More Data Can Hurt for Linear Regression: Sample-wise Double Descent

Paper by Preetum Nakkiran

Presentation by Gavin Brown

8/29/24

Setting:

Covariates: $x_i \sim \mathcal{N}(0, \mathbb{I}_d)$
True parameter: β satisfies $\|\beta\|_2 = 1$
Labels: $y_i \leftarrow \langle x_i , \beta \rangle + \mathcal{N}(0, \sigma^2)$

d dimensions, n examples

Estimator:

underparameterized

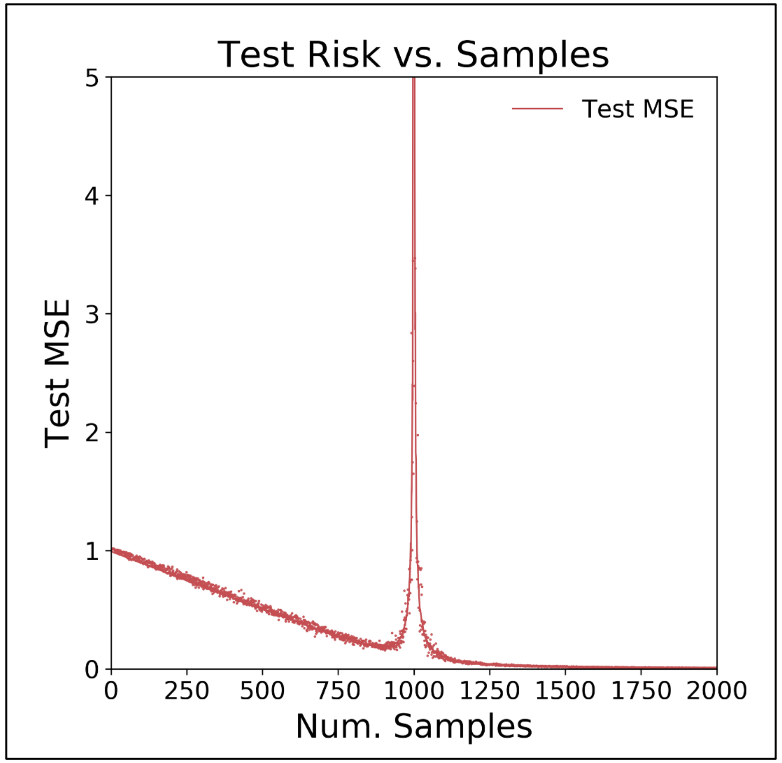
When $n \geq d$ $\hat{\beta} = (X^T X)^{-1} X^T y$
When $n < d$ $\hat{\beta} = X^\dagger y$

overparameterized

pseudoinverse
$X^\dagger = X^T (XX^T)^{-1}$

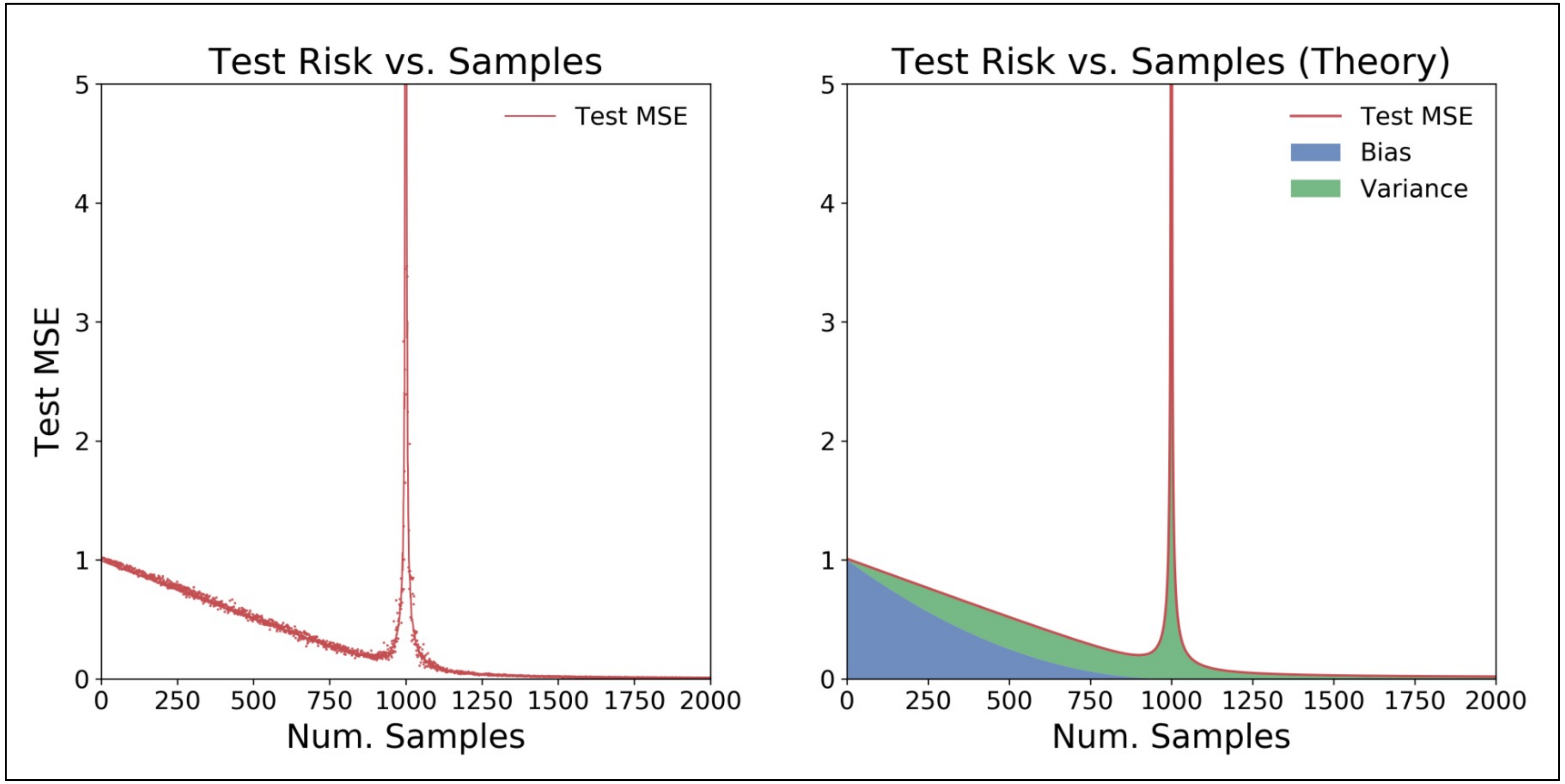
What happens?

Take $d = 1000, \sigma = 0.1$,
increase n



"We consider it extremely unsatisfying that the most popular technique in modern machine learning (training an overparameterized neural network with SGD) can be nonmonotonic in samples."

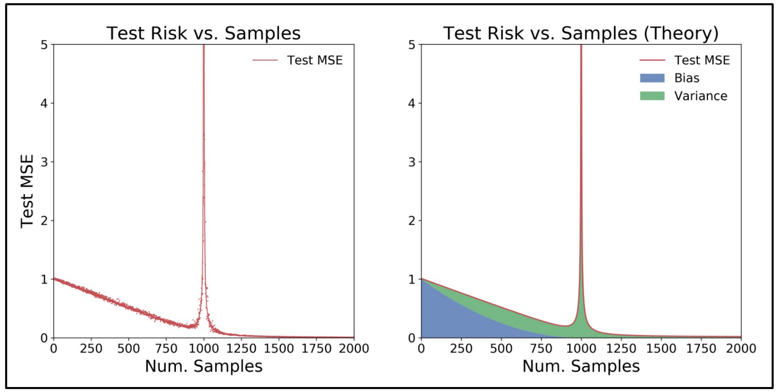
Variance is not monotone decreasing



Key Idea:

- when $n \ll d$ we have many interpolators: min-norm is “good” inductive bias
- when $n \approx d$ we have few interpolators: all have high norm

Analyzing Bias and Variance*



Test MSE
$$R(\hat{\beta}) := \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(\langle x, \hat{\beta}\rangle - y)^2]$$
$$= ||\hat{\beta} - \beta||^2 + \sigma^2$$

Excess Risk:
$$\overline{R}(\hat{\beta}) := ||\hat{\beta} - \beta||^2$$

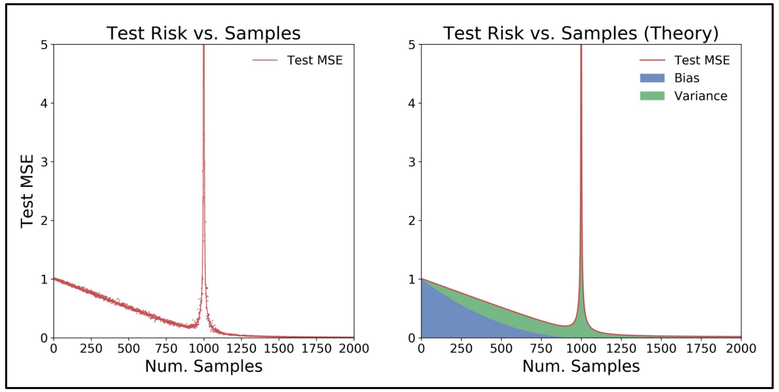
In expectation:
$$\mathop{\mathbb{E}}_{X,y}[\overline{R}(\hat{\beta}_{X,y})] = \mathop{\mathbb{E}}_{X,y}[||\hat{\beta} - \beta||^2] = \underbrace{||\beta - \mathbb{E}[\hat{\beta}]||^2}_{\text{Bias } B_n} + \underbrace{\mathbb{E}[||\hat{\beta} - \mathbb{E}[\hat{\beta}]||^2]}_{\text{Variance } V_n}$$

*This paper's analysis is similar to :

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. (2019). Surprises in high-dimensional ridgeless least squares interpolation.

Mei, S. and Montanari, A. (2019). The generalization error of random features regression: Precise asymptotics and double descent curve.

Expressions for Bias and Variance



$$\hat{\beta} = X^{\dagger} y = \begin{cases} X^T (XX^T)^{-1} y & \text{when } n \le d \\ (X^T X)^{-1} X^T y & \text{when } n > d \end{cases}$$

Lemma 1. *For $n \le d$, the bias and variance of the estimator $\hat{\beta} = X^{\dagger} y$ is*

$$B_n = \Big|\Big| \mathop{\mathbb{E}}_{X \sim \mathcal{D}^n} [Proj_{X^{\perp}}(\beta)] \Big|\Big|^2$$

$$V_n = \underbrace{\mathop{\mathbb{E}}_{X}[||Proj_X(\beta) - \mathop{\mathbb{E}}_{X}[Proj_X(\beta)]||^2]}_{(A)} + \sigma^2 \underbrace{\mathop{\mathbb{E}}_{X}[\mathrm{Tr}\big((XX^T)^{-1}\big)]}_{(B)}$$

this term blows up

For matrix X, projection onto rowspace:

$$Proj_X(\beta) = X^T (XX^T)^{-1} X \, \beta$$

Projection onto orthogonal complement:

$$Proj_{X^{\perp}}(\beta) = (\mathbb{I} - X^T (XX^T)^{-1} X) \, \beta$$

Aside: compare with *hat matrix*

$$H = X(X^T X)^{-1} X^T$$

projection onto columnspace

Intuition for high variance

Lemma 1. *For $n \le d$, the bias and variance of the estimator $\hat{\beta} = X^{\dagger} y$ is*

$$B_n = \|\mathop{\mathbb{E}}_{X \sim \mathcal{D}^n}[Proj_{X^{\perp}}(\beta)]\|^2$$

$$V_n = \underbrace{\mathop{\mathbb{E}}_{X}[\|Proj_X(\beta) - \mathop{\mathbb{E}}_{X}[Proj_X(\beta)]\|^2]}_{(A)} + \sigma^2 \underbrace{\mathop{\mathbb{E}}_{X}[\mathrm{Tr}((XX^T)^{-1})]}_{(B)}$$

B blows up because X becomes poorly conditioned as $n \to d$

Thought experiment: add one "good" sample to $n = d - 1$ "good" samples.

scaled basis vectors

$x_1, \dots, x_n = \sqrt{d} e_1, \dots, \sqrt{d} e_{d-1}$

write: $x_{n+1} = (g_1, g_2) \in \mathbb{R}^{d-1} \times \mathbb{R}$

$x_{n+1} \sim \mathcal{N}(0, \mathbb{I}_d)$

Old data matrix: $X = [\sqrt{d}\mathbb{I}_{d-1} \quad 0]$

New data matrix: $X_{n+1} = \begin{bmatrix} \sqrt{d}\mathbb{I}_{d-1} & 0 \\ g_1 & g_2 \end{bmatrix}$

<u>Claim 1:</u> X has no small, non-zero singular values.

<u>Claim 2:</u> X_{n+1} has a small non-zero singular value whp.

<u>Proof:</u> Let $v^T = [g_1 \quad -\sqrt{d}]$ $\quad \|v\|^2 \approx 2d$

But

$$v^T X_{n+1} = [g_1 \quad -\sqrt{d}] \begin{bmatrix} \sqrt{d}\,\mathbb{I}_{d-1} & 0 \\ g_1 & g_2 \end{bmatrix} = [0 \quad -\sqrt{d} g_2]$$

So $\|v^T X_{n+1}\|^2 \approx d \approx \frac{1}{2}\|v\|^2$ $\square$

Appendix

Proof of Lemma 1, Bias

$$\hat{\beta} = X^{\dagger}y = \begin{cases} X^T(XX^T)^{-1}y & \text{when } n \le d \\ (X^TX)^{-1}X^Ty & \text{when } n > d \end{cases}$$

Lemma 1. *For $n \le d$, the bias and variance of the estimator $\hat{\beta} = X^{\dagger}y$ is*

$$B_n = ||\mathop{\mathbb{E}}_{X\sim\mathcal{D}^n}[Proj_{X^\perp}(\beta)]||^2$$
$$V_n = \underbrace{\mathop{\mathbb{E}}_{X}[||Proj_X(\beta) - \mathop{\mathbb{E}}_{X}[Proj_X(\beta)]||^2]}_{(A)} + \sigma^2 \underbrace{\mathop{\mathbb{E}}_{X}[\mathrm{Tr}\big((XX^T)^{-1}\big)]}_{(B)}$$

Proof. **Bias.** Note that

$$\begin{aligned} \beta - \mathbb{E}[\hat{\beta}] &= \beta - \mathop{\mathbb{E}}_{X,\eta}[X^T(XX^T)^{-1}(X\beta + \eta)] \\ &= \mathop{\mathbb{E}}_{X}[(I - X^T(XX^T)^{-1}X)\beta] \\ &= \mathop{\mathbb{E}}_{X}[Proj_{X^\perp}(\beta)] \end{aligned}$$

Thus the bias is

$$\begin{aligned} B_n &= ||\beta - \mathbb{E}[\hat{\beta}]||^2 \\ &= ||E_{X_n}[Proj_{X_n^\perp}(\beta)]||^2 \end{aligned}$$

Proof of Lemma 1, Variance

$$\hat{\beta} = X^\dagger y = \begin{cases} X^T(XX^T)^{-1}y & \text{when } n \leq d \\ (X^TX)^{-1}X^Ty & \text{when } n > d \end{cases}$$

Lemma 1. *For $n \leq d$, the bias and variance of the estimator $\hat{\beta} = X^\dagger y$ is*

$$B_n = || \mathop{\mathbb{E}}_{X\sim\mathcal{D}^n}[Proj_{X^\perp}(\beta)]||^2$$

$$V_n = \underbrace{\mathop{\mathbb{E}}_{X}[||Proj_X(\beta) - \mathop{\mathbb{E}}_{X}[Proj_X(\beta)]||^2]}_{(A)} + \sigma^2 \underbrace{\mathop{\mathbb{E}}_{X}[\mathrm{Tr}\big((XX^T)^{-1}\big)]}_{(B)}$$

<u>Proof</u>

$$\begin{aligned} V_n &= \mathop{\mathbb{E}}_{\hat{\beta}}[||\hat{\beta} - \mathbb{E}[\hat{\beta}]||^2] \\ &= \mathop{\mathbb{E}}_{X,\eta}[||X^T(XX^T)^{-1}(X\beta+\eta) - \mathop{\mathbb{E}}_{X}[X^T(XX^T)^{-1}X\beta]||^2] \\ &= \mathop{\mathbb{E}}_{X,\eta}[||(S-\overline{S})\beta + X^T(XX^T)^{-1}\eta||^2] \qquad (S := X^T(XX^T)^{-1}X, \overline{S} := \mathbb{E}[S]) \\ &= \mathop{\mathbb{E}}_{X}[||(S-\overline{S})\beta||^2] + \mathop{\mathbb{E}}_{X,\eta}[||X^T(XX^T)^{-1}\eta||^2] \\ &= \mathop{\mathbb{E}}_{X}[||(S-\overline{S})\beta||^2] + \sigma^2 Tr((XX^T)^{-1}) \end{aligned}$$

Notice that S is projection onto the rowspace of X, i.e. $S = Proj_X$. Thus,

$$V_n := \mathop{\mathbb{E}}_{X}[||Proj_X(\beta) - \mathop{\mathbb{E}}_{X}[Proj_X(\beta)]||^2] + \sigma^2 Tr((XX^T)^{-1})$$

$\square$

Approximate Asymptotics for Bias and Variance

Claim 1 (Overparameterized Risk)**.** *Let* $\gamma := \frac{n}{d} < 1$ *be the underparameterization ratio. The bias and variance are:*

$$B_n = (1-\gamma)^2 ||\beta||^2 \tag{5}$$

$$V_n \approx \gamma(1-\gamma)||\beta||^2 + \sigma^2 \frac{\gamma}{1-\gamma} \tag{6}$$

And thus the expected excess risk for $\gamma < 1$ *is:*

$$\mathbb{E}[\overline{R}(\hat{\beta})] \approx (1-\gamma)||\beta||^2 + \sigma^2 \frac{\gamma}{1-\gamma} \tag{7}$$

$$= (1-\frac{n}{d})||\beta||^2 + \sigma^2 \frac{n}{d-n} \tag{8}$$

Claim 2 (Underparameterized Risk, [Hastie et al., 2019])**.** *Let* $\gamma := \frac{n}{d} > 1$ *be the underparameterization ratio. The bias and variance are:*

$$B_n = 0 \quad , \quad V_n \approx \frac{\sigma^2}{\gamma - 1}$$