Coffee Talk #3

September 2, 2021

Ozgur Taylan Turan

Optimal Regularization Can Mitigate Double Descent[1]

[1] P. Nakkiran, P. Venkat, S. Kakade, and T. Ma (2020). "Optimal Regularization Can Mitigate Double Descent". In: ISSN: 2331-8422. arXiv: 2003.01897

Why This Paper?

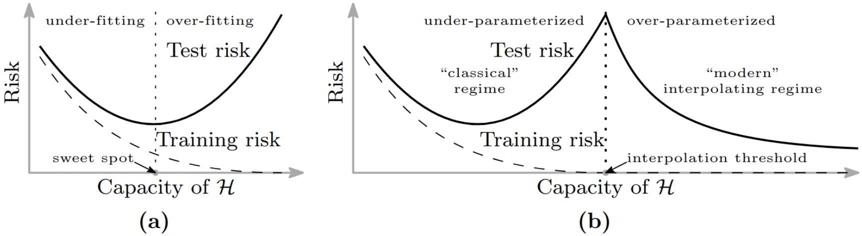
Journey after reading Marco & Tom's 2019 article[2]

[2] M. Loog, T. Viering, and A. Mey (2019). "Minimizers of the empirical risk and risk monotonicity". In: *Advances in Neural Information Processing Systems 32*.NeurIPS. ISSN: 10495258. arXiv: 1907.05476

Aim

Show theoretically and empirically optimal regularization can ensure monotonicity for sample and model size under certain assumptions!

Double Descent



[3]

- See Marco and other PR collogues work[4] for a detailed history of this behaviour.

---

[3] M. Belkin, D. Hsu, S. Ma, and S. Mandal (2019). "Reconciling modern machine-learning practice and the classical bias–variance trade-off". In: *Proceedings of the National Academy of Sciences of the United States of America* 116.32, pp. 15849–15854. ISSN: 10916490. DOI: 10.1073/pnas.1903070116. arXiv: arXiv:1812.11118v2

[4] M. Loog, T. Viering, A. Mey, J. H. Krijthe, and D. M. Tax (2020). "A brief prehistory of double descent". In: *Proceedings of the National Academy of Sciences of the United States of America* 117.20, pp. 10625–10626. ISSN: 10916490. DOI: 10.1073/pnas.2001875117. arXiv: 2004.04328

Why do we care?

- Potential gap in understanding of generalization. (performance on new data)
- We want monotonic behaving models with respect to model compleixty and data.

Aim

When does optimally tuned regularization mitigate the double descent phenomenon?

P.S. This question implicitly assumes that double descent is observed mostly for under-regularized models.

Remarks

- Claims are regarding the empirical test risk.
- Theoretical results are derived under the assumption that the covariance of the data is isotropic.

Ridge Regression-A

- For input $x \in \mathbb{R}^d$ generated from $\mathcal{N}(0, I_d)$ output is $y = \langle x, \beta^* \rangle + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2)$
- With the aim to learn $f_\beta(x) = \langle x, \beta \rangle$ with n training samples drawn i.i.d. from $\mathcal{D}$ which is the joint dist. of (x, y) by minimizing population mean-squared error $R(\beta) := \underset{(x,y)\sim\mathcal{D}}{\mathbb{E}}[(\langle x, \beta \rangle - y)^2]$
- with input matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ and output vector $\boldsymbol{y} \in \mathbb{R}^n$
- Ridge estimator is given by,

$$\hat{\beta}_{n,\lambda} = \underset{\beta}{\text{argmin}} \, ||\boldsymbol{X}\beta - \boldsymbol{y}||^2 + \lambda||\beta||^2 \quad (1)$$

$$= (\boldsymbol{X}^T\boldsymbol{X} + \lambda I_d)^{-1}\boldsymbol{X}^T\boldsymbol{y} \quad (2)$$

Ridge Regression-B

- Optimal ridge parameter for n samples is given by,

$$\lambda_n^{\text{opt}} = \underset{\lambda}{\text{argmin}}\, \bar{R}(\hat{\beta}_{n,\lambda}) \tag{3}$$

where, $\bar{R}(\hat{\beta}_{n,\lambda}) = \underset{\boldsymbol{X},\boldsymbol{y}\sim\mathcal{D}^n}{\mathbb{E}}[R(\hat{\beta}_n(\boldsymbol{X},\boldsymbol{y}))]$

Sample Monotonicity in Ridge Reg.

- The expected risk of optimally regularized well-specified isotropic(/non-isotropic?) linear reg. is monotonic in samples. $\rightarrow \bar{R}(\hat{\beta}^{\text{opt}}_{n+1}) \leq \bar{R}(\hat{\beta}^{\text{opt}}_{n})$

Closed form solution? So, ...

The Basic Idea

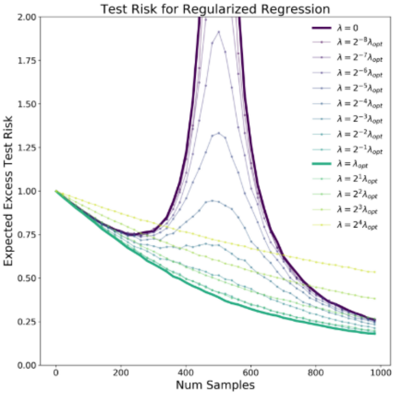
Noting that all the notation with $\sim$ on correspond to $n+1$ samples and the rest belongs to n samples.

- Let γ be the singular values of $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ which are distributed with Γ_n.
- Isotropy of x and exploiting interlacing between Γ_n and Γ_{n+1} allows,

$\bar{R}(\hat{\beta}^{\text{opt}}_{n}) = \mathop{\mathbb{E}}_{\Gamma_n}\left[\sum_{i=1}^{d} \frac{\sigma^2}{\gamma_i^2 + d\sigma^2/||\beta^*||^2}\right] + \sigma^2$. Noting that interlacing ensures $\gamma_i \leq \tilde{\gamma}_i$

Sample Monotonicity in Ridge Reg.

$$\mathop{\mathbb{E}}_{\Gamma_n}\left[\sum_{i=1}^{d}\frac{\sigma^2}{\gamma_i^2+d\sigma^2/||\beta^*||^2}\right]\leq\mathop{\mathbb{E}}_{\Gamma_{n+1}}\left[\sum_{i=1}^{d}\frac{\sigma^2}{\tilde{\gamma}_i^2+d\sigma^2/||\beta^*||^2}\right]$$



Model Monotonicity Ridge Reg.

Remark

- Only for this section it is assumed that the covariates live in p-dimensional space, but the regression model is employed after projection to a d-dimensional space ($\boldsymbol{X} \in \mathbb{R}^{n \times p}$). Then, $\tilde{\boldsymbol{X}}\boldsymbol{P}^T$ where $\boldsymbol{P} \in \mathbb{R}^{d \times p}$ is a random orthonormal matrix.

- Risk of the estimator $\bar{R}(\hat{\beta}) = \mathop{\mathbb{E}}_{\boldsymbol{P}} \mathop{\mathbb{E}}_{\tilde{\boldsymbol{X}}, \boldsymbol{y} \sim \mathcal{D}^n} [R_P(\hat{\beta}_n(\tilde{\boldsymbol{X}}, \boldsymbol{y}))]$
- In similar fashion $\bar{R}(\hat{\beta}^{\text{opt}}_{d+1}) \leq \bar{R}(\hat{\beta}^{\text{opt}}_{d})$

Similar emprical results on Random ReLU Features and CNN's.

Conclusions

- Certain linear models optimal $\ell 2$ regularization can prevent non-monotonic behaviour.
- Investigation of more complicated and nonlinear models?