Some mathematical problems related to artificial intellegence

Vladimir Temlyakov

Steklov Institute of Mathematics, 15 February 2024

Classical sampling recovery problem

Let Ω be a compact subset of $\mathbb{R}^d$. Suppose that $f \in C(\Omega)$ is a function that provides the mapping $x \to y$, $x \in \Omega$, $y \in \mathbb{R}$. Assume that we are given the data: $x_i \to y_i$, $i = 1, \dots m$. How to find f? Extra ingredients and assumptions:

1. $f \in W$
2. what to do if there is no $f \in W$ such that $f(x_i) = y_i$, $i = 1, \dots m$
3. x_i are chosen randomly
4. y_i are corrupted with noise

Problem setting

Let $X \subset \mathbb{R}^d$, $Y \subset \mathbb{R}$ be Borel sets, ρ be a Borel probability measure on $Z = X \times Y$. For $f : X \to Y$ define the error

$$\mathcal{E}(f) := \mathcal{E}_2(f) := \int_Z (f(x) - y)^2 d\rho.$$

Consider ρ_X - the marginal probability measure on X (for $S \subset X$, $\rho_X(S) := \rho(S \times Y)$). Define

$$f_\rho(x) := \mathbb{E}(y|x)$$

to be a conditional expectation of y.
The function f_ρ is known in statistics as the regression function of ρ. It is clear that if $f_\rho \in L_2(\rho_X)$ then it minimizes the error $\mathcal{E}(f)$ over all $f \in L_2(\rho_X)$: $\mathcal{E}(f_\rho) \leq \mathcal{E}(f)$, $f \in L_2(\rho_X)$. Thus, in the sense of error $\mathcal{E}(\cdot)$ the regression function f_ρ is the best to describe the relation between inputs $x \in X$ and outputs $y \in Y$.

Optimization problem

Also define a more general error

$$\mathcal{E}_p(f) := \int_Z |f(x) - y|^p d\rho \qquad 1 \le p < \infty.$$

Note that the following optimization problem

$$\inf_{f \in L_p(\rho_X)} \mathcal{E}_p(f)$$

is a convex optimization problem over the Banach space $L_p(\rho_X)$ or over the Hilbert space $L_2(\rho_X)$ in the case $p = 2$.
In the case of $p = 2$ our goal in terms of nonparametric statistics (distribution-free theory of regression) is the following.
Given: (x_i, y_i), $i = 1, \dots, m$, independent identically distributed according to ρ, $|y| \le M$ a.e.
Find a good estimator $\hat{f}$ for f_ρ with the error measured as $\mathbb{E}(\|f_\rho - \hat{f}\|^2_{L_2(\rho_X)})$.

Learning theory setting

Our setting is similar to the setting of the distribution-free regression problem. The goal is to find an estimator $f_{\mathbf{z}}$, on the base of given data $\mathbf{z} = ((x_1, y_1), \dots, (x_m, y_m))$ that approximates f_ρ (or its projection) well with high probability. We assume that (x_i, y_i), $i = 1, \dots, m$ are independent and distributed according to ρ. As in the distribution-free theory of regression we measure the error in the $L_2(\rho_X)$ norm.

We note that a standard setting in the distribution-free theory of regression (see the book Györfy, Kohler, Krzyzak and Walk (2002)) involves the expectation as a measure of quality of an estimator. An important new feature of the setting in learning theory formulated in Cucker and Smale (2001) is the following. They propose to study systematically the probability distribution function

$$\rho^m\{\mathbf{z} : \|f_\rho - f_{\mathbf{z}}\|_{L_2(\rho_X)} \geq \eta\}$$

instead of the expectation.

Mathematical formulation

There are several important ingredients in mathematical formulation of the learning problem. In our formulation we follow the way that has become standard in approximation theory and based on the concept of optimal method. We begin with a class $\mathcal{M}$ of admissible measures ρ. Usually, we impose restrictions on ρ in the form of restrictions on the regression function f_ρ: $f_\rho \in \Theta$. Then the first step is to find an optimal estimator for a given class Θ of priors (we assume $f_\rho \in \Theta$). In regression theory a usual way to evaluate performance of an estimator $f_{\mathbf{z}}$ is by studying its convergence in expectation, i.e. the rate of decay of the quantity $\mathbb{E}(\|f_\rho - f_{\mathbf{z}}\|^2_{L_2(\rho_X)})$ as the sample size m increases. Here the expectation is taken with respect to the product measure ρ^m defined on Z^m. We note that

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) = \|f_{\mathbf{z}} - f_\rho\|^2_{L_2(\rho_X)}$$

Assumptions

An important question in finding an optimal $f_{\mathbf{z}}$ is the following:
How to describe the class Θ of priors?
In other words, what characteristics of Θ govern, say, the optimal rate of decay of $\mathbb{E}(\|f_\rho - f_{\mathbf{z}}\|^2_{L_2(\rho_X)})$ for $f_\rho \in \Theta$?
As we already mentioned above a more accurate and more delicate way of evaluating performance of $f_{\mathbf{z}}$ has been pushed forward in Cucker and Smale (2001).

A historical comment

Previous works in statistics and learning theory (see Barron (1991), Barron, Birgé and Massart (1999), Barron, Cohen, Dahmen and DeVore (2008), Binev, Cohen, Dahmen, DeVore and V. Temlyakov (2005), Cucker and Smale (2001), DeVore, Kerkyacharian, Picard and Temlyakov (2004), DeVore, Kerkyacharian, Picard and Temlyakov (2006), Györfy, Kohler, Krzyzak and Walk (2002), Konyagin and Temlyakov (2004), Konyagin and Temlyakov (2007), Lugosi (2002), Temlyakov (2008), Vapnik (1998), Van de Geer (2000)) indicate that the compactness characteristics of Θ play a fundamental role in the above problem. It is convenient for us to express compactness of Θ in terms of the entropy numbers.

Kolmogorov width and entropy

For two sets A and B in a Banch space X define the best approximation of A by B (the deviation of A from B)

$$d(A,B) := \sup_{a\in A} \inf_{b\in B} \|a-b\|.$$

Kolmogorov width of a centrally symmetric compact set $W \subset X$:

$$d_n(W,X) := \inf_{B-\textit{linear subspace}, \dim B \le n} d(W,B).$$

Entropy numbers of a compact set $W \subset X$:

$$\varepsilon_k(W,X) := \inf_{B-\textit{finite set of points of cardinality } |B|\le 2^k} d(W,B).$$

Error and confidence

We have already mentioned above that the study of the probability distribution function $\rho^m\{\mathbf{z} : \|f_\rho - f_{\mathbf{z}}\|_{L_2(\rho_X)} \geq \eta\}$ is a more difficult and delicate problem than the study of the expectation $\mathbb{E}(\|f_\rho - f_{\mathbf{z}}\|^2_{L_2(\rho_X)})$. We encounter this difficulty even at the level of formulation of a problem. The reason for this is that the probability distribution function provides control of two characteristics: η – the error of estimation and $1 - \rho^m\{\mathbf{z} : \|f_\rho - f_{\mathbf{z}}\|_{L_2(\rho_X)} \geq \eta\}$ – the confidence of the error η. Therefore, we need a mathematical formulation of the above discussed problems of optimal estimators.

Accuracy confidence function

We proposed (see DeVore, Kerkyacharian, Picard and Temlyakov (2006)) to study the following function that we call the accuracy confidence function. Let a set $\mathcal{M}$ of admissible measures ρ, and a sequence $E := \{E(m)\}_{m=1}^{\infty}$ of allowed classes $E(m)$ of estimators be given. For $m \in \mathbb{N}$, $\eta > 0$ we define

$$\mathbf{AC}_m(\mathcal{M}, E, \eta) := \inf_{E_m \in E(m)} \sup_{\rho \in \mathcal{M}} \rho^m\{\mathbf{z} : \|f_\rho - f_{\mathbf{z}}\|_{L_2(\rho_X)} \geq \eta\}$$

where E_m is an estimator that maps $\mathbf{z} \to f_{\mathbf{z}}$. For example, $E(m)$ could be a class of all estimators, a class of linear estimators of the form

$$f_{\mathbf{z}} = \sum_{i=1}^{m} w_i(x_1, \dots, x_m, x) y_i,$$

or a specific estimator. In the case $E(m)$ is the set of all estimators, $m = 1, 2, \dots$, we write $\mathbf{AC}_m(\mathcal{M}, \eta)$.

Main theorem

We let μ be any Borel probability measure defined on X and let $\mathcal{M}(\Theta, \mu)$ denote the set of all $\rho \in \mathcal{M}(\Theta)$ such that $\rho_X = \mu$, $|y| \le 1$, where $\mathcal{M}(\Theta) = \{\rho : f_\rho \in \Theta\}$. Here is a result from DeVore, Kerkyacharian, Picard and Temlyakov (2006).

Theorem (DKPT)

Let μ be a Borel probability measure on X. Assume $r > 0$ and Θ is a compact subset of $L_2(\mu)$ such that $\Theta \subset \frac{1}{4} U(\mathcal{C}(X))$ and

$$\varepsilon_n(\Theta, L_2(\mu)) \asymp n^{-r}.$$

Then there exist $\delta_0 > 0$ and $\eta_m^- \le \eta_m^+$, $\eta_m^- \asymp \eta_m^+ \asymp m^{-\frac{r}{1+2r}}$ such that the following two relations hold

$$\mathbf{AC}_m(\mathcal{M}(\Theta, \mu), \eta) \ge \delta_0 \quad \text{for} \quad \eta \le \eta_m^-$$

$$C_1 e^{-c_1(r) m \eta^2} \le \mathbf{AC}_m(\mathcal{M}(\Theta, \mu), \eta) \le e^{-c_2 m \eta^2} \quad \text{for} \quad \eta \ge \eta_m^+.$$

Some comments

Let us now make some conclusions. First of all, the above theorem shows that the entropy numbers $\epsilon_n(\Theta, L_2(\mu))$ are the right characteristic of the class Θ in the estimation problem. The behavior of the sequence $\{\epsilon_n(\Theta, L_2(\mu))\}$ determines the behavior of the sequence $\{\mathbf{AC}_m(\mathcal{M}(\Theta, \mu), \eta)\}$ of the **AC**-functions.
Secondly, proof of that theorem points out that the optimal (in the sense of order) estimator can be always constructed as a Least Squares Estimator.
The above theorem discovers a new phenomenon – sharp phase transition. The behavior of the accuracy confidence function changes dramatically within the critical interval $[\eta_m^-, \eta_m^+]$. It drops from a constant δ_0 to an exponentially small quantity $\exp(-cm^{1/(1+2r)})$. One may also call the interval $[\eta_m^-, \eta_m^+]$ the interval of phase transition.

Least Squares Estimators

Let $W \subset \mathcal{C}(X)$ be a compact subset of $\mathcal{C}(X)$. As an estimator $f_{\mathbf{z}}$ for the data $\mathbf{z} = ((x_1, y_1), \ldots, (x_m, y_m))$ we take

$$f_{\mathbf{z},W} := \arg\min_{f \in W} \mathcal{E}_{\mathbf{z}}(f),$$

where

$$\mathcal{E}_{\mathbf{z}}(f) := \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - y_i)^2.$$

Theorem (DKPT) justifies the step of replacement the original optimization problem

$$\inf_{f \in \Theta} \mathcal{E}(f) \quad \text{by the discrete problem} \quad \inf_{f \in W} \mathcal{E}_{\mathbf{z}}(f)$$

with appropriately chosen W. In particular, we can take $W = \Theta$.

Defect function

Let $f \in L_2(\rho_X)$. The defect function of f is

$$L_{\mathbf{z}}(f) := L_{\mathbf{z},\rho}(f) := \mathcal{E}(f) - \mathcal{E}_{\mathbf{z}}(f); \quad \mathbf{z} = (z_1, \dots, z_m), \quad z_i = (x_i, y_i).$$

We are interested in estimating $L_{\mathbf{z}}(f)$ for functions f coming from a given class W.
We assume that ρ and W satisfy the following condition.

$$\forall f \in W, \quad f : X \to Y, \quad |f(x) - y| \le M \quad \text{a.e.} \tag{1}$$

An illustrating result

Theorem (KT; Konyagin and Temlyakov (2004))

Assume that ρ, W satisfy (1) and W is such that

$$\sum_{n=1}^{\infty} n^{-1/2}\varepsilon_n(W) = \infty.$$

For $\eta > 0$ define $J := J(\eta/M)$ as the minimal j satisfying $\varepsilon_{2^j} \le \eta/(8M)$ and

$$S_J := \sum_{j=1}^{J} 2^{(j+1)/2}\varepsilon_{2^{j-1}}.$$

Then for m, η satisfying $m(\eta/S_J)^2 \ge 480M^2$ we have

$$\rho^m\{\mathbf{z} : \sup_{f\in W} |L_{\mathbf{z}}(f)| \ge \eta\} \le C(M,\varepsilon(W))\exp(-c(M)m(\eta/S_J)^2).$$

Sampling discretization with absolute error

On this way we arrive at the following discretization problem, which is important by itself.
Let $W \subset L_q(\Omega, \mu)$, $1 \le q < \infty$, be a class of continuous on Ω functions. We are interested in estimating the following optimal errors of discretization of the L_q norm of functions from W

Sampling discretization with absolute error

On this way we arrive at the following discretization problem, which is important by itself.
Let $W \subset L_q(\Omega, \mu)$, $1 \le q < \infty$, be a class of continuous on Ω functions. We are interested in estimating the following optimal errors of discretization of the L_q norm of functions from W

$$er_m(W, L_q) := \inf_{\xi^1,\dots,\xi^m} \sup_{f \in W} \left| \|f\|_q^q - \frac{1}{m} \sum_{j=1}^{m} |f(\xi^j)|^q \right|.$$

General theorem

One can derive the following result from the above Theorem KT.

Theorem (T1; VT, 2018, 2022)

Assume that a class of real functions W is such that for all $f \in W$ we have $\|f\|_\infty \le M$ with some constant M. Also assume that the entropy numbers of W in the uniform norm L_∞ satisfy the condition

$$\varepsilon_n(W) \le n^{-r}(\log(n+1))^b, \qquad r \in (0,1/2), \quad b \ge 0, \quad n \in \mathbb{N}.$$

Then

$$er_m(W) := er_m(W, L_2) \le C(M,r,b)m^{-r}(\log(m+1))^b, \quad m \in \mathbb{N}.$$

Comments

A theorem alike Theorem T1 might be another way to justify the step of replacement the original optimization problem $\inf_{f\in\Theta}\mathcal{E}(f)$ by the discrete problem $\inf_{f\in W}\mathcal{E}_{\mathbf{z}}(f)$.
Theorem T1 is a rather general theorem, which connects the behavior of absolute errors of discretization with the rate of decay of the entropy numbers.

Comments

A theorem alike Theorem T1 might be another way to justify the step of replacement the original optimization problem $\inf_{f\in\Theta}\mathcal{E}(f)$ by the discrete problem $\inf_{f\in W}\mathcal{E}_{\mathbf{z}}(f)$.
Theorem T1 is a rather general theorem, which connects the behavior of absolute errors of discretization with the rate of decay of the entropy numbers.

- We impose a restriction $r < 1/2$ in Theorem T1 because the probabilistic technique from the supervised learning theory has a natural limitation to $r \leq 1/2$.

Comments

A theorem alike Theorem T1 might be another way to justify the step of replacement the original optimization problem $\inf_{f\in\Theta}\mathcal{E}(f)$ by the discrete problem $\inf_{f\in W}\mathcal{E}_{\mathbf{z}}(f)$.
Theorem T1 is a rather general theorem, which connects the behavior of absolute errors of discretization with the rate of decay of the entropy numbers.

- We impose a restriction $r < 1/2$ in Theorem T1 because the probabilistic technique from the supervised learning theory has a natural limitation to $r \leq 1/2$.
- It would be interesting to understand if Theorem T1 holds for $r \geq 1/2$.

Comments

A theorem alike Theorem T1 might be another way to justify the step of replacement the original optimization problem $\inf_{f\in\Theta}\mathcal{E}(f)$ by the discrete problem $\inf_{f\in W}\mathcal{E}_{\mathbf{z}}(f)$.
Theorem T1 is a rather general theorem, which connects the behavior of absolute errors of discretization with the rate of decay of the entropy numbers.

- We impose a restriction $r < 1/2$ in Theorem T1 because the probabilistic technique from the supervised learning theory has a natural limitation to $r \leq 1/2$.
- It would be interesting to understand if Theorem T1 holds for $r \geq 1/2$.

We point out that in applications to the Machine Learning we are interested in the randomised version of Theorem T1 – we would like to have it for $\xi^1, \ldots, \xi^m$ being independent random variables distributed according to measure μ. Moreover, μ is unknown!

Main problem

We now arrive at a critical problem of Learning Theory (Machine Learning):
How to choose W, which is usually called hypothesis space?
Here are some standard options.

- Choose $W = \Theta$ to be a smoothness class. In this case we assume that $f_\rho \in \Theta$.

Main problem

We now arrive at a critical problem of Learning Theory (Machine Learning):
How to choose W, which is usually called hypothesis space?
Here are some standard options.

- Choose $W = \Theta$ to be a smoothness class. In this case we assume that $f_\rho \in \Theta$.
- Assume that $f_\rho \in \Theta$ and take W as a simpler than Θ class, which well approximates Θ, for instance,
  (a) a ball of a finite-dimensional subspace,
  (b) a collection of n-term approximants with respect to a given finite system of functions,
  (c) a specific manifold parametrized by a finite number of parameters.

Main problem

We now arrive at a critical problem of Learning Theory (Machine Learning):
How to choose W, which is usually called hypothesis space?
Here are some standard options.

- Choose $W = \Theta$ to be a smoothness class. In this case we assume that $f_\rho \in \Theta$.
- Assume that $f_\rho \in \Theta$ and take W as a simpler than Θ class, which well approximates Θ, for instance,
  (a) a ball of a finite-dimensional subspace,
  (b) a collection of n-term approximants with respect to a given finite system of functions,
  (c) a specific manifold parametrized by a finite number of parameters.
- The above (a)–(c) sets without any assumptions on f_ρ.

Sparse approximation

A typical problem of sparse approximation is the following. Let X be a Banach space with norm $\|\cdot\|$ and $\mathcal{D}$ be a set of elements of X. For a given $\mathcal{D}$ consider the set of all m-term linear combinations with respect to $\mathcal{D}$ (m-sparse with respect to $\mathcal{D}$ elements):

$$\Sigma_m(\mathcal{D}) := \{x \in X : x = \sum_{i=1}^{m} c_i g_i, \quad g_i \in \mathcal{D}\}.$$

We are interested in approximation of a given $f \in X$ by elements of $\Sigma_m(\mathcal{D})$. The best we can do is to get the error

$$\sigma_m(f, \mathcal{D}) := \inf_{x \in \Sigma_m(\mathcal{D})} \|f - x\|. \tag{2}$$

Greedy algorithms in approximation theory are designed to provide a simple way to build good approximants of f from $\Sigma_m(\mathcal{D})$. Clearly, we have an optimization problem of $E_f(x) := \|f - x\|$ over the manifold $\Sigma_m(\mathcal{D})$.

Sparse optimization

A typical problem of convex optimization is to find an approximate solution x_0 to the problem

$$\inf_x E(x) \tag{3}$$

under assumption that E is a convex function. In the case that we are optimizing over the whole space X, it is called an unconstrained optimization problem. In many cases we are interested either in optimizing over x of special structure (for instance, $x \in \Sigma_m(\mathcal{D})$, as above) or in optimizing over x from a given domain D (constrained optimization problem). Greedy algorithms are used for finding an approximate solution of special structure for problem (3).

Sparse optimization continue

Usually in convex optimization, the function E is defined on a finite dimensional space $\mathbb{R}^d$. An important argument that motivates us to study this problem in the infinite dimensional setting is that in many contemporary data management applications an ambient space $\mathbb{R}^d$ involves a large dimension d and we would like to obtain bounds on the convergence rate independent of the dimension d. Our results for infinite dimensional spaces provide such bounds on the convergence rate. Thus, we consider a convex function E defined on a Banach space X. It is known that in many engineering applications researchers are interested in an approximate solution of problem (3) as a linear combination of a few elements from a given system $\mathcal{D}$ of elements. There is an increasing interest in building such sparse approximate solutions using different greedy-type algorithms.

Greedy approximation in Banach spaces 1

We begin with a brief description of greedy approximation methods in Banach spaces. Let X be a Banach space with norm $\|\cdot\|$. We say that a set of elements (functions) $\mathcal{D}$ from X is a dictionary if each $g \in \mathcal{D}$ has norm bounded by one ($\|g\| \le 1$) and the closure of $\operatorname{span} \mathcal{D}$ is X. A symmetrized dictionary is defined as

$$\mathcal{D}^{\pm} := \{\pm g, g \in \mathcal{D}\}.$$

We denote the closure (in X) of the convex hull of $\mathcal{D}^{\pm}$ by $A_1(\mathcal{D})$. In other words $A_1(\mathcal{D})$ is the closure of $\operatorname{conv}(\mathcal{D}^{\pm})$. We use this notation because it has become a standard notation in relevant greedy approximation literature.

Greedy approximation in Banach spaces 2

For a nonzero element $f \in X$ we let F_f denote a norming (peak) functional for f that is a functional with the following properties $\|F_f\| = 1$, $F_f(f) = \|f\|$. The existence of such a functional is guaranteed by the Hahn-Banach theorem. The norming functional F_f is a linear functional (in other words is an element of the dual to X space X^*) which can be explicitly written in some cases. In a Hilbert space F_f can be identified with $f\|f\|^{-1}$. In the real L_p, $1 < p < \infty$, it can be identified with $f|f|^{p-2}\|f\|_p^{1-p}$. We describe a typical greedy algorithm which uses a norming functional. We call this family of algorithms dual greedy algorithms. Let $\tau := \{t_k\}_{k=1}^{\infty}$ be a given weakness sequence of nonnegative numbers $t_k \le 1$, $k = 1, \dots$. We define the Weak Chebyshev Greedy Algorithm (WCGA) that is a generalization for Banach spaces of the Weak Orthogonal Greedy Algorithm (Weak Orthogonal Matching Pursuit).

Weak Chebyshev Greedy Algorithm (WCGA). We define $f_0^c := f_0^{c,\tau} := f$. Then for each $m \geq 1$ we have the following inductive definition.

(1) $\varphi_m^c := \varphi_m^{c,\tau} \in \mathcal{D}$ is any element satisfying

$$|F_{f_{m-1}^c}(\varphi_m^c)| \geq t_m \sup_{g \in \mathcal{D}} |F_{f_{m-1}^c}(g)|.$$

(2) Define

$$\Phi_m := \Phi_m^\tau := \operatorname{span}\{\varphi_j^c\}_{j=1}^m,$$

and define $G_m^c := G_m^{c,\tau}$ to be the best approximant to f from Φ_m.

(3) Let

$$f_m^c := f_m^{c,\tau} := f - G_m^c.$$

Greedy approximation in Banach spaces 4

In case $t_m = 1$ in (1) we assume that such φ_m^c exists. The index c in the notation refers to Chebyshev. We use the name Chebyshev in this algorithm because at step (2) of the algorithm we use best approximation operator which bears the name of the Chebyshev projection or the Chebyshev operator. In the case of Hilbert space the Chebyshev projection is the orthogonal projection and it is reflected in the name of the algorithm. We use notation f_m for the residual of the algorithm after m iterations. This standard in approximation theory notation is justified by the fact that we interpret f as a residual after 0 iterations and iterate the algorithm replacing f_0 by f_1, f_2, and so on. In signal processing the residual after $m$iterations is often denoted by r_m or r^m.

Greedy approximation in Banach spaces 5

For a Banach space X we define the modulus of smoothness

$$\varrho(u) := \sup_{\|x\|=\|y\|=1} (\frac{1}{2}(\|x+uy\| + \|x-uy\|) - 1).$$

The uniformly smooth Banach space is the one with the property

$$\lim_{u\to 0} \varrho(u)/u = 0.$$

Rate of convergence

It is known that the WCGA converges in any uniformly smooth Banach space under mild conditions on the weakness sequence $\{t_k\}$, for instance, $t_k = t$, $k = 1, 2, \dots$, $t > 0$, guarantees such convergence. The following theorem provides rate of convergence.

Theorem (VT (2001))

Let X be a uniformly smooth Banach space with modulus of smoothness $\varrho(u) \le \gamma u^q$, $1 < q \le 2$. Take a number $\varepsilon \ge 0$ and two elements f, f^ε from X such that

$$\|f - f^\varepsilon\| \le \varepsilon, \quad f^\varepsilon / B \in A_1(\mathcal{D}),$$

with some number $B = C(f, \varepsilon, \mathcal{D}, X) > 0$. Then, for the WCGA we have ($p := q/(q-1)$)

$$\|f_m^{c,\tau}\| \le \max\left(2\varepsilon, C(q,\gamma)(B+\varepsilon)(1+\sum_{k=1}^m t_k^p)^{-1/p}\right).$$

Comments

The above Theorem VT simultaneously takes care of two issues: noisy data and approximation in an interpolation space. In order to apply it for noisy data we interpret f as a noisy version of a signal and f^ε as a noiseless version of a signal. Then, assumption $f^\varepsilon/B \in A_1(\mathcal{D})$ describes our smoothness assumption on the noiseless signal. Theorem VT can be applied for approximation of f under assumption that f belongs to one of interpolation spaces between X and the space generated by the $A_1(\mathcal{D})$-norm (atomic norm).

In particular, Theorem VT gives the upper bound for the convergence rate for all $f_0 \in A_1(\mathcal{D})$. It is in style of the worst case setting for the class $A_1(\mathcal{D})$.

Examples of dictionaries 1

Perhaps the first example of sparse approximation with respect to a dictionary was considered by E. Schmidt (1906), who studied the approximation of functions $f(x, y)$ of two variables by bilinear forms,

$$\sum_{i=1}^{m} u_i(x) v_i(y),$$

in $L_2([0,1]^2)$. This problem is closely connected with properties of the integral operator

$$J_f(g) := \int_0^1 f(x, y) g(y) dy$$

with kernel $f(x, y)$. Schmidt gave an expansion (known as the Schmidt expansion)

$$f(x, y) = \sum_{j=1}^{\infty} s_j(J_f) \phi_j(x) \psi_j(y).$$

Examples of dictionaries 2

In the above Schmidt expansion $\{s_j(J_f)\}$ is a nonincreasing sequence of singular numbers of J_f, i.e. $s_j(J_f) := \lambda_j(J_f^* J_f)^{1/2}$, where $\{\lambda_j(A)\}$ is a sequence of eigenvalues of an operator A, and J_f^* is the adjoint operator to J_f. The two sequences $\{\phi_j(x)\}$ and $\{\psi_j(y)\}$ form orthonormal sequences of eigenfunctions of the operators $J_f J_f^*$ and $J_f^* J_f$, respectively.
He also proved that

$$\|f(x,y) - \sum_{j=1}^{m} s_j(J_f)\phi_j(x)\psi_j(y)\|_{L_2}$$

$$= \inf_{u_j, v_j \in L_2,\ j=1,\dots,m} \|f(x,y) - \sum_{j=1}^{m} u_j(x)v_j(y)\|_{L_2}.$$

Examples of dictionaries 3

It was understood later that the above best bilinear approximation can be realized by the following greedy algorithm. Assume c_j, $u_j(x)$, $v_j(y)$, $\|u_j\|_{L_2} = \|v_j\|_{L_2} = 1$, $j = 1, \dots, m-1$, have been constructed after $m-1$ steps of the algorithm. At the mth step we choose c_m, $u_m(x)$, $v_m(y)$, $\|u_m\|_{L_2} = \|v_m\|_{L_2} = 1$, to minimize

$$\|f(x,y) - \sum_{j=1}^{m} c_j u_j(x) v_j(y)\|_{L_2}.$$

We call this type of algorithm the Pure Greedy Algorithm (PGA).

Examples of dictionaries 4

Another problem of this type which is well known in statistics is the projection pursuit regression problem. The problem is to approximate in L_2 a given function $f \in L_2$ by a sum of ridge functions, i.e. by

$$\sum_{j=1}^{m} r_j(\omega_j \cdot x), \quad x, \omega_j \in \mathbb{R}^d, \quad j = 1, \dots, m,$$

where r_j, $j = 1, \dots, m$, are univariate functions. The following greedy-type algorithm (projection pursuit) was proposed in Friedman and Stuetzle (1981) to solve this problem.

Assume functions $r_1, \dots, r_{m-1}$ and vectors $\omega_1, \dots, \omega_{m-1}$ have been determined after $m-1$ steps of the algorithm. Choose at mth step a unit vector ω_m and a function r_m to minimize the error

$$\|f(x) - \sum_{j=1}^{m} r_j(\omega_j \cdot x)\|_{L_2}.$$

This is one more example of the Pure Greedy Algorithm.

Examples of dictionaries 5

The following version of the above dictionary of ridge functions is important in the theory of neural networks. We fix a univariate function $\sigma(t)$, $t \in \mathbb{R}$. Usually, this function takes values in $(0,1)$ and increases. Then as a dictionary we consider

$$\{g(\mathbf{x}) : g(\mathbf{x}) = \sigma(\omega \cdot \mathbf{x}), \quad \mathbf{x} \in \Omega \subset \mathbb{R}^d, \quad \omega \in \mathbb{R}^d, \|g\|_2 = 1\}.$$

Constructions based on this dictionary are called shallow neural networks or neural networks with one lair.

We build approximating manifolds inductively. Let $\mathbf{x} \in \Omega \subset \mathbb{R}^d$.
Fix a univariate function $h(t)$. A popular one in the $ReLU$ function: $ReLU(t) = 0$ for $t < 0$ and $ReLu(t) = t$ for $t \geq 0$. ReLU = Rectified Linear Unit.
Take numbers $s, n \in \mathbb{N}$ and build s-term approximants of depth n (neural network with n lairs). In the capacity of parameters take n matrices A_1 of size $s \times d$, $A_2, \ldots, A_n$ of size $s \times s$ and vectors $\mathbf{b}^1, \ldots, \mathbf{b}^n, \mathbf{c}$ from $\mathbb{R}^s$.
At the first step define $\mathbf{y}^1 \in \mathbb{R}^s$

$$\mathbf{y}^1 := h(A_1\mathbf{x} + \mathbf{b}^1) := (h((A_1\mathbf{x})_1 + \mathbf{b}^1_1), \ldots, h((A_1\mathbf{x})_s + \mathbf{b}^1_s))^T.$$

Note that $\mathbf{y}^1$ is a function on $\mathbf{x}$.

At the kth step ($k = 2, \dots, n$) define

$$\mathbf{y}^k := h(A_k \mathbf{y}^{k-1} + \mathbf{b}^k)$$

$$:= (h((A_k \mathbf{y}^{k-1})_1 + \mathbf{b}_1^k), \dots, h((A_1 \mathbf{y}^k)_s + \mathbf{b}_s^k))^T.$$

Finally, after the nth step we define

$$g_n(\mathbf{x}) := \langle \mathbf{c}, \mathbf{y}^n \rangle = \sum_{j=1}^{s} c_j y_j^n.$$

Thus we build a manifold, which is described by the following parameters: n matrices A_1 of size $s \times d$, $A_2, \dots, A_n$ of size $s \times s$ and vectors $\mathbf{b}^1, \dots, \mathbf{b}^n, \mathbf{c}$ from $\mathbb{R}^s$.

The End

Thank you!