STAT 946 — Deep Learning Theory

Lecture 9: Random Projections, Effective Regularization, and Double Descent

University of Waterloo

Date: October 1

Lecturer: Mufan Li **Scribe:** ChatGPT

Reference. These notes follow Francis Bach's paper *High-dimensional analysis of double descent for linear regression with random projections* [1].

1 Setting and key random matrix objects

We consider a random design linear model

$$y = X\theta_\star + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n), \tag{1}$$

where $X \in \mathbb{R}^{n\times d}$ has i.i.d. mean-zero, unit-variance sub-Gaussian entries (often taken as standard Gaussian). Let

$$\widehat{\Sigma} \coloneqq \frac{1}{n} X^\top X \in \mathbb{R}^{d\times d}, \qquad K \coloneqq \frac{1}{n} X X^\top \in \mathbb{R}^{n\times n}. \tag{2}$$

Both matrices share the same nonzero eigenvalues, but they may have different numbers of zero eigenvalues when $d \neq n$.

Stieltjes transform / resolvent. For $z \in \mathbb{C} \setminus \mathbb{R}_+$, define

$$\widehat{\varphi}(z) \coloneqq \frac{1}{n} \operatorname{Tr}\big[(K - z\mathbf{I}_n)^{-1}\big]. \tag{3}$$

In the high-dimensional regime $n, d \to \infty$ with $d/n \to \gamma > 0$, $\widehat{\varphi}(z)$ converges to a deterministic limit

$$\varphi(z) = \int \frac{1}{x - z}\, \rho(\mathrm{d}x), \tag{4}$$

the Stieltjes transform of the limiting spectral measure ρ of K.

2 Asymptotic equivalence and "effective regularization"

One key message from random matrix theory is that, for ridge-type spectral functionals, a random-design empirical covariance behaves like the population covariance, but with an *inflated* regularization parameter.

Proposition 1 (Ridge degrees of freedom: asymptotic equivalence)**.** *Let $\Sigma \coloneqq \mathbb{E}[xx^\top]$ denote the population covariance, and let $\widehat{\Sigma} = \frac{1}{n} X^\top X$. For $\lambda > 0$ and in the high-dimensional limit, we have the asymptotic equivalence*

$$\operatorname{Tr}\Big(\widehat{\Sigma}\,(\widehat{\Sigma} + \lambda \mathbf{I}_d)^{-1}\Big) \;\sim\; \operatorname{Tr}\Big(\Sigma\,(\Sigma + \kappa(\lambda)\mathbf{I}_d)^{-1}\Big), \tag{5}$$

where the effective regularization $\kappa(\lambda)$ *is given by*

$$\kappa(\lambda) = \frac{1}{\varphi(-\lambda)}. \tag{6}$$

Remark 1 (Interpretation)**.** *Even if we fit with ridge parameter λ, the random design induces an additional shift from λ to $\kappa(\lambda)$. In over-parameterized regimes, $\kappa(\lambda)$ can stay bounded away from 0 even as $\lambda \downarrow 0$ ("self-induced" or "implicit" regularization).*

2.1 Self-consistency equation in the isotropic case

To see where $\kappa(\lambda) = 1/\varphi(-\lambda)$ comes from, it is helpful to look at the isotropic setting $\Sigma = \mathbf{I}_d$. Let $\gamma = d/n$. The limiting Stieltjes transform $\varphi(z)$ satisfies the self-consistency equation

$$\frac{1}{\varphi(z)} + z = \frac{\gamma}{1+\varphi(z)}. \tag{7}$$

Equivalently, φ solves the quadratic

$$z\varphi(z)^2 - (\gamma - 1 - z)\varphi(z) + 1 = 0, \tag{8}$$

with the branch chosen so that $\mathrm{Im}(\varphi(z)) > 0$ when $\mathrm{Im}(z) > 0$.

Sketch of the trace manipulations. Set $z = -\lambda$ and note the algebraic identity

$$\widehat{\Sigma}(\widehat{\Sigma} - z\mathbf{I}_d)^{-1} = \mathbf{I}_d + z(\widehat{\Sigma} - z\mathbf{I}_d)^{-1}. \tag{9}$$

Taking traces gives

$$\mathrm{Tr}\big(\widehat{\Sigma}(\widehat{\Sigma} - z\mathbf{I}_d)^{-1}\big) = d + z\,\mathrm{Tr}\big((\widehat{\Sigma} - z\mathbf{I}_d)^{-1}\big). \tag{10}$$

Next relate $\mathrm{Tr}((\widehat{\Sigma} - z\mathbf{I}_d)^{-1})$ (a $d \times d$ resolvent) to $\mathrm{Tr}((K - z\mathbf{I}_n)^{-1})$ (an $n \times n$ resolvent). Because $\widehat{\Sigma}$ and K share the same nonzero eigenvalues and only differ in their number of zeros, one gets the correction

$$\mathrm{Tr}\big((\widehat{\Sigma} - z\mathbf{I}_d)^{-1}\big) = n\,\widehat{\varphi}(z) + \frac{n-d}{z}. \tag{11}$$

Plugging back and simplifying yields

$$\mathrm{Tr}\big(\widehat{\Sigma}(\widehat{\Sigma} - z\mathbf{I}_d)^{-1}\big) = n\big(1 + z\widehat{\varphi}(z)\big). \tag{12}$$

Random matrix theory then shows that the same quantity is asymptotically equivalent to $\mathrm{Tr}\big(\mathbf{I}_d(\mathbf{I}_d + \kappa\mathbf{I}_d)^{-1}\big) = d/(1+\kappa)$ with $\kappa = 1/\varphi(z)$, which is precisely Eq. (7).

3 Random projections as a (linear) one-hidden-layer network

Now introduce a random projection matrix (random first-layer weights)

$$S \in \mathbb{R}^{d\times m}, \qquad S_{ij} \text{ i.i.d. mean 0, variance 1.} \tag{13}$$

Given S, we fit a linear model in the projected feature space $x \mapsto S^\top x$:

$$\widehat{y} = XS\widehat{\eta}, \tag{14}$$

where $\widehat{\eta} \in \mathbb{R}^m$ are the trainable second-layer weights. We study the minimum-norm interpolating solution (a ridge limit):

$$\widehat{\eta} := \lim_{\lambda\downarrow 0} \arg\min_{\eta\in\mathbb{R}^m} \|y - XS\eta\|_2^2 + \lambda\|\eta\|_2^2 = ((XS)^\top(XS))^\dagger (XS)^\top y, \tag{15}$$

and map back to the original parameter space via

$$\widehat{\theta} = S\widehat{\eta} \in \mathbb{R}^d. \tag{16}$$

This is a linearized "two-layer" model: the first layer is random and fixed, and the second layer is trained.

3.1 Risk decomposition

In the isotropic case ($\Sigma = \mathbf{I}_d$), we consider the squared parameter error

$$R(\widehat{\theta}) = \|\widehat{\theta} - \theta_\star\|_2^2. \tag{17}$$

Conditioned on X and S, the expectation over noise decomposes into variance and squared bias:

$$R^{(\mathrm{var})}(\widehat{\theta}) \coloneqq \mathbb{E}_\varepsilon\big[R(\widehat{\theta}) \;\big|\; X, S, \theta_\star = 0\big], \tag{18}$$

$$R^{(\mathrm{bias})}(\widehat{\theta}) \coloneqq R(\widehat{\theta}) \text{ when } \sigma = 0. \tag{19}$$

4 High-dimensional limits and double descent (isotropic case)

Let

$$\gamma \coloneqq \frac{d}{n}, \qquad \delta \coloneqq \frac{m}{n}, \tag{20}$$

and take $n, d, m \to \infty$ with γ, δ converging to positive constants. The curve $\delta \mapsto \mathbb{E}_\varepsilon R(\widehat{\theta})$ exhibits *double descent* as m crosses n (i.e., $\delta = 1$).

4.1 Under-parameterized regime: $m < n$ ($\delta < 1$)

In this regime, $XS \in \mathbb{R}^{n \times m}$ has full column rank with high probability, and the variance term is the familiar Wishart expression

$$\mathbb{E}_\varepsilon R^{(\mathrm{var})}(\widehat{\theta}) \;\sim\; \sigma^2 \, \frac{m}{n - m} = \sigma^2 \, \frac{\delta}{1 - \delta}. \tag{21}$$

The bias term depends on the alignment of $\theta_\star$ with the data covariance; in the simplest isotropic model it simplifies to

$$R^{(\mathrm{bias})}(\widehat{\theta}) \;\sim\; \frac{\gamma - \delta}{\gamma(1 - \delta)} \, \|\theta_\star\|_2^2. \tag{22}$$

Both terms blow up as $\delta \uparrow 1$.

4.2 Over-parameterized regime: $m > n$ and $d > n$ ($\delta > 1$, $\gamma > 1$)

When $m > n$, there are infinitely many interpolating solutions in the feature space; the minimum-norm choice acts like an implicit regularizer. In the isotropic setting, the high-dimensional limits simplify to

$$\mathbb{E}_\varepsilon R^{(\mathrm{var})}(\widehat{\theta}) \;\sim\; \sigma^2 \left(\frac{1}{\gamma - 1} + \frac{1}{\delta - 1} \right), \tag{23}$$

$$R^{(\mathrm{bias})}(\widehat{\theta}) \;\sim\; \left(1 - \frac{1}{\gamma}\right) \frac{\delta}{\delta - 1} \, \|\theta_\star\|_2^2. \tag{24}$$

As $\delta \downarrow 1$ both terms diverge; as $\delta \to \infty$, the variance approaches $\sigma^2/(\gamma - 1)$ and the bias approaches $(1 - 1/\gamma)\|\theta_\star\|_2^2$.

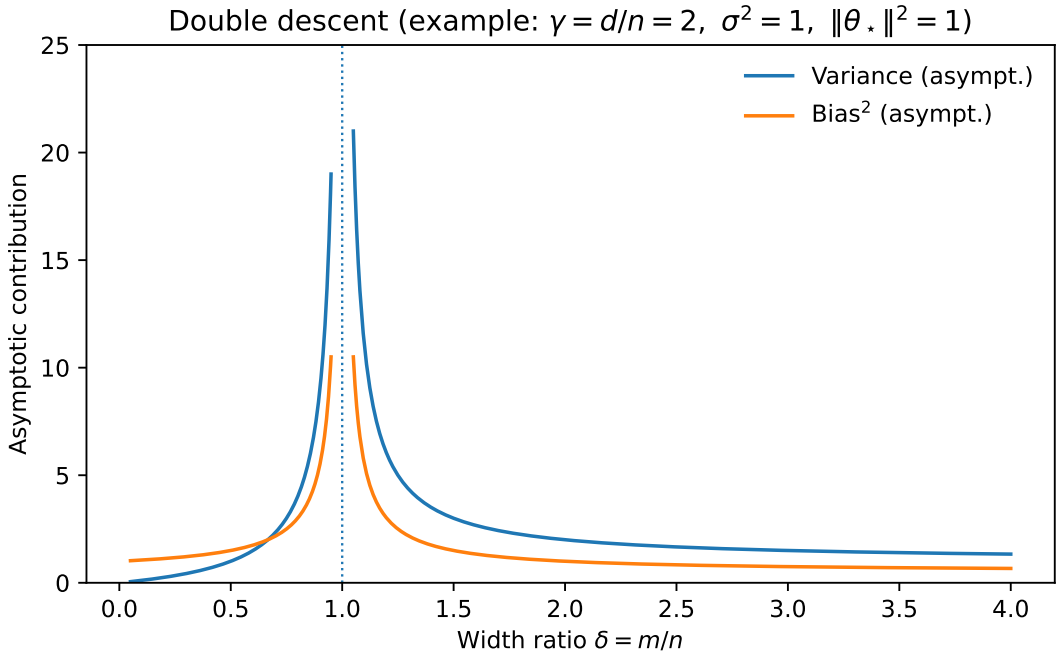

Figure 1: Illustration of the *double descent* phenomenon in the asymptotic equivalents for the variance and squared-bias contributions as functions of the width ratio $\delta = m/n$ (vertical dotted line: $\delta = 1$). The plotted curves use the piecewise limits stated in the text. In this example, we set $\gamma = d/n = 2$, $\sigma^2 = 1$, and $\|\theta_\star\|_2^2 = 1$ (other choices simply rescale the curves).

5 Takeaway: effective regularization in the interpolating regime

The random matrix equivalences can be summarized as follows. In ridge-type trace functionals, the random feature map behaves like a population covariance with a shifted regularization parameter. In particular, in isotropic settings, for the random-projection covariance one obtains the limiting behavior

$$\lim_{\lambda \downarrow 0} \kappa(\lambda) = \begin{cases} 0, & \delta \le 1, \\ \delta - 1, & \delta > 1, \end{cases} \tag{25}$$

which highlights that over-parameterization ($m > n$) induces a nontrivial effective regularization even when the explicit ridge penalty vanishes.

Next time. Feature learning perspectives (mean field / neural tangent limits).

References

[1] Francis Bach. *High-dimensional analysis of double descent for linear regression with random projections.* arXiv:2303.01372, 2023. https://arxiv.org/abs/2303.01372.