Università di Genova | MaLGa

High-capacity hypothesis spaces in modern statistical learning

Master thesis by Luca Wellmeier

Supervisor: Prof. Marco Formentin

Co-supervisor: Prof. Ernesto De Vito (MaLGa | UniGe)

Co-supervisor: Prof. Lorenzo Rosasco (MaLGa | UniGe)

Statistical learning theory

The theoretical foundation of machine learning.

$$Y = f^*(X) + \varepsilon$$

- Random *input variable* X taking values in some measurable space $\mathcal{X}$.
- Random *output variable* Y taking values in $\mathbb{R}$.
- *Regression function* $f^* : \mathcal{X} \to \mathbb{R}$.
  - Usually very complicated and unknown!
- *Additive noise* term ε.
  - mean $\mathbb{E}[\varepsilon] = 0$, variance $\sigma^2 = \mathbb{E}[\varepsilon^2] < \infty$ and ε is independent of X.

Goal: Find *estimator* f so that $f \approx f^*$

$$Y = f^*(X) + \varepsilon$$

- Need performance measure: The *risk* (w.r.t. the squared loss) of a proposed estimator $f\colon \mathcal{X} \to \mathbb{R}$ is defined as

$$R(f) := \mathbb{E}\left[(Y - f(X))^2\right] = \mathbb{E}\left[(f^*(X) - f(X))^2\right] + \sigma^2.$$

- Choose a space $\mathcal{H}$ of functions $\mathcal{X} \to \mathbb{R}$; the *hypothesis space*.
- Our goal: solve

$$\min_{f\in\mathcal{H}} R(f)$$

- **Problem**: Hopeless... computing true risk requires knowing f^*.
- **Idea**: learn f^* from observations instead!

Università di Genova | MaLGa

- Let $Z_1 = (X_1, Y_1), Z_2 = (X_2, Y_2), \ldots$ be IID copies of $Z = (X, Y)$
- Consider instead the empirical risk

$$\hat{R}(f) = \hat{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2.$$

Principle (Empirical Risk Minimization (ERM))

Learning methods should be designed so to produce (approximate) solutions of the problem

$$\min_{f \in \mathcal{H}} \hat{R}(f)$$

Is ERM any good? A classical defect bound

- $\mathcal{X}$ a compact set and $\mathcal{H}$ compact subset of $C(X)$
- Assume there is $M > 0$ such that a.s. $|Y - f(X)| \leq M$ for all $f \in \mathcal{H}$
- For all $\varepsilon > 0$

$$\text{Prob}\left\{\sup_{f \in H} |\hat{R}(f) - R(f)| \leq \varepsilon\right\} \geq 1 - 2C_1 \exp\left(-\frac{n\varepsilon^2}{4(C_2 + M^2\varepsilon/3)}\right)$$

  - $C_1 = C_1(H, \varepsilon, M)$ is the minimum number of balls of radius $\varepsilon/(8M)$ needed to cover H
  - $C_2 = C_2(H) = \sup_{f \in H} \text{Var}[f(X) - Y]$

- larger hypothesis spaces $\implies$ smaller ε possible but confidence gets worse $\implies$ tradeoff!
- Let $\hat{f}$ be a learned estimator. Decompose risk according to tradeoff:

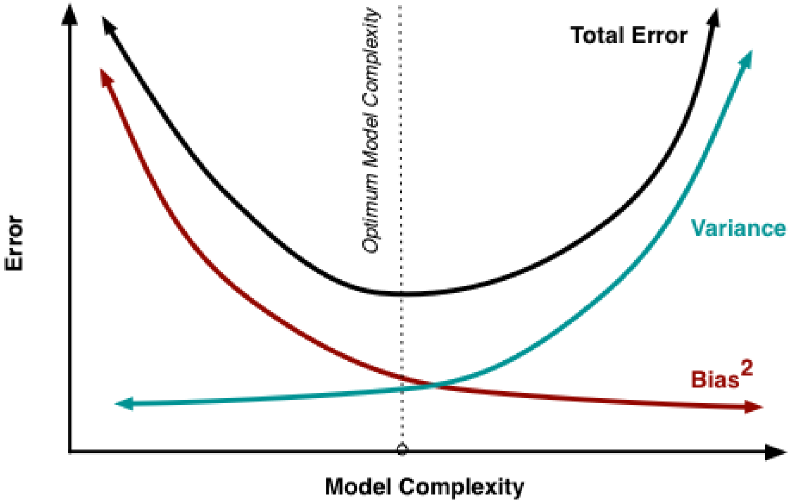
$$\mathbb{E}[R(\hat{f})] = \mathbf{B}^2 + \mathbf{V} + \sigma^2$$

with

$$\mathbf{B}^2 = \mathbb{E}_X\left[\left|\mathbb{E}_{Z_1,\ldots,Z_n}[\hat{f}(X)] - f^*(X)\right|^2\right],$$

$$\mathbf{V} = \mathbb{E}\left[\left|\hat{f}(X) - \mathbb{E}_{Z_1,\ldots,Z_n}[\hat{f}(X)]\right|^2\right].$$

- high bias corresponds to *underfitting*, high variance to *overfitting*



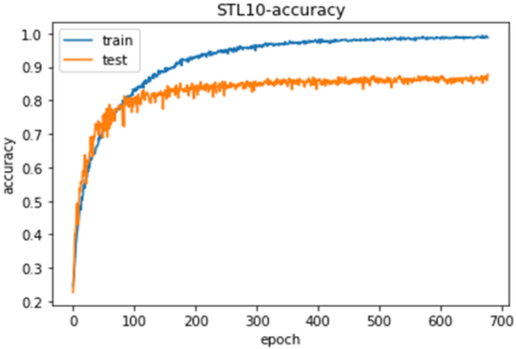
Università di Genova | MaLGa

Regularization

- Classical cure: regularization. Instead of minimizing the empirical risk, minimize
$$\hat{R}(f) + \lambda \|f\|_{\mathcal{H}}^2.$$
  - norm in $\mathcal{H}$ interpreted as complexity measure of an estimator
  - additional term acts as complexity penalty $\implies$ minimizing favors simplier solutions
  - Limits the ”reachable” size of the hypothesis space

Enter: deep learning

- highly over-paramterized architectures that perform best even when (almost) interpolating noisy data



- classical bounds become void $\implies$ new perspectives needed

Reproducing kernel Hilbert spaces

Definition

Let $\mathcal{H}$ be a Hilbert space of functions. If $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is such that

1. $k_x = k(\cdot, x) \in \mathcal{H}$ and
2. for all $x \in \mathcal{X}$ and $f \in \mathcal{H}$ we have the *reproducing property* $f(x) = \langle f, k_x \rangle$,

then k is a *reproducing kernel* of $\mathcal{H}$ and $\mathcal{H}$ is called reproducing kernel Hilbert space (RKHS).

Important properties:

- Equivalent definition: the evaluation functionals in $\mathcal{H}$ is continuous.
- Reproducing kernels are unique.
- All Gram matrices $[k(x_i, x_j)]_{ij}$ are symmetric and PSD.

Kernel ridge(less) regression

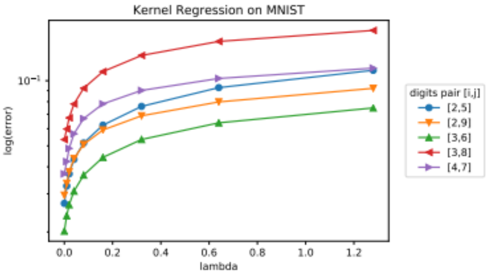
- *sampling operator* $\hat{S}\colon \mathcal{H} \to \mathbb{R}^n$ defined component-wise by $(\hat{S}f)_i := f(X_i)$
- adjoint is the *realization operator* $\hat{S}^* c = \sum_{i=1}^n c_i k_{X_i}$
- Representer theorem: any solution to $\min_{f\in H} \hat{R}(f) + \lambda \|f\|_{\mathcal{H}}^2$ admits the explicit form $f = \hat{S}^* c$ for some $c \in \mathbb{R}^n$
- Least squares theory allows us to give explicit solutions

$$\begin{aligned} \hat{f}_\lambda &= (\hat{\Sigma} + \lambda I)^{-1} \hat{S}^* \hat{Y} \\ &= \hat{S}^* (\hat{K} + \lambda I)^{-1} \hat{Y}, \end{aligned}$$

with covariance $\hat{\Sigma} = \hat{S}^* \hat{S}$ and kernel matrix $\hat{K} = \hat{S}\hat{S}^*$ (gram matrix).

Kernel methods as a study proxy

- Very complicated hypothesis spaces (e.g. certain Sobolev spaces).
- No iterative training needed: can apply linear least squares.
- Interpolation works well in many cases: e.g. Laplacian kernel $k(x, x') = \exp(-\|x - x'\|)$



- The neural tangent kernel encodes the learning behavior of gradient descent in infinite-width ReLU neural nets
  - same RKHS as the Laplacian kernel (a Sobolev space)!

Università di Genova | MaLGa

Mercer kernels on the torus

- the d-dimensional torus $\mathbb{T}^d := \mathbb{R}^d/\mathbb{Z}^d$ parameterized by $[0,1]^d$
- ONB of $L^2(\mathbb{T}^d)$:

$$e_{\mathbf{k}} : \mathbb{T}^d \to \mathbb{C}, x \mapsto \exp(2\pi i \mathbf{k} \cdot x) \quad \mathbf{k} \in \mathbb{Z}^d$$

- toral Mercer kernel are of the form

$$k(x, x') = \sum_{\mathbf{k}} \lambda_{\mathbf{k}} e_{\mathbf{k}}(x') \overline{e_{\mathbf{k}}(x)}.$$

  such that

$$\lambda_{\mathbf{k}} \geq 0, \quad \lambda_{\mathbf{k}} = \lambda_{-\mathbf{k}}, \quad \sum_{\mathbf{k}} \lambda_{\mathbf{k}} < \infty.$$

- Examples ($d = 1$):
  - Dirichlet kernel: $\lambda_k = 1$ for all $k \in \Lambda = [-R, R] \cap \mathbb{Z}$.
  - Sobolev kernel: $\lambda_k = |k|^{-2s}$ for real $s > \frac{1}{2}$

Explicit risk decomposition

If $f^* \in \mathcal{H}$, we have the following risk decomposition for toral Mercer kernels and kernel ridge(less) regression

$$\mathbb{E}[R(\hat{f}_\lambda) \mid Z_1, \ldots, Z_n] = \hat{\mathbf{B}}^2 + \hat{\mathbf{V}} + \sigma^2$$

with bias term

$$\hat{\mathbf{B}}^2 = \|\Sigma^{1/2}\hat{Q}(\lambda)f_H\|_{\mathcal{H}}^2, \qquad \hat{Q}(\lambda) = I - (\hat{\Sigma} + \lambda I)^{-1}\hat{\Sigma},$$

and variance term

$$\hat{\mathbf{V}} = \frac{\sigma^2}{n} \operatorname{trace}\left(\Sigma(\hat{\Sigma} + \lambda I)^{-2}\hat{\Sigma}\right).$$

Double descent