Double Descent and High-Dimensional Orthogonality

Overview of Double Descent

- As model complexity or feature dimension p increases, test error shows: descent $\rightarrow$ peak $\rightarrow$ second descent.
- Commonly observed in linear regression when increasing number of features p.
- Peak at $p \approx n$ (interpolation threshold): $X^\top X$ nearly singular, variance explosion.
- For $p \gg n$, minimum-norm solution is selected; high-dimensional orthogonality reduces variance $\rightarrow$ second descent.

Variance Explosion and Reduction in Linear Regression

- Model: $y = X\beta + \varepsilon,\ \varepsilon \sim \mathcal{N}(0, \sigma^2 I)$.
- $p < n$: $\hat{\beta} = (X^\top X)^{-1} X^\top y$. As $p \to n$, smallest eigenvalue of $X^\top X$ shrinks $\to$ variance increases.
- $p > n$: infinitely many solutions; gradient descent and least squares tend to pick the minimum-norm one (implicit regularization).
- In high dimensions, new features are nearly orthogonal to existing feature space, keeping coefficient norms small.

2D Case ($p = 2$): Uniform Around a Circle

- Random points on a unit circle (radius 1) are uniformly distributed in direction over $[0^\circ, 180^\circ]$.
- Fix the first vector pointing to the right (0°).
- The probability the second vector lies within $90^\circ \pm 10^\circ$:

$$\frac{20^\circ}{180^\circ} = \frac{1}{9} \approx 0.111.$$

- Right angles occur, but acute and obtuse angles are equally common.

3D Case ($p = 3$): Equatorial Band Advantage

- Points are uniformly distributed on the surface of a unit sphere (S^2).
- Surface area element:

$$dA = R^2 \sin\theta \, d\theta \, d\phi$$

  (θ: polar angle).
- Area of a latitude band between θ and $\theta + d\theta$:

$$A(\theta, \theta + d\theta) = \int_0^{2\pi} R^2 \sin\theta \, d\phi \, d\theta = 2\pi R^2 \sin\theta \, d\theta.$$

- $\sin\theta$ is maximized at $\theta = \pi/2$ (equator) $\Rightarrow$ equatorial band has the largest area.

Angle Concentration in 3D

- PDF of the angle $\theta \in [0, \pi]$:

$$f_3(\theta) = \frac{1}{2} \sin \theta.$$

- Probability of $90^\circ \pm 10^\circ$:

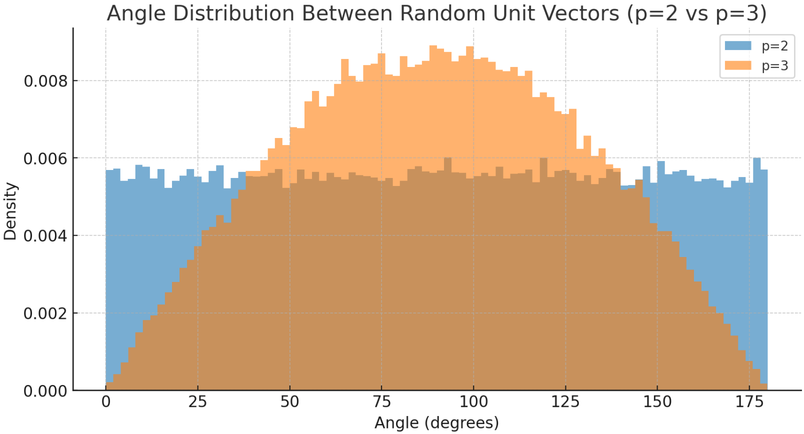
$$\int_{80^\circ}^{100^\circ} \frac{1}{2} \sin \theta \, d\theta = \frac{1}{2}(\cos 80^\circ - \cos 100^\circ) \approx 0.1736,$$

  larger than 0.111 in 2D.
- Equator's area dominance directly translates to higher probability of near-orthogonal angles.

Intuitive Comparison

- 2D: Directions are uniform on a circle; 90° is not special.
- 3D: Directions on a sphere; most of the surface lies near the equator, so angles cluster near 90°.
- As dimension increases, "right angle" becomes the norm.

Empirical Angle Distributions (p=2 vs p=3)



- $p = 2$: Almost uniform over angles.
- $p = 3$: Peak near 90°, low near $0^\circ, 180^\circ$.
- Generated by many random unit vectors.

Generalization to p Dimensions

- Angle PDF on the $(p-1)$-sphere:

$$f_p(\theta) = C_p \sin^{p-2}\theta, \quad C_p = \frac{\Gamma(\frac{p}{2})}{\sqrt{\pi}\,\Gamma(\frac{p-1}{2})}.$$

- As p increases, $\sin^{p-2}\theta$ peaks sharply at $\theta = \pi/2$, concentrating mass near 90°.
- Approx.: $\cos\theta \sim \mathcal{N}(0, 1/p)$; variance shrinks as $1/p$.

Probability of $90^\circ \pm 10^\circ$ for Various p

p	$\mathrm{P}(\lvert\theta - 90^\circ\rvert \leq 10^\circ)$
2	0.1111
3	0.1736
4	≈ 0.2200
10	≈ 0.3904
100	≈ 0.9175

- Higher $p \Rightarrow$ almost all pairs are near-orthogonal.
- At $p = 100$, almost everything lies within $90^\circ \pm 10^\circ$.

High-Dimensional Orthogonality and Double Descent

- At $p \approx n$: $X^\top X$ ill-conditioned, variance explodes (peak).
- For $p \gg n$: New features are nearly orthogonal to existing space. Minimum-norm solution keeps coefficient norm small.
- Orthogonality reduces noise amplification, lowering variance $\rightarrow$ second descent.

Practical Note for Real Data

- Real data populations often non-isotropic (latent factor correlations) $\rightarrow$ orthogonality effect weaker.
- Whitening (PCA/ZCA), ICA, or self-supervised learning can promote isotropy.
- Large latent dimension in intermediate layers + normalization/decorrelation regularizers can help.

Liu's Double Descent and the Hyper-High-Dimensional Factor Hypothesis

Qingfeng Liu

Background of the Hypothesis

- Real-world phenomena are determined by a vast number of nearly independent **hyper-high-dimensional factors**.
- Observable features are limited and cannot fully capture these underlying factors directly.
- Prediction has two main strategies:
  1. Reconstruct the hyper-high-dimensional factors from the features, then predict using them.
  2. If reconstruction is impossible, approximate the mapping with a complex function.

Why So Many Parameters Are Needed

1. **Increased Basis for High-Dimensional Representation** To represent independent factors, we need many orthogonal basis vectors, directly increasing parameter count.
2. **Curse of Dimensionality in Nonlinear Approximation** Capturing factor interactions requires deep networks or a large number of nodes.
3. **Reconstruction of Compressed Information** Observed features are projections of the original factors, and a high degree of model freedom is required to recover lost information.

Connection to Double Descent

- At $p \approx n$ (number of features close to sample size), $X^\top X$ becomes ill-conditioned, variance explodes (first peak).
- In the $p \gg n$ regime, new features are almost orthogonal to the existing space, keeping coefficient norms small (implicit regularization).
- Once the model has enough parameters to approximate the hyper-high-dimensional factors, test error enters the second descent.

Liu's Hypothesis (Summary)

Core Idea

Real-world phenomena consist of hyper-high-dimensional independent factors.
To predict from a finite set of observed features, we need a large number of parameters to reconstruct or approximate the factor space.

- High-dimensional orthogonality enables variance reduction in the $p \gg n$ regime.
- The second descent aligns with achieving sufficient factor reconstruction.
- For real data, preprocessing (whitening, ICA, etc.) can enhance factor independence.

Double Descent: Second Descent Essence and Replica-Trick Assumptions

August 11, 2025

Introduction

- This is precisely the essence of the **second descent** in double descent.
- Increasing capacity (number of parameters or feature dimension p) can improve generalization due to the mechanisms detailed next.

Mechanism (1): Interpolation Threshold

- Near $p \approx n$ (features $\approx$ samples): $X^\top X$ is nearly singular (ill-conditioned) $\Rightarrow$ **variance explosion** $\Rightarrow$ test error peaks (first peak).
- For $p > n$: the solution is non-unique; gradient descent / least squares tend to the **minimum-norm solution** (implicit regularization).

Mechanism (2): Near-Orthogonality in High Dimensions

- With very large feature dimension p, new feature vectors are **almost orthogonal** to the span of existing ones.
- This suppresses the injection of spurious noise into coefficient estimates $\Rightarrow$ estimator variance decreases.
- As capacity increases further, overfitting becomes less likely and test error drops again.

Mechanism (3): Positive Effect of Larger Capacity

- Models with many parameters can cover function families closer to the true mapping.
- Combined with high-dimensional near-orthogonality, this yields **high expressivity with low variance**.

Mechanism (4): Intuitive Flow

Capacity increase $\Rightarrow$ Overfitting peak at $p \approx n$ $\Rightarrow$ High-dimensional orth

Universal Approximation vs. Practice

Can "not-so-deep" models approximate complex functions?

- **Universal Approximation Theorem**: with non-linear activations, a single hidden layer of sufficient width can approximate any continuous function.
- In practice: required width can be enormous; optimization can be unstable; sample complexity can be high.
- **Depth buys efficiency**: hierarchical composition often reduces parameters for the same accuracy.

Shallow vs. Deep in Practice

- **Shallow can suffice**: smooth/low-frequency targets with weak interactions; strong inductive bias aligned with the task.
- **Deep is preferable**: non-smooth, multi-scale, high-order interactions (especially in high p).
- Deep nets can form large, quasi-isotropic latent spaces internally, leveraging near-orthogonality.

Angle Concentration: 2D vs 3D (Intuition)

- **2D**: directions uniform on a circle $\Rightarrow 90^\circ$ is not special.
- **3D**: sphere surface area element $dA = R^2 \sin\theta \, d\theta \, d\phi$ peaks at the equator ($\theta = \pi/2$).
- On S^{p-1}: angle pdf $f_p(\theta) = C_p \sin^{p-2}\theta \Rightarrow$ mass concentrates near 90° as p grows.

$$C_p = \frac{\Gamma(\frac{p}{2})}{\sqrt{\pi}\,\Gamma(\frac{p-1}{2})}, \qquad \cos\theta \approx \mathcal{N}\left(0, \frac{1}{p}\right).$$

(Optional) Empirical Angle Distributions

- $p = 2$: near-uniform over $[0^\circ, 180^\circ]$.
- $p = 3$: strong peak near 90°; $0^\circ/180^\circ$ are rare.

Thermodynamic Limit & Replica Trick (Overview)

- **Thermodynamic limit**: $n \to \infty$, $p \to \infty$, ratio $\alpha = p/n$ fixed.
- **Replica trick**: compute $\mathbb{E}[\log Z]$ via $\mathbb{E}[\log Z] = \lim_{m \to 0} \frac{\mathbb{E}[Z^m] - 1}{m}$.
- Yields analytic error curves matching large-scale simulations: reproduces the first peak and the second descent.

Replica Assumptions (1): Data Distribution

- Samples $x_i \in \mathbb{R}^p$ are i.i.d.
- Typically isotropic Gaussian:

$$x_i \sim \mathcal{N}(0, I_p),$$

  enabling clean high-dimensional geometry (near-orthogonality) and analyzable random-matrix spectra.
- Some works allow known, diagonalizable $\Sigma \neq I_p$ under mild spectral conditions.

Replica Assumptions (2): Label Generation

- Linear teacher–student model:

$$y_i = x_i^\top \beta^* + \epsilon_i,$$

where ϵ_i is Gaussian noise, independent of x_i.

- β^* often assumed i.i.d., zero-mean (Gaussian for tractability).

Replica Assumptions (3): Parameter Scaling

- Thermodynamic limit: $n \to \infty, \ p \to \infty$ with fixed $\alpha = p/n$.
- Enables tools like the Marčenko–Pastur distribution to describe eigenvalue spectra.

Replica Assumptions (4): Learning Algorithm

- Typically least squares (possibly ridge-regularized).
- Or gradient descent converging to the **minimum-norm solution**.
- Quadratic losses/penalties ensure closed-form expectations.

Replica Assumptions (5): Mathematical Technique

- Assume the validity of the replica limit exchange:

$$\mathbb{E}[\log Z] = \lim_{m \to 0} \frac{\mathbb{E}[Z^m] - 1}{m}.$$

- **Replica Symmetry (RS)** assumed; when RS breaks, solutions become more involved.

Summary of Assumptions and Scope

- Analytic formulas rely mainly on:
  1. High-dimensional limit + isotropic Gaussian (or rotation-invariant) features.
  2. Simple solvable estimators (linear/ridge; minimum-norm bias).
- If real data violate these (strong correlations, heavy tails, nonlinearities), treat the analytic curve as an *approximation/guide*.

Operational Tips: Using the Second Descent Safely

- Standardize/whiten features; reduce correlations; monitor the spectrum/condition of $X^\top X$.
- Expect a peak near $p \approx n$; in $p \gg n$, leverage the minimum-norm bias.
- Encourage near-orthogonality: larger latent p, normalization, decorrelation regularizers.
- Choose capacity to cover the function class; control variance via explicit/implicit regularization and early stopping.

One-Page Recap: Why Capacity Can Help

1. Crossing the interpolation threshold $\Rightarrow$ minimum-norm solutions dominate.
2. High-dimensional near-orthogonality suppresses variance.
3. Larger capacity better matches the target function class.

$\Rightarrow$ **Second descent**: test error decreases again as capacity increases.