Impact of Positional Encoding: Clean and Adversarial Rademacher Complexity for Transformers Under In-Context Regression

Weiyi He
Michigan State University

Yue Xing
Michigan State University

Abstract

Positional encoding (PE) is a core architectural component of Transformers, yet its impact on the Transformer's generalization and robustness remains unclear. In this work, we provide the first generalization analysis for a single-layer Transformer under in-context regression that explicitly accounts for a completely trainable PE module. Our result shows that PE systematically enlarges the generalization gap. Extending to the adversarial setting, we derive the adversarial Rademacher generalization bound. We find that the gap between models with and without PE is magnified under attack, demonstrating that PE amplifies the vulnerability of models. Our bounds are empirically validated by a simulation study. Together, this work establishes a new framework for understanding the clean and adversarial generalization in ICL with PE.

1 INTRODUCTION

Large language models (LLMs), built upon the basic Transformer architecture (Vaswani et al., 2017; Brown et al., 2020; Radford et al., 2019), have demonstrated remarkable capabilities of in-context learning (ICL), allowing them to perform new tasks based on a few examples provided in a prompt, without any parameter updates (Brown et al., 2020; Dong et al., 2022). This powerful few-shot learning paradigm has shown impressive empirical performance across various complex tasks (Liang et al., 2022; Raventós et al., 2023).

The success of ICL has motivated several theoretical studies to explain the underlying mechanism of ICL (Li et al., 2023; Deng et al., 2023; Zhang et al., 2024). These theoretical works often focus on understanding how Transformers can learn from examples by framing the process as an implicit implementation of classic algorithms (Von Oswald et al., 2023; Ahn et al., 2023; Xie et al., 2021). However, these analyses often rely on key simplifications. For example, the model in Cui et al. (2024) only considers the basic attention score in their formulation. The Positional Encoding (PE), which is an important component for encoding positional information, is missing. While some empirical works have investigated properties of PEs (Gao et al., 2024; Çavşi Zaim and Yolacan, 2025), a formal generalization theory is still lacking.

Furthermore, the urgency of filling this theoretical gap is also underscored by the adversarial vulnerability of LLMs. Despite the power of ICL, these models are highly vulnerable to carefully crafted adversarial attacks that can degrade performance or bypass safety alignments (Shayegani et al., 2023; Zou et al., 2023). For instance, the jailbreaking attack can trick the models into generating harmful or toxic content (Wei et al., 2023), while prompt injection can hijack the model's behavior for malicious purposes (Liu et al., 2023). Other literature also develops poisoning attacks empirically on ICL examples, e.g., He et al. (2025).

Observing the above limitations, this work provides a unified framework for ICL generalization based on Rademacher complexity (RC). Our framework extends the foundational theory of ICL to two crucial dimensions: the impact of trainable PEs on the model complexity, and the model's adversarial generalization under adversarial attacks.

Our contributions are summarized as follows:

First, we provide the generalization bound for a family of one-layer Transformers, **in the ICL regression setting,** that explicitly includes a completely trainable PE module. To address the technical challenge of analyzing PE parameters that are deeply embedded within the non-linear self-attention mechanism, prior theoretical work often absorbs these parameters into

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

a single, uniform norm bound (Edelman et al., 2022; Li, 2025). Our analysis quantifies the increase in the model's Rademacher complexity from PE.

Second, we extend our analysis to adversarial generalization by deriving the Adversarial Rademacher Complexity (ARC) bounds for the Transformer. Our approach builds on the work of Xiao et al. (2022a), which connects robustness to the covering number of the function class. A technical challenge is that the core tools in Xiao et al. (2022a) are designed for classification tasks. To handle the MSE loss in ICL regression, we employ the surrogate loss method, a general technique to bound the complex adversarial loss with a simpler function (Khim and Loh, 2018; Yin et al., 2019; Gao and Wang, 2021). Our derived bounds reveal that the difference between the bounds with/without PE is larger under adversarial attack, indicating that Transformers with PE are more vulnerable.

Finally, we validate our theoretical findings with experiments. The results show that Transformers with a trainable PE have a consistently larger generalization gap and are more sensitive to adversarial attacks on in-context examples. We also observe that the gap increases with stronger attacks but decreases as the context length grows, confirming that longer contexts still improve robustness even under adversarial attacks.

2 RELATED WORKS

Theoretical Analysis of ICL. Many theoretical works aim to demystify the mechanisms behind ICL. Garg et al. (2022) models the Transformer as an implicit algorithm that performs in-context learning of linear functions. Subsequent works have extended this framework with different focuses. Xing et al. (2024) studied the role of each component in the transformer architecture to learn from the unstructured data. The work of Cui et al. (2024) demonstrated that multi-head attention with a substantial embedding dimension performs better than single-head attention. Other related works can be found in Min et al. (2021); Chen et al. (2022); Zhang et al. (2022). This line of work typically involves analyzing a sufficiently large amount of data and examining how the Transformer performs in the ideal case.

Generalization Bounds for Transformers. The generalization ability of Transformers has been studied through various theoretical perspectives. One line of work relies on the uniform bounds for the input data matrix. Their bounds may depend on the length of the sequence (Edelman et al., 2022) or be independent of it under a certain format of covering numbers (Trauger and Tewari, 2024). Other approaches focus on specific settings, such as margin-based bounds for binary classification (Wei et al., 2022). In contrast, our work provides an analysis specifically tailored to the ICL setting. By adopting a specific Gaussian assumption, we can derive a more fine-grained bound that explicitly captures how the generalization gap decays with the context length t.

Adversarial Generalization. Adversarial generalization is a broad field that utilizes various tools to provide theoretical insights. These include methods based on VC-dimension (Montasser et al., 2019; Attias et al., 2022), algorithmic stability (Xing et al., 2021; Xiao et al., 2022b), and distributional robustness (Sinha et al., 2017). Moreover, Xiao et al. (2022a) provided the bound of Adversarial Rademacher Complexity (ARC) for deep neural networks based on the covering numbers. However, the adversarial generalization for the Transformers on ICL based on ARC remains unexplored, and our work provides the ARC results for the considered ICL setting.

3 PRELIMINARIES

3.1 ICL Task and Transformer Architecture

To mathematically define ICL, instead of merely passing a query $x_q \in \mathbb{R}^d$ to the transformer to make a prediction, ICL passes a prompt, i.e., a few examples with their labels $\{(x_i, y_i)\}_{i=1,\dots,t}$ together with the query x_q, to the transformer. Using the prompt in the format of

$$X = \begin{pmatrix} x_1 & x_2 & \dots & x_t & x_q \\ y_1 & y_2 & \dots & y_t & 0 \end{pmatrix}^\top \in \mathbb{R}^{(t+1)\times(d+1)}. \quad (1)$$

The transformer can learn from the examples to infer the prediction for x_q.

Following Trauger and Tewari (2024), we define the simplified single-layer, single-head Transformer architecture as follows. Let the input sequence be $X \in \mathbb{R}^{(t+1)\times(d+1)}$ as defined in Eq. (1). The model parameters θ consist of the trainable weight matrices $W_{\text{in}} \in \mathbb{R}^{(d+1)\times d_m}$, $W_Q, W_K, W_V \in \mathbb{R}^{d_m \times d_m}$, and $W_c \in \mathbb{R}^{d_m}$, where d_m is the embedding dimension. Following Edelman et al. (2022), we bound these matrices by some constants $\|W_c^T\|_{1,\infty} \le B_{W_c}, \|W_v^T\|_{1,\infty} \le B_{W_v}$.

The model first computes content embeddings via a read-in matrix W_{in}. To provide the model with sequence order information, we add a PE matrix $P \in \mathbb{R}^{(t+1)\times d_m}$. The final input representation H is the sum of content and positional embeddings:

$$H = XW_{\text{in}} + P. \quad (2)$$

In this work, we specifically analyze the case where P is a completely trainable parameter and $\|P\|_F \le B_P$.

For simplicity, we combine the query and key matrices into $W_{QK} = W_Q W_K^\top \in \mathbb{R}^{d_m \times d_m}$. Let σ be an L_σ-Lipschitz activation function. The attention head computes an output representation matrix $H' \in \mathbb{R}^{(t+1)\times d_m}$ as:

$$H' = \sigma\left(\text{RowSoftmax}\left(\frac{HW_{QK}H^\top}{\sqrt{d_m}}\right)HW_V\right). \quad (3)$$

The final prediction of the model, f_θ, is a linear readout from the last row of H', which corresponds to the query token. Let $h'_q \in \mathbb{R}^{d_m}$ be the last row of H'. The prediction is given by $f_\theta(X) = W_c^\top h'_q$.

3.2 Training and Adversarial Loss

We consider a Transformer parameterized by $\theta \in \Theta$. The model is trained to minimize the standard Mean Squared Error (MSE) loss. For a given prompt containing context $S_t = \{(x_i, y_i)\}_{i=1}^t$ and a query x_q, the model produces a prediction $\hat{y}_q$. Then we minimize the expected risk:

$$L(\theta) = \mathbb{E}_{S_t, x_q, y_q}\left(\hat{y}_q(S_t, x_q; \theta) - y_q\right)^2. \quad (4)$$

In the adversarial setting, a perturbed input matrix X' is constructed as:

$$X' = \begin{pmatrix} x'_1 & x'_2 & \dots & x'_t & x'_q \\ y_1 & y_2 & \dots & y_t & 0 \end{pmatrix}^\top \in \mathbb{R}^{(t+1)\times(d+1)}, \quad (5)$$

where each x'_i is a perturbed version of the original x_i and the attacked prompt is then $S'_t = \{(x'_i, y_i)\}_{i=1}^t$. We consider a global threat model where the total perturbation on the input matrix is bounded by $\|X' - X\|_F \le \varepsilon$. This defines the formal definition of the adversarial loss:

$$\tilde{\ell}_\theta(X, y_q) \triangleq \sup_{X' \text{ s.t. } \|X'-X\|_F \le \varepsilon} \left(\hat{y}_q(X'; \theta) - y_q\right)^2, \quad (6)$$

where $\hat{y}_q(X'; \theta)$ denotes the model's prediction for the query based on the entire perturbed input matrix X'.

3.3 Generalization Framework

Our theoretical analysis is based on the standard framework of generalization bounds. We aim to bound the generalization gap, $\sup_{h\in\mathcal{H}}(R(h) - \hat{R}_S(h))$, where $R(h) = \mathbb{E}_{(x,y)}[\ell(h(x), y)]$ is the true risk and $\hat{R}_S(h) = \frac{1}{m}\sum_{i=1}^m \ell(h(x_i), y_i)$ is the empirical risk over a sample S of size m. The primary tool is the Rademacher complexity, which quantifies the capacity of a function class. The following result connects the two concepts together (Shalev-Shwartz and Ben-David, 2014):

Proposition 3.1 (Generalization Bound). *Let $\mathcal{H}$ be a hypothesis class and ℓ be a loss function. If the magnitude of our loss function is bounded above by c, with probability at least $1-\delta$:*

$$\sup_{h\in\mathcal{H}}\left(R(h) - \hat{R}_S(h)\right) \le 2\,\text{Rad}_m(\ell\circ\mathcal{H}) + 4c\sqrt{\frac{2\log(4/\delta)}{m}},$$

where $\ell \circ \mathcal{H}$ is the class of functions formed by composing the loss with the hypotheses.

We provide the remaining definitions in Appendix A.

4 MAIN RESULTS

In this section, we present the standard and robustness generalization bound for the Transformer. We first state the data generation model as follows.

Assumption 4.1 (**Data Generation Model**). In each prompt, the example pairs $\{(x_i, y_i)\}_{i=1}^t$ and the query (x_q, y_q) are i.i.d. samples from the following noiseless linear regression model:

- The input $x \sim \mathcal{N}(0, I_d)$.
- The output $y = \mu^\top x$.
- The true coefficients $\mu \in \mathbb{R}^d$ are the same for all samples within a prompt but are drawn independently for each prompt from $\mu \sim \mathcal{N}(0, I_d/d)$.

The Assumption 4.1 follows from Cui et al. (2024) and Zhang et al. (2024) for the data generation model. This standard setting creates a clean and analytically tractable environment, allowing us to focus on the mechanism of in-context learning itself. Our proof can also be extended to more general sub-Gaussian distributions (see Appendix D for details).

4.1 RC for Transformer without PE

In this section, we derive the generalization bound for a one-layer Transformer in the standard ICL setting. The analysis is grounded in the Dudley integral framework, which connects the Rademacher complexity of a function class to its geometry via covering numbers.

We first present the final theorem:

Theorem 4.1 (RC for Transformer without PE). *Under Assumption 4.1, using the architecture defined in Section 3.1, let the context S_t be sampled i.i.d. from the data distribution. Let $W^*_{QK}(S_t)$ be the ideal attention matrix for the context S_t, and W^*_{QK} be a global ideal attention matrix for the transformer. Assuming that the trained model's attention matrix W_{QK} is close to W^*_{QK}, that is $\|W_{QK} - W^*_{QK}\|_F \le \gamma$, for a constant $\gamma > 0$. Then, with probability at least $1-\delta$ over the choice of S_t, there exists a constant L_f and a tolerance r (dependent of γ) such that the Rademacher complexity of the $\mathcal{F}_{S_t}$ is bounded by*

$$\text{Rad}_S(\mathcal{F}_{S_t}) \lesssim O\left(\frac{L_f\sqrt{rD}\sqrt{\log t/r}}{\sqrt{mt}}\right), \quad (7)$$

where $D = d \cdot d_m + d_m^2$.

This theorem provides the generalization bound for ICL regression that explicitly shows a decay with context length t. In contrast to prior work that provides sequence-length-independent bounds (Trauger and Tewari, 2024), our result captures the core learning dynamics of ICL.

There are multiple steps to prove Theorem 4.1. We start from the following lemma.

Lemma 4.1 (Generalization Bound via Dudley's Integral)**.** *Let* $\mathcal{F}_{S_t}$ *be the function class of a Transformer constrained by* S_t*, and* $D_{S_t} = \mathrm{Diam}(\mathcal{F}_{S_t})$*. Then its empirical Rademacher complexity is bounded by:*

$$\mathrm{Rad}_S(\mathcal{F}_{S_t}) \lesssim \frac{1}{\sqrt{m}} \int_0^{D_{S_t}} \sqrt{\log N(\mathcal{F}_{S_t}, \|\cdot\|_{m,2}, \varepsilon)}\, d\varepsilon, \tag{8}$$

where $\|\cdot\|_{m,2}$ *is the empirical* 2*-norm and* $N(\cdot)$ *is the covering number. The diameter of a function class* $\mathcal{F}$ *is defined as* $\mathrm{Diam}(\mathcal{F}) \triangleq \sup_{f_1, f_2 \in \mathcal{F}} \|f_1 - f_2\|_{m,2}$.

While Lemma 4.1 provides a general bound, the geometry of the function class $\mathcal{F}_{S_t}$ is complicated to analyze directly. To overcome this challenge, we first apply standard contraction results to peel out the W_c, σ, and W_V. This reduces the problem to the attention block that only depends on $(W_{\mathrm{in}}, W_{QK})$. Our key insight is to analyze the effective parameter space Θ_{S_t}, as defined in Definition 4.1. Finally, we obtain the corresponding covering number of Θ_{S_t} in Lemma 4.2, and connect it to the ICL solution space via Lemma 4.3 and Lemma 4.4.

Definition 4.1 (Effective Parameter Space)**.** Under the conditions of Theorem 4.1, we define the effective parameter space as

$$\Theta_{S_t} := \left\{ (W_{\mathrm{in}}, W_{QK}) : \|W_{QK} - W_{QK}^*(S_t)\|_F \le \gamma_{\mathrm{eff}} \right\},$$

Let $\Delta_t := \|W_{QK}^*(S_t) - W_{QK}^*\|_F$, as $\Delta_t \lesssim O(\sqrt{d/t})$ with high probability (see Appendix C.1), for any fixed $t_{\min}$ and all $t \ge t_{\min}$, we can set constant $\gamma_{\mathrm{eff}} = \gamma + \Delta^*$ and $\Delta^* := \sup_{t \ge t_{\min}} \Delta_t$.

With the above definition, we can formally define $\mathcal{F}_{S_t} := \{f_\theta : \theta \in \Theta_{S_t}\}$. Then we can connect the properties of Θ_{S_t} back to the $\mathcal{F}_{S_t}$ via a standard Lipschitz continuity assumption, and obtain the covering number of $\mathcal{F}_{S_t}$ as in the following lemma:

Lemma 4.2 (Covering Number of the Parameter Space)**.** *Under the conditions of Theorem 4.1 and Definition 4.1. Assume that for any* $\theta_1, \theta_2 \in \Theta$*, there is*

$$\|f_{\theta_1} - f_{\theta_2}\|_{m,2} \le L_f \|\theta_1 - \theta_2\|_2, \tag{9}$$

where L_f *is a Lipschitz constant determined by the architecture of Transformer (Appendix C.2). As the definition of* Θ_{S_t}*, the total number of parameters is therefore* $D = d \cdot d_m + d_m^2$*. Let* $D_t = \mathrm{Diam}(\Theta_{S_t})$*, the covering number is bounded by:*

$$\log N(\Theta_{S_t}, \|\cdot\|_2, \varepsilon) \le D \log(3D_t/\varepsilon), \tag{10}$$

where the diameter for a set W *is defined as* $\mathrm{Diam}(W) = \sup\{\|w_1 - w_2\|_2 : w_1, w_2 \in W\}$.

A direct consequence is that the geometry of $\mathcal{F}_{S_t}$ can be controlled by the geometry of Θ_{S_t}:

$$\begin{aligned} N(\mathcal{F}_{S_t}, \|\cdot\|_{m,2}, \varepsilon) &\le N(\Theta_{S_t}, \|\cdot\|_2, \varepsilon/L_f), \\ \mathrm{Diam}(\mathcal{F}_{S_t}) &\le L_f \cdot \mathrm{Diam}(\Theta_{S_t}). \end{aligned} \tag{11}$$

Inequalities (11) allow us to bound the Dudley integral by analyzing the parameter space, whose covering number can be bounded using standard results.

After obtaining the covering number w.r.t. Θ_{S_t}, the final step is to connect Θ_{S_t} and its diameter D_t back to ICL. To do this, following recent ICL literature, we introduce the concept of an effective linear weight, w_{eff}, which approximates the Transformer's ability as a simple linear model. To be specific, we can take the Transformer's ICL process for linear regression as finding an effective linear weight w_{eff} such that the prediction for a query x_q is approximately $w_{\mathrm{eff}}^\top x_q$, i.e.,

$$w_{\mathrm{eff}} := \arg\min_{w \in \mathcal{R}^d} \mathbb{E}_{x_q \sim N(0, I_d)} \left(\hat{y}_q - w^\top x_q\right)^2. \tag{12}$$

Given the above definition, we are interested in the solution space W_{S_t}, which is the set of all such w_{eff} generated by the allowable parameters in Θ_{S_t}. We formalize the geometry of W_{S_t} as follows.

Lemma 4.3. *Under the conditions of Theorem 4.1, suppose the model is on a high-probability "good event"* $\mathcal{G}_t$ *where the data matrix* $X_t \in \mathbb{R}^{t \times d}$ *is well-conditioned as justified in Appendix C.2. Given the solution space* W_{S_t} *and the model's parameter space* Θ_{S_t}*, we denote the mapping* $G : \theta \mapsto w_{\mathrm{eff}}$*. Assume* G *is inverse Lipschitz continuous where there exists a constant* $L_G > 0$*, independent of* t*, such that for any* $\theta_1, \theta_2 \in \Theta_{S_t}$*:*

$$\|\theta_1 - \theta_2\|_2 \le L_G \|G(\theta_1) - G(\theta_2)\|_2. \tag{13}$$

If we model the set of effective linear weights W_{S_t} *as a tolerance ellipsoid:*

$$W_{S_t} = \left\{ w : (w - \hat{w})^\top (X_t^\top X_t)(w - \hat{w}) \le r \right\}, \tag{14}$$

where $\hat{w} \in \mathbb{R}^d$ *is the OLS solution and* $r > 0$ *is a tolerance constant. Then for* $t > d$*,*

$$\mathrm{Diam}(\Theta_{S_t}) \lesssim L_G \,\mathrm{Diam}(W_{S_t}) \lesssim O\left(L_G \sqrt{r/t}\right). \tag{15}$$

Finally we provide some justifications on the existence of L_G and the size of r in Eq. (15):

Lemma 4.4 (Justification for L_G and r)**.** *Under the conditions of Theorem 4.1 and Lemma 4.3, assume that G is C^1 on a compact set $K \in \Theta_{S_t}$ and there exist constants $c_G, c_U > 0$ such that for all $\theta \in K$ the Jacobian $J_G(\theta)$ has full row rank d and satisfies $c_G \le \sigma_{\min}(J_G(\theta)) \le \sigma_{\max}(J_G(\theta)) \le c_U$. Then $L_G \le M(c_G, c_U)$. Moreover, there exists a constant $C_r > 0$ (independent of t and related to γ_{eff}) and one can choose a tolerance r such that $r \le C_r$.*

See the justification of all the lemmas in Appendix C.2 and the full proof of Theorem 4.1 in Appendix C.3.

4.2 RC for Transformer with PE

While Section 4.1 gives a bound for a Transformer without PE, the standard architectures universally include PEs to handle sequence order. Although in the context of ICL, the order of (x_i, y_i) in the prompt does not influence the solution for linear regression, adding PEs in the model increases the number of parameters, thus increasing the parameter space, and our focus is on quantifying the potential negative impact of PE on the generalization bound. Given the above, our next step is to understand how incorporating completely trainable parameters impacts the generalization bounds.

To consider PE, we employ a function class decomposition. We model any function f_{PE} from the class $\mathcal{F}_{\text{PE}}$ as the sum of a function from the no-PE class $\mathcal{F}_{\text{NoPE}}$ and a bias term f_{bias}:

$$f_{\text{PE}} = f_{\text{NoPE}} + f_{\text{bias}},$$

where f_{bias} belongs to the class $\mathcal{F}_{\text{bias}} \triangleq \{f_{\text{PE}} - f_{\text{NoPE}}\}$. This implies $\mathcal{F}_{\text{PE}} \subseteq \mathcal{F}_{\text{NoPE}} + \mathcal{F}_{\text{bias}}$. The subadditivity of the Rademacher complexity then allows us to bound the total complexity as:

$$\text{Rad}_S(\mathcal{F}_{\text{PE}}) \le \text{Rad}_S(\mathcal{F}_{\text{NoPE}}) + \text{Rad}_S(\mathcal{F}_{\text{bias}}). \quad (16)$$

This decomposition simplifies the analysis by only bounding the complexity of the bias class, $\text{Rad}_S(\mathcal{F}_{\text{bias}})$, which depends on the magnitude of the functional difference introduced by the PE. We formalize this with the following assumption.

Assumption 4.2 (Bounded PE Effect)**.** Let $w_{\text{PE}}, \eta_{\text{PE}}$ and $w_{\text{NoPE}}, \eta_{\text{NoPE}}$ be the effective linear weights and error for models with and without PE, respectively. We assume the difference for the weight is bounded by a constant $C_{PE} > 0$:

$$\|w_{\text{PE}} - w_{\text{NoPE}}\|_2 \le C_{PE}. \quad (17)$$

The above assumption supposes that adding the bounded, trainable PE matrix to the input embeddings results in a correspondingly bounded change to the model's effective linear weight.

To be specific, the function class $\mathcal{F}_{\text{bias}}$ represents the difference in the models' outputs, $f_{\text{PE}}(X) - f_{\text{NoPE}}(X)$. This difference can be decomposed into a linear component, $(w_{\text{PE}} - w_{\text{NoPE}})^\top x_q$, and a non-linear residual component, $\zeta(X) = \eta_{\text{PE}} - \eta_{\text{NoPE}}$. In our analysis, we focus on the dominant linear effect. This is justified because the residual term $\zeta(X)$ is a lower-order effect. As we argue in Appendix C.4, the difference $\zeta(X)$ is bounded by a t-independent constant related to γ_{eff}. We therefore analyze the complexity of the linear function class: $\mathcal{F}_{\text{bias}} = \{x \mapsto (\Delta w)^\top x \mid \|\Delta w\|_2 \le C_{PE}\}$.

With this framework, we can get the result with PE.

Theorem 4.2 (RC with PE)**.** *With Eq.(17) and Assumption 4.2, let $\mathcal{F}_{S_t,PE}$ be the function class of a Transformer with a completely trainable PE. The Rademacher complexity for $\mathcal{F}_{S_t,PE}$ is bounded as:*

$$\text{Rad}_S(\mathcal{F}_{S_t,PE}) \lesssim O\left(\frac{L_f\sqrt{rD}\sqrt{\log t/r}}{\sqrt{mt}}\right) + \frac{C_{PE}\sqrt{d}}{\sqrt{m}}. \quad (18)$$

The bound in Theorem 4.2 can be interpreted as an upward shift of the No-PE complexity. $C_{PE}\sqrt{d}/\sqrt{m}$ represents the cost incurred by the PE module. Moreover, when t becomes larger, $C_{PE}\sqrt{d}/\sqrt{m}$ dominates. This implies that with sufficiently long context, the model's generalization ability is no longer limited by the information in the context, but by the intrinsic property raised by the architectural complexity of the PE. The proof can be found in Appendix C.4.

4.3 ARC for Transformer

In this section, we extend our analysis by considering the adversarial robustness of the Transformer in ICL. The primary challenge in this setting is the max operator defined in A.3, which makes direct analysis difficult. To bypass this, we introduce the surrogate loss framework. We reveal that the adversarial attack increases the complexity bound by a factor Φ, which quantifies the degree of the attack. Finally, we also extend this result with a trainable PE module.

The Surrogate Loss Framework. We follow a general method by defining a simpler surrogate loss that upper bounds the true adversarial loss, then analyzing the complexity of this surrogate.

Proposition 4.1 (Surrogate Loss Upper Bound)**.** *Assume the model function $f_\theta(X)$ is L_x-Lipschitz with respect to its input X. Then the true adversarial loss $\tilde{\ell}_\theta$ is upper-bounded by the surrogate loss ℓ_{rob}:*

$$\tilde{\ell}_\theta(X, y) \le \ell_{rob}(f_\theta(X), y), \quad (19)$$

where $\ell_{rob}(f_\theta(X), y) \triangleq (|f_\theta(X) - y| + L_x\varepsilon)^2$.

By the monotonicity of the Rademacher complexity, this proposition immediately implies that $\mathrm{Rad}_{S'}(\tilde{\mathcal{L}}_{\mathcal{F}_{S'_t}}) \leq \mathrm{Rad}_{S'}(\mathcal{L}_{\mathrm{rob}})$, where $\mathcal{L}_{\mathrm{rob}}$ is the class of surrogate loss functions. The problem is now reduced to bounding the complexity of this surrogate class. The surrogate loss has a compositional structure $h(f(x), y)$, where $h(z, y) = (|z - y| + L_x\varepsilon)^2$. This structure allows us to apply Lemma B.4 to relate the complexity of the loss class back to the complexity of the original function class $\mathcal{F}_{S_t}$.

Proposition 4.2 (Complexity Bound via Surrogate)**.** *Let $\tilde{\mathcal{L}}_{\mathcal{F}_{S'_t}}$ be the adversarial loss class, where each element is $\tilde{\ell}_\theta(X, y)$ as defined in Eq.(6). Suppose the model's output is bounded by $|f_\theta(X)| \leq M$. Also, as shown in Lemma C.2, the response is bounded by $|y| \leq M_y$ with high probability. Then there exists a Lipschitz constant L_h such that the Rademacher complexity of the adversarial loss class is bounded by the complexity of the standard function class:*

$$\mathrm{Rad}_{S'}(\tilde{\mathcal{L}}_{\mathcal{F}_{S'_t}}) \leq L_h \cdot \mathrm{Rad}_{S'}(\mathcal{F}_{S'_t}), \tag{20}$$

where $L_h = 2((M + M_y) + L_x\varepsilon)$.

The Impact of Attack on Solution Space. Under adversarial attack, the geometry of the solution space also changes. Following the same paradigm in Section 4.1, as detailed below, we find that the diameter of the perturbed solution space can be expressed as the clean diameter scaled by a penalty term.

Proposition 4.3 (Diameter of the Perturbed Solution Space)**.** *The diameter of the adversarially perturbed solution space $W_{S'_t}$ is bounded by*

$$\mathrm{Diam}(W_{S'_t}) \lesssim O\left(\sqrt{r/t}\right) \cdot \Phi(\varepsilon, t, d), \tag{21}$$

where $\Phi(\varepsilon, t, d) = \frac{1}{1-\sqrt{d/t}-\varepsilon/\sqrt{t}}$. This bound is meaningful only when $\varepsilon < \sqrt{t} - \sqrt{d}$.

Similar to Eq. (15), Φ arises from analyzing the geometry of the attacked solution space $W_{S'_t}$, but with a perturbation according to Lemma B.5. As $\Phi > 1$, the attack will enlarge the diameter of W'_{S_t}, which in turn enlarges the Rademacher complexity. When the attack strength $\varepsilon \to 0$, the factor Φ approaches 1, and the bound in Eq. (21) recovers the clean bound from Eq. (15). Moreover, Φ is a decreasing function of t, which reveals that increasing the context length enhances adversarial robustness by reducing the attack's effect. The full proof for all the Propositions can be found in Appendix C.5.

Then we extend the Lemma 4.3 to the adversarial setting as follows. In order to connect $W_{S'_t}$ and $\Theta_{S'_t}$, we make an assumption for the mapping.

Assumption 4.3 (Robust Inverse Lipschitz Continuity)**.** Let $G_{\mathrm{adv}} : \theta \mapsto w'_{\mathrm{eff}}$, where w'_{eff} is computed based on the perturbed context S'_t. We assume this mapping is also inverse Lipschitz continuous with a constant L'_G:

$$\|\theta_1 - \theta_2\|_2 \leq L'_G \|G_{\mathrm{adv}}(\theta_1) - G_{\mathrm{adv}}(\theta_2)\|_2. \tag{22}$$

Assumption 4.3 extends the stability property of the parameter-to-solution-space mapping G to the adversarial setting. This is extended from Lemma 4.3 (justified in Lemma 4.4). We posit that a small, ε-bounded perturbation should only cause a correspondingly small, bounded degradation in the geometric properties of the mapping G. This ensures the inverse Lipschitz property is preserved, potentially with a slightly larger constant L'_G.

Given the above assumption, we finally present the adversarial RC result for Transformers without PE (Theorem 4.3) and with PE (Theorem 4.4):

Theorem 4.3 (ARC without PE)**.** *Under Assumption 4.3, if $\varepsilon < \sqrt{t} - \sqrt{d}$, its Adversarial Rademacher complexity is bounded by:*

$$\mathrm{Rad}_{S'}(\tilde{\mathcal{L}}_{\mathcal{F}_{S'_t}}) \lesssim O\left(\frac{L_h L_f \sqrt{rD}\sqrt{\log t/r}}{\sqrt{mt}}\right) \cdot \Phi(\varepsilon, t, d), \tag{23}$$

where $\Phi(\varepsilon, t, d)$ is defined in Proposition 4.3.

The proof combines the results from the preceding propositions. First, by Proposition 4.2, we bound the ARC of the loss class by the Rademacher complexity of the function class, scaled by L_h. We then bound the complexity of the function class $\mathcal{F}_{S'_t}$ using the Dudley integral framework as in the standard case, but now applying it to the adversarially perturbed parameter space $\Theta_{S'_t}$. The decaying diameter of this space is a result of Proposition 4.3, which introduces the Φ into the final bound. The full proof and the definition of $\Theta_{S'_t}$ is in Appendix C.6.

Following the same decomposition method used for the standard RC bound, we can derive the ARC for the model with PE.

Theorem 4.4 (ARC with PE)**.** *Let $\tilde{\mathcal{L}}_{\mathcal{F}_{S'_t,PE}}$ be the adversarial loss class for a Transformer with PE. Under Assumption 4.2 and 4.3, when $\varepsilon < \sqrt{t} - \sqrt{d}$, its Adversarial Rademacher complexity is bounded by:*

$$\mathrm{Rad}_{S'}(\tilde{\mathcal{L}}_{\mathcal{F}_{S'_t,PE}}) \lesssim L_h \cdot \left[\left(\frac{C_{PE}\sqrt{d}}{\sqrt{m}} + \frac{C_{PE}\varepsilon}{\sqrt{m}}\right) + \left(O\left(\frac{L_f\sqrt{rD}\sqrt{\log t/r}}{\sqrt{mt}}\right) \cdot \Phi(\varepsilon, t, d)\right)\right]. \tag{24}$$

This bound reveals two distinct sources of adversarial attacks. Compared to Eq. (23), $C_{PE}\varepsilon/\sqrt{m}$ implies that for stronger attacks, perturbing positional information itself may become a more significant threat than disturbing the data content. The proof of Theorem 4.4 can be found in Appendix C.7.

With Theorem 4.1, 4.2, 4.3 and 4.4, we can formalize the impact of PE. Define the complexity gap as $\Delta = \text{Rad(PE)} - \text{Rad(noPE)}$. In the clean setting, this gap is $\Delta_{\text{clean}} \approx C_{PE}\sqrt{d}/\sqrt{m}$. In the adversarial setting, the gap widens due to an additional attack-dependent term: $\Delta_{\text{adv}} \approx C_{PE}\sqrt{d}/\sqrt{m} + C_{PE}\varepsilon/\sqrt{m}$. Since $\Delta_{\text{adv}} = \Delta_{\text{clean}} + O(\varepsilon/\sqrt{m})$, it shows that the complexity cost of PE is magnified under attack, exposing Transformers with PE to be more vulnerable.

5 SIMULATIONS

In this section, we conduct simulations starting from training the transformer. We implemented the experiments based on Garg et al. (2022) with modifications.

5.1 Setup

Our data generation model follows Assumption 4.1 with data dimension $d = 5$. Following the experiment setting and sample size in Trauger and Tewari (2024), we generate a total number of 10,300 tasks, which are split into a training set of 309 tasks and a validation set of 9,991 tasks. For the model architecture, we use a single-layer, single-head Transformer with a read-in layer and an embedding dimension of $d_m = 1024$.

The training is based on Garg et al. (2022) but with a key modification to enable the measurement of a stable empirical risk. Instead of sampling new prompts at each training step, we create a **fixed training set**: for each of the 309 tasks in our training split, we pre-sample a single prompt of a fixed length t. The model is then trained for 3000 epochs to a stable training loss on this static dataset on a single NVIDIA A6000 GPU. Moreover, we set the batch size to the full training set size (309). This choice eliminates the stochasticity of mini-batch sampling, allowing for a cleaner measurement of the empirical risk and generalization gap. We use the Adam optimizer with a learning rate of 1×10^{-5}. The loss is the Mean Squared Error (MSE) computed only on the final query token of the prompt, $(\hat{y}_q - y_q)^2$. For the positional encoding, we scale the standard PE initialization by a factor of 25. This factor was chosen empirically to bring the initial norm of the PE into the same order of magnitude as the content embeddings, as in Eq.(2).

For the evaluation, we sample a large number of t-length prompts from the unseen tasks and evaluate the loss on the final query token. The generalization gap is defined as the difference between the true risk on the validation set and the empirical risk on the fixed training set.

To evaluate adversarial generalization, we use the Projected Gradient Descent (PGD) attack (Madry et al., 2017). We maximize the MSE loss on the final query prediction by adding a perturbation to the prompt as in Eq. (5). We constrain the perturbation using the L_2 norm, with a budget of ε. For our experiments, we use an attack configuration of $k = 40$ iterations with a step size of $\alpha = \varepsilon/10$.

In the following experiments, each configuration is run 6 times with different random seeds, and we report the mean generalization gap. In alignment with our theory (Lemma 4.3), we focus on context lengths $t > d = 5$.

5.2 Standard Generalization Bounds

In the experiment, we first validate our theoretical claim that incorporating a trainable PE module increases the model's complexity and thus its generalization gap. Figure 1 compares the generalization gap of the model with and without the PE module. Similar to Theorem 4.1 and 4.2, in Figure 1, the model with a PE module consistently exhibits a larger generalization gap across all context lengths t. Furthermore, both curves demonstrate the t decay trend in our analysis, showing that the model successfully utilizes more examples to improve the generalization bound.

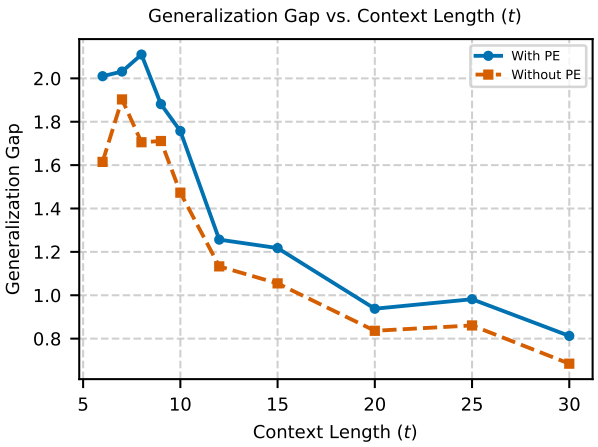

Figure 1: Mean Generalization Gap vs. Context Length (t) for models with and without Positional Encodings (PE).

In addition, we observe a transition phase for small context lengths (e.g., $t \leq 10$), where the gap has a large variation. When the number of examples t is

close to the dimension of the data d, the problem is less constrained, leading to a greater variance in the learned solution and thus a more unstable generalization gap. We noticed that this is also evidence of the "double descent" in modern machine learning (Belkin et al., 2019; Nakkiran et al., 2021). To be specific, in our framework, the Transformer implicitly learns an effective linear model of dimension d as in Eq. (12), which acts as the model complexity (p) in classic double descent. Then the error peaks at the interpolation threshold, where $p \approx n$, which is directly analogous to $d \approx t$ in our setting. Finally, as t grows much larger than d, the system becomes over-determined and the gap decay becomes stable.

5.3 Adversarial Generalization Bounds

PE vs No-PE under Attack. We now extend the comparison to the adversarial setting. Figure 2 shows the attacked generalization gap (for a fixed attack strength $\varepsilon = 0.02$) for models with and without PE.

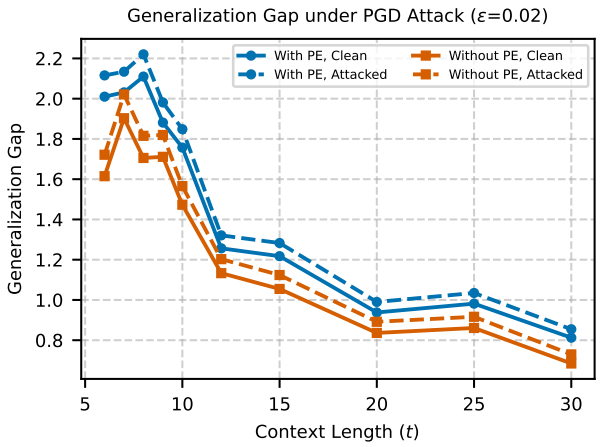

Figure 2: Mean Attacked Generalization Gap vs. Context Length (t) under a fixed PGD attack ($\varepsilon = 0.02$).

This result is also consistent with our theoretical analysis as Theorem 4.3. Our adversarial generalization bound is characterized by the term $\Phi(\varepsilon, t, d)$, which is an increasing function of the attack strength ε under $\varepsilon \leq \sqrt{t} - \sqrt{d}$. For any $\varepsilon \geq 0$, the term $\Phi(\varepsilon, t, d)$ becomes greater than 1, thus amplifying the overall complexity bound. This directly explains why the generalization gap is systematically larger under adversarial attack compared to the clean setting. Moreover, following Theorem 4.4, the inclusion of a trainable PE module explicitly increases the generalization gap, which is the same as the standard setting.

Different Attack Strengths. Finally, we provide empirical evidence for our main theoretical contribution on ARC. Figure 3 plots the attacked generalization gap versus context length t, with each curve representing a different attack strength ε.

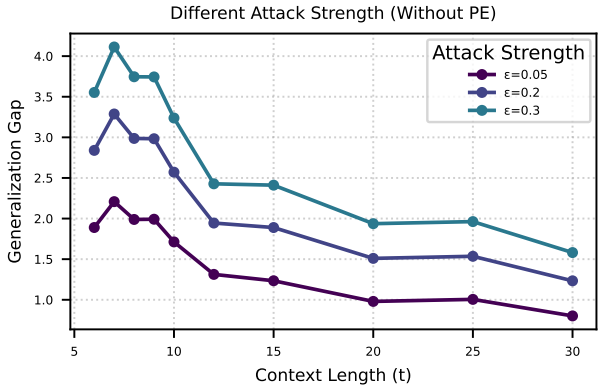

Figure 3: Mean Attacked Generalization Gap vs. Context Length (t) for different attack strengths (ε).

For every attack strength ε (e.g., 0.05, 0.2, 0.3), the generalization gap decreases as the context length increases. This demonstrates that the ICL mechanism remains effective in utilizing longer contexts to enhance generalization ability, even under adversarial attacks. Moreover, as ε increases, the generalization gap also increases, which is aligned with Theorem 4.3.

5.4 Comparison with RoPE

While our primary focus is on the impact of learnable parameters in PE, we also implement fixed encodings such as Rotary Position Encodings (RoPE) for comparison. Our theory suggests that since RoPE lacks learnable positional parameters, its complexity should be similar as the No-PE case. As detailed in Appendix E, our experiments confirm this: RoPE shows a generalization gap comparable to the "No-PE" baseline, which is consistently lower than that of Trainable PE. This further validates that the increased generalization gap discussed in Section 5.2 is indeed a consequence of the parameter complexity introduced by trainable PEs.

6 CONCLUSION

In this work, we provide a theoretical analysis of the impact of trainable Positional Encodings (PE) on the generalization and adversarial robustness of a one-layer Transformer performing ICL. Our analysis demonstrates that the increased model complexity from a trainable PE leads to a wider generalization gap. Furthermore, by deriving the ARC bounds for this setting, we quantified how this vulnerability is

amplified under adversarial attack. Our theoretical findings were validated by empirical evidence.

There are several future directions. First, we only consider the simplest linear regression task. One direction is to extend the problem into other, more complicated tasks and even into real-world applications. Second, a natural extension is to consider understanding how positional encodings scale in multi-layer or deep Transformer architectures. Finally, with the complexity terms such as L_f and D identified in our bounds, one may develop new regularization techniques that directly penalize these terms to design more robust models.

Acknowledgments

Weiyi He is supported by Open Philanthropy. Yue Xing is supported by NSF DMS 2515194, Open Philanthropy, NVIDIA Academic Grant Program and Google Cloud Research Credit.

References

Ahn, K., Cheng, X., Daneshmand, H., and Sra, S. (2023). Transformers learn to implement preconditioned gradient descent for in-context learning. *Advances in Neural Information Processing Systems*, 36:45614–45650.

Attias, I., Kontorovich, A., and Mansour, Y. (2022). Improved generalization bounds for adversarially robust learning. *Journal of Machine Learning Research*, 23(175):1–31.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Çavşi Zaim, H. and Yolacan, E. N. (2025). Fpe–transformer: A feature positional encoding-based transformer model for attack detection. *Applied Sciences*, 15(3):1252.

Chen, M., Du, J., Pasunuru, R., Mihaylov, T., Iyer, S., Stoyanov, V., and Kozareva, Z. (2022). Improving in-context few-shot learning via self-supervised training. *arXiv preprint arXiv:2205.01703*.

Cui, Y., Ren, J., He, P., Tang, J., and Xing, Y. (2024). Superiority of multi-head attention in in-context linear regression. *arXiv preprint arXiv:2401.17426*.

Deng, Y., Li, Z., and Song, Z. (2023). Attention scheme inspired softmax regression. *arXiv preprint arXiv:2304.10411*.

Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., Xia, H., Xu, J., Wu, Z., Liu, T., et al. (2022). A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Edelman, B. L., Goel, S., Kakade, S., and Zhang, C. (2022). Inductive biases and variable creation in self-attention mechanisms. In *International Conference on Machine Learning*, pages 5793–5831. PMLR.

Gao, Q. and Wang, X. (2021). Theoretical investigation of generalization bounds for adversarial learning of deep neural networks. *Journal of Statistical Theory and Practice*, 15(2):51.

Gao, S., Zhou, H., Chen, T., He, M., Xu, R., and Li, J. (2024). Pe-attack: On the universal positional embedding vulnerability in transformer-based models. *IEEE Transactions on Information Forensics and Security*, 19:9359–9373.

Garg, S., Tsipras, D., Liang, P. S., and Valiant, G. (2022). What can transformers learn in-context? a case study of simple function classes. *Advances in neural information processing systems*, 35:30583–30598.

He, P., Xu, H., Xing, Y., Liu, H., Yamada, M., and Tang, J. (2025). Data poisoning for in-context learning. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1680–1700.

Khim, J. and Loh, P.-L. (2018). Adversarial risk bounds via function transformation. *arXiv preprint arXiv:1810.09519*.

Li, Y. (2025). Theoretical analysis of positional encodings in transformer models: Impact on expressiveness and generalization. *arXiv preprint arXiv:2506.06398*.

Li, Y., Ildiz, M. E., Papailiopoulos, D., and Oymak, S. (2023). Transformers as algorithms: Generalization and stability in in-context learning. In *International conference on machine learning*, pages 19565–19594. PMLR.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., et al. (2022). Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Liu, Y., Deng, G., Li, Y., Wang, K., Wang, Z., Wang, X., Zhang, T., Liu, Y., Wang, H., Zheng, Y., et al. (2023). Prompt injection attack against llm-integrated applications. *arXiv preprint arXiv:2306.05499*.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2017). Towards deep learning mod-

els resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Min, S., Lewis, M., Zettlemoyer, L., and Hajishirzi, H. (2021). Metaicl: Learning to learn in context. *arXiv preprint arXiv:2110.15943*.

Montasser, O., Hanneke, S., and Srebro, N. (2019). Vc classes are adversarially robustly learnable, but only improperly. In *Conference on Learning Theory*, pages 2512–2530. PMLR.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. (2021). Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Raventós, A., Paul, M., Chen, F., and Ganguli, S. (2023). Pretraining task diversity and the emergence of non-bayesian in-context learning for regression. *Advances in neural information processing systems*, 36:14228–14246.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Shayegani, E., Mamun, M. A. A., Fu, Y., Zaree, P., Dong, Y., and Abu-Ghazaleh, N. (2023). Survey of vulnerabilities in large language models revealed by adversarial attacks. *arXiv preprint arXiv:2310.10844*.

Sinha, A., Namkoong, H., and Duchi, J. C. (2017). Certifiable distributional robustness with principled adversarial training. corr, abs/1710.10571. *arXiv preprint arXiv:1710.10571*.

Trauger, J. and Tewari, A. (2024). Sequence length independent norm-based generalization bounds for transformers. In *International Conference on Artificial Intelligence and Statistics*, pages 1405–1413. PMLR.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

Von Oswald, J., Niklasson, E., Randazzo, E., Sacramento, J., Mordvintsev, A., Zhmoginov, A., and Vladymyrov, M. (2023). Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pages 35151–35174. PMLR.

Wei, A., Haghtalab, N., and Steinhardt, J. (2023). Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36:80079–80110.

Wei, C., Chen, Y., and Ma, T. (2022). Statistically meaningful approximation: a case study on approximating turing machines with transformers. *Advances in Neural Information Processing Systems*, 35:12071–12083.

Xiao, J., Fan, Y., Sun, R., and Luo, Z.-Q. (2022a). Adversarial rademacher complexity of deep neural networks. *arXiv preprint arXiv:2211.14966*.

Xiao, J., Fan, Y., Sun, R., Wang, J., and Luo, Z.-Q. (2022b). Stability analysis and generalization bounds of adversarial training. *Advances in Neural Information Processing Systems*, 35:15446–15459.

Xie, S. M., Raghunathan, A., Liang, P., and Ma, T. (2021). An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*.

Xing, Y., Lin, X., Xu, C., Suh, N., Song, Q., and Cheng, G. (2024). Theoretical understanding of in-context learning in shallow transformers with unstructured data. *arXiv preprint arXiv:2402.00743*.

Xing, Y., Song, Q., and Cheng, G. (2021). On the algorithmic stability of adversarial training. *Advances in neural information processing systems*, 34:26523–26535.

Yin, D., Kannan, R., and Bartlett, P. (2019). Rademacher complexity for adversarially robust generalization. In *International conference on machine learning*, pages 7085–7094. PMLR.

Zhang, R., Frei, S., and Bartlett, P. L. (2024). Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55.

Zhang, Y., Feng, S., and Tan, C. (2022). Active example selection for in-context learning. *arXiv preprint arXiv:2211.04486*.

Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J. Z., and Fredrikson, M. (2023). Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]

(b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]

(c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:

(a) Statements of the full set of assumptions of all theoretical results. [Yes]

(b) Complete proofs of all theoretical results. [Yes]

(c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:

(a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]

(b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

(a) Citations of the creator If your work uses existing assets. [Not Applicable]

(b) The license information of the assets, if applicable. [Not Applicable]

(c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

(d) Information about consent from data providers/curators. [Not Applicable]

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. [Not Applicable]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendix

Contents

A Definitions

Definition A.1 (Covering Numbers)**.** Let $(\mathcal{F}, d)$ be a metric space. An ε-cover of a subset $A \subseteq \mathcal{F}$ is a set $\{c_1, \ldots, c_N\}$ such that for every $f \in A$, there exists some c_i with $d(f, c_i) \leq \varepsilon$. The covering number, $N(A, d, \varepsilon)$, is the size of the smallest possible ε-cover.

Definition A.2 (Rademacher Complexity (RC))**.** Let $\mathcal{F}$ be a class of real-valued functions and $S = \{x_1, \ldots, x_m\}$ be a fixed sample set. The empirical Rademacher complexity of $\mathcal{F}$ on S is:

$$\mathrm{Rad}_S(\mathcal{F}) \triangleq \frac{1}{m}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sigma_i f(x_i)\right],$$

where σ_i are i.i.d. and take values ± 1 each with half probability and $\sigma = (\sigma_1, \ldots, \sigma_n)$.

Definition A.3 (Adversarial Rademacher Complexity (ARC))**.** Let $\mathcal{F}$ be a function class and ℓ be a loss function. The adversarial loss is the worst-case loss under a perturbation of budget ε:

$$\tilde{\ell}_f(x, y) \triangleq \sup_{\|x'-x\|\leq\varepsilon} \ell(f(x'), y).$$

The Adversarial Rademacher Complexity (ARC) is the Rademacher complexity of the corresponding adversarial loss class $\tilde{\mathcal{L}}_\mathcal{F} = \{(x, y) \mapsto \tilde{\ell}_f(x, y) \mid f \in \mathcal{F}\}$:

$$\mathrm{ARC}_S(\mathcal{F}) \triangleq \mathrm{Rad}_S(\tilde{\mathcal{L}}_\mathcal{F}) = \frac{1}{m}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sigma_i\tilde{\ell}_f(x_i, y_i)\right].$$

Definition A.4. Let $p, q, r, d \in \mathbb{N}$ and let $W \in \mathbb{R}^{r\times d}$. We define three standard matrix norms used throughout our analysis.

The Operator Norm (Induced Norm): The $L_q \to L_p$ operator norm, denoted $\|W\|_{q\to p}$, measures the maximum amplification factor that the matrix W applies to any vector x. It is defined as the supremum:

$$\|W\|_{q\to p} \triangleq \sup_{x\neq 0}\frac{\|Wx\|_p}{\|x\|_q}.$$

A special case of this is the spectral norm, written as $\|W\|_2$, which corresponds to the $2 \to 2$ norm and is equivalent to the largest singular value of W.

The Mixed (q, p) Norm: The $\|W\|_{q,p}$ norm is calculated in a two-step process. First, the ℓ_q-norm of each of the d columns of W is computed. Then, the ℓ_p-norm is taken on the resulting d-dimensional vector of column norms:

$$\|W\|_{q,p} \triangleq \left\|[\|W_{:,1}\|_q, \ldots, \|W_{:,d}\|_q]^\top\right\|_p.$$

The Frobenius Norm: The Frobenius norm, $\|W\|_F$, is the square root of the sum of all squared entries in the matrix, analogous to the Euclidean norm for vectors:

$$\|W\|_F \triangleq \sqrt{\sum_{i=1}^{r}\sum_{j=1}^{d}W_{ij}^2}.$$

This norm is also equivalent to the Euclidean norm of the matrix's singular values. Therefore, we have $\|W\|_2 \leq \|W\|_F$ for any matrix W .

B Useful Lemmas

In this section, we list some useful lemmas, which are classic results and can be found in Vershynin (2018).

Lemma B.1 (Diameter of an Ellipsoid). *For the ellipsoid $\{w : (w-w_0)^\top A(w-w_0) \le r\}$ where A is symmetric positive definite, the diameter under the Euclidean norm is*

$$2\sqrt{r/\lambda_{\min}(A)},$$

where $\lambda_{\min}(A)$ is the smallest eigenvalue of A.

Lemma B.2 (Lower Bound for Minimum Singular Value). *Let A be an $n \times d$ random matrix with entries i.i.d. $\mathcal{N}(0,1)$ and $n > d$. Then for every $k > 0$,*

$$\mathbb{P}\left(\sigma_{\min}(A) \le \sqrt{n} - \sqrt{d} - k\right) \le e^{-k^2/2}.$$

Lemma B.3 (Covering Number for Balls). *Let $\Theta \subset V = \mathbb{R}^d$. Then*

$$\left(\frac{1}{\varepsilon}\right)^d \frac{\mathrm{vol}(\Theta)}{\mathrm{vol}(B)} \le N(\Theta, ||\cdot||_2, \varepsilon) \le \left(\frac{3}{\varepsilon}\right)^d \frac{\mathrm{vol}(\Theta)}{\mathrm{vol}(B)},$$

where B is the unit norm ball.

Lemma B.4 (Talagrand's Contraction Lemma (General Vector Form)). *Let $\mathcal{A} \subseteq \mathbb{R}^m$ be a set of vectors. For each $i = 1, \dots, m$, let $\phi_i : \mathbb{R} \to \mathbb{R}$ be a ρ-Lipschitz function. Define the mapping $\Phi : \mathbb{R}^m \to \mathbb{R}^m$ which applies each function coordinate-wise:*

$$\Phi(a) = (\phi_1(a_1), \phi_2(a_2), \dots, \phi_m(a_m)) \quad \text{for any} \quad a = (a_1, \dots, a_m) \in \mathcal{A}.$$

Let $\Phi \circ \mathcal{A} \triangleq \{\Phi(a) : a \in \mathcal{A}\}$ be the set of transformed vectors. Then the empirical Rademacher complexity of the transformed set is bounded by the complexity of the original set:

$$\mathrm{Rad}_m(\Phi \circ \mathcal{A}) \le \rho \cdot \mathrm{Rad}_m(\mathcal{A}).$$

Remark B.1. In many machine learning applications, all coordinate functions are the same, i.e., $\phi_1 = \phi_2 = \cdots = \phi_m = \phi$. The lemma still holds with ρ being the Lipschitz constant of the single function ϕ.

Lemma B.5 (Weyl Inequality). *For any matrix A and B, we have the following Weyl inequality:*

$$|\sigma_i(A+B) - \sigma_i(A)| \le \|B\|_2,$$

where σ_i is the singular value. A direct corollary is that

$$\sigma_{min}(A+B) \ge \sigma_{min}(A) - ||B||_2,$$

where $||B||_2$ is the spectral norm and $\sigma_{\min}$ is the smallest singular value.

C Proofs

C.1 Justification for Definition 4.1

Proof. Under the condition of Theorem 4.1, we have

$$\|W_{QK} - W^*_{QK}(S_t)\|_F \le \|W_{QK} - W^*_{QK}\|_F + \|W^*_{QK} - W^*_{QK}(S_t)\|_F \le \gamma + \Delta_t.$$

Then we prove $\Delta_t \lesssim O(\sqrt{d/t})$ with high probability.

Let $G_t := \frac{1}{t}\sum_{i=1}^t x_i x_i^\top$ with $x_i \overset{i.i.d.}{\sim} \mathcal{N}(0, I_d)$, and let

$$W^*_{QK}(S_t) = \Psi(G_t), \qquad W^*_{QK} = \Psi(I_d).$$

We only require Ψ to be locally Lipschitz in the spectral norm near I_d.

Assumption C.1. There exists $L_\Psi > 0$ and a neighborhood $\mathcal{U}$ of I_d such that for all $G, G' \in \mathcal{U}$,

$$\left\|\Psi(G) - \Psi(G')\right\|_F \le L_\Psi \left\|G - G'\right\|_2.$$

Assumption C.1 is mild and holds for a broad class of smooth matrix functionals. It can be justified via Fréchet differentiability of Ψ and a uniform bound on $\|\nabla\Psi\|$ in a neighborhood of I_d.

Proposition C.1. *Under Assumption C.1, for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\Delta_t := \left\|W^*_{QK}(S_t) - W^*_{QK}\right\|_F \le L_\Psi \left\|G_t - I_d\right\|_2 \le C\, L_\Psi \left(\sqrt{\frac{d}{t}} + \sqrt{\frac{\log(1/\delta)}{t}}\right),$$

where $C > 0$ is a constant. In particular, $\Delta_t = O(\sqrt{d/t})$.

Proof. The first inequality follows immediately from Assumption C.1 by taking $G = G_t$ and $G' = I_d$:

$$\Delta_t = \left\|\Psi(G_t) - \Psi(I_d)\right\|_F \le L_\Psi \left\|G_t - I_d\right\|_2.$$

For the second inequality, note that $tG_t \sim \text{Wishart}(d, t)$ with mean I_d. Standard spectral concentration for sample covariance matrices yields, for some absolute constant $C > 0$ and any $\delta \in (0,1)$,

$$\left\|G_t - I_d\right\|_2 \le C\left(\sqrt{\frac{d}{t}} + \sqrt{\frac{\log(1/\delta)}{t}}\right),$$

with probability at least $1-\delta$. Combining the two inequalities to get the results. □

Proposition C.1 shows that the only term depending on the context length t is factor $\sqrt{d/t}$ (up to a standard $\sqrt{\log(1/\delta)/t}$). All other quantities (L_Ψ, C) are constants and do not affect the decay profile in t. Hence, if we fix any $t_{\min}$ and work on the high-probability event, for all $t \ge t_{\min}$, we may set

$$\Delta_* := \sup_{t \ge t_{\min}} \Delta_t = O\Big(L_\Psi \sqrt{d/t_{\min}}\Big),$$

and absorb Δ_* into $\gamma_{\text{eff}} = \gamma + \Delta_*$. This ensures that all subsequent constants remain t-independent. □

C.2 Proof of Lemmas in Section 4.1

In this section, we first prove the lemmas used in Section 4.1, which are Lemma 4.1, Lemma 4.2, Lemma 4.3 and Lemma 4.4.

Proof of Lemma 4.1

Proof. This bound follows from the classical Dudley entropy integral (Vershynin, 2018). □

Proof of Lemma 4.2

Proof. We prove Lemma 4.2 in three steps: (i) a peeling reduction by contraction, (ii) a Lipschitz link from parameters to the attention output, and (iii) covering numbers for Euclidean balls and the translation from parameters to functions.

For the single-layer, single-head Transformer defined in Section 3.1, we analyze the parameter pair $\theta = (W_{\text{in}}, W_{QK}) \in \Theta_{S_t}$ defined in Definition 4.1:

$$\Theta_{S_t} = \Big\{(W_{\text{in}}, W_{QK}) \ : \ \|W_{QK} - W^*_{QK}(S_t)\|_F \le \gamma\Big\}.$$

Other components (W_V, W_c) are trained but uniformly bounded as in Section 3.1: $\|W_c^\top\|_{1,\infty} \le B_{W_c}$ and $\|W_V^\top\|_{1,\infty} \le B_{W_V}$, and σ is L_σ-Lipschitz.

(i) Peeling via contraction. Apply Lemma B.4 layer by layer: the composition of a bounded linear readout (W_c), a Lipschitz activation (σ), and a bounded linear map (W_V) gives a global multiplicative constant $B_{W_c} L_\sigma B_{W_V}$ relating the Rademacher complexity of the full function class and that of the attention block. Formally,

$$\text{Rad}_S(\mathcal{F}_{S_t}) \le B_{W_c}\, L_\sigma\, B_{W_V}\ \ \text{Rad}_S(\mathcal{U}_{S_t}),$$

where

$$\mathcal{U}_{S_t} = \{X \mapsto h_{\text{attn}}(X; W_{\text{in}}, W_{QK}) : (W_{\text{in}}, W_{QK}) \in \Theta_{S_t}\}, \qquad h_{\text{attn}} = \text{RowSoftmax}\left(\frac{HW_{QK}H^\top}{\sqrt{d_m}}\right) H$$

with $H = XW_{\text{in}}$.

(ii) Lipschitz link for the attention block. We show that the map $(W_{\text{in}}, W_{QK}) \mapsto h_{\text{attn}}$ is Lipschitz. Let $\theta_1 = (W_{\text{in},1}, W_{QK,1})$ and $\theta_2 = (W_{\text{in},2}, W_{QK,2})$ in Θ_{S_t}. Set $H_i = XW_{\text{in},i}$ and

$$M_i = \frac{1}{\sqrt{d_m}} H_i W_{QK,i} H_i^\top, \qquad A_i = \text{RowSoftmax}(M_i), \qquad i = 1, 2.$$

In the empirical norm, we have

$$\|h_{\text{attn}}(\cdot;\theta_1) - h_{\text{attn}}(\cdot;\theta_2)\|_{m,2} \ \le\ L_{\text{attn}}\ \|\theta_1 - \theta_2\|_2,$$

with L_{attn} depending on uniform bounds of $\|H\|_2$ and $\|W_{QK}\|_2$ over Θ_{S_t}. (These are controlled on a high-probability well-conditioned event $\mathcal{G}_t$ under the Gaussian design; see Proof of Lemma 4.3.)

Combining with the peeling constant from step (i), we obtain the Lipschitz link between functions and parameters claimed in the main text:

$$\|f_{\theta_1} - f_{\theta_2}\|_{m,2} \ \le\ B_{W_c}\, L_\sigma\, B_{W_V}\, L_{\text{attn}}\ \|\theta_1 - \theta_2\|_2 \ \equiv\ L_f\ \|\theta_1 - \theta_2\|_2. \tag{C.1}$$

(iii) Covering numbers and translation. Let $D = d \cdot d_m + d_m^2$ be the ambient dimension of $\theta = (W_{\text{in}}, W_{QK})$, and $D_t = \text{Diam}(\Theta_{S_t})$. Standard Euclidean ball covering from Lemma B.3: gives

$$\log N(\Theta_{S_t}, \|\cdot\|_2, \varepsilon) \le D \log\left(\frac{3D_t}{\varepsilon}\right).$$

Now (C.1) implies the covering translation: if $\{\theta_k\}$ is an (ε/L_f)-cover of Θ_{S_t} in $\|\cdot\|_2$, then $\{f_{\theta_k}\}$ is an ε-cover of $\mathcal{F}_{S_t}$ in $\|\cdot\|_{m,2}$, hence

$$N\big(\mathcal{F}_{S_t}, \|\cdot\|_{m,2}, \varepsilon\big) \ \le\ N\big(\Theta_{S_t}, \|\cdot\|_2, \varepsilon/L_f\big).$$

Similarly, taking the supremum over pairs gives $\text{Diam}(\mathcal{F}_{S_t}) \le L_f\ \text{Diam}(\Theta_{S_t})$, where $L_f = B_{W_c}\, L_\sigma\, B_{W_V}\, L_{\text{attn}}$.

□

Proof of Lemma 4.3

Proof. First, we justify the high probability "good event" $\mathcal{G}_t$ as mentioned in Lemma 4.3. We define the tail event $\mathcal{T}_t$ as

$$\mathcal{T}_t := \left\{ S_t : \sigma_{\min}(X_t) \le c_1\sqrt{t} - c_2\sqrt{d} \right\}.$$

Standard results from random matrix theory (Vershynin, 2018) guarantee that for i.i.d. Gaussian data, the probability of this tail event decays exponentially with t:

$$\mathbb{P}(\mathcal{T}_t) \le 2d\exp(-c_3 t).$$

Our analysis is therefore conditioned on the complementary $\mathcal{G}_t = \mathcal{T}_t^c$.

For any $\theta_1, \theta_2 \in \Theta_{S_t}$, let $w_1 = G(\theta_1)$ and $w_2 = G(\theta_2)$. By definition, $\|w_1 - w_2\|_2 \le \mathrm{Diam}(W_{S_t})$. Applying the Eq. (13):

$$\|\theta_1 - \theta_2\|_2 \le L_G\|w_1 - w_2\|_2 \le L_G \cdot \mathrm{Diam}(W_{S_t}).$$

Taking the supremum over all pairs (θ_1, θ_2) yields $\mathrm{Diam}(\Theta_{S_t}) \le L_G \cdot \mathrm{Diam}(W_{S_t})$.

Then applying Lemma B.2 to X_t with $n = t$, we have for any $\delta > 0$:

$$\mathbb{P}\left(\sigma_{\min}(X_t) \le \sqrt{t} - \sqrt{d} - \sqrt{2\log(1/\delta)}\right) \le \delta.$$

Then there exists $K_1 > 0$ (depending on d and δ) such that with probability at least $1 - \delta$:

$$\sigma_{\min}(X_t) \ge K_1\sqrt{t}.$$

By Lemma B.1, the diameter satisfies:

$$\mathrm{Diam}(W_{S_t}) = \frac{2\sqrt{r}}{\sigma_{\min}(X_t)} \le \frac{2\sqrt{r}}{K_1\sqrt{t}}.$$

Thus, with high probability:

$$\mathrm{Diam}(W_{S_t}) \lesssim O(\sqrt{r/t}).$$

Therefore, we now have the result that $\mathrm{Diam}(\Theta_{S_t}) \lesssim L_G\,\mathrm{Diam}(W_{S_t}) \lesssim O\left(L_G\sqrt{r/t}\right)$. □

Proof of Lemma 4.4

Proof. Denote by $\hat{w}$ the OLS solution on (X_t, y) as in Lemma 4.3, and by

$$\widehat{y}_t^{\mathrm{lin}}(\theta) := X_t\, w_{\mathrm{eff}}(\theta) \ \in \mathbb{R}^t$$

the effective linear prediction vector induced by the Transformer parameters θ on the same inputs $\{x_i\}_{i=1}^t$. Further, define the model prediction vector on the context by

$$g_t(\theta) := \big(g(\theta; S_t, x_1), \ldots, g(\theta; S_t, x_t)\big)^\top,$$

where $g(\theta; S_t, x)$ is the scalar prediction produced by the Transformer when queried at x under the prompt S_t. Introduce the fitting error $e_f := y - g_t(\theta)$ and the nonlinearity residual $\eta := g_t(\theta) - \widehat{y}_t^{\mathrm{lin}}(\theta)$ so that $y - \widehat{y}_t^{\mathrm{lin}}(\theta) = e_f + \eta$.

Since $(w - \hat{w})^\top(X_t^\top X_t)(w - \hat{w}) \le r$, we will first bound:

$$r^* := \left\|X_t\big(w_{\mathrm{eff}}(\theta) - \hat{w}\big)\right\|_2^2.$$

By the OLS Pythagorean identity (orthogonal projection onto $\mathrm{span}(X_t)$), for any $w \in \mathbb{R}^d$ one has

$$\|y - X_t w\|_2^2 = \|y - X_t\hat{w}\|_2^2 \ + \|X_t(\hat{w} - w)\|_2^2.$$

Apply this with $w = w_{\text{eff}}(\theta)$ and rearrange to get

$$\|X_t\big(w_{\text{eff}}(\theta) - \hat{w}\big)\|_2^2 = \|y - X_t w_{\text{eff}}(\theta)\|_2^2 \ - \|y - X_t\hat{w}\|_2^2.$$

Since the OLS residual norm is nonnegative, $\|y - X_t\hat{w}\|_2^2 \geq 0$, we deduce

$$r^* = \|X_t\big(w_{\text{eff}}(\theta) - \hat{w}\big)\|_2^2 \leq \|y - X_t w_{\text{eff}}(\theta)\|_2^2.$$

By the definition of $g_t(\theta)$ and $\widehat{y}_t^{\text{lin}}(\theta)$, we have the exact decomposition

$$y - X_t w_{\text{eff}}(\theta) = \big(y - g_t(\theta)\big) + \big(g_t(\theta) - X_t w_{\text{eff}}(\theta)\big) = e_f + \eta.$$

This yields

$$r^* \leq \|e_f + \eta\|_2^2 \leq 2\|e_f\|_2^2 + 2\|\eta\|_2^2.$$

As shown in Proof C.4, $\|\eta\|_2 \lesssim O(\gamma_{\text{eff}})$. In particular, since the empirical fitting error is small, we can bound it as $\|e_f\|_2^2 \leq C_{\text{fit}}$, then $r^* \leq \ 2C_{\text{fit}} + 2O(\gamma_{\text{eff}}) := C_r$, a constant independent of t. Specifically, let $r = C_r$ to get the final result.

Fix $\theta_0 \in K$ and set $y_0 := G(\theta_0)$. Since $J_G(\theta_0)$ has full row rank d, by the Constant Rank Theorem there exist neighborhoods $V \subseteq \Theta_{S_t}$ of θ_0 and $U \subseteq \mathbb{R}^d$ of y_0, and C^1 diffeomorphisms

$$\phi : V \to \phi(V) \subseteq \mathbb{R}^D, \qquad \psi : U \to \psi(U) \subseteq \mathbb{R}^d,$$

such that in local coordinates the map G takes the standard projection form

$$(\psi \circ G \circ \phi^{-1})(x_1, \ldots, x_D) = (x_1, \ldots, x_d).$$

Let $\iota : \mathbb{R}^d \to \mathbb{R}^D$ be the standard embedding $\iota(z_1, \ldots, z_d) = (z_1, \ldots, z_d, 0, \ldots, 0)$. Define a local right inverse $s : U \to V$ by

$$s(y) = \phi^{-1}\big(\iota(\psi(y))\big).$$

Then $G(s(y)) = y$ for all $y \in U$.

By the chain rule,

$$J_s(y) = J_{\phi^{-1}}\big(\iota(\psi(y))\big) \cdot J_\iota\big(\psi(y)\big) \cdot J_\psi(y).$$

We first note $\|J_\iota\|_2 = 1$. We claim that there exists a constant $M = M(c_G, c_U)$ such that on V and U,

$$\|J_\phi(\theta)\|_2, \ \|J_\phi(\theta)^{-1}\|_2, \ \|J_\psi(y)\|_2, \ \|J_\psi(y)^{-1}\|_2 \leq M.$$

Indeed, differentiating the normal form $\psi \circ G \circ \phi^{-1} = (x_1, \ldots, x_d)$ yields

$$J_\psi\big(G(\theta)\big) J_G(\theta) J_{\phi^{-1}}(\phi(\theta)) = [\, I_d \ \ 0\,],$$

hence

$$J_G(\theta) = J_\psi\big(G(\theta)\big)^{-1} [\, I_d \ \ 0\,] J_\phi(\theta).$$

Taking operator norms and using $\|[I_d \ 0]\|_2 = 1$, together with the uniform bounds $\|J_G(\theta)\|_2 \leq c_U$ and $\|J_G(\theta)^\dagger\|_2 \leq 1/c_G$ (since $\sigma_{\min}(J_G(\theta)) \geq c_G$), standard perturbation arguments for products imply that $J_\phi, J_\phi^{-1}, J_\psi, J_\psi^{-1}$ are uniformly bounded in operator norm on sufficiently small neighborhoods V, U, with bounds depending only on (c_G, c_U). Thus the claimed $M(c_G, c_U)$ exists.

Therefore,

$$\|J_s(y)\|_2 \leq \big\|J_{\phi^{-1}}\big(\iota(\psi(y))\big)\big\|_2 \cdot \|J_\iota\|_2 \cdot \|J_\psi(y)\|_2 \leq M \cdot 1 \cdot M = M^2.$$

Since s is C^1 on U and $\sup_{y \in U} \|J_s(y)\|_2 \leq M^2$, the mean value theorem implies s is Lipschitz on U with $\text{Lip}(s; U) \leq M^2$. Set $L_G := M^2$ to complete the proof. □

C.3 Proof of Theorem 4.1

Proof. Before we prove the final Theorem, we first state a technical lemma:

Lemma C.1 (Integral Bound). *For $D_t \lesssim O(\sqrt{r/t})$, the core integral from Dudley's bound satisfies:*

$$\int_0^{D_t} \sqrt{\log\left(\frac{KD_t}{\varepsilon}\right)}\, d\varepsilon \lesssim D_t\sqrt{\log(1/D_t)} \lesssim O\left(\frac{\sqrt{r\log t/r}}{\sqrt{t}}\right),$$

where K is a constant.

Proof. We use the standard inequality $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for $a, b \ge 0$.

$$\begin{aligned}\int_0^{D_t} \sqrt{\log(K) + \log(D_t/\varepsilon)}\, d\varepsilon &\le \int_0^{D_t} \left(\sqrt{\log(K)} + \sqrt{\log(D_t/\varepsilon)}\right) d\varepsilon \\ &= D_t\sqrt{\log(K)} + \int_0^{D_t} \sqrt{\log(D_t/\varepsilon)}\, d\varepsilon.\end{aligned}$$

The second term is a classic integral. By substituting $u = \varepsilon/D_t$, it becomes $D_t\int_0^1 \sqrt{\log(1/u)}\, du = D_t \cdot C$, where $C = \sqrt{\pi}/2$. A widely used, slightly looser bound that retains the dependency on D_t is $O(D_t\sqrt{\log(1/D_t)})$. As this term dominates $D_t\sqrt{\log(K)}$ for small D_t, the bound holds. Substituting $D_t \propto \sqrt{r/t}$ yields $\log(1/D_t) \propto \log(\sqrt{t/r}) = \frac{1}{2}\log(t/r)$. The result follows. □

Then we start our proof from Eq. (8) and first apply inequalities (11).

$$\begin{aligned}\mathrm{Rad}_S(\mathcal{F}_{S_t}) &\lesssim \frac{1}{\sqrt{m}} \int_0^{\mathrm{Diam}(\mathcal{F}_{S_t})} \sqrt{\log N(\mathcal{F}_{S_t}, \|\cdot\|_{m,2}, \varepsilon)}\, d\varepsilon \\ &\le \frac{1}{\sqrt{m}} \int_0^{L_f D_t} \sqrt{\log N(\Theta_{S_t}, \|\cdot\|_2, \varepsilon/L_f)}\, d\varepsilon\end{aligned}$$

Now, we perform a change of variables in the integral. Let $u = \varepsilon/L_f$. This implies $\varepsilon = u \cdot L_f$ and $d\varepsilon = L_f\, du$. The integration limits for u become 0 to D_t.

$$\begin{aligned}\mathrm{Rad}_S(\mathcal{F}_{S_t}) &\lesssim \frac{1}{\sqrt{m}} \int_0^{D_t} \sqrt{\log N(\Theta_{S_t}, \|\cdot\|_2, u)} \cdot (L_f\, du) \\ &= \frac{L_f}{\sqrt{m}} \int_0^{D_t} \sqrt{\log N(\Theta_{S_t}, \|\cdot\|_2, u)}\, du \\ &\lesssim \frac{L_f}{\sqrt{m}} \int_0^{D_t} \sqrt{D\log\left(\frac{3D_t}{\varepsilon}\right)}\, d\varepsilon \quad \text{(by Eq. (10))} \\ &= \frac{L_f\sqrt{D}}{\sqrt{m}} \int_0^{D_t} \sqrt{\log\left(\frac{3D_t}{\varepsilon}\right)}\, d\varepsilon \\ &\lesssim \frac{L_f\sqrt{rD}}{\sqrt{m}} \cdot O\left(\frac{\sqrt{\log t/r}}{\sqrt{t}}\right) \quad \text{(by Lemma C.1)},\end{aligned}$$

which derives Theorem 4.1. □

C.4 Proof of Theorem 4.2

Proof. Following the architecture of the Transformer, define

$$f_{\text{bias}}(X) = \Delta w^\top x_q + \zeta(X),$$

where $\Delta w := w_{\text{PE}} - w_{\text{NoPE}}$ and $\zeta(X)$ is the non-linear error. From Assumption 4.2, $\|\Delta w\|_2 \le C_{\text{PE}}$. We will leave the discussion for $\zeta(X)$ to the end of this section and focus on the main linear effect first.

For Theorem 4.2, we only need to bound the Rademacher Complexity for $\mathcal{F}_{bias} = \{X \mapsto (\Delta w)^\top x_q \mid \|\Delta w\|_2 \leq C_{PE}\}$. Following the definition of Rademacher Complexity and on sample $S = \{X^j\}_{j=1}^m$, we have:

$$\begin{aligned}
\mathrm{Rad}_S(\mathcal{F}_{\text{bias}}) &= \frac{1}{m}\mathbb{E}_\sigma\left[\sup_{\|\Delta w\|_2 \leq C_{PE}} \sum_{j=1}^m \sigma_j (\Delta w)^\top x_q^{(j)}\right] \\
&= \frac{1}{m}\mathbb{E}_\sigma\left[\sup_{\|\Delta w\|_2 \leq C_{PE}} (\Delta w)^\top \left(\sum_{j=1}^m \sigma_j x_q^{(j)}\right)\right] \\
&\leq \frac{C_{PE}}{m}\mathbb{E}_\sigma \Big\| \sum_{j=1}^m \sigma_j x_q^{(j)} \Big\|_2 \\
&\leq \frac{C_{PE}}{m}\sqrt{\mathbb{E}_\sigma \Big\| \sum_{j=1}^m \sigma_j x_q^{(j)} \Big\|_2^2} = \frac{C_{PE}}{m}\sqrt{\Big\| \sum_{j=1}^m x_q^{(j)} \Big\|_2^2}.
\end{aligned}$$

For $\big\| \sum_{j=1}^m x_q^{(j)} \big\|_2^2$, we know from the concentration inequality that with high probability $1-\delta$,

$$\Big\| \sum_{j=1}^m x_q^{(j)} \Big\|_2^2 \leq \mathbb{E}\Big\| \sum_{j=1}^m x_q^{(j)} \Big\|_2^2 = md + O(\delta).$$

For simplicity, we take the result that

$$\mathrm{Rad}_S(\mathcal{F}_{\text{bias}}) \lesssim \frac{C_{PE}\sqrt{d}}{\sqrt{m}},$$

which gives the result.

Then we show that $\zeta(X)$ is bounded by $O(\gamma_{\text{eff}})$. According to the Definition 4.1, let $\|W_{\text{in}}\| \leq B_{\text{in}}$. If for any $\theta \in \Theta_{S_t}$, we let $\theta = \theta^* + \Delta\theta$, where θ^* is defined in Lemma 4.4, then

$$\|\Delta\theta\|_2 \leq \|W_{\text{in}} - W_{\text{in}}^*\|_F + \|W_{QK} - W_{QK}^*(S_t)\|_F \leq O(\gamma_{\text{eff}}),$$

with B_{in} absorbed in $O(\gamma_{\text{eff}})$. Then from Lemma 4.2, we have

$$\|f_\theta - f_\theta^*\|_{m,2} \leq L_f\|\theta - \theta^*\|_2 = L_f\|\Delta\theta\|_2 \lesssim O(\gamma_{\text{eff}}).$$

Also, for the effective linear weight,

$$\|w_{\text{eff}}^\theta - w_{\text{eff}}^{\theta^*}\|_2 \leq L_w\|\theta - \theta^*\|_2 \lesssim O(\gamma_{\text{eff}}).$$

Let $\zeta(X) = \eta_{\text{PE}}(X) - \eta_{\text{NoPE}}(X)$ and $\eta_\theta(X) = f_\theta(X) - (w_{\text{eff}}^\theta)^\top x_q$, then we have decomposition

$$\eta_\theta(X) = f_\theta(X) - f_{\theta^*}(X) - (w_{\text{eff}}^\theta - w_{\text{eff}}^{\theta^*})^\top x_q + [f_{\theta^*}(X) - (w_{\text{eff}}^{\theta^*})^\top x_q],$$

while the third term is 0 according to the definition of θ^*.

Hence, taking the empirical norm,

$$\|\eta_\theta(X)\|_{m,2} \leq \|f_\theta - f_\theta^*\|_{m,2} + \|w_{\text{eff}}^\theta - w_{\text{eff}}^{\theta^*}\|_2 \cdot \|x_q\|_{m,2} \lesssim O(\gamma_{\text{eff}}).$$

Finally,

$$\|\zeta(X)\|_{m,2} = \|\eta_{\text{PE}}(X) - \eta_{\text{NoPE}}(X)\|_{m,2} \leq \|\eta_{\text{PE}}(X)\|_{m,2} + \|\eta_{\text{NoPE}}(X)\|_{m,2} \leq O(\gamma_{\text{eff}}).$$

This completes the proof. Therefore, we can focus on the main linear effect in the following. □

C.5 Proof of Propositions in Section 4.2

In this section, we first prove the propositions in Section 4.2, which are Proposition 4.1, Proposition 4.2, and Proposition 4.3.

Proof of Proposition 4.1

Proof. For any perturbation X' such that $\|X' - X\|_F \leq \varepsilon$, we use the triangle inequality on $|f_\theta(X') - y|$:

$$\begin{aligned} |f_\theta(X') - y| &= |(f_\theta(X') - f_\theta(X)) + (f_\theta(X) - y)| \\ &\leq |f_\theta(X') - f_\theta(X)| + |f_\theta(X) - y| \\ &\leq L_x \|X' - X\|_F + |f_\theta(X) - y| \\ &\leq L_x \varepsilon + |f_\theta(X) - y|. \end{aligned}$$

By squaring it we can get:

$$\sup_{\|X' - X\| \leq \varepsilon} (f_\theta(X') - y)^2 \leq \left(|f_\theta(X) - y| + L_x \varepsilon\right)^2.$$

Thus, $\tilde{l}_\theta(X, y) \leq \ell_{\text{rob}}(f_\theta(X), y)$. □

Proof of Proposition 4.2

Proof. The Rademacher complexity is monotonic. Since every function in $\tilde{\mathcal{L}}_{\mathcal{F}_{S'_t}}$ is upper-bounded by a corresponding function in $\mathcal{L}_{\text{rob}}$, we have:

$$\text{Rad}_{S'}(\mathcal{L}_{\tilde{\mathcal{L}}_\mathcal{F}}) \leq \text{Rad}_{S'}(\mathcal{L}_{\text{rob}}).$$

The surrogate loss class $\mathcal{L}_{\text{rob}}$ has the structure $\{(x, y) \mapsto h(f(x), y) \mid f \in \mathcal{F}\}$, where $h(z, y) = (|z - y| + L_f \varepsilon)^2$. Assuming bounded outputs, h is L_h-Lipschitz with respect to its first argument z. By Lemma B.4:

$$\text{Rad}_{S'}(\mathcal{L}_{\text{rob}}) = \text{Rad}_{S'}(h \circ \mathcal{F}) \leq L_h \cdot \text{Rad}_{S'}(\mathcal{F}).$$

Combining the inequalities gives the desired result. For L_h, follow that $L_h = \sup_z |h'(z)|$, we then need to find $|h'(z)|$. Easily to know that $|h'(z)| = 2(|z - y| + L_x \varepsilon)$, which can then be used for the final L_h, that is $L_h = 2((M + M_y) + L_x \varepsilon)$.

Finally, we provide evidence that the response y in Proposition 4.2 is bounded with high probability.

Lemma C.2 (High-Probability Bound on Responses)**.** *Under the data generation model in Assumption 4.1, where $\mu \sim \mathcal{N}(0, I_d/d)$ and $y = \mu^\top x$ with $x \sim \mathcal{N}(0, I_d)$, for any failure probability $\delta_y \in (0, 1)$, there exists a constant M_y such that:*

$$\mathbb{P}_{w,x}(|y| > M_y) \leq \delta_y.$$

Proof. The response y has two sources of randomness: the true weight vector μ and the input x. Our proof proceeds in two steps.

Since $\mu \sim \mathcal{N}(0, I_d/d)$, each component $\mu_i \sim \mathcal{N}(0, 1/d)$. Thus, $d \cdot \|\mu\|_2^2 = \sum_{i=1}^d (\mu_i \sqrt{d})^2$ follows a Chi-squared distribution with d degrees of freedom, χ_d^2. A standard concentration inequality for Chi-squared variables states that for any $\delta_\mu \in (0, 1)$, $\|\mu\|_2^2 \leq (1 + \sqrt{2\log(1/\delta_\mu)/d})^2$ with probability at least $1 - \delta_\mu$ over the draw of μ. For simplicity, this means we can find a constant C_μ (e.g., $C_\mu = 2$ for large d) such that $\|\mu\|_2^2 \leq C_\mu$ with overwhelmingly high probability.

We now condition on the high-probability event that $\|\mu\|_2^2 \leq C_\mu$. Conditioned on a fixed μ satisfying this, $y = \mu^\top x$ is a zero-mean Gaussian random variable (as it is a linear combination of Gaussians). Its variance is:

$$\text{Var}(y|\mu) = \text{Var}(\mu^\top x|\mu) = \|\mu\|_2^2 \triangleq \sigma_y^2.$$

Since $\|\mu\|_2^2 \leq C_\mu$, we have $\sigma_y^2 \leq C_\mu$. Let $Z = y/\sigma_y \sim \mathcal{N}(0, 1)$. The standard tail bound gives

$$\mathbb{P}(|y| > k\sigma_y) \leq 2e^{-k^2/2}.$$

To make this probability less than a desired δ'_y, we can choose $k = \sqrt{2\log(2/\delta'_y)}$. This gives us a bound on y:

$$M_y = k\sigma_y \leq \sqrt{C_\mu}\sqrt{2\log(2/\delta'_y)}.$$

By a union bound, the probability that either $\|\mu\|_2^2 > C_\mu$ or $|y| > M_y$ is at most $\delta_\mu + \delta'_y$. By setting these failure probabilities appropriately, we can ensure the total failure probability is less than any given δ_y. Thus, the statement holds with high probability over the joint distribution of μ and x. □

□

Proof of Proposition 4.3

Proof. Let $X'_t = X_t + \Delta_X$, then following Wely Inequality in Lemma B.5, we have

$$\sigma_{\min}(X'_t) = \sigma_{\min}(X_t + \Delta_X) \geq \sigma_{\min}(X_t) - \|\Delta_X\|_2.$$

From the definition in Eq. (6), that is $\|X' - X\|_F \leq \varepsilon$, we will get $|\Delta_X\|_2 \leq \|\Delta_X\|_F \leq \varepsilon$, then

$$\sigma_{\min}(X'_t) = \sigma_{\min}(X_t + \Delta_X) \geq \sigma_{\min}(X_t) - \varepsilon \geq \sqrt{t} - \sqrt{d} - \varepsilon.$$

and then

$$\mathrm{Diam}(W'_{S_t}) \lesssim \frac{2\sqrt{r}}{\sqrt{t} - \sqrt{d} - \varepsilon} = \frac{2\sqrt{r}}{\sqrt{t}} \cdot \frac{\sqrt{t}}{\sqrt{t} - \sqrt{d} - \varepsilon}.$$

Therefore, we have the definition for Φ, that is

$$\Phi(\varepsilon, t, d) = \frac{\sqrt{t}}{\sqrt{t} - \sqrt{d} - \varepsilon} = \frac{1}{1 - \sqrt{d/t} - \varepsilon/\sqrt{t}}.$$

□

C.6 Proof of Theorem 4.3

Proof. From Proposition 4.2, we have

$$\mathrm{Rad}_{S'}(\tilde{\mathcal{L}}_{\mathcal{F}_{S'_t}}) \leq L_h \cdot \mathrm{Rad}_{S'}(\mathcal{F}_{S'_t}).$$

Then we follow the proof of Theorem 4.1 (C.3), but with $D'_t = \mathrm{Diam}(\Theta'_{S_t})$. Θ'_{S_t} is defined similarly to Θ_{S_t} in Definition 4.1 with S'_t:

$$\Theta_{S'_t} := \left\{(W_{\text{in}}, W_{QK}) : \|W_{QK} - W^*_{QK}(S'_t)\|_F \leq \gamma'_{\text{eff}}\right\},$$

Let $\Delta'_t := \|W^*_{QK}(S'_t) - W^*_{QK}\|_F$. Following Appendix C.1, to show that $\Delta'_t \lesssim O(\sqrt{d/t})$ with high probability, since

$$\|G'_t - I_d\|_2 \leq \|G_t - I_d\|_2 + \|G'_t - G_t\|_2,$$

we only need to consider $\|G'_t - G_t\|_2$, where $G'_t = \frac{1}{t}\sum_{i=1}^t x'_i x_i'^\top$. Let $\delta_i = x'_i - x_i$, then

$$\begin{aligned}
\|G'_t - G_t\|_2 &= \|\frac{1}{t}\sum_{i=1}^t (x_i\delta_i^\top + \delta_i x_i^\top + \delta_i\delta_i^\top)\|_2 \\
&\leq \frac{1}{t}\sum_{i=1}^t \|\delta_i\|_2^2 + \frac{2}{t}\sum_{i=1}^t \|\delta_i\|_2\|x_i\|_2 \\
&\leq \frac{\varepsilon^2}{t} + \frac{2}{t}\left(\sum_{i=1}^t \|\delta_i\|_2^2\right)^{\frac{1}{2}}\left(\sum_{i=1}^t \|x_i\|_2^2\right)^{\frac{1}{2}} \\
&\leq \frac{\varepsilon^2}{t} + \frac{2\sqrt{d}}{\sqrt{t}}\varepsilon.
\end{aligned}$$

This gives the result. Therefore similarly, for any fixed $t_{\min}$ and all $t \geq t_{\min}$, we can set constant $\gamma'_{\text{eff}} = \gamma + \Delta'^*$ and $\Delta'^* := \sup_{t \geq t_{\min}} \Delta'_t$.

Then it is easy to know

$$\begin{aligned}\mathrm{Rad}_{S'}(\mathcal{F}_{S'_t}) &\lesssim \frac{L_f}{\sqrt{m}}\int_0^{D'_t}\sqrt{\log N(\Theta'_{S_t}, \|\cdot\|_2, u)}\,du \\ &\lesssim \frac{L_f}{\sqrt{m}}\int_0^{D'_t}\sqrt{D\log\left(\frac{3D'_t}{\varepsilon}\right)}\,d\varepsilon \\ &\lesssim \frac{L_f\sqrt{rD}}{\sqrt{m}}\cdot\Phi\cdot O\left(\frac{\sqrt{\log t/r}}{\sqrt{t}}\right),\end{aligned}$$

which completes the proof. □

C.7 Proof of Theorem 4.4

Proof. From Proposition 4.2, we have

$$\mathrm{Rad}_{S'}(\tilde{\mathcal{L}}_{\mathcal{F}_{S'_t,\mathrm{PE}}}) \le L_h\cdot\mathrm{Rad}_{S'}(\mathcal{F}_{S'_t,\mathrm{PE}}).$$

Following the function class decomposition in Eq. (16), we have

$$\mathrm{Rad}_{S'}(\mathcal{F}_{S'_t,\mathrm{PE}}) \le \mathrm{Rad}_{S'}(\mathcal{F}_{S'_t}) + \mathrm{Rad}_{S'}(\mathcal{F}_{\mathrm{bias}}).$$

Note that the bias function itself does not change under adversarial attack, it is only determined by the presence of PE. What changes is only the evaluation set (from S_t to S'_t), which affects the empirical Rademacher complexity. Similarly, we only need to bound the bias term $\mathrm{Rad}_{S'}(\mathcal{F}_{\mathrm{bias}})$, and we consider

$$f_{\mathrm{bias}}(X') = \Delta w^\top x'_q + \zeta(X').$$

Follow the same process in proof C.4, for the attacked sample $S' = \{X'^{(j)}\}_{j=1}^m$, we have its empirical Rademacher complexity as

$$\mathrm{Rad}_{S'}(\mathcal{F}_{\mathrm{bias}}) = \frac{1}{m}\mathbb{E}_\sigma\left[\sup_{\|\Delta w\|_2\le C_{PE}}\sum_{j=1}^m\sigma_j(\Delta w)^\top x_q'^{(j)}\right] \le \frac{C_{PE}}{m}\mathbb{E}_\sigma\Big\|\sum_{j=1}^m\sigma_j x_q'^{(j)}\Big\|_2.$$

Let $\delta_q^{(j)} = x_q'^{(j)} - x_q^{(j)}$, therefore

$$\|\delta_q^{(j)}\|_2 = \|x_q'^{(j)} - x_q^{(j)}\|_2 \le \|X' - X\|_2 \le \|X' - X\|_F \le \varepsilon.$$

Then we have

$$\Big\|\sum_{j=1}^m\sigma_j x_q'^{(j)}\Big\|_2 \le \Big\|\sum_{j=1}^m\sigma_j x_q^{(j)}\Big\|_2 + \Big\|\sum_{j=1}^m\sigma_j\delta_q'^{(j)}\Big\|_2.$$

Since

$$\mathbb{E}_\sigma\Big\|\sum_{j=1}^m\sigma_j\delta_q'^{(j)}\Big\|_2 \le \left(\sum_{j=1}^m\|\delta_q^{(j)}\|_2^2\right)^{\frac{1}{2}} \le \sqrt{m}\varepsilon,$$

we get the result:

$$\mathrm{Rad}_{S'}(\mathcal{F}_{\mathrm{bias}}) \lesssim \frac{C_{\mathrm{PE}}\sqrt{d}}{\sqrt{m}} + \frac{C_{\mathrm{PE}}\varepsilon}{\sqrt{m}}.$$

Finally

$$\begin{aligned}\mathrm{Rad}_{S'}(\tilde{\mathcal{L}}_{\mathcal{F}_{S'_t,\mathrm{PE}}}) \lesssim L_h\cdot\Bigg[&\left(\frac{C_{PE}\sqrt{d}}{\sqrt{m}} + \frac{C_{PE}\varepsilon}{\sqrt{m}}\right) \\ &+ \left(O\left(\frac{L_f\sqrt{rD}\sqrt{\log t/r}}{\sqrt{mt}}\right)\cdot\Phi(\varepsilon,t,d)\right)\Bigg].\end{aligned}$$

□

D Extension to sub-Gaussian designs

In this section, we claim that our proof can be extended to sub-Gaussian distributions. Importantly, all dependence on the context length t and sample size m remains unchanged; only constants are modified. We first provide the definition for sub-Gaussian distribution.

Definition D.1. A zero-mean random variable X is sub-Gaussian if there exists $K > 0$ such that

$$\mathbb{P}(|X| > t) \leq 2\exp\left(-\frac{t^2}{K^2}\right) \quad \forall t > 0,$$

and its sub-Gaussian norm is denoted $\|X\|_{\psi_2} = K$.

This class includes Bernoulli variables, bounded variables (e.g. normalized images, token embeddings), mixtures of Gaussians, and many real-world data-generating processes. Thus extending our analysis to sub-Gaussian designs significantly broadens its applicability. We extend Assumption 4.1 as follows.

Assumption D.1 (sub-Gaussian design)**.** The input vectors $\{x_i\}_{i=1}^t$ are i.i.d., zero-mean, $\mathbb{E}[x_i x_i^\top] = I_d$, and sub-Gaussian with $\|x_i\|_{\psi_2} \leq K$ for some constant K.

D.1 Minimum singular value of the data matrix

Lemma 4.3 relies on the lower bound $\sigma_{\min}(X_t) \gtrsim \sqrt{t} - \sqrt{d}$ as shown in C.2. For sub-Gaussian designs, a similar bound holds with modified constants.

Lemma D.1 (Minimum singular value under sub-Gaussian design)**.** *Under Assumption D.1, there exist positive constants $c_1(K), c_2(K)$ depending only on the sub-Gaussian norm K such that with probability at least $1 - 2\exp(-c_2 t)$,*

$$\sigma_{\min}(X_t) \;\geq\; c_1(K)\big(\sqrt{t} - \sqrt{d}\big).$$

This is a direct result of standard sub-Gaussian random matrix theory (Vershynin, 2018). Therefore the term $1/\sigma_{\min}(X_t)$ used in our generalization proofs preserves its $1/\sqrt{t}$ dependence exactly.

D.2 Concentration of quadratic norms

The only place where the Gaussian assumption enters the proof of Theorem 4.2 is the concentration of $\left\|\sum_{j=1}^m \sigma_j x_q^{(j)}\right\|_2^2$ (or equivalently $\|\sum x^{(j)}\|_2^2$ for simplicity, see Appendix C.4). The results in Appendix C.4 also hold for Sub-Gaussian vectors, introducing only an additional multiplicative constant related to K into the complexity term $C_{\mathrm{PE}}\sqrt{d/m}$.

In conclusion, our theory does not rely essentially on Gaussianity; all key results extend to sub-Gaussian distributions with only constant-factor changes. Thus our clean and adversarial RC bounds are robust to a wide variety of realistic input distributions.

E Comparison with Non-Trainable Positional Encodings

In this section, we extend our analysis to non-trainable positional encodings, specifically Rotary Positional Embeddings (RoPE). We discuss how RoPE fits into our theoretical framework and provide empirical comparisons with Trainable PE.

E.1 Theoretical Perspective

Our main theoretical results (Theorem 4.1 and Theorem 4.2) quantify the "cost of learnability" for positional information. We focus on completely trainable PE as a "worst-case" baseline because it represents the maximum flexibility in the hypothesis space. From this perspective, our results serve as a theoretical upper bound. In contrast, RoPE applies a fixed rotation to the query and key vectors. Under our function class decomposition (Eq. (16)), the bias term $\mathcal{F}_{\text{bias}}$ for RoPE becomes a fixed transformation with no additional learnable parameters. Therefore, the Rademacher complexity of this bias term vanishes ($Rad_S(\mathcal{F}_{bias}) \approx 0$), and RoPE should exhibit a much tighter generalization gap than completely trainable PE, comparable to the "No-PE" baseline.

E.2 Empirical Verification

To validate this theoretical prediction, we conducted experiments comparing RoPE, Trainable PE, and No-PE across different context lengths t. The results are presented in Table 1.

Table 1: Generalization Gap Comparison across Different Context Lengths t.

t	**Without PE**	**Trainable PE**	**RoPE**
6	1.6149	2.0100	1.5722
7	1.9030	2.0314	1.9081
8	1.7055	2.1100	1.7434
9	1.7111	1.8812	1.7406
10	1.4730	1.7571	1.5002
12	1.1331	1.2565	1.2060
15	1.0547	1.2176	1.0804
20	0.8362	0.9379	0.9969
25	0.8601	0.9817	0.8666
30	0.6845	0.8126	0.6610

As shown in Table 1, RoPE attains a generalization gap very similar to the "Without PE" baseline, and both are smaller than that of Trainable PE (except for an outlier at $t = 20$). For example, at $t = 30$, while the gap for Trainable PE remains at 0.8126, RoPE drops to 0.6610, effectively matching the Without PE baseline (0.6845). This empirically confirms our hypothesis that the widened gap observed in our main results is primarily driven by the parameter complexity of the trainable PE module.