$D \subseteq \mathbb{R}^{n_0} \times \mathbb{R}^{k}$, $\mathcal{X} = \{x \mid (x,y) \in D\}$, $\mathcal{Y} = \{y \mid (x,y) \in D\}$.

feed-forward network: $\begin{cases} h^{l+1} = x^{l} W^{l+1} + b^{l+1} \\ x^{l+1} = \phi(h^{l+1}) \end{cases}$, $\begin{cases} W^{l+1}_{ij} = \frac{\sigma_w}{\sqrt{n_l}} \omega^{l}_{ij} \\ b^{l}_{j} = \sigma_b \beta^{l}_{j} \end{cases}$ (NTK parameterization)

at init, $\omega^{l}_{ij}, \beta^{l}_{j} \sim N(0,1)$

$\theta^{l} \equiv \mathrm{vec}(\{W^{l}, b^{l}\}) \in \mathbb{R}^{(n_l(n_{l-1}+1))}$, $\theta = \mathrm{vec}\left(\bigcup_{l=1}^{L+1} \theta^{l}\right)$ (vector of all network parameters)

$f_t(x) \equiv h^{L+1}(x) \in \mathbb{R}^{k}$ (logits at time t)

$\ell(\hat{y}, y): \mathbb{R}^{k} \times \mathbb{R}^{k} \to \mathbb{R}$, $\ell(\hat{y}, y) = \frac{1}{2}\|\hat{y} - y\|_2^2$ (MSE loss)

We want to solve $\min_{\theta} \mathcal{L} = \min_{\theta} \sum_{(x,y) \in D} \ell(f_t(x,\theta), y)$

For continuous time GD, $\theta_{t+1} \leftarrow \theta_t - \eta \nabla_\theta \mathcal{L}$ where $\nabla_\theta \mathcal{L} = \nabla_\theta f_t(\mathcal{X})^T \nabla_{f_t(\mathcal{X})} \mathcal{L}$

$\Longrightarrow \Delta\theta_t = -\eta \nabla_\theta \mathcal{L}$

$$\frac{d\theta}{dt} = -\eta \nabla_\theta f_t(\mathcal{X})^T \nabla_{f_t(\mathcal{X})} \mathcal{L}$$

$$\frac{df_t(\mathcal{X})}{dt} = \nabla_\theta f_t(\mathcal{X}) \cdot \frac{d\theta}{dt} = -\eta \nabla_\theta f_t(\mathcal{X}) \nabla_\theta f_t(\mathcal{X})^T \nabla_{f_t(\mathcal{X})} \mathcal{L}$$

$= -\eta \hat{\Theta}_t(\mathcal{X}, \mathcal{X}) \nabla_{f_t(\mathcal{X})} \mathcal{L}$ where $f_t(\mathcal{X}) = \mathrm{vec}\left([f_t(x)]_{x \in \mathcal{X}}\right) \in \mathbb{R}^{k|D| \times 1}$ (concatenated logits for all examples)

and $\hat{\Theta}(\mathcal{X}, \mathcal{X}) \in \mathbb{R}^{k|D| \times k|D|}$ s.t. $\hat{\Theta}(\mathcal{X}, \mathcal{X}) = \nabla_\theta f_t(\mathcal{X}) \nabla_\theta f_t(\mathcal{X})^T = \sum_{l=1}^{L+1} \nabla_{\theta^l} f_t(\mathcal{X}) \nabla_{\theta^l} f_t(\mathcal{X})^T$

(tangent kernel at time t)

$$\frac{df_t(\mathcal{X})}{dt} = \sum_{l=1}^{L+1} \frac{\partial f_t(\mathcal{X})}{\partial \theta^{l}} \frac{d\theta^{l}}{dt} = \nabla_\theta f_t(\mathcal{X}) \frac{d\theta}{dt}$$

$$f_t^{lin}(x) \equiv \underbrace{f_0(x)}_{\text{const.}} + \underbrace{\nabla_\theta f_0(x)\big|_{\theta=\theta_0} \omega_t}_{\text{change to init. value during training}} \quad (\omega_t \equiv \theta_t - \theta_0)$$

$\longrightarrow \frac{d\omega_t}{dt} = -\eta \nabla_\theta f_0(\mathcal{X})^T \nabla_{f_t^{lin}(\mathcal{X})} \mathcal{L}$ since $\Delta\omega_t = \theta_t - \theta_0 = \Delta t \cdot \frac{d\theta}{dt}(0) = -\Delta t \, \eta \nabla_\theta f_0(\mathcal{X})^T \nabla_{f_t^{lin}(\mathcal{X})} \mathcal{L}$

$$\frac{df_t^{lin}(\mathcal{X})}{dt} = -\eta \hat{\Theta}_0(\mathcal{X}, \mathcal{X}) \nabla_{f_t^{lin}(\mathcal{X})} \mathcal{L}$$

Thm 2.1 (informal)

Assume: $n_1 = \cdots = n_L = n$, $\lambda_{min}(\Theta) > 0$

For GD with learning rate $\eta < \eta_{critical} = \frac{2}{\lambda_{min}(\Theta) + \lambda_{max}(\Theta)}$,

$\forall x \in \mathbb{R}^{n_0}$ s.t. $\|x\|_2 \le 1$,

with probability arbitrarily close to 1 over random init.,

$$\sup_{t \ge 0} \|f_t(x) - f_t^{lin}(x)\|_2, \ \sup_{t \ge 0} \frac{\|\theta_t - \theta_0\|_2}{\sqrt{n}}, \ \sup_{t \ge 0} \|\hat{\Theta}_t - \hat{\Theta}_0\|_F = O\left(\frac{1}{\sqrt{n}}\right) \quad \text{as} \quad n \to \infty$$

intuition.

- in the 'lazy' regime, individual weights barely change, but they collectively conspire to provide a finite change in the final output (big picture)
  - $\longrightarrow$ the network is perfectly described by a first-order approximation because the individual weight changes are so small.
    - this is the intuition for why the empirical NTK ($\hat{\Theta}_t$) should stay constant throughout training
    - stability of NTK is easier to show tabula rasa since there is no cumulation
- real network : $\dot{f}_t = -\eta \hat{\Theta}_t \nabla \mathcal{L}$ (empirical NTK)

  linear approx. : $\dot{f}_t^{lin} = -\eta \hat{\Theta}_0 \nabla \mathcal{L}$ (const. NTK from init)

  $\longrightarrow$ difference b/w the two is entirely determined by $\|\hat{\Theta}_t - \hat{\Theta}_0\|_F$ (error for derivatives)

  $\Longrightarrow \|f_t - f_t^{lin}\| \le$ some function of $\|\hat{\Theta}_t - \hat{\Theta}_0\|$

  using Gronwall's Inequality-type arguments (bounding functions using pre-existing bounds for the derivatives)

$$u'(t) \le \beta(t) u(t) \Longrightarrow u(t) \le u(0) \exp\left(\int_0^t \beta(s)\, ds\right)$$

$$\underbrace{u(t)}_{\propto \, dist(f_t, f_t^{lin})} \le \underbrace{\alpha(t)}_{\propto \int_0^t \|\hat{\Theta}_s - \hat{\Theta}_0\|_{op} \cdot [\text{training error of } f^{lin} \text{ at time } s] ds} + \int_0^t \underbrace{\beta(s)}_{\propto \|\hat{\Theta}_t - \hat{\Theta}_0\|_{op}} u(s)\, ds \Longrightarrow u(t) \le \alpha(t) \exp\left(\int_0^t \beta(s)\, ds\right) = \text{some function of } \|\hat{\Theta}_t - \hat{\Theta}_0\|$$

the integral is compounding error