Generalization Bounds

Theoretical Foundations of Deep Learning

Matteo Mazzarelli

December 17, 2024

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Introduction

Why Study Generalization?

- **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.

Defining Generalization

- **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.

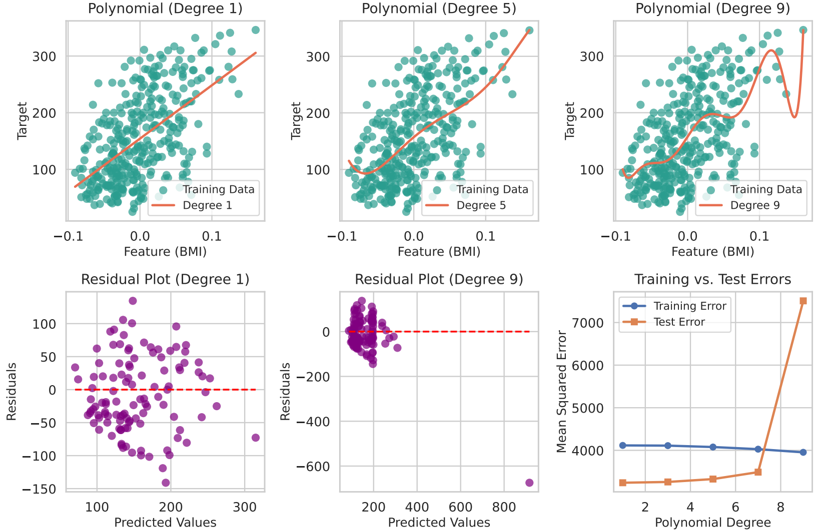

Figure 1: Overfitting Phenomenon in Polynomial Regression

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.

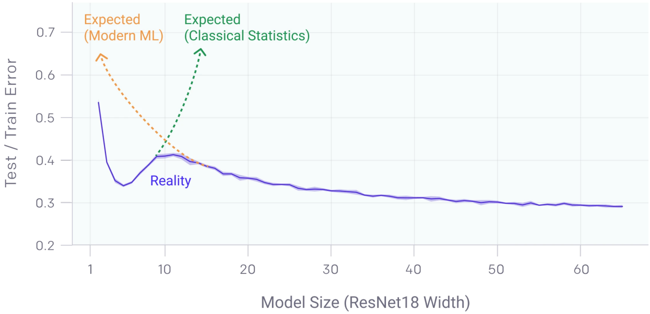

Figure 2: Double Descent phenomenon in a Residual Neural Network[2]

Classical Bounds

Generalization Bounds

- **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

- **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]:

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

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

Hoeffding's Inequality: Convergence

- **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 $\exp(-2n\varepsilon^2)$ term.

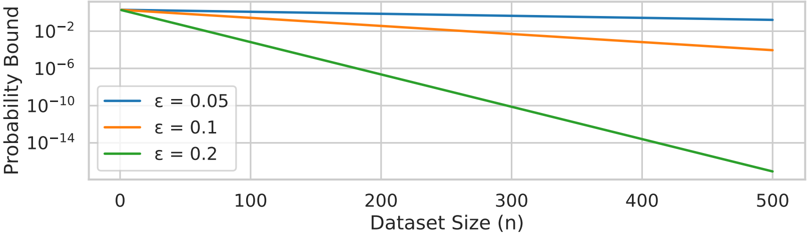

Figure 3: Hoeffding Bound Convergence Rate

Hoeffding's Inequality: Interpretation

- **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.

Union Bound

- **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 Maths

- **Expression**[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: Interpretation

- **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

Why Advanced Bounds?

- **Classical Bounds** give us a good starting point, but they can be loose.
- **Goal**: Tighter bounds that better reflect real-world performance.
- **How?**: By measuring model complexity in more sophisticated ways.

VC Dimension

- **Growth Function** ($\Pi_{\mathcal{H}}(n)$): How many ways can a model class ($\mathcal{H}$) label n data points?
  - More ways = more complex.
  - For small n, $\Pi_{\mathcal{H}}(n) = 2^n$.
- **Shattering**: A model class *shatters* a dataset if it can label it in *every possible way*.

VC Dimension: Definition

- **VC Dimension (d_{VC})**: The size of the *largest* dataset a model class can shatter.
- **Example**: Linear classifiers in 2D have $d_{\mathrm{VC}} = 3$. They can shatter 3 points but not 4 (in all configurations).

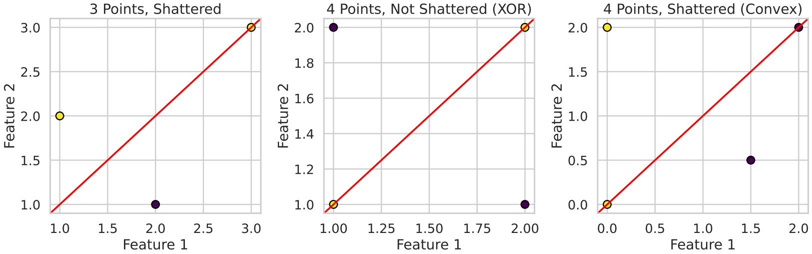

Figure 4: VC Dimension of Linear Classifiers in 2D

VC Generalization Bound

- **Formula**[5]:

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

  - $R(h)$: True error.
  - $R_{\text{emp}}(h)$: Training error.
  - d_{VC}: VC dimension.
  - n: Dataset size.
  - δ: Confidence parameter.

VC Bound: Interpretation

- **Higher VC Dimension**:
  - More complex model, looser bound, higher risk of overfitting.
- **Larger Dataset**:
  - Tighter bound, better generalization.

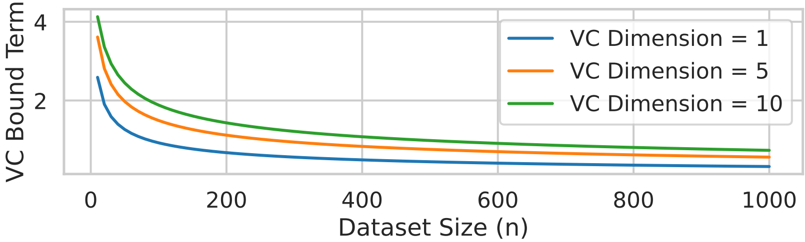

Figure 5: Approximation of the VC Generalization Bound

Distribution-Based Bounds

- **VC theory** often considers the *worst-case* scenario.
- **New Idea**: Use information about the **data distribution** for tighter bounds.
- **Example**: Support Vector Machines (SVMs).
  - **Margin**: Distance from the decision boundary to the nearest data points.
  - Larger margin = better generalization.
- **Benefit**: More realistic bounds reflecting real-world performance.

More Measures of Complexity

- **Why?**: VC dimension can be too pessimistic.
- **Goal**: More nuanced measures, especially for things like neural networks.

Table 2: Further ways to measure complexity[6]

Measure	Description	Key Idea
Covering Numbers	How many "balls" cover the hypothesis space?	Smaller = simpler = tighter bounds
Rademacher Complexity	How well can the model fit random noise?	Lower = less prone to overfitting

Conclusions

Key Takeaways I

- **Generalization** is crucial: We want models to work on **unseen data**, not just the training set.
- **Overfitting** is a risk: More complex models can memorize the training data but fail to generalize.
- **Classical Bounds** highlight the importance of:
  - **Dataset size**: More data leads to better generalization.
  - **Model complexity**: Simpler models (smaller hypothesis spaces) are safer.

Key Takeaways II

- **Advanced Bounds** offer a refined view:
  - **VC Dimension**: Measures a model's ability to shatter data. Higher VC dimension means more complexity.
  - **Distribution-Based**: Leverage data properties for tighter bounds.
- **The Goal**: Balance model expressiveness with the risk of overfitting by controlling complexity and leveraging insights from the data distribution.

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. Vapnik V. N. (1995). *The nature of statistical learning theory*. Springer.
6. Bousquet O., Boucheron S., & Lugosi G. (2003). Introduction to statistical learning theory. *Advanced Lectures on Machine Learning*.