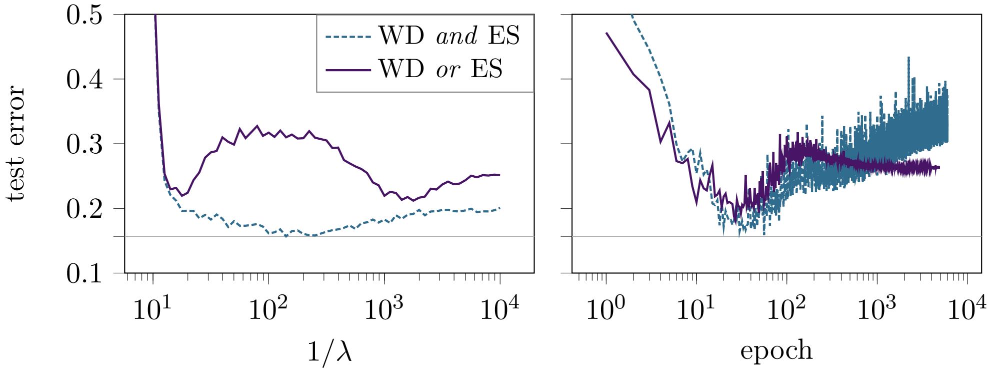

Figure 5: Comparison of individually or jointly applied regularization by early stopping and weight decay: Test performance of the 5-layer convolution network when trained on the CIFAR-10 dataset with 20% label noise. **Left:** Performance as a function of the regularization strength for training with (*solid*) weight decay only—WD—and (*dashed*) weight decay together with early stopping—WD and ES. **Left:** Performance as a function of the training epochs for (*solid*) standard training and (*dashed*) training with weight decay. **Both:** Better performance is achieved by jointly utilizing weight decay and early stopping—WD and ES.

A Double descent behavior of deep networks in the presence of both weight decay and early stopping

Here, we expand on the results provided in Figure 1 and show that both regularization-wise and epoch-wise double descent can be eliminated by employing additional forms of regularization. Specifically, in Figure 5, our results show that utilizing early stopping eliminates regularization-wise double descent, whereas utilizing (tuned) weight decay eliminates the corresponding epoch-wise double descent. Note that performance achieved in the case where early stopping and weight decay are used together is much better than that obtained by using either weight decay or early stopping alone.

B Double descent as a function of dropout regularization

Our results showcasing the double descent behavior as a function of the ℓ_2 regularization strength motivates the investigation of other types of regularization and whether double descent also occurs for other explicit regularization methods. In Figure 6, we show the test error of the 5-layer CNN with dropout added after the activations of each layer trained on the noisy CIFAR-10. The test error exhibits a U-shaped curve as a function of the dropout probability with optimal dropout probability $p_{dropout} = 0.4$.

C Discussion and proof statements for linear ridge regression

C.1 Intuition for the risk expression (2)

We first provide intuition on why the risk is governed by the risk expression given in (2).

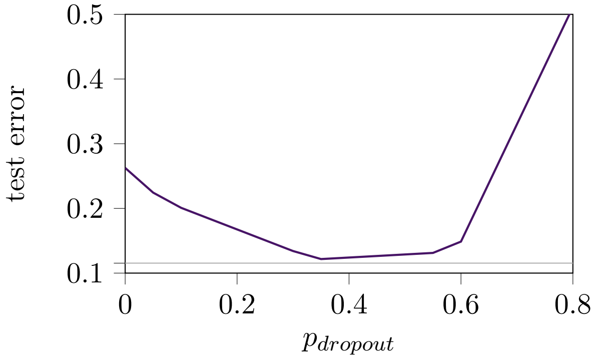

Figure 6: Test performance of the 5-layer convolution network as a function of the dropout probability when trained on the CIFAR-10 dataset with 20% label noise.

First, note that the risk of the resulting estimator can be written as a function of the variances of the features, σ_i^2, and of the coefficients of the underlying true linear model, $\boldsymbol{\theta}^* = [\theta_1^*, \ldots, \theta_d^*]$, as

$$R(\hat{\boldsymbol{\theta}}_\lambda) = \sigma^2 + \sum_{i=1}^{d} \sigma_i^2 (\theta_i^* - \hat{\theta}_{\lambda,i})^2. \tag{7}$$

which follows from noting that z and $\mathbf{x}$ are independently drawn.

Next, note that we aim to find the estimator which minimizes the ell_2-regularized MSE loss

$$\mathcal{L}_\lambda(\boldsymbol{\theta}) = \frac{1}{2}\|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|_2^2 + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2.$$

Recall that, as introduced in Section 3.1 , the matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ contains the scaled training feature vectors $\frac{1}{\sqrt{n}}\mathbf{x}_1, \ldots, \frac{1}{\sqrt{n}}\mathbf{x}_n$ as rows, and $\mathbf{y} = \frac{1}{\sqrt{n}}[y_1, \ldots, y_n]$ are the corresponding scaled responses. Then, the solution of the ℓ_2 regularized problem can be found by simply setting the gradient of the loss function to zero and solving for $\boldsymbol{\theta}$, which yields

$$\boldsymbol{\theta}_\lambda - \boldsymbol{\theta}^* = ((\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X} - \mathbf{I})\boldsymbol{\theta}^* + (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{z},$$

where $\mathbf{z} = [z_1, \ldots, z_n]$ is the noise. As we formalize below, in the under-parameterized regime where $n \gg d$, we have that $\mathbf{X}^T\mathbf{X} \approx \boldsymbol{\Sigma}^2$. Therefore the original solution is close to the proximal solution $\tilde{\boldsymbol{\theta}}_\lambda$ defined by

$$\tilde{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^* = ((\boldsymbol{\Sigma}^T\boldsymbol{\Sigma} + \lambda\mathbf{I})^{-1}\boldsymbol{\Sigma}^T\boldsymbol{\Sigma} - \mathbf{I})\boldsymbol{\theta}^* + (\boldsymbol{\Sigma}^T\boldsymbol{\Sigma} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{z}, \tag{8}$$

The proximal solution is close to the original solution obtained by solving for the minimizer of the ℓ_2-regularized loss function. Note that, from (8), we get, for the i-th entry of $\tilde{\boldsymbol{\theta}}_\lambda$

$$\tilde{\boldsymbol{\theta}}_{\lambda,i} - \boldsymbol{\theta}_i^* = \tilde{\mathbf{x}}_i^T\mathbf{z}\frac{1}{\sigma_i^2 + \lambda} - \frac{\lambda}{\sigma_i^2 + \lambda}\boldsymbol{\theta}_i^*,$$

where $\tilde{\mathbf{x}}_i$ is the i-th *column* of $\mathbf{X}$ (not the i-th example/feature vector!). Next note that, $\mathbb{E}\left[(\tilde{\mathbf{x}}_i^T\mathbf{z})^2\right] \approx \sigma^2\sigma_i^2$ because the entries of $\mathbf{z}$ are $\mathcal{N}(0, \sigma^2)$ distributed, and the entries of $\tilde{\mathbf{x}}_i$ are $1/\sqrt{n}\mathcal{N}(0, \sigma_i^2)$

distributed. Using this expectation in the solution $\tilde{\boldsymbol{\theta}}_\lambda$, and evaluating the resulting risk of those iterates via the formula for the risk given by (7) yields the risk expression (2). The proof of Theorem 1 in this appendix makes this intuition precise by formally bounding the difference of the proximal solution $\tilde{\boldsymbol{\theta}}_\lambda$ to the original solution $\boldsymbol{\theta}_\lambda$.

C.2 Proof of Theorem 1

In this section, we provide the formal proof for Theorem 1.

The difference between the two risk terms can be further dissected into two separate terms:

$$\left|R(\boldsymbol{\theta}_\lambda) - \bar{R}(\tilde{\boldsymbol{\theta}}_\lambda)\right| \leq \left|R(\boldsymbol{\theta}_\lambda) - R(\tilde{\boldsymbol{\theta}}_\lambda)\right| + \left|R(\tilde{\boldsymbol{\theta}}_\lambda) - \bar{R}(\tilde{\boldsymbol{\theta}}_\lambda)\right|. \tag{9}$$

We bound the two terms on the RHS of (9) separately. We first provide a bound for the first term with the lemma below.

Lemma 1. *Define $\tilde{\mathbf{X}}$ so that $\mathbf{X} = \tilde{\mathbf{X}}\boldsymbol{\Sigma}$. Suppose that $\left\|\mathbf{I} - \tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right\| \leq \epsilon$, with $\epsilon \leq (\min_i \sigma_i^2 + \lambda)/2$ Then*

$$\left|R(\boldsymbol{\theta}_\lambda) - R(\tilde{\boldsymbol{\theta}}_\lambda)\right| \leq 4\epsilon^2 \left(\frac{\max_i \sigma_i^4}{\min_i(\sigma_i^2 + \lambda)^2}\right)^2 \left(\left(\frac{\min_i \sigma_i^2 + \lambda}{\max_i \sigma_i^2} + 1\right) \left\|\boldsymbol{\Sigma}\boldsymbol{\theta}^*\right\|_2 + \left\|\tilde{\mathbf{X}}^T\mathbf{z}\right\|_2\right)^2 \tag{10}$$

We apply the lemma by first verifying its condition by referring to the derivations in [HY21, Lemma 1]. Note that the entries of the matrix $\tilde{\mathbf{X}}$ are iid Gaussians drawn from $\mathcal{N}(0, 1/n)$, and the same concentration inequality from [FR13, Chapter 9] results in, for any $\beta \in (0, 1)$,

$$\mathrm{P}\left[\left\|\mathbf{I} - \tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right\| \geq \beta\right] \leq e^{-\frac{n\beta^2}{15} + 4d}.$$

With $\beta = \sqrt{\frac{75d}{n}}$ we obtain that, with probability at least $1 - e^{-d}$,

$$\left\|\mathbf{I} - \tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right\| \leq \sqrt{75\frac{d}{n}}.$$

We next bound $\left\|\tilde{\mathbf{X}}^T\mathbf{z}\right\|_2$ with high probability:

Lemma 2. *With $\tilde{\mathbf{X}}$ previously defined such that $\mathbf{X} = \tilde{\mathbf{X}}\boldsymbol{\Sigma}$, with probability at least $1 - 2d(e^{-\beta^2/2} + e^{-n/8})$,*

$$\left\|\tilde{\mathbf{X}}^T\mathbf{z}\right\|_2 \leq 2\frac{d}{\sqrt{n}}\sigma\beta$$

Applying the lemma with $\beta^2 = 10\log(d)$, we obtain that with probability at least $1 - 2d^{-5} - 2de^{-n/8} - e^{-d}$ we have

$$\left|R(\boldsymbol{\theta}_\lambda) - R(\tilde{\boldsymbol{\theta}}_\lambda)\right| \leq 4\frac{75d}{n} \left(\frac{\max_i \sigma_i^4}{\min_i(\sigma_i^2 + \lambda)^2}\right)^2 \left(\left(\frac{\min_i \sigma_i^2 + \lambda}{\max_i \sigma_i^2} + 1\right) \left\|\boldsymbol{\Sigma}\boldsymbol{\theta}^*\right\|_2 + 2\frac{d}{\sqrt{n}}\sigma 10\log d\right)^2$$

We finally bound the second term in (9):

Lemma 3. *Provided that $d/n \leq \max_i((\sigma_i + \lambda)/\sigma_i^2)^4$, with probability at least $1 - 4e^{-\frac{\beta^2}{8}}$, we have that*

$$\left|R(\tilde{\boldsymbol{\theta}}_\lambda) - \bar{R}(\tilde{\boldsymbol{\theta}}_\lambda)\right| \leq \frac{\sigma^2}{n}\beta 3\sqrt{d}, \tag{11}$$

with $\bar{R}(\tilde{\boldsymbol{\theta}}_\lambda)$ as defined in (2).

For the proof of Lemma 3 we refer the reader to the proof of [HY21, Lemma 2] and note that (3) can be obtained by following the same steps with the additional assumption regarding the underparameterization as stated in Lemma 3.

We note that the assumption of the lemma is generally satisfied as we operate in the underparameterized regime and poses no strict restriction on the setup. Applying the two bounds (10) and (11) to the RHS of the bound (9) concludes the proof. The remainder of the proof is devoted to proving Lemma 1.

C.3 Proof of Lemma 1

Recall that the solutions of the original and closely related problem are given by

$$\begin{aligned}
\boldsymbol{\theta}_\lambda - \boldsymbol{\theta}^* &= ((\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X} - \mathbf{I})\boldsymbol{\theta}^* + (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{z},\\
\tilde{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^* &= ((\boldsymbol{\Sigma}^2 + \lambda\mathbf{I})^{-1}\boldsymbol{\Sigma}^2 - \mathbf{I})\boldsymbol{\theta}^* + (\boldsymbol{\Sigma}^2 + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{z}.
\end{aligned}$$

Note that $\mathbf{X} = \tilde{\mathbf{X}}\boldsymbol{\Sigma}$, where we defined $\tilde{\mathbf{X}}$ which has iid Gaussian entries $\mathcal{N}(0, 1/n)$. With this notation, and using that $\boldsymbol{\Sigma}$ is diagonal and therefore commutes with symmetric matrices, we obtain the following expressions for the residuals of the two solutions:

$$\begin{aligned}
\boldsymbol{\Sigma}\boldsymbol{\theta}_\lambda - \boldsymbol{\Sigma}\boldsymbol{\theta}^* &= \boldsymbol{\Sigma}((\boldsymbol{\Sigma}^2\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} + \lambda\mathbf{I})^{-1}\boldsymbol{\Sigma}^2\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} - \mathbf{I})\boldsymbol{\theta}^* + (\boldsymbol{\Sigma}^2\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} + \lambda\mathbf{I})^{-1}\boldsymbol{\Sigma}^2\tilde{\mathbf{X}}^T\mathbf{z},\\
\boldsymbol{\Sigma}\tilde{\boldsymbol{\theta}}_\lambda - \boldsymbol{\Sigma}\boldsymbol{\theta}^* &= \boldsymbol{\Sigma}((\boldsymbol{\Sigma}^2 + \lambda\mathbf{I})^{-1}\boldsymbol{\Sigma}^2 - \mathbf{I})\boldsymbol{\theta}^* + (\boldsymbol{\Sigma}^2 + \lambda\mathbf{I})^{-1}\boldsymbol{\Sigma}^2\tilde{\mathbf{X}}^T\mathbf{z}.
\end{aligned}$$

The difference between the residuals is

$$\begin{aligned}
\boldsymbol{\Sigma}\boldsymbol{\theta}_\lambda - \boldsymbol{\Sigma}\tilde{\boldsymbol{\theta}}_\lambda = \boldsymbol{\Sigma}^2&((\boldsymbol{\Sigma}^2\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} + \lambda\mathbf{I})^{-1}\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} - (\boldsymbol{\Sigma}^2 + \lambda\mathbf{I})^{-1})\boldsymbol{\Sigma}\boldsymbol{\theta}^*\\
&+ \boldsymbol{\Sigma}^2((\boldsymbol{\Sigma}^2\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} + \lambda\mathbf{I})^{-1} - (\boldsymbol{\Sigma}^2 + \lambda\mathbf{I})^{-1})\tilde{\mathbf{X}}^T\mathbf{z}.
\end{aligned}$$

$$\begin{aligned}
&= \boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}^2\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} + \lambda\mathbf{I})^{-1}(I - \tilde{\mathbf{X}}^T\tilde{\mathbf{X}})\boldsymbol{\Sigma}\boldsymbol{\theta}^*\\
&\quad + \boldsymbol{\Sigma}^2((\boldsymbol{\Sigma}^2\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} + \lambda\mathbf{I})^{-1} - (\boldsymbol{\Sigma}^2 + \lambda\mathbf{I})^{-1})(\boldsymbol{\Sigma}\boldsymbol{\theta}^* - \tilde{\mathbf{X}}^T\mathbf{z}).
\end{aligned}$$

Where, we added and subtracted $\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}^2\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} + \lambda\mathbf{I})^{-1}\boldsymbol{\Sigma}\boldsymbol{\theta}^*$ and re-arranged the terms. We bound the norm of the difference between the residuals $\left\|\boldsymbol{\Sigma}\boldsymbol{\theta}_\lambda - \boldsymbol{\Sigma}\tilde{\boldsymbol{\theta}}_\lambda\right\|_2$ by applying Cauchy-Schwarz inequality to the corresponding terms of the RHS of the equation above. We have, for the first term,

$$\begin{aligned}
\left\|\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}^2\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} + \lambda\mathbf{I})^{-1}(I - \tilde{\mathbf{X}}^T\tilde{\mathbf{X}})\boldsymbol{\Sigma}\boldsymbol{\theta}^*\right\| &\leq \left\|\boldsymbol{\Sigma}^2\right\|\left\|(I - \tilde{\mathbf{X}}^T\tilde{\mathbf{X}})\right\|\left\|(\boldsymbol{\Sigma}^2\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} + \lambda\mathbf{I})^{-1}\right\|\left\|\boldsymbol{\Sigma}\boldsymbol{\theta}^*\right\|_2\\
&\leq \max_i \sigma_i^2 \epsilon \frac{1}{\min_i \sigma_i^2(1-\epsilon) + \lambda}\left\|\boldsymbol{\Sigma}\boldsymbol{\theta}^*\right\|_2\\
&\overset{\text{(i)}}{\leq} 2\epsilon\frac{\max_i \sigma_i^2}{\min_i \sigma_i^2 + \lambda}\left\|\boldsymbol{\Sigma}\boldsymbol{\theta}^*\right\|_2
\end{aligned}$$

where we used $1-\epsilon \le \|\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\| \le 1+\epsilon$ and (i) follows by the assumption $\epsilon \le min_i(\sigma_i^2+\lambda)/2$ both of which follow from the conditions of the lemma.

We next bound the norm of the second term in the difference between the residuals. We have,

$$\begin{aligned}
&\left\|\boldsymbol{\Sigma}^2((\boldsymbol{\Sigma}^2\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}+\lambda\mathbf{I})^{-1}-(\boldsymbol{\Sigma}^2+\lambda\mathbf{I})^{-1})(\boldsymbol{\Sigma}\boldsymbol{\theta}^*-\tilde{\mathbf{X}}^T\mathbf{z})\right\| \\
&\qquad\le \left\|\boldsymbol{\Sigma}^2\right\|\left\|(\boldsymbol{\Sigma}^2\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}+\lambda\mathbf{I})^{-1}-(\boldsymbol{\Sigma}^2+\lambda I)^{-1}\right\|\left\|\boldsymbol{\Sigma}\boldsymbol{\theta}^*-\tilde{\mathbf{X}}^T\mathbf{z}\right\|_2 \\
&\qquad\overset{\text{(i)}}{\le} \max_i \sigma_i^2\left\|(\boldsymbol{\Sigma}^2\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}+\lambda\mathbf{I})^{-1}\right\|\left\|\boldsymbol{\Sigma}^2(\mathbf{I}-\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})\right\|\left\|(\boldsymbol{\Sigma}^2+\lambda\mathbf{I})^{-1}\right\|\left\|\boldsymbol{\Sigma}\boldsymbol{\theta}^*-\tilde{\mathbf{X}}^T\mathbf{z}\right\|_2 \\
&\qquad\le \max_i \sigma_i^2\frac{1}{\min_i(\sigma_i^2(1-\epsilon)+\lambda)}\frac{1}{\min_i(\sigma_i^2+\lambda)}\left\|\boldsymbol{\Sigma}^2\right\|\left\|\mathbf{I}-\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right\|\left\|\boldsymbol{\Sigma}\boldsymbol{\theta}^*-\tilde{\mathbf{X}}^T\mathbf{z}\right\|_2 \\
&\qquad\le 2\epsilon\frac{\max_i \sigma_i^4}{\min_i(\sigma_i^2+\lambda)^2}\left(\|\boldsymbol{\Sigma}\boldsymbol{\theta}^*\|_2+\|\tilde{\mathbf{X}}^T\mathbf{z}\|_2\right)
\end{aligned}$$

where the last inequality follows by the assumption $\epsilon \le min_i(\sigma_i^2+\lambda)/2$, and (i) follows by noting that the matrix $\boldsymbol{\Sigma}^2\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}+\lambda\mathbf{I}$ can be viewed as a perturbation of the non-singular matrix $\boldsymbol{\Sigma}^2+\lambda\mathbf{I}$ such that $\boldsymbol{\Sigma}^2\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}+\lambda\mathbf{I} = (\boldsymbol{\Sigma}^2+\lambda\mathbf{I})-\boldsymbol{\Sigma}^2(\mathbf{I}-\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})$, and applying a standard bound from the literature (see [HJ12, Chapter 5, Equation 5.8.1]) on the difference of the inverse of the two matrices. Combining the two bounds yields (10), which concludes the proof.

C.4 Proof of Lemma 2

We have

$$\left\|\tilde{\mathbf{X}}^T\mathbf{z}\right\|_2 = \left|\sum_{l=1}^{d}(\tilde{\mathbf{x}}_l^T\mathbf{z})^2\right|^{1/2} \le \sum_{l=1}^{d}\|\tilde{\mathbf{x}}_l^T\mathbf{z}\|_2$$

Conditioned on $\mathbf{z}$, the random variable $\tilde{\mathbf{x}}_i^T\mathbf{z}$ is zero-mean Gaussian with variance $\|\mathbf{z}\|_2/n$. Thus, $\mathrm{P}\left[|\tilde{\mathbf{x}}_i^T\mathbf{z}| \ge \frac{\|\mathbf{z}\|_2}{\sqrt{n}}\beta\right] \le 2e^{-\beta^2/2}$. Moreover, as provided in (13), with probability at least $1-2e^{-n/8}$, $\|\mathbf{z}\|_2^2 \le 2\sigma^2$. Combining the two with the union bound, we obtain

$$\mathrm{P}\left[|\tilde{\mathbf{x}}_i^T\mathbf{z}|^2 \ge \frac{2\sigma^2}{n}\beta^2\right] \le 2e^{-\beta^2/2}+2e^{-n/8}.$$

Utilizing the union bound again, we obtain

$$\left|\tilde{\mathbf{x}}_l^T\mathbf{z}\right| \le 2\frac{d}{\sqrt{n}}\sigma\beta$$

which holds with probability at least $1-2d(e^{-\beta^2/2}+e^{-n/8})$.

C.5 Proof of Lemma 3

For proving Lemma 3, we follow a similar argument to [HY21, Lemma 3]. We have

$$R(\tilde{\boldsymbol{\theta}}_\lambda) = \sigma^2 + \sum_{i=1}^{d} \sigma_i^2 \underbrace{\left(\sigma_i \theta_i^* \frac{\lambda}{\sigma_i^2 + \lambda} + \frac{\sigma_i}{\sigma_i^2 + \lambda} \tilde{\mathbf{x}}_i^T \mathbf{z}\right)^2}_{Z_i}.$$

Where, $\sum_{i=1}^d Z_i$ corresponds to an off-centered chi-squared distribution with the Z_i. The random variable Z_i, conditioned on $\mathbf{z}$, is a squared Gaussian with variance upper bounded by $\frac{\|\mathbf{z}\|_2}{\sqrt{n}}$ and has expectation

$$\mathbb{E}\left[Z_i\right] = \sigma_i^2 (\theta_i^*)^2 \left(\frac{\lambda}{\sigma_i^2 + \lambda}\right)^2 + \frac{\|z\|_2^2}{n} \left(\frac{\sigma_i}{\sigma_i^2 + \lambda}\right)^2$$

By a standard concentration inequality of sub-exponential random variables (see e.g. [Wai19, Chapter 2, Equation 2.21]), we get, for $\beta \in (0, \sqrt{d})$ and conditioned on $\mathbf{z}$, that the event

$$\mathcal{E}_1 = \left\{ \left| \sum_{i=1}^{d} (Z_i - \mathbb{E}\left[Z_i\right]) \right| \leq \frac{\|\mathbf{z}\|_2^2}{n} \sqrt{d} \beta \right\} \tag{12}$$

occurs with probability at least $1 - 2e^{-\frac{\beta^2}{8}}$. With the same standard concentration inequality for sub-exponential random variables, we have that the event

$$\mathcal{E}_2 = \left\{ \left| \|\mathbf{z}\|_2^2 - \sigma^2 \right| \leq \frac{\sigma^2 \beta}{\sqrt{n}} \right\} \tag{13}$$

also occurs with probability at least $1 - 2e^{-\frac{\beta^2}{8}}$. By the union bound, both events hold simultaneously with probability at least $1 - 4e^{-\frac{\beta^2}{8}}$. On both events, we have that

$$\begin{aligned}
\left| R(\tilde{\boldsymbol{\theta}}^t) - \bar{R}(\tilde{\boldsymbol{\theta}}^t) \right| &= \left| \sum_{i=1}^{d} (Z_i - \mathbb{E}\left[Z_i\right]) + \frac{1}{n} \left(\|\mathbf{z}\|_2^2 - \sigma^2 \sigma_i^2 \right) \left(\frac{\sigma_i}{\sigma_i + \lambda} \right)^2 \right| \\
&\leq \left| \sum_{i=1}^{d} (Z_i - \mathbb{E}\left[Z_i\right]) \right| + \frac{d}{n} \max_i \left[\left(\frac{\sigma_i}{\sigma_i + \lambda} \right)^2 |\|\mathbf{z}\|_2^2 - \sigma^2 \sigma_i^2| \right] \\
&\leq \frac{\|\mathbf{z}\|_2^2}{n} \sqrt{d} \beta + \frac{d}{n} \frac{1}{\sqrt{n}} \sigma^2 \beta \max_i \left[\left(\frac{\sigma_i}{\sigma_i + \lambda} \right)^2 \sigma_i^2 \right] \\
&\leq \frac{2\sigma^2}{n} \sqrt{d} \beta + \frac{d}{n} \frac{1}{\sqrt{n}} \sigma^2 \beta \max_i \left[\left(\frac{\sigma_i}{\sigma_i + \lambda} \right)^2 \sigma_i^2 \right] \\
&\leq \frac{2\sigma^2}{n} \sqrt{d} \beta + \frac{d}{n} \frac{1}{\sqrt{n}} \sigma^2 \beta \max_i \left(\frac{\sigma_i^2}{\sigma_i + \lambda} \right)^2 \\
&\overset{\text{(i)}}{\leq} \frac{\sigma^2}{n} \beta 3 \sqrt{d}.
\end{aligned}$$

where (i) follows from the assumption $d/n \leq \max_i ((\sigma_i + \lambda)/\sigma_i^2)^4$, which concludes the proof of our lemma.

C.6 Proof of Proposition 1

Here, we provide the formal proof for Proposition 1.

Note that we consider the generalized ridge regression problem, but with a diagonal regularization matrix $\mathbf{\Lambda}$ (i.e. Tikhonov regularization). Specifically, $\mathbf{\Lambda}$ is the $\mathbb{R}^{d\times d}$ diagonal matrix containing regularization parameters $\sqrt{\lambda_i}$ pertaining to each different features along its diagonal.

It then directly follows from the proof of Theorem 1 in Section C.2, by simply replacing $\lambda\mathbf{I}$ with $\mathbf{\Lambda}^{1/2}$, that the risk for the above generalized ridge regression problem is well estimated by the following expression:

$$\bar{R}(\tilde{\boldsymbol{\theta}}_{\mathbf{\Lambda}}) = \sigma^2 + \sum_{i=1}^{d} \underbrace{\sigma_i^2\theta_{i,*}^2\left(\frac{\lambda_i}{\sigma_i^2+\lambda_i}\right)^2 + \frac{\sigma^2}{n}\sigma_i^2\left(\frac{\sigma_i}{\sigma_i^2+\lambda_i}\right)^2}_{V_i(\mathbf{\Lambda})}, \tag{14}$$

We consider the set of values $\{\lambda_1,\ldots,\lambda_d\}$ that minimizes the risk expression in (14). Since $\bar{R}(\tilde{\boldsymbol{\theta}}_{\mathbf{\Lambda}})$ contains a summation of terms pertaining to each feature, we take the derivative of $\bar{R}(\tilde{\boldsymbol{\theta}}_{\mathbf{\Lambda}})$ with respect to λ_i:

$$\begin{aligned}
\frac{\partial}{\partial\lambda_i}\bar{R}(\tilde{\boldsymbol{\theta}}_{\mathbf{\Lambda}}) &= \frac{\partial}{\partial\lambda_i}\left(\sigma^2 + \sum_{j=1}^{d} V_j(\mathbf{\Lambda})\right)\\
&= \frac{\partial V_i(\mathbf{\Lambda})}{\partial\lambda_i}\\
&= 2\sigma_i^2\theta_{i,*}^2\left(\frac{\lambda_i}{\sigma_i^2+\lambda_i}\right)\frac{(\sigma_i^2+\lambda_i)-\lambda_i}{(\sigma_i^2+\lambda_i)^2} - 2\frac{\sigma^2}{n}\sigma_i^2\left(\frac{\sigma_i}{\sigma_i^2+\lambda_i}\right)\frac{\sigma_i}{(\sigma_i^2+\lambda_i)^2}\\
&= \frac{2\sigma_i^4\theta_{i,*}^2\lambda_i - 2\sigma^2\sigma_i^4/n}{(\sigma_i^2+\lambda_i)^3}.
\end{aligned}$$

Setting it above to 0, we get

$$\lambda_i = \frac{\sigma^2}{n}\theta_{i,*}^{-2}. \tag{15}$$

Plugging this back into the expression at (14), we get the risk at the optimal scaling as

$$\begin{aligned}
\bar{R}(\tilde{\boldsymbol{\theta}}_{\mathbf{\Lambda}_{opt}}) &= \sigma^2 + \sum_{i=1}^{d}\sigma_i^2\theta_{i,*}^2\frac{\sigma^4}{n^2}\theta_{i,*}^{-4}\left(\frac{1}{\sigma_i^2+\frac{\sigma^2}{n}\theta_{i,*}^{-2}}\right)^2 + \frac{\sigma^2}{n}\sigma_i^2\left(\frac{\sigma_i}{\sigma_i^2+\frac{\sigma^2}{n}\theta_{i,*}^{-2}}\right)^2\\
&= \sigma^2 + \sum_{i=1}^{d}\frac{\sigma^2}{n}\sigma_i^2\left(\frac{\sigma_i}{\sigma_i^2+\frac{\sigma^2}{n}\theta_{i,*}^{-2}}\right)^2\left(\frac{\sigma^2}{n}\theta_{i,*}^{-2}+\sigma_i^2\right)\\
&= \sigma^2 + \frac{\sigma^2}{n}\sum_{i=1}^{d}\frac{\sigma_i^2}{\sigma_i^2+\frac{\sigma^2}{n}\theta_{i,*}^{-2}}.
\end{aligned}$$

C.7 Proof of Proposition 2

Proof of Proposition 2 follows directly by equating the terms in the summation of the risk expression given in (8) for the generalized ridge regression problem and the risk expression of the early-stopped least squares given in (5), as studied in Heckel and Yilmaz [HY21].

It is straightforward to see that the terms inside the respective summations become equal when λ_i are chosen as $\lambda_i = \frac{\sigma_i^2}{1-(1-\eta_i\sigma_i^2)^t} - \sigma_i^2$.

D Details of how double descent occurs outside the linear regime in neural networks

In this section, we discuss in more detail how the individual parameters of a network with p many parameters trained by applying gradient descent with stepsize η to the ℓ_2-regularized least-squares loss with regularization strength λ change across gradient descent iterations.

Note that for an overparameterized network, the network Jacobian $\mathbf{J} \in \mathbb{R}^{n\times p}$ is a wide matrix that typically has full row rank (albeit the small singular values can be very small). Let $\mathbf{J} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ be the singular value decomposition of the Jacobian, where $\mathbf{V} \in \mathbb{R}^{p\times n}$ are the right-singular vectors. Note that only the directions of the parameter vector $\boldsymbol{\theta} \in \mathbb{R}^p$ that align with the right-singular vectors $\mathbf{V}$ impact the predictions of the linear model of the network, however the parameter vector also changes in the directions of the orthogonal complement of the right singular vectors, denoted by $\mathbf{V}_\perp \in \mathbb{R}^{p\times(p-n)}$, due to the ℓ_2-penalty. Specifically, with $\tilde{\mathbf{V}}^T = [\mathbf{V}^T, \mathbf{V}_\perp^T]$, the parameter update $\boldsymbol{\theta}_t$ at gradient iteration t takes the form

$$\begin{aligned}
\boldsymbol{\theta}_t &= \tilde{\mathbf{V}}\left(\mathbf{I} - \eta\begin{bmatrix}\mathbf{\Sigma}^2+\lambda\mathbf{I} & 0\\ 0 & \lambda\mathbf{I}\end{bmatrix}\right)^t \tilde{\mathbf{V}}^T\boldsymbol{\theta}_0 + \eta\sum_{\tau=0}^{t-1}\tilde{\mathbf{V}}\left(\begin{bmatrix}\mathbf{\Sigma}^2+\lambda\mathbf{I} & 0\\ 0 & \lambda\mathbf{I}\end{bmatrix}\right)^\tau \tilde{\mathbf{V}}^T\mathbf{J}^T\mathbf{y}\\
&= \tilde{\mathbf{V}}\left(\mathbf{I} - \eta\begin{bmatrix}\mathbf{\Sigma}^2+\lambda\mathbf{I} & 0\\ 0 & \lambda\mathbf{I}\end{bmatrix}\right)^t \tilde{\mathbf{V}}^T\boldsymbol{\theta}_0 + \mathbf{V}\mathrm{diag}(\ldots,\frac{\sigma_i}{\sigma_i^2+\lambda}(1-(1-\eta(\sigma_i^2+\lambda))^t),\ldots)\mathbf{U}^T\mathbf{y}
\end{aligned}$$

Then, the norm of the change in the parameters that is relevant to fitting the data is

$$\left\|\mathbf{V}^T(\boldsymbol{\theta}_t - \boldsymbol{\theta}_0)\right\|_2^2 = \sum_i^n (1-(1-\eta(\sigma_i^2+\lambda))^t)^2\left(-\frac{1}{\sigma_i}\langle\mathbf{u}_i, \mathbf{J}\theta_0\rangle + \frac{\sigma_i}{\sigma_i^2+\lambda}\langle\mathbf{u}_i,\mathbf{y}\rangle\right)^2. \tag{16}$$

Note that the convergence rate for the above depends primarily on the smallest singular value $\sigma_{\min}$. For a sufficiently small stepsize, we have $(1-\eta(\sigma_i^2+\lambda))^t \approx \exp(-\eta t(\sigma_i^2+\lambda))$, which means that this part converges when $\exp(-\eta t(\sigma_{\min}^2+\lambda))$ gets close to zero. This is the part that is relevant to fitting the data and if initialized appropriately, this change is not more than $O(n)$.

We next consider the change of the coefficient vector that is not relevant to fitting the training data:

$$\begin{aligned}
\left\|\mathbf{V}_\perp^T(\boldsymbol{\theta}_t - \boldsymbol{\theta}_0)\right\|_2^2 &= (1-(1-\eta\lambda)^t)^2\left\|\mathbf{V}_\perp^T\boldsymbol{\theta}_0\right\|_2^2\\
&\approx (1-e^{-\eta\lambda t})^2 O(p).
\end{aligned} \tag{17}$$

Therefore, the change in the coefficients for any λ is on the order of p, and hence is not contained within a small radius around the initialization, where the NTK approximation accurately captures the dynamics of the nonlinear network, unless $1 - e^{-\eta\lambda t}$ is very small (see Figure 7 (left) for an illustration).

In order to observe how this translates to the relationship between the smallest singular value of the network Jacobian $\sigma_{\min}$, and λ, consider the following assumption on $1 - e^{-\eta\lambda t}$ being sufficiently small as parameterized by a small number δ, i.e. $1 - e^{-\eta\lambda t} \leq \delta$. We then have $\lambda \leq \frac{-1}{\eta t} \ln(1-\delta) \approx \frac{\delta}{\eta t}$. Note that we are also interested in the training regime until the network is close to convergence. This occurs when $\exp(-\eta t(\sigma_{\min}^2 + \lambda)) \approx 0$ or $\exp(-\eta t(\sigma_{\min}^2 + \lambda)) \leq \epsilon$ for small ϵ. This in turn leads to the condition $\sigma_{\min}^2 \geq \frac{1/\epsilon - \delta}{\eta t}$.

Based on these conditions on the $\sigma_{\min}$ and λ, in order for the change in the parameters to be confined in a small radius around the network initialization, we need $\sigma_{\min}^2 \gg \lambda$. Based on our empirical observations, in the regime where double descent is observed, λ is much greater than $\sigma_{\min}^2$ and the above condition does not hold.

While in this section we study how the parameters of a network change throughout the training for any λ with respect to a fixed kernel, a similar result was shown for how the associated neural tangent kernel changes across gradient flow time t (iterations) with respect to λ (see [LG20, Theorem 1]). Specifically, Lewkowycz and Gur-Ari [LG20] have shown that, when gradient flow is applied to the ℓ_2-regularized MSE loss, the singular values of the kernel decay exponentially from the initialization with respect to λt, whereas the singular vectors remain static. This is in agreement with our discussion that $\sigma_{\min}^2 \gg \lambda$ is needed for a fixed kernel at initialization to accurately capture the training dynamics of the non-linear network throughout the course of the gradient descent.

Lastly, we show that even for small λ, the linearization (or NTK approximation) is not a good approximation for the network in a setup where regularization-wise double descent occurs. Specifically, when the disparity between the variances across the features of the data is sufficiently large to yield double descent, the change in the parameters of the network is large even for small λ. This can be seen in Figure 7 (right) for a two layer neural network. As indicated by the blue curve here, in the setting where the underlying data structure has differently scaled features and double descent is observed, the parameters change significantly from the initialization early on during the training even at smaller regularization strength. Note that, based on the decay of the kernel, this is not projected to occur until $t \sim 10^3$ for $\lambda = 0.001$ given in this example.

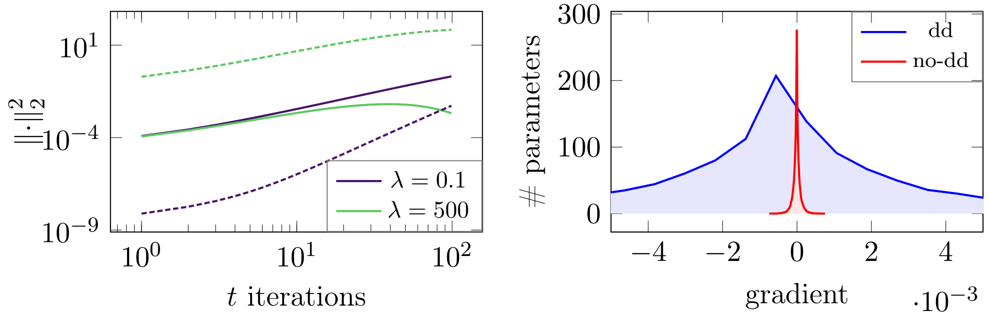

Figure 7: **Left:** The norm of the change in the parameters that is relevant to fitting the data (*solid*) and not relevant to fitting the data (*dashed*) for large and small values of λ. The results show that the parameters primarily change in the directions that are not relevant for fitting the data when λ becomes larger. This moves the neural network outside of the NTK regime (see SM D for details). **Right:** Distribution of the gradients corresponding to the first layer parameters of the network at the first gradient iteration ($t = 1$) for $\lambda = 0.001$. The red curve (scaled back ∼3 times for the sake of visualization) corresponds to the data setup where the difference in the scales of the data features is suppressed, hence resulting in no double descent behavior. The blue curve corresponds to the setting where the features are scaled as discussed before with double descent present as a function of the regularization strength. The results indicate that the dynamics of the network is different from the very beginning for the two regimes even for small λ.