Asymptotics of Learning with Deep Structured (Random) Features

Dominik Schröder [1] Daniil Dmitriev [1,2] Hugo Cui [3] Bruno Loureiro [4]

[1]Department of Mathematics, ETH Zurich [2]ETH AI Center [3]Statistical Physics Of Computation lab, Institute of Physics, EPFL [4]Département d'Informatique, École Normale Supérieure - PSL & CNRS

Motivation

The statistics of trained neural network weights carry important information about their generalization performance and inductive bias. Recent work suggests that *rainbow networks*, random networks with weights drawn from an ensemble with matching statistics, can retain a comparable performance to the original trained neural networks [Gut+23]. Motivated by this empirical observation, in this work we provide an exact asymptotic description of the generalization performance of Gaussian rainbow networks in the proportional high-dimensional regime.

Our results shed light on the inductive bias of this class of networks by precisely characterizing how the structured weights impact the performance, while also highlighting the limitations of this description of trained networks.

Ridge Regression

- Supervised learning task with i.i.d. training data $(x_i, y_i)_{i\in[n]}$, $x_i \in \mathbf{R}^p$ such that
$$\mathbf{E}\,x = 0, \quad \mathbf{E}\,y = 0$$
and
$$\mathbf{E}\,xx^\top = \Omega \in \mathbf{R}^{p\times p}, \quad \mathbf{E}\,y^2 = \sigma^2 \in \mathbf{R}_+, \quad \mathbf{E}\,xy = \phi \in \mathbf{R}^p.$$
- Regularized ridge regression with explicit solution
$$\hat\theta_\gamma = \arg\min_\theta \left(\frac{1}{n}\sum_i \left(y_i - \theta^\top x_i\right)^2 + \gamma\|\theta\|_2^2\right) = G(\gamma)\frac{Xy}{n}$$
in terms of the resolvent $G(\gamma)$ of the sample-covariance matrix $\widehat\Omega$,
$$\widehat\Omega := \frac{XX^\top}{n} = \frac{1}{n}\sum_i x_i x_i^\top, \quad G(\gamma) := \left(\widehat\Omega + \gamma\right)^{-1}, \quad X := (x_1, \cdots, x_n) \in \mathbf{R}^{p\times n}.$$
- Generalization error
$$\mathcal{E}_{\mathrm{gen}}(\gamma) = \mathbf{E}\left(y - \hat\theta_\gamma^\top x\right)^2 = \sigma^2 - 2\phi^\top G(\gamma)\frac{Xy}{n} + \frac{y^\top X^\top}{n}G(\gamma)\Omega G(\gamma)\frac{Xy}{n}$$

Background: Marchenko-Pastur

The resolvent $G(\gamma)$ of the sample covariance matrix $\widehat\Omega$ is defined as
$$G(\gamma) := \left(\widehat\Omega + \gamma I\right)^{-1}$$
The deterministic equivalent to $G(\gamma)$ is given by
$$G(\gamma) \approx \frac{\kappa(\gamma)}{\gamma}\left(\Omega + \kappa(\gamma)\right)^{-1},$$
where the effective regularization $\kappa(\gamma)$ is the unique solution to the equation
$$\kappa(\gamma) = \gamma + \frac{\kappa(\gamma)}{n}\operatorname{Tr}\Omega\left(\Omega + \kappa(\gamma)\right)^{-1}.$$
and satisfies the bounds
$$\gamma \le \kappa(\gamma) \le \min\left\{\gamma + \frac{\operatorname{Tr}\Omega}{n}, \frac{\gamma}{1 - \frac{1}{n}\operatorname{rank}\Omega}\right\}$$
and the asymptotics $\kappa(0+) = \lim_{\gamma\to 0}\kappa(\gamma) = 0$ for $\operatorname{rank}(\Omega) \le n$, while $\kappa(0+) > 0$ for $\operatorname{rank}\Omega > n$.

Assumptions

- Concentration. We assume that scalar Lipschitz functions of the feature matrix X are sub-Gaussian. (Example. Lipschitz functions of Gaussian random vectors)
- Comparable dimensions We assume that $\max\{n, k, p\} \ll (\min\{n, k, p\})^{3/2}$, somewhat relaxing the usual proportionality assumption.

Theorem (Generalization Error Equivalent)

The generalization error $\mathcal{E}_{\mathrm{gen}}$ is asymptotically equal to the deterministic quantity
$$\mathcal{E}_{\mathrm{gen}}^{\mathrm{rmt}}(\gamma) := \frac{\sigma^2 - \phi^\top\frac{\Omega + 2\kappa(\gamma)}{(\Omega+\kappa(\gamma))^2}\phi}{1 - \frac{1}{n}\operatorname{Tr}\Omega^2(\Omega + \kappa(\gamma))^{-2}}. \tag{1}$$

Special Case: Linear Labels

In the case of linearly generated labels
$$y = \theta_*^\top x + \epsilon, \quad \mathbf{E}\,\epsilon = 0, \quad \mathbf{E}\,\epsilon^2 = \delta^2$$
the deterministic equivalent (1) admits the bias variance decomposition
$$\mathcal{E}_{\mathrm{gen}}^{\mathrm{rmt}}(\gamma) = \frac{\delta^2 + \kappa(\gamma)^2\theta_*^\top\frac{\Omega}{(\Omega+\kappa(\gamma))^2}\theta_*}{1 - \frac{1}{n}\operatorname{Tr}\Omega^2(\Omega + \kappa(\gamma))^{-2}} \tag{2}$$

Main Case of Interest: Feature Ridge Regression

Consider the teacher-student setting, i.e. when
$$x = \varphi(\mathsf{x}), \quad y = \theta_*^\top\varphi_*(\mathsf{x}) + \epsilon, \quad \mathbf{E}\,\epsilon = 0, \quad \mathbf{E}\,\epsilon^2 = \delta^2,$$
for some random vector $\mathsf{x} \in \mathbf{R}^d$, and feature maps $\varphi\colon \mathbf{R}^d \to \mathbf{R}^p$, $\varphi_*\colon \mathbf{R}^d \to \mathbf{R}^k$ s.t.
$$\mathbf{E}\,\varphi(\mathsf{x}) = 0, \quad \mathbf{E}\,\varphi_*(\mathsf{x}) = 0$$
and
$$\mathbf{E}\,\varphi(\mathsf{x})\varphi(\mathsf{x})^\top = \Omega, \quad \mathbf{E}\,\varphi(\mathsf{x})\varphi_*(\mathsf{x})^\top = \Phi, \quad \mathbf{E}\,\varphi_*(\mathsf{x})\varphi_*(\mathsf{x})^\top = \Psi,$$
where $\varphi(\mathsf{x})$ is called the student network and $\varphi(\mathsf{x})$ is called the teacher network.

The deterministic equivalent (1) admits the bias variance decomposition
$$\mathcal{E}_{\mathrm{gen}}^{\mathrm{rmt}}(\gamma) = \frac{\delta^2 + \theta_*^\top\left(\Psi - \Phi^\top\frac{\Omega + 2\kappa(\gamma)}{(\Omega+\kappa(\gamma))^2}\Phi\right)\theta_*}{1 - \frac{1}{n}\operatorname{Tr}\Omega^2(\Omega + \kappa(\gamma))^{-2}} \tag{3}$$

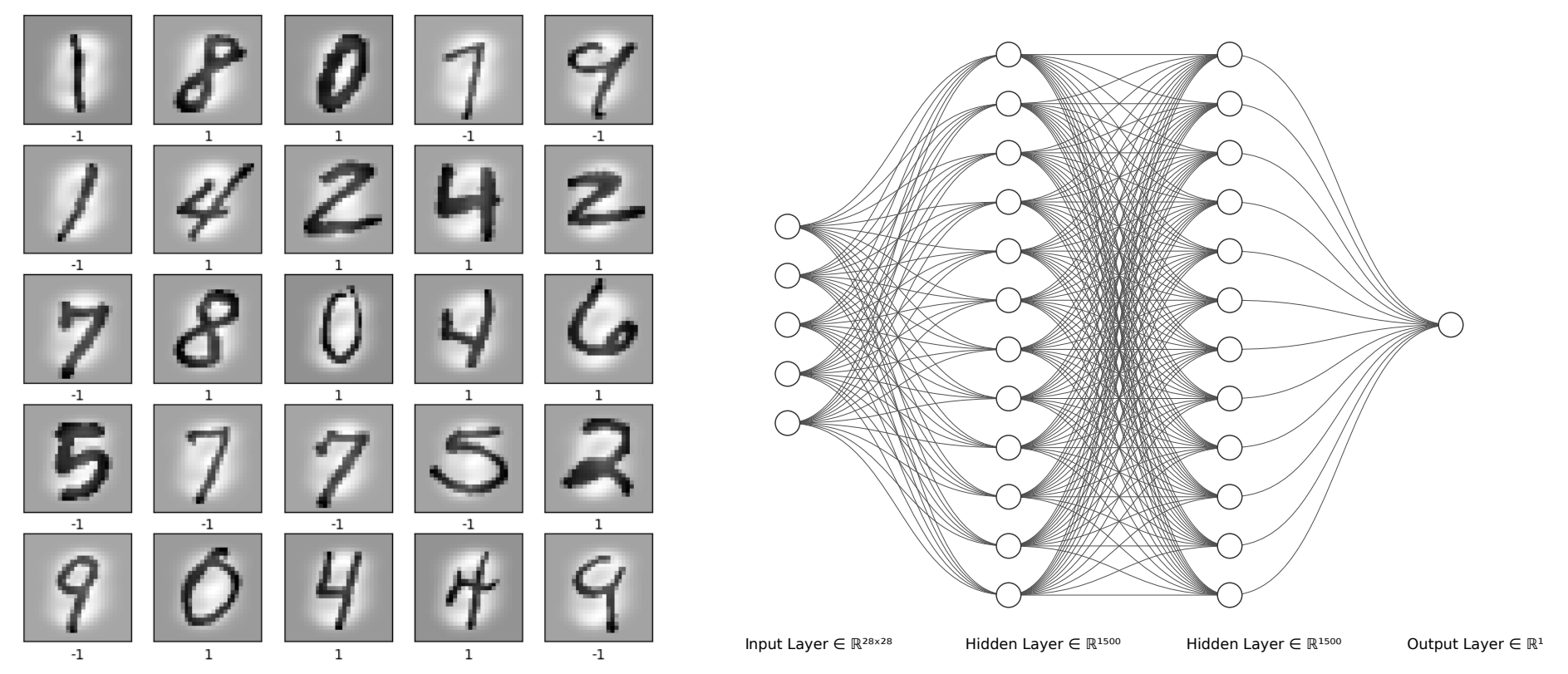
In certain cases (see Linearization), it is possible to simplify the expression above by replacing Ω, Ψ and Φ with easy-to-compute approximations.

Previous results

- The linear label case reduces to ordinary linear regression, see e.g. [Bac24]
- Our result confirms Conjecture 1 of [Lou+22]
- Independently and concurrently to the current work [LP23] obtained similar results under different assumptions. Most importantly [LP23] considers one-layer unstructured random feature models and computes the *empirical generalization error* for a deterministic data set, while we consider general Lipschitz features of random data, and compute the generalization error
- In the unstructured random feature model [MMM22; AP20] obtained an expression for the generalization error under the assumption that the target model is linear or rotationally invariant.

Case study: Gaussian rainbow networks

We take the MNIST dataset and normalize the images to have (empirical) mean zero. We split the dataset into for parts $\mathcal{D}_{\mathrm{NN}}, \mathcal{D}_{\mathrm{reg}}, \mathcal{D}_{\mathrm{test}}, \mathcal{D}_{\mathrm{cov}}$. Then we train a simple two-hidden layer neural network
$$\varphi(x) = \operatorname{relu}(W_2\operatorname{relu}(W_1 x)), \quad f(x) = w_3^\top\varphi(x)$$
to recognize whether a given digit is even or odd, using $\mathcal{D}_{\mathrm{NN}}$



During training we save the weights W_1, W_2 to obtain a sequence of feature maps φ_t. Then we use $\mathcal{D}_{cov}$ to empirically estimate[a] Ω and ϕ, choosing $\mathcal{D}_{\mathrm{NN}}$ large enough so that $|\mathcal{D}_{\mathrm{NN}}| \gg p$. Then we use $\mathcal{D}_{\mathrm{reg}}$ to perform a feature regression with the trained features and evaluate the generalization performance using $\mathcal{D}_{\mathrm{test}}$.

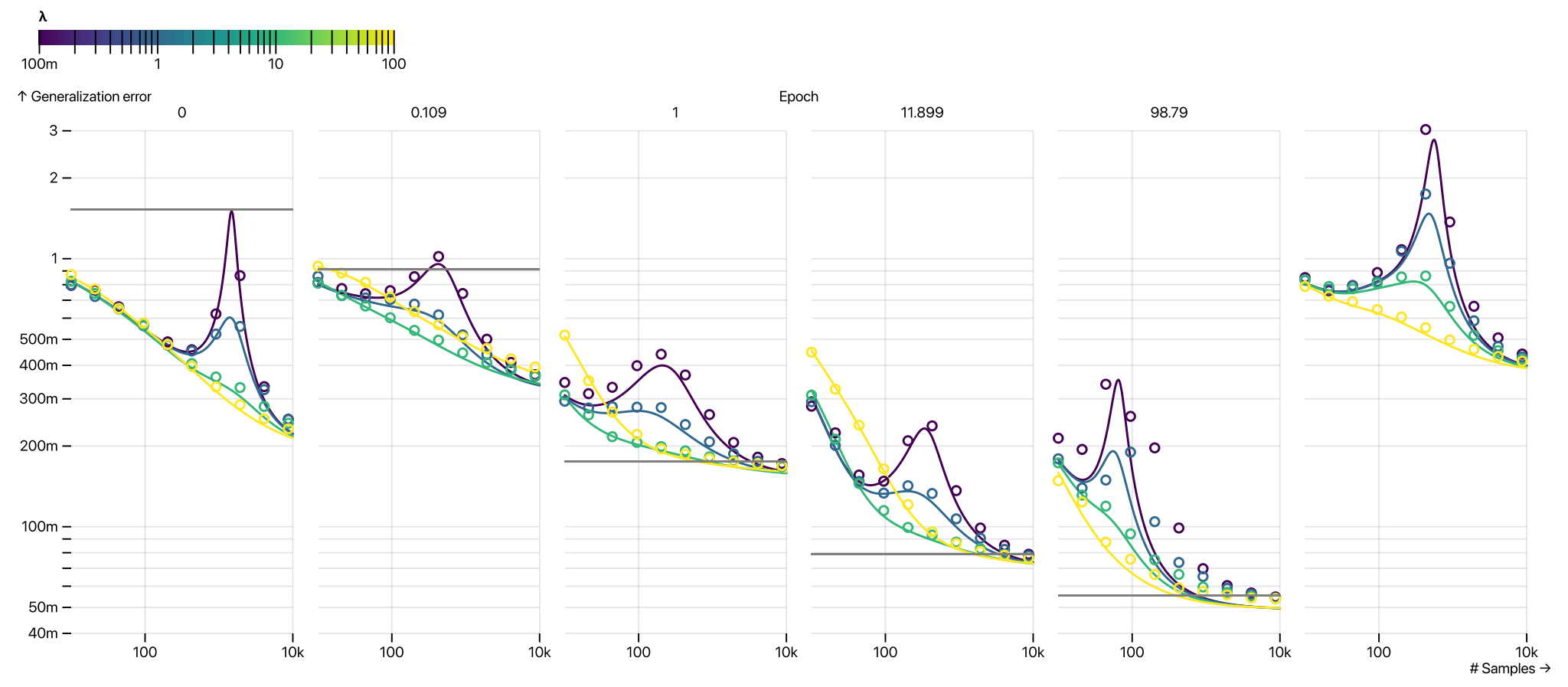


Figure 1. Generalization error of feature ridge regression during training, compared with linear regression. The deterministic equivalent remains an excellent approximation throughout the training process. Note that for the specific task even the untrained network outperforms linear regression.

So far we have looked at fixed regularization. Using the deterministic equivalent we can also determine the optimal regularization in order to obtain the following results:

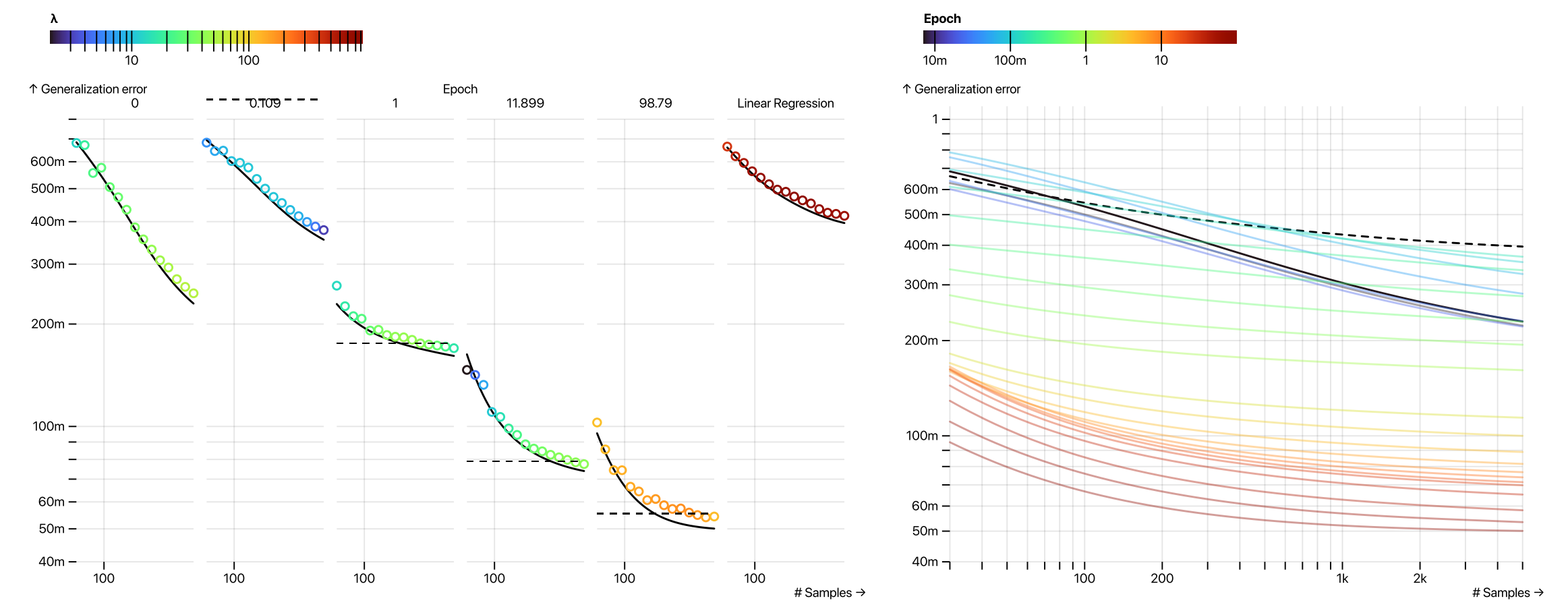


Figure 2. Generalization error of feature ridge regression during training at optimal regularization, compared with linear regression. In the right plot the dashed line represents linear regression.

[a] Note that $\sigma = 1$ by label choice

Linearization

Let $(W_\ell), (V_\ell)$ be two collection of matrices with widths p_ℓ. Consider two networks:
$$\varphi(\mathsf{x}) = \varphi_L(W_L\varphi_{L-1}(\dots\varphi_1(W_1 x))) \text{ and } \varphi_*(\mathsf{x}) = \widetilde\varphi_L(V_L\widetilde\varphi_{L-1}(\dots\widetilde\varphi_1(V_1 x))),$$
the student and teacher networks respectively. Assume that rows of W_ℓ and V_ℓ are i.i.d. $\sim w_\ell$ and $\sim v_\ell$ respectively. Define
$$C_\ell := p_\ell\,\mathbf{E}\,w_\ell w_\ell^\top \quad \widetilde C_\ell := p_\ell\,\mathbf{E}\,v_\ell v_\ell^\top \quad \check C_\ell := p_\ell\,\mathbf{E}\,w_\ell v_\ell^\top.$$

Equation (3) defines a deterministic equivalent $\mathcal{E}_{\mathrm{gen}}^{\mathrm{rmt}}(\gamma)$ for the generalization error as a function of $\Omega_L = \mathbf{E}_{\mathsf{x}}\,\varphi(\mathsf{x})\varphi(\mathsf{x})^\top, \Psi_L = \mathbf{E}_{\mathsf{x}}\,\varphi_*(\mathsf{x})\varphi_*(\mathsf{x})^\top$ and $\Phi_L = \mathbf{E}_{\mathsf{x}}\,\varphi(\mathsf{x})\varphi_*(\mathsf{x})^\top$.

These expectations are hard to compute because of non-linearities of $\varphi(\mathsf{x}), \varphi_*(\mathsf{x})$. We obtain the following linearizations:
$$\begin{aligned}\Omega_\ell^{\mathrm{lin}} &= (\kappa_\ell^1)^2 W_\ell\Omega_{\ell-1}^{\mathrm{lin}}W_\ell^\top + (\kappa_\ell^*)^2 I,\\ \Psi_\ell^{\mathrm{lin}} &= (\widetilde\kappa_\ell^1)^2 V_\ell\Psi_{\ell-1}^{\mathrm{lin}}V_\ell^\top + (\widetilde\kappa_\ell^*)^2 I,\\ \Phi_\ell^{\mathrm{lin}} &= \kappa_\ell^1\widetilde\kappa_\ell^1 W_\ell\Phi_{\ell-1}^{\mathrm{lin}}V_\ell^\top + (\check\kappa_\ell^*)^2 I,\end{aligned} \tag{4}$$
where $(\kappa_\ell^1, \kappa_\ell^*, \widetilde\kappa_\ell^1, \widetilde\kappa_\ell^*, \check\kappa_\ell^1)$ are some simple functions of $\varphi_\ell, \widetilde\varphi_\ell$:
$$\begin{aligned}\kappa_\ell^1 &:= \mathbf{E}\,\varphi_\ell'(N_\ell), & \widetilde\kappa_\ell^1 &:= \mathbf{E}\,\widetilde\varphi_\ell'(\widetilde N_\ell),\\ \kappa_\ell^* &:= \sqrt{\mathbf{E}[\varphi_\ell(N_\ell)^2] - r_\ell(\kappa_\ell^1)^2}, & \widetilde\kappa_\ell^* &:= \sqrt{\mathbf{E}[\widetilde\varphi_\ell(\widetilde N_\ell)^2] - \widetilde r_\ell(\widetilde\kappa_\ell^1)^2}\\ \check\kappa_\ell^* &:= \sqrt{\mathbf{E}[\varphi_\ell(N_\ell)\widetilde\varphi_\ell(\widetilde N_\ell)] - \check r_\ell\kappa_\ell^1\widetilde\kappa_\ell^1},\end{aligned} \tag{5}$$
where $N_\ell, \widetilde N_\ell$ are jointly mean-zero Gaussians with
$$\mathbf{E}\,N_\ell^2 = r_\ell := \operatorname{Tr}[C_\ell\Omega_{\ell-1}^{\mathrm{lin}}] \quad \mathbf{E}\,\widetilde N_\ell^2 = \widetilde r_\ell := \operatorname{Tr}[\widetilde C_\ell\Psi_{\ell-1}^{\mathrm{lin}}] \quad \mathbf{E}\,N_\ell\widetilde N_\ell = \check r_\ell := \operatorname{Tr}[\check C_\ell\Phi_{\ell-1}^{\mathrm{lin}}].$$

Theorem (Linearization for 2-layered networks)

We prove that, under some assumptions, for $L = 2$,
$$\|\Omega - \Omega_2^{\mathrm{lin}}\|_F + \|\Psi - \Psi_2^{\mathrm{lin}}\|_F + \|\Phi - \Phi_2^{\mathrm{lin}}\|_F \prec 1. \tag{6}$$

Proof technique: We use Wiener chaos expansion, the generalization of Hermite expansion, to decompose random variables $F = F(\mathsf{x})$:
$$F = \mathbf{E}\,F + \sum_{p\ge 1} I_p\left(\frac{\mathbf{E}\,D^p F}{p!}\right),$$
where I_p is the multiple integral and D^p is the p-th Malliavin derivative.

We also use Stein's method, which allows to prove that
$$d_W(F, N) \lesssim \mathbf{E}\,|\,\mathbf{E}\,F^2 - \langle DF, -DL^{-1}F\rangle|,$$
where $F := w^\top\varphi_1(Wx), N \sim \mathcal{N}(0, \mathbf{E}\,F^2)$ and L^{-1} is the pseudo-inverse of the generator of the Ornstein-Uhlenbeck semigroup.

Conjecture: We conjecture that Equation (6) holds for any fixed depth $L \ge 2$.

References

[AP20] Ben Adlam and Jeffrey Pennington. "The Neural Tangent Kernel in High Dimensions: Triple Descent and a Multi-Scale Theory of Generalization". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 13–18 Jul 2020, pp. 74–84. URL: https://proceedings.mlr.press/v119/adlam20a.html.

[Bac24] Francis Bach. "High-Dimensional Analysis of Double Descent for Linear Regression with Random Projections". In: *SIAM Journal on Mathematics of Data Science* 6.1 (2024), pp. 26–50. DOI: 10.1137/23M1558781. eprint: https://doi.org/10.1137/23M1558781. URL: https://doi.org/10.1137/23M1558781.

[Gut+23] Florentin Guth et al. *A Rainbow in Deep Network Black Boxes*. 2023. arXiv: 2305.18512 [cs.LG]. URL: https://arxiv.org/abs/2305.18512.

[Lou+22] Bruno Loureiro et al. "Learning curves of generic features maps for realistic datasets with a teacher-student model". In: *J. Stat. Mech. Theory Exp.* 2022.11 (2022), Paper No. 114001, 78. DOI: 10.1088/1742-5468/ac9825.

[LP23] Hugo Latourelle-Vigeant and Elliot Paquette. "Matrix Dyson equation for correlated linearizations and test error of random features regression". In: *arXiv preprint arXiv:2312.09194* (2023).

[MMM22] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. "Generalization error of random feature and kernel methods: hypercontractivity and kernel matrix concentration". In: *Appl. Comput. Harmon. Anal.* 59 (2022), pp. 3–84. ISSN: 1063-5203. DOI: 10.1016/j.acha.2021.12.003.

| arXiv:2402.13999

ICML 2024 Vienna

github:feature-ridge-regression |