Title: Scaling Limits of Bayesian Inference with Deep Neural Networks

Speakers: Boris Hanin

Series: Machine Learning Initiative

Date: April 19, 2024 - 2:30 PM

URL: https://pirsa.org/24040103

Abstract: Large neural networks are often studied analytically through scaling limits: regimes in which some structural network parameters (e.g. depth, width, number of training datapoints, and so on) tend to infinity. Such limits are challenging to identify and study in part because the limits as these structural parameters diverge typically do not commute. I will present some recent and ongoing work with Alexander Zlofaka (MIT), in which we provide the first solvable models of learning - in this case by Bayesian inference - with neural networks where the depth, width, and number of datapoints can all be large.

---

Zoom link
Bayesian Inference with Deep Shaped Networks
(joint w/ Alexander Zlokapa)
Data: \( D = \{ (x_\mu, y_\mu) \mid \mu = 1, \ldots, P \} \) with \( x_\mu \in \mathbb{R}^{N_0} \) and \( y_\mu \in \mathbb{R} \)

Model: \( f(x; \theta) = W^{(L+1)} \sigma W^{(L)} \cdots \sigma W^{(1)} x \), \( W^{(l)} \in \mathbb{R}^{N_{l-1} \times N_l} \), \( \sigma(v_n) = (v_n, 0) \)

Learning Rule: \( A: (D, f) \rightarrow \theta_{\text{learned}} \)

Q1: Analysis of \( A \) when \( P, N_0, L \gg 1 \)

Q2: Adaptation of \( f(x; \theta_{\text{learned}}) \) to \( D \) ?

Q3: Alignment of \( f, D, A \)
This Talk: Q1 - Q3 for

- $\sigma(t) = t + \frac{\Psi}{t} t^3 \approx \sqrt{L} \Psi_{\text{odd}}(t/L)$
- $\mathcal{A} = \text{Bayesian Inference} \quad e \quad T=0 \quad P < N_0$
- $N_{L+1} = 1$

$\Psi = 0$: (Deep Linear)

$f(x; \Theta) = W^{(L+1)} W^{(L)} \ldots W^{(1)} x = \Theta^T x$

Prior: $\Theta \sim \mathcal{P}_{\text{prior}} \implies W^{(1)} \sim \mathcal{N}(0, \sqrt{N_{L-1}}) \text{ iid}$
$N_{\xi_2} = N$

Free Prob. $\lambda_{prior} = 0$

\[ \frac{L}{N} = \lambda_{prior} \]

$\lambda_{prior} = \frac{\text{effective depth}}{\text{under prior}}$

$\lambda_{prior} = \infty$

$\|\Theta\| \rightarrow 0$

Ergodic
posterior: \[ P_{post}(\theta | D, N_\theta, L) = \lim_{\beta \to \infty} P_{prior}(\theta | N_\theta, L) \exp^{-\beta L(\theta | D)} \]

\[ L(\theta | D) = \frac{1}{2P} \sum_{\mu=1}^{P} (f(x_{\mu}; \theta) - y_{\mu})^2 \]

NB: \[ P_{post}(\theta) \propto P_{prior}(\theta) \int \{ \theta | x_\mu = y_\mu \ A_\mu \} \]

This Talk

\[ \cdot \sigma(t) \]

\[ \cdot A = \]

\[ \gamma = 0: (D) \]

\[ f(x; \theta) = \]

prior:
Posterior:
\[ p_{post}(\theta | D, N_e, L) = \lim_{\beta \to \infty} \frac{P_{prior}(\theta | N_e, L) \exp(\beta L(\theta | D))}{Z_\beta(D | N_e, L)} \]

\[ L(\theta | D) = \frac{1}{2P} \sum_{i=1}^{P} (f(x_i; \theta) - y_i)^2 \]

NB: \[ P_{post}(\theta) \propto P_{prior}(\theta) \int \{ \theta \mid x = y \} \forall \theta \]

\[ \Theta \sim P_{post} \iff \Theta = \Theta^* + \hat{\Theta} \]

\[ f(x; \theta) = \text{prior} \]

\[ \sigma(t) \]

\[ A = \]

\[ y = 0 : (D) \]

\[ \text{learned scalp} \]
Thm. (H-Zlokapa) \( E_{\text{post}} [\exp \{ -i \tau f(x, \Theta) \}] \) exactly computable via Miejer G Functions.

\[ P_{\text{post}} (\lambda_{\text{post}}, \| \Theta_* \|, d_o) \]

\[ P, N, L \to \infty \]
\[ P/N_0 \to d_o < 1 \]

\[ L \cdot P \]
\[ \to \lambda_{\text{post}} \]

"effective depth of post"
$\Psi \neq 0: \begin{cases} f(x; \Theta) = W^{(L+1)} \sigma W^{(L)} \ldots \sigma W^{(1)} x \\ W^{(a)} = N_a \times N_{a-1}, \quad \sigma(t) = t + \frac{\Psi}{t} t^3 \end{cases}$

If $L, N_0, \Psi \to \infty$ but $\frac{LP}{N} \to 0$, then model is equivalent to linear model with features

$x \in \mathbb{R}^{N_0} \rightarrow e^{\Psi (1 - 2\Psi \|x\|^2)^{-1/2}} x \rightarrow (1 + \|x\|^2)^{-1/2} x = x$
\[ f(x; \Theta, \psi) = f(x; \Theta, \frac{L^p}{N} = 0) \] 

\[ + C_{xy} \frac{L^p}{N} \sum_{\mu=1}^{P} y_{\mu} \alpha_{\mu} \left( \hat{x}^T \hat{x} - \hat{x}_{\mu}^T \hat{x}_{\mu} \right) + O\left( \frac{1}{N^2} \right) \] 

\[ (\hat{x})_{ll} = \frac{P}{\sum_{\mu=1}^{P}} \alpha_{\mu} \hat{x}_{\mu} \] 

\[ \hat{\Sigma} = \frac{1}{P} \sum_{\mu=1}^{P} \hat{x}_{\mu} \hat{x}_{\mu}^T \]
2. 1st order in $\frac{1}{2} \theta^2$ correction to kernel regime is cubic in $2x$. 

\{ (x_0, x) \mid x_0^2 = 1 + \|x\|^2 \geq 3 \}

$\psi > 0$

$(1 - \|x\|^2)^{1/2}(1, x)$

$\psi = 0$

$\{ \xi \times \mathbb{R}^N \}$

$(\hat{x})_i = \frac{\sum_i \xi_i x_i}{\sum_i \xi_i}$