TRAINING-FREE UNIVERSAL APPROXIMATION BY PROMPTING RANDOM TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

How expressive is prompting a transformer? Answering this question is important for disentangling the roles of prompting, architecture, and pretraining in transformer models, and for determining whether task-specific behavior must be stored in model weights or can instead be induced at inference time through the prompt. We show, in an approximation-theoretic sense, that pretraining is not strictly necessary: a single-layer softmax attention network with random, untrained weights can approximate any Hölder function on a compact manifold when steered by an appropriate soft prompt. Building on the connection between softmax attention and kernel methods, we construct, for each target function, an explicit query-independent soft prompt by solving linear systems that match attention logits to Gaussian kernel exponents, enabling the frozen transformer to emulate the classical Nadaraya–Watson kernel estimator. This construction requires only a mild rank condition on the weights, which we show holds almost surely under Gaussian initialization. The prompted network inherits the theoretical guarantees of kernel regression, leading to universal approximation results with minimax-optimal rates governed by the manifold's intrinsic dimension. Moreover, we quantify the cost of prompting, revealing a tradeoff among the norms of the constructed prompt tokens, prompt length, and hidden dimension. Numerical experiments corroborate the constructions and predicted rates.

1 INTRODUCTION

The transformer architecture (Vaswani et al., 2017) has seen tremendous real-world success, most notably through the rise of the large language model (LLM). Given the prohibitively expensive pretraining costs of these massive models, significant efforts have been dedicated to getting more out of existing pretrained models, rather than retraining from scratch. Ideally, general-purpose transformers could generalize to diverse domains, incorporate new information, or even self-improve. At the architectural level, (weight) *fine-tuning* and its many variants have proved popular, which work by adapting a subset of weights to different downstream tasks (Hu et al., 2022).

The focus of this work is to explore *prompt engineering*, broadly construed, which changes the behavior of the model through its inputs (Sahoo et al., 2024). Altering the prompt has emerged as a lightweight, flexible method of adapting pretrained transformers to new tasks at inference time. The most straightforward method directly edits the human-language input fed into the model, known as *hard prompting*. Some examples include providing examples in the prompt (*few-shot prompting*, Brown et al. (2020)) or requiring the model to think step-by-step (*Chain-of-Thought*, Wei et al. (2022)). While these work well heuristically, hard prompting is formally a discrete optimization problem, which makes it difficult from an optimization standpoint (Shin et al., 2020). More amenable to mathematical techniques are *soft prompting* (or *prompt tuning*, Lester et al. (2021)), which optimizes a prefix in real-valued embedding space, and *prefix-tuning* (Li & Liang, 2021), which appends trainable tokens to the keys and values of every attention head. Compared to weight-based tuning, these methods tend to require fewer trainable parameters and are easier to edit on the fly (Lester et al., 2021; Li & Liang, 2021). Despite widespread adoption, theoretical guarantees for prompt engineering are still developing.

Questions traditionally asked of neural network weights can equally be asked of prompts. For example, it is natural to wonder how expressive prompting methods are, especially in comparison to weight-based methods (Petrov et al., 2024b; Meyer et al., 2025). Viewing transformers through the lens of control theory (Luo et al., 2023), the expressivity of prompting corresponds to understanding reachability (Bhargava et al., 2023). Our work focuses on soft prompting in particular, where we formulate the question of expressivity formally through the ideas of approximation theory. One particularly relevant question is that of *universal approximation*: can a transformer with fixed weights be steered to approximate any function in some class simply by changing its prompt? In certain settings, this has been answered in the affirmative. Wang et al. (2023) and Hu et al. (2025) show prompting can universally approximate Lipschitz

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

sequence-to-sequence functions, and Petrov et al. (2024a) derive corresponding quantitative approximation rates for prompting a single attention head model. Nakada et al. (2025) study a mechanism behind prompting-based universal approximation, showing prompts can program a transformer to emulate ReLU networks at inference. All these works demonstrate the potential of prompt-based methods, but suffer from a limitation common to approximation-theoretic works: the results rely on particular transformer weights that the pretraining process is not guaranteed to find or even approximate. Understanding the training dynamics of transformers is an important (and difficult) question, which is only beginning to be studied (Ahn et al., 2023; Oymak et al., 2023; Huang et al., 2024; Zhang et al., 2024).

We consider function regression, but eschew pretraining completely by studying random transformers with frozen, untrained weights. Our work culminates in Theorem 6, which concerns a very simple transformer model consisting of an affine embedding layer, one softmax attention head, and an affine decoding layer, all of which have random Gaussian weights. Even in this extreme setting, we constructively show that for any given Hölder function in some class, there almost surely exists a soft prompt which, when prepended to any query, makes the transformer approximate the function evaluated at that query, with approximation rate scaling with prompt length. Moreover, we track the constructed soft prompt magnitude, exposing a tradeoff involving prompt length and hidden dimension.

Our approximation approach is straightforward and intuitive, building on connections between the attention mechanism (Bahdanau et al., 2015; Vaswani et al., 2017) and kernel methods (Tsai et al., 2019; Shen et al., 2026; Ching et al., 2026). In particular, we show that transformers can reproduce the Nadaraya-Watson kernel estimator (Nadaraya, 1964; Watson, 1964) in their forward pass, similar to Shen et al. (2026). Whereas they choose specific weights to algebraically build the estimator through the transformer blocks, we show that an appropriate soft prompt can already force a single softmax attention layer to closely match the estimator's functional form. The soft prompt is derived as the solution to specific linear systems, and we specify a mild rank condition on the weights (Condition 1) under which a solution exists. Using standard techniques from random matrix theory, we show that random Gaussian initializations satisfy this condition almost surely.

Working through the well-studied Nadaraya-Watson estimator allows us to borrow existing approximation theory bounds for kernel regression. We extend the approximation of the kernel estimator through the soft prompt to full quantitative universal approximation-type theorems for prompting with rate of convergence (Jackson-type), with probabilistic guarantees in the nondeterministic cases. Our results also inherit the beneficial properties of the kernel estimator, such as minimax-optimal rates for noisy regression of Hölder functions on manifolds (Bickel & Li, 2007) up to logarithmic factors, with primary dependence only on the manifold's intrinsic dimension.

We summarize our results as follows:

1. Given a single-layer softmax attention transformer whose weights satisfy a mild rank condition, and any set of n support points, we construct a soft prompt of prompt length n under which the transformer's output approximates the NW kernel estimator built on those points (Theorem 1). The soft prompt tokens are the minimum-norm solutions of linear systems matching attention logits to Gaussian kernel exponents; we use *logit shifting* to suppress the unwanted query self-attention term at an exponential rate.
2. We show that the rank condition for the weights is generic, satisfied by Gaussian initializations almost surely, which allows the construction to work with almost every random-attention transformer (Theorem 3). Furthermore, we extend to fully random transformers, where embedding and decoding become affine maps with Gaussian weights (Theorem 5). We conclude similar results in this setting.
3. In each setting, we produce corresponding quantitative universal approximation theorems for Hölder functions on a manifold (Theorems 2, 4 and 6). By tuning the kernel bandwidth and logit shift, we achieve the minimax-optimal noisy approximation rate of $\tilde{O}(n^{-2\alpha/(2\alpha+d_X)})$, where α is the Hölder exponent and d_X is the intrinsic manifold dimension.
4. We track the costs of our constructions in terms of the magnitude of the soft prompt tokens, and how it relates to the prompt length n and hidden dimension d_{h}. Prompt token norms grow as $O(n^{2/(2\alpha+d_X)})$ in the deterministic parameter case with fixed embedding and decoding, improving to $O(n^{2/(2\alpha+d_X)}/\sqrt{d_{\mathrm{h}}}+1)$ with random attention, and $O(n^{2/(2\alpha+d_X)}/d_{\mathrm{h}}+\sqrt{d_{\mathrm{h}}})$ for fully random transformers.

1.1 Other Related Works and Notation

Further discussion of related works can be found in Appendix A.1. Table 1 summarizes the notation of the variables and constants in this work. We use $\|\cdot\|$ to denote the ℓ^2-norm for vectors and the spectral norm for matrices. Boldface letters are reserved for vectors, lowercase regular letters are used for scalars, and matrices are denoted by uppercase regular letters. Throughout, we use the terms *weights* and *parameters* for the entries of the transformer's maps.

2 PRELIMINARIES

We begin by formally introducing our problem, along with the relevant notation and model details.

2.1 ARCHITECTURE

We consider the following hypothesis class ($\mathcal{G}$) of transformer networks: single-layer softmax-attention networks with fixed embedding and decoding maps. We define each component individually below.

Let d_{in} be the input dimension. The embedding maps each input token into a high-dimensional space of hidden dimension d_{h} via a homogeneous embedding with additional zero padding, arranged as a row vector. In the deterministic setting (Section 4.1), we need only that $d_{\text{h}} \geq d_{\text{in}} + 2$; in the later analysis on random versions (Sections 4.2 and 4.3) we will assume $d_{\text{h}} \gg d_{\text{in}} + 2$.

In this setting, our (hard) prompt will only consist of a single token $\mathbf{x} \in \mathbb{R}^{d_{\text{in}}}$, the query point we are predicting the function at, and we denote its embedding as

$$E(\mathbf{x}) = \mathbf{x}_{\text{emb}} \coloneqq [\mathbf{x}^T, 1, 0, \ldots, 0] \in \mathbb{R}^{1 \times d_{\text{h}}}. \tag{2.1}$$

Softmax (self-)attention A acting on a matrix $H \in \mathbb{R}^{(n+1) \times d_{\text{h}}}$ is defined as

$$A(H) = \text{softmax}\left(\frac{HW_Q(HW_K)^T}{\sqrt{d_{\text{h}}}}\right) HW_V \tag{2.2}$$

where softmax acts row-wise, with the query, key, and value matrices $W_Q, W_K, W_V \in \mathbb{R}^{d_{\text{h}} \times d_{\text{h}}}$.[1] Note that A outputs a matrix of the same size. We only consider single-head attention.

Since we are only interested in the transformer's prediction for the query $\mathbf{x}$, the decoder D only looks at the last row, where it reads out the bottom-right entry.

The transformer hypothesis class is defined as all compositions of these three components:

$$\mathcal{G} = \left\{G = D \circ A \circ E \mid W_Q, W_K, W_V \in \mathbb{R}^{d_{\text{h}} \times d_{\text{h}}}\right\}.$$

Notice the embedding and decoding are fixed in the current setup; in Section 4.3, we consider models where these components are also variable.

2.2 PROBLEM SETUP

We are interested in deriving a universal approximation-type theorem for prompting, which we formulate here. Let $G = D \circ A \circ E$ be a transformer network in the aforementioned class $\mathcal{G}$. We denote by $\widetilde{G} = D \circ A$, the part of the network after embedding. We prepend a soft prompt $P = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_n] \in \mathbb{R}^{d_{\text{h}} \times n}$ to the embedded $\mathbf{x}$ as

$$H_{P,\mathbf{x}} = \left[\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_n, \mathbf{x}_{\text{emb}}^T\right]^T \in \mathbb{R}^{(n+1) \times d_{\text{h}}}. \tag{2.3}$$

Our objective is: given some function f described as noisy samples,

$$\text{find a soft prompt} \qquad P \in \mathbb{R}^{d_{\text{h}} \times n} \qquad \text{such that} \qquad \widetilde{G}(H_{P,\mathbf{x}}) \approx f(\mathbf{x}) \quad \text{for all} \quad \mathbf{x} \in X.$$

The soft prompt should depend on the function being regressed (in particular, the noisy input-output samples of it), but be independent of the specific query.

3 PROOF STRATEGY

In this section, we give a comprehensive overview of our proof strategy, expanding on the informal description provided in the introduction. We will emphasize the novel aspects without much detail on standard techniques; complete formal theorem statements can be found in Section 4 and their rigorous proofs in Appendix C.

[1] We refer to these as the *attention parameters* to differentiate them from *attention weights*, the standard term for the post-softmax coefficients.

Recall regression with the *Nadaraya-Watson (NW) kernel estimator*. Assuming we have an empirical description of f as noisy input-output pairs: using an unnormalized Gaussian kernel with bandwidth $h > 0$, with n support points $\{\mathbf{x}_i \in X\}_{i=1}^n$ and corresponding target values $\{y_i \coloneqq f(\mathbf{x}_i) + \varepsilon_i\}_{i=1}^n$, the noisy empirical NW kernel estimator is defined on an $\mathbf{x} \in X$ as

$$\widehat{\mathcal{K}}_h(\mathbf{x}) = \frac{\sum_{i=1}^n \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}_i\|^2}{2h^2}\right) y_i}{\sum_{i=1}^n \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}_i\|^2}{2h^2}\right)} = \frac{\sum_{i=1}^n \exp\left(\frac{\mathbf{x}^T\mathbf{x}_i}{h^2} - \frac{\|\mathbf{x}_i\|^2}{2h^2}\right) y_i}{\sum_{i=1}^n \exp\left(\frac{\mathbf{x}^T\mathbf{x}_i}{h^2} - \frac{\|\mathbf{x}_i\|^2}{2h^2}\right)}, \tag{3.1}$$

where the second equality gives an equivalent formulation that follows from expanding the square. The kernel estimator approximates $f(\mathbf{x})$ as a distance-weighted average of the function evaluated at the support points. We first show how and when a transformer can be prompted to emulate this estimator.

Given a transformer network $G \in \mathcal{G}$ with attention parameters W_Q, W_K, W_V and an input of the form (2.3), its output can be expanded according to the definitions in Section 2.1 as follows:

$$\widetilde{G}(H_{P,\mathbf{x}}) = \frac{\sum_{i=1}^n \exp(\mathbf{x}_{\mathrm{emb}} W_Q W_K^T \mathbf{p}_i/\sqrt{d_{\mathrm{h}}}) \mathbf{p}_i^T W_V \mathbf{e}_{\sqrt{d_{\mathrm{h}}}} + \exp(\mathbf{x}_{\mathrm{emb}} W_Q W_K^T \mathbf{x}_{\mathrm{emb}}^T/\sqrt{d_{\mathrm{h}}}) \mathbf{x}_{\mathrm{emb}} W_V \mathbf{e}_{\sqrt{d_{\mathrm{h}}}}}{\sum_{i=1}^n \exp(\mathbf{x}_{\mathrm{emb}} W_Q W_K^T \mathbf{p}_i/\sqrt{d_{\mathrm{h}}}) + \exp(\mathbf{x}_{\mathrm{emb}} W_Q W_K^T \mathbf{x}_{\mathrm{emb}}^T/\sqrt{d_{\mathrm{h}}})}. \tag{3.2}$$

Observe the likeness of the left part of the fraction in (3.2) with the NW kernel estimator (3.1). Softmax attention is directly a normalized kernel smoother, so this similarity is not coincidental. This crucial observation inspires a straightforward construction: we pick the soft prompt vectors $\mathbf{p}_i$ such that (3.2) approximates (3.1), matching the attention logits with the Gaussian kernel exponents term-by-term.

There is an additional part remaining in the transformer output, corresponding to the self-attention contribution of the query token with itself, which does not appear in (3.1). We exponentially suppress this term with *logit shifting*, introducing an additional variable $\beta \geq 0$ that additively inflates the relevant attention logits: for $\beta \geq 0$, we have

$$\begin{aligned}
&\frac{\sum_{i=1}^n \exp(\mathbf{x}_{\mathrm{emb}} W_Q W_K^T \mathbf{p}_i/\sqrt{d_{\mathrm{h}}} + \beta) \mathbf{p}_i^T W_V \mathbf{e}_{d_{\mathrm{h}}} + \exp(\mathbf{x}_{\mathrm{emb}} W_Q W_K^T \mathbf{x}_{\mathrm{emb}}^T/\sqrt{d_{\mathrm{h}}}) \mathbf{x}_{\mathrm{emb}} W_V \mathbf{e}_{d_{\mathrm{h}}}}{\sum_{i=1}^n \exp(\mathbf{x}_{\mathrm{emb}} W_Q W_K^T \mathbf{p}_i/\sqrt{d_{\mathrm{h}}} + \beta) + \exp(\mathbf{x}_{\mathrm{emb}} W_Q W_K^T \mathbf{x}_{\mathrm{emb}}^T/\sqrt{d_{\mathrm{h}}})} \\
&= \frac{\sum_{i=1}^n \exp(\mathbf{x}_{\mathrm{emb}} W_Q W_K^T \mathbf{p}_i/\sqrt{d_{\mathrm{h}}}) \mathbf{p}_i^T W_V \mathbf{e}_{d_{\mathrm{h}}} + \frac{1}{e^\beta} \exp(\mathbf{x}_{\mathrm{emb}} W_Q W_K^T \mathbf{x}_{\mathrm{emb}}^T/\sqrt{d_{\mathrm{h}}}) \mathbf{x}_{\mathrm{emb}} W_V \mathbf{e}_{d_{\mathrm{h}}}}{\sum_{i=1}^n \exp(\mathbf{x}_{\mathrm{emb}} W_Q W_K^T \mathbf{p}_i/\sqrt{d_{\mathrm{h}}}) + \frac{1}{e^\beta} \exp(\mathbf{x}_{\mathrm{emb}} W_Q W_K^T \mathbf{x}_{\mathrm{emb}}^T/\sqrt{d_{\mathrm{h}}})}.
\end{aligned}$$

Any such logit shift can be incorporated into the soft prompt, at the cost of larger prompt tokens. Towards the above desiderata, we propose that the soft prompt tokens should simultaneously satisfy

$$\frac{1}{\sqrt{d_{\mathrm{h}}}} \mathbf{x}_{\mathrm{emb}} W_Q W_K^T \mathbf{p}_i = \frac{\mathbf{x}^T \mathbf{x}_i}{h^2} - \frac{\|\mathbf{x}_i\|^2}{2h^2} + \beta \qquad \text{and} \qquad \mathbf{p}_i^T W_V \mathbf{e}_{d_{\mathrm{h}}} = y_i.$$

Denote by $W \coloneqq \frac{1}{\sqrt{d_{\mathrm{h}}}} W_Q W_K^T$ and $\mathbf{w}_V \coloneqq W_V \mathbf{e}_{d_{\mathrm{h}}}$. This gives rise to the following n linear systems, one for each $\mathbf{p}_i$:

$$\underbrace{\begin{bmatrix} W_{1:d_{\mathrm{in}}} \\ W_{d_{\mathrm{in}}+1} \\ \mathbf{w}_V^T \end{bmatrix}}_{\widetilde{W}} \mathbf{p}_i = \underbrace{\begin{bmatrix} \frac{\mathbf{x}_i}{h^2} \\ -\frac{\|\mathbf{x}_i\|^2}{2h^2} + \beta \\ y_i \end{bmatrix}}_{\mathbf{v}_i},$$

where $W_{1:d_{\mathrm{in}}}$ denotes its first d_{in} rows and $W_{d_{\mathrm{in}}+1}$ is the $(d_{\mathrm{in}}+1)$-th row. To construct a concrete soft prompt P^*, we take the minimum-norm solutions via the Moore-Penrose pseudoinverse.

A sufficient condition guaranteeing the existence of solutions, and thus of an appropriate soft prompt, is the coefficient matrix $\widetilde{W}$ being full row rank. We formalize this as a condition on the attention parameters.

Condition 1 (Solvability of prompt)**.** *The attention parameters of G are such that $\widetilde{W}$ has full row rank.*

This condition is stronger than is necessary for the existence of a sufficient soft prompt from this procedure, which only requires $\mathbf{v}_i$ to be in the column space of $\widetilde{W}$. We use it to additionally track the prompt norm of the construction, and have a general sample-independent condition. Furthermore, we show that Condition 1 is not overly restrictive; rather, it is generic, satisfied almost surely by Gaussian initializations of the attention parameters. This is what allows us to prove these results for random transformers, as opposed to one with particular weights.

We impose standard assumptions to derive our explicit bounds. In particular, we restrict our study to Hölder functions on a compact manifold, with bounded label noise.

Assumption 1 (Domain). *The domain X is a compact d_X-dimensional Riemannian manifold ($1 \leq d_X \leq d_{\text{in}}$) with positive reach $\tau_X > 0$, isometrically embedded in $\mathbb{R}^{d_{\text{in}}}$, satisfying $\|\mathbf{x}\| \leq R$ for all $\mathbf{x} \in X$.*

Assumption 2 (Function class). *The function class is restricted to $f \in C_L^\alpha(X)$, consisting of α-Hölder functions $X \to \mathbb{R}$ with Hölder constant L, where $0 < \alpha \leq 1$.*

Assumption 3 (Label noise). *Labels are observed with noise: $y_i = f(\mathbf{x}_i) + \varepsilon_i$, where $\varepsilon_1, \ldots, \varepsilon_n$ are independent, mean-zero, and satisfy $|\varepsilon_i| \leq S_{\text{noise}}$ almost surely, independent of the support points $\{\mathbf{x}_i\}_{i=1}^n$ and the query x.*

For a fixed f, let $S_f \coloneqq \max_{\mathbf{x} \in X} |f(\mathbf{x})|$. Denote by S the combined upper bound $S \coloneqq S_f + S_{\text{noise}}$.

With this, we are able to formally show that, given a fixed transformer and any fixed set of kernel support points, we may construct a soft prompt such that the transformer output approximates the output of the empirical kernel estimator. Furthermore, we derive the rate of approximation and a bound on the norm of the constructed soft prompt tokens. We prove analogous results in three settings: deterministic weights (fixed weights satisfying Condition 1), random attention parameters, and fully random transformers. The last setting uses random affine embedding and decoding, and thus differs in architecture from Section 2.1. The details of the corresponding proof also differ, but it essentially follows the same overall approach as described here.

Leveraging existing results in approximation theory for kernel regression, we extend our constructions in each setting to complete Jackson-type universal approximation results. Introducing the noiseless empirical kernel estimator and the noiseless integral kernel estimator as additional intermediate terms, the total squared approximation error is bounded by the following decomposition:

$$4\underbrace{\left(\widetilde{G}(H_{P^*,\mathbf{x}}) - \widehat{\mathcal{K}}_h(\mathbf{x})\right)^2}_{\text{prompt construction}} + 4\underbrace{\left(\widehat{\mathcal{K}}_h(\mathbf{x}) - \overline{\mathcal{K}}_h(\mathbf{x})\right)^2}_{\text{noise}} + 4\underbrace{\left(\overline{\mathcal{K}}_h(\mathbf{x}) - \mathcal{K}_h(\mathbf{x})\right)^2}_{\text{variance}} + 4\underbrace{\left(\mathcal{K}_h(\mathbf{x}) - f(\mathbf{x})\right)^2}_{\text{bias}}.$$

Formal definitions, along with the lemmas bounding the relevant terms, can be found in Appendix B. Picking the kernel bandwidth and logit shift to balance the terms in each respective setting, we conclude minimax-optimal rates for noisy regression of Hölder functions on manifolds and quantify the respective effects on prompt token norms.

Remark 1. Our framework focuses on regressing scalar-valued functions (Assumption 2) to clearly demonstrate the relevant mechanisms at play. It is straightforward to extend our results to the vector-valued setting, essentially by doing coordinate-wise kernel regression. We give a sketch of how Theorem 6 would generalize in Appendix D.

4 Main Results

We formally state here our theorems in each setting, with all complete proofs deferred to Appendix C.

4.1 Deterministic Weights

Throughout this subsection, we require $d_{\text{h}} \geq d_{\text{in}} + 2$. Our first theorem states that under the aforementioned assumptions and conditions, prompting can make a transformer emulate the NW kernel estimator on any fixed set of support points at a specific rate. The proof can be found in Appendix C.1.

Theorem 1 (Prompting can approximate NW kernel estimator). *Let X satisfy Assumption 1 and let $f : X \to \mathbb{R}$ satisfy Assumption 2. Let $n \geq 1$ and fix support points $\{\mathbf{x}_i \in X\}_{i=1}^n$ with labels satisfying Assumption 3; let $\widehat{\mathcal{K}}_h(\mathbf{x})$ denote the Nadaraya-Watson kernel estimator evaluated at $\mathbf{x}$, constructed using these support points and bandwidth $h > 0$. Let $G \in \mathcal{G}$ be a transformer network whose attention parameters satisfy Condition 1. For any logit shift $\beta \geq 0$, there exists a corresponding soft prompt P^* of length n such that*

$$\sup_{\mathbf{x} \in X} \left(\widetilde{G}(H_{P^*,\mathbf{x}}) - \widehat{\mathcal{K}}_h(\mathbf{x})\right)^2 = O\left(\frac{1}{n^2 e^{2\beta}} \cdot \exp\left(\frac{3R^2}{h^2}\right)\right),$$

where $O(\cdot)$ depends on $\|W\|, \|\mathbf{w}_V\|, R$, and S. Furthermore, the soft prompt tokens satisfy

$$\|\mathbf{p}_i^*\| = O(\beta + h^{-2})$$

for all $1 \leq i \leq n$, where $O(\cdot)$ hides a dependence on $\sigma_{\min}(\widetilde{W})^{-1}, R$, and S.

The logit shift suppresses the unwanted query self-attention term at an exponential rate; without it ($\beta = 0$), there is still quadratic decay in prompt length, but this is insufficient for optimal rates in the upcoming universal approximation theorems.

As described at the end of Section 3, with standard techniques and results from approximation theory, the previous theorem can be extended to derive complete universal approximation error rates for regression. Mirroring the in-context learning setup, we study regression with sampled support points (as opposed to fixed designs) and a sampled query.

Assumption 4 (Distribution of support points and query)**.** *Support points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ and the query $\mathbf{x}$ are drawn iid from the uniform distribution on X, denoted $\mathcal{U}(X)$.*

Our universal approximation theorem for prompting in the deterministic case is stated below, showing that under certain assumptions, any function can be approximated at a specific rate by prompting. The proof is deferred to Appendix C.2. Note that by our approach of going through the NW kernel estimator, we also inherit its positive qualities: we recover the minimax-optimal convergence rate for noisy nonparametric regression on manifolds (up to logarithmic factors), which is independent of the ambient dimension d_{in} (up to constants, through R).

Theorem 2 (Universal approximation for prompting)**.** *Let X satisfy Assumption 1 and $G \in \mathcal{G}$ be a transformer network whose attention parameters satisfy Condition 1. Suppose the support points and query satisfy Assumption 4, and the labels satisfy Assumption 3. For any f satisfying Assumption 2, there exists a map assigning to each realization of samples $\{\mathbf{x}_i, y_i = f(\mathbf{x}_i) + \varepsilon_i\}_{i=1}^n$ a soft prompt P^* of length n, independent of the query $\mathbf{x}$, such that*

$$\mathop{\mathbb{E}}_{\mathbf{x}_i, \varepsilon_i, \mathbf{x}} (\widetilde{G}(H_{P^*,\mathbf{x}}) - f(\mathbf{x}))^2 = O\left(n^{-\frac{2\alpha}{2\alpha + d_X}} (\log n)^{1+\frac{3d_X}{2}}\right).$$

Moreover, uniformly over all realizations of samples, the tokens of P^* have norm bounded as

$$\|\mathbf{p}_i^*\| = O(n^{\frac{2}{2\alpha+d_X}}).$$

4.2 Random Attention Parameters

Instead of directly assuming particular parameters which satisfy Condition 1, we consider attention parameters with Gaussian entries, and show that these satisfy the condition almost surely. This lets us extend the universal approximation theorem for prompting to these random transformers.

Assumption 5 (Random attention parameters)**.** *The entries of the attention parameters W_K, W_Q, and W_V are drawn iid from Gaussian distributions $\mathcal{N}(0, \sigma_{W_K}^2)$, $\mathcal{N}(0, \sigma_{W_Q}^2)$, and $\mathcal{N}(0, \sigma_{W_V}^2)$ respectively.*

Unless otherwise stated, our numerical experiments initialize all frozen random parameters according to $\mathcal{N}(0, 1)$, and our theoretical derivations reflect this choice (see Remark 2 for further discussion).

While our previous deterministic results were for any hidden dimension satisfying $d_{\text{h}} \geq d_{\text{in}} + 2$, in the random setting we consider the large-width limit where $d_{\text{h}} \gg d_{\text{in}} + 2$, which is common in practice (Brown et al., 2020). We first formulate a random version of Theorem 1. Note that in the random setting, the quantities $\|W\|$ and $\|\mathbf{w}_V\|$, which were treated as constants in Theorem 1, are in fact random variables. We adjust our theorem accordingly to make this explicit; the proof is found in Appendix C.3.

Theorem 3 (Prompting NW kernel estimator from random attention transformer)**.** *Let X satisfy Assumption 1, $f : X \to \mathbb{R}$ satisfy Assumption 2, and labels satisfy Assumption 3. Let $G \in \mathcal{G}$ be a transformer network whose attention parameters satisfy Assumption 5. With probability 1 over the initialization of attention parameters, for any prompt length $n \geq 1$ and support points $\{\mathbf{x}_i \in X\}_{i=1}^n$, bandwidth $h > 0$, and logit shift $\beta \geq 0$, there exists a soft prompt P^* of length n such that*

$$\sup_{\mathbf{x} \in X} \left(\widetilde{G}(H_{P^*,\mathbf{x}}) - \widehat{\mathcal{K}}_h(\mathbf{x})\right)^2 \leq \frac{C_G^2}{n^2 e^{2\beta}} \cdot \exp\left(\frac{3R^2}{h^2}\right),$$

where $C_G = \exp\left((R^2+1)\|W\|\right)\left(\sqrt{R^2+1}\|\mathbf{w}_V\| + S\right)$ is an almost surely finite random constant depending on the realized parameters. Moreover, for any $\delta > 0$, the norm of the constructed soft prompt tokens $\|\mathbf{p}_i^*\|$ is bounded as

$$\|\mathbf{p}_i^*\| = O\left(\frac{\beta + h^{-2}}{\sqrt{d_{\text{h}}}}\right)$$

with probability at least $1 - \delta$ over the initialization of the attention parameters and uniformly over the realizations of the samples, where the hidden constant can depend on δ.

We get a universal approximation theorem in the random setting analogous to Theorem 2 as a relatively direct corollary, since the remaining approximation theory results on kernel regression are independent of transformers. The details are deferred to Appendix C.4.

Theorem 4 (Universal approximation for prompting random attention transformer)**.** *Let $G \in \mathcal{G}$ be a transformer network whose attention parameters satisfy Assumption 5. Suppose X satisfies Assumption 1 and fix an f satisfying Assumption 2. Suppose the support points and query satisfy Assumption 4, and labels satisfy Assumption 3. For almost every initialization of the attention parameters, there exists a map, depending on the realized parameters, assigning to each realization of samples $\{\mathbf{x}_i, y_i = f(\mathbf{x}_i) + \varepsilon_i\}_{i=1}^n$ a soft prompt P^* of length n, independent of the query $\mathbf{x}$, such that*

$$\mathop{\mathbb{E}}_{\mathbf{x}_i,\varepsilon_i,\mathbf{x}}(\widetilde{G}(H_{P^*,\mathbf{x}}) - f(\mathbf{x}))^2 = O\left(n^{-\frac{2\alpha}{2\alpha+d_X}}(\log n)^{1+\frac{3d_X}{2}}\right),$$

conditional on the realized parameters. Furthermore, for any $\delta > 0$, with probability at least $1-\delta$ over the attention parameters, the soft prompt tokens satisfy

$$\|\mathbf{p}_i^*\| = O\left(\frac{n^{\frac{2}{2\alpha+d_X}}}{\sqrt{d_{\mathrm{h}}}} + 1\right)$$

for all $1 \le i \le n$, uniformly over the realizations of the samples.

Compared to Theorem 2, the additional $O(1)$ cost in token norm comes from the fact that the specific logit shift is chosen adaptively to the realized network (in particular, the realization of the random attention parameters).

4.3 Random Embedding and Decoding

For fully random transformers, we expand our transformer class. The fixed embedding and decoding defined in Section 2.1 were used primarily for clarity and concreteness. Here, we redefine them as affine maps, which will also be initialized randomly alongside the attention parameters.

The embedding is now given by

$$E_{\mathrm{aff}}(\mathbf{x}) = \mathbf{x}_{\mathrm{emb}} \coloneqq \mathbf{x}^T U_E + \mathbf{b}_E^T, \tag{4.1}$$

where $U_E \in \mathbb{R}^{d_{\mathrm{in}} \times d_{\mathrm{h}}}$ and $\mathbf{b}_E \in \mathbb{R}^{d_{\mathrm{h}}}$. Similarly, letting H_{n+1} denote the last row of a matrix H, the decoding is redefined as

$$D_{\mathrm{aff}}(H) = H_{n+1}\mathbf{u}_D + b_D,$$

where $\mathbf{u}_D \in \mathbb{R}^{d_{\mathrm{h}}}$ and $b_D \in \mathbb{R}$. The bias is necessary in the embedding, but optional for decoding, as we will see in the proof of Theorem 5. Abusing notation, we define $\widetilde{G} = D_{\mathrm{aff}} \circ A$ like before.

We still consider a transformer model with a single attention layer. Our new transformer class is given by

$$\mathcal{G}_{\mathrm{aff}} = \left\{G = D_{\mathrm{aff}} \circ A \circ E_{\mathrm{aff}} \mid W_Q, W_K, W_V \in \mathbb{R}^{d_{\mathrm{h}} \times d_{\mathrm{h}}}, U_E \in \mathbb{R}^{d_{\mathrm{in}} \times d_{\mathrm{h}}}, \mathbf{b}_E, \mathbf{u}_D \in \mathbb{R}^{d_{\mathrm{h}}}, b_D \in \mathbb{R}\right\}.$$

We refer to the totality of the weights as the *transformer weights*, to distinguish them from just the attention parameters as before. We have a similar randomness assumption for the new embedding and decoding parameters.

Assumption 6 (Random embedding and decoding parameters)**.** *The embedding parameters U_E and $\mathbf{b}_E$ have entries distributed iid according to Gaussian distributions $\mathcal{N}(0, \sigma_{U_E}^2)$ and $\mathcal{N}(0, \sigma_{\mathbf{b}_E}^2)$ respectively. The decoding parameter $\mathbf{u}_D$ has entries distributed iid according to $\mathcal{N}(0, \sigma_{\mathbf{u}_D}^2)$ and $b_D \sim \mathcal{N}(0, \sigma_{b_D}^2)$.*

Even in the fully random setting, prompting can still reconstruct the NW kernel estimator, as the following theorem shows. The proof can be found in Appendix C.5. Again, we assume $d_{\mathrm{h}} \gg d_{\mathrm{in}} + 2$.

Theorem 5 (Prompting NW kernel estimator for fully random transformer)**.** *Let X satisfy Assumption 1 and $f : X \to \mathbb{R}$ satisfy Assumption 2. Assume labels satisfy Assumption 3. Fix $G \in \mathcal{G}_{\mathrm{aff}}$ whose attention parameters satisfy Assumption 5 and embedding and decoding parameters satisfy Assumption 6. With probability 1 over the initialization of the transformer weights, for any prompt length $n \ge 1$, set of support points $\{\mathbf{x}_i \in X\}_{i=1}^n$, bandwidth $h > 0$, and logit shift $\beta \ge 0$, there exists a soft prompt P^* of length n such that*

$$\sup_{\mathbf{x} \in X}\left(\widetilde{G}(H_{P^*,\mathbf{x}}) - \widehat{\mathcal{K}}_h(\mathbf{x})\right)^2 \le \frac{C_{G,\mathrm{aff}}^2}{n^2 e^{2\beta}} \cdot \exp\left(\frac{3R^2}{h^2}\right),$$

where

$$C_{G,\mathrm{aff}} \coloneqq \exp\left((R\|U_E\| + \|\mathbf{b}_E\|)^2 \|W\|\right)\left((R\|U_E\| + \|\mathbf{b}_E\|)\|W_V\|\|\mathbf{u}_D\| + |b_D| + S\right)$$

is an almost surely finite random constant depending on the realized transformer weights. Moreover, for any $\delta > 0$, the constructed soft prompt has tokens $\mathbf{p}_i^*$ satisfying

$$\|\mathbf{p}_i^*\| = O\left(\frac{\beta + h^{-2}}{d_{\mathrm{h}}}\right),$$

with probability at least $1-\delta$, uniformly over the realizations of the samples.

Using this result gives the universal approximation theorem in the fully random case, proved in Appendix C.6.

Theorem 6 (Universal approximation for prompting fully random transformer)**.** *Let $G \in \mathcal{G}_{\mathrm{aff}}$ be a transformer network whose attention parameters satisfy Assumption 5 and whose embedding and decoding parameters satisfy Assumption 6. Suppose X satisfies Assumption 1 and fix an f satisfying Assumption 2. Suppose the support points and query satisfy Assumption 4 with labels satisfying Assumption 3. For almost every initialization of the parameters, there exists a map assigning to each realization of samples $\{\mathbf{x}_i, y_i = f(\mathbf{x}_i) + \varepsilon_i\}_{i=1}^n$ a soft prompt P^* of length n, independent of the query $\mathbf{x}$, such that*

$$\mathop{\mathbb{E}}_{\mathbf{x}_i,\varepsilon_i,\mathbf{x}}(\widetilde{G}(H_{P^*,\mathbf{x}}) - f(\mathbf{x}))^2 = O\left(n^{-\frac{2\alpha}{2\alpha+d_X}}(\log n)^{1+\frac{3d_X}{2}}\right),$$

conditional on the realized parameters. Furthermore, for any $\delta > 0$, with probability at least $1-\delta$ over the transformer weights, the soft prompt tokens satisfy

$$\|\mathbf{p}_i^*\| = O\left(\frac{n^{\frac{2}{2\alpha+d_X}}}{d_{\mathrm{h}}} + \sqrt{d_{\mathrm{h}}}\right)$$

for all $1 \le i \le n$, uniformly over the realizations of the samples.

In contrast with Theorem 4, the random embedding and decoding improves the pseudoinverse scaling to $1/d_{\mathrm{h}}$, but also result in a larger additive term $\sqrt{d_{\mathrm{h}}}$ from the adaptive logit shift.

5 Numerical Experiments

We perform numerical experiments on synthetic datasets to validate our theoretical findings. We report here the main scaling result, with additional experiments in Appendix E. Our model architectures remain faithful to the settings where our results were derived, using a single softmax attention layer with no feedforward component. Our experiments involve the two architectural variants used in Sections 4.2 and 4.3 respectively:

- **Architecture A.** One-layer softmax attention with fixed embedding and decoding, exactly as specified in Section 2.1. In particular, we use a homogeneous embedding with additional zero padding and a readout decoder. The attention parameters are randomly initialized and not trained.
- **Architecture B.** One-layer softmax attention with variable affine embedding and decoding, as described at the beginning of Section 4.3. All transformer weights (attention, embedding, decoding) are randomly initialized and not trained.

As previously noted, all random parameters that will not be trained are initialized according to $\mathcal{N}(0,1)$ (see Remark 2 for further discussion). Only in the baseline comparisons for the regression rate experiment do we use width-dependent initializations. We perform all softmax computations in log-space. All experiments were run over 10 seeds.

Our domain is the unit sphere $X = S^2 \subset \mathbb{R}^3$ with $d_X = 2, d_{\mathrm{in}} = 3$, $R = 1$, and $\tau_X = 1$. Support points are drawn iid uniformly on X. Queries are either also sampled uniformly, or set over a 2,000-point Fibonacci lattice (when testing suprema). The test function is a unit-norm combination of degree-4 real spherical harmonics, drawn once by sampling a coefficient vector for the nine basis functions and fixed throughout. We find that the sampled function f has supremum norm $S_f \approx 2.6$, computed on a 200,000-point Fibonacci lattice. Spherical harmonics are smooth, and in particular $\alpha = 1$. Labels are generated with uniform noise $\mathcal{U}([-1,1])$.

Minimax regression rate. We verify Theorems 4 and 6, the predicted minimax convergence rate for universal approximation. Exactly as we did in the proofs, we pick h and β by explicitly computing the constants C_G and $C_{G,\mathrm{aff}}$, and directly constructing the soft prompts using the pseudoinverses.

We compare against four benchmarks, the last three of which are architecture dependent (A or B): (B1) the oracle (empirical) NW estimator; (B2) random frozen weights, trained prompt; (B3) no prompt, trained weights; (B4) frozen random prompt, trained weights.

For all benchmarks, parameters to be trained are initialized with variance $1/d_{\mathrm{h}}$ while everything else uses the standard variance 1 initialization. The trained parameters use full-batch Adam for 5000 steps, with learning rates chosen from $\{10^{-3}, 10^{-2}, 10^{-1}\}$ by MSE on a separate 500 sample validation set. The results are shown in Fig. 1.

In nearly all cases, the error is similar for small n, since the number of samples is insufficient to accurately capture the function. For both architectures, our closed-form construction of the prompt is virtually indistinguishable from the

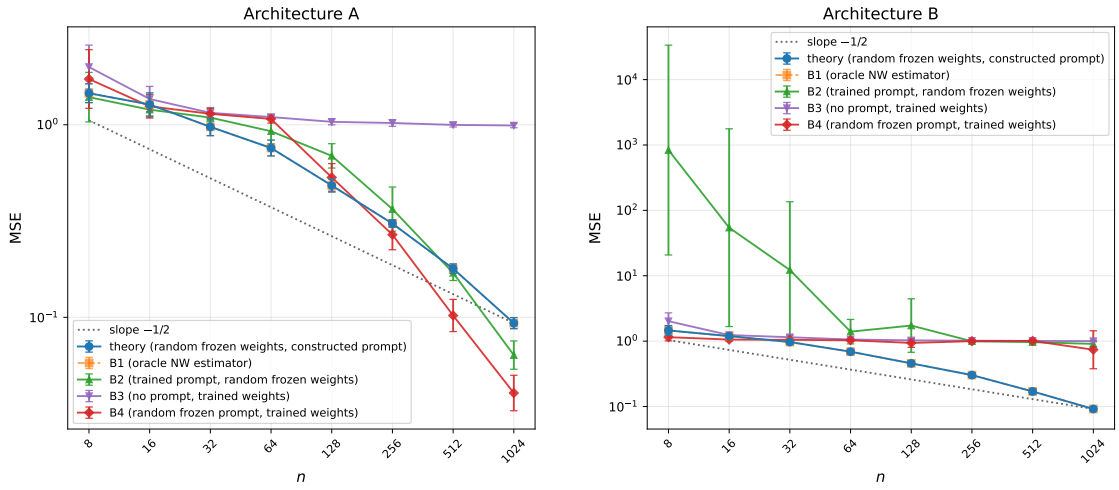


Figure 1: Log-log plots of MSE versus soft prompt length n for both architectures and various baselines, with fixed $d_{\mathrm{h}} = 256$ throughout. h and β are set according to the theory: $h = n^{-1/4}$ and $\beta = 3\sqrt{n}/2 + \log_+ C_G$ (or $\beta = 3\sqrt{n}/2 + \log_+ C_{G,\mathrm{aff}}$ for Architecture B). MSE is computed over 1000 uniformly sampled test queries per seed. Lines show geometric means and error bars are 95% Student-t CIs (computed in log-space).

empirical kernel estimator; comparing with the predicted minimax slope of $n^{-1/2}$ supports the upper bound for sample size/prompt length scaling. Benchmark 3 (weight training alone with no prompt) maintains MSE ≈ 1 throughout. This is unsurprising, since a single softmax attention head on a single token collapses to an affine map, which has insufficient capacity no matter how it is tuned.

In Architecture A, Benchmark 2 (frozen weights, trained prompt) shows slightly sharper scaling than our constructed prompts; the construction is sufficient, but likely not tight. Benchmark 4 (frozen prompt, trained weights) performs remarkably well, with seemingly better scaling than Benchmark 2, although not by a visibly significant amount. This highlights the utility of random context as a computational scratchpad for attention, an interesting idea that we leave for future work.

In Architecture B, Benchmarks 2 and 4 do not learn anything meaningful, with Benchmark 2 even performing worse than the flat Benchmark 3 at smaller n. This is almost certainly a numerical failure, since the unit-variance initializations for the frozen components scale the attention logits to a size that is computationally infeasible for gradient-based methods. We still expect our construction not to be tight, and this just shows that there are settings where an explicit construction can outperform direct computational optimization.

6 Conclusion

We showed that a single-layer softmax attention network with frozen random weights is a universal approximator under soft prompting: for a Hölder function on a manifold, we give an explicit construction of a soft prompt, as the minimum-norm solution to a linear system, that steers the network to emulate the Nadaraya-Watson kernel estimator, inheriting minimax-optimal rates that depend primarily on intrinsic dimension. The required rank condition on the weights holds almost surely under Gaussian initialization. We tracked the required soft prompt norm under three regimes: deterministic attention parameters, random attention parameters, and fully random transformers.

Our results demonstrate the expressivity of prompting in an extreme case, leveraging the computational power of a single random attention head. The result is based on an explicit construction that depends on the realized transformer, directly connecting softmax attention and kernel methods through the prompt.

Future work Our work has focused on a heavily simplified transformer model with a straightforward prompt construction. Can we incorporate other elements of transformers, such as multi-head attention, deeper networks, feedforward components, residual connections, causal masking, and layer norm? Are more efficient prompt constructions, with smaller norms or fewer tokens, possible?

AI USE STATEMENT

In this work, we used generative AI tools to help develop theoretical models and conceptual frameworks, formulate mathematical claims, provide ingredients for proving mathematical claims, assist in the development and verification of mathematical proofs, refine hypotheses, provide feedback on research methodology and experiments, and implement experimental methods. We have not used generative AI tools to generate synthetic datasets, assist with translation, clean and reformat datasets, support qualitative and thematic data analysis, or interpret results. Additionally, we used generative AI tools for coding the numerical experiments, brainstorming, searching for information and relevant literature, and providing suggestions on language and organization to improve clarity.

We have reviewed all AI-assisted work. All mathematical claims, proof ingredients, and derivations were thoroughly checked by the authors. Suggested references and related works have been verified against their original sources. All edits to the manuscript suggested by generative AI tools were carefully considered and implemented manually by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers Learn to Implement Preconditioned Gradient Descent for In-Context Learning. In *Advances in Neural Information Processing Systems*, 2023.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What Learning Algorithm Is In-Context Learning? Investigations with Linear Models. In *International Conference on Learning Representations*, 2023.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. In *International Conference on Learning Representations*, 2015.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as Statisticians: Provable In-Context Learning with In-Context Algorithm Selection. In *Advances in Neural Information Processing Systems*, 2023.

Aman Bhargava, Cameron Witkowski, Shi-Zhuo Looi, and Matt Thomson. What's the Magic Word? A Control Theory of LLM Prompting, 2023. URL https://arxiv.org/abs/2310.04444.

Peter J. Bickel and Bo Li. Local Polynomial Regression on Unknown Manifolds. *Lecture Notes-Monograph Series*, 54:177–186, 2007. ISSN 07492170.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, et al. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, 2020.

Michelle Ching, Ioana Popescu, Nico Smith, Tianyi Ma, William G. Underwood, and Richard J. Samworth. Efficient and Minimax Optimal In-context Nonparametric Regression with Transformers. In *International Conference on Machine Learning*, 2026.

Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, David Benjamin Belanger, Lucy J Colwell, and Adrian Weller. Rethinking Attention with Performers. In *International Conference on Learning Representations*, 2021.

George Cybenko. Approximation by Superpositions of a Sigmoidal Function. *Mathematics of Control, Signals and Systems*, 2:303–314, 1989.

Kenneth R. Davidson and Stanislaw J. Szarek. Chapter 8 - Local Operator Theory, Random Matrices and Banach Spaces. In W.B. Johnson and J. Lindenstrauss (eds.), *Handbook of the Geometry of Banach Spaces*, volume 1 of *Handbook of the Geometry of Banach Spaces*, pp. 317–366. Elsevier Science B.V., 2001. doi: https://doi.org/10.1016/S1874-5849(01)80010-3.

Yihe Dong, Lorenzo Noci, Mikhail Khodak, and Mufan Li. Is Random Attention Sufficient for Sequence Modeling? Disentangling Trainable Components in the Transformer, 2025. URL https://arxiv.org/abs/2506.01115.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What Can Transformers Learn In-Context? A Case Study of Simple Function Classes. In *Advances in Neural Information Processing Systems*, 2022.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and Mike Titterington (eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR.

Chi Han, Ziqi Wang, Han Zhao, and Heng Ji. Understanding Emergent In-Context Learning from a Kernel Regression Perspective. *Transactions on Machine Learning Research*, 2025.

Alex Havrilla and Wenjing Liao. Predicting Scaling Laws with Statistical and Approximation Theory for Transformer Neural Networks on Intrinsically Low-Dimensional Data. In *Advances in Neural Information Processing Systems*, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1026–1034, 2015. doi: 10.1109/ICCV.2015.123.

Alexander Hsu, Zhaiming Shen, Wenjing Liao, and Rongjie Lai. Understanding In-Context Learning for Nonlinear Regression with Transformers: Attention as Featurizer, 2026. URL https://arxiv.org/abs/2605.05176.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2022.

Jerry Yao-Chieh Hu, Wei-Po Wang, Ammar Gilani, Chenyang Li, Zhao Song, and Han Liu. Fundamental Limits of Prompt Tuning Transformers: Universality, Capacity and Efficiency. In *International Conference on Learning Representations*, 2025.

Jerry Yao-Chieh Hu, Hude Liu, Jennifer Yuntong Zhang, and Han Liu. In-Context Algorithm Emulation in Fixed-Weight Transformers. In *The Fourteenth International Conference on Learning Representations*, 2026.

Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew. Extreme Learning Machine: Theory and Applications. *Neurocomputing*, 70(1-3):489–501, 2006.

Yu Huang, Yuan Cheng, and Yingbin Liang. In-context Convergence of Transformers. In *Forty-first International Conference on Machine Learning*, 2024.

Boris Igelnik and Yoh-Han Pao. Stochastic Choice of Basis Functions in Adaptive Function Approximation and the Functional-Link Net. *IEEE Transactions on Neural Networks*, 6(6):1320–1329, 1995.

Herbert Jaeger. The "Echo State" Approach to Analysing and Training Recurrent Neural Networks. Technical Report GMD Report 148, German National Research Center for Information Technology, 2001.

Haotian Jiang and Qianxiao Li. Approximation Rate of the Transformer Architecture for Sequence Modeling. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 68926–68955. Curran Associates, Inc., 2024. doi: 10.52202/079017-2202.

Tokio Kajitsuka and Issei Sato. Are Transformers with One Layer Self-Attention Using Low-Rank Weight Matrices Universal Approximators? In *The Twelfth International Conference on Learning Representations*, 2024.

Brian Lester, Rami Al-Rfou, and Noah Constant. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

Xiang Lisa Li and Percy Liang. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021.

Yifan Luo, Yiming Tang, Chengfeng Shen, Zhennan Zhou, and Bin Dong. Prompt Engineering Through the Lens of Optimal Control, 2023. URL https://arxiv.org/abs/2310.14201.

Wolfgang Maass, Thomas Natschläger, and Henry Markram. Real-Time Computing Without Stable States: A New Framework for Neural Computation Based on Perturbations. *Neural Computation*, 14(11):2531–2560, 2002.

Maxime Meyer, Mario Michelessa, Caroline Chaux, and Vincent Y. F. Tan. Memory Limitations of Prompt Tuning in Transformers, 2025. URL https://arxiv.org/abs/2509.00421.

Elizbar A. Nadaraya. On Estimating Regression. *Theory of Probability and Its Applications*, 9(1):141–142, 1964.

Ryumei Nakada, Wenlong Ji, Tianxi Cai, James Zou, and Linjun Zhang. A Theoretical Framework for Prompt Engineering: Approximating Smooth Functions with Transformer Prompts, 2025. URL https://arxiv.org/abs/2503.20561.

Partha Niyogi, Stephen Smale, and Shmuel Weinberger. Finding the Homology of Submanifolds with High Confidence from Random Samples. *Discrete & Computational Geometry*, 39(1):419–441, March 2008. ISSN 1432-0444. doi: 10.1007/s00454-008-9053-2.

Samet Oymak, Ankit Singh Rawat, Mahdi Soltanolkotabi, and Christos Thrampoulidis. On the Role of Attention in Prompt-tuning. In *International Conference on Machine Learning*, 2023.

Hao Peng, Nikolaos Pappas, Dani Yogatama, Roy Schwartz, Noah Smith, and Lingpeng Kong. Random Feature Attention. In *International Conference on Learning Representations*, 2021.

Aleksandar Petrov, Philip H. S. Torr, and Adel Bibi. Prompting a Pretrained Transformer Can Be a Universal Approximator. In *International Conference on Machine Learning*, 2024a.

Aleksandar Petrov, Philip H. S. Torr, and Adel Bibi. When Do Prompting and Prefix-Tuning Work? A Theory of Capabilities and Limitations. In *International Conference on Learning Representations*, 2024b.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Ali Rahimi and Benjamin Recht. Random Features for Large-Scale Kernel Machines. In *Advances in Neural Information Processing Systems*, 2007.

Ali Rahimi and Benjamin Recht. Weighted Sums of Random Kitchen Sinks: Replacing minimization with randomization in learning. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou (eds.), *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications, 2024. URL https://arxiv.org/abs/2402.07927.

Sheng Shen, Alexei Baevski, Ari S. Morcos, Kurt Keutzer, Michael Auli, and Douwe Kiela. Reservoir Transformers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021a.

Sheng Shen, Zhewei Yao, Douwe Kiela, Kurt Keutzer, and Michael W. Mahoney. What's Hidden in a One-layer Randomly Weighted Transformer? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021b.

Zhaiming Shen, Alexander Hsu, Rongjie Lai, and Wenjing Liao. Understanding In-Context Learning on Structured Manifolds: Bridging Attention to Kernel Methods. In *The Fourteenth International Conference on Learning Representations*, 2026.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020.

Shokichi Takakura and Taiji Suzuki. Approximation and Estimation Ability of Transformers for Sequence-to-Sequence Functions with Infinite Dimensional Input. In *International Conference on Machine Learning*, 2023.

Yao-Hung Hubert Tsai, Shaojie Bai, Makoto Yamada, Louis-Philippe Morency, and Ruslan Salakhutdinov. Transformer Dissection: An Unified Understanding for Transformer's Attention via the Lens of Kernel. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, 2017.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2 edition, 2026.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, Joao Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers Learn In-Context by Gradient Descent. In *International Conference on Machine Learning*, 2023.

Haiyu Wang and Yuanyuan Lin. Prompt Tuning Transformers for Data Memorization. In *Advances in Neural Information Processing Systems*, 2025.

Yihan Wang, Jatin Chauhan, Wei Wang, and Cho-Jui Hsieh. Universality and Limitations of Prompt Tuning. In *Advances in Neural Information Processing Systems*, 2023.

Geoffrey S. Watson. Smooth Regression Analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 26(4):359–372, 1964.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, 2022.

Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94:103–114, 2017. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2017.07.002.

Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J. Reddi, and Sanjiv Kumar. Are Transformers Universal Approximators of Sequence-to-Sequence Functions? In *International Conference on Learning Representations*, 2020.

Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett. Trained Transformers Learn Linear Models In-Context. *Journal of Machine Learning Research*, 25(49):1–55, 2024.

Ziqian Zhong and Jacob Andreas. Algorithmic Capabilities of Random Transformers. In *Advances in Neural Information Processing Systems*, 2024.

A Supplementary Materials

A.1 Other Related Work

Single-task and in-context learning Traditional models are studied primarily in the setting of *single-task learning*, where each task or function requires its own tailored model. The ability of neural networks to approximate a single function is the foundation of approximation theory in machine learning, and has been studied extensively for various architectures (Cybenko, 1989; Yarotsky, 2017). For transformers in particular, single-task universal approximation was first shown by Yun et al. (2020), and later extended to other settings (Takakura & Suzuki, 2023; Havrilla & Liao, 2024; Jiang & Li, 2024). Relevant to our setting is the expressivity of models with only one attention layer, which has been studied by Kajitsuka & Sato (2024) and Jiang & Li (2024).

In contrast, transformers additionally demonstrate a remarkable phenomenon known as *in-context learning* (ICL), where a transformer is able to process and apply information provided in the prompt, entirely at inference time with frozen weights (Radford et al., 2019; Brown et al., 2020). The theoretical setting usually considers prompts consisting of a function, described as input-output pairs, which forms the *context*, followed by a separate *query* (Garg et al., 2022). ICL can be thought of as the transformer's ability to approximate an operator which maps such prompts to the function evaluated on the query. It must infer the relevant function from the context, then apply it to the query, and output the result, all in the same forward pass. A popular mechanistic explanation of ICL is that transformers, with the appropriate weights, are able to execute certain statistical algorithms at inference, such as gradient descent (Akyürek et al., 2023; von Oswald et al., 2023; Ahn et al., 2023), direct least-squares (Akyürek et al., 2023; Bai et al., 2023), polynomial and spline regression (Hsu et al., 2026), and kernel methods (Han et al., 2025; Shen et al., 2026). Related to our work, Hu et al. (2026) show that algorithm parameters can be encoded into the prompt, such that the transformer then emulates the corresponding algorithm in-context.

While our universal approximation theorems are presented in the language of single-task learning, we overlap heavily with the ICL setting. Most notably, all our results involve first fixing a transformer (i.e., a particular realization of the weights), which we then use to regress any function of a certain class by only adapting the soft prompt. The soft prompt is constructed based on the specific realization of the weights and function, but is independent of the chosen query. Much like context in the theoretical ICL setting, it is prepended to the query, which then steers the transformer to perform kernel regression. The main difference is that ICL assumes a fixed algebraic form of the input sequence (input-output pairs followed by query) and chooses the weights to process those, whereas we fix the weights and find a task-dependent soft prompt to achieve the same goal.

Randomness in machine learning Many techniques in machine learning leverage randomness. Beyond stochasticity in training, frozen random parameters have proven surprisingly effective and insightful across many domains, while also saving on the computational cost of training. Rahimi & Recht (2007; 2008) demonstrated the expressivity of random features in kernel methods, allowing kernel machines to scale to large datasets. Similar ideas of random representations were explored in neural networks: for feedforward networks, approaches like *Extreme Learning Machines* (Huang et al., 2006) and *Random Vector Functional Link* models (Igelnik & Pao, 1995) demonstrated that a single trainable linear readout following frozen random layers is sufficient for universal approximation. For sequential data, *reservoir computing* paradigms apply the same strategy to recurrent networks, projecting data into high-dimensional space through a fixed, untrained network (the *reservoir*) and only training a simple readout layer on top of that (Jaeger, 2001; Maass et al., 2002).

Random and partially random transformers have received growing attention. Choromanski et al. (2021) and Peng et al. (2021) introduce linear-complexity variants of softmax attention by approximation with random features. Shen et al. (2021b) demonstrate one-layer random-weight transformers contain performant subnetworks. Building on reservoir computing ideas, Shen et al. (2021a) study the effectiveness of interleaving random layers in a transformer, and Zhong & Andreas (2024) show training only the embeddings of an otherwise random transformer suffices for some nontrivial tasks. Dong et al. (2025) discover that random frozen attention performs competitively on language modeling, which they use to analyze the roles of different transformer components. Related to our setting, Wang & Lin (2025) observed prompt-tuned randomly initialized transformers can effectively memorize finite datasets.

Our work relies on randomness and mirrors many of the aforementioned ideas. Much like reservoir computing, we also borrow the expressivity of a black box, where the entire frozen untrained transformer network is the reservoir. Instead of training a readout layer, we only control the input to steer the transformer's behavior, and demonstrate that this is expressive enough. Moreover, beyond savings in pretraining cost, we also show how random standard normal weights can lead to smaller soft prompt norms by more effective use of hidden dimension.

A.2 Summary of Notation

Table 1: Summary of Notation

Notation	Description
X	Domain (manifold)
d_X	Intrinsic manifold dimension
R	$\max_{\mathbf{x}\in X} \|\mathbf{x}\|$
τ_X	Reach of X
d_{in}	Input dimension = ambient dimension
d_{h}	Hidden dimension
$\mathbf{x}_{\text{emb}}$	Embedding vector of $\mathbf{x}$
$\mathcal{G}$	Transformer class
G	A transformer
$\widetilde{G}$	The transformer G sans embedding
α	Hölder exponent
L	Hölder constant
S_f	$\max_{\mathbf{x}\in X} \|f(\mathbf{x})\|$
S_{noise}	Output noise bound $\|\varepsilon_i\| \le S_{\text{noise}}$
S	Combined bound $S_f + S_{\text{noise}}$
h	Kernel bandwidth
β	Logit shift
n	Number of support points = prompt length

B Definitions and Lemmas

For the universal approximation theorem (Theorem 2), we decompose the total approximation error by introducing some more intermediate terms (beyond Theorem 1), which we define here. First, to consider the error induced by noise, define the *noiseless* empirical kernel estimator analogously to (3.1):

$$\overline{\mathcal{K}}_h(\mathbf{x}) = \frac{\sum_{i=1}^n \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}_i\|^2}{2h^2}\right) f(\mathbf{x}_i)}{\sum_{i=1}^n \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}_i\|^2}{2h^2}\right)}.$$

We may bound the difference between the noisy and noiseless kernel estimators with the following lemma.

Lemma 1 (Noise term of kernel estimator)**.** *Suppose X satisfies Assumption 1, the support points and query satisfy Assumption 4, and the labels satisfy Assumption 3. Then for any $\mathbf{x} \in X$ and $0 < h \le \tau_X/2$,*

$$\mathop{\mathbb{E}}_{\mathbf{x}_i,\varepsilon_i}\left(\widehat{\mathcal{K}}_h(\mathbf{x}) - \overline{\mathcal{K}}_h(\mathbf{x})\right)^2 = O\left(\frac{S_{\text{noise}}^2}{nh^{d_X}}\right),$$

where $O(\cdot)$ hides dependence on d_X and τ_X.

Proof. Expanding out definitions, notice that the core of the estimator involving $f(\mathbf{x}_i)$ cancels out, and we are only left with the noise contribution:

$$\widehat{\mathcal{K}}_h(\mathbf{x}) - \overline{\mathcal{K}}_h(\mathbf{x}) = \frac{\sum_{i=1}^n \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}_i\|^2}{2h^2}\right) \varepsilon_i}{\sum_{i=1}^n \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}_i\|^2}{2h^2}\right)}.$$

Squaring and conditioning on the support points and query, note that Assumption 3 implies the cross-terms $\mathbb{E}[\varepsilon_i \varepsilon_j]$ vanish, leaving

$$\mathbb{E}\left[\left(\frac{\sum_{i=1}^n \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}_i\|^2}{2h^2}\right)\varepsilon_i}{\sum_{i=1}^n \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}_i\|^2}{2h^2}\right)}\right)^2 \middle| \{\mathbf{x}_i\}_{i=1}^n, \mathbf{x}\right] = \frac{\sum_{i=1}^n \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}_i\|^2}{2h^2}\right)^2 \mathbb{E}(\varepsilon_i^2)}{\left(\sum_{i=1}^n \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}_i\|^2}{2h^2}\right)\right)^2}$$
$$\leq S_{\text{noise}}^2 \cdot \min\left(1, \frac{1}{\sum_{i=1}^n \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}_i\|^2}{2h^2}\right)}\right),$$

by Assumption 3 and since $0 < \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}_i\|^2}{2h^2}\right) \leq 1$.

Each support point $\mathbf{x}_i$ such that $\|\mathbf{x} - \mathbf{x}_i\| \leq h$ will contribute at least $e^{-1/2}$ to the denominator; letting m denote the number of such points, the sum can be bounded as

$$\sum_{i=1}^n \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}_i\|^2}{2h^2}\right) \geq e^{-1/2} \cdot m.$$

We then have

$$\min\left(1, \frac{1}{\sum_{i=1}^n \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}_i\|^2}{2h^2}\right)}\right) \leq \min\left(1, \frac{e^{1/2}}{m}\right) \leq \frac{2e^{1/2}}{1+m}.$$

By Assumption 4 the support points are iid uniform on X (and independent of $\mathbf{x}$), thus when conditioned on the query, m is binomial with n trials and success probability $\mathbb{P}_{\mathbf{t}\sim\mathcal{U}(X)}(\|\mathbf{x} - \mathbf{t}\| \leq h)$. The first inverse moment identity for binomial variables yields

$$\mathbb{E}\left(\frac{2e^{1/2}}{1+m} \mid \mathbf{x}\right) \leq \frac{2e^{1/2}}{(n+1)\mathbb{P}_{\mathbf{t}\sim\mathcal{U}(X)}(\|\mathbf{x}-\mathbf{t}\| \leq h)}.$$

Finally, standard small-ball probability bounds (see, e.g., Niyogi et al. (2008)) give that $\mathbb{P}_{\mathbf{t}\sim\mathcal{U}(X)}(\|\mathbf{x} - \mathbf{t}\| \leq h) = \Omega(h^{d_X})$ for any $\mathbf{x} \in X$ and $0 < h \leq \tau_X/2$, where the hidden constant depends only on d_X and τ_X. Combining the above, we may conclude the result:

$$\mathop{\mathbb{E}}_{\mathbf{x}_i,\varepsilon_i}\left(\widehat{\mathcal{K}}_h(\mathbf{x}) - \overline{\mathcal{K}}_h(\mathbf{x})\right)^2 = O\left(\frac{S_{\text{noise}}^2}{nh^{d_X}}\right)$$

for any $\mathbf{x} \in X$. □

The remaining two components in the error decomposition are the finite-sample gap and kernel regression bias, both of which involve the noiseless *integral* kernel estimator:

$$\mathcal{K}_h(\mathbf{x}) = \frac{\int_X \exp\left(-\frac{\|\mathbf{x}-\mathbf{t}\|^2}{2h^2}\right) f(\mathbf{t})\,\mathrm{d}\mathbf{t}}{\int_X \exp\left(-\frac{\|\mathbf{x}-\mathbf{t}\|^2}{2h^2}\right)\,\mathrm{d}\mathbf{t}}.$$

For these, we recall some results regarding noiseless kernel regression on manifolds, directly adapted from Shen et al. (2026) (specifically, their Lemmas 3 and 2 respectively). The proofs are deferred to their work.

Lemma 2 (Variance of kernel estimator)**.** *Suppose X and f satisfy Assumption 1 and Assumption 2 respectively, and suppose we have support points $\{\mathbf{x}_i\}_{i=1}^n$ and a query $\mathbf{x}$ satisfying Assumption 4. For any $\delta \in (0, 1]$, we have that*

$$\left|\overline{\mathcal{K}}_h(\mathbf{x}) - \mathcal{K}_h(\mathbf{x})\right| = O\left(\left(\log\left(h^{-1}\right)\right)^{3d_X/4} \cdot \sqrt{\frac{\log(4/\delta)}{nh^{d_X}}}\right)$$

with probability at least $1-\delta$, where $h > 0$ is the kernel bandwidth. Here, $O(\cdot)$ hides dependence on d_X, S_f, and τ_X, where the dependency on d_X can be $d_X^{d_X/2}$ in the worst case.

Lemma 3 (Bias of kernel regression on manifold)**.** *Suppose X and f satisfy Assumption 1 and Assumption 2 respectively. Uniformly over X, the integral kernel estimator $\mathcal{K}_h$ for f with bandwidth $h > 0$ satisfies*

$$|\mathcal{K}_h(\mathbf{x}) - f(\mathbf{x})| = O(h^\alpha \log(h^{-1})).$$

The $O(\cdot)$ hides dependence on d_X, L, S_f, and τ_X, with at most exponential dependence on d_X.

Finally, necessary for the theorems in the random settings, we recall here a useful result in random matrix theory regarding the singular values of random matrices with Gaussian entries; for standard references, see Davidson & Szarek (2001) or Vershynin (2026).

Lemma 4 (Nonasymptotic tail bounds on singular values of random Gaussian matrix)**.** *Let $A \in \mathbb{R}^{M\times N}$ with $M \geq N$ be a random matrix with iid $\mathcal{N}(0,1)$ entries. Then for every $t \geq 0$,*

$$\mathbb{P}\left(\sigma_{\min}(A) \geq \sqrt{M} - \sqrt{N} - t\right) \geq 1 - e^{-t^2/2}$$

and

$$\mathbb{P}\left(\sigma_{\max}(A) \leq \sqrt{M} + \sqrt{N} + t\right) \geq 1 - e^{-t^2/2}.$$

We will usually apply this result to wide matrices, noting that a matrix and its transpose have the same singular values. Some of our applications will involve iid $\mathcal{N}(0,\sigma^2)$ entries; in those cases both bounds scale by σ.

C Deferred Proofs

C.1 Proof of Theorem 1

Proof. Expanding the definition of attention (2.2), we first find that the last row of $A(H_{P,\mathbf{x}})$ is

$$A(H_{P,\mathbf{x}})_{n+1} = \frac{\sum_{i=1}^n \exp\left(\frac{1}{\sqrt{d_\mathrm{h}}}\mathbf{x}_{\mathrm{emb}}W_QW_K^T\mathbf{p}_i\right)\mathbf{p}_i^TW_V + \exp\left(\frac{1}{\sqrt{d_\mathrm{h}}}\mathbf{x}_{\mathrm{emb}}W_QW_K^T\mathbf{x}_{\mathrm{emb}}^T\right)\mathbf{x}_{\mathrm{emb}}W_V}{\sum_{i=1}^n \exp\left(\frac{1}{\sqrt{d_\mathrm{h}}}\mathbf{x}_{\mathrm{emb}}W_QW_K^T\mathbf{p}_i\right) + \exp\left(\frac{1}{\sqrt{d_\mathrm{h}}}\mathbf{x}_{\mathrm{emb}}W_QW_K^T\mathbf{x}_{\mathrm{emb}}^T\right)}. \tag{C.1}$$

To get the whole $\widetilde{G}(H_{P,\mathbf{x}})$ we decode the attention output, reading off the last coordinate. Mathematically, this corresponds to right multiplying the above output by the d_h-th standard basis vector $\mathbf{e}_{d_\mathrm{h}} = [0,\ldots,0,1]^T$.

Note that the attention contribution of $\mathbf{x}_{\mathrm{emb}}$ with itself is independent of $\mathbf{p}_i$. Let us temporarily ignore this part, leaving the following output after decoding:

$$\frac{\sum_{i=1}^n \exp\left(\frac{1}{\sqrt{d_\mathrm{h}}}\mathbf{x}_{\mathrm{emb}}W_QW_K^T\mathbf{p}_i\right)\mathbf{p}_i^TW_V\mathbf{e}_{d_\mathrm{h}}}{\sum_{i=1}^n \exp\left(\frac{1}{\sqrt{d_\mathrm{h}}}\mathbf{x}_{\mathrm{emb}}W_QW_K^T\mathbf{p}_i\right)}. \tag{C.2}$$

We recall the kernel estimator (3.1) here:

$$\widehat{\mathcal{K}}_h(\mathbf{x}) = \frac{\sum_{i=1}^n \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}_i\|^2}{2h^2}\right)y_i}{\sum_{i=1}^n \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}_i\|^2}{2h^2}\right)}.$$

We rewrite this in a more convenient form. For each i, the exponents expand as

$$-\frac{\|\mathbf{x}-\mathbf{x}_i\|^2}{2h^2} = -\frac{\|\mathbf{x}\|^2}{2h^2} + \frac{\mathbf{x}^T\mathbf{x}_i}{h^2} - \frac{\|\mathbf{x}_i\|^2}{2h^2},$$

whereupon we may split off the first part from the exponentiation:

$$\exp\left(-\frac{\|\mathbf{x}-\mathbf{x}_i\|^2}{2h^2}\right) = \exp\left(-\frac{\|\mathbf{x}\|^2}{2h^2}\right)\exp\left(\frac{\mathbf{x}^T\mathbf{x}_i}{h^2} - \frac{\|\mathbf{x}_i\|^2}{2h^2}\right).$$

The first term is independent of the index i, thus may be factored out of the numerator and denominator and canceled. This leaves the mathematically equivalent formula of the kernel estimator:

$$\widehat{\mathcal{K}}_h(\mathbf{x}) = \frac{\sum_{i=1}^n \exp\left(\frac{\mathbf{x}^T\mathbf{x}_i}{h^2} - \frac{\|\mathbf{x}_i\|^2}{2h^2}\right)y_i}{\sum_{i=1}^n \exp\left(\frac{\mathbf{x}^T\mathbf{x}_i}{h^2} - \frac{\|\mathbf{x}_i\|^2}{2h^2}\right)},$$

where $y_i = f(\mathbf{x}_i) + \varepsilon_i$.

Comparing the attention output in (C.2) to the above formula, we see that the soft prompt should satisfy two properties. First, we should have $\mathbf{p}_i^T W_V \mathbf{e}_{d_\mathrm{h}} = y_i$ for all i, i.e. the projected values align with the observed target values. $W_V \mathbf{e}_{d_\mathrm{h}}$ is the last column of W_V, which we denote $\mathbf{w}_V$. We rewrite the condition as $\mathbf{w}_V^T \mathbf{p}_i = y_i$.

We also want to match the kernel part, which involves matching the attention logits with the exponents for all i, i.e.

$$\frac{1}{\sqrt{d_\mathrm{h}}} \mathbf{x}_\mathrm{emb} W_Q W_K^T \mathbf{p}_i = \frac{\mathbf{x}^T \mathbf{x}_i}{h^2} - \frac{\|\mathbf{x}_i\|^2}{2h^2}. \tag{C.3}$$

Denote by $W = \frac{1}{\sqrt{d_\mathrm{h}}} W_Q W_K^T$. Note that due to the zero padding in $\mathbf{x}_\mathrm{emb}$, we effectively only consider the first $d_\mathrm{in} + 1$ rows of W. We denote the first d_in rows as $W_{1:d_\mathrm{in}}$, and the $(d_\mathrm{in}+1)$-th row as $W_{d_\mathrm{in}+1}$. The attention logit rewrites as

$$\mathbf{x}_\mathrm{emb} W \mathbf{p}_i = \mathbf{x}^T W_{1:d_\mathrm{in}} \mathbf{p}_i + 1 \cdot W_{d_\mathrm{in}+1} \mathbf{p}_i.$$

We match respective parts with the RHS of (C.3):

$$\mathbf{x}^T (W_{1:d_\mathrm{in}} \mathbf{p}_i) = \frac{\mathbf{x}^T \mathbf{x}_i}{h^2},$$

for which it suffices to enforce

$$W_{1:d_\mathrm{in}} \mathbf{p}_i = \frac{\mathbf{x}_i}{h^2}$$

and

$$W_{d_\mathrm{in}+1} \mathbf{p}_i = -\frac{\|\mathbf{x}_i\|^2}{2h^2}. \tag{C.4}$$

This is the purpose of the homogeneous coordinate: the bias allows us to absorb scalar shifts. It also allows us to handle the contribution of the query's attention with itself, which we have hitherto ignored, using logit shifting—recalibrating attention by artificially inflating certain logits. Adding decoding and rewriting equation (C.1) with our new notation, we get

$$\widetilde{G}(H_{P,\mathbf{x}}) = \frac{\sum_{i=1}^n \exp(\mathbf{x}_\mathrm{emb} W \mathbf{p}_i) \mathbf{p}_i^T \mathbf{w}_V + \exp(\mathbf{x}_\mathrm{emb} W \mathbf{x}_\mathrm{emb}^T) \mathbf{x}_\mathrm{emb} \mathbf{w}_V}{\sum_{i=1}^n \exp(\mathbf{x}_\mathrm{emb} W \mathbf{p}_i) + \exp(\mathbf{x}_\mathrm{emb} W \mathbf{x}_\mathrm{emb}^T)}.$$

Notice the effect of adding a constant $\beta \geq 0$ to the logits which involve the soft prompt tokens:

$$\begin{aligned}
&\frac{\sum_{i=1}^n \exp(\mathbf{x}_\mathrm{emb} W \mathbf{p}_i + \beta) \mathbf{p}_i^T \mathbf{w}_V + \exp(\mathbf{x}_\mathrm{emb} W \mathbf{x}_\mathrm{emb}^T) \mathbf{x}_\mathrm{emb} \mathbf{w}_V}{\sum_{i=1}^n \exp(\mathbf{x}_\mathrm{emb} W \mathbf{p}_i + \beta) + \exp(\mathbf{x}_\mathrm{emb} W \mathbf{x}_\mathrm{emb}^T)} \\
&= \frac{e^\beta \sum_{i=1}^n \exp(\mathbf{x}_\mathrm{emb} W \mathbf{p}_i) \mathbf{p}_i^T \mathbf{w}_V + \exp(\mathbf{x}_\mathrm{emb} W \mathbf{x}_\mathrm{emb}^T) \mathbf{x}_\mathrm{emb} \mathbf{w}_V}{e^\beta \sum_{i=1}^n \exp(\mathbf{x}_\mathrm{emb} W \mathbf{p}_i) + \exp(\mathbf{x}_\mathrm{emb} W \mathbf{x}_\mathrm{emb}^T)} \\
&= \frac{\sum_{i=1}^n \exp(\mathbf{x}_\mathrm{emb} W \mathbf{p}_i) \mathbf{p}_i^T \mathbf{w}_V + \frac{1}{e^\beta} \exp(\mathbf{x}_\mathrm{emb} W \mathbf{x}_\mathrm{emb}^T) \mathbf{x}_\mathrm{emb} \mathbf{w}_V}{\sum_{i=1}^n \exp(\mathbf{x}_\mathrm{emb} W \mathbf{p}_i) + \frac{1}{e^\beta} \exp(\mathbf{x}_\mathrm{emb} W \mathbf{x}_\mathrm{emb}^T)}.
\end{aligned}$$

For large β, the attention contribution of the query with itself is negligible. We may realize this constant shift in the bias term by modifying condition (C.4) to

$$W_{d_\mathrm{in}+1} \mathbf{p}_i = -\frac{\|\mathbf{x}_i\|^2}{2h^2} + \beta.$$

This comes at the cost of a larger (in magnitude) soft prompt; we will formally bound this later. Adding the additional value matching condition $\mathbf{w}_V^T \mathbf{p}_i = y_i$ from before, for each $\mathbf{p}_i$ we are left with a linear system $\widetilde{W} \mathbf{p}_i = \mathbf{v}_i$, where

$$\widetilde{W} = \begin{bmatrix} W_{1:d_\mathrm{in}} \\ W_{d_\mathrm{in}+1} \\ \mathbf{w}_V^T \end{bmatrix} \qquad \text{and} \qquad \mathbf{v}_i = \begin{bmatrix} \frac{\mathbf{x}_i}{h^2} \\ -\frac{\|\mathbf{x}_i\|^2}{2h^2} + \beta \\ y_i \end{bmatrix}. \tag{C.5}$$

By Condition 1, the matrix $\widetilde{W}$ has full row rank, thus the system has at least one solution for any target vector $\mathbf{v}_i$. We construct the minimum-norm solution by the Moore-Penrose pseudoinverse $\widetilde{W}^\dagger$:

$$\mathbf{p}_i^* = \widetilde{W}^\dagger \mathbf{v}_i \coloneqq \widetilde{W}^T (\widetilde{W}\widetilde{W}^T)^{-1} \mathbf{v}_i.$$

Assembling the soft prompt tokens, we get a soft prompt P^* such that

$$\widetilde{G}(H_{P^*,\mathbf{x}}) = \frac{\sum_{i=1}^n \exp\left(\frac{\mathbf{x}^T\mathbf{x}_i}{h^2} - \frac{\|\mathbf{x}_i\|^2}{2h^2}\right) y_i + \frac{1}{e^\beta}\exp(\mathbf{x}_{\text{emb}} W \mathbf{x}_{\text{emb}}^T)\mathbf{x}_{\text{emb}}\mathbf{w}_V}{\sum_{i=1}^n \exp\left(\frac{\mathbf{x}^T\mathbf{x}_i}{h^2} - \frac{\|\mathbf{x}_i\|^2}{2h^2}\right) + \frac{1}{e^\beta}\exp(\mathbf{x}_{\text{emb}} W \mathbf{x}_{\text{emb}}^T)}.$$

It remains to uniformly bound the difference $\left|\widetilde{G}(H_{P^*,\mathbf{x}}) - \widehat{\mathcal{K}}_h(\mathbf{x})\right|$. Factoring out the denominator, the difference simplifies to

$$\begin{aligned}
\left|\widetilde{G}(H_{P^*,\mathbf{x}}) - \widehat{\mathcal{K}}_h(\mathbf{x})\right| &= \frac{\left|\frac{1}{e^\beta}\exp(\mathbf{x}_{\text{emb}} W \mathbf{x}_{\text{emb}}^T)\mathbf{x}_{\text{emb}}\mathbf{w}_V - \widehat{\mathcal{K}}_h(\mathbf{x})\frac{1}{e^\beta}\exp(\mathbf{x}_{\text{emb}} W \mathbf{x}_{\text{emb}}^T)\right|}{\sum_{i=1}^n \exp\left(\frac{\mathbf{x}^T\mathbf{x}_i}{h^2} - \frac{\|\mathbf{x}_i\|^2}{2h^2}\right) + \frac{1}{e^\beta}\exp(\mathbf{x}_{\text{emb}} W \mathbf{x}_{\text{emb}}^T)} \\
&\le \frac{\frac{1}{e^\beta}\left(\left|\exp(\mathbf{x}_{\text{emb}} W \mathbf{x}_{\text{emb}}^T)\right| |\mathbf{x}_{\text{emb}}\mathbf{w}_V| + \left|\widehat{\mathcal{K}}_h(\mathbf{x})\right| \left|\exp(\mathbf{x}_{\text{emb}} W \mathbf{x}_{\text{emb}}^T)\right|\right)}{\sum_{i=1}^n \exp\left(\frac{\mathbf{x}^T\mathbf{x}_i}{h^2} - \frac{\|\mathbf{x}_i\|^2}{2h^2}\right) + \frac{1}{e^\beta}\exp(\mathbf{x}_{\text{emb}} W \mathbf{x}_{\text{emb}}^T)} \\
&\le \frac{1}{e^\beta} \cdot \frac{\left|\exp(\mathbf{x}_{\text{emb}} W \mathbf{x}_{\text{emb}}^T)\right| \left(|\mathbf{x}_{\text{emb}}\mathbf{w}_V| + \left|\widehat{\mathcal{K}}_h(\mathbf{x})\right|\right)}{\sum_{i=1}^n \exp\left(\frac{\mathbf{x}^T\mathbf{x}_i}{h^2} - \frac{\|\mathbf{x}_i\|^2}{2h^2}\right)}
\end{aligned} \tag{C.6}$$

where the first inequality comes from the triangle inequality, and the second from discarding the positive term in the denominator.

We establish bounds on each component. Recall, in particular, Assumptions 1 to 3. It follows from the norm bound on the query and support points that

$$\frac{\mathbf{x}^T\mathbf{x}_i}{h^2} - \frac{\|\mathbf{x}_i\|^2}{2h^2} \ge -\frac{R^2}{h^2} - \frac{R^2}{2h^2} = -\frac{3R^2}{2h^2},$$

from which we may conclude the following bound on the denominator of the kernel estimator:

$$\sum_{i=1}^n \exp\left(\frac{\mathbf{x}^T\mathbf{x}_i}{h^2} - \frac{\|\mathbf{x}_i\|^2}{2h^2}\right) \ge n\exp\left(-\frac{3R^2}{2h^2}\right) > 0.$$

The homogeneous embedding of $\mathbf{x}$ specified in (2.1) satisfies $\|\mathbf{x}_{\text{emb}}\| \le \sqrt{R^2+1}$. We have that

$$|\mathbf{x}_{\text{emb}} W \mathbf{x}_{\text{emb}}^T| \le \|\mathbf{x}_{\text{emb}}\|^2 \|W\| \le (R^2+1)\cdot\|W\|$$

and

$$|\mathbf{x}_{\text{emb}}\mathbf{w}_V| \le \|\mathbf{x}_{\text{emb}}\|\|\mathbf{w}_V\| \le \sqrt{R^2+1}\cdot\|\mathbf{w}_V\|.$$

Exponentiating everywhere gives the corresponding bounds.

Finally, the kernel estimator is a convex combination of target values y_i, thus its absolute value is bounded above by

$$\left|\widehat{\mathcal{K}}_h(\mathbf{x})\right| \le \max_{1\le i\le n} |y_i| \le S.$$

Plugging everything back into (C.6), we arrive at the final bound

$$\sup_{\mathbf{x}\in X}\left|\widetilde{G}(H_{P^*,\mathbf{x}}) - \widehat{\mathcal{K}}_h(\mathbf{x})\right| \le \frac{1}{ne^\beta}\exp\left(\frac{3R^2}{2h^2}\right)\exp\left((R^2+1)\|W\|\right)\left(\sqrt{R^2+1}\|\mathbf{w}_V\| + S\right). \tag{C.7}$$

Squaring both sides and hiding constants gives our desired result:

$$\sup_{\mathbf{x}\in X}\left(\widetilde{G}(H_{P^*,\mathbf{x}}) - \widehat{\mathcal{K}}_h(\mathbf{x})\right)^2 = O\left(\frac{1}{n^2e^{2\beta}}\cdot\exp\left(\frac{3R^2}{h^2}\right)\right).$$

Asymptotically, we have quadratic decay in terms of soft prompt length n, exponential decay in terms of the logit shift β, and exponential growth in inverse kernel bandwidth h.

To bound the norm of the constructed soft prompt tokens, we first derive a uniform bound on the $\mathbf{v}_i$, which were defined in (C.5). By Assumption 1, $\|\mathbf{x}_i\| \le R$, so $\|\mathbf{x}_i/h^2\| \le R/h^2$. Similarly, taking absolute value of the second term yields

$$\left| -\frac{\|\mathbf{x}_i\|^2}{2h^2} + \beta \right| \le \frac{R^2}{2h^2} + \beta,$$

since β, h, and R are nonnegative. Lastly, by definition, $|y_i| \le S$. Together, we have

$$\|\mathbf{v}_i\| \le \sqrt{\frac{R^2}{h^4} + \left(\beta + \frac{R^2}{2h^2}\right)^2 + S^2}$$

for any $1 \le i \le n$.

Since $\widetilde{W} \in \mathbb{R}^{(d_{\text{in}}+2)\times d_{\text{h}}}$ is wide with full row rank (Condition 1), its smallest singular value $\sigma_{\min}(\widetilde{W})$ is guaranteed to be strictly positive. The Moore-Penrose pseudoinverse has spectral norm equal to the reciprocal of this smallest singular value:

$$\|\widetilde{W}^\dagger\| = \frac{1}{\sigma_{\min}(\widetilde{W})}.$$

Submultiplicativity then yields our total bound: for all $1 \le i \le n$,

$$\|\mathbf{p}_i^*\| \le \|\widetilde{W}^\dagger\|\|\mathbf{v}_i\| \le \frac{1}{\sigma_{\min}(\widetilde{W})}\sqrt{\frac{R^2}{h^4} + \left(\beta + \frac{R^2}{2h^2}\right)^2 + S^2} = O(\beta + h^{-2}).$$

□

C.2 Proof of Theorem 2

Proof. Fix a query $\mathbf{x} \in X$ and a set of support points $\{\mathbf{x}_i\}_{i=1}^n$. Inserting intermediate terms leads to the following decomposition:

$$\begin{aligned}
&\left(\widetilde{G}(H_{P^*,\mathbf{x}}) - f(\mathbf{x})\right)^2 \\
&\le 4\underbrace{\left(\widetilde{G}(H_{P^*,\mathbf{x}}) - \widehat{\mathcal{K}}_h(\mathbf{x})\right)^2}_{\text{I}} + 4\underbrace{\left(\widehat{\mathcal{K}}_h(\mathbf{x}) - \overline{\mathcal{K}}_h(\mathbf{x})\right)^2}_{\text{II}} + 4\underbrace{\left(\overline{\mathcal{K}}_h(\mathbf{x}) - \mathcal{K}_h(\mathbf{x})\right)^2}_{\text{III}} + 4\underbrace{\left(\mathcal{K}_h(\mathbf{x}) - f(\mathbf{x})\right)^2}_{\text{IV}}.
\end{aligned}$$

We bound the four terms separately. The first term I is essentially Theorem 1, which was proved by the explicit construction of an appropriate soft prompt. Terms II, III, and IV are bounded by Lemmas 1, 2 and 3 respectively.

To recover minimax-optimal rates, we choose the kernel bandwidth and logit shift as

$$h = n^{-\frac{1}{2\alpha+d_X}} \qquad \text{and} \qquad \beta = \frac{3R^2}{2} \cdot n^{\frac{2}{2\alpha+d_X}}$$

respectively. Since we are proving an asymptotic (in n) bound, we may assume $n > (2/\tau_X)^{2\alpha+d_X}$, so that $0 < h < \min(1/2, \tau_X/2)$.

Recall from (the proof of) Theorem 1, we have

$$\sup_{\mathbf{x}\in X}\left(\widetilde{G}(H_{P^*,\mathbf{x}}) - \widehat{\mathcal{K}}_h(\mathbf{x})\right)^2 = O\left(\frac{1}{n^2 e^{2\beta}} \cdot \exp\left(\frac{3R^2}{h^2}\right)\right).$$

This result holds independently of the specific choice of support points, and the supremum over the query implies the same bound holds in expectation. Plugging in our choice of h and β, the exponents cancel exactly, leaving

$$\mathop{\mathbb{E}}_{\mathbf{x}_i,\mathbf{x}\sim\mathcal{U}(X)}\left(\widetilde{G}(H_{P^*,\mathbf{x}}) - \widehat{\mathcal{K}}_h(\mathbf{x})\right)^2 = O(n^{-2}). \tag{C.8}$$

By Lemma 1, term II is directly bounded as

$$\mathop{\mathbb{E}}_{\mathbf{x}_i,\varepsilon_i}\left(\widehat{\mathcal{K}}_h(\mathbf{x})-\overline{\mathcal{K}}_h(\mathbf{x})\right)^2 = O\left(\frac{S_{\text{noise}}^2}{nh^{d_X}}\right).$$

Substituting the choice for h and hiding S_{noise},

$$\mathop{\mathbb{E}}_{\mathbf{x}_i,\varepsilon_i}\left(\widehat{\mathcal{K}}_h(\mathbf{x})-\overline{\mathcal{K}}_h(\mathbf{x})\right)^2 = O\left(n^{-2\alpha/(2\alpha+d_X)}\right). \tag{C.9}$$

For term III, by Lemma 2,

$$\left|\overline{\mathcal{K}}_h(\mathbf{x})-\mathcal{K}_h(\mathbf{x})\right| = O\left(\left(\log\left(h^{-1}\right)\right)^{3d_X/4}\cdot\sqrt{\frac{\log(4/\delta)}{nh^{d_X}}}\right)$$

with probability at least $1-\delta$, for any $\delta>0$. By definition of the kernel, we have the uniform bound $2S$ (recall $S=S_f+S_{\text{noise}}$), which we may use on the failure probability δ to get an expectation bound:

$$\mathop{\mathbb{E}}_{\mathbf{x}_i,\mathbf{x}\sim\mathcal{U}(X)}\left(\overline{\mathcal{K}}_h(\mathbf{x})-\mathcal{K}_h(\mathbf{x})\right)^2 \le (1-\delta)\cdot O\left(\left(\log\left(h^{-1}\right)\right)^{3d_X/2}\cdot\frac{\log(4/\delta)}{nh^{d_X}}\right)+\delta(2S)^2.$$

Recall that $\alpha\le 1$. We explicitly set $\delta=4h^2$, whereupon the bound simplifies to

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathbf{x}_i,\mathbf{x}}\left(\overline{\mathcal{K}}_h(\mathbf{x})-\mathcal{K}_h(\mathbf{x})\right)^2 &= O\left(\frac{(\log(h^{-1}))^{1+\frac{3d_X}{2}}}{nh^{d_X}}\right)\\ &= O\left(n^{-\frac{2\alpha}{2\alpha+d_X}}(\log n)^{1+\frac{3d_X}{2}}\right)\end{aligned} \tag{C.10}$$

The first equality uses that $0<h<1/2$; the second line follows from actually substituting $h=n^{-\frac{1}{2\alpha+d_X}}$.

Finally, Lemma 3 directly gives us a bound on term IV, which inherently does not involve support points:

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{U}(X)}\left(\mathcal{K}_h(\mathbf{x})-f(\mathbf{x})\right)^2 = O(h^{2\alpha}(\log(h^{-1}))^2).$$

Substituting in our choice of h yields

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{U}(X)}\left(\mathcal{K}_h(\mathbf{x})-f(\mathbf{x})\right)^2 = O\left(n^{-\frac{2\alpha}{2\alpha+d_X}}(\log n)^2\right). \tag{C.11}$$

Assembling (C.8), (C.9), (C.10), and (C.11) together, we arrive at the total bound:

$$\begin{aligned}&\mathop{\mathbb{E}}_{\mathbf{x}_i,\varepsilon_i,\mathbf{x}}(\widetilde{G}(H_{P^*,\mathbf{x}})-f(\mathbf{x}))^2\\ &= O(n^{-2})+O\left(n^{-\frac{2\alpha}{2\alpha+d_X}}\right)+O\left(n^{-\frac{2\alpha}{2\alpha+d_X}}(\log n)^{1+\frac{3d_X}{2}}\right)+O\left(n^{-\frac{2\alpha}{2\alpha+d_X}}(\log n)^2\right).\end{aligned}$$

The first term decays the fastest. For the other three terms, since $d_X\ge 1$, the exponents compare as $1+3d_X/2>2>0$ and the middle term slightly dominates:

$$\mathop{\mathbb{E}}_{\mathbf{x}_i,\varepsilon_i,\mathbf{x}}(\widetilde{G}(H_{P^*,\mathbf{x}})-f(\mathbf{x}))^2 = O\left(n^{-\frac{2\alpha}{2\alpha+d_X}}(\log n)^{1+\frac{3d_X}{2}}\right).$$

The norm bound for the soft prompt tokens follows directly by substituting our choice of h and β into the bound from Theorem 1 and hiding constants. □

C.3 Proof of Theorem 3

Proof. We prove existence first, then derive the corresponding norm bound.

Existence. By Theorem 1, it suffices to show that Assumption 5 implies $\widetilde{W}$ satisfies the full row rank condition of Condition 1 with probability 1.

Recall $\widetilde{W} \in \mathbb{R}^{(d_{\text{in}}+2)\times d_{\text{h}}}$ is defined by vertically concatenating the first $d_{\text{in}}+1$ rows of the query-key matrix $W = W_Q W_K^T/\sqrt{d_{\text{h}}}$ with the last column of the value matrix W_V transposed:

$$\widetilde{W} = \begin{bmatrix} W_{1:d_{\text{in}}} \\ W_{d_{\text{in}}+1} \\ \mathbf{w}_V^T \end{bmatrix}.$$

We first show that the key matrix W_K is invertible almost surely. This is a standard result for Gaussian random matrices (which W_K is, by Assumption 5), but we include a proof for completeness. The determinant of W_K is a polynomial in d_{h}^2 variables, and is clearly not identically zero. Its vanishing locus forms a lower-dimensional algebraic variety, thus has Lebesgue measure zero in $\mathbb{R}^{d_{\text{h}}^2}$. Integrating the joint PDF over the measure zero set yields $\mathbb{P}(\det(W_K) = 0) = 0$, and W_K is invertible almost surely.

Next, consider the wide matrix $W_{Q,1:d_{\text{in}}+1} \in \mathbb{R}^{(d_{\text{in}}+1)\times d_{\text{h}}}$, the first $d_{\text{in}}+1$ rows of W_Q. By Assumption 5, following the same argument as above, any $((d_{\text{in}}+1)\times(d_{\text{in}}+1))$-submatrix has nonzero determinant almost surely. Hence $\operatorname{rank}(W_{Q,1:d_{\text{in}}+1}) = d_{\text{in}}+1$ almost surely.

The transpose of an invertible matrix is invertible, and right-multiplication by an invertible matrix (or a nonzero scalar) preserves row rank. Therefore, we may conclude that the first $d_{\text{in}}+1$ rows of W have rank $d_{\text{in}}+1$ almost surely. Let $\mathcal{V}$ denote that row space.

Finally, W_V is initialized independently of W_Q and W_K, thus its last column $\mathbf{w}_V$ is a random vector independent of $\mathcal{V}$. Noting the dimensions, $\mathcal{V}$ must be a proper subspace of $\mathbb{R}^{d_{\text{h}}}$, thus has Lebesgue measure zero in $\mathbb{R}^{d_{\text{h}}}$. Integrating the respective PDF over a set of measure zero, we find $\mathbb{P}(\mathbf{w}_V \in \mathcal{V} \mid \mathcal{V}) = 0$, thus concatenating $\mathbf{w}_V$ will increase rank almost surely.

Put together, we conclude that $\operatorname{rank}(\widetilde{W}) = d_{\text{in}}+2$ (the condition of Condition 1) with probability 1, after which we may directly invoke the construction with the Moore-Penrose pseudoinverse from Theorem 1, which yields the desired soft prompt and corresponding bounds:

$$\sup_{\mathbf{x}\in X}\left|\widetilde{G}(H_{P^*,\mathbf{x}}) - \widehat{\mathcal{K}}_h(\mathbf{x})\right| \le \frac{1}{ne^{\beta}}\exp\left(\frac{3R^2}{2h^2}\right)\exp\left((R^2+1)\|W\|\right)\left(\sqrt{R^2+1}\|\mathbf{w}_V\|+S\right).$$

A small caveat, as noted above: in the random case, the last two terms which involve $\|W\|$ and $\|\mathbf{w}_V\|$ are no longer constants, but random variables (which grow like $\sqrt{d_{\text{h}}}$, by Lemma 4). Accordingly, we define a new explicit random constant

$$C_G \coloneqq \exp\left((R^2+1)\|W\|\right)\left(\sqrt{R^2+1}\|\mathbf{w}_V\|+S\right), \tag{C.12}$$

which is almost surely finite. With this, we rewrite our error as

$$\sup_{\mathbf{x}\in X}\left|\widetilde{G}(H_{P^*,\mathbf{x}}) - \widehat{\mathcal{K}}_h(\mathbf{x})\right| \le \frac{C_G}{ne^{\beta}}\exp\left(\frac{3R^2}{2h^2}\right).$$

Squaring both sides gives the squared error bound in the theorem.

Norm bound. Recall from the proof of Theorem 1, we have that

$$\|\mathbf{p}_i^*\| \le \|\widetilde{W}^{\dagger}\|\|\mathbf{v}_i\| \le \frac{1}{\sigma_{\min}(\widetilde{W})}\sqrt{\frac{R^2}{h^4}+\left(\beta+\frac{R^2}{2h^2}\right)^2+S^2}.$$

The only component that changes in the random attention case is the smallest singular value $\sigma_{\min}(\widetilde{W})$, which is now a random variable we must bound in probability. Fix $\delta > 0$.

We first consider $W_{1:d_{\text{in}}+1}$, the first $d_{\text{in}}+1$ rows of $\widetilde{W}$. By definition of W, it can be written as

$$W_{1:d_{\text{in}}+1} = \frac{1}{\sqrt{d_{\text{h}}}} W_{Q,1:d_{\text{in}}+1} W_K^T.$$

Note that $W_{Q,1:d_{\text{in}}+1} \in \mathbb{R}^{(d_{\text{in}}+1)\times d_{\text{h}}}$ is a wide Gaussian matrix. Invoking Lemma 4, for any $t > 0$,

$$\mathbb{P}(\sigma_{\min}(W_{Q,1:d_{\text{in}}+1}) \geq \sigma_{W_Q}(\sqrt{d_{\text{h}}} - \sqrt{d_{\text{in}}+1} - t)) \geq 1 - e^{-t^2/2}.$$

Picking $t = \sqrt{2\log(2/\delta)}$ and with $d_{\text{h}} \gg d_{\text{in}}$ (at least $d_{\text{h}} \geq 8(d_{\text{in}} + 1 + 2\log(2/\delta))$), we have

$$\sigma_{\min}(W_{Q,1:d_{\text{in}}+1}) = \Omega\left(\sigma_{W_Q}\sqrt{d_{\text{h}}}\right)$$

with probability at least $1 - \delta/2$.

Conditioned on $W_{Q,1:d_{\text{in}}+1}$, the product $W_{Q,1:d_{\text{in}}+1}W_K^T/\sqrt{d_{\text{h}}}$ is a random matrix with independent Gaussian columns. The final row of $\widetilde{W}$ is $\mathbf{w}_V^T$, which is completely independent of W_Q and W_K. Together, $\widetilde{W}$ has the block-diagonal covariance structure

$$\Sigma := \begin{bmatrix} \frac{\sigma_{W_K}^2}{d_{\text{h}}} W_{Q,1:d_{\text{in}}+1}W_{Q,1:d_{\text{in}}+1}^T & 0 \\ 0 & \sigma_{W_V}^2 \end{bmatrix}.$$

Conditioned on W_Q, we may write (reparameterization trick)

$$\widetilde{W} \overset{d}{=} \Sigma^{1/2} Z,$$

where $Z \in \mathbb{R}^{(d_{\text{in}}+2)\times d_{\text{h}}}$ is a random matrix with iid standard normal entries.

Since $\Sigma^{1/2}$ is square and invertible and Z is wide with full row rank almost surely, supermultiplicativity of the minimum singular value implies

$$\sigma_{\min}(\widetilde{W}) \geq \sigma_{\min}(\Sigma^{1/2})\sigma_{\min}(Z).$$

For block-diagonal matrices, their minimum singular value is simply the minimum of the singular values for each respective block:

$$\sigma_{\min}(\Sigma^{1/2}) = \min\left(\frac{\sigma_{W_K}}{\sqrt{d_{\text{h}}}}\sigma_{\min}(W_{Q,1:d_{\text{in}}+1}), \sigma_{W_V}\right).$$

On the aforementioned high probability event, this reduces to

$$\sigma_{\min}(\Sigma^{1/2}) \geq \min(c_1\sigma_{W_Q}\sigma_{W_K}, \sigma_{W_V})$$

Note $Z \in \mathbb{R}^{(d_{\text{in}}+2)\times d_{\text{h}}}$ is a wide matrix since $d_{\text{h}} \gg d_{\text{in}} + 2$. As above, by Lemma 4 we deduce that

$$\sigma_{\min}(Z) \geq \Omega\left(\sqrt{d_{\text{h}}}\right)$$

with probability at least $1 - \delta/2$.

Union bounding with the first high probability event, we have that with probability at least $1 - \delta$,

$$\sigma_{\min}(\widetilde{W}) \geq \sigma_{\min}(\Sigma^{1/2})\,\sigma_{\min}(Z) = \Omega\left(\min(\sigma_{W_Q}\sigma_{W_K}, \sigma_{W_V})\sqrt{d_{\text{h}}}\right).$$

Putting everything together, we conclude

$$\|\mathbf{p}_i^*\| \;\leq\; \frac{\|\mathbf{v}_i\|}{\sigma_{\min}(\widetilde{W})} \;\leq\; \frac{\sqrt{\dfrac{R^2}{h^4} + \left(\beta + \dfrac{R^2}{2h^2}\right)^2 + S^2}}{\sigma_{\min}(\widetilde{W})} \;=\; O\left(\frac{\beta + h^{-2}}{\sqrt{d_{\text{h}}}}\right)$$

with probability at least $1 - \delta$. □

C.4 Proof of Theorem 4

Proof. We have the same four-term decomposition from Theorem 2. Terms II, III, and IV are inherent to kernel regression, holding independently of the transformer. We use Theorem 3 to bound term I in the random case. Recall (C.7) from the proof of Theorem 1:

$$\sup_{\mathbf{x}\in X}\left|\widetilde{G}(H_{P^*,\mathbf{x}}) - \widehat{\mathcal{K}}_h(\mathbf{x})\right| \leq \frac{1}{ne^{\beta}}\exp\left(\frac{3R^2}{2h^2}\right)\exp\left((R^2+1)\|W\|\right)\left(\sqrt{R^2+1}\|\mathbf{w}_V\| + S\right).$$

In the proof of Theorem 2, we chose $h = n^{-\frac{1}{2\alpha+d_X}}$ and correspondingly set $\beta = \frac{3R^2}{2} \cdot n^{\frac{2}{2\alpha+d_X}}$; the remaining constant was absorbed into the $O(\cdot)$. In the random attention case this constant is a random variable, which we defined as C_G in (C.12), and will in fact depend exponentially on $\sqrt{d_\mathrm{h}}$. To deal with this, we may redefine β to exactly absorb it:

$$\beta = \frac{3R^2}{2} \cdot n^{\frac{2}{2\alpha+d_X}} + \log_+ C_G.$$

We restrict to $\log_+ x \coloneqq \max(\log x, 0)$, the positive part of the logarithm, to avoid edge cases which cause negative β. Substituting in this new β gives the desired $\leq n^{-2}$ bound for term I almost surely. The remaining steps follow the proof of Theorem 2, mutatis mutandis.

Fix $\delta > 0$. The token norm bound follows by inserting h and β into the token bound from Theorem 3, to which we allocate failure probability $\delta/2$. By Lemma 4, with high probability we have that

$$\|W\| \leq \frac{1}{\sqrt{d_\mathrm{h}}} \|W_Q\| \|W_K\| = O(\sigma_{W_Q} \sigma_{W_K} \sqrt{d_\mathrm{h}})$$

and $\|\mathbf{w}_V\| = O(\sigma_{W_V} \sqrt{d_\mathrm{h}})$, whence we find

$$\log_+ \left(\sqrt{R^2+1} \|\mathbf{w}_V\| + S \right) = O(\log(\sigma_{W_V} \sqrt{d_\mathrm{h}})).$$

We distribute the other $\delta/2$ failure probability over these events. Consequently, the change in β only results in an added $O(1)$ term to the norm bound (hiding the variance dependence):

$$\frac{\log_+ C_G}{\sqrt{d_\mathrm{h}}} \leq \frac{(R^2+1)\|W\| + \log_+ \left(\sqrt{R^2+1} \|\mathbf{w}_V\| + S \right)}{\sqrt{d_\mathrm{h}}} = O(1)$$

with probability at least $1 - \delta/2$. The union bound gives the claimed token norm bound with probability at least $1 - \delta$. □

C.5 Proof of Theorem 5

Proof. We closely follow the proof structure of Theorem 1, adapting it to this random setting. Recall that the attention logit for each prompt token $\mathbf{p}_i$ looks like $\mathbf{x}_\mathrm{emb} W \mathbf{p}_i$, where $W = \frac{1}{\sqrt{d_\mathrm{h}}} W_Q W_K^T$. Substituting our new affine embedding defined in (4.1), the logit expands as

$$\mathbf{x}_\mathrm{emb} W \mathbf{p}_i = (\mathbf{x}^T U_E + \mathbf{b}_E^T) W \mathbf{p}_i = \mathbf{x}^T (U_E W) \mathbf{p}_i + \mathbf{b}_E^T W \mathbf{p}_i.$$

Further recall that to mimic the NW kernel estimator (3.1), we want this logit to match $\frac{\mathbf{x}^T \mathbf{x}_i}{h^2} - \frac{\|\mathbf{x}_i\|^2}{2h^2} + \beta$, which yields the following two conditions:

$$U_E W \mathbf{p}_i = \frac{\mathbf{x}_i}{h^2} \qquad \text{and} \qquad \mathbf{b}_E^T W \mathbf{p}_i = -\frac{\|\mathbf{x}_i\|^2}{2h^2} + \beta.$$

We also wanted to match the value part, which is affected by the new decoding. The unnormalized contribution of a prompt token $\mathbf{p}_i$ to the output is $\mathbf{p}_i^T W_V \mathbf{u}_D + b_D$, which suggests the following constraint:

$$(W_V \mathbf{u}_D)^T \mathbf{p}_i = y_i - b_D.$$

We consolidate the conditions into a single system $\widetilde{W}_\mathrm{aff} \mathbf{p}_i = \mathbf{v}_i'$, where

$$\widetilde{W}_\mathrm{aff} = \begin{bmatrix} U_E W \\ \mathbf{b}_E^T W \\ (W_V \mathbf{u}_D)^T \end{bmatrix} \in \mathbb{R}^{(d_\mathrm{in}+2) \times d_\mathrm{h}} \qquad \text{and} \qquad \mathbf{v}_i' = \begin{bmatrix} \frac{\mathbf{x}_i}{h^2} \\ -\frac{\|\mathbf{x}_i\|^2}{2h^2} + \beta \\ y_i - b_D \end{bmatrix} \in \mathbb{R}^{d_\mathrm{in}+2}. \tag{C.13}$$

Like in the proof of Theorem 3, it suffices to show $\widetilde{W}_\mathrm{aff}$ has full row rank almost surely. Since W_Q and W_K are initialized with independent Gaussian entries, W has rank d_h almost surely. Similarly, the distributions of U_E and $\mathbf{b}_E$ imply

$$\begin{bmatrix} U_E \\ \mathbf{b}_E^T \end{bmatrix} \in \mathbb{R}^{(d_\mathrm{in}+1) \times d_\mathrm{h}}$$

has full row rank $d_{\text{in}} + 1$ almost surely, which is preserved under right multiplication by the full-rank W.

Lastly, consider $W_V \mathbf{u}_D$, a nondegenerate (since $\mathbf{u}_D \neq 0$ almost surely) random vector independent of the previous rows of $\widetilde{W}_{\text{aff}}$. The previous rows span a proper subspace of $\mathbb{R}^{d_{\text{h}}}$ (Lebesgue measure 0), thus the probability that $W_V \mathbf{u}_D$ is in the subspace is 0, meaning its inclusion strictly increases the rank.

We conclude the system has at least one solution; we construct P^* using the Moore-Penrose pseudoinverse. It remains to uniformly bound the difference $\left|\widetilde{G}(H_{P^*,\mathbf{x}}) - \widehat{\mathcal{K}}_h(\mathbf{x})\right|$. As we did in (C.6), we expand our attention output with the constructed prompt and split it into individual terms:

$$\left|\widetilde{G}(H_{P^*,\mathbf{x}}) - \widehat{\mathcal{K}}_h(\mathbf{x})\right| \leq \frac{1}{e^\beta} \cdot \frac{\left|\exp(\mathbf{x}_{\text{emb}} W \mathbf{x}_{\text{emb}}^T)\right| \left(|\mathbf{x}_{\text{emb}} W_V \mathbf{u}_D + b_D| + \left|\widehat{\mathcal{K}}_h(\mathbf{x})\right|\right)}{\sum_{i=1}^n \exp\left(\frac{\mathbf{x}^T \mathbf{x}_i}{h^2} - \frac{\|\mathbf{x}_i\|^2}{2h^2}\right)}. \tag{C.14}$$

We retain the bounds on the deterministic parts, i.e.

$$\sum_{i=1}^n \exp\left(\frac{\mathbf{x}^T \mathbf{x}_i}{h^2} - \frac{\|\mathbf{x}_i\|^2}{2h^2}\right) \geq n \exp\left(-\frac{3R^2}{2h^2}\right) \quad \text{and} \quad \left|\widehat{\mathcal{K}}_h(\mathbf{x})\right| \leq S.$$

The norm of the embedded query is bounded as

$$\|\mathbf{x}_{\text{emb}}\| = \|\mathbf{x}^T U_E + \mathbf{b}_E^T\| \leq R\|U_E\| + \|\mathbf{b}_E\|,$$

from which it follows that

$$|\mathbf{x}_{\text{emb}} W \mathbf{x}_{\text{emb}}^T| \leq \|\mathbf{x}_{\text{emb}}\|^2 \|W\| \leq (R\|U_E\| + \|\mathbf{b}_E\|)^2 \|W\|.$$

The decoding part is bounded similarly:

$$|\mathbf{x}_{\text{emb}} W_V \mathbf{u}_D + b_D| \leq \|\mathbf{x}_{\text{emb}}\| \|W_V\| \|\mathbf{u}_D\| + |b_D| \leq (R\|U_E\| + \|\mathbf{b}_E\|) \|W_V\| \|\mathbf{u}_D\| + |b_D|.$$

As before, we define a new variable to summarize the random constants, which is almost surely finite:

$$C_{G,\text{aff}} \coloneqq \exp\left((R\|U_E\| + \|\mathbf{b}_E\|)^2 \|W\|\right) \left((R\|U_E\| + \|\mathbf{b}_E\|) \|W_V\| \|\mathbf{u}_D\| + |b_D| + S\right). \tag{C.15}$$

Though we do not use it here, note that by Lemma 4 we may derive

$$C_{G,\text{aff}} = \exp\left(O(d_{\text{h}}^{1.5})\right) \tag{C.16}$$

with high probability. As before, we will absorb $C_{G,\text{aff}}$ into the logit shift later in the corresponding universal approximation theorem.

Plugging everything into (C.14) and squaring both sides, the final bound simplifies compactly as

$$\sup_{\mathbf{x} \in X} \left(\widetilde{G}(H_{P^*,\mathbf{x}}) - \widehat{\mathcal{K}}_h(\mathbf{x})\right)^2 \leq \frac{C_{G,\text{aff}}^2}{n^2 e^{2\beta}} \cdot \exp\left(\frac{3R^2}{h^2}\right).$$

Norm bound. As in the proof of Theorem 1, the minimum-norm solution satisfies

$$\|\mathbf{p}_i^*\| \leq \|\widetilde{W}_{\text{aff}}^\dagger\| \|\mathbf{v}_i'\| = \frac{\|\mathbf{v}_i'\|}{\sigma_{\min}(\widetilde{W}_{\text{aff}})}.$$

Compared to (C.5), the only coordinate in $\mathbf{v}_i'$ that differs is the last, which is now $y_i - b_D$ (versus just y_i). A standard Gaussian tail bound yields $|b_D| \leq \sigma_{b_D} \sqrt{2\log(10/\delta)}$ with probability at least $1 - \delta/5$. This does not change the asymptotic bound: on this event,

$$\|\mathbf{v}_i'\| \leq \sqrt{\frac{R^2}{h^4} + \left(\beta + \frac{R^2}{2h^2}\right)^2 + \left(S + \sigma_{b_D}\sqrt{2\log(10/\delta)}\right)^2} = O(\beta + h^{-2})$$

for all $1 \leq i \leq n$, where the hidden constant may depend on δ.

For the smallest singular value, we first write

$$M \coloneqq \begin{bmatrix} U_E \\ \mathbf{b}_E^T \end{bmatrix} = \Lambda Z_1 \in \mathbb{R}^{(d_{\text{in}}+1)\times d_{\text{h}}}$$

where $\Lambda \coloneqq \text{diag}(\sigma_{U_E}, \ldots, \sigma_{U_E}, \sigma_{\mathbf{b}_E})$ and $Z_1 \in \mathbb{R}^{(d_{\text{in}}+1)\times d_{\text{h}}}$ has iid standard normal entries. The first $d_{\text{in}}+1$ rows of $\widetilde{W}_{\text{aff}}$ can then be written as

$$MW = \frac{1}{\sqrt{d_{\text{h}}}} MW_Q W_K^T,$$

and the j-th column of MW can be written

$$\frac{1}{\sqrt{d_{\text{h}}}} MW_Q W_{K,j}^T,$$

where $W_{K,j}$ denotes the j-th row of W_K. The $W_{K,j}$ are iid random vectors with covariance $\sigma_{W_K}^2 I_{d_{\text{h}}}$. Similarly, the j-th entry of the last row of $\widetilde{W}_{\text{aff}}$ is $W_{V,j}\mathbf{u}_D$ where $W_{V,j}$ is the j-th row of W_V. These entries are iid from $\mathcal{N}(0, \sigma_{W_V}^2 \|\mathbf{u}_D\|^2)$, independent of W_K. Together, conditioned on M, W_Q, and $\mathbf{u}_D$, the columns of $\widetilde{W}_{\text{aff}}$ are iid Gaussian with block-diagonal covariance

$$\Sigma = \begin{bmatrix} \frac{\sigma_{W_K}^2}{d_{\text{h}}}(MW_Q)(MW_Q)^T & 0 \\ 0 & \sigma_{W_V}^2 \|\mathbf{u}_D\|^2 \end{bmatrix}.$$

Again, we write $\widetilde{W}_{\text{aff}} \overset{d}{=} \Sigma^{1/2} Z$ for $Z \in \mathbb{R}^{(d_{\text{in}}+2)\times d_{\text{h}}}$ with iid standard normal entries. By supermultiplicativity of the minimum singular value in this case and the block-diagonal structure,

$$\sigma_{\min}(\widetilde{W}_{\text{aff}}) \geq \min\left(\frac{\sigma_{W_K}}{\sqrt{d_{\text{h}}}} \sigma_{\min}(MW_Q),\ \sigma_{W_V}\|\mathbf{u}_D\| \right) \cdot \sigma_{\min}(Z). \tag{C.17}$$

It remains to lower bound each random quantity. Conditioning on M, the columns of MW_Q are iid $\mathcal{N}(0, \sigma_{W_Q}^2 MM^T)$, so $MW_Q \overset{d}{=} \sigma_{W_Q}(MM^T)^{1/2} Z_2$ with standard Gaussian $Z_2 \in \mathbb{R}^{(d_{\text{in}}+1)\times d_{\text{h}}}$, giving

$$\sigma_{\min}(MW_Q) \geq \sigma_{W_Q}\, \sigma_{\min}(M)\, \sigma_{\min}(Z_2) \geq \sigma_{W_Q} \min(\sigma_{U_E}, \sigma_{\mathbf{b}_E})\, \sigma_{\min}(Z_1)\, \sigma_{\min}(Z_2).$$

Each of Z_1, Z_2, and Z is a wide standard Gaussian matrix; we apply Lemma 4 individually. This yields

$$\sigma_{\min}(Z_1) = \Omega\left(\sqrt{d_{\text{h}}}\right), \quad \sigma_{\min}(Z_2) = \Omega\left(\sqrt{d_{\text{h}}}\right), \quad \sigma_{\min}(Z) = \Omega\left(\sqrt{d_{\text{h}}}\right),$$

each with probability at least $1 - \delta/5$, where the constants may depend on δ.

Finally, Lemma 4 also gives $\|\mathbf{u}_D\| = \Omega\left(\sigma_{\mathbf{u}_D}\sqrt{d_{\text{h}}}\right)$ with probability at least $1 - \delta/5$. Union bounding over all five events and substituting into (C.17), we find

$$\sigma_{\min}(\widetilde{W}_{\text{aff}}) = \Omega\left(\min\left(\sigma_{W_Q}\sigma_{W_K} \min(\sigma_{U_E}, \sigma_{\mathbf{b}_E}), \sigma_{W_V}\sigma_{\mathbf{u}_D}\right) \cdot d_{\text{h}}\right)$$

with probability at least $1 - \delta$.

Putting the two bounds together, we arrive at the overall bound: with probability at least $1 - \delta$,

$$\|\mathbf{p}_i^*\| \leq \frac{\sqrt{\frac{R^2}{h^4} + \left(\beta + \frac{R^2}{2h^2}\right)^2 + \left(S + \sigma_{b_D}\sqrt{2\log(10/\delta)}\right)^2}}{\sigma_{\min}(\widetilde{W}_{\text{aff}})} = O\left(\frac{\beta + h^{-2}}{d_{\text{h}}}\right).$$

□

C.6 Proof of Theorem 6

Proof. The proof follows the same approach as Theorem 4. In the four-term decomposition from the proof of Theorem 2, the latter three are bounded independently of the transformer, so still hold here.

For term I, we use the above result of Theorem 5:

$$\sup_{\mathbf{x}\in X}\left(\widetilde{G}(H_{P^*,\mathbf{x}}) - \widehat{\mathcal{K}}_h(\mathbf{x})\right)^2 \leq \frac{C_{G,\text{aff}}^2}{n^2 e^{2\beta}} \cdot \exp\left(\frac{3R^2}{h^2}\right).$$

Like we did in Theorem 4, we still set $h = n^{-\frac{1}{2\alpha+d_X}}$, but change β to absorb the random constant defined in (C.15):

$$\beta = \frac{3R^2}{2} \cdot n^{\frac{2}{2\alpha+d_X}} + \log_+ C_{G,\text{aff}}.$$

Substituting these choices, we are left with exactly

$$\sup_{\mathbf{x}\in X} \left(\widetilde{G}(H_{P^*,\mathbf{x}}) - \widehat{\mathcal{K}}_h(\mathbf{x})\right)^2 \le \frac{1}{n^2}$$

almost surely. Assembling the four terms exactly as in Theorem 2, we get our desired bound.

We compute our token bound the same way as in Theorem 4, but the different β changes our conclusion slightly. Fix $\delta > 0$; from Theorem 5 we have that

$$\|\mathbf{p}_i^*\| = O\left(\frac{\beta + h^{-2}}{d_\text{h}}\right)$$

with probability at least $1 - \delta/2$. As mentioned in (C.16), using standard Gaussian norm bounds we find that with probability at least $1 - \delta/2$,

$$C_{G,\text{aff}} = \exp\left(O(d_\text{h}^{1.5})\right),$$

which means

$$\beta = O\left(n^{\frac{2}{2\alpha+d_X}} + d_\text{h}^{1.5}\right).$$

Therefore, with probability at least $1 - \delta$ over the initialization of the transformer weights,

$$\|\mathbf{p}_i^*\| = O\left(\frac{n^{\frac{2}{2\alpha+d_X}}}{d_\text{h}} + \sqrt{d_\text{h}}\right)$$

uniformly over the realizations of the samples. □

D VECTOR-VALUED EXTENSION

We give a sketch of how Theorem 6 and its proof in Appendix C.6 would generalize to the vector-valued case, noting that the same strategy would apply to the other settings as well. Ultimately, we get the same asymptotic rates as before, where the hidden constant now also depends on the output dimension. We first require an analogue of Theorem 5.

Let $f : X \to \mathbb{R}^{d_\text{out}}$ where $d_\text{out} \ge 1$. Suppose $f = (f_1, f_2, \ldots, f_{d_\text{out}})$ where each $f_j : X \to \mathbb{R}$. We now require each f_j to satisfy Assumption 2, i.e. $f_j \in C_L^\alpha(X)$ for $1 \le j \le d_\text{out}$. Our observations are now of the form

$$\mathbf{y}_i = f(\mathbf{x}_i) + \boldsymbol{\varepsilon}_i,$$

and similar to before, we assume noise is independent across samples and from the support/query points. Moreover, we also assume mean-zero and bounded noise: either $\|\boldsymbol{\varepsilon}_i\| \le S_\text{noise}$ or $\|\boldsymbol{\varepsilon}_i\|_\infty \le S_\text{noise}$ almost surely (there would only be a difference in the noise term, which is ultimately dominated anyway).

The only thing that changes in the architecture is we replace the scalar decoder with a vector one:

$$D_\text{aff}(H) = U_D^T H_{n+1}^T + \mathbf{b}_D$$

where $U_D \in \mathbb{R}^{d_\text{h} \times d_\text{out}}$ and $\mathbf{b}_D \in \mathbb{R}^{d_\text{out}}$ have Gaussian entries.

In building the soft prompt, the logit-to-kernel-exponent matching condition stays the same. The value matching condition is now

$$(W_V U_D)^T \mathbf{p}_i = \mathbf{y}_i - \mathbf{b}_D,$$

which leads to the linear systems

$$\underbrace{\begin{bmatrix} U_E W \\ \mathbf{b}_E^T W \\ (W_V U_D)^T \end{bmatrix}}_{\widetilde{W}'_\text{aff} \in R^{(d_\text{in}+1+d_\text{out}) \times d_\text{h}}} \quad \mathbf{p_i} = \begin{bmatrix} \frac{\mathbf{x}_i}{h^2} \\ -\frac{\|\mathbf{x}_i\|^2}{2h^2} + \beta \\ \mathbf{y}_i - \mathbf{b}_D \end{bmatrix}.$$

Additional width is required to accommodate the extra constraints, where we need $d_\text{h} \ge d_\text{in} + 1 + d_\text{out}$ (but again, we consider the wide regime $d_\text{h} \gg d_\text{in} + 1 + d_\text{out}$ for the prompt norm bounds).

The random matrix arguments extend directly, showing that $\widetilde{W}'_{\text{aff}}$ has full row rank almost surely. We take the minimum-norm solution P^* as before, which gives coordinate-wise approximation of the noisy empirical kernel. Picking h and β as before (adjusting the random constant in β appropriately), we get the same rate of convergence for each coordinate; summing up over $1 \le j \le d_{\text{out}}$ yields

$$\mathop{\mathbb{E}}_{\mathbf{x}_i,\boldsymbol{\varepsilon}_i,\mathbf{x}}\left[\|\widetilde{G}(H_{P^*,\mathbf{x}}) - f(\mathbf{x})\|^2\right] = O\left(d_{\text{out}} \cdot n^{-\frac{2\alpha}{2\alpha+d_X}} (\log n)^{1+\frac{3d_X}{2}}\right)$$

conditional on almost every initialization, and d_{out} may be absorbed into the constant to get the same rate as before.

For the prompt norm, we have a bound like

$$\|\mathbf{p}_i^*\| \le \frac{\sqrt{\frac{R^2}{h^4} + \left(\beta + \frac{R^2}{2h^2}\right)^2 + \|\mathbf{y}_i - \mathbf{b}_D\|^2}}{\sigma_{\min}(\widetilde{W}'_{\text{aff}})}.$$

Both $\|\mathbf{y}_i\|$ and $\|b_D\|$ grow on the order of $\sqrt{d_{\text{out}}}$, and for large d_{h} we still have $\sigma_{\min}(\widetilde{W}'_{\text{aff}}) = \Omega(d_{\text{h}})$ with high probability. Thus our prompt tokens have norm

$$\|\mathbf{p}_i^*\| = O\left(\frac{\beta + h^{-2} + \sqrt{d_{\text{out}}}}{d_{\text{h}}}\right) = O\left(\frac{n^{\frac{2}{2\alpha+d_X}}}{d_{\text{h}}} + \sqrt{d_{\text{h}}}\right).$$

E Experimental Details and Additional Numerical Experiments

We first discuss our choice of $\mathcal{N}(0,1)$ initializations for the frozen random transformer parameters.

Remark 2. Since our networks are not trained, we do not use width-dependent initialization schemes which are necessary for gradient-based learning, such as Xavier (Glorot & Bengio, 2010) or Kaiming (He et al., 2015). Note that such initialization variances, typically on the order of $\Theta(1/d_{\text{h}})$, will change the asymptotic prompt norm bounds throughout Sections 4.2 and 4.3. For completeness, we faithfully track the variance dependencies throughout the intermediate derivations, up until the final asymptotic bounds.

We also run and present additional experiments to verify intermediate results, using the same settings as specified in Section 5.

Rank and $\sigma_{\min}$ scaling. Sweeping over hidden dimension d_{h}, we check that the relevant coefficient matrices, $\widetilde{W}$ and $\widetilde{W}_{\text{aff}}$ (defined in (C.5) and (C.13) respectively), have minimum singular values which scale as $\Omega(\sqrt{d_{\text{h}}})$ and $\Omega(d_{\text{h}})$ respectively, and are full rank. The results are shown in Fig. 2.

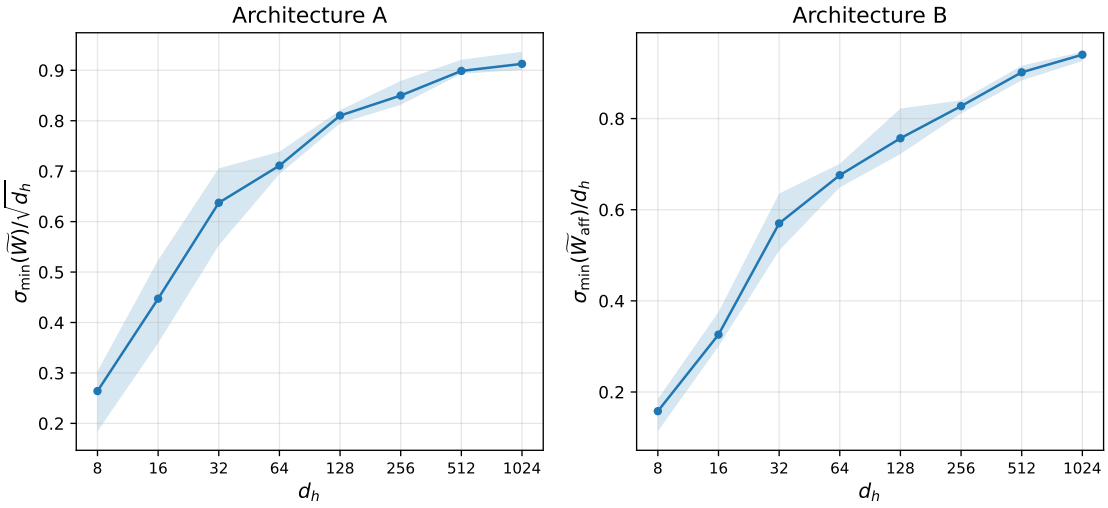
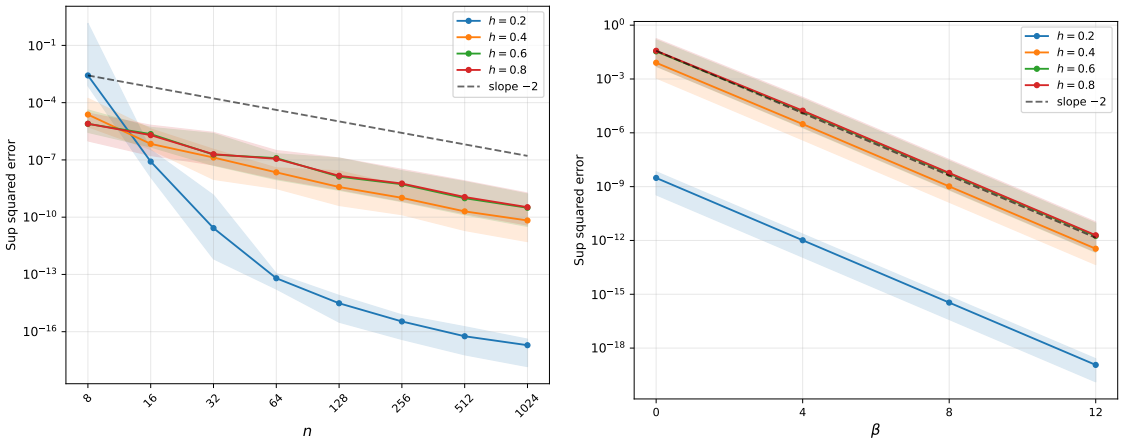


Figure 2: Semi-log plot of normalized minimum singular value versus d_{h} for both architectures. Architecture A normalizes by $\sqrt{d_{\text{h}}}$ and Architecture B normalizes by d_{h}. Solid lines show medians and shaded regions show IQR.

As expected, Fig. 2 exhibits respective normalized minimum singular values concentrating on a positive constant at a shrinking speed. The absence of visible points close to zero supports the almost surely full rank condition.

Kernel approximation. We individually test the main mechanism of our work: the explicit closed-form construction of a soft prompt to approximate the NW kernel estimator, following Theorem 3. Across different kernel bandwidths h, we observe the scaling in prompt length n and logit shift β, reporting supremum squared error over a 2000-point Fibonacci lattice. We test only for Architecture A; the inflated norms of Architecture B cause large fluctuations across seeds which obscure the scaling. The results are shown in Fig. 3.



(a) Log-log plot of error versus sample size/prompt length n, with fixed $\beta = 8$ and $d_{\text{h}} = 256$.

(b) Semi-log plot of error versus logit shift β, with fixed $n = 256$ and $d_{\text{h}} = 256$.

Figure 3: Scaling of supremum squared error between construction and oracle NW estimator across different bandwidths. Solid lines show medians and shaded regions show IQR.

Fig. 3a supports the predicted scaling in n at the rate n^{-2}; we observe an interesting phenomenon for small bandwidth $h = 0.2$, where there is an initial steeper drop for small n before aligning with the predicted rate. This is likely attributable to some queries having insufficient local support points in the small n and h regime, so the query self-attention term in the transformer output dominates. In Fig. 3b, we see the scaling matches almost exactly across all bandwidths (our fixed n is sufficiently large to avoid the aforementioned behavior for $h = 0.2$).

Token norm bound. We report the largest norm of the constructed soft prompt tokens, with results shown in Fig. 4. Again, we only test Architecture A here.

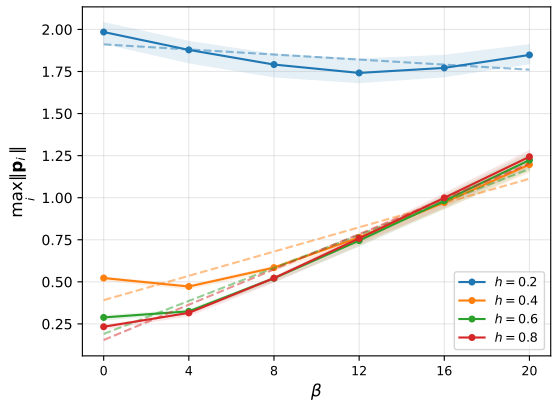


Figure 4: Max token norm versus β across various bandwidths h. Fixed $d_{\text{h}} = 256$ and $n = 256$. Solid lines show medians and shaded regions show IQR.

Fig. 4 verifies the predicted linear norm growth in logit shift; we again observe nonmonotonic behavior with small bandwidth $h = 0.2$, explainable by the cancellation of the particular prompt coordinate $\beta - h^{-2}/2 = \beta - 12.5$. The other bandwidths demonstrate similar behavior but are not as obvious on the plot (for example, $h = 0.4$ dips down at around 4).

Cost of adaptive logit shift. Finally, we test the predicted prompt norm tradeoffs in d_{h}, as specified in Theorems 4 and 6. In particular, we see the impact of the adaptive logit shift between Architecture A, which involves C_G, and Architecture B, which instead uses $C_{G,\mathrm{aff}}$. The results are presented in Fig. 5.

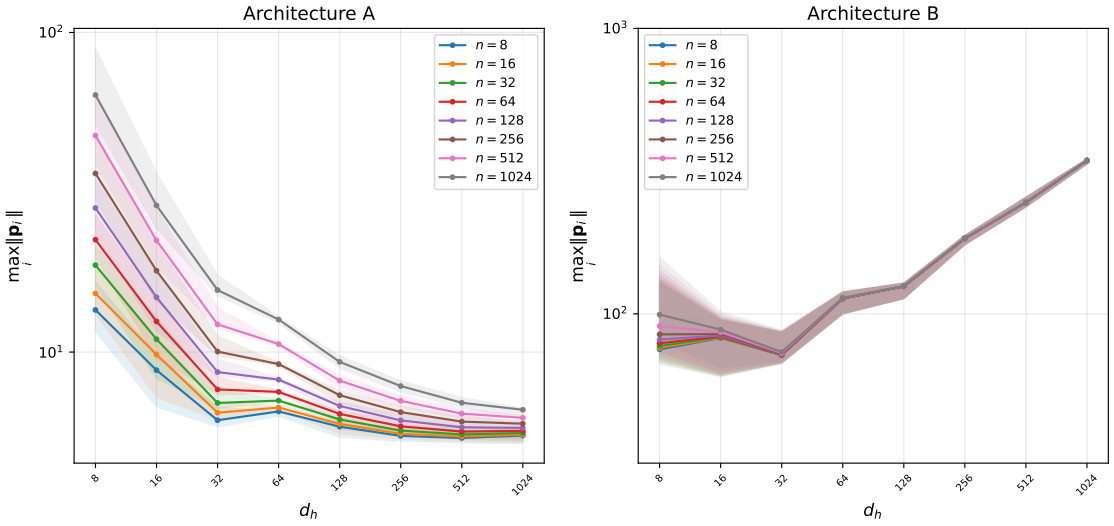


Figure 5: Log-log plots of maximum prompt token norm of constructed soft prompt versus d_{h} for both architectures across various n. We use the same explicitly computed choices for h and β as in Fig. 1. Lines show medians and shaded regions show IQR.

The prompt norm size in Architecture A is essentially monotone decreasing and flattening, aligning with the predicted $O(d_{\mathrm{h}}^{-1/2} + 1)$ rate. In the Architecture B result, the rate of $O(d_{\mathrm{h}}^{-1} + \sqrt{d_{\mathrm{h}}})$ manifests as a U-shaped curve, where the size initially decreases before rising linearly (on the log-log scale) after a certain threshold.