Regularization methods for large scale machine learning

Lorenzo Rosasco

March 7, 2017

Abstract

After recalling an inverse problems perspective on supervised learning, we discuss regularization methods for large scale machine learning. In particular, we derive and contrast different regularization schemes. Starting from classic Tikhonov regularization, we then introduce iterative regularization, a.k.a. early stopping, and discuss different variants including accelerated and stochastic versions. Finally we discuss projection with regularization and introduce stochastic extensions. Our discussion shows how, while the different methods are grounded in common estimation principles, their computational properties are different. Iterative regularization allows to combine statistical and time complexities. While regularization with stochastic projections allows to simultaneously control statistical, time and space complexity. These latter properties makes these method particularly suited to large scale setting.

Contents

1 Inverse problems

Inverse problems provide a general framework to describe a variety of applied problems. A linear inverse problem is defined by a linear equation

$$Af = g \tag{1}$$

where $A : \mathcal{H} \to \mathcal{G}$ is a linear continuos operator between Hilbert spaces and $f \in \mathcal{H}, g \in \mathcal{G}$. Given A and g, the problem is to recover f. This problem is often ill-posed, that is:

- a solution might not exist $g \notin \text{Range}(A)$,
- it might now be unique $\text{Ker}(A) \neq \emptyset$,
- it might not depend continuously to the datum g .

The question is how to find well-posed approximate solutions to the above problem. The first two requirements can be fixed considering

$$H_0 = \operatorname*{argmin}_{f \in \mathcal{H}} \big\| Af - g \big\|^2$$

which is not empty under the weaker condition $Pg \in \text{Range}(A)$, and letting,

$$f^\dagger = \operatorname*{argmin}_{f \in \mathcal{H}_0} \big\| f \big\|.$$

The function $f^\dagger$ can be shown to be unique and is called pseudo-solution or Moore-Penrose solution, since it can be shown that $f^\dagger = A^\dagger g$, where $A^\dagger : \mathcal{G} \to \mathcal{H}$ denotes the Moore-Penrose pseudo-inverse. The latter is typically not continuous and the question is how to derive approximations to ensure stability. This question is particularly important as data in practice might be affected by noise. A common way to formalize this idea is assuming to be given g_δ rather than g where

$$\big\| g - g_\delta \big\| \leq \delta$$

and $\delta > 0$ is seen as a noise level.

Regularization theory provides a general framework to derive stable solutions. Broadly speaking regularization refers to a sequence of solutions that converge to $f^\dagger$ and is stable to noise. A classic example is Tikhonov regularization given by

$$f_\delta^\lambda = \operatorname*{argmin}_{f \in \mathcal{H}} \big\| Af - g_\delta \big\|^2 + \lambda \|f\|_{\mathcal{H}}^2, \quad \lambda > 0$$

Classical results in regularization theory show that if λ is chosen as function λ_δ of delta such that

$$\lim_{\delta \to 0} \lambda_\delta = 0, \qquad \lim_{\delta \to 0} \frac{\delta}{\lambda_\delta} = 0,$$

then

$$\lim_{\delta\to 0}\left\| f_\delta^{\lambda_\delta} - f^\dagger \right\| = 0$$

In the following, we introduce the problem of supervised learning, show how it can be seen as an inverse problem and discuss how regularization techniques can be adapted to the learning setting.

2 Statistical Learning Theory

In this section we briefly introduce the problem of supervised learning.

Supervised learning is concerned with the problem of learning a function from random samples. More precisely, consider a probability space $\mathcal{X}$ and assume ρ to be a probability measure on $\mathcal{Z} = \mathcal{X} \times \mathbb{R}$, such that for all measurable functions $h : \mathcal{X} \times \mathbb{R} \to \mathbb{R}$,

$$\int f(x,y)d\rho(x,y) = \int f(x,y)d\rho_{\mathcal{X}}(x)\rho(y|x) \tag{2}$$

where $\rho_{\mathcal{X}}$ is called the marginal measure on $\mathcal{X}$ and $\rho(\cdot\,|x)$ the conditional probability measure on $\mathbb{R}$ for almost all $x \in \mathcal{X}$. Let

$$L^2(\mathcal{X},\rho_{\mathcal{X}}) = \left\{ f : \mathcal{X} \to \mathbb{R} \mid \left\|f\right\|_\rho^2 = \int |f(x)|^2 d\rho_{\mathcal{X}}(x) < \infty \right\}$$

The function of interest is the *regression function*, defined by

$$f_\rho(x) = \int y\rho(y|x)$$

for almost all $x \in \mathcal{X}$. The distribution ρ, hence the regression function, are fixed but known only through a set $\mathbf{z}_n = (x_1,y_1),\ldots,(x_n,y_n) \in Z^n$ sampled independently and identically according to ρ. Given $\mathbf{z}$, the goal is find an estimate f_n of the regression function f_ρ. Assuming $f_\rho \in L^2(\mathcal{X},\rho_{\mathcal{X}})$, a natural metric to measure the quality of the estimate is the norm in $L^2(\mathcal{X},\rho_{\mathcal{X}})$. Indeed, if we consider

$$\left\|f_n - f_\rho\right\|_\rho^2 = \int |f_n(x) - f_\rho(x)|^2 d\rho_{\mathcal{X}}(x)$$

points that are more likely to be sampled will have more influence on the error. We will se that other error measures are also possible. The above quantity is stochastic through its dependence to the data-set $\mathbf{z}$. In statistical learning theory the focus is on studying the convergence as well as explicit bounds on the probability

$$\rho^n\left\{ \mathbf{z}_n \in Z^n \;\middle|\; \left\|f_n - f_\rho\right\|_\rho^2 \geq \epsilon \right\}$$

for all $\epsilon \in (0,\infty)$. We next discuss how the above problem can be reformulated as a linear inverse problems. We first add one remark and discuss two basic examples of the above framework.

Remark 1 (Risk Minimization)**.** *It is standard in statistical learning to view the regression function as a solution of a stochastic optimization problem. Indeed, if $\int y^2\rho(y|x) < \infty$ is finite, then the so called expected risk,*

$$\mathcal{E} : L^2(\mathcal{X},\rho_{\mathcal{X}}) \to \mathbb{R}, \qquad \mathcal{E}(f) = \int d\rho(x,y)(y - f(x))^2, \tag{3}$$

is well defined, continuous and convex. A direct computation shows that f_ρ is the minimizer of the expected risk on $L^2(\mathcal{X}, \rho_\mathcal{X})$ and moreover the following equality holds for all $f \in L^2(\mathcal{X}, \rho_\mathcal{X})$,

$$\left\|f - f_\rho\right\|_\rho^2 = \mathcal{E}(f) - \mathcal{E}(f_\rho).$$

The above quantity is called the excess expected risk.

Example 1 (Regression)**.** *For all $i = 1, \dots, n$, $n \in \mathbb{N}$, let x_i be a sequence of random points in $\mathcal{X}$, for example $\mathcal{X} = \mathbb{R}^d$, $d \in \mathbb{N}$, and ϵ_i a sequence of random numbers with zero mean, bounded variance and possibly dependent on x_i. Given a function $f_* : \mathcal{X} \to \mathbb{R}$, assume*

$$y_i = f_*(x_i) + \epsilon_i, \quad i = 1, \dots, n. \tag{4}$$

In other words, data are samples of a function corrupted with noise and evaluated at random locations. The above is the classical model for regression. It is a special case of the general framework in this section where $f_\rho = f_*$ and the conditional distribution is defined by the noise distribution.

Example 2 (Binary Classification)**.** *Consider the case where the conditional distribution is supported on $\{-1, 1\}$, that is it corresponds to the pair of point masses $\rho(1|x), \rho(-1|x)$ for almost all $x \in \mathcal{X}$. In this case, the natural error measure is the misclassification risk*

$$R(f) = \rho\{(x, y) \in Z \mid f(x)y < 0\}$$

that is the expected number of misclassifications. In this setting, it is not hard to show that the misclassification risk is minimized by the so called Bayes decision rule $b_\rho = sign(f_\rho - 1/2)$. and moreover

$$R(f) - R(b_\rho) \leq \left\|f - f_\rho\right\|_\rho.$$

This latter observation justifies the use of least squares for classification problems.

3 Learning as an inverse problem

In this section we reformulate the problem of learning with least squares as linear inverse problems under a suitable data model.

As a starter note that, considering the empirical data, it is well known that problem (4) could be formulated as a linear inverse problem with discrete data. Indeed, in this case it is natural to consider the candidate functions to belong to a Hilbert space of functions $\mathcal{H}$ where the evaluation functionals $f \mapsto f(x)$ are continuous, for all $x \in X$. Then, Rietz representation theorem, ensures that for all $x \in \mathcal{X}$ there exists a function $K_x \in \mathcal{H}$ such that

$$f(x) = \langle f, K_x \rangle_\mathcal{H}$$

where $\langle \cdot, \cdot \rangle_\mathcal{H}$ is the inner product in $\mathcal{H}$.

If we define the *sampling operator*,

$$S_n : \mathcal{H} \to \mathbb{R}^n, \qquad (S_n f)^i = \langle f, K_{x_i} \rangle_\mathcal{H}, \quad i = 1, \dots, n,$$

then problem (4) can be written as the linear inverse problem corresponding to finding $f \in \mathcal{H}$ such that

$$S_n f = \mathbf{y} \tag{5}$$

for $\mathbf{y} = (y_1, \ldots, y_n)$.

While the above is a promising start, it essentially corresponds to a "noisy" inverse problem, in the sense that we have only an empirical problem based on data. While this problem is the basis for practical algorithms, it is not clear how it relates to the problem of estimating the regression function, which is the target of learning. The question is then if the problem of estimating the regression function can be written as a linear inverse problem, of which problem (5) is an empirical instantiation. Roughly speaking the answer follows considering the *infinite data* limit of problem (5).

Assume the reproducing kernel K to be measurable and the operator

$$S_\rho : \mathcal{H} \to L^2(\mathcal{X}, \rho_\mathcal{X}), \qquad (S_\rho f)(x) = \langle f, K_x \rangle_\mathcal{H}, \quad \rho_\mathcal{X} - \text{almost surely},$$

to be bounded. Then, consider the linear inverse problem

$$S_\rho f = f_\rho. \tag{6}$$

The above inverse problem can be seen as the one corresponding to estimating the regression function. We add three remarks to illustrate the above discussion.

Remark 2 (Risk and Moore Penrose Solution)**.** *First, in words, the above inverse problem corresponds to looking for a function in $\mathcal{H}$ providing a good approximation of the regression function. This problem is typically ill-posed, in particular note that generally the regression function does not belong to $\mathcal{H}$. The associated least squares problem is*

$$\min_{f \in \mathcal{H}} \left\| S_\rho f - f_\rho \right\|_\rho^2 \tag{7}$$

which in light of Remark 1 corresponds to considering

$$\min_{f \in \mathcal{H}} \mathcal{E}(f).$$

The solutions of the above problem, if any, are the set of generalized solutions of problem (6)*. If this set is not empty, then we denote by $f_\mathcal{H}^\dagger$ the Moore-Penrose solution, that is the generalized solution with minimal (RKHS) norm. Such a solution often replaces the regression function as the target of learning. Note that $f_\mathcal{H}^\dagger$ can be written as $f_\mathcal{H}^\dagger = S_\rho^\dagger f_\rho$ and contrasted to the Moore-Penrose solution $f_\mathcal{H}^\dagger = S_n^\dagger \mathbf{y}$ of problem* (5)*.*

Remark 3 (Empirical and population problems)**.** *Second, if $\rho_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ is the empirical measure on the data, then we can identify $\mathbb{R}^n$ with $L^2(\mathcal{X}, \rho_n)$, and S_ρ reduces to S_n if we replace ρ by ρ_n. Developing this latter observation we can view problem* (6) *as the ideal inverse problem we would wish to solve, and to problem* (5) *as a corresponding empirical problem. It is important to note that unlike classical inverse problems, here the operators defining the two problems have same domains but different ranges. We will see in the following how the distance (noise) between the two problems can be quantified. Provided with the above connection we next introduce and analyze a class of regularized methods for learning.*

Remark 4 (Noise and sampling)**.** *Following the above remark, problem* (5) *can be seen as a noisy randomly discretized version of Problem* (6)*. Note however, that it is not immediately clear how this idea can be formalized since the operators defininining the two problems have different range (*$\mathbf{y}$ *is a vector and f_ρ a function!). One idea is to consider the normal equations associated to the two problems that is*

$$S_\rho^* S_n f = S_\rho^* \mathbf{y}, \qquad S_\rho^* S_\rho f = S_\rho^* f_\rho.$$

This suggests to consider

$$\|S_\rho^* f_\rho - S_\rho^* \mathbf{y}\|, \qquad \|S_\rho^* S_\rho - S_\rho^* S_n\|.$$

The above quantities provides a measure of the perturbation due to random noise and random sampling. As seen in the following they will play a role similar to the noise level in classical inverse problems.

Remark 5 (Connection to compressed sensing and linear regression)**.** *Note that the sampling operator can be seen as a collection of measurements defined by random linear functionals. This suggests a connection to classical linear regression and compressed sensing. Indeed, if we consider the linear kernel, then problem* (5) *can be written as*

$$X_n w = \mathbf{y}$$

where X_n is the n by d data matrix, $y_i = x_i^\top w_* + \epsilon_i$ and w_* is a parameter to be estiamted. Unlike in compressed sensing, the source of randomness in the sampling operator lies in the nature of the data and it is not a design choice.

Remark 6 (Kernels and RKHS)**.** *The space $\mathcal{H}$ is called reproducing kernel Hilbert space (RKHS) and the function $K(x, x') = \langle K_x, K_x' \rangle_\mathcal{H}$, $x, x' \in X$ reproducing kernel. It can be shown that $\mathcal{H}$ is the closure of the span span$\{K_x \mid x \in \mathcal{X}\}$. The list of examples of kernels and RKHS is endless. We provide three examples.*

- **Linear**. *Let $\mathcal{X} = \mathbb{R}^d$ and consider the kernel $K(x, x') = x^\top x'$, for all $x, x' \in \mathcal{X}$. The corresponding RKHS is the space of linear functions on $\mathcal{X}$.*
- **Finite dictionaries**. *Consider $\{\phi_i : \mathcal{X} \to \mathbb{R} \mid i = 1, \dots, p\}$ and the kernel $K(x, x') = \sum_{j=1}^p \phi_j(x)\phi_j(x')$ for all $x, x' \in \mathcal{X}$. The corresponding RKHS is $\mathcal{H} = \{f : \mathcal{X} \to \mathbb{R} \ : \ \exists w \in \mathbb{R}^p \ \text{ such that } \ f = \sum_{j=1}^p w^j \phi_j\}$*
- **Gaussian**. *Let $\mathcal{X} = \mathbb{R}^d$ and consider the kernel $K(x, x') = e^{-\left\|x - x'\right\|^2 \gamma}$, for all $x, x' \in \mathcal{X}$. The corresponding RKHS can be seen as a subspace $\mathcal{H}$ of $L^2(\mathcal{X}) = \{f : \mathcal{X} \to \mathbb{R} \ : \ \int |f(x)|^2| < \infty\}$ such that, for all $f \in \mathcal{H}$*

  $$\int |\tilde{f}(\omega)|^2 e^{\frac{\omega^2}{\gamma}} < \infty$$

  where $\tilde{f}$ denotes the Fourier Transform of f.

The above discussion raises at least two lines of questions. The first concerns, the nature of the inverse problem describing supervised learning. We investigate this in the first section, analyzing the operators defining the problem. The second

4 An interlude: operators defined by the kernels

The above discussion can be further elucidated considering in details the operators defined by the kernel.

4.1 Population kernel operators

We begin noting that functions in $\mathcal{H}$ are defined over the whole space $\mathcal{X}$, while functions in $L^2(\mathcal{X}, \rho_\mathcal{X})$ are defined on $\mathcal{X}_\rho$, the support of the distribution $\rho_\mathcal{X}$ which can be strictly contained in $\mathcal{X}$. Indeed, it is

often interesting to think of $\mathcal{X}$ as a high dimensional Euclidean space and $\mathcal{X}_\rho$ as smaller set, for example a low dimensional sub-manifold.

In this view, the operator S_ρ can be seen as restriction operator, that given a function defined over the whole space $\mathcal{X}$ provides a restriction to $\mathcal{X}_\rho$. The corresponding adjoint operator $S_n^* : L^2(\mathcal{X}, \rho_\mathcal{X}) \to \mathcal{H}$ can be shown to have the following form

$$S^* g = \int d\rho_\mathcal{X}(x) K_x g, \qquad \forall g \in L^2(\mathcal{X}, \rho_\mathcal{X})$$

and can be seen as an extension operator. Given a function g defined on $\mathcal{X}_\rho$ it provides an harmonic extension on the whole space $\mathcal{X}$ defined by the kernel K. The operator $L_\rho = S_\rho S_\rho^* : L^2(\mathcal{X}, \rho_\mathcal{X}) \to L^2(\mathcal{X}, \rho_\mathcal{X})$ is the integral operator defined by the kernel

$$L_\rho g(x) = \int d\rho_\mathcal{X} K(x, x') g(x') d\rho_\mathcal{X}(x'), \qquad \forall g \in L^2(\mathcal{X}, \rho_\mathcal{X}), \tag{8}$$

and $\rho_\mathcal{X}-$almost everywhere. The operator $T_\rho = S_\rho^* S_\rho : \mathcal{H} \to \mathcal{H}$ can be written as

$$T_\rho = \int d\rho_\mathcal{X} K_x \otimes K_x,$$

where $K_x \otimes K_x = \langle K_x, \cdot \rangle_\mathcal{H} K_x$, so that

$$\langle T_\rho f, f' \rangle_\mathcal{H} = \int d\rho_\mathcal{X} f(x) f'(x), \qquad \forall f, f' \in \mathcal{H}.$$

As discussed below T_ρ can be seen as a suitable covariance operator.

Remark 7 (Properties of the kernel operators)**.** *If the kernel is bounded, that it there exists $\kappa > 0$ such that*

$$K(x, x') \leq \kappa^2$$

$\rho_\mathcal{X}-$ almost everywhere, then all the above operators are well defined. The operators L_ρ, T_ρ are positive, self-adjoint and trace class and the operators S_ρ, S_ρ^* are Hilbert-Schmidt.

4.2 Empirical kernel operators

The operator S_n is called sampling operator. Given a function in $\mathcal{H}$ it evaluates the function at the training set inputs. The corresponding adjoint operator $S_n^* : \mathbb{R}^n \to \mathcal{H}$ can be shown to have the following form

$$S_n^* c = \frac{1}{n} \sum_{i=1}^n K_{x_i} c^i, \qquad \forall c \in \mathbb{R}^n.$$

As discussed above $\mathbb{R}^n$ can be identified with $L^2(\mathcal{X}, \rho_n)$, whereas the latter can be seen as space of functions defined on the training set inputs. In this view, we can identify c $f(x_1), \ldots, f(x_n)$, the action of S_n^* can be seen as an extension operator providing the value of the functions outside of the training set inputs. Such an operator is said to provide an *out-of-sample extension*. The operator $L_n = S_n S_n^* : L^2(\mathcal{X}, \hat{\rho_\mathcal{X}}) \to L^2(\mathcal{X}, \hat{\rho_\mathcal{X}})$ can be written as

$$(L_n c)^i = \frac{1}{n} \sum_{j=1}^n K(x_i, x_j) c^j.$$

The above operator can be seen as discretization of the integral operator in (8) and in particular as a so called N ystrom approximation. The operator $T_n = S_n^* S_n : \mathcal{H} \to \mathcal{H}$ can be written as

$$T_n = \frac{1}{n} \sum_{j=1}^{n} K_{x_i} \otimes K_{x_i},$$

so that

$$\left\langle T_n f, f' \right\rangle_{\mathcal{H}} = \frac{1}{n} \sum_{j=1}^{n} f(x_i) f'(x_i), \qquad \forall f, f' \in \mathcal{H}.$$

As discussed below T_n can be seen as a suitable empirical covariance operator.

4.3 The linear kernel case

The above operators takes a familiar form if we consider the linear kernel. In this case the RKHS can be identified with $\mathbb{R}^d$ and the sampling operator S_n with the n by d data matrix X_n whose rows are the training set input points. The adjoint S_n^* is the transpose of X_n (multiplied by $1/n$) and $S_n^* S_n$ is the empirical covariance matrix[1]

$$\Sigma_n = \frac{1}{n} X_n^\top X_n = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^\top.$$

In the population case the only operator that have a familiar form is $S_\rho^* S_\rho$ that can be seen as the population covariance

$$\Sigma = \mathbb{E}[\frac{1}{n} X_n^\top X_n] = \int d\rho_{\mathcal{X}}(x) x x^\top.$$

5 Tikhonov regularization

Following, the connection discussed before, consider the family of variational problems,

$$\min_{f \in \mathcal{H}} \left\| S_n f - \mathbf{y} \right\|_n^2 + \lambda \|f\|_{\mathcal{H}}^2, \quad \lambda > 0, \tag{9}$$

where $\|\cdot\|_{\mathcal{H}}$ is the norm in $\mathcal{H}$, $\left\|\cdot\right\|_n$ the norm in $\mathbb{R}^n$ (normalized by 1/n) and it is easy to see that

$$\left\| S_n f - \mathbf{y} \right\|_n^2 = \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2. \tag{10}$$

A direct computation shows that the minimizer of problem (9) is given by

$$f_n^\lambda = (S_n^* S_n + \lambda I)^{-1} S_n^* \mathbf{y}, \qquad \forall \lambda > 0, \tag{11}$$

where $S_n^* : \mathcal{H} \to L^2(\mathcal{X}, \hat{\rho_{\mathcal{X}}})$ is the adjoint of the sampling operator. Note that, while the sampling operator is finite rank, in general, the above expression is not directly applicable. However, the following simple lemma holds.

[1] Or rather the second moment matrix.

Lemma 1. *For all $\lambda > 0$, let f_n^λ be defined as in* (11)*, then*

$$f_n^\lambda(x) = \sum_{i=1}^n K(x, x_i)c_i, \qquad \mathbf{c} = (K_n + \lambda n I)^{-1}\mathbf{y}, \tag{12}$$

where $\mathbf{c} = (c_1, \dots, c_n)$ and

$$K_n : L^2(\mathcal{X}, \hat{\rho_\mathcal{X}}) \to L^2(\mathcal{X}, \hat{\rho_\mathcal{X}}), \quad (K_n)_{i,j} = 1/nK(x_i, x_j), \quad \forall i, j = 1, \dots, n.$$

Proof. Note that

$$(S_n^* S_n + \lambda I)^{-1} S_n^* = S_n^*(S_n S_n^* + \lambda I)^{-1},$$

so that $f_n^\lambda = S_n^*(S_n S_n^* + \lambda I)^{-1}\mathbf{y}$. Recalling that $S_n S_n^* = \frac{1}{n}K_n$ and that $S_n^* c = \frac{1}{n}\sum_{i=1}^n K_{x_i} c_i$, for all $c \in \mathbb{R}^n$, by a direct computation we can write

$$(S_n S_n^* + \lambda I)^{-1} = n(K_n + \lambda n I)^{-1} = n\mathbf{c}$$

and

$$f_n^\lambda(x) = \left\langle K_x, f_n^\lambda \right\rangle_\mathcal{H} = \langle K_x, S_n^* n\mathbf{c}\rangle_\mathcal{H} = n \langle S_n K_x, \mathbf{c}\rangle_n = \sum_{i=1}^n K(x, x_i)c_i$$

□

5.1 Error analysis for Tikhonov regularization

We next provide an error analysis for Tikhonov regularization.

We make a few simplifying assumptions. We assume $K(x, x) \leq 1$ for all $x \in \mathcal{X}$, and assume the regression model

$$y_i = f_\mathcal{H}^\dagger(x_i) + \epsilon_i$$

where x_i are i.i.d. random vectors and ϵ_i zero mean random number smaller than 1. Note that this mean that

$$S_\rho f_\mathcal{H}^\dagger = f_\rho.$$

The main results os this section are the following two theorems.

Theorem 1. *Under the above assumptions, the following results holds:*

- *for all $\lambda > 0$, there exists constant c, C not depending on n, λ such that with probability at least $1 - Ce^{-\tau}$*

$$\|f_n^\lambda - f_\mathcal{H}^\dagger\|_\mathcal{H} \leq c\frac{\tau}{\lambda\sqrt{n}} + \|f^\lambda - f_\mathcal{H}^\dagger\|_\mathcal{H},$$

- *If λ is chosen as a function λ_n of the number of points so that*

$$\lim_{n\to\infty} \lambda_n = 0, \qquad \lim_{n\to\infty} \frac{1}{\lambda_n\sqrt{n}} = 0,$$

then

$$\lim_{n\to\infty} \|f_\delta^{\lambda_n} - f_\mathcal{H}^\dagger\|_\mathcal{H} = 0$$

almost surely.

Theorem 2. *Under the same assumption of the previous theorem, if*

$$f_{\mathcal{H}}^{\dagger} \in Range((S_{\rho}^{*}S_{\rho})^{-r}), \quad r > 0, \tag{13}$$

then the following results holds:

- *For all $\lambda > 0$, there exist constants c, c', C, C' not depending on n, λ such that with probability at least $1 - Ce^{-\tau}$*

$$\|f_n^{\lambda} - f_{\mathcal{H}}^{\dagger}\|_{\mathcal{H}} \leq c\frac{\tau}{\lambda\sqrt{n}} + c'\lambda^r$$

  if $0 \leq r \leq 1$.

- *If λ is chosen as a $\lambda_n = n^{-\frac{1}{2(r+1)}}$, there exist constant c, C not depending on n, λ such that with probability at least $1 - Ce^{-\tau}$*

$$\|f_{\delta}^{\lambda_n} - f_{\mathcal{H}}^{\dagger}\|_{\mathcal{H}} \leq cn^{-\frac{r}{2(r+1)}}$$

To derive the above bound we first consider a suitable error decomposition and then study the various error terms.

5.2 Error decomposition

The idea is to fist study the difference $\|f_n^{\lambda} - f_{\mathcal{H}}^{\dagger}\|_{\mathcal{H}}$ for any $\lambda > 0$, and then derive a suitable choice for λ. The idea is to decompose such an error into several terms. We begin by considering

$$f^{\lambda} = (S_{\rho}^{*}S_{\rho} + \lambda I)^{-1}S_{\rho}^{*}f_{\rho} \tag{14}$$

which is the unique solution of the problem

$$\min_{f \in \mathcal{H}} \|S_{\rho}f - f_{\rho}\|_{\rho}^{2} + \lambda\|f\|_{\mathcal{H}}^{2}$$

Then, we have the following equation

$$f_n^{\lambda} - f_{\mathcal{H}}^{\dagger} = f_n^{\lambda} - f^{\lambda} + f^{\lambda} - f_{\mathcal{H}}^{\dagger}.$$

In the above expression:

- The term $a^{\lambda} = f^{\lambda} - f_{\mathcal{H}}^{\dagger}$ does not depend on the data but only on the distribution and is called approximation error.

- The term $s_n^{\lambda} = f_n^{\lambda} - f^{\lambda}$ depends on the data, is stochastic, and is called variance, estimation or sample error.

We study this two terms next.

5.3 Approximation error

Combining with (**??**) we have

$$a^\lambda = f^\lambda - f^\dagger_{\mathcal{H}} = ((S^*_\rho S_\rho + \lambda I)^{-1} S^*_\rho S_\rho - I) f^\dagger_{\mathcal{H}} = \lambda (S^*_\rho S_\rho + \lambda I)^{-1} f^\dagger_{\mathcal{H}}.$$

A first question is whether the approximation error converges to zero, and indeed from the above equation it is possible to show that

$$\lim_{\lambda \to 0} \|a^\lambda\|_{\mathcal{H}} = 0. \tag{15}$$

A second question we can ask is if it possible to derive the rate of convergence for (15). This latter question can be answered positively only under further assumption. A standard assumption is given by the source condition (13). The source condition can be illustrated considering the eigen-system $(\sigma_j, v_j)_{j=1}^\infty$ of the operator $S^*_\rho S_\rho$. Indeed, (13) can written as

$$\|(S^*_\rho S_\rho)^{-r} f^\dagger\|^2_{\mathcal{H}} = \sum_{j=1}^\infty \frac{|\langle f^\dagger, v_j \rangle_{\mathcal{H}}|^2}{\sigma_j^{2r}} < \infty$$

which can be seen as weak form of *sparsity* on the dictionary $(v_j)_{j=1}^\infty$. The coefficients of the pseudo-solution with respect to the dictionary $(v_j)_{j=1}^\infty$ need be decreasing with respect to the eigen-values $(\sigma_j)_{j=1}^\infty$.

Indeed, the following result holds

$$\|a^\lambda\|_{\mathcal{H}} \leq \lambda^r \|(S^*_\rho S_\rho)^{-r} f^\dagger\|_{\mathcal{H}}$$

if $0 \leq r \leq 1$ and

$$\|a^\lambda\|_{\mathcal{H}} \leq \lambda \|(S^*_\rho S_\rho)^{-1} f^\dagger\|_{\mathcal{H}}$$

5.4 Sample error

Consider s^λ_n and use the explicit form of f^λ_n, f^λ to get

$$s^\lambda_n = (S^*_n S_n + \lambda I)^{-1} S^*_n \mathbf{y} - (S^*_\rho S_\rho + \lambda I)^{-1} S^*_\rho f_\rho$$

the idea is to further decompose the above expression to isolate the perturbations due to noise and random sampling. We add and subtract $(S^*_n S_n + \lambda I)^{-1} S^*_n S_n f^\dagger_{\mathcal{H}}$ so that

$$s^\lambda_n = (S^*_n S_n + \lambda I)^{-1} (S^*_n \mathbf{y} - S^*_n S_n f^\dagger_{\mathcal{H}}) + [(S^*_n S_n + \lambda I)^{-1} S^*_n S_n - (S^*_\rho S_\rho + \lambda I)^{-1} S^*_\rho S_\rho) f^\dagger_{\mathcal{H}}]$$

where we used the fact that $S^*_\rho f_\rho = S^*_\rho S_\rho f^\dagger_{\mathcal{H}}$. The study of the above expression is based on two analytic and two probabilistic inequalities. Indeed,

- using the spectral theorem and the definition of operator norm

$$\left\|(S^*_n S_n + \lambda I)^{-1}\right\| \leq \frac{1}{\lambda};$$

- a result in functional analysis allows to study Lipschitz continuity of spectral functions

$$\left\|[(S^*_n S_n + \lambda I)^{-1} S^*_n S_n - (S^*_\rho S_\rho + \lambda I)^{-1} S^*_\rho S_\rho)]\right\| \leq \frac{1}{\lambda} \left\|S^*_n S_n - S^*_\rho S_\rho\right\|$$

  where $\frac{1}{\lambda}$ is the Lipschitz constant of the function of the real valued function $(\sigma + \lambda)^{-1}\sigma$.

Then

$$\|s_n^\lambda\|_{\mathcal{H}} = \frac{1}{\lambda}\|S_n^*\mathbf{y} - S_n^*S_n f_{\mathcal{H}}^\dagger\|_{\mathcal{H}} + \frac{1}{\lambda}\big\|S_n^*S_n - S_\rho^*S_\rho\big\|\|f_{\mathcal{H}}^\dagger\|_{\mathcal{H}}$$

The study of the latter terms follows applying Höeffidng inequality(**??**) for the random variables in Hilbert space defined by:

$$\xi_i = K_{x_i} y_i$$

with valued in $\mathcal{H}$, and the random variable

$$\zeta_i = K_{x_i} \otimes K_{x_i}$$

seen as a Hilbert Schmidt operator.

5.5 Deriving the final bound

Combining the above bounds we have for some universal constant c

$$\|f_n^\lambda - f_{\mathcal{H}}^\dagger\|_{\mathcal{H}} \leq c\frac{\tau}{\lambda\sqrt{n}}(1 + \|f_{\mathcal{H}}^\dagger\|_{\mathcal{H}}) + \|f^\lambda - f_{\mathcal{H}}^\dagger\|_{\mathcal{H}},$$

which easily allows to derive Theorem **??**, using Borel-Cantelli Lemma. Further assuming a source condition we have

$$\|f_n^\lambda - f_{\mathcal{H}}^\dagger\|_{\mathcal{H}} \leq c\frac{\tau}{\lambda\sqrt{n}}(1 + \|f_{\mathcal{H}}^\dagger\|_{\mathcal{H}}) + \lambda^r\|(S_\rho^*S_\rho)^{-r} f_{\mathcal{H}}^\dagger\|_{\mathcal{H}}$$

if $0 \leq r \leq 1$.

which allows to derive Theorem **??**.

6 From Tikhonov to iterative regularization

In this section, we consider the algorithm defined by the following sequence

$$f_n^j = f_n^{j-1} - 2\frac{\eta}{n}\sum_{i=1}^n S_n^*(S_n f_n^{j-1} - \mathbf{y}), \qquad j = 1, \dots t-1 \tag{16}$$

where $f_n^0 = 0$, $\eta > 0$ and $t \in \mathbb{N}$. The above iteration can be seen to be the gradient descent iteration of the empirical error (10). It is called Landweber iteration in the context of inverse problems. Following the same reasoning as in Lemma 2 we have the following result providing a numerical realization for the above method.

Lemma 2. *For all $t \in \mathbb{N}$, let f_n^λ be defined as in* (16)*, then*

$$f_n^t(x) = \sum_{i=1}^n K(x, x_i)c_i^t, \qquad \mathbf{c}^{t+1} = \mathbf{c}^t - \frac{\eta}{n}(K_n\mathbf{c}^t - \mathbf{y}) \tag{17}$$

where $\mathbf{c}^t = (c_1^t, \dots, c_n^t)$ and $\mathbf{c}^0 = 0$

The following result allows to draw a connection to Tikhonov regularization and shed light on the regularization properties of Landweber iteration.

Lemma 3. *The iteration in* (16) *can be written as*

$$f_n^t = \eta \sum_{j=0}^{t-1} (I - \eta S_n^* S_n)^j S_n^* \mathbf{y}.$$

The proof of the above results follows from a basic induction argument. It shows that Landweber iteration can be seen as the linear operator $G_t = \sum_{j=0}^{t-1} (I - \eta S_n^* S_n)^j$ applied to $S_n^* \mathbf{y}$. If η is chosen so that

$$\left\| I - \eta S_n^* S_n \right\| < 1 \tag{18}$$

then

$$\eta \sum_{j=0}^{\infty} (I - \eta S_n^* S_n)^j = (S_n^* S_n)^{-1}$$

where $(S_n^* S_n)^{-1}$ is assumed to exists for the sake of simplicity[2] Then, if the step-size is chosen to satisfy (18), the operator corresponding to Landweber iteration can be seen as truncated series expansion. The only free parameter is the number of iterations which corresponds to the number of terms in such an expansion. It is easy to see that the condition number of the operator G_t is controlled by t, the bigger t the larger is the condition number. Indeed, the operators

$$(S_n^* S_n + \lambda I)^{-1}, \quad \eta \sum_{j=0}^{\infty} (I - \eta S_n^* S_n)^j$$

are similar and one can consider roughly a correspondence $t \sim 1/\lambda$. The number of iteration t acts as the regularization parameter for Landweber iteration.

Landweber iteration and iterative regularization. Indeed, Landweber iteration is an instance of so called iterative regularization, sometimes called early stopping regularization. The remarkable property of these class of method is that they couple computational and learning (statistical) properties. The number of iterations controls at the same time the stability, and hence the learning properties, of the solution as well the computational requirements. More computations are needed if the data can be exploited, whereas fewer computations must be considered to ensure stability when data are poor or scarce.

The above reasoning is made for precise by the following result.

Theorem 3.

The proof of the above result follows the same line of the one of Theorem (**??**).

A regularization view on optimization Another interesting aspect of the above discussion is that it provides a different perspectives on optimization methods in the context of machine learning. The classical optimization perspective would be to consider the convergence properties of the gradient iteration (16) to a minimizer of the empirical error (10). The above discussion provides an alternative point of view, by looking at gradient descent from a regularization perspectives. The iteration (16) is only an empirical

[2] More generally it can be shown that

$$\eta \sum_{j=0}^{\infty} (I - \eta S_n^* S_n)^j S_n^* = S_n^\dagger$$

iteration whereas the ideal objective is to solve (7). From this perspective, early stopping is needed to ensure a stable solution can be learned given finite data.

Following this discussion, it is natural to ask whether other optimization methods can also be analyzed within a regularization framework. This is indeed, the case as we discuss in the following. However before doing this we pro

6.1 Accelerated iterative regularization

A key problem in optimization is to find fast methods to minimize an objective function of interest. The literature on the topic is vast and here we discuss two ideas which have been considered in the context of machine learning.

Nesterov acceleration. The first is the so called Nesterov acceleration of the gradient method defining Landweber iteration. In out context it defines the following iteration

$$f_n^j = f_n^{j-1} - \eta S_n^*(S_n h_n^{j-1} - \mathbf{y}), \qquad h_n^{j-1} = f_n^{j-1} + \alpha_j(f_n^{j-1} + f_n^{j-2})$$

for $f_n^0 = f_n^{-1} = 0$ and

$$\eta \leq 1$$
$$\alpha_j = \frac{j-1}{j+\beta}, \quad \beta \geq 1.$$

The ν-method This method is also known as Chebychev method is given $\nu > 0$ by

$$f_n^j = f_n^{j-1} - \omega_t S_n^*(S_n h_n^{j-1} - \mathbf{y}) + \alpha_j(f_n^{j-1} + f_n^{j-2})$$

for $f_n^0 = f_n^{-1} = 0$ and $\omega_1 =$

$$\eta_j = 4\frac{(2j+2\nu-1)(j+\nu-1)}{(j+2\nu-1)(2j+4\nu-1)},$$
$$\alpha_j = \frac{(j-1)(2j-3)(2j+2\nu-1)}{(j+2\nu-1)(2j+4\nu-1)(2j+2\nu-3)},$$

Remark 8 (Numerical realization)**.** *The numerical realization of the above methods can be derived analogously to Tikhonov regularization and Landwber iteration.*

Error bounds The proof of the corresponding error bounds can also be proved following similar arguments, to obtain

$$\|f_n^\lambda - f_\mathcal{H}^\dagger\|_\mathcal{H} \leq c\frac{\tau t^2}{\sqrt{n}} + c'\frac{1}{t^{2r}}$$

if $0 \leq r \leq r_*$, and where $r_* = 1/2$ for Nesterov acceleration, and $r_* = \nu - 1/2$ for the ν-method.

The remarkable properties of the above method is that they yield again the same optimal bound, but now the regularization parameter is t^2, so that a more aggressive stopping rule

$$t_n = \sqrt{n^{\frac{1}{2(r+1)}}}$$

is allowed!

6.2 Incremental and stochastic Iterative Regularization

Here we consider incremental optimization techniques defined by the following iteration

$$f_n^j = f_n^{j-1} - \eta_t K_{x_{p(j)}}(f_n^{j-1}(x_{p(j)}) - y_{p(j)}).$$

Compared to Landweber iteration only a pair of input-output is used to compute a point-wise gradient in each iteration.

Remark 9 (Numerical realization)**.** *The numerical realization of the above methods can be derived analogously to Tikhonov regularization and Landwber iteration.*

Error bounds The proof of the corresponding error bounds is more complex than in the above cases and some care is needed. However, the final bound can be shown to be the same as Landweber iteration, suggesting that there is no gain in considering incremental techniques!

A Basic mathematical facts

A.1 Basic functional analysis

Let $A : \mathcal{H} \to \mathcal{G}$

- Cauchy-Schwartz inequality

$$\forall f, f' \in \mathcal{H} \qquad \left\langle f, f' \right\rangle_{\mathcal{H}} \leq \|f\|_{\mathcal{H}} \|f'\|_{\mathcal{H}}$$

- Operator norm

$$\left\|A\right\| = \sup_{f \in \mathcal{H}} \frac{\left\|Af\right\|}{\|f\|_{\mathcal{H}}}$$

- Hilbert Schmidt operator

$$\left\|A\right\|_2 = \mathrm{Trace}(A^*A) < \infty$$

  recalling that $\mathrm{Trace}(A) = \sum_j \sigma_j^{1/2} = \sum_j \left\langle (A^*A)^{1/2} e_j, e_j \right\rangle$. The norm $\left\|A\right\|_2$ is called Hilbert-Schmidt or Frobenius norm.

- Trace class operator

$$\left\|A\right\|_1 = \mathrm{Trace}(A) < \infty$$

A.2 Singular system

Recall that if $A : \mathcal{H} \to \mathcal{G}$ is a linear, compact operator than there is a corresponding singular system $(\sigma_j; u_j, v_j)$ such that for all j

$$A^*Av_j = \sigma_j v_j, \quad AA^*u_j = \sigma_j u_i$$

and

$$Av_j = \sigma_j^{1/2} u_j, \quad A^*u_j = \sigma_j^{1/2} v_i.$$

The singular values $\sigma_j^{1/2}$ can be ordered in a decreasing fashion and have an accumulation point at zero. The singular vectors $(v_j)_{j=1}^{\infty}$ and $(u_j)_{j=1}^{\infty}$ provide orthonormal basis for $\mathcal{H}$ and $\mathcal{G}$ respectively.

A.3 Spectral theorem

The spectral theorem provides an explicit expression of A in terms of it singular systems, indeed

$$A = \sum_j \sigma^{1/2} u_j \otimes v_j$$

where $u_j \otimes v_j : \mathcal{H} \to \mathcal{G}$ is the rank one operator defined by $u_j \otimes v_j f = \langle v_j, f \rangle_{\mathcal{H}} u_j$.

B Exercises

Risks and norms

- Prove that for all $f \in L^2(\mathcal{X}, \rho_{\mathcal{X}})$, $\|f - f_\rho\|_\rho^2 = \mathcal{E}(f) - \mathcal{E}(f_\rho)$.
- Prove that for all $f \in \mathcal{H}$, $\|f\|_\rho = \|(S_\rho^* S_\rho)^{1/2} f\|_{\mathcal{H}}$.
- Prove that for all $f \in L^2(\mathcal{X}, \rho_{\mathcal{X}})$, $R(f) - R(b_\rho) \leq \|f - f_\rho\|_\rho$.

Kernel operators

- Derive the explicit form of the operators $S_\rho, S_\rho^*, S_\rho^* S_\rho, S_\rho S_\rho^*$ and their empirical counter parts.
- Derive the explicit form of this operators in the the case of the linear kernel.
- Compute $\mathrm{Tr}(S_\rho^* S_\rho)$ and $\|S_\rho^* S_\rho\|_2$.

Operator estimates and related inequalities

- Let A be a positive symmetric matrix. Estimate $\|(A + \lambda I)^{-1}\|$.
- Compute the Lipschitz constant of the function $\sigma \mapsto \sigma/(\sigma + \lambda)$.
- Compute the maximum of the function $\sigma \mapsto \lambda \sigma^r/(\sigma + \lambda)$.
- Let A be a positive symmetric matrix bounded by 1. Estimate $\|\sum_{j=0}^{t-1} (I - A)^j\|$
- Compute the Lipschitz constant of the function $\sigma \mapsto \sigma \sum_{j=0}^{t-1} (I - \sigma)^j$.
- Compute the maximum of the function $\sigma \mapsto (1 - \sigma)^t \sigma^r$.

Concentration inequalties Recall the following inequality for random vectors. If $z_1, \ldots, z_n$ are i.i.d. random vectors in $\mathbb{R}^d$ bounded by 1 and with mean μ, with probability at least $1 - e^{-\tau}$, $\tau > 1$,

$$\|\frac{1}{n} \sum_{i=1}^n z_i - \mu\| \leq \frac{c\tau}{\sqrt{n}}, \tag{19}$$

where c is a small numerical constants. For $w_* \in \mathbb{R}^d$, and let $i = 1, \ldots, n$, let $y_i = w_*^\top x_i + \epsilon_i$, where x_i are i.i.d. random vectors with norm bounded by 1, and ϵ_i i.i.d. zero mean random numbers smaller than 1. Let X_n be the n by d matrix with rows $x_1, \ldots, x_n$ and $\mathbf{y} = (y_1, \ldots, y_n)$. Use (19) to estimate

- $\|\frac{X_n^\top \mathbf{y}}{n} - \frac{X_n^\top X_n w_*}{n}\|$, and
- $\|\frac{X_n^\top X_n}{n} - \mathbb{E}[\frac{X_n^\top X_n}{n}]\|_2$ where $\|\cdot\|_2$ is the Frobenius norm.