http://www.centreborelli.fr

Introduction to Statistical Learning

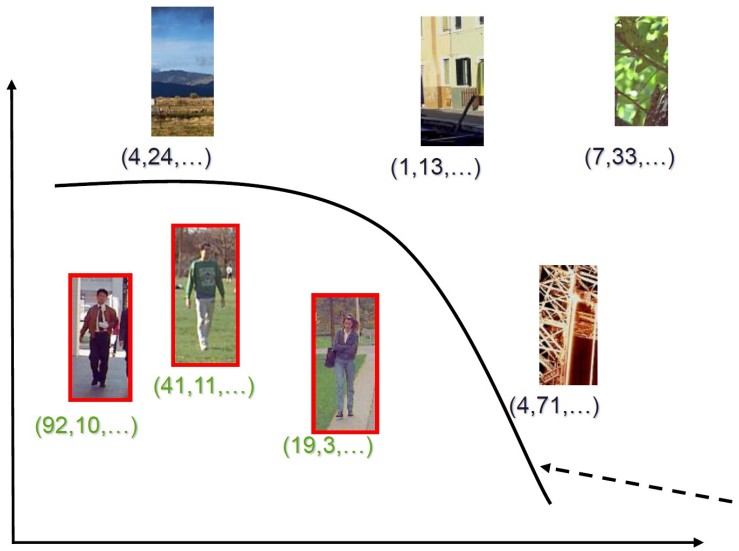
Nicolas Vayatis

nicolas.vayatis@ens-paris-saclay.fr

The goal of supervised machine learning

Finding a function

- Example : Pedestrian detection from video cameras

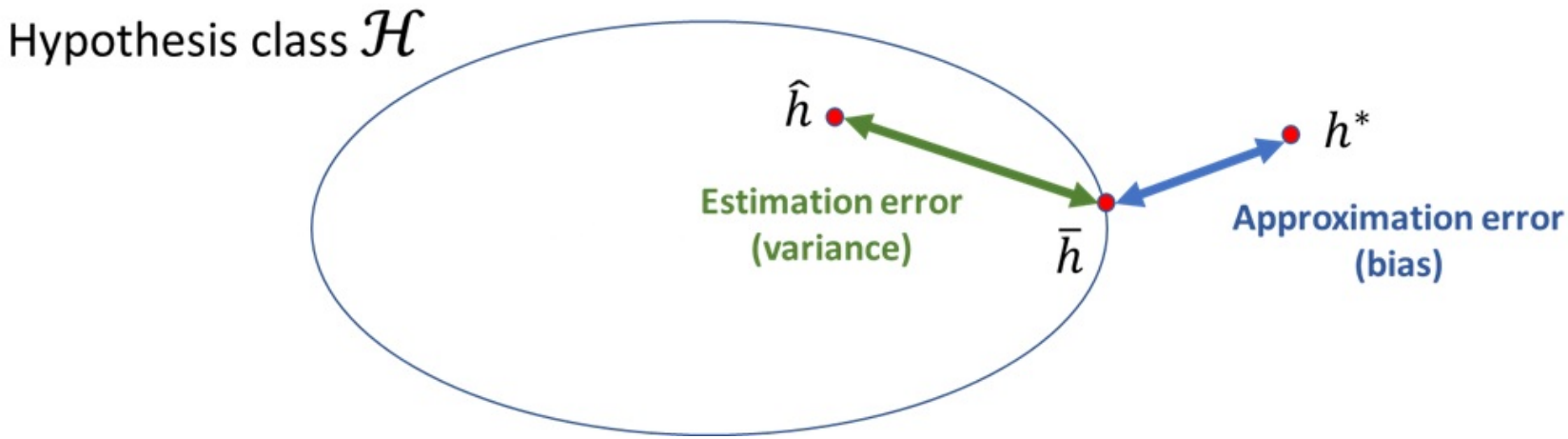


- What is the search space for such a function?

The art of machine learning

Solving the "bias-variance" trade-off

- Distance between solution provided by a learning method and the optimal solution (function): sum of *Approximation error* and *Estimation error*



- Learning a function amounts to:
  (a) chosing a search space (design process),
  (b) estimating the best function in this space (training process).

The three paradigms of ML

1. Local methods: based on grouping and local voting (or averaging)
   - k-Nearest-Neighbors
   - Kernel rules
   - Decision trees
2. Global methods: based on function optimization
   - Regularized regression (Ridge, LASSO...)
   - Support Vector Machines
   - Boosting
   - Feedforward neural networks
3. Ensemble methods: based on resampling and aggregation
   - Bagging
   - Boosting
   - Random forests

Shallow vs. Deep Learning

- Shallow learning: often relates to Tikohnov's regularization

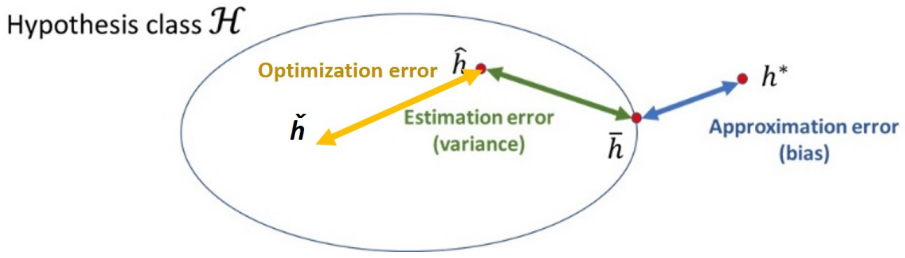
$$\min_{h \in \mathcal{H}} \left(\frac{1}{n} \sum_{i=1}^{n} \ell(h(X_i), Y_i) + \lambda_n \cdot \mathrm{pen}(h, n) \right)$$

  - The penalty controls the variance term (Occam's razzor)
  - It may also induce a desired structure of the function (e.g. sparsity).

- Deep Learning:

  - Universal approximators (zero bias)
  - No penalty term in the optimization but lots of tricks in the implementation which amount to *implicit regularization*

Bias-variance revisited



$$\begin{aligned}\mathcal{E} &= \mathbb{E}\left[E(f^*_{\mathcal{F}}) - E(f^*)\right] + \mathbb{E}\left[E(f_n) - E(f^*_{\mathcal{F}})\right] + \mathbb{E}\left[E(\tilde{f}_n) - E(f_n)\right] \\ &= \mathcal{E}_{\text{app}} + \mathcal{E}_{\text{est}} + \mathcal{E}_{\text{opt}}.\end{aligned}$$

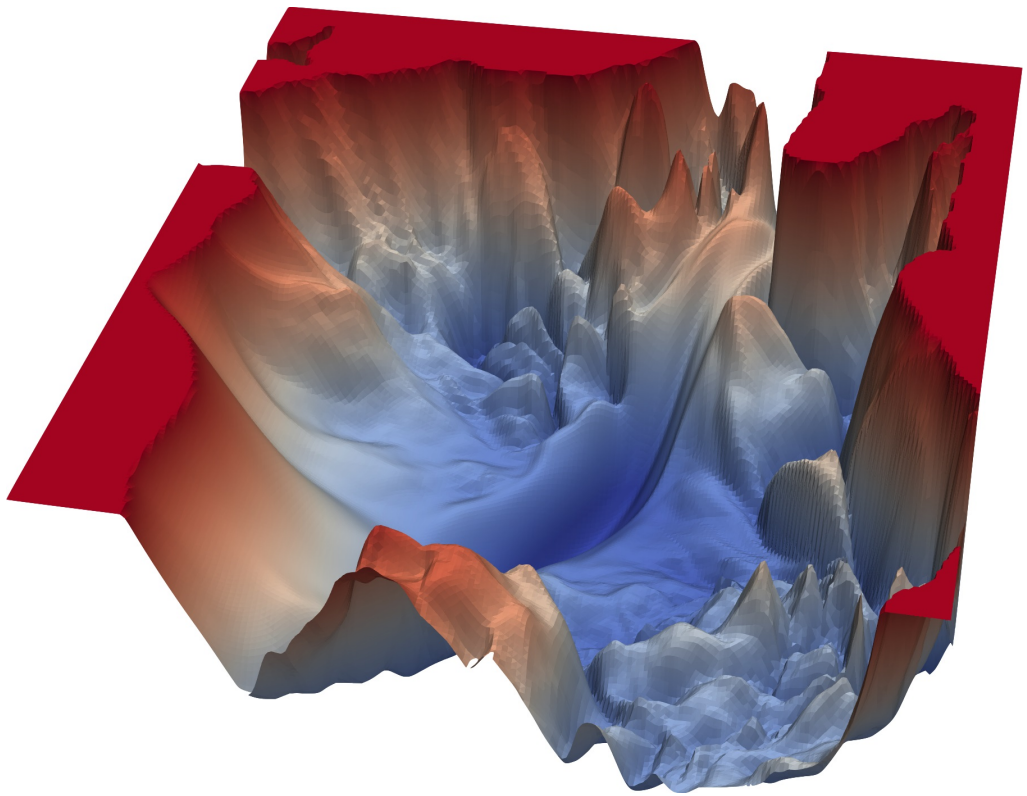
Trade-off wrt: Search space $\mathcal{F}$, sample size n, numerical tolerance ρ

		$\mathcal{F}$	n	ρ
$\mathcal{E}_{\text{app}}$	(approximation error)	$\searrow$		
$\mathcal{E}_{\text{est}}$	(estimation error)	$\nearrow$	$\searrow$	
$\mathcal{E}_{\text{opt}}$	(optimization error)	$\cdots$	$\cdots$	$\nearrow$
T	(computation time)	$\nearrow$	$\nearrow$	$\searrow$

[The trade-offs of Large Scale Learning, L. Bottou, O. Bousquet, 2011]

The loss landscape of Deep Learning

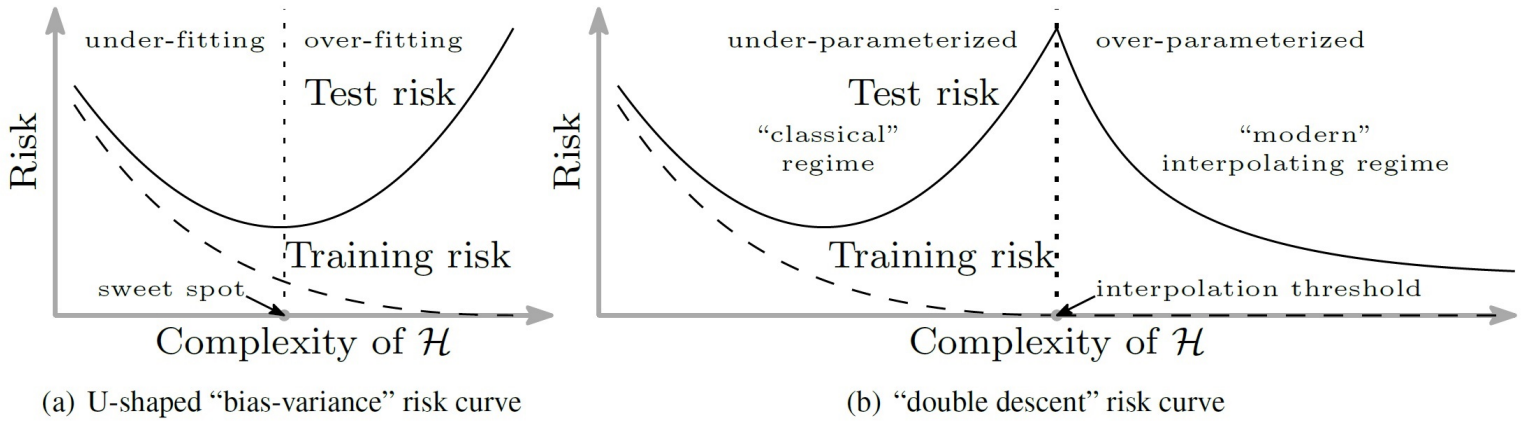
View on a 56-layer neural network without skip-connection



From [Visualizing the Loss Landscape of Neural Nets,
H. Li, Z. Xu1, G. Taylor, C. Studer, T. Goldstein, 2018]

The theory of a double descent risk curve

How Deep Learning (and random forests) avoid overfitting



(a) U-shaped "bias-variance" risk curve

(b) "double descent" risk curve

From [Reconciling modern machine learning and the bias-variance trade-off, M. Belkin, D. Hsu, S. Ma, S. Mandal, 2018]

Practical information

- **Course website:** `http://nvayatis.perso.math.cnrs.fr/ISLcourse-2020.html`
  - Tuesday morning 11am-1pm / ENS Paris-Saclay / Amphi Lagrange (1Z14)
  - 6 courses + 4 exercise sessions
  - Office hours: Tuesday 1pm-2pm
- **Evaluation:**
  - Two mandatory exams: Mid-term exam M + final exam F
  - Final grade G = max (F ; (F+M)/2)