Contrasting random and learned features in deep Bayesian linear regression

Jacob Zavatone-Veth

HARVARD
Department of Physics

jzv.io

10 August 2023

Statistical Physics and Machine Learning Back Together Again, Cargèse

How does representational flexibility affect generalization in overparameterized models?

We’ll study this in a tractable set of toy models: deep linear nets

This talk is based on:

- Z-V, Tong, & Pehlevan, *PRE,* 2022
- Z-V & Pehlevan, *SciPost Physics Core,* 2023
- Z-V & Pehlevan, arXiv:2303.00564

William Tong

Cengiz Pehlevan

Deep linear models

Parameterize a simple linear regressor by a matrix product

$$g(\mathbf{x};\Theta) = (\mathbf{U}_1 \cdots \mathbf{U}_L \mathbf{v})^\top \mathbf{x}$$

where

$$\mathbf{U}_\ell \in \mathbb{R}^{n_{\ell-1} \times n_\ell}$$

For simplicity, consider the task of fitting data $\mathcal{D} = \{(\mathbf{x}_\mu, y_\mu)\}_{\mu=1}^{p}$ generated by a noisy teacher

$$\mathbf{x}_\mu \sim_{iid} \mathcal{N}(\mathbf{0}, \mathbf{I}_{n_o})$$
$$y_\mu = w_*^\top x + \xi_\mu$$
$$\xi_\mu \sim \mathcal{N}(0, \eta^2)$$

Fix a Gaussian prior

$$(U_\ell)_{ij} \sim_{iid} \mathcal{N}(0, \sigma_\ell^2 / n_{\ell-1})$$
$$(v)_i \sim_{iid} \mathcal{N}(0, \sigma_{L+1}^2 / n_L)$$

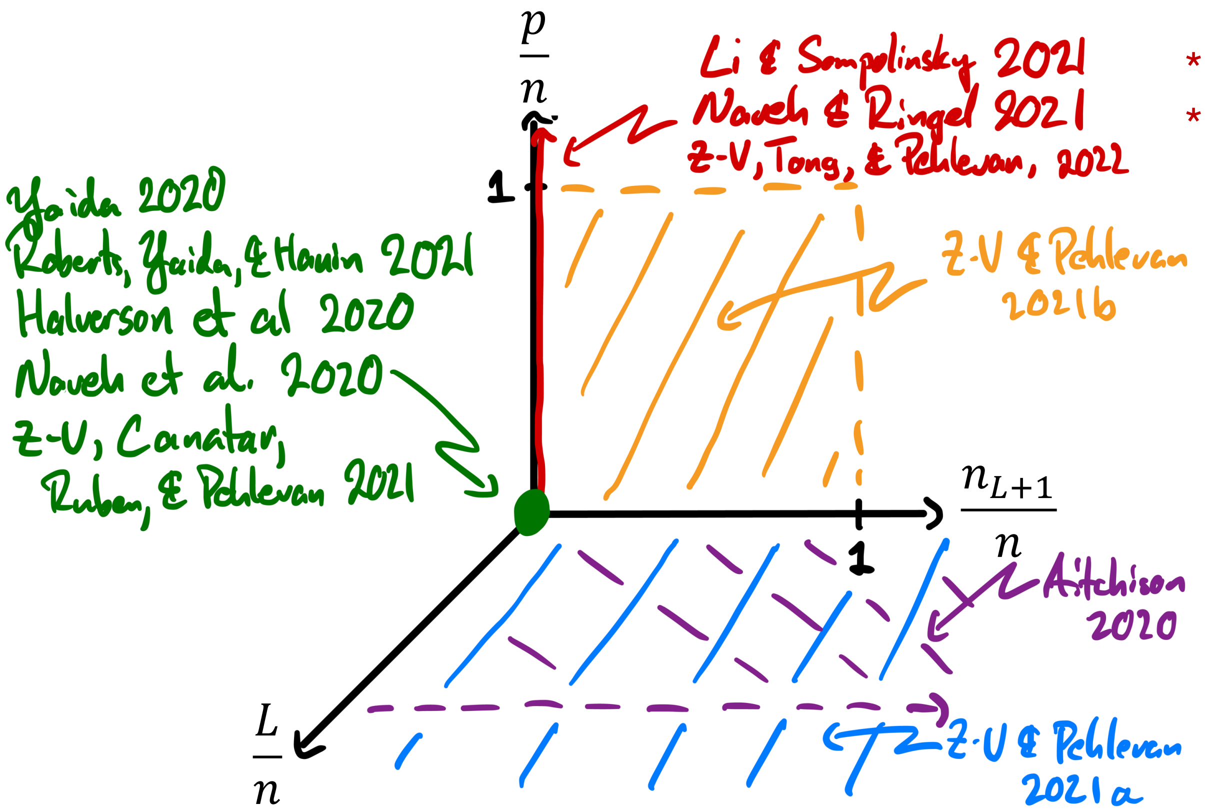
and likelihood

$$p(\mathcal{D}|\Theta) \propto \exp\left[-\frac{\beta}{2}\sum_{\mu=1}^{p}[g(x_\mu, \theta) - y_\mu]^2\right]$$

Scaling limits for deep Bayesian neural networks

At width $n \gg 1$:

Hanin & Zlokapa 2023



Li & Sompolinsky 2021
Naveh & Ringel 2021
Z-V, Tong, & Pehlevan, 2022

Z-V & Pehlevan 2021b

Yaida 2020
Roberts, Yaida, & Hanin 2021
Halverson et al 2020
Naveh et al. 2020
Z-V, Canatar,
Ruben, & Pehlevan 2021

Aitchison 2020

Z-V & Pehlevan 2021a

*Cui et al. 2023
*Seroussi et al. 2023

How does generalization performance change depending on whether we train the hidden weights $\mathbf{U}_\ell$ or fix them?

Estimators: Gibbs, MMSE, ridgeless

Writing $\langle\cdot\rangle_\beta$ for expectation with respect to the resulting posterior, we can consider various estimators:

- MMSE:

$$\epsilon_{MMSE} = \frac{1}{2}\left\|\langle \mathbf{U}_1 \cdots \mathbf{U}_L \mathbf{v}\rangle_\beta - \mathbf{w}_*\right\|^2$$

- Gibbs:

$$\epsilon_{Gibbs} = \frac{1}{2}\langle\|\mathbf{U}_1 \cdots \mathbf{U}_L \mathbf{v} - \mathbf{w}_*\|^2\rangle_\beta = \epsilon_{MMSE} + \frac{1}{2}\left\langle\left\|\mathbf{U}_1 \cdots \mathbf{U}_L \mathbf{v} - \langle \mathbf{U}_1 \cdots \mathbf{U}_L \mathbf{v}\rangle_\beta\right\|^2\right\rangle_\beta$$

Because the MMSE for deep linear networks is mostly trivial (Boris’ talk), we will consider the latter, focusing on the zero—temperature limit $\beta \to \infty$.

For the random feature model, the zero-temperature MMSE coincides with the ridgless/minimum-norm interpolator.

Replica trick for product random matrix problems

Concretely, for fixed depth L, we consider the proportional limit of width and dataset size:

$$n_0, n_1, \dots, n_L, p \to \infty$$

with

$$\frac{n_\ell}{p} \to \alpha_\ell \in (0, \infty)$$

In this limit, it is straightforward to compute the limiting generalization error using the replica trick, with the relevant order parameters being the overlaps

$$C_\ell^{ab} = \mathbb{E}\big\langle (\mathbf{U}_{\ell+1}^a \cdots \mathbf{U}_L^a \mathbf{v}^a)^\top \big(\mathbf{U}_{\ell+1}^b \cdots \mathbf{U}_L^b \mathbf{v}^b\big)\big\rangle_\beta$$

for $\ell = 1, \dots, L$ and the overlaps

$$Q^{ab} = \mathbb{E}\big\langle (\mathbf{U}_1^a \cdots \mathbf{U}_L^a \mathbf{v}^a - \mathbf{w}_*)^\top \big(\mathbf{U}_1^b \cdots \mathbf{U}_L^b \mathbf{v}^b - \mathbf{w}_*\big)\big\rangle_\beta$$

We will assume replica-symmetry throughout, which should be exact for the random feature model.

A more or less identical setup affords a straightforward replica approach to other product random matrix computations, e.g. of the Stieltjes transform of the eigenspectrum.

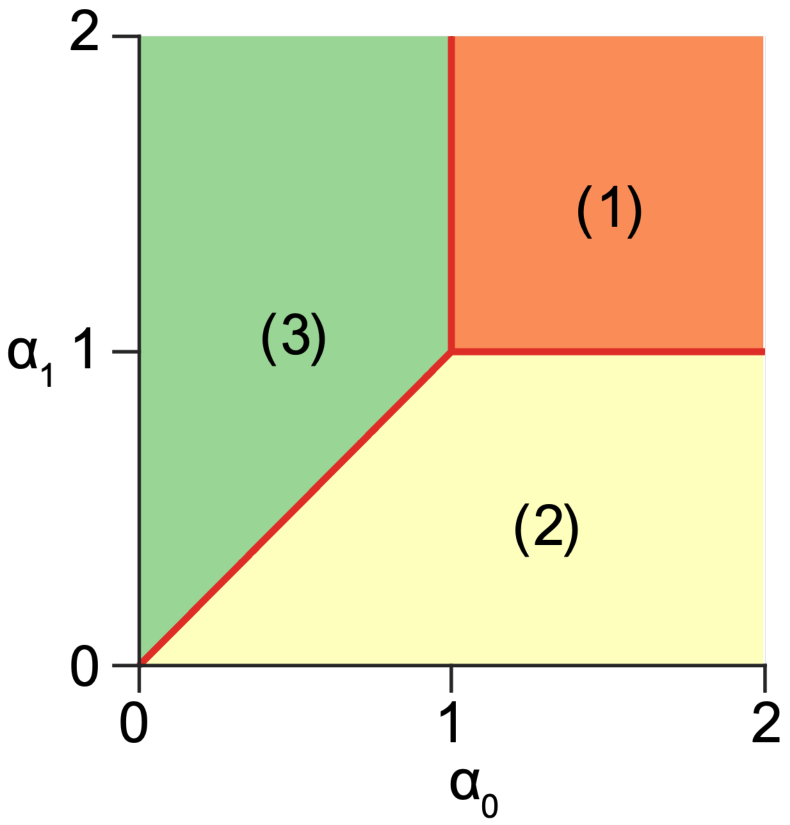
See also Dupic & Péres Castillo 2014 for a related computation using the cavity method

Phase diagram of overparameterization

The phase diagram of this model at zero temperature depends only on the input density α_0 and the minimum layer width

$$\alpha_{\min} = \min\{\alpha_1, \dots, \alpha_L\}$$

1. If $\alpha_0, \alpha_{\min} > 1$, the model is overparameterized
2. If $\alpha_{\min} < 1$ and $\alpha_{\min} < \alpha_0$, the model is bottlenecked
3. If $\alpha_0 < 1$ and $\alpha_{\min} > \alpha_0$, the model is overdetermined



Asymptotic learning curves

For a deep RFM, we have

$$\epsilon_{RF} = \begin{cases} \left(1 - \frac{1}{\alpha_0}\right)\left(1 + \sum_{\ell=1}^{L} \frac{1}{\alpha_\ell - 1}\right) + \left(\sum_{\ell=0}^{L} \frac{1}{\alpha_\ell - 1}\right)\eta^2 + \left(\prod_{\ell=1}^{L}(1-\alpha_\ell)\right)\sigma^2, & \alpha_0, \alpha_{min} > 1 \\ \frac{1 - \alpha_{min}/\alpha_0}{1 - \alpha_{min}} + \frac{\alpha_{min}}{1 - \alpha_{min}}\eta^2 & \alpha_{min} < 1, \alpha_{min} < \alpha_0 \\ \frac{\alpha_0}{1 - \alpha_0}\eta^2, & \alpha_0 < 1, \alpha_0 < \alpha_m \end{cases}$$

For a deep NN, we have

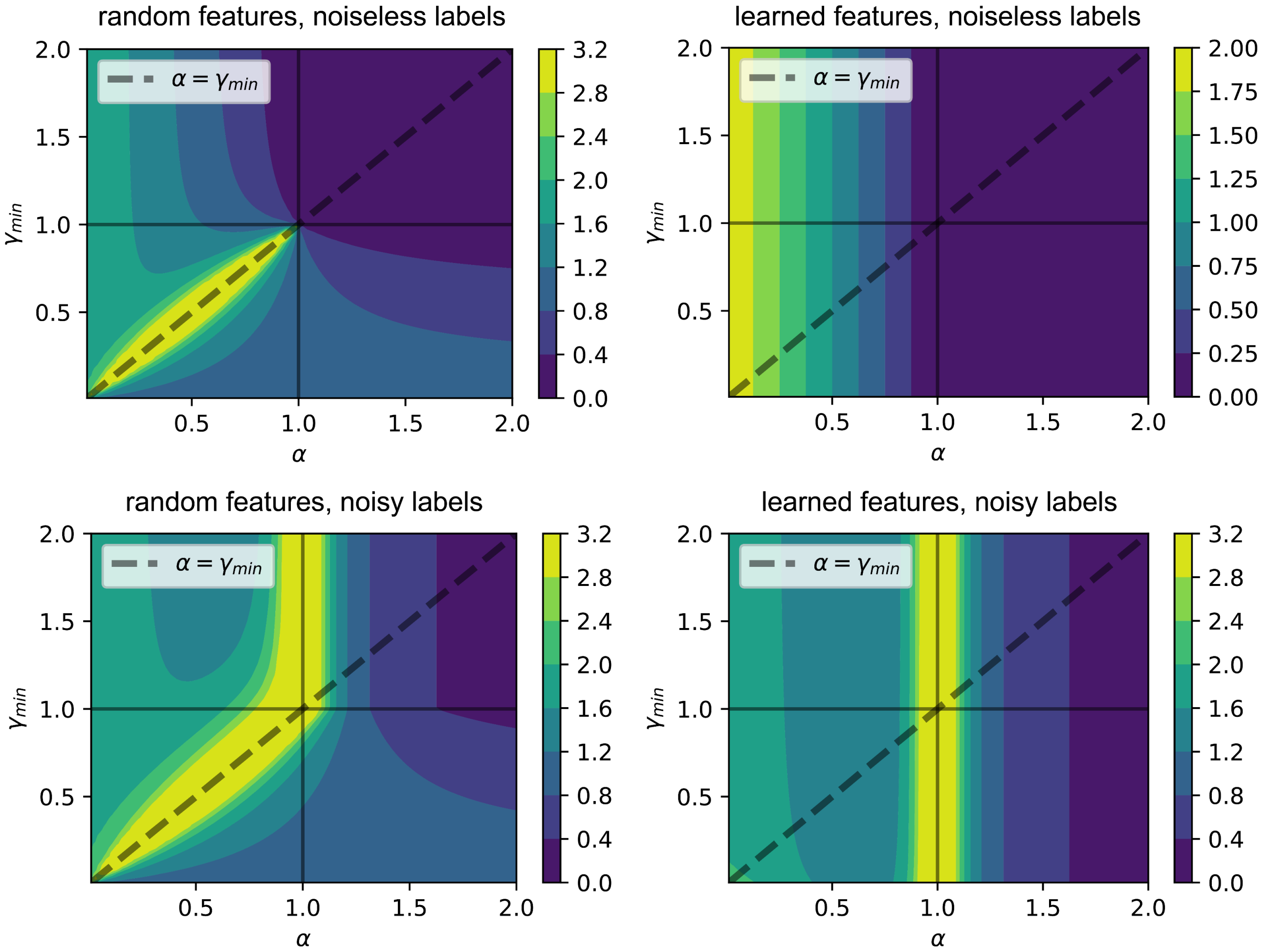
$$\epsilon_{NN} = \begin{cases} \left(1 - \frac{1}{\alpha_0}\right) + \frac{1}{\alpha_0 - 1}\eta^2 + \left(1 - \frac{1}{\alpha_0}\right)z, & \alpha_0 > 1 \\ \frac{\alpha_0}{1 - \alpha_0}\eta^2, & \alpha_0 < 1 \end{cases}$$

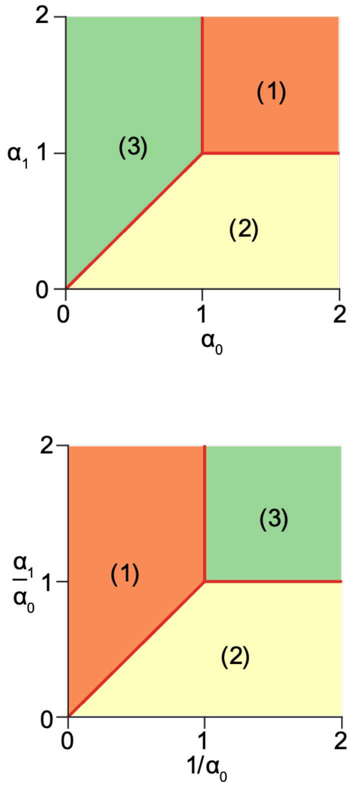
where z solves

$$z^{L+1} = \sigma^2 \prod_{\ell=1}^{L}\left[\left(1 - \frac{1}{\alpha_\ell}\right)z + \frac{1}{\alpha_\ell}\left(1 + \frac{\alpha_0}{1 - \alpha_0}\eta^2\right)\right]$$

Representational flexibility accommodates bottlenecks

$$\gamma_{\min} = \frac{\alpha_{\min}}{\alpha_0}$$

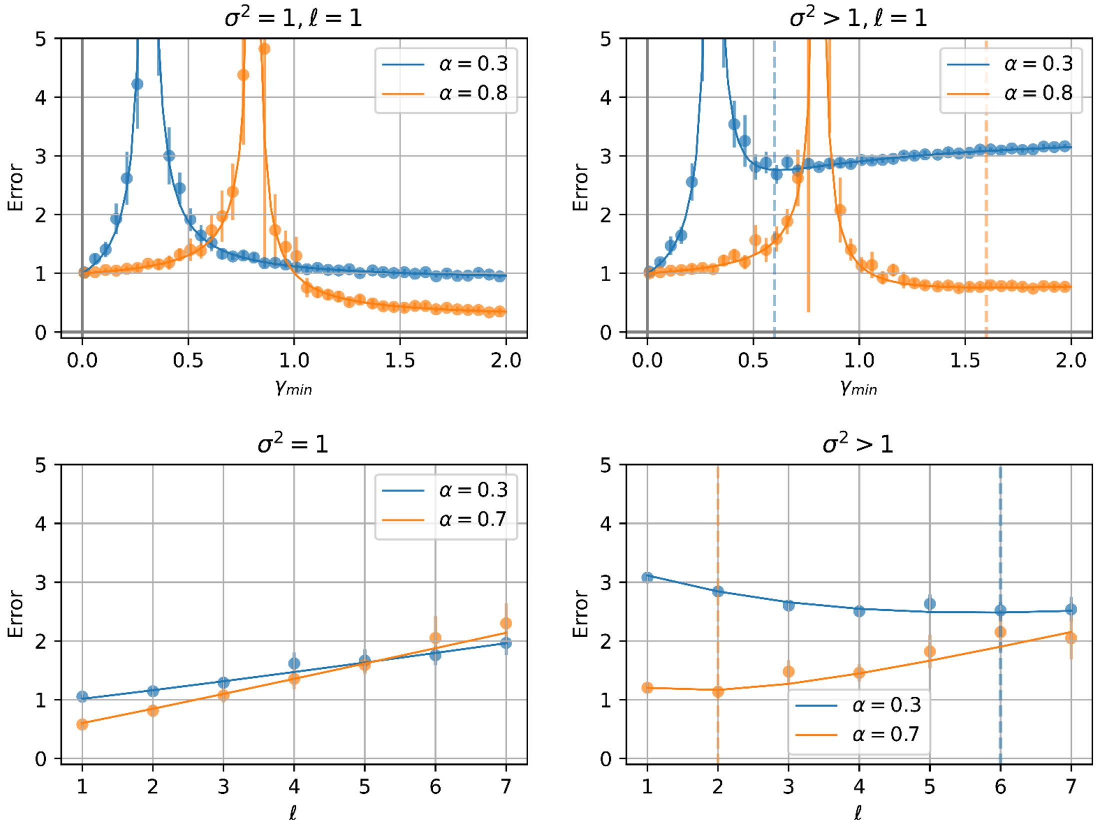




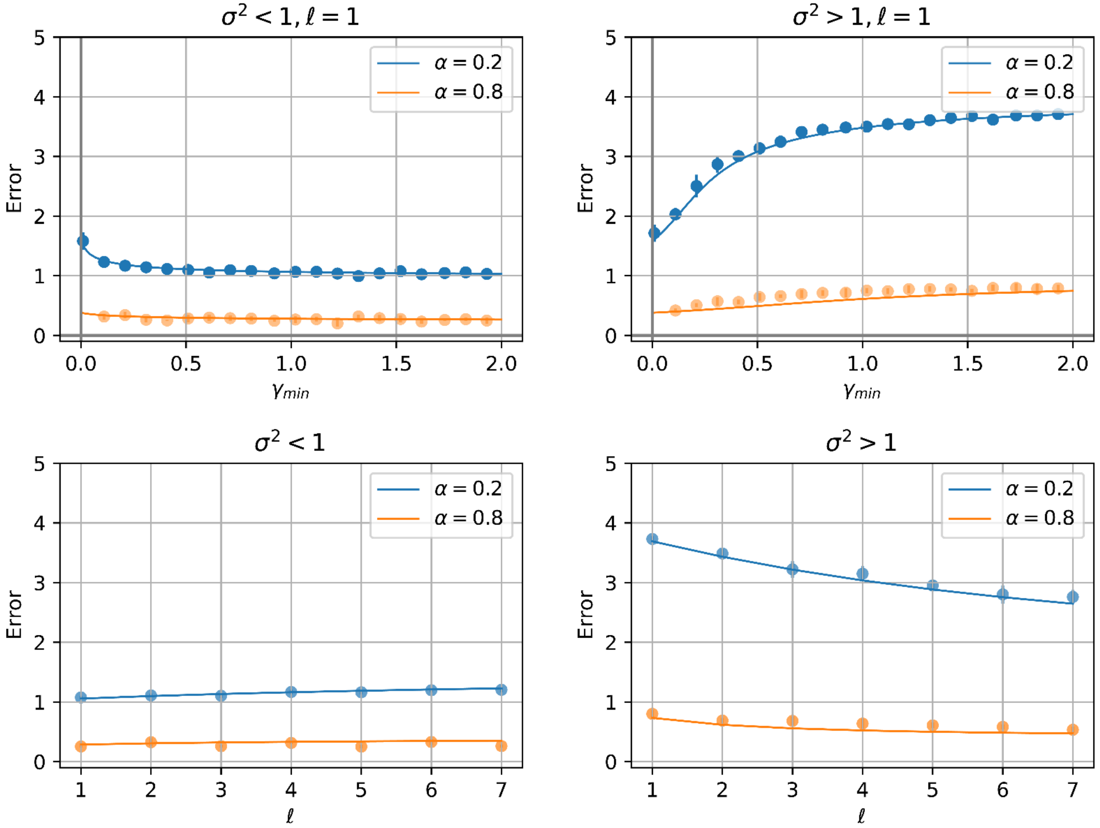
Models with flexible features do not have $\epsilon \to \infty$ when hidden layer width is comparable to input dimension

Optimal architecture depends on target-prior mismatch

RF models can have a non-trivial optimal width and depth

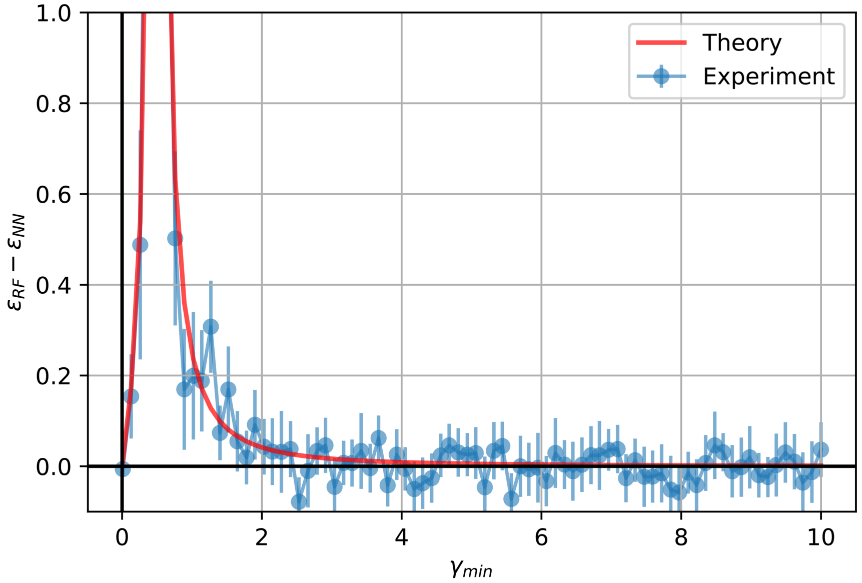


NNs should always be as wide or narrow as possible



Here, $\sigma^2 = \mathbb{E}_{\text{prior}}\|\mathbf{U}_1 \cdots \mathbf{U}_L \mathbf{v}\|^2$

Representational flexibility has a sub-leading effect



At large equal widths $\alpha_1 = \cdots = \alpha_L \gg 1$,

$$\epsilon_{RF} - \epsilon_{NN} \propto L(L+1)\frac{1}{\alpha_1^2} + \mathcal{O}\left(\frac{1}{\alpha_1^3}\right)$$

Thus, leading-order perturbation theory cannot resolve feature learning effects. The leading-order correction to generalization is due entirely to variance.

See also Atanasov, Bordelon, Sainathan, & Pehlevan, *ICLR* 2023 for work on the MLE

How does structure in data and weight priors affect this picture?

A deep structured Gaussian feature model

Now consider data $\mathcal{D} = \{(\mathbf{x}_\mu, y_\mu)\}_{\mu=1}^{p}$ generated by a noisy, correlated teacher

$$\mathbf{x}_\mu \sim_{iid} \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_0)$$
$$y_\mu = w_*^\top x + \xi_\mu$$
$$\xi_\mu \sim \mathcal{N}(0, \eta^2)$$

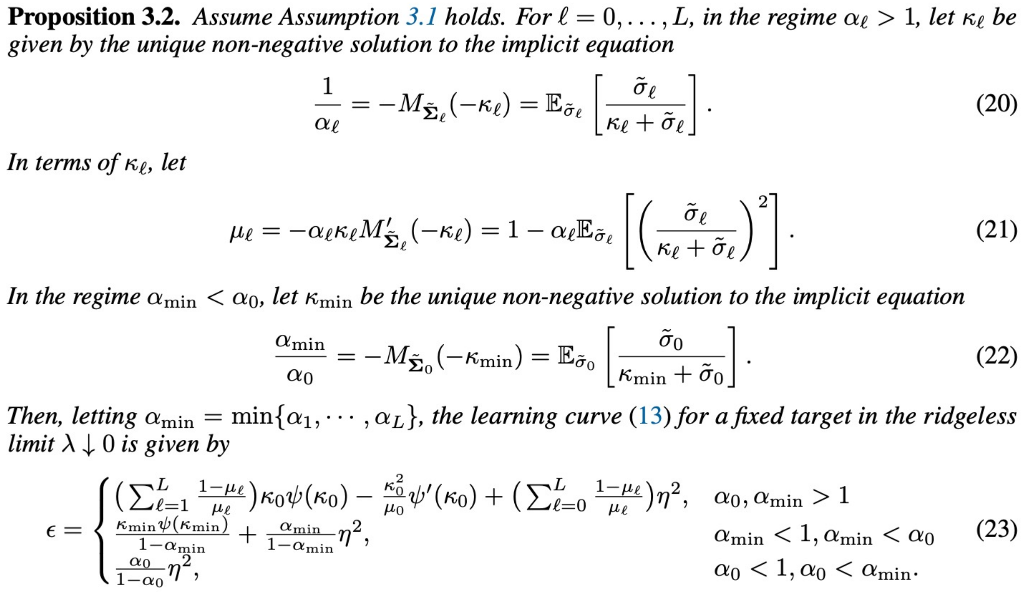
Choose structured matrix-Gaussian priors

$$\mathbf{U}_\ell \sim \mathcal{MN}(\mathbf{0}, \mathbf{\Gamma}_\ell, \mathbf{\Sigma}_\ell / n_{\ell-1})$$
$$\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{L+1} / n_L)$$

How does this structure affect generalization?

Asymptotic learning curves for structured data and weights

Again, we can derive asymptotic formulas for the generalization error for random feature models and deep linear networks, this time in terms of the spectral generating functions of the weight and data covariances.



Proposition 3.2. *Assume Assumption 3.1 holds. For $\ell = 0, \dots, L$, in the regime $\alpha_\ell > 1$, let κ_ℓ be given by the unique non-negative solution to the implicit equation*

$$\frac{1}{\alpha_\ell} = -M_{\tilde{\mathbf{\Sigma}}_\ell}(-\kappa_\ell) = \mathbb{E}_{\tilde{\sigma}_\ell}\left[\frac{\tilde{\sigma}_\ell}{\kappa_\ell + \tilde{\sigma}_\ell}\right]. \tag{20}$$

In terms of κ_ℓ, let

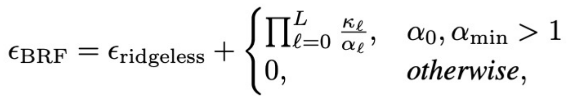
$$\mu_\ell = -\alpha_\ell \kappa_\ell M'_{\tilde{\mathbf{\Sigma}}_\ell}(-\kappa_\ell) = 1 - \alpha_\ell \mathbb{E}_{\tilde{\sigma}_\ell}\left[\left(\frac{\tilde{\sigma}_\ell}{\kappa_\ell + \tilde{\sigma}_\ell}\right)^2\right]. \tag{21}$$

In the regime $\alpha_{\min} < \alpha_0$, let $\kappa_{\min}$ be the unique non-negative solution to the implicit equation

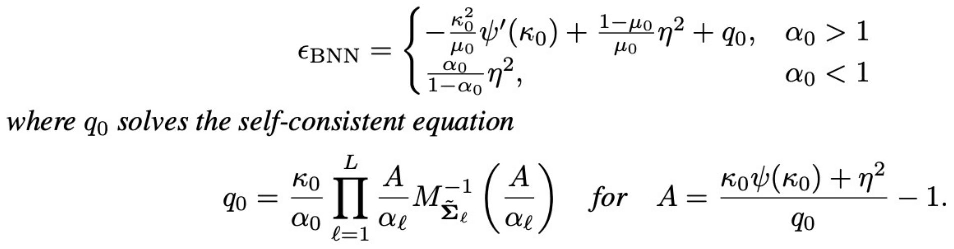
$$\frac{\alpha_{\min}}{\alpha_0} = -M_{\tilde{\mathbf{\Sigma}}_0}(-\kappa_{\min}) = \mathbb{E}_{\tilde{\sigma}_0}\left[\frac{\tilde{\sigma}_0}{\kappa_{\min} + \tilde{\sigma}_0}\right]. \tag{22}$$

Then, letting $\alpha_{\min} = \min\{\alpha_1, \cdots, \alpha_L\}$, the learning curve (13) for a fixed target in the ridgeless limit $\lambda \downarrow 0$ is given by

$$\epsilon = \begin{cases} \left(\sum_{\ell=1}^{L} \frac{1-\mu_\ell}{\mu_\ell}\right)\kappa_0\psi(\kappa_0) - \frac{\kappa_0^2}{\mu_0}\psi'(\kappa_0) + \left(\sum_{\ell=0}^{L} \frac{1-\mu_\ell}{\mu_\ell}\right)\eta^2, & \alpha_0, \alpha_{\min} > 1 \\ \frac{\kappa_{\min}\psi(\kappa_{\min})}{1-\alpha_{\min}} + \frac{\alpha_{\min}}{1-\alpha_{\min}}\eta^2, & \alpha_{\min} < 1, \alpha_{\min} < \alpha_0 \\ \frac{\alpha_0}{1-\alpha_0}\eta^2, & \alpha_0 < 1, \alpha_0 < \alpha_{\min}. \end{cases} \tag{23}$$



$$\epsilon_{\mathrm{BRF}} = \epsilon_{\mathrm{ridgeless}} + \begin{cases} \prod_{\ell=0}^{L} \frac{\kappa_\ell}{\alpha_\ell}, & \alpha_0, \alpha_{\min} > 1 \\ 0, & \textit{otherwise}, \end{cases}$$

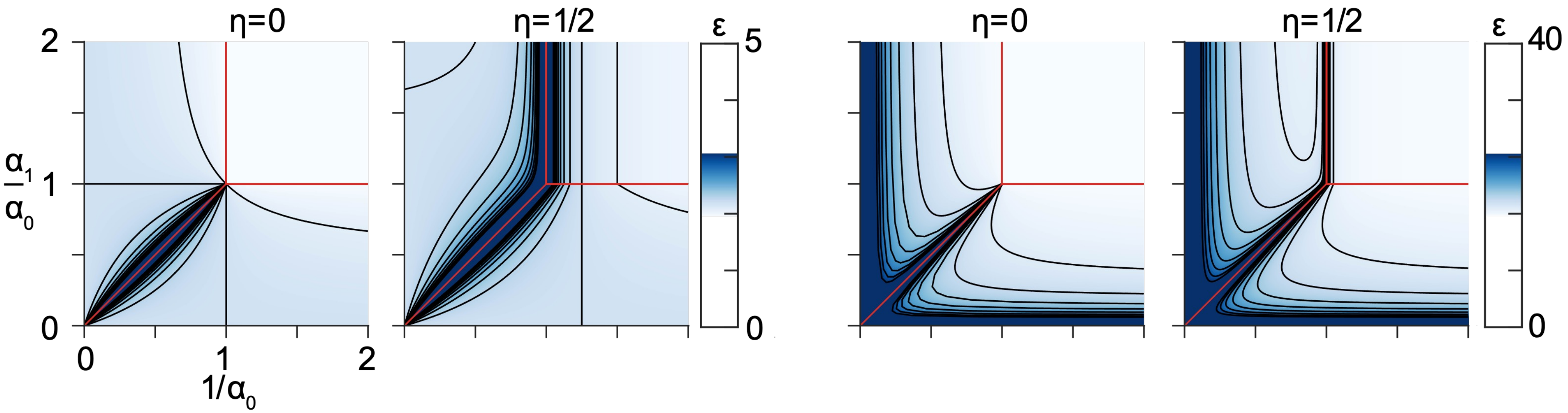


$$\epsilon_{\mathrm{BNN}} = \begin{cases} -\frac{\kappa_0^2}{\mu_0}\psi'(\kappa_0) + \frac{1-\mu_0}{\mu_0}\eta^2 + q_0, & \alpha_0 > 1 \\ \frac{\alpha_0}{1-\alpha_0}\eta^2, & \alpha_0 < 1 \end{cases}$$

where q_0 solves the self-consistent equation

$$q_0 = \frac{\kappa_0}{\alpha_0}\prod_{\ell=1}^{L}\frac{A}{\alpha_\ell}M_{\tilde{\mathbf{\Sigma}}_\ell}^{-1}\left(\frac{A}{\alpha_\ell}\right) \quad \textit{for} \quad A = \frac{\kappa_0\psi(\kappa_0) + \eta^2}{q_0} - 1.$$

Structure does not alter many qualitative phenomena


Unstructured data & weights, random features

Power-law spectra, random features

The scaling of the NN-RF gap remains subleading in width, i.e., of $\mathcal{O}(\alpha^{-2})$.

Structure beyond the first layer is generally detrimental*

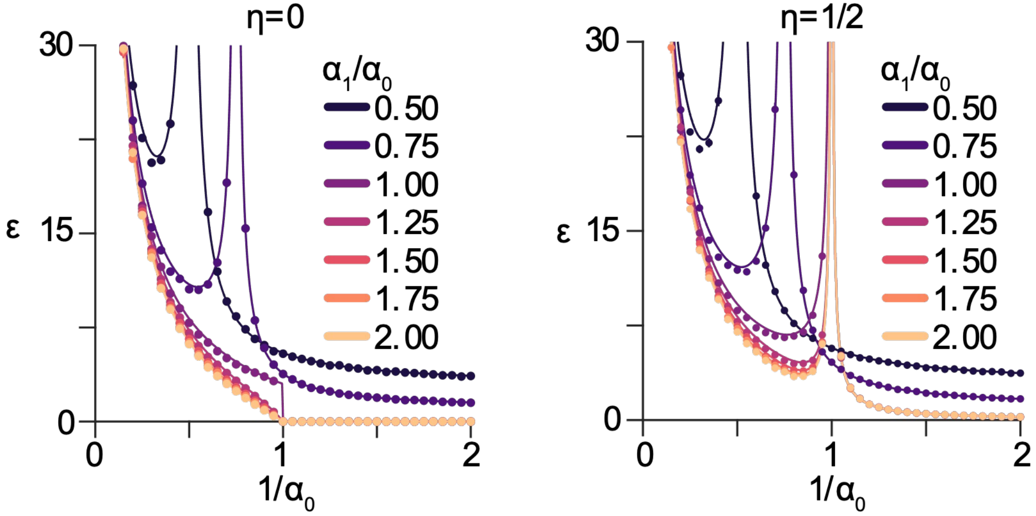
*at small prior variance

In the limit of zero prior variance, structure in the weights is never helpful for random feature models, and doesn’t affect fully-trained networks.

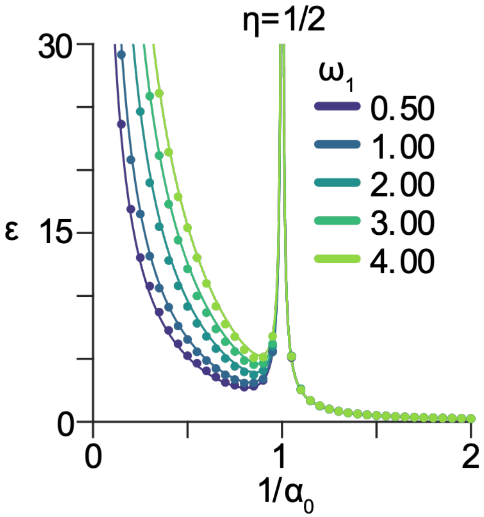
At large prior variance, structure can be helpful for both classes of models, as it can reduce posterior variance.

Example: ridgeless regression with power-law spectra

As a tractable example, consider power-law spectra $\sigma_{\ell,j} \propto j^{-(1+\omega_\ell)}$



The weight spectra do not affect the asymptotic scaling law for ridgeless interpolation: $\epsilon \sim \alpha_0^{\omega_0}$



Conclusions & open problems

- In the weakly-coupled/product RMT regime $L = O(1)$, $p, n_0, \dots, n_d \to \infty$, $n_\ell = \Theta(p)$, deep linear networks are easily studied using replica approaches, even in the presence of structure.
- Leading-order perturbation theory is likely to pick up only variance-related effects rather than feature learning benefits.
- **Open problem:** Establish a rigorous proof for the RS prediction for the generalization error. For unstructured weights, Hanin & Zlokapa's results offer a proof.
- **Open problem:** Can we tackle priors with correlation across layers?
- **Open question:** Can we go beyond Gaussian priors?
- **Open problem:** In what limits can we go beyond linear networks? **See Hugo's talk**

Acknowledgements

Pehlevan group:

Cengiz Pehlevan
Alexander Atanasov
Blake Bordelon
Abdulkadir Canatar (now at Flatiron)
Hamza Chaudhry
Aidan Duncan
Matthew Farrell
Anindita Maiti
Paul Masset (McGill starting 2024)
Shanshan Qin
Benjamin Ruben
Julian Rubinfien (Yale '24 → Harvard Physics)
Sabarish Sainathan
William Tong
Nikhil Vayas
Ningjing Xia
Sheng Yang

NSF

NSF SIMONS CENTER HARVARD
MATHEMATICAL & STATISTICAL ANALYSIS OF BIOLOGY

Google Research

See jzv.io for paper links