Universality in neural networks

Dan Mikulincer
MIT

Based on joint works with Ronen Eldan, Tselil Schramm
and Itay Glazer

Setting

- We consider a randomly initialized two-layered neural network,

$$N_k(x) := \frac{\pm 1}{\sqrt{k}} \sum_{i=1}^{k} \sigma\left(w_i \cdot x\right),$$

  where $x \in \mathbb{R}^n$, σ is non-linear and $\{w_i\}_{i=1}^k$ are *i.i.d.* $\mathcal{N}\left(0, \mathrm{I}_n\right)$.
- In 1995, it was observed by Neal that as $k \to \infty$, the law of the random function N_k tends to a Gaussian process, on the sphere.
- A Gaussian process is a random function $\mathcal{G}(x)$, such that all of its finite-dimensional marginals are jointly Gaussian.

Setting

- We consider a randomly initialized two-layered neural network,

$$N_k(x) := \frac{\pm 1}{\sqrt{k}} \sum_{i=1}^{k} \sigma\left(w_i \cdot x\right),$$

  where $x \in \mathbb{R}^n$, σ is non-linear and $\{w_i\}_{i=1}^k$ are *i.i.d.* $\mathcal{N}\left(0, \mathrm{I}_n\right)$.
- In 1995, it was observed by Neal that as $k \to \infty$, the law of the random function N_k tends to a Gaussian process, on the sphere.
- A Gaussian process is a random function $\mathcal{G}(x)$, such that all of its finite-dimensional marginals are jointly Gaussian.

Setting

- We consider a randomly initialized two-layered neural network,

$$N_k(x) := \frac{\pm 1}{\sqrt{k}} \sum_{i=1}^{k} \sigma\left(w_i \cdot x\right),$$

 where $x \in \mathbb{R}^n$, σ is non-linear and $\{w_i\}_{i=1}^{k}$ are *i.i.d.* $\mathcal{N}\left(0, \mathrm{I}_n\right)$.
- In 1995, it was observed by Neal that as $k \to \infty$, the law of the random function N_k tends to a Gaussian process, on the sphere.
- A Gaussian process is a random function $\mathcal{G}(x)$, such that all of its finite-dimensional marginals are jointly Gaussian.

Setting

- We consider a randomly initialized two-layered neural network,

$$N_k(x) := \frac{\pm 1}{\sqrt{k}} \sum_{i=1}^{k} \sigma\left(w_i \cdot x\right),$$

 where $x \in \mathbb{R}^n$, σ is non-linear and $\{w_i\}_{i=1}^k$ are *i.i.d.* $\mathcal{N}\left(0, \mathrm{I}_n\right)$.
- In 1995, it was observed by Neal that as $k \to \infty$, the law of the random function N_k tends to a Gaussian process, on the sphere.
- A Gaussian process is a random function $\mathcal{G}(x)$, such that all of its finite-dimensional marginals are jointly Gaussian.

Background

- There have been many works on this topic since Neal's original result.
- However, most previous results where either:
    1. Asymptotic - dealt with the limit.
    2. Finite-dimensional - If $\{x_i\}_{i=1}^{M} \subset \mathbb{R}^n$, then $\{N_k(x_i)\}_{i=1}^{M}$ is approximately Gaussian in $\mathbb{R}^M$
- We provide non-asymptotic convergence bounds in a functions space.

Background

- There have been many works on this topic since Neal's original result.
- However, most previous results where either:
  1. Asymptotic - dealt with the limit.
  2. Finite-dimensional - If $\{x_i\}_{i=1}^{M} \subset \mathbb{R}^n$, then $\{N_k(x_i)\}_{i=1}^{M}$ is approximately Gaussian in $\mathbb{R}^M$
- We provide non-asymptotic convergence bounds in a functions space.

Background

- There have been many works on this topic since Neal's original result.
- However, most previous results where either:
  1. Asymptotic - dealt with the limit.
  2. Finite-dimensional - If $\{x_i\}_{i=1}^M \subset \mathbb{R}^n$, then $\{N_k(x_i)\}_{i=1}^M$ is approximately Gaussian in $\mathbb{R}^M$
- We provide non-asymptotic convergence bounds in a functions space.

Background

- There have been many works on this topic since Neal's original result.
- However, most previous results where either:
  1. Asymptotic - dealt with the limit.
  2. Finite-dimensional - If $\{x_i\}_{i=1}^M \subset \mathbb{R}^n$, then $\{N_k(x_i)\}_{i=1}^M$ is approximately Gaussian in $\mathbb{R}^M$
- We provide non-asymptotic convergence bounds in a functions space.

Background

- There have been many works on this topic since Neal's original result.
- However, most previous results where either:
  1. Asymptotic - dealt with the limit.
  2. Finite-dimensional - If $\{x_i\}_{i=1}^{M} \subset \mathbb{R}^n$, then $\{N_k(x_i)\}_{i=1}^{M}$ is approximately Gaussian in $\mathbb{R}^M$
- We provide non-asymptotic convergence bounds in a functions space.

Background

- There have been many works on this topic since Neal's original result.
- However, most previous results where either:
    1. Asymptotic - dealt with the limit.
    2. Finite-dimensional - If $\{x_i\}_{i=1}^M \subset \mathbb{R}^n$, then $\{N_k(x_i)\}_{i=1}^M$ is approximately Gaussian in $\mathbb{R}^M$
- We provide non-asymptotic convergence bounds in a functions space.

A Metric on $L^2(\mathbb{S}^{n-1})$

To state our results, we define the following transportation metrics between random elements of $L^2(\mathbb{S}^{n-1})$:

$$\mathcal{WF}_2(\mathcal{F},\mathcal{F}') = \inf_{(\mathcal{F},\mathcal{F}')} \left(\int_{\mathbb{S}^{n-1}} \mathbb{E}\left[|\mathcal{F}(x) - \mathcal{F}'(x)|^2 \right] dx \right)^{\frac{1}{2}},$$

and

$$\mathcal{WF}_\infty(\mathcal{F},\mathcal{F}') = \inf_{(\mathcal{F},\mathcal{F}')} \mathbb{E}\left[\sup_{x\in\mathbb{S}^{n-1}} |\mathcal{F}(x) - \mathcal{F}'(x)| \right].$$

Results

For any reasonable activation σ, we establish bounds on the rate of convergence, $\mathcal{WF}_2(N_k, \mathcal{G}) \xrightarrow{k\to\infty} 0$.

If σ is polynomial, then our bounds are typically better and hold for the stronger $\mathcal{WF}_\infty$ metric.

For example, if $\{x_i\}_{i=1}^M \subset \mathbb{R}^n$, we can conclude that $\{N_k(x_i)\}_{i=1}^M$ converges to a Gaussian in a rate which is independent from M.

Results

For any reasonable activation σ, we establish bounds on the rate of convergence, $\mathcal{WF}_2(N_k, \mathcal{G}) \xrightarrow{k\to\infty} 0$.

If σ is polynomial, then our bounds are typically better and hold for the stronger $\mathcal{WF}_\infty$ metric.

For example, if $\{x_i\}_{i=1}^M \subset \mathbb{R}^n$, we can conclude that $\{N_k(x_i)\}_{i=1}^M$ converges to a Gaussian in a rate which is independent from M.

Results

For any reasonable activation σ, we establish bounds on the rate of convergence, $\mathcal{WF}_2(N_k, \mathcal{G}) \xrightarrow{k\to\infty} 0$.

If σ is polynomial, then our bounds are typically better and hold for the stronger $\mathcal{WF}_\infty$ metric.

For example, if $\{x_i\}_{i=1}^M \subset \mathbb{R}^n$, we can conclude that $\{N_k(x_i)\}_{i=1}^M$ converges to a Gaussian in a rate which is independent from M.

Plan

Main challenge: Known convergence rates of the high-dimensional CLT tend to deteriorate with the dimension.

Crucial observation: If σ is a polynomial, the same is also true for N_k. Hence, it is supported on a finite dimensional space of $L^2(\mathbb{S}^{n-1})$.

Our plan:

- For polynomials, embed N_k in a finite dimensional Euclidean space and invoke known CLT results.
- For general σ, we approximate by a polynomial and reduce to the previous case.

Plan

Main challenge: Known convergence rates of the high-dimensional CLT tend to deteriorate with the dimension.

Crucial observation: If σ is a polynomial, the same is also true for N_k. Hence, it is supported on a finite dimensional space of $L^2(\mathbb{S}^{n-1})$.

Our plan:

- For polynomials, embed N_k in a finite dimensional Euclidean space and invoke known CLT results.
- For general σ, we approximate by a polynomial and reduce to the previous case.

Plan

Main challenge: Known convergence rates of the high-dimensional CLT tend to deteriorate with the dimension.

Crucial observation: If σ is a polynomial, the same is also true for N_k. Hence, it is supported on a finite dimensional space of $L^2(\mathbb{S}^{n-1})$.

Our plan:

- For polynomials, embed N_k in a finite dimensional Euclidean space and invoke known CLT results.
- For general σ, we approximate by a polynomial and reduce to the previous case.

Plan

Main challenge: Known convergence rates of the high-dimensional CLT tend to deteriorate with the dimension.

Crucial observation: If σ is a polynomial, the same is also true for N_k. Hence, it is supported on a finite dimensional space of $L^2(\mathbb{S}^{n-1})$.

Our plan:

- For polynomials, embed N_k in a finite dimensional Euclidean space and invoke known CLT results.
- For general σ, we approximate by a polynomial and reduce to the previous case.

The Embedding

For now, suppose that $\sigma(t) = t^d$, for some $d \in \mathbb{N}$. Now, recall the following identity of tensor products $\langle v, u\rangle^d = \langle v^{\otimes d}, u^{\otimes d}\rangle$.
So,

$$N_k(x) = \frac{\pm 1}{\sqrt{k}} \sum_{\ell=1}^{k} (w_\ell \cdot x)^d$$

$$= \frac{\pm 1}{\sqrt{k}} \sum_{\ell=1}^{k} \langle w_\ell^{\otimes d}, x^{\otimes d}\rangle = \left\langle \frac{\pm 1}{\sqrt{k}} \sum_{\ell=1}^{k} w_\ell^{\otimes d}, x^{\otimes d} \right\rangle.$$

The Embedding

For now, suppose that $\sigma(t) = t^d$, for some $d \in \mathbb{N}$. Now, recall the following identity of tensor products $\langle v, u \rangle^d = \langle v^{\otimes d}, u^{\otimes d} \rangle$.
So,

$$N_k(x) = \frac{\pm 1}{\sqrt{k}} \sum_{\ell=1}^{k} (w_\ell \cdot x)^d$$
$$= \frac{\pm 1}{\sqrt{k}} \sum_{\ell=1}^{k} \langle w_\ell^{\otimes d}, x^{\otimes d} \rangle = \left\langle \frac{\pm 1}{\sqrt{k}} \sum_{\ell=1}^{k} w_\ell^{\otimes d}, x^{\otimes d} \right\rangle .$$

The Embedding

If G is any Gaussian vector in $(\mathbb{R}^n)^{\otimes d}$, we can define a Gaussian process $\mathcal{G}(x) = \left\langle G, x^{\otimes d} \right\rangle$.

Now,

$$\begin{aligned}
\mathcal{WF}_\infty(N_k, \mathcal{G}) &\leq \mathbb{E}\left[\sup_{x \in \mathbb{S}^{n-1}} |N_k(x) - \mathcal{G}(x)|\right] \\
&= \mathbb{E}\left[\sup_{x \in \mathbb{S}^{n-1}} \left\langle \frac{\pm 1}{\sqrt{k}} \sum_{\ell=1}^{k} w_\ell^{\otimes d} - G, x^{\otimes d} \right\rangle\right] \\
&\leq \mathbb{E}\left[\left\| \frac{\pm 1}{\sqrt{k}} \sum_{\ell=1}^{k} w_\ell^{\otimes d} - G \right\|^2\right]
\end{aligned}$$

The Embedding

If G is any Gaussian vector in $(\mathbb{R}^n)^{\otimes d}$, we can define a Gaussian process $\mathcal{G}(x) = \left\langle G, x^{\otimes d} \right\rangle$.

Now,

$$
\begin{aligned}
\mathcal{WF}_\infty(N_k, \mathcal{G}) &\le \mathbb{E}\left[\sup_{x \in \mathbb{S}^{n-1}} |N_k(x) - \mathcal{G}(x)|\right] \\
&= \mathbb{E}\left[\sup_{x \in \mathbb{S}^{n-1}} \left\langle \frac{\pm 1}{\sqrt{k}} \sum_{\ell=1}^{k} w_\ell^{\otimes d} - G, x^{\otimes d} \right\rangle\right] \\
&\le \mathbb{E}\left[\left\| \frac{\pm 1}{\sqrt{k}} \sum_{\ell=1}^{k} w_\ell^{\otimes d} - G \right\|^2\right]
\end{aligned}
$$

A Central Limit Theorem for Neural Networks

So, to control $\mathcal{WF}_\infty(N_k, \mathcal{G})$, it is enough to understand

$$\mathbb{E}\left[\left\|\frac{\pm 1}{\sqrt{k}}\sum_{\ell=1}^{k} w_\ell^{\otimes d} - G\right\|^2\right].$$

By using a tailored CLT for tensor powers, we then prove:

Theorem

Suppose that $\sigma(t) = t^d$. Then, there exists a Gaussian process $\mathcal{G}$, such that

$$\mathcal{WF}_\infty(N_k, \mathcal{G}) \leq \sqrt{\frac{n^{2.5d-1.5}}{k}}.$$

A Central Limit Theorem for Neural Networks

So, to control $\mathcal{WF}_\infty(N_k, \mathcal{G})$, it is enough to understand

$$\mathbb{E}\left[\left\|\frac{\pm 1}{\sqrt{k}}\sum_{\ell=1}^{k} w_\ell^{\otimes d} - G\right\|^2\right].$$

By using a tailored CLT for tensor powers, we then prove:

Theorem

Suppose that $\sigma(t) = t^d$. Then, there exists a Gaussian process $\mathcal{G}$, such that

$$\mathcal{WF}_\infty(N_k, \mathcal{G}) \leq \sqrt{\frac{n^{2.5d-1.5}}{k}}.$$

A Central Limit Theorem for Neural Networks

Theorem

Suppose that $\sigma(t) = t^d$. Then, there exists a Gaussian process $\mathcal{G}$, such that

$$\mathcal{WF}_\infty(N_k, \mathcal{G}) \leq \sqrt{\frac{n^{2.5d-1.5}}{k}}.$$

Remarks:

- The proof requires to bound the eigenvalues of $\mathrm{Cov}\left(w^{\otimes d}\right)$. (Recently generalized to other measures)
- A similar proof applies for general polynomials $\sigma(t) = \sum_{i=0}^{d} a_i x^i$. (Greatly improved by Adam Klukowski)

General Activations

For general σ, we may still write, for p_d a degree d polynomial,

$$\sigma = p_d + (\sigma - p_d).$$

It makes sense to minimize $\|p_d - \sigma\|_{L^2(\gamma)}$ and we take p_d to be the Hermite approximation of σ.

Thus, the following quantity $R_\sigma(d) := \|p_d - \sigma\|_{L^2(\gamma)}$ is fundamental.

General Activations

For general σ, we may still write, for p_d a degree d polynomial,

$$\sigma = p_d + (\sigma - p_d).$$

It makes sense to minimize $\|p_d - \sigma\|_{L^2(\gamma)}$ and we take p_d to be the Hermite approximation of σ.

Thus, the following quantity $R_\sigma(d) := \|p_d - \sigma\|_{L^2(\gamma)}$ is fundamental.

General Activations

For general σ, we may still write, for p_d a degree d polynomial,

$$\sigma = p_d + (\sigma - p_d).$$

It makes sense to minimize $\|p_d - \sigma\|_{L^2(\gamma)}$ and we take p_d to be the Hermite approximation of σ.

Thus, the following quantity $R_\sigma(d) := \|p_d - \sigma\|_{L^2(\gamma)}$ is fundamental.

General Activations

By optimizing over the value of $R_\sigma(d)$ and the bound for polynomial activations we prove:

Theorem

Suppose that $\|\sigma\|_{L^2(\gamma)} < \infty$. *Then,*

$$\mathcal{WF}_2(N_k, \mathcal{G}) \lesssim \sqrt{\frac{1}{k^{\frac{1}{6}}} + R_\sigma\left(\frac{\log(k)}{\log(n)}\right)}.$$

General Activations - Specific Examples

Theorem

Suppose that $\sigma = \mathrm{ReLU}$. *Then,*

$$\mathcal{WF}_2(N_k, \mathcal{G}) \lesssim \frac{\log(n)}{\log(k)}.$$

Theorem

Suppose that $\sigma = \tanh$. *Then,*

$$\mathcal{WF}_2(N_k, \mathcal{G}) \lesssim e^{-\sqrt{\frac{\log(k)}{\log(n)}}}.$$

General Activations - Specific Examples

Theorem

Suppose that $\sigma = \mathrm{ReLU}$. *Then,*

$$\mathcal{WF}_2(N_k, \mathcal{G}) \lesssim \frac{\log(n)}{\log(k)}.$$

Theorem

Suppose that $\sigma = \tanh$. *Then,*

$$\mathcal{WF}_2(N_k, \mathcal{G}) \lesssim e^{-\sqrt{\frac{\log(k)}{\log(n)}}}.$$

Thank You