# STAT 946:

## Lecture 11: Feature Learning

October 21, 2025 • Scribed by: **Edward Chang**

---

## Contents



# 1 Feature Learning

## 1.1 High-level points

- **Non-kernel learning**.
- Non-linear regime (beyond fixed NTK).

## 1.2 Recall: 1-layer network (NTK scaling)

We consider a single-hidden-layer network under NTK scaling

$$f(x;\theta) \;=\; \frac{1}{\sqrt{n}}\, W_1\, \phi\big(W_0 x\big), \qquad x \in \mathbb{R}^{n_0},\; W_0 \in \mathbb{R}^{n\times n_0},\; W_1 \in \mathbb{R}^{1\times n}, \tag{1}$$

with width $m$ and elementwise nonlinearity $\phi$.

## 1.3 Loss and training

Given data $\{(x^\alpha, y^\alpha)\}_{\alpha=1}^n$, use the squared loss

$$L(\theta) \;=\; \frac{1}{2m}\sum_{i=1}^{m} \big(f(x^\alpha;\theta) - y^\alpha\big)^2. \tag{2}$$

A gradient step with constant Learning Rate $\eta > 0$ is

$$\theta_{k+1} \;=\; \theta_k \;-\; \eta\, \nabla L(\theta_k). \tag{3}$$

### 1.4 Hidden features

Write the hidden *features* at step $k$ as

$$h^{\alpha}(\theta) \;=\; W_0 x^{\alpha} \in \mathbb{R}^m. \tag{4}$$

### 1.5 Feature dynamics

For a training point $x^n$, define

$$\Delta h^{\alpha} \;:=\; h^{(\alpha)}(\theta_1) \;-\; h^{(\alpha)}(\theta_0). \tag{5}$$

A gradient step on $W_0$ yields a change of the form

$$\Delta h^{\alpha} \;=\; -\,\frac{\eta}{m\sqrt{n}} \sum_{\beta=1}^{N} \big(f^{\beta} - y^{\beta}\big)\, \langle x^{\alpha}, x^{\beta}\rangle\, W_1^{\top} \;\otimes \phi'\big(h^{\beta}\big), \tag{6}$$

We note that as $n \to \infty$,

$$\eta, m, f^{\beta}, y^{\beta}, x^{\alpha}, x^{\beta}, W_1, \phi'(h^{\beta}) = \mathcal{O}(1)$$

hence,

$$\Delta h^{\alpha} \;=\; \mathcal{O}\Big(\tfrac{1}{\sqrt{n}}\Big) \tag{7}$$

Now, consider back to NTK. The (discrete) flow can be written with the NTK:

$$\Delta f^{\alpha} = \frac{\eta}{m} K^{(\alpha)}(f-y) = \mathcal{O}(1) \tag{8}$$

But we need $\Delta h^{\alpha} = \mathcal{O}(1)$ instead of $\mathcal{O}(n^{-1/2})$ for NTK to evolve.

## 2 Scaling Tweaks

### 2.1 Naive modification

Set the learning rate to

$$\eta \;=\; \eta_0\sqrt{n}.$$

Then

$$\Delta h^{\alpha} = \mathcal{O}(1), \qquad \Delta f^{\alpha} = \mathcal{O}(\sqrt{n}) \;\; \text{(diverges)}.$$

*Remark:* for cross-entropy loss this can still work (as noted in class).

### 2.2 Compromise: scale down the output

Use

$$f(x,\theta) \;=\; \frac{1}{n}\, W_1\, \phi(W_0 x). \tag{9}$$

Under this scaling,

$$\Delta h^{k} = \mathcal{O}\Big(\tfrac{1}{\sqrt{n}}\Big), \qquad \Delta f^{k} = \mathcal{O}\Big(\tfrac{1}{\sqrt{n}}\Big),$$

so changes are controlled (*cf.* the original where $\Delta h^k = \mathcal{O}(1)$ but $\Delta f^k = \mathcal{O}(\sqrt{n})$).

### 2.3 Mean-field parameterization

Pre-factor is $1/n$; choose the learning rate

$$\eta \;=\; \eta_0\, n,$$

which *accelerates the hidden layer* and escapes the strict kernel regime:

$$\Delta h^k = \mathcal{O}(1), \qquad \Delta f^k = \mathcal{O}(1).$$

### 2.4 Comparison

| | **Prefactor** | **LR** | **Init sd** |
|---|---|---|---|
| NTK (W1) | $1/\sqrt{n}$ (1) | 1 $(1/\sqrt{n})$ | 1 $(1/\sqrt{n})$ |
| Mean field | $1/n$ | $n$ | 1 |

In bracket is what people use in practice.

#### Remarks

- "ABC reparameterization"
- NTK linearized dynamics:
$$\Delta f \;\approx\; -\frac{\eta}{n}\, K\,(f-y),$$
which stays kernel-like unless features move. In the mean-field scaling, the *state space* dynamics are genuinely parameter-driven (nonlinear) rather than purely kernel.
- Full-parameter update: $\theta_{k+1} = \theta_k - \eta \nabla L(\theta_k)$, vs. NTK linearized in function space $f$ (with neurons summarized via $h, g$ **[(placeholders to match board notation)]**).

## 3 Mean-Field ODE

### 3.1 Network as an integral against an empirical measure

$$f(x;\theta) \;=\; \frac{1}{n}\sum_{i=1}^{n} w_{1,i}\, \phi\big(\langle w_{0,i}, x\rangle\big) \;=\; \int u(\langle w, x\rangle)\, d\rho^{(n)}(w,u), \tag{10}$$

where the *empirical measure* on parameter space is

$$\rho^{(n)} \;=\; \frac{1}{n}\sum_{i=1}^{n} \delta_{(w_{0,i}, w_{1,i})}.$$

### 3.2 Gradient flow (finite width)

Denote the population risk $L(\theta)$. Gradient flow is

$$\frac{d}{dt}\theta(t) = \partial_t \theta(t) \;=\; -\,\eta\, \nabla L\big(\theta(t)\big).$$

At the particle level, this induces ODEs for each $(w_{0,i}, w_{1,i})$ (indices suppressed for clarity):

$$\partial_t w_{1,i} = -\,\frac{1}{m}\sum_{\beta=1}^{n} \big(f^\beta(t) - y^\beta\big)\, \phi\big(\langle w_{0,i}, x^\beta\rangle\big),$$

$$\partial_t w_{0,i} = -\,\frac{1}{m}\sum_{\beta=1}^{n} \big(f^\beta(t) - y^\beta\big)\, W_1\, \phi'\big(\langle w_{0,i}, x^\beta\rangle\big)\, x^\beta.$$

Equivalently, the empirical measure $\rho_t^{(n)}$ evolves by transporting each particle according to a vector field

$$\partial_t \theta_i(t) \;=\; b\big(\theta_i(t)\;, \rho_t^{(n)}\big),$$

this is called *mean-field ODE* (McKean-Vlason).

### 3.3 Propagation of chaos

As $n \to \infty$,

$$\rho_t^{(n)} \implies \rho_t, \qquad \mathcal{L}\big(\theta_i(t)\big) \implies \rho_t,$$

where $\rho_t$ mean field measure, $\rho_t^{(n)}$ all particles, $L(\theta_i(t))$ contains one particle. This is quite surprising.
Stronger Notation:

$$\text{for each fixed } k: \quad \mathcal{L}\big(\theta_1(t), \ldots, \theta_k(t)\big) \implies \rho_t^{\otimes k} \quad \text{(asymptotic } k\text{-particle independence).}$$

A quantitative bound (for a suitable metric, e.g. Wasserstein-2) takes the form

$$W_2\Big(\mathcal{L}\big(\theta_1(t), \ldots, \theta_k(t)\big),\, \rho_t^{\otimes k}\Big) \;\lesssim\; \frac{1}{n} e^{c(t, n_0)}.$$

## 4 Mean-Field PDE

### 4.1 Test functions and weak derivatives

1. We generally cannot differentiate $\rho_t^{(n)}, \delta_\theta(t)$ .
2. The standard trick is to use a test function $q \in C_c^\infty(\mathbb{R}^{n_0+1})$ (smooth with compact support).

Informally, let's pretend the following works. (weak derivative)

$$\int q(\theta) \nabla_\theta \rho_t^{(n)}(\theta) d\theta \overset{IBP}{=} -\int \nabla_\theta q(\theta) d\rho_t^{(n)}(\theta)$$

Now consider,

$$\partial_t \int q(\theta)\, \rho_t^{(n)}(d\theta) \;=\; \frac{1}{n} \sum_{i=1}^{n} \langle \nabla_\theta q(\theta_i(t)),\, \partial_t \theta_i(t) \rangle \;=\; \int \langle \nabla_\theta q(\theta),\, b(\theta\;,\, \rho_t^{(n)}) \rangle\, \rho_t^{(n)}(d\theta).$$

$$= \int \nabla_\theta q(\theta) b(\theta, \rho_t^{(n)}) d\rho_t^{(n)}$$

Further, imagine this works:

$$\int \langle \nabla_\theta q(\theta),\, b(\theta \,|\, \rho_t^{(n)}) \rangle\, \rho_t^{(n)}(d\theta) \;\overset{IBP}{=}\; \int q(\theta) \Big[ -\operatorname{div}_\theta\big(b(\theta \,|\, \rho_t^{(n)})\, \rho_t^{(n)}(\theta)\big) \Big]\, d\theta$$

$$= \int q(\theta) \partial_t \rho_t^{(n)}(\theta) d\theta$$

And this is valid for any $q$. Hence $\rho_t^{(n)}$ is a *weak solution* of the continuity/transport PDE.
We define the first order non-linear "Transport Equation":

$$\boxed{\begin{aligned} &\partial_t \rho_t = -\operatorname{div}\Big(b(\theta\,,\, \rho_t)\, \rho_t\Big), \\ &\rho_0 = \rho_0^{(n)}. \end{aligned}} \tag{11}$$

This is a first-order, nonlinear **transport equation** (the mean-field PDE).

**References:**

1. Mei, Montanari, & Nguyen (2018).
2. Chizat & Bach (2018).
3. Nitanda & Suzuki (2017).
4. Rotskoff, Vanden-Eijnden (2018).
5. Sirignano & Spiliopoulos (2018).