Deterministic equivalent and error universality of deep random features learning

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

[1] ETH Zurich [2] EPFL [3] ENS PSL

* Equal contribution

Check it out!

Problem Setting

Let $(x^\mu, y^\mu) \in \mathbb{R}^d \times \mathcal{Y}$, $\mu \in [n]$, with $x^\mu \overset{iid}{\sim} \mathcal{N}(0_d, \Omega_0)$ and $y^\mu = f_\star(\mathrm{x}^\mu)$ a (random) target function. We consider a generalised linear estimation:

$$\hat{y} = \sigma\left(\frac{\theta^\top \varphi(x)}{\sqrt{k}}\right), \tag{1}$$

with *deep random features* (dRF):

$$\varphi(x) := \underbrace{(\varphi_L \circ \varphi_{L-1} \circ \cdots \circ \varphi_2 \circ \varphi_1)}_{L}(x), \tag{2}$$

where the post-activations are given by:

$$\varphi_\ell(h) = \sigma_\ell\left(\frac{1}{\sqrt{k_{\ell-1}}} W_\ell \cdot h\right), \quad \ell \in [L]. \tag{3}$$

The entries of $\{W_\ell \in \mathbb{R}^{k_\ell \times k_{\ell-1}}\}_{\ell \in [L]}$ are $(W_\ell)_{ij} \overset{iid}{\sim} \mathcal{N}(0, \Delta_\ell)$.

Sample covariance matrices

Sample covariance matrix $\widehat{\Sigma} := \mathcal{X}\mathcal{X}^\top / n \in \mathbb{R}^{d \times d}$ for $\mathcal{X} := (x_1, \dots, x_n)$, corresponding to the *population covariance matrix* Σ.
Gram matrix $\check{\Sigma} := \mathcal{X}^\top \mathcal{X}/n \in \mathbb{R}^{n \times n}$ has the same non-zero eigenvalues.
In the regime $d \sim n \gg 1$ the empirical spectral density $\mu(\widehat{\Sigma}) := d^{-1}\sum_{\lambda \in \mathrm{Spec}(\widehat{\Sigma})} \delta_\lambda$ of $\widehat{\Sigma}$ is approximately equal to the *free multiplicative convolution* of $\mu(\Sigma)$ and a Marchenko-Pastur distribution μ^c_{MP} with $c = d/n$,

$$\mu(\widehat{\Sigma}) \approx \mu(\Sigma) \boxtimes \mu^{d/n}_{\mathrm{MP}}. \tag{4}$$

The free multiplicative convolution $\mu \boxtimes \mu^c_{\mathrm{MP}}$ is the unique distribution ν whose Stieltjes transform $m = m_\nu(z) := \int (x - z)^{-1} \mathrm{d}\nu(x)$ satisfies the scalar *self-consistent equation*

$$zm = \frac{z}{1 - c - czm} m_\mu\left(\frac{z}{1 - c - czm}\right). \tag{5}$$

Gaussian universality of the test error

$\theta \in \mathbb{R}^k$ is obtained via the regularized *empirical risk minimization*:

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^k}{\mathrm{argmin}}\left[\sum_{\mu=1}^n \ell(y^\mu, \theta^\top \varphi(\mathrm{x}^\mu)) + \frac{\lambda}{2}||\theta||^2\right], \tag{6}$$

where $\ell : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_+$ is a convex loss function.
We assume that the labels are generated by a deep random neural network:

$$f_\star(\mathrm{x}^\mu) = \sigma^\star\left(\frac{\theta_\star^\top \varphi^\star(\mathrm{x}^\mu)}{\sqrt{k^\star}}\right).$$

Here, $\theta_\star \in \mathbb{R}^{k^\star}$ and $\varphi^\star$ denotes composition $\varphi^\star_{L^\star} \circ \dots \circ \varphi^\star_1$:

$$\varphi^\star_\ell(\mathrm{x}) = \sigma^\star_\ell\left(\frac{1}{\sqrt{k^\star_{\ell-1}}} W^\star_\ell \cdot \mathrm{x}\right).$$

The matched setting $\varphi = \varphi^\star$ with the readout layer trained using a square loss corresponds to $\mathcal{Y} = \mathbb{R}$, $\ell(y, \hat{y}) = {}^1\!/_2(y - \hat{y})^2$. Here (6) equals to

$$\hat{\theta} = {}^1\!/\!\sqrt{k}(\lambda I_k + {}^1\!/_k X_L X_L^\top)^{-1} X_L y \tag{7}$$

where $X_L \in \mathbb{R}^{k \times n}$ contains last layer features column-wise and $y \in \mathbb{R}^n$.

Deterministic equivalents

The relationship (4) between the asymptotic spectra of Σ and $\widehat{\Sigma}, \check{\Sigma}$ extends to eigenvectors, and the resolvents $\widehat{G}(z) := (\widehat{\Sigma} - z)^{-1}$, $\check{G}(z) := (\check{\Sigma} - z)^{-1}$ are asymptotically equal to *deterministic equivalents*.

Iteration over one layer

Consider a data matrix $X_0 \in \mathbb{R}^{d \times n}$ and $X_1 := \sigma_1(W_1 X_0/\sqrt{d}) \in \mathbb{R}^{k_1 \times n}$. Furthermore, assume that the Gram matrix concentrates as

$$\left\|\frac{X_0^\top X_0}{d} - r_1 I\right\|_{\max} \prec \frac{1}{\sqrt{n}}, \quad \left\|\frac{X_0}{\sqrt{d}}\right\| \prec 1 \tag{8}$$

for some positive constant r_1. For any deterministic A and Lipschitz-continuous σ_1, for any $z \in \mathbb{C} \backslash \mathbb{R}_+$ (denoting $\langle A \rangle := {}^{\mathrm{Tr} A}\!/_n$ for $A \in \mathbb{R}^{n \times n}$)

$$\left|\left\langle A\left[\left(\frac{X_1^\top X_1}{k_1} - z\right)^{-1} - \left(c_1(z)\frac{X_0^\top X_0}{d} + c_2(z)\right)^{-1}\right]\right\rangle\right| \prec \frac{\langle AA^* \rangle^{1/2}}{\sqrt{n}},$$

and similar result holds for the matrix ${}^{X_1 X_1^\top}\!/_{k_1}$ Furthermore, Assumption (8) holds true with X_0, r_1 replaced by X_1, r_2, respectively.

Proof idea

The proof follows from the following sequence of approximations

$$\left(\frac{X_1^\top X_1}{k_1} - z\right)^{-1} \approx \left(c_0(z)\Sigma_X - z\right)^{-1} \approx \left(c_1(z)\frac{X_0^\top X_0}{d} + c_2(z)\right)^{-1}, \tag{9}$$

where $\Sigma_X := \mathbb{E}_{w \sim \mathcal{N}(0,I)} \sigma\left(\frac{X_0^\top w}{\sqrt{d}}\right)\sigma\left(\frac{w^\top X_0}{\sqrt{d}}\right) \in \mathbb{R}^{n \times n}$. The first approximation follows from [1], and the second one from Hermite series expansion.
The proposition can be iterated over arbitrary finite number of layers L

$$\left(\frac{X_L^\top X_L}{k_L} - z\right)^{-1} \approx \left(c_1'\frac{X_{L-1}^\top X_{L-1}}{k_{L-1}} + c_2'\right)^{-1} \approx \dots \approx \left(c_1\frac{X_0^\top X_0}{d} + c_2\right)^{-1}, \tag{10}$$

where c_1, c_1', c_2, c_2' are some functions of $z \in \mathbb{C} \setminus \mathbb{R}_+$.
In the proof, we assume fixed X_0 and random W_ℓ, leading to Σ_X. This approach facilitates iteration over the layers and appears in [2]. Another view is to consider $X_\ell X_\ell^\top/n$ as a sample covariance matrix with population covariance $\Omega_\ell := \mathbb{E}_{X_0}\frac{X_\ell X_\ell^\top}{n}$ since the matrix X_ℓ conditioned on $W_1, \dots, W_\ell$ has independent columns. The matrices are related as the population covariance and Gram matrices. We also derive a heuristic formula for Ω_ℓ.

Ridge universality of matched target

Let $\lambda > 0$. In the limit $n \sim d \sim k_\ell \gg 1$, under Assumption (8), the asymptotic test error of the ridge estimator (7) on the target (1) with $L = L^\star$ and $\varphi^\star_\ell = \varphi_\ell$ and additive $\mathcal{N}(0, \Delta)$ noise is given by:

$$\epsilon_g(\hat{\theta}) \xrightarrow{k \to \infty} \epsilon_g^\star = \Delta\left(\langle \Omega_L \rangle \widetilde{m}_L(-\lambda) + 1\right) - \lambda(\lambda - \Delta)\langle \Omega_L \rangle \partial_\lambda \widetilde{m}_L(-\lambda)$$

where $\widetilde{m}_L$ can be recursively computed.
This implies Gaussian universality of this model, since (1) agrees with the asymptotic test error of data $\mathrm{x} \sim \mathcal{N}(0_d, \Omega_L)$ derived in [3].

General case

The same result is shown to hold numerically for a much wider class of models, i.e., they belong to the *Gaussian universality class*. There is an equivalent Gaussian covariate model consisting of doing generalized linear estimation on $\check{\mathcal{D}} = \{v^\mu, \check{y}^\mu\}_{\mu \in [n]}$ with labels $\check{y}^\mu = f_\star({}^1\!/\!\sqrt{k^\star}\theta_\star^\top u^\mu)$ and:

$$(u, v) \sim \mathcal{N}\begin{pmatrix} \Psi_{L^\star} & \Phi_{L^\star L} \\ \Phi_{L^\star L}^\top & \Omega_L \end{pmatrix} \tag{11}$$

where $\Phi \in \mathbb{R}^{k^\star \times k}$ and $\Psi \in \mathbb{R}^{k^\star \times k^\star}$ are the covariances between the model and target features and the target variance respectively. This provides an analogous contribution as [4] to the case of multi-layer random features.

Depth-induced implicit regularization

An insightful takeaway is that the activations in dRF (2) share the same population statistics as the activations in a deep *noisy* linear network

$$\varphi_\ell^{\mathrm{lin}}(\mathrm{x}) = \kappa_1^\ell \frac{W_\ell^\top \mathrm{x}}{\sqrt{k_{\ell-1}}} + \kappa_*^\ell \xi_\ell, \tag{12}$$

where $\xi_\ell \sim \mathcal{N}(0_{k_\ell}, I_{k_\ell})$ is a Gaussian noise term.

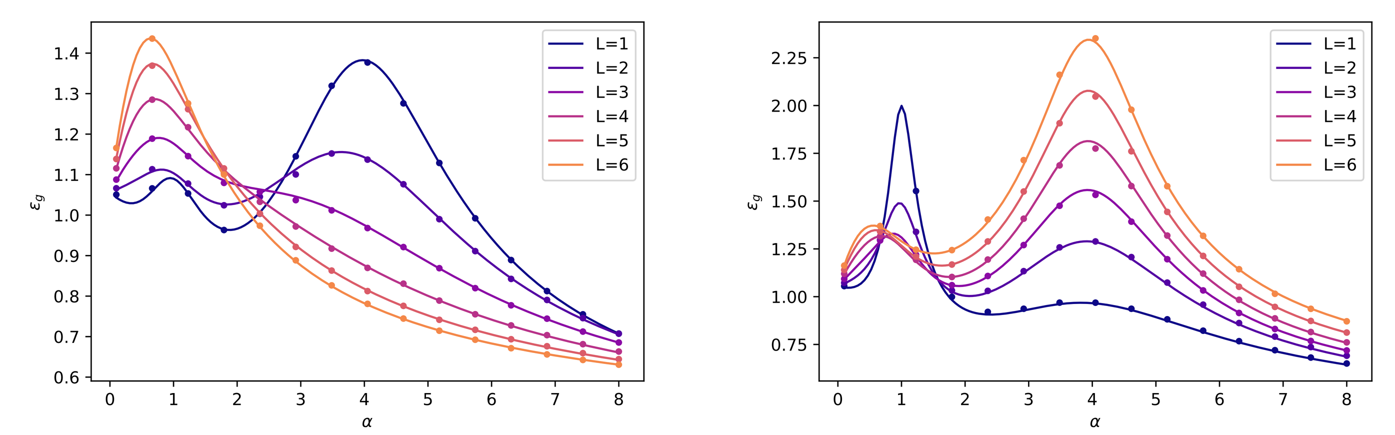

Figure 1: Learning curves for ridge regression on a 1-hidden layer target function ($\gamma_1^\star = 2$, $\sigma_1^\star = \mathrm{sign}$) using a $L-$hidden layers learner with widths $\gamma_1 = \dots = \gamma_L = 4$ and $\sigma_{1,\dots,L} = \tanh$ activation (left) or $\sigma_{1,\dots,L}(x) = 1.1 \times \mathrm{sign}(x) \times \min(2, |x|)$ clipped linear activation (right), for depths $1 \le L \le 6$. The regularization is $\lambda = 0.001$. *Solid lines* represent theoretical curves, while numerical simulations are indicated by *dots*. Two peaks, linear and non-linear, appear at $\alpha = {}^n\!/_d = 1$ and $\alpha = \gamma = 4$ respectively.

There exists an interplay between the two peaks, with higher noise ξ_L both helping to mitigate the linear peak, and aggravating the non-linear peak. The depth of the network plays a role in that it modulates the amplitudes of the signal part and the noise part.

References

[1] C. Chouard. Quantitative deterministic equivalent of sample covariance matrices with a general dependence structure. arXiv preprint arXiv:2211.13044. 2022.
[2] Z. Fan, Z. Wang. Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks. NeurIPS 2020.
[3] E. Dobriban, S. Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. The Annals of Statistics. 2018.
[4] F. Gerace, B. Loureiro, F. Krzakala, M. Mézard, L. Zdeborová. Generalisation error in learning with random features and the hidden manifold model. ICML 2020.