STAT 946 – Lecture 19: Feature Learning in the Proportional Limit

Course Notes by Marty Mukherjee

November 24, 2025

Big Picture

These notes follow Li-Noci-Hanin ("Feature Learning in the Proportional Limit") and study how neural tangent kernels and feature learning behave when both the depth d and width n grow with a fixed ratio $d/n \to \bar{\tau}$. The main goals of this lecture (I think) are:

- Understand the role of *Gaussian matrix conditioning* in the backward pass.
- Derive stochastic differential equations (SDEs) for the kernels G_l (Backward Feature) , Φ_l (Forward Feature) , Q_l (Omnidirectional).
- Interpret these SDEs as concrete evidence of *feature learning* and *hyperparameter transfer*.

Feature Learning in the Proportional Limit

(In Progress, Li, Noci, Hanin)

Throughout, we consider a dataset $\{(x^\alpha, y^\alpha)\}_{\alpha=1}^m$ and a fully-connected network of depth d and width n. We keep the notation as close as possible to the original paper.

1 Network Parameterization and Training Dynamics

Network and Activation Parameterization

We consider the network

$$\begin{aligned} f(x^\alpha;\theta) = h_{d+1}^\alpha &= \frac{1}{\sqrt{n}} W_d h_d^\alpha, \\ h_{l+1}^\alpha &= \sqrt{\frac{c}{n}} W_l \phi_s(h_l^\alpha), \quad l = 0, \dots, d-1, \\ h_1^\alpha &= \frac{1}{\sqrt{n}} W_0 x^\alpha, \end{aligned}$$

with weights $W_{l,ij} \sim \mathcal{N}(0,1)$ and a leaky ReLU activation

$$\phi_s(x) = s_+ \max(x,0) + s_- \min(x,0), \qquad s_\pm = 1 + \frac{c_\pm}{\sqrt{n}}.$$

We define the normalization constant

$$c^{-1} = \mathbb{E}\big[\phi_s(w)^2\big], \quad w \sim \mathcal{N}(0,1).$$

The loss and gradient flow dynamics are

$$\begin{aligned} L(\theta) &= \sum_{\alpha=1}^{m} \ell(f^\alpha, y^\alpha), \\ \partial_t \theta(t) &= -\eta \nabla_\theta L(\theta(t)) \quad \text{(NTK scaling)}. \end{aligned}$$

This parameterization is chosen so that pre-activations stay $\mathcal{O}(1)$ as $n \to \infty$ and the limit dynamics are non-trivial.

1.1 Backward Variables

Backward Pass Variables

For each sample α and layer l, we define two backward-pass objects:

- Post-activation gradient (with respect to h_l^α):

$$g_l^\alpha = \sqrt{\frac{n}{c}} \frac{\partial f^\alpha}{\partial h_l^\alpha},$$

- Pre-activation gradient (with respect to W_l):

$$z_l^\alpha = \frac{1}{\sqrt{n}} W_l^T g_{l+1}^\alpha.$$

These will be the building blocks of the layer-wise kernels.

2 Revisiting NTK in Deep Models

We first rewrite the backward pass more explicitly. The post-activation backward variable is

$$g_l^\alpha = \sqrt{\frac{n}{c}} \frac{\partial f^\alpha}{\partial h_l^\alpha}.$$

The pre-activation backward variable can be written as

$$\begin{aligned} z_l^\alpha &= \frac{1}{\sqrt{n}} W_l^T g_{l+1}^\alpha \\ &= \sqrt{\frac{c}{n}} W_l^T \operatorname{diag}\left(\phi_s'(h_{l+1}^\alpha)\right) z_{l+1}^\alpha \\ &=: \sqrt{\frac{c}{n}} W_l^T D_{l+1}^\alpha z_{l+1}^\alpha, \end{aligned}$$

where $D_{l+1}^\alpha := \operatorname{diag}(\phi_s'(h_{l+1}^\alpha))$. The evolution of $f^\alpha(t) := f(x^\alpha; \theta(t))$ under gradient flow is

$$\begin{aligned} \partial_t f(x^\alpha; \theta(t)) &= \langle \nabla_\theta f^\alpha, \partial_t \theta \rangle \\ &= \langle \nabla_\theta f^\alpha, -\eta \nabla_\theta L(\theta) \rangle \\ &= \sum_{l=0}^{d} \left\langle \nabla_{W_l} f^\alpha, -\eta \sum_{\beta=1}^{m} \partial_f \ell(f^\beta, y^\beta) \nabla_{W_l} f^\beta \right\rangle \\ &= \sum_{l=0}^{d} \sum_{\beta=1}^{m} \Delta^\beta \left\langle \nabla_{W_l} f^\alpha, \nabla_{W_l} f^\beta \right\rangle, \end{aligned}$$

where

$$\Delta^\beta := -\eta \, \partial_f \ell(f^\beta, y^\beta).$$

Using the chain rule and the parameterization,

$$\nabla_{W_l} f^\alpha = \frac{\partial f^\alpha}{\partial h_{l+1}^\alpha} \nabla_{W_l} \left(\sqrt{\frac{c}{n}} W_l \phi_s(h_l^\alpha) \right) = \sqrt{\frac{c}{n}} g_{l+1}^\alpha \sqrt{\frac{c}{n}} (\phi_l^\alpha)^T,$$

and thus

$$\left\langle \nabla_{W_l} f^\alpha, \nabla_{W_l} f^\beta \right\rangle = \frac{c}{n} \langle g_{l+1}^\alpha, g_{l+1}^\beta \rangle \frac{c}{n} \langle \phi_l^\alpha, \phi_l^\beta \rangle =: G_{l+1}^{\alpha\beta} \, \Phi_l^{\alpha\beta}.$$

Therefore, the NTK can be written as

$$\partial_t f^\alpha = \eta \sum_{l=0}^{d} \sum_{\beta=1}^{m} \Delta^\beta G_{l+1}^{\alpha\beta} \Phi_l^{\alpha\beta}.$$

Remark (I think this is true)

$G_{l+1}^{\alpha\beta}$ captures the *backward* feature geometry, while $\Phi_l^{\alpha\beta}$ captures the *forward* feature geometry.

Under the proportional limit $n, d \to \infty$,

$$\sum_{\beta=1}^{m} \Delta^{\beta} G_{l+1}^{\alpha\beta} \Phi_l^{\alpha\beta} = \Theta(1), \qquad \sum_{l=0}^{d} (\cdots) = \Theta(d),$$

so to keep $\partial_t f^{\alpha} = \Theta(1)$ we need

$$\eta = \Theta\left(\frac{1}{d}\right).$$

Guiding Questions.

- Can we obtain *feature learning* in this proportional limit?
- Is there room for *hyperparameter (HP) transfer*?
- Strategy: Study the dynamics of the feature kernel Φ_t.

3 Conditional Distributions in Forward and Backward Passes

Forward Distribution at Initialization

At initialization, with $W_{l,ij} \sim \mathcal{N}(0,1)$,

$$\left[h_{l+1}^{\alpha}\right]_{\alpha=1}^{m} \big| \mathcal{F}_l \overset{d}{=} \mathcal{N}\big(0, \Phi_l \otimes I_n\big),$$

for some Gram matrix $\Phi_l \in \mathbb{R}^{m \times m}$. Understanding the evolution of Φ_l is therefore important to understanding feature learning.

However, once we start training, the weights are no longer i.i.d. Gaussian. We must condition on the joint information from the forward and backward passes.

Gaussian Matrix Conditioning

For a Gaussian weight matrix W, the joint conditioning

$$W \,\Big|\, (W\phi,\, g^T W)$$

admits an explicit decomposition

$$W \,\big|\, W\phi, g^T W \;\overset{d}{=}\; P_g W + W P_\phi - P_g W P_\phi + P_g^\perp \tilde{W} P_\phi^\perp,$$

where

$$\begin{aligned}
\phi &:= [\phi^1, \ldots, \phi^m] \in \mathbb{R}^{n\times m},\\
g &:= [g^1, \ldots, g^m] \in \mathbb{R}^{n\times m},\\
P_\phi &:= \phi(\phi^T\phi)^{-1}\phi^T \quad \text{(projection onto span}(\phi)),\\
\tilde{W} &\text{ is an independent copy of } W.
\end{aligned}$$

This decomposition tells us how W_l changes under conditioning on forward/backward layers.

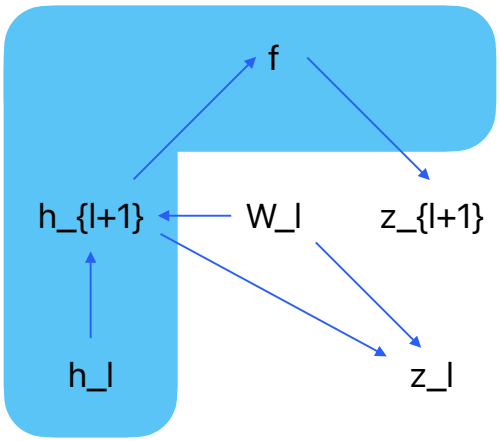
We revisit the backward pass:

$$z_l^\alpha = \sqrt{\frac{c}{n}} W_l^T \operatorname{diag}\left(\phi_s'(h_{l+1}^\alpha)\right) z_{l+1}^\alpha.$$

Note that both h_{l+1}^α and z_{l+1}^α depend on W_l, so conditioning can be done.

4 Filtration and Layer-wise Conditioning

Consider the graphical model:



As done in the NTK literature, we perform *conditioning* on the shaded area. While this simplifies the backward pass, it has the drawback that W_l is no longer Gaussian (or at least not elementwise iid).

Backward Filtration

To formalize what we condition on at each layer, we introduce the filtration

$$\mathcal{F}_l^z := \sigma\left(\{h_k^\alpha\}_{k\in[d+1],\alpha\in[m]},\ \{z_k^\alpha\}_{k\geq l,\alpha\in[m]}\right),$$

where $k \geq l$ because the backward pass propagates from the output layer backwards.

Then we can write:

$$\begin{aligned} W_l \big| \mathcal{F}_{l+1}^z &= W_l \big| W_l\phi_l \\ &\overset{d}{=} W_l P_{\phi_l} + \tilde{W}_l P_{\phi_l}^\perp, \end{aligned}$$

and hence

$$\begin{aligned} z_l^\alpha \big| \mathcal{F}_{l+1}^z &\overset{d}{=} \frac{1}{\sqrt{n}}\big(P_{\phi_l} W_l^T + P_{\phi_l}^\perp \tilde{W}_l^T\big) g_{l+1}^\alpha \\ &\overset{d}{=} \frac{1}{\sqrt{n}}\phi_l(\phi_l^T\phi_l)^{-1}\phi_l^T W_l^T g_{l+1}^\alpha + \frac{1}{\sqrt{n}} P_{\phi_l}^\perp \tilde{W}_l^T g_{l+1}^\alpha \\ &\overset{d}{=} \frac{1}{\sqrt{n}}\phi_l \frac{c}{n}\Phi_l^{-1}\phi_l^T W_l^T g_{l+1}^\alpha + P_{\phi_l}^\perp \tilde{z}_l^\alpha. \end{aligned}$$

Auxiliary Gaussian Variables

We introduce the notation

$$\begin{aligned} [\tilde{z}_l^\alpha]_{\alpha=1}^m &\sim \mathcal{N}(0, G_{l+1}\otimes I_n), \\ [P_{\phi_l}^\perp \tilde{z}_l^\alpha]_{\alpha=1}^m &\sim \mathcal{N}(0, G_{l+1}\otimes P_{\phi_l}^\perp), \end{aligned}$$

so that the residual term is explicitly Gaussian and orthogonal to the span of ϕ_l.

5 The Omnidirectional (?) Kernel Q_l and Its SDE

We now collect some identities that relate ϕ_l, h_{l+1}, g_{l+1}, and z_l.

Identities Connecting Forward and Backward Variables

Using the definitions above, we have

$$\phi_l^T W_l^T = h_{l+1}^T \sqrt{\frac{n}{c}},$$
$$W_l^T g_{l+1} = z_l \sqrt{\frac{n}{c}},$$
$$\phi_l^T W_l g_{l+1}^\alpha = \sqrt{\frac{n}{c}}\, h_{l+1}^T g_{l+1}^\alpha =: \sqrt{\frac{n}{c}} \langle h_{l+1}^\beta, g_{l+1}^\beta \rangle.$$

Omnidirectional Kernel Q_l

We define the omnidirectional kernel

$$Q_l^{\alpha\beta} := \sqrt{\frac{c}{n}} \langle h_l^\alpha, g_l^\beta \rangle.$$

5.1 Terminal Condition for Q_l

At the output layer, we get a clean terminal condition.

Terminal Condition for Q_{d+1}

At $l = d + 1$,

$$\begin{aligned} Q_{d+1}^{\alpha\beta} &= \sqrt{\frac{c}{n}} \langle h_{d+1}^\alpha, g_{d+1}^\beta \rangle \\ &= \sqrt{\frac{c}{n}} \left\langle f^\alpha, \sqrt{\frac{n}{c}} \frac{\partial f^\beta}{\partial h_{d+1}^\beta} \right\rangle \\ &= \sqrt{\frac{c}{n}} \langle f^\alpha, 1 \rangle \quad \text{(final linear layer)} \\ &= f^\alpha = \Theta(1). \end{aligned}$$

So $Q_{d+1}^{\alpha\beta}$ is of order one and encodes the initial alignment of outputs f^α.

5.2 Layer-wise Evolution of Q_l

For a (non-output) layer l,

$$\begin{aligned} Q_l^{\alpha\beta} &= \sqrt{\frac{c}{n}} \langle h_l^\alpha, g_l^\beta \rangle \\ &= \sqrt{\frac{c}{n}} \langle h_l^\alpha, \sqrt{\frac{c}{n}} D_l^\beta W_l^T g_{l+1}^\beta \rangle \\ &= \sqrt{\frac{c}{n}}\, h_l^{\alpha,T} \sqrt{\frac{c}{n}} D_l^\beta W_l^T g_{l+1}^\beta \\ &= \sqrt{\frac{c}{n}} \left\langle \sqrt{\frac{c}{n}} W_l D_l^\beta h_l^\alpha, g_{l+1}^\beta \right\rangle. \end{aligned}$$

On the other hand, the forward dynamics satisfy

$$\begin{aligned} h_{l+1}^{\alpha} &= \sqrt{\frac{c}{n}} W_l \phi_s(h_l^{\alpha}) \\ &= \sqrt{\frac{c}{n}} W_l \operatorname{diag}\left(\phi_s'(h_l^{\alpha})\right) h_l^{\alpha}. \end{aligned}$$

Combining these expressions, we can rewrite

$$\begin{aligned} Q_l^{\alpha\beta} &= \sqrt{\frac{c}{n}} \Big\langle \sqrt{\frac{c}{n}} W_l \big(D_l^{\alpha} h_l^{\alpha} + (D_l^{\beta} - D_l^{\alpha}) h_l^{\alpha}\big), g_{l+1}^{\beta} \Big\rangle \\ &= Q_{l+1}^{\alpha\beta} + \frac{1}{\sqrt{n}} \xi_{l+1}^{\alpha\beta}, \end{aligned}$$

for a noise term $\xi_{l+1}^{\alpha\beta}$.

SDE for the Omnidirectional Kernel Q

In the proportional limit, the layer index l rescales to a continuous "depth time" τ, and the recurrence above converges to a backward Itô SDE

$$dQ_\tau = \Sigma_\tau^{Q,1/2} \cdot dB_\tau^Q, \qquad Q_{\bar{\tau}}^{\alpha\beta} = f^{\alpha},$$

for some covariance operator Σ_τ^Q and Brownian motion B^Q indexed by (α, β).

6 Dynamics of G_l and Φ_l and the NTK Limit

For the kernel $G_l^{\alpha\beta}$, the conditional expectation is given (exercise) by

$$\mathbb{E}\big[G_l^{\alpha\beta} \mid \mathcal{F}_{l+1}^z\big] = G_{l+1}^{\alpha\beta} - \frac{m}{n} G_{l+1}^{\alpha\beta} + \frac{1}{n} Q_{l+1}^{:\alpha,T} \Phi_l^{-1} Q_{l+1}^{:\beta} + \frac{1}{n}\Big(K_\circ(\rho_l^{\alpha\beta}) - \tfrac{c_+^2 - c_-^2}{2}\Big) G_{l+1}^{\alpha\beta} + \mathcal{O}(n^{-3/2}),$$

where

$$K_\circ(\rho_l^{\alpha\beta}) = \mathbb{E}\big[\phi_s'(w^{\alpha})\phi_s'(w^{\beta})\big], \quad \begin{pmatrix} w^{\alpha} \\ w^{\beta} \end{pmatrix} \sim \mathcal{N}\Big(0, \begin{bmatrix} 1 & \rho_l^{\alpha\beta} \\ 0 & 1 \end{bmatrix}\Big).$$

Covariance of G_l

The conditional covariance satisfies

$$\operatorname{Cov}\left(G_l^{\alpha\beta}, G_l^{\gamma\delta} \mid \mathcal{F}_{l+1}^z\right) = \frac{1}{n}\big(G_l^{\alpha\gamma} G_l^{\beta\delta} + G_l^{\alpha\delta} G_l^{\beta\gamma}\big) + \mathcal{O}(n^{-3/2}).$$

Defining

$$\Sigma(G_{l+1})_{\alpha\beta,\gamma\delta} := G_l^{\alpha\gamma} G_l^{\beta\delta} + G_l^{\alpha\delta} G_l^{\beta\gamma},$$

one obtains another SDE in the limit.

6.1 NTK Limit

NTK as an Integral over Depth

The empirical NTK is

$$\frac{1}{d}\langle \nabla_\theta f^\alpha, \nabla_\theta f^\beta \rangle = \frac{1}{d}\sum_{l=0}^{d} G_{l+1}^{\alpha\beta}\Phi_l^{\alpha\beta}.$$

In the proportional limit $d/n \to \bar{\tau}$, this converges to

$$\frac{1}{\bar{\tau}}\int_0^{\bar{\tau}} G_\tau^{\alpha\beta}\Phi_\tau^{\alpha\beta}\, d\tau.$$

Moreover, one can organize the Brownian motions so that

- B^Φ is independent of $\{B^Q, B^G\}$,
- B^Q and B^G can be arranged to be independent in many regimes (but not always).

7 First Steps Toward Feature Learning Dynamics

The next step is to track the time evolution of the feature kernel

$$\Phi_d^{\alpha\beta} = \frac{c}{n}\langle \phi_s(h_d^\alpha), \phi_s(h_d^\beta)\rangle.$$

Differentiating Φ_d

Differentiating in training time t,

$$\begin{aligned}\partial_t \Phi_d^{\alpha\beta} &= \partial_t\left[\frac{c}{n}\langle \phi_d^\alpha, \phi_d^\beta\rangle\right] \\ &= \frac{c}{n}\langle \partial_t\phi_d^\alpha, \phi_d^\beta\rangle + \frac{c}{n}\langle \phi_d^\alpha, \partial_t\phi_d^\beta\rangle.\end{aligned}$$

To understand this, we need a recursion for $\partial_t h_l^\alpha$.

From the parameterization,

$$\partial_t h_{l+1}^\alpha = \sqrt{\frac{c}{n}}\,\partial_t W_l\,\phi_l^\alpha + \sqrt{\frac{c}{n}}\,W_l D_l^\alpha \partial_t h_l^\alpha,$$

which is a *forward recursion* in l.

New Inner Products for Feature Dynamics

Define the inner products

$$R_l^h := \frac{c}{n}\langle \partial_t h_l^\alpha, h_l^\beta \rangle,$$
$$R_l^z := \sqrt{\frac{c}{n}}\langle \partial_t h_l^\alpha, z_l^\beta \rangle,$$
$$\dot{\Phi}_l := \frac{c}{n}\langle \partial_t h_l^\alpha, \partial_t h_l^\beta \rangle.$$

In the proportional limit, these satisfy SDEs of the form

$$dR_\tau^{z,\alpha\beta} = \frac{\eta_0}{\bar{\tau}} \sum_{\mu=1}^{m} \Delta^\mu \Phi_\tau^{\alpha\mu} G_\tau^{\beta\mu}\, d\tau + \cdots\, dB_\tau^{R^z},$$

which already looks like a *feature-learning* term that couples Φ and G via the loss gradients.

Similarly,

$$dR_\tau^\phi = \left[R_\tau^{z\alpha;T}(\cdots) + R_\tau^\phi(\cdots)\right] d\tau + \cdots\, dB_\tau, \qquad R_0^\phi = 0.$$

In particular, if $R_\tau^z \equiv 0$, then $R_\tau^\phi \equiv 0$, and thus

$$\partial_t \Phi_\tau^{\alpha\beta} = R_\tau^{\phi,\alpha\beta} + R_\tau^{\phi,\beta\alpha}$$

vanishes. So non-trivial feature learning requires non-zero couplings R_τ^z.

Conceptual Takeaways

Conclusion

- The proportional limit $d/n \to \bar{\tau}$ leads to SDEs for the kernel triplet $(G_\tau, \Phi_\tau, Q_\tau)$.
- The term
$$\frac{1}{\bar{\tau}} \int_0^{\bar{\tau}} G_\tau^{\alpha\beta} \Phi_\tau^{\alpha\beta}\, d\tau$$
can be viewed as an average over random feature maps along depth.
- Non-zero R_τ^z drives the evolution of Φ_τ, which means *feature learning* and *hyperparameter transfer* occur in deep and wide networks.