A general framework for the mean field limit of multilayer neural networks

Huy Tuan Pham
Stanford University

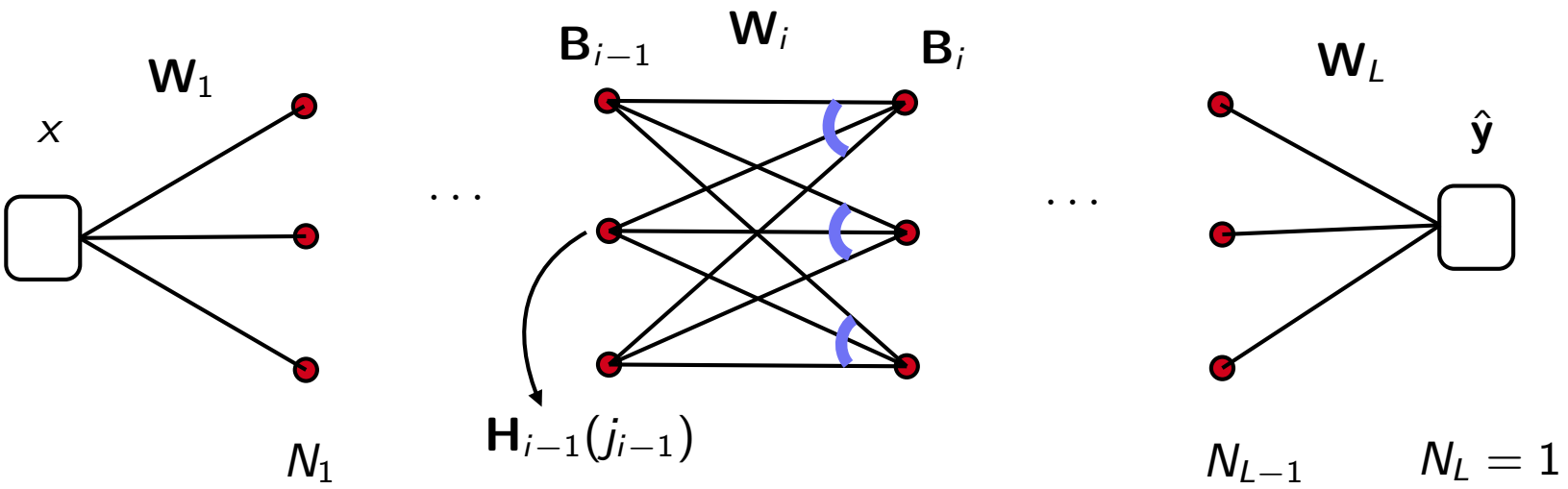
(Joint with Phan-Minh Nguyen)

One World Seminar on the Mathematics of Machine Learning
July 25, 2020

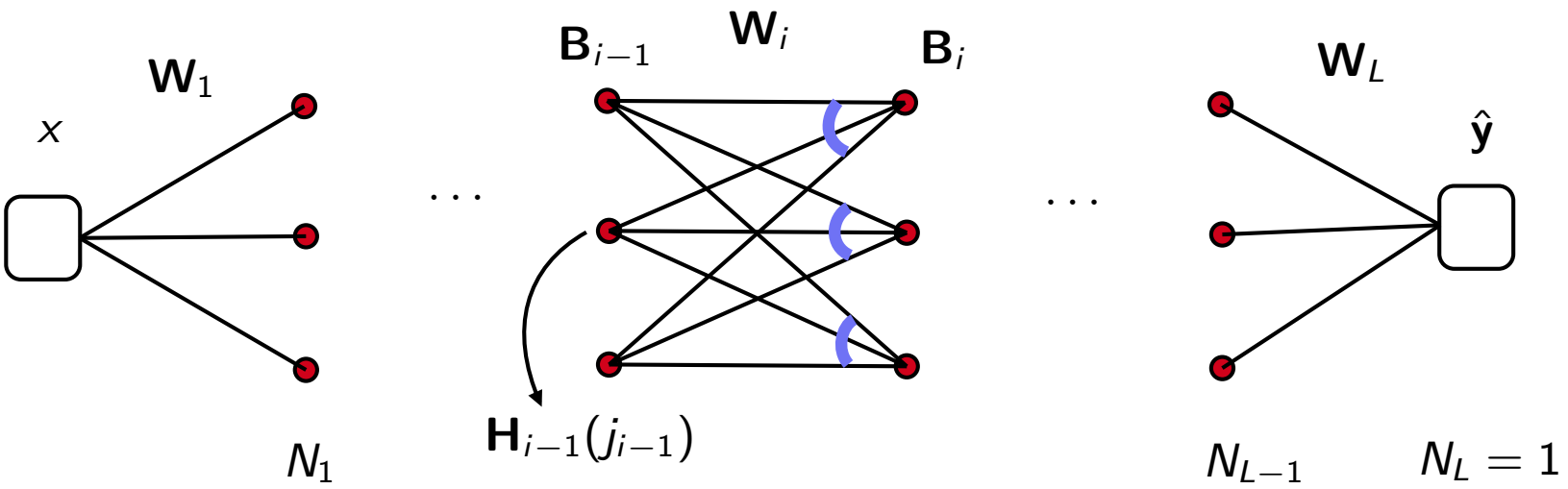
The setting

- Data distribution:
$$(y, x) \sim \mathbb{P}.$$
- Model:
$$\hat{y} = f(x; \theta).$$
We are interested in the class of models given by deep neural networks.
- Optimization problem:
$$\min_{\theta \in \Theta} \mathbb{E}[\ell(\hat{y}, y)].$$
We are interested in solving the optimization problem using stochastic gradient descent.

Fully connected L-layer neural network



Fully connected L-layer neural network



The model:

$$\mathbf{H}_1(x, j_1) = \phi_1(\mathbf{w}_1(j_1), x), \qquad \forall j_1 \in [N_1],$$

$$\mathbf{H}_i(x, j_i) = \frac{1}{N_{i-1}} \sum_{j_{i-1}=1}^{N_{i-1}} \phi_i(\mathbf{w}_i(j_{i-1}, j_i), \mathbf{b}_i(j_i), \mathbf{H}_{i-1}(x, j_{i-1})),$$

$$\forall j_i \in [N_i], i \in [2, L],$$

$$\hat{\mathbf{y}}(x) = \phi_{L+1}(\mathbf{H}_L(x, 1)).$$

Fully connected L-layer neural network

We assume $N_L = 1$.

Fully connected L-layer neural network

We assume $N_L = 1$.
A standard choice of ϕ (Fully-connected neural network with bias):

We assume $N_L = 1$.
A standard choice of ϕ (Fully-connected neural network with bias):

- $$\mathbf{H}_1(x) = \mathbf{W}_1 x \qquad \forall j_1 \in [N_1],$$
$$\mathbf{H}_1(x, j_1) = \mathbf{w}_1(j_1) \cdot x,$$
- $$\mathbf{H}_i(x) = \mathbf{B}_i + \frac{1}{N_{i-1}} \mathbf{W}_i \sigma_{i-1}(\mathbf{H}_{i-1}(x)) \qquad \forall j_i \in [N_i], i \in [2, L],$$
$$\mathbf{H}_i(x, j_i) = \frac{1}{N_{i-1}} \sum_{j_{i-1}=1}^{N_{i-1}} \mathbf{b}_i(j_i) + \mathbf{w}_i(j_{i-1}, j_i) \sigma_{i-1}(\mathbf{H}_{i-1}(x, j_{i-1})),$$
- $$\hat{\mathbf{y}}(x) = \sigma_L(\mathbf{H}_L(x, 1)),$$

where

$$\mathbf{W}_1 \in \mathbb{R}^{N_1 \times d}, \qquad \mathbf{W}_i \in \mathbb{R}^{N_{i-1} \times N_i}, \ \mathbf{B}_i \in \mathbb{R}^{N_i} \qquad \forall i \in [2, L].$$

Stochastic gradient descent

$$\mathbf{w}_1(t+1, j_1) = \mathbf{w}_1(t, j_1) - \varepsilon N_1 \nabla_{\mathbf{w}_1(j_1)} \ell(\hat{\mathbf{y}}(t, x(t)), y(t)),$$
$$\mathbf{w}_i(t+1, j_{i-1}, j_i) = \mathbf{w}_i(t, j_{i-1}, j_i) - \varepsilon N_{i-1} N_i \nabla_{\mathbf{w}_i(j_{i-1}, j_i)} \ell(\hat{\mathbf{y}}(t, x(t)), y(t)),$$
$$\mathbf{b}_i(t+1, j_i) = \mathbf{b}_i(t, j_i) - \varepsilon N_i \nabla_{\mathbf{b}_i(j_i)} \ell(\hat{\mathbf{y}}(t, x(t)), y(t)).$$

$$\mathbf{w}_1(t+1, j_1) = \mathbf{w}_1(t, j_1) - \varepsilon N_1 \nabla_{\mathbf{w}_1(j_1)} \ell(\hat{\mathbf{y}}(t, x(t)), y(t)),$$
$$\mathbf{w}_i(t+1, j_{i-1}, j_i) = \mathbf{w}_i(t, j_{i-1}, j_i) - \varepsilon N_{i-1} N_i \nabla_{\mathbf{w}_i(j_{i-1}, j_i)} \ell(\hat{\mathbf{y}}(t, x(t)), y(t)),$$
$$\mathbf{b}_i(t+1, j_i) = \mathbf{b}_i(t, j_i) - \varepsilon N_i \nabla_{\mathbf{b}_i(j_i)} \ell(\hat{\mathbf{y}}(t, x(t)), y(t)).$$

Remarks.

$$\mathbf{w}_1(t+1, j_1) = \mathbf{w}_1(t, j_1) - \varepsilon N_1 \nabla_{\mathbf{w}_1(j_1)} \ell(\hat{\mathbf{y}}(t, x(t)), y(t)),$$
$$\mathbf{w}_i(t+1, j_{i-1}, j_i) = \mathbf{w}_i(t, j_{i-1}, j_i) - \varepsilon N_{i-1} N_i \nabla_{\mathbf{w}_i(j_{i-1}, j_i)} \ell(\hat{\mathbf{y}}(t, x(t)), y(t)),$$
$$\mathbf{b}_i(t+1, j_i) = \mathbf{b}_i(t, j_i) - \varepsilon N_i \nabla_{\mathbf{b}_i(j_i)} \ell(\hat{\mathbf{y}}(t, x(t)), y(t)).$$

Remarks.

- The gradients are scaled so that weight movement is of order 1.

$$\mathbf{w}_1(t+1, j_1) = \mathbf{w}_1(t, j_1) - \varepsilon N_1 \nabla_{\mathbf{w}_1(j_1)} \ell(\hat{\mathbf{y}}(t, x(t)), y(t)),$$
$$\mathbf{w}_i(t+1, j_{i-1}, j_i) = \mathbf{w}_i(t, j_{i-1}, j_i) - \varepsilon N_{i-1} N_i \nabla_{\mathbf{w}_i(j_{i-1}, j_i)} \ell(\hat{\mathbf{y}}(t, x(t)), y(t)),$$
$$\mathbf{b}_i(t+1, j_i) = \mathbf{b}_i(t, j_i) - \varepsilon N_i \nabla_{\mathbf{b}_i(j_i)} \ell(\hat{\mathbf{y}}(t, x(t)), y(t)).$$

Remarks.

- The gradients are scaled so that weight movement is of order 1.
- The derivation of the mean field limit applies to more general dynamics.

In the standard model:

$$N_1 \nabla_{\mathbf{w}_1(j_1)} \ell(\hat{\mathbf{y}}(t, x(t)), y(t)) = \Delta_1^{\mathbf{H}}(y, x, j_1) x,$$

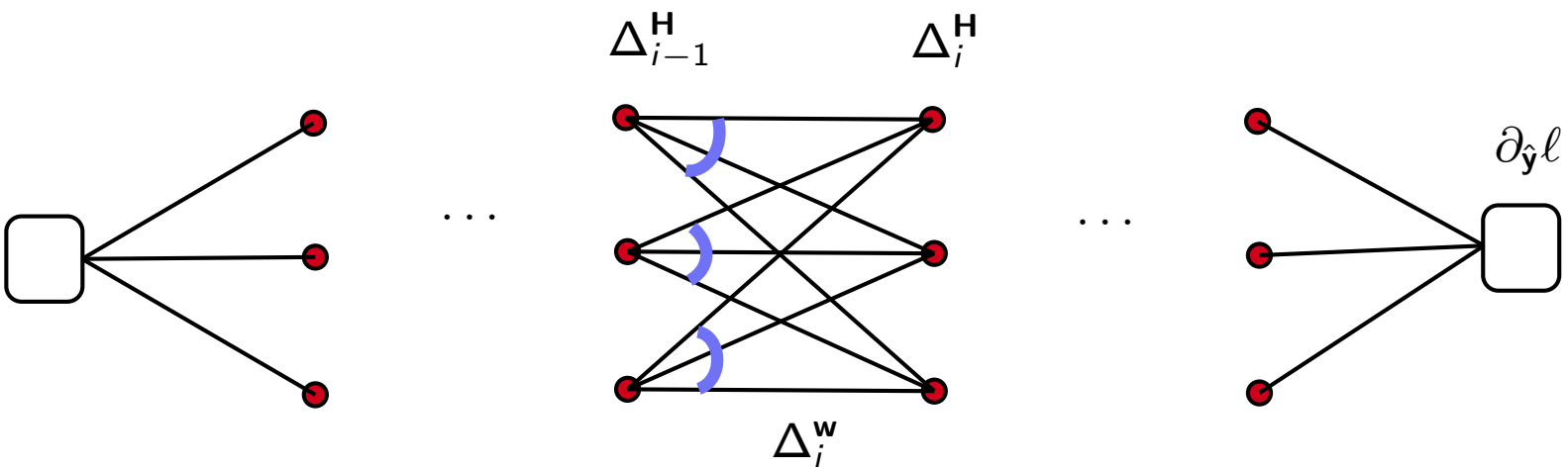
$$N_{i-1} N_i \nabla_{\mathbf{w}_i(j_{i-1}, j_i)} \ell(\hat{\mathbf{y}}(t, x(t)), y(t)) = \Delta_i^{\mathbf{H}}(y, x, j_i) \sigma_{i-1}(\mathbf{H}_{i-1}(x, j_{i-1})),$$

$$N_i \nabla_{\mathbf{b}_i(j_i)} \ell(\hat{\mathbf{y}}(t, x(t)), y(t)) = \Delta_i^{\mathbf{H}}(y, x, j_i),$$

where

$$\Delta_L^{\mathbf{H}}(y, x, j_L) = \partial_{\hat{\mathbf{y}}} \ell(y, \hat{\mathbf{y}}(x)) \sigma_L'(\mathbf{H}_L(x, j_L)),$$

$$\Delta_{i-1}^{\mathbf{H}}(y, x, j_{i-1}) = \frac{1}{N_i} \sum_{j_i=1}^{N_i} \Delta_i^{\mathbf{H}}(y, x, j_i) \mathbf{w}_i(t, j_{i-1}, j_i) \sigma_{i-1}'(\mathbf{H}_{i-1}(x, j_{i-1})).$$



In the general model:

$$N_1 \nabla_{\mathbf{w}_1(j_1)} \ell(\hat{\mathbf{y}}(t, x(t)), y(t)) = \Delta_1^{\mathbf{H}}(y, x, j_1) x,$$

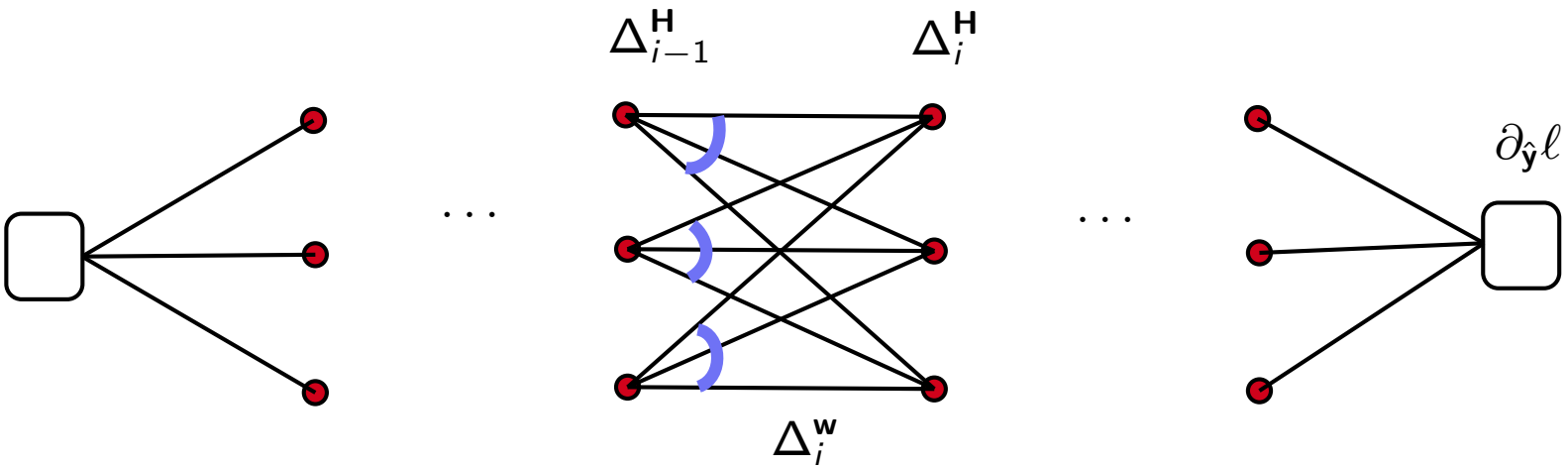
$$N_{i-1} N_i \nabla_{\mathbf{w}_i(j_{i-1}, j_i)} \ell(\hat{\mathbf{y}}(t, x(t)), y(t)) = \Delta_i^{\mathbf{H}}(y, x, j_i) \sigma_{i-1}(\mathbf{H}_{i-1}(x, j_{i-1})),$$

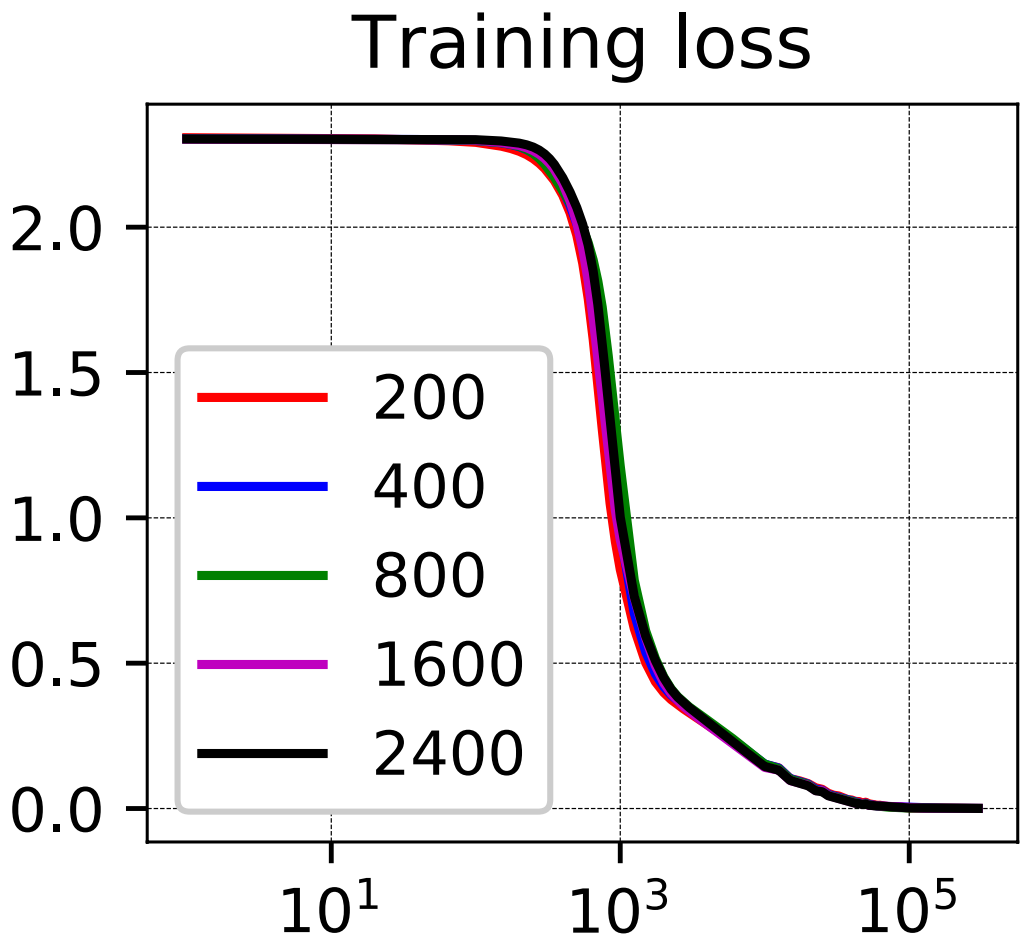
$$N_i \nabla_{\mathbf{b}_i(j_i)} \ell(\hat{\mathbf{y}}(t, x(t)), y(t)) = \Delta_i^{\mathbf{H}}(y, x, j_i),$$

where

$$\Delta_L^{\mathbf{H}}(y, x, j_L) = \partial_{\hat{\mathbf{y}}} \ell(y, \hat{\mathbf{y}}(x)) \sigma_L'(\mathbf{H}_L(x, j_L)),$$

$$\Delta_{i-1}^{\mathbf{H}}(y, x, j_{i-1}) = \frac{1}{N_i} \sum_{j_i=1}^{N_i} \psi(\Delta_i^{\mathbf{H}}(y, x, j_i), \mathbf{w}_i(t, j_{i-1}, j_i), \mathbf{H}_{i-1}(x, j_{i-1})).$$





MNIST, 3-layer nets, $N_1 = N_2 = 200, 400, \ldots$

In the large width limit, the mean field limit is obtained in the setting of i.i.d. initialization by Araújo-Oliveira-Yukimura and Sirignano-Spiliopoulos. The i.i.d. initialization plays an important role, and the limit formulation hinges on certain degeneracy in the dynamics.

In the large width limit, the mean field limit is obtained in the setting of i.i.d. initialization by Araújo-Oliveira-Yukimura and Sirignano-Spiliopoulos. The i.i.d. initialization plays an important role, and the limit formulation hinges on certain degeneracy in the dynamics.

In the general setting, Nguyen sets up a different formulation of the mean field limit.

In the large width limit, the mean field limit is obtained in the setting of i.i.d. initialization by Araújo-Oliveira-Yukimura and Sirignano-Spiliopoulos. The i.i.d. initialization plays an important role, and the limit formulation hinges on certain degeneracy in the dynamics.

In the general setting, Nguyen sets up a different formulation of the mean field limit.

- Our formulation of the mean field limit applies to general initialization distributions where there is no degeneracy in the dynamics.

A primer on the framework for the mean field limit

In the large width limit, the mean field limit is obtained in the setting of i.i.d. initialization by Araújo-Oliveira-Yukimura and Sirignano-Spiliopoulos. The i.i.d. initialization plays an important role, and the limit formulation hinges on certain degeneracy in the dynamics.

In the general setting, Nguyen sets up a different formulation of the mean field limit.

- Our formulation of the mean field limit applies to general initialization distributions where there is no degeneracy in the dynamics.
- The limit formulation is based on an infinite width representation that factors out the symmetries of the neurons in the hidden layer, and applies to general dynamics on systems with mean-field interactions.

In the large width limit, the mean field limit is obtained in the setting of i.i.d. initialization by Araújo-Oliveira-Yukimura and Sirignano-Spiliopoulos. The i.i.d. initialization plays an important role, and the limit formulation hinges on certain degeneracy in the dynamics.

In the general setting, Nguyen sets up a different formulation of the mean field limit.

- Our formulation of the mean field limit applies to general initialization distributions where there is no degeneracy in the dynamics.
- The limit formulation is based on an infinite width representation that factors out the symmetries of the neurons in the hidden layer, and applies to general dynamics on systems with mean-field interactions.
- Specializing to i.i.d. initializations, the degeneracy properties of the dynamics can be readily recovered from the framework.

D. Araújo, R. Oliveira, D. Yukimura, arXiv:1906.00193 (2019).
J. Sirignano, K. Spiliopoulos, arXiv:1903.04440 (2019).
P.-M. Nguyen, arXiv:1902.02880 (2019).

'Neuronal embedding'

Plan of the talk

1. Description of the infinite width limit

Plan of the talk

1. Description of the infinite width limit

2. Connecting finite width networks and the infinite width limit via neuronal embedding

Plan of the talk

1. Description of the infinite width limit
2. Connecting finite width networks and the infinite width limit via neuronal embedding
3. Symmetries in the mean field limit

Plan of the talk

1. Description of the infinite width limit
2. Connecting finite width networks and the infinite width limit via neuronal embedding
3. Symmetries in the mean field limit
4. Global convergence: Three-layer network

Plan of the talk

1. Description of the infinite width limit
2. Connecting finite width networks and the infinite width limit via neuronal embedding
3. Symmetries in the mean field limit
4. Global convergence: Three-layer network
5. Global convergence: General L-layer network

Plan of the talk

1. Description of the infinite width limit
2. Connecting finite width networks and the infinite width limit via neuronal embedding
3. Symmetries in the mean field limit
4. Global convergence: Three-layer network
5. Global convergence: General L-layer network

Infinite width limit - Revisit the two-layer case

- Two-layer neural network:

$$\hat{\mathbf{y}}(x) = \frac{1}{N} \sum_{i=1}^{N} \sigma(\mathbf{w}_i, x).$$

- Two-layer neural network:

$$\hat{\mathbf{y}}(x) = \frac{1}{N}\sum_{i=1}^{N}\sigma(\mathbf{w}_i, x).$$

- Symmetry group S_N: For $\pi \in S_N$,

$$\hat{\mathbf{y}}(x) = \frac{1}{N}\sum_{i=1}^{N}\sigma(\mathbf{w}_i, x) = \frac{1}{N}\sum_{i=1}^{N}\sigma(\mathbf{w}_{\pi(i)}, x).$$

- Two-layer neural network:

$$\hat{\mathbf{y}}(x) = \frac{1}{N}\sum_{i=1}^{N}\sigma(\mathbf{w}_i, x).$$

- Symmetry group S_N: For $\pi \in S_N$,

$$\hat{\mathbf{y}}(x) = \frac{1}{N}\sum_{i=1}^{N}\sigma(\mathbf{w}_i, x) = \frac{1}{N}\sum_{i=1}^{N}\sigma(\mathbf{w}_{\pi(i)}, x).$$

- Equivalent description:

$$\hat{\mathbf{y}}(x) = \mathbf{E}_{w\sim\rho_N}[\sigma(w, x)],$$

where $\rho_N = \frac{1}{N}\sum_{i=1}^{N}\delta_{\mathbf{w}_i}$ is the empirical measure of the neurons.

The infinite width limit of a two-layer network can be captured by a general distribution over $\mathbb{R}^d$.

$$\hat{y}(x) = \mathbf{E}_{w \sim \rho_N}[\sigma(w, x)] \longrightarrow \hat{y} = \mathbf{E}_{w \sim \rho}[\sigma(w, x)].$$

The infinite width limit of a two-layer network can be captured by a general distribution over $\mathbb{R}^d$.

$$\hat{y}(x) = \mathbf{E}_{w\sim\rho_N}[\sigma(w,x)] \longrightarrow \hat{y} = \mathbf{E}_{w\sim\rho}[\sigma(w,x)].$$

This does not generalize to networks with more than two layers:

- For L-layer networks, the symmetry group acts as follows: For permutation matrices $\Pi_1, \ldots, \Pi_{L-1}$,

$$\hat{\mathbf{y}}(x; \mathbf{W}_1, \ldots, \mathbf{W}_L) = \hat{\mathbf{y}}(x; \Pi_1\mathbf{W}_1, \Pi_2\mathbf{W}_2\Pi_1^T, \ldots, \mathbf{W}_L\Pi_{L-1}^T).$$

- Difficult to factor out symmetry to a distributional representation due to complex interaction with $\mathbf{W}_2, \ldots, \mathbf{W}_{L-1}$.

Constructing the infinite width network - Two-layer case

Claim

There exists a probability space (Ω, P) such that any two-layer network can be written as

$$\hat{\mathbf{y}}(x) = \mathbf{E}_{C\sim P}[\sigma(w(C), x)],$$

for a measurable function w on (Ω, P).

Finite width:

$$\mathbf{H}_i(x, j_i) = \frac{1}{N_{i-1}} \sum_{j_{i-1}=1}^{N_{i-1}} \phi_i(\mathbf{b}_i(j_i), \mathbf{w}_i(j_{i-1}, j_i), \mathbf{H}_{i-1}(x, j_{i-1})),$$

$$\hat{\mathbf{y}}(x) = \phi_{L+1}(\mathbf{H}_L(x, 1)).$$

Finite width:

$$\mathbf{H}_i(x, j_i) = \frac{1}{N_{i-1}} \sum_{j_{i-1}=1}^{N_{i-1}} \phi_i(\mathbf{b}_i(j_i), \mathbf{w}_i(j_{i-1}, j_i), \mathbf{H}_{i-1}(x, j_{i-1})),$$

$$\hat{\mathbf{y}}(x) = \phi_{L+1}(\mathbf{H}_L(x, 1)).$$

Infinite width:
For $(c_1, \ldots, c_L)$ in some probability space $(\Omega_1 \times \cdots \times \Omega_L, P_1 \times \cdots \times P_L)$ with $\Omega_L = \{1\}$, and functions $w_i : \Omega_{i-1} \times \Omega_i \to \mathbb{W}_i$, $b_i : \Omega_i \to \mathbb{B}_i$, let

$$H_i(x, c_i) = \mathbb{E}_{C_{i-1} \sim P_{i-1}}[\phi_i(b_i(c_i), w_i(C_{i-1}, c_i), H_{i-1}(x, C_{i-1}))],$$

$$\hat{y}(x; w, b) = \phi_{L+1}(H_L(x, c_L)).$$

Claim

There exists a probability space $(\Omega_1 \times \cdots \times \Omega_L, P_1 \times \cdots \times P_L)$ and measurable functions $w_i : \Omega_{i-1} \times \Omega_i \to \mathbb{W}_i$, $b_i : \Omega_i \to \mathbb{B}_i$, such that any depth L network can be written as

$$\hat{\mathbf{y}}(x) = \hat{y}(x; w, b).$$

Claim

There exists a probability space $(\Omega_1 \times \cdots \times \Omega_L, P_1 \times \cdots \times P_L)$ and measurable functions $w_i : \Omega_{i-1} \times \Omega_i \to \mathbb{W}_i$, $b_i : \Omega_i \to \mathbb{B}_i$, such that any depth L network can be written as

$$\hat{\mathbf{y}}(x) = \hat{y}(x; w, b).$$

We call the space $(\Omega_1 \times \cdots \times \Omega_L, P_1 \times \cdots \times P_L)$ a **neuronal emsemble**.

Claim

There exists a probability space $(\Omega_1 \times \cdots \times \Omega_L, P_1 \times \cdots \times P_L)$ and measurable functions $w_i : \Omega_{i-1} \times \Omega_i \to \mathbb{W}_i$, $b_i : \Omega_i \to \mathbb{B}_i$, such that any depth L network can be written as

$$\hat{\mathbf{y}}(x) = \hat{y}(x; w, b).$$

We call the space $(\Omega_1 \times \cdots \times \Omega_L, P_1 \times \cdots \times P_L)$ a **neuronal emsemble**.

This defines a class of depth-L networks with arbitrary (or infinite) width.

Claim

There exists a probability space $(\Omega_1 \times \cdots \times \Omega_L, P_1 \times \cdots \times P_L)$ and measurable functions $w_i : \Omega_{i-1} \times \Omega_i \to \mathbb{W}_i$, $b_i : \Omega_i \to \mathbb{B}_i$, such that any depth L network can be written as

$$\hat{\mathbf{y}}(x) = \hat{y}(x; w, b).$$

We call the space $(\Omega_1 \times \cdots \times \Omega_L, P_1 \times \cdots \times P_L)$ a **neuronal emsemble**.

This defines a class of depth-L networks with arbitrary (or infinite) width.

This talk: focus on optimization aspect (would be interesting to study generalization, approximation, etc.).

Dynamics of the infinite width network

$$
\begin{aligned}
\partial_t w_i(t, c_{i-1}, c_i) &= -\mathbb{E}_{(y,x)}[\nabla_{w_i(c_{i-1},c_i)}\ell(\hat{y}, y)] \\
&= -\mathbb{E}_{(y,x)}[\mathbb{E}_{C_{i+1}\sim P_{i+1}}[\psi_i^w(w_i(t, c_{i-1}, c_i), w_{i+1}(t, c_i, C_{i+1}), \\
&\qquad\qquad H_{i+1}(x, C_{i+1}), H_{i-1}(x, c_{i-1}))]]. \\
\partial_t b_i(t, c_i) &= -\mathbb{E}_{(y,x)}[\nabla_{b_i(c_i)}\ell(\hat{y}, y)] \\
&= -\mathbb{E}_{(y,x)}[\mathbb{E}_{C_{i+1}\sim P_{i+1}}[\psi_i^b(b_i(t, c_i), w_{i+1}(t, c_i, C_{i+1}), H_{i+1}(x, C_{i+1}))]].
\end{aligned}
$$

In the standard model:

$$\partial_t w_i(t, c_{i-1}, c_i) = -\mathbb{E}_{(y,x)}[\Delta_i^w(y, x, c_{i-1}, c_i)],$$
$$\partial_t b_i(t, c_i) = -\mathbb{E}_{(y,x)}[\Delta_i^b(y, x, c_i)],$$

where

$$\Delta_L^H(y, x, c_L) = \partial_{\hat{y}}\ell(\hat{y}, y)\phi_L'(H_L(x, c_L)),$$
$$\Delta_{i-1}^H(y, x, c_{i-1}) = \mathbb{E}_{C_i}[\Delta_i^H(y, x, C_i)w_i(t, c_{i-1}, C_i)\phi_{i-1}'(H_{i-1}(x, c_{i-1}))],$$
$$\Delta_i^w(y, x, c_{i-1}, c_i) = \Delta_i^H(y, x, c_i)\phi_{i-1}(H_{i-1}(x, c_{i-1})),$$
$$\Delta_i^b(y, x, c_i) = \Delta_i^H(y, x, c_i),$$
$$\Delta_1^w(y, x, c_1) = \Delta_1^H(y, x, c_1)x.$$

Plan of the talk

1. Description of the infinite width limit
2. Connecting finite width networks and the infinite width limit via neuronal embedding
3. Symmetries in the mean field limit
4. Global convergence: Three-layer network
5. Global convergence: General L-layer network

Sampling finite width networks from an infinite width network

Sampling procedure

Sampling procedure

- Consider a neuronal ensemble $(\Omega_1 \times \cdots \times \Omega_L, P_1 \times \cdots \times P_L)$ and measurable functions

$$w_i^0 : \Omega_{i-1} \times \Omega_i \to \mathbb{W}_i, \qquad b_i^0 : \Omega_i \to \mathbb{B}_i.$$

Sampling procedure

- Consider a neuronal ensemble $(\Omega_1 \times \cdots \times \Omega_L, P_1 \times \cdots \times P_L)$ and measurable functions

$$w_i^0 : \Omega_{i-1} \times \Omega_i \to \mathbb{W}_i, \qquad b_i^0 : \Omega_i \to \mathbb{B}_i.$$

- For $i \in [L]$ and $j_i = 1, 2, \ldots, N_i$, sample $C_i(j_i)$ independently at random from (Ω_i, P_i).

Sampling procedure

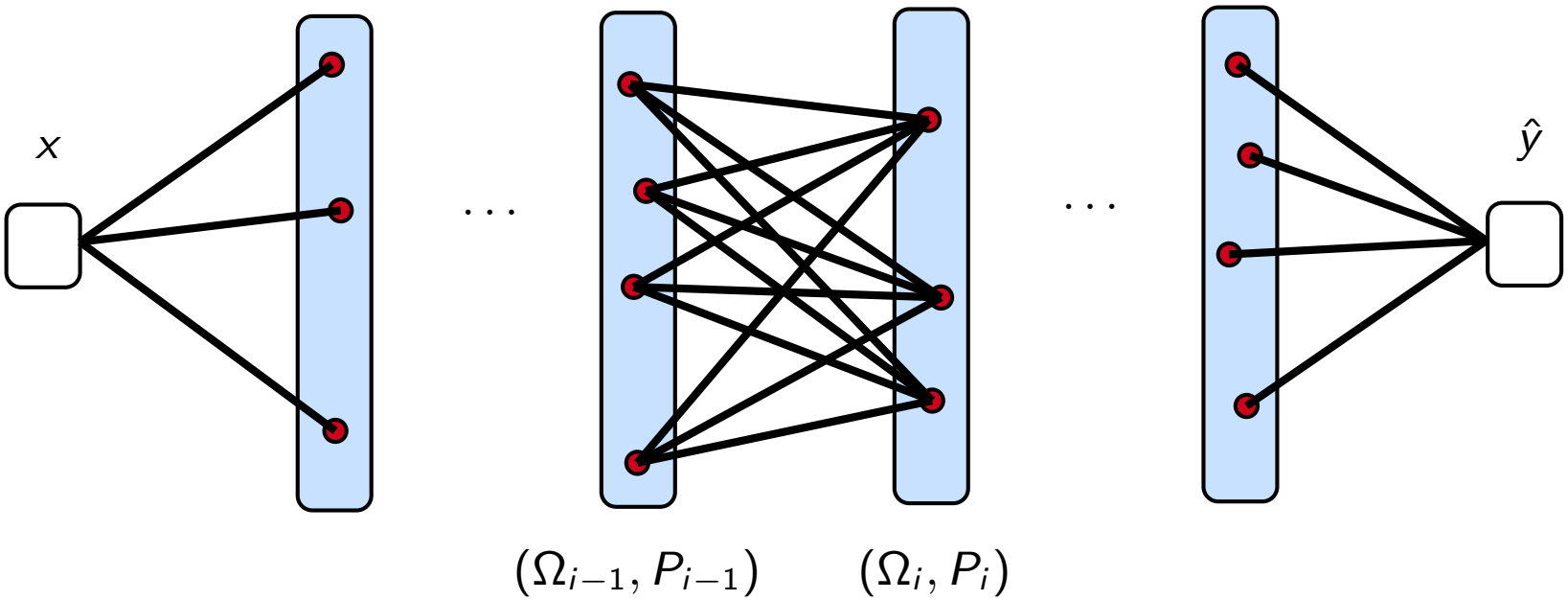
- Consider a neuronal ensemble $(\Omega_1 \times \cdots \times \Omega_L, P_1 \times \cdots \times P_L)$ and measurable functions

$$w_i^0 : \Omega_{i-1} \times \Omega_i \to \mathbb{W}_i, \qquad b_i^0 : \Omega_i \to \mathbb{B}_i.$$

- For $i \in [L]$ and $j_i = 1, 2, \ldots, N_i$, sample $C_i(j_i)$ independently at random from (Ω_i, P_i).
- Construct the neural network of widths $(N_1, \ldots, N_L)$ with weights initialized at

$$\mathbf{w}_i(0, j_{i-1}, j_i) = w_i^0(C_{i-1}(j_{i-1}), C_i(j_i)), \qquad \mathbf{b}_i(0, j_i) = b_i^0(C_i(j_i)).$$

Sampling finite width networks from an infinite width network



Neuronal embedding

Definition (Neuronal embedding)

Consider an L-layer neural network with initialization distribution ρ. A *neuronal embedding* is a tuple

$$((\Omega_1 \times \cdots \times \Omega_L, P_1 \times \cdots \times P_L), (w_i^0, b_i^0)_{i=1}^L)$$

such that following the sampling procedure for this tuple, the constructed finite width network has the same initialization distribution as ρ.

Mean field limit

- The neuronal embedding allows to embed all finite size networks into an infinite width limit.

- The neuronal embedding allows to embed all finite size networks into an infinite width limit.
- Consider a neuronal embedding

$$(\Omega_1 \times \cdots \times \Omega_L, P_1 \times \cdots \times P_L, (w_i^0, b_i^0)_{i=1}^L).$$

- The neuronal embedding allows to embed all finite size networks into an infinite width limit.
- Consider a neuronal embedding

$$(\Omega_1 \times \cdots \times \Omega_L, P_1 \times \cdots \times P_L, (w_i^0, b_i^0)_{i=1}^L).$$

- Run the infinite width network dynamics initialized at $(w_i^0, b_i^0)_{i=1}^L$. We refer to $(w_i(t, c_{i-1}, c_i), b_i(t, c_i))_{i=1}^L$ as the mean field limit.

Connecting the mean field limit and finite width networks

- Consider the sampled neural network with hidden widths N_i:

$$\mathbf{w}_i(0, j_{i-1}, j_i) = w_i^0(C_{i-1}(j_{i-1}), C_i(j_i)), \qquad \mathbf{b}_i(0, j_i) = b_i^0(C_i(j_i)).$$

- Consider the sampled neural network with hidden widths N_i:

$$\mathbf{w}_i(0, j_{i-1}, j_i) = w_i^0(C_{i-1}(j_{i-1}), C_i(j_i)), \qquad \mathbf{b}_i(0, j_i) = b_i^0(C_i(j_i)).$$

- Run stochastic gradient descent on the network with step-size ε in time T/ε.

- Consider the sampled neural network with hidden widths N_i:

$$\mathbf{w}_i(0, j_{i-1}, j_i) = w_i^0(C_{i-1}(j_{i-1}), C_i(j_i)), \qquad \mathbf{b}_i(0, j_i) = b_i^0(C_i(j_i)).$$

- Run stochastic gradient descent on the network with step-size ε in time T/ε.
- Let

$$\begin{aligned}
&\|\mathbf{w} - w\|_{t,\infty} \\
&= \max_{i \in [L], j_{i-1} \in [N_{i-1}], j_i \in [N_i]} \|\mathbf{w}_i(t, j_{i-1}, j_i) - w_i(\varepsilon t, C_{i-1}(j_{i-1}), C_i(j_i))\|, \\
&\|\mathbf{b} - b\|_{t,\infty} \\
&= \max_{i \in [L], j_i \in [N_i]} \|\mathbf{b}_i(t, j_i) - b_i(\varepsilon t, C_i(j_i))\|.
\end{aligned}$$

Theorem (Nguyen-P. 2020)

Under some regularity properties of the activations, then with probability at least $1-\delta$*, we have*

$$\sup_{t \leq T/\varepsilon} \|\mathbf{w} - w\|_{t,\infty} = O_T\left(\sqrt{\log(\max N_i/\delta)}\left(\frac{1}{\sqrt{\min N_i}} + \sqrt{\varepsilon}\right)\right).$$

Plan of the talk

1. Description of the infinite width limit
2. Connecting finite width networks and the infinite width limit via neuronal embedding
3. Symmetries in the mean field limit
4. Global convergence: Three-layer network
5. Global convergence: General L-layer network

Theorem (Nguyen-P. 2020)

There exists a neuronal embedding of the neural network whose weights are i.i.d. with law $\mathbf{w}_i \sim \rho_i^w$, $\mathbf{b}_i \sim \rho_i^b$.

Theorem (Nguyen-P. 2020)

There exists a neuronal embedding of the neural network whose weights are i.i.d. with law $\mathbf{w}_i \sim \rho_i^w$, $\mathbf{b}_i \sim \rho_i^b$.

By specializing the general framework to initialization distributions with certain symmetries, we can also characterize the symmetries in the mean field limit.

Theorem - simplified (Nguyen-P. 2020)

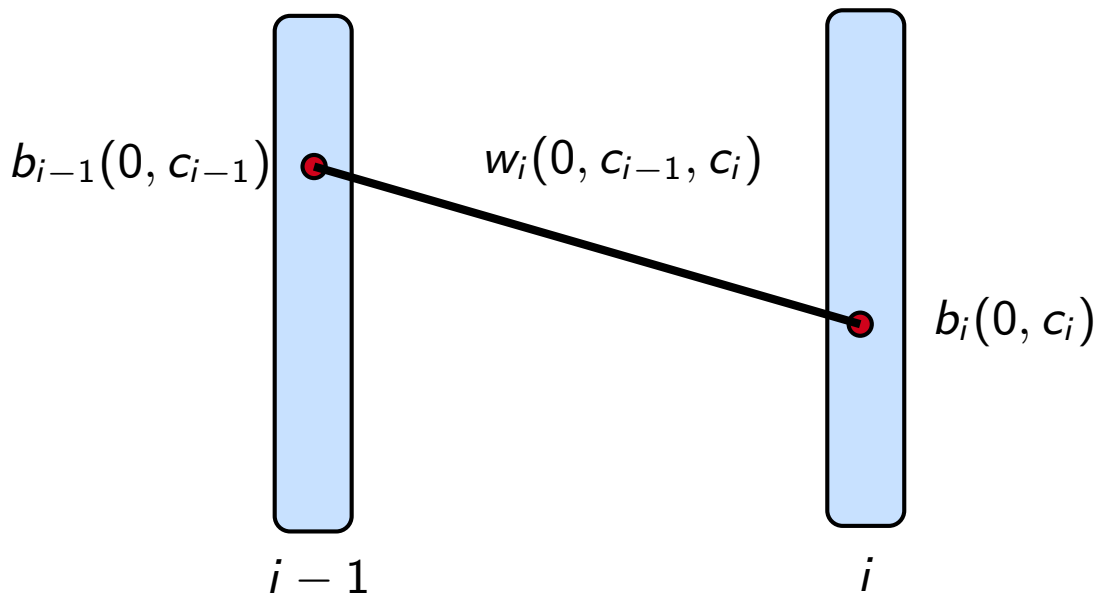
Consider the neural network whose weights are i.i.d. with law $\mathbf{w}_i \sim \rho_i^w$, $\mathbf{b}_i \sim \rho_i^b$. Then for $3 \leq i \leq L-2$:

- The middle layer weight $w_i(t, c_{i-1}, c_i)$ is a Borel function of its initial value $w_i(0, c_{i-1}, c_i)$ and the initial biases $b_{i-1}(0, c_{i-1}), b_i(0, c_i)$.
- The middle layer bias $b_i(t, c_i)$ is a Borel function of its initial value $b_i(0, c_i)$.

Theorem - simplified (Nguyen-P. 2020)

Consider the neural network whose weights are i.i.d. with law $\mathbf{w}_i \sim \rho_i^w$, $\mathbf{b}_i \sim \rho_i^b$. Then for $3 \leq i \leq L-2$:

- The middle layer weight $w_i(t, c_{i-1}, c_i)$ is a Borel function of its initial value $w_i(0, c_{i-1}, c_i)$ and the initial biases $b_{i-1}(0, c_{i-1}), b_i(0, c_i)$.
- The middle layer bias $b_i(t, c_i)$ is a Borel function of its initial value $b_i(0, c_i)$.



Theorem (Nguyen-P. 2020)

- For $3 \leq i \leq L-2$,

$$w_i(t, c_{i-1}, c_i) = w_i^*(t, w_i(0, c_{i-1}, c_i), b_{i-1}^0(c_{i-1}), b_i^0(c_i)),$$
$$b_i(t, c_i) = b_i^*(t, b_i(0, c_i)).$$

- For $i = 1, L$,

$$w_1(t, c_1) = w_1^*(t, w_1(0, c_1), b_1(0, c_1)),$$
$$b_1(t, c_1) = b_1^*(t, b_1(0, c_1)),$$
$$w_L(t, c_{L-1}, c_L) = w_L^*(t, w_L(0, c_{L-1}, c_L), b_{L-1}(0, c_{L-1}), b_L(0, c_L)),$$
$$b_L(t, c_L) = b_L^*(t, b_L(0, c_L)).$$

- For $i = 2, L-1$,

$$w_2(t, c_1, c_2) = w_2^*(t, w_2(0, c_1, c_2), b_2(0, c_2), w_1(0, c_1), b_1(0, c_1)),$$
$$b_2(t, c_2) = b_2^*(t, b_2(0, c_2)),$$
$$w_{L-1}(t, c_{L-2}, c_{L-1}) = w_{L-1}^*(t, w_{L-1}(0, c_{L-2}, c_{L-1}), b_{L-1}(0, c_{L-1}), w_L(0, c_{L-1}, c_L), b_{L-2}(0, c_{L-2})),$$
$$b_{L-1}(t, c_{L-1}) = b_{L-1}^*(t, b_{L-1}(0, c_{L-1}), w_L(0, c_{L-1}, c_L)).$$

Symmetries in the mean field limit

Remark. The Borel functions w^*, b^* can be obtained by solving a self-contained system of ODEs.

Remark. The Borel functions w^*, b^* can be obtained by solving a self-contained system of ODEs.

Corollary (Nguyen-P. 2020)

If the biases are initialized to be constants, then the weight w_i at time t only depends on its value at initialization for $3 \leq i \leq L-2$. Thus, for any positive time t, the middle layer weights $w_i(t, c_{i-1}, c_i)$ for $3 \leq i \leq L-2$ remain i.i.d. random variables.
Furthermore, the pre-activations $H_i(t, c_i)$ are equal for all $2 \leq i \leq L-2$.

Corollary (Nguyen-P. 2020)

If we assume the standard model of the neural network with the biases initialized to be constants, then there exists functions $\Delta_i^* : [0, \infty) \to \mathbb{R}$ such that for $3 \leq i \leq L-2$,

$$w_i(t, c_{i-1}, c_i) = w_i(0, c_{i-1}, c_i) + \Delta_i^*(t).$$

Corollary (Nguyen-P. 2020)

If we assume the standard model of the neural network with the biases initialized to be constants, then there exists functions $\Delta_i^* : [0, \infty) \to \mathbb{R}$ such that for $3 \leq i \leq L-2$,

$$w_i(t, c_{i-1}, c_i) = w_i(0, c_{i-1}, c_i) + \Delta_i^*(t).$$

Thus, in the standard architecture with no bias, each of the middle layers of the network degenerates to a single translation parameter.

Plan of the talk

1. Description of the infinite width limit
2. Connecting finite width networks and the infinite width limit via neuronal embedding
3. Symmetries in the mean field limit
4. Global convergence: Three-layer network
5. Global convergence: General L-layer network

Global convergence - Three-layer network

Consider the standard model of a three-layer network with i.i.d. initialization law.

Consider the standard model of a three-layer network with i.i.d. initialization law.

Theorem (Global convergence of the mean field dynamics, Nguyen-P. 2020)

Assume some regularity properties of the activations, and

- $\partial_{\hat{y}} \ell(\hat{y}, y) = 0 \Rightarrow \ell(\hat{y}, y) = 0$.
- *Diversity: At initialization,* w_1 *has full support in* $\mathbb{R}^d$.
- *Universal approximation:* $\{f_w(x) = \sigma_1(w \cdot x)\}_{w \in \mathbb{R}^d}$ *is dense in* $L^2(\mathbb{P}_x)$.
- $y = y(x)$.
- $w_1(t), w_2(t), w_3(t)$ *converges in appropriate sense as* $t \to \infty$.

Consider the standard model of a three-layer network with i.i.d. initialization law.

Theorem (Global convergence of the mean field dynamics, Nguyen-P. 2020)

Assume some regularity properties of the activations, and

- $\partial_{\hat{y}} \ell(\hat{y}, y) = 0 \Rightarrow \ell(\hat{y}, y) = 0$.
- *Diversity: At initialization,* w_1 *has full support in* $\mathbb{R}^d$.
- *Universal approximation:* $\{f_w(x) = \sigma_1(w \cdot x)\}_{w \in \mathbb{R}^d}$ *is dense in* $L^2(\mathbb{P}_x)$.
- $y = y(x)$.
- $w_1(t), w_2(t), w_3(t)$ *converges in appropriate sense as* $t \to \infty$.

Then under the mean field dynamics

$$\mathbb{E}_{(y,x)}[\ell(\hat{y}, y)] \to 0 \ \ as \ t \to \infty.$$

Convergence assumptions:

Convergence assumptions:

- We assume that the gradient of the second layer weights $\partial_t w_2(t, C_1, C_2) \to 0$ uniformly, and appropriate convergence in moments.

Convergence assumptions:

- We assume that the gradient of the second layer weights $\partial_t w_2(t, C_1, C_2) \to 0$ uniformly, and appropriate convergence in moments.
- If $\mathbb{E}_{(y,x)}[\ell(\hat{y}, y)] \to 0$, then $\partial_t w_2(t, C_1, C_2) \to 0$ uniformly.

Convergence assumptions:

- We assume that the gradient of the second layer weights $\partial_t w_2(t, C_1, C_2) \to 0$ uniformly, and appropriate convergence in moments.
- If $\mathbb{E}_{(y,x)}[\ell(\hat{y}, y)] \to 0$, then $\partial_t w_2(t, C_1, C_2) \to 0$ uniformly.

Remarks.

Convergence assumptions:

- We assume that the gradient of the second layer weights $\partial_t w_2(t, C_1, C_2) \to 0$ uniformly, and appropriate convergence in moments.
- If $\mathbb{E}_{(y,x)}[\ell(\hat{y}, y)] \to 0$, then $\partial_t w_2(t, C_1, C_2) \to 0$ uniformly.

Remarks.

- The loss ℓ does not need to be convex.

Convergence assumptions:

- We assume that the gradient of the second layer weights $\partial_t w_2(t, C_1, C_2) \to 0$ uniformly, and appropriate convergence in moments.
- If $\mathbb{E}_{(y,x)}[\ell(\hat{y}, y)] \to 0$, then $\partial_t w_2(t, C_1, C_2) \to 0$ uniformly.

Remarks.

- The loss ℓ does not need to be convex.
- No assumption on the limiting distribution of the mean field dynamics.

Global convergence - Three-layer network

Idea:

Idea:

- Our first insight is to look at the second layer weights:

$$\mathbb{E}_{(y,x)}[\partial_{\hat{y}}\ell(\hat{y}, y)F(C_2)\sigma_1(w_1(t, C_1) \cdot x)] \to 0.$$

Idea:

- Our first insight is to look at the second layer weights:

$$\mathbb{E}_{(y,x)}[\partial_{\hat{y}}\ell(\hat{y},y)F(C_2)\sigma_1(w_1(t,C_1)\cdot x)] \to 0.$$

- **Key step.** Use symmetries of the mean field dynamics and topological invariance, show that $w_1(t, C_1)$ has full support for all finite t (but not necessarily at $t = \infty$).

Idea:

- Our first insight is to look at the second layer weights:

$$\mathbb{E}_{(y,x)}[\partial_{\hat{y}}\ell(\hat{y},y)F(C_2)\sigma_1(w_1(t,C_1)\cdot x)] \to 0.$$

- **Key step.** Use symmetries of the mean field dynamics and topological invariance, show that $w_1(t, C_1)$ has full support for all finite t (but not necessarily at $t = \infty$).
- By denseness of $x \mapsto \sigma_1(w_1 \cdot x)$ in $L^2(P_x)$ and since $w_1(t, C_1)$ has full support, under mild assumptions so F is bounded away from 0, we obtain that for a.e. (y, x),

$$\partial_{\hat{y}}\ell(\hat{y},y) = 0$$

Hence, for a.e. (y, x), $\ell(\hat{y}, y) = 0$.

Idea:

- Our first insight is to look at the second layer weights:

$$\mathbb{E}_{(y,x)}[\partial_{\hat{y}}\ell(\hat{y},y)F(C_2)\sigma_1(w_1(t,C_1)\cdot x)] \to 0. \quad \text{(Uniform convergence.)}$$

- **Key step.** Use symmetries of the mean field dynamics and topological invariance, show that $w_1(t, C_1)$ has full support for all finite t (but not necessarily at $t = \infty$). (Topological invariance.)
- By denseness of $x \mapsto \sigma_1(w_1 \cdot x)$ in $L^2(P_x)$ and since $w_1(t, C_1)$ has full support, under mild assumptions so F is bounded away from 0, we obtain that for a.e. (y, x),

$$\partial_{\hat{y}}\ell(\hat{y},y) = 0$$

Hence, for a.e. (y, x), $\ell(\hat{y}, y) = 0$.

Plan of the talk

1. Description of the infinite width limit
2. Connecting finite width networks and the infinite width limit via neuronal embedding
3. Symmetries in the mean field limit
4. Global convergence: Three-layer network
5. Global convergence: General L-layer network

- The global convergence guarantee of three-layer network relies crucially on the diversity of the neurons in the first layer.

- The global convergence guarantee of three-layer network relies crucially on the diversity of the neurons in the first layer.
- Problem: Under i.i.d. initialization, the middle layers degenerate to a single "effective" neuron.

- The global convergence guarantee of three-layer network relies crucially on the diversity of the neurons in the first layer.
- Problem: Under i.i.d. initialization, the middle layers degenerate to a single "effective" neuron.

Question

Can we obtain global convergence for deep neural networks without changing the architecture?

Consider the standard model of an L-layer neural network.

Theorem (Global convergence of the mean field dynamics, Nguyen-P. 2020)

Assume some regularity properties of the activations, and

- $\partial_{\hat{y}}\ell(\hat{y}, y) = 0 \Rightarrow \ell(\hat{y}, y) = 0$.
- **Bidirectional diversity:** $(w_1(0, C_1), w_2(0, C_1, c_2))_{C_1 \sim P_1}$ *has full support in* $\mathbb{R}^d \times L^2(P_2)$ *and* $(w_i(0, c_{i-1}, C_i), w_{i+1}(0, C_i, c_{i+1}))_{C_i \sim P_i}$ *has full support in* $L^2(P_{i-1}) \times L^2(P_{i+1})$ *for all* $2 \leq i \leq L-1$.
- *Universal approximation:* $\{f_w(x) = \sigma_1(w \cdot x)\}_{w \in \mathbb{R}^d}$ *is dense in* $L^2(\mathbb{P}_x)$.
- $y = y(x)$.
- $w_i(t)$ *converges in appropriate sense as* $t \to \infty$.

Consider the standard model of an L-layer neural network.

Theorem (Global convergence of the mean field dynamics, Nguyen-P. 2020)

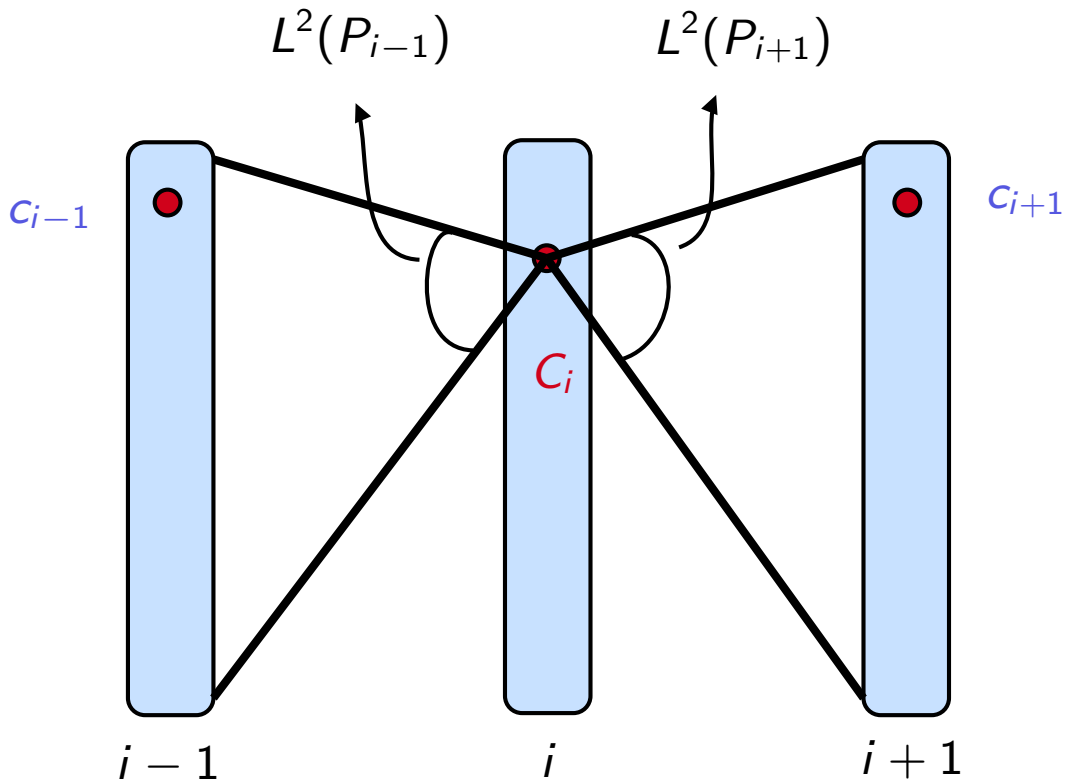
Assume some regularity properties of the activations, and

- $\partial_{\hat{y}} \ell(\hat{y}, y) = 0 \Rightarrow \ell(\hat{y}, y) = 0$.
- **Bidirectional diversity:** $(w_1(0, C_1), w_2(0, C_1, c_2))_{C_1 \sim P_1}$ *has full support in* $\mathbb{R}^d \times L^2(P_2)$ *and* $(w_i(0, c_{i-1}, C_i), w_{i+1}(0, C_i, c_{i+1}))_{C_i \sim P_i}$ *has full support in* $L^2(P_{i-1}) \times L^2(P_{i+1})$ *for all* $2 \leq i \leq L-1$.
- *Universal approximation:* $\{f_w(x) = \sigma_1(w \cdot x)\}_{w \in \mathbb{R}^d}$ *is dense in* $L^2(\mathbb{P}_x)$.
- $y = y(x)$.
- $w_i(t)$ *converges in appropriate sense as* $t \to \infty$.

Then under the mean field dynamics

$$\mathbb{E}_{(y,x)}[\ell(\hat{y}, y)] \to 0 \ as \ t \to \infty.$$

The bidirectional diversity condition:



Convergence assumptions:

Convergence assumptions:

- We assume that the gradient of the last layer weights $\partial_t w_L(t, C_{L-1}, C_L) \to 0$ uniformly, and appropriate convergence in moments.

Convergence assumptions:

- We assume that the gradient of the last layer weights $\partial_t w_L(t, C_{L-1}, C_L) \to 0$ uniformly, and appropriate convergence in moments.
- If $\mathbb{E}_{(y,x)}[\ell(\hat{y}, y)] \to 0$, then $\partial_t w_L(t, C_{L-1}, C_L) \to 0$ uniformly.

Convergence assumptions:

- We assume that the gradient of the last layer weights $\partial_t w_L(t, C_{L-1}, C_L) \to 0$ uniformly, and appropriate convergence in moments.
- If $\mathbb{E}_{(y,x)}[\ell(\hat{y}, y)] \to 0$, then $\partial_t w_L(t, C_{L-1}, C_L) \to 0$ uniformly.

Remarks.

Convergence assumptions:

- We assume that the gradient of the last layer weights $\partial_t w_L(t, C_{L-1}, C_L) \to 0$ uniformly, and appropriate convergence in moments.
- If $\mathbb{E}_{(y,x)}[\ell(\hat{y}, y)] \to 0$, then $\partial_t w_L(t, C_{L-1}, C_L) \to 0$ uniformly.

Remarks.

- Global convergence guarantee applies to gradient descent on the fully connected L-layer neural network with no modification to the architecture.

Convergence assumptions:

- We assume that the gradient of the last layer weights $\partial_t w_L(t, C_{L-1}, C_L) \to 0$ uniformly, and appropriate convergence in moments.
- If $\mathbb{E}_{(y,x)}[\ell(\hat{y}, y)] \to 0$, then $\partial_t w_L(t, C_{L-1}, C_L) \to 0$ uniformly.

Remarks.

- Global convergence guarantee applies to gradient descent on the fully connected L-layer neural network with no modification to the architecture.
- The bidirectional diversity condition necessitates correlated weight initialization distribution to avoid the degeneracy of the mean field dynamics of i.i.d. initializations.

Global convergence - L-layer neural network

Idea:

Idea:

- Consider the gradient of the last layer weights:

$$\mathbb{E}_{(y,x)}[\partial_{\hat{y}}\ell(\hat{y},y)\sigma_L'(H_L(t,x,C_L))\sigma_{L-1}(H_{L-1}(t,x,C_{L-1}))] \to 0. \quad \text{(Uniform conv.)}$$

Idea:

- Consider the gradient of the last layer weights:

$$\mathbb{E}_{(y,x)}[\partial_{\hat{y}}\ell(\hat{y},y)\sigma_L'(H_L(t,x,C_L))\sigma_{L-1}(H_{L-1}(t,x,C_{L-1}))] \to 0. \quad \text{(Uniform conv.)}$$

- **Key step.** Using a flow argument, we show that at all finite time t: $(w_1(t,C_1), w_2(t,C_1,c_2))_{C_1\sim P_1}$ has full support in $\mathbb{R}^d \times L^2(P_2)$ and $(w_i(t,c_{i-1},C_i), w_{i+1}(t,C_i,c_{i+1}))_{C_i\sim P_i}$ has full support in $L^2(P_{i-1}) \times L^2(P_{i+1})$ for all $2 \le i \le L-1$. (Topological invariance.)

Idea:

- Consider the gradient of the last layer weights:

$$\mathbb{E}_{(y,x)}[\partial_{\hat{y}}\ell(\hat{y},y)\sigma_L'(H_L(t,x,C_L))\sigma_{L-1}(H_{L-1}(t,x,C_{L-1}))] \to 0. \quad \text{(Uniform conv.)}$$

- **Key step.** Using a flow argument, we show that at all finite time t: $(w_1(t,C_1), w_2(t,C_1,c_2))_{C_1\sim P_1}$ has full support in $\mathbb{R}^d \times L^2(P_2)$ and $(w_i(t,c_{i-1},C_i), w_{i+1}(t,C_i,c_{i+1}))_{C_i\sim P_i}$ has full support in $L^2(P_{i-1}) \times L^2(P_{i+1})$ for all $2 \le i \le L-1$. (Topological invariance.)
- The diversity of the weights can be used to show that $H_i(t,x,C_i)$ has full support in $L^2(\mathbb{P}_x)$ for all $i \in [2, L-1]$. Thus, $\sigma_{L-1}(H_{L-1}(t,x,C_{L-1}))$ is dense in $L^2(\mathbb{P}_x)$.

Idea:

- Consider the gradient of the last layer weights:

$$\mathbb{E}_{(y,x)}[\partial_{\hat{y}}\ell(\hat{y},y)\sigma_L'(H_L(t,x,C_L))\sigma_{L-1}(H_{L-1}(t,x,C_{L-1}))] \to 0. \quad \text{(Uniform conv.)}$$

- **Key step.** Using a flow argument, we show that at all finite time t: $(w_1(t,C_1), w_2(t,C_1,c_2))_{C_1\sim P_1}$ has full support in $\mathbb{R}^d \times L^2(P_2)$ and $(w_i(t,c_{i-1},C_i), w_{i+1}(t,C_i,c_{i+1}))_{C_i\sim P_i}$ has full support in $L^2(P_{i-1}) \times L^2(P_{i+1})$ for all $2 \le i \le L-1$. (Topological invariance.)
- The diversity of the weights can be used to show that $H_i(t,x,C_i)$ has full support in $L^2(\mathbb{P}_x)$ for all $i \in [2, L-1]$. Thus, $\sigma_{L-1}(H_{L-1}(t,x,C_{L-1}))$ is dense in $L^2(\mathbb{P}_x)$.
- The conclusion follows combining the convergence condition with the denseness of $\sigma_{L-1}(H_{L-1}(t,x,C_{L-1}))$.

Global convergence III - General scheme

Our insight on a general scheme for a global convergence mechanism:

Global convergence III - General scheme

Our insight on a general scheme for a global convergence mechanism:

- Uniform convergence of the gradient update of an appropriate layer's weights.

Global convergence III - General scheme

Our insight on a general scheme for a global convergence mechanism:

- Uniform convergence of the gradient update of an appropriate layer's weights.
- Diversity: assumed at initialization, shown to hold at any finite time and across depth (topological invariance).

Global convergence III - General scheme

Our insight on a general scheme for a global convergence mechanism:

- Uniform convergence of the gradient update of an appropriate layer's weights.
- Diversity: assumed at initialization, shown to hold at any finite time and across depth (topological invariance).
- Universal approximation.

Global convergence III - General scheme

Our insight on a general scheme for a global convergence mechanism:

- Uniform convergence of the gradient update of an appropriate layer's weights.
- Diversity: assumed at initialization, shown to hold at any finite time and across depth (topological invariance).
- Universal approximation.
- Uniform convergence, diversity and universal approximation together imply global convergence (without convexity).

Global convergence III - General scheme

Our insight on a general scheme for a global convergence mechanism:

- Uniform convergence of the gradient update of an appropriate layer's weights.
- Diversity: assumed at initialization, shown to hold at any finite time and across depth (topological invariance).
- Universal approximation.
- Uniform convergence, diversity and universal approximation together imply global convergence (without convexity).

Remarks. The diversity condition is inspired by Chizat-Bach 2018, where global convergence is shown for two-layer networks by a different mechanism.

Conclusion

Conclusion

1. By constructing a **neuronal embedding** for the initialization distribution, we obtain a mean field limit that tracks the gradient descent dynamics of large width networks under mean field scaling.

Conclusion

1. By constructing a **neuronal embedding** for the initialization distribution, we obtain a mean field limit that tracks the gradient descent dynamics of large width networks under mean field scaling.
2. The mean field limit applies to general initialization distributions. When specializing to i.i.d. initializations, it recovers the degeneracy properties of the dynamics.

Conclusion

1. By constructing a **neuronal embedding** for the initialization distribution, we obtain a mean field limit that tracks the gradient descent dynamics of large width networks under mean field scaling.
2. The mean field limit applies to general initialization distributions. When specializing to i.i.d. initializations, it recovers the degeneracy properties of the dynamics.
3. The mean field limit can be used to understand global convergence guarantees of stochastic gradient descent.

Conclusion

1. By constructing a **neuronal embedding** for the initialization distribution, we obtain a mean field limit that tracks the gradient descent dynamics of large width networks under mean field scaling.
2. The mean field limit applies to general initialization distributions. When specializing to i.i.d. initializations, it recovers the degeneracy properties of the dynamics.
3. The mean field limit can be used to understand global convergence guarantees of stochastic gradient descent.
4. For deep fully connected networks, one can obtain global convergence guarantee assuming that the initialization distribution satisfies certain diversity condition that avoids the degeneracy of i.i.d. initializations.

Future directions

Future directions

- Quantitative rate of convergence.
- Uniform-in-time coupling with the mean field limit.
- Quantitative effect of depth.

References

A Rigorous Framework for the Mean Field Limit of Multilayer Neural Networks, Phan-Minh Nguyen, Huy Tuan Pham, Jan 2020. arXiv:2001.11443.

A Note on the Global Convergence of Multilayer Neural Networks in the Mean Field Regime, Huy Tuan Pham, Phan-Minh Nguyen, Jun 2020. arXiv:2006.09355.

Thank you!