Benign Overfitting in Distributed Learning

Nicole Mücke (TU Braunschweig)
Enrico Reiss, Markus Klein, Jonas Rungenhagen (Univ. Potsdam)

SIMAI 2020+2021
MS 55 Mathematics of Machine Learning

What is DL ?

- large size of training datasets generally offers improvement in model performance, however the training process becomes computationally expensive and time consuming

  Example 1: Kernel Ridge Regression requires matrix inversion with costs $\mathcal{O}(n^3)$ in time and $\mathcal{O}(n^2)$ in memory [ZDW13, MB18]

What is DL ?

- large size of training datasets generally offers improvement in model performance, however the training process becomes computationally expensive and time consuming

  Example 1: Kernel Ridge Regression requires matrix inversion with costs $\mathcal{O}(n^3)$ in time and $\mathcal{O}(n^2)$ in memory [ZDW13, MB18]

  Example 2: training a state-of-the-art ResNet-50 model (in 90 epochs) on the ImageNet dataset with a Nvidia Tesla V100 GPU requires about two days [WWS+20]

What is DL ?

- large size of training datasets generally offers improvement in model performance, however the training process becomes computationally expensive and time consuming

  Example 1: Kernel Ridge Regression requires matrix inversion with costs $\mathcal{O}(n^3)$ in time and $\mathcal{O}(n^2)$ in memory [ZDW13, MB18]

  Example 2: training a state-of-the-art ResNet-50 model (in 90 epochs) on the ImageNet dataset with a Nvidia Tesla V100 GPU requires about two days [WWS+20]

- distributed learning (DL) is a very common strategy to reduce the overall training time by exploiting multiple computing devices
  $\longrightarrow$ datasets are partitioned over machines, which compute locally, and communicate short messages
- communication often the bottleneck
- here: focus on communication efficient methods

Random-Design Linear Regression in Hilbert Spaces

Model: $Y = \langle \beta^*, X \rangle + \varepsilon$

- random pair $(X, Y) \in \mathcal{H} \times \mathbb{R}$ with unknown joint distribution P
- noise $\varepsilon \in \mathbb{R}$ with $\mathbb{E}[\varepsilon | X] = 0$
- β^* minimizes prediction risk

$$\mathcal{R}(\beta^*) = \min_{\beta \in \mathcal{H}} \mathcal{R}(\beta) \,, \quad \mathcal{R}(\beta) := \mathbb{E}[(\langle \beta, X \rangle - Y)^2] \tag{1}$$

note: the optimal predictor f^* among all (measurable) functions is the regression function

$$f^*(x) = \mathbb{E}[Y | X = x]$$

Random-Design Linear Regression in Hilbert Spaces

Model: $Y = \langle \beta^*, X \rangle + \varepsilon$

- random pair $(X, Y) \in \mathcal{H} \times \mathbb{R}$ with unknown joint distribution P
- noise $\varepsilon \in \mathbb{R}$ with $\mathbb{E}[\varepsilon | X] = 0$
- β^* minimizes prediction risk

$$\mathcal{R}(\beta^*) = \min_{\beta \in \mathcal{H}} \mathcal{R}(\beta) \,, \quad \mathcal{R}(\beta) := \mathbb{E}[(\langle \beta, X \rangle - Y)^2] \tag{1}$$

note: the optimal predictor f^* among all (measurable) functions is the regression function

$$f^*(x) = \mathbb{E}[Y | X = x]$$

Assumption 1:

Our model is well-specified, i.e. $f^*(x) = \langle \beta^*, x \rangle$.

Distributed Regression

- Data: $D = D_1 \cup \ldots \cup D_M, \quad \text{size } |D_m| = \frac{n}{M}, \quad m = 1, \ldots, M$

$$D_m := \{(x_m^{(1)}, y_m^{(1)}), \ldots, (x_m^{(\frac{n}{M})}, y_m^{(\frac{n}{M})})\} \subset \mathcal{H} \times \mathbb{R}$$

  write as data matrix $\mathrm{X}_m \in \mathcal{L}(\mathcal{H}, \mathbb{R}^n)$, $\mathrm{Y}_m \in \mathbb{R}^{\frac{n}{M}}$

Distributed Regression

- Data: $D = D_1 \cup \ldots \cup D_M$, size $|D_m| = \frac{n}{M}$, $m = 1, \ldots, M$

$$D_m := \{(x_m^{(1)}, y_m^{(1)}), \ldots, (x_m^{(\frac{n}{M})}, y_m^{(\frac{n}{M})})\} \subset \mathcal{H} \times \mathbb{R}$$

  write as data matrix $\mathrm{X}_m \in \mathcal{L}(\mathcal{H}, \mathbb{R}^n)$, $\mathrm{Y}_m \in \mathbb{R}^{\frac{n}{M}}$

- local minimum norm interpolator: $\hat{\beta}_m$ solves

$$\min_{\beta \in \mathcal{H}} ||\beta|| \quad \text{such that} \quad ||\mathrm{X}_m \beta - \mathrm{Y}_m||^2 = \min_{\tilde{\beta} \in \mathcal{H}} ||\mathrm{X}_m \tilde{\beta} - \mathrm{Y}_m||^2$$

- equivalently: minimum norm solution to the normal equations (see [EHN96])

$$\hat{\beta}_m = \arg\min_{\beta} \{\beta \ : \ X_m^T X_m \beta = X_m^T \mathrm{Y}_m\} = \mathrm{X}_m^\dagger \mathrm{Y}_m$$

  note: if $X_m^T X_m$ is invertible, then $\mathrm{X}_m^\dagger = X_m^T (X_m^T X_m)^{-1}$

Distributed Regression

- Data: $D = D_1 \cup ... \cup D_M$, size $|D_m| = \frac{n}{M}$, $m = 1, ..., M$

$$D_m := \{(x_m^{(1)}, y_m^{(1)}), ..., (x_m^{(\frac{n}{M})}, y_m^{(\frac{n}{M})})\} \subset \mathcal{H} \times \mathbb{R}$$

  write as data matrix $\mathrm{X}_m \in \mathcal{L}(\mathcal{H}, \mathbb{R}^n)$, $\mathrm{Y}_m \in \mathbb{R}^{\frac{n}{M}}$

- local minimum norm interpolator: $\hat{\beta}_m$ solves

$$\min_{\beta \in \mathcal{H}} ||\beta|| \quad \text{such that} \quad ||\mathrm{X}_m \beta - \mathrm{Y}_m||^2 = \min_{\tilde{\beta} \in \mathcal{H}} ||\mathrm{X}_m \tilde{\beta} - \mathrm{Y}_m||^2$$

- equivalently: minimum norm solution to the normal equations (see [EHN96])

$$\hat{\beta}_m = \arg\min_{\beta} \{\beta \ : \ X_m^T X_m \beta = X_m^T \mathrm{Y}_m\} = \mathrm{X}_m^\dagger \mathrm{Y}_m$$

  note: if $X_m^T X_m$ is invertible, then $\mathrm{X}_m^\dagger = X_m^T (X_m^T X_m)^{-1}$

- final estimator: $\bar{\beta}_M = \frac{1}{M} \sum_{j=1}^M \hat{\beta}_m$

Question:

What is the statistical performance of $\bar{\beta}_M$ compared to the single machine problem?

Definition 1:

The covariance operator $\Sigma : \mathcal{H} \to \mathcal{H}$ is defined as $\Sigma := \mathbb{E}[\langle \cdot, X \rangle X]$.

If $\mathbb{E}[||X||^2] < \infty$, then Σ is nuclear [KL17]. The eigenvalues are denoted in descending order: $\lambda_1 \geq \lambda_2 \geq \dots$, the corresponding eigenvectors are denoted $(v_j)_j$.

Excess Risk: $\mathcal{R}(\bar{\beta}_M) - \mathcal{R}(\beta^*) = ||\Sigma^{\frac{1}{2}}(\bar{\beta}_M - \beta^*)||^2$

Definition 1:

The covariance operator $\Sigma : \mathcal{H} \to \mathcal{H}$ is defined as $\Sigma := \mathbb{E}[\langle \cdot, X \rangle X]$.

If $\mathbb{E}[||X||^2] < \infty$, then Σ is nuclear [KL17]. The eigenvalues are denoted in descending order: $\lambda_1 \geq \lambda_2 \geq \dots$, the corresponding eigenvectors are denoted $(v_j)_j$.

Excess Risk: $\mathcal{R}(\bar{\beta}_M) - \mathcal{R}(\beta^*) = ||\Sigma^{\frac{1}{2}}(\bar{\beta}_M - \beta^*)||^2$

note: When do minimum norm predictors generalize ?

[CL20], [TB20], [BLLT20], [MVSS20], [KLS20], [RMR20], [LR+20], [Bel21], [BMR21], ...

... find criterion on the eigenvalues of Σ to ensure that overfitting is benign.
Overparameterization is essential.

Hardness of the Learning Problem

Source Condition: increasing source function $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$ describes how coefficients of β^* vary along the eigenvectors of Σ, see [BPR07, BM18, RMR20]

Assumption 2:

Let β^* be randomly sampled with mean $\mathbb{E}_{\beta^*}[\beta^*] = 0$ and covariance $\mathbb{E}_{\beta^*}[\langle \cdot, \beta^* \rangle \beta^*] = \Phi(\Sigma)$.

Hardness of the Learning Problem

Source Condition: increasing source function $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$ describes how coefficients of β^* vary along the eigenvectors of Σ, see [BPR07, BM18, RMR20]

Assumption 2:

Let β^* be randomly sampled with mean $\mathbb{E}_{\beta^*}[\beta^*] = 0$ and covariance $\mathbb{E}_{\beta^*}[\langle \cdot, \beta^* \rangle \beta^*] = \Phi(\Sigma)$.

- expected contribution of j-th direction to the signal is given by

$$\mathbb{E}_{\beta^*}\left[\left|\left| \left\langle \Sigma^{\frac{1}{2}} \beta^*, v_j \right\rangle v_j \right|\right|^2 \right] = \lambda_j \Phi(\lambda_j)$$

- when Φ is increasing, strength along direction v_j decays fast for decreasing λ_j and principle components with larger eigenvalues carry more signal ("low-dim problem", "sparsity")

Hardness of the Learning Problem

Source Condition: increasing source function $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$ describes how coefficients of β^* vary along the eigenvectors of Σ, see [BPR07, BM18, RMR20]

Assumption 2:

Let β^* be randomly sampled with mean $\mathbb{E}_{\beta^*}[\beta^*] = 0$ and covariance $\mathbb{E}_{\beta^*}[\langle \cdot, \beta^* \rangle \beta^*] = \Phi(\Sigma)$.

- expected contribution of j-th direction to the signal is given by

$$\mathbb{E}_{\beta^*}\left[\left|\left| \left\langle \Sigma^{\frac{1}{2}} \beta^*, v_j \right\rangle v_j \right|\right|^2 \right] = \lambda_j \Phi(\lambda_j)$$

- when Φ is increasing, strength along direction v_j decays fast for decreasing λ_j and principle components with larger eigenvalues carry more signal ("low-dim problem", "sparsity")

Example: $\Phi(x) = x^r$, $r > 0$, thus: the larger r the easier the problem

Distributional Assumptions

Assumption 3:

1. $z = \Sigma^{-\frac{1}{2}} x$ has components that are independent and σ_x-subgaussian.
2. The noise $\varepsilon = Y - \langle \beta^*, X \rangle$ is σ-subgaussian conditionally on X.
   Note: This implies our model is well-defined.
3. The projection of the local data X_m on the space orthogonal to any eigenvector spans a space of dimension $\frac{n}{M}$. E.g.: $\text{rank}(\Sigma) > \frac{n}{M}$.

Theorem 1 [MRKR21]:

Suppose A1, A2, A3 are satisfied and let $\tau \leq \frac{n}{M}$, $c > 0$. There exists a $k^* \leq \frac{n}{M}$ such that the excess risk satisfies with probability at least $1 - Me^{-\frac{n}{M}c}$

$$\mathbb{E}_{\beta^*}[||\Sigma^{1/2}(\bar{\beta}_M - \beta^*)||^2] \lesssim \widehat{\text{Bias}}(\bar{\beta}_M) + \widehat{\text{Var}}(\bar{\beta}_M) \ ,$$

Theorem 1 [MRKR21]:

Suppose A1, A2, A3 are satisfied and let $\tau \leq \frac{n}{M}$, $c > 0$. There exists a $k^* \leq \frac{n}{M}$ such that the excess risk satisfies with probability at least $1 - Me^{-\frac{n}{M}c}$

$$\mathbb{E}_{\beta^*}[||\Sigma^{1/2}(\bar{\beta}_M - \beta^*)||^2] \lesssim \widehat{\text{Bias}}(\bar{\beta}_M) + \widehat{\text{Var}}(\bar{\beta}_M) \ ,$$

where

$$\widehat{\text{Bias}}(\bar{\beta}_M) = \tau \ ||\Sigma^{1+r}|| \sqrt{\frac{M}{n}} \ ,$$

$$\widehat{\text{Var}}(\bar{\beta}_M) = \tau\sigma^2 \left(\frac{k^*}{n} + \frac{n}{M^2} \frac{\sum_{j>k^*} \lambda_j^2}{\left(\sum_{j>k^*} \lambda_j\right)^2} \right) .$$

note: For $M = 1$ we recover the variance from [BLLT20] while improving the bias bound. The variance is optimal.

Corollary 1 [MRKR21]:

Assume $\lambda_j = c_n\, j^{-(1+\gamma_n)}$. Then, with probability at least $1 - Me^{-\tau^2}$

$$\mathbb{E}_{\beta^*}[||\Sigma^{1/2}(\bar{\beta}_M - \beta^*)||^2] \;\lesssim\; c_n^{1+r}\sqrt{\frac{M}{n}} + \sigma^2\,\frac{\gamma_n}{M}\,.$$

If $\gamma_n = c_n = 1/\log(n)$ and

$$M_n \;\leq\; \sigma^{4/3}\, n^{\frac{1}{3}}\, \log(n)^{\frac{2}{3}r}\,,$$

then for any n sufficiently large

$$\mathbb{E}_{\beta^*}[||\Sigma^{1/2}(\bar{\beta}_{M_n} - \beta^*)||^2] \;\simeq\; 2\sigma^2\,\frac{1}{M_n\log(n)}\,.$$

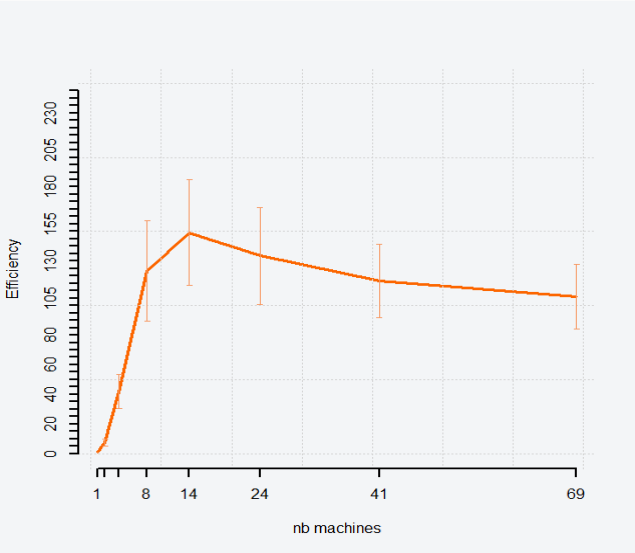
note: The excess risk converges to zero and overfitting is benign!

Corollary 2 [MRKR21]:

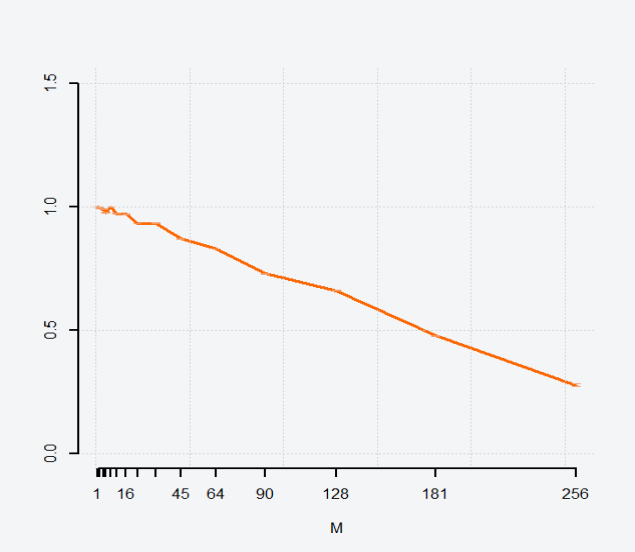
Assumptions as above. Then for any n sufficiently large, the relative prediction efficiency is

$$\mathrm{Eff}(M_n) \;=\; \frac{\mathbb{E}_{\beta^*}[||\Sigma^{1/2}(\bar{\beta}_1 - \beta^*)||^2]}{\mathbb{E}_{\beta^*}[||\Sigma^{1/2}(\bar{\beta}_{M_n} - \beta^*)||^2]} \;=\; \mathcal{O}(M_n)\,.$$

linear increase in efficiency until a threshold



linear loss in efficiency for distributed OLS, underparameterized case $d = 10$, $n = 8000$



see e.g. [RN16]

References i

Mikhail Belkin.
Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation.
arXiv preprint arXiv:2105.14368, 2021.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler.
Benign overfitting in linear regression.
Proceedings of the National Academy of Sciences, 117(48):30063–30070, 2020.

Gilles Blanchard and Nicole Mücke.
Optimal rates for regularization of statistical inverse learning problems.
Foundations of Computational Mathematics, 18(4):971–1013, 2018.

Peter L Bartlett, Andrea Montanari, and Alexander Rakhlin.
Deep learning: a statistical viewpoint.
arXiv preprint arXiv:2103.09177, 2021.

References ii

Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco.
On regularization algorithms in learning theory.
Journal of complexity, 23(1):52–72, 2007.

Geoffrey Chinot and Matthieu Lerasle.
Benign overfitting in the large deviation regime.
arXiv preprint arXiv:2003.05838, 2020.

Heinz Werner Engl, Martin Hanke, and Andreas Neubauer.
Regularization of inverse problems, volume 375.
Springer Science & Business Media, 1996.

Vladimir Koltchinskii and Karim Lounici.
Concentration inequalities and moment bounds for sample covariance operators.
Bernoulli, 23(1):110–133, 2017.

References iii

Dmitry Kobak, Jonathan Lomond, and Benoit Sanchez.
The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization.
Journal of Machine Learning Research, 21(169):1–16, 2020.

Tengyuan Liang, Alexander Rakhlin, et al.
Just interpolate: Kernel ridgeless regression can generalize.
Annals of Statistics, 48(3):1329–1347, 2020.

Nicole Mücke and Gilles Blanchard.
Parallelizing spectrally regularized kernel algorithms.
The Journal of Machine Learning Research, 19(1):1069–1097, 2018.

Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai.
Harmless interpolation of noisy data in regression.
IEEE Journal on Selected Areas in Information Theory, 1(1):67–83, 2020.

References iv

Dominic Richards, Jaouad Mourtada, and Lorenzo Rosasco.
Asymptotics of ridge (less) regression under general source condition.
arXiv preprint arXiv:2006.06386, 2020.

Jonathan D Rosenblatt and Boaz Nadler.
On the optimality of averaging in distributed statistical learning.
Information and Inference: A Journal of the IMA, 5(4):379–404, 2016.

Alexander Tsigler and Peter L Bartlett.
Benign overfitting in ridge regression.
arXiv preprint arXiv:2009.14286, 2020.

References v

Yuxin Wang, Qiang Wang, Shaohuai Shi, Xin He, Zhenheng Tang, Kaiyong Zhao, and Xiaowen Chu.
Benchmarking the performance and energy efficiency of ai accelerators for ai training.
In *2020 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGRID)*, pages 744–751. IEEE, 2020.

Yuchen Zhang, John Duchi, and Martin Wainwright.
Divide and conquer kernel ridge regression.
In *Conference on learning theory*, pages 592–617. PMLR, 2013.