Theory of the Double Descent Phenomena in High-Dimensional Linear Regression

Using the Replica Method from Statistical Mechanics

Charles H. Martin, PhD
Calculation Consulting
San Francisco, CA
charles@calculationconsulting.com

Contents

1 Introduction and Overview **2**

2 Problem Setup and Notation **2**
2.1 Teacher Model (Data Generation) . . . 2
2.2 Student Model (Learning Procedure) . . . 3
2.3 Test Error (Generalization Error) . . . 3
2.4 Large-Dimensional Limit . . . 3

3 Replica Formalism: Partition Function and Replication **4**
3.1 Defining a Partition Function . . . 4
3.2 Replication: k Copies of the System . . . 4
3.3 Explicit Form of $E(\mathbf{w}^a)$ in the Teacher–Student Setup . . . 5

4 Averaging Over the Random Design X **5**
4.1 Notation: The Design Matrix $\mathbf{X}$ . . . 5
4.2 Quadratic Forms in $\mathbf{X}$ . . . 5
4.3 Hubbard–Stratonovich (HS) Transform or Random Matrix Identities . . . 6
4.4 A Sketch of the HS Decoupling Step . . . 6

5 Replica Order Parameters and Saddle-Point Equations **6**
5.1 Defining Overlaps in Replica Space . . . 6
5.2 Self-Consistency from the Free Energy (Saddle Point) . . . 7

6 Ridgeless Limit ($\beta \to \infty$) and the Double Descent **7**

7 Explicit Final Expressions for the Generalization Error **8**
7.1 Schematic Form in the Noise-Free Case . . . 8
7.2 Noise $\sigma^2 > 0$ . . . 8
7.3 General Lessons . . . 8

8 Putting It All Together: Step-by-Step Summary **8**

1 Introduction and Overview

These notes provide a full, detailed, and self-contained derivation of the *double descent* phenomenon in high-dimensional linear regression using the replica method from statistical mechanics. The aim is to show *every step* that goes into a typical replica-style calculation, assuming only that the reader has a basic familiarity with Gaussian integrals, elementary linear algebra, and some broad knowledge of mean-field methods in physics.

The **double descent** effect refers to a non-monotonic dependence of test (generalization) error on model complexity or parameter dimension p. In a simple linear-regression setting, when the number of parameters p is close to the number of data points n, there is a large peak (traditionally viewed as "overfitting") but then, somewhat surprisingly, the test error *goes back down* again as p becomes much larger than n.

We demonstrate how the *replica method* captures this phenomenon by computing the *typical* generalization error in the limit $n, p \to \infty$ at fixed ratio $\alpha = p/n$. We will see that at $\alpha = 1$ the error develops a significant peak (a divergence when noise is absent), signifying the interpolation threshold, followed by a second descent for $\alpha > 1$.

2 Problem Setup and Notation

We focus on a *teacher–student* linear regression problem with random design:

2.1 Teacher Model (Data Generation)

- There is a **true parameter** (teacher) vector $\mathbf{w}^\star \in \mathbb{R}^p$.
- We have n training samples $(\mathbf{x}_\mu, y_\mu)$, labeled by $\mu = 1, 2, \ldots, n$.
- Each $\mathbf{x}_\mu \in \mathbb{R}^p$ is drawn i.i.d. from a Gaussian distribution with mean zero and covariance $\frac{1}{p}\mathbf{I}_p$, i.e.
$$\mathbf{x}_\mu \sim \mathcal{N}\big(\mathbf{0}, \tfrac{1}{p}\mathbf{I}_p\big).$$
- The training label is generated as
$$y_\mu \;=\; \mathbf{x}_\mu^\top \mathbf{w}^\star \;+\; \varepsilon_\mu,$$
where ε_μ is (optional) noise, often taken i.i.d. $\mathcal{N}(0, \sigma^2)$, with σ^2 being the noise variance. In the noise-free case, $\sigma^2 = 0$.

2.2 Student Model (Learning Procedure)

We fit a vector $\mathbf{w} \in \mathbb{R}^p$ to the training data by *ordinary least squares* (OLS). That is, we minimize the sum of squared errors:

$$E(\mathbf{w}) \;=\; \sum_{\mu=1}^{n} \Big(y_\mu \;-\; \mathbf{x}_\mu^\top \mathbf{w} \Big)^2 .$$

The solution that minimizes $E(\mathbf{w})$ is

$$\hat{\mathbf{w}} \;=\; \arg\min_{\mathbf{w}} E(\mathbf{w}).$$

In the under-parameterized regime ($p < n$) and when $\mathbf{X}$ is full column rank, the classical closed-form is

$$\hat{\mathbf{w}} \;=\; \big(\mathbf{X}^\top \mathbf{X}\big)^{-1} \mathbf{X}^\top \mathbf{y},$$

where $\mathbf{X}$ is the $n \times p$ design matrix whose μ-th row is $\mathbf{x}_\mu^\top$ and $\mathbf{y} = (y_1, \ldots, y_n)^\top$.

When $p > n$, $\mathbf{X}^\top \mathbf{X}$ is rank-deficient (or nearly so). However, *ridgeless* regression can still find a solution that perfectly interpolates the data (often given by the Moore–Penrose pseudoinverse).

2.3 Test Error (Generalization Error)

We define the *test error* (or generalization error) by

$$E_{\text{test}}(\hat{\mathbf{w}}) \;=\; \mathbb{E}_{\mathbf{x}}\Big[\big(\mathbf{x}^\top \hat{\mathbf{w}} - \mathbf{x}^\top \mathbf{w}^\star \big)^2 \Big],$$

where the expectation is with respect to a fresh sample $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \frac{1}{p}\mathbf{I}_p)$. In isotropic settings, we often find

$$E_{\text{test}}(\hat{\mathbf{w}}) \;=\; \frac{1}{p} \, \|\hat{\mathbf{w}} \;-\; \mathbf{w}^\star\|^2 \;\;+\;\; (\text{possible extra term if } \sigma^2 > 0).$$

2.4 Large-Dimensional Limit

We consider the high-dimensional limit

$$n \to \infty, \quad p \to \infty, \quad \text{with} \quad \alpha \;=\; \frac{p}{n} \text{ held fixed.}$$

We will see that E_{test} becomes a function of α and other parameters (e.g. the noise variance σ^2). The **replica method** lets us compute the *typical* $E_{\text{test}}(\alpha)$ when $\mathbf{x}_\mu$ and $\mathbf{w}^\star$ are random.

3 Replica Formalism: Partition Function and Replication

3.1 Defining a Partition Function

To analyze the properties of the *minimum* of $E(\mathbf{w})$, we can embed the problem in a statistical-mechanics framework by introducing a "Gibbs measure" at inverse temperature β:

$$Z \;=\; \int \mathrm{d}\mathbf{w}\, \exp\bigl[-\beta\, E(\mathbf{w})\bigr].$$

In the limit $\beta \to \infty$, the integral is dominated by the global minima of $E(\mathbf{w})$. Hence, studying Z (or $\ln Z$) as $\beta \to \infty$ is effectively studying the *behavior of the minimizer* $\hat{\mathbf{w}}$.

However, *we also need to average over the random data* $\{\mathbf{x}_\mu, \varepsilon_\mu\}$. Typically, one writes $\langle Z^k \rangle$ to denote the average of Z^k over the data. Then one tries to evaluate

$$\bigl\langle \ln Z \bigr\rangle \;=\; \lim_{k\to 0} \frac{\langle Z^k \rangle - 1}{k}.$$

This is the **replica trick**: we compute $\langle Z^k \rangle$ for integer k, then analytically continue to $k \to 0$.

3.2 Replication: k Copies of the System

We define

$$Z^k \;=\; \left(\int \mathrm{d}\mathbf{w}\, \exp[-\beta\, E(\mathbf{w})] \right)^k \;=\; \int \prod_{a=1}^{k} \mathrm{d}\mathbf{w}^a\, \exp\Bigl[-\beta \sum_{a=1}^{k} E(\mathbf{w}^a)\Bigr].$$

Hence we have k *replicas* of the weight vector, labeled by $a = 1, 2, \ldots, k$.

Next, we take the average over $\mathbf{X}$ and ε_μ:

$$\langle Z^k \rangle \;=\; \int (\mathrm{d}\mathbf{X})\, (\mathrm{d}\varepsilon)\; p(\mathbf{X}, \varepsilon) \prod_{a=1}^{k} \exp\Bigl[-\beta E(\mathbf{w}^a)\Bigr].$$

We want to handle $p(\mathbf{X}, \varepsilon)$, the probability density of the data. Typically:

- $\mathbf{X}$ has rows $\mathbf{x}_\mu \sim \mathcal{N}(0, \frac{1}{p}\mathbf{I}_p)$ i.i.d.,
- $\varepsilon_\mu \sim \mathcal{N}(0, \sigma^2)$ i.i.d.,
- $y_\mu = \mathbf{x}_\mu^\top \mathbf{w}^\star + \varepsilon_\mu$.

3.3 Explicit Form of $E(\mathbf{w}^a)$ in the Teacher–Student Setup

Recall

$$E(\mathbf{w}^a) \;=\; \sum_{\mu=1}^{n} \big(y_\mu - \mathbf{x}_\mu^\top \mathbf{w}^a\big)^2.$$

Since $y_\mu = \mathbf{x}_\mu^\top \mathbf{w}^\star + \varepsilon_\mu$, we have

$$E(\mathbf{w}^a) = \sum_{\mu=1}^{n} \big(\mathbf{x}_\mu^\top \mathbf{w}^\star + \varepsilon_\mu - \mathbf{x}_\mu^\top \mathbf{w}^a\big)^2.$$

Thus

$$E(\mathbf{w}^a) = \sum_{\mu=1}^{n} \big(\mathbf{x}_\mu^\top (\mathbf{w}^\star - \mathbf{w}^a) + \varepsilon_\mu\big)^2.$$

Hence,

$$\sum_{a=1}^{k} E(\mathbf{w}^a) \;=\; \sum_{a=1}^{k} \sum_{\mu=1}^{n} \Big[\, \mathbf{x}_\mu^\top (\mathbf{w}^\star - \mathbf{w}^a) \;+\; \varepsilon_\mu \Big]^2.$$

4 Averaging Over the Random Design X

4.1 Notation: The Design Matrix X

We have an $n \times p$ matrix $\mathbf{X}$, each row is $\mathbf{x}_\mu^\top$, with

$$\mathbf{x}_\mu \sim \mathcal{N}\big(\mathbf{0}, \tfrac{1}{p}\mathbf{I}_p\big).$$

Hence $x_{\mu i} \sim \mathcal{N}(0, \frac{1}{p})$ i.i.d. The measure for $\mathbf{X}$ (ignoring normalizing constants) is

$$\mathrm{d}\mathbf{X}\, \exp\Big[-\tfrac{p}{2}\, \mathrm{Tr}(\mathbf{X}^\top \mathbf{X})\Big].$$

(More precisely, we have a factor $\exp[-\frac{p}{2} \sum_{\mu,i} x_{\mu i}^2]$, matching variance $\frac{1}{p}$.)

4.2 Quadratic Forms in X

Observe that

$$\sum_{\mu=1}^{n} \Big[\mathbf{x}_\mu^\top (\mathbf{w}^\star - \mathbf{w}^a)\Big]^2 \;=\; \|\mathbf{X}(\mathbf{w}^\star - \mathbf{w}^a)\|^2 \;=\; (\mathbf{w}^\star - \mathbf{w}^a)^\top \mathbf{X}^\top \mathbf{X}\, (\mathbf{w}^\star - \mathbf{w}^a).$$

In the replicated exponent, we have $\sum_{a=1}^{k} \|\mathbf{X}(\mathbf{w}^\star - \mathbf{w}^a)\|^2$.

4.3 Hubbard–Stratonovich (HS) Transform or Random Matrix Identities

To carry out $\int \mathrm{d}\mathbf{X} \exp[\dots]$, one typically uses either:

1. **A known random matrix integral**, e.g. Wishart integrals for large n, p, or
2. **HS transform**, which introduces auxiliary variables to decouple the product in $\mathbf{X}$ from $(\mathbf{w}^\star - \mathbf{w}^a)$.

In either approach, the end result is that *the dependence on* $\mathbf{X}$ *becomes encapsulated by macroscopic order parameters* (like norms and overlaps).

4.4 A Sketch of the HS Decoupling Step

For example, a typical HS formula for the factor

$$\exp\Big[-\sum_{a=1}^{k} \|\mathbf{X}(\mathbf{w}^\star - \mathbf{w}^a)\|^2\Big]$$

is (omitting normalization constants):

$$\exp\Big[-\sum_{a} \|\mathbf{X}(\mathbf{w}^\star - \mathbf{w}^a)\|^2\Big] = \int \prod_{a=1}^{k} \mathrm{d}\mathbf{z}^a \, \exp\Big[-\tfrac{1}{2}\sum_{a} \|\mathbf{z}^a\|^2 \; + \; i \sum_{a} \mathbf{z}^{a\top} \mathbf{X} \, (\mathbf{w}^\star - \mathbf{w}^a)\Big].$$

Then one can integrate over $\mathbf{X}$ row by row, using its Gaussian measure. Finally, we see that the integral organizes itself in terms of the dot products among $(\mathbf{w}^\star - \mathbf{w}^a)$, revealing our *order parameters*.

5 Replica Order Parameters and Saddle-Point Equations

5.1 Defining Overlaps in Replica Space

A standard approach is to define the following *replica-symmetric* order parameters. For $a, b \in \{1, \dots, k\}$,

$$q^a \;=\; \frac{1}{p}\,\|\mathbf{w}^a\|^2, \quad Q^{ab} \;=\; \frac{1}{p}\,\mathbf{w}^a \cdot \mathbf{w}^b, \quad m^a \;=\; \frac{1}{p}\,\mathbf{w}^a \cdot \mathbf{w}^\star.$$

Often, by *replica symmetry* (RS), we have $q^a = q$, $m^a = m$ for all a, and $Q^{ab} = Q$ for $a \neq b$. Then we only need to solve for a small number of scalar variables $\{q, Q, m\}$.

5.2 Self-Consistency from the Free Energy (Saddle Point)

After integrating out $\mathbf{X}$ and ε_μ, one obtains an expression like

$$\langle Z^k \rangle \;\approx\; \int \mathrm{d}\{q, m, Q, \dots\} \; \exp\big[n\,\Phi(q, m, Q, \dots)\big].$$

Then in the large n limit, by a saddle-point approximation, the integral is dominated by the stationary point of Φ. This yields **saddle-point equations**

$$\frac{\partial \Phi}{\partial q} = 0, \quad \frac{\partial \Phi}{\partial m} = 0, \quad \frac{\partial \Phi}{\partial Q} = 0, \; \dots$$

In the limit $k \to 0$, these define the typical (most probable) values of q, m, Q that correspond to the typical solution $\hat{\mathbf{w}}$.

Crucially, from q, m, we get

$$\|\mathbf{w}^a - \mathbf{w}^\star\|^2 \;=\; p\,q + \|\mathbf{w}^\star\|^2 - 2\,p\,m,$$

so the *test error* for each replica is (in isotropic design):

$$E_{\text{test}} \;=\; \frac{1}{p}\,\|\mathbf{w}^a - \mathbf{w}^\star\|^2 \;=\; q + \frac{\|\mathbf{w}^\star\|^2}{p} \;-\; 2\,m.$$

Hence, once the saddle-point $(q, m, Q, \dots)$ is found, we can read off $E_{\text{test}}(\alpha)$.

6 Ridgeless Limit ($\beta \to \infty$) and the Double Descent

In ridgeless (or ℓ_2-regularization-free) regression, we effectively take $\beta \to \infty$, so that the partition function Z is dominated by the *global minimum* of $E(\mathbf{w})$. In that case, the replica analysis simplifies. One typically finds:

$$\min_{\mathbf{w}} E(\mathbf{w}) \;=\; \begin{cases} \text{some finite value} & (p < n), \\ 0 & (p > n, \text{ i.e. the data can be perfectly fit}). \end{cases}$$

The self-consistent equations reveal that:

- For $\alpha = p/n < 1$, the model is under-parameterized; classical least-squares formulas suggest a certain decreasing function of α for the test error.
- As $\alpha \to 1$, $\mathbf{X}^\top\mathbf{X}$ becomes nearly singular, the variance of $\hat{\mathbf{w}}$ blows up, and the test error *peaks*.
- For $\alpha > 1$, the solution *interpolates* perfectly, yet the *typical* generalization error begins to decrease again.

This yields the **double descent** curve: a large spike around $\alpha = 1$, with lower error on both sides.

7 Explicit Final Expressions for the Generalization Error

7.1 Schematic Form in the Noise-Free Case

When $\sigma^2 = 0$, a well-known result from random matrix theory / replica for (ridgeless) linear regression states that the typical test error has a divergence near $\alpha = 1$. A classic form is:

$$E_{\text{test}}(\alpha) \;=\; \begin{cases} \dfrac{1}{1-\alpha} & \alpha < 1, \\[2ex] \dfrac{1}{\alpha - 1} & \alpha > 1, \end{cases}$$

modulo prefactors depending on $\|\mathbf{w}^\star\|^2$ (and ignoring additive constants). This simple formula is not the full universal expression but captures the *qualitative* blow-up around $\alpha = 1$ and the descent afterward. More refined versions incorporate the exact norm $\|\mathbf{w}^\star\|^2$ and other details.

7.2 Noise $\sigma^2 > 0$

With nonzero noise, $\sigma^2 > 0$, the peak remains but is finite; the error does not literally diverge. The essence is the same: as α crosses 1, we see a large bump in the test error, then a second descent for $\alpha > 1$.

7.3 General Lessons

Replica theory yields closed-form or semi-closed-form solutions for

$$E_{\text{test}}(\alpha) \;=\; \left\langle \tfrac{1}{p} \|\hat{\mathbf{w}} - \mathbf{w}^\star\|^2 \right\rangle,$$

averaged over random design. One finds that:

1. E_{test} is typically a non-monotonic function of α.
2. A large spike (first descent $\to$ high peak $\to$ second descent) appears around $\alpha = 1$.
3. For large $\alpha \gg 1$, the solution is heavily over-parameterized but paradoxically can generalize better in a *typical-case* sense than solutions at moderate α.

8 Putting It All Together: Step-by-Step Summary

Here we collect the derivation in a *logical order* without skipping:

1. **Define the regression problem**:

$$y_\mu = \mathbf{x}_\mu^\top \mathbf{w}^\star + \varepsilon_\mu, \quad E(\mathbf{w}) = \sum_{\mu=1}^{n} (\mathbf{x}_\mu^\top \mathbf{w}^\star + \varepsilon_\mu - \mathbf{x}_\mu^\top \mathbf{w})^2.$$

2. **Partition function**:

$$Z = \int \mathrm{d}\mathbf{w}\, \exp[-\beta\, E(\mathbf{w})].$$

3. **Replicate** k times:

$$Z^k = \int \prod_{a=1}^{k} \mathrm{d}\mathbf{w}^a\, \exp\Big[-\beta \sum_{a=1}^{k} E(\mathbf{w}^a)\Big].$$

4. **Average over data** $\mathbf{X}, \varepsilon$:

$$\langle Z^k \rangle = \int \mathrm{d}\mathbf{X}\, \mathrm{d}\varepsilon\; p(\mathbf{X}, \varepsilon)\, \big[\dots\big].$$

   Here

$$p(\mathbf{X}) \;\propto\; \exp\big[-\tfrac{p}{2} \operatorname{Tr}(\mathbf{X}^\top \mathbf{X})\big],$$

   assuming $x_{\mu i} \sim \mathcal{N}(0, \frac{1}{p})$, and similarly for ε_μ.

5. **Rewriting the cost**:

$$\sum_{a=1}^{k} E(\mathbf{w}^a) = \sum_{a=1}^{k} \sum_{\mu=1}^{n} \Big(\mathbf{x}_\mu^\top (\mathbf{w}^\star - \mathbf{w}^a) + \varepsilon_\mu\Big)^2.$$

6. **Decoupling the integrals**: Use either the Hubbard–Stratonovich transformation or known random-matrix expansions. This is where one obtains the expression in terms of *overlaps* among $(\mathbf{w}^\star - \mathbf{w}^a)$.

7. **Introduce order parameters** m^a, q^a, Q^{ab} for the norms and overlaps in replica space. Under *replica symmetry*, reduce to m, q, Q.

8. **Saddle-point approximation**: In the large-n, p limit, $\langle Z^k \rangle$ is dominated by the extremum of an exponent with $n\, \Phi(\dots)$, leading to self-consistent equations for (m, q, Q).

9. **Take** $k \to 0$ **limit**: Obtain $\langle \ln Z \rangle$ from $\langle Z^k \rangle$. In the $\beta \to \infty$ limit (ridgeless), interpret solutions that correspond to the *global* minima.

10. **Result for test error**:

$$E_{\text{test}} = \frac{1}{p}\, \|\hat{\mathbf{w}} - \mathbf{w}^\star\|^2 = q + \frac{\|\mathbf{w}^\star\|^2}{p} - 2\, m$$

    (plus possibly a term if $\sigma^2 > 0$).

11. **Double Descent**: The solution for $E_{\text{test}}(\alpha)$ has a characteristic peak near $\alpha = 1$ and lower values for $\alpha \ll 1$ or $\alpha \gg 1$.

9 Conclusion and Interpretation

The **replica method** provides a systematic tool to compute *typical-case* performance of high-dimensional regression under random design. By:

- Defining a partition function from the cost,
- Replicating and averaging over $\mathbf{X}$,
- Introducing order parameters (overlaps) in replica space,
- Solving the saddle-point equations in the large-n, p limit,

we derive a closed-form expression (or set of self-consistent equations) for the *mean generalization error* $\langle E_{\text{test}} \rangle$ as a function of $\alpha = p/n$.

This computation shows a *peak* or *divergence* around $\alpha = 1$, capturing the so-called **double descent** phenomenon: a first descent for $\alpha < 1$, a large spike near $\alpha = 1$, and a second descent for $\alpha > 1$. This result *cannot* be seen through purely classical bias–variance trade-off arguments but arises naturally from the replica/typical-case analysis (or equivalently from advanced random-matrix approaches).

Key References for Further Reading:

- H. Nishimori, *Statistical Physics of Spin Glasses and Information Processing*, Oxford, 2001.
- M. Mézard and A. Montanari, *Information, Physics, and Computation*, Oxford, 2009.
- Works of Manfred Opper and coauthors from the 1990s and 2000s on statistical mechanics of learning.
- The more recent wave of papers on "double descent" in linear models and neural networks (e.g. Belkin, Hsu, and Mitra, 2019).

End of Derivation.