Goal To discuss how a class of regularization methods originally designed for solving ill-posed inverse problems, give rise to regularized learning algorithms. These algorithms are kernel methods that can be easily implemented and have a common derivation, but different computational and theoretical properties.
Plan

- From ERM to Tikhonov regularization.
- Linear ill-posed problems and stability.
- Spectral Regularization and Filtering.
- Example of Algorithms.
training set $S = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$.

$X$ is the $n$ by $d$ input matrix.

$Y = (Y_1, \ldots, Y_n)$ is the output vector.

$k$ denotes the kernel function, $K$ the $n$ by $n$ kernel matrix with entries $K_{ij} = k(X_i, X_j)$ and $\mathcal{H}$ the RKHS with kernel $k$.

RLS estimator solves

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2 + \lambda \| f \|^2_{\mathcal{H}}.$$
We have seen that RKHS allow us to write the RLS estimator in the form

\[ f^\lambda_S(X) = \sum_{i=1}^{n} c_i k(X, X_i) \]

with

\[ (K + n\lambda I)c = Y \]

where \( c = (c_1, \ldots, c_n) \).
Similarly we can prove that the solution of empirical risk minimization

\[
\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2
\]

can be written as

\[
f_S(X) = \sum_{i=1}^{n} c_i k(X, X_i)
\]

where the coefficients satisfy

\[
Kc = Y.
\]
We observed that adding a penalization term can be interpreted as way to to control smoothness and avoid overfitting

\[
\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2 \Rightarrow \min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2 + \lambda \|f\|_{\mathcal{H}}^2.
\]
Now we can observe that adding a penalty has an effect from a numerical point of view:

\[ Kc = Y \Rightarrow (K + n\lambda I)c = Y \]

It stabilizes a possibly ill-conditioned matrix inversion problem.

This is the point of view of regularization for (ill-posed) inverse problems.
Hadamard introduced the definition of ill-posedness. Ill-posed problems are typically inverse problems.

If \( g \in G \) and \( f \in F \), with \( G, F \) Hilbert spaces, a linear, continuous operator \( L \), consider the equation

\[
g = Lf.
\]

The direct problem is is to compute \( g \) given \( f \); the inverse problem is to compute \( f \) given the data \( g \).

The inverse problem of finding \( f \) is well-posed when

- the solution exists,
- is unique and
- is stable, that is depends continuously on the initial data \( g \).

Otherwise the problem is ill-posed.
Hadamard introduced the definition of ill-posedness. Ill-posed problems are typically inverse problems. If \( g \in G \) and \( f \in F \), with \( G, F \) Hilbert spaces, a linear, continuous operator \( L \), consider the equation

\[
g = Lf.
\]

The direct problem is is to compute \( g \) given \( f \); the inverse problem is to compute \( f \) given the data \( g \). The inverse problem of finding \( f \) is well-posed when

- the solution exists,
- is unique and
- is stable, that is depends continuously on the initial data \( g \).

Otherwise the problem is ill-posed.
Hadamard introduced the definition of ill-posedness. Ill-posed problems are typically inverse problems. If $g \in G$ and $f \in F$, with $G, F$ Hilbert spaces, a linear, continuous operator $L$, consider the equation

$$g = Lf.$$

The direct problem is to compute $g$ given $f$; the inverse problem is to compute $f$ given the data $g$. The inverse problem of finding $f$ is well-posed when

- the solution exists,
- is unique and
- is stable, that is depends continuously on the initial data $g$.

Otherwise the problem is ill-posed.
In the finite dimensional case the main problem is numerical stability.

For example, in the learning setting the kernel matrix can be decomposed as $K = Q \Sigma Q^t$, with $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_n)$, $\sigma_1 \geq \sigma_2 \geq \ldots \sigma_n \geq 0$ and $q_1, \ldots, q_n$ are the corresponding eigenvectors. Then

$$c = K^{-1} Y = Q \Sigma^{-1} Q^t Y = \sum_{i=1}^{n} \frac{1}{\sigma_i} < q_i, Y > q_i.$$

In correspondence of small eigenvalues, small perturbations of the data can cause large changes in the solution. The problem is ill-conditioned.
In the finite dimensional case the main problem is numerical stability.

For example, in the learning setting the kernel matrix can be decomposed as $K = Q\Sigma Q^t$, with $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_n)$, $\sigma_1 \geq \sigma_2 \geq \ldots \sigma_n \geq 0$ and $q_1, \ldots, q_n$ are the corresponding eigenvectors.

Then

$$c = K^{-1} Y = Q\Sigma^{-1} Q^t Y = \sum_{i=1}^{n} \frac{1}{\sigma_i} < q_i, Y > q_i.$$ 

In correspondence of small eigenvalues, small perturbations of the data can cause large changes in the solution. The problem is ill-conditioned.
Regularization as a Filter

For Tikhonov regularization

\[ c = (K + n\lambda I)^{-1} Y = Q(\Sigma + n\lambda I)^{-1} Q^t Y = \sum_{i=1}^{n} \frac{1}{\sigma_i + n\lambda} < q_i, Y > q_i. \]

Regularization filters out the undesired components. For \( \sigma \gg \lambda n \), then \( \frac{1}{\sigma_i + n\lambda} \sim \frac{1}{\sigma_i} \). For \( \sigma \ll \lambda n \), then \( \frac{1}{\sigma_i + n\lambda} \sim \frac{1}{\lambda n} \).
Note that we can look at a scalar function $G_\lambda(\sigma)$ as a function on the kernel matrix.

Using the eigen-decomposition of $K$ we can define

$$G_\lambda(K) = QG_\lambda(\Sigma)Q^T,$$

meaning

$$G_\lambda(K)Y = \sum_{i=1}^{n} G_\lambda(\sigma_i) < q_i, Y > q_i.$$

For Tikhonov

$$G_\lambda(\sigma) = \frac{1}{\sigma + n\lambda}.$$
In the inverse problems literature many algorithms are known besides Tikhonov regularization.

Each algorithm is defined by a suitable filter function $G_\lambda$.

This class of algorithms is known collectively as spectral regularization.

Algorithms are not necessarily based on penalized empirical risk minimization.
The idea of using regularization from inverse problems in statistics (see Wahba) and machine learning (see Poggio and Girosi) is now well known.

Ideas coming from inverse problems regarded mostly the use of Tikhonov regularization.

The notion of filter function was studied in machine learning and gave a connection between function approximation in signal processing and approximation theory. The work of Poggio and Girosi enlightened the relation between neural network, radial basis function and regularization.

Filtering was typically used to define a penalty for Tikhonov regularization, in the following it is used to define algorithms different though similar to Tikhonov regularization.
Besides Tikhonov regularization we consider the following algorithms:

- Gradient Descent or Landweber Iteration or L2 Boosting
- \(\nu\)-method, accelerated Landweber.
- Iterated Tikhonov
- Truncated Singular Value Decomposition (TSVD) Principal Component Regression (PCR)

The spectral filtering perspective leads to a unified framework.
Not every scalar function defines a regularization scheme.

Roughly speaking a good filter function must have the following properties:

- as \( \lambda \) goes to 0, \( G_\lambda(\sigma) \rightarrow 1/\sigma \) so that
  \[
  G_\lambda(K) \rightarrow K^{-1}.
  \]

- \( \lambda \) controls the magnitude of the (smaller) eigenvalues of \( G_\lambda(K) \).
We can define a class of Kernel Methods as follows.

Spectral Regularization

We look for estimators

$$f_S^\lambda(X) = \sum_{i=1}^n c_i k(X, X_i)$$

where

$$c = G_\lambda(K) Y.$$
Consider the (Landweber) iteration:

\[
\text{set } c^0 = 0 \\
\text{for } i = 1, \ldots, t - 1 \\
\quad c^i = c^{i-1} + \eta (Y - Kc^{i-1})
\]

If the largest eigenvalue of $K$ is smaller than $n$ the above iteration converges if we choose the step-size $\eta = 2/n$.

The above iteration can be seen as the minimization of the empirical risk

\[
\frac{1}{n} \| Y - Kc \|_2^2
\]

via gradient descent.
Gradient Descent as Spectral Filtering

Note that $c^0 = 0$, $c^1 = \eta Y$,

$$c^2 = \eta Y + \eta (I - \eta K) Y$$

$$c^3 = \eta Y + \eta (I - \eta K) Y + \eta (Y - K(\eta Y + \eta (I - \eta K) Y))$$

$$= \eta Y + \eta (I - \eta K) Y + \eta (I - 2\eta K + \eta^2 K^2) Y$$

One can prove by induction that the solution at the $t$–th iteration is given by

$$c = \eta \sum_{i=0}^{t-1} (I - \eta K)^i Y.$$ 

The filter function is

$$G_{\lambda}(\sigma) = \eta \sum_{i=0}^{t-1} (I - \eta \sigma)^i.$$
Note that $\sum_{i \geq 0} x^i = 1/(1 - x)$, also holds replacing $x$ with the a matrix. If we consider the kernel matrix (or rather $I - \eta K$) we get

$$K^{-1} = \eta \sum_{i=0}^{\infty} (I - \eta K)^i Y \sim \eta \sum_{i=0}^{t-1} (I - \eta K)^i Y.$$ 

The filter function of Landweber iteration corresponds to a truncated power expansion of $K^{-1}$. 

L. Rosasco
Spectral Regularization
The regularization parameter is the number of iteration. Roughly speaking $t \sim 1/\lambda$.

- Large values of $t$ correspond to minimization of the empirical risk and tend to overfit.
- Small values of $t$ tends to oversmooth, recall we start from $c = 0$.

Early stopping of the iteration has a regularization effect.
Gradient Descent at Work

L. Rosasco

Spectral Regularization
Gradient Descent at Work

L. Rosasco
Spectral Regularization
Landweber iteration (or gradient descent) has been rediscovered in statistics with name of $L_2$ Boosting.

**Boosting**

- Then name *Boosting* denotes a large class of methods building estimators as linear (convex) combinations of weak learners.
- Many boosting algorithms can be seen as gradient descent minimization of the empirical risk on the linear span of some basis function.

For Landweber iteration the weak learners are $k(X_i, \cdot), i = 1, \ldots, n$. 
The so called $\nu$-method or accelerated Landweber iteration can be thought as an accelerated version of gradient descent.

The filter function is $G_t(\sigma) = p_t(\sigma)$ with $p_t$ a polynomial of degree $t - 1$.

The regularization parameter (think of $1/\lambda$) is $\sqrt{t}$ (rather than $t$): fewer iterations are needed to attain a solution.
The method is implemented by the following iteration.

**Gradient Descent**

1. Set $c_0 = 0$
2. $\omega_1 = (4\nu + 2)/(4\nu + 1)$
3. $c_1 = c_0 + \frac{\omega_1}{n} (Y - Kc_0)$

For $i = 2, \ldots, t - 1$

- $c_i = c_{i-1} + u_i (c_{i-1} - c_{i-2}) + \frac{\omega_i}{n} (Y - Kc_{i-1})$
- $u_i = \frac{(i-1)(2i-3)(2i+2\nu-1)}{(i+2\nu-1)(2i+4\nu-1)(2i+2\nu-3)}$
- $\omega_i = 4 \frac{(2i+2\nu-1)(i+\nu-1)}{(i+2\nu-1)(2i+4\nu-1)}$
The following method can be seen a combination of Tikhonov regularization and gradient descent.

**Gradient descent**

\[
\text{set } c_0 = 0 \\
\text{for } i = 0, \ldots, t - 1 \\
(K + n\lambda I)c_i = Y + n\lambda c_{i-1}
\]

The filter function is:

\[
G_\lambda(\sigma) = \frac{(\sigma + \lambda)^t - \lambda^t}{\sigma(\sigma + \lambda)^t}.
\]
Both the number of iteration and $\lambda$ can be seen as regularization parameters.

It can be used to enforce more smoothness on the solution.

Tikhonov regularization suffers from a *saturation* effect: it cannot exploit the regularity of the solution beyond a certain critical value.
This method is one of the oldest regularization techniques and is also called spectral cut-off.

Given the eigen-decomposition $K = Q \Sigma Q^t$, a regularized inverse of the kernel matrix is built discarding all the eigenvalues before the prescribed threshold $\lambda n$.

It is described by the filter function $G_{\lambda}(\sigma) = 1/\sigma$ if $\sigma \leq \lambda/n$ and 0 otherwise.
Principal component Analysis is a well known dimensionality reduction technique often used as preprocessing in learning.

Assuming centered data, $X^t X$ is the covariance matrix and its eigenvectors $(V^j)^d_{j=1}$ are the principal components.

PCA amounts to map each example $X_i$ in

$$\tilde{X}_i = (X_i^t V^1, \ldots, X_i^t V^m)$$

where $m < \min\{n, d\}$.

notation: $X_i^t$ is the transpose of the first row (example) of $X$. 

L. Rosasco  Spectral Regularization
The above algorithm can be written using only the linear kernel matrix $XX^t$ and its eigenvectors $(U^i_i)_{i=1}^n$.

The eigenvalues of $XX^t$ and $X^tX$ are the same and

$$V^i = \frac{1}{\sqrt{\sigma_i}} X^t U^i_j.$$

Then

$$\tilde{X}_i = (\frac{1}{\sqrt{\sigma_i}} \sum_{j=1}^n U^i_j X^t_i X_j), \ldots, \frac{1}{\sqrt{\sigma_n}} \sum_{j=1}^n U^m_j X^t_i X_j).$$

Note that $X^t_i X_j = k(X_i, X_j)$. 

We can perform a non linear principal component analysis, namely KPCA, by choosing non linear kernel functions.

Using $K = Q \Sigma Q^T$ we can rewrite the projection in vector notation.

If we let $\Sigma_M = diag(\sigma_1, \cdots, \sigma_m, 0, \cdots, 0)$ then the projected data matrix $\tilde{X}$ is

$$\tilde{X} = KQ\Sigma_m^{-1/2}$$
ERM on the projected data

\[
\min_{\beta \in \mathbb{R}^m} \left\| Y - \beta \tilde{X} \right\|_n^2,
\]

is equivalent to perform truncated singular values decomposition on the original problem.

Representer Theorem tells us that

\[
\beta^t \tilde{X}_i = \sum_{j=1}^{n} \tilde{X}_j^t \tilde{X}_i c_j
\]

with

\[
c = (\tilde{X} \tilde{X}^t)^{-1} Y.
\]
Using $\tilde{X} = KQ\Sigma_m^{-1/2}$ we get

$$\tilde{X}\tilde{X}^t = Q\Sigma Q^t Q\Sigma_m^{-1/2} \Sigma_m^{-1/2} Q^t Q\Sigma Q^t = Q\Sigma_m Q^t.$$ 

so that

$$c = Q\Sigma_m^{-1} Q^t Y = G_\lambda(K) Y,$$

where $G_\lambda$ is the filter function of TSVD.

The two procedure are equivalent. The regularization parameter is the eigenvalue threshold in one case and the number of components kept in the other case.
Projection Regularizes!

Doing KPCA and then RLS is redundant.

If data are centered Spectral regularization (also Tikhonov) can see as filtered projection on the principal components.
we have seen that

$$G_\lambda(K) \to K^{-1} \text{ if } \lambda \to 0$$

anyway usually, we DON’T want to solve

$$Kc = Y$$

since it would simply correspond to an over-fitting solution

**stability vs generalization**

how can we show that **stability** ensures **generalization**?
It is useful to consider what happens if we know the true distribution.

**integral operator**

for $n$ large enough

$$\frac{1}{n} K \sim L_k f(s) = \int_X k(x, s)f(x)p(x)dx$$

**the ideal problem**

for $n$ large enough we have

$$Kc = Y \sim L_k f = L_k f_\rho$$

where $f_\rho$ is the regression (target) function defined by

$$f_\rho(x) = \int_Y yp(y|x)dy$$
it can be shown that which is the least squares problem associated to $L_k f = L_k f_\rho$.

tikhonov regularization in this case is simply

or equivalently

$$f^\lambda = (L_k f + \lambda I)^{-1} L_k f_\rho$$
Fourier Decomposition of the Regression Function

Fourier decomposition of $f_\rho$ and $f^\lambda$

If we diagonalize $L_k$ to get the eigensystem $(t_i, \phi_i)_i$ we can write

$$f_\rho = \sum_i \langle f_\rho, \phi_i \rangle \phi_i$$

Perturbations affect high order components. Tikhonov regularization can be written as

$$f^\lambda = \sum_i \frac{t_i}{t_i + \lambda} \langle f_\rho, \phi_i \rangle \phi_i$$

Sampling IS a perturbation

Stabilizing the problem with respect to random discretization (sampling) we can recover $f_\rho$
Iterative methods perform matrix vector multiplication $O(n^2)$ at each iteration and the regularization parameter is the number of iteration itself.

There is not a closed form for leave one out error.

Parameter tuning is different from method to method.

- Compared to RLS in iterative and projected methods the regularization parameter is naturally discrete.
- TSVD has a natural range for the search of the regularization parameter.
- For TSVD the regularization parameter can be interpreted in terms of dimensionality reduction.
Many different principles lead to regularization: penalized minimization, iterative optimization, projection. The common intuition is that they enforce stability of the solution.

All the methods are implicitly based on the use of square loss. For other loss function different notion of stability can be used.