Generalization Bounds

Theoretical Foundations of Deep Learning

Matteo Mazzarelli

December 17, 2024

LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

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 $f : X \to Y$ 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

Demonstrating Overfitting

- **Objective**:
  - Show how increasing model complexity (polynomial degree) leads to overfitting.
- **Dataset**:
  - Using the scikit-learn **Diabetes** dataset with a single feature (BMI) and a quantitative response variable indicating disease progression (Target)[1].
- **Approach**:
  1. Fit polynomial regression models of varying degrees.
  2. Visualize polynomial fits on the training data.
  3. Examine the fits' residuals to see how errors behave.
  4. Plot training vs. test errors to highlight overfitting.

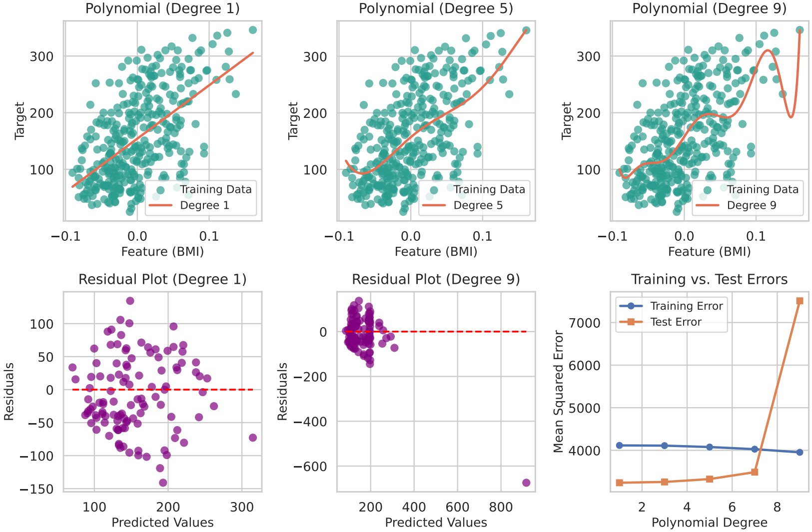

Figure 1: Overfitting Phenomenon in Polynomial Regression

Double Descent

- Modern machine learning introduces a fascinating twist: **Double Descent**, where increasing model complexity can lead to improved generalization after an initial overfitting phase.

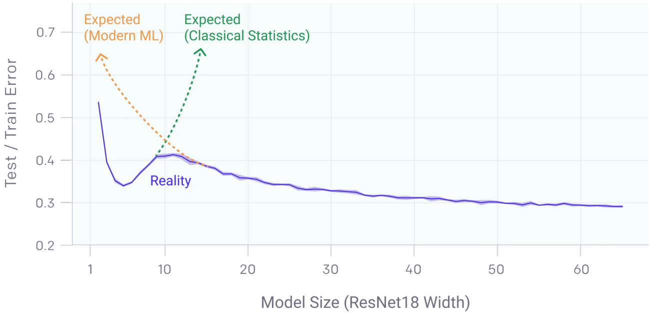

Figure 2: Double Descent phenomenon in a Residual Neural Network[2]

Classical Bounds

Generalization Bounds: Bridging the Gap

- **Goal**: Predict a model's performance on **unseen data**.
- **Generalization Bounds** provide theoretical guarantees, linking:
  - **Generalization Error**: Error on unseen data.
  - **Empirical Risk**: Error on training data.
  - **Model Complexity**: Model's flexibility.
- **Why They Matter**: They help understand the trade-offs between:
  - **Accuracy**: How well the model fits the data.
  - **Complexity**: Ability to model intricate patterns.
  - **Data Size**: Amount of data needed for reliable learning.

Hoeffding's Inequality: A Foundation

- **What it is**: A probabilistic tool that helps estimate how well a model will generalize.
- **Focus**: Quantifies the difference between **empirical risk** (training error) and **generalization error** (true error) for a *single, fixed model*.

Hoeffding's Inequality: The Math

- **Formula**[3,4]:

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

  - $R(h)$: True error on unseen data.
  - $R_{\text{emp}}(h)$: Error on training data.
  - ε: Error tolerance.
  - m: Dataset size.
- **Interpretation**: The probability of a large difference between true error and training error decreases **exponentially** with:
  - **Larger datasets** (m).
  - **Smaller error tolerance** (ε).

Convergence: How Fast Does It Happen?

- **Rate of Convergence**: How quickly the training error becomes a good estimate of the true error as we get more data.
- **Hoeffding's Formula** shows **faster convergence** with larger datasets due to the $\exp(-2m\varepsilon^2)$ term.

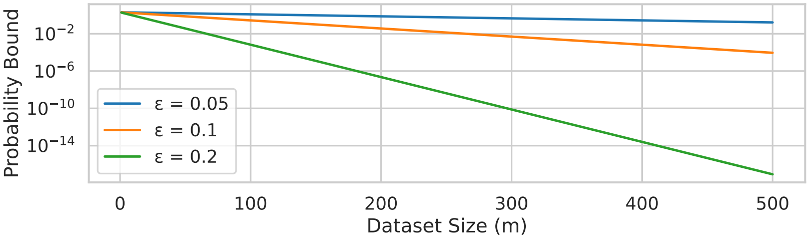

Figure 3: Hoeffding Bound Convergence Rate

Interpreting Hoeffding's Inequality

- **Meaning**: With more data, training error becomes a better predictor of true error.
- **Practical Implication**: For a fixed model, training performance is a good indicator of unseen data performance, and this improves with dataset size.
- **Limitations**: We usually pick the best model from many, not just one. Hoeffding doesn't account for how complex the model class is.

The Union Bound: Handling Multiple Models

- **What it does**: Extends bounds like Hoeffding's to work when choosing from **many models** (a hypothesis space $\mathcal{H}$).
- **Main Idea**: Considers the chance that *at least one* model in $\mathcal{H}$ has a large difference between training and true error.

Union Bound: The Formula

- **Expression**[3,4]:

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

- **Breakdown**:
  - $\sup_{h\in\mathcal{H}}$: Account for the worst-case scenario across all hypotheses.
  - $\sum_{h\in\mathcal{H}}$: Sums up probabilities of large error differences for each model.

Union Bound: Key Implications

- **Larger Model Space**: The more models we consider, the looser the bound becomes.

Table 1: Trade-off: Hypothesis Space vs. Bound & Capacity

Hypothesis Space Size	Bound	Model Capacity
Small	Tighter	Limited
Large	Looser	Higher

Moving Forward

- **Challenge**: Real-world model spaces are often infinite or too large.
- **Solution**: We need ways to measure model complexity that go beyond counting.
- **Next**: Exploring **complexity measures** for more practical generalization bounds.

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**:

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

    - $R(h)$: Generalization error.
    - $R_{\text{emp}}(h)$: Empirical risk.
    - d_{VC}: VC dimension.
    - m: Dataset size.
    - δ: 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.

Conclusions

References

1. Pedregosa F., Varoquaux G., & et al. (2011). *Scikit-learn: Machine learning in python, diabetes dataset*. https://scikit-learn.org/1.5/modules/generated/sklearn.datasets.load_diabetes.html
2. Nakkiran P., Kaplun G., & et al. (2019). *Deep double descent: Where bigger models and more data hurt*. https://arxiv.org/abs/1912.02292
3. Mohri M., Rostamizadeh A., & Talwalkar A. (2012). *Foundations of machine learning*. MIT Press.
4. Samir M. (2016). *A gentle introduction to statistical learning theory*. https://mostafa-samir.github.io/ml-theory-pt2/.
5. Bousquet O., Boucheron S., & Lugosi G. (2003). Introduction to statistical learning theory. *Advanced Lectures on Machine Learning*.
6. Vapnik V. N. (1995). *The nature of statistical learning theory*. Springer.