A Glimpse of Double Descent

Introduction

In classical learning theory, increasing model complexity beyond a certain point is expected to cause overfitting and thus higher test error. Yet in modern overparameterized models, we often observe *double descent*: as model complexity grows, the test error first decreases, then spikes dramatically near the point where the model has just enough parameters to fit the training data (the *interpolation threshold*), and finally decreases again as the model becomes even more overparameterized. In this note, we analyze this phenomenon in the simplest setting: ordinary linear regression. We focus purely on the mathematical analysis of double descent in linear models, using only linear algebra and basic probability, and provide precise conditions under which test error will "spike" at the interpolation threshold. All results are presented in a theorem-lemma-proof format. We assume familiarity with the notion that gradient descent in overparameterized linear regression converges to the minimum-norm least-squares solution (the Moore-Penrose pseudoinverse solution).

1 Problem Setup and Notation

Consider a training dataset of N examples $(\mathbf{x}_n, y_n)_{n=1}^N$, where each feature vector $\mathbf{x}_n \in \mathbb{R}^P$ and target $y_n \in \mathbb{R}$. We assemble the feature vectors into a design matrix $X \in \mathbb{R}^{N\times P}$ whose n-th row is $\mathbf{x}_n^T$, and we write $Y \in \mathbb{R}^{N\times 1}$ for the column vector of targets. We assume a linear model for the true data-generating process:

$$y_n = x_n^T\beta^* + \xi_n, \qquad n = 1, \dots, N,$$

where $\beta^* \in \mathbb{R}^P$ is the (unknown) true parameter vector and ξ_n is independent noise with mean 0 and variance σ^2. In vector form,

$$Y = X\beta^* + \xi\,,$$

where $\xi \in \mathbb{R}^N$ is the noise vector (with $\mathbb{E}[\xi] = 0$ and $\mathrm{Cov}(\xi) = \sigma^2 I_N$).

In linear regression, we seek an estimator $\hat{\beta} \in \mathbb{R}^P$ minimizing the training mean squared error. Depending on the relationship between N and P, the solution may or may not be unique. We will use the following terminology:

Definition 1 (Underparameterized vs. Overparameterized)**.** *We say the model is* underparameterized *if $P < N$ (fewer parameters than data points) and* overparameterized *if $P > N$ (more parameters than data points). The* interpolation threshold *is the boundary case $P = N$. In general, when $P \geq N$, a linear model can exactly interpolate (fit) any set of N training points.*

In the underparameterized regime ($P < N$) with full column rank P, the empirical risk minimization problem has a unique solution given by the ordinary least squares (OLS) formula:

$$\hat{\beta}_{\text{under}} = \arg\min_{\beta\in\mathbb{R}^P} \sum_{n=1}^{N} (x_n^T\beta - y_n)^2 \;=\; (X^TX)^{-1}X^TY\,,$$

assuming X^TX is invertible. In the overparameterized case ($P > N$), the least-squares problem is underdetermined (infinitely many interpolating solutions achieve zero training error). In practice, however, gradient descent or any small-norm favoring algorithm will converge to the *minimum ℓ_2-norm* interpolating solution:

$$\hat{\beta}_{\text{over}} = \arg\min_{\beta\in\mathbb{R}^P} \{\|\beta\|_2 : X\beta = Y\}\,.$$

This minimum-norm solution can be written in closed-form using the Moore-Penrose pseudoinverse:

$$\hat{\beta}_{\text{over}} = X^{+}Y = X^{T}(XX^{T})^{-1}Y\,,$$

where $X^{+} \in \mathbb{R}^{P\times N}$ denotes the pseudoinverse of X. (For $P = N$, if X is invertible, this coincides with the usual solution $\hat{\beta} = X^{-1}Y$.) In summary, the estimator we analyze is

$$\hat{\beta} \;=\; \begin{cases} (X^{T}X)^{-1}X^{T}Y, & P \le N \text{ (unique OLS solution)}, \\ X^{T}(XX^{T})^{-1}Y, & P \ge N \text{ (minimum-norm interpolating solution)}\,. \end{cases}$$

Our goal is to analyze the generalization performance of $\hat{\beta}$. We measure this via the test mean squared error (MSE), defined as

$$E_{\text{test}} = E_{\text{new}}\big[\,(y_{\text{new}} - x_{\text{new}}^{T}\hat{\beta})^{2}\,\big]\,,$$

where the expectation is over a new test example $(\mathbf{x}_{\text{new}}, y_{\text{new}})$ drawn from the same distribution as the training data (and independent of the training set). By the law of total expectation (averaging also over the training set and noise in $\hat{\beta}$), we can decompose the expected test error into *bias*, *variance*, and irreducible error components. We formalize this next.

2 Bias–Variance Decomposition of Test Error

First, note that the optimal predictor (Bayes regressor) for this problem is $f^{*}(\mathbf{x}) = \mathbf{x}^{T}\beta^{*}$, which achieves an irreducible error $\mathbb{E}[(y_{\text{new}} - f^{*}(\mathbf{x}_{\text{new}}))^{2}] = \sigma^{2}$ due to the noise ξ. Our estimator $\hat{\beta}$ will generally not exactly equal β^{*}, either due to statistical estimation error or, if $P > N$, due to unidentifiable components of β^{*}. We denote the *bias* of the estimator (as a vector) by

$$\mathbb{E}[\hat{\beta}] - \beta^{*}\,,$$

the difference between the expected estimated parameters and the true parameters, where the expectation is taken over the randomness in the training process (sample noise and/or random sampling of training examples). The *variance* is captured by the covariance matrix $\mathrm{Cov}(\hat{\beta})$. These induce corresponding bias and variance in the predictions on a test point.

The following lemma gives the bias–variance decomposition for the test mean squared error.

Lemma 1 (Bias–Variance Decomposition)**.** *For any fixed test feature vector $\mathbf{x} \in \mathbb{R}^{P}$,*

$$\mathbb{E}\Big[(y - x^{T}\hat{\beta})^{2}\Big] = \sigma^{2} + \underbrace{\big(x^{T}(\mathbb{E}[\hat{\beta}] - \beta^{*})\big)^{2}}_{(bias)^{2}} + \underbrace{x^{T}\,\mathrm{Cov}(\hat{\beta})\,x}_{variance}\,,$$

where the expectation is over the randomness in the training set (including noise) and $y = \mathbf{x}^{T}\beta^{*} + $ (noise) is the true label for $\mathbf{x}$. In particular, σ^{2} is the irreducible noise term, the squared bias term represents error due to any systematic difference between $\hat{\beta}$ and β^{*}, and the variance term represents error due to estimation variance of $\hat{\beta}$.

Proof. By adding and subtracting the optimal prediction $\mathbf{x}^{T}\beta^{*}$ inside the square and expanding, we obtain:

$$(y - x^{T}\hat{\beta})^{2} = \underbrace{(y - x^{T}\beta^{*})^{2}}_{\text{noise + Bayes error}} + (x^{T}\beta^{*} - x^{T}\hat{\beta})^{2} + 2\,(y - x^{T}\beta^{*})\,(x^{T}\beta^{*} - x^{T}\hat{\beta})\,.$$

Taking expectation and using the fact that $\mathbb{E}[\,y \mid \mathbf{x}\,] = \mathbf{x}^{T}\beta^{*}$ and $\mathbb{E}[\,y - \mathbf{x}^{T}\beta^{*} \mid \mathbf{x}\,] = 0$, the cross term vanishes. Thus

$$\mathbb{E}\big[(y - x^{T}\hat{\beta})^{2} \,\big|\, x\big] = \mathbb{E}\big[(y - x^{T}\beta^{*})^{2} \,\big|\, x\big] + \mathbb{E}\big[(x^{T}\beta^{*} - x^{T}\hat{\beta})^{2} \,\big|\, x\big]\,.$$

Since $\mathbb{E}[(y - \mathbf{x}^{T}\beta^{*})^{2} \mid \mathbf{x}] = \mathrm{Var}(y \mid \mathbf{x}) = \sigma^{2}$ by assumption, we have

$$\mathbb{E}[(y - x^{T}\hat{\beta})^{2}] = \sigma^{2} + \mathbb{E}\big[(x^{T}(\beta^{*} - \hat{\beta}))^{2}\big]\,.$$

The second term is the mean squared error in predicting $\mathbf{x}^T\beta^*$ by $\mathbf{x}^T\hat{\beta}$. We can decompose it as

$$\mathbb{E}\big[(x^T(\beta^* - \hat{\beta}))^2\big] = \mathbb{E}\Big[\big(x^T(\beta^* - \mathbb{E}[\hat{\beta}]) + x^T(\mathbb{E}[\hat{\beta}] - \hat{\beta})\big)^2\Big]\,.$$

Expanding this and noting that $\mathbf{x}^T(\beta^* - \mathbb{E}[\hat{\beta}])$ is constant with respect to the inner expectation and $\mathbb{E}\big[x^T(\mathbb{E}[\hat{\beta}] - \hat{\beta})\big] = 0$, we get

$$\mathbb{E}\big[(x^T(\beta^* - \hat{\beta}))^2\big] = \big(x^T(\beta^* - \mathbb{E}[\hat{\beta}])\big)^2 + x^T\,\mathrm{Cov}(\hat{\beta})\,x\,.$$

This is exactly the bias^2 + variance decomposition of the prediction error (conditioned on $\mathbf{x}$). Combining with the irreducible noise σ^2 term yields the stated result. □

Lemma 1 tells us that to understand test error, we need to characterize the bias $\mathbb{E}[\hat{\beta}] - \beta^*$ and the covariance of $\hat{\beta}$. We now derive these for our linear regression estimator. To facilitate this analysis in a unified way for both under- and overparameterized cases, it is useful to work with the singular value decomposition (SVD) of the design matrix X.

3 SVD Analysis of the Estimator

Let the (thin) singular value decomposition of X be

$$X = U\Sigma V^T\,,$$

where $U \in \mathbb{R}^{N\times r}$, $V \in \mathbb{R}^{P\times r}$ have orthonormal columns, and $\Sigma \in \mathbb{R}^{r\times r}$ is diagonal with positive entries. Here $r = \mathrm{rank}(X) = \min\{N, P\}$ (we assume X has full rank in the smaller dimension, as is typical when data are in general position). We write the singular values as $\sigma_1, \sigma_2, \ldots, \sigma_r$ (assumed without loss of generality in nonincreasing order $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$), and denote by $u_i \in \mathbb{R}^N$ and $v_i \in \mathbb{R}^P$ the i-th columns of U and V, respectively (the left and right singular vectors). Note that $\{v_1, \ldots, v_r\}$ forms an orthonormal basis for the column space of X^T (which is $\mathbb{R}^P$ when $P \leq N$, and a subspace of $\mathbb{R}^P$ of dimension r when $P > N$). We define $P_r = VV^T$, which is the orthogonal projector onto the subspace spanned by $\{v_1, \ldots, v_r\}$ (the *row space* of X, equivalently the column space of X^T).

Using this decomposition, one can express the estimator conveniently as follows:

Lemma 2 (Formula for $\hat{\beta}$ via SVD)**.** *For both the underparameterized case ($P \leq N$) and overparameterized case ($P \geq N$), the estimator can be written as*

$$\hat{\beta} = V\Sigma^+U^TY\,,$$

where $\Sigma^+ \in \mathbb{R}^{r\times r}$ is the diagonal matrix with entries σ_i^{-1}. Moreover, this yields the explicit decomposition

$$\hat{\beta} = P_r\beta^* + \sum_{i=1}^{r} \frac{u_i^T\xi}{\sigma_i}\,v_i\,.$$

In particular, $\mathbb{E}[\hat{\beta} \mid X] = P_r\beta^*$.

Proof. Using the SVD $X = U\Sigma V^T$, the normal equations for the minimizer of $\|X\beta - Y\|_2^2$ (with an appropriate minimal-norm constraint if needed) yield the pseudoinverse solution $\hat{\beta} = X^+Y = V\Sigma^+U^TY$. This formula covers both regimes: if $P \leq N$ and X has full column rank, then $r = P$ and indeed $X^+ = (X^TX)^{-1}X^T$; if $P \geq N$ and X has full row rank, then $r = N$ and $X^+ = X^T(XX^T)^{-1}$. Now substitute $Y = X\beta^* + \xi = U\Sigma V^T\beta^* + \xi$. We have

$$U^TY = U^T(U\Sigma V^T\beta^* + \xi) = \Sigma V^T\beta^* + U^T\xi\,,$$

using $U^TU = I_r$. Therefore,

$$\hat{\beta} = V\Sigma^+(\Sigma V^T\beta^* + U^T\xi) = VV^T\beta^* + V\Sigma^+U^T\xi\,.$$

Noting that $VV^T = P_r$ and writing $V\Sigma^+U^T\xi = \sum_{i=1}^r \sigma_i^{-1}(u_i^T\xi)\,v_i$ gives the stated result. Taking expectation over the noise ξ (conditional on X) yields $\mathbb{E}[\hat{\beta} \mid X] = P_r\beta^*$, since $\mathbb{E}[\,u_i^T\xi\,] = 0$ for each i. □

Lemma 2 provides a clear interpretation. The estimator $\hat{\beta}$ is the sum of two parts: (i) a *signal recovery* term $P_r\beta^*$, which is the projection of the true parameter onto the subspace of $\mathbb{R}^P$ that the training data can actually "see" (the row space of X), and (ii) a *noise-fitting* term $\sum_{i=1}^{r} \frac{u_i^T \xi}{\sigma_i} v_i$, which is a linear combination of the right singular vectors weighted by the noise components $u_i^T \xi$ scaled by $1/\sigma_i$. Note that if the model is underparameterized ($P \leq N$ so that $r = P$ and $P_r = I_P$), then $P_r\beta^* = \beta^*$ and the estimator is unbiased ($\mathbb{E}[\hat{\beta}] = \beta^*$). In contrast, in the overparameterized case ($P > N$, so $r = N < P$), P_r is a strict projector (rank N), and $\hat{\beta}$ is generally biased because $P_r\beta^*$ may differ from β^* (the components of β^* in the nullspace of X cannot be recovered and are effectively set to zero by the minimum-norm solution).

Using the above decomposition of $\hat{\beta}$, we can compute the covariance as well:

Lemma 3 (Bias and Variance of $\hat{\beta}$)**.** *For the estimator $\hat{\beta}$, the bias (conditional on X) is*

$$\mathbb{E}[\hat{\beta}] - \beta^* = -(I - P_r)\,\beta^*\,,$$

and the covariance matrix is

$$\mathrm{Cov}(\hat{\beta} \mid X) = \sigma^2 \sum_{i=1}^{r} \frac{1}{\sigma_i^2}\, v_i v_i^T\,.$$

Equivalently, for any direction v_i in the row space of X, $\mathrm{Var}(v_i^T \hat{\beta} \mid X) = \sigma^2/\sigma_i^2$, and for any w orthogonal to all v_i (i.e. w in the nullspace of X), $\mathrm{Var}(w^T \hat{\beta} \mid X) = 0$.

Proof. From Lemma 2, $\mathbb{E}[\hat{\beta} \mid X] = P_r\beta^*$, so $\mathbb{E}[\hat{\beta} \mid X] - \beta^* = (P_r - I)\beta^* = -(I - P_r)\beta^*$. For the covariance, again from Lemma 2 we have

$$\hat{\beta} - \mathbb{E}[\hat{\beta} \mid X] = \sum_{i=1}^{r} \frac{u_i^T \xi}{\sigma_i}\, v_i\,.$$

The coefficients $u_i^T \xi$ are independent zero-mean random variables with $\mathbb{E}[(u_i^T \xi)^2] = \sigma^2$ (because u_i has norm 1). Thus conditional on X, the variance along each v_i direction is σ^2/σ_i^2. In matrix form, since $\{v_1, \ldots, v_r\}$ are orthonormal, this gives

$$\mathrm{Cov}(\hat{\beta} \mid X) = \sum_{i=1}^{r} \mathrm{Var}\Big(\frac{u_i^T \xi}{\sigma_i}\Big) v_i v_i^T = \sum_{i=1}^{r} \frac{\sigma^2}{\sigma_i^2} v_i v_i^T\,,$$

as claimed. (For directions w orthogonal to all v_i, clearly $w^T \hat{\beta} = 0$ always since $\hat{\beta}$ lies in the span of $\{v_i\}$, so the variance is zero in those directions.) □

We now have the necessary ingredients to express the test error in terms of the singular values and singular vectors of X. Consider a fresh test example $\mathbf{x}_{\text{new}} \in \mathbb{R}^P$ with true label $y_{\text{new}} = \mathbf{x}_{\text{new}}^T \beta^* + \xi_{\text{new}}$ (with $\xi_{\text{new}} \sim \mathcal{N}(0, \sigma^2)$ independent of the training noise). We are interested in the mean squared prediction error $\mathbb{E}[(y_{\text{new}} - \mathbf{x}_{\text{new}}^T \hat{\beta})^2]$. Using Lemma 1 and plugging in the bias and covariance from Lemma 3, we obtain:

$$\begin{aligned}
\mathbb{E}\Big[(y_{\text{new}} - \mathbf{x}_{\text{new}}^T \hat{\beta})^2\Big] &= \sigma^2 + \Big(\mathbf{x}_{\text{new}}^T (P_r - I)\beta^*\Big)^2 + \mathbf{x}_{\text{new}}^T\, \mathrm{Cov}(\hat{\beta})\, \mathbf{x}_{\text{new}} \\
&= \sigma^2 + \Big(\mathbf{x}_{\text{new}}^T (P_r - I)\beta^*\Big)^2 + \sigma^2 \sum_{i=1}^{r} \frac{(\mathbf{x}_{\text{new}}^T v_i)^2}{\sigma_i^2}\,. \qquad (1)
\end{aligned}$$

The three terms in (1) correspond to: (i) the irreducible noise σ^2, (ii) the squared bias term due to any part of β^* lying in the nullspace of X (i.e. $(I - P_r)\beta^*$) which the model could not learn, and (iii) the variance term arising from noise in the training data being amplified by the parameter estimation (note this variance term can be equivalently written as $\sum_{i=1}^{r} \frac{\sigma^2}{\sigma_i^2} (\mathbf{x}_{\text{new}}^T v_i)^2$ as shown, or as $\sigma^2\, \mathbf{x}_{\text{new}}^T (X^T X)^{-1} \mathbf{x}_{\text{new}}$ in the underparameterized case, etc.).

We can now understand qualitatively why *double descent* occurs. In the underparameterized regime ($P < N$), the bias term in (1) is typically large (if P is much smaller than the true intrinsic dimensionality of the data, the model is misspecified and underfits), but the variance term is relatively small (since all σ_i are

bounded away from zero when $P \ll N$). As P increases (adding more parameters/features), the bias term decreases (the model can fit the data better, reducing systematic error), but the smallest singular value σ_r also decreases, causing the variance term to increase. The worst case occurs near $P \approx N$, where the design matrix X becomes almost rank-deficient and $\sigma_r \to 0$, causing an enormous explosion in the variance term. Exactly at $P = N$ (if X is exactly square and nonsingular) one can fit the data perfectly with no bias, but the condition number of X is typically very large, so the estimator is extremely sensitive to noise. This is the peak of test error at the interpolation threshold. When P increases beyond N, the variance term actually starts to *decrease* again, meanwhile, the bias remains low. Thus the test error can decrease again for $P > N$. This non-monotonic behavior as a function of P is the double descent phenomenon.