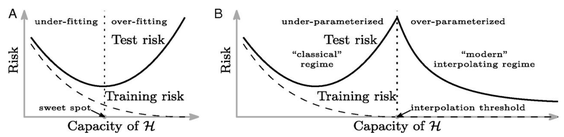
Dimensionality reduction, regularization, and generalization in overparameterized regressions

Ningyuan (Teresa) Huang, David W. Hogg, Soledad Villar

Johns Hopkins Mathematical Institute for Data Science

Simons Foundation

Introduction: double-descent



- **Generalization** behavior of large-capacity models.
- Ordinary least squares (OLS) min-norm estimator:
  - **(1) Peaking**: the risk (expected out-of-sample prediction error) can grow arbitrarily when the number of parameters **p** approaches the number of samples **n**;
  - **(2) Benefit of overparameterization**: The risk decreases with **p** for p > n, sometimes achieving a lower value than the lowest risk for p < n.
- **Regularization**:
  - ✓ Theory: ridge regression avoids the peaking (Hastie et al. 2020);
  - ✓ In practice: deep neural networks are trained with weight decay, dropouts, early stopping, ensemble, etc.

Research questions

- **RQ1: Can we avoid the peaking phenomenon using dimensionality reduction?**
  - Perform PCA on input features before fitting regression (principal component regression, or PCA-OLS);
  - Improve robustness against data-poisoning attacks.
- **RQ2: Is overparameterization necessary for good generalization?**
  - Compare PCA-OLS with other projection methods., some of which can overparameterize.

Problem setup

- Linear model and Gaussian data:

$$(x_i, \varepsilon_i) \sim P_x \times P_\varepsilon = \mathcal{N}(0_p, C_{xx}) \times \mathcal{N}(0, \sigma^2)$$

$$y_i = x_i^\top \beta + \varepsilon_i$$

- Training data: $(x_i, y_i)_{i=1}^n \in \mathbb{R}^p \times \mathbb{R}$
- Test risk (conditional on training data):

$$\mathcal{R}(\hat{\beta}|X) = \mathbb{E}_{x_* \sim P_x}((x_*^\top \hat{\beta} - x_*^\top \beta)^2 | X)$$

- OLS min-norm estimator and its risk:

$$\hat{\beta} = \arg\min_\beta \|X\beta - Y\|_2 = X^\dagger Y \qquad Y \in \mathbb{R}^{n\times 1},\ X \in \mathbb{R}^{n\times p}$$

$$\mathcal{R}(\hat{\beta}|X) = \beta^\top \Pi_{X_\perp} C_{xx} \Pi_{X_\perp} \beta + \frac{\sigma^2}{n} \operatorname{tr}\left((\frac{1}{n} X^\top X)^\dagger C_{xx} \right) + \sigma^2$$

- PCA-OLS risk:

$$\mathcal{R}(\hat{\beta}_{\mathrm{P}}|X) = \underbrace{\beta^\top \Pi_{X_{\mathrm{PCA}\perp}} C_{xx} \Pi_{X_{\mathrm{PCA}\perp}} \beta}_{\text{bias squared}} + \underbrace{\frac{\sigma^2}{n} \operatorname{tr}\left((\frac{1}{n} X_{\mathrm{PCA}}^\top X_{\mathrm{PCA}})^\dagger C_{xx} \right)}_{\text{variance}} + \sigma^2$$

Main Results I: PCA-OLS avoids the peaking phenomenon

Theorem: non-asymptotic risk bound for PCA-OLS

Let $x_i \sim \mathcal{N}(0_p, C_{xx})$ $i = 1, \ldots, n$, *and*

$$y_i = x_i^\top \beta + \epsilon$$

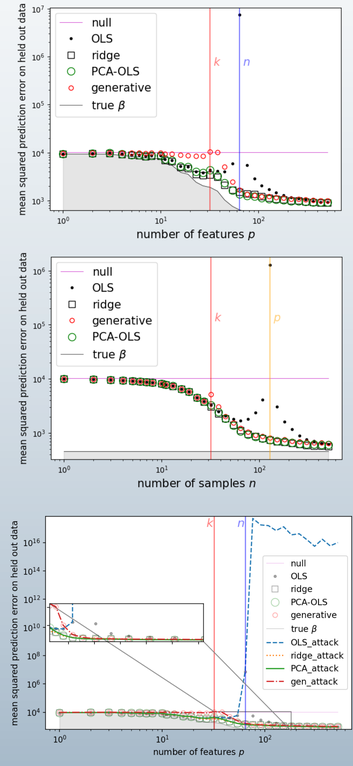
where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. *Let* c, t *be some constants,* λ_1 *be the largest eigenvalue of* C_{xx}, *and* $r_0(C_{xx}) := \frac{\operatorname{tr}(C_{xx})}{\lambda_1}$ *be the effective rank. Let* $M = c\lambda_1 \max\left\{ \sqrt{\frac{r_0(C_{xx})}{n}}, \frac{r_0(C_{xx})}{n}, \frac{t}{n} \right\}$, *and assume* $M < \lambda_k$. *Then with probability greater than* $1 - e^{-t}$ *we have*

$$\mathcal{R}(\hat{\beta}_{PCA\text{-}OLS-k} \mid X) = \mathbb{B} + \mathbb{V} + \sigma^2, \tag{1}$$

$$\lambda_p \|\Pi_{X_{PCA\perp}} \beta\|^2 \le \mathbb{B} \le \|\beta\|^2 \left(M + \lambda_{k+1} \right), \tag{2}$$

$$\frac{\sigma^2}{n} \frac{k\lambda_p}{\lambda_1 + M} \le \mathbb{V} \le \frac{\sigma^2}{n} \frac{k\lambda_1}{\lambda_k - M}, \tag{3}$$

where $\|\cdot\|$ *is the 2-norm for vectors, and* k *denotes the rank-*k *PCA with* $k < \min\{n, p\}$.



Consequences: robustness against data-poisoning attacks

Attacker: add a noisy pair to the training data set.

$$\tilde{X} = \begin{bmatrix} X \\ x_0 \end{bmatrix} \in \mathbb{R}^{(n+1)\times p}, \quad \tilde{Y} = \begin{bmatrix} Y \\ y_0 \end{bmatrix} \in \mathbb{R}^{n+1}$$

Attack: $\max_{\|x_0\| \le \epsilon, \|y_0\| \le \epsilon} \|\tilde{Y} - \tilde{X}\beta\|^2$

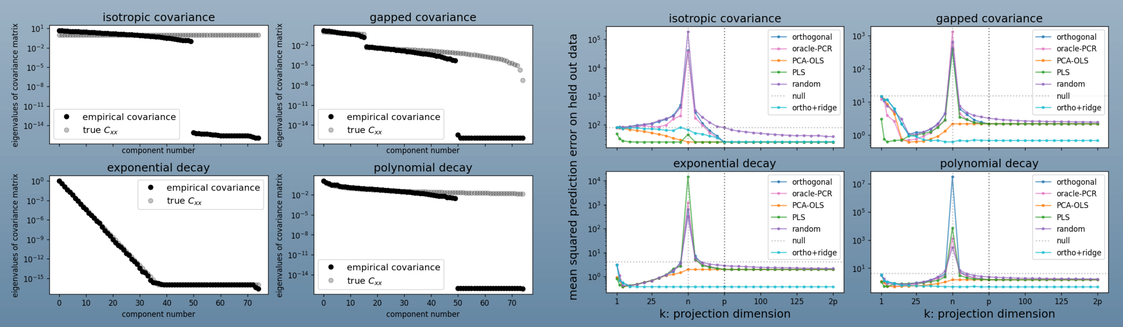
$$\hat{\beta}_{poison} = \tilde{X}^\top (\tilde{X}\tilde{X}^\top)^{-1} \tilde{Y} = \tilde{X}^\top \begin{bmatrix} XX^\top & Xx_0 \\ x_0^\top X^\top & x_0^\top x_0 \end{bmatrix}^{-1} \tilde{Y}$$

Proposition In the overparameterized regime, OLS is arbitrarily sensitive to data-poisoning attacks. The risk tends to infinity when the additional adversary feature x_0 is arbitrarily close to the column space of X.

Proposition In the underparameterized regime, OLS is arbitrarily sensitive to data-poisoning attack if the smallest eigenvalue of $X^\top X$ is smaller than the attack strength ϵ.

Corollary If the k-th largest eigenvalue of $X^\top X$ is much greater than ϵ, then PCA-OLS is robust to data-poisoning attack.

Main Results II: Overparameterization is not necessary for good generalization



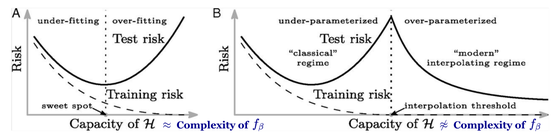
Framework: projection methods (recast as 2-layer neural network):

- 1st layer: projection (not trained);
- 2nd layer: regression (trained).

Without regularization, overparameterized projection methods (i.e., projection matrix chosen independent of training data) do not outperform PCA-OLS.

First layer projection	$\Pi \in \mathbb{R}^{p\times k}$	Reference
Random Gaussian	$\Pi = [w_1, \cdots, w_k]$, where $w_i \overset{i.i.d}{\sim} \mathcal{N}(0, p^{-1} I_p)$	Ba et al. (2020)
Random orthogonal	$\Pi^\top \Pi = I_k$	Lin and Dobriban (2021)
Oracle-PCR	$[I_k, 0_{k,p-k}]^\top$ (first k cols)	Xu and Hsu (2019)
Partial least squares	$[X^\top Y, (X^\top X) X^\top Y, \cdots, (X^\top X)^{k-1} X^\top Y]$	Cook and Forzani (2020)
PCA-OLS	$[\tilde{u}_1, \tilde{u}_2, \cdots, \tilde{u}_k]$ (first k PCs)	Huang et al. (2021)

Rethinking double-descent



- **What is the correct metric on the x-axis?**
  - Capacity of the hypothesis class ≠ complexity of the learned predictor;
- Existing generalization metrics depend on:
  - Hypothesis class;
  - Optimization landscape;
  - Stability and robustness.
- **Open problem**: metric of learned predictor that works for
  - ✓ Parametric model: linear regression, deep neural networks;
  - ✓ Nonparametric model: random forest, nearest neighbors.

Summary

- Regularization via dimensionality reduction:
  - ✓ Removes peaking phenomenon at the interpolation threshold;
  - ✓ Improves robustness against data poisoning attacks.
- Overparameterization is not necessary for good generalization:
  - Double-descent curves are more meaningful when plotted across different methods.
- Overparameterization + Regularization can work well:
  - Explicit regularization: dimensionality reduction, ridge penalty;
  - Implicit regularization: induced from optimization.

References

- N. Huang, D. W. Hogg, S. Villar. *Dimensionality reduction, regularization, and generalization in overparameterized regressions*, arXiv:2011.11477.
- M. Belkin, D. Hsu, S. Ma, S. Mandal. *Reconciling modern machine-learning practice and the classical bias–variance trade-off*, PNAS 2019.
- T. Hastie, A. Montanari, S. Rosset, R. J. Tibshirani, *Surprises in High-Dimensional Ridgeless Least Squares Interpolation*, arXiv:1903.08560.
- J. Xu, D. W. Hsu, *On the number of variables to use in principal component regression*, NeurIPS 2019.
- J. Ba et. al., *Generalization of Two-layer Neural Networks: An Asymptotic Viewpoint*, ICLR 2020
- C. Lin, E. Doriban, *What causes the test error? Going beyond bias-variance via ANOVA*, JMLR 2021
- R. D. Dennis, L. Forzani, *Partial least squares prediction in high-dimensional regression*, Ann. Stats. 2019

For more details, please refer to our paper here: