Overfitting, regularization, and complexity

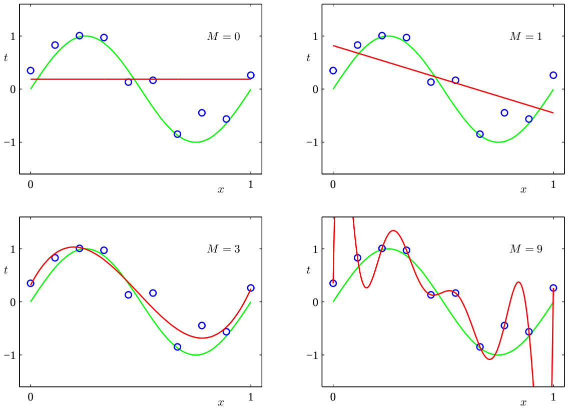
MATH-412 - Statistical Machine Learning

Polynomial regression and overfitting

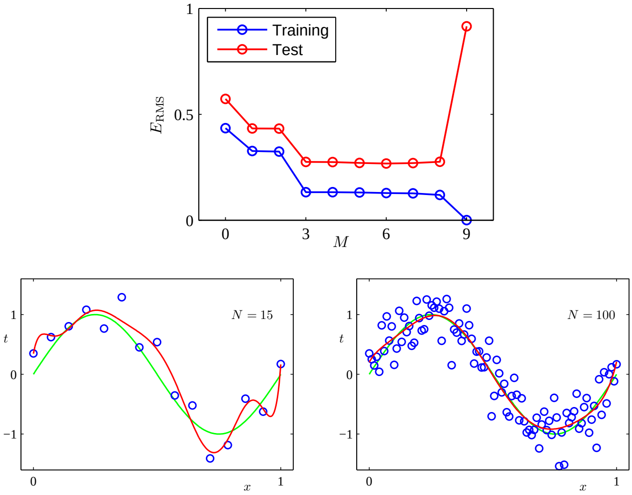
Polynomial regression : an instance of linear regression

Model of the form $Y = w_0 + w_1 X + w_2 X^2 + \ldots + w_p X^p + \varepsilon$

$$\min_{\boldsymbol{w}} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - (w_0 + w_1 x_i + w_2 x_i^2 + \ldots + w_p x_i^p)\right)^2 \quad \text{with} \quad p = M$$



Overfitting : symptoms and characteristics



Overfitting and generalization

In ML/stats we care about the generalization ability of the predictor

Fitting perfectly the data does not always entail *lack of generalization*...

- e.g., deep neural networks in computer vision.

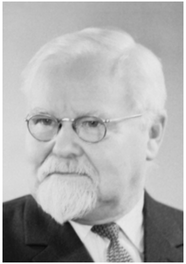
But fitting perfectly the data is a problem if

- the data is noisy and the model "fits the noise" or
- to be able to fit the data training the model learned is too "complex".

How do we measure **complexity**?...

Regularization

Tikhonov regularization



Andrey N. Tikhonov
(1906 - 1993)

$$\min_{f \in S} \widehat{\mathcal{R}}_n(f) + \lambda \|f\|^2$$

- λ is the regularization coefficient or hyperparameter

Is the problem now well-posed ?

If $\widehat{\mathcal{R}}_n$ is convex

$\Rightarrow$ The objective is strongly convex and coercive for any $\lambda > 0$

$\Rightarrow$ The solution exists and is unique.

$\Rightarrow$ $\lambda \mapsto \widehat{f}_\lambda$ is a continuous function

If $\widehat{\mathcal{R}}_n$ is bounded below

$\Rightarrow$ The objective is coercive for any $\lambda > 0$

$\Rightarrow$ At least a solution exists

If $\widehat{\mathcal{R}}_n$ is $\mathcal{C}^2$ with bounded curvature

$\Rightarrow$ Regularization eliminates small local minima.

Ridge regression

Is obtained by applying Tikhonov regularization to OLS regression.

$$\min_{\boldsymbol{w}\in\mathbb{R}^p} \frac{1}{2n}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}\|_2^2 + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2$$

- Normal equation

$$\boxed{\left(\tfrac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}+\lambda\boldsymbol{I}\right)\boldsymbol{w} = \tfrac{1}{n}\boldsymbol{X}^\top\boldsymbol{y}}$$

- Thus with unique solution :

$$\hat{\boldsymbol{w}}^{(\mathsf{ridge})} = \tfrac{1}{n}\left(\tfrac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}+\lambda\boldsymbol{I}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{y}$$

- Shrinkage effect

Linear vs affine regression and regularization

$$f_{\boldsymbol{w}}(\mathbf{x}) = \boldsymbol{w}^\top \mathbf{x} \qquad \text{vs} \qquad f_{\boldsymbol{w},b}(\mathbf{x}) = \boldsymbol{w}^\top \mathbf{x} + b = \widetilde{\boldsymbol{w}}^\top \widetilde{\mathbf{x}}$$

With

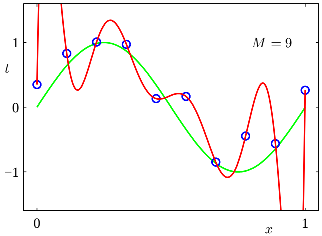
$$\widetilde{\boldsymbol{w}} = \begin{bmatrix} \boldsymbol{w} \\ b \end{bmatrix} \quad \text{and} \quad \widetilde{\mathbf{x}} = \begin{bmatrix} \mathbf{x} \\ 1 \end{bmatrix}$$

- ... an affine model in dimension p is a linear model in dimension $p+1$
- Working with $(\boldsymbol{w}, b)$ vs $\widetilde{\boldsymbol{w}}$ is equivalent when we don't regularize and otherwise not, because usually b is not regularized :

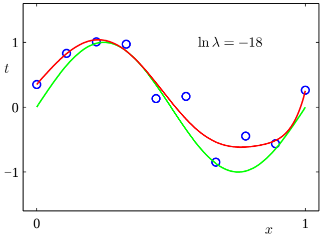
$$\min_{\boldsymbol{w} \in \mathbb{R}^p} \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w} + b\mathbf{1}\|_2^2 + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2$$

Polynomial regression with ridge

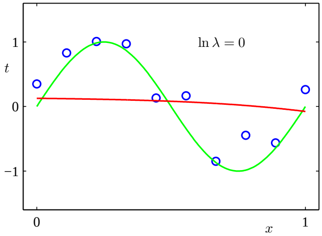
No regularization



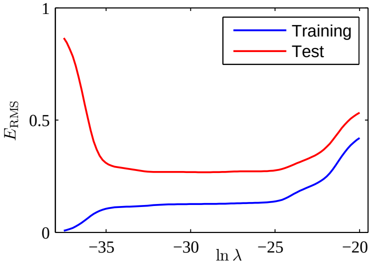
$10^{-8} \leq \lambda \leq 10^{-7}$



$\lambda = 1$



Regression with :
$n = 10$ examples and
a polynomial of degree $M = 9$,
but with ridge regularization.



Complexity

Controlling the complexity of the hypothesis space

Explicit control

- number of variables
- max degree for polynomial functions
- degree and # of knots for spline functions
- max resolution in wavelet approximations.
- bandwidth in RKHS

Implicit control

- with regularization,
- using Bayesian formulations
- via the learning/optimization algorithm
- randomization
- ...

The complexity of the predictor often results from a compromise between fitting and increasing complexity.

Problem of model selection : How to choose the right level of complexity ?

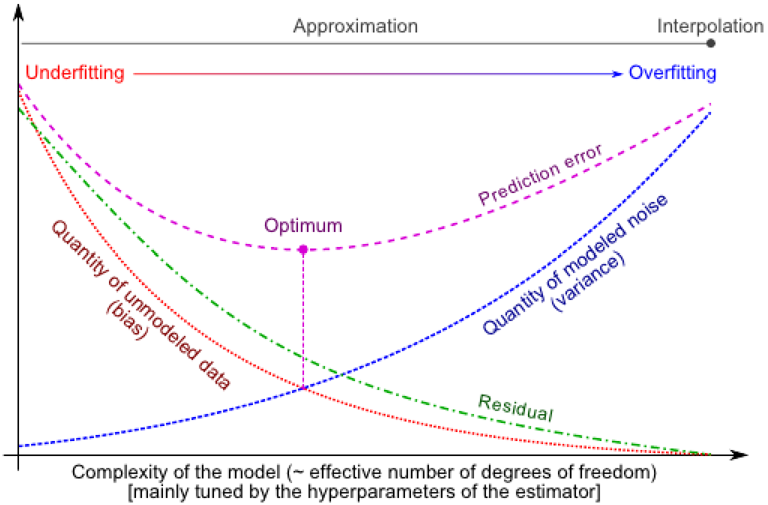
Risk decomposition : approximation-estimation trade-off

- $f^* =$ target function
- $f_S^* = \mathrm{argmin}_{f \in S} \mathcal{R}(f)$
- $\hat{f}_S =$ predictor/estimator in S

$$\underbrace{\mathcal{R}(\widehat{f}_S) - \mathcal{R}(f^*)}_{\text{excess risk}} = \underbrace{\mathcal{R}(\widehat{f}_S) - \mathcal{R}(f_S^*)}_{\text{estimation error}} + \underbrace{\mathcal{R}(f_S^*) - \mathcal{R}(f^*)}_{\text{approximation error}}$$

- Sometimes also called "bias-variance tradeoff"

Approximation-estimation tradeoff



The view that there is a necessarily a compromise between
fitting well the training data
and
learning a too complex model to generalize
has been challenged by neural networks...