Bayesian Interpolation with Deep Linear Networks

Boris Hanin (Princeton ORFE) and Alexander Zlokapa (MIT Physics)

PNAS 2023 (arXiv:2212.14457)

Motivating Questions

Motivating Questions

Given **Model.** $z(x,\theta) = W^{(L+1)}\sigma W^{(L)} \cdots \sigma W^{(1)} x, \quad W^{(\ell)} \ - \ N_\ell \times N_{\ell-1}$

Motivating Questions

Given

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

Data. $\mathcal{D} \; = \; \{(x_i, y_i), \quad i = 1, \ldots, P\}$

Motivating Questions

Given

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

Data. $\mathcal{D} \; = \; \{(x_i, y_i), \quad i = 1, \ldots, P\}$

Alg. $\mathcal{A} : \quad (z, \mathcal{D}) \; \mapsto \; \theta_{\text{learned}}$

Motivating Questions

Given

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

Data. $\mathcal{D} \; = \; \{(x_i, y_i), \quad i = 1, \ldots, P\}$

Alg. $\mathcal{A} : \quad (z, \mathcal{D}) \; \mapsto \; \theta_{\text{learned}}$

Q1. How do $P,\, L,\, N_\ell$ jointly affect learning?

Motivating Questions

Given

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

Data. $\mathcal{D} \; = \; \{(x_i, y_i), \quad i = 1, \ldots, P\}$

Alg. $\mathcal{A}: \quad (z, \mathcal{D}) \; \mapsto \; \theta_{\text{learned}}$

Q1. How do P, L, N_ℓ jointly affect learning?

- $\underline{L=0}:$ $(x_i^T x_j)_{1\leq i,j\leq P}$ depends on P/N_0(Marchenko-Pastur)

Motivating Questions

Given

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

Data. $\mathcal{D} \;=\; \{(x_i, y_i), \quad i = 1, \ldots, P\}$

Alg. $\mathcal{A}: \quad (z, \mathcal{D}) \;\mapsto\; \theta_{\text{learned}}$

Q1. How do $P,\ L,\ N_\ell$ jointly affect learning?

- $\underline{L=0}$: $\quad (x_i^T x_j)_{1\leq i,j\leq P}$ depends on $\quad P/N_0$(Marchenko-Pastur)
- $\underline{P<\infty}$: Deviation from linear model depends on

$$\frac{1}{N_1} + \cdots + \frac{1}{N_L} \simeq \frac{L}{N}$$

Motivating Questions

Given

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

Data. $\mathcal{D} \; = \; \{(x_i, y_i), \quad i = 1, \ldots, P\}$

Alg. $\mathcal{A}: \quad (z, \mathcal{D}) \; \mapsto \; \theta_{\text{learned}}$

Q1. How do $P, \; L, \; N_\ell$ jointly affect learning?

Q2. How does one analyze learning in non-linear models?

Motivating Questions

Given

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

Data. $\mathcal{D} \; = \; \{(x_i, y_i), \quad i = 1, \ldots, P\}$

Alg. $\mathcal{A}: \quad (z, \mathcal{D}) \; \mapsto \; \theta_{\text{learned}}$

Q1. How do $P,\ L,\ N_\ell$ jointly affect learning?

Q2. How does one analyze learning in non-linear models?

Q3. How do models make predictions on test data?

This Work: Q1 - Q3 for Linear Networks

This Work: Q1 - Q3 for Linear Networks

Model. Set $N_{L+1} = 1$ and take identity non-linearity

$$z^{(L+1)}(x, \theta) = W^{(L+1)} W^{(L)} \cdots W^{(1)} x = \theta^T x$$

This Work: Q1 - Q3 for Linear Networks

Model. Set $N_{L+1} = 1$ and take identity non-linearity

$$z^{(L+1)}(x, \theta) = W^{(L+1)} W^{(L)} \cdots W^{(1)} x = \theta^T x$$

Data. $\mathcal{D} \;=\; \{(x_i, y_i), \quad i = 1, \ldots, P\}$

This Work: Q1 - Q3 for Linear Networks

Model. Set $N_{L+1} = 1$ and take identity non-linearity

$$z^{(L+1)}(x, \theta) = W^{(L+1)} W^{(L)} \cdots W^{(1)} x = \theta^T x$$

Data. $\mathcal{D} = \{(x_i, y_i), \quad i = 1, \ldots, P\}$

Alg. Bayesian inference at zero temperature:

This Work: Q1 - Q3 for Linear Networks

Model. Set $N_{L+1} = 1$ and take identity non-linearity

$$z^{(L+1)}(x, \theta) = W^{(L+1)} W^{(L)} \cdots W^{(1)} x = \theta^T x$$

Data. $\mathcal{D} \;=\; \{(x_i, y_i), \quad i = 1, \ldots, P\}$

Alg. Bayesian inference at zero temperature:

- $\theta \sim \mathbb{P}_{\text{prior}} \quad \Longleftrightarrow \quad W_{ij}^{(\ell)} \sim \mathcal{N}(0, \sigma^2 / N_{\ell-1})$

This Work: Q1 - Q3 for Linear Networks

Model. Set $N_{L+1} = 1$ and take identity non-linearity

$$z^{(L+1)}(x, \theta) = W^{(L+1)} W^{(L)} \cdots W^{(1)} x = \theta^T x$$

Data. $\mathcal{D} \ = \ \{(x_i, y_i), \quad i = 1, \ldots, P\}$

Alg. Bayesian inference at zero temperature:

- $\theta \sim \mathbb{P}_{\text{prior}} \quad \Longleftrightarrow \quad W_{ij}^{(\ell)} \sim \mathcal{N}(0, \sigma^2 / N_{\ell-1})$
- $-\log \mathbb{P}_\beta(\mathcal{D} \mid \theta) = \beta \sum_i (z(x_i; \theta) - y_i)^2$

This Work: Q1 - Q3 for Linear Networks

Model. Set $N_{L+1} = 1$ and take identity non-linearity

$$z^{(L+1)}(x, \theta) = W^{(L+1)} W^{(L)} \cdots W^{(1)} x = \theta^T x$$

Data. $\mathcal{D} \;=\; \{(x_i, y_i), \quad i = 1, \ldots, P\}$

Alg. Bayesian inference at zero temperature:

- $\theta \sim \mathbb{P}_{\mathrm{prior}} \quad \Longleftrightarrow \quad W_{ij}^{(\ell)} \sim \mathcal{N}(0, \sigma^2 / N_{\ell-1})$
- $-\log \mathbb{P}_\beta(\mathcal{D} \mid \theta) = \beta \sum_i (z(x_i; \theta) - y_i)^2$
- $\mathbb{P}_{\mathrm{post}}(\theta \mid \mathcal{D}) = \lim_{\beta \to \infty} Z_\beta^{-1}(\mathcal{D}) \mathbb{P}_{\mathrm{prior}}(\theta) \mathbb{P}_\beta(\mathcal{D} \mid \theta)$

Key Theorems, Informally

Key Theorems, Informally

T1. Bayesian inference is exactly solvable

Key Theorems, Informally

T1. Bayesian inference is exactly solvable

T2. Effective depth of prior: $$\frac{1}{N_1} + \cdots + \frac{1}{N_L} \simeq \frac{L}{N}$$

Effective depth of posterior: $$P\left(\frac{1}{N_1} + \cdots + \frac{1}{N_L}\right) \simeq \frac{PL}{N}$$

Key Theorems, Informally

T1. Bayesian inference is exactly solvable

T2. Effective depth of prior:

$$\frac{1}{N_1} + \cdots + \frac{1}{N_L} \simeq \frac{L}{N}$$

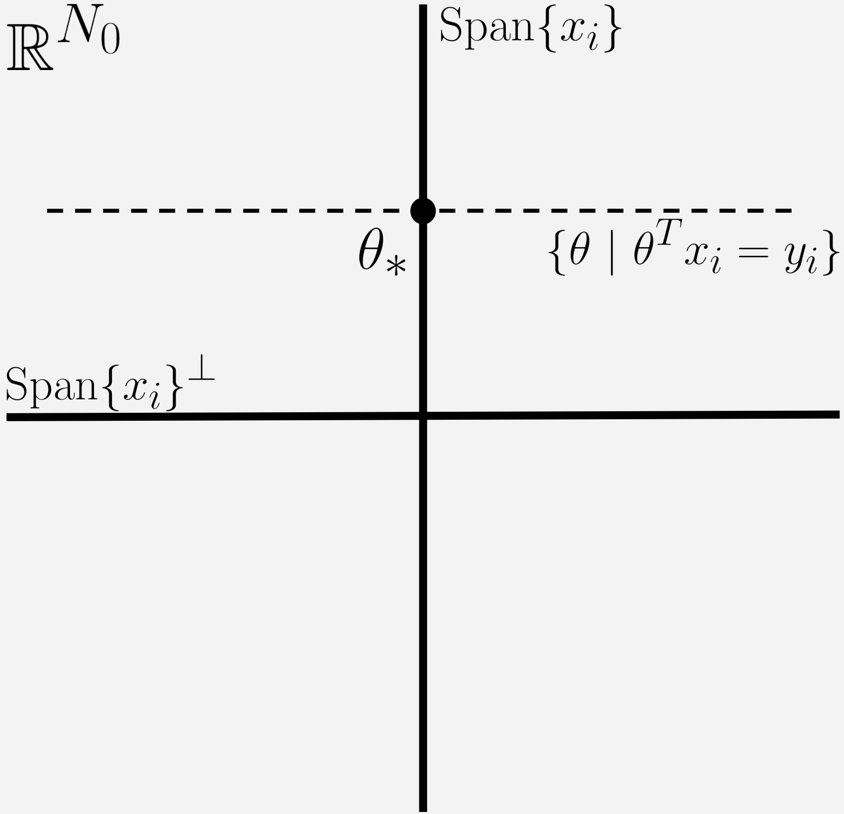
Effective depth of posterior:

$$P\left(\frac{1}{N_1} + \cdots + \frac{1}{N_L}\right) \simeq \frac{PL}{N}$$

T3. Deep networks with universal priors learn same posteriors as shallow networks with optimal data-dependent priors

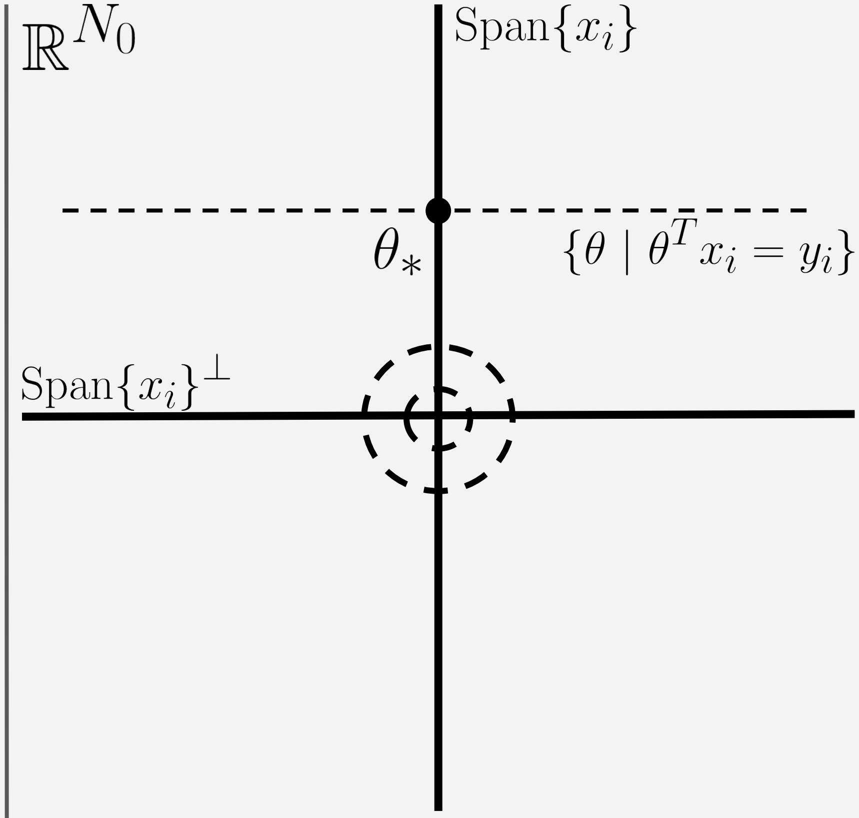
Structure of Prior/Posterior

Structure of Prior/Posterior



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

Structure of Prior/Posterior

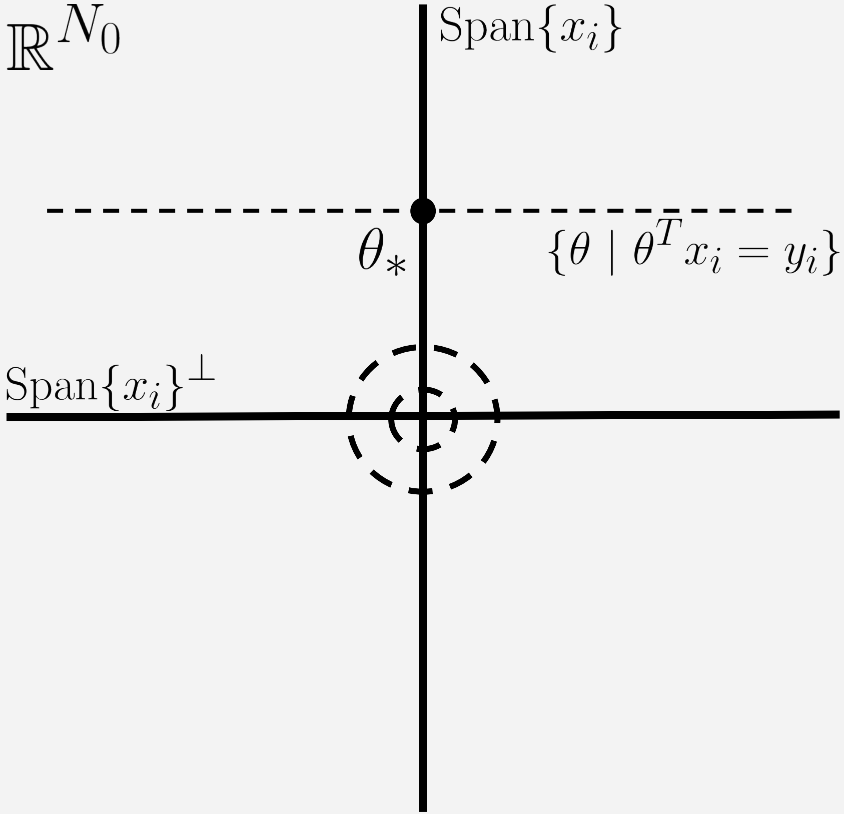


$$z(x;\theta) = W^{(L+1)} \cdots W^{(1)} x = \theta^T x$$

Structure of Prior/Posterior

1 $\mathbb{P}_{\text{post}}(\theta \mid L, N, \mathcal{D}, \sigma^2)$

$$= \lim_{\beta\to\infty} \frac{\mathbb{P}_{\text{prior}}(\theta \mid L, N, \sigma^2) e^{-\beta \mathcal{L}_{\mathcal{D}}(\theta)}}{Z_\beta\left(\mathcal{D} \mid L, N, \sigma^2\right)}$$



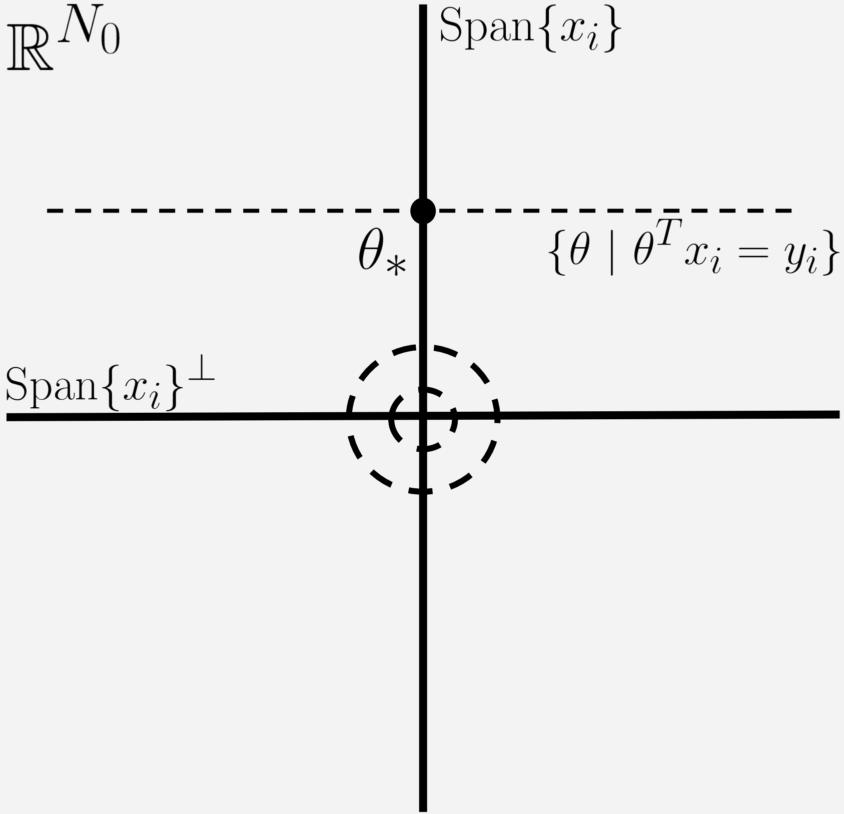
$$z(x;\theta) = W^{(L+1)} \cdots W^{(1)} x = \theta^T x$$

Structure of Prior/Posterior

1 $\mathbb{P}_{\text{post}}(\theta \mid L, N, \mathcal{D}, \sigma^2)$

$$= \lim_{\beta \to \infty} \frac{\mathbb{P}_{\text{prior}}(\theta \mid L, N, \sigma^2) e^{-\beta \mathcal{L}_{\mathcal{D}}(\theta)}}{Z_\beta\left(\mathcal{D} \mid L, N, \sigma^2\right)}$$

$$\propto \mathbb{P}_{\text{prior}}(\theta \mid \theta^T x_i = y_i)$$



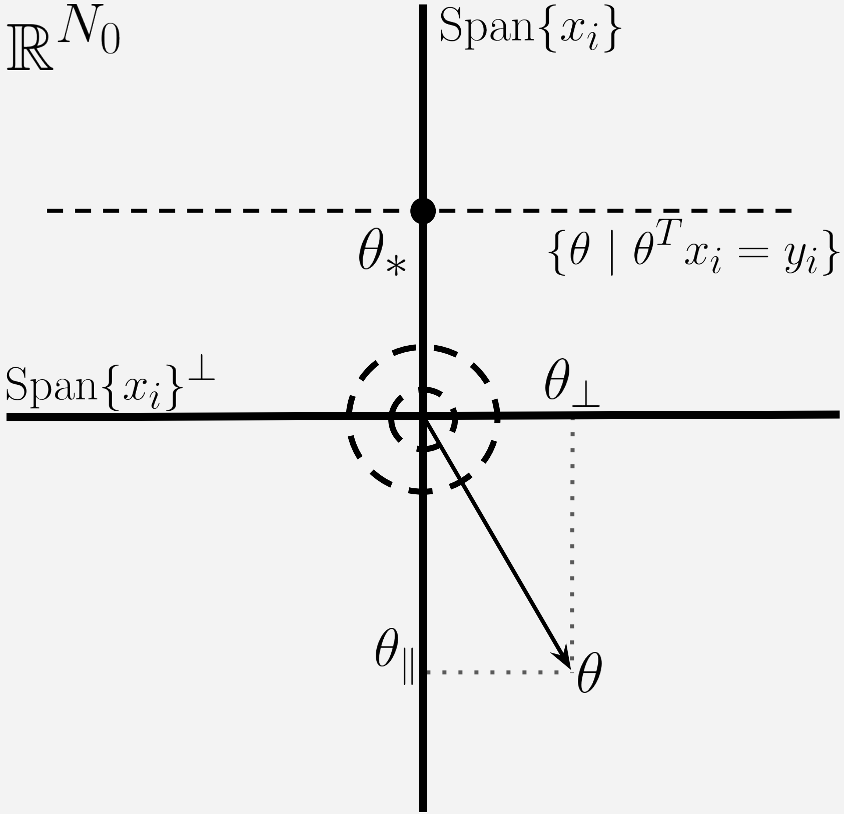
$$z(x;\theta) = W^{(L+1)} \cdots W^{(1)} x = \theta^T x$$

Structure of Prior/Posterior

1 $\mathbb{P}_{\text{post}}(\theta \mid L, N, \mathcal{D}, \sigma^2)$

$$= \lim_{\beta \to \infty} \frac{\mathbb{P}_{\text{prior}}(\theta \mid L, N, \sigma^2) e^{-\beta \mathcal{L}_{\mathcal{D}}(\theta)}}{Z_\beta(\mathcal{D} \mid L, N, \sigma^2)}$$

$$\propto \mathbb{P}_{\text{prior}}(\theta \mid \theta^T x_i = y_i)$$



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

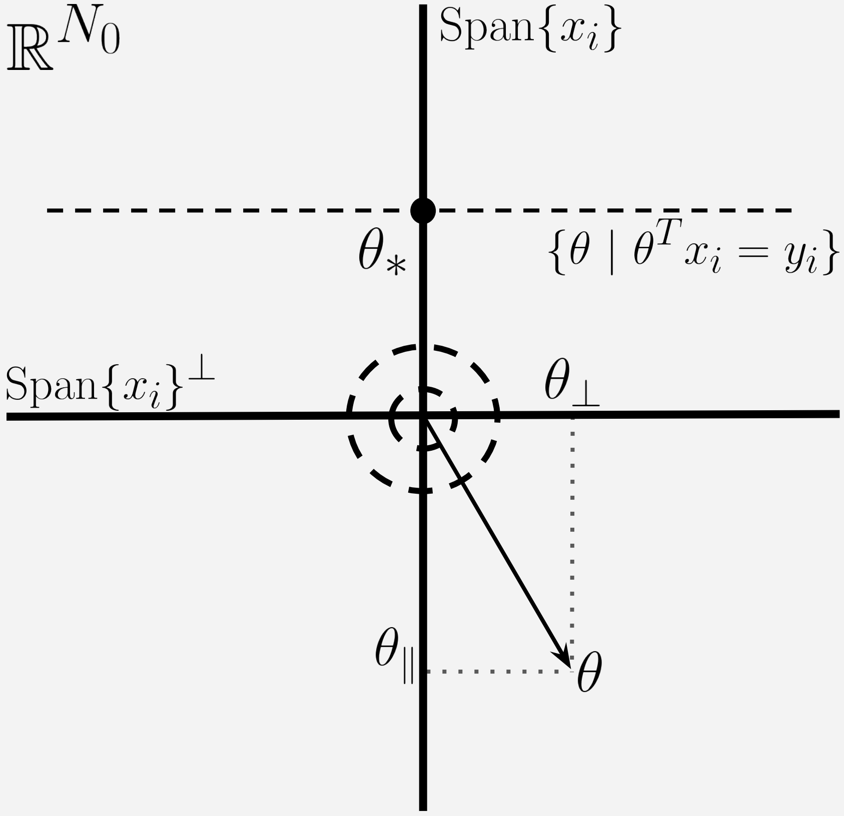
Structure of Prior/Posterior

1

$$\mathbb{P}_{\text{post}}(\theta \mid L, N, \mathcal{D}, \sigma^2) = \lim_{\beta\to\infty} \frac{\mathbb{P}_{\text{prior}}(\theta \mid L, N, \sigma^2) e^{-\beta \mathcal{L}_{\mathcal{D}}(\theta)}}{Z_\beta(\mathcal{D} \mid L, N, \sigma^2)} \propto \mathbb{P}_{\text{prior}}(\theta \mid \theta^T x_i = y_i)$$

2

$$\Rightarrow \theta = \theta_* + ||\theta_\perp|| \cdot \text{unif}$$



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

Structure of Prior/Posterior

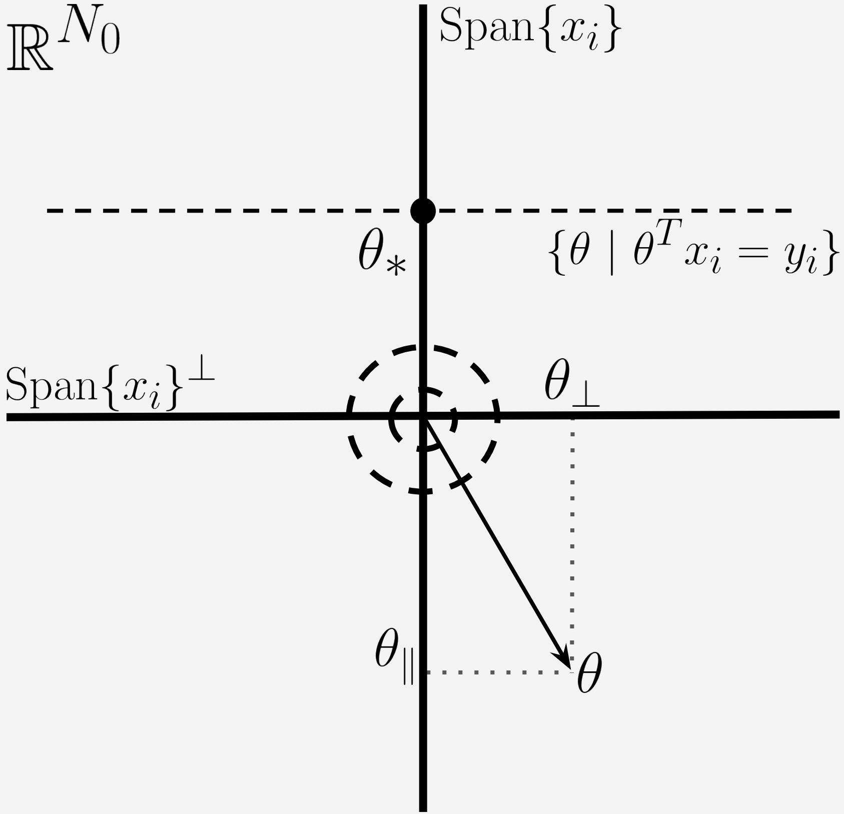
1 $\mathbb{P}_{\text{post}}(\theta \,|\, L, N, \mathcal{D}, \sigma^2)$

$$= \lim_{\beta\to\infty} \frac{\mathbb{P}_{\text{prior}}(\theta \,|\, L, N, \sigma^2) e^{-\beta \mathcal{L}_{\mathcal{D}}(\theta)}}{Z_\beta\left(\mathcal{D} \,|\, L, N, \sigma^2\right)}$$

$$\propto \mathbb{P}_{\text{prior}}(\theta \,|\, \theta^T x_i = y_i)$$

2 $\Rightarrow \theta = \theta_* + ||\theta_\perp|| \cdot \text{unif}$

3 $||\theta_\perp||$ is a learnable data-dependent scale for predictions in new directions



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

Evidence and Posterior Via G-Functions

Thm. Write $Z(t) = \mathbb{E}_{\mathrm{post}}\left[\exp\left\{-it\cdot\theta\right\}\right].$ Then,

$$Z(t) = \left(\frac{4\pi}{||\theta_*||^2}\right)^{\frac{P}{2}} \times e^{-i\langle\theta_*,\, t_{||}\rangle} \times \prod_{\ell=1}^{L}\Gamma\left(\frac{N_\ell}{2}\right)^{-1}$$

$$\times \sum_{k=0}^{\infty}\frac{(-||t^{\perp}||^2 M)^k}{k!}\, G_{L+1}\left(\frac{||\theta_*||^2}{4M}\,\middle|\,\frac{P}{2}, \frac{N_1}{2}+k, \ldots, \frac{N_L}{2}+k\right)$$

where $M = \prod_{\ell=0}^{L} 2\sigma^2/N_\ell$ and

$$G_\ell(z\,|\,b_1,\ldots,b_\ell) = \frac{1}{2\pi i}\int_{-i\infty}^{i\infty} ds\; z^s \prod_{j=1}^{\ell}\Gamma\left(1+b_j-s\right)$$

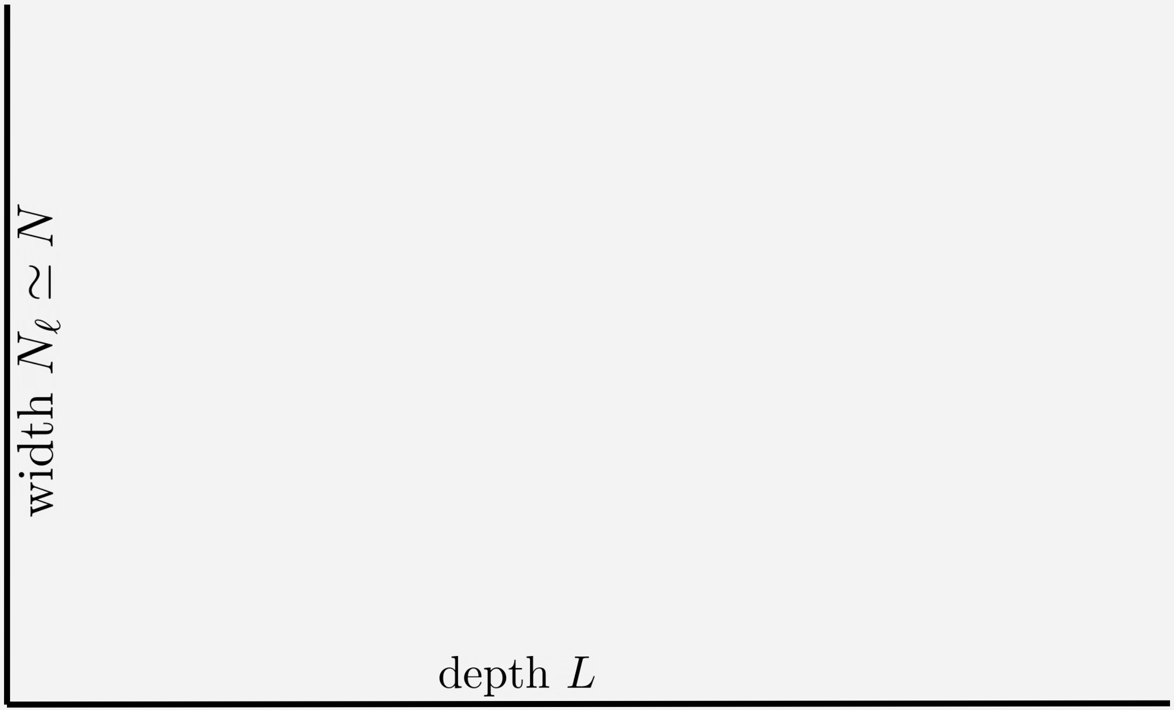
Phase Diagram for Deep Linear Posteriors

Phase Diagram for Deep Linear Posteriors

Setup

$$\begin{cases} \textbf{model}\text{: } W^{(L+1)}\cdots W^{(1)}x \\ \textbf{prior}\text{: } W_{ij}^{(\ell)} \sim \mathcal{N}(0, 1/N_{\ell-1}) \\ \textbf{likelihood}\text{: } \text{MSE on } \mathcal{D} = \{(x_i, y_i)\}_{i=1}^{P} \\ \textbf{regime}\text{: } P, N_\ell \to \infty,\ P/N_0 \to \alpha_0 < 1 \\ \textbf{prior depth}\text{: } L/N = \gamma \\ \textbf{post. depth}\text{: } PL/N = \lambda \end{cases}$$

Phase Diagram for Deep Linear Posteriors

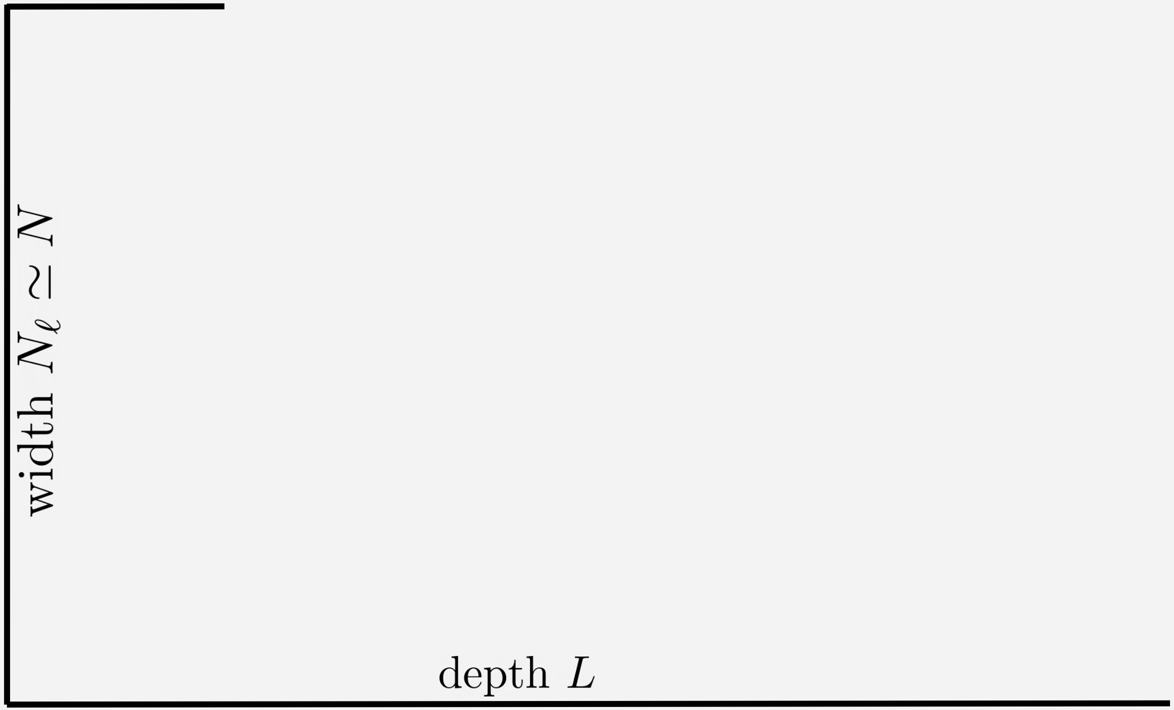


Setup

$$\begin{cases} \textbf{model}\text{: } W^{(L+1)} \cdots W^{(1)} x \\ \textbf{prior}\text{: } W_{ij}^{(\ell)} \sim \mathcal{N}(0, 1/N_{\ell-1}) \\ \textbf{likelihood}\text{: MSE on } \mathcal{D} = \{(x_i, y_i)\}_{i=1}^{P} \\ \textbf{regime}\text{: } P, N_\ell \to \infty,\ P/N_0 \to \alpha_0 < 1 \\ \textbf{prior depth}\text{: } L/N = \gamma \\ \textbf{post. depth}\text{: } PL/N = \lambda \end{cases}$$

Phase Diagram for Deep Linear Posteriors

$\underline{\lambda, \gamma = 0}$



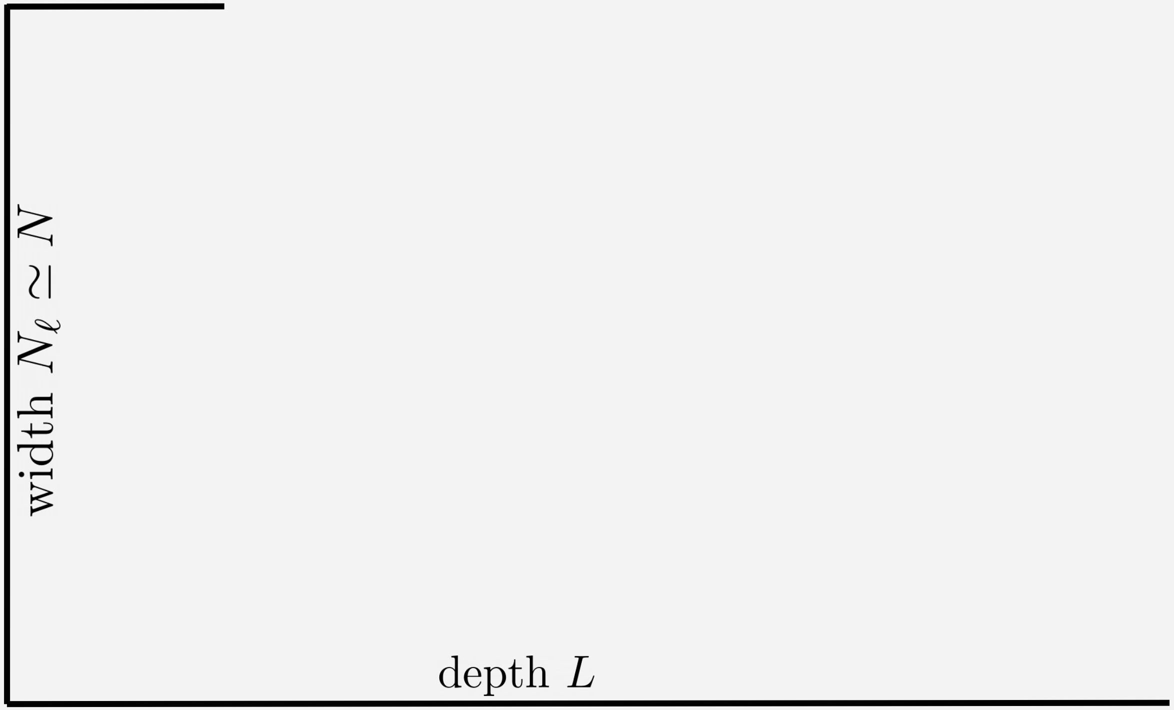
Setup

$$\begin{cases} \textbf{model}\text{: } W^{(L+1)} \cdots W^{(1)} x \\ \textbf{prior}\text{: } W_{ij}^{(\ell)} \sim \mathcal{N}(0, 1/N_{\ell-1}) \\ \textbf{likelihood}\text{: MSE on } \mathcal{D} = \{(x_i, y_i)\}_{i=1}^{P} \\ \textbf{regime}\text{: } P, N_\ell \to \infty,\ P/N_0 \to \alpha_0 < 1 \\ \textbf{prior depth}\text{: } L/N = \gamma \\ \textbf{post. depth}\text{: } PL/N = \lambda \end{cases}$$

Phase Diagram for Deep Linear Posteriors

$\lambda, \gamma = 0$

- $L = 0$



Setup

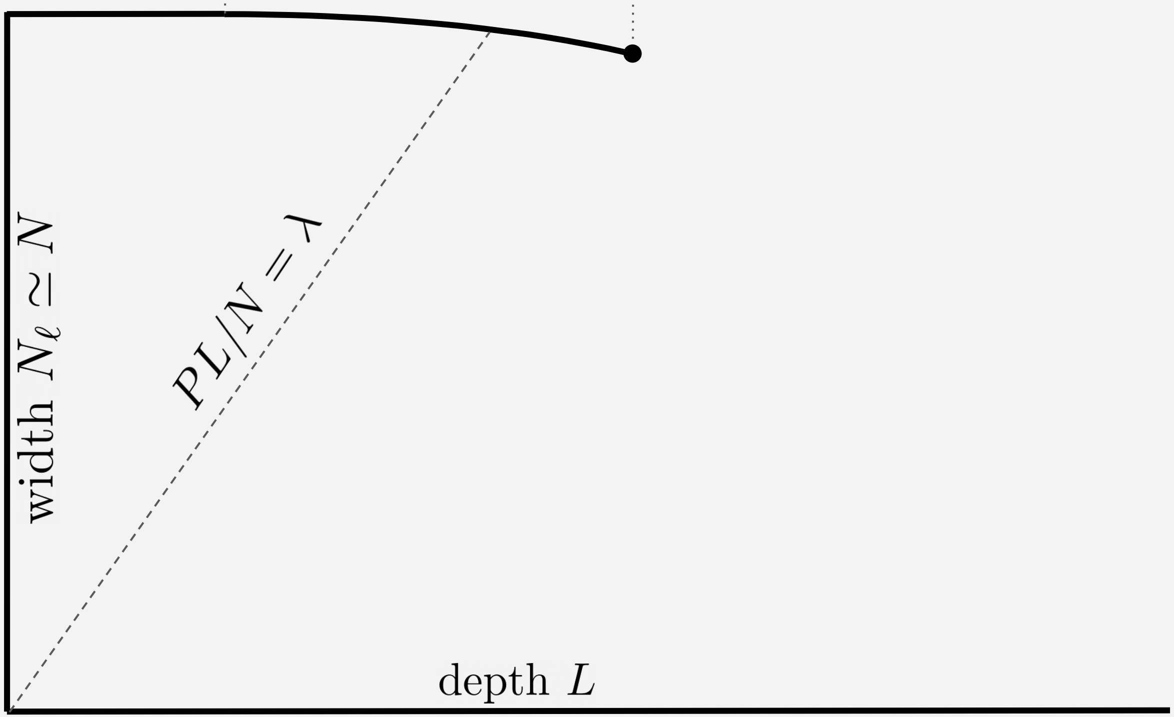
- **model**: $W^{(L+1)} \cdots W^{(1)} x$
- **prior**: $W_{ij}^{(\ell)} \sim \mathcal{N}(0, 1/N_{\ell-1})$
- **likelihood**: MSE on $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{P}$
- **regime**: $P, N_\ell \to \infty$, $P/N_0 \to \alpha_0 < 1$
- **prior depth**: $L/N = \gamma$
- **post. depth**: $PL/N = \lambda$

Phase Diagram for Deep Linear Posteriors

$\underline{\lambda, \gamma = 0}$

- $L = 0$

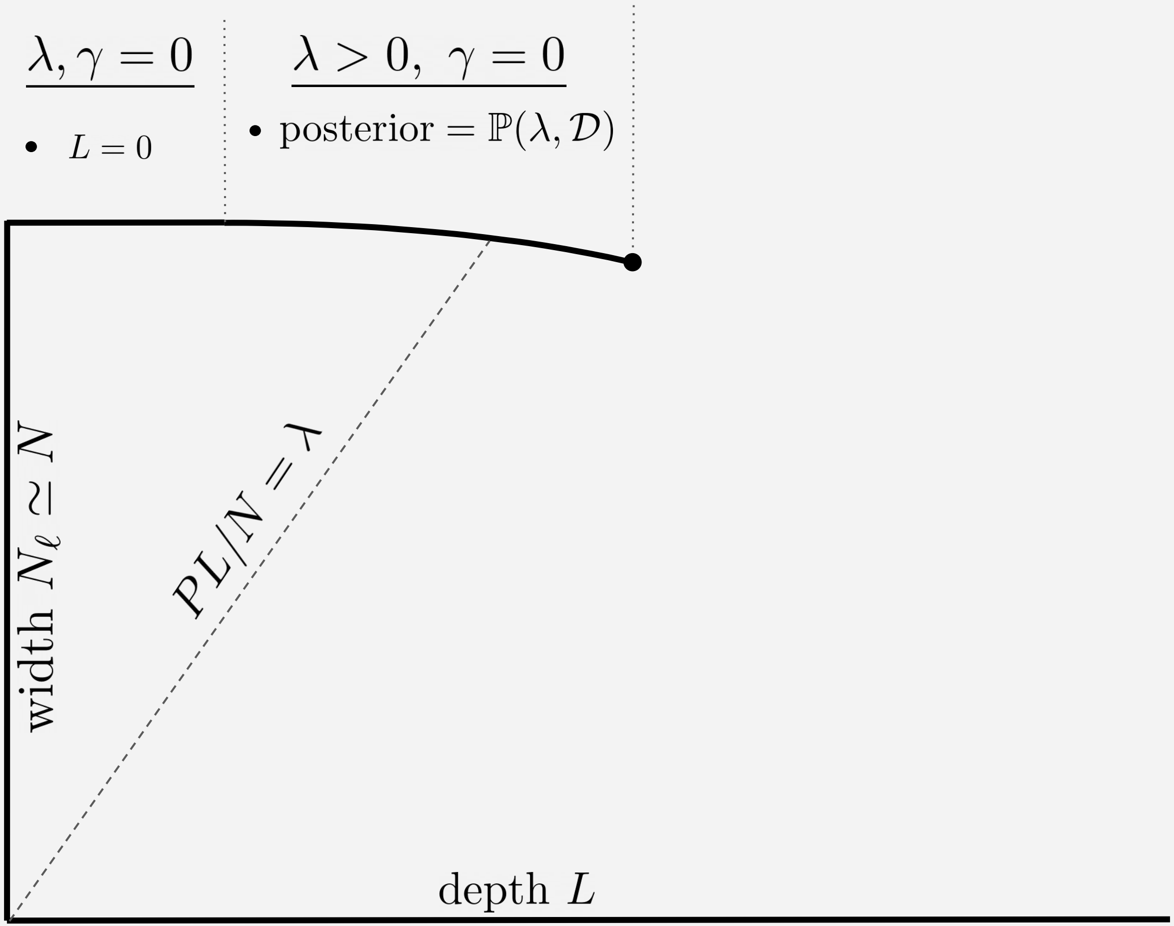
$\underline{\lambda > 0,\ \gamma = 0}$



Setup

- **model**: $W^{(L+1)} \cdots W^{(1)} x$
- **prior**: $W_{ij}^{(\ell)} \sim \mathcal{N}(0, 1/N_{\ell-1})$
- **likelihood**: MSE on $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{P}$
- **regime**: $P, N_\ell \to \infty$, $P/N_0 \to \alpha_0 < 1$
- **prior depth**: $L/N = \gamma$
- **post. depth**: $PL/N = \lambda$

Phase Diagram for Deep Linear Posteriors



Setup

- **model**: $W^{(L+1)} \cdots W^{(1)} x$
- **prior**: $W_{ij}^{(\ell)} \sim \mathcal{N}(0, 1/N_{\ell-1})$
- **likelihood**: MSE on $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{P}$
- **regime**: $P, N_\ell \to \infty,\ P/N_0 \to \alpha_0 < 1$
- **prior depth**: $L/N = \gamma$
- **post. depth**: $PL/N = \lambda$

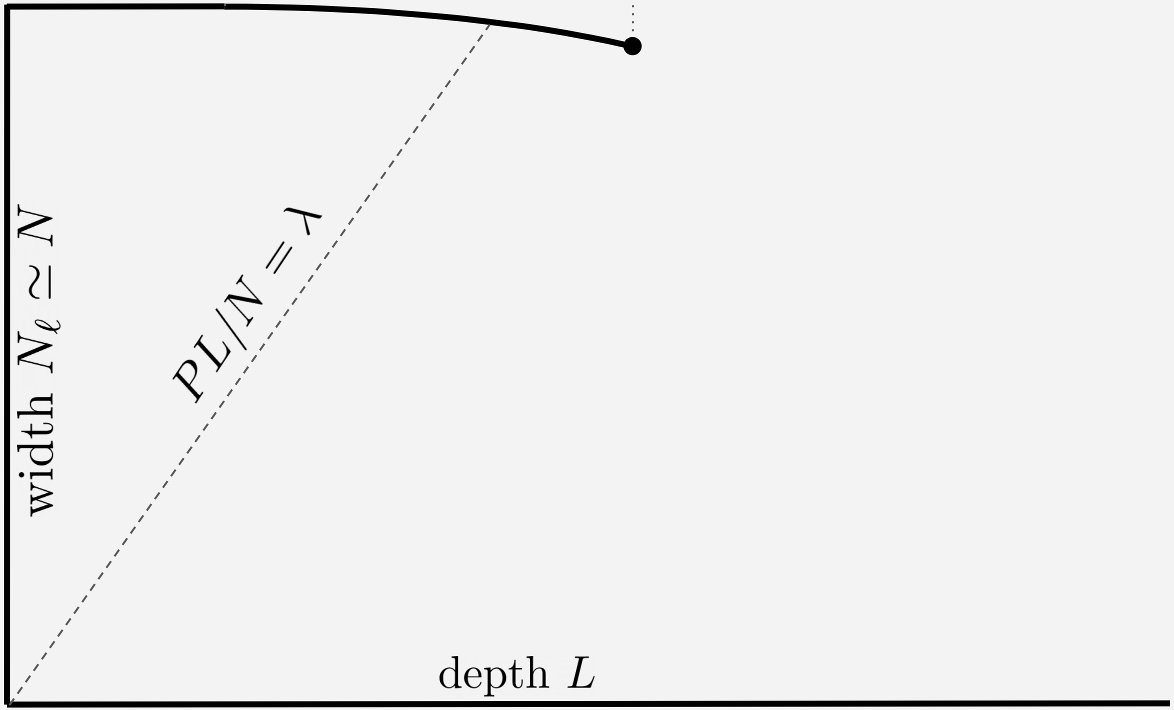
Phase Diagram for Deep Linear Posteriors

$\underline{\lambda, \gamma = 0}$

- $L = 0$

$\underline{\lambda > 0,\ \gamma = 0}$

- posterior $= \mathbb{P}(\lambda, \mathcal{D})$
- $\partial\,\text{evidence}/\partial\lambda > 0$



Setup

- **model**: $W^{(L+1)} \cdots W^{(1)} x$
- **prior**: $W^{(\ell)}_{ij} \sim \mathcal{N}(0, 1/N_{\ell-1})$
- **likelihood**: MSE on $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{P}$
- **regime**: $P, N_\ell \to \infty,\ P/N_0 \to \alpha_0 < 1$
- **prior depth**: $L/N = \gamma$
- **post. depth**: $PL/N = \lambda$

Phase Diagram for Deep Linear Posteriors

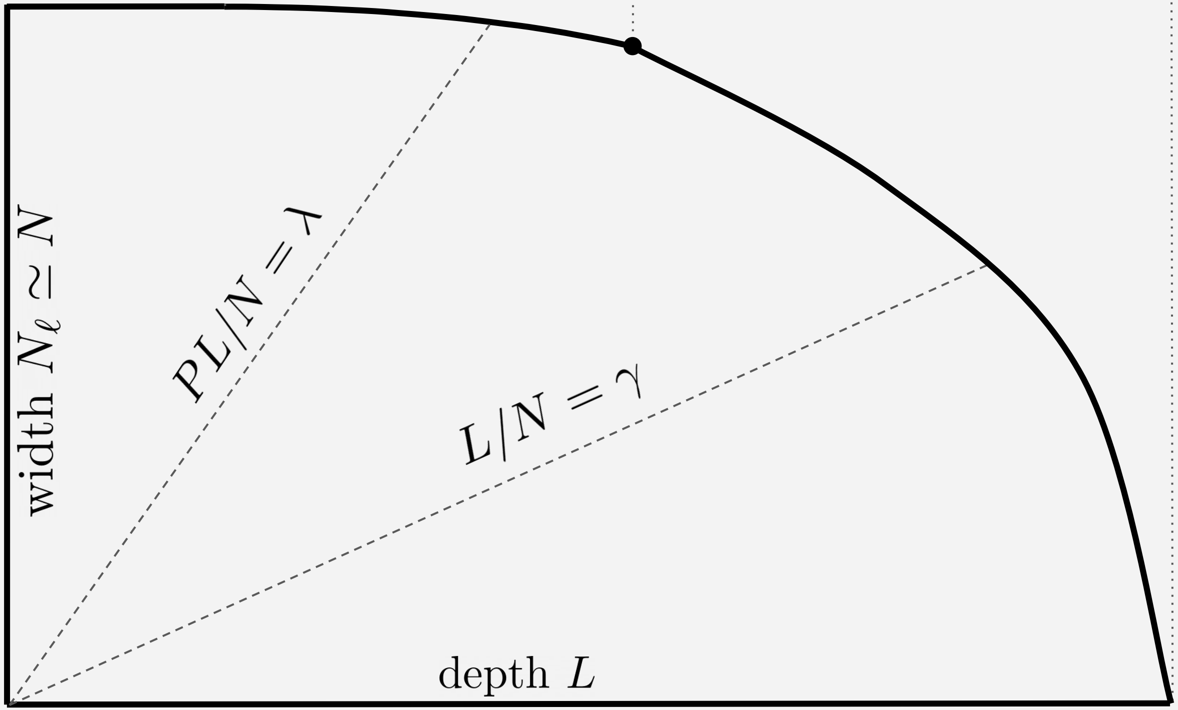
$\lambda, \gamma = 0$

- $L = 0$

$\lambda > 0,\ \gamma = 0$

- posterior $= \mathbb{P}(\lambda, \mathcal{D})$
- $\partial\,\text{evidence}/\partial\lambda > 0$

$\lambda = \infty, \gamma > 0$



Setup

- **model**: $W^{(L+1)} \cdots W^{(1)} x$
- **prior**: $W_{ij}^{(\ell)} \sim \mathcal{N}(0, 1/N_{\ell-1})$
- **likelihood**: MSE on $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{P}$
- **regime**: $P, N_\ell \to \infty,\ P/N_0 \to \alpha_0 < 1$
- **prior depth**: $L/N = \gamma$
- **post. depth**: $PL/N = \lambda$

Phase Diagram for Deep Linear Posteriors

$\lambda, \gamma = 0$

- $L = 0$

$\lambda > 0, \ \gamma = 0$

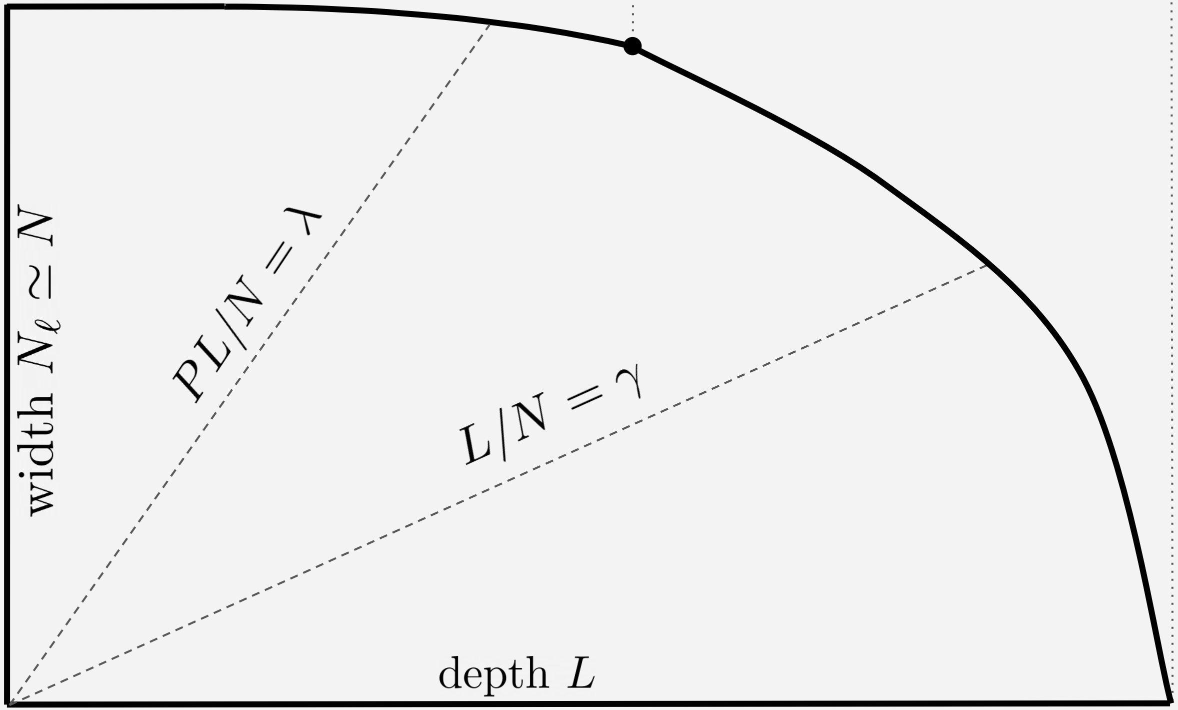
- posterior $= \mathbb{P}(\lambda, \mathcal{D})$
- $\partial\,$evidence$/\partial\lambda > 0$

$\lambda = \infty, \gamma > 0$

- universal, max evidence posterior $\mathbb{P}^*(\mathcal{D})$

Setup

$$\begin{cases} \textbf{model}\text{: } W^{(L+1)} \cdots W^{(1)} x \\ \textbf{prior}\text{: } W_{ij}^{(\ell)} \sim \mathcal{N}(0, 1/N_{\ell-1}) \\ \textbf{likelihood}\text{: MSE on } \mathcal{D} = \{(x_i, y_i)\}_{i=1}^{P} \\ \textbf{regime}\text{: } P, N_\ell \to \infty,\ P/N_0 \to \alpha_0 < 1 \\ \textbf{prior depth}\text{: } L/N = \gamma \\ \textbf{post. depth}\text{: } PL/N = \lambda \end{cases}$$



Phase Diagram for Deep Linear Posteriors

$\lambda, \gamma = 0$

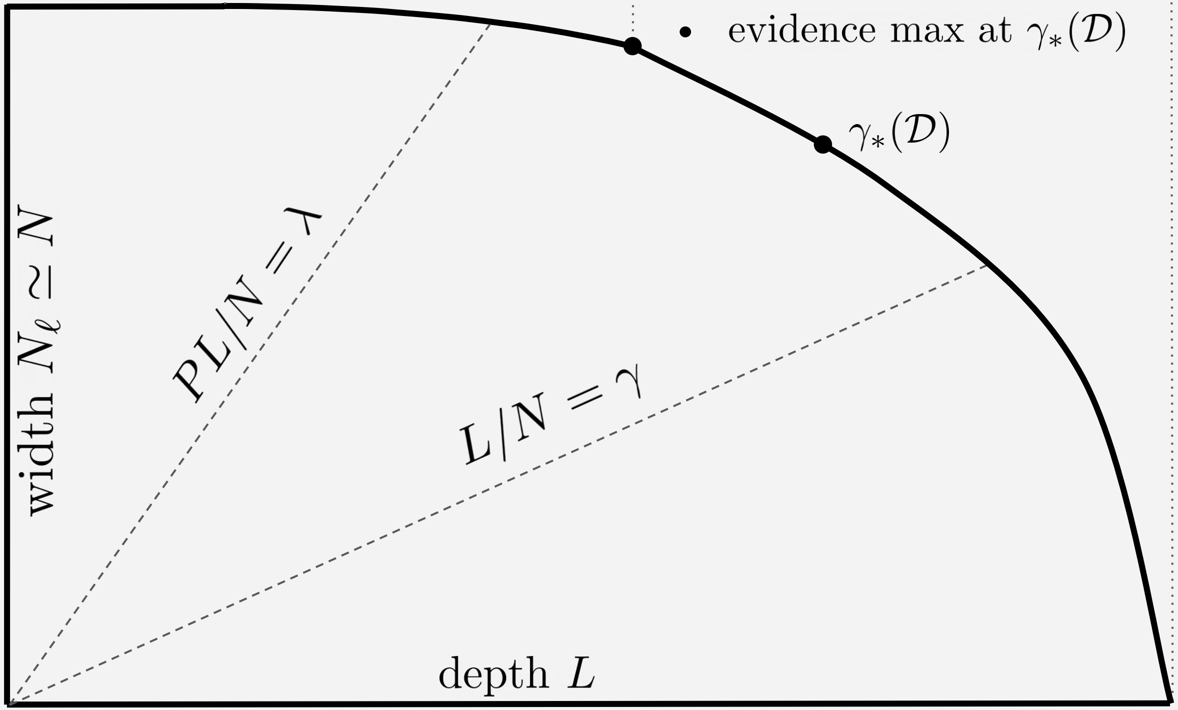
- $L = 0$

$\lambda > 0,\ \gamma = 0$

- posterior $= \mathbb{P}(\lambda, \mathcal{D})$
- $\partial\,$evidence$/\partial\lambda > 0$

$\lambda = \infty, \gamma > 0$

- universal, max evidence posterior $\mathbb{P}^*(\mathcal{D})$
- evidence max at $\gamma_*(\mathcal{D})$



Setup

$$\begin{cases} \textbf{model}: W^{(L+1)} \cdots W^{(1)} x \\ \textbf{prior}: W^{(\ell)}_{ij} \sim \mathcal{N}(0, 1/N_{\ell-1}) \\ \textbf{likelihood}: \text{MSE on } \mathcal{D} = \{(x_i, y_i)\}_{i=1}^P \\ \textbf{regime}: P, N_\ell \to \infty,\ P/N_0 \to \alpha_0 < 1 \\ \textbf{prior depth}: L/N = \gamma \\ \textbf{post. depth}: PL/N = \lambda \end{cases}$$

Deep Linear Networks Learn Optimal Features

Thm. Define $\sigma_* := \mathrm{argmax}_\sigma \lim_{\substack{P, N_0 \to \infty \\ P/N_0 \to \alpha_0}} Z_\infty(\mathcal{D} \mid L, N_\ell, \sigma)$

Then,

$$\lim_{\substack{P, N_0 \to \infty \\ P/N_0 \to \alpha_0}} \mathbb{P}_{\text{post}}(||\theta_\perp|| \mid L, N_\ell, \sigma = \sigma_*) = \delta_{\kappa^*}$$

for $\kappa^* = \kappa^*(\alpha_0, ||\theta_*||)$. Moreover,

$$\lim_{\substack{P, N_0 \to \infty \\ P/N_0 \to \alpha_0 \\ PL/N \to \lambda}} \mathbb{P}_{\text{post}}(||\theta_\perp|| \mid L, N_\ell, \sigma = 1) = \delta_{\kappa(\lambda)}$$

where

$$\kappa(\lambda) \to \begin{cases} \kappa^*, & \lambda \to \infty \\ L = 0 \text{ post}, & \lambda \to 0 \end{cases}.$$