STAT 946 - Lecture 12: Mean Field Networks and Multilayer Feature Learning

Course Notes by Marty Mukherjee

October 22, 2025

Contents

1 Preliminaries

Equivalence Relation (Obtained from Wikipedia)

A binary relation $\sim$ on a set X is said to be an equivalence relation if it is reflexive, symmetric and transitive. That is, for all $a, b, c \in X$:

- $a \sim a$ (reflexivity)
- $a \sim b \iff b \sim a$ (symmetry)
- $a \sim b, b \sim c \implies a \sim c$ (transitivity)

Equivalence Class

The equivalence class of $a \in X$ under $\sim$, denoted $[a]$, is defined as $[a] = \{x \in X : x \sim a\}$.

Quotient Set

The set of equivalence classes of X is called the quotient set of X, and denoted by

$$Q = X/\sim := \{[x] : x \in X\}$$

The map $\pi : X \to Q$ is called the quotient map if $\pi(x) = \pi(y) \iff x \sim y$.

Examples

1. Modulus: Let $X = \mathbb{R}$. Let $x \sim y$ if $x = y + k$, where k is an integer. An appropriate quotient map is $\pi : x \to x - \lfloor x \rfloor$
2. Permutation invariance: Let $X = \{[\theta_1, ..., \theta_n] : \theta_i \in \Theta, i = 1, ..., n\}$, where Θ is some arbitrary class. $[\theta_1, ..., \theta_n] \sim [\theta_{\sigma(1)}, ..., \theta_{\sigma(n)}]$ for all $\sigma \in S_n$ (σ is a permutation of indices). An appropriate quotient map is $\pi : [\theta_1, ..., \theta_n] \to \frac{1}{n} \sum_{i=1}^{n} \delta_{\theta_i}$

On Permutation Invariance

Permutation Invariance is the same thing as:

- graph isomorphism (CS)
- indistinguishable particles (Physics)
- exchangeability (Statistics)

2 Biggest Contribution of Deep Learning Theory: Hyperparameter Transfer

Neural Network Formulations [3]

Consider a one-hidden-layer network. In the NTK literature, it is expressed as

$$f(x; \theta) = \boxed{\frac{1}{\sqrt{n}}} W_1 \varphi(W_0 x),$$

and the learning rate scales as

$$\eta = \eta_0.$$

In the Mean Field (feature learning) literature, it is expressed as

$$f(x; \theta) = \boxed{\frac{1}{n}} W_1 \varphi(W_0 x),$$

and the learning rate scales as

$$\eta = \eta_0 \boxed{n}.$$

In both formulations, the parameters (W_1, W_0) are sampled i.i.d elementwise from a law $\mathcal{N}(0, 1)$.

Core Idea: If you have a number of different LR, and plot the training loss [2]:

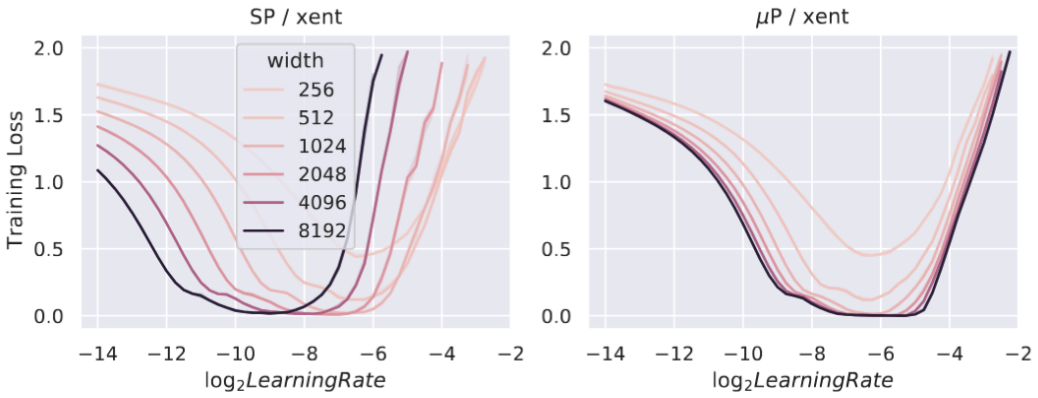


Figure 3: MLP width different hidden sizes trained for 20 epoch on CIFAR-10 using SGD. Left uses standard parametrization (SP); right uses maximal update parametrization (μP). μP networks exhibit better learning rate stability than their SP counterparts.

In standard parameterization settings (left), there is not only scaling, but also translation as the number of parameters (n) increases. Generally, the optimal learning rate decreases linearly with $\log n$.

Goal: Estimate the best learning rate for large values n

Problem: We can't train large models multiple times with different learning rates due to computational limitations.

- Sweep $L(\eta_0)$ for small values of n and model $\eta_0^*(n) = -c_1 \log n + c_2$.
- For large n, set the learning rate to $\eta_0^*(n)$

Of course, this is not ideal, as it requires extrapolating beyond the sweeping domain.

Alternatively, if we incorporate feature learning (right), all optimal learning rates are roughly the same. We don't know why this happens! (Open problem)

Maximal Update Parameterization (MUP/μP)

MUP rescales parameters to maintain update magnitudes as width grows, steering training away from the strict NTK regime and enabling nontrivial feature learning at infinite width.

Zero-shot Transfer (Informal)

Empirically, optimal learning rates (LR) for different widths align after an affine transform of the loss axis, allowing LR predicted at small width to extrapolate to large width. Incorporating feature learning sharpens this alignment and suggests a near-constant LR band across widths [2].

Remarks

All theories so far have finite training time T.

In Practice

A pragmatic rule of thumb is roughly 20 tokens (data points) per parameter ($\approx dn^2$) before significant overfitting arises.

Online SGD Budget

With streaming data (no reuse), the number of training iterations is proportional to the number of weights: $T \propto dn^2$

Brief intuition on the results obtained by feature learning

Let h be the hidden layer (features).

1. **NTK:** $\Delta h = \Theta(n^{-1/2})$
2. **Mean Field:** $\Delta h = \Theta(1)$

The optimal learning rate η^* shifts to compensate for Δh.

3 Multilayer Feature Learning

Recall Mean Field Distributions

Consider a multilayer network

$$\begin{aligned} f_n(x;\theta) &= \boxed{\frac{1}{\sqrt{n}}}\frac{1}{\sqrt{n}} W_d \phi(h_d), \\ h_{l+1} &= \frac{1}{\sqrt{n}} W_l \phi(h_l), \\ h_1 &= \frac{1}{n_0} W_0 x, \\ \eta &= \eta_0 \boxed{n}, \end{aligned}$$

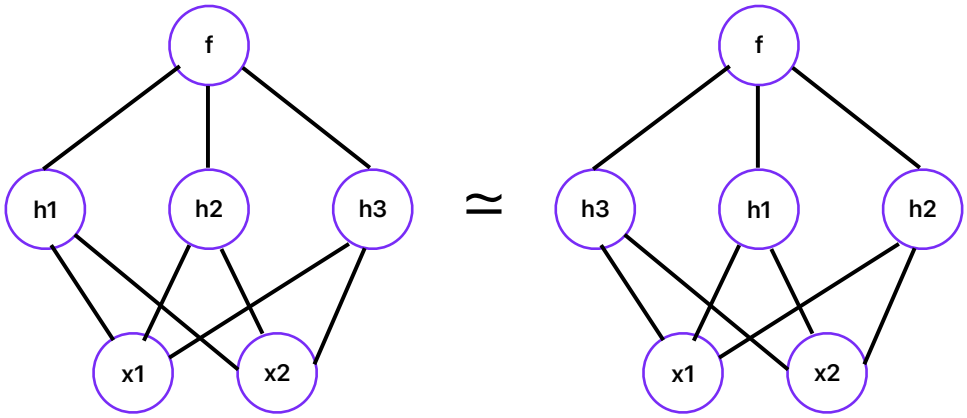
where $\boxed{\cdot}$ denotes additional terms in the mean field formulation. Consider the empirical distribution

$$\rho_t^{(n)} = \frac{1}{n}\sum_{i=1}^{n} \delta_{(W_{1,i}(t), W_{0,i}(t))}.$$

$\rho_t^{(n)}$ is the permutation invariance quotient map.

3.1 Permuted Neural Networks are Indistinguishable

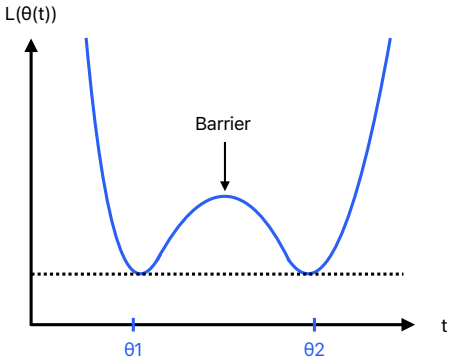
Intuition: Draw a NN and permute the hidden layer [1].



These two NNs are indistinguishable: We obtain the final layer by adding the hidden layers

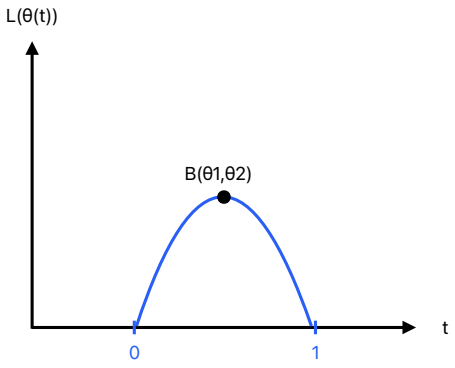
$$\begin{aligned} f(x;\theta) &= \frac{1}{n}\sum_{i=1}^{n} W_{1,i}\phi(\langle W_{0,i}, x\rangle) \\ &= \frac{1}{n}\sum_{i=1}^{n} W_{1,\sigma(i)}\phi(\langle W_{0,\sigma(i)}, x\rangle), \end{aligned}$$

where $\sigma \in S_n$ is a permutation of $[n]$. Consider a loss curve of a NN under a non-convex landscape $L(\theta)$.

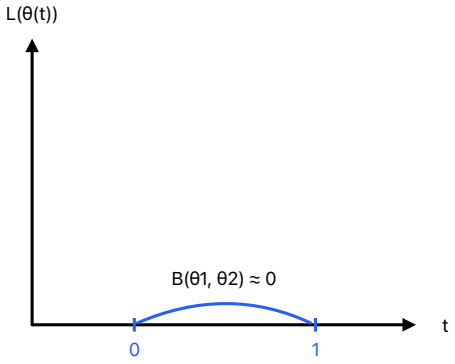


The NN converges twice in this train run at θ_1 and θ_2. WLOG set $L(\theta_1) = L(\theta_2) = 0$.

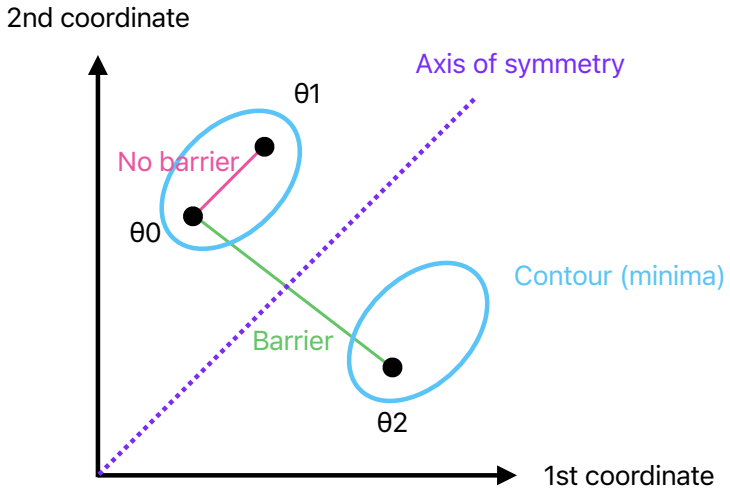
Let us consider an interpolation $\theta(t) = \theta_1 + t(\theta_2 - \theta_1)$ and plot the loss along this interpolation. Define $B(\theta_1, \theta_2) = \sup_{t \in [0,1]} L(\theta_1 + t(\theta_2 - \theta_1))$ as the barrier function.



You can find an optimal permutation of the weights such that $B(\theta_1, \sigma(\theta_2)) \approx 0$. Note: You are not in the same training landscape.



We obtain the following intuition from a contour plot that exploits the permutation invariance of the NN.



Conjecture

For all parameters θ_0, θ_1 that converge under SGD, there exists a layer-wise permutation σ such that $B(\theta_0, \sigma(\theta_1)) \approx 0$.
The proof (or counter-argument) of this is an open problem!

References

[1] Roozbeh Entezari et al. The role of permutation invariance in linear mode connectivity of neural networks. In *ICML*, 2021.

[2] Greg Yang et al. Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer. In *ICLR*, 2022.

[3] Greg Yang and Edward J. Hu. Tensor programs iv: Feature learning in infinite-width neural networks. In *NeurIPS*, 2021.