

RegML2017@SIMULA Oslo
Class 4
Early Stopping and Spectral Regularization

Lorenzo Rosasco
UNIGE-MIT-IIT

June 28, 2016

Learning problem

Solve

$$\min_w \mathcal{E}(w), \quad \mathcal{E}(w) = \int d\rho(x, y) L(w^\top x, y)$$

given $(x_1, y_1), \dots, (x_n, y_n)$

Beyond linear models: non-linear features and kernels

Regularization by penalization

Replace

$$\min_w \mathcal{E}(w) \quad \text{by} \quad \min_w \underbrace{\widehat{\mathcal{E}}(w) + \lambda \|w\|^2}_{\widehat{\mathcal{E}}_\lambda(w)}$$

- ▶ $\widehat{\mathcal{E}}(w) = \frac{1}{n} \sum_{i=1}^n L(w^\top x_i, y_i)$
- ▶ $\lambda > 0$ regularization parameter

Loss functions and computational methods

- ▶ Logistic loss

$$\log(1 + e^{-yw^\top x})$$

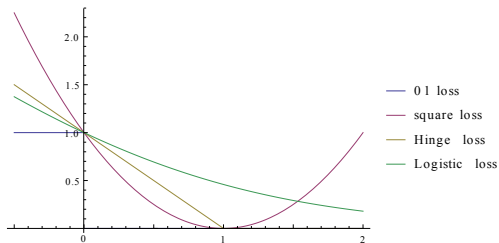
- ▶ Hinge loss

$$|1 - yw^\top x|_+$$

$$w_{t+1} = w_t - \gamma_t \nabla \hat{\mathcal{E}}_\lambda(w_t)$$

...

Square loss



$$(y - w^T x)^2 = (1 - yw^T x)^2$$

Ridge regression / Tikhonov regression

$$\hat{\mathcal{E}}_\lambda(w) = \frac{1}{n} \underbrace{\|\hat{X}w - \hat{y}\|^2}_{\text{Smooth and strongly convex}} + \lambda \|w\|^2$$

- ▶ \hat{X} $n \times d$ data matrix
- ▶ \hat{y} $n \times 1$ output vector.

Ridge regression / Tikhonov regression

$$\hat{\mathcal{E}}_\lambda(w) = \underbrace{\frac{1}{n} \|\hat{X}w - \hat{y}\|^2 + \lambda \|w\|^2}_{\text{Smooth and strongly convex}}$$

- ▶ \hat{X} $n \times d$ data matrix
- ▶ \hat{y} $n \times 1$ output vector.

$$\nabla \hat{\mathcal{E}}_\lambda(w) = \frac{2}{n} \hat{X}^\top (\hat{X}w - \hat{y}) + 2\lambda w = 0$$

$$\implies (\hat{X}^\top \hat{X} + \lambda n I)w = \hat{X}^\top \hat{y}$$

Linear systems

$$(\hat{X}^\top \hat{X} + \lambda n I)w = \hat{X}^\top \hat{y}$$

- ▶ nd^2 to form $\hat{X}^\top \hat{X}$
- ▶ roughly d^3 to solve the linear system

Representer theorem for square loss

$$f(x) = x^\top w \quad \implies \quad f(x) = \sum_{i=1}^n x^\top x_i c_i$$

Representer theorem for square loss

$$f(x) = x^\top w \quad \implies \quad f(x) = \sum_{i=1}^n x^\top x_i c_i$$

Proof:

Using SVD of \hat{X} ...

$$w = (\hat{X}^\top \hat{X} + \lambda n I)^{-1} \hat{X}^\top \hat{y} = \hat{X}^\top \underbrace{(\hat{X} \hat{X}^\top + \lambda n I)^{-1} \hat{y}}_c$$

$$\implies w = \hat{X}^\top c = \sum_{i=1}^n x_i c_i$$

Beyond linear models

$$f(x) = x^\top w = \sum_{i=1}^n x^\top x_i c_i,$$

$$w = (\hat{X}^\top \hat{X} + \lambda n I)^{-1} \hat{X}^\top \hat{y}, \quad c = (\hat{X} \hat{X}^\top + \lambda n I)^{-1} \hat{y}$$

- ▶ non-linear features

$$x \mapsto \phi(x) = (\phi_1(x), \dots, \phi_n(x)), \quad f(x) = \phi(x)^\top w$$

- ▶ non-linear kernels

$$\hat{X} \hat{X}^\top = \hat{K}, \quad f(x) = \sum_{i=1}^n K(x, x_i) c_i.$$

Interlude: linear systems and stability

$$\begin{aligned} Aw &= y, & A &= \text{diag}(a_1, \dots, a_d), & c &= \frac{a_1}{a_d} < \infty, \\ w &= A^{-1}y, & A^{-1} &= \text{diag}(a_1^{-1}, \dots, a_d^{-1}) \end{aligned}$$

More generally

$$\begin{aligned} A &= U\Sigma U^\top, & \Sigma &= \text{diag}(\sigma_1, \dots, \sigma_d) \\ A^{-1} &= U\Sigma^{-1}U^\top, & \Sigma^{-1} &= \text{diag}(\sigma_1^{-1}, \dots, \sigma_d^{-1}) \end{aligned}$$

Tikhonov Regularization

$$\frac{1}{n} \sum_{i=1}^n (y_i - w^\top x_i)^2 \quad \mapsto \quad \frac{1}{n} \sum_{i=1}^n (y_i - w^\top x_i)^2 + \lambda \|w\|^2$$

$$\hat{X}^\top \hat{X} w = \hat{X}^\top \hat{y} \quad \mapsto \quad (\hat{X}^\top \hat{X} + \lambda n I) w = \hat{X}^\top \hat{y}$$

Overfitting and numerical stability

Beyond Tikhonov: TSVD

$$\hat{X}^\top \hat{X} = V \Sigma V^\top, \quad w_M = (\hat{X}^\top \hat{X})_M^{-1} \hat{X}^\top \hat{y}$$

- ▶ $(\hat{X}^\top \hat{X})_M^{-1} = V \Sigma_M^{-1} V^\top$
- ▶ $\Sigma_M^{-1} = \text{diag}(\sigma_1^{-1}, \dots, \sigma_M^{-1}, 0, \dots, 0)$

Also known as principal component regression (PCR)

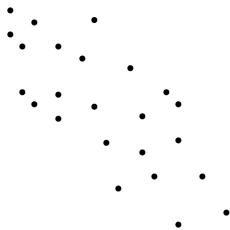
Principal component analysis (PCA)

Dimensionality reduction

$$\hat{X}^T \hat{X} = V \Sigma V^T$$

Eigenfunctions are
directions, of

- ▶ maximum variance
- ▶ best reconstruction



TSVD and PCA

$$TSVD \Leftrightarrow PCA + ERM$$

Regularization by projection

TSVD/PCR beyond linearity

Non-linear function

$$f(x) = \sum_{i=1}^p w_i \phi_i(x) = \Phi(x)^\top w$$

with

$$w = (\widehat{\Phi}^\top \widehat{\Phi})_M^{-1} \widehat{\Phi}^\top \hat{y}$$

Let $\widehat{\Phi} = (\Phi(x_1), \dots, \Phi(x_n))^\top \in \mathbb{R}^{n \times p}$.

$$\widehat{\Phi}^\top \widehat{\Phi} = V \Sigma V^\top, \quad (\widehat{\Phi}^\top \widehat{\Phi})_M^{-1} = V \Sigma_M^{-1} V^\top$$

$$\Sigma = \text{diag}(\sigma_1, \dots, \sigma_p), \quad \Sigma_M^{-1} = \text{diag}(\sigma_1^{-1}, \dots, \sigma_M^{-1}, 0, \dots)$$

TSVD/PCR with kernels

$$f(x) = \sum_{i=1}^n K(x, x_i) c_i, \quad c = (\hat{K})_M^{-1} \hat{y}$$

$$\hat{K}_{ij} = K(x_i, x_j), \quad \hat{K} = U \Sigma U^\top, \quad \Sigma = (\sigma_1, \dots, \sigma_n),$$

$$\hat{K}_M^{-1} = U \Sigma_M^{-1} U^\top, \quad \Sigma_M^{-1} = (\sigma_1^{-1}, \dots, \sigma_M^{-1}, 0, \dots),$$

Early stopping regularization

Other example of regularization: Early stopping of an iterative procedure applied to noisy data.

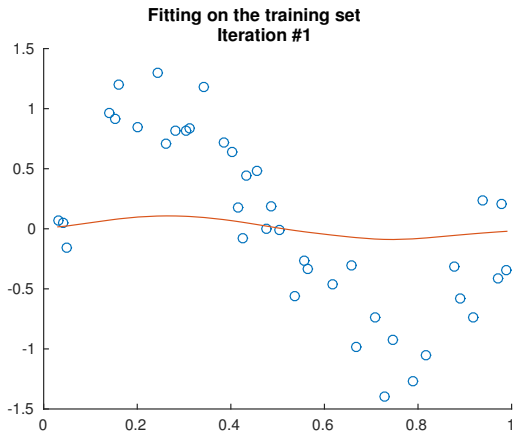
Gradient descent for square loss

$$w_{t+1} = w_t - \gamma \hat{X}^\top (\hat{X} w_t - \hat{y})$$

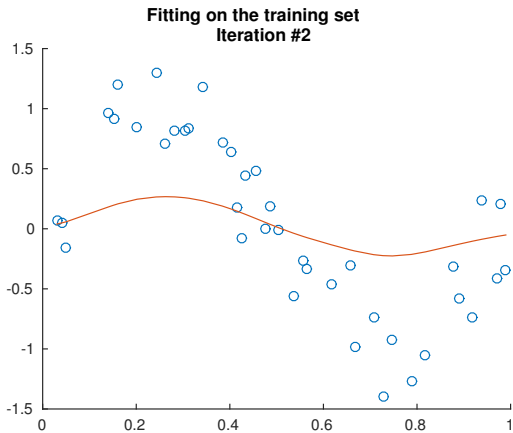
$$\sum_{i=1}^n (y_i - w^\top x_i)^2 = \|\hat{X} w - \hat{y}\|^2$$

- ▶ no penalty
- ▶ stepsize chosen a priori $\gamma = \frac{2}{\|\hat{X}^\top \hat{X}\|}$

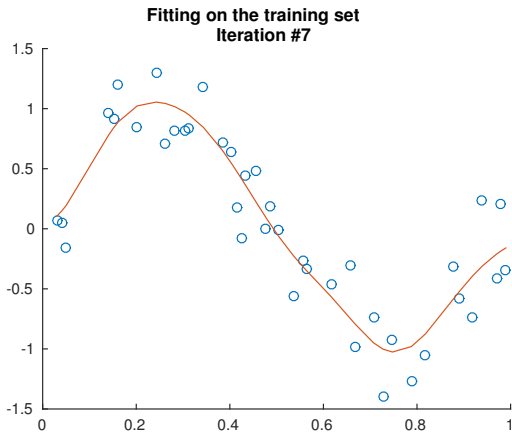
Early stopping at work



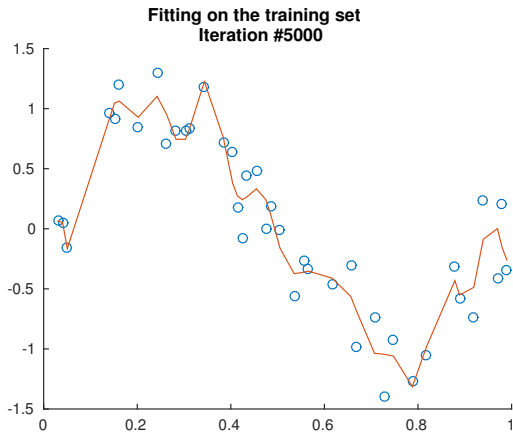
Early stopping at work



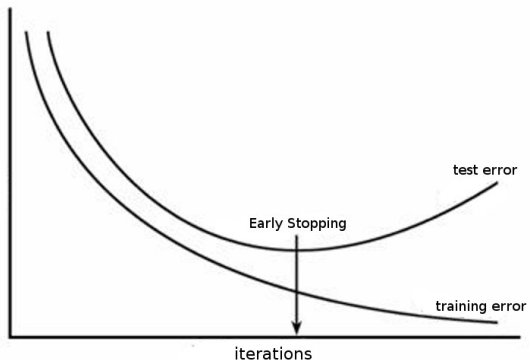
Early stopping at work



Early stopping at work



Semi-convergence



$$\min_w \mathcal{E}(w) \quad \text{vs} \quad \min_w \hat{\mathcal{E}}(w)$$

Connection to Tikhonov or TSVD

$$\begin{aligned}w_{t+1} &= w_t - \gamma \hat{X}^\top (\hat{X} w_t - \hat{y}) \\ &= (I - \gamma \hat{X}^\top \hat{X}) w_t + \gamma \hat{X}^\top \hat{y}\end{aligned}$$

by induction

$$w_t = \gamma \underbrace{\sum_{j=0}^{t-1} (I - \gamma \hat{X}^\top \hat{X})^j \hat{X}^\top \hat{y}}_{\text{Truncated power series}}$$

Neumann series

$$\gamma \sum_{j=0}^{t-1} (I - \gamma \hat{X}^\top \hat{X})^j$$

Neumann series

$$\gamma \sum_{j=0}^{t-1} (I - \gamma \hat{X}^\top \hat{X})^j$$

► $|a| < 1$

$$(1 - a)^{-1} = \sum_{j=0}^{\infty} a^j \quad \Longrightarrow \quad a^{-1} = \sum_{j=0}^{\infty} (1 - a)^j$$

Neumann series

$$\gamma \sum_{j=0}^{t-1} (I - \gamma \hat{X}^\top \hat{X})^j$$

- ▶ $|a| < 1$

$$(1 - a)^{-1} = \sum_{j=0}^{\infty} a^j \quad \Longrightarrow \quad a^{-1} = \sum_{j=0}^{\infty} (1 - a)^j$$

- ▶ $A \in \mathbb{R}^{d \times d}$, $\|A\| < 1$, invertible

$$A^{-1} = \sum_{j=0}^{\infty} (I - A)^j$$

Stable matrix inversion

Truncated Neumann Series

$$(\hat{X}^\top \hat{X})^{-1} = \gamma \sum_{j=0}^{\infty} (I - \gamma \hat{X}^\top \hat{X})^j \approx \gamma \sum_{j=0}^{t-1} (I - \gamma \hat{X}^\top \hat{X})^j$$

Note that

$$\gamma \sum_{j=0}^{t-1} (I - \gamma \hat{X}^\top \hat{X})^j = (I - (I - \gamma \hat{X}^\top \hat{X})^t) (\hat{X}^\top \hat{X})^{-1}$$

and compare to

$$(\hat{X}^\top \hat{X} + \lambda n I)^{-1}$$

Spectral filtering

Different instances of the same principle.

- ▶ Tikhonov

$$w_t = (\hat{X}^\top \hat{X} + \lambda n I)^{-1} \hat{X}^\top \hat{y}$$

- ▶ Early Stopping

$$w_t = \gamma \sum_{j=0}^{t-1} (I - \gamma \hat{X}^\top \hat{X})^j \hat{X}^\top \hat{y}$$

- ▶ TSVD

$$w_t = (\hat{X}^\top \hat{X})_M^{-1} \hat{X}^\top \hat{y}$$

Statistics and optimization

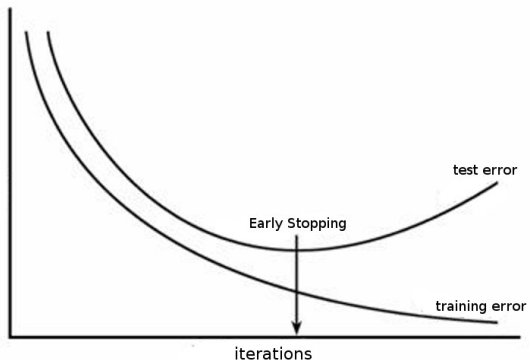
$$w_t = \gamma \sum_{j=0}^{t-1} (I - \gamma \hat{X}^\top \hat{X})^j \hat{X}^\top \hat{y}$$

The difference is in the computations

$$w_{t+1} = w_t - \gamma \hat{X}^\top (\hat{X} w_t - \hat{y})$$

- ▶ Tikhonov - $O(nd^2 + d^3)$
- ▶ TSVD - $O(nd^2 + d^2 M)$
- ▶ GD - $O(ndt)$

Regularization path and warm restart



$$\min_w \mathcal{E}(w) \quad \text{vs} \quad \min_w \hat{\mathcal{E}}(w)$$

Beyond linear models

Non-linear function

$$f(x) = \sum_{i=1}^p w_i \phi_i(x) = \Phi(x)^\top w$$

- ▶ Replace x by $\Phi(x) = (\phi_1(x), \dots, \phi_p(x))$
- ▶ Replace \hat{X} by

$$\hat{\Phi} = (\Phi(x_1), \dots, \Phi(x_n))^\top \in \mathbb{R}^{n \times p}.$$

$$w_{t+1} = w_t - \gamma \hat{\Phi}^\top (\hat{\Phi} w_t - \hat{y})$$

Computational cost $O(npt)$.

Early-stopping and kernels

$$f(x) = \sum_{i=1}^n K(x_i, x) c_i$$

By induction

$$c_{t+1} = c_t - \gamma(\widehat{K}c_t - \hat{y})$$

$$\widehat{K}_{ij} = K(x_i, x_j)$$

Computational Complexity $O(n^2t)$.

What about other loss functions?

- ▶ PCA + ERM

- ▶ Gradient / Subgradient Descent. Iterations for regularization, not only optimization!

Going big...

Bottleneck of Kernel methods

Memory

\hat{K} is $O(n^2)$

Approaches to large scale

- ▶ (Random) features - find $\tilde{\Phi} : X \rightarrow \mathbb{R}^M$, with $M \ll n$ s.t.

$$K(x, x') \approx \tilde{\Phi}(x)^\top \tilde{\Phi}(x')$$

Approaches to large scale

- ▶ (Random) features - find $\tilde{\Phi} : X \rightarrow \mathbb{R}^M$, with $M \ll n$ s.t.

$$K(x, x') \approx \tilde{\Phi}(x)^\top \tilde{\Phi}(x')$$

- ▶ Subsampling (Nyström) - replace

$$f(x) = \sum_{i=1}^n K(x, x_i) c_i \quad \text{by} \quad f(x) = \sum_{i=1}^M K(x, \tilde{x}_i) c_i$$

\tilde{x}_i subsampled from training set, M

Approaches to large scale

- ▶ (Random) features - find $\tilde{\Phi} : X \rightarrow \mathbb{R}^M$, with $M \ll n$ s.t.

$$K(x, x') \approx \tilde{\Phi}(x)^\top \tilde{\Phi}(x')$$

- ▶ Subsampling (Nyström) - replace

$$f(x) = \sum_{i=1}^n K(x, x_i) c_i \quad \text{by} \quad f(x) = \sum_{i=1}^M K(x, \tilde{x}_i) c_i$$

\tilde{x}_i subsampled from training set, M

- ▶ Greedy!
- ▶ Neural Nets

This class

Regularization beyond penalization

- ▶ Regularization by projection
- ▶ Regularization by early stopping

Next class

Multioutput learning

- ▶ Multitask learning
- ▶ Vector valued learning
- ▶ Multiclass learning