

1 A Derivation of Bayes-optimal ridge estimator for w

2 Given a set of context samples, we derive the Bayes-optimal ridge estimator for w . We begin by
 3 placing a Gaussian prior on w , assumed to be a random vector with distribution $w \sim \mathcal{N}(\mu_0, \Sigma_0)$,
 4 where μ_0 denotes the prior mean and Σ_0 the prior covariance matrix. Let the observed context data
 5 consist of centered input vectors $\bar{X} = [\bar{x}_1, \dots, \bar{x}_{l-1}]^T$ and centered labels $\bar{y} = [\bar{y}_1, \dots, \bar{y}_{l-1}]^T$.
 6 Under the assumption of i.i.d. Gaussian noise with variance σ^2 , the likelihood of the data given w is
 7 expressed as

$$p(\bar{y}|\bar{X}, w) = \prod_{i=1}^{l-1} p(\bar{y}_i|\bar{x}_i, w) = \prod_{i=1}^{l-1} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\bar{y}_i - w^T \bar{x}_i)^2}{2\sigma^2}\right), \quad (S1)$$

$$p(\bar{y}|\bar{X}, w) \propto \exp\left(-\frac{1}{2\sigma^2}(\bar{y} - \bar{X}w)^T(\bar{y} - \bar{X}w)\right), \quad (S2)$$

8 where \propto denotes proportionality, and the second expression reformulates the likelihood compactly in
 9 matrix notation.

10 Applying Bayes' theorem, the posterior distribution of w conditioned on the observed data is
 11 proportional to the product of the likelihood and the prior:

$$p(w|\bar{y}, \bar{X}) \propto p(\bar{y}|\bar{X}, w) p(w). \quad (S3)$$

12 Substituting the previously derived expressions for the likelihood and the prior yields:

$$p(w|\bar{y}, \bar{X}) \propto \exp\left(-\frac{1}{2\sigma^2}(\bar{y} - \bar{X}w)^T(\bar{y} - \bar{X}w)\right) \exp\left(-\frac{1}{2}(w - \mu_0)^T \Sigma_0^{-1}(w - \mu_0)\right). \quad (S4)$$

13 To determine the form of the posterior distribution, we complete the square in the exponent by
 14 collecting all terms involving w . Expanding the exponent in the joint expression from above, we
 15 obtain:

$$-\frac{1}{2\sigma^2}(\bar{y}^T \bar{y} - 2\bar{y}^T \bar{X}w + w^T \bar{X}^T \bar{X}w) - \frac{1}{2}(w^T \Sigma_0^{-1}w - 2\mu_0^T \Sigma_0^{-1}w + \mu_0^T \Sigma_0^{-1}\mu_0). \quad (S5)$$

16 Grouping the quadratic and linear terms in w , we arrive at:

$$-\frac{1}{2}w^T \left(\frac{\bar{X}^T \bar{X}}{\sigma^2} + \Sigma_0^{-1}\right) w + w^T \left(\frac{\bar{X}^T \bar{y}}{\sigma^2} + \Sigma_0^{-1}\mu_0\right) + \text{terms independent of } w. \quad (S6)$$

17 Defining the posterior precision and linear coefficient terms as $\Sigma_l^{-1} = \frac{\bar{X}^T \bar{X}}{\sigma^2} + \Sigma_0^{-1}$ and $b_l =$
 18 $\frac{\bar{X}^T \bar{y}}{\sigma^2} + \Sigma_0^{-1}\mu_0$, the exponent can be rewritten as

$$-\frac{1}{2}w^T \Sigma_l^{-1}w + w^T b_l = -\frac{1}{2}(w - \mu_l)^T \Sigma_l^{-1}(w - \mu_l) + \text{const}, \quad (S7)$$

19 where $\mu_l = \Sigma_l b_l$ denotes the posterior mean. Expanding this expression gives:

$$\mu_l = \left(\frac{\bar{X}^T \bar{X}}{\sigma^2} + \Sigma_0^{-1}\right)^{-1} \left(\frac{\bar{X}^T \bar{y}}{\sigma^2} + \Sigma_0^{-1}\mu_0\right). \quad (S8)$$

20 Hence, the posterior distribution of w given the observed data is Gaussian:

$$w | \bar{y}, \bar{X} \sim \mathcal{N}(\mu_l, \Sigma_l), \quad (S9)$$

21 where μ_l is the posterior mean and Σ_l is the posterior covariance matrix.

22 Under a squared error loss, the Bayes-optimal estimator corresponds to the posterior mean. Therefore,
 23 the Bayes-optimal estimate of w is given by

$$\hat{w}_{\text{Ridge}} = \mathbb{E}[w | \bar{y}, \bar{X}] = \mu_l = \left(\frac{\bar{X}^T \bar{X}}{\sigma^2} + \Sigma_0^{-1}\right)^{-1} \left(\frac{\bar{X}^T \bar{y}}{\sigma^2} + \Sigma_0^{-1}\mu_0\right). \quad (S10)$$

24 This expression defines the Bayes-optimal ridge estimator, derived under the assumptions of a
 25 Gaussian prior $w \sim \mathcal{N}(\mu_0, \Sigma_0)$ and additive Gaussian noise in the labels. The result minimizes the
 26 expected squared error with respect to the posterior distribution.

27 B Derivation of linearized softmax

28 The softmax function $\text{softmax} : \mathbb{R}^l \rightarrow \mathbb{R}^l$ is defined component-wise as

$$\text{softmax}(\mathbf{z})_i := \frac{e^{z_i}}{\sum_{j=1}^l e^{z_j}} \quad \forall i \in \{1, 2, \dots, l\}. \quad (\text{S11})$$

29 To obtain a linear approximation, we apply a first-order Taylor series expansion around the origin
30 $\mathbf{z} = \mathbf{0}$:

$$\text{softmax}(\mathbf{z}) \approx \text{softmax}(\mathbf{0}) + J_{\text{softmax}}(\mathbf{0})(\mathbf{z} - \mathbf{0}), \quad (\text{S12})$$

31 where $J_{\text{softmax}}(\mathbf{0})$ denotes the Jacobian matrix of the softmax function evaluated at $\mathbf{z} = \mathbf{0}$. We now
32 evaluate the first term in the expansion:

$$\text{softmax}(\mathbf{0}) = \frac{e^0}{\sum_{j=1}^l e^0} \mathbf{1} = \frac{1}{\sum_{j=1}^l 1} \mathbf{1} = \frac{1}{l} \mathbf{1}. \quad (\text{S13})$$

33 Next, we compute the Jacobian matrix $J_{\text{softmax}}(\mathbf{0})$ by evaluating its entries:

$$J_{\text{softmax}}(\mathbf{0})_{ii} = \text{softmax}(\mathbf{0})_i (1 - \text{softmax}(\mathbf{0})_i) = \frac{l-1}{l^2}, \quad \forall i \in \{1, 2, \dots, l\}, \quad (\text{S14})$$

$$J_{\text{softmax}}(\mathbf{0})_{ij} = -\text{softmax}(\mathbf{0})_i \cdot \text{softmax}(\mathbf{0})_j = -\frac{1}{l^2}, \quad \forall i \neq j. \quad (\text{S15})$$

34 Combining these expressions, we obtain a compact matrix representation:

$$J_{\text{softmax}}(\mathbf{0}) = \frac{1}{l} \mathbf{I} - \frac{1}{l^2} \mathbf{1}\mathbf{1}^T. \quad (\text{S16})$$

35 We can now write the linearized softmax function as

$$\text{softmax}(\mathbf{z}) \approx \text{softmax}(\mathbf{0}) + J_{\text{softmax}}(\mathbf{0})\mathbf{z}, \quad (\text{S17})$$

$$= \frac{1}{l} \mathbf{1} + \frac{1}{l} \mathbf{z} - \frac{1}{l^2} \mathbf{1}\mathbf{1}^T \mathbf{z}, \quad (\text{S18})$$

$$= \left(\frac{1}{l} - \frac{1}{l^2} \sum_{j=1}^l z_j \right) \mathbf{1} + \frac{1}{l} \mathbf{z}, \quad (\text{S19})$$

36 which yields the linearized attention formulation in Eq. (5). From a practical perspective, such
37 linearized attention mechanisms have been empirically evaluated and shown to attain performance
38 comparable to that of the standard softmax attention [11].

39 C Linear vs. linearized attention for in-context learning

40 Here, we highlight the distinction between linear attention and linearized attention in the context of
41 the linear regression problem defined in Eq. (2). Analytically, the key difference lies in the fact that
42 the linearized attention model operates on centered input data, whereas the linear attention model
43 uses raw data without centering. Apart from this centering step, both mechanisms are equivalent,
44 except that linearized attention includes an additional bias term (b_{Att} in Eq. (10)). However, this bias
45 term is inconsequential in our linear regression setting and does not affect the predictive outcome.

46 Thus, the data-centering operation is the principal differentiator in our analysis. Specifically, linear
47 attention’s omission of centering makes it sensitive to shifts in the input mean, whereas linearized
48 attention remains robust under such transformations. We illustrate this phenomenon in Figure S1,
49 where we simulate a shift in the input mean at test time. The results demonstrate that linear attention
50 fails to recover Bayes-optimal performance under mean shift, indicating its limitations for in-context
51 learning in this setting. In contrast, linearized attention successfully compensates for the mean shift
52 and achieves Bayes-optimal performance as the number of context points l increases.

53 Therefore, in the presence of possible distributional shifts—particularly in the input mean—linearized
54 attention offers a more robust and theoretically grounded alternative to linear attention for in-context
55 learning.

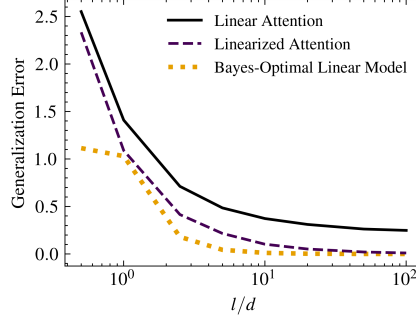


Figure S1: **Comparison of linear and linearized attention under a shift in input mean.** The plot illustrates the impact of a test-time shift in input mean on the performance of linear attention and linearized attention. While linear attention degrades under the distribution shift and fails to recover the Bayes-optimal performance, linearized attention remains robust and asymptotically matches the Bayes-optimal predictor as the number of context length l increases.

56 D Expanded form of linearized attention

57 Using block matrix notation, the prediction from the linearized attention model can be expanded as:

$$\hat{y}(\mathbf{Z}; \mathbf{V}, \mathbf{M}) = A_{d+1,l}, \quad (\text{S20})$$

$$= \frac{1}{l} \mathbf{V}_{d+1,:} \mathbf{Z} \left(\frac{(\mathbf{KZ})^T (\mathbf{QZ}_{:,l})}{\tau} - \frac{1}{l} \sum_{j=1}^l \frac{(\mathbf{KZ}_{:,j})^T (\mathbf{QZ}_{:,l})}{\tau} + \mathbf{1} \right), \quad (\text{S21})$$

$$= \frac{1}{l} [\mathbf{v}_{21}^T \quad \mathbf{v}_{22}] \mathbf{Z} \left(\frac{\mathbf{Z}^T \mathbf{M} \mathbf{Z}_{:,l}}{\tau} - \frac{1}{l} \sum_{j=1}^l \frac{(\mathbf{Z}_{:,j})^T \mathbf{M} \mathbf{Z}_{:,l}}{\tau} + \mathbf{1} \right), \quad (\text{S22})$$

$$= \frac{1}{l} [\mathbf{v}_{21}^T \quad \mathbf{v}_{22}] [\mathbf{X} \quad \mathbf{y}]^T \left(\frac{[\mathbf{X} \quad \mathbf{y}] \mathbf{M} [\mathbf{x}_l^T \quad 0]^T}{\tau} - \frac{1}{l} \sum_{j=1}^l \frac{[\mathbf{x}_j^T \quad y_j] \mathbf{M} [\mathbf{x}_l^T \quad 0]^T}{\tau} + \mathbf{1} \right), \quad (\text{S23})$$

$$= \frac{1}{l} [\mathbf{v}_{21}^T \quad \mathbf{v}_{22}] [\mathbf{X} \quad \mathbf{y}]^T \left(\frac{1}{\tau} [\mathbf{X} - \mathbf{1} \mathbf{s}_x^T \quad \mathbf{y} - s_y \mathbf{1}] \begin{bmatrix} \mathbf{M}_{11} & * \\ \mathbf{m}_{21}^T & * \end{bmatrix} [\mathbf{x}_l^T \quad 0]^T + \mathbf{1} \right), \quad (\text{S24})$$

$$= \frac{1}{l} [\mathbf{v}_{21}^T \quad \mathbf{v}_{22}] [\mathbf{X} \quad \mathbf{y}]^T \left(\frac{1}{\tau} (\mathbf{X} - \mathbf{1} \mathbf{s}_x^T) \mathbf{M}_{11} \mathbf{x}_l + \frac{1}{\tau} (\mathbf{y} - s_y \mathbf{1}) \mathbf{m}_{21}^T \mathbf{x}_l + \mathbf{1} \right), \quad (\text{S25})$$

$$= \frac{1}{l} (\mathbf{v}_{21}^T \mathbf{X}^T + \mathbf{v}_{22} \mathbf{y}^T) \left(\frac{1}{\tau} (\mathbf{X} - \mathbf{1} \mathbf{s}_x^T) \mathbf{M}_{11} \mathbf{x}_l + \frac{1}{\tau} (\mathbf{y} - s_y \mathbf{1}) \mathbf{m}_{21}^T \mathbf{x}_l + \mathbf{1} \right), \quad (\text{S26})$$

$$= \frac{1}{\tau} \left(\mathbf{v}_{21}^T \left(\frac{\mathbf{X}^T \mathbf{X}}{l} - \mathbf{s}_x \mathbf{s}_x^T \right) + \mathbf{v}_{22} \left(\frac{\mathbf{y}^T \mathbf{X}}{l} - s_y \mathbf{s}_x^T \right) \right) \mathbf{M}_{11} \mathbf{x}_l, \\ + \frac{1}{\tau} \left(\mathbf{v}_{21}^T \left(\frac{\mathbf{X}^T \mathbf{y}}{l} - s_y \mathbf{s}_x \right) + \mathbf{v}_{22} \left(\frac{\mathbf{y}^T \mathbf{y}}{l} - s_y^2 \right) \right) \mathbf{m}_{21}^T \mathbf{x}_l + \mathbf{v}_{21}^T \mathbf{s}_x + \mathbf{v}_{22} s_y, \quad (\text{S27})$$

$$= \frac{1}{\tau} (\mathbf{v}_{21}^T \mathbf{C}_{xx} + \mathbf{v}_{22} \mathbf{C}_{xy}^T) \mathbf{M}_{11} \mathbf{x}_l + \frac{1}{\tau} (\mathbf{v}_{21}^T \mathbf{C}_{xy} + \mathbf{v}_{22} \mathbf{C}_{yy}) \mathbf{m}_{21}^T \mathbf{x}_l + \mathbf{v}_{21}^T \mathbf{s}_x + \mathbf{v}_{22} s_y, \quad (\text{S28})$$

$$= \frac{1}{\tau} ((\mathbf{v}_{21}^T \mathbf{C}_{xx} + \mathbf{v}_{22} \mathbf{C}_{xy}^T) \mathbf{M}_{11} + (\mathbf{v}_{21}^T \mathbf{C}_{xy} + \mathbf{v}_{22} \mathbf{C}_{yy}) \mathbf{m}_{21}^T) \mathbf{x}_l + \mathbf{v}_{21}^T \mathbf{s}_x + \mathbf{v}_{22} s_y, \quad (\text{S29})$$

58 where the summary statistics are defined as:

$$\mathbf{s}_x := \frac{1}{l} \sum_{i=1}^l \mathbf{x}_i, \quad s_y := \frac{1}{l} \sum_{i=1}^{l-1} y_i, \\ \mathbf{C}_{xx} := \frac{1}{l} \sum_{i=1}^l \mathbf{x}_i \mathbf{x}_i^T - \mathbf{s}_x \mathbf{s}_x^T, \quad \mathbf{C}_{xy} := \frac{1}{l} \sum_{i=1}^{l-1} y_i \mathbf{x}_i - s_y \mathbf{s}_x, \quad \mathbf{C}_{yy} := \frac{1}{l} \sum_{i=1}^{l-1} y_i^2 - s_y^2.$$

59 Then, we define

$$\hat{\mathbf{w}}_{Att}(\mathbf{C}_{xx}, \mathbf{C}_{xy}, \mathbf{C}_{yy}; \mathbf{M}, \mathbf{V}) = \mathbf{M}_{11}^T (\mathbf{C}_{xx} \mathbf{v}_{21} + v_{22} \mathbf{C}_{xy}) + (\mathbf{v}_{21}^T \mathbf{C}_{xy} + v_{22} \mathbf{C}_{yy}) \mathbf{m}_{21}, \quad (\text{S30})$$

$$b_{Att}(\mathbf{s}_x, s_y; \mathbf{V}) = \mathbf{v}_{21}^T \mathbf{s}_x + v_{22} s_y, \quad (\text{S31})$$

60 which allows us to write

$$\hat{y}(\mathbf{Z}; \mathbf{V}, \mathbf{M}) = \frac{1}{\tau} \hat{\mathbf{w}}_{Att}(\mathbf{C}_{xx}, \mathbf{C}_{xy}, \mathbf{C}_{yy}; \mathbf{M}, \mathbf{V})^T \mathbf{x}_l + b_{Att}(\mathbf{s}_x, s_y; \mathbf{V}). \quad (\text{S32})$$

61 **E Derivation of the pretraining for ICL by mimicking the Bayes-optimal** 62 **estimator**

63 Here, we derive the pretraining of the linearized attention model by mimicking the Bayes-optimal
64 ridge estimator (9). Recall that the prediction of the linearized attention model is

$$\hat{y}(\mathbf{Z}; \mathbf{V}, \mathbf{M}) = \frac{1}{\tau} \hat{\mathbf{w}}_{Att}(\mathbf{C}_{xx}, \mathbf{C}_{xy}, \mathbf{C}_{yy}; \mathbf{M}, \mathbf{V})^T \mathbf{x}_l + b_{Att}(\mathbf{s}_x, s_y; \mathbf{V}), \quad (\text{S33})$$

65 which is derived in Appendix D. Furthermore, the Bayes-optimal ridge regression model's prediction
66 is

$$\hat{y}_{Bayes} = \hat{\mathbf{w}}_{Bayes}^T \mathbf{x}_l. \quad (\text{S34})$$

67 Therefore, we select the parameters \mathbf{M} and \mathbf{V} such that

$$\hat{\mathbf{w}}_{Att}(\mathbf{C}_{xx}, \mathbf{C}_{xy}, \mathbf{C}_{yy}; \mathbf{M}, \mathbf{V}) \approx \hat{\mathbf{w}}_{Bayes}, \quad b_{Att}(\mathbf{s}_x, s_y; \mathbf{V}) \approx 0, \quad (\text{S35})$$

68 which makes the prediction of the linearized attention model approximately equal to that of the
69 Bayes-optimal regression. Furthermore, we consider $\tau = 1$ for the pretraining. Let's first focus on
70 $\hat{\mathbf{w}}_{Att}(\mathbf{C}_{xx}, \mathbf{C}_{xy}, \mathbf{C}_{yy}; \mathbf{M}, \mathbf{V})$ as follows

$$\begin{aligned} \hat{\mathbf{w}}_{Att}(\mathbf{C}_{xx}, \mathbf{C}_{xy}, \mathbf{C}_{yy}; \mathbf{M}, \mathbf{V}) &= (\mathbf{M}_{11}^T (\mathbf{C}_{xx} \mathbf{v}_{21} + v_{22} \mathbf{C}_{xy}) + (\mathbf{v}_{21}^T \mathbf{C}_{xy} + v_{22} \mathbf{C}_{yy}) \mathbf{m}_{21}), \end{aligned} \quad (\text{S36})$$

$$= \left(\mathbf{M}_{11}^T \left(\frac{\bar{\mathbf{X}}^T \bar{\mathbf{X}}}{l} \mathbf{v}_{21} + v_{22} \frac{\bar{\mathbf{X}}^T \bar{\mathbf{y}}}{l} \right) + \left(\mathbf{v}_{21}^T \frac{\bar{\mathbf{X}}^T \bar{\mathbf{y}}}{l} + v_{22} \frac{\bar{\mathbf{y}}^T \bar{\mathbf{y}}}{l} \right) \mathbf{m}_{21} \right). \quad (\text{S37})$$

71 To reach the last line, we use the fact that $\mathbf{C}_{xx} := \mathbf{X}^T \mathbf{X} / l - \mathbf{s}_x \mathbf{s}_x^T = \bar{\mathbf{X}}^T \bar{\mathbf{X}} / l$, $\mathbf{C}_{xy} := \mathbf{X}^T \mathbf{y} / l -$
72 $s_y \mathbf{s}_x = \bar{\mathbf{X}}^T \bar{\mathbf{y}} / l$ and $\mathbf{C}_{yy} := \mathbf{y}^T \mathbf{y} / l - s_y^2 = \bar{\mathbf{y}}^T \bar{\mathbf{y}} / l$, where $\bar{\mathbf{X}} := \mathbf{X} - \mathbf{s}_x \mathbf{s}_x^T$ and $\bar{\mathbf{y}} := \mathbf{y} - s_y \mathbf{1}$ denote centered input
73 matrix and centered label vector. Now, recall that the Bayes-optimal ridge estimator is

$$\hat{\mathbf{w}}_{Bayes} = \left(\frac{\bar{\mathbf{X}}^T \bar{\mathbf{X}}}{\sigma^2} + \Sigma_w^{-1} \right)^{-1} \left(\frac{\bar{\mathbf{X}}^T \bar{\mathbf{y}}}{\sigma^2} + \Sigma_w^{-1} \boldsymbol{\mu}_w \right), \quad (\text{S38})$$

74 as derived in Appendix A. Looking at equations (S38) and (S37) together, we can see that setting the
75 parameters as follows would make $\hat{\mathbf{w}}_{Att} = \hat{\mathbf{w}}_{Bayes}$ hold

$$\mathbf{M}_{11} = \frac{l}{\sigma^2} \left(\frac{\bar{\mathbf{X}}^T \bar{\mathbf{X}}}{\sigma^2} + \Sigma_w^{-1} \right)^{-1}, \quad \mathbf{v}_{21} = \frac{\sigma^2}{l} \left(\frac{\bar{\mathbf{X}}^T \bar{\mathbf{X}}}{l} \right)^{-1} \Sigma_w^{-1} \boldsymbol{\mu}_w, \quad \mathbf{m}_{21} = \mathbf{0}, \quad v_{22} = 1. \quad (\text{S39})$$

76 However, while Bayes-optimal estimator $\hat{\mathbf{w}}_{Bayes}$ is different for each sample, the attention model
77 should be pretrained and fixed. Thus, we replace $\bar{\mathbf{X}}^T \bar{\mathbf{X}}$ in (S39) with $\hat{\mathbf{X}}^T \hat{\mathbf{X}} / m$ as follows, where
78 $\hat{\mathbf{X}} \in \mathbb{R}^{ml \times d}$ is the centred input matrix including all the (pre)training data consisting of ml samples.

$$\mathbf{M}_{11} = \frac{l}{\sigma^2} \left(\frac{\hat{\mathbf{X}}^T \hat{\mathbf{X}}}{m \sigma^2} + \Sigma_w^{-1} \right)^{-1}, \quad \mathbf{v}_{21} = \frac{\sigma^2}{l} \left(\frac{\hat{\mathbf{X}}^T \hat{\mathbf{X}}}{ml} \right)^{-1} \Sigma_w^{-1} \boldsymbol{\mu}_w, \quad \mathbf{m}_{21} = \mathbf{0}, \quad v_{22} = 1. \quad (\text{S40})$$

79 In practice, the variance of noise σ^2 , the mean $\boldsymbol{\mu}_w$, and covariance Σ_w of the task vectors are
80 unknown. Yet, we can use their estimates based on the (pre)training data.

81 Now, we can focus on making $b_{Att}(\mathbf{s}_x, \mathbf{s}_y; \mathbf{V}) \approx 0$ hold as follows

$$b_{Att}(\mathbf{s}_x, \mathbf{s}_y; \mathbf{V}) = \mathbf{v}_{21}^T \mathbf{s}_x + v_{22} s_y, \quad (\text{S41})$$

82 where \mathbf{s}_x and s_y are based on data so we have no control over them. Instead, by using Assumptions
83 3.1 and 3.2, we can choose \mathbf{v}_{21} and v_{22} such that $b_{Att} \rightarrow 0$ as $l, d \rightarrow \infty$. Note that Assumption
84 3.1 makes $\mathbf{v}_{21}^T \mathbf{s}_x + v_{22} s_y$ bounded with high probability for \mathbf{v}_{21} and v_{22} given in (S40). Therefore,
85 multiplying \mathbf{v}_{21}, v_{22} given in (S40) with $1/d$ would make $b_{Att} \rightarrow 0$ as $d \rightarrow \infty$. To fix the impact of
86 the multiplication for $\hat{\mathbf{w}}_{Att}$, we can multiply \mathbf{M}_{11} with d as well. So, by applying the mentioned
87 multiplications, we reach the following pretrained parameters mimicking the Bayes-optimal regression
88 model

$$\mathbf{M}_{11} = \frac{dl}{\sigma^2} \left(\frac{\hat{\mathbf{X}}^T \hat{\mathbf{X}}}{m\sigma^2} + \Sigma_w^{-1} \right)^{-1}, \quad \mathbf{v}_{21} = \frac{\sigma^2}{dl} \left(\frac{\hat{\mathbf{X}}^T \hat{\mathbf{X}}}{ml} \right)^{-1} \Sigma_w^{-1} \boldsymbol{\mu}_w, \quad \mathbf{m}_{21} = \mathbf{0}, \quad v_{22} = \frac{1}{d}. \quad (\text{S42})$$

89 F Characterization of generalization error for ICL under distribution shift

90 Here, we characterize the generalization error for in-context learning under distribution shift, given
91 that \mathbf{M} and \mathbf{V} are pretrained and fixed. So, the impact of pretraining distribution \mathcal{D}^{train} is captured
92 by \mathbf{M} and \mathbf{V} . Suppose that \mathcal{D}^{test} denotes the test distribution. To avoid additional notations, here,
93 we again use $\boldsymbol{\mu}_x, \boldsymbol{\mu}_w, \Sigma_x, \Sigma_w, \sigma^2$ to denote means and covariances for input and task vectors and
94 noise variance for the inference (test). However, note that these can be different from those used for
95 pretraining. We begin studying the generalization error defined in (8) as follows

$$\mathcal{G}(\mathbf{V}, \mathbf{M}) := \mathbb{E}_{(\mathbf{Z}, y_l) \sim \mathcal{D}^{test}} \left[(y_l - \hat{y}(\mathbf{Z}; \mathbf{V}, \mathbf{M}))^2 \right], \quad (\text{S43})$$

$$= \mathbb{E}_{(\mathbf{Z}, y_l) \sim \mathcal{D}^{test}} \left[\left(\frac{1}{\tau} \hat{\mathbf{w}}_{Att}(\mathbf{C}_{xx}, \mathbf{C}_{xy}, \mathbf{C}_{yy}; \mathbf{M}, \mathbf{V})^T \mathbf{x}_l + b_{Att}(\mathbf{s}_x, \mathbf{s}_y; \mathbf{V}) - y_l \right)^2 \right], \quad (\text{S44})$$

$$= \mathbb{E}_{(\mathbf{Z}, y_l) \sim \mathcal{D}^{test}} \left[\left(\frac{1}{\tau} (\mathbf{M}_{11}^T (\mathbf{C}_{xx} \mathbf{v}_{21} + v_{22} \mathbf{C}_{xy}))^T \mathbf{x}_l - y_l \right)^2 \right], \quad (\text{S45})$$

96 where we use the parameters from pretraining (S42) together with Assumptions 3.1 and 3.2 to reach
97 the last line. Then,

$$\mathcal{G}(\mathbf{V}, \mathbf{M}) = \mathbb{E}_{(\mathbf{Z}, y_l) \sim \mathcal{D}^{test}} \left[\left(\frac{1}{\tau} (\mathbf{M}_{11}^T (\mathbf{C}_{xx} \mathbf{v}_{21} + v_{22} \mathbf{C}_{xy}))^T \mathbf{x}_l - y_l \right)^2 \right], \quad (\text{S46})$$

$$= \mathbb{E} \left[\left(\frac{1}{\tau} \left(\mathbf{M}_{11}^T \left(\frac{1}{l} \sum_{i \leq l} \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^T \mathbf{v}_{21} + v_{22} \frac{1}{l} \sum_{i \leq l-1} \bar{\mathbf{x}}_i (\bar{\mathbf{x}}_i^T \mathbf{w} + \epsilon_i) \right) \right)^T \mathbf{x}_l - \mathbf{w}^T \mathbf{x}_l - \epsilon_l \right)^2 \right], \quad (\text{S47})$$

$$= \mathbb{E} \left[\left(\frac{1}{\tau} \left(\mathbf{M}_{11}^T \left(\frac{1}{l} \sum_{i \leq l} \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^T \mathbf{v}_{21} + v_{22} \frac{1}{l} \sum_{i \leq l-1} \bar{\mathbf{x}}_i (\bar{\mathbf{x}}_i^T \mathbf{w} + \epsilon_i) \right) \right)^T \mathbf{x}_l - \mathbf{w}^T \mathbf{x}_l \right)^2 \right] + \sigma^2 \quad (\text{S48})$$

98 where $\bar{\mathbf{x}}_i := \mathbf{x}_i - \mathbf{s}_x = \mathbf{x}_i - \frac{1}{l} \sum_{i \leq l} \mathbf{x}_i$ and we use $\epsilon_l \sim \mathcal{N}(0, \sigma^2)$ to reach the final line. We
99 continue by defining

$$\mathbf{w}_{diff} := \frac{1}{\tau} \mathbf{M}_{11}^T \left(\frac{1}{l} \sum_{i \leq l-1} \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^T \mathbf{v}_{21} + v_{22} \frac{1}{l} \sum_{i \leq l-1} \bar{\mathbf{x}}_i (\bar{\mathbf{x}}_i^T \mathbf{w} + \epsilon_i) \right) - \mathbf{w}, \quad (\text{S49})$$

100 which allows us to write

$$\mathcal{G}(\mathbf{V}, \mathbf{M}) = \mathbb{E} \left[(\mathbf{w}_{diff}^T \mathbf{x}_l)^2 \right] + \sigma^2, \quad (\text{S50})$$

$$= \mathbb{E} \left[\mathbf{w}_{diff}^T \mathbb{E}_{\mathbf{x}_l} [\mathbf{x}_l \mathbf{x}_l^T] \mathbf{w}_{diff} \right] + \sigma^2, \quad (\text{S51})$$

$$= \mathbb{E} \left[\mathbf{w}_{diff}^T (\boldsymbol{\mu}_x \boldsymbol{\mu}_x^T + \Sigma_x) \mathbf{w}_{diff} \right] + \sigma^2, \quad (\text{S52})$$

101 by the law of total expectation since \mathbf{w}_{diff} is independent of \mathbf{x}_l . Note that when writing (S50), we
 102 safely ignore terms with $(1/l)\bar{\mathbf{x}}_l\bar{\mathbf{x}}_l^T\mathbf{v}_{21}$ in (S48) since they vanish as $l \rightarrow \infty$ by Assumptions 3.1-3.2
 103 and 4.5. Letting $\mathbf{A} := \boldsymbol{\mu}_x\boldsymbol{\mu}_x^T + \boldsymbol{\Sigma}_x$, we write

$$\mathcal{G}(\mathbf{V}, \mathbf{M}) = \mathbb{E}[\mathbf{w}_{diff}^T \mathbf{A} \mathbf{w}_{diff}] + \sigma^2, \quad (\text{S53})$$

$$= \mathbb{E}[\text{Tr}(\mathbf{w}_{diff}^T \mathbf{A} \mathbf{w}_{diff})] + \sigma^2, \quad (\text{S54})$$

$$= \mathbb{E}[\text{Tr}(\mathbf{A} \mathbf{w}_{diff} \mathbf{w}_{diff}^T)] + \sigma^2, \quad (\text{S55})$$

$$= \text{Tr}(\mathbf{A} \mathbb{E}[\mathbf{w}_{diff} \mathbf{w}_{diff}^T]) + \sigma^2, \quad (\text{S56})$$

104 where we first apply the cyclic property of trace and then use the linearity of expectation and trace to
 105 reach the last line. Now, we need to calculate $\mathbb{E}[\mathbf{w}_{diff} \mathbf{w}_{diff}^T]$, for which we first take the expectation
 106 over \mathbf{w} . To do so, we rewrite \mathbf{w}_{diff} as

$$\mathbf{w}_{diff} = \underbrace{\frac{1}{\tau} \mathbf{M}_{11}^T \left(\frac{1}{l} \sum_{i \leq l-1} \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^T \mathbf{v}_{21} + v_{22} \frac{1}{l} \sum_{i \leq l-1} \bar{\mathbf{x}}_i \epsilon_i \right)}_{\mathbf{e}} + \underbrace{\left(\frac{v_{22}}{\tau} \mathbf{M}_{11}^T \frac{1}{l} \sum_{i \leq l-1} \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^T - \mathbf{I} \right)}_{\mathbf{D}} \mathbf{w}, \quad (\text{S57})$$

$$= \mathbf{e} + \mathbf{D} \mathbf{w}, \quad (\text{S58})$$

107 where we define

$$\mathbf{e} := \frac{1}{\tau} \mathbf{M}_{11}^T \left(\frac{1}{l} \sum_{i \leq l-1} \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^T \mathbf{v}_{21} + v_{22} \frac{1}{l} \sum_{i \leq l-1} \bar{\mathbf{x}}_i \epsilon_i \right), \quad (\text{S59})$$

$$\mathbf{D} := \left(\frac{v_{22}}{\tau} \mathbf{M}_{11}^T \frac{1}{l} \sum_{i \leq l-1} \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^T - \mathbf{I} \right). \quad (\text{S60})$$

108 Since \mathbf{e} and \mathbf{D} are independent of \mathbf{w} , we can easily calculate $\mathbb{E}_{\mathbf{w}}[\mathbf{w}_{diff} \mathbf{w}_{diff}^T]$ as follows

$$\mathbb{E}[\mathbb{E}_{\mathbf{w}}[\mathbf{w}_{diff} \mathbf{w}_{diff}^T]] = \mathbb{E}[\mathbb{E}_{\mathbf{w}}[(\mathbf{e} + \mathbf{D} \mathbf{w})(\mathbf{e} + \mathbf{D} \mathbf{w})^T]], \quad (\text{S61})$$

$$= \mathbb{E}[\mathbf{e} \mathbf{e}^T] + \mathbb{E}[\mathbf{e} \boldsymbol{\mu}_w^T \mathbf{D}^T] + \mathbb{E}[\mathbf{D} \boldsymbol{\mu}_w \mathbf{e}^T] + \mathbb{E}[\mathbf{D}(\boldsymbol{\mu}_x \boldsymbol{\mu}_x^T + \boldsymbol{\Sigma}_w) \mathbf{D}^T], \quad (\text{S62})$$

$$= \mathbb{E}[\mathbf{e} \mathbf{e}^T] + \mathbb{E}[\mathbf{D} \boldsymbol{\mu}_w \mathbf{e}^T]^T + \mathbb{E}[\mathbf{D} \boldsymbol{\mu}_w \mathbf{e}^T] + \mathbb{E}[\mathbf{D} \mathbf{B} \mathbf{D}^T], \quad (\text{S63})$$

109 where we first apply the law of total expectation, then take the expectation over \mathbf{w} and finally, we
 110 define $\mathbf{B} := \boldsymbol{\mu}_x \boldsymbol{\mu}_x^T + \boldsymbol{\Sigma}_w$ to reach the last line. Note that $\boldsymbol{\mu}_w$ and \mathbf{B} are fixed while \mathbf{e} and \mathbf{D} are
 111 random in the last line. Therefore, we are required to calculate the three expectations that appeared in
 112 (S63).

113 Before getting into the calculations of the aforementioned expectations, we provide the following
 114 lemma that is useful for the calculation of the expectations.

115 **Lemma F.1.** *Let $\bar{\mathbf{x}} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where $\bar{\mathbf{x}} \in \mathbb{R}^d$. Let $\bar{\mathbf{x}}_i$ be $l-1$ independent samples of $\bar{\mathbf{x}}$ for
 116 $i = 1, \dots, l-1$. Furthermore, let \mathbf{A} be a fixed $d \times d$ matrix. Then, the following holds*

$$\mathbb{E} \left[\left(\frac{1}{l} \sum_{i \leq l-1} \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^T \right) \mathbf{A} \left(\frac{1}{l} \sum_{i \leq l-1} \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^T \right) \right] = \frac{l-1}{l} \boldsymbol{\Sigma} \mathbf{A} \boldsymbol{\Sigma} + \frac{1}{l} \boldsymbol{\Sigma} \mathbf{A}^T \boldsymbol{\Sigma} + \frac{1}{l} \text{Tr}(\mathbf{A} \boldsymbol{\Sigma}) \boldsymbol{\Sigma}. \quad (\text{S64})$$

117 *Proof.* This is proven by using Isserlis' theorem [12] in Appendix G. □

118 Note that our inputs $\bar{\mathbf{x}}_i$ are centered, i.e., $\bar{\mathbf{x}}_i = \mathbf{x}_i - \frac{1}{l} \sum_{i \leq l} \mathbf{x}_i$, so their distribution is $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_x)$ as
 119 $l \rightarrow \infty$. Therefore, Lemma F.1 is directly applicable in our setting.

120 Next, we start the calculations of the expectations in (S63) with $\mathbb{E}[ee^T]$ as follows

$$\mathbb{E}[ee^T] = \frac{1}{\tau^2} \mathbf{M}_{11}^T \mathbb{E} \left[\left(\frac{1}{l} \sum_{i \leq l-1} \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^T \mathbf{v}_{21} + v_{22} \frac{1}{l} \sum_{i \leq l-1} \bar{\mathbf{x}}_i \epsilon_i \right) \cdot \left(\frac{1}{l} \sum_{i \leq l-1} \mathbf{v}_{21}^T \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^T + v_{22} \frac{1}{l} \sum_{i \leq l-1} \bar{\mathbf{x}}_i^T \epsilon_i \right) \right] \mathbf{M}_{11}, \quad (\text{S65})$$

$$= \frac{1}{\tau^2} \mathbf{M}_{11}^T \left(\mathbb{E} \left[\left(\frac{1}{l} \sum_{i \leq l-1} \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^T \mathbf{v}_{21} \right) \left(\frac{1}{l} \sum_{i \leq l-1} \mathbf{v}_{21}^T \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^T \right) + \left(v_{22} \frac{1}{l} \sum_{i \leq l-1} \bar{\mathbf{x}}_i \epsilon_i \right) \left(v_{22} \frac{1}{l} \sum_{i \leq l-1} \bar{\mathbf{x}}_i^T \epsilon_i \right) \right] \right) \mathbf{M}_{11}, \quad (\text{S66})$$

$$= \frac{1}{\tau^2} \mathbf{M}_{11}^T \left(\mathbb{E} \left[\left(\frac{1}{l} \sum_{i \leq l-1} \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^T \right) \mathbf{v}_{21} \mathbf{v}_{21}^T \left(\frac{1}{l} \sum_{i \leq l-1} \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^T \right) + \left(v_{22}^2 \frac{\sigma^2}{l^2} \sum_{i \leq l-1} \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^T \right) \right] \right) \mathbf{M}_{11}, \quad (\text{S67})$$

$$= \frac{1}{\tau^2} \mathbf{M}_{11}^T \left(\Sigma_x \mathbf{v}_{21} \mathbf{v}_{21}^T \Sigma_x + \frac{1}{l} \text{Tr}(\mathbf{v}_{21} \mathbf{v}_{21}^T \Sigma_x) \Sigma_x + v_{22}^2 \frac{\sigma^2(l-1)}{l^2} \Sigma_x \right) \mathbf{M}_{11}, \quad (\text{S68})$$

$$= \frac{1}{\tau^2} \mathbf{M}_{11}^T \left(v_{22}^2 \frac{\sigma^2}{l} \Sigma_x \right) \mathbf{M}_{11}, \quad (\text{S69})$$

121 where we first use the independence of the random variables and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ to simplify the
 122 equation. Then, we apply Lemma F.1 and use the fact that $\mathbb{E}[\bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^T] = \Sigma_x$ to get the penultimate line.
 123 Finally, we drop the vanishing terms and simplify the result using Assumptions 3.1-3.2 and 4.5 in
 124 order to reach the last line.

125 We continue with the calculation of $\mathbb{E}[\mathbf{D}\mu_w e^T]$ as

$$\mathbb{E}[\mathbf{D}\mu_w e^T] = \frac{1}{\tau} \mathbb{E} \left[\left(\frac{v_{22}}{\tau} \mathbf{M}_{11}^T \frac{1}{l} \sum_{i \leq l-1} \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^T - \mathbf{I} \right) \mu_w \left(\frac{1}{l} \sum_{i \leq l-1} \mathbf{v}_{21}^T \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^T + v_{22} \frac{1}{l} \sum_{i \leq l-1} \bar{\mathbf{x}}_i^T \epsilon_i \right) \right] \mathbf{M}_{11}, \quad (\text{S70})$$

$$= \frac{1}{\tau} \mathbb{E} \left[\left(\frac{v_{22}}{\tau} \mathbf{M}_{11}^T \frac{1}{l} \sum_{i \leq l-1} \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^T - \mathbf{I} \right) \mu_w \left(\frac{1}{l} \sum_{i \leq l-1} \mathbf{v}_{21}^T \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^T \right) \right] \mathbf{M}_{11}, \quad (\text{S71})$$

$$= \frac{1}{\tau} \frac{v_{22}}{\tau} \mathbf{M}_{11}^T \mathbb{E} \left[\left(\frac{1}{l} \sum_{i \leq l-1} \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^T \right) \mu_w \mathbf{v}_{21}^T \left(\frac{1}{l} \sum_{i \leq l-1} \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^T \right) \right] \mathbf{M}_{11} - \mu_w \mathbf{v}_{21}^T \mathbb{E} \left[\frac{1}{l} \sum_{i \leq l-1} \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^T \right] \mathbf{M}_{11}, \quad (\text{S72})$$

$$= \frac{1}{\tau} \frac{v_{22}}{\tau} \mathbf{M}_{11}^T \left(\Sigma_x \mu_w \mathbf{v}_{21}^T \Sigma_x + \frac{1}{l} \Sigma_x \mathbf{v}_{21} \mu_w^T \Sigma_x + \frac{1}{l} \text{Tr}(\mu_w \mathbf{v}_{21}^T \Sigma_x) \Sigma_x \right) \mathbf{M}_{11} - \frac{1}{\tau} \frac{l-1}{l} \mu_w \mathbf{v}_{21}^T \Sigma_x \mathbf{M}_{11}, \quad (\text{S73})$$

$$= \frac{v_{22}}{\tau^2} \mathbf{M}_{11}^T \left(\Sigma_x \mu_w \mathbf{v}_{21}^T \Sigma_x + \frac{1}{l} \text{Tr}(\mu_w \mathbf{v}_{21}^T \Sigma_x) \Sigma_x \right) \mathbf{M}_{11} - \frac{1}{\tau} \mu_w \mathbf{v}_{21}^T \Sigma_x \mathbf{M}_{11}, \quad (\text{S74})$$

126 where we again first use the independence of the random variables and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Then, we
 127 apply basic algebraic manipulations. To reach the penultimate line, we utilize Lemma F.1 together
 128 with the fact that $\mathbb{E}[\bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^T] = \Sigma_x$. Using Assumptions 3.1-3.2 and 4.5, we reach the last line.

129 Finally, we calculate $\mathbb{E}[DBD^T]$ as follows

$$\mathbb{E}[DBD^T] = \mathbb{E}\left[\left(\frac{v_{22}}{\tau}M_{11}^T\frac{1}{l}\sum_{i\leq l-1}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T - \mathbf{I}\right)\mathbf{B}\left(\frac{v_{22}}{\tau}\frac{1}{l}\sum_{i\leq l-1}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^TM_{11} - \mathbf{I}\right)\right], \quad (\text{S75})$$

$$= \mathbb{E}\left[\left(\frac{v_{22}}{\tau}M_{11}^T\frac{1}{l}\sum_{i\leq l-1}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T\right)\mathbf{B}\left(\frac{v_{22}}{\tau}\frac{1}{l}\sum_{i\leq l-1}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^TM_{11}\right)\right] - \mathbb{E}\left[\left(\frac{v_{22}}{\tau}M_{11}^T\frac{1}{l}\sum_{i\leq l-1}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T\right)\mathbf{B}\right] - \mathbb{E}\left[\mathbf{B}\left(\frac{v_{22}}{\tau}\frac{1}{l}\sum_{i\leq l-1}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^TM_{11}\right)\right] + \mathbf{B}, \quad (\text{S76})$$

$$= \frac{v_{22}^2}{\tau^2}M_{11}^T\mathbb{E}\left[\left(\frac{1}{l}\sum_{i\leq l-1}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T\right)\mathbf{B}\left(\frac{1}{l}\sum_{i\leq l-1}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T\right)\right]M_{11} - \frac{v_{22}}{\tau}M_{11}^T\mathbb{E}\left[\left(\frac{1}{l}\sum_{i\leq l-1}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T\right)\right]\mathbf{B} - \frac{v_{22}}{\tau}\mathbf{B}\mathbb{E}\left[\left(\frac{1}{l}\sum_{i\leq l-1}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T\right)\right]M_{11} + \mathbf{B}, \quad (\text{S77})$$

$$= \frac{v_{22}^2}{\tau^2}M_{11}^T\left(\Sigma_x\mathbf{B}\Sigma_x + \frac{1}{l}\text{Tr}(\mathbf{B}\Sigma_x)\Sigma_x\right)M_{11} - \frac{v_{22}}{\tau}\frac{l-1}{l}M_{11}^T\Sigma_x\mathbf{B} - \frac{v_{22}}{\tau}\frac{l-1}{l}\mathbf{B}\Sigma_xM_{11} + \mathbf{B}, \quad (\text{S78})$$

$$= \frac{v_{22}^2}{\tau^2}M_{11}^T\left(\Sigma_x\mathbf{B}\Sigma_x + \frac{1}{l}\text{Tr}(\mathbf{B}\Sigma_x)\Sigma_x\right)M_{11} - \frac{v_{22}}{\tau}M_{11}^T\Sigma_x\mathbf{B} - \frac{v_{22}}{\tau}\mathbf{B}\Sigma_xM_{11} + \mathbf{B}, \quad (\text{S79})$$

130 where we first do basic algebraic manipulations. Then, we use Lemma F.1 and $\mathbb{E}[\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T] = \Sigma_x$ to get
131 the penultimate line. For the final line, we utilize $l \rightarrow \infty$ by Assumption 3.2.

132 Putting the found expectation results into (S63), we get

$$\mathbb{E}[\mathbb{E}_w[\mathbf{w}_{diff}\mathbf{w}_{diff}^T]] = \mathbb{E}[\mathbf{e}\mathbf{e}^T] + \mathbb{E}[\mathbf{D}\boldsymbol{\mu}_w\mathbf{e}^T]^T + \mathbb{E}[\mathbf{D}\boldsymbol{\mu}_w\mathbf{e}^T] + \mathbb{E}[DBD^T], \quad (\text{S80})$$

$$= \frac{1}{\tau^2}M_{11}^T\mathbf{F}_1M_{11} - \frac{1}{\tau}\mathbf{F}_2M_{11} + \frac{1}{\tau}M_{11}^T\mathbf{F}_2^T + \mathbf{B}. \quad (\text{S81})$$

133 where matrices \mathbf{F}_1 and \mathbf{F}_2 are defined as

$$\mathbf{F}_1 := v_{22}^2\frac{\sigma^2}{l}\Sigma_x + v_{22}\left(\Sigma_x\boldsymbol{\mu}_w\mathbf{v}_{21}^T\Sigma_x + \frac{1}{l}\text{Tr}(\boldsymbol{\mu}_w\mathbf{v}_{21}^T\Sigma_x)\Sigma_x\right) \quad (\text{S82})$$

$$+ v_{22}\left(\Sigma_x\boldsymbol{\mu}_w\mathbf{v}_{21}^T\Sigma_x + \frac{1}{l}\text{Tr}(\boldsymbol{\mu}_w\mathbf{v}_{21}^T\Sigma_x)\Sigma_x\right)^T + v_{22}^2\left(\Sigma_x\mathbf{B}\Sigma_x + \frac{1}{l}\text{Tr}(\mathbf{B}\Sigma_x)\Sigma_x\right), \\ = \left(\Sigma_x\hat{\mathbf{B}} + \left(v_{22}^2\frac{\sigma^2}{l} + \frac{1}{l}\text{Tr}(\hat{\mathbf{B}}\Sigma_x)\right)\mathbf{I}\right)\Sigma_x, \quad (\text{S83})$$

$$\mathbf{F}_2 := \boldsymbol{\mu}_w\mathbf{v}_{21}^T\Sigma_x + v_{22}\mathbf{B}\Sigma_x = (\boldsymbol{\mu}_w\mathbf{v}_{21}^T + v_{22}\mathbf{B})\Sigma_x, \quad (\text{S84})$$

134 with $\hat{\mathbf{B}} := v_{22}\boldsymbol{\mu}_w\mathbf{v}_{21}^T + v_{22}\mathbf{v}_{21}\boldsymbol{\mu}_w^T + v_{22}^2\mathbf{B}$.

135 Going back to generalization error in (S56), we have

$$\mathcal{G}(\mathbf{V}, \mathbf{M}) = \text{Tr}(\mathbf{A}\mathbb{E}[\mathbf{w}_{diff}\mathbf{w}_{diff}^T]) + \sigma^2, \quad (\text{S85})$$

$$= \text{Tr}\left(\mathbf{A}\left(\frac{1}{\tau^2}M_{11}^T\mathbf{F}_1M_{11} - \frac{1}{\tau}\mathbf{F}_2M_{11} + \frac{1}{\tau}M_{11}^T\mathbf{F}_2^T + \mathbf{B}\right)\right) + \sigma^2, \quad (\text{S86})$$

136 where $\mathbf{F}_1 = \left(\Sigma_x\hat{\mathbf{B}} + \frac{1}{l}\left(v_{22}^2\sigma^2 + \text{Tr}(\hat{\mathbf{B}}\Sigma_x)\right)\mathbf{I}\right)\Sigma_x$, and $\mathbf{F}_2 = (\boldsymbol{\mu}_w\mathbf{v}_{21}^T + v_{22}\mathbf{B})\Sigma_x$. Further-
137 more, $\hat{\mathbf{B}}$ is defined as $\hat{\mathbf{B}} := v_{22}\boldsymbol{\mu}_w\mathbf{v}_{21}^T + v_{22}\mathbf{v}_{21}\boldsymbol{\mu}_w^T + v_{22}^2\mathbf{B}$.

138 G Proof of Lemma F.1

139 We first restate the lemma as follows.

140 Let $\bar{\mathbf{x}} \sim \mathcal{N}(0, \Sigma)$, where $\bar{\mathbf{x}} \in \mathbb{R}^d$. Let $\bar{\mathbf{x}}_i$ be l independent samples of $\bar{\mathbf{x}}$ for $i = 1, \dots, l$. Let \mathbf{A} be a
141 fixed $d \times d$ matrix. Then, the following holds

$$\mathbb{E} \left[\left(\frac{1}{l} \sum_{i=1}^l \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^T \right) \mathbf{A} \left(\frac{1}{l} \sum_{i=1}^l \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^T \right) \right] = \Sigma \mathbf{A} \Sigma + \frac{1}{l} \Sigma \mathbf{A}^T \Sigma + \frac{1}{l} \text{Tr}(\mathbf{A} \Sigma) \Sigma. \quad (\text{S87})$$

142 *Proof.* Let $\mathbf{S}_x = \frac{1}{l} \sum_{i=1}^l \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^T$. First, note that $E[\bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^T] = \Sigma$ since $\bar{\mathbf{x}}_i \sim \mathcal{N}(0, \Sigma)$.

143 Thus, $\mathbb{E}[\mathbf{S}_x] = \frac{1}{l} \sum_{i=1}^l \mathbb{E}[\bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^T] = \frac{1}{l} \sum_{i=1}^l \Sigma = \Sigma$. We have

$$\mathbf{S}_x \mathbf{A} \mathbf{S}_x = \frac{1}{l^2} \sum_{i=1}^l \sum_{j=1}^l \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^T \mathbf{A} \bar{\mathbf{x}}_j \bar{\mathbf{x}}_j^T \quad (\text{S88})$$

144 Taking the expectation, we get

$$\mathbb{E}[\mathbf{S}_x \mathbf{A} \mathbf{S}_x] = \frac{1}{l^2} \sum_{i=1}^l \sum_{j=1}^l \mathbb{E}[\bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^T \mathbf{A} \bar{\mathbf{x}}_j \bar{\mathbf{x}}_j^T] \quad (\text{S89})$$

145 When $i \neq j$, $\bar{\mathbf{x}}_i$ and $\bar{\mathbf{x}}_j$ are independent, so

$$\mathbb{E}[\bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^T \mathbf{A} \bar{\mathbf{x}}_j \bar{\mathbf{x}}_j^T] = \mathbb{E}[\bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^T] \mathbf{A} \mathbb{E}[\bar{\mathbf{x}}_j \bar{\mathbf{x}}_j^T] = \Sigma \mathbf{A} \Sigma \quad (\text{S90})$$

146 When $i = j$,

$$\mathbb{E}[\bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^T \mathbf{A} \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^T] = \mathbb{E}[\bar{\mathbf{x}} \bar{\mathbf{x}}^T \mathbf{A} \bar{\mathbf{x}} \bar{\mathbf{x}}^T] \quad (\text{S91})$$

147 Let $\bar{\mathbf{x}} = [x_1, x_2, \dots, x_d]^T$. Then, from Isserlis' theorem [12], we have

$$\mathbb{E}[x_i x_j x_k x_l] = \Sigma_{ij} \Sigma_{kl} + \Sigma_{ik} \Sigma_{jl} + \Sigma_{il} \Sigma_{jk} \quad (\text{S92})$$

148 Let $\mathbf{A} = [a_{ij}]$. Then, $\bar{\mathbf{x}}^T \mathbf{A} \bar{\mathbf{x}} = \sum_{i,j} a_{ij} x_i x_j$. Thus, we reach

$$\bar{\mathbf{x}} \bar{\mathbf{x}}^T \mathbf{A} \bar{\mathbf{x}} \bar{\mathbf{x}}^T = \bar{\mathbf{x}} \bar{\mathbf{x}}^T \sum_{i,j} a_{ij} x_i x_j, \quad (\text{S93})$$

$$\mathbb{E}[\bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^T \mathbf{A} \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^T] = \text{Tr}(\mathbf{A} \Sigma) \Sigma + \Sigma \mathbf{A} \Sigma + \Sigma \mathbf{A}^T \Sigma. \quad (\text{S94})$$

149 There are l^2 terms in the double sum. l terms are of the form $\mathbb{E}[\bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^T \mathbf{A} \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^T]$ and $l^2 - l$ terms are of
150 the form $\Sigma \mathbf{A} \Sigma$. Therefore, we can write

$$\mathbb{E}[\mathbf{S}_x \mathbf{A} \mathbf{S}_x] = \frac{1}{l^2} [l(\text{Tr}(\mathbf{A} \Sigma) \Sigma + \Sigma \mathbf{A} \Sigma + \Sigma \mathbf{A}^T \Sigma) + l(l-1) \Sigma \mathbf{A} \Sigma], \quad (\text{S95})$$

$$= \frac{1}{l} (\text{Tr}(\mathbf{A} \Sigma) \Sigma + \Sigma \mathbf{A} \Sigma + \Sigma \mathbf{A}^T \Sigma) + \frac{l-1}{l} \Sigma \mathbf{A} \Sigma, \quad (\text{S96})$$

$$= \Sigma \mathbf{A} \Sigma + \frac{1}{l} \Sigma \mathbf{A}^T \Sigma + \frac{1}{l} \text{Tr}(\mathbf{A} \Sigma) \Sigma, \quad (\text{S97})$$

151 which completes the proof. \square

152 H Analysis of optimal temperature for ICL under distribution shift

153 Here, we find the optimal temperature minimizing the generalization error. First, recall that we have
154 the following generalization error.

$$\mathcal{G}(\mathbf{V}, \mathbf{M}) = \frac{1}{\tau^2} \text{Tr}(\mathbf{A} \mathbf{M}_{11}^T \mathbf{F}_1 \mathbf{M}_{11}) - \frac{1}{\tau} \text{Tr}(\mathbf{A} (\mathbf{F}_2 \mathbf{M}_{11} + \mathbf{M}_{11}^T \mathbf{F}_2^T)) + \text{Tr}(\mathbf{A} \mathbf{B}) + \sigma^2, \quad (\text{S98})$$

155 as specified in Theorem 4.6. So, we can express the generalization error as,

$$\mathcal{G}(\tau; \mathbf{V}, \mathbf{M}) = \frac{a}{\tau^2} - \frac{b}{\tau} + c, \quad (\text{S99})$$

156 where $a := \text{Tr}(\mathbf{A}\mathbf{M}_{11}^T\mathbf{F}_1\mathbf{M}_{11})$, $b := \text{Tr}(\mathbf{A}(\mathbf{F}_2\mathbf{M}_{11} + \mathbf{M}_{11}^T\mathbf{F}_2^T))$, and $c = \text{Tr}(\mathbf{A}\mathbf{B}) + \sigma^2$. There-
157 fore, we have the following optimization problem

$$\tau_{\text{optimal}} := \arg \min_{\tau} \mathcal{G}(\tau; \mathbf{V}, \mathbf{M}), \quad (\text{S100})$$

$$= \arg \min_{\tau} \left\{ \frac{a}{\tau^2} - \frac{b}{\tau} + c \right\}. \quad (\text{S101})$$

158 To find the optimal value of τ that minimizes the given function, we can take the derivative of the
159 expression with respect to τ and set it to zero. From now on, we consider generalization error as a
160 function of τ , written as $\mathcal{G}(\tau)$.

161 Next, find the derivative of $\mathcal{G}(\tau)$ with respect to τ as

$$\mathcal{G}'(\tau) = -2a\tau^{-3} + b\tau^{-2}. \quad (\text{S102})$$

162 To find the critical points, set $\mathcal{G}'(\tau) = 0$ as follows

$$\mathcal{G}'(\tau) = -2a\tau^{-3} + b\tau^{-2} = 0, \quad (\text{S103})$$

163 Solving this equation for τ , we reach the following critical point

$$\tau = \frac{2a}{b}. \quad (\text{S104})$$

164 Now, we need to check if this is a minimum by taking the second derivative, which is

$$\mathcal{G}''(\tau) = 6a\tau^{-4} - 2b\tau^{-3}. \quad (\text{S105})$$

165 Evaluate $\mathcal{G}''(\tau)$ at $\tau = \frac{2a}{b}$ as follows

$$\mathcal{G}''\left(\frac{2a}{b}\right) = 6a\left(\frac{2a}{b}\right)^{-4} - 2b\left(\frac{2a}{b}\right)^{-3} = 6a\left(\frac{b^4}{16a^4}\right) - 2b\left(\frac{b^3}{8a^3}\right) = \frac{b^4}{8a^3}. \quad (\text{S106})$$

166 Since $a, b > 0$, we reach $\mathcal{G}''\left(\frac{2a}{b}\right) = \frac{b^4}{8a^3} > 0$, which means the function has a minimum at $\tau = \frac{2a}{b}$.
167 Therefore, $\tau_{\text{optimal}} = \frac{2a}{b}$ is the solution minimizing the generalization error $\mathcal{G}(\tau)$. Writing a, b back
168 into the optimal solution, we get

$$\tau_{\text{optimal}} = \frac{2\text{Tr}(\mathbf{A}\mathbf{M}_{11}^T\mathbf{F}_1\mathbf{M}_{11})}{\text{Tr}(\mathbf{A}(\mathbf{F}_2\mathbf{M}_{11} + \mathbf{M}_{11}^T\mathbf{F}_2^T))}, \quad (\text{S107})$$

169 which concludes our derivation of the optimal temperature τ_{optimal} .

170 I Experimental details and GPT-2 experiments

171 This section provides details of our experiments involving the GPT-2 model and other large language
172 models.¹

173 I.1 GPT-2: Transformer with MLP layers

174 After testing the linearized attention model, we explore whether the optimal temperature benefits
175 more complex Transformer models in linear regression tasks. We experiment with the GPT-2 model
176 [23] under a shift in input covariance, as shown in Figure S2. We observe that distribution shift
177 significantly degrades performance, as noted in prior work [9, 35], even causing nonmonotonic

¹The code for our experimental results will be released before the camera-ready version of this work.

178 generalization error with respect to context length l . However, the optimal temperature mitigates this
 179 nonmonotonicity and improves generalization performance for in-context learning.

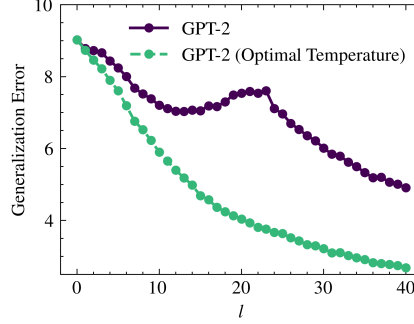


Figure S2: Experiments on GPT-2 [23] under a shift in input covariance. GPT-2 exemplifies the Transformer architecture [27], incorporating multi-layer perceptron layers and multi-head softmax self-attention. The model used here is pretrained by [9] on linear regression tasks defined in (2). We consider a shift from $\Sigma_x^{train} = \mathbf{I}$ to $\Sigma_x^{test} = 3\mathbf{I}$ in the input covariance. The attention temperature at each layer is scaled as $\tau\sqrt{d_k}$, where d_k is the key dimension, to ensure dimension-independent temperature τ values.

180 I.2 Details of the GPT-2 experiments in Figure S2

181 For the GPT-2 experiments, we use the standard GPT2 architecture [23], as implemented in HuggingFace [32]. We utilize the pretrained model by [9]. The training data is the same as ours while
 182 the training procedure is slightly different. Their loss function is auto-regressive, so the objective
 183 function is the average over the full length of the context sequence, where the full context length
 184 is $l = 40$. We also use the embedding method employed during the training [9]. Input dimension
 185 is $d = 20$, number of layers is 12, and number of heads is 8. We run the GPT-2 experiments on an
 186 NVIDIA Tesla V100 GPU, and it takes approximately 10 minutes to finish.

188 I.3 Details of the LLM experiments in Figure 3

189 For our experiments with large language models, we use Llama2-7B model [26] and SCIQ dataset
 190 [31]. The dataset consists of science questions with additional supporting information. We generate
 191 in-context learning (ICL) problems from the dataset following a prior work [8]. The in-context
 192 demonstrations are selected based on the TopK [17] retrieval technique so that the demonstrations are
 193 relevant to the test question. An ICL sample generated from the SCIQ dataset is illustrated in Table S1.
 194 For the distribution shift, following [8], we consider in-context demonstrations with noisy labels that
 195 are wrong but related to the original labels (see Appendix I.4 for the motivation). Table S2 illustrates
 196 an example of in-context demonstration with a noisy label. Note that the noisy ratio is the ratio of
 197 noisy-labeled demonstrations to all demonstrations. Thus, 0.6 means 60% of the demonstrations are
 198 with noisy labels. For our experiments, we modify and use the code base provided by [8], which also
 199 utilized Huggingface [32] and OpenICL [34]. We run the LLM experiments on an NVIDIA A40
 200 GPU, and a single Monte Carlo run for each plot (in Figure 3) takes a couple of hours to complete.

201 I.4 Why in-context demonstrations with noisy labels as an example of distribution shift?

202 At first glance, the relationship between introducing noisy labels to the in-context demonstrations and
 203 its effect as a distribution shift can look unclear. Indeed, measuring distribution shifts from pretraining
 204 to test time for pretrained LLMs is challenging since the datasets used for the pretraining are complex
 205 mixtures of different data sources [26]. However, we conjecture that high perplexity, an empirical
 206 measure of uncertainty in generating new tokens with an LLM, can indicate a distribution shift for
 207 pretrained LLMs, and it has been shown that the noisy demonstrations lead to higher perplexity
 208 [8]. Particularly, any input that exists (or relates to existing text) in the training set would lead to
 209 generation with high confidence (low perplexity), while any input that contradicts texts in the training
 210 set would result in generation with low confidence (high perplexity). Since noisy demonstrations

are expected to contradict the existing text in the training set, such noisy demonstrations lead to high perplexity. Therefore, the noisy labels introduced into the in-context demonstrations can be considered as a distribution shift from pretraining to test time.

In-context demonstration 1

Support: Cells are organized into tissues, tissues are organized into organs.
Question: What is considered the smallest unit of the organ?
Answer: Cells

In-context demonstration 2

Support: ... four basic types of tissue: connective, muscle, nervous, and epithelial.
Question: The four basic types of tissue are epithelial, muscle, connective, and what?
Answer: nervous

⋮

Test example

Support: All forms of life are built of at least one cell. A cell is the basic unit of life.
Question: What are the smallest structural and functional units of all living organisms?
Output: ???

Table S1: A sample illustration of in-context learning on the SCIQ dataset.

Setting	In-context demonstration
True Label	Support: Cells are organized into tissues, tissues are organized into organs. Question: What is considered the smallest unit of the organ? Label: Cells
Noisy Label	Support: Cells are organized into tissues, tissues are organized into organs. Question: What is considered the smallest unit of the organ? Label: tissues

Table S2: An example of a true label vs. a relevant but noisy label. A relevant label is related to the question but is not necessarily true. Therefore, relevant labels can be considered noisy labels.

References

- [1] Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.
- [2] Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations*, 2023.
- [3] Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.
- [4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [5] Xiangyu Chen, Qinghao Hu, Kaidong Li, Cuncong Zhong, and Guanghui Wang. Accumulated trivial attention matters in vision transformers on small datasets. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3984–3992, 2023.

- [6] Krzysztof Marcin Choromanski, Valerii Likhoshesterov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, David Benjamin Belanger, Lucy J Colwell, and Adrian Weller. Rethinking attention with performers. In *International Conference on Learning Representations*, 2021.
- [7] Deqing Fu, Tianqi CHEN, Robin Jia, and Vatsal Sharan. Transformers learn higher-order optimization methods for in-context learning: A study with linear models, 2024.
- [8] Hongfu Gao, Feipeng Zhang, Wenyu Jiang, Jun Shu, Feng Zheng, and Hongxin Wei. On the noise robustness of in-context learning for text generation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- [9] Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598, 2022.
- [10] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. <http://www.deeplearningbook.org>.
- [11] Dongchen Han, Yifan Pu, Zhuofan Xia, Yizeng Han, Xuran Pan, Xiu Li, Jiwen Lu, Shiji Song, and Gao Huang. Bridging the divide: Reconsidering softmax and linear attention. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- [12] L. Isserlis. On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, 12(1/2):134–139, 1918.
- [13] Seung Hoon Lee, Seunghyun Lee, and Byung Cheol Song. Vision transformer for small-size datasets. *arXiv preprint arXiv:2112.13492*, 2021.
- [14] Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. Transformers as algorithms: Generalization and stability in in-context learning. In *International Conference on Machine Learning*, pages 19565–19594. PMLR, 2023.
- [15] Yingcong Li, Ankit Singh Rawat, and Samet Oymak. Fine-grained analysis of in-context linear estimation: Data, architecture, and beyond. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- [16] Junyang Lin, Xu Sun, Xuancheng Ren, Muyu Li, and Qi Su. Learning when to concentrate or divert attention: Self-adaptive attention temperature for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2985–2990, 2018.
- [17] Jiachang Liu, Dinghan Shen, Yizhe Zhang, William B Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for gpt-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, 2022.
- [18] Arvind V. Mahankali, Tatsunori Hashimoto, and Tengyu Ma. One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. In *The Twelfth International Conference on Learning Representations*, 2024.
- [19] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. <https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html>.
- [20] Core Francisco Park, Ekdeep Singh Lubana, Itamar Pres, and Hidenori Tanaka. Competition dynamics shape algorithmic phases of in-context learning. *arXiv preprint arXiv:2412.01003*, 2024.

- [21] Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. YaRN: Efficient context window extension of large language models. In *The Twelfth International Conference on Learning Representations*, 2024.
- [22] Zhen Qin, Weixuan Sun, Hui Deng, Dongxu Li, Yunshen Wei, Baohong Lv, Junjie Yan, Lingpeng Kong, and Yiran Zhong. cosformer: Rethinking softmax in attention. In *International Conference on Learning Representations*, 2022.
- [23] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI*, 2019.
- [24] Allan Raventós, Mansheej Paul, Feng Chen, and Surya Ganguli. Pretraining task diversity and the emergence of non-bayesian in-context learning for regression. *Advances in Neural Information Processing Systems*, 36, 2024.
- [25] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.
- [26] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- [27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 2017.
- [28] Petar Veličković, Christos Perivolaropoulos, Federico Barbero, and Razvan Pascanu. softmax is not enough (for sharp out-of-distribution). *arXiv preprint arXiv:2410.01104*, 2024.
- [29] Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, Joao Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 35151–35174. PMLR, 23–29 Jul 2023.
- [30] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022.
- [31] Johannes Welbl, Nelson F Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, 2017.
- [32] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.
- [33] Jingfeng Wu, Difan Zou, Zixiang Chen, Vladimir Braverman, Quanquan Gu, and Peter Bartlett. How many pretraining tasks are needed for in-context learning of linear regression? In *The Twelfth International Conference on Learning Representations*, 2024.
- [34] Zhenyu Wu, YaoXiang Wang, Jiacheng Ye, Jiangtao Feng, Jingjing Xu, Yu Qiao, and Zhiyong Wu. Openicl: An open-source framework for in-context learning. *arXiv preprint arXiv:2303.02913*, 2023.
- [35] Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024.

- 328 [36] Shengqiang Zhang, Xingxing Zhang, Hangbo Bao, and Furu Wei. Attention temperature matters
329 in abstractive summarization distillation. In *Proceedings of the 60th Annual Meeting of the*
330 *Association for Computational Linguistics (Volume 1: Long Papers)*, pages 127–141, 2022.
- 331 [37] Yixiong Zou, Ran Ma, Yuhua Li, and Ruixuan Li. Attention temperature matters in vit-based
332 cross-domain few-shot learning. In *Neural Information Processing Systems*, 2024.