Rate of Convergence of Polynomial Networks to Gaussian Processes

Adam Klukowski AK2028@CANTAB.AC.UK
Huawei Noah's Ark Lab

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

We examine one-hidden-layer neural networks with random weights. It is well-known that in the limit of infinitely many neurons they simplify to Gaussian processes. For networks with a polynomial activation, we demonstrate that the rate of this convergence in 2-Wasserstein metric is $O(n^{-\frac{1}{2}})$, where n is the number of hidden neurons. We suspect this rate is asymptotically sharp. We improve the known convergence rate for other activations, to power-law in n for ReLU and inverse-square-root up to logarithmic factors for erf. We explore the interplay between spherical harmonics, Stein kernels and optimal transport in the non-isotropic setting.

Keywords: Deep learning theory, wide neural networks, Gaussian processes, CLT, neural priors

1. Introduction

We are concerned with a 1-hidden-layer neural network $\mathcal{P}_n$ of width n. This is a random function from the sphere $\sqrt{d}S^{d-1} = \left\{x \in \mathbb{R}^d : \|x\|^2 = d\right\}$ to $\mathbb{R}$, defined by

$$\mathcal{P}_n(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} s_i \phi\left(\frac{w_i . x}{\sqrt{d}}\right)$$

where ϕ is a fixed function $\mathbb{R} \to \mathbb{R}$ called activation. The randomness of $\mathcal{P}_n$ comes from the weights s_i, w_i, which are random variables taking values in $\mathbb{R}$ and $\mathbb{R}^d$ respectively. We assume they are independent and identically distributed across i.

Under relatively mild regularity conditions, $\mathcal{P}_n$ converges to a Gaussian process (GP) as $n \to \infty$. This fact was first noticed in Neal (1996), and discussed in greater generality in Hanin (2021). A number of works aimed to go beyond the limit and understand the phenomena in its neighbourhood. The distribution of preactivations[1] was studied perturbatively in Yaida (2020). The behaviour of observables – scalars summarizing the distribution, analogous to the moments of a random variable – was analysed in Dyer and Gur-Ari (2020) and Aitken and Gur-Ari (2020). Here we investigate the rate of convergence from the angle of functional metrics. This line of work was initiated in Eldan et al. (2021), where it was shown that $\mathcal{P}_n$ is within $O\left(n^{-\frac{1}{6}}\right)$ from a GP in the ∞-Wasserstein distance when ϕ is a polynomial. We extend their method and obtain a rate of $O(n^{-\frac{1}{2}})$ in the 2-Wasserstein metric, which is asymptotically sharp.

This work is closely tied to Stein's method and optimal transport. Given a vector-valued random variable X, a Stein kernel for X is any matrix-valued function τ satisfying $\mathbb{E}[X . f(X)] = \mathbb{E}\langle \tau(X), \nabla f \rangle_{HS}$ for any test function f, where $\langle \bullet, \bullet \rangle_{HS}$ is the Hilbert-Schmidt inner product. They

1. Intermediate vectors, or neuron outputs, in deep networks comprising multiple stacked affine maps and coordinate-wise activations

were introduced in Stein (1986); for an overview see for example Mijoule et al. (2018) or Azmoodeh et al. (2021). They are strongly related to quantitative and multi-dimensional forms of CLT Courtade et al. (2019). It was shown in Ledoux et al. (2015) using the Orstein-Uhlenbeck diffusion semigroup that the Wasserstein distance to a Gaussian can be controlled by the discrepancy between the kernel and a constant matrix. Here we show how to exploit the rich symmetries of spherical harmonics to construct Stein kernels. Also, we explore generalizations of diffusive methods Ledoux et al. (2015),Otto and Villani (2000) to relate Wasserstein distance and Stein discrepancy in a non-isotropic[2] setting.

Organization: In the rest of this section we define the notation and then describe the main results. Section 2 relates random functions to vector-valued random variables. We prove our main result about polynomial networks in section 3, and analyse general activations in section 4. Spherical harmonics are explained in appendix A, and Stein kernels and optimal transport are described in appendix B.

1.1. Notation

We denote the inner product and induced norm of vectors as

$$u.v = \sum_i u_i v_i \qquad\qquad \|u\|^2 = u.u$$

Both will often be accompanied by normalizing factors. We will mostly be working on the sphere $\sqrt{d}S^{d-1} = \{x \in \mathbb{R}^d : \|x\|^2 = d\}$, and we denote the uniform distribution on it by $\sqrt{d}US^{d-1}$.

We denote the Hilbert-Schmidt product of matrices as

$$\langle A, B\rangle_{HS} = \text{Tr } AB^\top = \sum_{i,j} A_{i,j}B_{i,j}$$

The metric suitable for comparing random objects is the 2-Wasserstein distance, defined for random vectors X, Y and random functions f, g as

$$W_2(X,Y)^2 = \mathbb{E}\left[\left\|X - Y\right\|^2\right] \qquad\qquad W_2(f,g)^2 = \mathbb{E}\int_{\sqrt{d}S^{d-1}} \left|f(x) - g(x)\right|^2 dx$$

When the two vectors do not share a common probability space, we define $\mathcal{W}_2(X,Y) = \inf_{(X,Y)} W_2(X,Y)$, where the infimum is taken over all couplings (joint distributions having X, Y as marginals); the definition for random functions is analogous.

Random function $\mathcal{G}$ is a Gaussian process (GP) if the vector $\big(\mathcal{G}(x)\big)_{x\in X}$ has multivariate normal distribution for any finite set of arguments X.

We will make heavy use of spherical harmonics – a set of functions $Y_{l,m} : \sqrt{d}S^{d-1} \to \mathbb{R}$ indexed by $l \in \mathbb{N}_0, 1 \le m \le \text{đ}_l$. They are discussed in detail in appendix A. Their key property is orthonormality, meaning

$$\int_{\sqrt{d}S^{d-1}} Y_{l,m}(x)Y_{l',m'}(x)dx = \delta_{ll'}\delta_{mm'}$$

also, they span the Hilbert space of square-integrable functions $\mathcal{L}^2\big(\sqrt{d}S^{d-1}\big)$. They give rise to the orthogonal family of Gegenbauer polynomials P_l.

2. In this paper by isotropic variables we mean those with covariance matrix equal to identity

1.2. Overview of results and main ideas

Our main result (with some technicalities omitted) is

Theorem 1 (simplified). *Assume that s_i satisfies $\mathbb{E}[s^2] = 1$, w_i are uniformly distributed on the sphere $w_i \sim \sqrt{d}US^{d-1}$, and the activation ϕ is a polynomial. Then there exists a Gaussian process $\mathcal{G}$ such that*

$$\mathcal{W}_2(\mathcal{P}_n, \mathcal{G}) \leq \frac{C}{\sqrt{n}}$$

where $C^2 = O\left((d + \deg\phi)^d \cdot \mathbb{E}\left[s^4\right] \cdot \mathbb{E}\left[\phi'\left(\mathcal{N}(0,1)\right)^2\right]\right)$.

The precise statement is theorem 1 in section 3. By approximating the activation with polynomials (section 4) we obtain

Theorem 2 (simplified). *Assume $\mathbb{P}(s_i = 1) = \mathbb{P}(s_i = -1) = \frac{1}{2}$, $w_i \sim \sqrt{d}US^{d-1}$. For the rectified linear unit*[3] *activation $\phi = \mathrm{ReLU}$ we have*

$$\mathcal{W}_2\left(\mathcal{P}_n, \mathcal{G}\right) \leq 7n^{-\frac{3}{2(2d-1)}}$$

while for the error function[4] *$\phi = \mathrm{erf}$ we have*

$$\mathcal{W}_2\left(\mathcal{P}_n, \mathcal{G}\right) \leq C^d (\log n)^{\frac{d-2}{2}} \cdot n^{-\frac{1}{2}}$$

The first main idea is to find a Hilbert space V together with an embedding $E : \sqrt{d}S^{d-1} \hookrightarrow V$ satisfying $\phi\left(\frac{w.x}{\sqrt{d}}\right) = E(w).E(x)$, and use it to express the neural network as an inner product in V as

$$\mathcal{P}_n(x) = \underbrace{\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} s_i E(w_i)\right)}_{\tilde{w}} .E(x) \tag{1}$$

This description separates "random" from "function" – the first bracket $\tilde{w}$ does not depend on the argument x, while the second is a deterministic function of x.

The next step is approximating the first factor $\tilde{w}$ by a multivariate normal, using a variant of quantitative CLT. This can then be translated this into an approximation of the network $\mathcal{P}_n$ by a Gaussian process.

In Eldan et al. (2021) an embedding into $V = \left(\mathbb{R}^d\right)^{\otimes 0} \oplus \cdots \oplus \left(\mathbb{R}^d\right)^{\otimes \deg\phi}$ was obtained by expanding all monomials $(w.x)^k$ of $\phi(w.x)$. Here we use an expansion in the basis of spherical harmonics. This approach gives a simple covariance structure of the random vector $\tilde{w}$ – its matrix is diagonal with explicit eigenvalues. In fact, this expansion enables us to isometrically translate the problem into a question about countably-dimensional random vectors. Then we employ the machinery of Stein kernels, and construct one by leveraging the geometry of spherical harmonics.

2. Harmonic decomposition

Here we exhibit a Gaussianity-preserving linear Wasserstein-isometry between random functions and random vectors. Any random function f on the sphere can be expanded in the basis of spherical

3. $\mathrm{ReLU}(x) = \max\{0, x\}$
4. $\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-y^2} dy$

harmonics, giving a $\mathbb{R}^{\sharp_0+\sharp_1+\cdots}$-valued random variable

$$X_{l,m} = \int_{\sqrt{d}S^{d-1}} Y_{l,m}(x) f(x) dx$$

Conversely, every $\mathbb{R}^{\sharp_0+\sharp_1+\cdots}$-valued random variable $X_{l,m}$ naturally defines a random function via

$$f(x) = \sum_{l=0}^{\infty} \sum_{m=1}^{\sharp_l} X_{l,m} Y_{l,m}(x)$$

Spherical harmonics form a complete orthonormal basis, so these transformations are mutually inverse. When restricted to sufficiently nice spaces, they define correspondences

$$\text{random functions on } \sqrt{d}S^{d-1} \quad \longleftrightarrow \quad \mathbb{R}^{\sharp_0+\sharp_1+\cdots}\text{-valued random variables} \tag{2}$$

$$\mathbb{E}\int_{\sqrt{d}S^{d-1}} \left|f_1(x) - f_2(x)\right|^2 dx \quad \longleftrightarrow \quad \mathbb{E}\|X^{(1)} - X^{(2)}\|^2 \tag{3}$$

$$\text{Gaussian processeses} \quad \longleftrightarrow \quad \text{multivariate normal variables} \tag{4}$$

$$\text{NNs } \mathcal{P}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} s_i \phi\left(\frac{w_i . x}{\sqrt{d}}\right) \quad \longleftrightarrow \quad \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\hat{\phi}_l}{\sqrt{\sharp_l}} s_i Y_{l,m}(w_i) \tag{5}$$

Line 3 states that we are dealing with an isometry with respect to appropriate 2-Wasserstein metrics. This is a consequence of the orthonormality of harmonics

$$\begin{aligned}
&\mathbb{E}\int_{\sqrt{d}S^{d-1}} \left|f_1(x) - f_2(x)\right|^2 dx = \mathbb{E}\int \left[\sum_{l,m} \left(X^{(1)}_{l,m} - X^{(2)}_{l,m}\right) Y_{l,m}(x)\right]^2 dx = \\
&= \mathbb{E} \sum_{l,m,l',m'} \left(X^{(1)}_{l,m} - X^{(2)}_{l,m}\right)\left(X^{(1)}_{l',m'} - X^{(2)}_{l',m'}\right) \int Y_{l,m}(x) Y_{l',m'}(x) dx = \\
&= \mathbb{E} \sum_{l,m} \left(X^{(1)}_{l,m} - X^{(2)}_{l,m}\right)^2 = \mathbb{E}\left\|X^{(1)} - X^{(2)}\right\|^2
\end{aligned}$$

Preservation of Gaussianity 4 holds because the maps are linear.

In equation 5, the coefficients $\hat{\phi}_l$ come from the expansion $\phi = \sum_{l=0}^{\infty} \hat{\phi}_l P_l$ of the activation function ϕ into Gegenbauer polynomials P_l. Equation 17 from appendix A.2 states that their value at a dot product is expressible in terms of spherical harmonics as

$$P_l\left(\frac{w . x}{\sqrt{d}}\right) = \frac{1}{\sqrt{\sharp_l}} \sum_{m=1}^{\sharp_l} Y_{l,m}(w) Y_{l,m}(x)$$

this allows us to interpret the network as an Euclidean inner product in an enlarged space

$$\begin{aligned}
\mathcal{P}_n(x) &= \sum_{l=0}^{\infty} \sum_{m=1}^{\sharp_l} \sum_{i=1}^{n} \frac{1}{\sqrt{n}} \cdot \frac{\hat{\phi}_l}{\sqrt{\sharp_l}} \cdot s_i Y_{l,m}(w_i) Y_{l,m}(x) = \\
&= \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\hat{\phi}_l}{\sqrt{\sharp_l}} \cdot s_i Y_{l,m}(w_i)\right)_{l,m} \cdot \left(Y_{l,m}(x)\right)_{l,m}
\end{aligned} \tag{6}$$

3. Polynomial networks

Theorem 1 *Assume that the weights s_i obey $\mathbb{E}[s^2]=1$, the weights w_i are distributed uniformly on the sphere $w_i \sim \sqrt{d}US^{d-1}$, and the activation ϕ is a polynomial of degree k satisfying $\mathbb{E}\left[\phi(x_1)\middle| x\sim\sqrt{d}US^{d-1}\right]=0$. Then, for each n, there exists a Gaussian process $\mathcal{G}$ such that*

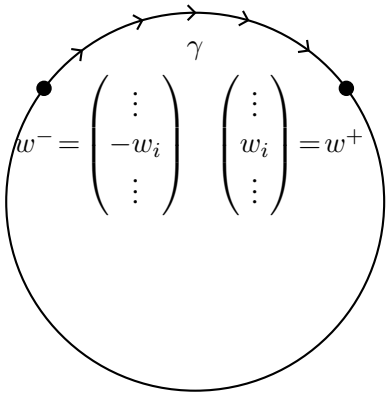
$$\mathcal{W}_2(\mathcal{P}_n,\mathcal{G})\le\frac{C}{\sqrt{n}}$$

where

$$C^2=\frac{6d(d+k)^{d-2}}{(d-1)!}\cdot\mathbb{E}\left[s^4\right]\mathbb{E}\left[\phi'\left(\mathcal{N}(0,1)\right)^2\right]+\mathrm{var}\left[s^2\right]\mathbb{E}\left[\phi(x_1)^2\middle| x\sim\sqrt{d}US^{d-1}\right]$$

Idea of proof: Note it is enough to bound the distance between the random bracket from equation 6 and a Gaussian. We achieve this by exhibiting a Stein kernel for the random variable $\frac{\hat{\phi}_l}{\sqrt{d_l}}Y_{l,m}(w)\ \middle|\ w\sim\sqrt{d}US^{d-1}$.
Our construction for $l=1$ is illustrated on the right. Let f be a test function, and recall that $Y_{1,i}(w)=w_i$. For each i, we pair up the points w^+,w^- that differ only by the sign of the i-th coordinate

$$\mathbb{E}[w_i f(w)]=\tfrac{1}{2}\mathbb{E}\left[|w_i|\left(f(w^+)-f(w^-)\right)\right]$$



We join them with the shortest curve γ, and apply the fundamental theorem of calculus to the difference

$$f(w^+)-f(w^-)=\int_\gamma \nabla f.d\gamma=\mathop{\mathbb{E}}_{w\in\gamma}\left[\dot{\gamma}(w).\nabla f(w)\right]$$

Averaging over the sphere gives an equation of the form

$$\mathbb{E}[w_i f(w)]=\mathbb{E}\left[(\text{some vector field}).\nabla f\right]$$

Which is precisely the form of a Stein kernel.
It is not immediately clear how to generalize this construction beyond $l=1$. However, it turns out that the vector field we obtain is precisely the gradient of $Y_{1,i}$ tangent to the sphere $\sqrt{d}S^{d-1}$. This interpretation makes sense for any l,m. Thus, what we actually do is the calculation of average derivative of test functions in the direction of $\nabla Y_{l,m}$. It turns out that every spherical harmonic except $Y_{l,m}$ is annihilated.

Once we construct the kernel, we compute its Stein discrepancy S using identities from appendix A.1. We finish by invoking lemmas from appendix B to extract a bound on the Wasserstein distance from the discrepancy.

3.1. Proof of theorem 1

3.1.1. NOTATION AND QUOTED RESULTS

For $1\le l\le k$ denote

$$\tilde{Y}_{l,m}(w)=\frac{\hat{\phi}_l}{\sqrt{d_l}}Y_{l,m}(w)$$

This is an embedding $\tilde{Y} : \sqrt{d}S^{d-1} \hookrightarrow \mathbb{R}^{\not{d}_1+\cdots+\not{d}_k}$, whose left inverse is the projection onto the first $\not{d}_1 = d$ coordinates. We will be concerned with the random variable $\tilde{y} = \tilde{Y}(w)\big|w \sim \sqrt{d}US^{d-1}$.

We will need the rotation matrices R^α_{ab} that act on the basis vectors e_i as

$$\begin{aligned} R^\alpha_{ab}e_a &= \cos\alpha\ e_a - \sin\alpha\ e_b \\ R^\alpha_{ab}e_b &= \sin\alpha\ e_a + \cos\alpha\ a_b \\ R^\alpha_{ab}e_c &= e_c \quad \text{when} \quad c \notin \{a,b\} \end{aligned}$$

and the operators

$$\partial_r \overset{\text{def}}{=} w.\nabla = \sum_{i=1}^{d} w_i\partial_i \qquad L_{ab} \overset{\text{def}}{=} w_a\partial_b - w_b\partial_a \qquad L^2 \overset{\text{def}}{=} \sum_{a<b} L_{ab}^2$$

Finally we recall the following identities from appendix A

$$\begin{aligned} \big(L_{ab}f\big)(x) &= -\,\partial_\alpha f(R^\alpha_{ab}x)\Big|_{\alpha=0} && \text{(equation 14)} \\ r^2\nabla^2 &= L^2 + \partial_r(\partial_r + d - 2) && \text{(equation 15)} \\ 0 &= (d-t^2)P_l''(t) - (d-1)tP_l'(t) + l(l+d-2)P_l(t) && \text{(equation 19)} \end{aligned}$$

Where $\nabla^2 = \sum_{i=1}^d \partial_i^2$ is the Laplacian.

3.1.2. CONSTRUCTION OF THE KERNEL

We want to build a Stein kernel for $\tilde{y}$. Consider a test function $f : \mathbb{R}^{\not{d}_1+\cdots+\not{d}_k} \to \mathbb{R}$. We would like to understand

$$\mathbb{E}\left[\tilde{y}_{l,m}f(\tilde{y})\right] = \tfrac{\hat{\phi}_l}{\sqrt{\not{d}_l}}\mathbb{E}\left[Y_{l,m}(w)f(\tilde{Y}(w))\right]$$

The second expectation is simply the coefficient standing next to $Y_{l,m}$ in the harmonic expansion of $f \circ \tilde{Y}$. We will temporarily move from $\mathbb{R}^{\not{d}_1+\cdots+\not{d}_k}$ with the test function f to $\mathbb{R}^d \supseteq \sqrt{d}S^{d-1}$ with the test function $f \circ \tilde{Y}$. As promised, consider the tangent gradient

$$\nabla Y_{l,m} - \tfrac{l}{d}Y_{l,m}w \qquad \text{and the corresponding operator} \qquad \left(\nabla Y_{l,m} - \tfrac{l}{d}Y_{l,m}w\right).\nabla \tag{7}$$

Viewed in $\mathbb{R}^d$, $Y_{l,m}$ is a homogeneous polynomial of degree l, so $w.\nabla Y_{l,m} = lY_{l,m}$. Hence the operator annihilates $r^2 = \|w\|^2$, so this vector field is tangent to the sphere $\sqrt{d}S^{d-1}$.

Let us look at how does this vector field affect harmonic expansions. Remembering equation 15 and $\mathbb{E}[L^2 g] = 0$, we can rewrite the action of 7 on spherical harmonics as

$$\begin{aligned} &\mathbb{E}\Big[\left(\nabla Y_{l,m} - \tfrac{l}{d}Y_{l,m}w\right).\nabla Y_{l',m'}\Big|w \sim \sqrt{d}US^{d-1}\Big] = \\ &= \mathbb{E}\Big[\tfrac{1}{2d}r^2\nabla^2\big(Y_{l,m}Y_{l',m'}\big) - \tfrac{l}{d}Y_{l,m}\partial_r Y_{l',m'}\Big] = \\ &= \mathbb{E}\Big[\tfrac{1}{2d}\big(L^2 + \partial_r(\partial_r+d-2)\big)\big(Y_{l,m}Y_{l',m'}\big) - \tfrac{ll'}{2d}Y_{l,m}Y_{l',m'}\Big] = \\ &= \Big(\tfrac{(l+l')(l+l'+d-2)}{2d} - \tfrac{ll'}{d}\Big)\,\mathbb{E}\left[Y_{l,m}Y_{l',m'}\right] = \\ &= \tfrac{l^2+l'^2+(l+l')(d-2)}{2d}\delta_{ll'}\delta_{mm'} \end{aligned}$$

In expectation, this vector field annihilates every spherical harmonic other than $Y_{l,m}$ itself, which is sent to $\frac{l(l+d-2)}{d}$. Therefore we can filter the coefficients of $f \circ \tilde{Y}$ using the identity

$$\mathbb{E}\left[Y_{l,m}(w)(f\circ\tilde{Y})(w)\right] = \tfrac{d}{l(l+d-2)}\mathbb{E}\left[\left(\nabla Y_{l,m} - \tfrac{l}{d}Y_{l,m}w\right).\nabla(f\circ\tilde{Y})\right]$$

Now we need to return from the sphere $\sqrt{d}US^{d-1}$ and go back to $\mathbb{R}^{\sharp_1+\cdots+\sharp_k}$. We do it using chain rule

$$\left(\nabla Y_{l,m} - \tfrac{l}{d}Y_{l,m}w\right).\nabla(f\circ\tilde{Y}) = \sum_{l',m'}\tfrac{\hat{\phi}_l}{\sqrt{\sharp_l}}\left(\nabla Y_{l,m} - \tfrac{l}{d}Y_{l,m}w\right).\nabla Y_{l',m'}\partial_{l',m'}f$$

Therefore

$$\mathbb{E}\left[\tilde{y}_{l,m}f(\tilde{y})\right] = \mathbb{E}\left[\sum_{l',m'}\tau_{l,m;l',m'}\partial_{l',m'}f(\tilde{y})\right]$$

where $$\tau_{l,m;l',m'}\left(\tilde{Y}(w)\right) = \tfrac{\hat{\phi}_l\hat{\phi}_{l'}}{\sqrt{\sharp_l\sharp_{l'}}}\tfrac{d}{l(l+d-2)}\left(\nabla Y_{l,m} - \tfrac{l}{d}Y_{l,m}w\right).\nabla Y_{l',m'}$$

This means that τ is a Stein kernel for $\tilde{y}$.

3.1.3. HILBERT-SCHMIDT NORM OF BLOCKS

Let us rewrite the kernel as

$$\tau_{l,m;l',m'} = \tfrac{\hat{\phi}_l\hat{\phi}_{l'}}{\sqrt{\sharp_l\sharp_{l'}}}\tfrac{1}{2l(l+d-2)}\left[r^2\nabla^2\left(Y_{l,m}Y_{l',m'}\right) - 2ll'Y_{l,m}Y_{l',m'}\right]$$

expanding the Laplacian according to equation 15 we get

$$\begin{aligned}
&r^2\nabla^2\left(Y_{l,m}Y_{l',m'}\right) - ll'Y_{l,m}Y_{l',m'} = \\
&\quad = -\left(l(l+d-2)+l'(l'+d-2)\right)Y_{l,m}Y_{l',m'} + (l+l')(l+l'+d-2)Y_{l,m}Y_{l',m'} + \\
&\qquad + \sum_{a,b}L_{ab}Y_{l,m}\cdot L_{ab}Y_{l',m'} - 2ll'Y_{l,m}Y_{l',m'} = \\
&\quad = \sum_{a,b}L_{ab}Y_{l,m}\cdot L_{ab}Y_{l',m'}
\end{aligned}$$

which means

$$\tau_{l,m;l',m'} = \tfrac{\hat{\phi}_l\hat{\phi}_{l'}}{\sqrt{\sharp_l\sharp_{l'}}}\tfrac{1}{2l(l+d-2)}\sum_{a,b}L_{ab}Y_{l,m}\cdot L_{ab}Y_{l',m'}$$

We will calculate the Hilbert-Schmidt norm of (l,l')-block of τ. We have

$$\begin{aligned}
&\sum_{m,m'}\left(\sum_{a,b}L_{ab}Y_{l,m}\cdot L_{ab}Y_{l',m'}\right)^2 = \\
&\quad = \sum_{a,b,c,d,m,m'}L_{ab}Y_{l,m}L_{ab}Y_{l',m'}L_{cd}Y_{l,m}L_{cd}Y_{l',m'} = \\
&\quad = \sum_{a,b,c,d}\left(\sum_{m}L_{ab}Y_{l,m}L_{cd}Y_{l,m}\right)\left(\sum_{m'}L_{ab}Y_{l',m'}L_{cd}Y_{l',m'}\right)
\end{aligned}$$

We will calculate the sums for fixed l. Denote $(S_{ab})_{ij} = \left(\partial_\alpha R^\alpha_{ab}\big|_{\alpha=0}\right)_{ij} = \delta_{ai}\delta_{bj} - \delta_{aj}\delta_{bi}$. Recalling equation 14, we can compute the action of L-operators

$$\sum_m L_{ab}Y_{l,m}L_{cd}Y_{l,m} = \sum_m \partial_\alpha Y_{l,m}\big(R^\alpha_{ab}x\big)\Big|_{\alpha=0}\partial_\beta Y_{l,m}\big(R^\beta_{cd}x\big)\Big|_{\beta=0} =$$
$$=\partial_\alpha\partial_\beta \sum_m Y_{l,m}\big(R^\alpha_{ab}x\big)Y_{l,m}\big(R^\beta_{cd}x\big)\Big|_{\alpha=\beta=0} =$$
$$=\sqrt{\sharp_l}\partial_\alpha\partial_\beta P_l\Big(\tfrac{1}{\sqrt{d}}x^\top R^{-\alpha}_{ab}R^\beta_{cd}x\Big)\Big|_{\alpha=\beta=0} =$$
$$=\tfrac{\sqrt{\sharp_l}}{d}P_l''(\sqrt{d})\cdot\partial_\alpha x^\top R^{-\alpha}_{ab}x\Big|_{\alpha=0}\cdot\partial_\beta x^\top R^\beta_{cd}x\Big|_{\beta=0} + \sqrt{\tfrac{\sharp_l}{d}}P_l'(\sqrt{d})\cdot x^\top\partial_\alpha R^{-\alpha}_{ab}\Big|_{\alpha=0}\partial_\beta R^\beta_{cd}\Big|_{\beta=0}x =$$
$$=-\tfrac{\sqrt{\sharp_l}}{d}P_l''(\sqrt{d})\cdot x^\top S_{ab}x\cdot x^\top S_{cd}x - \sqrt{\tfrac{\sharp_l}{d}}P_l'(\sqrt{d})\cdot x^\top S_{ab}S_{cd}x =$$
$$=-\sqrt{\tfrac{\sharp_l}{d}}P_l'(\sqrt{d})\Big(\delta_{ad}x_bx_c - \delta_{bd}x_ax_c - \delta_{ac}x_bx_d + \delta_{bc}x_ax_d\Big)$$

This gives

$$\sum_{m,m'}\tau^2_{l,m;l',m'} = \tfrac{\hat\phi_l^2\hat\phi_{l'}^2}{\sharp_l\sharp_{l'}}\tfrac{1}{4l^2(l+d-2)^2}\tfrac{\sqrt{\sharp_l\sharp_{l'}}}{d}P_l'(\sqrt{d})P_{l'}'(\sqrt{d})\sum_{a,b,c,d}\Big(\delta_{ad}x_bx_c - \delta_{bd}x_ax_c - \delta_{ac}x_bx_d + \delta_{bc}x_ax_d\Big)^2 =$$
$$=\tfrac{\hat\phi_l^2\hat\phi_{l'}^2}{\sqrt{\sharp_l\sharp_{l'}}}\tfrac{1}{4dl^2(l+d-2)^2}P_l'(\sqrt{d})P_{l'}'(\sqrt{d})\cdot 4(d-1)r^4$$

We substitute $x=\sqrt{d}$ into the differential equation 19 for Gegenbauer polynomials to deduce

$$P_l'(\sqrt{d}) = \tfrac{l(l+d-2)}{(d-1)\sqrt{d}}P_l(\sqrt{d}) = \tfrac{l(l+d-2)}{(d-1)\sqrt{d}}\sqrt{\sharp_l}$$

Finally

$$\sum_{m,m'}\tau^2_{l,m;l',m'} = \tfrac{\hat\phi_l^2\hat\phi_{l'}^2}{d-1}\tfrac{l'(l'+d-2)}{l(l+d-2)} \tag{8}$$

3.1.4. FINAL BOUND

Let $\Sigma_{l,m;l',m'} = \delta_{ll'}\delta_{mm'}\frac{\hat\phi_l^2}{\sharp_l}$. Equation 8 gives

$$S\Big(\tilde y,\Sigma\Big)^2 \le \big\|\Sigma^{-\frac12}\tau\big\|^2_{HS} = \sum_{l,l'}\tfrac{\sharp_l}{\hat\phi_l^2}\cdot\tfrac{\hat\phi_l^2\hat\phi_{l'}^2}{d-1}\tfrac{l'(l'+d-2)}{l(l+d-2)} = \tfrac{1}{d-1}\left(\sum_{l=1}^k\tfrac{\sharp_l}{l(l+d-2)}\right)\left(\sum_{l=1}^k\hat\phi_l^2 l(l+d-2)\right) \tag{9}$$

Substituting $\sharp_l = \binom{d+l-1}{d-1} - \binom{d+l-3}{d-1}$, for $d\ge 4$ we can bound the first term by

$$\sum_{l=1}^k\tfrac{\sharp_l}{l(l+d-2)} = \sum_{l=1}^k\tfrac{d^2+2dl-3d-3l+2}{l+d-2}\cdot\tfrac{(d+l-3)\dots(l+1)}{(d-1)!} \le k\cdot 2d\cdot\tfrac{(d+k)^{d-3}}{(d-1)!} \le \tfrac{2d(d+k)^{d-2}}{(d-1)!}$$

We can check by hand that this also holds for $d=2,3$. The second term of 9 can be simplified by recalling the orthonormality of P_l with respect to the density of single coordinate (equation 18 from appendix A.2)

$$\int_{-\sqrt{d}}^{\sqrt{d}}P_l(t)P_{l'}(t)\xi(t)dt = \delta_{ll'} \qquad \text{where} \qquad \xi(t) = \tfrac{\Gamma\left(\frac{d}{2}\right)}{\Gamma\left(\frac{d-1}{2}\right)\sqrt{\pi d}}\left(1-\tfrac{t^2}{d}\right)^{\frac{d-3}{2}}$$

which gives

$$\sum_{l=0}^{\infty} \hat{\phi}_l^2 l(l+d-2) = \int_{-\sqrt{d}}^{\sqrt{d}} \left(\sum_{l=0}^{\infty} \hat{\phi}_l P_l \right) \left(\sum_{l=0}^{\infty} \hat{\phi}_l l(l+d-2) P_l \right) \xi(t) dt \tag{10}$$

Recalling the ODE 19 for Gegenbauer polynomials, we note

$$\sum_{l=0}^{\infty} \hat{\phi}_l l(l+d-2) P_l = -(d-t^2)\phi'' + (d-1)t\phi' = -d\left(1-\tfrac{t^2}{d}\right)^{-\frac{d-3}{2}} \left(\left(1-\tfrac{t^2}{d}\right)^{\frac{d-1}{2}} \phi' \right)'$$

substituting this relation and the explicit form of ξ yields

$$\begin{aligned}
\sum_{l=0}^{\infty} \hat{\phi}_l^2 l(l+d-2) =& -\tfrac{\Gamma\left(\frac{d}{2}\right)}{\Gamma\left(\frac{d-1}{2}\right)} \sqrt{\tfrac{d}{\pi}} \int_{-\sqrt{d}}^{\sqrt{d}} \phi(t) \cdot \left(\left(1-\tfrac{t^2}{d}\right)^{\frac{d-1}{2}} \phi' \right)' dt = \\
=& \tfrac{\Gamma\left(\frac{d}{2}\right)}{\Gamma\left(\frac{d-1}{2}\right)} \sqrt{\tfrac{d}{\pi}} \int_{-\sqrt{d}}^{\sqrt{d}} \phi'(t)^2 \left(1-\tfrac{t^2}{d}\right)^{\frac{d-1}{2}} dt \leq \\
\leq& \sqrt{\tfrac{d-1}{2}} \sqrt{\tfrac{d}{\pi}} \int_{-\sqrt{d}}^{\sqrt{d}} \phi'(t)^2 \cdot 2e^{-\frac{t^2}{2}} dt \leq 2\sqrt{d(d-1)} \mathbb{E}\left[\phi'\left(\mathcal{N}(0,1)\right)^2 \right]
\end{aligned}$$

Finally, equation 9 becomes

$$S(\tilde{y}, \Sigma)^2 \leq \tfrac{6d(d+k)^{d-2}}{(d-1)!} \cdot \mathbb{E}\left[\phi'\left(\mathcal{N}(0,1)\right)^2 \right]$$

Now we only need to translate the discrepancy into Wasserstein distance. Lemma 5 implies

$$S(s\tilde{y}, \Sigma)^2 \leq \mathbb{E}\left[s^4\right] S(\tilde{y}, \Sigma)^2 + \operatorname{var}\left[s^2\right] \cdot \|\tilde{y}\|^2$$

The norm of $\tilde{y}$ is $\sum_{l=1}^{k} \hat{\phi}_l^2 \leq \sum_{l=1}^{\infty} \hat{\phi}_l^2 = \mathbb{E}\left[\phi(x_1)^2 \middle| x \sim \sqrt{d} U S^{d-1} \right]$. By corollary 6,

$$\mathcal{W}_2 \left(\tfrac{1}{\sqrt{n}} \sum_{i=1}^{n} s_i \tilde{y}_i, \mathcal{N}(0, \Sigma) \right) \leq \tfrac{1}{\sqrt{n}} S\left(s\tilde{y}, \Sigma\right)$$

And according to the theory from section 2, this translates isometrically to a distance between $\mathcal{P}_n$ and some Gaussian process.

■

4. Non-polynomial activations

Here we obtain approximations of networks with ReLU and erf activations. We do this by truncating the expansion of the activation function into (normalized) Hermite polynomials h_l. This is a family of polynomials orthonormal with respect to the Gaussian weight $\frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$. They can be viewed as the "limit" of Gegenbauer polynomials as $d \to \infty$. We will make use of the generating function for the normalized Hermite polynomials

$$e^{tx - \frac{t^2}{2}} = \sum_{l=0}^{\infty} \frac{h_l(x) t^l}{\sqrt{l!}} \tag{11}$$

Theorem 2 *Assume $d \geq 3$, $s_i \sim U\{-1,1\}$, $w_i \sim \sqrt{d}US^{d-1}$. With $\phi =$ ReLU, for sufficiently large n there exists a Gaussian process $\mathcal{G}$ satisfying*

$$\mathcal{W}_2\left(\mathcal{P}_n, \mathcal{G}\right) \leq 7n^{-\frac{3}{2(2d-1)}}$$

while with $\phi =$ erf

$$\mathcal{W}_2\left(\mathcal{P}_n, \mathcal{G}\right) \leq \sqrt{\frac{e}{\log\frac{3}{2}}}^d \cdot \frac{(\log n)^{\frac{d-2}{2}}}{\sqrt{n}}$$

4.1. Proof

4.1.1. GENERAL ACTIVATIONS

Let $\phi = \sum_{l=0}^{\infty} a_l h_l$ be the expansion of ϕ in the basis of normalized Hermite polynomials. Denote the truncations as $\overline{\phi} = \sum_{l=0}^{k} a_l h_l$ and $\overline{\mathcal{P}}_n(x) = \frac{1}{\sqrt{n}} \sum_{i=1} s_i \overline{\phi}\left(\frac{w_i \cdot x}{\sqrt{d}}\right)$. Then a simple calculation shows

$$\mathcal{W}_2\left(\mathcal{P}_n, \overline{\mathcal{P}}_n\right)^2 = \int_{-\sqrt{d}}^{\sqrt{d}} \left(\phi(t) - \overline{\phi}(t)\right)^2 \xi(t) dt$$

Since $\xi(t) \leq \frac{5}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$ (which follows from the explicit formula 16 in appendix A.1), this is at most

$$5 \int_{-\infty}^{\infty} \left(\phi(t) - \overline{\phi}(t)\right)^2 \cdot e^{-\frac{t^2}{2}} \frac{dt}{\sqrt{2\pi}} = 5 \sum_{l=k+1}^{\infty} a_l^2$$

By theorem 1, the truncated network $\overline{\mathcal{P}}_n$ can be approximated by some Gaussian process $\mathcal{G}$ as

$$\mathcal{W}_2\left(\overline{\mathcal{P}}_n, \mathcal{G}\right) \leq \sqrt{\frac{6d(d+k)^{d-2}}{n(d-1)!} \cdot \mathbb{E}\left[\phi'(x)^2 \middle| x \sim \mathcal{N}(0,1)\right]}$$

Using the triangle inequality, and simplifying $\frac{d(d+k)^{d-2}}{(d-1)!} < \frac{d^{d-1}k^{d-2}}{(d-1)!} < e^{d-1}k^{d-2}$, we obtain

$$\mathcal{W}_2\left(\mathcal{P}_n, \mathcal{G}\right) \leq \sqrt{5 \sum_{l=k+1}^{\infty} a_l^2} + \sqrt{\frac{3e^d k^{d-2}}{n} \cdot \mathbb{E}\left[\phi'(x)^2 \middle| x \sim \mathcal{N}(0,1)\right]} \tag{12}$$

4.1.2. RELU

Using the equation 11, the coefficients of ReLU satisfy

$$\sum_{l=0}^{\infty} \frac{a_l t^l}{\sqrt{l!}} = \int_0^{\infty} x e^{-\frac{(x-t)^2}{2}} \frac{dx}{\sqrt{2\pi}} = \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}} + \frac{t}{2} + \frac{t}{\sqrt{2\pi}} \int_0^t e^{-\frac{x^2}{2}} dx = \frac{1}{\sqrt{2\pi}} + \frac{t}{2} + \frac{1}{\sqrt{2\pi}} \sum_{l=1}^{\infty} \frac{(-1)^{l-1} t^{2l}}{l! \cdot 2^l \cdot (2l-1)}$$

Which means

$$a_0 = \frac{1}{\sqrt{2\pi}} \qquad a_1 = \frac{1}{2} \qquad a_l = \frac{(-1)^{\frac{l}{2}-1}\sqrt{l!}}{\sqrt{2\pi} \cdot \left(\frac{l}{2}!\right) \cdot 2^{\frac{l}{2}} \cdot (l-1)} \cdot \mathbb{1}_{2|l} \qquad \text{for } l > 1$$

By Stirling's formula $a_l^2 \sim \frac{1}{\pi\sqrt{2\pi}} l^{-\frac{5}{2}}$. Therefore for large enough l we have $a_l^2 < \frac{1}{7} l^{-\frac{5}{2}}$, and as a consequence $\sum_{l=k+1}^{\infty} a_l^2 < \frac{1}{7} \sum_{l=k+1}^{\infty} l^{-\frac{5}{2}} < \frac{1}{7} \int_k^{\infty} l^{-\frac{5}{2}} dl = \frac{2}{21} k^{-\frac{3}{2}}$. Inequality 12 becomes

$$\mathcal{W}_2\left(\mathcal{P}_n, \mathcal{G}\right) \leq k^{-\frac{3}{4}} + \sqrt{\frac{2e^d k^{d-2}}{n}}$$

Picking $\frac{1}{3} n^{\frac{2}{2d-1}} < k < e^{-1} n^{\frac{2}{2d-1}}$ makes the two terms be of comparable order, and gives

$$\mathcal{W}_2\left(\mathcal{P}_n, \mathcal{G}\right) \leq 3n^{-\frac{3}{2(2d-1)}} + \sqrt{2e^d e^{-d+2} n^{-\frac{3}{2d-1}}} < 7n^{-\frac{3}{2(2d-1)}}$$

4.1.3. ERF

Again, using the generating function 11 we find

$$\frac{\partial}{\partial t} \int_{-\infty}^{\infty} \operatorname{erf}(x) \cdot e^{tx - \frac{t^2}{2}} \cdot e^{-\frac{x^2}{2}} \frac{dx}{\sqrt{2\pi}} = \sum_{l=0}^{\infty} \frac{l t^{l-1}}{\sqrt{l!}} a_l$$

On the other hand

$$\begin{aligned}
\frac{\partial}{\partial t} \int_{-\infty}^{\infty} \operatorname{erf}(x) \cdot e^{-\frac{(x-t)^2}{2}} \frac{dx}{\sqrt{2\pi}} &= \int_{-\infty}^{\infty} \operatorname{erf}(x) \cdot \left(- \frac{\partial}{\partial x} e^{-\frac{(x-t)^2}{2}} \right) \frac{dx}{\sqrt{2\pi}} = \\
&= - \operatorname{erf}(x) e^{-\frac{(x-t)^2}{2}} \Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \operatorname{erf}'(x) \cdot e^{-\frac{(x-t)^2}{2}} \frac{dx}{\sqrt{2\pi}} = \\
&= \frac{\sqrt{2}}{\pi} \int_{-\infty}^{\infty} e^{-\frac{3x^2}{2} + xt - \frac{t^2}{2}} dx = \frac{2}{\sqrt{3\pi}} e^{-\frac{t^2}{3}}
\end{aligned}$$

Comparing the coefficients we obtain $a_l = \frac{2(-1)^{\frac{l-1}{2}} \sqrt{(l-1)!}}{\sqrt{\pi l} \sqrt{3}^l \left(\frac{l-1}{2}!\right)} \cdot \mathbb{1}_{2 \nmid l}$.

From Stirling's formula $a_l^2 \sim \left(\frac{2}{\pi l}\right)^{\frac{3}{2}} \left(\frac{2}{3}\right)^l$, so eventually $a_l^2 < \left(\frac{2}{3}\right)^l$ and $\sum_{l=k+1}^{\infty} a_l^2 < 2\left(\frac{2}{3}\right)^k$. Then equation 12 together with $|\operatorname{erf}'| \leq 1$ give

$$\mathcal{W}_2\left(\mathcal{P}_n, \mathcal{G}\right) \leq \sqrt{10} \left(\frac{2}{3}\right)^{\frac{k}{2}} + \sqrt{\frac{3e^d k^{d-2}}{n}}$$

Setting $k \sim \frac{\log n}{\log \frac{3}{2}}$ completes the proof.

■

5. Discussion

We have demonstrated that one-hidden-layer neural networks with polynomial activation approach GPs at the rate $O(n^{-\frac{1}{2}})$ in 2-Wasserstein distance. A natural question to ask is how far can our result be generalized. Can the condition of a polynomial activation be dropped? Can we retain a polynomial dependence on the input dimension from the bounds in Eldan et al. (2021)? How about p-Wasserstein metrics for $p > 2$? One source of difficulty with these questions seems to originate from Σ^{-1} in the definition of S in lemma 7. This factor does not appear in the isotropic case,

and for general covariances one may define S in a few different ways and still obtain bounds on $\mathcal{W}$. However, dimensional analysis suggests that Σ^{-1} is more than just an artifact of a particular wording of Cauchy-Schwarz inequality: if we scale every variable by λ, then $\mathcal{W}(X, \mathcal{N})^2$ scales like λ^2 but plain $\|\tau - \Sigma\|_{HS}^2$ scales like λ^4. In our proof, construction of the kernel and relating the distance to discrepancy are largely independent, so we hope that deeper understanding of the relationship between Wasserstein distance and Stein discrepancy will allow to improve our result for little extra effort.

In classical CLT the convergence of a normalized sum to a Gaussian is not faster than $O(n^{-\frac{1}{2}})$, provided that the variables being averaged have non-zero fourth cumulant. Therefore the bound from theorem 1 is asymptotically sharp in n. As a concrete example, using notation from thm 1, one can take $\mathbb{P}(s = -1) = \mathbb{P}(s = 1) = \frac{1}{2}$ and $\phi = \mathrm{id}$. Then the 2-Wasserstein distance of the resulting neural network $\mathcal{P}_n$ to any GP is not smaller that the minimal distance of its coefficient $X_{1,1}$ to a normal random variable. It is not too difficult to see that $X_{1,1} = \frac{1}{\sqrt{nd}} \sum_{i=1}^{n} (w_i)_1$, where $(w_i)_1$ is the first coordinate of w_i. The cumulant of $(w_i)_1$ is equal $\mathbb{E}[(w_i)_1^4] - 3\mathbb{E}[(w_i)_1^2]^2 = -\frac{6}{d+2}$, so the cumulant of $X_{1,1}$ is $-\frac{6}{nd(d+2)}$. Since cumulant of any Gaussian is zero, it is possible to lower-bound the distance to a normal in terms of the cumulant, which then shows that the rate $O(n^{-\frac{1}{2}})$ cannot be improved.

Acknowledgments

The author would like to thank prof. Guang Cheng for introducing this problem to him, and Dr Hafiz Tiomoko Ali and Dr Diego Granziol for comments and feedback on the manuscript.

References

Kyle Aitken and Guy Gur-Ari. On the asymptotics of wide networks with polynomial activations. *arXiv preprint arXiv:2006.06687*, 2020.

Ehsan Azmoodeh, Giovanni Peccati, and Xiaochuan Yang. Malliavin–stein method: a survey of some recent developments. *Modern Stochastics: Theory and Applications*, 8(2):141–177, 2021.

Thomas A Courtade, Max Fathi, and Ashwin Pananjady. Existence of stein kernels under a spectral gap, and discrepancy bounds. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 55, pages 777–790. Institut Henri Poincaré, 2019.

Ethan Dyer and Guy Gur-Ari. Asymptotics of wide networks from feynman diagrams. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1gFvANKDS.

Ronen Eldan, Dan Mikulincer, and Tselil Schramm. Non-asymptotic approximations of neural networks by gaussian processes. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 1754–1775. PMLR, 15–19 Aug 2021. URL https://proceedings.mlr.press/v134/eldan21a.html.

Boris Hanin. Random neural networks in the infinite width limit as gaussian processes. *arXiv preprint arXiv:2107.01562*, 2021.

Michel Ledoux, Ivan Nourdin, and Giovanni Peccati. Stein's method, logarithmic sobolev and transport inequalities. *Geometric and Functional Analysis*, 25(1):256–306, 2015.

Guillaume Mijoule, Gesine Reinert, and Yvik Swan. Stein operators, kernels and discrepancies for multivariate continuous distributions. *arXiv preprint arXiv:1806.03478*, 2018.

Radford M Neal. Priors for infinite networks. In *Bayesian Learning for Neural Networks*, pages 29–53. Springer, 1996.

Felix Otto and Cédric Villani. Generalization of an inequality by talagrand and links with the logarithmic sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000.

Charles Stein. Approximate computation of expectations. IMS, 1986.

Sho Yaida. Non-gaussian processes and neural networks at finite widths. In *Mathematical and Scientific Machine Learning*, pages 165–192. PMLR, 2020.

Appendix A. Spherical harmonics

There are at least three equivalent ways to think about spherical harmonics

- Algebraic: harmonic (i.e. $\nabla^2 Y = 0$) homogeneous polynomials in d variables
- Representation-theoretic: irreducible representations of $SO(d)$
- Analytic: basis of the Hilbert space $\mathcal{L}^2(\sqrt{d}S^{d-1})$ of functions on the sphere

In the discussion below, we will typically start with the algebraic picture, viewing polynomials as elements of the ring $\mathbb{R}[X_1, \dots, X_d]$ and operators L_{ab} as $\mathbb{R}$-linear derivations[5] over this ring. Then we proceed to the analytic picture – remind ourselves that polynomials can be treated as functions on the sphere $\sqrt{d}S^{d-1}$, and think of L_{ab} as infinitesimal generators of rotations; we translate the algebraic results and explore the consequences of acquiring an inner product. Representation-theoretic picture will be present in the background and manifest itself whenever we talk about the symmetries of spherical harmonics.

A.1. Rotations and operators

Special orthogonal group $SO(d)$ acts on points from $\sqrt{d}S^{d-1}$ by $R : x \mapsto Rx$, preserving geometry $x_1 . x_2$. This induces an action on functions $\mathcal{L}^2(\sqrt{d}S^{d-1})$ by $R : f \mapsto f \circ R^{-1}$, which preserves the inner product

$$\int_{\sqrt{d}S^{d-1}} f_1(x) f_2(x) dx = \mathbb{E}\big[f_1(x) f_2(x) \big| x \sim \sqrt{d}US^{d-1}\big] \qquad \text{with normalization } \int 1 dx = 1 \tag{13}$$

5. Operator L is a derivation if it satisfies the Leibniz rule $L(fg) = Lf \cdot g + f \cdot Lg$

The group is generated by rotations of the form $R^\alpha_{ab} = \exp(\alpha S_{ab})$ for anti-symmetric matrices $(S_{ab})_{ij} = \delta_{ai}\delta_{bj} - \delta_{aj}\delta_{bi}$. They act on the basis vectors as

$$\begin{aligned} R^\alpha_{ab} e_a &= \cos\alpha\ e_a - \sin\alpha\ e_b \\ R^\alpha_{ab} e_b &= \sin\alpha\ e_a + \cos\alpha\ a_b \\ R^\alpha_{ab} e_c &= e_c \quad \text{when} \quad c \notin \{a,b\} \end{aligned}$$

The infinitesimal generators of such rotations are

$$\partial_\alpha R^\alpha_{ab} f\Big|_{\alpha=0} = L_{ab} f \qquad \text{where} \qquad L_{ab} = X_a\partial_b - X_b\partial_a \tag{14}$$

In particular, $\mathbb{E}[L_{ab}f(x)|x \sim \sqrt{d}US^{d-1}] = \partial_\alpha \mathbb{E}[R^\alpha_{ab} f]\big|_{\alpha=0} = 0$.

Define the Laplace-Beltrami operator as

$$L^2 \stackrel{\text{def}}{=} \sum_{a<b} L^2_{ab} = \tfrac{1}{2}\sum_{a,b} L^2_{ab}$$

it is straightforward to verify

$$r^2\nabla^2 = L^2 + \partial_r(\partial_r + d - 2) \tag{15}$$

where

$$\partial_r \stackrel{\text{def}}{=} x\cdot\nabla = \sum_{i=1}^d X_i\partial_i$$

Let us note the algebraic properties of these operators. The simplest is ∂_r – it multiplies a polynomial by its degree. Operators L_{ab} are derivations annihilating $r^2 \stackrel{\text{def}}{=} X_1^2 + \cdots + X_d^2$, so Laplace-Beltrami operator satisfies $L^2(r^2 f) = r^2 L^2 f$, and by identity 15 is multiplies harmonic homogeneous polynomials of degree l by $-l(l+d-2)$. Both ∂_r and L^2 are invariant under rotations (equivalently, commute with each L_{ab}).

Now we describe their basic analytic properties. The operators L_{ab} are tangent[6] to the sphere, so they and L^2 have well-defined restrictions[7] to $\mathcal{L}^2(\sqrt{d}S^{d-1})$. Inner product of functions is invariant under rotations and L_{ab} obey Leibniz rule, so they are anti-self-adjoint

$$\int_{\sqrt{d}S^{d-1}} L_{ab}f_1 \cdot f_2 dx = -\int_{\sqrt{d}S^{d-1}} f_1 \cdot L_{ab} f_2 dx$$

As a consequence, L^2 is self-adjoint (with respect to 13). Also, L_{ab} annihilate the constant function, so again $\mathbb{E}[L_{ab}f(x)|x \sim \sqrt{d}US^{d-1}] = 0$.

Finally let us note a geometric fact about the sphere. It will be usefult later to know the distribution of a single coordinate x_1 when we draw x uniformly from the sphere $\sqrt{d}S^{d-1}$. Its density is supported on the interval $[-\sqrt{d}, \sqrt{d}]$ and equals

$$\xi(x) = \frac{\Gamma\left(\frac{d}{2}\right)}{\Gamma\left(\frac{d-1}{2}\right)\sqrt{\pi d}}\left(1 - \frac{x^2}{d}\right)^{\frac{d-3}{2}} \tag{16}$$

6. In the sense $L_{ab}r^2 = 0$, or $L_{ab} = v\cdot\nabla$ where the vector field v is tangent to $\sqrt{d}S^{d-1}$

7. Any derivation J on $\mathbb{R}[X_1,\ldots,X_d]$ annihilating r^2 gives rise to an operator on $\mathcal{L}^2(\sqrt{d}S^{d-1})$ as follows. For a function f that is the restriction of a polynomial F to $\sqrt{d}S^{d-1}$ we send $f \mapsto J(F)\big|_{\sqrt{d}S^{d-1}}$. This is well defined, because if $F\big|_{\sqrt{d}S^{d-1}} = F'\big|_{\sqrt{d}S^{d-1}}$ then $F - F' = (r^2 - d)G$ for some G, so $J(F-F')$ is the zero function on $\sqrt{d}S^{d-1}$. Conversely, a differential operator $\sum_{i=1}^d P_i\partial_i$ acting on $\mathcal{L}^2(\sqrt{d}S^{d-1})$ with $P_i \in \mathbb{R}[X_1,\ldots,X_d]$ can be naturally reinterpreted as a derivation on $\mathbb{R}[X_1,\ldots,X_d]$.

One way to see this is by noting that $\frac{x_1^2}{d}$ for $x \sim \sqrt{d}US^{d-1}$ has the same distribution as $\frac{z_1^2}{\|z\|^2}$ for $z \sim \mathcal{N}(0, I_d)$, that is B $\left(\frac{1}{2}, \frac{d-1}{2}\right)$.

A.2. Spherical harmonics

Lemma 3 *Every homogeneous polynomial $f \in \mathbb{R}[X_1, \ldots, X_d]$ can be uniquely written as*

$$f = f_0 + r^2 f_1 + r^4 f_2 + \ldots$$

where f_i are homogeneous harmonic polynomials.

Proof We proceed by induction on $l = \deg f$. For $l = 0, 1$ the statement is trivial. For $l \geq 1$, by inductive assumption we may write

$$\nabla^2 f = g_0 + r^2 g_1 + \cdots + r^{2\left\lfloor \frac{l-2}{2} \right\rfloor} g_{\left\lfloor \frac{l-2}{2} \right\rfloor}$$

for harmonic g_i of degree $l - 2 - 2i$. Now, construct

$$g = \sum_{i=0}^{\left\lfloor \frac{l-2}{2} \right\rfloor} \frac{r^{2i} g_i}{2(i+1)(2l - 2i - 4 + d)}$$

then, either by writing $\nabla^2(r^2 \bullet) = L^2 + \partial_r^2 + (d+2)\partial_r + 2d$ and recalling the eigenvalues of L^2 and ∂_r, or by direct calculation, we can see that

$$\nabla^2(r^2 g) = \sum_{i=0}^{\left\lfloor \frac{l-2}{2} \right\rfloor} r^{2i} g_i = \nabla^2 f$$

Therefore, the Laplacian of $f - r^2 g$ is zero, and f can be decomposed as

$$f = \underbrace{\left(f - r^2 g\right)}_{f_0} + \sum_{i=0}^{\left\lfloor \frac{l-2}{2} \right\rfloor} r^{2i+2} \frac{g_i}{2(i+1)(2l - 2i - 4 + d)}$$

To see uniqueness, note that each factor is an eigen-element of L^2 with a different eigenvalue. ∎

Let us denote the space of degree l homogeneous harmonic polynomials as H_l. By the lemma 3 above, we have

$$\{\deg\text{-}l \text{ homog polys}\} = H_l \ \oplus \ r^2\{\deg\text{-}(l-2) \text{ homog polys}\} = H_l \oplus r^2 H_{l-2} \oplus r^4 H_{l-4} \oplus \ldots$$

this allows to deduce their dimensions

$$\text{ đ}_l \stackrel{\text{def}}{=} \dim H_l = \binom{d+l-1}{d-1} - \binom{d+l-3}{d-1}$$

Harmonicity and homogeneity of given degree are preserved by rotations, so each H_l is closed under $SO(d)$, and each $r^{2k} H_l$ is a subrepresentation of $SO(d)$ inside $\mathbb{R}[X_1, \ldots, X_d]$. Note that by

equation 15 we have $H_l = \ker L^2 + l(l+d-2)$ (in algebraic sense, with X_i considered as abstract symbols).

Now let us think about restrictions of polynomials $\mathbb{R}[X_1, \dots, X_n]$ to $\sqrt{d}S^{d-1}$. By Stone-Weierstrass theorem, they are dense in $C(\sqrt{d}S^{d-1}, \mathbb{R})$ with supremum norm. Thus with ℓ_2-norm we must have

$$\mathcal{L}^2\big(\sqrt{d}S^{d-1}\big) = \overline{\bigoplus_{l=0}^{\infty} H_l}$$

with each H_l closed under $SO(d)$. Also, $H_l = \ker L^2 + l(l+d-2)$ (in analytic sense, with H_l considered as functions on the sphere and L^2 as a second-order differential operator); the operator L^2 is self-adjoint, so different H_l are orthogonal.

We take spherical harmonics $Y_{l,1}, \dots, Y_{l,\sharp_l}$ to be any orthonormal basis of H_l. Then

$$\mathbb{E}\big[Y_{l,m}(w)Y_{l',m'}(w)\big|w \sim \sqrt{d}US^{d-1}\big] = \delta_{ll'}\delta_{mm'}$$

Each H_l comes with a representation ρ of $SO(d)$

$$RY_{l,m} = Y_{l,m}(R^{-1}x) = \sum_{m'=1}^{\sharp_l} \rho(R)_{m,m'}Y_{l,m'}(x)$$

Such matrices $\rho(R)$ are also orthogonal, which follows from the invariance of the inner product:

$$\begin{aligned}\big(\rho(R)\rho(R)^\top\big)_{m,m''} =& \sum_{m'} \rho(R)_{m,m'}\rho(R)_{m'',m'} = \sum_{m',m'''} \rho(R)_{m,m'}\rho(R)_{m'',m'''}\big\langle Y_{l,m'}, Y_{l,m'''}\big\rangle = \\ =& \big\langle RY_{l,m}, RY_{l,m''}\big\rangle = \big\langle Y_{l,m}, Y_{l,m''}\big\rangle = \delta_{m,m''}\end{aligned}$$

Now we look at the relation between spherical harmonics at different points, which will eventually lead to Gegenbauer polynomials. Consider

$$\check{P}_{l,x}(x') \stackrel{\text{def}}{=} \tfrac{1}{\sqrt{\sharp_l}} \sum_{m=1}^{\sharp_l} Y_{l,m}(x)Y_{l,m}(x')$$

By construction $\check{P}_{l,x} \in H_l$. Also, for any rotation $R \in SO(d)$ we have

$$\begin{aligned}\check{P}_{l,Rx}(Rx') =& \tfrac{1}{\sqrt{\sharp_l}} Y_{l,:}(Rx)^\top Y_{l,:}(Rx') = \\ =& \tfrac{1}{\sqrt{\sharp_l}} Y_{l,:}(x)^\top \rho(R^{-1})^\top \rho(R^{-1}) Y_{l,:}(x') = \\ =& \tfrac{1}{\sqrt{\sharp_l}} Y_{l,:}(x)^\top Y_{l,:}(x') = \\ =& \check{P}_{l,x}(x')\end{aligned}$$

Therefore $\check{P}_{l,x}(x')$ depends only on the angle between x, x' and not on their absolute position on the sphere, i.e. $P_{l,x}(x') = P_l\left(\frac{x \cdot x'}{\sqrt{d}}\right)$ for some function P_l; it must be a polynomial of degree at most l. This gives us the key identity

$$P_l\left(\tfrac{x \cdot x'}{\sqrt{d}}\right) = \tfrac{1}{\sqrt{\sharp_l}} \sum_{m=1}^{\sharp_l} Y_{l,m}(x)Y_{l,m}(x') \tag{17}$$

The P_l are called Gegenbauer polynomials[8]. They are the unique (up to scaling) functions for which the map $x \mapsto P_l(x_1)$ belongs to H_l. Orthogonality of the spaces H_l for different l means that Gegenbauer polynomials are orthogonal with respect to the single-coordinate density ξ from equation 16

$$\begin{aligned}
P_l(\sqrt{d}) =& \tfrac{1}{\sqrt{\not{d}_l}} \sum_{m=1}^{\not{d}_l} \mathbb{E}\left[Y_{l,m}(x)Y_{l,m}(x)\right] = \sqrt{\not{d}_l} \\
\int_{-\sqrt{d}}^{\sqrt{d}} P_l(t)P_{l'}(t)\cdot\xi(t)dt =& \mathbb{E}\left[P_l(X_i)P_{l'}(X_i)\right] = \\
=& \tfrac{1}{\sqrt{\not{d}_l \not{d}_{l'}}} \sum_{m,m'} Y_{l,m}(\sqrt{d}e_i)Y_{l',m'}(\sqrt{d}e_i)\mathbb{E}\left[Y_{l,m}(x)Y_{l',m'}(x)\right] = \\
=& \tfrac{1}{\sqrt{\not{d}_l}}\delta_{ll'}P_l(\sqrt{d}) = \delta_{ll'}
\end{aligned} \tag{18}$$

therefore P_l can be computed by Gram-Schmidt orthonormalization of $\{t^0, t^1, t^2, \dots\}$ with respect to the the density of a single coordinate ξ.

Finally we exhibit an ODE for P_l. Observe that $r^l P_l\left(\frac{X_1\sqrt{d}}{r}\right)$ is a homogeneous degree-l harmonic polynomial. After tidying up the harmonicity condition we obtain

$$0 = \tfrac{1}{r^{l-2}}\nabla^2\left(r^l P_l\left(\tfrac{X_1\sqrt{d}}{r}\right)\right) = (d-t^2)P_l''(t) - (d-1)tP_l'(t) + l(l+d-2)P_l(t) \tag{19}$$

Example 1 *For $d = 2$ this construction is precisely the Fourier analysis. We work over $\sqrt{2}S^1 = \{(x_1, x_2) : x_1^2 + x_2^2 = 2\}$, parameterized as $x_1 = \sqrt{2}\cos\theta, x_2 = \sqrt{2}\sin\theta$. Harmonic subspaces are*

$$\begin{aligned}
H_0 &= \operatorname{span}\{Y_{0,1} = 1\} && \text{with} && \not{d}_0 = 1 \\
H_l &= \operatorname{span}\{Y_{l,1} = \sqrt{2}\cos l\theta, Y_{l,2} = \sqrt{2}\sin l\theta\} && \text{with} && \not{d}_l = 2
\end{aligned}$$

spherical harmonics are restrictions of polynomials

$$\begin{aligned}
Y_{l,1} =& 2^{\frac{1-l}{2}}\Re(X_1 + iX_2)^l = \tfrac{r^l}{\sqrt{2}^{l-1}}\cos l\theta \\
Y_{l,2} =& 2^{\frac{1-l}{2}}\Im(X_1 + iX_2)^l = \tfrac{r^l}{\sqrt{2}^{l-1}}\sin l\theta
\end{aligned}$$

There is only one rotation generator

$$\begin{aligned}
L_{12} =& \partial_\theta = X_1\partial_2 - X_2\partial_1 \\
L^2 =& \partial_\theta^2 = X_1^2\partial_2^2 + X_2^2\partial_1^2 - 2X_1X_2\partial_1\partial_2 - X_1\partial_1 - X_2\partial_2
\end{aligned}$$

and the Laplace-Beltrami operator L^2 acts on H_l as a multiplication by $-l^2$.

Gegenbauer polynomials are characterized by

$$P_l\big(\sqrt{2}\cos(\theta - \theta')\big) = \sqrt{2}\cos l\theta\cos l\theta' + \sqrt{2}\sin l\theta\sin l\theta' = \sqrt{2}\cos l(\theta - \theta')$$

i.e. are rescaled Chebyshev polynomials. They are orthonormal with respect to $\xi(t) = \frac{dt}{\pi\sqrt{2-t^2}} = \frac{d\theta}{\pi}$.

8. Different scaling/normalization conventions are used the literature

Appendix B. Stein kernels

We say that τ is a Stein kernel for random variable X if for each $f \in C_c^\infty$ we have

$$\mathbb{E}\big[X.f(X)\big] = \mathbb{E}\big\langle \tau(X), \mathrm{Jac} f(X)\big\rangle_{HS}$$

where $(\mathrm{Jac} f)_{ab} = \frac{\partial f_a}{\partial X_b}$ is the Jacobian of f, and $\langle A, B\rangle_{HS} = \mathrm{Tr}\ AB^\top$ is the Hilbert-Schmidt product.

One can show that a constant matrix Σ is a Stein kernel for X if and only if $X \sim \mathcal{N}(0, \Sigma)$ (this statement is known as Stein's lemma). It turns out that the difference between τ and Σ can be used to bound the Wasserstein distance between X and $\mathcal{N}(0, \Sigma)$ (see lemma 7). The measure of deviation is called Stein discrepancy, and in the isotropic case it is defined as $S(X, \mathrm{Id}) = \inf_\tau \mathbb{E}\|\tau(X) - \mathrm{Id}\|_{HS}^2$. We will be working with non-isotropic random variables, and following the formulation of lemma 7 we generalize the Stein discrepancy as

$$S(X, \Sigma) \stackrel{\text{def}}{=} \inf_\tau \sqrt{\mathbb{E}\big\|\Sigma^{-\frac{1}{2}}(\tau(X) - \Sigma)\big\|_{HS}^2}$$

However, note that other generalizations to non-isotropic case are also possible, and modifying the last part of the proof[9] of 7 can give bounds of a different form.

Substituting $f(X) = X_i e_j$ we see that $\mathbb{E}\tau = \mathbb{E}XX^\top = \mathrm{cov}[X]$. Therefore, Stein discrepancy can also be viewed as a measure of variance of τ. Intuitively, as we average independent copies of X, we can expect the variance to decrease and τ to approach its expectation, leading to central limit theorem. This intuition is formalized in corollary 6; a stronger result – that $nS\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i\right)$ is non-increasing in n – was proved in Courtade et al. (2019).

B.1. Addition and scaling

Lemma 4 *Suppose $\tau_1, \dots, \tau_n$ are Stein kernels for independent $X_1, \dots, X_n$, and write $\bar{X} = \sum_{i=1}^n X_i$. Then*

$$\tau(x) \stackrel{def}{=} \mathbb{E}\left[\sum_i \tau_i(X_i) \middle| \bar{X} = x\right]$$

is a Stein kernel for $\bar{X}$. If X_i have the same covariance Σ, then $S\big(\bar{X}, n\Sigma\big)^2 \le \frac{1}{n}\sum_{i=1}^n S\big(X_i, \Sigma\big)^2$.

Proof

$$\begin{aligned}
\mathbb{E}\left[\bar{X}.f\big(\bar{X}\big)\right] &= \sum_i \mathbb{E}\left[X_i.f\big(\bar{X}\big)\right] = \\
&= \sum_i \mathbb{E}\left\langle \tau_i\big(X_i\big), \mathrm{Jac} f\big(\bar{X}\big)\right\rangle_{HS} = \\
&= \mathop{\mathbb{E}}_{\bar{X}}\left\langle \mathbb{E}\left[\sum_i \tau_i(X_i)\middle|\bar{X}\right], \mathrm{Jac} f\big(\bar{X}\big)\right\rangle_{HS}
\end{aligned}$$

9. For example by rearranging the equation 25 before applying Cauchy-Schwarz inequality

If all covariances are equal then we have

$$\begin{aligned}
S\big(\bar{X}, n\Sigma\big)^2 = \mathop{\mathbb{E}}_{\bar{X}} \left\| (n\Sigma)^{-\frac{1}{2}} \left(\mathbb{E}\left[\sum_i \tau_i\big(X_i\big) \middle| \bar{X} \right] - n\Sigma \right) \right\|_{HS}^2 = \\
= \tfrac{1}{n} \mathop{\mathbb{E}}_{\bar{X}} \left\| \mathbb{E}\left[\sum_i \Sigma^{-\frac{1}{2}} \Big(\tau_i\big(X_i\big) - \Sigma \Big) \middle| \bar{X} \right] \right\|_{HS}^2 \leq \\
\leq \tfrac{1}{n} \mathbb{E} \left\| \sum_i \Sigma^{-\frac{1}{2}} \Big(\tau_i\big(X_i\big) - \Sigma \Big) \right\|_{HS}^2 = \\
= \tfrac{1}{n} \sum_i \mathbb{E} \left\| \Sigma^{-\frac{1}{2}} \Big(\tau_i\big(X_i\big) - \Sigma \Big) \right\|_{HS}^2
\end{aligned}$$

■

Lemma 5 *If τ is a Stein kernel for X and $Y = sX$, then $\tau'(y) = \mathbb{E}\big[s^2 \tau(X) \big| Y = y\big]$ is a Stein kernel for Y. Its discrepancy is at most*

$$S\big(sX, \mathbb{E}[s^2]\mathrm{cov}[X]\big)^2 \leq \tfrac{\mathbb{E}s^4}{\mathbb{E}s^2} \cdot S(X, \mathrm{cov}[X])^2 + \tfrac{\mathrm{var}[s^2]}{\mathbb{E}s^2} \cdot \mathbb{E}\|X\|^2$$

Proof

$$\begin{aligned}
\mathbb{E}\big[sX \boldsymbol{.} f(sX)\big] = \mathop{\mathbb{E}}_{X} \big[X \boldsymbol{.} \mathop{\mathbb{E}}_{s} \big[sf(sX) \big] \big] = \mathop{\mathbb{E}}_{X} \Big\langle \tau(X), \mathrm{Jac} \mathop{\mathbb{E}}_{s} \big[sf(sX) \big] \Big\rangle_{HS} = \\
= \mathop{\mathbb{E}}_{X} \Big\langle \tau(X), \mathop{\mathbb{E}}_{s} \big[s^2 \big(\mathrm{Jac} f \big)(sX) \big] \Big\rangle_{HS} = \mathop{\mathbb{E}}_{X,s} \Big\langle s^2 \tau(X), \big(\mathrm{Jac} f \big)(sX) \Big\rangle_{HS} = \\
= \mathop{\mathbb{E}}_{Y} \Big\langle \mathbb{E}\big[s^2 \tau(X) \big| Y \big], \mathrm{Jac} f(Y) \Big\rangle_{HS} = \mathbb{E} \Big\langle \tau'(Y), \mathrm{Jac} f(Y) \Big\rangle_{HS}
\end{aligned}$$

Now we will bound its discrepancy. Denote $\mathbb{E}s^2 = \sigma^2, \mathrm{cov}[X] = \Sigma$. Then

$$\begin{aligned}
S\big(sX, \mathbb{E}\big[s^2\big]\mathrm{cov}[X]\big)^2 \leq \mathop{\mathbb{E}}_{Y} \left\| \mathbb{E}\Big[\sigma^{-1} \Sigma^{-\frac{1}{2}} \big(s^2 \tau(X) - \sigma^2 \Sigma \big) \Big| Y \Big] \right\|_{HS}^2 \leq \\
\leq \mathbb{E} \left\| \sigma^{-1} \Sigma^{-\frac{1}{2}} \big(s^2 \tau(X) - \Sigma \big) \right\|_{HS}^2 = \\
= \sigma^{-2} \mathbb{E}\big[s^4\big] \mathbb{E} \left\| \Sigma^{-\frac{1}{2}} \tau(X) \right\|_{HS}^2 - \sigma^{-2} \left\| \Sigma^{\frac{1}{2}} \right\|_{HS}^2 = \\
= \tfrac{\mathbb{E}s^4}{\sigma^2} S\big(X, \Sigma\big)^2 + \tfrac{\mathbb{E}s^4 - \sigma^4}{\sigma^2} \left\| \Sigma^{\frac{1}{2}} \right\|_{HS}^2
\end{aligned}$$

and we simplify $\left\| \Sigma^{\frac{1}{2}} \right\|_{HS}^2 = \mathrm{Tr}\, \Sigma = \mathbb{E}\|X\|^2$.

■

Corollary 6 *Suppose X_i are iid with Stein kernel τ and covariance Σ, and let $\bar{X} = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i$ be the normalized sum. Then we get a quantitative central limit theorem by combining lemmas 7, 5 (for constant $s = \frac{1}{\sqrt{n}}$), and 4*

$$\mathcal{W}_2\big(\bar{X}, \mathcal{N}(0, \Sigma)\big) \leq S\big(\bar{X}, \Sigma\big) \leq \tfrac{1}{\sqrt{n}} S\Big(\textstyle\sum_{i=1}^n X_i, n\Sigma \Big) \leq \tfrac{1}{\sqrt{n}} S\big(X_i, \Sigma\big)$$

B.2. Wasserstein bound in non-isotropic case

Lemma 7 *Suppose that τ is a Stein kernel for random variable X, and Σ is a symmetric positive-definite matrix. Then*

$$\mathcal{W}_2(X, \mathcal{N}(0,\Sigma)) \le S(X,\Sigma) \qquad \textit{where} \qquad S(X,\Sigma)^2 = \mathbb{E}\|\Sigma^{-\frac{1}{2}}\big(\tau(X)-\Sigma\big)\|_{HS}^2$$

This proof is a compilation of Proposition 3.1 from Ledoux et al. (2015) and Lemma 2 from Otto and Villani (2000), additionally keeping track of the covariance matrix. It is based on interpolation of the heat flow along the Ornstein-Uhlenbeck semigroup.

B.3. Proof

Let μ_0, μ_∞ be measures/densities of $X, \mathcal{N}(0,\Sigma)$ respectively, living in D-dimensional space. We will tackle the case when X has a Radon-Nikodym derivative $h = \frac{d\mu_0}{d\mu_\infty}$ with respect to the target normal measure. The general case follows by an approximation argument – see Otto and Villani (2000).

B.3.1. HEAT FLOW SEMIGROUP

Introduce

$$X_t = e^{-t}X + \sqrt{1-e^{-2t}}\mathcal{N}(0,\Sigma)$$

Let μ_t be the measure of X_t and $h_t = \frac{d\mu_t}{d\mu_\infty}$. Define a vector field

$$v_t(x) = \mathbb{E}\left[\frac{dX_t}{dt}\middle| X_t = x\right] \qquad \text{or equivalently} \qquad v_t = -\Sigma\big(\nabla \log h_t\big) \tag{20}$$

Then the density μ_t satisfies the diffusion equation

$$\frac{\partial \mu_t}{\partial t} = -\nabla\cdot(\mu_t v_t) \tag{21}$$

A brute-force way to verify the equivalence of definitions in 20 and the diffusion equation 21 is to plug in the explicit formulas

$$\begin{aligned}
\mu_{0,t}(x_0,x_t) =& (1-e^{-2t})^{-\frac{D}{2}}\mu_0(x_0)\mu_\infty\left(\frac{x_t-e^{-t}x_0}{\sqrt{1-e^{-2t}}}\right)\\
\mu_t(x) =& (1-e^{-2t})^{-\frac{D}{2}}\int \mu_0(y)\mu_\infty\left(\frac{x_t-e^{-t}x_0}{\sqrt{1-e^{-2t}}}\right)dy\\
v_t(x) =& \mu_t(x)^{-1}(1-e^{-2t})^{-\frac{D}{2}}\int \mu_0(y)\mu_\infty\left(\frac{x_t-e^{-t}x_0}{\sqrt{1-e^{-2t}}}\right)\left(\frac{-e^{-2t}x+e^{-t}y}{1-e^{-2t}}\right)dy
\end{aligned} \tag{22}$$

As a consequence of the diffusion equation 21, the density μ_t is transported along the trajectories tangent to v_t. Intuitively, if the norm of v_t is small, then the density needs to "travel a short distance" to move from μ_0 to μ_∞. Formally, lemma 2 from Otto and Villani (2000) states

$$\begin{aligned}
\frac{d^+}{ds}W_2(\mu_t,\mu_{t+s}) \le& \sqrt{\mathbb{E}\|v_t(X_t)\|^2} = \sqrt{\int \|v_t(x)\|^2 d\mu_t(x)}\\
W_2(\mu_0,\mu_\infty) \le& \int_0^\infty \sqrt{\mathbb{E}\|v_t(X_t)\|^2}dt
\end{aligned} \tag{23}$$

In the next part of the proof we bound the flow norm as

$$\sqrt{\mathbb{E}\|v_t(X_t)\|^2} \leq \frac{e^{-2t}}{\sqrt{1-e^{-2t}}} S(X,\Sigma) \tag{24}$$

Substituting this to the inequality 23 and integrating completes the proof of the lemma.

This was all we need to construct the flow. In the next part of the proof we will need a few more properties. We start with changes in expectations under the semigroup. Define

$$P_t f(x) \stackrel{\text{def}}{=} \mathbb{E} f\left(e^{-t}x + \sqrt{1-e^{-2t}}\mathcal{N}(0,\Sigma)\right) = \int f\left(e^{-t}x + \sqrt{1-e^{-2t}}y\right) d\mu_\infty(y)$$

This is called Mehler's formula. It is straightforward to check $P_s P_t = P_{s+t}$ and $\mathbb{E} f(X_t) = \mathbb{E} P_t f(X_0)$. Gaussian integration by parts gives a PDE $\frac{\partial}{\partial t} P_t f = \mathcal{L} P_t f$, where $\mathcal{L} = \sum_{i,j} \Sigma_{ij} \partial_i \partial_j - x.\nabla$. Combining these yields

$$\int f \dot{\mu}_t dx = \frac{d}{ds} \mathbb{E} P_s f(X_t)\Big|_{s=0} = \mathbb{E}(\mathcal{L} f)(X_t) = \int f \left(\sum_{i,j} \Sigma_{ij} \partial_i \partial_j + x.\nabla + D \right) \mu_t dx$$

Thus we must have $\frac{\partial \mu_t}{\partial t} = \left(\sum_{i,j} \partial_i \partial_j + x.\nabla + D \right) \mu_t$, which turns out to be a restatement of 21.

An explicit calculation of $P_t h_0$ turns out to be equivalent to $\frac{\mu_t}{\mu_\infty}$ from formula 22, so $h_t = P_t h_0$. It also satisfies

$$\int f \cdot P_t g d\mu_\infty = \mathbb{E}\left[f(x) g(y) \middle| \begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} \Sigma & e^{-t}\Sigma \\ e^{-t}\Sigma & \Sigma \end{pmatrix}\right)\right] = \int P_t f \cdot g d\mu_\infty$$

and $\nabla P_t f = e^{-t} P_t \nabla f$. The diffusion operator satisfies $\int f \mathcal{L} g d\mu_\infty = -\int (\nabla f)^\top \Sigma (\nabla g) d\mu_\infty$.

B.3.2. BOUND ON THE FLOW NORM

This part of the proof is concerned with proving the inequality 24. We start from the transformations

$$\begin{aligned}
\int \|v_t\|^2 d\mu_t &= \int (\nabla \log h_t)^\top \Sigma^2 (\nabla h_t) d\mu_\infty = \\
&= e^{-t} \int (\nabla \log h_t)^\top \Sigma^2 (P_t \nabla h_0) d\mu_\infty = \\
&= \int (\nabla P_t \log h_t)^\top \Sigma^2 (\nabla h_0) \mu_\infty(x) dx = \\
&= -\int \nabla. \left(\mu_\infty \cdot \Sigma^2 \nabla P_t \log h_t\right) h_0 dx = \\
&= \int \left(x.\Sigma \nabla P_t \log h_t - \nabla.\Sigma^2 \nabla P_t \log h_t\right) h_0 \mu_\infty dx = \\
&= \int \left(x_i \Sigma_{ij} \partial_j P_t \log h_t - \Sigma_{ik}\Sigma_{kj} \partial_i \partial_j P_t \log h_t\right) d\mu_0 = \\
&= \int \left(\tau_{ik}(x) - \Sigma_{ik}\right) \Sigma_{kj} \partial_i \partial_j P_t \log h_t d\mu_0(x)
\end{aligned}$$

where in the last two lines we used the Einstein summation convention. We substitute the identity

$$\partial_i \partial_j P_t \log h_t = \frac{e^{-2t}}{\sqrt{1-e^{-2t}}} \int \left(\Sigma^{-1} y\right)_i \left(\partial_j \log h_t\right) \left(e^{-t}x + \sqrt{1-e^{-2t}}y\right) d\mu_\infty(y)$$

to get

$$\begin{aligned} \mathbb{E}\|v_t(X_t)\|^2 = \frac{e^{-2t}}{\sqrt{1-e^{-2t}}} \iint \left(y^\top \Sigma^{-1}\left(\tau(x)-\Sigma\right)\right) \times \left(\Sigma \nabla \log h_t\right) \times \\ \times \left(e^{-t}x + \sqrt{1-e^{-2t}}y\right) d\mu_0(x) d\mu_\infty(y) \end{aligned} \tag{25}$$

By Cauchy-Schwarz inequality the integral is at most

$$\sqrt{\iint \left\|y^\top \Sigma^{-1}\left(\tau(x)-\Sigma\right)\right\|^2 d\mu_0(x) d\mu_\infty(y)} \times \tag{26}$$

$$\times \sqrt{\iint \left\| \left(\Sigma \nabla \log h_t\right) \left(e^{-t}x + \sqrt{1-e^{-2t}}y\right)\right\|^2 d\mu_0(x) d\mu_\infty(y)} \tag{27}$$

Expression under the root in 26 equals

$$\begin{aligned} \iint y_i y_j \Sigma^{-1}_{ik} \Sigma^{-1}_{jl} \left(\tau(x)-\Sigma\right)_{km} \left(\tau(x)-\Sigma\right)_{lm} d\mu_0(x) d\mu_\infty(y) = \\ = \int \Sigma^{-1}_{kl} \left(\tau(x)-\Sigma\right)_{km} \left(\tau(x)-\Sigma\right)_{lm} d\mu_0(x) = \\ = \int \left\|\Sigma^{-\frac{1}{2}}\left(\tau(x)-\Sigma\right)\right\|^2_{HS} d\mu_0(x) = S(X,\Sigma)^2 \end{aligned}$$

while the expression under the root in 27 is

$$\int P_t\left(\|v_t\|^2\right) d\mu_0 = \int \|v_t\|^2 \cdot P_t h_0 d\mu_\infty = \int \|v_t\|^2 d\mu_t$$

These two simplifications allow to bound the equation 25 as

$$\mathbb{E}\|v_t(X_t)\|^2 \leq \tfrac{e^{-2t}}{\sqrt{1-e^{-2t}}} \cdot S\big(X,\Sigma\big) \cdot \sqrt{\mathbb{E}\|v_t(X_t)\|^2}$$

Which is equivalent to the inequality 24. Now, combining inequalities 23 with 24 completes the proof of the lemma.

■