Mathematics of Deep Learning, Summer Term 2020

Week 1

Deep Learning as Statistical Learning

Philipp Harms Lars Niemann

University of Freiburg

Overview of Week 1

1. Motivation for Deep Learning
2. Introduction to Statistical Learning
3. Empirical risk minimization and related algorithms
4. Error decompositions
5. Error trade-offs
6. Error bounds
7. Organizational Issues
8. Wrapup

Acknowledgement of Sources

Sources for this lecture:

- Frank Hutter and Joschka Boedecker (Department of Computer Science, Freiburg): Course on Deep Learning.
- Bousquet, Boucheron, and Lugosi (2003): Introduction to statistical learning theory.
- Vapnik (1999): An overview of statistical learning theory.

Mathematics of Deep Learning, Summer Term 2020

Week 1, Video 1

Motivation for Deep Learning

Philipp Harms Lars Niemann

University of Freiburg

Deep Learning in the News

The New York Times

Science

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

ENVIRONMENT | SPACE & COSMOS

Scientists See Promise in Deep-Learning

A voice recognition program translated a speech given by Richard F. Rashid Chinese.

By JOHN MARKOFF
Published: November 23, 2012

Using an artificial intelligence technique inspired by theories al the brain recognizes patterns, technology companies are reportin gains in fields as diverse as computer vision, speech recognitior

THE NEW YORKER

NEWS | CULTURE | BOOKS & FICTION | SCIENCE & TECH | BUSINESS | HUMOR | MAGAZINE | ARCHIVE | SUBSCRIBE

NOVEMBER 25, 2012

IS "DEEP LEARNING" A REVOLUTION IN ARTIFICIAL INTELLIGENCE?

BY GARY MARCUS

Can a new technique known as deep learning revolutionize artificial intelligence, as yesterday's front-page article at the New York *Times* suggests? There is good reason to be excited about deep learning, a sophisticated "machine learning" algorithm that far exceeds many of its predecessors in its abilities to recognize syllables and images. But there's also good reason to be skeptical. While the *Times* reports that "advances in an artificial intelligence technology that can recognize patterns offer

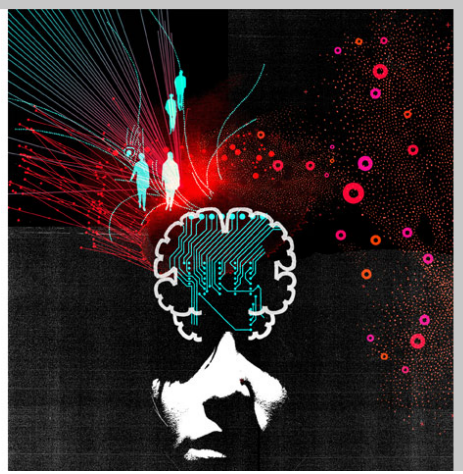
MIT Technology Review

10 BREAKTHROUGH TECHNOLOGIES 2013

Introduction | The 10 Technologies
Past Years

Deep Learning

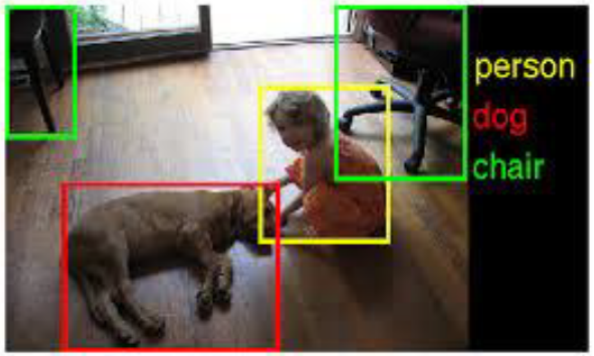
With massive amounts of computational power, machines can now recognize objects and translate speech in real time. Artificial intelligence is finally getting smart.



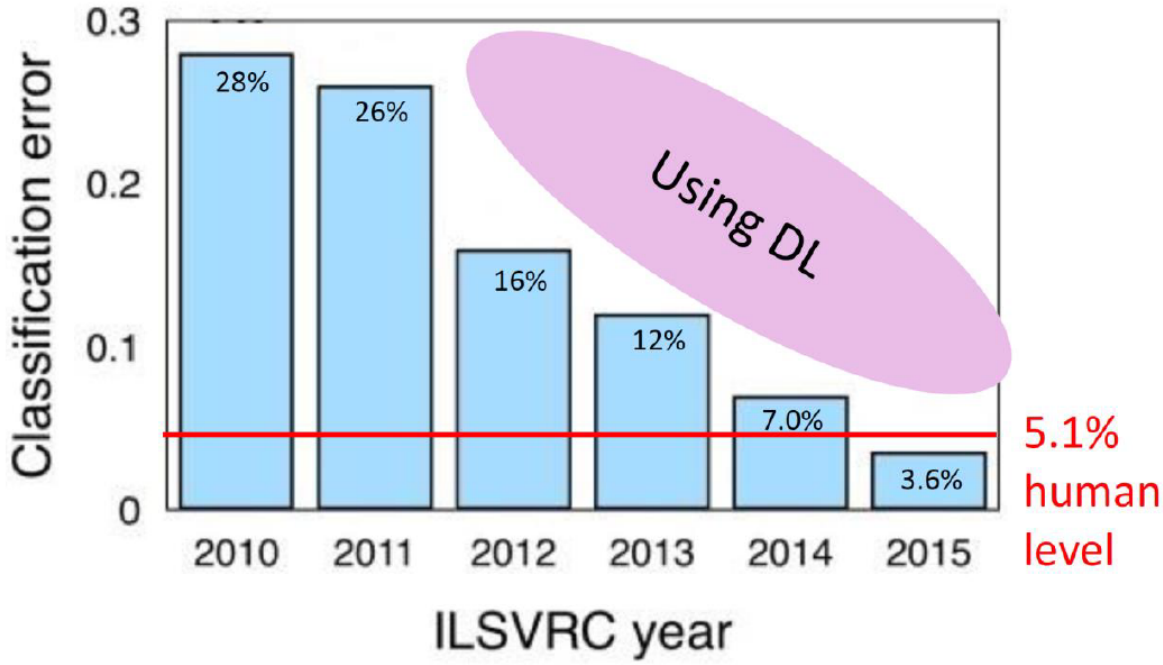
Deep Learning Revolutionized Computer Vision

- Excellent empirical results

Object recognition



Self-driving cars

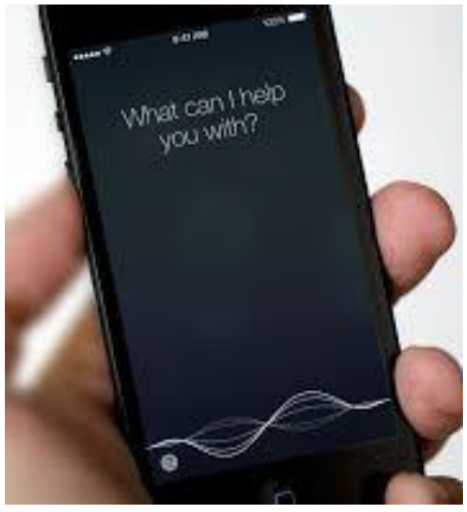


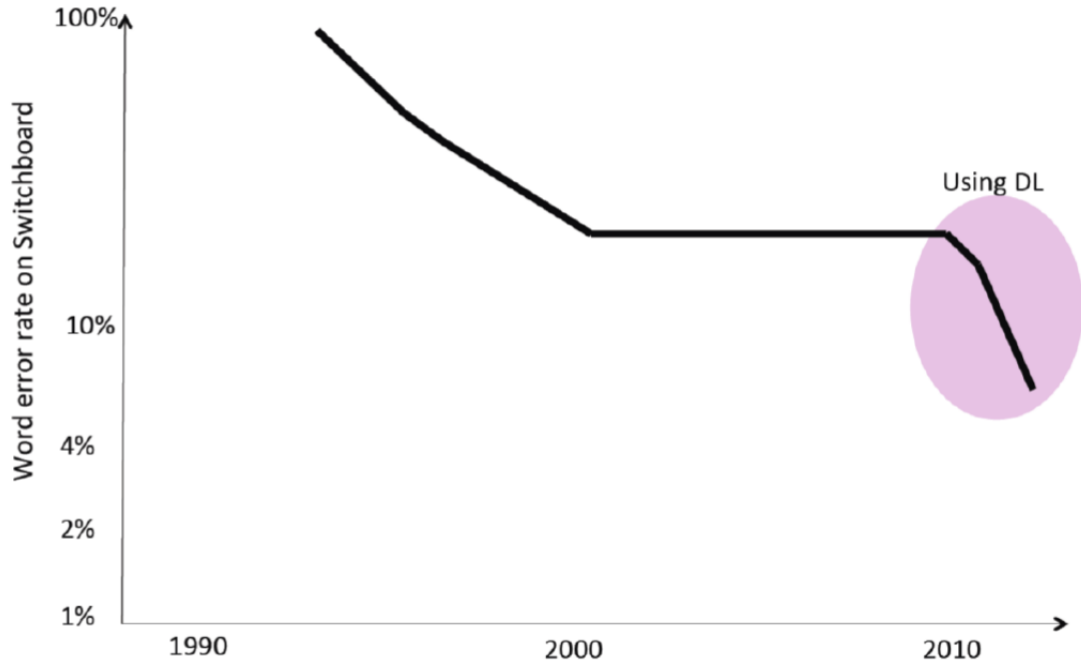
ILSVRC: ImageNet Large-Scale Visual Recognition Challenge

Deep Learning Revolutionized Speech Recognition

- Excellent empirical results

Speech recognition





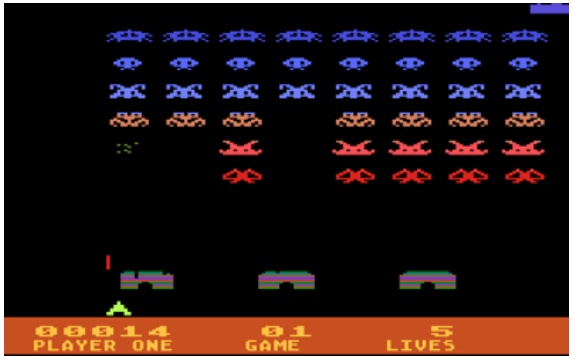
Auto-Translator

Image credit: Yoshua Bengio (data from Microsoft speech group)

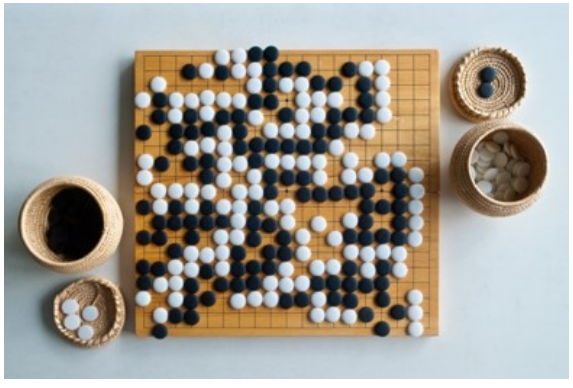
Deep Learning Goes Great with Reinforcement Learning

- Excellent empirical results obtained by deep reinforcement learning

  - Superhuman performance in playing Atari games
    [Mnih et al, Nature 2015]



  - Beating the world's best Go player
    [Silver et al, Nature 2016]



(Deep) Learning as A Different Way of Programming

- We don't understand how the human brain solves certain problems
  - Face recognition
  - Playing Atari games
  - Speech recognition
  - Picking the next move in the game of Go
- We can nevertheless learn these tasks from data/experience
- If the task changes, we simply re-train

Deep Learning Allows Many Branches of AI to Converge

- Deep learning is now the principle approach in many different branches of AI:
  - Computer vision
  - Speech recognition
  - Natural language processing
  - (Robotics)

- The same general techniques apply in all of these fields
  - Amazing potential for cross-fertilization
  - Fields that drifted apart for decades have largely converged again
  - E.g., in Freiburg:
    - close collaboration & joint reading group between machine learning, computer vision, robotics, neurorobotics, and robot learning

Further Reasons for the Popularity of Deep Learning

- Very quick to get good results for some problems
  - Deep learning can handle raw data (images, speech, text, etc)
  - Very well-engineered libraries handle the complex underpinnings (Tensorflow, Pytorch, ...)
  - Very little machine learning knowledge is required to get started
- Misconception: "it works like the brain"
- Neural networks are very flexible models – this is the main content of the lecture

Understanding deep learning

- Neural networks are excellent function approximators
  - They are dense in many function spaces; this is often called the universal approximation property [Cybenko, Hornik]
  - Approximation rates are known for many shallow and deep network architectures

- However, this only partially explains their success
  - Generalization capability is needed in addition to approximation capability
  - Deep learning performs better than the theory predicts; this is the oft-quoted unreasonable effectiveness of deep learning in artificial intelligence [Sejnowski]

- Many interesting mathematical questions remain
  - Mathematicians are ideally prepared for appreciating the abstract issues involved in high-dimensional data analysis [Donoho]

Questions to Answer for Yourself / Discuss with Friends

- Repetition: Why is deep learning so popular?
- Discussion:
  What might a mathematical theory of deep learning look like?
- Relation to your interests:
  What would you like to learn from this lecture?

Mathematics of Deep Learning, Summer Term 2020

Week 1, Video 2

Introduction to Statistical Learning

Philipp Harms Lars Niemann

University of Freiburg

Learning

Learning or, more precisely, inductive inference:

- Observe a phenomenon
- Construct a model of that phenomenon
- Make predictions using this model

Goals of learning theory and machine learning:

- Machine learning: automize inference
- Statistical learning theory: formalize inference

> *Nothing is more practical than a good theory.* *[Vapnik, Statistical Learning Theory 1998]*

Main assumption of statistical learning theory:

- Test and training data are iid.
- This distinguishes it from time series analysis (not independent) and transfer learning (not the same distribution).

Formalization

- Input and output spaces: measurable spaces $\mathcal{X}$ and $\mathcal{Y}$.
- Loss function: a measurable function $L\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$.
- Hypothesis class (aka. model class): a set H_0 of measurable functions $f\colon \mathcal{X} \to \mathcal{Y}$.
- Observations: independent random variables (X_1, Y_1), $\ldots$, (X_n, Y_n), defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, distributed according to a probability measure P on $\mathcal{X} \times \mathcal{Y}$.
- Objective: Find a function $f \in H_0$ with low or minimal risk (aka. test or generalization risk)

$$R(f) := \int L(f(x), y) P(dx, dy)$$

in the situation where P is unknown and the only information is contained in the observations.

Remarks

Applications:

- Regression: $\mathcal{Y} = \mathbb{R}$ and $L(y_1, y_2) = (y_1 - y_2)^2$.
- Classification: $\mathcal{Y} = \{0, 1\}$ and $L(y_1, y_2) = \mathbb{1}_{\{y_1 \neq y_2\}}$.

Useful hypothesis classes:

- Linear functions, polynomials, C^k functions, splines, or, as in deep learning, multilayer perceptrons.

Main challenge:

- The distribution P of the data and consequently also the risk functional R, which is to be minimized, are unknown.
- Otherwise this would be a standard optimization problem.

Questions to Answer for Yourself / Discuss with Friends

- Repetition: Describe the setup and goal of statistical learning theory.
- Discussion: Which aspects of machine learning are well-described by statistical learning theory? Which aren't?

Mathematics of Deep Learning, Summer Term 2020

Week 1, Video 3

Empirical risk minimization and related algorithms

Philipp Harms Lars Niemann

University of Freiburg

Risk versus empirical risk

Risk: Recall that...

- The objective in statistical learning theory is to minimize the risk

$$R(f) := \int L(f(x), y) P(dx, dy)$$

over all f in the hypothesis class H_0.

- The problem is that the distribution P of the data is unknown.

Empirical risk:

- As a substitute, define the empirical risk

$$R_n(f) := \frac{1}{n} \sum_{i=1}^{n} L(f(X_i), Y_i) = \int L(f(x), y) P_n(dx, dy),$$

where $P_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{(X_i, Y_i)}$ is the empirical measure.

Algorithms

Empirical risk minimization (aka. supervised learning):

$$f_n \in \operatorname*{arg\,min}_{f \in H_0} R_n(f).$$

Structural risk minimization:

$$f_n \in \operatorname*{arg\,min}_{\substack{k \in \mathbb{N} \\ f \in H_k}} R_n(f) + p(k, n),$$

for some increasing sequence $(H_k)_{k \in \mathbb{N}}$ of hypothesis classes and a penalty $p(k, n)$ for the size or capacity of the class.

Regularization:

$$f_n \in \operatorname*{arg\,min}_{f \in H_0} R_n(f) + \|f\|^2,$$

$$f_n \in \operatorname*{arg\,min}_{f \in H_0} R_n(f) + \|f\|^2 = \operatorname*{arg\,max}_{f \in H_0} e^{-R_n(f) - \|f\|^2},$$

for some suitable norm $\|\cdot\|$ (or some other form of penalty).

Algorithms (cont.)

Maximum likelihood:

$$f_n \in \arg\max_{f \in H_0} e^{-R_n(f)} p(f) = \arg\min_{f \in H_0} R_n(f) - \log p(f),$$

where $p\colon H_0 \to \mathbb{R}_+$ is a probability density with respect to some reference measure π on H_0.

Posterior mean:

$$f_n = \frac{1}{Z_n} \int_{H_0} f e^{-R_n(f)} p(f) \pi(df),$$

where $Z_n := \int_{H_0} e^{-R_n(f)} p(f) \pi(df)$ is a normalizing factor.

Gibbs sampling:

$$f_n \sim \frac{1}{Z_n} e^{-R_n} p\pi.$$

Questions to Answer for Yourself / Discuss with Friends

- Transfer (optimization): What algorithms could be used to solve the empirical risk minimization problem?
- Transfer (statistics): What do the law of large numbers and the central limit theorem say about the convergence of $R_n(f)$ to $R(f)$ for fixed $f \in H_0$?

Mathematics of Deep Learning, Summer Term 2020

Week 1, Video 4

Error decompositions

Philipp Harms Lars Niemann

University of Freiburg

Error decompositions

Notation: $\mathbb{E}$ and E denote expectations w/r to $\mathbb{P}$ and P, respectively, and:

- f^* solves $R(f^*) = \inf_{f\colon \mathcal{X}\to\mathcal{Y}} R(f)$,
- f_0 solves $R(f_0) = \inf_{f\in H_0} R(f)$, and
- f_n is an H_0-valued random variable.

Approximation and estimation error:

$$R(f_n) = \underbrace{R(f^*)}_{\text{statistical risk}} + \underbrace{\big(R(f_0) - R(f^*)\big)}_{\text{approximation error}} + \underbrace{\big(R(f_n) - R(f_0)\big)}_{\text{estimation error}}$$

Empirical risk and generalization error:

$$R(f_n) = \underbrace{R_n(f_n)}_{\text{empirical risk}} + \underbrace{\big(R(f_n) - R_n(f_n)\big)}_{\text{generalization error}}$$

Bias and variance: for $\mathcal{Y} = \mathbb{R}$ and $L(y_1, y_2) = (y_1 - y_2)^2$,

$$\mathbb{E}[R(f_n)] = \underbrace{R(f^*)}_{\text{statistical risk}} + E\Big[\underbrace{\mathbb{E}[f_n(x) - f^*(x)]^2}_{\text{bias}} + \underbrace{\mathbb{V}\mathrm{ar}[f_n(x)]}_{\text{variance}}\Big]$$

Proof of the bias-variance decomposition

Recall:

- $R(f^*) := \inf_{f\colon \mathcal{X}\to\mathcal{Y}} R(f)$.
- $\mathcal{Y} = \mathbb{R}$, $L(y_1, y_2) = (y_1 - y_2)^2$.

Mean-square optimality of the mean: $f^*(x) = E[y|x]$.

Conditional risk of f_n given (x, ω):

$$\begin{aligned} E[(f_n(x) - y)^2 \mid x] &= \mathrm{Var}[f_n(x) - y \mid x] + E[f_n(x) - y \mid x]^2 \\ &= E[(f^*(x) - y)^2 \mid x] + (f_n(x) - f^*(x))^2. \end{aligned}$$

Expected risk of f_n:

$$\begin{aligned} \mathbb{E}[R(f_n)] &= R(f^*) + E\big[\mathbb{E}[(f_n(x) - f^*(x))^2]\big] \\ &= R(f^*) + E\big[\mathbb{E}[f_n(x) - f^*(x)]^2 + \mathbb{V}\mathrm{ar}[f_n(x)]\big]. \end{aligned}$$

□

Questions to Answer for Yourself / Discuss with Friends

- Repetition: Visualize the approximation, estimation, and generalization error in a drawing.
- Discussion: Can you guess which error terms increase or decrease with respect to H_0 and n?

Mathematics of Deep Learning, Summer Term 2020

Week 1, Video 5

Error trade-offs

Philipp Harms Lars Niemann

University of Freiburg

Error trade-offs

Decompositions versus trade-offs

- A trade-off occurs when one term in an error decomposition increases while another term decreases with respect to a parameter.

Trade-offs in the choice of hypothesis class?

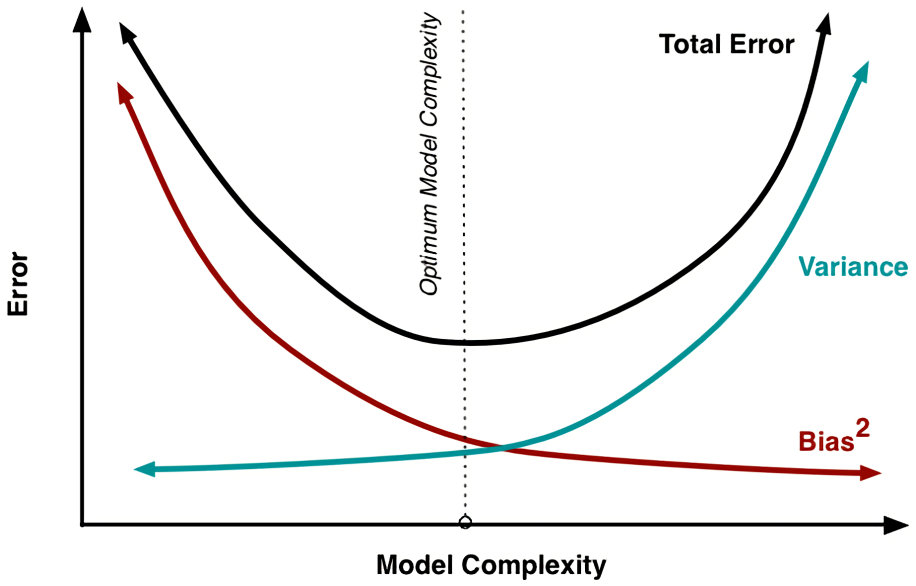
- In general, there is no trade-off in the above error decompositions with respect to H_0.
- However, there may be trade-offs with respect to H_0 in error bounds (as opposed to the error itself).

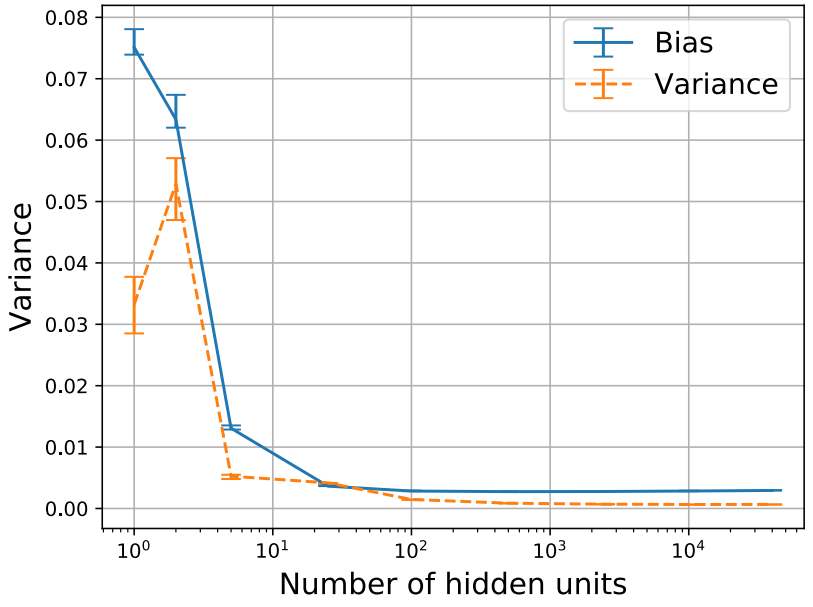
Example: bias-variance decomposition

- Conventional wisdom: The price to pay for achieving low bias is high variance—a trade-off in the choice of H_0. [Geman et al. 1992].
- However, this is false in over-parameterized regimes, which are common in modern machine learning applications (see next slide).

Example: bias-variance decomposition

Traditional view of the bias-variance trade-off (left) versus lack of any trade-off in MNIST character recognition using sufficiently wide ReLu networks (right).

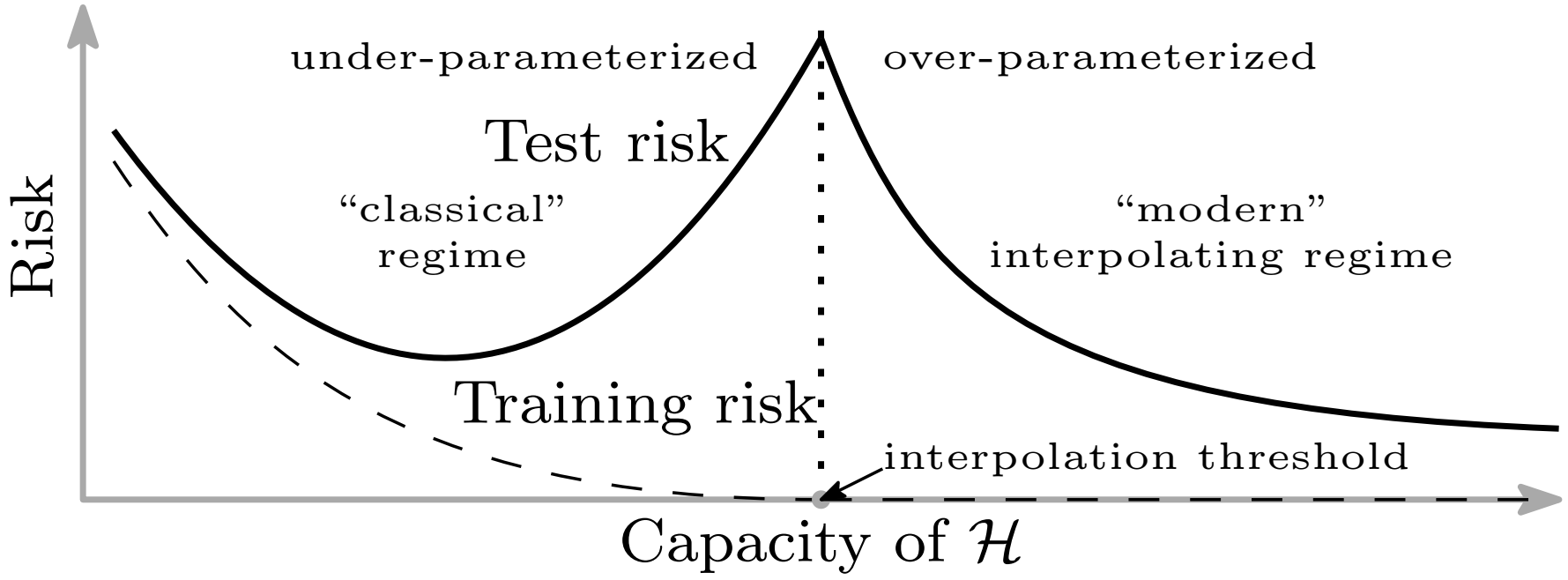




[Figures from Neal 2019]

Example: bias-variance decomposition (cont.)

Conjectured reconciliation: U-shaped risk curve in the underparameterized regime and decreasing risk in the overparameterized regime [Belkin e.a. 2019]



[Figure from Belkin e.a. 2019]

Questions to Answer for Yourself / Discuss with Friends

- Discussion: Can you think of a reason (or an example) why the variance might be decreasing in over-parameterized regimes?

Mathematics of Deep Learning, Summer Term 2020

Week 1, Video 6

Error bounds

Philipp Harms Lars Niemann

University of Freiburg

Bounding the approximation error

Notation:

- f^* solves $R(f^*) = \inf_{f\colon \mathcal{X}\to\mathcal{Y}} R(f)$, and
- f_0 solves $R(f_0) = \inf_{f\in H_0} R(f)$.

Approximation error: $R(f_0) - R(f^*)$

- Decreases when H_0 increases.
- Depends on how closely f^* can be approximated by functions in H_0.
- Is the main focus of this lecture.

Bound for quadratic loss functions:

$$\begin{aligned} 0 \le R(f_0) - R(f^*) &= E\big[(f_0(x) - y)^2 - (f^*(x) - y)^2\big] \\ &= E\big[(f_0(x) + f^*(x) - 2y)(f_0(x) - f^*(x))\big] \\ &\le E\big[|f_0(x) + f^*(x) - 2y|\big] \sup_{x\in\mathcal{X}} \big|f_0(x) - f^*(x)\big|. \end{aligned}$$

Bounding the generalization error

Notation:

- $R(f) = \int L(f(x), y) P(dx, dy)$,
- $R_n(f) = \int L(f(x), y) P_n(dx, dy)$, and
- f_n is a random element of H_0.

Generalization error: $R(f_n) - R_n(f_n)$

- Is the difference between a mean and an empirical mean:

$$R(f_n) - R_n(f_n) = \int L(f_n(x), y)(P - P_n)(dx, dy).$$

- Is of order $n^{-1/2}$ by the central limit theorem for fixed $f_n \equiv f$.

Uniform generalization error: $\sup_{f \in H_0} |R(f) - R_n(f)|$

- Increases when H_0 increases.
- Is the main focus of statistical learning theory.

Bounding the estimation error

Notation:

- $R(f) = \int L(f(x), y) P(dx, dy)$,
- $R_n(f) = \int L(f(x), y) P_n(dx, dy)$, and
- f_n is a random element of H_0.

Estimation error: $R(f_n) - R(f_0)$

- Is bounded by twice the uniform generalization error if f_n minimizes the empirical risk:

$$\cdots \leq \underbrace{R(f_n) - R_n(f_n)}_{\text{generalization error}} + \underbrace{R_n(f_n) - R_n(f_0)}_{\leq 0} + \underbrace{R_n(f_0) - R(f_0)}_{\text{generalization error}}.$$

A glimpse into statistical learning theory

Höffding's inequality: for any function $g\colon \mathcal{X} \times \mathcal{Y} \to [a,b]$, one has the Gaussian tail estimate

$$\mathbb{P}\big[|P_n g - Pg| > \epsilon\big] \leq 2\exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right), \qquad \epsilon > 0.$$

Uniform risk bound: given $H_0 = \{f_1, \ldots, f_N\}$, assume that the losses $g_i := L(f_i(\cdot), \cdot)$ take values in $[a,b]$ and estimate

$$\begin{aligned}\mathbb{P}\left[\max_{f\in H_0} |R_n f - Rf| > \epsilon\right] &= \mathbb{P}\left[\max_{i\in\{1,\ldots,N\}} |P_n g_i - Pg_i| > \epsilon\right] \\ &\leq \sum_{i=1}^{N} \mathbb{P}\big[|P_n g_i - Pg_i| > \epsilon\big] \leq 2N\exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right).\end{aligned}$$

A glimpse into statistical learning theory

Expected risk: deduce convergence of order $n^{-1/2}$ via

$$\mathbb{E}\left[\max_{f\in H_0}|R_n f - Rf|\right] = \int_0^\infty \mathbb{P}\left[\max_{f\in H_0}|R_n f - Rf| > \epsilon\right] d\epsilon$$
$$\leq N(b-a)\sqrt{\frac{\pi}{2n}}.$$

Note that the right-hand side depends on the size N of H_0.

Extension to infinite sets H_0: Approximate H_0 by finite sets of indicator functions; the error can be controlled by the Vapnik–Cervonenkis (VC) dimension of H_0 or other capacity measures.

Further topics: unbounded loss functions and capacity measures for specific hypothesis classes such as indicator functions or neural networks.

Caveat: deep learning performs better than predicted by this theory—once more, the unreasonable effectiveness of deep learning...

Questions to Answer for Yourself / Discuss with Friends

- Discussion: Can you spot any points where the error analysis of statistical learning theory might leave room for improvements?
- Suggestion: Read up on Höffding's inequality and related large deviations results or concentration inequalities.

Mathematics of Deep Learning, Summer Term 2020

Week 1, Video 7

Organizational Issues

Philipp Harms Lars Niemann

University of Freiburg

Team

- Philipp Harms: Lecturer, main contact for lectures
  `www.stochastik.uni-freiburg.de/professoren/harms/philipp-harms`
- Jakob Stiefel: Teaching Assistant, main contact for exercises
- Lars Niemann: Teaching Assistant
  `www.stochastik.uni-freiburg.de/mitarbeiter/niemann`

Web links

- Lecture homepage for general information: `www.stochastik.uni-freiburg.de/lehre/ss-2020/vorlesung-deep-learning-ss-2020`
- ILIAS for slides, videos, forum, and exercises: `ilias.uni-freiburg.de/goto.php?target=crs_1542865&client_id=unifreiburg`
- BigBlueButton: virtual meeting room `vHarms` with password `vHarms20206` at `www.math.uni-freiburg.de/lehre/virtuelle_veranstaltungen.html`. Supported Browsers include Chrome and Firefox on desktops and Chrome and Safari on mobiles.
- HisInOne for administrative issues

Outlook on the lecture

- Approximation theory for neural networks
  - shallow/deep
  - feed-forward/residual/recurrent
- Using methods from
  - functional analysis
  - harmonic analysis
  - differential geometry
  - probability theory
  - stochastic analysis
- Further topics
  - For example, generalization capability, auto-encoders, variational auto-encoders, adversarial networks, etc.
  - Depending on your interests and how we do time-wise

Relation to other deep learning courses in Freiburg

- This course: mathematical aspects of deep learning
- At the Mathematical Institute:
  - Angelika Rohde's seminar about the mathematical foundations of statistical learning: `www.stochastik.uni-freiburg.de/professoren/rohde/teaching`
  - Next term: Thorsten Schmidt's lecture on Machine Learning
- At the Department of Computer Science: in the groups on
  - Computer Vision
  - Machine Learning
  - Statistical Pattern Recognition
  - Artificial Intelligence

Parts of the course

- Short videos and slides:
  - Available on ILIAS every Tuesday night
- Live discussion and further reading:
  - Wednesdays 14:15-14:45 via BigBlueButton
- Forum:
  - Available on ILIAS for questions of all kinds
  - Please answer a question if you know the answer
- Graded exercises:
  - Mathematical and programming tasks
  - Solutions to be uploaded to ILIAS every two weeks
  - Collaboration in groups of two is allowed and encouraged.
  - Groups cannot be changed during the term.

Requirements and exam

- Requirements:
  - Solid background in probability theory and functional analysis
  - Basic knowledge in differential equations and stochastic analysis.
  - Basic programming skills
- Oral exam:
  - 50% of exercise points required for participation
  - Scope: content covered in the lecture, live discussions, and exercises
  - Focus on conceptual understanding rather than learning by heart

Resources for Python

- Python tutorials
  - Official tutorial:
    `https://docs.python.org/3/tutorial/index.html`
  - For beginners: `www.learnpython.org/`
  - For programmers:
    `http://stephensugden.com/crash_into_python/`
  - Many more: http://docs.python-guide.org/en/latest/intro/learning/
- Python libraries:
  - Numpy: `http://wiki.scipy.org/Tentative_NumPy_Tutorial`
  - SciPy: `http://docs.scipy.org/doc/scipy/reference/tutorial/`
  - Matplotlib: `http://matplotlib.org/users/beginner.html`

Mathematics of Deep Learning, Summer Term 2020

Week 1, Video 8

Wrapup

Philipp Harms Lars Niemann

University of Freiburg

Summary by learning goals

Having heard this lecture, you can now . . .

- Describe why deep learning is so popular
- Formulate the basic principles of statistical learning theory
- Understand deep learning in the context of statistical learning theory

Outlook on this week's discussion and reading session

- Discussion:
  - Questions and feedback, in both directions
  - Administrative and IT issues, if any
- Reading: related original literature
  - Sejnowski (2020): The unreasonable effectiveness of deep learning in artificial intelligence
  - Donoho (2000): High-Dimensional Data Analysis—the Curses and Blessings of Dimensionality
  - Vapnik (1999): An overview of statistical learning theory
- Preparation:
  - Watch the videos of the week