

# Generalization Bounds

## Theoretical Foundations of Deep Learning

Matteo Mazzarelli

December 17, 2024



# Introduction

# Motivation

- ▶ **Core Question:** How can models trained on limited data perform reliably on unseen scenarios?
- ▶ **Generalization** is a fundamental goal in machine learning: ensuring models extend their learned patterns to new, unseen data.
- ▶ A poorly generalized model risks:
  - ▶ **Overfitting:** Performing well on training data but poorly on unseen data.
  - ▶ **Underfitting:** Failing to capture the underlying patterns of the data.

# The Learning Problem

- ▶ **Supervised Learning:**
  - ▶ Goal: Learn a function  $f : X \rightarrow Y$  mapping inputs  $X$  to outputs  $Y$  based on labeled training data.
- ▶ **Key Question:** Can the learned function perform well on unseen data?
- ▶ **Generalization:**
  - ▶ Ability of a model to extend its learning beyond the training data.
  - ▶ **Central Problem** in machine learning: balancing *empirical performance* with *future predictions*.

# Overfitting

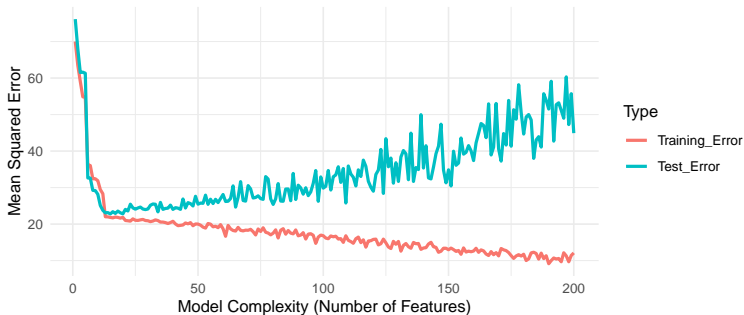
# Simulating Overfitting

- ▶ **Objective:**
  - ▶ Visualize the impact of model complexity on overfitting in a linear regression model.
- ▶ **Dataset**
  - ▶ The experiment uses the **Boston Housing dataset**, where the target variable is `medv` (median value of owner-occupied homes), and the 13 features represent housing characteristics.

- ▶ **Experimental Setup:**
  - ▶ **Model Complexity:**
    - ▶ Complexity is defined by the number of features included in the model.
    - ▶ Additional random features are generated to simulate increasing complexity beyond the real features in the dataset.
  - ▶ **Train-Test Split:**
    - ▶ The dataset is split into 70% training and 30% test data.
  - ▶ **Range of Complexity:**
    - ▶ Models are trained with 1 to 200 features. Beyond the actual features in the dataset, random noise features are added incrementally.

- ▶ **Procedure:** For each level of complexity:
  1. A subset of features (real and random) is used to train a linear regression model.
  2. Predictions are made on both the training and test datasets.
  3. Mean Squared Errors (MSE) are calculated for both datasets.
- ▶ **Results:**
  - ▶ Training error decreases consistently as model complexity increases.
  - ▶ Test error initially decreases but then increases, demonstrating the overfitting phenomenon.
- ▶ **Visualization:**
  - ▶ A line plot shows the relationship between model complexity (number of features) and mean squared error for both the training and test datasets.





**Figure 1:** Overfitting Phenomenon in Linear Regression

► **Highlights:**

- The **bias-variance tradeoff**.
- The point where overfitting begins, indicated by the divergence of training and test errors.

## Key Insights

- ▶ Increasing model complexity without consideration of the underlying data structure can lead to overfitting.
- ▶ Simple models that focus on the true underlying pattern often generalize better to unseen data.

# Double Descent

- ▶ However, modern machine learning introduces a fascinating twist: **Double Descent**, where increasing model complexity can sometimes lead to improved generalization after an initial overfitting phase.
- ▶ Unlike traditional models where increasing complexity leads to overfitting, further increasing the complexity (e.g., using overparameterized neural networks) can eventually reduce the generalization error after an initial peak.
- ▶ This challenges the classical view of overfitting and highlights the complex relationship between model complexity and generalization in modern machine learning.

## Classical Bounds

# Introducing Generalization Bounds

- ▶ **What Are Generalization Bounds?**
  - ▶ Theoretical tools offering guarantees about a model's performance on unseen data.
  - ▶ Relate:
    - ▶ **Generalization Error**: How well the model performs on unseen data.
    - ▶ **Empirical Risk**: Performance observed on training data.
    - ▶ **Model Complexity**: How expressive the model is.

► **Purpose:**

- Provide insights into the trade-offs between:
  - **Model Accuracy:** How well the model captures the data patterns.
  - **Model Complexity:** The expressiveness of the model and its ability to fit intricate patterns.
  - **Training Data Size:** How much data is required to achieve reliable generalization.

# Hoeffding's Inequality: A Starting Point

- ▶ **What is Hoeffding's Inequality?**
  - ▶ A fundamental result in probability theory used to bound the difference between the **empirical risk** and the **generalization error** for a fixed hypothesis.
  - ▶ Provides a way to measure how closely a model's performance on training data reflects its performance on unseen data.

# Mathematical Formulation of Hoeffding's Inequality

## ► Hoeffding's Inequality:

$$P(|R(h) - R_{\text{emp}}(h)| > \varepsilon) \leq 2 \exp(-2m\varepsilon^2)$$

- $R(h)$ : Generalization error (true performance on unseen data).
- $R_{\text{emp}}(h)$ : Empirical risk (error on training data).
- $\varepsilon$ : A small positive value (tolerance).
- $m$ : Size of the dataset.



## Key Insights

- ▶ The probability that the generalization error  $R(h)$  deviates significantly from the empirical risk  $R_{\text{emp}}(h)$  decreases **exponentially** with:
  - ▶ Larger dataset size  $m$ .
  - ▶ Smaller tolerance  $\varepsilon$ .

# Rates of Convergence

## ► What Are Rates of Convergence?

- Quantify how quickly the generalization error approaches the empirical risk as the dataset size  $m$  grows.
- Provide a guideline for determining the dataset size needed to achieve a desired level of generalization.
- In Hoeffding's inequality:

$$P(|R(h) - R_{\text{emp}}(h)| > \varepsilon) \leq 2 \exp(-2m\varepsilon^2)$$

- The **exponential term**  $\exp(-2m\varepsilon^2)$  shows that the convergence is faster with larger datasets.

## ► Key Factors:

- **Dataset Size ( $m$ )**: Larger datasets reduce the gap between  $R(h)$  and  $R_{\text{emp}}(h)$  more quickly.
- **Tolerance ( $\varepsilon$ )**: Smaller tolerances require larger datasets for the same level of confidence.

# Interpretation of Hoeffding's Inequality

## ► What Does It Mean?

- As the dataset size ( $m$ ) increases, the empirical risk becomes a more reliable indicator of the generalization error.
- For a fixed hypothesis, we can be confident that the performance observed on training data is close to what can be expected on unseen data.
- The **rate of convergence** shows how quickly this reliability improves as  $m$  grows.

## Key Insights

- ▶ Hoeffding's inequality gives a **quantitative guarantee** about the relationship between training performance and unseen data performance.
- ▶ Understanding convergence rates helps in planning how much data is needed for robust generalization.

# Limitations of Hoeffding's Inequality

- ▶ **Beyond Fixed Hypotheses:**
  - ▶ Hoeffding's inequality assumes a single, fixed hypothesis.
  - ▶ In practice, machine learning involves selecting the best hypothesis from a **large hypothesis class**  $\mathcal{H}$ , increasing the risk of overfitting.
- ▶ **Need for Complexity-Aware Bounds:**
  - ▶ Simple bounds like Hoeffding's do not consider the complexity of the hypothesis class, which influences generalization.

# The Union Bound

- ▶ **What is the Union Bound?**
  - ▶ A probability tool used to extend bounds like Hoeffding's inequality to apply across an entire hypothesis space  $\mathcal{H}$ .
  - ▶ Helps estimate the probability that **at least one hypothesis** in  $\mathcal{H}$  has a large generalization gap.
- ▶ **Key Idea:**
  - ▶ Instead of considering a single fixed hypothesis, the Union Bound aggregates the probabilities of generalization gaps over all hypotheses in  $\mathcal{H}$ .

# Formalization of The Union Bound

## ► Mathematical Expression:

$$P\left(\sup_{h \in \mathcal{H}} |R(h) - R_{\text{emp}}(h)| > \epsilon\right) \leq \sum_{h \in \mathcal{H}} P(|R(h) - R_{\text{emp}}(h)| > \epsilon)$$

- $\sup_{h \in \mathcal{H}}$ : The supremum ensures we account for the worst-case scenario across all hypotheses.
- $P(|R(h) - R_{\text{emp}}(h)| > \epsilon)$ : The probability of a significant generalization gap for each hypothesis.
- **How It Works:**
  - By summing up the probabilities for all hypotheses, the Union Bound provides a way to analyze the worst-case scenario over the hypothesis space.

# Implications of The Union Bound

- ▶ **Impact of Hypothesis Space Size:**
  - ▶ The bound depends directly on the **size of the hypothesis space**  $|\mathcal{H}|$ .
  - ▶ Larger hypothesis spaces increase the sum, making the bound looser.
- ▶ **Takeaway:**
  - ▶ The Union Bound highlights a trade-off:
    - ▶ **Small hypothesis space:** Tighter bounds, but limited model capacity.
    - ▶ **Large hypothesis space:** Higher capacity, but risk of overfitting and looser bounds.



# Transition to Advanced Bounds

- ▶ **Connection to Practical Learning:**
  - ▶ In practice, hypothesis spaces are often infinite or too large to enumerate explicitly. This motivates the need for alternative ways to measure hypothesis complexity.
- ▶ **From Simple to Sophisticated:**
  - ▶ The Union Bound provides a conceptual basis for understanding how hypothesis space size affects generalization.
  - ▶ Next, we delve into **complexity measures** that allow us to extend generalization bounds to more practical, infinite hypothesis spaces.

## Advanced Bounds

# Motivation for Advanced Bounds

- ▶ **Advanced bounds** address a variety of the limitations we have outlined by incorporating:
  - ▶ **VC Dimension:** A measure of the capacity or expressiveness of a hypothesis class. Higher VC dimensions indicate more complex models, which may require more data to generalize well.
  - ▶ **Rademacher Complexity:** A data-dependent measure of how well a hypothesis class can fit random noise in the training data. It captures both the hypothesis class and the specifics of the data distribution.

## ► Extending Convergence Rates:

- Advanced bounds refine the rates of convergence by linking the generalization error to:
  - The size of the dataset  $m$ .
  - The complexity of the hypothesis class (e.g., **VC dimension** or **Rademacher complexity**).
- For example, the generalization error is often bounded as:

$$R(h) - R_{\text{emp}}(h) \leq \mathcal{O} \left( \sqrt{\frac{\text{Complexity}(\mathcal{H})}{m}} \right)$$

- Larger datasets  $m$  reduce error, but higher complexity increases the required data for a desired level of generalization.

## ► Practical Implications:

- These bounds provide actionable insights for balancing model complexity and dataset size.

# Vapnik-Chervonenkis (VC) Theory

## ► Growth Function

- The **Growth Function** is a measure of the expressiveness of a hypothesis space  $\mathcal{H}$ .
- **Definition:**
  - The growth function,  $\Pi_{\mathcal{H}}(m)$ , is the maximum number of distinct ways a hypothesis space can label  $m$  data points.
- **Key Idea:**
  - A more expressive hypothesis space can label datasets in a greater number of ways, indicating higher complexity.
- **Growth Behavior:**
  - For small  $m$ ,  $\Pi_{\mathcal{H}}(m) = 2^m$ .
  - For larger  $m$ , the growth may be limited by the structure of  $\mathcal{H}$ .

## ▶ VC Dimension

▶ The **VC Dimension** is a scalar value that quantifies the capacity of a hypothesis space  $\mathcal{H}$ .

### ▶ Definition:

▶ The VC dimension  $d_{VC}$  is the size of the largest dataset that can be **shattered** by  $\mathcal{H}$ .

### ▶ Shattering:

▶ A dataset is shattered if every possible labeling of the dataset can be perfectly captured by hypotheses in  $\mathcal{H}$ .

## ▶ Examples:

▶ A linear classifier in 2D space has a VC dimension of 3 (it can shatter any 3 points, but not all configurations of 4 points).

# VC Generalization Bound

## ► What is the VC Generalization Bound?

- A theoretical result that connects the **generalization error** with the **empirical risk**, the **VC dimension**, and the size of the dataset.
- **Mathematical Formulation:**

$$R(h) \leq R_{\text{emp}}(h) + \sqrt{\frac{8d_{\text{VC}} \left( \ln \left( \frac{2m}{d_{\text{VC}}} \right) + 1 \right) + 8 \ln \left( \frac{4}{\delta} \right)}{m}}$$

- $R(h)$ : Generalization error.
- $R_{\text{emp}}(h)$ : Empirical risk.
- $d_{\text{VC}}$ : VC dimension.
- $m$ : Dataset size.
- $\delta$ : Confidence level ( $1 - \delta$  is the probability that the bound holds).

## Key Insights

- ▶ As  $d_{VC}$  increases (more complex hypothesis space):
  - ▶ The bound becomes looser, reflecting a higher risk of overfitting.
- ▶ As  $m$  increases (larger dataset size):
  - ▶ The bound tightens, improving generalization guarantees.



# Summing Up VC Theory

- ▶ **Expressiveness vs. Generalization:**
  - ▶ The VC dimension captures the **expressiveness** of a hypothesis space:
    - ▶ Higher  $d_{VC}$ : More complex, more expressive.
  - ▶ A balance is required to avoid overfitting (high complexity) or underfitting (low complexity).
- ▶ **Implications for Learning:**
  - ▶ The VC dimension helps understand:
    - ▶ Why simpler models often generalize better.
    - ▶ Why increasing data size improves generalization, especially for complex models.
- ▶ **Foundation for Algorithm Design:**
  - ▶ VC theory guides the development of learning algorithms by quantifying the trade-offs between hypothesis complexity, data size, and generalization performance.

# Distribution-Based Bounds

- ▶ **From General Bounds to Data-Driven Insights:**
  - ▶ Generalization bounds like Hoeffding's inequality and VC bounds rely on worst-case scenarios.
- ▶ **Distribution-Based Bounds:**
  - ▶ Leverage specific properties of the data distribution to achieve **tighter bounds**.
  - ▶ Exploit **data structure** to understand how well a model generalizes in practice.

# Example: Support Vector Machines (SVMs)

- ▶ **SVMs and Margin-Based Bounds:**
  - ▶ Support Vector Machines (SVMs) introduce the concept of a **margin**, the distance between the decision boundary and the nearest data points.
  - ▶ **Intuition:**
    - ▶ A larger margin indicates better separation between classes, leading to better generalization.
  - ▶ **Margin-Based Generalization Bounds:**
    - ▶ Generalization error decreases as the margin increases, even for infinite hypothesis spaces.

# Alternative Capacity Measures

- ▶ **Why Explore Alternative Measures?**
  - ▶ VC dimension assumes worst-case datasets, often leading to overly conservative bounds.
  - ▶ Alternative measures provide more nuanced insights into hypothesis space complexity, especially for modern machine learning models like neural networks.

► **Examples of Alternative Measures**

1. **Covering Numbers:** The minimum number of small “balls” needed to cover the hypothesis space under a certain metric. Smaller covering numbers indicate a simpler hypothesis space, leading to tighter generalization bounds.
2. **Rademacher Complexity:** Measures the ability of a hypothesis class to fit random noise. A lower Rademacher complexity indicates that the hypothesis space is less prone to overfitting.

► **Next Steps:**

- We want to explore how these theoretical tools are applied to modern machine learning methods.

## Key Insights

- ▶ These measures allow for tighter, data-adaptive generalization bounds, particularly useful for complex or large-scale models.
- ▶ There's no one-size-fits-all measure. The choice of capacity measure depends on:
  - ▶ The hypothesis space.
  - ▶ The structure of the data.
  - ▶ The learning algorithm.

# Conclusion

## References

- Bousquet, Olivier, Stéphane Boucheron, and Gábor Lugosi. 2003. "Introduction to Statistical Learning Theory." *Advanced Lectures on Machine Learning*.
- Mohri, Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar. 2012. *Foundations of Machine Learning*. MIT Press.
- Mostafa-Samir. 2016. "A Gentle Introduction to Statistical Learning Theory." <https://mostafa-samir.github.io/ml-theory-pt2/>.
- Vapnik, Vladimir N. 1998. *Statistical Learning Theory*. Wiley-Interscience.