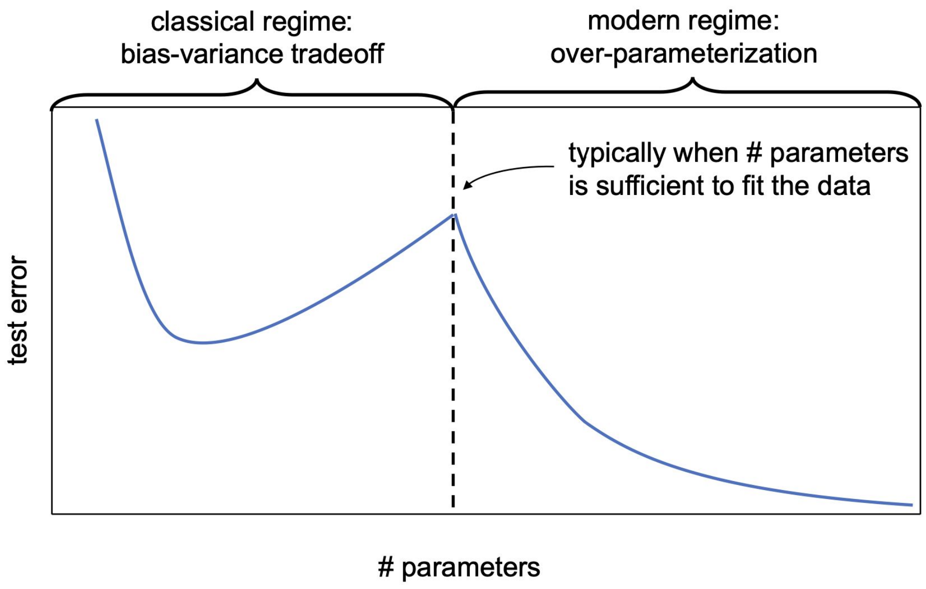
Double descent mitigated

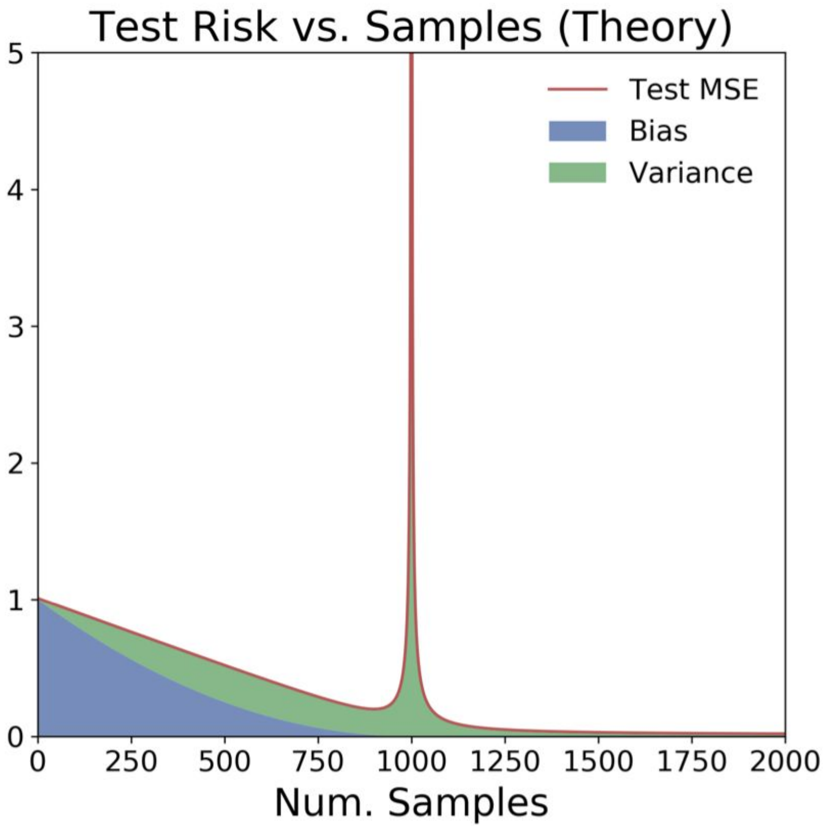
Artin Tajdini

Double descent

- Model size



- Sample size



- Epoch-wise

Regression setting

Ground truth: $\beta^* \in \mathbb{R}^d$

Samples (x_i , y_i) with noise $\epsilon \sim N(0, \sigma^2)$

$$x \sim \mathcal{N}(0, I_d) \qquad\qquad y := \langle x, \beta^* \rangle + \epsilon$$

Estimator: $\hat{\beta} := \text{argmin}_\beta \, ||X\beta - Y||^2 + \lambda ||\beta||^2$

Estimator risk:

$$R(\hat{\beta}) := \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(\langle x, \hat{\beta} \rangle - y)^2] = ||\hat{\beta} - \beta^*||_2^2 + \sigma^2$$

Optimal ridge

Expected risk for estimator function $\hat{\beta}_n(X, \vec{y})$

$$\overline{R}(\hat{\beta}_n) := \mathop{\mathbb{E}}_{X,y\sim\mathcal{D}^n}[R(\hat{\beta}_n(X, \vec{y}))]$$

Optimal ridge parameter: $\lambda_n^{\text{opt}} := \mathop{\text{argmin}}_{\lambda:\lambda\geq 0} \overline{R}(\hat{\beta}_{n,\lambda}))$ ($\lambda^* = \frac{d\sigma^2}{||\beta*||_2^2}$.)

Is regression in this setting with optimal ridge monotonic in sample/parameter size? Yes

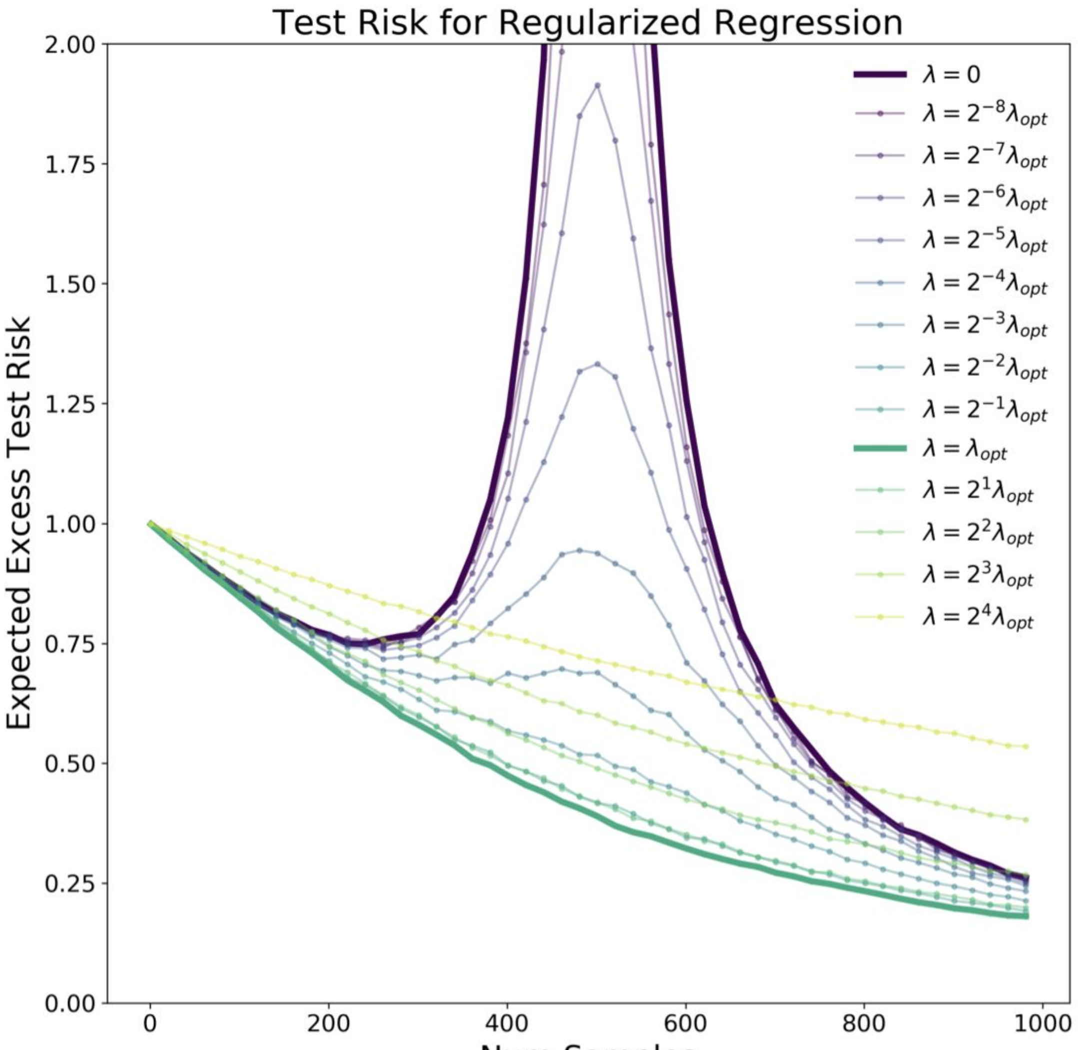
Sample size mitigation

Theorem 1: $\overline{R}(\hat{\beta}^{\text{opt}}_{n+1}) \leq \overline{R}(\hat{\beta}^{\text{opt}}_{n})$

for all n, d

Theorem 2 (over-regularized):

$$\forall \lambda \geq \lambda^* : \quad \overline{R}(\hat{\beta}_{n+1,\lambda}) \leq \overline{R}(\hat{\beta}_{n,\lambda})$$



Sample size mitigation

Proof of Theorem 1:

1. $\overline{R}(\hat{\beta}_{n,\lambda}) = \mathop{\mathbb{E}}_{(\gamma_1,\dots\gamma_d)\sim\Gamma_n}\left[\sum_{i=1}^{d} \frac{||\beta^*||_2^2\lambda^2/d + \sigma^2\gamma_i^2}{(\gamma_i^2+\lambda)^2}\right] + \sigma^2$ using SVD of X

2. $\overline{R}(\hat{\beta}_n^{\mathrm{opt}}) = \mathop{\mathbb{E}}_{(\gamma_1,\dots\gamma_d)\sim\Gamma_n}\left[\sum_{i=1}^{d} \frac{\sigma^2}{\gamma_i^2 + d\sigma^2/||\beta^*||_2^2}\right] + \sigma^2$

3. Cauchy interlacing theorem

Counterexample: non-Gaussian data distribution

$$(x, y) \sim \begin{cases} (\vec{e}_1, 1) & \text{w.p.} 1/2 \\ (\vec{e}_2, \pm A) & \text{w.p.} 1/2 \end{cases}$$ where A is uniform over [0, 10]

So $\beta^* = [1, 0]$

The first coordinate has optimal 0 ridge and second coordinate has optimal ∞ ridge

Theorem 4: with slight modification of above instance, we have

$$\overline{R}(\hat{\beta}^{\text{opt}}_{n=1}) < \overline{R}(\hat{\beta}^{\text{opt}}_{n=2})$$

Model size mitigation

Data comes from space of dim p. Project with random orthonormal matrix P to dim d

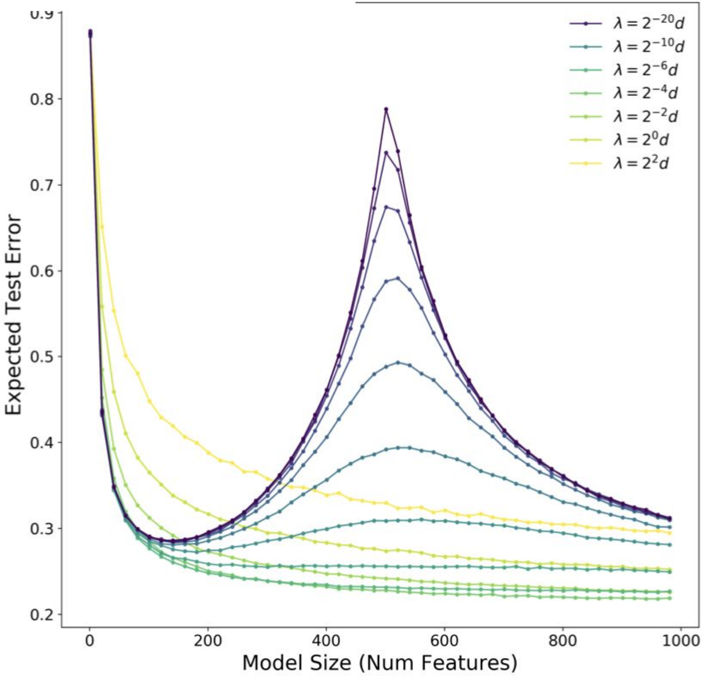
$$R_P(\hat{\beta}) := \mathop{\mathbb{E}}_{(\tilde{x},y)\sim\mathcal{D}}[(\langle \tilde{x}, \hat{\beta}\rangle - y)^2] = \mathop{\mathbb{E}}_{(x,y)}[(\langle Px, \hat{\beta}\rangle - y)^2]$$

Expected risk is now $\overline{R}(\hat{\beta}) := \mathop{\mathbb{E}}_{P} \mathop{\mathbb{E}}_{\tilde{X},\vec{y}\sim\mathcal{D}^n}[R_P(\hat{\beta}(\tilde{X},\vec{y})]$

where $\widetilde{X} = XP^T \in \mathbb{R}^{n\times d}$

Theorem 3: For all d ≤ p and n

$$\overline{R}(\hat{\beta}^{\text{opt}}_{d+1}) \leq \overline{R}(\hat{\beta}^{\text{opt}}_{d})$$



Model size mitigation

Proof:

1. $$\overline{R}(\hat{\beta}_{d,\lambda}) = \sigma^2 + (1 - \frac{d}{p})||\theta||_2^2 + \mathop{\mathbb{E}}_{(\gamma_1,\ldots,\gamma_m)\sim\Gamma_d}\left[\sum_{i=1}^{p} \frac{(\sigma^2 + \frac{p-d}{p}||\theta||_2^2)\gamma_i^2 + \frac{d}{p^2}||\theta||_2^2\lambda^2}{(\gamma_i^2+\lambda)^2}\right]$$

2. $\lambda_d^{\text{opt}} = \dfrac{p^2\widetilde{\sigma}^2}{d||\theta||_2^2}$ where $\widetilde{\sigma}^2 := \sigma^2 + \dfrac{p-d}{p}||\theta||_2^2$

3. Cauchy interlacing theorem

Epoch-wise mitigation(Stop early!)

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^{t} - \tfrac{1}{2}\mathrm{diag}(\boldsymbol{\eta})\nabla \hat{R}(\boldsymbol{\theta}^{t})$$

- Under Parameterized d << n: Risk at time t approximated by

$$\bar{R}(\tilde{\boldsymbol{\theta}}^t) := \sigma^2 + \sum_{i=1}^{d} \underbrace{\sigma_i^2(\theta_i^*)^2(1-\eta_i\sigma_i^2)^{2t} + \frac{\sigma^2}{n}(1-(1-\eta_i\sigma_i^2)^t)^2}_{U_i(t)},$$

- Over parameterized for two layer NN $f_{\mathbf{W},\mathbf{v}}(\mathbf{x}) = \frac{1}{\sqrt{k}}\mathrm{relu}(\mathbf{x}^T\mathbf{W})\mathbf{v}.$

Initialization: $[\mathbf{W}^0]_{i,j} \sim \mathcal{N}(0,\omega^2), \quad [\mathbf{v}^0]_i \sim \mathrm{Uniform}(\{-\nu,\nu\}).$

Theorem 2. *Let $\alpha > 0$ be the smallest eigenvalue of the Gram matrix $\mathbf{\Sigma}$, suppose that the network is sufficiently wide, i.e., $k \geq \Omega\left(\frac{n^{10}}{\alpha^{15}\min(\nu,\omega)}\right)$, and suppose the initialization scale parameters obey $\nu\omega \leq \alpha/\sqrt{32\log(2n/\delta)}$ and $\nu+\omega \leq 1$ for some $\delta \in (0,1)$. Then, with probability at least $1-\delta$, the risk of the network trained with gradient descent for t iterations is at most*

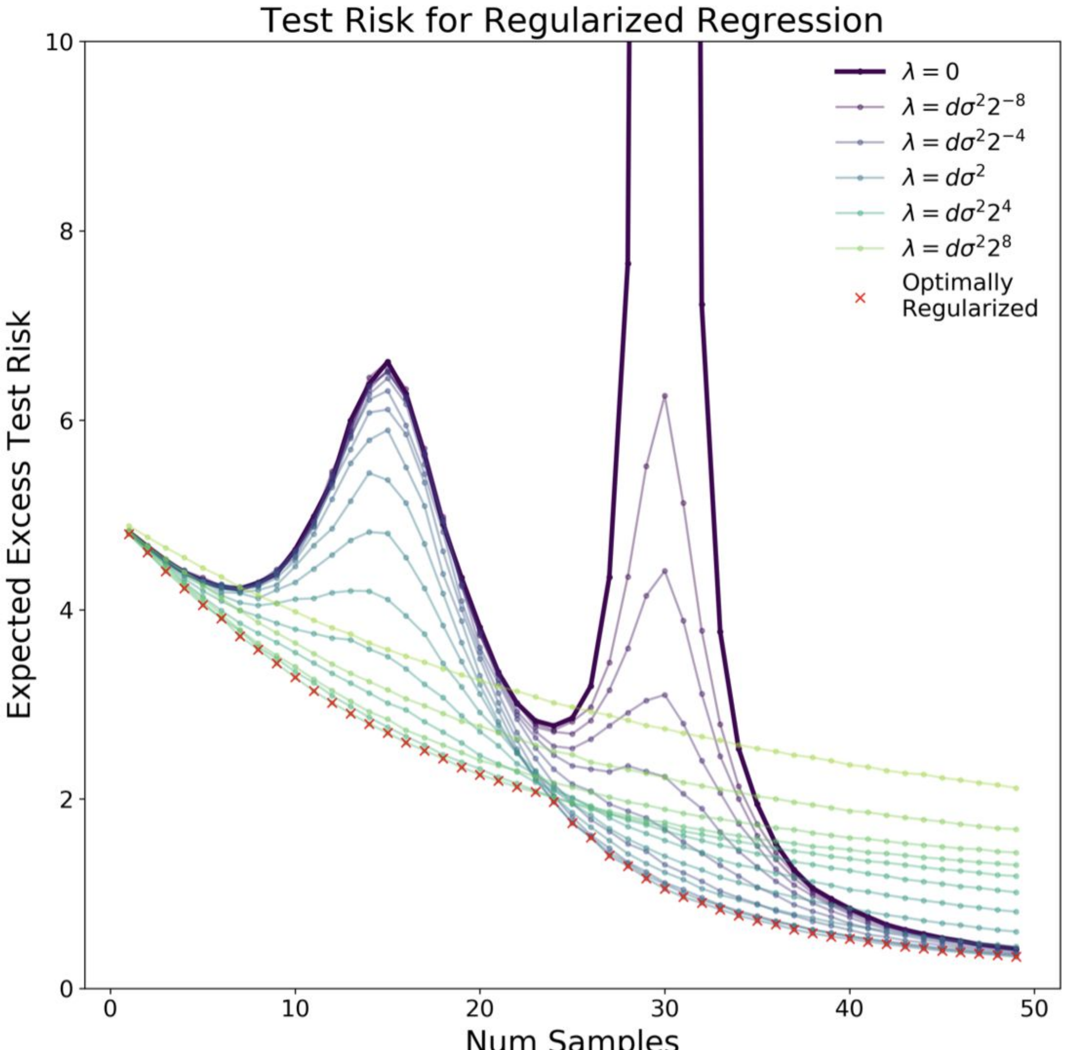
$$R(f_{\mathbf{W}_t,\mathbf{v}_t}) \leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}\langle\mathbf{u}_i,\mathbf{y}\rangle^2(1-\eta\sigma_i^2)^{2t}} + \sqrt{\frac{1}{n}\sum_{i=1}^{n}\langle\mathbf{u}_i,\mathbf{y}\rangle^2\frac{(1-(1-\eta\sigma_i^2)^t)^2}{\sigma_i^2}} + O(\frac{1}{\sqrt{n}}).$$

Open Problems

Proof for non-isotropic covariates?

Multiple descents based on eigenspace of $\boldsymbol{\Sigma}$

Nonlinear models?



Some References

[1] P. Nakkiran, P. Venkat, S. Kakade and T. Ma, “Optimal Regularization Can Mitigate Double Descent” in International Conference on Learning Representations, 2021.

[4] R. Heckel and F. Yılmaz, “Early Stopping in Deep Networks: Double Descent and How to Eliminate it” in International Conference on Learning Representations, 2021.

[2] T. Viering and M. Loog, "The Shape of Learning Curves: A Review" in IEEE Transactions on Pattern Analysis & Machine Intelligence, 2023.

[3] M.Loogetal et. al., “A Brief Prehistory Of Double Descent,” in Proceedings of the National Academy of Sciences, 2020