University of Insubria

DIPARTIMENTO DI SCIENZA E ALTA TECNOLOGIA
PhD in Computer Science and Mathematics of Calculus

Exam of Statistical Learning Theory and Applications

A Mathematical Panorama of One-Layer Architectures Learning Algorithms

Teachers: *S. Villa* and *L. Rosasco* PhD Student: *M. Tarsia*[*] Tutor: *D. Cassani* Advisor: *E. Mastrogiacomo*

Academic Year 2018−2019
14/09/2019

Introduction and Summary

The purpose of *learning*, or *supervised learning*, is to understand intelligence, that means to understand how the human brain works in order to build intelligent machines which learn from experience and improve.

The intrinsic issue of this field is in fact dealing with possibly many data and, consequently, the mathematics of learning theory originates from the idea to learn from a finite set of input-output pairs to perform a very specific task: that is, *training* and *learning from examples*.

For instance, one could have input points equal to multi-dimensional points of characterizing properties and output points equal to or -1 or $+1$ meaning respectively or bad or good according to some conventions.

Training sets. Let be $m \in \mathbb{N} \setminus \{\,0\,\}$ a cardinality or size for such a set, $d \in \mathbb{N} \setminus \{\,0\,\}$ a dimension for input points, $X \subseteq \mathbb{R}^d_{\boldsymbol{x}}$ an *input space* which is supposed to be a closed set, and $Y \subseteq \mathbb{R}_y$ an *output space* which can be imagined as $Y \equiv \mathbb{R}_y$. Then a *training set*, or *set of training points*, is a set $\mathrm{S}^m \equiv \mathrm{S}^m_{d,X,Y}$ of m points $\big\{(\boldsymbol{x}^i, y^i)\big\}_{i=1}^m$ in $X \times Y \subseteq \mathbb{R}^{d+1}_{(\boldsymbol{x},y)}$ called in fact *training points*:

$$\mathrm{S}^m \doteq \big\{(\boldsymbol{x}^i, y^i)\big\}_{i=1}^m.$$

The objective. *Training* with the following meaning: finding an optimal functional relationship between inputs and outputs or, more precisely, synthesizing an *empirical optimum*

$$\widehat{f} \equiv \widehat{f}_{\mathrm{S}^m} \colon X \to Y$$

which belongs to an appropriately structured *hypothesis space* $\mathcal{H}$ of functions and which best represents the more likely relation between the input points $\boldsymbol{x}^i$ and the corresponding output points y^i for each $i = 1, \ldots, m$.

The central question must becomes how and how much we could learn from a training set S^m or, equivalently, how much such a $\widehat{f}$ would be predictive which means well estimating outputs for previously unseen inputs.

The main key issue will be the tradeoff or conflict between the number m of training points and the complexity of the hypothesis space $\mathcal{H}$. The learning techniques to deal with all that are based on approximation theory, functional analysis and probability theory as well.

Heuristic idea. Nothing in known except the training set S^m and therefore, to achieve an empirical optimum $\widehat{f} \equiv \widehat{f}_{\mathrm{S}^m}$ within a suitable hypothesis space $\mathcal{H}$, a very natural approach should result to look for a functional combination of input-centered regular kernels $K(\boldsymbol{x}^i, \cdot\,) \equiv K_{\boldsymbol{x}^i}(\,\cdot\,)$ on X, preferably both finite and linear. Here the corresponding concept of optimality must involve all the training points through a reasonable choice of the error of the training set called *empirical risk* or *error* $\widehat{\mathcal{E}}(\,\cdot\,) \equiv \widehat{\mathcal{E}}_{\mathrm{S}^m}(\,\cdot\,)$ to minimize on $\mathcal{H}$, and all this in such a way that everything leads to a concrete and quite simply calculable well-posed formula.

[*] E-mail: `mtarsia1@uninsubria.it`. Web page: `https://www.uninsubria.it/hpp/marco.tarsia`.

In order to get $\widehat{f}$ as such a special linear combination, $\mathcal{H}$ will be built as a real Hilbert space of continuous functions from X to Y deeply related to K, namely the corresponding reproducing kernel Hilbert space or RKHS $(\mathcal{H}, \langle \cdot ; \cdot \rangle_{\mathcal{H}}) \equiv (\mathcal{H}_K, \langle \cdot ; \cdot \rangle_K)$ according to the Moore-Aronszajn theorem. Among other things, indeed, one could count also on the reproducing formula $f(\cdot) = \big\langle f ; K_{(\cdot)} \big\rangle_K$ valid on X for each $f \in \mathcal{H}_K$ from which immediately the intuition to select as empirical risk $\widehat{\mathcal{E}}$ to minimize on $\mathcal{H}_K$ a regularized quadratic-type functional to be able then to differentiate in the classical sense of Fréchet and conclude easily while in addition, at the same time, parameters will be modulated so that even computationally everything would be great.

Lastly, since the training points are of course intrinsecally affected by noises, the framework will be supposed to be dominated by an unknown underlying probability distribution $\rho \equiv \rho^m_{d,X,Y}$ which brings in a canonical way to the true *target function* $f_\rho \colon X \to Y$ and therefore to a technical notion of *true optimum* $f_{\widetilde{\mathcal{H}}_K} \colon X \to Y$ to introduce finally the so-called conditions of *consistency* for learning algorithms: these determine when the empirical minimizer is a rather good approximation, at least in probability as $m \uparrow +\infty$, of this true optimum.

Learning algorithms.

1. Fix a training set $\mathrm{S}^m \equiv \big\{(\boldsymbol{x}^i, y^i)\big\}_{i=1}^m$ and choose a not trivial Mercer's kernel $K(\cdot\,,\cdot)$ on X.

2. Place $\widehat{K} \equiv \big[K(\boldsymbol{x}^i, \boldsymbol{x}^j)\big]_{i,j=1}^m \in \mathbb{R}^{m\times m}$, $\boldsymbol{y} \equiv (y^1, \dots, y^m) \in \mathbb{R}^m$, $I \equiv I_m \in \mathbb{R}^{m\times m}$ the $m \times m$ identity matrix and, for each $\gamma \in \,]0, +\infty[$, compute the solution $\boldsymbol{c} \equiv \boldsymbol{c}_{\mathrm{S}^m,K,\gamma} \equiv (c^1, \dots, c^m) \in \mathbb{R}^m$ of the $m \times m$ linear system

$$\big(m\gamma I + \widehat{K}\big)\boldsymbol{c} = \boldsymbol{y}.$$

3. Define $\widehat{f} \colon X \to Y$ through the following formula: for each $\boldsymbol{x} \in X$,

$$\widehat{f}(\boldsymbol{x}) \overset{def}{=} \sum_{i=1}^m c^i K(\boldsymbol{x}^i, \boldsymbol{x}).$$

Some remarks.

- The square linear system above is always well-posed because, for each data and parameter in play, $m\gamma I + \widehat{K}$ is a symmetric and strictly positive-definite real matrix and moreover, if $m\gamma$ is taken large enough, then its $\|\cdot\|_2$-condition number becomes good enough: $\lambda_{\max}\big(m\gamma I + \widehat{K}\big)/\lambda_{\min}\big(m\gamma I + \widehat{K}\big) \to 1$ as $m\gamma \to +\infty$.

  The algorithms for solving it efficently constitute one of the most developed areas in numerical analysis.
- Data points are sufficient to describe the empirical optimum and, for any $i = 1, \dots, m$, $y^i = \widehat{f}(\boldsymbol{x}^i) + m\gamma c^i$.
- If there exists $\sigma \in \,]0, +\infty[$ such for which $K \equiv K_\sigma$ is a Gaussian kernel on X, then the learning algorithms approximate $\widehat{f}$ by a weighted superposition of Gaussian blobs centered at the location of one of the m input data and, of course, each weight is such as to minimize the designated regularized empirical error.

  The parameter σ, together with γ, control its degree of smoothing, of noise tolerance and of generalization. For instance, as $\sigma \downarrow 0$, $\widehat{f}$ becomes as a look-up table which cannot generalize because, for each $(\boldsymbol{x}, y) \in X \times Y$, it provides $y = y^i$ only when $\boldsymbol{x} = \boldsymbol{x}^i$, otherwise zero, for each $i = 1, \dots, m$.
- The learning algorithms perform really well in a multitude of applications involving regression as much as *binary classification*: when, for each $i = 1, \dots, m$, $y^i \in \{-1, +1\}$ and thus the predicted label is $\{-1, +1\}$ itself depending on the sign of $\widehat{f}$. Regression ones are surely the oldest and they typically involve fitting data in a small number of dimensions and not as happens for computer, graphics, robotics science, etc. Binary classification employment instead simply abounds in every field.

  Particularly interesting are financial applications too as, for instance, the estimation of the price of derivative sicurities such as stock options. In this case the algorithms replace the classical Black-Scholes equation by learning from historical data the map from an input space - volatility, underlying stock price, time to expiration of the option, ... - to the output space - the price of the option.

▶ About the whole continuation, we assume that objects as those we've discussed above are fixed once for all: a size m, a dimension d, a compact input space $X \subset \mathbb{R}^d_{\boldsymbol{x}}$, an output space Y, a training set $\mathrm{S}^m \equiv \big\{(\boldsymbol{x}^i, y^i)\big\}_{i=1}^m$, a C^∞ Mercer's kernel K on X and its RKHS $(\mathcal{H}_K, \langle \cdot ; \cdot \rangle_K)$ equipped thus also with $\|\cdot\|_K \equiv \langle \cdot ; \cdot \rangle_K^{1/2}$.

Optimality and Consistency

Empirical risks. For each $\gamma \in \,]0,+\infty[$, an *empirical risk* or *error* is a loss or price-to-pay functional $\widehat{\mathcal{E}} \equiv \widehat{\mathcal{E}}_{\mathrm{S}^m,K,\gamma}\colon \mathcal{H}_K \to [0,+\infty[$ defined as a Tikhonov regularized mean squared sum: for each $f \in \mathcal{H}_K$,

$$\widehat{\mathcal{E}}(f) \doteq \frac{1}{m}\sum_{i=1}^{m}\Big(f(\boldsymbol{x}^i) - y^i\Big)^2 + \gamma\,\|f\|_K^2.$$

Theorem (Empirical risk minimization (ERM))**.** *Let $\widehat{\mathcal{E}} \equiv \widehat{\mathcal{E}}_{\mathrm{S}^m,K,\gamma}$ be an empirical risk. Then there exists only one function $\widehat{f} \equiv \widehat{f}_{\mathrm{S}^m,K,\gamma} \in \mathcal{H}_K$ which is an empirical optimum in $\mathcal{H}_K$ with respect to $\widehat{\mathcal{E}}$ meaning that*

$$\widehat{\mathcal{E}}(\widehat{f}) = \inf_{f\in\mathcal{H}_K}\widehat{\mathcal{E}}(f) \equiv \min_{f\in\mathcal{H}_K}\widehat{\mathcal{E}}(f)$$

and moreover $\widehat{f}$ is uniquely determined by the following equation: for each $\boldsymbol{x} \in X$,

$$\widehat{f}(\boldsymbol{x}) = \sum_{i=1}^{m}\frac{y^i - \widehat{f}(\boldsymbol{x}^i)}{m\gamma}K(\boldsymbol{x}^i,\boldsymbol{x}).$$

Proof. Compute the classical Fréchet derivative of $\widehat{\mathcal{E}}(\,\cdot\,)$ on $\mathcal{H}_K$, applying it to a generic element $\bar{f} \in \mathcal{H}_K$ and setting all that equal to zero obtaining that, for each $\widehat{f} \in \mathcal{H}_K$ minimizer for $\widehat{\mathcal{E}}$,

$$\frac{1}{m}\sum_{i=1}^{m}\Big(\widehat{f}(\boldsymbol{x}^i) - y^i\Big)\bar{f}(\boldsymbol{x}^i) + \gamma\left\langle \widehat{f}\,;\bar{f}\right\rangle_K = 0$$

and now, for each $\boldsymbol{x} \in X$, choose $\bar{f} \equiv K_{\boldsymbol{x}}$ and use the reproducing formula $\widehat{f}(\boldsymbol{x}) = \left\langle \widehat{f}\,; K_{\boldsymbol{x}}\right\rangle_K$. □

Remark. A quite similar derivation holds for an arbitrary non-quadratic loss functional block $V(y, f(\boldsymbol{x}))$ instead of $(f(\boldsymbol{x}) - y)^2$, $(\boldsymbol{x}, y) \in X \times Y$, which yields in fact the same type of identity for $\widehat{f}$ while however the equation for the weights $c^1,\dots,c^m$ becomes in general non-linear depending on the exact form of V: for instance, it could happen that each c^i must be found by solving a quadratic programming problem as it happens in support-vectors machines, or SVM, which are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis.

Probabilistic setting. We agree that there exists an unknown probability measure $\rho \equiv \rho^m_{d,X,Y}$ defined on the class $\mathcal{B}(X\times Y)$ of the Borel sets in the product space $X \times Y \subset \mathbb{R}^{d+1}_{(\boldsymbol{x},y)}$ from which the data $\mathrm{S}^m \equiv \big\{(\boldsymbol{x}^i, y^i)\big\}_{i=1}^m$ is randomly drawn in such a way that $(X\times Y, \mathcal{B}(X\times Y), \rho)$ becomes a probability space and

$$(\boldsymbol{x}^i, y^i)_{i=1}^m \overset{\mathcal{L}}{=} \rho^{\otimes m}$$

with respect to a suitable starting probability space. We assume also that $\rho(\{\,\cdot\,\}\times Y) > 0$ holds on X and thus we introduce other fundamental probability measures which are canonically induced by ρ.

$\boldsymbol{\rho_{\boldsymbol{x}}}$**.** For each $\boldsymbol{x} \in X$, the *conditional* probability measure on $\{\boldsymbol{x}\}\times Y \subset X \times Y$: for any $B \subseteq Y$ Borel set,

$$\rho(B|\boldsymbol{x}) \coloneqq \rho_{\boldsymbol{x}}(\{\boldsymbol{x}\}\times B) \doteq \frac{\rho(\{\boldsymbol{x}\}\times B)}{\rho(\{\boldsymbol{x}\}\times Y)}.$$

$\boldsymbol{\rho_X}$**.** The *marginal* probability measure on X: for any $A \subseteq X$ Borel set, $\rho_X(A) \doteq \rho(A\times Y)$.

Target function. The *target function* $f_\rho\colon X \to Y$ is defined through $\rho_{\boldsymbol{x}}$ by placing, for each $\boldsymbol{x} \in X$,

$$f_\rho(\boldsymbol{x}) \doteq \int_{\{\boldsymbol{x}\}\times Y} y\,d\rho_{\boldsymbol{x}}(y) \equiv \int_Y y\,d\rho(y|\boldsymbol{x})$$

where naturally, for any $y \in Y$, $\rho(y|\boldsymbol{x}) \coloneqq \rho(\{y\}|\boldsymbol{x}) \equiv \rho(\{(\boldsymbol{x},y)\})/\rho(\{\boldsymbol{x}\}\times Y)$.

That f_ρ could be certainly regarded as the true input-output function reflecting our stochastic environment.

Hypothesis spaces. A convex space $\mathcal{H}$ of continuous functions from X to Y which, as a subset of $C(X;Y)$, is also compact with respect to the topology induced by the uniform norm $\|f\|_{\infty} \equiv \sup_{\boldsymbol{x}\in X}|f(\boldsymbol{x})|$, $f\in C(X;Y)$.

Let's dip into the following standard situation considering the usual inclusion operator $i_K\colon \mathcal{H}_K \hookrightarrow C(X;Y)$, which is compact as linear operator since K is supposed to be C^{∞} on X, and, for any arbitrary $R\in\,]0,+\infty[$, the closed ball $B_R \doteq \{\, g\in\mathcal{H}_K \mid \|g\|_K \le R \,\}$ in $\mathcal{H}_K$: then we could take $\mathcal{H}\equiv\widetilde{\mathcal{H}}_K$ as the $\|\cdot\|_{\infty}$-clousure

$$\widetilde{\mathcal{H}}_K \doteq \overline{i_K(B_R)}$$

(for each $f\in C(X;Y)$, $f\in\mathcal{H}\equiv\widetilde{\mathcal{H}}_K$ if and only if there exists an uniformly bounded sequence $(g_n)_n$ in $\mathcal{H}_K$ such for which $g_n\to f$ with respect to $\|\cdot\|_{\infty}$ as $n\to+\infty$).

Note. A way to build a C^{∞} Mercer's kernel on X: choose a completely monotonic function $\psi\colon\,]0,+\infty[\,\to\mathbb{R}_+$ (ψ is C^{∞} with $(-1)^n\psi^{(n)}\ge 0$ for any $n\in\mathbb{N}$) and thus define, for each $\boldsymbol{x},\boldsymbol{x'}\in X$, $K_{\psi}(\boldsymbol{x},\boldsymbol{x'})\doteq\psi(\|\boldsymbol{x}-\boldsymbol{x'}\|_2^2)$.

Expected risks. For each $R\in\,]0,+\infty[$, an *expected risk* or *error* is a mean squared integral loss functional $\mathcal{E}_{\rho}\equiv\mathcal{E}_{\rho,K,R}\colon\widetilde{\mathcal{H}}_K\to[0,+\infty[$ defined through f_{ρ} and ρ_X by placing, for each $f\in\widetilde{\mathcal{H}}_K$,

$$\mathcal{E}_{\rho}(f)\doteq\int_X\Big(f(\boldsymbol{x})-f_{\rho}(\boldsymbol{x})\Big)^2 d\rho_X(\boldsymbol{x}).$$

Remark. If we've to deal with a loss function $\ell\colon Y\times\mathbb{R}_a\to[0,+\infty[$ as it is the square loss $(y,a)\mapsto(a-y)^2$, $(y,a)\in Y\times\mathbb{R}$, and if we assume that ρ can be decomposed in such a way that worths

$$\int_{X\times Y}\ell(y,\star(\boldsymbol{x}))\,d\rho(\boldsymbol{x},y)=\int_X\left[\int_Y\ell(y,\star(\boldsymbol{x}))\,d\rho(y|\boldsymbol{x})\right]d\rho_X(\boldsymbol{x})$$

then the problem of minimizing this sort of expected risk becomes to minimize simply that functional inside the integral over X and thus, in the case of the square loss, it leads exactly to the mean $f_{\rho}(\,\cdot\,)$ of $\rho_{(\,\cdot\,)}$ on X.

Empirical consistency. The result below could be shown without particular difficulties.

Theorem (Empirical consistency)**.** *Let's fix $R\in\,]0,+\infty[$. Then the three following conditions hold.*

1. *There exists only one function $\widetilde{f}_{\mathrm{S}^m}\equiv\widetilde{f}_{\mathrm{S}^m,K,R}\in\widetilde{\mathcal{H}}_K$ such that*

$$\widetilde{f}_{\mathrm{S}^m}\in\operatorname*{arg\,min}_{f\in\widetilde{\mathcal{H}}_K}\left\{\frac{1}{m}\sum_{i=1}^m\Big(f(\boldsymbol{x}^i)-y^i\Big)^2\right\}.$$

2. *Let $\mathcal{E}_{\rho}\equiv\mathcal{E}_{\rho,K,R}$ be an expected risk. Then there exists only one function $f_{\widetilde{\mathcal{H}}_K}\equiv f_{\rho,K,R}\in\widetilde{\mathcal{H}}_K$ which is the* true optimum *in $\widetilde{\mathcal{H}}_K$ with respect to $\mathcal{E}_{\rho}$ meaning that*

$$\mathcal{E}_{\rho}(f_{\widetilde{\mathcal{H}}_K})=\inf_{f\in\widetilde{\mathcal{H}}_K}\mathcal{E}_{\rho}(f)\equiv\min_{f\in\widetilde{\mathcal{H}}_K}\mathcal{E}_{\rho}(f).$$

This number $\mathcal{A}_{\widetilde{\mathcal{H}}_K}\equiv\mathcal{A}_{\rho,K,R}\doteq\mathcal{E}_{\rho}(f_{\widetilde{\mathcal{H}}_K})\in[0,+\infty[$ is the approximation error *(corresponding to ρ and $\widetilde{\mathcal{H}}_K$). The non-negative random variable $\mathcal{S}_{\mathrm{S}^m}\equiv\mathcal{S}_{\mathrm{S}^m,\rho,K,R}\doteq\mathcal{E}_{\rho}(\widetilde{f}_{\mathrm{S}^m})-\mathcal{A}_{\widetilde{\mathcal{H}}_K}$ is the* sample error *corresponding to* S^m*.*

3. *Let $M\equiv M_{X,Y,\widetilde{\mathcal{H}}_K}\in\,]0,+\infty[$ be such that, for any $f\in\widetilde{\mathcal{H}}_K$ and for almost all $(\boldsymbol{x},y)\in X\times Y$, $|f(\boldsymbol{x})-y|\le M$ while, for any $\eta\in\,]0,+\infty[$, let $\operatorname{cov\#}(\eta)\equiv\operatorname{cov\#}(\widetilde{\mathcal{H}}_K,\eta)\in\mathbb{N}\setminus\{\,0\,\}$ denote the minimum number of open balls in $\widetilde{\mathcal{H}}_K$ of radius η necessary to cover $\widetilde{\mathcal{H}}_K$. Then $\mathcal{S}_{\mathrm{S}^m}\to 0$ in probability as $m\to+\infty$ because, more precisely,*

$$\rho^{\otimes m}\Big(\Big\{\,\mathrm{S}^m\cong(\boldsymbol{x}^i,y^i)_{i=1}^m\in(X\times Y)^m\;\Big|\;\mathcal{S}_{\mathrm{S}^m}\le\varepsilon\,\Big\}\Big)\ge 1-\operatorname{cov\#}(\varepsilon/24M)\exp(-m\varepsilon/288M^2).$$

Remark. There's also a good amount of studies concerning estimates for those covering numbers. Moreover, the writing $\mathcal{E}_{\rho}(\widetilde{f}_{\mathrm{S}^m})=\mathcal{A}_{\widetilde{\mathcal{H}}_K}+\mathcal{S}_{\mathrm{S}^m}$ is indirectly related to the well-known bias-variance decomposition.

Approximation error. Let's symbolize $L^2_{\rho_X} \equiv L^2_{\rho_X}(X;Y)$ and let's consider the Hilbert-Schmidt integral operator $T_K\colon L^2_{\rho_X} \to C(X;Y) < L^2_{\rho_X}$ corresponding to K defined, for any $f \in L^2_{\rho_X}$ and $\boldsymbol{x} \in X$, by the formula

$$(T_K f)(\boldsymbol{x}) \doteq \int_X f(\boldsymbol{x}')K(\boldsymbol{x},\boldsymbol{x}')\,d\rho_X(\boldsymbol{x}').$$

Observe that T_K is strictly positive meaning self-adjoint - for each $f,g \in L^2_{\rho_X}$, $\langle T_K f\,;g\rangle_{L^2_{\rho_X}} = \langle f\,;T_K g\rangle_{L^2_{\rho_X}}$ - and such that, for any $f \in L^2_{\rho_X} \setminus \{\,0\,\}$, $\langle T_K f\,;f\rangle_{L^2_{\rho_X}} > 0$, and also compact as operator: so let $S_K \coloneqq T_K^{1/2}$ be that unique strictly positive and compact operator $S_K\colon L^2_{\rho_X} \to C(X;Y)$ satisfying $S_K^2 \equiv S_K \circ S_K = T_K$. Then it could be shown that $S_K(L^2_{\rho_X}) = \mathcal{H}_K$ and rather that S_K is a Hilbert isomorphism between $L^2_{\rho_X}$ and $\mathcal{H}_K$ (ρ_X-free!) and thus, for any $s \in \,]0,+\infty[$, $S_K^{-s}\colon L^2_{\rho_X} \to \mathcal{H}_K$ results a Hilbert isomorphism as well with

$$\mathcal{G}_K \doteq \left\{\, g \in L^2_{\rho_X} \;\middle|\; \|S_K^{-s} g\|_{L^2_{\rho_X}} < +\infty \,\right\} = \mathcal{H}_K$$

which is canonically equipped with the inner product given, for any $g,h \in \mathcal{G}_K$, by $\langle g\,;h\rangle_{\mathcal{G}_K} \doteq \left\langle S_K^{-s} g\,;S_K^{-s} h\right\rangle_{L^2_{\rho_X}}$.

Remark. That $L^2_{\rho_X}$-norm above must be interpreted by the meaning of the classical spectral theorem for strictly positive and compact operators, which in fact we mention below.

Theorem (Spectral theorem for compact operators). *Let S be a compact operator on an infinite dimensional and separable real Hilbert space $\mathcal{H}$. Then there exists a complete orthonormal system or Hilbert basis $\{\varphi_n\}_{n\in\mathbb{N}}$ in $\mathcal{H}$ consisting of all the eigenvectors of S. More precisely, for each $n \in \mathbb{N}$, let $\lambda_n \in \mathbb{C}$ be the eigenvalue of S corresponding to the eigenvector φ_n of S. Then the four following conditions hold.*

a. *Either $\#\{\lambda_n\}_{n\in\mathbb{N}} < +\infty$, or $\lambda_n \to 0$ as $n \to +\infty$.*

b. *It's worth that* $\sup_{n\in\mathbb{N}}|\lambda_n| \equiv \max_{n\in\mathbb{N}}|\lambda_n| = \|S\| \equiv \sup_{\|h\|_{\mathcal{H}}=1}\|S(h)\|_{\mathcal{H}}$.

c. *If S is self-adjoint then, for any $n \in \mathbb{N}$, $\lambda_n \in \mathbb{R}$.*

d. *If S is positive (respectively strictly positive) then, for any $n \in \mathbb{N}$, $\lambda_n \geq 0$ (respectively $\lambda_n > 0$).*

So, if S is a positive and compact operator on such a $\mathcal{H}$, then is defined, for any $\tau \in [0,+\infty[$ and $(a_n)_{n\in\mathbb{N}} \subset \mathbb{R}$,

$$S^\tau\Bigl(\sum_{n=0}^{\infty} a_n\varphi_n\Bigr) \doteq \sum_{n=0}^{\infty} {\lambda_n}^\tau a_n\varphi_n$$

and thus, if S is strictly positive, the similar formula even for any $\tau \in \,]-\infty,0[$ but regarding only the subspace $\left\{\, \sum_{n=0}^{\infty} a_n\varphi_n \in \mathcal{H} \;\middle|\; \sum_{n=0}^{\infty}({\lambda_n}^\tau a_n)^2 < +\infty \,\right\}$ of $\mathcal{H}$, placing finally $\|S^\tau\star\|_{\mathcal{H}} \equiv +\infty$ outside of it.
The following estimate could be shown.

Theorem (Approximation error). *For any $R, s \in \,]0,+\infty[$ and $r \in \,]0,s[$,*

$$\mathcal{A}_{\widetilde{\mathcal{H}}_K} \equiv \left\| f_{\widetilde{\mathcal{H}}_K} - f_\rho \right\|^2_{L^2_{\rho_X}} \leq R^{-2r/(s-r)}\left\|S_K^{-r} f_\rho\right\|^{2s/(s-r)}_{L^2_{\rho_X}}.$$

Remark. Something can be said also about sample and approximation errors for the regularization algorithms.

A Bayesian interpretation. The data term can be seen as a model of Gaussian additive noise with the RKHS term as a prior probability on the set of parameters $\mathcal{H}_K$. In fact, for any $f \in \mathcal{H}_K$, let's set:

- $\mathbf{P}[f]$ as a prior probability of the random field f;
- $\mathbf{P}[\mathrm{S}^m|f]$ as a conditional probability of S^m given f or, namely, a likelihood of the model;
- $\mathbf{P}[f|\mathrm{S}^m]$ as the consequent conditional probability of f given S^m or, namely, the posterior probability.

Then the classical Bayes' theorem of inversing probability gives us that, for any $f \in \mathcal{H}_K$,

$$\mathbf{P}[f\,|\mathrm{S}^m] = \frac{\mathbf{P}[\mathrm{S}^m|f]\,\mathbf{P}[f]}{\mathbf{P}[\mathrm{S}^m]} \propto \mathbf{P}[\mathrm{S}^m|f]\,\mathbf{P}[f]$$

and so, if we choose $\mathbf{P}[f] \propto \exp(-\|f\|_K^2)$ and we take a normally distributed noise with standard deviation $\sigma \in \,]0,+\infty[$ meaning precisely that, for any $i = 1,\ldots,m$, $y^i \sim \mathcal{N}(f(\boldsymbol{x}^i),\sigma^2)$ with $y^1,\ldots,y^m$ independent, being thus in a situation of homoscedasticity, then finally

$$\mathbf{P}[f\,|\mathrm{S}^m] \propto \exp\biggl(-\frac{1}{2\sigma^2}\sum_{i=1}^{m}\Bigl(f(\boldsymbol{x}^i)-y^i\Bigr)^2 - \|f\|_K^2\biggr).$$

Now one of the most natural estimates from that posterior probability is the well-known maximum a posteriori estimate or MAP, taking indeed the most probable solution $\widehat{f}_{\mathrm{MAP}} \in \mathcal{H}_K$ given the prior and the data:

$$\widehat{f}_{\mathrm{MAP}} \in \operatorname*{arg\,max}_{f\in\mathcal{H}_K} \mathbf{P}[f\,|\mathrm{S}^m] \equiv \operatorname*{arg\,min}_{f\in\mathcal{H}_K}\Biggl\{\sum_{i=1}^{m}\Bigl(f(\boldsymbol{x}^i)-y^i\Bigr)^2 + 2\sigma^2\|f\|_K^2\Biggr\}$$

which leads in fact to the same $\widehat{f}$ as the regularization functional in the case of $2\sigma^2 \equiv m\gamma$.

Condition for learnability. A necessary and sufficient condition on a hypothesis space $\mathcal{H}$ which guarantee the empirical consistency is that $\mathcal{H}$ results an uniform Glivenko-Cantelli class of functions, in which case no specific topology is assumed for it: that is, a class $\mathcal{H}$ of functions $f\colon X \to Y$ such that, for any $\varepsilon > 0$,

$$\lim_{m\to+\infty}\sup_{\bar\rho_X\in\mathscr{P}_X}\bar\rho_X^{\otimes m}\Biggl(\Biggl\{\,\boldsymbol{x}^{\otimes m} := (\boldsymbol{x}^1,\ldots,\boldsymbol{x}^m)\in X^m \;\Bigg|\; \biggl|\int_X f(\boldsymbol{x})\,d\widehat{\rho}_{\boldsymbol{x}^{\otimes m}}(\boldsymbol{x}) - \int_X f(\boldsymbol{x})\,d\bar\rho_X(\boldsymbol{x})\biggr| > \varepsilon\,\Biggr\}\Biggr) = 0$$

where $\mathscr{P}_X$ is the family of every Borel probability measures $\bar\rho_X$ on X and where, for each $m \in \mathbb{N}\setminus\{\,0\,\}$ and $\boldsymbol{x}^{\otimes m} := (\boldsymbol{x}^1,\ldots,\boldsymbol{x}^m) \in X^m$, $\widehat{\rho}_{\boldsymbol{x}^{\otimes m}}$ is the empirical probability measure on X supported on the finite set $\bigl\{\boldsymbol{x}^1,\ldots,\boldsymbol{x}^m\bigr\} \subset X$ given, for any $A \subseteq X$ Borel set, by

$$\widehat{\rho}_{\boldsymbol{x}^{\otimes m}}(A) \equiv \frac{1}{m}\sum_{i=1}^{m}\delta_{\boldsymbol{x}^i}(A)$$

denoting with $\delta_{\boldsymbol{x}}(\,\cdot\,) \equiv \mathbb{1}_{(\,\cdot\,)}(\boldsymbol{x})$, $\boldsymbol{x} \in X$, the usual Dirac probability measure on X centered at $\boldsymbol{x}$.

Appendix

Mercer's kernels. A *Mercer's kernel* K on X is a function $K\colon X\times X \to \mathbb{R}$ which is symmetric, continuous and semipositive-definite in the following natural terms.

- Symmetry: for each $\boldsymbol{x},\boldsymbol{x}' \in X$, $K(\boldsymbol{x},\boldsymbol{x}') = K(\boldsymbol{x}',\boldsymbol{x})$.
- Continuity: with respect to the usual Euclidean norm $\|\cdot\|_2$ on $\mathbb{R}^{2d}$, for instance.
- Semipositiveness: for each $N \in \mathbb{N}\setminus\{\,0\,\}$, $\boldsymbol{u}^1,\ldots,\boldsymbol{u}^N \in X$ and $c^1,\ldots,c^N \in \mathbb{R}$,

$$\sum_{i,j=1}^{N} c^i c^j K(\boldsymbol{u}^i,\boldsymbol{u}^j) \geq 0$$

and in particular $\bigl[K(\boldsymbol{u}^i,\boldsymbol{u}^j)\bigr]_{i,j=1}^{N} \in \mathbb{R}^{N\times N}$ is a symmetric and semipositive-definite real matrix and thus, as such, each of its N eigenvalues is real and non-negative.

Example. Given $\sigma \in \,]0,+\infty[$, the *Gaussian kernel* $K \equiv K_\sigma$ on X defined by placing, for any $\boldsymbol{x},\boldsymbol{x}' \in X$,

$$K_\sigma(\boldsymbol{x},\boldsymbol{x}') \doteq \exp\biggl(-\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|_2^2}{2\sigma^2}\biggr).$$

Reproducing kernel Hilbert spaces. A *reproducing kernel Hilbert space* or RKHS $(\mathcal{H}, \langle \,\cdot\,;\,\cdot\,\rangle_{\mathcal{H}})$ is a real Hilbert space of continuous functions from X to $\mathbb{R}$ in which all the evaluation functionals result (linear and) continuous (in zero) belonging therefore in the dual space $\mathcal{H}'$: that is, for any $\boldsymbol{x} \in X$ and $f \in \mathcal{H}$,

$$|f(\boldsymbol{x})| = \mathcal{O}_{\boldsymbol{x}}(\,\|f\|_{\mathcal{H}}) \equiv \mathcal{O}_{\boldsymbol{x}}(\langle f\,;f\rangle_{\mathcal{H}}^{1/2}).$$

As a simple result of the classical Riesz-Fréchet representation theorem, any RKHS is associated with a Mercer's kernel K on X which in fact reproduces every function f in $\mathcal{H}$ in the sense that, for each $\boldsymbol{x} \in X$, the evaluation or centering of f at $\boldsymbol{x}$ could be performed by taking the inner product with the centered kernel $K(\boldsymbol{x}, \cdot\,) \equiv K_{\boldsymbol{x}}(\,\cdot\,)$ at $\boldsymbol{x}$ itself: so, for any $\boldsymbol{x} \in X$ and $f \in \mathcal{H}$,

$$f(\boldsymbol{x}) = \langle f\,; K_{\boldsymbol{x}}\rangle_{\mathcal{H}}$$

a kind of relation which actually implies that continuity the definition requires (Cauchy-Schwarz inequality). The interesting part is of course the viceversa to all this.

Theorem (Moore-Aronszajn)**.** *Let K be a Mercer's kernel on X. Then there exists an unique RKHS consisting of continuous functions from X to $\mathbb{R}$ for which K is a reproducing kernel.*

Proof. The existence is quite standard. The uniqueness is substantially based on the well-known decomposition theorem of a real Hilbert space into the direct sum of a closed vectorial subspace with its orthogonal space. □

Remark. It's not entirely straightforward to construct a real Hilbert space of functions which is not a RKHS.

Functional derivatives. A *functional* or *variational derivative* relates a change in an abstract functional to a change in functions on which the functional depends. When the set of the functions considered is a real Banach space, the functional derivative becomes known as the *Fréchet derivative* and extends the concept of total differentiability; when such a set is instead a more general locally convex topological real vector space, one uses the *Gâteaux derivative* which extends the concept of directional differentiability.

Frêchet derivative. Let be $(V, \|\cdot\|_V), (W, \|\cdot\|_W)$ two real Banach spaces, $U \subseteq V$ an open set and $u_0 \in U$. Then a functional $F\colon U \to W$ is *Fréchet differentiable at* u_0 if there exists a bounded linear operator $L \equiv L_{F,u_0}\colon V \to W$ (L is linear with $\sup_{v \in V\setminus\{\,0\,\}} \|L(v)\|_W / \|v\|_V \equiv \sup_{\|v\|_V = 1} \|L(v)\|_W < +\infty$) such that

$$\lim_{\|h\|_V \to 0} \frac{\|F(u_0 + h) - F(u_0) - L(h)\|_W}{\|h\|_V} = 0.$$

In that case such an operator L is unique: it's written $\mathrm{D}F(u_0)$ and called the *Fréchet derivative of F at u_0*.

Gâteaux derivative. Let be V, W two locally convex topological real vector spaces, $U \subseteq V$ an open set and $u_0 \in U$. Then a functional $F\colon U \to W$ is *Gâteaux differentiable at* u_0 if there exists an operator $G \equiv G_{F,u_0}\colon V \to W$ such that, for any direction $h \in V$ at u_0, the following limit in W holds:

$$\mathrm{D}F(u_0\,; h) \doteq G(h) = \lim_{t \to 0} \frac{F(u_0 + th) - F(u_0)}{t}.$$

Remark. Fréchet differentiability implies Gâteaux differentiability, but the viceversa cannot hold in general.

Tikhonov regularization. The most commonly employed method of regularization for ill-posed problems: that is, a procedure of adding information in order to solve such a problem or to prevent overfitting.

Ill-posedness. A well-posed problem in the classical sense of Hadamard has the peculiarities that a solution exists, it's unique and it changes continuously with respect to the initial conditions; problems which are not well-posed are termed ill-posed. Inverse problems, namely processes of calculating from a set of observations the causal factors which produced them, are often ill-posed. Well, if a problem results ill-posed, then it needs to be re-formulated for numerical treatment including typically some additional preference assumptions as, for instance, smoothness of solutions: such a process is known in fact as regularization.

Let's observe that, as a continuum model must often be discretized in order to obtain a numerical solution, even if a problem is well-posed it may still be ill-conditioned: when occours that an arbitrarily small variation in the initial data could lead to much bigger errors in the answers.

Overfitting. Overfitting is the production of an analysis which corresponds too closely or even exactly to a given set of data and may therefore fail to fit additional data or to predict reliably future observations as well: so an overfitted model is a statistical model that contains more parameters that can be justified by the data.

The essence of overfitting is to have unknowingly extracted some of the residual variation or noise as if that discrepancy represents a considerable portion of the underlying model structure, while we'd like instead to make sure we obtain something much more intrinsic to the situation. The possibility of doing so exists basically because the criterion used for selecting a model isn't the same as that used to establish its suitability: that is, there's a conflict between memorizing or observing data and learning or generalizing from a trend.

Viceversa, when a statistical model fails to capture enough of the underlying structure of the assigned data, we talk about underfitting: so an unfitted model is a model where some parameters or terms that would appear in a correctly specified model are missing. For instance, when fitting a linear model to non-linear data.

To be rigorous, within learning theory we should say respectively overtraining and undertraining.

Linear systems. Let be $m, p \in \mathbb{N} \setminus \{ 0 \}$, $A \in \mathbb{R}^{m \times p}$, $\boldsymbol{b} \in \mathbb{R}^m$ and let's focus on the $m \times p$ linear system

$$A\boldsymbol{x} = \boldsymbol{b}.$$

The very standard approach to determine such a $\boldsymbol{x} \in \mathbb{R}^p$ is given by the ordinary least squares linear regression or OLS which, as is well-known, intends to minimize on $\mathbb{R}^p$ the sum of the corresponding squared residuals $\|A\boldsymbol{x} - \boldsymbol{b}\|_2^2$, where $\|\cdot\|_2$ denotes the usual Euclidean norm on $\mathbb{R}^m$. However a whole series of conditions which involves m, p and A, as $m > p$ plus A injective, must be verified and consequently it could likely happen that the solution doesn't exist or that it isn't unique: in short words, that the problem is ill-posed.

Within such situations, OLS estimation leads to an overdetermined - overfitted - or more often an undetermined - underfitted - system of equations: think about the fact that most real-world phenomena have the effect of low-pass filters in the forward direction $A\colon \boldsymbol{x} \mapsto \boldsymbol{b}$ and thus, in solving the inverse problem, the inverse map $\boldsymbol{b} \mapsto \boldsymbol{x}$ operates conversely as a high-pass filter with the very undesirable tendency of amplifying noise as its singular values become larger. In addition, OLS implicity nullifies every element of the reconstructed version of $\boldsymbol{x}$ that is in the null-space or kernel of A, rather than allowing for a model to be used as a prior for $\boldsymbol{x}$.

In order to overcome all this giving preference to a particular solution $\widehat{\boldsymbol{x}} \in \mathbb{R}^p$ with some desirable properties, one includes a suitably chosen Tikhonov regularization matrix $\Gamma \in \mathbb{R}^{m \times p} \setminus \{ 0 \}$ in such a minimization:

$$\min_{\boldsymbol{x} \in \mathbb{R}^p} \left\{ \|A\boldsymbol{x} - \boldsymbol{b}\|_2^2 + \|\Gamma \boldsymbol{x}\|_2^2 \right\}.$$

Then there exists an unique solution $\widehat{\boldsymbol{x}} \equiv \widehat{\boldsymbol{x}}_\Gamma \in \mathbb{R}^p$ which is explicitly given by the following formula:

$$\widehat{\boldsymbol{x}} = \left(A^\mathsf{T} A + \Gamma^\mathsf{T} \Gamma\right)^{-1} A^\mathsf{T} \boldsymbol{b}.$$

This regularization improves also the conditioning of the problem, enabling a direct numerical solution, and guarantees the freedom to modulate its effect by the scale of Γ itself (for instance, $\widehat{\boldsymbol{x}}_{\Gamma \equiv 0}$ is the OLS solution).

We could think about Γ as the rectangular diagonal matrix with all $\alpha \in]0, +\infty[$ on the improper diagonal and in fact, for such a case, we conclude with the three following consequential remarks.

Singular values decomposition. Let's assume $m \geq p$ and let $V \in \mathbb{R}^{p \times p}$ and $U \in \mathbb{R}^{m \times m}$ be those unitary matrices ($V^\mathsf{T} \equiv V^{-1}$ and $U^\mathsf{T} \equiv U^{-1}$), and $\Sigma \in \mathbb{R}^{m \times p}$ be that rectangular diagonal matrix with non-negative diagonal components $\sigma_1, \ldots, \sigma_p \geq 0$, which compose the singular values decomposition or SVD of A:

$$A = U \Sigma V^\mathsf{T}.$$

The columns of V are the right-singular vectors of A, those of U are the left-singular vectors of A and the $\sigma_1, \ldots, \sigma_p$ are the singular values of A. The latter coincide exactly with the square roots of the eigenvalues of the symmetric matrix $A^\mathsf{T} A \in \mathbb{R}^{p \times p}$ and, geometrically, they correspond to the lengths of the semi-axes of the ellipsoid in $\mathbb{R}^m$ that is the image by A of the unit sphere in $\mathbb{R}^p$.

By this writing, it's rather immediate to discover that the Γ-regularized solution $\widehat{\boldsymbol{x}} \equiv \widehat{\boldsymbol{x}}_\Gamma \in \mathbb{R}^p$ becomes

$$\widehat{\boldsymbol{x}} = VDU^\mathsf{T}\boldsymbol{b}$$

where $D \in \mathbb{R}^{p\times m}$ is the rectangular diagonal matrix with diagonal components $\sigma_i/({\sigma_i}^2 + \alpha^2) \geq 0$, $i = 1, \ldots, p$. That shows clearly the effect of the Tikhonov parameter α on the condition number of the regularized problem:

$$\max_{i=1,\ldots,p} \frac{\sigma_i}{{\sigma_i}^2 + \alpha^2} \approx_{\alpha\uparrow+\infty} \min_{i=1,\ldots,p} \frac{\sigma_i}{{\sigma_i}^2 + \alpha^2} \approx_{\alpha\uparrow+\infty} 0.$$

Wiener's filter. So, if $r := \operatorname{rank}(A) \leq m \wedge p \equiv p$, and if $(\boldsymbol{u}^i)_{i=1}^r$ and $(\boldsymbol{v}^i)_{i=1}^r$ are r independent columns of U and V respectively, then the weights $w_i := {\sigma_i}^2/({\sigma_i}^2 + \alpha^2) \in [0, 1]$, $i = 1, \ldots, r$, allow us to get

$$\widehat{\boldsymbol{x}} = \sum_{i=1}^r w_i \frac{(\boldsymbol{u}^i)^\mathsf{T}\boldsymbol{b}}{\sigma_i}\boldsymbol{v}^i.$$

The Tikhonov factor. Finally, a possible approach to determine the optimal regularization parameter α relies on the so called leave-one-out cross-validation through which it could be shown that such a α minimizes the ratio between the sum of the squared residuals and the square of the number of degrees of freedom:

$$\frac{\|A\widehat{\boldsymbol{x}} - \boldsymbol{b}\|_2^2}{\left[\operatorname{tr}\Big(I_m - A\left(A^\mathsf{T}A + \alpha^2 I_p\right)^{-1}A^\mathsf{T}\Big)\right]^2}$$

expression which could be simplified using again the previous singular values decomposition of A.

References

[1] F. Cuker, S. Smale. *On the Mathematical Foundations of Learning.* Bulletin of the American Mathematical Society (New Series), Volume 39, Number 1, Pages 1-49, 2001.

[2] T. Poggio, S. Smale. *The Mathematics of Learning: Dealing with Data.* Notices of the American Mathematical Society, Volume 50, Number 5, Pages 537-544, 2003.