\sigma'_\perp(\mathbf{F}\tilde{\mathbf{H}}^T)\| + \alpha \|\mathbf{1}_{k \times d(d+1)}\|,$$

where $\alpha := \mathbb{E}_{z \sim \mathcal{N}(0,1)}[\sigma'(z)]$ and $\mathbf{1}_{k \times d(d+1)}$ denotes a matrix of all ones. Since $\|\mathbf{1}_{k \times d(d+1)}\| = \mathcal{O}(k)$, the result follows from part (i).

(iii) The term $\mathbf{w}^T\sigma(\mathbf{F}\tilde{\mathbf{H}}^T)$ is a Lipschitz function of Gaussian random variables, and thus obeys Gaussian concentration [40, Theorem 5.2.2]. Since $\|\mathbf{w}\| = \tilde{\mathcal{O}}(1)$ by Lemma C.2, the bound follows. \square

With all necessary high-probability bounds in place, we now derive a bound on the norm of the residual term Δ appearing in the gradient decomposition. Recall:

$$\|\Delta\| = \left\| \frac{1}{n\sqrt{k}} \left(\mathbf{w}\tilde{\mathbf{y}}^T \odot \sigma'_\perp(\mathbf{F}\tilde{\mathbf{H}}^T) \right) \tilde{\mathbf{H}} - \frac{1}{nk} \left(\mathbf{w}\mathbf{w}^T\sigma(\mathbf{F}\tilde{\mathbf{H}}^T) \odot \sigma'(\mathbf{F}\tilde{\mathbf{H}}^T) \right) \tilde{\mathbf{H}} \right\|, \quad (\text{S24})$$

$$\leq \frac{1}{n\sqrt{k}} \left\| \mathbf{w}\tilde{\mathbf{y}}^T \odot \sigma'_\perp(\mathbf{F}\tilde{\mathbf{H}}^T) \right\| \|\tilde{\mathbf{H}}\| + \frac{1}{nk} \left\| \mathbf{w}\mathbf{w}^T\sigma(\mathbf{F}\tilde{\mathbf{H}}^T) \odot \sigma'(\mathbf{F}\tilde{\mathbf{H}}^T) \right\| \|\tilde{\mathbf{H}}\|, \quad (\text{S25})$$

$$\leq \frac{1}{n\sqrt{k}} \|\mathbf{w}\|_\infty \|\tilde{\mathbf{y}}\|_\infty \|\sigma'_\perp(\mathbf{F}\tilde{\mathbf{H}}^T)\| \|\tilde{\mathbf{H}}\| + \frac{1}{nk} \|\mathbf{w}\|_\infty \|\mathbf{w}^T\sigma(\mathbf{F}\tilde{\mathbf{H}}^T)\|_\infty \|\sigma'(\mathbf{F}\tilde{\mathbf{H}}^T)\| \|\tilde{\mathbf{H}}\|, \quad (\text{S26})$$

$$= \tilde{\mathcal{O}}(k^{-\beta}), \quad (\text{S27})$$

which holds with high probability, where $\beta \in [0, 1]$ satisfies $\|\text{Cov}(\text{vec}(\mathbf{H}_Z))\| = \tilde{\mathcal{O}}(k^{-\beta+1})$ under Assumption 4.5.

To derive this bound, we first apply the triangle inequality, and then use the identity for Hadamard products:

$$\mathbf{a}\mathbf{b}^T \odot \mathbf{C} = \text{diag}(\mathbf{a}) \mathbf{C} \text{diag}(\mathbf{b}),$$

which allows us to factor out the vectors and simplify norm computations. The final result follows by applying high-probability bounds established in Lemmas C.1, C.2, and C.3.

Similarly, we can bound the norm of the spiked term $\mathbf{u}\mathbf{v}^T$:

$$\|\mathbf{u}\mathbf{v}^T\| = \|\mathbf{u}\| \|\mathbf{v}\| = \frac{\alpha}{n\sqrt{k}} \|\mathbf{w}\| \|\tilde{\mathbf{H}}\| \|\tilde{\mathbf{y}}\| = \tilde{\mathcal{O}}(k^{-\beta/2}), \quad (\text{S28})$$

which also holds with high probability.

Taken together, the bounds in (S27) and (S28) confirm that the leading contribution to the gradient matrix arises from the rank-one term uv^T , whereas Δ constitutes a negligible residual. This completes the proof of Lemma 4.11.

D Conditional Gaussian equivalence

With the results from the preceding appendices, we now establish a new result: a *conditional Gaussian equivalence*, which will play a key role in the proof of Theorem 4.12. To formulate this result, we first define a subspace relevant for the conditioning in the equivalence argument.

Lemma D.1 (Decomposition of $\hat{F}\text{vec}(\mathbf{H}_Z)$ conditioned on data source s). *Let s be a fixed data source. Define the subspace*

$$\mathbf{S} := [\mathbf{v}, \gamma_{s,1}, \gamma_{s,2}, \dots, \gamma_{s,r_s}],$$

where \mathbf{v} is the spiked direction from Lemma 4.11, and $\gamma_{s,1}, \dots, \gamma_{s,r_s}$ are the spiked directions of $\text{Cov}(\text{vec}(\mathbf{H}_Z))$ as specified in Assumption 4.4. Let \mathbf{P} and \mathbf{P}_\perp denote the orthogonal projection matrices onto $\text{span}(\mathbf{S})$ and its orthogonal complement, respectively. Then,

$$\hat{F}\text{vec}(\mathbf{H}_Z) = (\mathbf{F} + \eta\Delta)\mathbf{P}_\perp\text{vec}(\mathbf{H}_Z) + \hat{\mathbf{F}}\mathbf{S}\boldsymbol{\kappa}_s, \quad (\text{S29})$$

where $\boldsymbol{\kappa}_s := (\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}^T\text{vec}(\mathbf{H}_Z)$.

Proof. This decomposition follows directly from the definition of orthogonal projections and the linearity of matrix multiplication. \square

Lemma D.1 yields a decomposition of the transformed hidden state $\hat{F}\text{vec}(\mathbf{H}_Z)$ into two components: a “bulk” term that is conditionally Gaussian, and a structured term aligned with the low-rank subspace \mathbf{S} induced by spiked directions. Crucially, this implies that, conditional on the coefficient vector $\boldsymbol{\kappa}_s$, the transformation $\hat{F}\text{vec}(\mathbf{H}_Z)$ is sub-Gaussian. This motivates the following conditional Gaussian equivalence theorem, which approximates the nonlinear feature map $\sigma(\hat{F}\text{vec}(\mathbf{H}_Z))$ with a Gaussian counterpart conditioned on $(s, \boldsymbol{\kappa}_s)$.

Theorem D.2 (Conditional Gaussian equivalence). *Under the assumptions in Section 4.1, define the nonlinear feature map $\psi(\text{vec}(\mathbf{H}_Z)) := \sigma(\hat{F}\text{vec}(\mathbf{H}_Z))$, and let $\mathbf{o} := \mathbf{P}_\perp\text{vec}(\mathbf{H}_Z)$ be the projection onto the orthogonal complement of the subspace \mathbf{S} defined in Lemma D.1. Then, the following conditional Gaussian feature map is equivalent to $\psi(\text{vec}(\mathbf{H}_Z))$ in terms of ICL error; when conditioned on data source s and alignment vector $\boldsymbol{\kappa}_s$:*

$$\hat{\psi}(\text{vec}(\mathbf{H}_Z); s, \boldsymbol{\kappa}_s) := \boldsymbol{\nu}(s, \boldsymbol{\kappa}_s) + \boldsymbol{\Psi}(s, \boldsymbol{\kappa}_s)\mathbf{o} + \boldsymbol{\Phi}(s, \boldsymbol{\kappa}_s)^{1/2}\mathbf{g}, \quad (\text{S30})$$

where $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I})$, and

$$\begin{aligned} \boldsymbol{\nu}(s, \boldsymbol{\kappa}_s) &:= \mathbb{E} \left[\sigma(\hat{F}\text{vec}(\mathbf{H}_Z)) \mid s, \boldsymbol{\kappa}_s \right], \\ \boldsymbol{\Psi}(s, \boldsymbol{\kappa}_s) &:= \mathbb{E} \left[\sigma(\hat{F}\text{vec}(\mathbf{H}_Z))\mathbf{o}^T \mid s, \boldsymbol{\kappa}_s \right], \\ \boldsymbol{\Phi}(s, \boldsymbol{\kappa}_s) &:= \text{Cov} \left(\sigma(\hat{F}\text{vec}(\mathbf{H}_Z)) \mid s, \boldsymbol{\kappa}_s \right) - \boldsymbol{\Psi}(s, \boldsymbol{\kappa}_s)\boldsymbol{\Psi}(s, \boldsymbol{\kappa}_s)^T. \end{aligned}$$

Proof. The proof strategy follows the standard approach to Gaussian equivalence for random features as developed in [31, 11, 10, 13]. In particular, it is sufficient to establish a central limit theorem (CLT) for the bulk component $(\mathbf{F} + \eta\Delta)\mathbf{o}$, conditional on $s, \boldsymbol{\kappa}_s$. Once this CLT is in place, the remainder of the proof mirrors that of [13], and is omitted here for brevity.

Conditional CLT For any Lipschitz function $\zeta : \mathbb{R}^2 \rightarrow \mathbb{R}$, for all $s \in \{0, \dots, \mathcal{S} - 1\}$ and for all $\boldsymbol{\kappa}_s \in \mathbb{R}^{r_s+1}$,

$$\lim_{d,k \rightarrow \infty} \sup_{\tilde{\mathbf{w}}, \tilde{\boldsymbol{\xi}}} \left| \mathbb{E} \left[\zeta \left(\tilde{\mathbf{w}}^T \psi(\text{vec}(\mathbf{H}_Z)), \tilde{\boldsymbol{\xi}}^T \mathbf{x} \right) \mid s, \boldsymbol{\kappa}_s \right] - \mathbb{E} \left[\zeta \left(\tilde{\mathbf{w}}^T \hat{\psi}(\text{vec}(\mathbf{H}_Z)), \tilde{\boldsymbol{\xi}}^T \mathbf{x} \right) \mid s, \boldsymbol{\kappa}_s \right] \right| = 0, \quad (\text{S31})$$

where the supremum is taken over $\tilde{\mathbf{w}} \in \{\mathbf{w} \in \mathbb{R}^k \mid \|\mathbf{w}\| = \mathcal{O}(1), \|\mathbf{w}\|_\infty = \mathcal{O}(k^{-\epsilon})\}$ for some $\epsilon > 0$, and $\tilde{\boldsymbol{\xi}} \in \mathbb{R}^d$ satisfies $\|\tilde{\boldsymbol{\xi}}\| = 1/\|\text{Cov}(\mathbf{x} \mid s)\|^{1/2}$. Here, $k, d \rightarrow \infty$ such that $k/d^2 \in \mathbb{R}$, as specified in Assumption 4.1.

This conditional CLT establishes the equivalence of the original and Gaussian feature maps, $\psi(\text{vec}(\mathbf{H}_Z))$ and $\hat{\psi}(\text{vec}(\mathbf{H}_Z); s, \boldsymbol{\kappa}_s)$, in terms of their behavior under any test function ζ , conditional on s and $\boldsymbol{\kappa}_s$. Notably, the supremum ensures robustness to variation in task vectors, accounting for worst-case alignment.

We begin proving the conditional CLT by recalling the decomposition from Lemma D.1:

$$\hat{\mathbf{F}}\text{vec}(\mathbf{H}_Z) = (\mathbf{F} + \eta\boldsymbol{\Delta})\mathbf{P}_\perp\text{vec}(\mathbf{H}_Z) + \hat{\mathbf{F}}\mathbf{S}\boldsymbol{\kappa}_s.$$

Define $\mathbf{r} := (\mathbf{F} + \eta\boldsymbol{\Delta})\mathbf{P}_\perp\text{vec}(\mathbf{H}_Z)$ and $\mathbf{c} := \hat{\mathbf{F}}\mathbf{S}\boldsymbol{\kappa}_s$ so that $\hat{\mathbf{F}}\text{vec}(\mathbf{H}_Z) = \mathbf{r} + \mathbf{c}$. The random vector \mathbf{r} is approximately Gaussian because:

- $\|\boldsymbol{\Delta}\| \rightarrow 0$ asymptotically (see Appendix C),
- \mathbf{P}_\perp is an orthogonal projection ($\|\mathbf{P}_\perp\| = 1$),
- Lemma 4.9 shows that $\mathbf{F}\text{vec}(\mathbf{H}_Z)$ converges to a Gaussian distribution,
- Assumption 4.4 ensures \mathbf{P}_\perp eliminates spiked directions, leading to $\mathbf{r} \rightarrow \mathcal{N}(0, \mathbf{I}_k)$ almost surely.

Consequently, we may write $\mathbf{r} \stackrel{d}{=} \tilde{\mathbf{F}}\mathbf{q}$, where $\stackrel{d}{=}$ denotes equivalence in distribution, $\tilde{\mathbf{F}}$ is a random feature matrix satisfying the conditions in [21] and $\mathbf{q} \sim \mathcal{N}(0, \mathbf{I})$. Meanwhile, \mathbf{c} is deterministic conditional on $(\mathbf{F}, s, \boldsymbol{\kappa}_s)$. Define the neuron-wise conditional activation:

$$\tilde{\sigma}_{j|(s, \boldsymbol{\kappa}_s)}(\tilde{\mathbf{f}}_j^T \mathbf{q}) := \sigma_j(\tilde{\mathbf{f}}_j^T \mathbf{q} + c_j), \quad j = 1, \dots, k.$$

This representation allows us to view the feature map as a collection of neuron-specific activations with fixed shifts c_j . Importantly, the CLT for random features in [21] applies even when activations vary across neurons, a point further supported by related results in [11, 13].

Utilizing Corollary 4.10 and applying the central limit theorem for heterogeneous neuron activations [21, Theorem 2], we obtain convergence in distribution of $\psi(\text{vec}(\mathbf{H}_Z))$ to its Gaussian approximation $\hat{\psi}(\text{vec}(\mathbf{H}_Z); s, \boldsymbol{\kappa}_s)$. Furthermore, the odd activation function assumption required in [21] can be omitted here, since both feature maps share the same conditional covariance structure. See [31, 11, 10, 13] for related proofs and further technical details.

So far, we have established an equivalence with respect to the training error. While not stated explicitly earlier, this training error corresponds to that of ridge regression applied to the second-layer weights in our two-stage training setup. To extend this asymptotic equivalence to the ICL error defined in (8)—which is our main object of interest—we require an additional technical condition, formalized below.

Assumption D.3. Consider a perturbed optimization objective:

$$\mathcal{T}_n(c) := \min_{\mathbf{w} \in \mathcal{C}_k} \frac{1}{n} \sum_{j=1}^n \left(y_{\ell+1}^j - \frac{1}{\sqrt{k}} \mathbf{w}^T \sigma(\hat{\mathbf{F}}\text{vec}(\mathbf{H}_{Z^j})) \right)^2 + \lambda \|\mathbf{w}\|_2^2 + c \mathcal{E}(\mathbf{w}), \quad (\text{S32})$$

where $c \in \mathbb{R}$ is a scalar perturbation parameter, and $\mathcal{E}(\mathbf{w})$ denotes the ICL error from (8) associated with second-layer weight vector \mathbf{w} . The constraint set \mathcal{C}_k is the constraint set in the conditional CLT and it is defined as $\mathcal{C}_k := \{\mathbf{w} \in \mathbb{R}^k \mid \|\mathbf{w}\| = \mathcal{O}(1), \|\mathbf{w}\|_\infty = \mathcal{O}(k^{-\epsilon})\}$ for some $\epsilon > 0$. Then, there exists a constant $c^* > 0$ such that, for all $c \in [-c^*, c^*]$, the function $\mathcal{T}_n(c)$ converges pointwise to a limiting function $\mathcal{T}(c)$, which is differentiable at $c = 0$.

Although this assumption may appear somewhat artificial at first glance, it enables the use of convexity-based arguments in establishing generalization error equivalence (ICL errors in our case), as seen in prior work on Gaussian universality [11, Assumption 5]. Similar assumptions are also employed in related studies on asymptotic equivalence [21, 31, 13]. Importantly, this condition arises

primarily as a technical requirement of the proof method, rather than as a limitation on the broader applicability of the Gaussian equivalence results established in this work.

With Assumption D.3 in place, we are now able to extend the asymptotic equivalence result to the ICL error, thereby completing the proof of the conditional Gaussian equivalence, as established in [13]. \square

Theorem D.2 establishes an asymptotic equivalence between two feature maps with respect to the ICL error, under the condition that their first two conditional moments match. This result implies the existence of an equivalent (conditional) Gaussian model—namely, $\mathbf{w}^T \hat{\psi}(\text{vec}(\mathbf{H}_Z); s, \boldsymbol{\kappa}_s)$ —that can replace the original feature map without affecting ICL performance. Leveraging this equivalence, we now prove that the Transformer model with a nonlinear MLP is asymptotically equivalent to a polynomial model, as formalized in Theorem 4.12. The full proof is presented below.

E Equivalent polynomial model

We aim to establish the asymptotic equivalence between a Transformer model with a nonlinear MLP and the polynomial model described in Theorem 4.12. Our strategy relies on the conditional Gaussian equivalence result stated and proved in Appendix D (Theorem D.2). This result asserts that two models with matching conditional means and covariances yield the same ICL error (8). Thus, to prove equivalence, it suffices to demonstrate that the first two conditional moments of the two models coincide.

Fix a data source index s , and recall the orthogonal decomposition with respect to the subspace defined by the matrix \mathbf{S} :

$$\hat{\mathbf{F}}\text{vec}(\mathbf{H}_Z) = (\mathbf{F} + \eta\boldsymbol{\Delta})\mathbf{P}_\perp \text{vec}(\mathbf{H}_Z) + \hat{\mathbf{F}}\mathbf{S}\boldsymbol{\kappa}_s, \quad (\text{S33})$$

where $\boldsymbol{\kappa}_s := (\mathbf{S}^T \mathbf{S})^{-1} \mathbf{S}^T \text{vec}(\mathbf{H}_Z)$, as established in Lemma D.1. The first term, $(\mathbf{F} + \eta\boldsymbol{\Delta})\mathbf{P}_\perp \text{vec}(\mathbf{H}_Z)$, behaves asymptotically like a Gaussian random variable (see the proof of Theorem D.2), while the second term, $\hat{\mathbf{F}}\mathbf{S}\boldsymbol{\kappa}_s$, is deterministic conditional on \mathbf{F} , s , and $\boldsymbol{\kappa}_s$.

According to [13, Theorem 4], if conditional Gaussian equivalence holds and the deterministic component $\hat{\mathbf{F}}\mathbf{S}\boldsymbol{\kappa}_s$ vanishes at a rate of $\tilde{\mathcal{O}}(k^{-\delta})$ for some $\delta > 0$, then there exists a finite polynomial degree p such that the polynomial activation $\hat{\sigma}_p(\cdot)$ (as defined in Theorem 4.12) yields equivalent generalization performance to that of the original nonlinear activation $\sigma(\cdot)$. In their proof, the vanishing nature of the deterministic term is used to show that $\hat{\sigma}_p(\cdot)$ and $\sigma(\cdot)$ induce the same first two conditional moments, thereby ensuring asymptotic equivalence in generalization (ICL) error.

Similarly, in our setting, the vanishing nature of the term $\hat{\mathbf{F}}\mathbf{S}\boldsymbol{\kappa}_s$ —together with Assumption 4.8—implies the following bounds:

$$\left\| \mathbb{E} \left[\sigma(\hat{\mathbf{F}}\text{vec}(\mathbf{H}_Z)) \mid s, \boldsymbol{\kappa}_s \right] - \mathbb{E} \left[\hat{\sigma}_p(\hat{\mathbf{F}}\text{vec}(\mathbf{H}_Z)) \mid s, \boldsymbol{\kappa}_s \right] \right\| = o(1), \quad (\text{S34})$$

$$\left\| \mathbb{E} \left[\sigma(\hat{\mathbf{F}}\text{vec}(\mathbf{H}_Z)) \mathbf{o}^T \mid s, \boldsymbol{\kappa}_s \right] - \mathbb{E} \left[\hat{\sigma}_p(\hat{\mathbf{F}}\text{vec}(\mathbf{H}_Z)) \mathbf{o}^T \mid s, \boldsymbol{\kappa}_s \right] \right\|_F = o(1), \quad (\text{S35})$$

$$\left\| \text{Cov} \left(\sigma(\hat{\mathbf{F}}\text{vec}(\mathbf{H}_Z)) \mid s, \boldsymbol{\kappa}_s \right) - \text{Cov} \left(\hat{\sigma}_p(\hat{\mathbf{F}}\text{vec}(\mathbf{H}_Z)) \mid s, \boldsymbol{\kappa}_s \right) \right\| = o(1), \quad (\text{S36})$$

where $\mathbf{o} := \mathbf{P}_\perp \text{vec}(\mathbf{H}_Z)$. These bounds suffice to prove the equivalence between the original activation $\sigma(\cdot)$ and the polynomial approximation $\hat{\sigma}_p(\cdot)$, as explained by [13]. Therefore, it remains to verify that

$$|\hat{\mathbf{f}}_i^T \mathbf{S}\boldsymbol{\kappa}_s| = \tilde{\mathcal{O}}(k^{-\delta}) \quad \text{for all } i \in \{1, \dots, k\}, \quad (\text{S37})$$

for some $\delta > 0$, where $\hat{\mathbf{f}}_i$ denotes the i -th row of $\hat{\mathbf{F}}$.

To analyze $|\hat{\mathbf{f}}_i^T \mathbf{S}\boldsymbol{\kappa}_s|$, we begin by expanding the expression:

$$|\hat{\mathbf{f}}_i^T \mathbf{S}\boldsymbol{\kappa}_s| = \left| \hat{\mathbf{f}}_i^T \mathbf{S}(\mathbf{S}^T \mathbf{S})^{-1} \mathbf{S}^T \text{vec}(\mathbf{H}_Z) \right| \quad (\text{S38})$$

$$\leq \left| \hat{\mathbf{f}}_i^T \mathbf{S}(\mathbf{S}^T \mathbf{S})^{-1} \mathbf{S}^T \text{vec}(\mathbf{H}_Z) \right| + \eta \left| \mathbf{g}_i^T \mathbf{S}(\mathbf{S}^T \mathbf{S})^{-1} \mathbf{S}^T \text{vec}(\mathbf{H}_Z) \right|, \quad (\text{S39})$$

where we used the decomposition $\hat{f}_i = f_i + \eta g_i$, and g_i is the i -th row of the gradient matrix G defined in (7). The first term corresponds to the contribution from the randomly initialized feature matrix F , while the second accounts for the effect of the gradient update.

Finally, invoking the bounds derived in Appendix C, along with Assumptions 4.4 and 4.5, we conclude that each term vanishes at the desired rate. These assumptions ensure that the step size η and the influence of spiked directions in $\text{Cov}(\text{vec}(\mathbf{H}_Z))$ remain controlled, which completes the proof.

In summary, our results indicate that although the output of the attention layer, $\text{vec}(\mathbf{H}_Z)$, exhibits a heavy-tailed distribution [41], the application of the random matrix F attenuates the heavy tails. Moreover, training the first-layer weights F with a single gradient step has a negligible effect on the analysis. Together, these insights allow us to transfer known results from supervised learning with two-layer neural networks to the in-context learning setting of Transformers with nonlinear MLPs. This connection opens promising avenues for analyzing complex, realistic models using Gaussian equivalence techniques.

F Extension to inputs with non-zero mean

In the main body of our proofs, we have assumed zero-mean inputs, i.e., $\mu_{x,s} = \mathbf{0}$ for all data sources s , to streamline the exposition and focus on the core aspects of the Gaussian equivalence framework. However, in practical scenarios, input data often has non-zero mean, necessitating a generalization of our theoretical results to handle such cases.

The key observation enabling this extension is that the mean vector $\mu_{x,s}$ of each Gaussian data source introduces a structured, low-rank perturbation in the representation of the attention output. Formally, we have:

$$\mathbf{H}_Z \mid s = (\mathbf{x}_{\ell+1} \mid s - \mu_{x,s})\mathbf{b}^T + \mu_{x,s}\mathbf{b}^T, \quad (\text{S40})$$

which separates the stochastic (zero-mean) and deterministic (mean) components. Consequently, the vectorized attention output $\text{vec}(\mathbf{H}_Z)$ becomes non-zero-mean, but its non-central second moment still captures the combined statistical behavior of both components.

To rigorously extend our results to this more general setting, the following changes are required.

1. Substitution of covariances with second moments:

- In all analytical steps where the covariances of $\mathbf{x}_i \mid s$ and $\text{vec}(\mathbf{H}_Z)$ appear, we need to replace it with the corresponding non-central second moment:

$$\begin{aligned} \text{Cov}(\text{vec}(\mathbf{H}_Z)) &\rightsquigarrow \mathbb{E}[\text{vec}(\mathbf{H}_Z)\text{vec}(\mathbf{H}_Z)^T], \\ \Sigma_{x,s} = \text{Cov}(\mathbf{x}_i \mid s) &\rightsquigarrow \mathbb{E}[\mathbf{x}_i \mathbf{x}_i^T \mid s]. \end{aligned}$$

- This replacement maintains the validity of our results since the spike due to the mean is treated analogously to spikes in the spiked covariance model in Assumption 4.4.

2. Bounding mean-induced terms in gradient decomposition:

- The primary technical adjustment is required in the proof of the gradient decomposition (Appendix C).
- The decomposition must now include additional cross-terms involving $\mu_{x,s}$, which must be carefully bounded using concentration inequalities and structural assumptions on the data distribution (e.g., bounded mean norm, low-rank behavior).
- These bounds ensure that the mean-induced components do not asymptotically dominate or distort the gradient behavior established under the zero-mean assumption.

By implementing the above modifications—namely, using non-central second moments and bounding mean-induced gradient terms—we can extend our theoretical guarantees to the case of non-zero-mean Gaussian inputs. This generalization not only reinforces the robustness of our analysis but also broadens its relevance to real-world applications, where input means are rarely zero in practice.

G ICL errors per-source in the setting of Figure 2

For the sake of completeness, we illustrate the ICL errors per-source (corresponding to each source) in the setting of Figure 2. Mathematically, the per-source error is defined as $\mathbb{E}[(y_{\ell+1} - \hat{y})^2 | s = \hat{s}]$ where source indicator $\hat{s} \in \{0, 1\}$ since we consider settings with two different data sources. For all of the figures below, on the left, the ICL error corresponding to source 0, i.e., $\mathbb{E}[(y_{\ell+1} - \hat{y})^2 | s = 0]$, is plotted while the ICL error for source 1, i.e., $\mathbb{E}[(y_{\ell+1} - \hat{y})^2 | s = 1]$, is shown on the right. Figures 4, 5, and 6 illustrate the per-source ICL errors in the settings of Figure 2 (a), (b), and (c), respectively. Namely, Figure 4 shows the changes in the per-source ICL errors when varying the input covariance and Figure 5 displays the effects of changing the task covariance on the per-source ICL errors, while Figure 6 depicts the per-source ICL errors for different noise levels. In all of these cases, the first data source (source 0) is kept fixed while the aforementioned properties of the second data source are modified, as detailedly explained in the caption of Figure 2. The results indicate two primary points. First, per-source ICL errors for the Transformer (approximately) match those of the equivalent model, indicating that the equivalence specified by Theorem 4.12 is useful for studying per-source ICL errors as well. Second, the trends for the ICL errors (which is the average of the per-source errors) in Figure 2 are also observed for each of the per-source errors, providing further evidence for our conclusions.

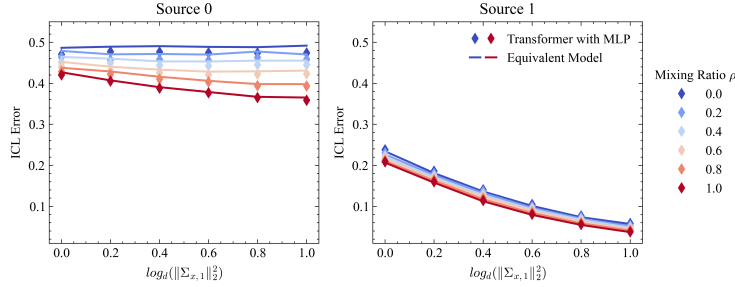


Figure 4: Per-source ICL errors in the case of Figure 2(a): impact of varying input covariance of source 1 on the per-source ICL errors.

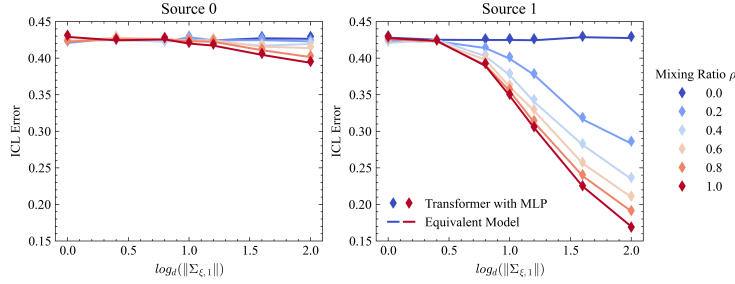


Figure 5: Per-source ICL errors in the case of Figure 2(b): effect of altering the task covariance of source 1 on the per-source ICL errors.

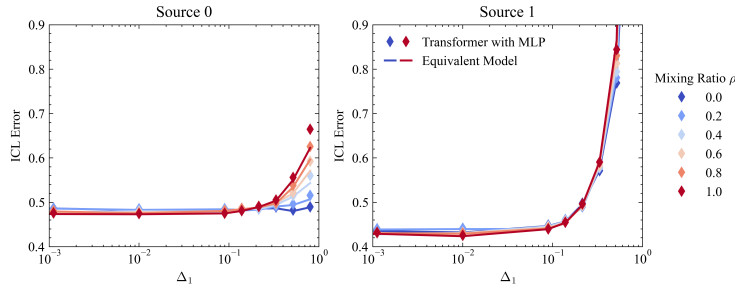


Figure 6: Per-source ICL errors in the case of Figure 2(c): result of changing noise level of source 1 on the per-source ICL errors.