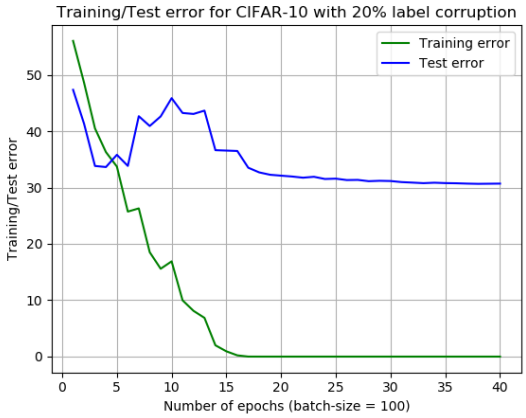
Project Progress Report

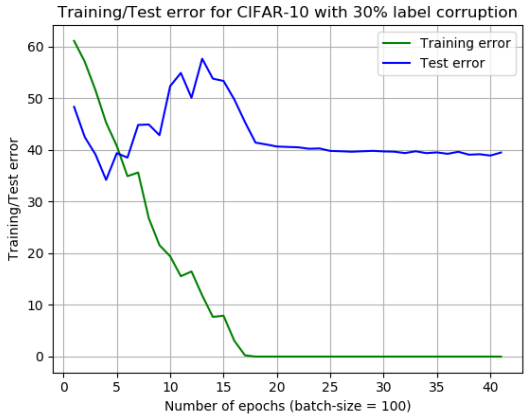
Rudrajit Das & Anish Acharya
University of Texas at Austin

1 Some Experimental Results

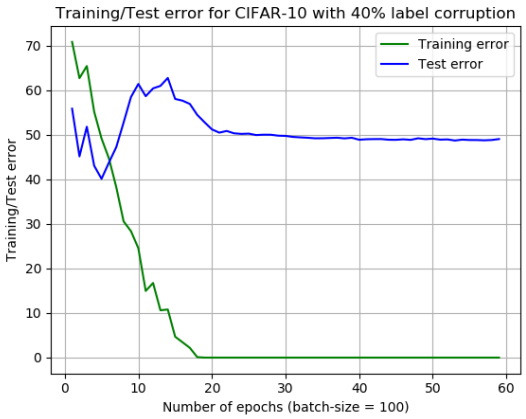
CIFAR-10:



(a) 20% labels corrupted



(b) 30% labels corrupted



(c) 40% labels corrupted

Figure 1: Double descent in CIFAR-10 as a function of the number of epochs (fixed model size and sample size) with different levels of label corruption.

We trained CIFAR-10 on ResNet-18 using SGD with a batch-size of 100. We applied weight decay=0.001. We used the following learning rate scheme: $\eta_i = \eta_0\left(1 - \frac{i}{200}\right)$, where

$\eta_0 = 0.005$. We obtained results for three different levels of label corruption - 20%, 30%, 40%. 20% label corruption means that the labels of 20% of the training examples were randomly flipped (to a label other than its true label). Observe epoch-wise double descent for the 3 cases. For 20% label corruption, the final test error (when training error is 0) is lower than the initial minimum in the test error (before it starts rising). But for 30% and 40% label corruption, the final test error (when training error is 0) is actually higher than the initial minimum in the test error.

Linear Regression:

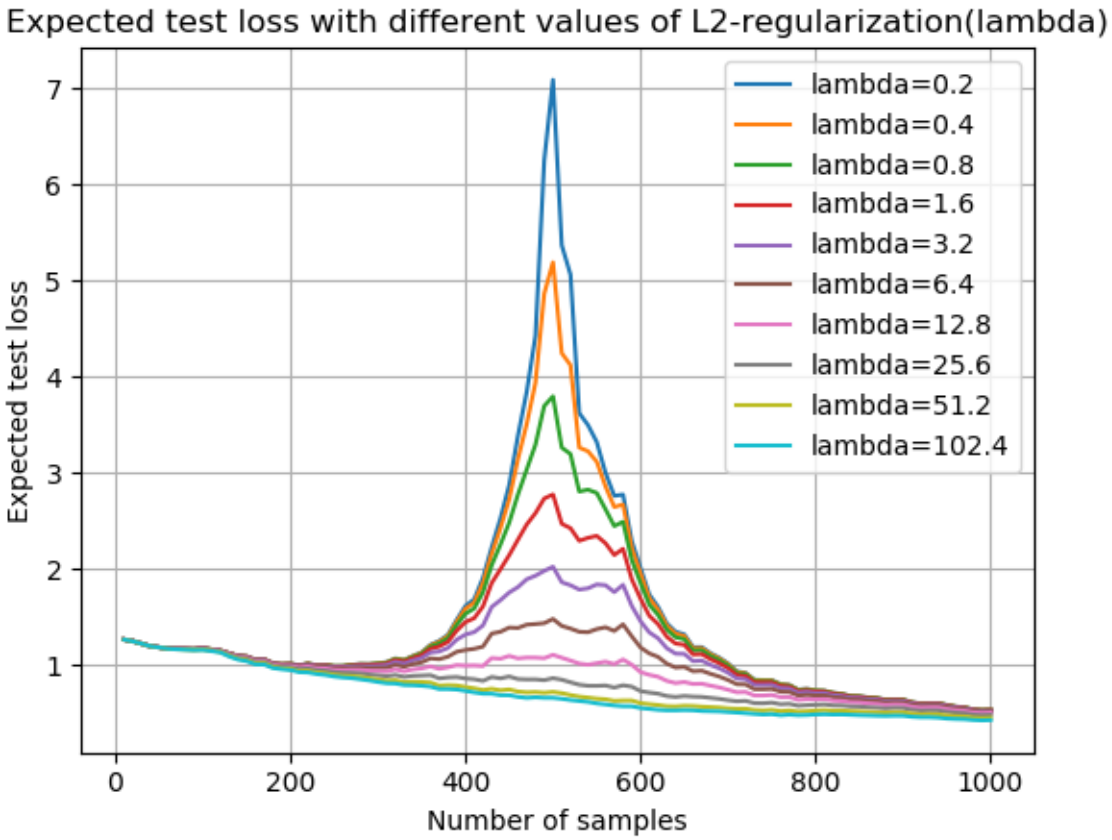


Figure 2: Sample-wise double descent in linear regression with different values of L2-regularizer.

We considered the following model - $y = \langle \boldsymbol{x}, \boldsymbol{\beta} \rangle + \epsilon$, where $\boldsymbol{x} \in \mathbb{R}^{500}$ and each component of $\boldsymbol{x}$ is sampled i.i.d from $\mathcal{N}(0, 1)$, $\boldsymbol{\beta} \in \mathbb{R}^{500}$ and $\|\boldsymbol{\beta}\| = 1$, $\epsilon \sim \mathcal{N}(0, 0.25)$. We consider the problem of estimation of $\boldsymbol{\beta}$ given multiple samples of $(\boldsymbol{x}, y)$ by solving a ridge-regression problem with different values of L2-regularizer (λ). For each value of λ, the number of training samples was varied from 10 to 1000 in steps of 10. The values of λ which we tried were $\{0.2, 0.4, 0.8, 1.6, 3.2, 6.4, 12.8, 25.6, 51.2, 102.4\}$. Figure 2 shows the variation of expected test loss with the number of samples for the aforementioned values of λ.

Observe that for small values of λ, there is a peak when the number of samples is around 500 (consistent with the results in [1]). However, for larger values of λ, there is no peak and instead, we have a monotonic decrease in the expected test loss. Thus, there is no double descent with larger values of λ. This forms the starting point to our discussion in Section 2.

An observation

It seems to us that the double descent phenomenon depends on the presence of noise in the training data – without the presence of non-negligible noise, double descent is hard to observe. Further, in most of the experiments of [2], they have not used any regularization at all.

2 Summary of very recent concurrent work

We discovered a recent Arxiv submission [3] which is closely based on our proposal. This was submitted to Arxiv after we submitted our proposal due to which this is not there in the list of our original references. [3] proves that optimally-tuned ℓ_2 regularization for some linear regression problems with isotropic data distribution results in monotonic test performance with either increasing sample size or model size.

Sample-wise monotonicity in ridge regression

To show monotonicity in test loss with respect to the number of training samples using an optimal value of regularizer, the authors consider the following linear regression problem:

$$y = \langle \boldsymbol{x}, \boldsymbol{\beta}^* \rangle + \epsilon \tag{1}$$

In (1), $\boldsymbol{x} \in \mathbb{R}^d$ and is drawn from $\mathcal{N}(\vec{\mathbf{0}}, \boldsymbol{I}_d)$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$, and $\boldsymbol{\beta}^* \in \mathbb{R}^d$. Let us denote the joint distribution of $(\boldsymbol{x}, y)$ by $\mathcal{D}$. We are given n input-output pairs $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ drawn from $\mathcal{D}$ and we wish to estimate $\boldsymbol{\beta}^*$. This is done by solving a ridge-regression problem with ℓ_2 regularizer $= \lambda$. Let us call this estimate $\widehat{\boldsymbol{\beta}}_{n,\lambda}$ which is given by:

$$\widehat{\boldsymbol{\beta}}_{n,\lambda} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \ \|\boldsymbol{y}_n - \boldsymbol{X}_n\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2 \tag{2}$$

In (2), $\boldsymbol{y}_n = [y_1, \ldots, y_n]^T$ and $\boldsymbol{X}_n = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^T$. The expected test set loss is:

$$\overline{R}(\widehat{\boldsymbol{\beta}}_{n,\lambda}) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[(y - \langle \boldsymbol{x}, \widehat{\boldsymbol{\beta}}_{n,\lambda}\rangle)^2] \tag{3}$$

Now let $\lambda_n^{\text{opt}} \triangleq \underset{\lambda}{\operatorname{argmin}}\ \overline{R}(\widehat{\boldsymbol{\beta}}_{n,\lambda})$. They show that this is equal to $d\sigma^2/\|\boldsymbol{\beta}^*\|^2 = \lambda^*$. Notice that this quantity is independent of n. Now define $\widehat{\boldsymbol{\beta}}_n^{\text{opt}} \triangleq \widehat{\boldsymbol{\beta}}_{n,\lambda^*}$, i.e.:

$$\widehat{\boldsymbol{\beta}}_n^{\text{opt}} \triangleq \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \ \|\boldsymbol{y}_n - \boldsymbol{X}_n\boldsymbol{\beta}\|^2 + \lambda^*\|\boldsymbol{\beta}\|^2 \tag{4}$$

Finally, they show that with λ^*:

$$\overline{R}(\widehat{\boldsymbol{\beta}}_{n+1,\lambda^*}) \leq \overline{R}(\widehat{\boldsymbol{\beta}}_{n,\lambda^*}) \implies \overline{R}(\widehat{\boldsymbol{\beta}}_{n+1}^{\text{opt}}) \leq \overline{R}(\widehat{\boldsymbol{\beta}}_n^{\text{opt}}) \tag{5}$$

Finally, with all $\lambda \geq \lambda^*$, sample-wise monotonicity still holds, i.e. $\overline{R}(\widehat{\boldsymbol{\beta}}_{n,\lambda}) \leq \overline{R}(\widehat{\boldsymbol{\beta}}_{n+1,\lambda})\ \forall \lambda \geq \lambda^*$.

Model-wise monotonicity in ridge regression

To show monotonicity in test loss with respect to the model-size using an optimal value of regularizer, the authors consider the following linear regression problem

$$y = \langle \boldsymbol{x}, \boldsymbol{\theta}^* \rangle + \epsilon \tag{6}$$

In (6), $\boldsymbol{x} \in \mathbb{R}^p$ is drawn from $\mathcal{N}(\vec{\mathbf{0}}, \boldsymbol{I}_p)$, $\boldsymbol{\theta}^* \in \mathbb{R}^p$, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$. But instead of using $\boldsymbol{x}$ as the input variable, they propose using $\tilde{\boldsymbol{x}} = \boldsymbol{P}\boldsymbol{x}$, where $\boldsymbol{P}$ is a random $d \times p$ orthonormal matrix with $d \leq p$ (so $\tilde{\boldsymbol{x}} \in \mathbb{R}^d$), as the input variable to the regression problem. Thus, d represents the model-size while $p \geq d$ is the dimension of the actual data. Let us denote the joint distribution of $(\tilde{\boldsymbol{x}}, y)$ by $\mathcal{D}$. Just as in the previous case, we solve a ridge-regression problem with n pairs of $(\tilde{\boldsymbol{x}}, y)$ which are $\{(\tilde{\boldsymbol{x}}_i, y_i)\}_{i=1}^n$:

$$\widehat{\boldsymbol{\beta}}_{d,\lambda} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\ \|\boldsymbol{y} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2 \tag{7}$$

In (7), $\boldsymbol{y} = [y_1, \ldots, y_n]^T$ and $\tilde{\boldsymbol{X}} = [\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_n]^T = \boldsymbol{X}\boldsymbol{P}^T$ ($\tilde{\boldsymbol{X}}$ is a $n \times d$ matrix while $\boldsymbol{X}$ is a $n \times p$ matrix). Note that $\boldsymbol{P}$ is fixed here. Then, the expected test set loss over $\mathcal{D}$ and the distribution of $\boldsymbol{P}$ is:

$$\overline{R}(\widehat{\boldsymbol{\beta}}_{d,\lambda}) = \mathbb{E}_{\boldsymbol{P}}[\mathbb{E}_{(\tilde{\boldsymbol{x}},y)\sim\mathcal{D}}[(y - \langle \tilde{\boldsymbol{x}}, \widehat{\boldsymbol{\beta}}_{n,\lambda}\rangle)^2]] = \mathbb{E}_{\boldsymbol{P}}[\mathbb{E}_{(\boldsymbol{x},y)}[(y - \langle \boldsymbol{P}\boldsymbol{x}, \widehat{\boldsymbol{\beta}}_{n,\lambda}\rangle)^2]] \tag{8}$$

Now let $\lambda_d^{\text{opt}} \triangleq \underset{\lambda}{\operatorname{argmin}}\ \overline{R}(\widehat{\boldsymbol{\beta}}_{d,\lambda})$. Just as in the earlier case, now define $\widehat{\boldsymbol{\beta}}_d^{\text{opt}} \triangleq \widehat{\boldsymbol{\beta}}_{d,\lambda_d^{\text{opt}}}$, i.e.:

$$\widehat{\boldsymbol{\beta}}_d^{\text{opt}} \triangleq \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\ \|\boldsymbol{y} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}\|^2 + \lambda_d^{\text{opt}}\|\boldsymbol{\beta}\|^2 \tag{9}$$

Finally, like the earlier case, they show that:

$$\overline{R}(\widehat{\boldsymbol{\beta}}_{d+1}^{\text{opt}}) \leq \overline{R}(\widehat{\boldsymbol{\beta}}_d^{\text{opt}}) \tag{10}$$

Limitations of this work: All the theoretical results have been derived for the special case of $\boldsymbol{x} \sim \mathcal{N}(\vec{\mathbf{0}}, \boldsymbol{I})$. Proving these results for the general case of $\boldsymbol{x} \sim \mathcal{N}(\vec{\mathbf{0}}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a PSD matrix, is much harder – the authors leave another conjecture which if proved would imply sample-wise monotonicity for the general case of non-isotropic data. Having said that, the authors provide some empirical evidence to suggest that sample-wise monotonicity holds even for the general case of non-isotropic data.

References

[1] Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *arXiv preprint arXiv:1903.07571*, 2019.

[2] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *arXiv preprint arXiv:1912.02292*, 2019.

[3] Preetum Nakkiran, Prayaag Venkat, Sham Kakade, and Tengyu Ma. Optimal regularization can mitigate double descent. *arXiv preprint arXiv:2003.01897*, 2020.