Title: Scaling Limits of Bayesian Inference with Deep Neural Networks

Speakers: Boris Hanin

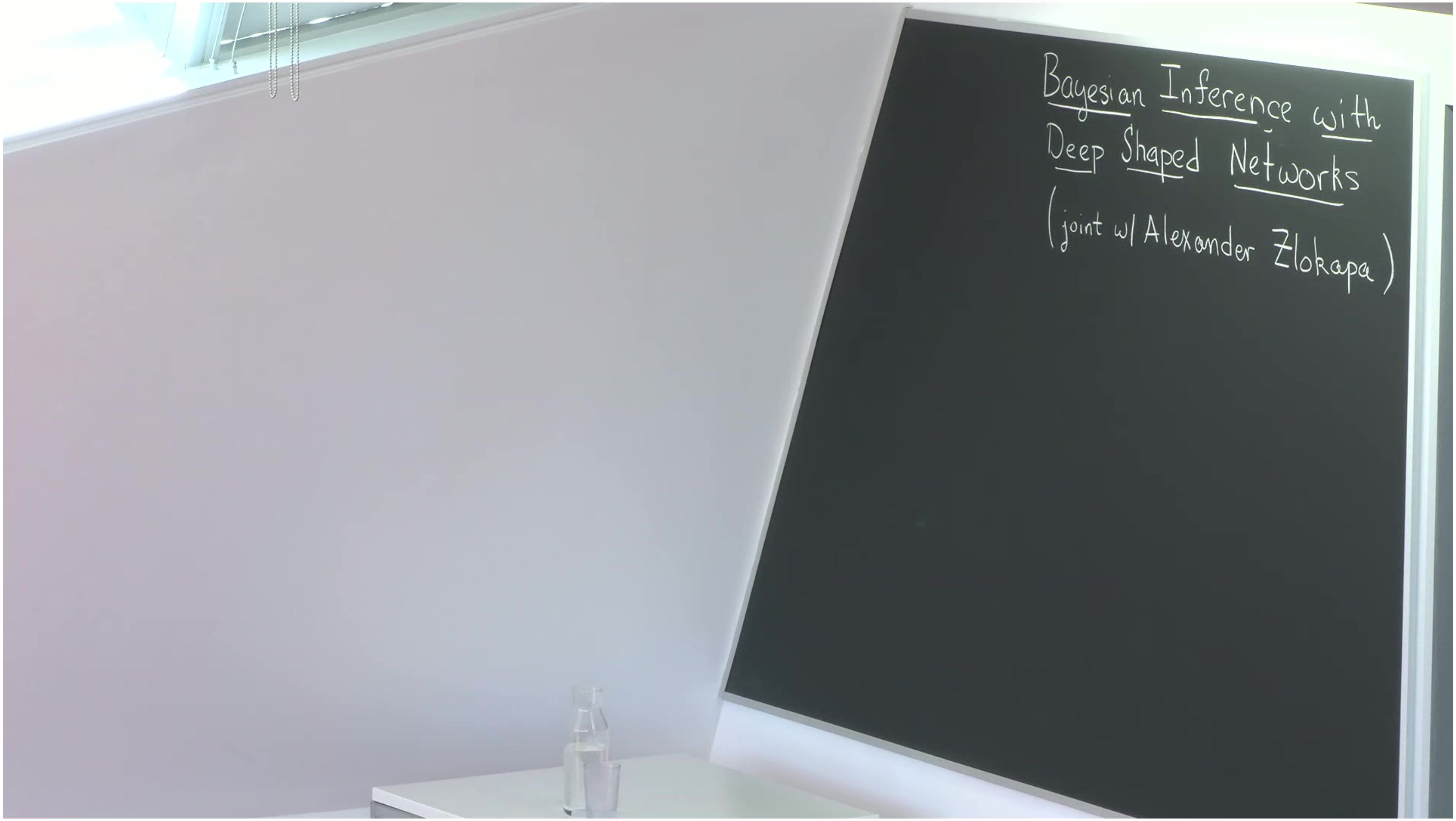
Series: Machine Learning Initiative

Date: April 19, 2024 - 2:30 PM

URL: https://pirsa.org/24040103

Abstract: Large neural networks are often studied analytically through scaling limits: regimes in which some structural network parameters (e.g. depth, width, number of training datapoints, and so on) tend to infinity. Such limits are challenging to identify and study in part because the limits as these structural parameters diverge typically do not commute. I will present some recent and ongoing work with Alexander Zlokapa (MIT), in which we provide the first solvable models of learning - in this case by Bayesian inference - with neural networks where the depth, width, and number of datapoints can all be large.

Zoom link


Bayesian Inference with
Deep Shaped Networks
(joint w/ Alexander Zlokapa)

Data: $\mathcal{D} = \{(x_\mu, y_\mu), \mu = 1, \ldots, P\}$ $\quad x_\mu \in \mathbb{R}^{N_0}, \; y_\mu \in \mathbb{R}$

Model: $f(x;\theta) = W^{(L+1)} \sigma W^{(L)} \cdots \sigma W^{(1)} x$, $W^{(\ell)} - N_\ell \times N_{\ell-1}$

$\sigma\begin{pmatrix} v_1 \\ \vdots \\ v_N \end{pmatrix} \equiv \begin{pmatrix} \sigma(v_1) \\ \vdots \\ \sigma(v_N) \end{pmatrix}$

Learning Rule: $\mathcal{A} : (\mathcal{D}, f) \longmapsto \theta_{\text{learned}}$

Q1: Analysis of $\mathcal{A}$ when $P, N_\ell, L \gg 1$

Q2: Adaption of $f(x;\theta_{\text{learned}})$ to $\mathcal{D}$?

Q3: Alignment of $f, \mathcal{D}, \mathcal{A}$

This Talk

• $\sigma(t$

• $\mathcal{A} =$

$\in \mathbb{R}$

$\times N_{\ell-1}$

This Talk: Q1 — Q3 for

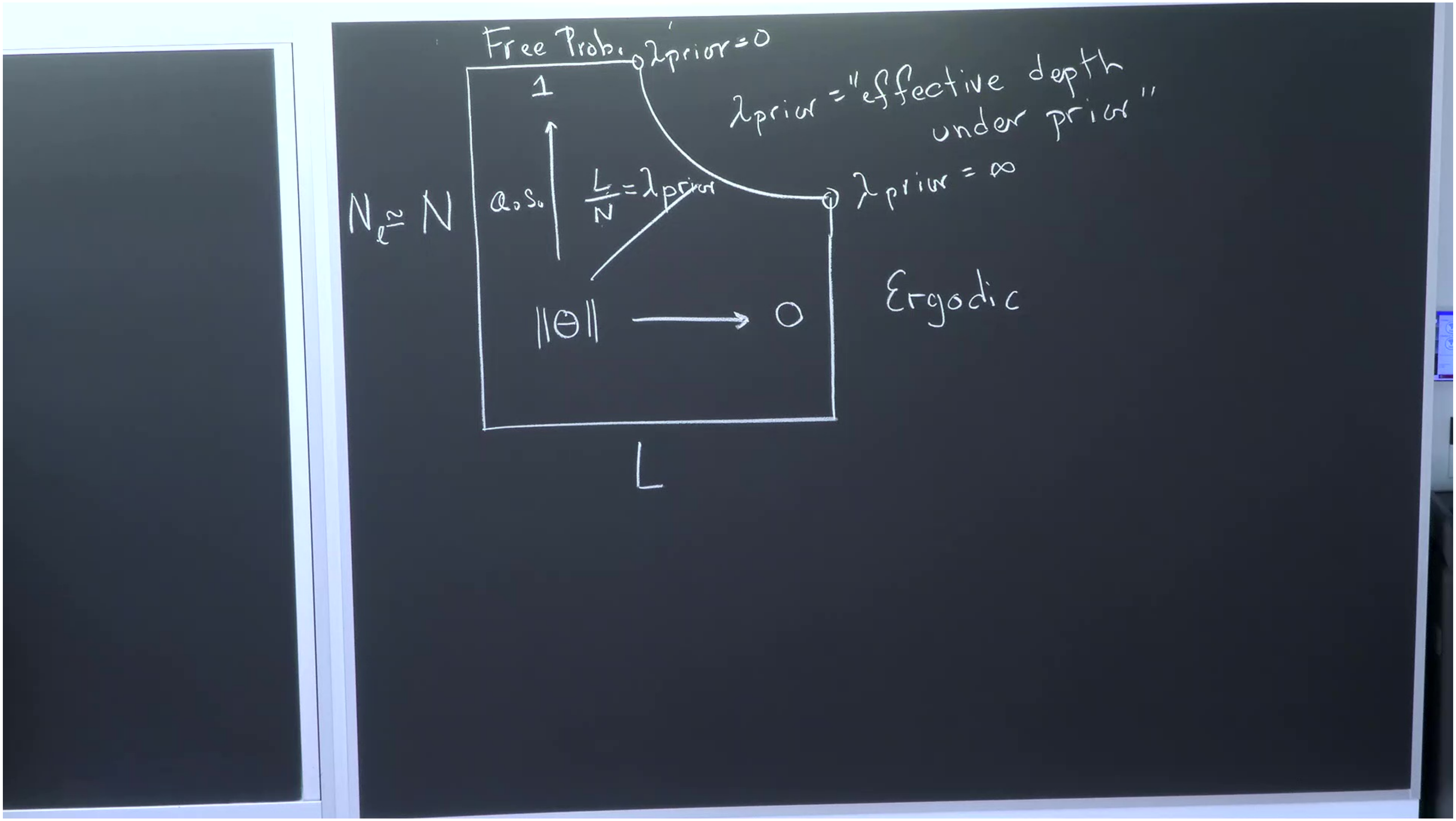
- $\sigma(t) = t + \frac{\Psi}{L} t^3 \simeq \sqrt{L}\, \varphi_{odd}(t/\sqrt{L})$
- A = Bayesian Inference @ $T=0$, $P < N_0$
  $N_{L+1} = 1$

$\Psi = 0$: (Deep Linear)

$$f(x;\theta) = W^{(L+1)} W^{(L)} \cdots W^{(1)} x = \Theta^T x$$

prior: $\Theta \sim \mathbb{P}_{prior} \iff W_{ij}^{(\ell)} \sim \mathcal{N}(0, 1/N_{\ell-1})$ iid

$\mathbb{R}^{N_0}$


Free Prob.
$\lambda_{prior} = 0$
1
λ_{prior} = "effective depth under prior"
$N_\ell \simeq N$
a.s.
$\frac{L}{N} = \lambda_{prior}$
$\lambda_{prior} = \infty$
$\|\Theta\|$
0
Ergodic
L

posterior:
$$\begin{cases} \mathbb{P}_{post}(\theta | D, N_e, L) = \lim\limits_{\beta \to \infty} \dfrac{\mathbb{P}_{prior}(\theta | N_e, L) \exp\{-\beta \mathcal{L}(\theta | D)\}}{Z_\beta(D | N_e, L)} \\ \mathcal{L}(\theta | D) \equiv \dfrac{1}{2P} \sum\limits_{\mu=1}^{P} \left(f(x_\mu; \theta) - y_\mu\right)^2 \end{cases}$$

NB: $\mathbb{P}_{post}(\theta) \propto \mathbb{P}_{prior}(\theta)\, \delta_{\{\theta^T x_\mu = y_\mu \forall \mu\}}$

This Talk

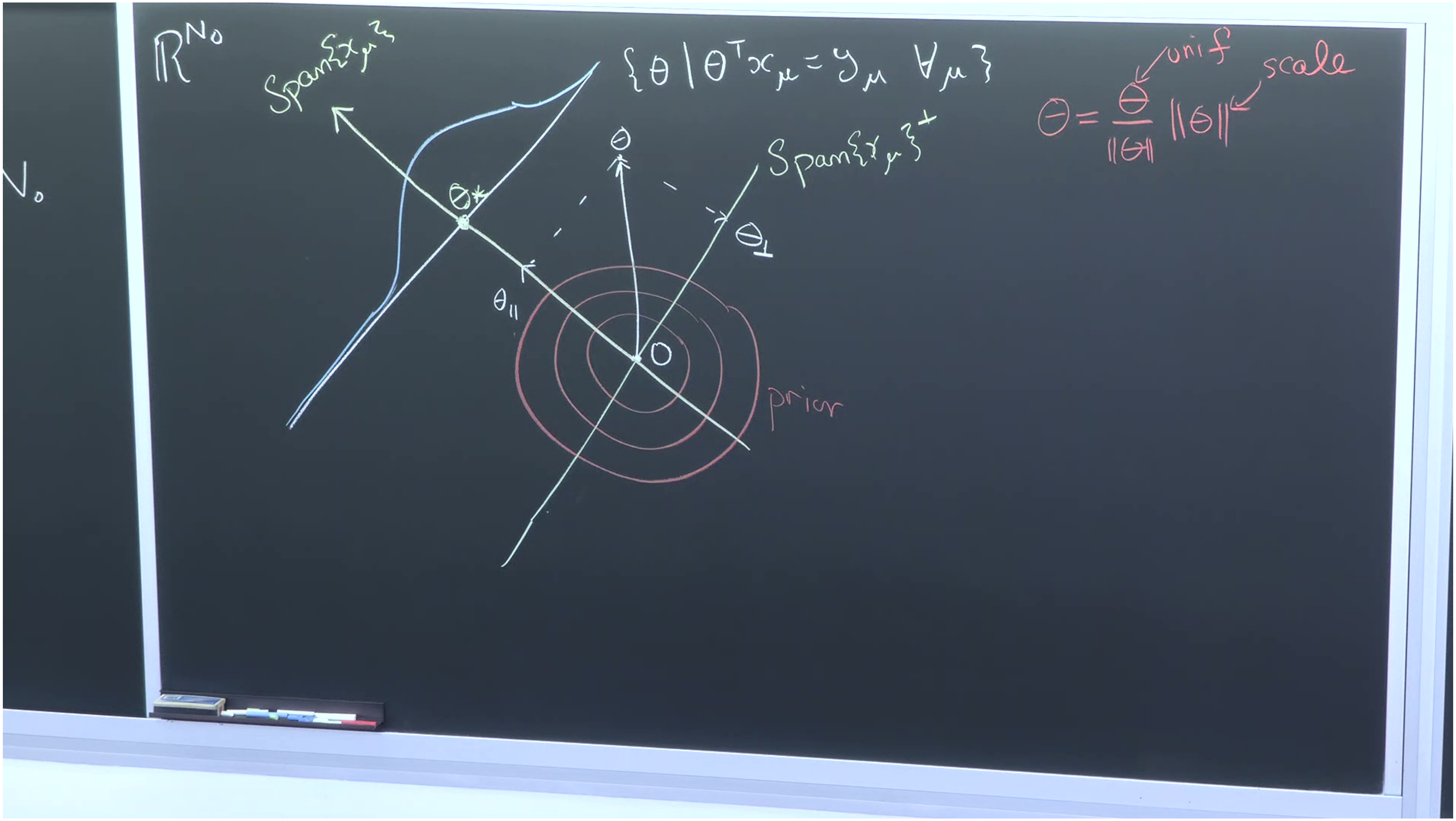
• $\sigma(t$

• $A =$

$\Upsilon = 0$: (

$f(x;\theta) =$

Prior:



$\mathbb{R}^{N_0}$
Span$\{x_\mu\}$
$\{\theta \mid \theta^T x_\mu = y_\mu \;\forall \mu\}$
unif
scale
$\theta = \frac{\theta}{\|\theta\|}\|\theta\|$
θ
Span$\{x_\mu\}^\perp$
θ^*
$\theta_\perp$
$\theta_\parallel$
0
prior

posterior: $\begin{cases} \mathbb{P}_{post}(\theta | D, N_e, L) = \lim_{\beta \to \infty} \dfrac{\mathbb{P}_{prior}(\theta | N_e, L) \exp\{-\beta \mathcal{L}(\theta | D)\}}{Z_\beta(D | N_e, L)} \\ \mathcal{L}(\theta | D) \equiv \dfrac{1}{2P} \sum_{\mu=1}^{P} \left(f(x_\mu; \theta) - y_\mu\right)^2 \end{cases}$

NB: $\mathbb{P}_{post}(\theta) \propto \mathbb{P}_{prior}(\theta)\, \delta_{\{\theta^T x_\mu = y_\mu \,\forall \mu\}}$

$\theta \sim \mathbb{P}_{post} \iff \theta = \theta_* + \underbrace{\dfrac{\theta_\perp}{\|\theta_\perp\|}}_{\text{unif}} \underbrace{\|\theta_\perp\|}_{\text{learned scale}}$

This Talk

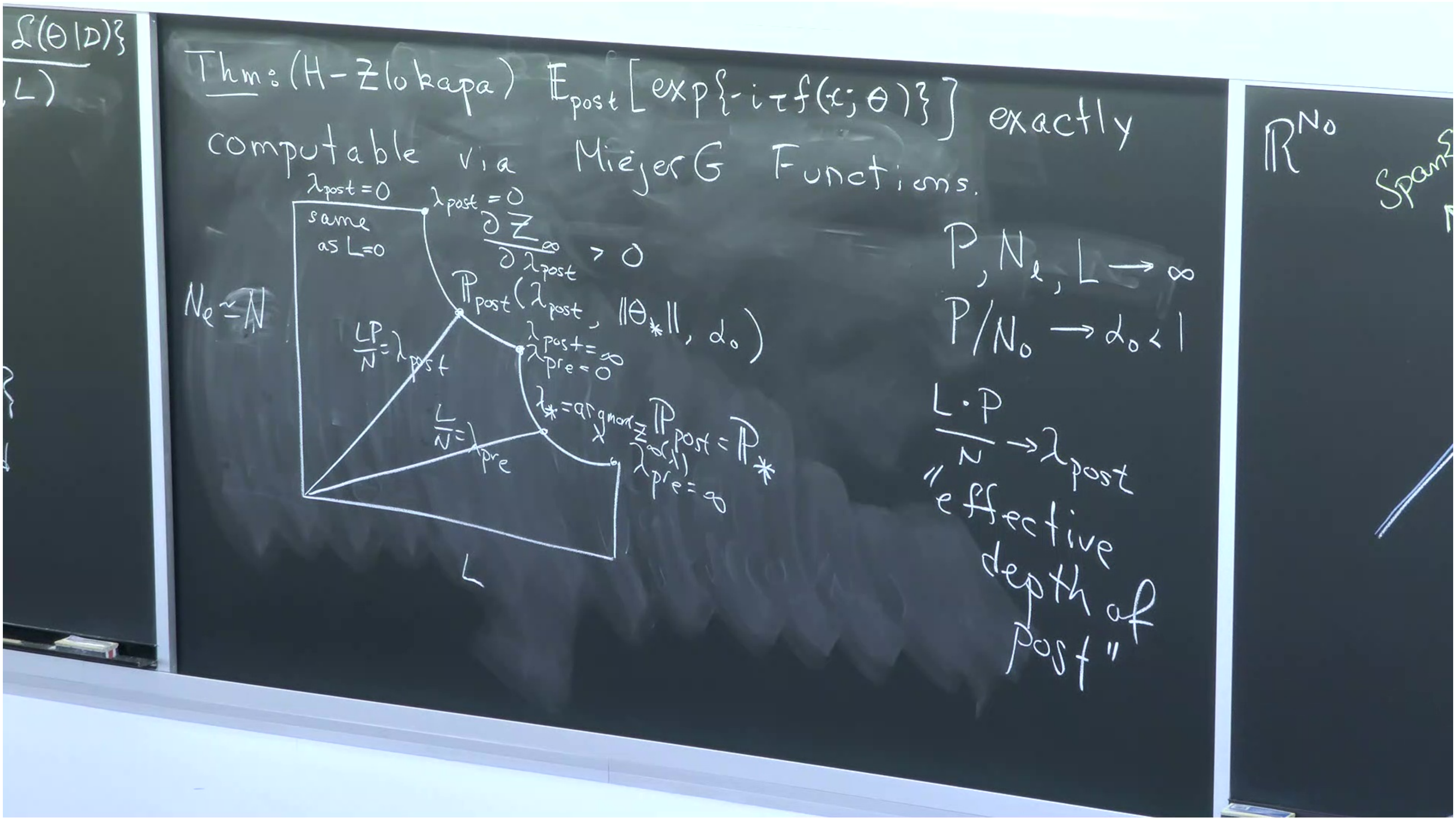
• $\sigma(t$

• $A =$

$\Psi = 0$: (

$f(x;\theta) =$

Prior:


Thm: (H-Zlokapa) $\mathbb{E}_{post}[\exp\{-i\tau f(x;\theta)\}]$ exactly computable via Miejer G Functions.
$\lambda_{post}=0$
same as $L=0$
$\lambda_{post}=0$
$\frac{\partial Z_\infty}{\partial \lambda_{post}} > 0$
$N_\ell \simeq N$
$\mathbb{P}_{post}(\lambda_{post}, \|\theta_*\|, \alpha_0)$
$\frac{LP}{N}=\lambda_{post}$
$\lambda_{post}=\infty$
$\lambda_{pre}=0$
$\frac{L}{N}=\lambda_{pre}$
$\lambda_* = \arg\max_\lambda \mathbb{P}_{post} = \mathbb{P}_*$
$\lambda_{pre}=\infty$
L
$P, N_\ell, L \to \infty$
$P/N_0 \to \alpha_0 < 1$
$\frac{L\cdot P}{N} \to \lambda_{post}$
"effective depth of post"
$\mathbb{R}^{N_0}$

$\psi \neq 0$: $\begin{cases} f(x;\theta) = W^{(L+1)}\sigma W^{(L)} \cdots \sigma W^{(1)} x \\ W^{(\ell)} \sim N_\ell \times N_{\ell-1}, \quad \sigma(t) = t + \frac{\psi}{L} t^3 \end{cases}$

① If $L, N_\ell, P \to \infty$ but $\frac{LP}{N} \longrightarrow 0$. Then model is equivalent to linear model with features

$$x \in \mathbb{R}^{N_0} \longrightarrow e^{\psi}\left(1-2\psi\|x\|^2\right)^{-1/2} x$$

$$\longmapsto \left(1 \pm \|x\|^2\right)^{-1/2} x = \hat{x}$$

$$f(x;\theta,\psi) = f\left(x;\theta,\tfrac{LP}{N}=0\right)$$

$$+\, C_\psi \frac{LP}{N} \sum_{\mu=1}^{P} y_\mu a_\mu \left(\hat{x}^T \hat{\Sigma} x - \hat{x}_\mu^T \hat{\Sigma} \hat{x}_\mu \right) + O\left(N^{-2}\right)$$

$$(\hat{x})_{\parallel} = \sum_{\mu=1}^{P} a_\mu \hat{x}_\mu \qquad \hat{\Sigma} \equiv \frac{1}{P} \sum_{\mu=1}^{P} \hat{x}_\mu \hat{x}_\mu^T$$

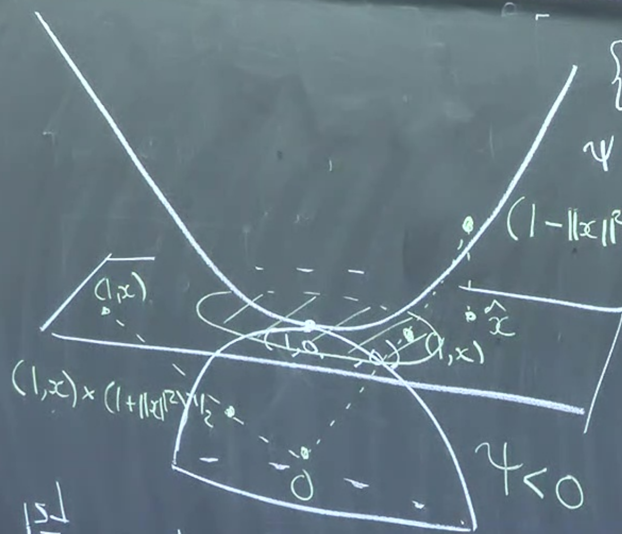
$$\vdots$$

model

tures

$\{(x_0, x) \mid x_0^2 = 1 + \|x\|^2\}$

$\psi > 0$



$\{1\} \times \mathbb{R}^{N_0}$ $\quad \psi = 0$

$\psi < 0$

② 1^{st} order in $\frac{1}{N}$ correction to kernel regime is cubic in $\hat{x}$

$f(x; \theta, \psi)$

$(\hat{x})_{\parallel} = \sum$