COMS 6998 Theoretical Foundations of LLMs Final Project: Provably Learning the Softmax Kernel of an Attention Head via the Method of Moments

Manuel Paez
map2332@columbia.edu

May 5th, 2025

Abstract

We study the problem of PAC learning the softmax kernel of a single-head attention mechanism, defined by $S(X) = \mathrm{softmax}(X\Theta X^\top)$, in the realizable setting, where $X \in \mathbb{R}^{k\times d}$ consists of a k-length sequence of d-dimensional tokens drawn from a standard Gaussian distribution and $\Theta \in \mathbb{R}^{d\times d}$ is an unknown attention matrix. Using the method of moments, we derive a Hermite polynomial expansion of the exponential kernel $\exp(x_i^\top \Theta x_j)$ and construct a truncated polynomial approximation to the softmax kernel. We present an efficient ℓ_2-regression algorithm that learns the Hermite coefficients of the numerator and denominator of the softmax expression. Under standard non-degeneracy conditions, we show that the learning algorithm has runs in time $\widetilde{O}(\mathrm{poly}(m, k, d, \|\Theta\|_F, \log(1/\epsilon))$ and has sample complexity$m = \widetilde{O}\left(\frac{(d^2\|\Theta\|_F^2 \log(1/\epsilon))^{kd} + \log(1/\delta)}{\epsilon^2}\right)$. Our results provide the first moment-based framework for learning attention kernels from Gaussian data with provable guarantees.

1 Introduction

Multi-head attention layers [BCB14, VSP+17] are central to the transformer architecture, a critical component of state-of-the-art large language models [BCE+23, AAA+23]. Given a sequence length k, attention matrices $\Theta_1, ..., \Theta_m \in \mathbb{R}^{d\times d}$ and projection matrices $W_1, ..., W_m \in \mathbb{R}^{d\times d}$, the multi-head attention layer $F : \mathbb{R}^{k\times d} \to \mathbb{R}^{k\times d}$ takes length-k sequences of d-dimensional tokens $X \in \mathbb{R}^{k\times d}$ via

$$F(X) := \sum_{i=1}^{m} \mathrm{softmax}(X\Theta_i X^T)XW_i \tag{1}$$

where each term $\mathrm{softmax}(X\Theta_i X^T)XW_i$ is referred to as an attention head.

However, the problem of provably learning such attention mechanisms has not been well-studied. [CL24] presents the first theoretical analysis of the problem in the realizable, distribution-specific PAC learning framework, centering its argument around using examples from the Boolean input distribution to sculpt a convex body containing the unknown parameters. In contrast, the *method of moments* [Pea94] - an approach for learning feed-forward networks by estimating correlations between the network output and certain polynomials in the input and setting up a tensor decomposition problem to exploit the moment structure [AGH+14, CLS20, CLLZ22, CN24, DK24] - has not yet been applied to attention mechanisms. The Gaussian distribution is particularly well-suited to moment-based techniques due to its algebraic structure and rotation invariance. We thus ask:

Question 1.1. Can the methods of moments approach be used to PAC learn a single-head attention mechanism from Gaussian examples in the realizable setting?

In this work, we consider the first step of this question by considering the problem of PAC learning the softmax kernel $S(X)$ of a single-head attention layer from Gaussian examples in the realizable setting:

$$S(X) := \text{softmax}(X\Theta X^T) \tag{2}$$

where $\Theta \in R^{d\times d}$ is an unknown attention matrix and $X \in R^{k\times d}$ is a matrix of length k-sequences of tokens of d-dimensions where each input token of $x_i = X_{i,:}$ is drawn i.i.d from a standard d-dimensional Gaussian measure $\gamma = N(0, I_d)$. Given examples $(X^{(1)}, S(X^{(1)})), ..., (X^{(m)}, S(X^{(m)}))$, the learning goal is to output an estimator $\widehat{S}$ such that $\mathbb{E}_{X\sim N(0,I_d)^{\otimes k}}[\|S(X) - \widehat{S}(X)\|_F^2 \leq \epsilon^2$ for some target error ϵ^2. Our main result is as follows:

Theorem 1.2 (Main Algorithmic Result)**.** *Let $S : \mathbb{R}^{d\times k} \to \mathbb{R}^{k\times d}$ be $S(X) = softmax(X\Theta X^T)$ of a single-head attention whose attention matrix Θ is non-degenerate in the sense of 4.1. Then, given at least $m = \frac{(d^2\|\Theta\|_F^2 \log(1/\epsilon))^{kd} + \log(1/\delta)}{\epsilon^2}$ examples $((X^{(1)}, S(X^{(1)})), ..., (X^{(m)}, S(X^{(m)}))$ for each row of $X_{i,:} \sim N(0, I_d)$, this is an algorithm that runs in time in time and with probability $1-\delta$ outputs estimate $\widehat{S}(X)$ that satisfies:*

$$\mathbb{E}_{X\sim N(0,I_d)^{\otimes k}}[\|S(X) - \widehat{S}(X)\|_F^2 \leq \epsilon^2 \tag{3}$$

2 Preliminaries

Given $n \in \mathbb{N}$, let $[n] := \{1, ..., n\}$. We use $x \cdot y$ to denote the standard inner product between $x, y \in \mathbb{R}^d$ and $\|x\|_2$ for the Euclidean norm of a vector $x \in \mathbb{R}^d$. For a matrix $M \in \mathbb{R}^{m\times n}$, let $\|M\|_F$ denote its Frobenius norm and $\|M\|_{op}$ denote its operator norm. We will denote by $\delta_{i,j}$ the Kronecker delta. Throughout the paper, we let $\otimes$ denote the tensor/Kronecker product. For a vector $x \in \mathbb{R}^d$, we denote by $x^{\otimes m}$ the m-th order tensor power of x. Let $\gamma = N(0, I_d)$ denote the standard Gaussian distribution on $\mathbb{R}^d$ with zero mean and identity covariance. Let $L^2(\mathbb{R}^d, \gamma)$ denote the space of square-integrable functions under γ, with inner product $\langle f, g\rangle = \mathbb{E}_{x\sim\gamma}[f(x)g(x)]$.

Hermite Analysis and Expansion The Lebesgue space $L^2(\mathbb{R}^d, \gamma)$ admits a complete orthonormal basis for normalized probabilist's Hermite polynomials, defined as follows:

Definition 2.1 (Normalized Probabilist's Hermite Polynomial (from [O'D14]))**.** For $j \in \mathbb{N}$, the $j - th$ probabilist's Hermite polynomial $H_j : \mathbb{R} \to \mathbb{R}$ are the uni-variate polynomials defined as

$$H_j(z) = \frac{(-1)^j}{p(z)} \cdot \frac{d^j}{dz^j} p(z)$$

where $p(z) = \frac{1}{2\pi} \exp(-z^2/2)$ is the standard Gaussian density. The normalized uni-variate probabilist's Hermite polynomials are $h_j := \frac{1}{\sqrt{j}} H_j$. For a multi-index polynomial $\alpha = (\alpha_1, ..., \alpha_d) \in N^d$, the normalized multivariate Hermite polynomial $h_\alpha : \mathbb{R}^n \to \mathbb{R}$:

$$h_\alpha(z) = \prod_{i=1}^{d} h_{\alpha_j}(z_j) \quad \text{with total degree } |\alpha| = \sum_{j=1}^{d} \alpha_j$$

These polynomials satisfy $\langle h_\alpha, h_\beta \rangle = \delta_{\alpha,\beta}$, forming an orthonormal basis for $L^2(\mathbb{R}^d, \gamma)$.

3 Approximation via Hermite Expansion

We analyze the Hermite polynomial expansion of the kernel function $g(x_i, x_j) := \exp(x_i^T \Theta x_j)$ under Gaussian inputs and use it to approximate the softmax matrix $S(X) = \text{softmax}(X\Theta X^T)$. Let $x_1, ..., x_k$ to be drawn i.i.d from a standard d-dimensional gaussian measure $N(0, I_d)$ and let X be the matrix with rows x_i^T. Then for the matrix $X\Theta X^T \in R^{k\times k}$, $[X\Theta X^T]_{i,j} := X_{i,:}\Theta X_{j,:}^T = x_i^T \Theta x_j$ for $\forall i, j \in [k]$. Each entry of the softmax matrix is:

$$S_{ij}(X) = \frac{\exp(x_i^T \Theta x_j)}{\sum_{l=1}^{k} \exp(x_i^T \Theta x_l)} = \frac{g(x_i, x_j)}{\sum_{l=1}^{k} g(x_i, x_l)}$$

with each row i of $S(X)$ has non-negative entries summing to 1, i.e. $\sum_{l=1}^{k} \exp(x_i^T \Theta x_l) = 1$.

3.1 Hermite Expansion of the Exponential Kernel

Since $g(x_i, x_j) = \exp(x_i^T \Theta x_j)$ is a smooth function of Gaussian inputs, it admits a Hermite expansion:

$$g(x_i, x_j) := \sum_{\alpha,\beta \in \mathbb{N}^d} c_{\alpha,\beta} h_\alpha(x_i) h_\beta(x_j)$$

where the Hermite coefficients $c_{\alpha,\beta}$ are given by :

$$c_{\alpha,\beta} := \mathbb{E}[\exp(x_i^T \Theta x_j) h_\alpha(x) h_\beta(x)] \tag{4}$$

Even Function under the Joint Transformation $g(x_i, x_j) := \exp(x_i^T \Theta x_j)$ is an even function under the joint probability density (x_i, x_j):

$$g(-x_i, -x_j) = \exp((-x_i)^T \Theta(-x_j)) = \exp(x_i \Theta x_j^T) = g(x_i, x_j)$$

We have the following lemma for the coefficients $c_{\alpha,\beta}$:

Lemma 3.1. *Let $x_i, x_j \sim N(0, I_d)$ be i.i.d sampled from the standard Gaussian, and let $\alpha, \beta \in \mathbb{N}^d$ be multi-indices. Define the Hermite coefficient $c_{\alpha,\beta}$ as in 4. If $|\alpha| + |\beta|$ is odd, then $c_{\alpha,\beta} = 0$ and if $|\alpha| + |\beta|$ is even, then:*

$$c_{\alpha,\beta} = \frac{1}{\sqrt{\alpha!\beta!}} \sum_{r_1,...,r_n=1} \sum_{s_1,...,s_n=1} \Big(\prod_{k=1}^{n} \Theta_{a_k,r_k} \Theta_{s_k,b_k} \Big) \cdot \mathbb{E}\Big[\Big(\prod_{k=1}^{n} x_{i,s_k} x_{j,r_k} \exp(x_i^T \Theta x_j) \Big) \Big] \tag{5}$$

where $\alpha = \sum_{k=1}^{n} e_{a_k}$ and $\beta = \sum_{k=1}^{n} e_{a_b}$ are the sums of the standard basis vectors e_{a_k}, e_{b_k} in $\mathbb{R}^d$.

Proof. We proceed with the odd-order coefficients and then the even-order coefficients for $c_{\alpha,\beta}$.

Odd-degree Case: Gaussian Parity The product of the Hermite polynomials satisfy:

$$h_\alpha(-x_i)h_\beta(-x_j) = (-1)^{|\alpha|+|\beta|}h_\alpha(x_i)h_\beta(x_j)$$

It follows that if $|\alpha| + |\beta| = 1 \mod 2$ (odd) and knowing $\exp(x_i^T\Theta x_j)$ is even, then the integration under the Symmetric Gaussian distribution is zero:

$$c_{\alpha,\beta} = \mathbb{E}[\exp(x_i^T\Theta x_j)h_\alpha(x_i)h_\beta(x_j)] = 0$$

Even-degree Case: Generalized Formula and Stein's Lemma Firstly, for $c_{0,0}$:

$$c_{0,0} = \mathbb{E}[\exp(x_i\Theta x_j) = \frac{1}{\sqrt{\det(I - \Theta\Theta^T)}} \quad \text{if } \|\Theta\|_{op} < 1$$

The full proof is in A.1. We now derive a recursive formula for the even-degree coefficients using Stein's lemma. For multi-indices $\alpha, \beta \in \mathbb{N}^d$ with $|\alpha| + |\beta| = 0 \mod 2$, we compute:

$$c_{\alpha,\beta} := \mathbb{E}\left[\exp(x_i^T\Theta x_j)h_\alpha(x_i)h_\beta(x_j)\right]$$

We will first show the recursive formula for $c_{\alpha+e_a,\beta}$. We use the recurrence relation for Hermite polynomials:

$$\sqrt{\alpha_a + 1}h_{\alpha+e_a}(x_i) = x_a h_\alpha(x_i) - \frac{\partial}{\partial x_a}h_\alpha(x_i)$$

Multiplying both sides by $\exp(x_i^T\Theta x_j)h_\beta(x_j)$, taking expectations, and rearranging gives:

$$\sqrt{\alpha_a + 1} \cdot c_{\alpha+e_a,\beta} = \mathbb{E}[x_{i,a}\exp(x_i^T\Theta x_j)h_\alpha(x_i)h_\beta(x_j)] - \mathbb{E}\left[\exp(x_i^T\Theta x_j)\frac{\partial}{\partial x_{i,a}}h_\alpha(x_i)h_\beta(x_j)\right]$$

Now apply Stein's Lemma to the first term:

$$\begin{aligned}\mathbb{E}[x_{i,a} \cdot (\exp(x_i^T\Theta x_j) \cdot h_\alpha(x_i))] &= \mathbb{E}\left[\frac{\partial}{\partial x_{i,a}}(\exp(x_i^T\Theta x_j) \cdot h_\alpha(x_i))\right] \\ &= \mathbb{E}\left[\exp(x_i^T\Theta x_j) \cdot \Big((\Theta x_j)_a \cdot h_\alpha(x_i) + \frac{\partial}{\partial x_{i,a}}h_\alpha(x_i)\Big)\right]\end{aligned}$$

Then for our expression for $c_{\alpha+e_a,\beta}$ we have

$$\begin{aligned}\sqrt{\alpha_a + 1} \cdot c_{\alpha+e_a,\beta} &= \mathbb{E}[\exp(x_i^T\Theta x_j) \cdot \Big((\Theta x_j)_a \cdot h_\alpha(x_i) + \frac{\partial}{\partial x_{i,a}}h_\alpha(x_i)\Big)h_\beta(x_j)] \\ &- \mathbb{E}\left[\exp(x_i^T\Theta x_j)\frac{\partial}{\partial x_{i,a}}h_\alpha(x_i)h_\beta(x_j)\right] \\ &= \mathbb{E}[(\Theta x_j)_a \cdot \exp(x_i^T\Theta x_j)h_\alpha(x_i)h_\beta(x_j)]\end{aligned}$$

Expressing $(\Theta x_j)_a = \sum_{b=1}^d \Theta_{ab}x_{j,b}$, we get:

$$c_{\alpha+e_a,\beta} = \frac{1}{\sqrt{\alpha_a + 1}}\sum_{b=1}^{d}\Theta_{ab} \cdot \mathbb{E}[x_{j,b}\exp(x_i^T\Theta x_j) \cdot h_\alpha(x_i)h_\beta(x_j)] \tag{6}$$

A similar expression for $c_{\alpha,\beta+e_b}$ can be derived by a symmetric argument; see A.2 for the proof:

$$c_{\alpha,\beta+e_b} = \frac{1}{\sqrt{\beta_b+1}} \sum_{a=1}^{d} \Theta_{ab} \cdot \mathbb{E}[x_{i,a} \exp(x_i^T \Theta x_j) \cdot h_\alpha(x_i) h_\beta(x_j)] \tag{7}$$

Given the recurrence formulas, we can now inductively compute all even-degree Hermite coefficients $c_{\alpha,\beta}$ using lower terms. Set $\alpha = \sum_{k=1}^{n} e_{a_k}$ and $\beta = \sum_{k=1}^{n} e_{b_k}$ so that $|\alpha| = |\beta| = n$ and the total degree is $|\alpha| + |\beta| = 2n$ for $n \in \mathbb{N}$. By induction on n, we want to show 5.
Base Case ($n = 0$): For $\alpha = \beta = 0$, the coefficient is $c_{0,0} = \mathbb{E}[\exp(x_i^T \Theta x_j)]$ matches 5.
Inductive Step: Assume 5 holds for multi-indices α', β' of degree $n-1$. We will now show that the formula holds for $\alpha = \alpha' + e_{a_n}, \beta = \beta' = e_{b_n}$. Using the recurrences:

$$\sqrt{\alpha_{a_n}} c_{\alpha,\beta'} = \sum_{r=1}^{d} \Theta_{a_n r} \, \mathbb{E}[x_{j,r} \exp(x_i^T \Theta x_j) \cdot h_\alpha(x_i) h_{\beta'}(x_j)]$$

$$\sqrt{\beta_{b_n}} c_{\alpha',\beta} = \sum_{s=1}^{d} \Theta_{s b_n} \, \mathbb{E}[x_{j,r} \exp(x_i^T \Theta x_j) \cdot h_{\alpha'}(x_i) h_\beta(x_j)$$

we combine both recursions to obtain:

$$\sqrt{\alpha!\beta!} c_{\alpha,\beta} = \sum_{r=1}^{d} \sum_{s=1}^{d} \Theta_{a_n r} \Theta_{s b_n} \cdot \mathbb{E}[x_{i,s} x_{j,r} \exp(x_i^T \Theta x_j) \cdot h_{\alpha'}(x_i) h_{\beta'}(x_j)]$$

By inductive hypothesis: $h_{\alpha'}(x_i) h_{\beta'}(x_j) = \sum_{r_1,\ldots,r_{n-1}} \sum_{s_1,\ldots,s_{n-1}} c' \cdot \left(\prod_{k=1}^{n-1} x_{i,s_k} x_{j,r_j} \right)$ with appropriate constants c'. Therefore, via recursion, we have that

$$c_{\alpha,\beta} = \frac{1}{\sqrt{\alpha!\beta!}} \sum_{r_1,\ldots,r_n=1}^{d} \sum_{s_1,\ldots,s_n=1}^{d} \left(\prod_{k=1}^{n} \Theta_{a_k,r_k} \Theta_{s_k,b_k} \right) \cdot \mathbb{E}\left[\left(\prod_{k=1}^{n} x_{i,s_k} x_{j,r_k} \right) \exp(x_i^T \Theta x_j) \right] \tag{8}$$

□

3.2 Truncation Approximation

We define the degree-N Hermite truncation $\widehat{g}^{(N)}(x_i, x_j)$ of $g(x_i, x_j)$ as

$$\widehat{g}^{(N)}(x_i, x_j) = \sum_{|\alpha|+|\beta| \le N} c_{\alpha,\beta} h_\alpha(x_i) h_\beta(x_j)$$

Let $\widehat{S}^{(N)}(X) \in \mathbb{R}^{k \times k}$ be the degree-N Hermite truncation of $S(X)$ with entries:

$$\widehat{S}_{ij}^{(N)}(X) := \frac{\widehat{g}^{(N)}(x_i, x_j)}{\sum_{l=1}^{k} \widehat{g}^{(N)}(x_i, x_l)}$$

The following theorem bounds the expected Frobenius error between the softmax matrix and its Hermite approximation.

Theorem 3.2. *Let the matrix $X \in R^{k \times d}$ be the k-sequence of d-dimensional tokens $x_1, \ldots, x_k \sim_{i.i.d}$*

$N(0, I_d)$. Let $S(X) = softmax(X\Theta X^T)$ and let $\widehat{S}(X) = \widehat{S}^{(N)}(X)$ be the degree-N Hermite truncation as above. Then for any $\epsilon > 0$, if $N = \Theta\big((d^2\|\Theta\|_F^2) \cdot \log(1/\epsilon)\big)$, we have

$$\mathbb{E}_{X\sim\gamma^{\otimes k}}\left[\|S(X) - \widehat{S}^{(N)}(X)\|_F^2\right] \leq \epsilon^2$$

We will show a proof sketch of the problem. The full proof is in Appendix B.6.

Proof Sketch. The proof sketch is as follows: we aim to bound:

$$\|S(X) - \widehat{S}^{(N)}(X)\|_F^2 = \sum_{i,j=1}^{k}\Big(\frac{g(x_i, g_j)}{Z_i} - \frac{\widehat{g}^{(N)}(x_i, x_j)}{\widehat{Z}_i^{(N)}}\Big)^2$$

where $g(x_i, x_j) = \exp(x_i^T\Theta x_j)$, $\widehat{g}^{(N)}(x_i, x_j)$ is g's degree-N Hermite truncated expansion, and $Z_i = \sum_{l=1}^{k} g(x_i, x_l)$, $\widehat{Z}_i^{(N)}$ is Z_i's degree -N Hermite truncated expansion. Using the inequality B.2:

$$\left|\frac{a}{b} - \frac{\widehat{a}}{\widehat{b}}\right| \leq 2\Big(\frac{|a-\widehat{a}|^2}{b^2} + \frac{a^2|\widehat{b}-\widehat{b}|^2}{b^2\widehat{b}^2}\Big)$$

we apply this to $a = g(x_i, x_j)$, $\widehat{a} = \widehat{g}^{(N)}(x_i, x_j)$, $b = Z_i, \widehat{b} = \widehat{Z}_i^{(N)}$. We get:

$$\|S(X) - \widehat{S}^{(N)}(X)\|_F^2 \leq 2\sum_{i,j=1}^{k}\left(\frac{(g(x_i, x_j) - \widehat{g}^{(N)}(x_i, x_j))^2}{Z_i^2} + \frac{g(x_i, x_j)^2(Z_i - \widehat{Z}_i^{(N)})^2}{Z_i^2 \cdot \widehat{Z}_i^{(N)2}}\right)$$

We know bound terms. From B.5, the point-wise approximation error is $|g(x_i, x_j) - \widehat{g}^{(N)}(x_i, x_j)| \leq \epsilon_N$ with $\epsilon_N^2 = O\Big(\frac{(d\|\Theta\|_F)^{2(N+1)}}{((N+1)!)^2}\Big)$. The partition function error satisfies $|Z_i - \widehat{Z}_i^{(N)}| \leq k\epsilon_N$. We also set $|x_i^T\Theta x_j| \leq M = O(\log(k))$. Using this, we have that each term is bounded by at most $O(\exp(6M)\epsilon_N^2/k^2$, and summing over k^2 terms gives:

$$\|S(X) - \widehat{S}^{(N)}(X)\|_F^2 \leq C \cdot \epsilon_N^2, \quad \text{where } C = O(\exp(6M))$$

To ensure $\|S(X) - \widehat{S}^{(N)}(X)\|_F^2 \leq \epsilon^2$, we require that $C \cdot \frac{(d\|\Theta\|_F)^{2(N+1)}}{(N+1)!} \leq \epsilon^2$ for some constant $C > 0$. Solving for N gives $N = \Theta(d^2\|\Theta\|_F^2 \log(1/\epsilon))$. This finishes the proof. □

4 Learning Algorithm

We now present a learning algorithm that estimates the softmax kernel $S(X) = \text{softmax}(X\Theta X^T)$ by learning the Hermite expansions of both the numerator and denominator using least-squares regression over Gaussian examples.

4.1 Assumptions

We assume the following throughout this section: 1. each input $x \in \mathbb{R}^{k\times d}$ consists of k tokens $x_1, x_2, ..., x_k$ which are sampled i.i.d from the standard gaussian $N(0, I_d)$, 2. the attention matrix $\Theta \in \mathbb{R}^{d\times d}$ satisfies $\|\Theta\|_{op} < 1$ to ensure convergence of the Hermite expansion. We also use the assumptions in C.3 for the sample complexity.

4.2 Learning via ℓ_2 regression

We aim to learn the Hermite coefficient tensor c_p and c_q for the numerator and denominator of

$$S_{ij}(X) \approx \frac{p^{(N)}(x_i, x_j)}{q^{(N)}(x_1, .., x_k)}$$

Let $p^{(N)}(x_i, x_j)$ be the degree-N Hermite truncated polynomial that approximates $\exp(x_i^T \Theta x_j)$:

$$p^{(N)}(x_i, x_j) = \sum_{|\alpha|+|\beta| \leq N} c_p[\alpha, \beta] \cdot h_\alpha(x_i) h_\beta(x_j) \approx \exp(x_i^T \Theta x_j)$$

with $c_p[\alpha, \beta]$ being the Hermite coefficients for $\exp(x_i^T \Theta x_j)$. Let $q^{(N)}(x_1, ..., x_k) = q^{(N)}(X)$ be the degree-N Hermite truncated polynomial that approximates $\sum_{l=1}^{k} \exp(x_i^T \Theta x_l)$:

$$q^{(N)}(x_1, ..., x_k) = \sum_{\sum_l |\lambda_l| \leq N} c_q[\lambda_1, ..., \lambda_k] \cdot \prod_{i=1}^{k} h_{\lambda_w}(x_w) \approx \sum_{l=1}^{k} \exp(x_i^T \Theta x_l)$$

with $c_q[\lambda_1, ..., \lambda_k]$ being the Hermite coefficients for $\sum_{l=1}^{k} \exp(x_i^T \Theta x_l)$. The algorithm is as follows:

4.3 Analysis of Algorithm 4.2

Firstly, the feature dimensions are $D_p = O(N^{2d})$ and $D_q = O(N^{kd})$ (See Appendix C.1 for analysis).

Runtime Analysis For Hermite feature construction, each sample $X^{(a)}$, computing $\phi_p^{(N)}(x_i, x_j)$ takes $O(k^2 D_p)$ times and computing $\phi_q^{(N)}(x_1, ..., x_k)$ takes $O(D_p)$. Then for all m samples, we have $O(m(k^2 D_p + D_q))$. For constructing the design matrix, we build mk^2 vectors in $v_{i,j}^{(a)} \in \mathbb{R}^{D_p + D_q}$, so it takes $O(mk^2(D_p + D_q))$. Lastly, the least squared solvers, solving $\min_{w \in \mathbb{R}^{D_p+D_q}} \|Vw\|_2^2$ has equations $V^T V$ constructed in runtime of $O(mk^2(D_p+D_q)^2)$ and solving them takes $O((D_p+D_q)^3)$, so runtime is $O(mk^2(D_p + D_q)^2 + (D_p + D_q)^3)$. Therefore, total runtime is

$$O(mk^2 D_p + mk^2(D_p + D_q) + mk^2(D_p + D_q)^2 + (D_p + D_q)^3)$$

From C.2, by plugging in $D_p = O(N^{2d})$, $D_q = O(N^{kd})$ and $N = Cd^2 \|\Theta\|_F^2 \log(1/\epsilon)$ for some constant $C > 0$, we obtain the total runtime to be

$$\widetilde{O}\Big(mk^2 \cdot \Big(d^2 \|\Theta\|_F^2 \log\Big(\frac{1}{\epsilon}\Big)\Big)^{2kd} + \Big(d^2 \|\Theta\|_F^2 \log\Big(\frac{1}{\epsilon}\Big)^{3kd}\Big)\Big) \tag{9}$$

Sample Complexity See Appendix A.2 for the sample complexity analysis. By standard generalization bounds for linear regression, and assuming 1) the features ϕ_p, ϕ_q are bounded (as Hermite polynomials are sub-exponential under Gaussians), and true coefficients c_p, c_q lie in a bounded norm ball (since Hermite coefficients decay exponentially), we obtain a uniform convergence guarantee: with probability $\geq 1 - \delta$, the empirical solutions satisfies:

$$\mathop{\mathbb{E}}_{X \sim \gamma^{\otimes k}} \big[S(X) - \widehat{S}^{(N)}(X)\big] \leq \frac{1}{m} \sum_{a=1}^{m} \|S(X^{(a)}) - \widehat{S}^{(N)}(X^{(a)})\|_F^2 + \widetilde{O}\Big(\frac{D_p + D_q}{m}\Big)$$

Algorithm 1 ℓ_2 algorithm for Softmax Kernel Approximation

```
Require: Samples $\{X^{(a)}, S(X^{(a)})\}_{a=1}^m$ where input $X^{(a)} = [x_1^{(a)}, ..., x_k^{(a)}] \in R^{k \times d}$ is i.i.d sampled from the standard Gaussian $x_i^{(a)} \sim N(0, I_d)$ and labels $S^{(a)}(X) = \text{softmax}(X^{(a)}\Theta X^{(a)T})$.
Require: Truncation degree $N = \Theta(d^2\|\Theta\|_F^2 \log(1/\epsilon))$ and non-degeneracy assumptions.
Ensure: Learn Hermite coefficients vectors $(c_p^*, c_q^*) \in \mathbb{R}^{D_p} \times \mathbb{R}^{D_q}$ for the numerator and denominator Hermite polynomial to learn estimator $\widehat{S}_{ij}(X) := \widehat{S}_{ij}^{(N)}(X) = \frac{p^{(N)}(x_i,x_j)}{q^{(N)}(x_1,...,x_k)} \quad \forall i, j \in [k]$.
  1. Construct Hermite Features: for each sample $a \in [m]$ and each $i, j \in [k]$, we define pairwise features $\phi_p(x_i^{(a)}, x_j^{(a)})$:
       $\phi_p(x_i^{(a)}, x_j^{(a)}) := [h_\alpha(x_i^{(a)})h_\beta(x_j^{(a)})] \in \mathbb{R}^{D_p}$
     and joint features:
       $\phi_q(x_1^{(a)}, ..., x_k^{(a)}) := \big[\prod_{l=1}^{k} h_{\lambda_w}(x_w^{(a)})\big]_{\sum_{w=1}^k |\lambda_w| \le N} \in \mathbb{R}^{D_q}$
     where $D_p, D_q$ are the respective feature dimensions.
  2. Define coefficient vectors $c_p \in \mathbb{R}^{D_p}, c_q \in \mathbb{R}^{D_q}$. Then define $p^{(N)}(x_i^{(a)}, x_j^{(a)}) = \langle c_p, \phi_p^{(a)}(x_i, x_j)\rangle$ and $q^{(N)}(x_1^{(a)}, ..., x_k^{(a)}) = \langle c_q, \phi_q(x_1^{(a)}, ..., x_k^{(a)})\rangle$.
  3. To enable linear regression, define for all $a \in [m], i, j \in [k]$:
       $v_{i,j}^{(a)} := \begin{bmatrix} -\phi_q(x_i^{(a)}, x_j^{(a)}) \\ S_{i,j} \cdot \phi_q(x_1^{(a)}, ..., x_k^{(a)}) \end{bmatrix} \in R^{D_p + D_q} \quad \text{and} \quad w := \begin{bmatrix} c_p \\ c_q \end{bmatrix} \in R^{D_p + D_q}$
     where $v_{i,j}^{(a)}$ is the final feature vector and $w$ is the combined coefficient vector.
  4. Set $\epsilon_{i,j}^{(a)} = \langle v_{i,j}^{(a)}, w\rangle = S_{ij}^{(a)} \cdot q^{(N)}(X^{(a)} - p^{(N)}(x_i i^{(a)}, x_j^{(a)}$. Define the least squared objective $\mathcal{L}(w)$:
       $\mathcal{L}(w) := \frac{1}{mk^2} \sum_{a=1}^{m} \sum_{i,j=1}^{k} \big(\epsilon_{i,j}^{(a)}\big)^2 = \frac{1}{mk^2}\|Vw\|_2^2$
     where $V \in \mathbb{R}^{mk^2 \times (D_p + D_q)}$ which stacks all the $v_{i,j}^{(a)}$ rows for $a = 1, ..., m$. Solve for $w^*$
       $w^* := \begin{bmatrix} c_p* \\ c_q* \end{bmatrix} = \arg\min_{w \in R^{D_p + D_q}} \mathcal{L}(w)$
  5. For new input $X \in \mathbb{R}^{k \times d}$, evaluate for all $i, j \in [k]$:
       $\widehat{S}_{ij}(X) = \frac{\langle c_p^*, \phi_p^{(N)}(x_i, x_j)\rangle}{\langle c_q^*, \phi_q^{(N)}(x_1, ..., x_k)\rangle}$
  6. Form $\widehat{S}(X) := \big[\widehat{S}_{ij}(x)\big]_{i,j=1}^k \in R^{k \times k}$
```

To ensure total error $\leq \epsilon^2$, set sample size: $m = \widetilde{O}\Big(\frac{d^2\|\Theta\|_F^2 \log(1/\epsilon))^{kd}+\log(1/\delta)}{\epsilon^2}\Big)$. which controls both truncation and generalization error.

4.3.1 Proof of Correctness

We will prove 1.2 will show that the algorithm under the realizable PAC learning setting outputs a hypothesis $\widehat{S}$ such that

$$\mathop{\mathbb{E}}_{X\sim\gamma^{\otimes k}}\left[\|S(X)-\widehat{S}(X)\|_F^2\right] \leq \epsilon^2$$

where $S(X) = \text{softmax}(X\Theta X^T) \in \mathbb{R}^{k\times k}$ and $\Theta \in \mathbb{R}^{d\times d}$ unknown.

Proof. The proof is broken down into four steps.

Step 1: Approximation via Truncated Hermite Expansion From 3.2, we know that the Hermite expansion truncated at degree $N = \Theta(d^2\|\Theta\|_F^2 \log(1/\epsilon))$ satisfies:

$$\mathbb{E}\left[\|S(X)-\widehat{S}^{(N)}(X)\|_F^2\right] \leq \epsilon^2$$

Thus, if we had access to the exact coefficients (c_p, c_q), the approximation error of the softmax kernel would be at most ϵ.

Step 2: Realizable Case and Polynomial Parameterization Under this approximation, each entry of the softmax is approximated by a ratio of two Hermite polynomials: $S_{ij}(X) \approx \frac{p^{(N)}(x_i,x_j)}{q^{(N)}(x_1,\ldots,x_k)}$. Both $p^{(N)}$ and $q^{(N)}$ are degree-N Hermite polynomials parameterized by coefficient vectors c_p and c_q. In the realizable case, this approximation is exact, and there exists a linear relationship:

$$S_{ij}\cdot q^{(N)}(x_1,..,x_k) = p^{(N)}(x_i,x_j)$$

which is used in the Algorithm provided that $q^{(N)}(x_1,..,x_k) \neq 0$.

Step 3: Least Square Recovery The algorithm constructs feature vectors ϕ_p and ϕ_q, and define a joint features $v_{i,j}^{(a)} \in R^{D_p+D_q}$ to ensure that

$$\langle v_{i,j}^{(a)}, w\rangle = S_{i,j}^{(a),N}\cdot q^{(a)N}(x_1,...,x_k) - p^{(a)N}(x_i,x_j)$$

It solves the least-squares objective via ℓ_2 regression: $L(w) = \frac{1}{mk^2}\sum_{a=1}\sum_{i,j=1}^{k}\left(\langle v_{i,j}^{(a)}, w\rangle\right)^2 = \frac{1}{mk^2}\|Vw\|_2^2$. In the realizable case, there exists a solution $w^* = (c_p*, c_q*)$ such that this loss is zero:

$$\begin{aligned} w^* &= \arg\min_{w\in\mathbb{R}^{D_p+D_q}} \frac{1}{mk^2}\|Vw\|_2^2 = \text{argmin}_w \|Vw\|_2^2 = \arg\min_w w^T V^T V w \\ &= \text{null space minimizer of } V \end{aligned}$$

Step 4: Generalization Bounds As shown in 4.3, using standard generalization bounds for linear regression and assuming assumptions of the features and true coefficients, we obtain the number of samples $m = \widetilde{O}\Big(\frac{(d^2\|\Theta\|_F^2 \log(1/\epsilon)^{kd}+\log(1/\delta)}{\epsilon^2}\Big)$ to ensure that $\mathbb{E}[(S(X)-\widehat{S}(X))^2] \leq \epsilon^2$.

Therefore, the algorithm correctly learns a low-error approximation $\widehat{S}(X)$ to the softmax kernel $\mathrm{softmax}(X\Theta X^T)$ under Gaussian inputs, in realizable settings, with provable bounds on sample complexity and runtime. □

5 Conclusion and Future Work

We presented a provable algorithm for learning the softmax kernel of an attention head using Hermite moment expansions and ℓ_2 regression. Under Gaussian inputs, the method achieves efficient runtime and sample complexity with formal guarantees. Extending this framework to multi-head attention remains open. This may require tensor decomposition techniques or convex formulations that disentangle multiple kernels jointly.

References

[AAA+23] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report (2023). *arXiv preprint arXiv:2303.08774*, 2023.

[AGH+14] Animashree Anandkumar, Rong Ge, Daniel J Hsu, Sham M Kakade, Matus Telgarsky, et al. Tensor decompositions for learning latent variable models. *J. Mach. Learn. Res.*, 15(1):2773–2832, 2014.

[BCB14] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[BCE+23] Sébastien Bubeck, Varun Chadrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.

[BM02] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[CL24] Sitan Chen and Yuanzhi Li. Provably learning a multi-head attention layer. *arXiv preprint arXiv:2402.04084*, 2024.

[CLLZ22] Sitan Chen, Jerry Li, Yuanzhi Li, and Anru R Zhang. Learning polynomial transformations. *arXiv preprint arXiv:2204.04209*, 2022.

[CLS20] Sitan Chen, Jerry Li, and Zhao Song. Learning mixtures of linear regressions in subexponential time via fourier moments. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 587–600, 2020.

[CN24] Sitan Chen and Shyam Narayanan. A faster and simpler algorithm for learning shallow networks. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 981–994. PMLR, 2024.

[DK24] Ilias Diakonikolas and Daniel M Kane. Efficiently learning one-hidden-layer relu networks via schurpolynomials. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 1364–1378. PMLR, 2024.

[O'D14] Ryan O'Donnell. *Analysis of boolean functions.* Cambridge University Press, 2014.

[Pea94] Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

A Full Results for 3.1

A.1 $c_{0,0}$ Hermite calculation

We derive an expression for $c_{0,0}$ that matches the general formula for even coefficients. Given $h_0(x_i) = h_0(x_j) = 1$, we have that

$$c_{0,0} = \mathbb{E}[\exp(x_i^T \Theta x_j)]$$

We condition on x_i and compute the inner expectation over x_j: $\mathbb{E}[\exp(x_i^T \Theta x_j)] = \mathbb{E}_{x_i}[\mathbb{E}_{x_j}[\exp(x_i^T \Theta x_j)|x_i]]$. For $\mathbb{E}_{x_j}[\exp(x_i^T \Theta x_j)|x_i]$, we have

$$\mathbb{E}_{x_j}[\exp(x_i^T \Theta x_j)|x_i] = \exp\big(\frac{1}{2}\|\Theta^T x_i\|^2\big) = \exp\big(\frac{1}{2} x_i^T \Theta\Theta^T x_i\big)$$

Then, taking the expectation over x_i, we have

$$\begin{aligned}\mathbb{E}_{x_i}\big[\exp\big(\frac{1}{2}x_i^T\Theta\Theta^T x_i\big)\big] &= \int_{\mathbb{R}^d} \exp\big(-\frac{1}{2}x_i^T(I-\Theta\Theta^T)x_i\big)\cdot\frac{1}{(2\pi)^{d/2}}\mathrm{d}x_i \\ &= \frac{1}{\sqrt{\det(I-\Theta\Theta^T)}} \quad \text{for } I-\Theta\Theta^T \succ 0\end{aligned}$$

Under the condition $\|\Theta\|_{op} < 1$, which ensures $I - \Theta\Theta^T \succ 0$, we have that

$$c_{0,0} = \mathbb{E}[\exp(x_i \Theta x_j) = \frac{1}{\sqrt{\det(I-\Theta\Theta^T)}}$$

A.2 Symmetric Argument for $c_{\alpha,\beta+e_b}$

For $c_{\alpha,\beta+e_b}$, we can use the following recurrence formula:

$$\sqrt{\beta_b+1}h_{\beta+e_b}(x) = x_b h_\beta(x) - \frac{\partial}{\partial x_b}h_\beta(x)$$

Multiplying the formula by $\exp(x_i^T\Theta x_j)h_\alpha(x_i)$, applying expectation, and rearranging, we have that:

$$\sqrt{\beta_b+1}\cdot c_{\alpha,\beta+e_b} = \mathbb{E}[x_{j,b}\exp(x_i^T\Theta x_j)h_\alpha(x_i)h_\beta(x_j)] - \mathbb{E}\left[\exp(x_i^T\Theta x_j)h_\alpha(x_i)\frac{\partial}{\partial x_{j,b}}h(x_j)\right]$$

Applying Stein's Lemma, we have that

$$\begin{aligned}\mathbb{E}[x_{j,b} \cdot (\exp(x_i^T \Theta x_j) \cdot h_\beta(x_j)] &= \mathbb{E}\left[\frac{\partial}{\partial x_{j,b}} h(x_j)(\exp(x_i^T \Theta x_j) \cdot h_\beta(x_j))\right] \\ &= \mathbb{E}\left[\exp(x_i^T \Theta x_j) \cdot \Big((\Theta^T x_i)_b h_\beta(x_j) + \frac{\partial}{\partial x_{j,b}} h_\beta(x_j)\Big)\right]\end{aligned}$$

Then for our expression for $c_{\alpha,\beta+e_b}$ we have

$$\begin{aligned}\sqrt{\beta_b+1} \cdot c_{\alpha,\beta+e_b} &= \mathbb{E}\left[\exp(x_i^T \Theta x_j) h_\alpha(x_i)\Big((\Theta^T x_i)_b h_\beta(x_j) + \frac{\partial}{\partial x_{j,b}} h_\beta(x_j)\Big)\right] \\ &- \mathbb{E}\left[\exp(x_i^T \Theta x_j) h_\alpha(x_i) \frac{\partial}{\partial x_{j,b}} h(x_j)\right] \\ &= \mathbb{E}[(\Theta^T x_i)_b \exp(x_i^T \Theta x_j) h_\alpha(x_i) \cdot h_\beta(x_j)]\end{aligned}$$

Using $(\Theta^T b) = \sum_{a=1}^d \Theta_{ab} x_{i,a}$, then $c_{\alpha,\beta+e_b}$ is:

$$c_{\alpha,\beta+e_b} = \frac{1}{\sqrt{\beta_b+1}} \sum_{a=1}^d \Theta_{ab} \cdot \mathbb{E}[x_{i,a} \exp(x_i^T \Theta x_j) \cdot h_\alpha(x_i) h_\beta(x_j)]$$

A.3 Potential Tensor Formulation for the Hermite Expansion

We devised a tensor formulation for $g(x_i, x_j)$ in case this work is extended beyond the scope of the class. We define the Hermite tensor as follows:

Definition A.1 (Normalized Hermite Tensor)**.** . Let $H^{(n)}(x) \in \left(\mathbb{R}^d\right)^{\otimes n}$ be the n-order normalized multivariate Hermite tensor such that

$$H^{(n)}(x) := \sum_{|\alpha|=n} \frac{1}{\sqrt{\alpha!}} H_\alpha(x) \cdot e_\alpha$$

where e_α is the canonical basis tensor.

We will require a few properties that follow from this definition. We will need to know that the entries of $H^{(n)}(x)$ form a useful Fourier basis of $L_2(\mathbb{R}^d, N(0, I_d))$. In particular, for non-negative integers m and n, we have that $\mathbb{E}_{x\sim N(0,I_d)}[H^{(n)}(x) \otimes h^{(m)}(x)]$ is 0 is $m \neq n$ and $Sym_n(I_{d^n})$ if $m = n$, where Sym_n is the symmetrization operation over the first n coordinates. From this we conclude that if T is a symmetric n-tensor, then $\mathbb{E}_{x\sim N(0,I_d)}[\langle H^{(n)}(x), T\rangle H^{(m)}(x)]$ is 0 if $m \neq n$ and T if $m = n$.

For $|\alpha| + |\beta| = 2n$, define the tensor $T^{(2n)} \in \left(\mathbb{R}^d\right)^{\otimes 2n}$ be the coefficient tensor with entries $c_{\alpha,\beta}$:

$$T^{(2n)} = \mathbb{E}[\exp(x_i^T \Theta x_j) \cdot H^{(n)}(x_i) \otimes H^{(n)}(x_j)] = \langle T^{(2n)}, e_\alpha \otimes e_\beta\rangle$$

which is a fully symmetric order $-2n$ tensor. We obtain

$$g(x_i, x_j) = \exp(x_i^T \Theta x_j) = \sum_{|\alpha|+|\beta|=2n} c_{\alpha,\beta} h_\alpha(x_i) h_\beta(x_j) = \sum_{n=0}^\infty \langle T^{2n}, H^{(n)}(x_i) \otimes H^{(n)}(x_j)\rangle$$

With the N-order Hermite truncation of $\exp(x_i^T\Theta x_j)$ being

$$\widetilde{g}^{(N)}(x_i,x_j)=\sum_{|\alpha|+|\beta|=2n}^{N}c_{\alpha,\beta}h_\alpha(x_i)h_\beta(x_j)=\sum_{n=0}^{N}\langle T^{2n},H^{(n)}(x_i)\otimes H^{(n)}(x_j)\rangle$$

B Results for Lemma 3.2

B.1 Sub-Exponential Distribution of $x_i^T\Theta x_j$

Given $x_1,...,x_k\sim_{i.i.d}N(0,I_d)$ and a fixed matrix $\Theta\in\mathbb{R}^{d\times d}$. $x_i^T\Theta x_j$ can be expressed as $x_i^T\Theta x_j=\sum_{a,b=1}^{d}\Theta_{a,b}(x_i)_a(x_j)b$ which is quadratic for $i=j$, and bilinear for $i\neq j$. For the bilinear case $(i\neq j)$, $x_i^T\Theta x_j$ is a sum of independent random variables $(x_i)_a(x_j)_b$ with weights Θ_{ab}. Each $(x_i)_a(x_j)_b$ is such that

$$\mathbb{E}[(x_i)_a(x_j)_b]=\mathbb{E}[(x_i)_a]\mathbb{E}[(x_j)_b]=0$$

$$\textbf{Var}[(x_i)_a(x_j)_b]=\mathbb{E}[((x_i)_a(x_j)_b)^2]=1\quad\text{since}\quad\mathbb{E}[(x_i)_a^2]=\mathbb{E}[(x_i)_b^2]=1$$

and thus

$$\textbf{Var}[x_i^T\Theta x_j]=\sum_{a,b}\Theta_{ab}^2=\|\Theta\|_F^2$$

For higher moments, since $(x_i)_a(x_j)_b\sim$ standardized product-normal, it is sub-exponential:

$$P(|(x_i)_a(x_j)_b|>t)\leq 2\exp(-ct)\quad\text{for some constant }c>0$$

Thus $x_i^T\Theta x_j$ is a weighted sum of independent sub-exponential random variable $\sim\|\Theta\|_{op}$ as each term $\Theta_{a,b}(x_i)_a(x_j)_b$ has sub-exponential norm $\sim|\Theta_{a,b}|$

B.2 Quotient Perturbation Inequality

Let $a,\widehat{a},b,\widehat{b}\in\mathbb{R}$ with $a,\widehat{a},b,\widehat{b},\neq 0$. Given $\frac{a}{b}-\frac{\overline{a}}{\overline{b}}$, we can expand it to obtain:

$$\frac{a}{b}-\frac{\overline{a}}{\overline{b}}=\frac{a\overline{b}-\overline{a}b}{b\overline{b}}=\frac{\overline{b}(a-\overline{a})+\overline{a}(\overline{b}-b)}{b\overline{b}}$$

For $(\overline{b}(a-\overline{a})+\overline{a}(\overline{b}-b))^2$, we have

$$(\overline{b}(a-\overline{a})+\overline{a}(\overline{b}-b))^2=(a-\overline{a})^2\overline{b}^2+2(a-\overline{a})\overline{a}\overline{b}(\overline{b}-b)+\overline{a}^2(\overline{b}-b)^2$$

Since AM-GM says $2ab\leq a^2+b^2$

$$\leq 2\overline{b}^2(a-\overline{a})^2+2\overline{a}^2(b-\overline{b})^2$$

With this, $|\frac{a}{b}-\frac{\overline{a}}{\overline{b}}|^2$ is bounded by

$$\begin{aligned}|\frac{a}{b}-\frac{\overline{a}}{\overline{b}}|^2&\leq\frac{2\overline{b}^2|a-\overline{a}|^2}{b^2\overline{b}^2}+\frac{2\overline{a}^2|b-\overline{b}|^2}{b^2\overline{b}^2}\\&=2\Big(\frac{|a-\overline{a}|^2}{b^2}+\frac{\overline{a}^2|b-\overline{b}|^2}{b^2\overline{b}^2}\Big)\end{aligned}$$

B.3 Concentration Inequality to bound $M = \max_{i,j} \exp(x_i^T \Theta x_j)$

Instead of $M = O(\log k)$ fade away as in the proof of Theorem 3.2 B.6, one can find a high-probability bound for M as follows:

Definition B.1 (Bernstein's Inequality)**.** For a sum of independent sub-exponential random variables with variance σ^2, and sub-exponential parameter b, we have:

$$\mathbb{P}(|z| \geq t) \leq 2\exp\Big(-c\Big(\frac{t^2}{\sigma^2} \wedge \frac{t}{b}\Big)\Big)$$

for an absolute constant $c > 0$

For our case, recall that $x_i^T \Theta x_j$ have a sub-exponential distribution with $\sigma^2 = \|\Theta\|_F^2$ with $b \lesssim \|\Theta\|_{op}$. Applying Bernsteins' Inequality and we obtain:

$$\mathbb{P}(|x_i^T \Theta x_j| \geq t) \leq 2\exp\Big(-c\Big(\frac{t^2}{\|\Theta\|_F^2} \wedge \frac{t}{\|\Theta\|_{op}^2}\Big)\Big)$$

Via union bound, we have

$$\mathbb{P}(\max_{i,j}|x_i^T \Theta x_j| \geq t) = \mathbb{P}(\exists i,j : |x_i^T \Theta x_j| \geq t) \leq 2k^2 \exp\Big(-c\Big(\frac{t^2}{\|\Theta\|_F^2} \wedge \frac{t}{\|\Theta\|_{op}^2}\Big)\Big)$$

We want to bound the RHS with $\leq \delta$ and let $t = M$.

$$\begin{aligned}
2k^2 \exp\Big(-c\Big(\frac{M^2}{\|\Theta\|_F^2} \wedge \frac{M}{\|\Theta\|_{op}^2}\Big)\Big) &\leq \delta \\
-c\Big(\frac{M^2}{\|\Theta\|_F^2} \wedge \frac{M}{\|\Theta\|_{op}^2}\Big) + \log(2k^2) &\leq \log\delta \\
\Big(\frac{M^2}{\|\Theta\|_F^2} \wedge \frac{M}{\|\Theta\|_{op}^2}\Big) + \log(2k^2) &\geq \frac{1}{c}\big(\log(2k^2) - \log(\delta)\big)
\end{aligned}$$

For the case where $M \leq \frac{\|\Theta\|_F^2}{\|\Theta\|_{op}}$:

$$M \geq \|\Theta\|_F \sqrt{\frac{1}{c}\log(2k^2/\delta)}$$

For the case where $M \geq \frac{\|\Theta\|_F^2}{\|\Theta\|_{op}}$:

$$M \geq \|\Theta\|_{op}\frac{1}{c}\log(2k^2/\delta)$$

Thus, with probability at least $1 - \delta$, we have:

$$M = \max_{i,j}|x_i^T \Theta x_j| \leq C\Big(\|\Theta\|_F\sqrt{\log(2k^2/\delta)} + \|\Theta\|_{op}\log(2k^2/\delta)\Big)$$

for some universal constant $C > 0$. If we use the assumption that $\|\Theta\|_{op} < 1$, we have that

$$M \leq O\Big(\|\Theta\|_F\sqrt{\log(2k^2/\delta)}\Big)$$

Assumption or not, this guarantees that $g(x_i, x_j), \widehat{g}^{(N)}(x_i, x_j) \in [\exp(-M), \exp(M)]$ and $Z_i, \widehat{Z}_i \in$

$[k\exp(-M), k\exp(M)]$. This was derived for future extensions of this project.

B.4 Bound for $|Z_i - \widehat{Z}_i^{(N)}|$

Consider $|Z_i - \widehat{Z}_i^{(N)}|$ and Cauchy-Schwartz Inequality: $(\sum_{j=1}^{k}|E_{i,j}|)^2 \leq k\sum_{j=1}^{k}|E_{ij}|^2$ We can rewrite this as

$$|Z_i - \widehat{Z}_i^{(N)}|^2 = \left|\sum_{l=1}^{k}(g(x_i,x_j) - \widehat{g}^{(N)}(x_i,x_j)\right|^2 \leq k\sum_{l=1}^{k}\left|g(x_i,x_j) - \widehat{g}^{(N)}(x_i,x_j)\right|^2$$

B.5 Bound for ϵ_N

Let us devise an upper bound for ϵ_N^2 where

$$\epsilon_N = \sup_{x_i,x_j\in\mathbb{R}^n}|\exp(x_i^T\Theta x_j) - \widehat{g}^{(N)}(x_i,x_j)\rangle|$$

We know that

$$\begin{aligned}\epsilon_N^2 = \sup_{x_i,x_j\in\mathbb{R}^n}|\exp(x_i^T\Theta x_j) - \widehat{g}^{(N)}(x_i,x_j)\rangle|^2 &= \sup_{x_i,x_j\in\mathbb{R}^n}\left|\exp(x_i^T\Theta x_j) - \sum_{|\alpha|+|\beta|\leq N}c_{\alpha,\beta}h_\alpha(x_i)h_\beta(x_j)\right|^2\\ &\leq \sum_{n=N+1}^{\infty}\sum_{|\alpha|=|\beta|=n}\frac{c_{\alpha,\beta}^2}{\alpha!\beta!}\end{aligned}$$

For $c_{\alpha,\beta}$ is the Hermite coefficients of $g(x_i,x_j) = \exp(x_i^T\Theta x_j)$. We know that for any $x_i, x_j \sim_{i.i.d} N(0, I_d)$, we have

$$\exp(x_i^T\Theta x_j) = \sum_{n=0}^{\infty}\frac{(x_i^T\Theta x_j)^n}{n!}$$

is sub-exponential from B.1. From the Hermite expansion properties for smooth functions, for $|\alpha| = |\beta| = n$, we have

$$|c_{\alpha,\beta}| \lesssim (C\|\Theta\|_F)^n$$

for some constant $C = C(d)$. Thus, $c_{\alpha,\beta}^2 \lesssim \frac{(C\|\Theta\|_F)^{2n}}{\alpha!\beta!}$ The number of multi-indices for pairs $(\alpha,\beta) \in \mathbb{N}^d \times \mathbb{N}^d$ with $|\alpha| = |\beta| = n$ is about

$$\#(\text{multi-indices } \alpha \text{ and } \beta) = \binom{n+d-1}{n}^2 \sim O(d^{2n})$$

Thus, we have that

$$\begin{aligned}\epsilon_N^2 = \sum_{n=N+1}^{\infty}\sum_{|\alpha|=|\beta|=n}\frac{c_{\alpha,\beta}^2}{\alpha!\beta!} &\lesssim \sum_{n=N+1}^{\infty}d^{2n}\cdot\frac{(C\|\Theta\|_F)^{2n}}{(n!)^2}\\ &= \sum_{n=N+1}^{\infty}\frac{(Cd\|\Theta\|_F)^{2n}}{(n!)^2}\end{aligned}$$

with C being a constant depending on dimension d and the smoothness of f. Apply Sterling's Inequality:

Definition B.2 (Stirling's Inequality)**.** $\forall n \geq 1$:

$$\sqrt{2\pi n}\Big(\frac{n}{e}\Big)^n \leq n! \leq e\sqrt{n}\Big(\frac{n}{e}\Big)^n$$

for $(n!)^2 \geq \left(\frac{n}{e}\right)^{2n}$, we have

$$\frac{(Cd\|\Theta\|_F)^{2n}}{(n!)^2} \leq (Cd\|\Theta\|_F)^{2n} \cdot \Big(\frac{n}{e}\Big)^n = (e^2C^2d^2\|\Theta\|_F^2)^n \cdot n^{-2n}$$

Because the terms decay rapidly due to n^{-2n}, the sum is dominated by its first term at $n = N+1$. Thus,

$$\epsilon_N^2(e^2C^2d^2\|\Theta\|_F^2)^{N+1}(N+1)^{-2(N+1)}$$

Therefore, we have

$$\epsilon_N^2 = O\Big(\frac{(d\|\Theta\|_F)^{2(N+1)}}{(N+1)!}\Big)$$

B.6 Proof of Theorem 3.2

Proof. The goal is to bound the approximation error for $\|S(X) - \widehat{S}(X)\|_F^2$. Let $S(X) = \text{softmax}(X\Theta X^T)$ and $\widehat{S}(X) = \widehat{S}^{(N)}(X)$ be the degree-N Hermite truncation expansion of $S(X)$. Let $g(x_i, x_j) = \exp(x_i^T\Theta x_j)$ and $\widehat{g}^{(N)}(x_i, x_j)^{(N)}$ be its degree-N Hermite truncation expansion of g. Let $Z_i = \sum_{l=1}^k g(x_i, x_l)$, and $\widehat{Z}_i^{(N)} = \sum_{l=1}^k \widehat{g}^{(N)}(x_i, x_l)$.

We can express $\|S(X) - \widehat{S}(X)\|_F^2$ as follows:

$$\begin{aligned}\|S(X) - \widehat{S}(X)\|_F^2 = \sum_{i,j=1}^k |S_{ij} - \widehat{S}_{ij}^{(N)}| &= \sum_{i,j=1}^k \Big(\frac{g(x_i, x_j)}{\sum_{l=1}^k g(x_i, x_l)} - \frac{\widehat{g}^{(N)}(x_i, x_j)}{\sum_{l=1}^k \widehat{g}^{(N)}(x_i, x_l)}\Big)^2 \\ &\leq 2\sum_{i,j=1}^k \Big(\frac{|g(x_i, x_j) - \widehat{g}^{(N)}(x_i, x_j)|^2}{Z_i^2} + \frac{\widehat{g}^{(N)}(x_i, x_j)^2}{Z_i^2\widehat{Z}_i^{(N)2}}|Z_i - \widehat{Z}_i^{(N)}|\Big)\end{aligned}$$

The last inequality derives from B.2. Consider $\forall x_i, x_j \sim_{i.i.d} N(0, I_d)$ we have $|x_i^T\Theta x_j| \leq M$, then $g(x_i, x_j), \widehat{g}^{(N)}(x_i, x_j) \in [\exp(-M), \exp(M)]$ which implies that $Z_i, \widehat{Z}_i^{(N)} \in [k\exp(-M), k\exp(M)]$. For the first term $\frac{|g(x_i, x_j) - \widehat{g}^{(N)}(x_i, x_j)|^2}{Z_i^2}$, we have

$$\frac{|g(x_i, x_j) - \widehat{g}^{(N)}(x_i, x_j)|^2}{Z_i^2} \leq \frac{|g(x_i, x_j) - \widehat{g}^{(N)}(x_i, x_j)|}{k^2\exp(-2M)} = \frac{\exp(2M)|g(x_i, x_j) - \widehat{g}^{(N)}(x_i, x_j)|}{k^2}$$

For the second term $\frac{\widehat{g}^{(N)}(x_i, x_j)^2}{Z_i^2\widehat{Z}_i^{(N)2}}|Z_i - \widehat{Z}_i^{(N)}|$, we have

$$\frac{\widehat{g}^{(N)}(x_i, x_j)^2}{Z_i^2\widehat{Z}_i^{(N)2}}|Z_i - \widehat{Z}_i^{(N)}| \leq \frac{\exp(2M)|Z_i - \widehat{Z}_i^{(N)}|}{k^4\exp(-4M)} = \frac{\exp(6M)|Z_i - \widehat{Z}_i^{(N)}|}{k^4}$$

Thus, $\|S-\widehat{S}|_F^2$ is upper bounded by

$$\begin{aligned}\|S-\widehat{S}|_F^2 &\leq 2\sum_{i=1}^{k}\sum_{j=1}^{k}\Big(\frac{\exp(2M)|g(x_i,x_j)-\widehat{g}^{(N)}(x_i,x_j)|}{k^2}+\frac{\exp(6M)|Z_i-\widehat{Z}_i^{(N)}|}{k^4}\Big)\\ &=2\Big(\Big(\frac{\exp(2M)}{k^2}+\frac{\exp(6M)}{k^3}\Big)\sum_{i=1}^{k}\sum_{j=1}^{k}|g(x_i,x_j)-\widehat{g}^{(N)}(x_i,x_j)|\Big)\\ &=2\Big(\frac{\exp(2M)}{k^2}+\frac{\exp(6M)}{k^3}\Big)\|G(X)-\widehat{G}^{(N)}(X)\|_F^2\end{aligned}$$

where $G(X):=\exp(X\Theta X^T)$ and $\widehat{G}^{(N)}(X)$ is the degree-N Hermite truncation of $G(X)$. Let us now bound $\|G(X)-\widehat{G}^{(N)}(X)\|_F^2$. Given $\widehat{g}^{(N)}(x_i,x_j)=\sum_{|\alpha|+|\beta|\leq N}c_{\alpha,\beta}h_\alpha(x_i)h_\beta(x_j)$, we can bound $\|G-\widehat{G}^{(N)}(x_i,x_j)\|_F^2$ with

$$\begin{aligned}\|G-\widehat{G}^{(N)}\|_F^2&=\sum_{i=1}^{k}\sum_{j=1}^{k}\big(g(x_i,x_j)-\sum_{|\alpha|+|\beta|\leq N}c_{\alpha,\beta}h_\alpha(x_i)h_\beta(x_j)\big)^2\\ &\leq k^2\cdot\epsilon_N^2\end{aligned}$$

with ϵ_N defined as:

$$\epsilon_N=\sup_{x_i,x_j\in\mathbb{R}^n}\Big|\exp(x_i^T\Theta x_j)-\sum_{|\alpha|+|\beta|\leq N}c_{\alpha,\beta}h_\alpha(x_i)h_\beta(x_j)\Big|$$

We obtain the upper bound $\epsilon_N^2\lesssim\frac{C(d\|\Theta\|_F)^{2(N+1)}}{((N+1)!)}$ for some constant $C>0$ is shown in B.5. Thus, for $\|G(X)-\widetilde{G}^N(X)\|_F^2$, we have that

$$\|G-\widetilde{G}^N\|_F^2\leq\frac{k^2C(d\|\Theta\|_F)^{2(N+1)}}{((N+1)!)}$$

We also know that $\frac{\exp(2M)}{k^2}+\frac{\exp(6M)}{k^3}\leq\frac{\exp(6M)}{k^2}$. Our bound for $\|S(X)-\widehat{S}^{(N)}(X)\|_F^2$ becomes:

$$\|S(X)-\widehat{S}^{(N)}(X)\|_F^2=O\Big(\exp(6M)\cdot\frac{(d\|\Theta\|_F)^{2(N+1)}}{(N+1)!}\Big)$$

Since we want $\|S(X)-\widehat{S}^{(N)}(X)\|_F^2\leq\epsilon^2$ for $\epsilon\in(0,1)$, we set the RHS to $\leq\epsilon^2$:

$$\|S(X)-\widehat{S}^N(X)\|_F^2\leq\epsilon^2\implies\exp(6M)\frac{(d^2\|\Theta\|_F^2)^{N+1}}{(N+1)!}\leq\epsilon^2$$

Set $A=(d^2\|\Theta\|_F^2)$ and $x=N+1$. Applying log to both sides, we have that

$$\begin{aligned}\log\Big(\exp(6M)\cdot\frac{(d^2\|\Theta\|_F^2)^{(N+1)}}{(N+1)!}\Big)=\log\Big(\exp(6M)\cdot\frac{A^x}{x!}\Big)&\leq2\log\epsilon\\ 6M+x\log A-\log(x!)&\leq2\log\epsilon\end{aligned}$$

Set $R=-\log\epsilon+6M$, which means we have $6M+x\log A-\log(x!)\leq-R$. From Sterling's

Approximation Inequality, we have that $\log(x!) \leq x \log x - x$. Then, we can rearrange and write

$$\begin{aligned} x \log A - \log(x!) \leq -R \quad &\implies \quad x \log A - (x \log x - x) \leq -R \\ &\implies \quad x(\log A - \log x + 1) \leq -R \\ &\implies \quad x(\log x - \log A - 1) \geq R \end{aligned}$$

Let's make the guess that $x - CA \log R$ for $C > 0$. We have that

$$\log(x) = \log(CA \log R) = \log C + \log A + \log \log R$$

$$\implies \log x - \log A - 1 = \log C + \log \log R - 1$$

Then for $x(\log x - \log A - 1)$ we have

$$x(\log x - \log A - 1) \leq CA \log R \cdot (\log \log R + \log C - 1) = \Theta\big(A \log R + \log \log R\big)$$

$$\implies R = \Theta\big(A \log R + \log \log R\big) \implies x = \Theta\Big(A \log\Big(\frac{1}{\epsilon}\Big)\Big)$$

Since $M = O(\log k)$ is worse than $\log(1/\epsilon) << 1$, we have that

$$N = \Theta\Big((d^2 \|\Theta\|_F^2) \cdot \log\Big(\frac{1}{\epsilon}\Big)\Big)$$

which finishes the proof. It is important to note that this proof does not need M from B.3, but an improved bound for N could be found using the results from B.3 which we did not show. □

C Learning Algorithm

C.1 Dimensions D_p, D_q

Firstly, $D_p := |\{(\alpha, \beta) \in \mathbb{N}^d \times \mathbb{N}^d : |\alpha| + |\beta| \leq N\}|$ is the dimension of the polynomial $p^{(N)}$ and $D_q := |\{(\lambda_1, ..., \lambda_k) : \sum_{w=1}^{k} |\lambda_w| \leq N \leq N\}|$ is the dimension of the polynomial $q^{(N)}$. Let us calculate the dimensions D_p, D_q of the Hermite feature basis. For D_p, we have that

$$\begin{aligned} D_p := |\{(\alpha, \beta) \in \mathbb{N}^d \times \mathbb{N}^d : |\alpha| + |\beta| \leq N\}| &= \sum_{m=0}^{N} \Big(\sum_{a+b=m} \#(\alpha \in \mathbb{N}^d : |\alpha| = a\} \cdot \#(\beta \in \mathbb{N}^d : |\beta| = b\} \Big) \\ &= \Big(\sum_{m=0}^{N} \sum_{s=0}^{N-m} \binom{d+m-1}{m} \cdot \binom{d+s-1}{s} \Big) \\ &\leq \binom{2d+N}{N} \approx O(N^{2d}) \end{aligned}$$

with the last inequality coming from applying Sterling's inequality.

$$D_q := |\{(\lambda_1, ..., \lambda_k) \in (\mathbb{N}^d)^k : \sum_{w=1}^{k} |\lambda_w| \leq N| = \sum_{m_1,...,m_k \leq N} \prod_{l=1}^{k} \binom{d + m_l - 1}{m_l}$$
$$\leq \binom{kd + N}{N} \approx O(N^{kd})$$

with the last inequality coming from applying Sterling's Inequality.

C.2 Run-Time Analysis

Firstly, from C.1 we know that $D_p = O(N^{2d})$ and $D_q = O(N^{kd})$ and $D = D_p + D_q = O(N^{kd})$ since $kd \geq 2$ for $k \geq 2$. From the total runtime of the algorithm we obtained, we can substitute in D_p and D_q to get

$$\begin{aligned} &O(mk^2 D_p + mk^2(D_p + D_q) + mk^2(D_p + D_q)^2 + (D_p + D_q)^3) \\ &= O\big(mk^2 N^{kd} + mk^2 N^{2d} + mk^2 N^{kd} + mk^2 N^{2kd} + N^{3kd}\big) \\ &= O\big(mk^2 N^{2kd} + N^{3kd}\big) \end{aligned}$$

With the last line having the highest-order terms kept. Now, from Lemma 3.2 we know that $N = \Theta\big(d^2\|\Theta\|_F^2 \log(1/\epsilon)\big) = Cd^2\|\Theta\|_F^2 \log(1/\epsilon)$ for some constant $C > 0$. Let us substitute in N into the runtime. For the first term $mk^2 N^{2kd}$, we obtain:

$$mk^2 N^{2kd} = mk^2\big(Cd^2\|\Theta\|_F^2 \log(1/\epsilon)\big) = mk^2 C^{2kd} d^{4kd} \|\Theta\|_F^{4kd} \big(\log(1/\epsilon)\big)^{2kd}$$

For the second term N^{3kd}, we obtain:

$$N^{3kd} = \big(Cd^2\|\Theta\|_F^2 \log(1/\epsilon)\big)^{3kd} = C^{3kd} d^{6kd} \|\Theta\|_F^{6kd} \big(\log(1/\epsilon)\big)^{3kd}$$

Therefore, we obtain

$$\begin{aligned} &O\Big(mk^2 C^{2kd} d^{4kd} \|\Theta\|_F^{4kd} \big(\log(1/\epsilon)\big)^{2kd} + C^{3kd} d^{6kd} \|\Theta\|_F^{6kd} \big(\log(1/\epsilon)\big)^{3kd}\Big) \\ &= \widetilde{O}\Big(mk^2 \cdot \Big(d^2\|\Theta\|_F^2 \log\Big(\frac{1}{\epsilon}\Big)\Big)^{2kd} + \Big(d^2\|\Theta\|_F^2 \log\Big(\frac{1}{\epsilon}\Big)^{3kd}\Big)\Big) \end{aligned} \tag{10}$$

to be the total runtime of the PAC learning algorithm.

C.3 Sample Complexity

To find the sample complexity bound for the PAC learning algorithm in the realizable setting, we will use Rademacher Complexity as defined in [BM02]. Let $\mathcal{F}_N$ be the class of functions (hypothesis class) defined by

$$\mathcal{F}_N := \left\{ (x_1, ..., x_k) \mapsto f_{i,j}(X) = \frac{\langle c_p, \phi_p(x_i, x_j)\rangle}{\langle c_q, \phi_q(x_1, ..., x_k)\rangle} \right\}$$

with the following assumptions: the coefficients are bounded,

$$\|c_p\|_2 \leq B_p, \quad \|c_q\|_2 \leq B_q$$

the features are bounded,

$$\|\phi_p(x_i, x_j)\|_2 \leq R_p, \quad \|\phi_q(x_1, ..., x_k)\|_2 \leq R_q,$$

and the denominator is bounded away from zero.

$$\langle c_q, \phi(x_1, ..., x_k)\rangle \geq \tau > 0$$

Let us make these further non-degeneracy assumptions:

Assumption C.1. $B_p, B_q = O(1)$ and $\tau = \Omega(1)$

Let us use Theorem 8 of [BM02] to obtain the sample complexity for PAC learning given our class $\mathcal{F}_N$:

Theorem C.2. *For a class $\mathcal{F}$ of real-valued functions with outputs in $[a, b]$ and loss $l(f, S)$ that is L-Lipschitz in f, the generalization error of the empirical minimizer satisfies:*

$$\forall f \in \mathcal{F}, \quad \mathbb{E}[l(f, S)] \leq \widehat{E}[l(f, S)] + 2L\mathcal{R}_m(\mathcal{F}) + 3M_R\sqrt{\frac{\log(2/\delta)}{2m}}$$

where $\mathcal{R}_m(\mathcal{F}) := \mathbb{E}_\sigma[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m \sigma_i f(z_i)]$ is the empirical Rademacher complexity, $l(f, S) = (S(X) - f(X))^2$ has Lipschitz constant $L = 2M_R$ if $f(X), S(X) \in [-M_R, M_R]$ and M_R is a bound on the function outputs.

We know that $|f_{i,j}(X)| \leq \frac{B_pR_p}{\tau}$ and thus $M_R = \frac{B_pR_p}{\tau}$ and $L = \frac{2B_pR_p}{\tau}$. Applying C.2, for all $f \in \mathcal{F}_N$, with probability $1 - \delta$ we have

$$\begin{aligned}\mathbb{E}\left[(S(X) - f(X))^2\right] &\leq \frac{1}{m}\sum_{a=1}^m (S(X^{(a)}) - f(X^{(a)}))^2 + \frac{4B_pR_p}{\tau} \cdot \widehat{\mathcal{R}}_m(\mathcal{F}) + 3\Big(\frac{B_pR_p}{\tau}\Big)\sqrt{\frac{\log(2/\delta)}{2m}} \\ &\leq \frac{1}{m}\sum_{a=1}^m (S(X^{(a)}) - f(X^{(a)}))^2 + O\Big(\frac{(B_pR_p + B_qR_q)^2}{\tau^2 m} + \frac{\log(1/\delta)}{m}\Big)\end{aligned}$$

as the empirical Rademacher complexity can be bounded by $\widehat{\mathcal{R}}_m(\mathcal{F}_N) \leq \frac{B_pR_p + B_qR_q}{\tau\sqrt{m}}$ and the bounds for L, R_M. Now, if the training loss is small (e.g., realizable case with good Hermite approximation), then with high probability $1 - \delta$ we have:

$$\mathbb{E}\left[(S(X) - f(X))^2\right] \leq O\Big(\frac{(B_pR_p + B_qR_q)^2 + \tau^2\log(1/\delta)}{\tau^2 m}\Big)$$

Solving for m to guarantee $\mathbb{E}\left[(S(X) - f(X))^2\right] \leq \epsilon^2$:

$$m = O\Big(\frac{(B_pR_p + B_qR_q)^2 + \tau^2\log(1/\delta)}{\tau^2\epsilon^2}\Big)$$

Now, let us bound R_p and R_q. For bounding R_p, note that each entry of $\phi_p^{(N)}(x_i, x_j)$ is of the form

$h_\alpha(x_i)h_\beta(x_j)$. Since $h_\alpha(x_i)$ is orthonormal with respect to $\gamma = N(0, I_d)$, and we truncate to total degree $\leq N$, we have that

$$\mathop{\mathbb{E}}_{x_i,x_j\sim\gamma}\big[\|\phi_p(x_i,x_j)\|_2^2\big] = \sum_{|\alpha|+|\eta|\leq N} \mathbb{E}[h_\alpha(x_i)^2]\,\mathbb{E}[h_\beta(x_i)^2] = D_p$$

By concentration of measure under Gaussian (see B.1), the norm is tightly concentrated near this mean. Thus, with high probability:

$$R_p = \sup_{x_i,x_j} \|\phi_p^{(N)}(x_i,x_j)\|_2 \leq C\sqrt{D_p} = O(N^d)$$

Bounding R_q is similar: given each entry of $\phi_q^{(N)}(x_1,...,x_k)$, we have that

$$\mathop{\mathbb{E}}_{X\sim\gamma}[\|\phi_q(X)\|_2] \leq C\sqrt{D_q} = O(N^{kd/2})$$

Given our non-degeneracy assumptions that $B_p, B_q = O(1)$ and $\tau = \Omega(1)$, with probability at least $1-\delta$, the number of training examples m must be

$$\begin{aligned} m = O\Big(\frac{(B_pR_p + B_qR_q)^2 + \tau^2\log(1/\delta)}{\tau^2\epsilon^2}\Big) &= O\Big(\frac{(R_p+R_q)^2 + \log(1/\delta)}{\epsilon^2}\Big) \\ &= O\Big(\frac{(N)^{kd} + \log(1/\delta)}{\epsilon^2}\Big) \end{aligned}$$

With the last inequality coming from $kd \geq 2d$ for $k \geq 2$ which implies that $N^{kd/2} >> N^d$ and thus $(R_p+R_q)^2 = O(N^{kd})$. Let us now use $N = Cd^2\|\Theta\|_F^2\log(1/\epsilon)$ for some constant $C > 0$. Therefore, we have that with probability at least $1-\delta$, the number of training examples m must satisfy

$$m = O\Big(\frac{(d^2\|\Theta\|_F^2\log(1/\epsilon))^{kd} + \log(1/\delta)}{\epsilon^2}\Big) \tag{11}$$

D Acknowledgments

I would like to thank Daniel Hsu for the helpful notes and direction for this problem.