1/28/2026 AMPDUP (NTK)

I. NTK

A. Intro.

1. fix consider fully-connected NN w/ layers $0, \dots, L$ (input, output)
   a. layers have $n_0, \dots, n_L$ neurons
   b. $\sigma : \mathbb{R} \xrightarrow{C^2} \mathbb{R}$, bdd 2nd derivative for nonlinearity
2. the realization fcn $F^{(L)} : \mathbb{R}^P \to \mathcal{F}$ for the NN maps $\theta \mapsto f_\theta$ (our NN)
   a. dim. $P = \sum_{\ell=0}^{L-1} (n_\ell + 1) n_{\ell+1}$ (bias)
3. parameters
   a. connection matrices $W^{(\ell)} \in \mathbb{R}^{n_{\ell+1} \times n_\ell}$
   b. bias vectors $b^{(\ell)} \in \mathbb{R}^{n_{\ell+1}}$
   c. initialize i.i.d. from $\mathcal{N}(0,1)$
4. consider ρ^{in} distrib on $\mathbb{R}^{n_0}$ and then define seminorm $\|\cdot\|_{\rho^{in}}$ by $\langle f, f \rangle_{\rho^{in}}^{\frac{1}{2}}$ for (literally just weighted L^2 norm)
   $$\langle f, g \rangle_{\rho^{in}} = \mathbb{E}_{x \sim \rho^{in}} (f(x)^T g(x))$$
   a. in particular, take ρ^{in} to be empirical distrib'n
   $$\rho^{in} = \frac{1}{N} \sum_{i=1}^{N} \delta_{x_i}$$
5. In particular,
   $$f_\theta(x) = \tilde{\alpha}^{(L)}(x; \theta)$$
   a. $\tilde{\alpha}^{(\ell)}(\cdot; \theta) : \mathbb{R}^{n_0} \to \mathbb{R}^{n_\ell}$ — preactivations
   b. $\alpha^{(\ell)}(\cdot; \theta) : \mathbb{R}^{n_0} \to \mathbb{R}^{n_\ell}$ — activations
   c. define recursively:
      i. $\alpha^{(0)}(x, \theta) = x$
      ii. $\tilde{\alpha}^{(\ell+1)}(x; \theta) = \frac{1}{\sqrt{n_\ell}} W^{(\ell)} \alpha^{(\ell)}(x; \theta) + \beta b^{(\ell)}$
      iii. $\alpha^{(\ell)}(x; \theta) = \sigma(\tilde{\alpha}^{(\ell)}(x; \theta))$
   d. here, σ is applied entry-wise
   e. β is a hyperparameter
   f. note by dividing by $\sqrt{n_\ell}$ helps when take width large.
   g. and β helps $b^{(\ell)}$ from dominating when n_ℓ is large if β is small enough

(* review: why do they divide by $\sqrt{n_\ell}$?)

B. Kernel gradient

1. a multi-dimensional kernel is a fcn $\mathbb{R}^{n_0} \times \mathbb{R}^{n_0} \to \mathbb{R}^{n_L \times n_L}$

$\exists B$ | such that for any $x, x' \in \mathbb{R}^{n_0}$,

$$K(x, x') = K(x', x)^T$$

a. equivalently, K is a symmetric tensor in $\mathcal{F} \otimes \mathcal{F}$

i. think of a basis for $\mathcal{F}$ as $\left\{ \begin{pmatrix} 0 \\ \delta_x \\ \vdots \\ 0 \end{pmatrix} \right\}$, then basis el't of $\mathcal{F} \otimes \mathcal{F}$ looks like $\begin{pmatrix} 0 & \cdots & 0 \\ & \delta_x \delta_{x'} & \\ 0 & & 0 \end{pmatrix}$

2. ~~then~~ given kernel, define bilinear map on $\mathcal{F}$ by

$$\langle f, g \rangle_K = \mathbb{E}_{x, x' \sim p^{in}} \left(f(x)^T K(x, x') g(x') \right)$$

independent

review LCTVS stuff

3. Def'n: K is pos. def. wrt $\|\cdot\|_{p^{in}}$ if

$$\|f\|_{p^{in}} > 0 \implies \|f\|_K > 0$$

4. $\mathcal{F}^* =$ dual of $\mathcal{F}$ wrt p^{in}

Since have inner product, can use Riesz rep'n?

a. is set of linear forms $\mu : \mathcal{F} \to \mathbb{R}$ of the form $\mu = \langle d, \cdot \rangle_{p^{in}}$ for some $d \in \mathcal{F}$

5. let $K_{i,\cdot}(x, \cdot)$ denote "partial application of the kernel"

6. define map $\Phi_K : \mathcal{F}^* \to \mathcal{F}$ by

$$\Phi_K(\mu)(x) = \mu(K_{i,\cdot}(x, \cdot))$$
$$= \langle d, K_{i,\cdot}(x, \cdot) \rangle_{p^{in}}$$

~~I think this is more of a "topology generated by a family of seminorms" thing...~~

where $\mu = \langle d, \cdot \rangle_{p^{in}}$

a. write $f_\mu = \Phi_K(\mu)$

7. now, cost funct'l C depends only on our data points $x_1, \ldots, x_N$

8. we can compute variational derivative of C — is an el't of $\mathcal{F}^*$

a. this derivative at $f_0 \in \mathcal{F}$ is denoted $\partial_f^{in} C|_{f_0}$

b. & we write $d|_{f_0}$ for el't of $\mathcal{F}$ s.t.

$$\partial_f^{in} C|_{f_0} = \langle d|_{f_0}, \cdot \rangle_{p^{in}}$$

9. define kernel gradient as

$$\nabla_K C|_{f_0} = \Phi_K(\partial_f^{in} C|_{f_0})$$

a. we have

$$\nabla_K C|_{f_0}(x) = \frac{1}{N} \sum_{j=1}^{N} K(x, x_j) d|_{f_0}(x_j)$$

I 8 10. Hence, could think of kernel gradient as ~~trying to find~~ approximating $\frac{\delta C}{\delta f}$ on space w/ ball measure than μ^{in} by using some kernel to add up its derivatives w/ Dirac measures

11. a fxn f follows the kernel gradient descent wrt K if it satisfies

$$\partial_t f(t) = -\nabla_K C|_{f(t)}$$

a. during such descent, cost evolves per

$$\partial_t C|_{f(t)} = -\langle dl_{f(t)}, \nabla_K C|_{f(t)} \rangle_{\mu^{in}} = -\|dl_{f(t)}\|_K^2$$

12. Thus, if C is convex & bdd from below & K is pos. def. wrt $\|\cdot\|_{\mu^{in}}$, then $f(t)$ will converge to a global min as $t \to \infty$.

13. ex.: Random fxn approximation

a. Pick some fxns $f^{(p)} \in \mathcal{F}$, $p = 1, \dots, P$

b. Use these to define a parametrization $F^{lin}: \mathbb{R}^P \to \mathcal{F}$ ~~$\partial_{\theta_p} F^{lin}(\theta) = \frac{1}{\sqrt{P}} f^{(p)}$~~ given by

$$F^{lin}(\theta) = \frac{1}{\sqrt{P}} \sum_{p=1}^{P} \theta_p f^{(p)}$$

c. Now, say params evolve through grad. descent:

~~d. then~~ $\partial_t \theta_p(t) = -\partial_{\theta_p} (C \circ F^{lin})(\theta(t))$

$$= -\frac{1}{\sqrt{P}} \partial_f^{in} C|_{F^{lin}(\theta(t))} \partial_{\theta_p} F^{lin}(\theta(t))$$

$$= -\frac{1}{\sqrt{P}} \partial_f^{in} C|_{F^{lin}(\theta(t))} f^{(p)} = -\frac{1}{\sqrt{P}} \langle dl_{F^{lin}_{\theta(t)}}, f^{(p)} \rangle_{\mu^{in}}$$

e. so $f^{lin}_{\theta(t)} = F^{lin}(\theta(t))$ evolves per

$$\partial_t f^{lin}_{\theta(t)} = \frac{1}{\sqrt{P}} \sum_{p=1}^{P} \partial_t \theta_p(t) f^{(p)} = -\frac{1}{P} \sum_{p=1}^{P} \langle dl_{f^{lin}_{\theta(t)}}, f^{(p)} \rangle_{\mu^{in}} f^{(p)}$$

f. but this is $-\nabla_{\tilde{K}} C$ for

$$\tilde{K} = \sum_{p=1}^{P} \frac{1}{P} f^{(p)} \otimes f^{(p)}$$

i.e., $\tilde{K}(x, x') = \frac{1}{P} \sum_{p=1}^{P} f_1^{(p)}(x) f_1^{(p)}(x')$

c. Neural tangent kernel (NTK)

1. In our real set up, we have

Recall $\partial_t \theta_p = -\partial_{\theta_p}(C \circ F^{(L)})(\theta(t))$

$= -\partial_f \hat{C}|_{F^{(L)}(\theta(t))} \partial_{\theta_p} F^{(L)}(\theta(t))$

$= -\langle d|_{F^{(L)}(\theta(t))}, \partial_{\theta_p} F^{(L)}(\theta(t)) \rangle_{p^{in}}$

b. so

$\partial_t f_{\theta(t)} = \sum_{p=1}^{P} \partial_t \theta_p \, \partial_{\theta_p} F^{(L)}(\theta(t))$

$= \sum_{p=1}^{P} -\langle d|_{F^{(L)}(\theta(t))}, \partial_{\theta_p} F^{(L)}(\theta(t)) \rangle_{p^{in}} \partial_{\theta_p} F^{(L)}(\theta(t))$

$= -\sum_{p=1}^{P} \sum_{i=1}^{n_L} \sum_{j=1}^{n_L} (d|_{F^{(L)}(\theta(t))}(x_j)) \partial_{\theta_p} F^{(L)}(\theta(t))(x_j) \partial_{\theta_p} F^{(L)}(\theta(t))$

$= -\partial_{\theta_p} F^{(L)}(\theta(t)) \partial_{\theta_p} F^{(L)}(\theta(t))(x_j)^T \, d|_{F^{(L)}(\theta(t))}(x_j)$

c. Hence, we set

THIS IS the NTK → $\Theta^{(L)}(\theta) = \sum_{p=1}^{P} \partial_{\theta_p} F^{(L)}(\theta) \otimes \partial_{\theta_p} F^{(L)}(\theta)$

d. $\sum_{j=1}^{n_L} f_{\theta,i}(x_j) = \Theta^{(L)}(\theta)(x, x_j)$, so

$\partial_t f_{\theta(t)} = -\partial_{\Theta^{(L)}} C|_{f_{\theta(t)}}$

2. Proposition: For a network of depth L at initialization w/ Lipschitz nonlinearity σ, in the limit $n_1, \ldots, n_{L-1} \to \infty$, the output functions $f_{\theta,k}$ $(k = 1, \ldots, n_L)$ tend (in law) to iid centered Gaussian processes of covariance $\Sigma^{(L)}$ defined by

a. $\Sigma^{(1)}(x, x') = \frac{1}{n_0} x^T x' + \beta^2$

b. $\Sigma^{(L+1)}(x, x') = \mathbb{E}_{f \sim \mathcal{N}(0, \Sigma^{(L)})}(\sigma(f(x))\sigma(f(x'))) + \beta^2$,

taking the expectation wrt a centered Gaussian process f of covariance $\Sigma^{(L)}$.

3. so:

a. $f_{\theta,k}$ is a fxn

b. but! since θ is randomly initialized, for any fixed input x, the value $f_{\theta,k}(x)$ may be seen as a RV

c. and we're just saying what the law of this RV is in limit as width $\to \infty$.

4. proof is by induction

a. In basis step, we have covariance $\frac{1}{n_0} x^T x' + \beta^2$

$$\Sigma^{(1)}(x,x') = \mathbb{E}_{\theta\sim\mathcal{N}(0,I)}\left(f_\theta(x) f_\theta(x')^T\right)$$
$$= \mathbb{E}_{\theta\sim\mathcal{N}(0,I)}\left(\frac{1}{\sqrt{n_0}}\left(W^{(0)}x+\beta b^{(0)}\right)\frac{1}{\sqrt{n_0}}\left(W^{(0)}x'+\beta b^{(0)}\right)^T\right)$$
$$= \frac{1}{n_0}\mathbb{E}\left(W^{(0)}x(W^{(0)}x')^T + W^{(0)}x'\beta b^{(0)T} + \beta b^{(0)}(W^{(0)}x')^T + \beta^2 b^{(0)}b^{(0)T}\right)$$

use iid $\mathcal{N}(0,1)$ assumption $\rightarrow = \frac{1}{n_0}x^Tx'I + 0 + 0 + \beta^2 I$

4. a. i. so $f_{\theta,b}$ are indepnt (diagonal matrix) cd all have the indicated covariance, as desired.

b. in induction step, suppose claim holds for L layers

c. so as $n_1,\dots,n_{L-1}\to\infty$, $\tilde{\alpha}_i^{(L)}$ tends to iid Gaussians w/ covariance $\Sigma^{(L)}$

d. so we must consider
$$f_{\theta,i} = \frac{1}{\sqrt{n_L}}W_i^{(L)}\alpha^{(L)} + \beta b_i^{(L)}$$

e. exactly like above, these have covariance
$$\Sigma^{(L+1)}(x,x') = \frac{1}{n_L}\alpha^{(L)}(x)^T\alpha^{(L)}(x') + \beta^2$$

f. so LLN: in limit, get
$$\Sigma^{(L+1)}(x,x') = \mathbb{E}_{f\sim\mathcal{N}(0,\Sigma^{(L)})}\left(\sigma(f(x))\sigma(f(x'))\right) + \beta^2$$

5. Thm: For a network of depth L at initialization w/ Lipschitz nonlinearity σ, in the limit as the widths $n_1,\dots,n_{L-1}\to\infty$, the NTK $\Theta^{(L)}$ converges in probability to a deterministic limiting kernel

Θ is unnecessary here

$$\Theta^{(L)} \to \Theta_\infty^{(L)}\otimes \mathrm{Id}_{n_L}$$

where scalar kernel $\Theta_\infty^{(L)}: \mathbb{R}^{n_0}\times\mathbb{R}^{n_0}\to\mathbb{R}$ is defined by

a. $\Theta_\infty^{(1)}(x,x') = \Sigma^{(1)}(x,x')$

b. $\Theta_\infty^{(L+1)}(x,x') = \Theta_\infty^{(L)}(x,x')\dot{\Sigma}^{(L+1)}(x,x') + \Sigma^{(L+1)}(x,x')$

for $\dot{\Sigma}^{(L+1)}(x,x') = \mathbb{E}_{f\sim\mathcal{N}(0,\Sigma^{(L)})}\left(\dot{\sigma}(f(x))\dot{\sigma}(f(x'))\right)$

i. with expectation wrt centered Gaussian process of covariance $\Sigma^{(L)}$

6. proof is by induction

a. for $L=1$, use
$$(\partial_{W_{ij}}F)_k = \delta_{ik}x_j\frac{1}{\sqrt{n_0}}$$

i. so $\partial_{W_{ij}}F\,(\partial_{W_{ij}}F)^T = \frac{1}{n_0}\delta_{ik}x_jx_j'$

ii. so $\Theta^{(1)}(\theta)(x,x') = \frac{1}{n_0}x^Tx'\,\mathrm{Id}_{n_1} + \beta^2 = \Sigma^{(1)}(x,x')$

If CLT for induction, say holds for L layers

c. as $n_1, \dots, n_{L-1} \to \infty$, $\tilde{z}_i^{(L)}$ goes to iid Gaussians w/ covariance $\Sigma^{(L)}$ per prev. proposition

d. Look at $\tilde{\theta}_p$ belongs to first L layers, then

$$\partial_{\tilde{\theta}_p} F^{(L+1)} = \frac{1}{\sqrt{n_L}} W^{(L+1)} D\sigma(\tilde{z}^{(L)}) \partial_{\tilde{\theta}_p} \tilde{z}^{(L)}$$

e. so for these parameters, get

$$\sum \partial_{\tilde{\theta}_p} F^{(L+1)}(x) (\partial_{\tilde{\theta}_p} F^{(L+1)}(x'))^T$$

$$= \sum \frac{1}{n_L} W^{(L+1)} D\sigma(\tilde{z}^{(L)}) \partial_{\tilde{\theta}_p} \tilde{z}^{(L)} \partial_{\tilde{\theta}_p} \tilde{z}^{(L)T} D\sigma(\tilde{z}^{(L)T}) W^{(L+1)T}$$

I.H. $\to$ $n_1, \dots, n_{L-1} \to \infty$: $\frac{1}{n_L} \Theta^{(L)}_{\theta_p}(x,x') W^{(L+1)} D\sigma(\tilde{z}^{(L)}) D\sigma(\tilde{z}^{(L)T}) W^{(L+1)T}$

LLN, Induce $\to n_{L+1} \to \infty$: $\Theta^{(L)}_{\infty}(x,x') \, Id \, \mathbb{E}_{(z,z') \sim \mathcal{N}(0, \Sigma^{(L)})} [\dot{\sigma}(z(x)) \dot{\sigma}(z(x'))]$

e. next, for other params, just use

$$\partial_{w_{ij}} F^{(L+1)} = \frac{1}{\sqrt{n_L}} \sigma(\tilde{z}^{(L)}_j(x)) e_i$$

f. so $\sum_{ij} \partial_{w_{ij}} F^{(L+1)} (\partial_{w_{ij}} F^{(L+1)})^T = \frac{1}{n_L} \sum_j \sigma(\tilde{z}^{(L)}_j(x)) \sigma(\tilde{z}^{(L)}_j(x')) \, Id$

g. use proposition, don't forget β_1^2, you're good.

D. Training

just added. 1. general def'n of training: update params per

$$\partial_t \theta_p(t) = \langle \partial_{\theta_p} F^{(L)}(\theta(t)), d_t \rangle_{p^{in}}$$

de

what does "stochastically bounded" mean?

$\to$ where $\int_0^T \|d_t\|_{p^{in}} dt$ is "stochastically bounded" as width goes to ∞

b. for gradient descent, take

$$d_t = -d|_{f(\theta(t))}$$

2. Thm: Assume σ Lipschitz, 2nd cts nonlinearity $\mathbb{R}$ on w/ bounded 2nd derivative. For any T s.t. $\int_0^T \|d_t\|_{p^{in}} dt$ stays stochastically bounded, as $n_1, \dots, n_{L-1} \to \infty$, we have, uniformly for $t \in [0,T]$,

$$\Theta^{(L)}(t) \to \Theta^{(L)}_{\infty} \otimes Id_{n_L}$$

a. As a consequence, in this limit, the dynamics of f_θ is described by the diff eq.

$$\partial_t f_{\theta(t)} = \Phi_{\Theta^{(L)}_{\infty} Id_{n_L}}(\langle d_t, \cdot \rangle_{p^{in}})$$

or $\partial_t f_{\theta(t)}(x) = \langle d_t, \Theta^{(L)}_{\theta}(x, \cdot) Id \rangle$

3. proof is by induction

a. for $L=1$, NTK doesn't depend on params, so it

by reverse triangle $\partial_t \|g(t)\| = \lim_{h\to 0} \frac{\|g(t+h)\| - \|g(t)\|}{h} \le \lim_{h\to 0} \frac{\|g(t+h)-g(t)\|}{h} = \|\partial_t g(t)\|$

constant.

I036. for other L, need a lemma

4. lemma: In setting of thm 2, for a network of depth $L+1$, for any $l = 1, \dots, L$ we have convergence in probability

$$\lim_{n_0\to\infty} \cdots \lim_{n_{L-1}\to\infty} \sup_{t\in[0,T]} \left\| \frac{1}{\sqrt{n_L}} \left(W^{(l)}(t) - W^{(l)}(0) \right) \right\|_{op} = 0$$

does not apply to $W^{(0)}$

4. proof

a. we have

$$\partial_t W_{ij}^{(l)} = \frac{1}{\sqrt{n_L}} \langle \alpha_i^{(l)} \theta_j, d_t^{(l+1)} \rangle_{in}$$
$$\partial_t \tilde{z}^{(l)} = \odot \ominus^{(l)} \left(\langle d_t^{(l)}, - \rangle_{in} \right)$$

for

$$d_t^{(l)} = \begin{cases} d_t & \text{if } l = L+1 \\ \dot\sigma(\tilde z^{(l)}) \frac{1}{\sqrt{n_l}} (W^{(l)})^T d_t^{(l+1)} & \text{if } l \le L \end{cases}$$

b. let $w^{(l)}(t) = \| \frac{1}{\sqrt{n_L}} W^{(l)}(t) \|_{op}$, $a^{(l)}(t) = \| \frac{1}{\sqrt{n_L}} \alpha^{(l)}(t) \|_{in}$

why ℓ^{in} instead of Euclidean norm? ↑ see the Lipschitz constant

c. then

$$\| d_t^{(l)} \|_{in} \le c\, w^{(l)}(t) \| d_t^{(l+1)} \|_{in} \le \cdots \le c^{L+1-l} \prod_{k=l}^{L} w^{(k)}(t) \, \|d_t\|_{in}$$

d. hidden sub networks can be written

$$\ominus^{(1)} = \frac{1}{n_0} \alpha^{(0)T} \alpha^{(0)} Id_{n_1} + \beta^2 Id_{n_1}$$
$$\ominus^{(l+1)} = \frac{1}{n_l} W^{(l)} \dot\sigma(\tilde z^{(l)}) \ominus^{(l)} \dot\sigma(\tilde z^{(l)})^T W^{(l)T} + \frac{1}{n_l} \alpha^{(l)T} \alpha^{(l)} Id_{n_{l+1}} + \beta^2 Id_{n_{l+1}}$$

does this induce ℓ^{in} norm on $\ominus$? ↑ yes:

$$\|C\|_{op} = \max_{\|f\|_{in}\le 1} \|Cf\|_{in} = \max_{\|f\|_{in}\le 1} \mathbb{E}_{x\sim\rho^{in}} \left((Cf)^T (Cf) \right)^{1/2}$$

e. so set bounds

$$\|\ominus^{(1)}\|_{op} \le (a^{(0)}(t))^2 + \beta^2$$
$$\|\ominus^{(l+1)}\|_{op} \le c^2 (w^{(l)}(t))^2 \|\ominus^{(l)}\|_{op} + (a^{(l)}(t))^2 + \beta^2$$
$$\le \cdots \le P\left(a^{(1)}, \dots, a^{(l)}, w^{(1)}, \dots, w^{(l)}\right)$$

for P some poly.

f. now, define

$$\tilde a^{(l)}(t) = \left\| \frac{1}{\sqrt{n_L}} \left(\tilde z^{(l)}(t) - \tilde z^{(l)}(0) \right) \right\|_{in}$$
$$\tilde w^{(l)}(t) = \left\| \frac{1}{\sqrt{n_L}} \left(W^{(l)}(t) - W^{(l)}(0) \right) \right\|_{op}$$
$$A(t) = \sum_{k=1}^{L} a^{(k)}(0) + c\, \tilde a^{k}(t) + w^{(0)}(0) + w^{(0)}(t)$$

i. note $a^{(l)}(t), w^{(l)}(t) \le A(t)$

g. next, by eq'ns for evolution of $W, \tilde z$, we set

304, $\partial_t \tilde{w}^{(\ell)}(t) \le \frac{1}{\sqrt{n_\ell}} a^{(\ell)}(t) \|d_t^{(\ell+1)}\|_{p,n}$

$\partial_t a^{(\ell)}(t) \le \frac{1}{\sqrt{n_\ell}} \|\Theta^{(\ell)}(t)\|_{op} \|d_t^{(\ell)}\|_{p,n}$

b. therefore

$$\partial_t A(t) \le \sum_{\ell=1}^{L} \frac{c}{\sqrt{n_\ell}} \|\Theta^{(\ell)}(t)\|_{op} \|d_t^{(\ell)}\|_{p,n} + \frac{1}{\sqrt{n_\ell}} a^{(\ell)}(t) \|d_t^{(\ell+1)}\|_{p,n}$$

$$\le \frac{1}{\sqrt{\min\{n_1,\dots,n_L\}}} \Sigma(w^{(1)}(t),\dots,w^{(L)}(t), a^{(1)}(t),\dots,a^{(L)}(t)) \|d_t\|_{p,n}$$

$$\le \frac{1}{\sqrt{\min\{n_1,\dots,n_L\}}} R(A(t)) \|d_t\|_{p,n}$$

for new poly's Σ, R

i. Now: use a Gronwall lemma

ii. then like $A(t) = \sum_{\ell=1}^{L} a^{(\ell)}(t) + w^{(\ell)}(t)$ is stochastically bounded as $n_\ell \to \infty$, and so is $\int_0^t \|d_s\|_{p,n}$

iii. bc. of $\min\{n_1,\dots,n_L\}$ part, we actually get that $A(t)$ stays uniformly bounded on $[0,T]$

iv. so, in order get $\partial_t A(t) \to 0$ uniformly as $A(t) \to A(0)$

v. roughly,

$$A(t) \le A(t) \le A(0) \exp(\underbrace{\dots}_{\to 0 \text{ as } n_1,\dots,n_L \to \infty})$$

5. very unintuitive: weights and activations approach constant as width $\to \infty$

a. "Lazy Training in Differentiable Programming" by Chizat et. al

b. "On the linearity of large non-linear models: when and why the tangent kernel is constant" by Liu et al.

c. both say it's like the NN approaches a linear NN

d. but 2nd does Hessian analysis

6. least squares regression

i. $C(f) = \frac{1}{2}\|f - f^*\|_{p,n}^2$ for target f^*

ii. things apply to this cost