Provable In-Context Learning of Linear Systems and Linear Elliptic PDEs with Transformers

Anonymous Author(s)
Affiliation
Address
email

Abstract

Foundation models for natural language processing, empowered by the transformer architecture, exhibit remarkable *in-context learning* (ICL) capabilities: pre-trained models can adapt to a downstream task by only conditioning on few-shot prompts without updating the weights of the models. Recently, transformer-based foundation models also emerged as universal tools for solving scientific problems, including especially partial differential equations (PDEs). However, the theoretical underpinnings of ICL-capabilities of these models still remain elusive. This work develops rigorous error analysis for transformer-based ICL of the solution operators associated to a family of linear elliptic PDEs. Specifically, we show that a linear transformer defined by a linear self-attention layer can provably learn in-context to invert linear systems arising from the spatial discretization of the PDEs. We derive theoretical scaling laws for the proposed linear transformers in terms of the size of the spatial discretization, the number of training tasks, the lengths of prompts used during training and inference, under both the in-domain generalization setting and various settings of distribution shifts. Empirically, we validate the ICL-capabilities of transformers through extensive numerical experiments.

1 Introduction

Foundation models (FMs) for natural language processing (NLP), exemplified by ChatGPT Achiam et al. [2023], have demonstrated unprecedented power in text generation tasks. From an architectural perspective, the main novelty of these models is the use of transformer-based neural networks Vaswani et al. [2017], which are distinguished from feedforward neural networks by their self-attention layers. Those transformer-based FMs, pre-trained on a broad range of tasks with large amounts of data, exhibit remarkable transferability to diverse downstream tasks with limited data Brown et al. [2020]. The success of of foundation models for NLP has recently sparked a large amount of work on building FMs in domain-specific scientific fields Batatia et al. [2023], Celaj et al. [2023], Méndez-Lucio et al. [2022]. Specifically, there is growing interest within the community of Scientific Machine Learning (SciML) in building scientific foundation models (SciFMs) to solve complex partial differential equations (PDEs) Subramanian et al. [2024], McCabe et al. [2023], Ye et al. [2024], Yang et al. [2023], Sun et al. [2024].

Traditional deep learning approaches for PDEs such as Physics-Informed Neural Networks Raissi et al. [2019] for learning solutions and neural operators Lu et al. [2019], Li et al. [2020] for learning solution operators need to be retrained from scratch for a different set of coefficients or different PDE systems. Instead, these SciFMs for PDEs, once pre-trained on large datasets of coefficients-solution pairs from multiple PDE systems, can be adapted to solving new PDE systems without training the model from scratch. Even more surprisingly, transformer-based FMs have demonstrated their **in-context learning (ICL)** capability in Achiam et al. [2023], Bubeck et al. [2023], Kirsch et al.

[2022] and in SciML Yang et al. [2023], Chen et al. [2024a], Yang and Osher [2024]: when given a prompt consisting of examples from a new learning task and a query, they are able to make correct predictions without updating their parameters. While the emergence of ICL has been deemed a paradigm shift in transformer-based FMs, its theoretical understandings remain underdeveloped.

The goal of this paper is to investigate the ICL capability of transformers for solving a class of linear elliptic PDEs and the associated linear systems. We are particularly interested in developing **neural scaling laws** that quantify the prediction risk of transformers as a function of the size of the training data, the model size, and other key parameters. Additionally, we aim to quantify the error incurred by **distribution shifts** between tasks and data used in pre-training and those in adaptation. As distribution shifts have been identified in Subramanian et al. [2024], McCabe et al. [2023], Ye et al. [2024], Yang et al. [2023] as a significant hurdle in the generalization capability of SciFMs, it is crucial to develop a rigorous theory of out-of-distribution generalization for SciFMs.

1.1 Main contributions.

We highlight our main contributions as follows:

- We formalize a framework for learning the solution operators of linear elliptic PDEs in-context. This is based on (1) reducing the infinite dimensional PDE problem into a problem of solving a finite dimensional linear system arising from spatial discretization of the PDE and (2) learning to invert the finite dimensional linear system in-context.
- We adopt transformers defined by single linear self-attention layers for ICL of the linear systems and establish a non-asymptotic generalization bound of ICL in terms of the discretization size, the number of pre-training tasks, and the lengths of prompts used in pre-training and downstream tasks; see Theorem 1. This bound further enables us to prove an H^1-error bound for learning the solution of PDEs; see Theorem 2.
- We examine the prediction risk error that arises due to shifts in downstream task and covariate distributions. Specifically, we introduce a novel concept of task diversity and demonstrate that pre-trained transformers can generalize to out-of-distribution settings when the pre-training task distribution is diverse; see Theorem 3. Additionally, we provide several sufficient conditions under which task diversity holds; see Theorem 4.
- We demonstrate the ICL ability of linear transformers through several numerical experiments.

1.2 Related work

ICL and FMs for PDE. Several transformer-based FMs for solving PDEs have been developed in Subramanian et al. [2024], McCabe et al. [2023], Ye et al. [2024], Sun et al. [2024] where the pre-trained transformers are adapted to downstream tasks with fine-tuning on additional datasets. The work Yang et al. [2023], Yang and Osher [2024] study the in-context operator learning of differential equations where the adaption of the pre-trained model is achieved by only conditioning on new prompts. While these empirical work show great transferabilities of SciFMs for solving PDEs, their theoretical guarantees are largely open. To the best of our knowledge, this work is the first to derive the theoretical error bounds of transformers for learning linear elliptic PDEs in context.

Theory of ICL for linear regression and other statistical models. The work Garg et al. [2022] provides theoretical understanding of the ability of transformers in learning simple functions in-context. In the follow-up works Akyürek et al. [2022], Von Oswald et al. [2023], it is shown by explicit construction of attention matrices that linear transformers can implement a single step of gradient descent when given a new in-context linear regression task, and numerical experiments supported that trained transformer indeed implement gradient descent on unseen tasks. Several recent works Mahankali et al. [2024], Zhang et al. [2023], Ahn et al. [2024] extend the results of Von Oswald et al. [2023] by proving that one step of gradient descent is indeed optimal for learning linear models in-context. These works are further complemented by ICL guarantees for learning nonlinear functions Bai et al. [2024], Cheng et al. [2023], Kim et al. [2024] and for reinforcement learning problems Lin et al. [2023].

Among the aforementioned works, the settings of Zhang et al. [2023], Ahn et al. [2024], Chen et al. [2024b] are closest to us. Our theoretical bound on the population risk extends the results of Zhang et al. [2023], Ahn et al. [2024] for the linear regression tasks to the tasks of inverting linear systems that are associated to elliptic PDEs. Our main novelty is that our results apply to a much larger class of task distributions, since our task matrices must respect the PDE structure. In particular, this leads to new and nontrivial results regarding task distribution shifts, whereas the effect of task distribution shifts is simple under the assumptions of the aforementioned works. We also provide sample complexity bounds with respect to the number of pre-training tasks, which have not addressed by the above works.

2 Problem set-up

2.1 In-context operator learning of linear elliptic PDEs

Consider the second-order strongly-elliptic PDE on a bounded Lipschitz domain $\Omega \subseteq \mathbb{R}^{d_0}$:

$$\begin{cases} \mathcal{L}_{a,V} u(x) := -\nabla \cdot \Big(a(x) \nabla u(x) \Big) + V(x) u(x) = f(x),\ x \in \Omega \\ u(x) = 0, x \in \partial\Omega. \end{cases} \tag{1}$$

where $a \in L^\infty(\Omega)$ is strictly positive, $V \in L^\infty(\Omega)$ is non-negative and $f \in \mathcal{X}_f \subset L^2(\Omega)$. By the standard well-posedness of the elliptic PDE, the solution $u \in \mathcal{X}_u \subset H_0^1(\Omega)$. We are interested in learning the linear solution operator $\Psi : f \to u \in \mathcal{X}_u$ in context Yang et al. [2023]. More specifically, at the training stage we are given a training dataset comprising N length-n prompts of source-solution pairs $\{(f_i^j, u_i^j)_{i=1}^n\}_{j=1}^N$, where $\{f_i^j\} \overset{i.i.d.}{\sim} P_f$ for some distribution P_f on the space of functions f, and u_i^j are the solutions corresponding to f_i^j and parameters $(a_j, V_j) \overset{i.i.d.}{\sim} P_a \times P_V$, where P_a and P_V are distributions on the coefficient a and V respectively. An ICL model, after pre-trained on the data above, is asked to predict the solution u for a new source term f conditioned on a new prompt $(f_i, u_i)_{i=1}^m$ which may or may not have the same distribution as the training prompts. Further, the prompt-length m in the downstream task may be different from the prompt-length n in the training.

While the ideal ICL problem above is stated for learning operators defined on infinite dimensional function spaces, a practical ICL model (e.g. a transformer) can only operate on finite dimensional data, which are typically observed in the form of finite dimensional projections or discrete evaluations. To be more concrete, let $\{\phi_k(x)\}_{k=1}^\infty$ be a basis on both $\mathcal{X}_u$ and $\mathcal{X}_f$, and define a truncated base set $\mathbf{\Phi}(x) := [\phi_1, \cdots, \phi_d(x)]$ for some $d < \infty$. An approximate solution $\tilde{u}$ to problem (1) can be constructed in the framework of Galerkin method: we seek $\tilde{u}(x) = \langle \mathbf{u}, \mathbf{\Phi}(x) \rangle$ where $\mathbf{u} \in \mathbb{R}^d$ solves the linear system $A\mathbf{u} = \mathbf{f}$, where the matrix $A = (A_{ij}) \in \mathbb{R}^{d \times d}$ and the right hand side $\mathbf{f} = (f_i) \in \mathbb{R}^d$ are defined by

$$A_{ij} = \langle \phi_j, \mathcal{L}_{a,V} \phi_i \rangle \text{ and } f_i = \langle f, \phi_i \rangle, i, j = 1, \cdots, d. \tag{2}$$

As quantitative discretization error bounds of PDEs are well established, e.g. for finite element methods Brenner and Scott [2007] and spectral methods Shen et al. [2011], this paper focuses on the error analysis of in-context learning of the finite dimensional linear systems defined by the matrix inversion A^{-1}, which will ultimately translate to estimation bounds for the PDEs.

2.2 ICL of linear systems

The consideration above reduces the original infinite dimensional in-context operator learning problem to the finite dimensional ICL problem of solving linear systems. To keep the framework more general, we make the following change of notations: $\mathbf{f} \to \mathbf{y}$ and $\mathbf{u} \to \mathbf{x}$. An ICL model operates on a prompt of n input-output pairs, denoted by $S := \{(\mathbf{y}_i, \mathbf{x}_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}^d$ with $\mathbf{x}_i = A^{-1}\mathbf{y}_i$ as well as a new query input $\mathbf{y}_{n+1} \in \mathbb{R}^d$. Given multiple prompts, the model aims to predict $\mathbf{x}_{n+1}$ corresponding to the new independent query input $\mathbf{y}_{n+1}$. Unlike in supervised learning, each prompt the model takes is drawn from a different data distribution. To be more precise, for $j = 1, \cdots, N$, we assume that the j-th prompt $S^{(j)} := \{(\mathbf{y}_i^{(j)}, \mathbf{x}_i^{(j)})\}_{i=1}^n$ is generated from the sources $\{\mathbf{y}_i^{(j)}\}_{i=1}^n \overset{i.i.d.}{\sim} p_{\mathbf{y}}$; the solutions $\mathbf{x}_i^{(j)}$ are associated to the j-th inversion task via $\mathbf{x}_i^{(j)} = (A^{(j)})^{-1}\mathbf{y}_i^{(j)}$ where the matrices $A^{(j)} \overset{i.i.d.}{\sim} p_A$. Informed by task matrices derived from discretizations of PDEs as illuminated in (2), we make the following assumption on the task distribution p_A.

Assumption 1. *The task distribution p_A is supported on the set of symmetric positive definite matrices, and there exist constants $c_A, C_A > 0$ such that the bounds $c_A^{-1}\mathbf{I}_d \prec A \prec C_A\mathbf{I}_d$ hold for all $A \in supp(p_A)$. The source term $\mathbf{y}$ follows a Gaussian distribution $N(0, \Sigma)$.*

Observe that Assumption 1 on A is very mild and holds for instance whenever the coefficient a is strictly positive and V is non-negative and bounded. We will make repeated use of the bounds[1]

$$\|A^{-1}\|_{\text{op}} \leq c_A, \ \|A\|_{\text{op}} \leq C_A, \ p_A - \text{a.s.} \tag{3}$$

The Gaussian assumption on the covariate $\mathbf{y}$ holds when we assume that the source term f of the PDE is drawn from a Gaussian measure $N(0, \Sigma_f)$, where $\Sigma_f : L^2(\Omega) \to L^2(\Omega)$ is bounded, in which case the covariance matrix Σ is defined by $\Sigma_{ij} = \langle \Sigma_f \phi_i, \phi_j \rangle_{L^2(\Omega)}$.

2.3 Linear transformer architecture for linear systems

Inspired by the recent line of work on ICL of linear functions, we consider a linear transformer defined by a single-layer linear self-attention layer for our ICL model. Following the standard convention, we encode the data of each prompt into a prompt matrix

$$Z = \begin{bmatrix} \mathbf{y}_1 & \cdots & \mathbf{y}_n & \mathbf{y}_{n+1} \\ \mathbf{x}_1 & \cdots & \mathbf{x}_n & 0 \end{bmatrix} \in \mathbb{R}^{D\times(n+1)}, \tag{4}$$

where $D = 2d$. For $\tilde{P}, \tilde{Q} \in \mathbb{R}^{D\times D}$, the linear self-attention module with parameters $\tilde{\theta} = (\tilde{P}, \tilde{Q})$ is given by

$$\text{Attn}_{\tilde{\theta}}(Z) = Z + \frac{1}{n}\tilde{P}ZMZ^T\tilde{Q}Z,$$

where $M = \begin{bmatrix} \mathbf{I}_n & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{(n+1)\times(n+1)}$ is a masking matrix to account for the asymmetry of the prompt matrix. Our definition of the self-attention module makes several simplifying assumptions compared to the standard definition in the literature, namely we merge the key and query matrices into a single matrix Q and we omit the softmax activation function. A transformer $f_{\tilde{\theta}}$ predicts a new label $\mathbf{x}$ for the downstream task by reading out the $\mathbf{x}$-component from the self-attention output, i.e.

$$f_{\tilde{\theta}}(Z) := [\text{Attn}_{\tilde{\theta}}(Z)]_{d+1:D,n+1} = \sum_{j=1}^{d} \langle \mathbf{e}_{d+j}, \text{Attn}_{\tilde{\theta}}(Z)\mathbf{e}_{n+1}\rangle \mathbf{e}_{d+j},$$

where $\mathbf{e}_i$ denotes the i^{th} standard basis vector. Since the output of the transformer only reads out the last d entries on the bottom right of the output of the self-attention layer, many blocks in $\tilde{P}$ and $\tilde{Q}$ do not actually play a role in the prediction defined by the transformer. More precisely, similar to Von Oswald et al. [2023], Zhang et al. [2023], Ahn et al. [2024], if we set $\tilde{P} = \begin{bmatrix} 0 & 0 \\ 0 & P \end{bmatrix}$ and $\tilde{Q} = \begin{bmatrix} Q & 0 \\ 0 & 0 \end{bmatrix}$ with $P, Q \in \mathbb{R}^{d\times d}$, then output of the transformer can be re-written in a compact form: with $\theta = (P, Q)$,

$$\text{TF}_\theta(Z) = PA^{-1}Y_nQ\mathbf{y},$$

where $Y_n := \frac{1}{n}\sum_{k=1}^{n} \mathbf{y}_k\mathbf{y}_k^T$ denotes the empirical covariance matrix associated to the in-context examples. We work with this simplified parameterization for the remainder of our theoretical analysis.

2.4 Generalization of ICL

Our goal is to find the attention matrices P and Q that minimize the *population risk* functional

$$\mathcal{R}_n(P, Q; n) = \mathbb{E}\left[\left\|\text{TF}_\theta(Z) - A^{-1}\mathbf{y}\right\|^2\right] = \mathbb{E}\left[\left\|PA^{-1}Y_nQ\mathbf{y} - A^{-1}\mathbf{y}\right\|^2\right], \tag{5}$$

[1]Most of our estimates involve bounds on the norm of A^{-1}, since it represents the 'solution operator' of the PDE. However, for technical reasons, we also require a bound on the norm of A.

where the expectation is taken over $A \sim p_A$, $\{\mathbf{y}, \mathbf{y}_1, \ldots, \mathbf{y}_n\} \sim N(0, \Sigma)^{\otimes n+1}$. Since we do not have access to the distribution on tasks, P and Q are instead trained by minimizing the corresponding *empirical risk* functional defined on N tasks:

$$\mathcal{R}_{n,N}(P,Q) = \frac{1}{N} \sum_{i=1}^{N} \left\| PA_i^{-1} Y_n^{(i)} Q\mathbf{y}_i - A_i^{-1}\mathbf{y}_i \right\|^2, \tag{6}$$

where $\{A_i\} \overset{i.i.d.}{\sim} p_A$, $\{\mathbf{y}_i\} \overset{i.i.d.}{\sim} N(0, \Sigma)$, and $Y_n^{(i)}$ is the empirical covariance matrix associated to the in-context examples $\{\mathbf{y}_1^{(i)}, \ldots, \mathbf{y}_n^{(i)}\}$ which are (jointly) independent from $\mathbf{y}_i$.

A pre-trained transformer is expected to make predictions on a downstream task that consists of a new length-m prompt $\{(\mathbf{y}_i, \mathbf{x}_i)\}_{i=1}^m = \{(\mathbf{y}_i, (A')^{-1}\mathbf{y}_i)\}_{i=1}^m$ and a new test sample $\mathbf{y}$, where the input samples $\{\mathbf{y}_i\}_{i=1}^n \cup \{\mathbf{y}\} \sim P'_{\mathbf{y}}$ and the matrix $A' \sim P'_A = N(0, \Sigma')$. Our primary interest is to bound the generalization performance (measured by the prediction risk) of the pre-trained transformer for the downstream task in two different scenarios.

- **In-domain generalization:** The distributions of tasks and of prompt data in the pre-training are the same as these in the downstream task ($P_{\mathbf{y}} = P'_{\mathbf{y}}$ and $P_A = P'_A$). Thus in-domain generalization measures the testing performance on unseen samples in the downstream task that do not appear in the training samples. The in-domain generalization error is defined by

$$\mathcal{R}_m(P,Q;m) = \mathbb{E}_{A\sim p_A,(y_1,\ldots,y_m,y)\sim N(0,\Sigma)^{\otimes(m+1)}} \left[\left\| PA^{-1}Y_m Q\mathbf{y} - A^{-1}\mathbf{y} \right\|^2 \right]. \tag{7}$$

- **Out-of-domain (OOD) generalization:** The distributions of tasks or within-task data in the pre-training are different from those in the downstream task, i.e. $P_{\mathbf{y}} \neq P'_{\mathbf{y}}$ or $P_A \neq P'_A$. Specifically, the OOD-generalization error with respect to the task distribution shift is defined by

$$\mathcal{R}_m^{p'_A}(P,Q;m) = \mathbb{E}_{A'\sim p'_A,(y_1,\ldots,y_m,y)\sim N(0,\Sigma)^{\otimes(m+1)}} \left[\left\| P(A')^{-1}Y_m Q\mathbf{y} - (A')^{-1}\mathbf{y} \right\|^2 \right]. \tag{8}$$

We also define the OOD-generalization error with respect to the covariate distribution shift by

$$\mathcal{R}_m^{\Sigma'}(P,Q;m) = \mathbb{E}_{A\sim p_A,(y_1,\ldots,y_m,y)\sim N(0,\Sigma')^{\otimes(m+1)}} \left[\left\| PA^{-1}Y_m Q\mathbf{y} - A^{-1}\mathbf{y} \right\|^2 \right]. \tag{9}$$

Notice that the prompt length m in the prediction risk need not equal the prompt length n in the model pre-training. We are particularly interested in quantifying the scaling laws of the generalization errors for the pre-trained transformer as the amount of data increases to infinity, i.e. $N, n, m \uparrow \infty$.

3 Theoretical results

3.1 Error bounds for in-domain generalization of learning linear systems

Our first result studies the generalization ability of the transformer obtained by empirical risk minimization over a set of norm-constrained transformers, where the error is measured by the prediction risk $\mathcal{R}_m$.

Theorem 1. *Let $\widehat{\theta} = (P_N, Q_N) \in \text{argmin}_{\|\theta\|\leq M}\mathcal{R}_{n,N}(\theta)$, where $\|\theta\| := \max\Big(\|P\|_{op}, \|Q\|_{op}\Big)$. Then for n sufficiently large and $m \leq n$, we have with probability $\geq 1 - \frac{1}{poly(N)}$,*

$$\mathcal{R}_m(\widehat{\theta}) \lesssim \frac{1}{m} + \frac{1}{n^2} + \frac{d^2}{\sqrt{N}},$$

where the implicit constants depend on M, the data covariance Σ, and the task distribution p_A, and we have omitted factors which are polylog in N.

The precise statement of Theorem 1 is given in Appendix B, where we discuss what happens when $m > n$. We refer to $m \leq n$ as the practical regime, since it is commonly satisfied by large pre-trained transformers. Notice that the prompt lengths during training and testing contribute different rates to the overall sample complexity bound, with the sequence length n during training contributing an $O(n^{-2})$ rate while the sequence length m at inference contributing an $O(m^{-1})$ rate; a similar phenomenon was observed in [Zhang et al., 2023, Theorem 4.2] for in-context linear regression.

3.2 Error bounds for in-domain generalization of learning elliptic PDEs

Building upon Theorem 1, we proceed to bound the ICL-generalization error for learning the elliptic PDE (1). Our next result provides a rather general upper bound on the ICL-generalization error for the PDE solution in terms of the spatial discretization error of the PDE and the ICL-generalization error in learning the finite linear systems associated to the discretization. The discretization error is typically fully determined by the number d of basis functions used in the Galerkin projection. The second term is bounded by Theorem 1. In the following result, let u denote the solution to the elliptic PDE specified by (1). We write u_d for a discrete approximation to u with the mesh size h and we write $\widehat{u}_d$ for the approximate solution obtained by solving a discrete linear system with a pre-trained transformer.

Theorem 2. *Let Φ' be the stiffness matrix defined by $\Phi'_{ij} = (\phi'_i, \phi'_j)_{L^2(\Omega)}$ and let Φ be the mass matrix defined by $\Phi_{ij} = (\phi_i, \phi_j)_{L^2(\Omega)}$. Assume that both matrices are symmetric and positive definite. Then,*

$$\mathbb{E}\|u - \widehat{u}_d\|^2_{H^1(\Omega)} \lesssim \mathbb{E}\|u - u_d\|^2_{H^1(\Omega)} + (1 + \lambda_{\max}(\Phi^{-1/2}\Phi'\Phi^{-1/2})) \cdot \mathcal{R}_m(\widehat{\theta}),$$

where $\widehat{\theta}$ is a minimizer of the empirical risk defined in Theorem 1 and $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue of a symmetric positive definite matrix.

Theorem 2 bounds the in-domain generalization error of ICL of the PDE as a sum of the discretization error of the PDE solver and the statistical error of learning the linear system associated to the discretization of the PDE. It is worth-noting that there is a trade-off between the two terms; the first term decreases as the number of basis functions (or fineness of the mesh) increases, while the prefactor $\lambda_{\max}(\Phi^{-1/2}\Phi'\Phi^{-1/2})$ in the second term can grow as the number of basis functions tends to infinity. The abstract bound established in Theorem 2 is agnostic to the choice of PDE discretization. We show in Appendix C how this result can be used to derive an explicit error estimate for the ICL in the context of a P^1-finite element discretization of the PDE in one dimension.

3.3 OOD-generalization under task distribution shift

Let $\widehat{\theta}$ denote the minimizer of the empirical risk $\mathcal{R}_{n,N}$ over the bounded set $\{\|\theta\| \leq M\}$ for some $M > 0$, and recall that the training tasks (modeled by A) are drawn from a distribution p_A. Let p'_A denote the distribution of A in the downstream tasks, and let $\mathcal{R}_m, \mathcal{R}'_m$ be the prediction risk functionals defined as in (8) where the expectations over tasks are taken with respect to p_A and p'_A respectively. We would like to bound the quantity $\mathcal{R}'_m(\widehat{\theta})$, which represents the test error of the trained transformer under a shift on the task distribution. We say that a pre-trained model $\widehat{\theta}$ ***achieves OOD generalization*** if its population risk with respect to the downstream task distribution p'_A converges to zero in probability: $\lim_{(m,n,N)\to\infty} \mathcal{R}'_m(\widehat{\theta}) \xrightarrow{\text{P}} 0$. In order to state our results on OOD generalization, we first introduce the following 'infinite-context' variant of the in-domain denoted by $\mathcal{R}_\infty$:

$$\mathcal{R}_\infty(\theta) = \mathbb{E}_{A\sim p_A}[\|(PA^{-1}\Sigma Q - A^{-1})\Sigma^{1/2}\|_F^2]. \tag{10}$$

We also define an OOD-generalization risk $\mathcal{R}'_\infty$ similar to above with p_A replaced by p'_A. We denote by $\mathcal{M}_\infty$ and $\mathcal{M}'_\infty$ the sets of minimizers of $\mathcal{R}_\infty$ and $\mathcal{R}'_\infty$ respectively. We are now able to define the key notion of task diversity.

Definition 1. *The pre-training task distribution p_A is **diverse** relative to the downstream task distribution p'_A if $\mathcal{M}_\infty \subseteq \mathcal{M}'_\infty$.*

The importance of task diversity has been observed in the prior work Tripuraneni et al. [2020] for transfer learning. Our notion of diversity differs from the previous notion in that we compare the set of minimizers of population losses instead of the loss values. Theorem 3 below shows that the task diversity, in the sense of Definition 1, is sufficient for the pre-trained transformer to achieve OOD-generalization.

Theorem 3. *Let p_A and p'_A denote the pre-training and downstream task distributions respectively, and suppose p_A is diverse relative to p'_A. Then, with $\widehat{\theta} \in \text{argmin}_{\|\theta\|\leq M}\mathcal{R}_{n,N}(\theta)$, we have*

$$\mathcal{R}'_m(\widehat{\theta}) \lesssim \mathcal{R}_m(\widehat{\theta}) + \frac{d(p_A, p'_A)}{m} + \mathit{dist}(\widehat{\theta}, \mathcal{M}_\infty)^2,$$

where $d(p_A, p'_A)$ is a discrepancy between the pre-training and downstream task distributions that satisfies $d(p_A, p'_A) = 0$ if $p_A = p'_A$.

The precise definition of the discrepancy $d(p_A, p'_A)$ is technical and can be found in the statement of Lemma 2 in the appendix. The OOD generalization error is bounded by a sum of three terms: the in-domain generalization error, the task-shift error, and the model error, the latter of which is captured by $\mathrm{dist}(\widehat{\theta}_n, \mathcal{M}_\infty)$. A salient feature of Theorem 3, compared to the prior ICL-generalization bound Mroueh [2023] under distribution shift, is that the task-shift error inherits a factor of m^{-1}, which elucidates the robustness of transformers under shifts in the task distribution. Theorem 3 also extends the prior OOD-generalization result of ICL for linear regression Zhang et al. [2023] to learning linear systems. However, unlike in the linear regression setting, the set of minimizers of the population risk in the linear system setting can vary substantially when the task distribution changes, we need the training tasks to be sufficient diverse compared to the downstream tasks in order to control the additional model error due to the change of the minimizers; see Appendix D for more details. We also note that Proposition 4 in the appendix shows that the minimizers of the empirical risk converge in probability to the minimizers of $\mathcal{R}_\infty$, thus guaranteeing that the bound in Theorem 3 is $o_P(1)$.

Since task diversity is sufficient to achieve OOD generalization, it is natural to ask what conditions on p_A and p'_A guarantee task diversity. The following result provides two sufficient conditions. We refer the readers to Appendix D for additional discussions on task diversity. To state the result, we recall that the notion of the centralizer $\mathcal{C}(\mathcal{S})$ of a subset $\mathcal{S} \subseteq \mathbb{R}^{d\times d} : \mathcal{C}(\mathcal{S}) = \{P \in \mathbb{R}^{d\times d} : PS = SP \ \forall S \in \mathcal{S}\}$.

Theorem 4. *Let p_A, p'_A be two distributions on the matrices A that satisfy Assumption 1. Then*

1. *If supp$(p'_A) \subseteq$ supp(p_A), then p_A is diverse relative to p'_A.*
2. *Define $\mathcal{S}(p_A) := \{A_1 A_2^{-1} : A_1, A_2 \in \text{supp}(p_A)\}$. If $\mathcal{C}(\mathcal{S}(p_A)) = \{c\mathbf{I} : c \in \mathbb{R}\}$, then p_A is diverse relative to any distribution p'_A.*

The first statement of Theorem 4 is a natural one: it says that the pre-training task distribution is diverse whenever the downstream task distribution is a 'subset' of it, in the sense of supports. The second condition is particularly interesting because it implies OOD-generalization (by Theorem 3) regardless of the downstream task distribution. The second condition based on the centralizer of the set $\mathcal{S}(p_A)$ is less obvious, but heuristically it enforces that the support of p_A must be large enough that the only matrices which can commute with all pairwise products in $\mathcal{S}(p_A)$ are scalars. Our empirical results suggest that the task distributions associated to elliptic PDE problems are diverse.

3.4 OOD-generalization under covariate distribution shift

We now study the OOD-generalization error due to the distribution shift with respect to the Gaussian *covariates* $\{\mathbf{y_1}, \ldots, \mathbf{y_n}\}$, i.e., the vectors at which a task matrix A is evaluated. The next proposition provides a quantitative upper bound for the generalization error in terms of the discrepancy between the covariance matrices. To simplify the proof, we use a Frobenius norm bound on the empirical risk minimizer. However, this choice of norm is not essential to the result.

Theorem 5. *Let $\Sigma = W\Lambda W^T$ and $\tilde{\Sigma} = \tilde{W}\tilde{\Lambda}\tilde{W}^T$ be the covariance matrices of Gaussian covariates used in the training and testing respectively. Let $(\widehat{P}, \widehat{Q})$ be minimizers of the empirical risk associated to covariates sampled from $N(0, \Sigma)$ and take $M > 0$ such that $\max\left(\|\widehat{P}\|_F, \|\widehat{Q}\|_F\right) \le M$. Then*

$$\mathcal{R}_m^{\tilde{\Sigma}}(\widehat{P}, \widehat{Q}) \lesssim \mathcal{R}_m^{\Sigma}(\widehat{P}, \widehat{Q}) + \|\Sigma - \tilde{\Sigma}\|_{op} + \frac{1}{m}\|W - \tilde{W}\|_{op}.$$

where the implicit constants depend on M, Σ, $\tilde{\Sigma}$, and the constant c_A defined in Assumption 1.

Theorem 5 states that the OOD-generalization error with respect to the covariate distribution shift is Lipschitz stable with respect to changes in the covariance matrix. However, unlike the case of task distribution shift, the covariate distribution shift error can not be mitigated by increasing the prompt-length in the downstream task; see also Figure 3. A similar phenomenon was observed in Zhang et al. [2023].

4 Numerical experiments

4.1 In-domain generalization

We first investigate numerically the neural scaling law of the transformer model for solving the linear system associated to the Galerkin discretization of the elliptic PDE (1) in the setting of in-domain generalization. More precisely, we consider the one dimensional elliptic PDE $(-\Delta + V(x))u(x) = f(x)$ on $\Omega = [0, 1]$ with Dirichlet boundary condition. We assume that the source $f \sim N(0, \mathbb{I})$, where $\mathbb{I}$ denotes the identity operator. We discretize the PDE using Galerkin projection under d sine bases. Further we assume that the potential V is uniform random field that is obtained by dividing the domain into $2d + 1$ sub-intervals and in each cell independently, the potential takes values uniformly in $[1, 2]$. In Figure 1: A-C, we demonstrate the empirical scaling law of the linear transformer for learning the discrete linear system by showing the log-log plots of the ℓ^2-errors as functions of the number of pre-training tasks N, the sequence length n during training and the sequence length m at inference. These numerical results suggest that the decaying rates of the prediction errors are $O(N^{-\frac{1}{2}}), O(n^{-2})$ and $O(m^{-1})$ respectively, which agree with the rates predicted in Theorem 1 in the practical regime $m \leq n$. We also demonstrate the ICL-generalization error for learning the PDE solutions. Figure 1:D shows that prediction error increases as d increases indicating that ICL of the linear system becomes harder as the d increases.

Figure 2:B shows the H^1-error curve between the numerical solution predicted by the ICL-model and the ground-truth as a function of the number of bases d, while fixing the prompt-lengths and the number of tasks. The U-shaped curve indicates the trade-off between the dimension of the discrete problem and the amount of data. More details on the experiment set-ups can be found in Appendix H.

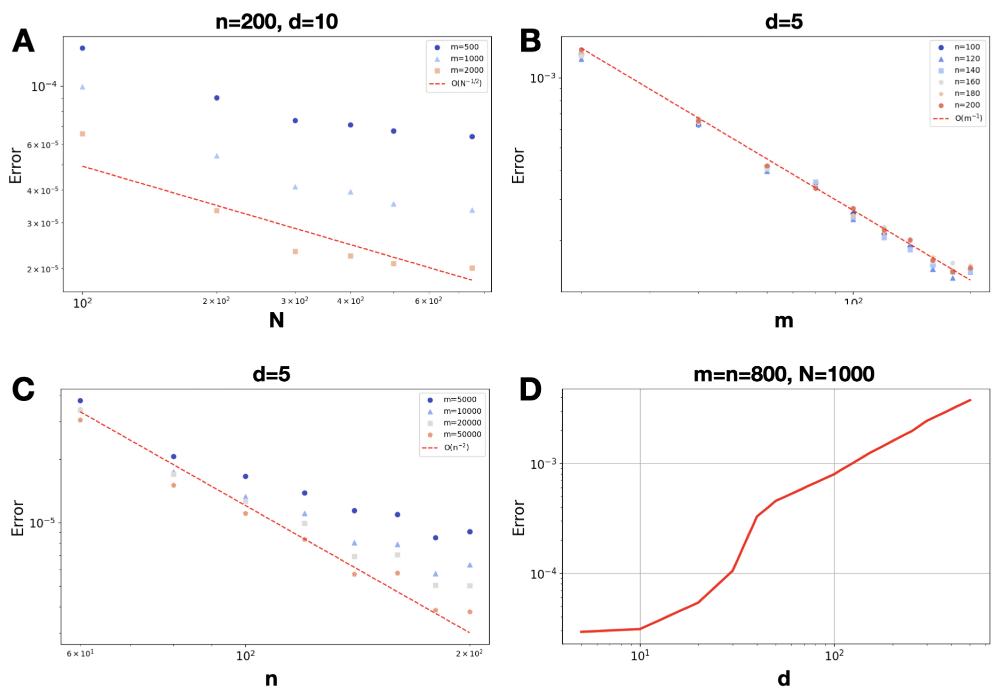

Figure 1: The figures A-D show the log-log plots for the ℓ^2-error of learning the linear system associated to the PDE discretization with respect to the number of tasks N, the prompt length n during training, the prompt length m during inference, and the dimension d of the linear system.

4.2 Out-of-domain generalization

Task shifts. We validate the ICL-capability of pre-trained transformers for learning the linear systems and PDEs under task distribution shifts. Specifically, for the PDE (1) in one dimension, we consider the task distribution shifts in a and V exclusively. To sample $a(x)$, we write $a(x) = e^{b(x)}$, where $b(x)$ is sampled from a centered normal distribution with covariance operator $-(\Delta + \tau\mathbb{I})^{-\alpha}$, for

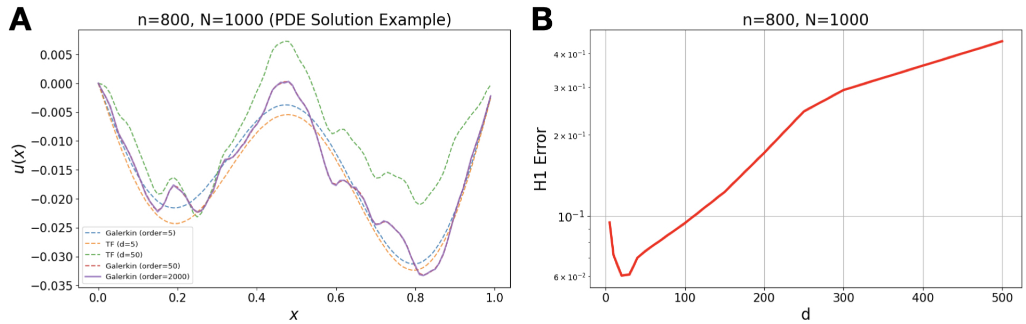

Figure 2: The left plot shows the PDE solution defined by the pre-trained transformer with the reference solution, obtained by Galerkin's method with 2000 basis functions. The right plot shows the H^1-error between the solution predicted by the transformer and reference solution with respect to the number of Galerkin basis functions d.

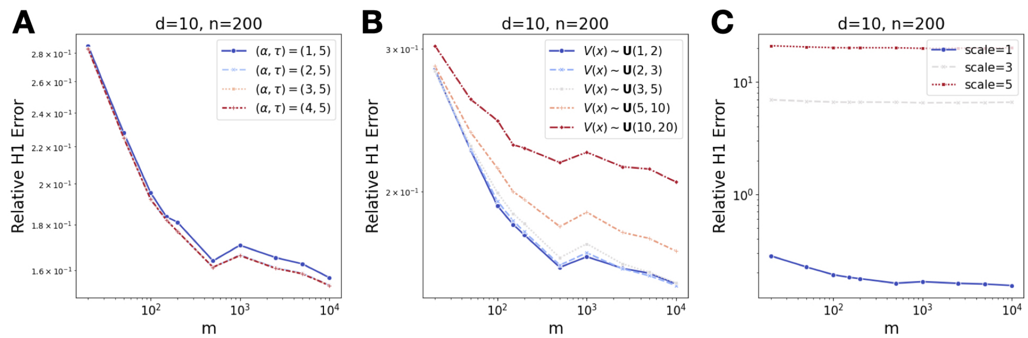

Figure 3: Figures A, B show the relative H^1-error under shifts on $a(x)$ and $V(x)$ respectively. Figure C shows the relative H^1-error under the covariate shift on the source term f.

$\alpha, \tau > 0$. During training, we set $\alpha = 3$ and $\tau = 5$, and during inference, we vary the values of α and τ according to Figure 3: A. We assume the potential V is piecewise constant on $2d+1$ subintervals and that the value of V on each cell is drawn according to the uniform distribution on $[a, b]$. During training, we set $[a, b] = [1, 2]$, and we vary the values of $[a, b]$ at inference according to Figure 3: B. For further details on the experimental setup, see in Appendix H. Figure 3: A shows that the pre-trained transformer can perform equally well on tasks on smoother a but perform slightly worse on tasks with less regular a. Figure 3: B shows the OOD-generalization errors increase as the distribution shift in V becomes stronger, but they decrease as the context length at inference increases, as predicted by Theorem 3.

Covariate shifts. Finally, we test the performance of the pre-trained transformer under covariate distribution shifts. Specifically, we train the model to solve the PDE (1), where the source term $f \sim N(0, C)$ for $C = (-\Delta + c\mathbb{I})^{-\beta}$, where $c, \beta > 0$ are fixed. Then, at inference, we consider solving the same PDE, but where the source term is defined by $N(0, 3C)$ or $N(0, 5C)$. Figure 3 show that the pre-trained transformers are not robust to covariate distribution shifts. We refer to Figure 5 in the appendix for additional numerical results for the covariant shifts in c and β.

5 Conclusion

In this work, we studied the ability of a transformer characterized by a single linear self-attention layer to in-context learn the solution operator of a linear elliptic PDE. We characterized the role of the number of pre-training task, the number of in-context examples during pre-training and testing, the mesh size, and various distribution shifts on the PDE coefficients in the overall PDE recovery error. We also provided thorough numerical experiments to demonstrate our theory. There are several natural extensions of this work, such as to nonlinear and time-dependent PDE problems. In these more complex settings, it is crucial to characterize the role that depth and nonlinearity play in the ability of transformers to approximate the PDE solution. We leave these directions to future work.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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*, 30, 2017.

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.

Ilyes Batatia, Philipp Benner, Yuan Chiang, Alin M Elena, Dávid P Kovács, Janosh Riebesell, Xavier R Advincula, Mark Asta, William J Baldwin, Noam Bernstein, et al. A foundation model for atomistic materials chemistry. *arXiv preprint arXiv:2401.00096*, 2023.

Albi Celaj, Alice Jiexin Gao, Tammy TY Lau, Erle M Holgersen, Alston Lo, Varun Lodaya, Christopher B Cole, Robert E Denroche, Carl Spickett, Omar Wagih, et al. An rna foundation model enables discovery of disease mechanisms and candidate therapeutics. *bioRxiv*, pages 2023–09, 2023.

Oscar Méndez-Lucio, Christos Nicolaou, and Berton Earnshaw. Mole: a molecular foundation model for drug discovery. *arXiv preprint arXiv:2211.02657*, 2022.

Shashank Subramanian, Peter Harrington, Kurt Keutzer, Wahid Bhimji, Dmitriy Morozov, Michael W Mahoney, and Amir Gholami. Towards foundation models for scientific machine learning: Characterizing scaling and transfer behavior. *Advances in Neural Information Processing Systems*, 36, 2024.

Michael McCabe, Bruno Régaldo-Saint Blancard, Liam Holden Parker, Ruben Ohana, Miles Cranmer, Alberto Bietti, Michael Eickenberg, Siavash Golkar, Geraud Krawezik, Francois Lanusse, et al. Multiple physics pretraining for physical surrogate models. *arXiv preprint arXiv:2310.02994*, 2023.

Zhanhong Ye, Xiang Huang, Leheng Chen, Hongsheng Liu, Zidong Wang, and Bin Dong. Pdeformer: Towards a foundation model for one-dimensional partial differential equations. *arXiv preprint arXiv:2402.12652*, 2024.

Liu Yang, Siting Liu, Tingwei Meng, and Stanley J Osher. In-context operator learning with data prompts for differential equation problems. *Proceedings of the National Academy of Sciences*, 120 (39):e2310142120, 2023.

Jingmin Sun, Yuxuan Liu, Zecheng Zhang, and Hayden Schaeffer. Towards a foundation model for partial differential equation: Multi-operator learning and extrapolation. *arXiv preprint arXiv:2404.12355*, 2024.

Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707, 2019.

Lu Lu, Pengzhan Jin, and George Em Karniadakis. Deeponet: Learning nonlinear operators for identifying differential equations based on the universal approximation theorem of operators. *arXiv preprint arXiv:1910.03193*, 2019.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv:2303.12712*, 2023.

Louis Kirsch, James Harrison, Jascha Sohl-Dickstein, and Luke Metz. General-purpose in-context learning by meta-learning transformers. *arXiv preprint arXiv:2212.04458*, 2022.

Wuyang Chen, Jialin Song, Pu Ren, Shashank Subramanian, Dmitriy Morozov, and Michael W Mahoney. Data-efficient operator learning via unsupervised pretraining and in-context learning. *arXiv preprint arXiv:2402.15734*, 2024a.

Liu Yang and Stanley J Osher. Pde generalization of in-context operator networks: A study on 1d scalar nonlinear conservation laws. *arXiv preprint arXiv:2401.07364*, 2024.

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.

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*, 2022.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pages 35151–35174. PMLR, 2023.

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.

Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. Trained transformers learn linear models in-context. *arXiv preprint arXiv:2306.09927*, 2023.

Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *Advances in neural information processing systems*, 36, 2024.

Xiang Cheng, Yuxin Chen, and Suvrit Sra. Transformers implement functional gradient descent to learn non-linear functions in context. *arXiv preprint arXiv:2312.06528*, 2023.

Juno Kim, Tai Nakamaki, and Taiji Suzuki. Transformers are minimax optimal nonparametric in-context learners. *arXiv preprint arXiv:2408.12186*, 2024.

Licong Lin, Yu Bai, and Song Mei. Transformers as decision makers: Provable in-context reinforcement learning via supervised pretraining. *arXiv preprint arXiv:2310.08566*, 2023.

Siyu Chen, Heejune Sheen, Tianhao Wang, and Zhuoran Yang. Training dynamics of multi-head softmax attention for in-context learning: Emergence, convergence, and optimality. *arXiv preprint arXiv:2402.19442*, 2024b.

Susanne Brenner and Ridgway Scott. *The Mathematical Theory of Finite Element Methods*, volume 15. Springer Science & Business Media, 2007.

Jie Shen, Tao Tang, and Li-Lian Wang. *Spectral methods: algorithms, analysis and applications*, volume 41. Springer Science & Business Media, 2011.

Nilesh Tripuraneni, Michael Jordan, and Chi Jin. On the theory of transfer learning: The importance of task diversity. *Advances in neural information processing systems*, 33:7852–7862, 2020.

Youssef Mroueh. Towards a statistical theory of learning to learn in-context with transformers. In *NeurIPS 2023 Workshop Optimal Transport and Machine Learning*, 2023.

Frank Cole and Yulong Lu. Score-based generative models break the curse of dimensionality in learning a family of sub-gaussian probability distributions. *arXiv preprint arXiv:2402.08082*, 2024.

Jiyoung Park, Ian Pelakh, and Stephan Wojtowytsch. Minimum norm interpolation by perceptra: Explicit regularization and implicit bias. *Advances in Neural Information Processing Systems*, 36, 2023.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Richard M Dudley. The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967.

Alexandre Ern and Jean-Luc Guermond. *Theory and practice of finite elements*, volume 159. Springer, 2004.

Daniele Boffi. Finite element approximation of eigenvalue problems. *Acta numerica*, 19:1–120, 2010.

Michael E Sander, Raja Giryes, Taiji Suzuki, Mathieu Blondel, and Gabriel Peyré. How do transformers perform in-context autoregressive learning? *arXiv preprint arXiv:2402.05787*, 2024.

Gilbert Strang. *Introduction to linear algebra*. SIAM, 2022.

Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. 2013.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

A Notation

Before delving into the proofs of our main results, we briefly go over all relevant notation:

- Physical dimension of PDE problem: d_0
- Dimension of task matrix for ICL: d
- Task matrix for ICL: A
- Covariates for ICL: $\{\mathbf{y_1}, \ldots, \mathbf{y_n}\}$
- Prompt matrix for ICL: Z
- Empirical covariance matrix of $\{\mathbf{y_1}, \ldots, \mathbf{y_n}\}$: Y_n
- Distribution on tasks: p_A
- Upper bound on largest eigenvalue of A^{-1} over $\mathrm{supp}(p_A)$: c_A
- Covariance operator of the distribution on $L^2(\Omega)$-valued covariates: Σ_f
- Covariance matrix of the distribution on $\mathbb{R}^d$-valued covariates: Σ
- Parameters of transformer: $\theta = (P, Q)$
- Prediction of the transformer with parameters θ: $\mathrm{TF}_\theta(Z)$
- Population risk for training: $\mathcal{R}_n$
- Population risk for inference: $\mathcal{R}_m$
- Empirical risk: $\mathcal{R}_{n,N}$
- "Infinite-context" population risk: $\mathcal{R}_\infty$
- Number of context examples per prompt during training: n
- Number of context examples per prompt during inference: m
- Number of pre-training tasks: N

B Proofs for Subsection 3.1

In this section we prove Theorem 1, which controls the (in-distribution) generalization error for in-context learning of linear systems in terms of the context length during training, the context length during inference, and the number of pre-training tasks. Before the proof, we present a more precise statement of the theorem.

Theorem 6 (Theorem 1, precise version)**.** *Let $\widehat{\theta} = (P_N, Q_N) \in argmin_{\|\theta\| \le M} \mathcal{R}_{n,N}(\theta)$, where $\|\theta\| := \max\left(\|P\|_{op}, \|Q\|_{op}\right)$. Then for n sufficiently large, we have with probability $\ge 1 - poly(N)$*

$$\begin{aligned}\mathcal{R}_m(\widehat{\theta}) \lesssim \frac{(c_A^2 + d)Tr(\Sigma)}{m} &+ \frac{c_A^2 C_A^4 \|\Sigma\|_{op}^2 \|\Sigma^{-1}\|_{op}^2 \left(1 + Tr_\Sigma(\mathbb{E}_{A\sim p_A}[A^{-2}])\right)^2 Tr(\Sigma)}{n^2} \\ &+ \frac{d^2 c_A^2 \|\Sigma\|_{op}^2 \max(1, \|\Sigma^{-1}\|_{op})^4}{\sqrt{N}} \\ &+ \max(1, \|\Sigma^{-1}\|_{op})^4 c_A^2 \max(Tr(\Sigma), \|\Sigma\|_{op}^2) Tr(\Sigma) \Big|\frac{1}{n} - \frac{1}{m}\Big|,\end{aligned} \tag{11}$$

where we have omitted factors which are polylog in N.

Remark 1. *We would like to comment on the possible suboptimality of the bound* (11)*. Specifically, the last term on the right side of* (11)*, which we term the "context mismatch error", is mainly due to our proof strategy and can likely be removed with a refined analysis. This term is not observed in our numerical experiments; see Figure 1. In the practical* [2] *regime where the length of the testing prompts is less than that of the training prompts (i.e. $m \le n$), we have $\left|\frac{1}{n} - \frac{1}{m}\right| \le \frac{1}{m}$, and hence the context-mismatch error is absorbed into the $O\left(\frac{1}{m}\right)$ term, leading to the following overall generalization bound*

$$\mathcal{R}_m(\widehat{\theta}) \lesssim \frac{1}{m} + \frac{1}{n^2} + \frac{1}{\sqrt{N}}. \tag{12}$$

Proof of Theorem 1. **Step 1 - error decomposition:** Throughout the proof, we use the notation $\theta = (P, Q)$ and $\|\theta\| = \max(\|P\|_{\mathrm{op}}, \|Q\|_{\mathrm{op}})$. Write $\ell(A, Y_n, \mathbf{y}; \theta) = \|(PA^{-1}Y_nQ - A^{-1})\mathbf{y}\|^2$, so that the risk functionals can be expressed as

$$\mathcal{R}_n(\theta) = \mathbb{E}_{A,Y_n,\mathbf{y}} \ell(A, Y_n, \mathbf{y}; \theta), \quad \mathcal{R}_{n,N}(\theta) = \frac{1}{N}\sum_{i=1}^N \ell(A_i, Y_n^{(i)}, \mathbf{y}_i; \theta).$$

Let us introduce an auxiliary parameters $t > 0$ – to be specified precisely at the end of the proof – and define the events

$$\mathcal{A}_t(Y_n, \mathbf{y}) = \left\{\|\mathbf{y}\| \le \sqrt{\mathrm{Tr}(\Sigma)} + t,\ \|Y_n\|_{\mathrm{op}} \le \|\Sigma\|_{\mathrm{op}}\left(1 + t + \sqrt{\frac{d}{n}}\right)\right\}.$$

Define the truncated loss function as $\ell^{R,t}(A, Y_n, \mathbf{y}; \theta) = \ell(A, Y_n, \mathbf{y}; \theta) \cdot 1\{\mathcal{A}_{R,t}\}(Y_n, \mathbf{y})$, and let $\mathcal{R}_n^t$, $\mathcal{R}_{n,N}^t$, and $\mathcal{R}_m^t$ denote the associated truncated risk functionals. Further, let θ^* denote a fixed parameter, to be specified later on. We decompose the generalization error into a sum of approximation error, statistical error conditioned on the data being bounded, and truncation error that

[2] The performance of GPTs is known to deteriorate when the test sequence length exceeds the train sequence length; Zhang et al. [2023] conjectures this phenomenon to be the result of positional encoding.

leverages the tail decay of the data distribution. In more detail, we have

$$\mathcal{R}_m(\widehat{\theta}) = \Big(\mathcal{R}_m(\widehat{\theta}) - \mathcal{R}^t_m(\widehat{\theta})\Big) + \Big(\mathcal{R}^t_m(\widehat{\theta}) - \mathcal{R}^t_{m,N}(\widehat{\theta})\Big) + \Big(\mathcal{R}^t_{m,N}(\widehat{\theta}) - \mathcal{R}^t_{m,N}(\theta^*)\Big) \tag{13}$$

$$+ \Big(\mathcal{R}^t_{m,N}(\theta^*) - \mathcal{R}^t_m(\theta^*)\Big) + \Big(\mathcal{R}^t_m(\theta^*) - \mathcal{R}_m(\theta^*)\Big) + \mathcal{R}_m(\theta^*) \tag{14}$$

$$\leq \sup_{\|\theta\|\leq M} \Big(\mathcal{R}_m(\theta) - \mathcal{R}^t_m(\theta)\Big) + 2 \sup_{\|\theta\|\leq M} \Big|\mathcal{R}^t_m(\theta) - \mathcal{R}^t_{m,N}(\theta)\Big| \tag{15}$$

$$+ \Big(\mathcal{R}^t_{m,N}(\widehat{\theta}) - \mathcal{R}^t_{m,N}(\theta^*)\Big) + \inf_{\|\theta^*\|\leq M} \mathcal{R}(\theta^*). \tag{16}$$

where we discarded the nonpositive term $\Big(\mathcal{R}^t(\theta^*) - \mathcal{R}(\theta^*)\Big)$. This decomposition mimics the standard decomposition of generalization error into approximation and statistical errors, with an additional term that arises from truncating the data. Similar techniques have recently been used in Cole and Lu [2024] and Park et al. [2023]. There is one more technical detail to be addressed. We would like to say that the term $\Big(\mathcal{R}^t_{m,N}(\widehat{\theta}) - \mathcal{R}^t_{m,N}(\theta^*)\Big)$ is nonpositive with high probability, as a consequence of the minimality of $\widehat{\theta}$. However, the parameter $\widehat{\theta}$ is a minimizer of the empirical risk $\mathcal{R}_{n,N}$ corresponding to the context length n during training, as opposed to the empirical risk $\mathcal{R}_{m,N}$ corresponding to the context length m during inference. However, it is easy to see that the following bound holds

$$\mathcal{R}^t_{m,N}(\widehat{\theta}) - \mathcal{R}^t_{m,N}(\theta^*) \leq 2 \sup_{\|\theta\|\leq M} \Big(\mathcal{R}^t_{m,N}(\theta) - \mathcal{R}^t_m(\theta)\Big) + 2 \sup_{\|\theta\|\leq M} \Big(\mathcal{R}^t_{n,N}(\theta) - \mathcal{R}^t_n(\theta)\Big) \tag{17}$$

$$+ \sup_{\|\theta\|\leq M} \Big(\mathcal{R}_m(\theta) - \mathcal{R}^t_m(\theta)\Big) + \sup_{\|\theta\|\leq M} \Big(\mathcal{R}_n(\theta) - \mathcal{R}^t_n(\theta)\Big) \tag{18}$$

$$+ 2 \sup_{\|\theta\|\leq M} \Big|\mathcal{R}_m(\theta) - \mathcal{R}_n(\theta)\Big| + \Big(\mathcal{R}^t_{n,N}(\widehat{\theta}) + \mathcal{R}^t_{n,N}(\theta^*)\Big). \tag{19}$$

Plugging the estimate 17 into the bound from 13 gives the final bound

$$\mathcal{R}_m(\widehat{\theta}) \leq \underbrace{2 \sup_{\|\theta\|\leq M} \Big(\mathcal{R}_m - \mathcal{R}^t_m\Big)(\theta) + \sup_{\|\theta\|\leq M} \Big(\mathcal{R}_n - \mathcal{R}^t_n\Big)(\theta)}_{\text{data truncation error}} \tag{20}$$

$$\underbrace{+ 4 \sup_{\|\theta\|\leq M} \Big(\mathcal{R}^t_m - \mathcal{R}^t_{m,N}\Big)(\theta) + 2 \sup_{\|\theta\|\leq M} \Big(\mathcal{R}^t_n - \mathcal{R}^t_{n,N}\Big)(\theta)}_{\text{statistical error}} \tag{21}$$

$$\underbrace{+ 2 \sup_{\|\theta\|\leq M} \Big|\mathcal{R}_m(\theta) - \mathcal{R}_n(\theta)\Big|}_{\text{context mismatch error}} + \underbrace{\Big(\mathcal{R}^t_{n,N}(\widehat{\theta}) - \mathcal{R}^t_{n,N}(\theta^*)\Big)}_{\leq 0 \text{ w.h.p.}} + \underbrace{\mathcal{R}_m(\theta^*)}_{\text{approx. error}} \tag{22}$$

$$= I + II + III + IV + V. \tag{23}$$

The plan of action is to bound term I using the tail decay of the data and term II using tools from empirical process theory; term III is controlled via Lemma 12; term IV can be shown to be nonpositive with high-probability, and term V, the approximation error, is controlled by Proposition 1.

Step 2 - bounding the truncation error: By Lemma 7 and Example 6.2 in Wainwright [2019], when $\mathbf{y} \sim N(0, \Sigma)$ and Y_n is the empirical covariance of iid samples from $N(0, \Sigma)$ we have

$$P(\mathcal{A}^c_t(Y_n, \mathbf{y})) \leq \exp\Big(-\frac{nt^2}{2}\Big) + \exp\Big(-\frac{t^2}{C\|\Sigma\|_{\text{op}}}\Big)$$

for some universal constant $C > 0$. Therefore, for any $\|\theta\| \le M$, we can apply the Cauchy-Schwarz inequality to obtain

$$\begin{aligned}
\mathcal{R}_m(\theta) - \mathcal{R}^t_m(\theta) &= \mathbb{E}\|(PA^{-1}Y_mQ - A^{-1})\mathbf{y}\|^2 \cdot 1\{\mathcal{A}^c_{R,t}(Y_m, \mathbf{y})\} \\
&\le \Big(\mathbb{E}\|(PA^{-1}Y_mQ - A^{-1})\mathbf{y}\|^4\Big)^{1/2} \cdot \mathbb{P}\Big(\mathcal{A}^c_{R,t}(Y_m, \mathbf{y})\Big)^{1/2} \\
&\le c_A^2\Big(M^2\Big(\mathbb{E}\|Y_n\|^4_{\text{op}}\Big)^{1/2} + 1\Big)\Big(\mathbb{E}\|\mathbf{y}\|^4\Big)^{1/2} \cdot \sqrt{\exp\Big(-\frac{mt^2}{2}\Big) + \exp\Big(-\frac{t^2}{C\|\Sigma\|_{\text{op}}}\Big)}.
\end{aligned}$$

This shows that the truncation error is quite mild, since R and t can be made large – in fact, we will see that the generalization error depends only poly-logarithmically on R. Analogous bounds hold for $\sup_{\|\theta\|\le M}\Big(\mathcal{R}_n - \mathcal{R}^t_n\Big)(\theta)$.

Step 3 - Reduction to bounded data: Note that by the union bound,

$$\mathcal{B}_{N,t} := \bigcap_{i=1}^N \mathcal{A}_t(Y_n^{(i)}, \mathbf{y_i})$$

satisfies

$$\mathbb{P}(\mathcal{B}_{N,R,t}) \ge 1 - N\Big(\exp\Big(-\frac{nt^2}{2}\Big) + \exp\Big(-\frac{t^2}{C\|\Sigma\|_{\text{op}}}\Big)\Big).$$

Moreover, on the event $\mathcal{B}_{N,t}$, we have $\ell(\cdot;\theta) = \ell^{R,t}(\cdot;\theta)$, and hence $\widehat{\theta} = \text{argmin}_{\|\theta\|\le M}\mathcal{R}^t_N(\theta)$. Therefore, if we restrict attention to the event $\mathcal{B}_{N,R,t}$, we may assume boundedness of the data, which is crucial to proving statistical error bounds, and the error term

$$IV = \Big(\mathcal{R}^t_N(\widehat{\theta}) - \mathcal{R}^t_N(\theta^*)\Big)$$

is nonpositive by the minimality of $\mathcal{R}^t_N(\widehat{\theta})$. For the remainder of the proof, we assume that the event $\mathcal{B}_{N,R,t}$ holds, i.e., all expectations taken are conditioned on the event $\mathcal{B}_{N,R,t}$.

Step 4 - bounding the statistical error: The statistical error is measured by

$$\begin{aligned}
&\sup_{\|\theta\|\le M}\Big|\mathcal{R}^t_n(\theta) - \mathcal{R}^t_{n,N}(\theta)\Big| \\
&= \sup_{\|\theta\|\le M}\Big|\mathbb{E}_{A,Y_n,\mathbf{y}}\|(PA^{-1}Y_nQ - A^{-1})\mathbf{y}\|^2 - \frac{1}{N}\sum_{i=1}^N \|(PA_i^{-1}Y_n^{(i)}Q - A_i^{-1})\mathbf{y_i}\|^2\Big|,
\end{aligned}$$

where the expectations over Y_n and $\mathbf{y}$ are over truncated versions of their original distributions. By a standard symmetrization argument, we have

$$\begin{aligned}
&\sup_{\|\theta\|\le M}\Big|\mathbb{E}_{A,Y_n,\mathbf{y}}[\|(PA^{-1}Y_nQ - A^{-1})\mathbf{y}\|^2] - \frac{1}{N}\sum_{i=1}^N \|(PA_i^{-1}Y_n^{(i)}Q - A_i^{-1})\mathbf{y_i}\|^2\Big| \\
&\le 2\mathbb{E}_{A_i,Y_n^{(i)},\mathbf{y_i}}\mathbb{E}_{\epsilon_i}\sup_{\|\theta\|\le M}\frac{1}{N}\sum_{i=1}^N \epsilon_i\|(PA_i^{-1}Y_n^{(i)}Q - A_i^{-1})\mathbf{y_i}\|^2 \\
&= 2\mathbb{E}_{A_i,Y_n^{(i)},\mathbf{y_i}}\mathbb{E}_{\epsilon_i}\sup_{\|\theta\|\le M}\frac{1}{N}\sum_{i=1}^N \epsilon_i\Big(\|PA_i^{-1}Y_n^{(i)}Q\mathbf{y_i}\|^2 + \|A_i^{-1}\mathbf{y_i}\|^2 - 2\langle PA_i^{-1}Y_n^{(i)}Q\mathbf{y_i}, A_i^{-1}\mathbf{y_i}\rangle\Big) \\
&\le 2\mathbb{E}_{A_i,Y_n^{(i)},\mathbf{y_i}}\mathbb{E}_{\epsilon_i}\sup_{\|\theta\|\le M}\frac{1}{N}\sum_{i=1}^N \epsilon_i\|PA_i^{-1}Y_n^{(i)}Q\mathbf{y_i}\|^2 \\
&+ 4\mathbb{E}_{A_i,Y_n^{(i)},\mathbf{y_i}}\mathbb{E}_{\epsilon_i}\sup_{\|\theta\|\le M}\frac{1}{N}\sum_{i=1}^N \epsilon_i\langle PA_i^{-1}Y_n^{(i)}Q\mathbf{y_i}, A_i^{-1}\mathbf{y_i}\rangle,
\end{aligned}$$

where the last inequality follows from the triangle inequality, noting that the term $\sum_{i=1}^N \epsilon_i\|A_i^{-1}\mathbf{y_i}\|^2$ is independent of θ and hence vanishes in the expectation over ϵ_i. Now, define the function classes

$$\begin{aligned}
\Theta_1(M) &= \{(A, Y_n, \mathbf{y}) \mapsto \|PA^{-1}Y_nQ\mathbf{y}\|^2 : \|\theta\| \le M\}, \\
\Theta_2(M) &= \{(A, Y_n, \mathbf{y}) \mapsto \langle PA^{-1}Y_nQ\mathbf{y}, A^{-1}\mathbf{y}\rangle : \|\theta\| \le M\}.
\end{aligned}$$

By Dudley's integral theorem Dudley [1967], it holds that

$$\mathbb{E}_{A_i, Y_n^{(i)}, \mathbf{y_i}} \mathbb{E}_{\epsilon_i} \sup_{\|\theta\| \le M} \frac{1}{N} \sum_{i=1}^{N} \epsilon_i \|PA_i^{-1} Y_n^{(i)} Q \mathbf{y_i}\|^2 \le \inf_{\epsilon > 0} \frac{12\sqrt{2}}{\sqrt{N}} \int_{\epsilon}^{D_1(M)} \sqrt{\log \mathcal{N}\Big(\Theta_1(M), \|\cdot\|_N, \tau\Big)} d\tau, \tag{24}$$

where $\mathcal{N}\Big(\Theta_1(M), \|\cdot\|_N, \tau\Big)$ is the τ-covering number of the function class $\Theta_1(M)$ with respect to the metric induced by the empirical L^2 norm $\|F\|_N^2 = \frac{1}{N}\sum_{i=1}^N F(A_i, Y_n^{(i)}, \mathbf{y_i})^2$ and

$$D_1(M) = \sup_{\|\theta\| \le M} \Big\| \|PA^{-1} Y_n Q \mathbf{y}\|^2 \Big\|_N.$$

Note the bound

$$\begin{aligned} D_1(M)^2 &= \sup_{\|\theta\| \le M} \frac{1}{N} \sum_{i=1}^{N} \|PA_i^{-1} Y_n^{(i)} Q \mathbf{y_i}\|^4 \\ &\le \frac{1}{N} \sum_{i=1}^{N} M^8 c_A^4 \|\Sigma\|_{\text{op}}^4 \Big(1 + t + \sqrt{\frac{d}{n}}\Big)^4 \Big(\sqrt{\text{Tr}(\Sigma)} + t\Big)^4 \end{aligned}$$

and hence $D_1(M) \le M^4 c_A^2 \|\Sigma\|_{\text{op}}^2 \Big(1 + t + \sqrt{\frac{d}{n}}\Big)^2 \Big(\sqrt{\text{Tr}(\Sigma)} + t\Big)^2$. Similarly, for $\theta_1 = (P_1, Q_1), \theta_2 = (P_2, Q_2)$, with $\|\theta_1\|, \|\theta_2\| \le M$, we have

$$\begin{aligned} \|\theta_1 - \theta_2\|_N^2 &= \frac{1}{N} \sum_{i=1}^{N} \|(P_1 - P_2) A_i^{-1} Y_n^{(i)} (Q_1 - Q_2) \mathbf{y_i}\|^4 \\ &\le 16 M^4 c_A^2 \|\Sigma\|_{\text{op}}^2 \Big(1 + t + \sqrt{\frac{d}{n}}\Big)^2 R^2 \cdot \frac{1}{N} \sum_{i=1}^{N} \|(P_1 - P_2) A_i^{-1} Y_n^{(i)} (Q_1 - Q_2)\|^2 \\ &\le M^4 c_A^4 \|\Sigma\|_{\text{op}}^4 \Big(1 + t + \sqrt{\frac{d}{n}}\Big)^4 \Big(\sqrt{\text{Tr}(\Sigma)} + t\Big)^4 \cdot \max\Big(\|P_1 - P_2\|_{\text{op}}^2, \|Q_1 - Q_2\|_{\text{op}}^2\Big). \end{aligned}$$

This shows that the metric induced by $\|\cdot\|_N$ is dominated by the metric $d(\theta_1, \theta_2) = \max\Big(\|P_1 - P_2\|_{\text{op}}, \|Q_1 - Q_2\|_{\text{op}}\Big)$, up to a factor of $M^2 c_A^2 \|\Sigma\|_{\text{op}}^2 \Big(1 + t + \sqrt{\frac{d}{n}}\Big)^2 \Big(\sqrt{\text{Tr}(\Sigma)} + t\Big)^2$. The covering number of the set $\{\|\theta\| \le M\}$ in the metric $d(\cdot, \cdot)$ is well-known, from which we conclude that

$$\log \mathcal{N}\Big(\Theta_1(M), \|\cdot\|_N, \tau\Big) \le 2d^2 \log\Big(M^2 c_A^2 \|\Sigma\|_{\text{op}}^2 \Big(1 + \frac{2}{\tau}\Big)\Big).$$

Optimizing over the choice of ϵ in Equation 24, this proves that

$$\mathbb{E}_{A_i, Y_n^{(i)}, \mathbf{y_i}} \mathbb{E}_{\epsilon_i} \sup_{\|\theta\| \le M} \frac{1}{N} \sum_{i=1}^{N} \epsilon_i \|PA_i^{-1} Y_n^{(i)} Q \mathbf{y_i}\|^2 \tag{25}$$

$$= O\Big(\frac{d^2 M^4 c_A^2 \|\Sigma\|_{\text{op}}^2 \Big(1 + t + \sqrt{\frac{d}{n}}\Big)^2 \Big(\sqrt{\text{Tr}(\Sigma)} + t\Big)^2}{\sqrt{N}}\Big), \tag{26}$$

where $O\Big(\cdot\Big)$ omits factors that are logarithmic in N. An analogous argument proves a bound of the same order on the quantity

$$\mathbb{E}_{A_i, Y_n^{(i)}, \mathbf{y_i}} \mathbb{E}_{\epsilon_i} \sup_{\|\theta\| \le M} \frac{1}{N} \sum_{i=1}^{N} \epsilon_i \langle PA_i^{-1} Y_n^{(i)} Q \mathbf{y_i}, A_i^{-1} \mathbf{y_i} \rangle,$$

which in turn bounds the statistical error

$$\sup_{\|\theta\| \le M} \Big| \mathcal{R}_n^t(\theta) - \mathcal{R}_{n,N}^t(\theta) \Big|$$

by the right-hand side of Equation 25. The same argument proves in analogous bound on the statistical error term

$$\sup_{\|\theta\|\le M}\Big|\mathcal{R}^t_m(\theta)-\mathcal{R}^t_{m,N}(\theta)\Big|,$$

where n is replaced by m in the bound of Equation 25.

Step 5: Bounding the context mismatch error The context mismatch error satisfies the bound

$$\sup_{\|\theta\|\le M}\Big|\mathcal{R}_m(\theta)-\mathcal{R}_n(\theta)\Big|\le 2M^4c_A^2\max(\mathrm{Tr}(\Sigma),\|\Sigma\|_{\mathrm{op}}^2)\mathrm{Tr}(\Sigma)\Big|\frac{1}{n}-\frac{1}{m}\Big|.$$

The proof of this fact is deferred to Lemma 12.

Step 6 - Approximation error: It remains to bound the approximation error term $\mathcal{R}(\theta^*)$. From Proposition 1, we have

$$\mathcal{R}_m(\theta^*)\le\frac{c_A^2\mathrm{Tr}(\Sigma)}{m}+\frac{c_A^6\|\Sigma^{-1}\|_{\mathrm{op}}^2\|\Sigma\|_{\mathrm{op}}^6\mathrm{Tr}(\Sigma)}{n^2}+O\Big(\frac{1}{mn}\Big)$$

for an appropriate choice of θ^*, where C_1 and C_2 depend only on the task and data distributions. Moreover, upon inspection of the proof of Proposition 1, we see that the $\theta^*=(\mathbf{I}_d,Q_n)$ that attains this error is an $O(1/n)$-perturbation of the pair $(\mathbf{I}_d,\Sigma^{-1})$. Thus, if n is sufficiently large, we are guaranteed that θ^* belongs in the set $\{\|\theta\|\le M\}$ for $M\ge 2\max(1,\|\Sigma^{-1}\|_{\mathrm{op}})$.

Step 7 - Balancing error terms: Putting everything together and applying the error decomposition from step 1, we have shown that [3]

$$\begin{aligned}\mathcal{R}_m(\widehat{\theta})&\lesssim c_A^2\Big(M^2\mathbb{E}[\|Y_n\|_{\mathrm{op}}^4]^{1/2}+1\Big)\mathbb{E}[\|\mathbf{y}\|^4]^{1/2}\cdot\sqrt{\exp\Big(-\frac{nt^2}{2}\Big)+\exp\Big(-\frac{t^2}{C\|\Sigma\|_{\mathrm{op}}}\Big)}\\&+\frac{d^2M^4c_A^2\|\Sigma\|_{\mathrm{op}}^2\Big(1+t+\sqrt{\frac{d}{n}}\Big)^2\Big(\sqrt{\mathrm{Tr}(\Sigma)}+t\Big)^2}{\sqrt{N}}+\frac{2\mathrm{Tr}(\mathbb{E}[A^{-2}]\Sigma)}{n},\end{aligned}$$

with probability at least

$$1-N\Big(\exp\Big(-\frac{nt^2}{2}\Big)+\exp\Big(-\frac{t^2}{C\|\Sigma\|_{\mathrm{op}}}\Big)\Big).$$

For a fixed $p>0$, we choose t such that

$$\Big(\exp\Big(-\frac{nt^2}{2}\Big)+\exp\Big(-\frac{t^2}{C\|\Sigma\|_{\mathrm{op}}}\Big)\Big)=\frac{1}{N^{p+1}}.$$

It is clear that such a t satisfies $t\lesssim\sqrt{p\log(N)}$. For such a t, we have, omitting universal constants and $\log(N)$ factors, that

$$\begin{aligned}\mathcal{R}_m(\widehat{\theta})&\lesssim\frac{c_A^2\mathrm{Tr}(\Sigma)}{m}+\frac{c_A^6\|\Sigma^{-1}\|_{\mathrm{op}}^2\|\Sigma\|_{\mathrm{op}}^6\mathrm{Tr}(\Sigma)}{n^2}+\sqrt{p}\frac{c_A^2\Big(M^2\mathbb{E}[\|Y_n\|_{\mathrm{op}}^4]^{1/2}+1\Big)\mathbb{E}[\|\mathbf{y}\|^4]^{1/2}}{N}\\&+\frac{d^2M^4c_A^2\|\Sigma\|_{\mathrm{op}}^2}{\sqrt{N}}+M^4c_A^2\max(\mathrm{Tr}(\Sigma),\|\Sigma\|_{\mathrm{op}}^2)\mathrm{Tr}(\Sigma)\Big|\frac{1}{n}-\frac{1}{m}\Big|,\quad\text{w.p.}\ge 1-\frac{2}{N^p}.\end{aligned}$$

We omit the third term from the final bound, since, asymptotically, it is dominated by the fourth term. □

We now present an important preliminary result, which gives an upper bound on $\inf_\theta\mathcal{R}_m(\theta)$, the minimal risk achieved by a transformer in the infinite-task limit. To motivate our result, we

[3] For simplicity, we have omitted the terms from the truncation and statistical errors which depend on m, as they do not change the order of the final bound with respect to m, n, or N.

first observe that for $\theta = (P, Q)$, the output of the transformer TF_θ at a prompt Z of length m corresponding to a task matrix A is

$$\mathrm{TF}_\theta(Z) = P\Big(\frac{1}{m}\sum_{i=1}^m \mathbf{x_i}\mathbf{y_i}^T\Big)Q\mathbf{y}.$$

Since $\mathbf{x_i} = A^{-1}\mathbf{y_i}$, we can equivalently write the prediction of the transformer as

$$\mathrm{TF}_\theta(Z) = PA^{-1}Y_mQ\mathbf{y},$$

where $Y_m = \frac{1}{m}\sum_{i=1}^m \mathbf{y_i}\mathbf{y_i}^T$ is the empirical covariance associated to the context vectors $\{\mathbf{y_1}, \ldots, \mathbf{y_m}\}$. Note that if we set $P = Id$ and $Q = \Sigma^{-1}$ to be the inverse of the data covariance matrix, then for sufficiently large m we have $\mathrm{TF}_\theta(Z) \approx A^{-1}\mathbf{y}$. This suggests that the transformer can learn to solve linear systems in a way that is extremely robust to shifts in the distribution on the task matrices. We note that similar choices of attention matrices have been studied in the linear regression setting (Ahn et al. [2024], Zhang et al. [2023]). Our result essentially employs the parameterization $P = \mathbf{Id}_d$ and $Q = \Sigma^{-1}$, but with an additional bias term to account for the fact that the sequence length n during training may differ from the sequence length m during inference.

Before stating our result precisely, let us define $B := \mathbb{E}_{A\sim p_A}[A^{-2}]$. In addition, recall the weighted trace of a matrix K with respect to the covariance $\Sigma = W\Lambda W^T$ defined by

$$\mathrm{Tr}_\Sigma(K) := \sum_{i=1}^d \sigma_i^2\langle K\varphi_i, \varphi_i\rangle,$$

where $\sigma_1^2, \ldots, \sigma_d^2$ are the eigenvalues of Σ and $\varphi_i = We_i$ are the eigenvectors. Note that the weighted trace is independent of the choice of eigenbasis.

Proposition 1. *With*

$$Q_n = B\Big(\frac{n+1}{n}\Sigma B + \frac{Tr_\Sigma(B)}{n}\Sigma\Big)^{-1},$$

we have

$$\mathcal{R}_m(\mathbf{I}_d, Q_n) \le \frac{(c_A^2 + d)Tr(\Sigma)}{m} + \frac{c_A^2 C_A^4\|\Sigma\|_{op}^2\|\Sigma^{-1}\|_{op}^2\Big(1 + Tr_\Sigma(B)\Big)^2 Tr(\Sigma)}{n^2} + O\Big(\frac{1}{mn}\Big).$$

Proof. By Lemma 8, we can write $Q_n = \Sigma^{-1} + \frac{1}{n}K$, where

$$\|K\|_{\mathrm{op}} \le \|\Sigma^{-1}\|_{\mathrm{op}}\|\Sigma\|_{\mathrm{op}}\Big(1 + \mathrm{Tr}_\Sigma(B)\Big)C_A^2. \tag{27}$$

It follows that

$$\begin{aligned}
\mathcal{R}_m(\mathbf{I}_d, Q_n) &= \mathbb{E}_{A,Y_m}[\mathrm{Tr}(A^{-1}(Y_mQ_n - \mathbf{I}_d)\Sigma(Q_nY_m - \mathbf{I}_d)A^{-1})] \\
&= \mathbb{E}_{Y_m}[\mathrm{Tr}(B(Y_mQ_n - \mathbf{I}_d)\Sigma(Q_n^TY_m - \mathbf{I}_d))], \quad B := \mathbb{E}[A^{-2}] \\
&= \mathrm{Tr}(B\Sigma) + \mathbb{E}_{Y_m}[\mathrm{Tr}(BY_mQ_n\Sigma Q_n^TY_m)] - \mathrm{Tr}(B\Sigma Q_n\Sigma) - \mathrm{Tr}(B\Sigma Q_n^T\Sigma) \\
&= \mathrm{Tr}(B\Sigma) + \mathrm{Tr}(B\Sigma Q_n\Sigma Q_n\Sigma) - \mathrm{Tr}(B\Sigma Q_n\Sigma) - \mathrm{Tr}(B\Sigma Q_n^T\Sigma) \\
&+ \frac{1}{m}\Big(\mathrm{Tr}\Big(B\Sigma Q_n\Sigma Q_n^T\Sigma\Big) + \mathrm{Tr}_\Sigma(Q_n\Sigma Q_n^T)\mathrm{Tr}(B\Sigma)\Big)
\end{aligned}$$

where the last equality follows from Lemma 4. Writing $Q_n = \Sigma^{-1} + \frac{1}{n}K$ and doing some simplifying algebra, we find that

$$\begin{aligned}
\mathcal{R}_m(\mathbf{I}_d, Q_n) &= \frac{1}{m}\Big(\mathrm{Tr}((B + \mathrm{Tr}_\Sigma(\Sigma^{-1}\mathbf{I_d})\Sigma)\Big) + \frac{1}{n^2}\mathrm{Tr}\Big(B\Sigma K\Sigma K^T\Sigma\Big) + O\Big(\frac{1}{mn}\Big) \\
&= \frac{1}{m}\Big(\mathrm{Tr}((B + d\mathbf{I_d})\Sigma)\Big) + \frac{1}{n^2}\mathrm{Tr}\Big(B\Sigma K\Sigma K^T\Sigma\Big) + O\Big(\frac{1}{mn}\Big),
\end{aligned}$$

where we used the fact that $\mathrm{Tr}_\Sigma(\Sigma^{-1}) = d$. Using the bound on the norm of K stated in Equation 27, and the fact that $\|B\|_{\mathrm{op}} \le c_A^2$, we have

$$\mathrm{Tr}(B\Sigma K\Sigma K^T\Sigma) \le c_A^2C_A^4\|\Sigma\|_{\mathrm{op}}^2\|\Sigma^{-1}\|_{\mathrm{op}}^2\Big(1 + \mathrm{Tr}_\Sigma(B)\Big)^2\mathrm{Tr}(\Sigma).$$

Similarly, the bound

$$\mathrm{Tr}((B + d\mathbf{I_d})\Sigma) \leq (c_A^2 + d)\mathrm{Tr}(\Sigma)$$

holds. We conclude that

$$\mathcal{R}_m(\mathbf{I}_d, Q_n) \leq \frac{(c_A^2 + d)\mathrm{Tr}(\Sigma)}{m} + \frac{c_A^2 C_A^4 \|\Sigma\|_{\mathrm{op}}^2 \|\Sigma^{-1}\|_{\mathrm{op}}^2 \Big(1 + \mathrm{Tr}_\Sigma(B)\Big)^2 \mathrm{Tr}(\Sigma)}{n^2} + O\Big(\frac{1}{mn}\Big).$$

□

To justify our ansatz for upper bounding the approximation error (i.e., how the matrix Q_n in Proposition 1 was chosen), we introduce the following lemma.

Lemma 1. *The minimizer of the functional $Q \mapsto \mathcal{R}_n(\mathbf{I}_d, Q)$ is given by*

$$Q_n = B\Big(\frac{n+1}{n}\Sigma B + \frac{Tr_\Sigma(B)}{n}\Sigma\Big)^{-1},$$

where $B = \mathbb{E}[A^{-2}]$ and $Tr_\Sigma(\cdot)$ denotes the Σ-weighted trace.

Proof. Let us recall the definition of the population risk functional

$$\mathcal{R}(\mathbf{I}_d, Q) = \mathbb{E}\Big[\Big\|A^{-1}\Big(Y_n Q - I\Big)y\Big\|^2\Big],$$

where $Y_n := \frac{1}{n}\sum_{i=1}^n \mathbf{y_i}\mathbf{y_i}^T$ denotes the empirical covariance of $\{\mathbf{y_i}\}_{i=1}^n$. Note that, conditioned on A and $\{\mathbf{y_i}\}_{i=1}^n$, $A^{-1}\Big(Y_n Q - I\Big)y$ is a centered Gaussian random vector with covariance $A^{-1}\Big(Y_n Q - I\Big)\Sigma\Big(QY_n - I\Big)A^{-1}$. In addition, since the task and data distributions are independent, we can replace the task by its expectation. It therefore holds that

$$\mathbb{E}\Big[\Big\|A^{-1}\Big(Y_n Q - I\Big)y\Big\|^2\Big] = \mathbb{E}_{Y_n}\Big[\mathrm{Tr}\Big(B\Big(Y_n Q - I\Big)\Sigma\Big(Q^T Y_n - I\Big)\Big)\Big].$$

Since this is a convex functional of Q, it suffices to characterize the critical point. Taking the derivative, we find that the critical point equation for the risk it

$$\nabla_Q \mathcal{R}(\mathbf{I}_d, Q) = \mathbb{E}_{Y_n}[\Sigma Q^T Y_n B Y_n + Y_n B Y_n Q \Sigma] - 2\Sigma B \Sigma = 0.$$

Using Lemma 4 to compute the expectation, we further rewrite the critical point equation as

$$\Big(\frac{n+1}{n}B\Sigma + \frac{\mathrm{Tr}_\Sigma(B)}{n}\Sigma\Big)Q + Q^T\Big(\frac{n+1}{n}\Sigma B + \frac{\mathrm{Tr}(\Sigma)}{n}\Sigma\Big) = 2B.$$

This equation is solved by the matrix Q_n defined in the statement of the Lemma. □

C Proofs and additional results for Subsection 3.2

In this section, we present a proof of Theorem 2 and provide an example of the PDE recovery error bound when the spatial discretization is defined by a P^1-finite element method.

Proof of Theorem 2. By the triangle inequality, we have

$$\mathbb{E}\Big[\|u - \widehat{u}_d\|_{H^1(\Omega)}^2\Big] \leq 2\mathbb{E}\Big[\|u - u_d\|_{H^1(\Omega)}^2\Big] + 2\mathbb{E}\Big[\|u_d - \widehat{u_d}\|_{H^1(\Omega)}^2\Big].$$

Notice that $\mathbb{E}\Big[\|\mathbf{u}_d - \widehat{\mathbf{u}}_d\|_{L^2(\Omega)}^2\Big] = \mathcal{R}_m(\widehat{\theta})$, where $\widehat{\theta}$ is as defined in the statement of Theorem 1. The desired estimate therefore follows, provided we can bound $\mathbb{E}\Big[\|u_d - \widehat{u_d}\|_{H^1(\Omega)}^2\Big]$ by a multiple of

$\mathbb{E}\Big[\|u_d - \widehat{u_d}\|^2_{L^2(\Omega)}\Big]$. For any $g = \sum_{k=1}^d c_k\phi_k \in \text{span}\{\phi_k\}_{k=1}^d$, we have

$$
\begin{aligned}
\|g\|^2_{H^1(\Omega)} &= \|g\|^2_{L^2(\Omega)} + \Big\|\sum_{k=1}^d c_k\phi_k'(x)\Big\|^2_{L^2(\Omega)}\\
&= c^T(\Phi + \Phi')c\\
&= \tilde{c}(\mathbf{I}_d + \Phi^{-1/2}\Phi'\Phi^{-1/2})\tilde{c}\\
&\leq (1 + \lambda_{\max}(\Phi^{-1/2}\Phi'\Phi^{-1/2}))\|\tilde{c}\|^2\\
&= (1 + \lambda_{\max}(\Phi^{-1/2}\Phi'\Phi^{-1/2}))\|g\|^2_{L^2(\Omega)},
\end{aligned}
$$

where $\tilde{c} = \Phi c$ We conclude that

$$\mathbb{E}\Big[\|u_d-\widehat{u_d}\|^2_{H^1(\Omega)}\Big] \leq (1+\lambda_{\max}(\Phi^{-1/2}\Phi'\Phi^{-1/2})\cdot\mathbb{E}\Big[\|u_d-\widehat{u_d}\|^2_{L^2(\Omega)}\Big] = 2\max_{1\leq k\leq d}\|\phi_k\|^2_{H^1(\Omega)}\cdot\mathcal{R}_m(\widehat{\theta}),$$

and therefore that

$$\mathbb{E}\Big[\|u - \widehat{u}_d\|^2_{H^1(\Omega)}\Big] \lesssim \mathbb{E}\Big[\|u - u_d\|^2_{H^1(\Omega)}\Big] + (1 + \lambda_{\max}(\Phi^{-1/2}\Phi'\Phi^{-1/2})\cdot\mathcal{R}_m(\widehat{\theta}).$$

$\square$

Example 1 (PDE recovery error with FEM discretization in 1D)**.** *Consider the elliptic PDE* (1) *on a unit interval $\Omega = [0, 1]$. Let $\mathcal{I}_k = [(k-1)j, kh]$ for $0 \leq k \leq d$ be the uniform mesh on Ω, where $h = d^{-1}$ is the mesh size. Let $P_1^h(\Omega)$ be the linear finite element space spanned by the P_1-finite element base functions $\{\phi_k\}_{k=0}^d$. Let $\mathbf{u}_h \in P_1^h(\Omega)$ denote the P_1-finite element approximation of the solution u. Suppose that Assumption 1 holds for the task distributions P_a, P_V and assume further that $a(x) \in C^1(\Omega)$ P_a-a.s and $V \in C(\Omega)$ P_v-a.s. Then by classical regularity estimates for elliptic PDEs, the solution $u \in H^2(\Omega)$ and satisfies $\|u\|_{H^2(\Omega)} \lesssim \|f\|_{L^2(\Omega)}$ up to a universal constant. Moreover, by Theorem 3.16 in Ern and Guermond [2004], the FEM-solution u_d satisfies the discretization error estimate*

$$\|u - u_d\|_{H^1(\Omega)} \lesssim h\|u\|_{H^2(\Omega)}.$$

It follows that

$$\mathbb{E}\Big[\|u - u_d\|^2_{H^1(\Omega)}\Big] \lesssim h^2\mathbb{E}[\|u\|^2_{H^2(\Omega)}] \lesssim h^2\mathbb{E}[\|f\|^2_{L^2(\Omega)}] = h^2 \mathit{Tr}(\Sigma_f),$$

where $\Sigma_f : L^2(\Omega) \to L^2(\Omega)$ is the covariance operator of $f \sim P_f$. In addition, it can be shown that for piecewise linear FEM on 1D, the stiffness and mass matrices satisfy $\lambda_{\max}(\Phi^{-1/2}\Phi'\Phi^{-1/2}) \lesssim h^{-2}$ (see e.g. equation (2.4) of Boffi [2010]). By Theorem 2, we conclude that in the practical regime that $m \leq n$, the PDE recovery error of the transformer is bounded by

$$\mathbb{E}\Big[\|u - \widehat{u}_h\|^2_{H^1(\Omega)}\Big] \lesssim h^2 + \frac{1}{h^2}\Big(\frac{1}{m} + \frac{C_A^4\|\Sigma^{-1}\|^2_{op}}{n^2} + \frac{d^2\|\Sigma^{-1}\|^4_{op}}{\sqrt{N}}\Big).$$

Note that the terms $\|\Sigma^{-1}\|_{op}$ and C_A^4 depend on the number of Galerkin basis functions d. For the matrix A corresponding to the FEM discretization, it can be shown that $C_A \lesssim h^{-2}$. In addition, when the covariance operator of the random source is given by $\Sigma_f = (-\Delta + I)^{-\alpha}$ for some $\alpha > 0$ which controls the smoothness of the source term, it follows from the inverse inequalities [Ern and Guermond, 2004, Lemma 12.1] that $\|\Sigma^{-1}\|_{op} \lesssim h^{2\alpha}$. Inserting this estimate to above leads to the final PDE recovery bound in terms of the mesh size h

$$\mathbb{E}\Big[\|u - \widehat{u}_h\|^2_{H^1(\Omega)}\Big] \lesssim h^2 + \frac{1}{h^2 m} + \frac{1}{h^{10+4\alpha}n^2} + \frac{1}{h^{4+8\alpha}\sqrt{N}}, \tag{28}$$

or equivalently in terms of the number of Galerkin basis functions d

$$\mathbb{E}\Big[\|u - \widehat{u}_h\|^2_{H^1(\Omega)}\Big] \lesssim \frac{1}{d^2} + \frac{d^2}{m} + \frac{d^{10+4\alpha}}{n^2} + \frac{d^{4+8\alpha}}{\sqrt{N}}. \tag{29}$$

Here, we have hidden all constants from the estimate of Theorem 1 that do not depend on the dimension d.

D Proofs and additional results for Subsection 3.3

We first state a more general version of Theorem 3, which does not assume that the pre-training task distribution is diverse relative to the downstream task distribution.

Theorem 7. *Let p_A and p'_A denote the pre-training and downstream task distributions respectively and assume both satisfy Assumption 1. Let $\mathcal{M}_\infty(p_A)$ and $\mathcal{M}_\infty(p'_A)$ denote the minimizers of $\mathcal{R}_\infty$ and $\mathcal{R}'_\infty$ respectively, and let $\widehat{\theta} \in \text{argmin}_{\|\theta\| \leq M} \mathcal{R}_{n,N}(\theta)$ denote the empirical risk minimizer. Then*

$$\mathcal{R}'_m(\widehat{\theta}) \lesssim \mathcal{R}_m(\widehat{\theta}) + \frac{d(p_A, p'_A)}{m} + \text{dist}(\widehat{\theta}, \mathcal{M}_\infty(p_A))^2 + \text{dist}(\widehat{\theta}, \mathcal{M}_\infty(p'_A))^2,$$

where $d(p_A, p'_A)$ is a distance between the distributions p_A and p'_A, and the implicit constants depend on M, Σ, and the constant c_A defined in Assumption 1.

Notice that Theorem 3 is a direct consequence of Theorem 7, because the assumption that p_A is diverse relative to p'_A implies that for any θ, $\text{dist}(\theta, \mathcal{M}_\infty(p'_A)) \leq \text{dist}(\theta, \mathcal{M}_\infty(p_A))$. The fourth term in the bound of Theorem 7, corresponding to $\text{dist}(\widehat{\theta}, \mathcal{M}_\infty(p'_A))^2$, is novel to the best of our knowledge, and it motivates the definition of task diversity. It highlights the hardness of learning general linear systems in-context, compared to learning linear regression models Zhang et al. [2023] or linear systems corresponding to diagonal matrices Chen et al. [2024b].

Proof of Theorem 7. Recall that $\widehat{\theta} \in \text{argmin}_{\|\theta\| \leq M} \mathcal{R}_{n,N}(\theta)$ is the ERM. Let $\theta_* = (P_*, Q_*)$ denote a projection of $\widehat{\theta}$ onto the set $\mathcal{M}_\infty$ and let $\theta'_* = (P'_*, Q'_*)$ denote a projection of $\widehat{\theta}$ onto $\mathcal{M}_\infty$. Let $\epsilon_1 = \|\widehat{\theta} - \theta_*\|$ and $\epsilon_2 = \|\widehat{\theta} - \theta'_*\|$. Then we have the error decomposition

$$\mathcal{R}'_m(\widehat{\theta}) = \mathcal{R}_m(\widehat{\theta}) + (\mathcal{R}'_m(\widehat{\theta}) - \mathcal{R}'_m(\theta'_*)) + (\mathcal{R}_m(\theta'_*) - \mathcal{R}_m(\theta_*)) + (\mathcal{R}_m(\theta_*) - \mathcal{R}_m(\widehat{\theta}))$$

Taking the infimum over all projections θ_* and θ'_* of $\widehat{\theta}$ onto $\mathcal{M}_\infty(p_A)$ and $\mathcal{M}_\infty(p'_A)$, followed by the supremum over $\widehat{\theta}$ in $\{\|\theta\| \leq M\}$, we arrive at the bound

$$\begin{aligned}\mathcal{R}'_m(\widehat{\theta}) \leq \mathcal{R}_m(\widehat{\theta}) &+ \sup_{\|\widehat{\theta}\| \leq M} \inf_{\theta_*, \theta'_*} |\mathcal{R}_m(\theta_*) - \mathcal{R}'_m(\theta'_*)| + \sup_{\|\theta_1\|, \|\theta_2\| \leq M, \|\theta_1 - \theta_2\| \leq \epsilon_2} |\mathcal{R}_m(\theta_1) - \mathcal{R}_m(\theta_2)| \\ &+ \sup_{\|\theta_1\|, \|\theta_2\| \leq M, \|\theta_1 - \theta_2\| \leq \epsilon_1} |\mathcal{R}'_m(\theta_1) - \mathcal{R}'_m(\theta_2)|.\end{aligned}$$

The second and third terms can be bounded using a simple Lipschitz continuity estimate. Note that for m sufficiently large and $\theta = (P, Q)$ with $\|\theta\| \leq M$, we have

$$\|(PA^{-1}Y_mQ - A^{-1})\Sigma^{1/2}\|_F^2 \lesssim c_A^2(1 + \|\Sigma\|_{\text{op}} M^2)^2 \text{Tr}(\Sigma)$$

for any $A \in \text{supp}(p_A)$. It follows that

$$R_m(\theta) = \mathbb{E}_{A \sim p_A, Y_m}[\|(PA^{-1}Y_mQ - A^{-1})\Sigma^{1/2}\|_F^2]$$

is $O\left(c_A^2(1 + \|\Sigma\|_{\text{op}} M^2)^2 \text{Tr}(\Sigma)\right)$-Lipschitz on $\{\|\theta\| \leq M\}$. We therefore have

$$\sup_{\|\theta_1\|, \|\theta_2\| \leq M, \|\theta_1 - \theta_2\| \leq \epsilon_1} |\mathcal{R}_m(\theta_1) - \mathcal{R}_m(\theta_2)| \lesssim \left(c_A^2(1 + \|\Sigma\|_{\text{op}} M^2)^2 \text{Tr}(\Sigma)\right) \epsilon_1^2.$$

An analogous bound holds for $\sup_{\|\theta_1\|, \|\theta_2\| \leq M, \|\theta_1 - \theta_2\| \leq \epsilon_2} |\mathcal{R}'_m(\theta_1) - \mathcal{R}'_m(\theta_2)|$, since the test distribution p'_A is also assumed to satisfy Assumption 1. To bound the term $|\mathcal{R}_m(\theta_*) - \mathcal{R}'_m(\theta'_*)|$, we recall by Lemma 5 that for any $\theta = (P, Q)$,

$$\mathcal{R}_m(\theta) = \mathcal{R}_\infty(\theta) + \frac{1}{m}\mathbb{E}_{A \sim p_A}\left[\text{Tr}(PA^{-1}\Sigma Q \Sigma Q^T \Sigma A^{-1} P^T) + \text{Tr}_\Sigma(Q\Sigma Q^T)\text{Tr}(PA^{-1}\Sigma A^{-1}P^T)\right]$$

and

$$\begin{aligned}\mathcal{R}'_m(\theta) = \mathcal{R}'_\infty(\theta) + \frac{1}{m}\mathbb{E}_{A \sim p'_A}\Big[&\text{Tr}(P(A')^{-1}\Sigma Q \Sigma Q^T \Sigma (A')^{-1} P^T) \\ &+ \text{Tr}_\Sigma(Q\Sigma Q^T)\text{Tr}(P(A')^{-1}\Sigma (A')^{-1}P^T)\Big].\end{aligned}$$

In particular, since $\theta_* \in \mathrm{argmin}_\theta \mathcal{R}_\infty(\theta)$ and $\theta'_* \in \mathrm{argmin}_\theta \mathcal{R}'_\infty(\theta)$, and each functional achieves 0 as its minimum value, we have

$$\begin{aligned}
|\mathcal{R}_m(\theta_*) - \mathcal{R}'_m(\theta'_*)| \leq \frac{1}{m}\Big|&\mathbb{E}_{A\sim p_A}\big[\mathrm{Tr}(P_* A^{-1}\Sigma Q_* \Sigma Q_*^T \Sigma A^{-1} P_*^T) \\
&+ \mathrm{Tr}_\Sigma(Q_*\Sigma Q_*^T)\mathrm{Tr}(P_* A^{-1}\Sigma A^{-1}P_*^T)\big] \\
&- \mathbb{E}_{A\sim p'_A}\big[\mathrm{Tr}(P'_*(A')^{-1}\Sigma Q'_*\Sigma (Q'_*)^T\Sigma (A')^{-1}(P'_*)^T) \\
&+ \mathrm{Tr}_\Sigma(Q'_*\Sigma(Q'_*)^T)\mathrm{Tr}(P'_*(A')^{-1}\Sigma(A')^{-1}(P'_*)^T)\big]\Big| \\
=: \frac{1}{m}&\left|\mathbb{E}_{A\sim p_A}[f(A;\theta_*)] - \mathbb{E}_{A'\sim p'_A}[f(A';\theta'_*)]\right|.
\end{aligned}$$

It follows that

$$\begin{aligned}
\sup_{\|\widehat{\theta}\|\leq M} \inf_{\theta_*,\theta'_*} |\mathcal{R}_m(\theta_*) - \mathcal{R}'_m(\theta'_*)| &\leq \frac{1}{m}\sup_{\|\widehat{\theta}\|\leq M} \inf_{\theta_*,\theta'_*} \left|\mathbb{E}_{A\sim p_A}[f(A;\theta_*)] - \mathbb{E}_{A'\sim p'_A}[f(A';\theta'_*)]\right| \\
&=: \frac{1}{m} d(p_A, p'_A),
\end{aligned}$$

where, again, the infimum is taken over all $\theta_* \in \mathrm{argmin}_{\theta\in\mathcal{M}_\infty(p_A)}\|\theta - \widehat{\theta}\|^2$ and $\theta'_* \in \mathrm{argmin}_{\theta'\in\mathcal{M}_\infty(p'_A)}\|\theta' - \widehat{\theta}\|^2$. Combining the estimates for each individual term in the error decomposition, we obtain the final bound in the statement of Theorem 7. The fact that the bound we have obtained tends to zero as the sample size $(m, n, N) \to \infty$ follows from examination of each term in the estimate: the in-domain generalization error $\mathcal{R}_m(\widehat{\theta})$ tends to zero in probability by Theorem 1, the term $\frac{d(p_A,p'_A)}{m}$ is deterministic and tends to zero as $m \to \infty$, and $\mathrm{dist}(\widehat{\theta}, \mathcal{M}_\infty)$ tends to zero as N and n tend to infinity, respectively, by Proposition 4. □

The discrepancy $d(p_A, p'_A)$ defined in the proof of Theorem 3 may not be a metric, but, crucially, it satisfies $d(p_A, p_A) = 0$. This ensures that the error term due to distribution shift in Theorem 3 vanishes when the pre-training and downstream tasks coincide. We give a simple proof of this fact below.

Lemma 2. *Let*

$$d(p_A, p'_A) = \sup_{\|\widehat{\theta}\|\leq M} \inf_{\theta_*,\theta'_*} \Big|\mathbb{E}_{A\sim p_A}[f(A;\theta_*)] - \mathbb{E}_{A'\sim p'_A}[f(A';\theta'_*)]\Big|,$$

where the infimum is taken over all projections θ_* and θ'_* of $\widehat{\theta}$ onto the sets $\mathcal{M}_\infty(p_A)$ and $\mathcal{M}_\infty(p'_A)$ respectively, and

$$f(A;\theta) = Tr(PA^{-1}\Sigma Q\Sigma Q^T\Sigma A^{-1}P^T) + Tr_\Sigma(Q\Sigma Q^T)Tr(PA^{-1}\Sigma A^{-1}P^T),\ \theta = (P, Q).$$

Then $d(p_A, p'_A) = 0$ if $p_A = p'_A$.

Proof. Note that we can upper bound $d(p_A, p_A)$ by

$$d(p_A, p_A) \leq \sup_{\|\widehat{\theta}\|\leq M} \inf_{\theta_*} \Big|\mathbb{E}_{A\sim p_A}[f(A;\theta_*)] - \mathbb{E}_{A\sim p_A}[f(A;\theta_*)]\Big|,$$

where the infimum is now taken only over all projections θ_* of $\widehat{\theta}$ onto $\mathcal{M}_\infty(p_A)$. Clearly we have

$$\Big|\mathbb{E}_{A\sim p_A}[f(A;\theta_*)] - \mathbb{E}_{A\sim p_A}[f(A;\theta_*)]\Big| = 0$$

for all θ_*, hence $d(p_A, p_A) \leq 0$. Since $d(p_A, p_A)$ is clearly non-negative, we conclude that $d(p_A, p_A) = 0$. □

The next proposition gives a characterization of the minimizers of the functionals $\mathcal{R}_\infty$ and $\mathcal{R}'_\infty$. Apart from being interesting in its own right, it is a key tool to prove Theorem 4.

Proposition 2. *Fix a task distribution p_A satisfying Assumption 1. Then $\theta = (P, Q)$ is a minimizer of $\mathcal{R}_\infty$ if and only if P commutes with all elements of the set $\{A_1 A_2^{-1} : A_1, A_2 \in supp(p_A)\}$ and Q is given by $Q = \Sigma^{-1} A_0 P^{-1} A_0^{-1}$ for any $A_0 \in supp(p_A)$.*

Proof of Proposition 2. Recall that

$$\mathcal{R}_\infty(\theta) = \mathbb{E}_{A\sim p_A}[\|(PA^{-1}\Sigma Q - A^{-1})\Sigma^{1/2}\|_F^2],\ \theta = (P, Q),$$

and $\mathcal{M}_\infty(p_A) = \text{argmin}_\theta \mathcal{R}_\infty(\theta)$. Let us first prove that for any p_A satisfying Assumption 1, $\theta \in \mathcal{M}_\infty(p_A)$ if and only if $PA^{-1}\Sigma Q = A^{-1}$ for all $A \in \text{supp}(p_A)$. Let us first observe that the minimum value of $\mathcal{R}_\infty$ is 0 - this is attained, for instance, at $P = \mathbf{I}_d$ and $Q = \Sigma^{-1}$. It is clear that if the equality $PA^{-1}\Sigma Q = A^{-1}$ holds over the support of p_A, then $\mathbb{E}_{A\sim p_A}[\|(PA^{-1}\Sigma Q - A^{-1})\Sigma^{1/2}\|_F^2] = 0$. Conversely, suppose (P, Q) satisfies $\mathbb{E}_{A\sim p_A}[\|(PA^{-1}\Sigma Q - A^{-1})\Sigma^{1/2}\|_F^2] = 0$. Fixing $A_0 \in \text{supp}(p_A)$ and $\epsilon > 0$, let $p_{A,\epsilon}(A_0)$ denote the normalized restriction of p_A to the ball of radius ϵ centered about A_0. Then the equality $\mathbb{E}_{A\sim p_A}[\|(PA^{-1}\Sigma Q - A^{-1})\Sigma^{1/2}\|_F^2] = 0$ implies that

$$\mathbb{E}_{A\sim p_{A,\epsilon}(A_0)}[\|(PA^{-1}\Sigma Q - A^{-1})\Sigma^{1/2}\|_F^2] = 0$$

for each $\epsilon > 0$. Since $p_{A,\epsilon}(A_0)$ converges weakly to the Dirac measure centered at A_0, we have that $\|(PA_0^{-1}\Sigma Q - A_0^{-1})\Sigma^{1/2}\|_F^2 = 0$, and hence that $PA_0^{-1}\Sigma Q = A_0^{-1}$. As A_0 was arbitrary, this concludes the first part of the proof.

Now, suppose $\theta = (P, Q)$ is a minimizer of $\mathcal{R}_\infty$. By the previous argument, this is equivalent to the system of equations $PA^{-1}\Sigma Q = A^{-1}$ holding simultaneously for all $A \in \text{supp}(p_A)$. In particular, for any fixed $A_0 \in \text{supp}(p_A)$, the equation $PA_0^{-1}\Sigma Q = A_0^{-1}$ can be solved for Q, yielding $Q = \Sigma^{-1} A_0 P^{-1} A_0^{-1}$. Since the matrix Q is constant, this implies that the function $A \mapsto AP^{-1}A^{-1}$ is a constant on the support of p_A. We have therefore shown that the minimizers of $\mathcal{R}_\infty$ can be completely characterized as $\{(P, \Sigma^{-1} A_0 P^{-1} A_0^{-1}) : P \in \mathbb{R}^{d\times d}\}$, where A_0 is any element of $\text{supp}(p_A)$. In addition, the fact that the function $A \mapsto AP^{-1}A^{-1}$ is constant on the support of p_A implies that P commutes with all products of the form $\{A_1 A_2^{-1} : A_1, A_2 \in \text{supp}(p_A)\}$. □

We now give a proof of Theorem 4.

Proof of Theorem 4. 1) This is a direct corollary of Proposition 2.

2) Let $\theta_* = (P_*, Q_*)$ be a minimizer of $\mathcal{R}_\infty$. Then Proposition 2 implies that $P_* \in \mathcal{C}(\mathcal{S}(p_A))$. Since the centralizer of $\mathcal{S}(p_A)$ is trivial by assumption, this implies that $P_* = c\mathbf{I}_d$ for some $c \in \mathbb{R}\backslash\{0\}$. Using the characterization of minimizers of $\mathcal{R}_\infty$ derived in Proposition 2, we have that Q_* solves the equation $cA^{-1}\Sigma Q_* = A^{-1}$ for all $A \in \text{supp}(p_A)$, and therefore $Q = c^{-1}\Sigma^{-1}$. □

The proof of Theorem 4 implies that if $\text{supp}(p_A)$ satisfies the condition that the centralizer of $\{A_1 A_2^{-1} : A_1, A_2 \in \text{supp}(p_j)\}$ is trivial, then all minimizers of $\mathcal{R}_\infty$ are of the form $\{(P, Q) = (c\mathbf{I}_d, c^{-1}\Sigma^{-1}) : c \neq 0\}$. In this case, it is worth noting that the discrepancy on task distributions $d(p_A, p'_A)$ defined in Theorem 3 admits a much simpler expression. We state this result as a Corollary below.

Corollary 1. *Under the assumption that the pre-training task distribution p_A satisfies the centralizer condition*

$$\mathcal{C}\big(\{A_1 A_2^{-1} : A_1, A_2 \in supp(p_j)\}\big) = \{c\mathbf{I}_d : c \in \mathbb{R}\},$$

the out-of-distribution generalization error admits the more tractable expression

$$\mathcal{R}'_m(\widehat{\theta}) = \mathcal{R}_m(\widehat{\theta}) + \frac{(d+1)\left|Tr\left(\left(\mathbb{E}_{A\sim p_A}[A^{-2}] - \mathbb{E}_{A'\sim p'_A}[(A')^{-2}]\right)\Sigma\right)\right|}{m} + dist(\widehat{\theta}, \mathcal{M}_\infty(p_A))^2.$$

In particular, the second term, reflecting the discrepancy between p_A and p'_A, depends only on the second moments of A^{-1} and $(A')^{-1}$.

Proof. By combining Theorems 3 and 4, we immediately derive the bound on the out-of-distribution generalization error

$$\mathcal{R}'_m(\widehat{\theta}) = \mathcal{R}_m(\widehat{\theta}) + \frac{d(p_A, p'_A)}{m} + \text{dist}(\widehat{\theta}, \mathcal{M}_\infty(p_A))^2,$$

where the distance $d(p_A, p'_A)$ is given by

$$d(p_A, p'_A) = |\mathcal{R}_m(\theta_*) - \mathcal{R}'_m(\theta_*)|\,,$$

and θ_* is defined as the projection of $\widehat{\theta}$ onto the $\mathcal{M}_\infty(p_A)$. Under our assumptions, we have $\mathcal{M}_\infty(p_A) = \{(c\mathbf{I}_d, c^{-1}\Sigma^{-1}) : c \in \mathbb{R}\backslash\{0\}\}$, and applying Lemma 6 to compute $\mathcal{R}_m(\theta_*)$ and $\mathcal{R}'_m(\theta_*)$, we obtain

$$d(p_A, p'_A) = (d+1)\left|\mathrm{Tr}\left(\left(\mathbb{E}_{A\sim p_A}[A^{-2}] - \mathbb{E}_{A'\sim p'_A}[(A')^{-2}]\right)\Sigma\right)\right|.$$

□

To conclude this section, we investigate the diversity of task distributions whose support consists of simultaneously diagonalizable matrices. The simultaneous-diagonalizability of task matrices has been used as a key assumption in the existing theoretical analysis of in-context learning of linear systems (Chen et al. [2024b]) and in the in-context learning of linear dynamical systems (Sander et al. [2024]). In addition, it is also relevant to the PDE setting: if the diffusion coefficient $a(x)$ and potential $V(x)$ are both constant, $a(x) \equiv a_0$, $V(x) \equiv v_0$, then the solution operator of the corresponding elliptic PDE is given by $\Big(-a_0\Delta + v_0\mathbf{I}\Big)^{-1}$, whose diagonalization is independent of the constants a_0 and v_0. It is therefore natural to ask whether such a task distribution is diverse in the sense of Definition 1.

Proposition 3. *Let p_A and p'_A denote the pre-training and downstream task distributions, and suppose that the matrices in supp(p_A) are simultaneously diagonalizable for a common orthogonal matrix U. Suppose additionally that there exist matrices $A_1, A_2 \in$ supp(p_A) and $A'_1 A'_2 \in$ supp(p'_A) such that $A_1A_2^{-1}$ and $A'_1(A'_2)^{-1}$ have no repeated eigenvalues.*

1. *If supp(p'_A) is also simultaneously diagonalizable with respect to U, then p_A is diverse relative to p'_A.*
2. *If there exist matrices $A'_3, A'_4 \in$ supp(p'_A) such that $A'_3(A'_4)^{-1}$ is not diagonalizable with respect to U, then p_A is not diverse relative to p'_A.*

Proposition 3 reveals that a simultaneously-diagonalizable task distribution cannot achieve out-of-distribution generalization under arbitrary shifts in the downstream task distribution; namely the downstream task distribution must also be simultaneously diagonalizable in the same basis. However, it also shows that, provided the pre-training and downstream task distributions are simultaneously diagonalizable, pre-trained transformers can generalize under arbitrary shifts on the distribution shifts on the eigenvalues of the task matrices. This provides a precise characterization of the diversity of a simultaneously diagonalizable task distribution.

Before proving Proposition 3, we first introduce a preliminary lemma.

Lemma 3. *Let p_A be a task distribution satisfying Assumption 1. Suppose that the support of p_A is simultaneously diagonalizable with a common orthogonal diagonalizing matrix $U \in \mathbb{R}^{d\times d}$. Assume in addition that there exist $A_1, A_2 \in$ supp(p_A) such that $A_1A_2^{-1}$ has distinct eigenvalues. Then $M_\infty(p_A) = \Theta_{U,\Sigma}$, where*

$$\Theta_{U,\Sigma} := \left\{(P, \Sigma^{-1}P^{-1}) : P = UDU^T,\ D = \mathit{diag}(\lambda_1, \ldots, \lambda_d)\right\}.$$

Proof. By Proposition 2, a parameter (P, Q) belongs to $\mathcal{M}_\infty(p_A)$ if and only if P commutes with all products of the form $\{A_iA_j^{-1} : A_i, A_j \in \mathrm{supp}(p_A)\}$, in which case Q is defined by $Q = \Sigma^{-1}A_0P^{-1}A_0^{-1}$ for any $A_0 \in \mathrm{supp}(p_A)$. Let $A_1, A_2 \in \mathrm{supp}(p_A)$ be as defined in the statement of the lemma. Since P and $A_1A_2^{-1}$ are commuting diagonalizing matrices and $A_1A_2^{-1}$ has no repeated eigenvalues (Strang [2022]), they must be simultaneously diagonalizable. This implies that P is diagonal in the basis U, and hence Q is given by $Q = \Sigma^{-1}A_0P^{-1}A_0^{-1} = \Sigma^{-1}P^{-1}$. □

Proof of Proposition 3. For 1), if the support of p'_A is also simultaneously diagonalizable with respect to U, then Lemma 3 implies that $\mathcal{M}_\infty(p_A) = \mathcal{M}_\infty(p'_A) = \Theta_{U,\Sigma}$, where $\Theta_{U,\Sigma}$, where $\Theta_{U,\Sigma}$ is as defined in the statement of Lemma 3. This proves that if the support of p'_A is also simultaneously diagonalizable with respect to U, then p_A is diverse.

For 2), we must find a minimizer of $\mathcal{R}_\infty$ which is not a minimizer of $\mathcal{R}'_\infty$. Consider the parameter $\theta = (P, \Sigma^{-1}P^{-1})$, where $P = UDU^T$ for D an invertible diagonal matrix with no repeated entries. By Lemma 3, θ is a minimizer of $\mathcal{R}_\infty$. Let $A'_3, A'_4 \in \text{supp}(p'_A)$ be such that $A'_3(A'_4)^{-1}$ is not diagonalizable with respect to U. Since $A'_3(A'_4)^{-1}$ and P are not simultaneously diagonalizable and P has no repeated eigenvalues (Strang [2022]), P does not commute with $A'_3(A'_4)^{-1}$. By Proposition 2, θ is therefore not a minimizer of $\mathcal{R}'_\infty$, completing the proof. □

E Proofs for Subsection 3.4

We begin by stating a more formal version of Theorem 5 where the constants are more explicit.

Theorem 8. *Let $\Sigma = W\Lambda W^T$ and $\tilde{\Sigma} = \tilde{W}\tilde{\Lambda}\tilde{W}^T$ be two covariance matrices, let $(\widehat{P}, \widehat{Q})$ be minimizers of the empirical risk when the in-context examples follow the distribution $N(0, \Sigma)$ and take $M > 0$ such that $\max\Big(\|\widehat{P}\|_F, \|\widehat{Q}\|_F\Big) \leq M$. Then*

$$\mathcal{R}_m^{\tilde{\Sigma}}(\widehat{P}, \widehat{Q}) \lesssim \mathcal{R}_m^{\Sigma}(\widehat{P}, \widehat{Q}) + c_A^2 M^4 \max(\|\Sigma\|_{op}, \|\tilde{\Sigma}\|_{op})^2 \|\Sigma - \tilde{\Sigma}\|_{op}$$
$$+ \frac{1}{m} \cdot c_A^2 M^4 \max(\|\Sigma\|_{op}, \|\tilde{\Sigma}\|_{op})^2 Tr(\tilde{\Sigma}) \Big(\|\Sigma - \tilde{\Sigma}\|_{op} + \|\Lambda - \tilde{\Lambda}\|_1 + \|W - \tilde{W}\|_{op}\Big).$$

Theorem 5 then follows from Theorem 8 by bounding $\|\Lambda - \tilde{\Lambda}\|_1 \lesssim \|\Sigma - \tilde{\Sigma}\|_{\text{op}}$, merging the term

$$\frac{1}{m} \cdot c_A^2 M^4 \max(\|\Sigma\|_{\text{op}}, \|\tilde{\Sigma}\|_{\text{op}})^2 \text{Tr}(\tilde{\Sigma}) \Big(\|\Sigma - \tilde{\Sigma}\|_{\text{op}} + \|\Lambda - \tilde{\Lambda}\|_1\Big)$$

into the second term, and omitting the constant factors.

Proof of Theorem 8. By the triangle inequality, we have

$$\mathcal{R}_m^{\tilde{\Sigma}}(\widehat{P}, \widehat{Q}) \leq \mathcal{R}_m^{\Sigma}(\widehat{P}, \widehat{Q}) + \sup_{\|P\|_{\text{op}}, \|Q\|_{\text{op}} \leq M} \Big|\mathcal{R}_m^{\tilde{\Sigma}}(P, Q) - \mathcal{R}_m^{\Sigma}(P, Q)\Big|. \tag{30}$$

It therefore suffices to bound the second term. From the proof of Proposition 1, we know that

$$\mathcal{R}_m^{\Sigma}(P, Q) = \mathbb{E}_A\Big[\frac{m+1}{m}\text{Tr}(PA^{-1}\Sigma Q\Sigma Q^T\Sigma A^{-1}P^T + \frac{\text{Tr}_\Sigma(Q\Sigma Q^T)}{m}\text{Tr}(PA^{-1}\Sigma A^{-1}P^T)\Big] \tag{31}$$
$$+ \mathbb{E}_A\Big[\text{Tr}(A^{-1}\Sigma A^{-1}) - \text{Tr}(PA^{-1}\Sigma Q\Sigma A^{-1}) - \text{Tr}(A^{-1}\Sigma Q^T\Sigma A^{-1}P^T)\Big]. \tag{32}$$

Similarly, we have

$$\mathcal{R}_m^{\tilde{\Sigma}}(P, Q) = \mathbb{E}_A\Big[\frac{m+1}{m}\text{Tr}(PA^{-1}\tilde{\Sigma} Q\tilde{\Sigma} Q^T\tilde{\Sigma} A^{-1}P^T + \frac{\text{Tr}_{\tilde{\Sigma}}(Q\tilde{\Sigma} Q^T)}{m}\text{Tr}(PA^{-1}\tilde{\Sigma} A^{-1}P^T)\Big] \tag{33}$$
$$+ \mathbb{E}_A\Big[\text{Tr}(A^{-1}\tilde{\Sigma} A^{-1}) - \text{Tr}(PA^{-1}\tilde{\Sigma} Q\tilde{\Sigma} A^{-1}) - \text{Tr}(A^{-1}\tilde{\Sigma} Q^T\tilde{\Sigma} A^{-1}P^T)\Big]. \tag{34}$$

We seek to bound the difference $\Big|\mathcal{R}_m^{\Sigma}(\theta) - \mathcal{R}_m^{\tilde{\Sigma}}(\theta)\Big|$ by bounding the respective differences of each term appearing in the expressions for $\mathcal{R}_m^{\Sigma}$ and $\mathcal{R}_m^{\tilde{\Sigma}}$. By a simple applications of Hölder's inequality and the triangle inequality, we see that

$$\begin{aligned}
\mathbb{E}_A \text{Tr}(PA^{-1}(\Sigma Q\Sigma - \tilde{\Sigma} Q\tilde{\Sigma})A^{-1}) &\leq \mathbb{E}_A\|A^{-1}PA^{-1}\|_F \|\Sigma Q\Sigma - \tilde{\Sigma}Q\tilde{\Sigma}\|_F \\
&\leq c_A^2\|P\|_F\Big(\|(\Sigma - \tilde{\Sigma})Q\Sigma\|_F + \|\tilde{\Sigma}Q(\Sigma - \tilde{\Sigma})\|_F\Big) \\
&\leq c_A^2\|P\|_F\Big(\|Q\Sigma\|_F + \|\tilde{\Sigma}Q\|_F\Big)\|\Sigma - \tilde{\Sigma}\|_{\text{op}} \\
&\leq 2c_A^2\|P\|_F\|Q_f \max(\|\Sigma\|_{\text{op}}, \|\tilde{\Sigma}\|_{\text{op}})\|\Sigma - \tilde{\Sigma}\|_{\text{op}} \\
&= 2c_A^2 M^2 \max(\|\Sigma\|_{\text{op}}, \|\tilde{\Sigma}\|_{\text{op}})\|\Sigma - \tilde{\Sigma}\|_{\text{op}}.
\end{aligned}$$

Analogous arguments can be used to prove the bounds

$$\mathbb{E}_A \mathrm{Tr}(A^{-1}(\Sigma Q^T\Sigma - \tilde{\Sigma}Q^T\tilde{\Sigma})A^{-1}P^T) \le 2c_A^2 M^2 \max(\|\Sigma\|_{\text{op}}, \|\tilde{\Sigma}\|_{\text{op}})\|\Sigma - \tilde{\Sigma}\|_{\text{op}},$$
$$\mathbb{E}_A \mathrm{Tr}(A^{-1}(\Sigma - \tilde{\Sigma})A^{-1}) \le c_A^2 \|\Sigma - \tilde{\Sigma}\|_{\text{op}}$$

and

$$\mathbb{E}_A \mathrm{Tr}(PA^{-1}(\Sigma Q\Sigma Q^T\Sigma - \tilde{\Sigma}Q\tilde{\Sigma}Q^T\tilde{\Sigma})A^{-1}P^T) \le c_A^2 M^4 \max(\|\Sigma\|_{\text{op}}, \|\tilde{\Sigma}\|_{\text{op}})^2\|\Sigma - \tilde{\Sigma}\|_{\text{op}}.$$

Notice that the term above dominates each of the preceding three terms. For the final term, we have

$$\begin{aligned}
&\mathrm{Tr}_\Sigma(Q\Sigma Q^T)\mathrm{Tr}(PA^{-1}\Sigma A^{-1}P^T) - \mathrm{Tr}_{\tilde{\Sigma}}(Q\tilde{\Sigma}Q^T)\mathrm{Tr}(PA^{-1}\tilde{\Sigma}A^{-1}P^T)\\
&\le \left|\mathrm{Tr}_\Sigma(Q\Sigma Q^T) - \mathrm{Tr}_{\tilde{\Sigma}}(Q\tilde{\Sigma}Q^T)\right|\left|\mathrm{Tr}(PA^{-1}\Sigma A^{-1}P^T)\right|\\
&\quad + \left|\mathrm{Tr}_{\tilde{\Sigma}}(Q\tilde{\Sigma}Q^T)\right|\left|\mathrm{Tr}(PA^{-1}(\Sigma - \tilde{\Sigma})A^{-1}P^T)\right|.
\end{aligned}$$

By Lemma 10 and Holder's inequality, the second term satisfies

$$\left|\mathrm{Tr}_{\tilde{\Sigma}}(Q\tilde{\Sigma}Q^T)\right|\left|\mathrm{Tr}(PA^{-1}(\Sigma - \tilde{\Sigma})A^{-1}P^T)\right| \le c_A^2 M^4 \|\tilde{\Sigma}\|_{\text{op}}\mathrm{Tr}(\tilde{\Sigma}) \cdot \|\Sigma - \tilde{\Sigma}\|_{\text{op}}.$$

Similarly, using Lemma 11, the first term satisfies

$$\begin{aligned}
&\left|\mathrm{Tr}_\Sigma(Q\Sigma Q^T) - \mathrm{Tr}_{\tilde{\Sigma}}(Q\tilde{\Sigma}Q^T)\right|\left|\mathrm{Tr}(PA^{-1}\Sigma A^{-1}P^T)\right|\\
&\le c_A^2 M^4 \|\Sigma\|_{\text{op}}\Big(\mathrm{Tr}(\tilde{\Sigma})\|\Sigma - \tilde{\Sigma}\|_{\text{op}} + \|\Sigma\|_{\text{op}}\Big(\|\Lambda - \tilde{\Lambda}\|_1 + 2\mathrm{Tr}(\tilde{\Sigma})\|W - \tilde{W}\|_{\text{op}}\Big)\Big)
\end{aligned}$$

Combining the estimates for each individual term and taking the supremum over the all P, Q with Frobenius norm bounded by M yields the final bound

$$\begin{aligned}
\mathcal{R}_m^{\tilde{\Sigma}}(\widehat{P}, \widehat{Q}) &\lesssim \mathcal{R}_m^{\Sigma}(\widehat{P}, \widehat{Q}) + c_A^2 M^4 \max(\|\Sigma\|_{\text{op}}, \|\tilde{\Sigma}\|_{\text{op}})^2\|\Sigma - \tilde{\Sigma}\|_{\text{op}}\\
&\quad + \frac{1}{m} \cdot c_A^2 M^4 \max(\|\Sigma\|_{\text{op}}, \|\tilde{\Sigma}\|_{\text{op}})^2 \mathrm{Tr}(\tilde{\Sigma})\Big(\|\Sigma - \tilde{\Sigma}\|_{\text{op}} + \|\Lambda - \tilde{\Lambda}\|_1 + \|W - \tilde{W}\|_{\text{op}}\Big).
\end{aligned}$$

□

F Discussion on dependence of constants on dimension

It is important to consider the dependence of the constants appearing in Theorem 1 on the dimension of the linear system. Recall that in the PDE setting, the dimension d corresponds to the number of basis functions used in Galerkin's method, and hence the true PDE solution is only recovered in the limit $d \to \infty$.

Since the solution operator of the PDE is a bounded operator on $L^2(\Omega)$, the norm of the inverse A^{-1} is uniformly bounded in d, and hence the constant $c_A = \sup_{A \in \text{supp}(p_A)} \|A^{-1}\|_{\text{op}}$ is dimension-independent. Similarly, constants involving the norm of the covariance Σ are dimension-independent, since we always have

$$\|\Sigma\|_{\text{op}} \le \|\Sigma_f\|_{\text{op}}, \quad \mathrm{Tr}(\Sigma) \le \mathrm{Tr}(\Sigma_f),$$

where Σ_f is the covariance of the source f on the infinite-dimensional space. However, the constant $C_A = \sup_{A \in \text{supp}(p_A)} \|A\|_{\text{op}}$ is unbounded as $d \to \infty$, because the limiting forward operator is unbounded on $L^2(\Omega)$. Similarly, the constant $\|\Sigma^{-1}\|_{\text{op}}$ is unbounded as $d \to \infty$. The precise growth of these constants depends on the distributions on the coefficients of the PDE; as a prototypical example, we have $\|A\|_{\text{op}} = O(d^2)$ for the Laplace operator under FEM discretization in 1D. It is thus important to consider the trade-offs between discretization and generalization error with respect to the dimension d; this is explored in Example 1 for the specific case of FEM discretization.

G Auxiliary lemmas

We make frequent use of the following lemma to compute expectations of products of empirical covariance matrices.

Lemma 4. *Let $\{y_1, \dots, y_n\} \subseteq \mathbb{R}^d$ be iid samples from $N(0, \Sigma)$ and assume that $\Sigma = W\Lambda W^T$, where $\Lambda = diag(\sigma_1^2, \dots, \sigma_d^2)$. Let $Y_n = \frac{1}{n}\sum_{k=1}^n y_k y_k^T$ associated to $\{y_1, \dots, y_n\}$ and let $K \in \mathbb{R}^{d\times d}$ denote a deterministic symmetric matrix. Then*

$$\mathbb{E}[Y_n K Y_n] = \frac{n+1}{n}\Sigma K \Sigma + \frac{Tr_\Sigma(K)}{n}\Sigma,$$

where $Tr_\Sigma(K) := \sum_{\ell=1}^d \sigma_\ell^2 \langle K\varphi_\ell, \varphi_\ell\rangle$ and $\varphi_\ell := We_\ell$ denote the eigenvectors of Σ.

Proof. Let us first consider the case that $W = \mathbf{I}_d$, so that the covariance is diagonal with entries $\sigma_1^2, \dots, \sigma_d^2$. Observe that

$$\begin{aligned}\mathbb{E}[(Y_n K Y_n)_{ij}] = \mathbb{E}\Bigg[\sum_{\ell,\ell'=1}^d \frac{1}{n^2}\Bigg(&\sum_{k\neq k'} \langle y_k, e_i\rangle\langle y_{k'}, e_j\rangle\langle y_k, e_\ell\rangle\langle y_{k'}, e_{\ell'},\rangle K_{\ell,\ell'} \\ &+ \sum_{k=1}^n \langle e_i, y_k\rangle\langle e_j, y_k\rangle\langle e_\ell, y_k\rangle\langle e_{\ell'}, y_k\rangle K_{\ell,\ell'}\Bigg)\Bigg].\end{aligned}$$

When $i \neq j$, we compute that

$$\sum_{\ell,\ell'=1}^d \mathbb{E}\Big[\langle y_k, e_i\rangle\langle y_{k'}, e_j\rangle\langle y_k, e_\ell\rangle\langle y_{k'}, e_{\ell'},\rangle K_{\ell,\ell'}\Big] = \sigma_i^2\sigma_j^2 K_{i,j}$$

and

$$\sum_{\ell,\ell'=1}^d \mathbb{E}\Big[\langle y_k, e_i\rangle\langle y_k, e_j\rangle\langle y_k, e_\ell\rangle\langle y_k, e_{\ell'},\rangle K_{\ell,\ell'}\Big] = 2\sigma_i^2\sigma_j^2 K_{i,j}.$$

On the other hand, for $i = j$, we have

$$\sum_{\ell,\ell'=1}^d \mathbb{E}\Big[\langle y_k, e_i\rangle\langle y_{k'}, e_i\rangle\langle y_k, e_\ell\rangle\langle y_{k'}, e_{\ell'},\rangle K_{\ell,\ell'}\Big] = \sigma_i^4 K_{i,i}$$

and

$$\sum_{\ell,\ell'=1}^d \mathbb{E}\Big[\langle y_k, e_i\rangle^2\langle y_k, e_\ell\rangle\langle y_k, e_{\ell'},\rangle K_{\ell,\ell'}\Big] = 2\sigma_i^4 K_{i,i} + \sigma_i^2\sum_{\ell=1}^d \sigma_\ell^2 K_{\ell,\ell}.$$

Putting everything together, we have shown that

$$\mathbb{E}(Y_n K Y_n)_{i,j}] = \frac{n+1}{n}\sigma_i^2\sigma_j^2 K_{i,j} + \delta_{ij}\cdot\frac{\mathrm{Tr}_\Sigma(K)}{n}\sigma_i^2.$$

The result then follows since $(\Sigma K \Sigma)_{i,j} = \sigma_i^2\sigma_j^2 K_{i,j}$. For general covariance $\Sigma = W\Lambda W^T$, we have $Y_n K Y_n = W(Z_n W^T K W Z_n)W^T$, where Z_n is the empirical covariance matrix associated to $\{W^T y_1, \dots, W^T y_n\}$. Noting that $W^T y \sim N(0, \Lambda)$ for $y \sim N(0, \Sigma)$, we can apply the above result to $W^T K W$:

$$\begin{aligned}\mathbb{E}[Y_n K Y_n] &= W\mathbb{E}[Z_n(W^T K W)Z_n]W^T \\ &= W\Big(\frac{n+1}{n}\Lambda W^T K W \Lambda + \frac{\mathrm{Tr}_\Sigma(K)}{n}\Lambda\Big)W^T \\ &= \frac{n+1}{n}\Sigma K \Sigma + \frac{\mathrm{Tr}_\Sigma(K)}{n}\Sigma.\end{aligned}$$

□

We quickly put Lemma 4 to work to give a tractable expression for the population risk.

Lemma 5. *For $\theta = (P, Q)$, we have*

$$\begin{aligned}\mathcal{R}_n(\theta) := \mathbb{E}_{A,Y_n}[\|(PA^{-1}Y_n Q - A^{-1})\Sigma^{1/2}\|_F^2] &= \mathbb{E}_A[\|(PA^{-1}\Sigma Q - A^{-1})\Sigma^{1/2}\|_F^2] \\ &+ \frac{1}{n}\mathbb{E}_A\Big[Tr(PA^{-1}\Sigma Q \Sigma Q^T \Sigma A^{-1}P^T) + Tr_\Sigma(Q\Sigma Q^T)Tr(PA^{-1}\Sigma A^{-1}P^T)\Big].\end{aligned}$$

Proof. This follows from a direct computation of the expectation with respect to Y_n :

$$\begin{aligned}
&\mathbb{E}_{A,Y_n}[\|(PA^{-1}Y_nQ - A^{-1})\Sigma^{1/2}\|_F^2] = \mathbb{E}_{A,Y_n}[\text{Tr}((PA^{-1}Y_nQ - A^{-1})\Sigma(Q^TY_nA^{-1}P^T - A^{-1}))] \\
&= \mathbb{E}_{A,Y_n}[\text{Tr}(A^{-1}\Sigma A^{-1} + PA^{-1}Y_nQ\Sigma Q^TY_nA^{-1}P^T - PA^{-1}Y_nQ\Sigma A^{-1} - A^{-1}\Sigma Q^TY_nA^{-1}P^T)] \\
&= \mathbb{E}_A[\text{Tr}(A^{-1}\Sigma A^{-1} - PA^{-1}\Sigma Q\Sigma A^{-1} - A^{-1}\Sigma Q^T\Sigma A^{-1}P^T] \\
&\qquad + \mathbb{E}_{A,Y_n}[\text{Tr}(PA^{-1}Y_nQ\Sigma Q^TY_nA^{-1}P^T)] \\
&= \mathbb{E}_A[\text{Tr}(A^{-1}\Sigma A^{-1} - PA^{-1}\Sigma Q\Sigma A^{-1} - A^{-1}\Sigma Q^T\Sigma A^{-1}P^T] \\
&+ \frac{n+1}{n}\mathbb{E}_A[\text{Tr}(PA^{-1}\Sigma Q\Sigma Q^T\Sigma A^{-1}P^T)] + \frac{1}{n}\mathbb{E}_A[\text{Tr}_\Sigma(Q\Sigma Q^T)\text{Tr}(PA^{-1}\Sigma A^{-1}P^T)] \\
&= \mathbb{E}_A[\|(PA^{-1}\Sigma Q - A^{-1})\Sigma^{1/2}\|_F^2] \\
&\qquad + \frac{1}{n}\mathbb{E}_A\Big[\text{Tr}(PA^{-1}\Sigma Q\Sigma Q^T\Sigma A^{-1}P^T) + \text{Tr}_\Sigma(Q\Sigma Q^T)\text{Tr}(PA^{-1}\Sigma A^{-1}P^T)\Big],
\end{aligned}$$

where we used Lemma 4 to compute the expectation over Y_n in the second-to-last line. □

It will also be useful to derive a simpler expression for the population risk $\mathcal{R}_m(\theta)$ when θ belongs to the set $\Theta_\Sigma = \{(c\mathbf{I}_d, c^{-1}\Sigma^{-1}) : c \in \mathbb{R}\backslash\{0\}\}$.

Lemma 6. *Let $P = c\mathbf{I}_d$, $Q = c^{-1}\Sigma^{-1}$ for $c \in \mathbb{R}\backslash\{0\}$. Then*

$$\mathcal{R}_m(\theta) = \frac{d+1}{n}\mathbb{E}_A\Big[Tr\Big(A^{-1}\Sigma A^{-1}\Big)\Big].$$

Proof. Using Lemma 4 to compute the expectations defining $\mathcal{R}_m$, we have

$$\begin{aligned}
\mathcal{R}_m(\theta) &= \mathbb{E}_A[\text{Tr}(A^{-1}\Sigma A^{-1} - PA^{-1}\Sigma Q\Sigma A^{-1} - A^{-1}\Sigma Q^T\Sigma A^{-1}P^T] \\
&\quad + \frac{n+1}{n}\mathbb{E}_A[\text{Tr}(PA^{-1}\Sigma Q\Sigma Q^T\Sigma A^{-1}P^T)] + \frac{1}{n}\mathbb{E}_A[\text{Tr}_\Sigma(Q\Sigma Q^T)\text{Tr}(PA^{-1}\Sigma A^{-1}P^T)].
\end{aligned}$$

Since $P = c\mathbf{I}_d$ and $Q = c^{-1}\Sigma^{-1}$, we have that $PA^{-1}\Sigma Q\Sigma A^{-1}$, $A^{-1}\Sigma Q^T\Sigma A^{-1}P^T$, and $PA^{-1}\Sigma Q\Sigma Q^T\Sigma A^{-1}P^T$ are all equal to $A^{-1}\Sigma A^{-1}$, and

$$\mathbb{E}_A\text{Tr}_\Sigma(Q\Sigma Q^T)\text{Tr}(PA^{-1}\Sigma A^{-1}P^T) = \mathbb{E}_A\text{Tr}_\Sigma(\Sigma^{-1})\text{Tr}(A^{-1}\Sigma A^{-1}).$$

Therefore, after some algebra, the population risk simplifies to

$$\mathcal{R}_m(\theta) = \frac{1+\text{Tr}_\Sigma(\Sigma^{-1})}{n}\mathbb{E}_A\Big[\text{Tr}\Big(A^{-1}\Sigma A^{-1}\Big)\Big].$$

Noting that $\text{Tr}_\Sigma(\Sigma^{-1}) = d$, we conclude the expression for $\mathcal{R}_m(\theta)$ as stated in the lemma. □

We quote the following result from Theorem 2.1 of Rudelson and Vershynin [2013].

Lemma 7. *[Gaussian concentration bound] Let $y \sim N(0,\Sigma)$. Then*

$$\mathbb{P}\Big\{\|y\| \geq \sqrt{Tr(\Sigma)} + t\Big\} \leq 2\exp\Big(-\frac{t^2}{C\|\Sigma\|_{op}}\Big),$$

where $C > 0$ is a constant independent of Σ and d.

We use the following result to control the error between Q_n and Σ^{-1}.

Lemma 8.

Let $Q_n = B\Big(\frac{n+1}{n}B\Sigma + \frac{Tr_\Sigma(B)}{n}\Sigma\Big)^{-1}$ be as defined in Lemma 1. Assume that n satisfies

$$\frac{\|\Sigma^{-1}\|_{op}\Big\|\Sigma\Big(\mathbf{I}_d + Tr_\Sigma(B)B^{-1}\Big)\Big\|_{op}}{n} \leq \frac{1}{2}.$$

Then we can write

$$Q_n = \Sigma^{-1} + \frac{1}{n}\mathcal{E}_1,$$

where $\mathcal{E}_1$ satisfies

$$\|\mathcal{E}_1\| \lesssim \|\Sigma^{-1}\|_{op}\|\Sigma\|_{op}\Big(1 + Tr_\Sigma(B)\Big)C_A^2.$$

Proof. Using some algebra, we find

$$\begin{aligned}
Q_n &= B\Big(\frac{n+1}{n} B\Sigma + \frac{\mathrm{Tr}_\Sigma(B)}{n}\Sigma\Big)^{-1} \\
&= \Big(\frac{n+1}{n}\Sigma + \frac{\mathrm{Tr}_\Sigma(B)}{n}\Sigma B^{-1}\Big)^{-1} \\
&= \Big(\Sigma + \frac{1}{n}\Sigma\Big(\mathbf{Id} + \mathrm{Tr}_\Sigma(B)B^{-1}\Big)\Big)^{-1}.
\end{aligned}$$

By Lemma 9, we have

$$\|Q_n - \Sigma^{-1}\|_{\text{op}} \le \|\Sigma^{-1}\|_{\text{op}} \cdot \frac{\epsilon^*}{1-\epsilon^*},$$

where

$$\epsilon^* = \frac{\|\Sigma^{-1}\|_{\text{op}}\Big\|\Sigma\Big(\mathbf{I}_d + \mathrm{Tr}_\Sigma(B)B^{-1}\Big)\Big\|_{\text{op}}}{n}.$$

This gives the final bound

$$\|Q_n - \Sigma^{-1}\|_{\text{op}} \lesssim \frac{\|\Sigma^{-1}\|_{\text{op}}\Big\|\Sigma\Big(\mathbf{I}_d + \mathrm{Tr}_\Sigma(B)B^{-1}\Big)\Big\|_{\text{op}}}{n} \le \frac{\|\Sigma^{-1}\|_{\text{op}}\|\Sigma\|_{\text{op}}\Big(1 + \mathrm{Tr}_\Sigma(B)\|B^{-1}\|_{\text{op}}\Big)}{n},$$

Here, we used the bound $\frac{\epsilon}{1-\epsilon} \lesssim \epsilon$ which holds for ϵ sufficiently small; in particular, for $\epsilon \in (0, 1/2)$, we have $\frac{\epsilon}{1-\epsilon} \le 2\epsilon$. □

The following result, used to bound the inverse of a perturbed matrix, is a standard application of matrix power series.

Lemma 9. *Suppose that A is an invertible $d \times d$ matrix and $D \in \mathbb{R}^{d\times d}$ satisfies $\|D\|_{op} \le \frac{\epsilon}{\|A^{-1}\|_{op}}$ for some $\epsilon < 1$. Then*

$$\|(A+D)^{-1} - A^{-1}\|_{op} \le \|A^{-1}\|_{op} \cdot \frac{\epsilon}{1-\epsilon}.$$

Proof. Note that $A + D = (\mathbf{I}_d + DA^{-1})A$. Under our assumption on D, we have $\|DA^{-1}\|_{\text{op}} \le \|D\|_{\text{op}}\|A^{-1}\|_{\text{op}} < 1$, which implies the series expansion

$$(I + DA^{-1})^{-1} = \sum_{k=0}^{\infty}(-DA^{-1})^k.$$

It follows that

$$\begin{aligned}
(A+D)^{-1} &= \Big(\Big(I + DA^{-1}\Big)A\Big)^{-1} \\
&= A^{-1}\Big(I + DA^{-1}\Big)^{-1} \\
&= A^{-1}\sum_{k=0}(-DA^{-1})^k.
\end{aligned}$$

In turn, this gives the bound

$$\begin{aligned}
|(A+D)^{-1} - A^{-1}\|_{\text{op}} &= \Big\|A^{-1}\sum_{k=1}^{\infty}(-DA^{-1})^k\Big\|_{\text{op}} \\
&\le \|A^{-1}\|_{\text{op}}\sum_{k=1}^{\infty}\|DA^{-1}\|_{\text{op}}^k \\
&\le \|A^{-1}\|_{\text{op}}\sum_{k=1}^{\infty}\epsilon^k \\
&= \|A^{-1}\|_{\text{op}}\frac{\epsilon}{1-\epsilon}.
\end{aligned}$$

□

Recall that for a positive definite matrix $\Sigma = W\Lambda W^T$ and a symmetric matrix K,

$$\mathrm{Tr}_\Sigma(K) = \sum_{i=1}^{d} \sigma_i^2 \langle K\varphi_i, \varphi_i\rangle,$$

where $\sigma_1^2, \ldots, \sigma_d^2$ are the eigenvalues of Σ and $\varphi_i = We_i$ are the eigenvectors of Σ.

Lemma 10. *For any symmetric matrix K, we have*

$$Tr_\Sigma(K) \leq \|K\|_{op} Tr(\Sigma).$$

Proof. For each $1 \leq i \leq d$, we have $\langle K\varphi_i, \varphi_i\rangle \leq \|K\varphi_i\|\|\varphi_i\| \leq \|K\|_{\mathrm{op}}$. Therefore,

$$\mathrm{Tr}_\Sigma(K) = \sum_{i=1}^{d} \sigma_i^2 \langle K\varphi_i, \varphi_i\rangle \leq \|K\|_{\mathrm{op}} \sum_{i=1}^{d} \sigma_i^2 = \|K\|_{\mathrm{op}} \mathrm{Tr}(\Sigma).$$

□

In order to prove Theorem 5, we also need the following stability bound of $\mathrm{Tr}_\Sigma(K)$ with respect to perturbations of both Σ and K.

Lemma 11. *Let $\Sigma = W\Lambda W^T$ and $\tilde{\Sigma} = \tilde{W}\tilde{\Lambda}\tilde{W}^T$ be two symmetric positive definite matrices and $K, \tilde{K}$ two symmetric matrices, let $\{\sigma_i^2\}_{i=1}^d$ and $\{\tilde{\sigma}_i^2\}_{i=1}^d$ be the respective eigenvalues of Σ and $\tilde{\Sigma}$ and let $\{\varphi_i\}_{i=1}^d$ and $\{\tilde{\varphi}_i\}_{i=1}^d$ be the respective eigenvectors. Then*

$$\Big| Tr_\Sigma(K) - Tr_{\tilde{\Sigma}}\tilde{K} \Big| \leq Tr(\tilde{\Sigma})\|K - \tilde{K}\|_{op} + \|K\|_{op}\Big(\|\Lambda - \tilde{\Lambda}\|_1 + 2Tr(\tilde{\Sigma})\|W - \tilde{W}\|_{op}\Big).$$

Proof. We have

$$\mathrm{Tr}_\Sigma(K) - \mathrm{Tr}_{\tilde{\Sigma}}(\tilde{K}) \leq \Big|\mathrm{Tr}_\Sigma(K) - \mathrm{Tr}_{\tilde{\Sigma}}(K)\Big| + \Big|\mathrm{Tr}_{\tilde{\Sigma}}(K - \tilde{K})\Big|. \tag{35}$$

□

The second term in 35 can be bounded by an application of Lemma 10, which yields

$$\Big|\mathrm{Tr}_{\tilde{\Sigma}}(K - \tilde{K})\Big| \leq \mathrm{Tr}(\tilde{\Sigma})\|K - \tilde{K}\|_{\mathrm{op}}.$$

To bound the first term in 35, we first use the estimate

$$\Big|\mathrm{Tr}_\Sigma(K) - \mathrm{Tr}_{\tilde{\Sigma}}(K)\Big| \leq \Big|\sum_{i=1}^{d}\Big(\sigma_i^2 - \tilde{\sigma}_i^2\Big)\langle K\varphi_i, \varphi_i\rangle\Big| + \Big|\sum_{i=1}^{d}\tilde{\sigma}_i^2\Big(\langle K(\varphi_i - \tilde{\varphi}_i), \varphi_i\rangle + \langle K\tilde{\varphi}_i, \varphi_i - \tilde{\varphi}_i\rangle\Big)\Big|.$$

The first term above can be bounded by

$$\Big|\sum_{i=1}^{d}\Big(\sigma_i^2 - \tilde{\sigma}_i^2\Big)\langle K\varphi_i, \varphi_i\rangle\Big| \leq \|K\|_{\mathrm{op}} \cdot \sum_{i=1}^{d}\Big|\sigma_i^2 - \tilde{\sigma}_i^2\Big| = \|K\|_{\mathrm{op}} \cdot \|\Lambda - \tilde{\Lambda}\|_1. \tag{36}$$

To bound the second term in 36, note that for any $1 \leq i \leq d$, we have

$$\langle K(\varphi_i - \tilde{\varphi}_i, \varphi_i\rangle \leq \|K\|_{\mathrm{op}}\|\varphi_i - \varphi_i\| \leq \|K\|_{\mathrm{op}}\|W - \tilde{W}\|_{\mathrm{op}},$$

and similarly $\langle K\tilde{\varphi}, \varphi - \tilde{\varphi}\rangle \leq \|K\|_{\mathrm{op}}\|W - \tilde{W}\|_{\mathrm{op}}$. It therefore holds that

$$\Big|\sum_{i=1}^{d}\tilde{\sigma}_i^2\Big(\langle K(\varphi_i - \tilde{\varphi}_i), \varphi_i\rangle + \langle K\tilde{\varphi}_i, \varphi_i - \tilde{\varphi}_i\rangle\Big)\Big| \leq 2\|K\|_{\mathrm{op}}\mathrm{Tr}(\tilde{\Sigma})\|W - \tilde{W}\|_{\mathrm{op}}.$$

Combining all terms yields the final estimate

$$\Big|\mathrm{Tr}_\Sigma(K) - \mathrm{Tr}_{\tilde{\Sigma}}\tilde{K}\Big| \leq \mathrm{Tr}(\tilde{\Sigma})\|K - \tilde{K}\|_{\mathrm{op}} + \|K\|_{\mathrm{op}}\Big(\|\Lambda - \tilde{\Lambda}\|_1 + 2\mathrm{Tr}(\tilde{\Sigma})\|W - \tilde{W}\|_{\mathrm{op}}\Big).$$

The following lemma bounds the 'context mismatch error', which arises in the proof of Theorem 1.

Lemma 12. *The bound*

$$\sup_{\|\theta\|\leq M}\Big|\mathcal{R}_m(\theta)-\mathcal{R}_n(\theta)\Big| \leq 2M^4c_A^2\max(Tr(\Sigma),\|\Sigma\|_{op}^2)Tr(\Sigma)\Big|\frac{1}{n}-\frac{1}{m}\Big|$$

holds.

Proof. Denote $\theta=(P,Q)$. Recall that, as a direct consequence of Lemma 5, we have

$$\begin{aligned}\mathcal{R}_n(\theta) &= \mathbb{E}_A\big[\mathrm{Tr}(A^{-1}\Sigma A^{-1}) - \mathrm{Tr}(PA^{-1}\Sigma Q\Sigma A^{-1}) - \mathrm{Tr}(A^{-1}\Sigma Q^T\Sigma A^{-1}P^T) \\ &\quad + \frac{n+1}{n}\mathrm{Tr}(PA^{-1}\Sigma Q\Sigma Q^T\Sigma A^{-1}P^T) + \frac{\mathrm{Tr}_\Sigma(Q\Sigma Q^T)}{n}\mathrm{Tr}(PA^{-1}\Sigma A^{-1}P^T)\big],\end{aligned}$$

An analogous expression holds for $\mathcal{R}_m(\theta)$. Therefore, for θ satisfying $\|\theta\|=\max(\|P\|_{\mathrm{op}},\|Q\|_{\mathrm{op}})\leq M$, we have the bound

$$\begin{aligned}\Big|\mathcal{R}_m(\theta)-\mathcal{R}_n(\theta)\Big| &= \Big|\frac{1}{n}-\frac{1}{m}\Big|\Big|\mathbb{E}_A\Big[\mathrm{Tr}(PA^{-1}\Sigma Q\Sigma Q^T\Sigma A^{-1}P^T)+\mathrm{Tr}_\Sigma(Q\Sigma Q^T)\mathrm{Tr}(PA^{-1}\Sigma A^{-1}P^T)\Big]\Big| \\ &\leq \Big|\frac{1}{n}-\frac{1}{m}\Big|\cdot 2M^4c_A^2\max(\mathrm{Tr}(\Sigma),\|\Sigma\|_{\mathrm{op}}^2)\mathrm{Tr}(\Sigma).\end{aligned}$$

□

The following lemma is an adaptation of Wald's consistency theorem of M-estimators [Van der Vaart, 2000, Theorem 5.14]. We use it to prove the convergence in probability of empirical risk minimizers to population risk minimizers.

Lemma 13. *Let $\theta\in\mathbb{R}^m$, $x\in\mathbb{R}^d$, and suppose $\ell(\cdot,\cdot):\mathbb{R}^d\times\mathbb{R}^m\to[0,\infty)$ is lower semicontinuous in θ. Let $m_0=min_\theta\mathbb{E}[\ell(x,\theta)]$ for some fixed distribution on x, and let $\Theta_0=argmin_\theta\mathbb{E}[\ell(x,\theta)]$. Let $\{\theta_N\}_{N\in\mathbb{N}}$ be a collection of estimators such that $\sup_N\|\theta_N\|<\infty$ and*

$$m_0-\mathbb{E}_N[\ell(x,\theta_0)]=o_P(1)$$

Then $dist(\theta_N,\Theta_0)\xrightarrow{\mathrm{P}}0$.

Proposition 4. *For any sequence $\{\widehat{\theta}_{n,N}\}_{n,N\in\mathbb{N}}$ of minimizers of the empirical risk $\mathcal{R}_{n,N}$ with $\sup_N\|\widehat{\theta}_{n,N}\|<\infty$ for all n, we have*

$$\lim_{n\to\infty}\lim_{N to\infty} dist(\widehat{\theta}_{n,N},\mathcal{M}_\infty)=0,\ \textit{in probability.}$$

Proof. For each fixed $n\in\mathbb{N}$. we can apply Lemma 13 to the empirical risk minimizer $\widehat{\theta}_{n,N}$. In this context, the condition of the lemma amounts to the condition that $\mathcal{R}_n(\theta_*)-\mathcal{R}_{n,N}(\widehat{\theta}_{n,N})=o_P(1)$, for any $\theta_*\in\mathrm{argmin}_\theta\mathcal{R}_n$, which is satisfied since

$$\mathcal{R}_n(\theta_*)-\mathcal{R}_{n,N}(\widehat{\theta}_{n,N})=\Big(\mathcal{R}_n(\theta_*)-\mathcal{R}_{n,N}(\theta_*)\Big)+\Big(\mathcal{R}_{n,N}(\theta_*)-\mathcal{R}_{n,N}(\widehat{\theta}_{n,N})\Big).$$

The first term tends to zero in probability by the law of large numbers, and the second term is non-negative by the minimality of $\widehat{\theta}_{n,N}$. This proves that

$$\lim_{N to\infty}\mathrm{dist}(\widehat{\theta}_{n,N},\mathcal{M}_n)=0,\ \text{in probability},$$

where $\mathcal{M}_n=\mathrm{argmin}_\theta\mathcal{R}_n(\theta)$. Consequently, since $\mathcal{R}_n$ and $\mathcal{R}_\infty$ are polynomials in θ such that the coefficients of $\mathcal{R}_n$ converge to the coefficients of $\mathcal{R}_\infty$ as $n\to\infty$, we have by the triangle inequality that

$$\lim_{n\to\infty}\lim_{N\to\infty}\mathrm{dist}(\widehat{\theta}_{n,N},\mathcal{M}_\infty)=0,\ \text{in probability}.$$

□

H Experimental setup

H.1 In-domain generalization

We recapitulate the experimental set-up described in Subsection 4.1 for our in-domain experiments. We consider the one dimensional elliptic PDE $(-\Delta+V(x))u(x)=f(x)$ on $\Omega=[0,1]$ with Dirichlet boundary condition. We assume that the source term is a Gaussian white noise, i.e. $f=N(0,\mathbb{I})$, where $\mathbb{I}$ denotes the identity operator. We discretize the PDE using Galerkin projection onto the sine basis $\phi_k(x)=\sin(k\pi x)$, $k\in\{1,\dots,d\}$. Furthermore, we assume that the potential V is uniform random field that is obtained by dividing the domain into $2d+1$ sub-intervals and in each cell independently, the potential V takes values uniformly in $[1,2]$. This leads to the linear system $A\mathbf{u}=\mathbf{f}$, where $\mathbf{f}\sim N(0,\mathbf{I}_d)$ and

$$A_{ij}=k^2\pi^2\delta_{ij}+\langle\phi_i,V\phi_j\rangle_{L^2}.$$

The prompts used for pre-training are then built on observations of the form $((\mathbf{f}_1,A^{-1}\mathbf{f}_1),\dots,(\mathbf{f}_n,A^{-1}\mathbf{f}_n))$.

H.2 Out-of-domain generalization

For out-of-domain generalization, we consider the PDE defined by $-\nabla\cdot(a(x)\nabla u(x))+V(x)u(x)=f(x)$ on $[0,1]$ with Dirichlet boundary conditions.

Task shifts: During both training and inference, we assume that f is a centered Gaussian with covariance operator defined by $(-\Delta+c\mathbb{I})^{-\beta}$ for some fixed $c,\beta>0$. We parameterize $a(x)$ as a *log-normal random field*, i.e., we write $a(x)=e^{b(x)}$, where $b(x)$ is sampled from an infinite-dimensional Gaussian measure $N(0,C_{\alpha,\tau})$, where $C_\alpha=(-\Delta+\tau\mathbb{I})^{-\alpha}$. The parameter α governs the smoothness of the field. During training, we set $\alpha=3,\tau=5$, and during inference we use $\alpha=1,2,4$. For V, we assume during training that V is piecewise constant, and the constant values are iid according to the uniform distribution $U(1,2)$. During inference, we shift the distribution on the pieces of V to $U(3,4)$, $U(5,10)$, and $U(10,20)$.

Covariate shifts: We train the model to solve the PDE (1), where the source term is defined by a Gaussian measure $N(0,C)$ for $C=(-\Delta+c\mathbb{I})^{-\beta}$, where $c=\beta=1$ Then, at inference, we consider solving the same PDE, but where the source term is defined by $N(0,3C)$ or $N(0,5C)$; see Figure 3: C. We also consider covariate shifts defined by changing the parameters c and β in the covariance; see Figure 5 in Appendix I.

I Additional numerical results

In this section, we present some additional numerics. The plots in Figure 4: A.1-C.1 are identical to those in Figure 1: A-C, but Figure 1: A.2 - C.2 also show the slopes of the log-log plots as a function of the sample size. This makes it easier to compare the empirical scaling laws with those derived in Theorem 1. Figure 5 depicts the relative H^1-error of the pre-trained transformer under covariate shifts with respect to a set of parameters in the covariance operator that are different from the one discussed in Section 4.2. More precisely, we recall that the source term f is sampled from a centered Gaussian measure on $L^2([0,1])$ with covariance operator given by $(-\Delta+c\mathbb{I})^{-\beta}$. During training, we set the parameters of the covariance as $\beta=c=1$. We then shift the parameters of the covariance during inference, as defined by the legend of Figure 5: A. Figures 5: B shows the heat map of the relative H^1-error with respect to the parameters α and τ. Note that the shift on the covariance operator of f defined in Figure 5 differs from the shift defined in Figure 3: C, where the shift on the covariance operator was defined by constant multiplication. Both cases validate Theorem 5 and provide further evidence that pre-trained transformers are not robust with respect to covariate shifts. In particular, the prediction errors are more sensitive to the shifts in the amplitude of field and the smoothness parameter β than the shift in the shift parameter c. Figure 6 complements Figure 3 with an additional heat map of the relative H^1-error under tasks shift in the diffusion coefficient a with respect to the parameters α and τ. Figure 6 shows that the prediction errors under task shifts remain decently small in a wide range of parameter shifts.

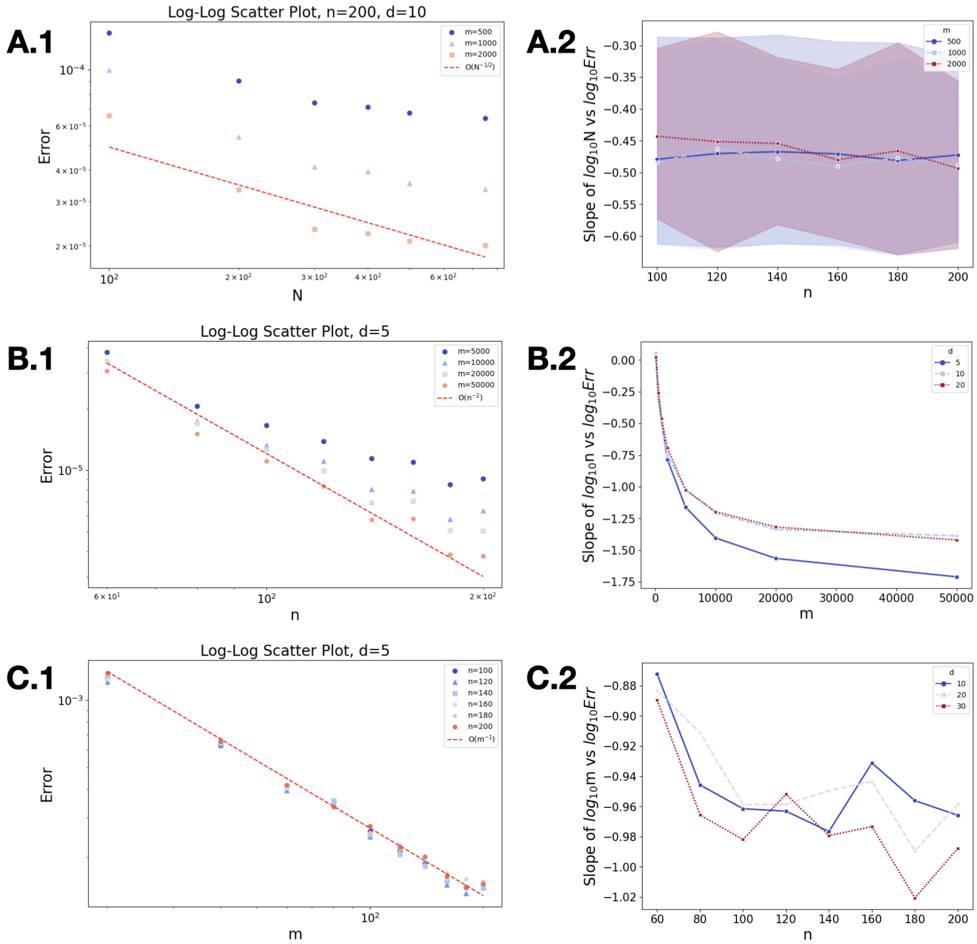

Figure 4: Plots A.1-C.1 are identical to those shown in Figure 1. Plots A.2-C.2 show the slopes of the error curves in the left column as functions of various sample sizes.

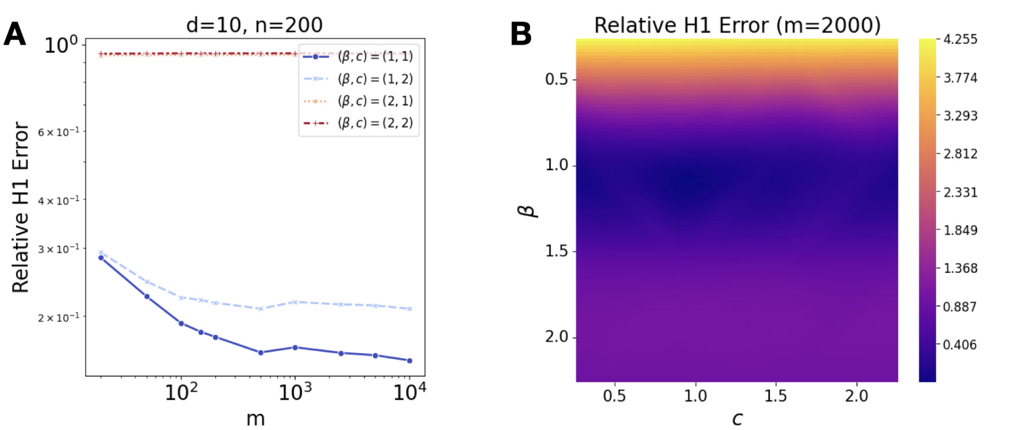

Figure 5: The figures show the relative H^1-error of learning the linear systems under covariate shifts in the covariance operator $C = (-\Delta + c\mathbb{I})^{-\beta}$ with respect to the parameters c and β. During training, we set $c = \beta = 1$. Figure A plots the error curves corresponding to four parameter pairs (β, c) as a function of the testing prompt length. Figure B plots the errors for the data corresponding to a wide range of parameter pairs.

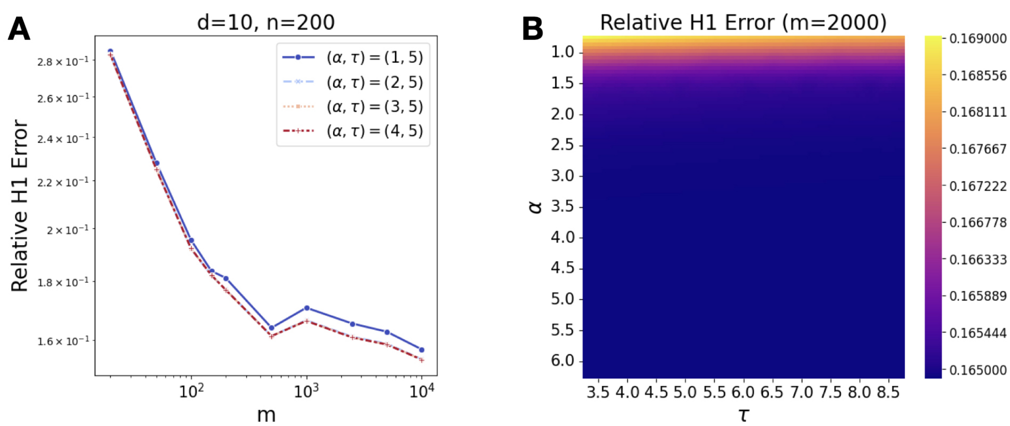

Figure 6: Figure A shows the relative H^1 error as a function of the prompt length under shifts on the distribution of $a(x)$ (the training distribution is $a(x) = e^{b(x)}$ with $b(x) \sim N(0, (-\Delta + \tau\mathbb{I})^{-\alpha})$, $\alpha = 3$ and $\tau = 5$). Figure B shows the corresponding heat map for the relative H^1 error with respect to the parameters α and τ.