

# Lecture 2

## learning and regularization from interpolation to approximation

Stéphane Canu and Cheng Soon Ong

`stephane.canu@insa-rouen.fr`

`asi.insa-rouen.fr/~scanu`

RSISE ANU - NICTA, Canberra - INSA, Rouen

RSISE, SML group - Laboratoire PSI

# Appetizer...matrix factorization - $K = U^T D U$

$K$  is a regular matrix, eigenvectors  $\phi_i$  and eigenvalues  $\mu_i$   $K\phi_i = \mu_i\phi_i$   $i = 1, m$

$$K \underbrace{(\phi_1 \dots \phi_i \dots \phi_m)}_U = \underbrace{(\phi_1 \dots \phi_i \dots \phi_m)}_U \underbrace{\begin{pmatrix} \mu_1 & 0 & \dots & 0 \\ 0 & \dots & \mu_i & \dots & 0 \\ 0 & \dots & 0 & \dots & \mu_m \end{pmatrix}}_D$$

$U$  is an orthogonal matrix (eigen vectors orthonormal)

$D$  is a diagonal matrix

$$U^T U = I \Leftrightarrow U^T = U^{-1}$$

$$K U = U D \Leftrightarrow K U U^T = U D U^T$$

use it :

$$\begin{aligned} K\alpha = \mathbf{y} &\Leftrightarrow U D U^T \alpha = \mathbf{y} \\ &\Leftrightarrow U^T U D U^T \alpha = U^T \mathbf{y} \\ &\Leftrightarrow D U^T \alpha = U^T \mathbf{y} \\ &\Leftrightarrow U^T \alpha = D^{-1} U^T \mathbf{y} \quad \text{\color{red} } D \text{ is diagonal} \\ &\Leftrightarrow U U^T \alpha = U D^{-1} U^T \mathbf{y} \quad \Leftrightarrow \alpha = U D^{-1} U^T \mathbf{y} \end{aligned}$$

# Previously on functional learning and regularization

- interpolation : given a sample  $S_m = \{x_i \in \Omega, y_i \in \mathbb{R}, \quad i = 1, m\}$   
find  $f \in \mathcal{H}$  such that  $Tf = \mathbf{y}$  i.e.  $f(x_i) = y_i, \quad i = 1, m$
- define  $\mathcal{H}$  hypothesis space
  - $\mathcal{H} \subset \mathbb{R}^\Omega$  pointwise define functions
  - $T$  is continuous  $\Leftrightarrow$  there exists a kernel  $\kappa(x, x')$
  - $\forall f \in \mathcal{H}, \forall x_i \in \mathbb{R}, \quad f(x_i) = \langle \kappa(x_i, \cdot), f(\cdot) \rangle_{\mathcal{H}}$  reproducing property
- $\mathcal{H}$  is a r.h.s.  $\Leftrightarrow \kappa(x, x') \Leftrightarrow$  operator  $S : L^2 \longrightarrow \mathbb{R}^\Omega$
- $Tf = \mathbf{y} \Leftrightarrow K\alpha = \mathbf{y}$
- ready for more about regularization

# Road map

---

- ill posed problem, regularization and Learning
  1. splines
  2. the interpolation problem
  3. well posed and ill posed problem
  4. regularization framework -the regularization path
  5. Tikhonov regularization
  6. other regularization methods
  7. regularization analysis through filter functions
- Iterative algorithm and regularization
- Semi convergence - the proof

# ill posed problems

Let  $H_1$  and  $H_2$  be two normed sets. Let  $T$  be some linear operator from  $H_1$  to  $H_2$

**problem  $\mathcal{P}$**  given  $T$  and  $y \in H_2$ , find  $f \in H_1$  such that  $Tf = y$

## Definition (Well posed problem)

the problem  $\mathcal{P}$  is well posed if its solution

- exists
- is unique
- is stable ( $\|f - f_t\| \leq C\|y - y_t\|$ ) - the solution depends continuously on the data

## Definition (ill posed problem)

the problem  $\mathcal{P}$  is ill posed if its solution violated one of the above requirement

## Definition (Regularized solution)

Let  $\lambda \in I \subseteq \mathbb{R}^+$ . A regularized solution of the **problem  $\mathcal{P}$**  is a sequence  $f_\lambda$  of solutions of a sequence of **well posed** problems  $\mathcal{P}_\lambda$  (called the regularized problems) such that  $f_\lambda \xrightarrow{\lambda \rightarrow 0} f$ .

what are the regularization strategies ?

# Regularization strategies

- The initial problem : (almost) variational - minimal norm formulation

$$Tf = \mathbf{y} \quad \Leftrightarrow \quad \min_{f \in \mathcal{H}} \frac{1}{2} \|Tf - \mathbf{y}\|^2$$

- penalization : Tikhonov regularization

$$\mathcal{P}_\lambda : \min_{f \in \mathcal{H}} \|Tf - \mathbf{y}\|^2 + \lambda \|f\|_{\mathcal{H}}^2$$

- explicit subspace methods

- $\mathcal{H}_1 \subset \mathcal{H}_2 \subset \dots \subset \mathcal{H}_n = \mathcal{H}$ , e.g. Truncated spectral factorization

$$\mathcal{H}_k = \text{span}\{\phi_1, \dots, \phi_k\}$$

$$\mathcal{P}_k : \min_{f \in \mathcal{H}_k} \|Tf - \mathbf{y}\|^2$$

- iterative approaches

- gradient iterations (Landweber-Friedman) (fixed given stepsize  $\rho$ )

$$f_k = f_{k-1} - \rho \nabla_f (\|Tf_{k-1} - \mathbf{y}\|^2)$$

- Krylov subspace (conjugate gradient type)

find a sequence of iteration polynomial  $q_k$ ,  $f_k = q_{k-1}(T)\mathbf{y}$

# Regularization strategies

- The initial problem : (almost) variational - minimal norm formulation

$$Tf = \mathbf{y} \quad \Leftrightarrow \quad \min_{f \in \mathcal{H}} \frac{1}{2} \|Tf - \mathbf{y}\|^2$$

- penalization : Tikhonov regularization

$$\mathcal{P}_\lambda : \min_{f \in \mathcal{H}} \|Tf - \mathbf{y}\|^2 + \lambda \|f\|_{\mathcal{H}}^2$$

$$f_\lambda \xrightarrow{\lambda \rightarrow 0} f$$

- explicit subspace methods

- $\mathcal{H}_1 \subset \mathcal{H}_2 \subset \dots \subset \mathcal{H}_n = \mathcal{H}$ , e.g. Truncated spectral factorization

$$\mathcal{H}_k = \text{span}\{\phi_1, \dots, \phi_k\}$$

$$\mathcal{P}_k : \min_{f \in \mathcal{H}_k} \|Tf - \mathbf{y}\|^2$$

- iterative approaches

- gradient iterations (Landweber-Friedman) (fixed given stepsize  $\rho$ )

$$f_k = f_{k-1} - \rho \nabla_f (\|Tf_{k-1} - \mathbf{y}\|^2)$$

$$f_k \xrightarrow{k \rightarrow \infty} f$$

- Krylov subspace (conjugate gradient type)

find a sequence of iteration polynomial  $q_k$ ,

$$f_k = q_{k-1}(T)\mathbf{y}$$

# Regularization strategies for the interpolation problem

## ■ The initial problem : variational and minimal norm formulations

$$\min_{f \in \mathcal{H}} \|f\|_{\mathcal{H}}, \text{ with } Tf = \mathbf{y} \quad \Leftrightarrow \quad \text{solve } K\alpha = \mathbf{y} \text{ in } \mathbb{R}^m$$

$$K\alpha = \mathbf{y} \quad \Leftrightarrow \quad \min_{\alpha} \frac{1}{2} \alpha^{\top} K \alpha - K \alpha^{\top} \mathbf{y} \quad \Leftrightarrow \quad \min_{\alpha} \frac{1}{2} \|K\alpha - \mathbf{y}\|^2$$

## ■ penalization : Tikhonov regularization

$$\mathcal{P}_{\lambda} : \min_{f \in \mathcal{H}} \|Tf - \mathbf{y}\|^2 + \lambda \|f\|_{\mathcal{H}}^2 \quad \mathcal{P}_{\lambda} : \min_{\alpha} \|K\alpha - \mathbf{y}\|^2 + \lambda \|\alpha\|_{\mathbb{R}^m}^2$$

## ■ explicit subspace methods

- Truncated spectral factorization of  $K$ ,  $\mathcal{H}_k = \text{span}\{\phi_1, \dots, \phi_k\}$

$$\mathcal{P}_k : \min_{\alpha \in \mathcal{H}_k} \|K\alpha - \mathbf{y}\|^2$$

## ■ iterative approaches

- gradient iterations (Landweber-Friedman) (fixed given stepsize  $\rho$ )

$$\alpha_k = \alpha_{k-1} - \rho(K\alpha_{k-1} - \mathbf{y})$$

- Krylov subspace (conjugate gradient and minimal residual)

find a sequence of iteration polynomial  $q_k$ ,  $\alpha_k = q_{k-1}(K)\mathbf{y}$

# Tikhonov regularization of the interpolation problem

Given  $\mathcal{H}$  and  $S_m = \{\mathbf{x}_1, y_1, \dots, \mathbf{x}_m, y_m\}$ , find  $f \in \mathcal{H}$  such that  $Tf = \mathbf{y}$  ( $f(\mathbf{x}_i) = y_i$ )

## Definition (Tikhonov regularization - 3 equivalent ways)

given  $\lambda_1$   $\min_{f \in \mathcal{H}} \|f\|_{\mathcal{H}}^2$  such that  $\|Tf - \mathbf{y}\|_{\mathbf{R}^m}^2 < \lambda_1$

$\min_{f \in \mathcal{H}} \|Tf - \mathbf{y}\|_{\mathbf{R}^m}^2$  such that  $\|f\|_{\mathcal{H}}^2 < \lambda_2$

using Lagrange multiplier  $\min_{f \in \mathcal{H}} R_\lambda(f)$  with  $R_\lambda(f) = \sum_{i=1}^m (f(x_i) - y_i)^2 + \frac{1}{\lambda} \|f\|_{\mathcal{H}}^2$

**The solution**  $f(x) = \sum_{i=1}^m \alpha_i \kappa(x_i, x)$   
 $\alpha_i = \lambda(f(x_i) - y_i) \Leftrightarrow \frac{1}{\lambda} \alpha_i = \sum_{j=1}^m \alpha_j \kappa(x_j, x_i) - y_i$

using matrix notations

$$\frac{1}{\lambda} \alpha = K\alpha - \mathbf{y} \Leftrightarrow (\mathbf{K} + \mathbf{1}/\lambda \mathbf{I})\alpha = \mathbf{y}$$

**Tikhonov regularization  $\Leftrightarrow$  Preconditioning matrix  $\mathbf{K}$**

# filter functions to analyse the effect of regularization

- $K$  is a regular matrix, eigenvectors  $\phi_i$  and eigenvalues  $\mu_i$

$$K\phi_i = \mu_i\phi_i \quad i = 1, m$$

- eigen values are normalized  $0 \leq \phi_i \leq 1$

- $U$  is an orthogonal matrix (eigen vectors orthonormal)

- $D$  is a diagonal matrix

- factorization

$$K = UDU^T \quad K\alpha = \mathbf{y} \Leftrightarrow \alpha = \mathbf{U}\mathbf{D}^{-1}\mathbf{U}^T\mathbf{y}$$

- regularization through filtering the eigen values

$$\alpha = \mathbf{U}\varphi(\mathbf{D})\mathbf{D}^{-1}\mathbf{U}^T\mathbf{y}$$

- two extreme cases  $\varphi(\mu) = 0$  (cancelation)  $\varphi(\mu) = 1$  (no filter)

$\varphi(\mu)$  is called the filter function

# filter functions

## Tikhonov regularization

$$\begin{aligned}(K + \lambda I)\alpha = \mathbf{y} &\Leftrightarrow (UDU^\top + \lambda UU^\top)\alpha = \mathbf{y} \\ &\Leftrightarrow U(D + \lambda I)U^\top \alpha = \mathbf{y} \\ &\Leftrightarrow \alpha = U(D + \lambda I)^{-1}U^\top \mathbf{y} \quad \Leftrightarrow \quad \alpha = U \underbrace{(D + \lambda I)^{-1} D D^{-1}}_{\varphi_\lambda(\mu) = \frac{\mu}{\mu + \lambda}} U^\top \mathbf{y}\end{aligned}$$

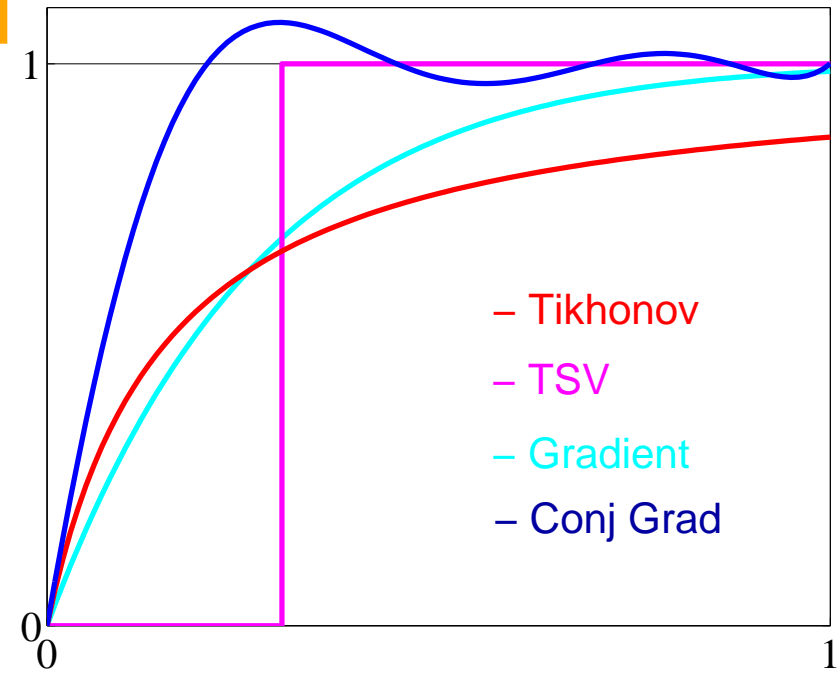
## Gradient iterations (Landweber-Fridman)

$$\begin{aligned}\alpha_{k+1} &= \alpha_k - \rho(K\alpha_k - \mathbf{y}) \\ &= (I - \rho K)\alpha_k - \rho \mathbf{y} \\ &= (I - \rho K)^2 \alpha_{k-1} - (I - \rho K)\rho \mathbf{y} - \rho \mathbf{y} && \text{some tedious algebra...} \\ &= (I - (I - \rho K)^k)K^{-1}\mathbf{y} \\ &= UI - (I - \rho D)^k D^{-1}U^\top \mathbf{y} && = U \underbrace{(I - (I - \rho D)^k)}_{\varphi_{k+1}(\mu) = 1 - (1 - \rho\mu)^k} D^{-1}U^\top \mathbf{y}\end{aligned}$$

## Krylov subspace (conjugate gradient)

$$\begin{aligned}\alpha_{k+1} &= q_k(K)\mathbf{y} \\ &= Uq_k(D)U^\top \mathbf{y} = U \underbrace{q_k(D)D}_{\varphi_{k+1}(\mu) = q_k(\mu)\mu} D^{-1}U^\top \mathbf{y}\end{aligned}$$

# Plotting filter functions (illustration)



example of behavior of the filter functions  $\phi_\lambda(\mu)$  for different regularization methods

$$\varphi_\lambda(\mu) = \frac{\mu}{\mu + \lambda}$$

$$\lambda = 0.25$$

$$\varphi_\lambda(\mu) = \begin{cases} 1 & \text{if } \mu > \lambda \\ 0 & \text{else} \end{cases}$$

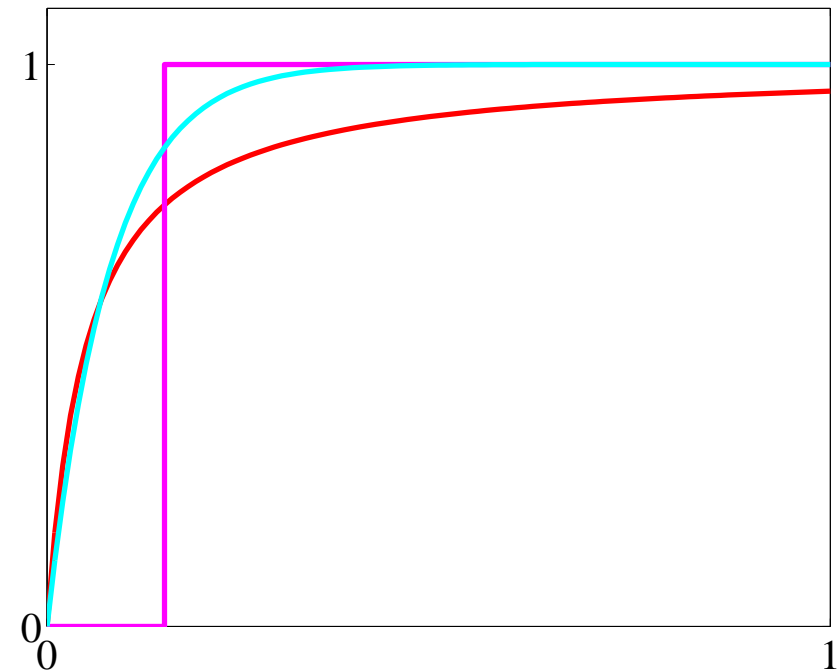
$$\lambda = 0.4$$

$$\varphi_{k+1}(\mu) = 1 - (1 - \rho\mu)^k$$

$$\rho = 0.25, k = 15$$

$$\varphi_{k+1}(\mu) = q_k(\mu)\mu$$

$$k = 5$$



same filter functions with less regularization

$$\lambda = 0.05, k = 50$$

# Summarize

---

- the regularization methods produce a **sequence** of well posed problems whose solutions  $f_\lambda$  are of **increasing complexity**
- this sequence of solutions **converges** toward the interpolating function
- these methods differ through the **regularization path** they follow

What about approximation ?

# Learning as a minimum prediction error

## ■ Definitions

- inputs :  $X \in \mathbb{R}^d$
- output :  $Y \in \mathbb{R}$
- cost :  $C(f, x, y) = (f(x) - y)^2$
- unknown :  $\mathbb{P}(x, y)$
- risk :  $R(f) = \mathbb{E}(C(f, X, Y)) = \mathbb{E}((f(X) - Y)^2)$

## ■ Theoretical problem

$$\text{Problem : } \min_f R(f) \quad \text{Solution : } t(x) = \mathbb{E}(Y | X = x)$$

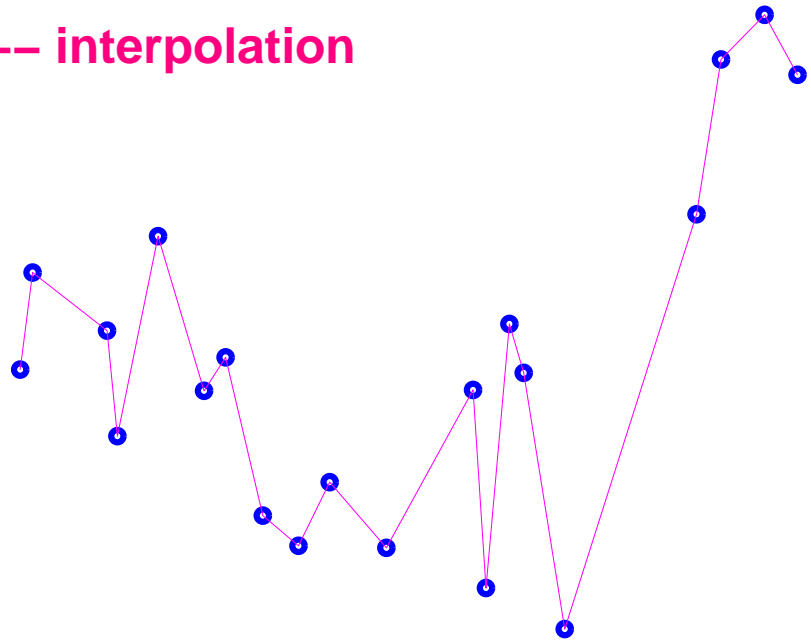
## ■ Learning algorithm :

given a sample  $S_m = \{(x_i, y_i), i = 1, m\}$  i.i.d. from  $\mathbb{P}(x, y)$ ,  
estimate function  $t$

learning with noise = approximate  $(x_i, y_i), i = 1, m$

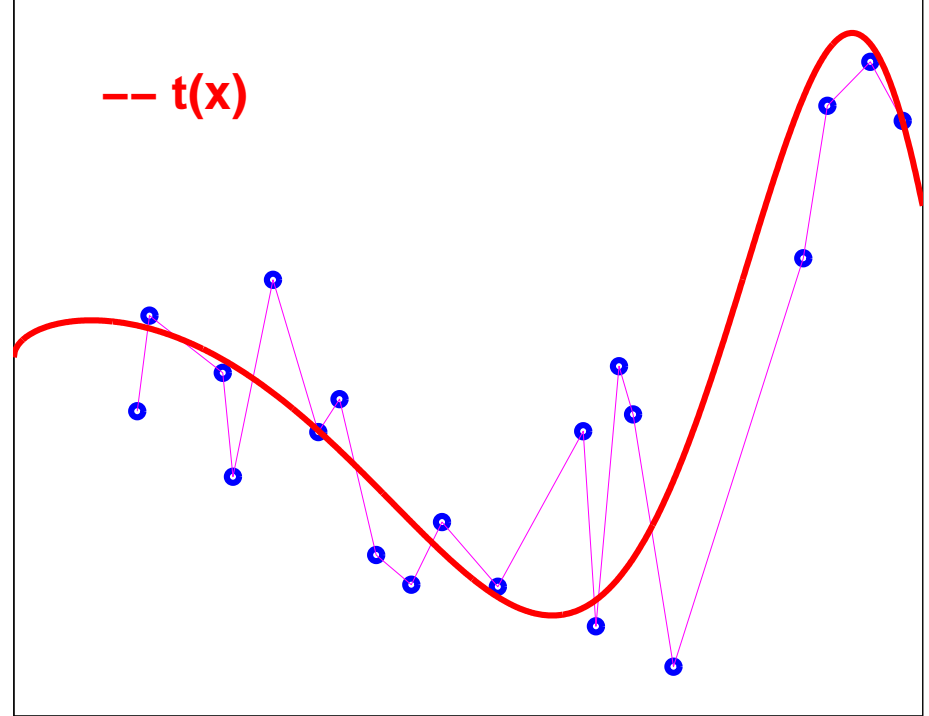
# Approximation and regularization

-- interpolation



data, interpolation

--  $t(x)$



data, interpolation and the target function

$$y_i = t(x) + \varepsilon_i$$

the **approximation** is in the **regularization path** toward interpolation

Approximation = early stopping in the regularization path

# Conclusion

- Regularization and interpolation
  - penalisation - Tikhonov
  - subset - TSD
  - iterative methods
- approximation via **early stopping** in the regularization path
  - penalisation Tikhonov : find the "good"  $\lambda$ , complexity  $\mathcal{O}((\#\lambda)m^3)$
  - subset - TSD : find the good size  $k$
  - iterative methods : iterate to  $k$ , complexity  $\mathcal{O}(km^2)$ ,  $k \ll m$
- iterative methods (CG, MR, Krylov subspace)
  - regularizing
  - fast
  - generic