On stochastic optimization and statistical learning in reproducing kernel hilbert spaces by support vector machines (SVM)

Norkin, V.I.; Keyzer, M.A.

published in
Informatica
2009

document version
Publisher's PDF, also known as Version of record

INFORMATICA, 2009, Vol. 20, No. 2, 273–292

On Stochastic Optimization and Statistical Learning in Reproducing Kernel Hilbert Spaces by Support Vector Machines (SVM)

Vladimir NORKIN
Glushkov Institute of Cybernetics, National Academy of Sciences of Ukraine
Glushkov avenue 40, 03187 Kiev, Ukraine
e-mail: norkin@i.com.ua

Michiel KEYZER
Center for World Food Studies (SOW-VU), VU University of Amsterdam
De Boelelaan 1105, 1081 HV, Amsterdam, The Netherlands
e-mail: m.a.keyzer@sow.vu.nl

Received: August 2008; accepted: February 2009

Abstract. The paper studies stochastic optimization problems in Reproducing Kernel Hilbert Spaces (RKHS). The objective function of such problems is a mathematical expectation functional depending on decision rules (or strategies), i.e. on functions of observed random parameters. Feasible rules are restricted to belong to a RKHS. This kind of problems arises in on-line decision making and in statistical learning theory. We solve the problem by sample average approximation combined with Tihonov's regularization and establish sufficient conditions for uniform convergence of approximate solutions with probability one, jointly with a rule for downward adjustment of the regularization factor with increasing sample size.

Keywords: stochastic optimization, empirical risk minimization, decision rule, Reproducing Kernel Hilbert Spaces (RKHS), support vector machine (SVM), regularization, convergence.

1. Introduction

Stochastic optimization deals with decision making models and problems under probabilistic uncertainty of the future, see Ruszczynski and Shapiro (2003). Its decision performance criterion usually has the form of a probability functional or a mathematical expectation over random factors and depends on decision parameters. One of the difficulties in such problems is that the probabilistic distribution of random factors is not known completely and may be given, for example, by a set of observations. A main field of application is to find optimal decision rules (also called strategies), i.e., a priori established actions as functions of revealed situations, on the basis of their expected performance, see Ermoliev (1976), Yudin (1979; Ch. 5).

Mathematically similar problems arise in mathematical statistics (regression) and in statistical learning theory, see Vapnik (1998). In statistical leaning theory they are

treated as infinite dimensional optimization problems in the so-called Reproducing Kernel Hilbert Spaces (RKHS), a setting known as kernel learning. RKHS is a very specific Hilbert space, see Aronszajn (1950), Cucker and Smale (2001), for example, strong convergence in norm in this space implies uniform convergence of functions. RKHS is completely defined by its reproducing kernel, i.e., a function of two vector variables such that under one fixed variable (1) the resulting kernel function, as a function of the other variable, is an element of the space and (2) in inner products with other functions the kernel functions act as Dirac's delta-function. It appears that functions in RKHS are approximated by linear combinations of kernel functions. If the kernel is continuous then all functions in RKHS are continuous. In statistical learning theory kernel is interpreted as a measure of similarity of two vectors (situations). Because of these properties, RKHS' considerably differ from spaces of measurable/integrable functions and are of great interest for decision theory under uncertainty.

There are two basic approaches to solving optimization problems in RKHS. The first one combines Tihonov's regularization (Tihonov and Arsenin, 1977) with sample average approximation (regularized empirical risk minimization; Vapnik (1998), Scholkopf and Smola (2002). A solution is sought as a linear combination of kernel functions associated with some subset of observations (support vectors). Due to the so-called Representer Theorem (Kimeldorf and Wahba, 1970), see also Scholkopf and Smola (2002; Theorem 4.2, P. 90), the original optimization problem is reduced to a finite dimensional one. This approach is also known as Support Vector Method/Machine (SVM). An important and debatable issue in this approach is the choice of the regularization parameter reflecting the trade-off between quality of fit to data and complexity of the decision model. The other optimization approach studies sequential (on-line or stochastic approximation) procedures for (quadratic) unconstrained optimization problem in RKHS, see Smale and Zhou (2005), Smale and Yao (2006), where convergence in probability in RKHS-norm is established, that implies uniform convergence in probability.

Vapnik's (1998) convergence analysis of the regularized empirical risk minimization method is based on the uniform functional law of large numbers, which in turn is guaranteed by finiteness of VC-dimension of a corresponding functional class. Assuming finite VC-dimension and fixed regularization parameter, Vapnik (1998) proves uniform convergence of regularized empirical risk approximations to the true regularized risk, leaving out of consideration, however, the convergence of the corresponding sequence of minimizers. Alternatively, other approaches to SVM-convergence analysis that do not rely on VC-dimension, other capacity measures and uniform law of large numbers, make use of stability property of learning algorithms (Bousquet and Elisseeff, 2002; Mukherjee *et al.*, 2006; Takeuchi *et al.*, 2006) and probabilistic concentration measure (large deviation) inequalities (McDiarmid, 1989), but again for fixed regularization parameter.

In the statistical context, regularized empirical least squares risk minimization naturally produces estimators that converge to the regression function in mean norm (and hence in probability; Cucker and Smale, 2001; Cucker and Smale, 2002; Poggio and Smale, 2003; De Vito *et al.*, 2005).

In the stochastic optimization context, Yudin (1974; 1979) studied quadratic stochastic optimization problems over square integrable decision rules and proposed dual meth-

ods for their solution. Raik (1972) studied stochastic optimization problems with probabilistic functionals defined on continuous decision rules and applied Galerkin solution technique. Ermoliev (1976), Ermoliev and Yastremski (1979), Ermoliev and Leonardi (1982) studied optimality conditions for convex stochastic optimization problems with square integrable decision rules and measurability constraints. Chancelier and SOWG (2006) consider stochastic optimization problems over L_p-strategies subject to a specific σ-field measurability constraint and propose a discrete approximation technique for their solution. In this respect, the present paper considers stochastic optimization problems on a more narrow class of strategies, namely from Hilbert space with a reproducing kernel.

Remark that for finite dimensional stochastic optimization problems there is a variety of solution techniques, in particular the sample average approximation method, which in this case need not require regularization, see Ruszczynski and Shapiro (2003). But for the infinite dimensional stochastic optimization problems the situation is different, as regularization may be necessary both for the validation of convergence and for numerical implementation.

In the present paper, we extend the Support Vector Method (SVM) approach that was primarily designed for statistical learning to the more general problem of solving ball-constrained stochastic optimization problems in RKHS. The original problem is replaced by a sample average approximation plus a quadratic regularization term multiplied by a scalar, the regularization parameter. We exploit the fact that a solution of this approximation problem can be found in a kernel form, i.e., as a finite linear combination of kernel functions associated with a sample of observations. Consequently, the implementation of SV-method reduces to a finite dimensional optimization. We study asymptotic convergence of the SV-method under the assumption that regularization parameter $\lambda(m)$ depends on the sample size m and tends to zero, $\lambda(m) \to 0$, as $m \to \infty$. Our main asymptotic convergence result is that if the regularization parameter goes to zero proportionally to $(\ln m)/\sqrt[4]{m}$ as the sample size $m \to \infty$, then corresponding kernel solutions uniformly converge to a true solution with probability one, and the rate of mean error decrease is of the same order, $(\ln m)/\sqrt[4]{m}$. Here the true solution is a (unique) normal minimizer (or a normal solution in terms of Tihonov and Arsenin (1977)) of the original functional minimization problem. As a byproduct we obtain convergence results for SVM-regression and SVM-classification. For a binary classification problem, convergence of a misclassification risk to its absolute minimum holds true with probability one if the regularization factor goes to zero proportionally to $(\ln m)/\sqrt{m}$ as $m \to \infty$. The present paper is an extended version of Norkin and Keyzer (2008).

The paper proceeds as follows. In Section 2 we introduce notations and present necessary facts on Reproducing Kernel Hilbert Spaces. Section 3 formulates stochastic optimization problem in RKHS and describes a computational approximation scheme. Section 4 establishes sufficient conditions of consistency of approximations. Section 5 specializes the results for a kernel regression and binary classification problems. Section 6 concludes. Proofs are presented in Appendix.

2. Reproducing Kernel Hilbert Spaces (RKHS)

To introduce notations we start with basic facts concerning Reproducing Kernel Hilbert Spaces, see Aronszajn (1950); Cucker and Smale (2001); Scholkopf and Smola (2002).

DEFINITION 1. A Hilbert space $H(\Omega)$ of functions defined on a closed set $\Omega \subseteq R^n$ with inner product $\langle \cdot\,, \cdot \rangle$ is said to be a Reproducing Kernel Hilbert Space (RKHS) if there exists a kernel functional $k(\cdot\,, \cdot\,)$ defined on $\Omega \times \Omega$ with the properties:

(i) $k(\cdot, \omega) \in H(\Omega), \forall\, \omega \in \Omega$;

(ii) $f(\omega) = \langle f, k(\cdot, \omega) \rangle, \forall\, f \in H(\Omega)$ and $\forall\, \omega \in \Omega$ (this is a so-called reproducing property of the kernel $k(\cdot\,, \cdot\,)$).

A Reproducing Kernel Hilbert Space with kernel k is denoted as $H_k(\Omega)$ or H_k for short. Corresponding inner (scalar) product and norm in RKHS are denoted as $\langle \cdot\,, \cdot \rangle_k$ and $\| \cdot \|_k = \langle \cdot, \cdot \rangle_k^{1/2}$, respectively; R^n is the n-dimensional real vector space.

PROPOSITION 1 (nonparametric representation of RKHS). In RKHS $H_k(\Omega)$ the set of all finite linear combinations

$$\left\{ f(\omega) = \sum_i \alpha_i k(\omega, \omega^i), \quad \alpha_i \in R^1, \quad \omega^i \in \Omega \right\}$$

constitutes a dense subset.

We remark that the kernel reproducing property ensures that

$$\Big\| \sum_i \alpha_i k(\omega, \omega^i) \Big\|_k^2 = \sum_{i,j} \alpha_i \alpha_j k(\omega^j, \omega^i).$$

PROPOSITION 2. If kernel $k(\omega, \bar{\omega})$ is continuous in $(\omega, \bar{\omega}) \in \Omega \times \Omega$ with compact Ω, then the corresponding RKHS $H_k(\Omega)$ consists of continuous functions.

DEFINITION 2 (Mercer kernel). A function $k(\omega, \bar{\omega})$, $\omega, \bar{\omega} \in \Omega$ is called Mercer kernel if $k(\cdot\,, \cdot\,)$ is continuous and symmetric, and for any finite set of points $\{\omega^i \in \Omega\}$ matrix with entries $\{k(\omega^i, \omega^j)\}$ is positive semidefinite.

PROPOSITION 3 (parametric representation of RKHS). For any Mercer kernel $k(\cdot\,, \cdot\,)$ there exists Reproducing Kernel Hilbert Space $H_k(\Omega)$, defined by this kernel according to Definition 1.

The following propositions establish relations between norms $\|f\|_k$ and $\|f\|_\infty = \sup_{\omega \in \Omega} |f(\omega)|$.

PROPOSITION 4. If $\sup_{\omega \in \Omega} |k(\omega, \omega)| \leqslant K^2 < \infty$, then $\|f\|_\infty \leqslant K \|f\|_k$ and hence (strong) convergence $f^m \to f$ in H_k-norm implies a uniform convergence $f^m \Rightarrow f$ on Ω, $m \to \infty$.

PROPOSITION 5. If $0 < \varepsilon \leqslant k(\omega, \bar{\omega}) \leqslant K^2$ for all $\omega, \bar{\omega} \in \Omega$, and $\lim_{n\to\infty} \|f - f_n\|_k = 0$, $f_n(\omega) = \sum_{i=1}^{I_n} \alpha_{ni} k(\omega, \omega^{n\,i})$, $\alpha_{ni} \geqslant 0$, then $\|f\|_k \leqslant (K/\varepsilon)\|f\|_\infty$.

Proof. Since

$$f_n(\omega) = \sum_{i=1}^{I_n} \alpha_{ni} k(\omega, \omega^{n\,i}) \geqslant \varepsilon \sum_{i=1}^{I_n} \alpha_{ni}$$

then

$$\sum_{i=1}^{I_n} \alpha_{ni} \leqslant \|f_n\|_\infty / \varepsilon$$

and hence

$$\|f_n\|_k = \Bigg(\sum_{i,j=1}^{I_n} \alpha_{ni}\alpha_{nj} k(\omega^{n\,i}, \omega^{nj}) \Bigg)^{1/2} \leqslant K \sum_{i=1}^{I_n} \alpha_{ni} \leqslant \big(K/\varepsilon\big)\|f_n\|_\infty.$$

By Proposition 4 $\lim_{n\to\infty} \|f - f_n\|_k = 0$ implies $\lim_{n\to\infty} \|f - f_n\|_\infty = 0$, so

$$\|f\|_k = \lim_{n\to\infty} \|f_n\|_k \leqslant (K/\varepsilon) \lim_{n\to\infty} \|f_n\|_\infty = \big(K/\varepsilon\big)\|f\|_\infty.$$

3. Stochastic Optimization in Reproducing Kernel Hilbert Spaces

Stochastic optimization problem in decision rules has the form:

$$R(f) := E_\omega c(\omega, f(\omega)) := \int_\Omega c(\omega, f(\omega)) P(\mathrm{d}\omega) \to \min_{f(\cdot)\in F}, \tag{1}$$

where $c(\cdot\,,\cdot)\colon \Omega \times R^1 \to R^1$ is a nonnegative performance function; F is some functional class of feasible decision rules (strategies) $f\colon \Omega \to R^1$; $P(\cdot)$ is a probability distribution on the set $\Omega \subseteq R^n$; (Ω, B_Ω, P) is a probability space. Let $F^* \subset F$ denote a set of minimizers in (1). The distribution $P(\cdot)$ is unknown, only a set of independent observations $S_m = \{\omega^i \in \Omega\}_{\,i=1}^{\,m}$ is available. In what follows, the feasible set F be a closed convex subset of a RKHS with kernel $k\colon \Omega \times \Omega \to R^1$, for example, F can be a ball in H_k.

A possible interpretation of problem (1) would be as follows. Suppose that an expert has to prepare rules of behavior in an uncertain environment having only a probabilistic knowledge on the future states of the environment, namely he/she has to suggest in advance before the state ω of the environment is known a decision rule $f(\cdot)$ which is applied when the state ω is revealed. Optimal rule $f(\cdot)$ is found by optimization of the mean value $R(f) = Ec(\omega, f(\omega))$ of the performance criterion $c(\omega, f(\omega))$. Rules have to be selected from class F.

ASSUMPTION A (existence of solution).

(i) Either $F = H_k$ and problem (1) has at least one solution, or

(ii) $F = B_k = \{f \in H_k \colon \|f\|_k \leqslant C^*\}$.

ASSUMPTION B (properties of the performance function and the kernel). Assume that

(i) the performance function $c(\omega, \cdot)\colon R^1 \to R^1$ in (1) is nonnegative, convex and Lipschitzian with constant L uniformly in $\omega \in \Omega$; $c(\cdot, y)$ is continuous on Ω for any $y \in R^1$;

(ii) $\sup_{\omega\in\Omega} c(\omega, 0) \leqslant C < +\infty$;

(iii) $\sup_{\omega\in\Omega} |k(\omega, \omega)| \leqslant K^2 < \infty$.

Assumption B(i) covers piecewise linear and quadratic loss functions for bounded Ω. Requirement B(ii) is fulfilled if $c(\omega, 0)$ is continuous and Ω is bounded. Assumption B(iii) implies fulfillment of Proposition 4. Assumptions A(ii), B(i) jointly guarantee existence of solution in (1), see Ekeland and Temam (1976; Section 2.1).

Since in Assumptions A and B functions $f \in H_k(\Omega)$ and $c(\cdot\,, \cdot)\colon \Omega \times R^1 \to R^1$ are continuous, $c(\omega, f(\omega)) \leqslant C + LK\|f\|_k$, function $c(\omega, f(\omega))$ is measurable (even continuous) and bounded, thus functional $R(f)$ is well defined.

Since the distribution $P(\cdot)$ is not known and a decision rule problem is generally ill-posed, we apply regularization and instead of (1) solve the associated regularized sample average approximation problem:

$$\frac{1}{m}\sum_{i=1}^{m} c(\omega^i, f(\omega^i)) + \lambda\|f\|_k^2 \to \inf_{f(\cdot)\in F}, \tag{2}$$

where $\{\omega^i\}_{i=1}^m$ are independently sampled from $P(\cdot)$. If $c(\omega, \cdot)$ is convex then objective function in (2) is strictly convex with bounded level sets, so (2) has a unique solution $f_m^\lambda(\omega)$ in $F \subseteq H_k$. As in Assumptions A and B, solutions $f_m^\lambda(\cdot)$ continuously depend on the whole sample $\{\omega^i\}_{i=1}^m$, $f_m^\lambda(\cdot)$ can be considered random variables with values in H_k and defined on a countable product of the original probability space (Ω, B_Ω, P).

By the Representer Theorem (Kimeldorf and Wahba, 1970), see also Scholkopf and Smola (2002; Theorem 4.2, p. 90), a solution of problem (2) can be written in kernel form:

$$f_m^\lambda(\omega) = \sum_{j=1}^{m} \alpha_j k(\omega, \omega^j), \tag{3}$$

where coefficients $\alpha^m = \{\alpha_j\}_{j=1}^m$ are real numbers. Solutions of form (3) are called kernel minimizers. The Representer Theorem plays the key role in solution of problem (2), because it enables us, by substituting (3) into (2) and noting that $\|f_m^\lambda\|_k^2 = \sum_{i,j=1}^m \alpha_i\alpha_j k(\omega^i, \omega^j)$, to reduce the infinite dimensional optimization problem (2) to a finite dimensional one. In case $F = H_k$, problem (2) becomes

$$\frac{1}{m}\sum_{i=1}^{m} c\left(\omega^i, \sum_{j=1}^{m} \alpha_j k(\omega^i, \omega^j)\right) + \lambda \sum_{i,j=1}^{m} \alpha_i\alpha_j k(\omega^i, \omega^j) \to \min_{\{\alpha_1,\ldots,\alpha_m\}}. \tag{4}$$

Computational details for solving (4) for piece-wise linear function $c(\cdot\,,\cdot\,)$ can be found, for example, in Scholkopf and Smola (2002); Keyzer (2005). If $F = B_k$ in (2), then problem (4) is supplemented by an additional ball (quadratic) constraint $\sum_{i,j=1}^{m} \alpha_i \alpha_j k(\omega^i, \omega^j) \leqslant (C^*)^2$, which may be bounding or not. Mark that the presence of the ball constraint in (2) does not influence on the validity of the Representer Theorem.

Our first result establishes bounds on the expected functional value at the kernel minimizer f_m^λ. In what follows functional random variables f_m^λ are considered on a countable product of copies of the original probability space (Ω, B_Ω, P).

Theorem 1 (bounds on the expected functional value). *Under Assumptions* A(i), B, *for any* $\lambda > 0$ *and* m:

$$E_m R\big(f_m^\lambda\big) \leqslant R(f^*) + 2\frac{2C + L\big\|f^*\big\|_\infty}{\sqrt{m}} + \frac{LK(5LK + 2\sqrt{\lambda C})}{\lambda\sqrt{m}} + \lambda\big\|f^*\big\|_k^2, \quad (5)$$

where expectation E_m *is taken over i.i.d. sample* $\{\omega^1, \ldots, \omega^m\}$, f^* *is a solution of* (1).

The theorem guarantees convergence in mean of $R(f_m^\lambda)$ to the minimum value $R(f^*)$ if $\lambda(m) \to 0$ and $\sqrt{m}\lambda(m) \to 0$ as $m \to \infty$.

If Assumption A(ii) holds, i.e., $F = B_k$, the unknown constants $\|f^*\|_k$ and $\|f^*\|_\infty$ in (5) can be safely replaced by C^* and KC^*, respectively, while the term $\sqrt{\lambda C}$ in inequality (5) should be replaced by λC^*. In this case, the optimal regularization parameter, minimizing the right hand side of (5), becomes equal to $\lambda^*(m) = \frac{\sqrt{5}KL}{C^* \sqrt[4]{m}}$.

4. Strong Consistency (Convergence) of Kernel Minimizers

We are now ready to establish conditions of consistency of kernel solutions (3), expressed as convergence of f_m^λ to some $f^* \in F^*$, as $\lambda = \lambda(m) \to 0$ and $m \to \infty$, for appropriate specification of $\lambda = \lambda(m)$.

DEFINITION 3. Solution $f^* \in F^*$ is called a normal minimizer (in terms of Tihonov and Arsenin (1977) if it has a minimal norm, $\|f^*\|_k = \min_{f \in F^*} \|f\|_k$.

The following two theorems give sufficient conditions of uniform convergence of kernel minimizers $f_m^{\lambda(m)}$ to the normal minimizer $f^* \in F^*$ of (1), i.e., with probability one

$$\lim_{m\to\infty} \sup_{\omega\in\Omega} \big|f_m^{\lambda(m)}(\omega) - f^*(\omega)\big| = 0.$$

Theorem 2 (sufficient conditions for strong consistency of kernel minimizers). *Assume that Assumptions A, B hold and* $\lim_{m\to\infty} \lambda(m) = 0$.

If $\lim_{m\to\infty} m\lambda^2(m)\big/ \ln m = \infty$, *then* $\lim_{m\to\infty} R(f_m^{\lambda(m)}) = R(f^*)$ *with probability one.*

If $\lim_{m\to\infty} m\lambda^4(m)/\ln m = \infty$, *then* $R(f_m^{\lambda(m)}) \to R(f^*)$ *and minimizers* $f_m^{\lambda(m)}$ *of* (2) *uniformly converge to the normal minimizer* f^* *of* (1), *with probability one as* $m \to +\infty$.

Remark that for ill-posed problems relation $\lim_{m\to\infty} R(f_m^{\lambda(m)}) = R(f^*)$ generally does not imply convergence of estimators, $f_m^{\lambda(m)} \to F^*$. Thus, to guarantee consistency of kernel solutions regularization parameter $\lambda = \lambda(m)$ in expressions (2), (4) should go to zero slower than $\sqrt[4]{(\ln m)/m}$ as $m \to +\infty$.

Theorem 3 (strong uniform consistency of kernel minimizers). *Assume that Assumptions* A(i), B *hold. For regularization parameters* $\lambda(m) = \Lambda(\ln m)^\varepsilon/m^{1/4}$, $\Lambda > 0$, $1/4 < \varepsilon \leqslant 1$, *error term* $[R(f_m^{\lambda(m)}) - R(f^*)]$ *tends to zero and corresponding solutions* $f_m^{\lambda(m)}$ *uniformly in* $\omega \in \Omega$ *converge to the normal minimizer* f^*, *with probability one as* $m \to +\infty$. *For the mean error the following estimate holds true*

$$E_m R\big(f_m^{\lambda(m)}\big) - R(f^*) \leqslant \frac{4C + 2L\|f^*\|_\infty}{\sqrt{m}} + \frac{LK(5LK + 2\sqrt{2\Lambda C})}{\Lambda(\ln m)^\varepsilon \sqrt[4]{m}} + \\ + \frac{\|f^*\|_k^2 \Lambda(\ln m)^\varepsilon}{\sqrt[4]{m}}. \tag{6}$$

Error estimates (6) show that the expected error asymptotically goes to zero not slower than $const \cdot (\ln m)^\varepsilon/m^{1/4}$, $1/4 < \varepsilon \leqslant 1$.

To obtain a similar estimate under Assumption A(ii), one has in inequality (6) to replace term $\sqrt{2\Lambda C}$ by $2\Lambda C^*$, while the unknown constants $\|f^*\|_k$ and $\|f^*\|_\infty$ in (6) can be safely replaced by C^* and KC^*, respectively.

5. Applications to Statistical Learning

Since functionals used in the statistical learning literature are particular cases of form (1), the results of previous sections can be applied to kernel regression and classification.

5.1. *Kernel Learning for a Regression Function*

Particular cases of problem (1) arise in mathematical statistics. Suppose that problem (1) has the following structure:

$$R(f) := E_{\omega_1,\omega_2} c\big(\omega_2, f(\omega_1)\big) \\ = \int\limits_{\Omega_1 \times R^1} c\big(\omega_2, f(\omega_1)\big) p(\omega_1, \omega_2) \mathrm{d}\omega_1 \mathrm{d}\omega_2 \to \min_{f(\cdot)\in F(\Omega_1)},$$

where $\omega = (\omega_1 \in \Omega_1 \subseteq R^n, \omega_2 \in R^1)$; $c(\cdot\,,\cdot)$: $\Omega_2 \times R^1 \to R^1$; $f(\cdot)$: $\Omega_1 \to R^1$; $p(\omega_1, \omega_2)$ is a density on $\Omega_1 \times R^1$; $F(\Omega_1)$ is a set of measurable functions defined on

a compact domain Ω_1. It is well known (see Gyorfi *et al.*, 2002) that (1) achieves its unconstrained minimum over measurable dependences $f(\omega_1)$ and the quadratic loss function

$$c(\omega_2, f(\omega_1)) = (\omega_2 - f(\omega_1))^2 \tag{7}$$

at the conditional mean of the density $p(\omega_1, \omega_2)$ under fixed ω_1:

$$f^*(\omega_1) = \int\limits_{R^1} \omega_2 p(\omega_1, \omega_2) \mathrm{d}\omega_2 \Big/ \int\limits_{R^1} p(\omega_1, \omega_2) \mathrm{d}\omega_2.$$

The median and other ν-quantiles $f_\nu^*(\omega_1)$ of the conditional distribution function

$$\Phi_{\omega_1}(t) = \int\limits_{-\infty}^{t} p(\omega_1, \omega_2) \mathrm{d}\omega_2 \Big/ \int\limits_{-\infty}^{+\infty} p(\omega_1, \omega_2) \mathrm{d}\omega_2$$

can be found by solving (1) with the integrand

$$c(\omega_2, f(\omega_1)) = \max\big\{(1-\nu)(f(\omega_1) - \omega_2), \nu\big(\omega_2 - f(\omega_1)\big)\big\}, \tag{8}$$

$0 < \nu < 1$, see for example, Koenker and Bassett (1978); Ermolaev and Yastremski (1979; p. 95); Ermolaen and Leonardi (1982); Ruszczynski and Shapiro (2003; p. 2); Takeuchi *et al.* (2006). The median is obtained for $\nu = 1/2$. Solutions $f^*(\omega_1)$, $f_\nu^*(\omega_1)$ are not necessarily general measurable functions, they may be continuous, smooth and so on. Hence constrained minimization in (1) with loss functions (7), (8) for sufficiently broad class F leads to the same solution. Of course, it is not straightforward to check whether a conditional mean and quantiles of a given distribution belong to a particular RKHS, or to postulate a suitable RKHS containing these characteristics of the distribution. This actually is one of the main problems in kernel learning theory, commonly referred to as the problem of "learning the kernel". In practice, it is solved by selection of the kernel from some parametric family $\{k_\theta,\ \theta \in \Theta\}$ (jointly with the value of the regularization parameter λ) so as to provide the best performance on the test data of the regression rule $f_m^{\lambda,\theta}(\cdot)$ identified on training data. For example, the family $\{k_\theta,\ \theta \in \Theta\}$ can be a convex combination of several different kernels.

Under Assumptions A(i) and $\lim_{m\to\infty} m\lambda^4(m)/\ln m = \infty$ kernel minimizers for (7) and (8) by Theorem 2 uniformly converge with probability one to the mean regression and quantile regression functions respectively.

Rate of convergence of regression estimators is given by (5), (6). The right hand side of (5), (6) as a function of the sample size m is called an enveloping learning curve; under optimal regularization parameter it is a second order polynom of $(1/\sqrt[4]{m})$ with unknown coefficients. The curve and hence unknowns $\|f^*\|_\infty$, $\|f^*\|_k$ can be statistically estimated. In Brumen *et al.* (2007) the learning curve is first estimated for relatively small m, and subsequently used for performance assessment for large m. It is seen from (5), (6) that

norm $\|f^*\|_\infty$ and especially the unknown norm $\|f^*\|_k$ of the exact minimizer f^* are important (unknown) characteristics of the problem under consideration. Since space H_k is the Reproducing Kernel Hilbert Space constructed on the basis of kernel $k(\cdot\,,\cdot\,)$, by Proposition 4 we have $\|f^*\|_k \geqslant \|f^*\|_\infty/K$, and since the lower the norm $\|f^*\|_k$ the tighter the bounds (5), (6), we may conclude that the kernel has to be selected in such a way that this norm multiplied by K becomes minimal. Clearly, as f^* is unknown, this cannot be done exactly but the result confirms that finding a kernel that closely fits a particular class of functions may be important.

5.2. *Consistency of Kernel Binary Classifiers*

A binary classification problem can be reduced to the risk minimization problem (1) with the absolute deviation loss function $c(\omega_2, f(\omega_1)) = |f(\omega_1) - \omega_2|$ as follows, see Devroye *et al.* (1996; p. 20). For any given function $f(\omega_1)$ a binary classification rule is defined as

$$g_f(\omega_1) = \begin{cases} 1, & f(\omega_1) > 1/2, \\ 0, & \text{otherwise.} \end{cases}$$

Quality of this decision rule g_f can be measured by Bayesian risk, the probability $P\{g_f(\omega_1) \neq \omega_2\}$ of misclassification, where $\omega_2 \in \{0, 1\}$. It is well known that Bayesian risk achieves its minimal value P^* on the decision rule g_η defined by a conditional probability function $\eta(\omega_1) = P\{\omega_2 = 1|\omega_1\}$, which is, however, unknown. Hence decision function $f(\omega_1)$ has to be searched in some class of functions H, for example, in some Reproducing Kernel Hilbet Space H_k. If $g_\eta(\omega_1) \in H_k$ then the following relation holds true

$$P\big\{g_f(\omega_1) \neq \omega_2\big\} - P^* \leqslant 2\Big(E\big|f(\omega_1) - \omega_2\big| - \min_{f \in H_k} E\big|f(\omega_1) - \omega_2\big|\Big). \tag{9}$$

In the statistical learning literature, decision function is found by solving regularized empirical risk minimization problem (2) with absolute deviation loss function $c(\omega_2, f(\omega_1)) = |f(\omega_1) - \omega_2|$. Consequently, solution $f(\omega_1) = f_m^\lambda(\omega_1)$ and corresponding kernel binary classifier

$$g_m^\lambda(\omega_1) := g_{f_m^\lambda}(\omega_1) = \begin{cases} 1, & f_m^\lambda(\omega_1) > 1/2, \\ 0, & \text{otherwise,} \end{cases} \tag{10}$$

satisfy relation (9).

If $\lambda(m)$ is such that $\lim_{m\to\infty} m\lambda^2(m)\big/\ln m = \infty$, then under Assumptions A(i), by (9) and Theorem 2 kernel binary classifier (10) is strongly consistent, i.e.,

$$\lim_{m\to\infty} P\big\{g_m^{\lambda(m)}(\omega_1) \neq \omega_2\big\} = P^* \tag{11}$$

with probability one. The rate of convergence of $E_m P\{g_m^{\lambda(m)}(\omega_1) \neq \omega_2\}$ to P^* in view of (9) follows from Theorems 2, 3.

Another approach to binary classification, based on direct Byaesian approximation of $\eta(\omega_1)$, is presented in Sergienko *et al.* (2008), where a comparison is also given with empirical risk minimization.

6. Conclusions

We have presented and studied a computational framework for solving a certain class of stochastic optimization problems in (infinite dimensional) Reproducing Kernel Hilbert Spaces (RKHS), namely for finding optimal decision rules (or strategies) under probabilistic uncertainty. This framework stems from statistical learning theory, where it is known as Support Vector Method/Machine (SVM). The framework combines sample average approximation of the expectation performance criterion with Tihonov's regularization and reduces the problem to solution of a finite dimensional convex optimization. The paper studies asymptotic properties of approximate solutions for sample size growing to infinity. It establishes sufficient conditions and a rule for downward reduction of the regularization term, which guarantee almost sure convergence of the sequence of minimizers and indicate the rate of convergence in mean. In the best case this rate is inversely proportional to the fourth power root of the sample size. These convergence results directly apply to Support Vector Method/Machine in statistical learning applications.

Appendix: Proofs

Vapnik's (1998) approach to the analysis of convergence of empirical risk minimizers f_m^λ of (2) to a true risk minimizer f^* of (1) consists of establishing conditions for uniform convergence of the empirical risk functional $R_m(f)$ to the true risk functional $R(f)$, $f \in F$. Specifically, define random variables

$$\delta_m = \sup_{f\in F} \big|R_m(f) - R(f)\big| = \sup_{f\in F} \left| \frac{1}{m} \sum_{i=1}^{m} c\big(\omega^i, f(\omega^i)\big) - E_\omega c\big(\omega, f(\omega)\big) \right|.$$

If $\delta_m/\lambda(m) \to 0$ in some probabilistic sense and $\lambda(m) \to 0$, then solutions $f_m^{\lambda(m)}$ of regularized problem (2) converge in the same probabilistic sense to the normal minimizer f^* in the strong topology of the corresponding Hilbert space, see Tihonov and Arsenin (1977) and Vapnik (1998). We remark that the uniform convergence of functions $R_m(\cdot) \to R(\cdot)$ is in general not necessary for convergence of minimizers, as it is sufficient epi-convergence of $R_m(\cdot)$ to $R(\cdot)$ (Rockafellar and Wets, 1998).

In our analysis we follow a different approach based on stability properties of regularized minimizers. Bousquet and Elisseeff (2002) and Mukherjee *et al.* (2006) highlight the importance of stability for the consistency of learning algorithms. We obtain the convergence results by enclosing within probabilistic bounds the deviation of risk $R\,(f) = E_\omega c(\omega, f(\omega))$ at a single point $f = f_m^{\lambda(m)}$ from its minimal possible value

$R(f^*) = \inf_{f\in F} R(f)$, and by ensuring that these bounds become (exponentially) tighter with rising number of observations, hence deriving $\lim_{m\to\infty}(R(f_m^{\lambda(m)}) - R(f^*)) = 0$ with probability one. Since in parallel with $R(f)$ a norm $\|f\|_k$ is minimized, we obtain convergence of kernel minimizers $f_m^{\lambda(m)}$ to the normal minimizer of problem (1). Uniform convergence of $f_m^{\lambda(m)}$ follows from properties of RKHS with bounded kernel (Proposition 4).

Remark that the related papers by Bosquet and Elisseeff (2002), Takeuchi *et al.* (2006), Scholkopf and Smola (2002, Section 12.1) Scholkopf and Smola (2002; Section 12.1) study convergence of $R(f_m^\lambda) + \lambda\|f_m^\lambda\|^2$ to $\min_f[R(f) + \lambda\|f\|^2]$ under fixed λ, and do not discuss the issue of convergence of f_m^λ to some solution f^* as $\lambda \to 0$.

A1. Bounds

First we prove a number of lemmas.

Lemma 1 (bounds on $\|f_m^\lambda\|_k$ and $\|f_m^\lambda\|_\infty = \sup_{\omega\in\Omega}|f_m^\lambda(\omega)|$). *Under Assumptions A(i), B the following inequalities hold*:

$$\big\|f_m^\lambda\big\|_k^2 \leqslant C/\lambda, \quad \big\|f_m^\lambda\big\|_\infty^2 \leqslant K^2\big\|f_m^\lambda\big\|_k^2 \leqslant K^2C/\lambda.$$

Proof. By optimality of f_m^λ for (3) and by B(i), B(ii), we have

$$\begin{aligned}\lambda\big\|f_m^\lambda\big\|_k^2 &\leqslant \frac{1}{m}\sum_{i=1}^m c]\big(\omega^i, f_m^\lambda(\omega^i)\big) + \lambda\big\|f_m^\lambda\big\|_k^2 \\ &\leqslant \frac{1}{m}\sum_{i=1}^m c(\omega^i, 0) \leqslant \sup_{\omega\in\Omega} c(\omega, 0) \leqslant C. \end{aligned} \tag{12}$$

By the Cauchy–Schwarz inequality and B(iii), we have for any ω

$$\big|f_m^\lambda(\omega)\big| = \big|\big\langle f_m^\lambda(\cdot), k(\omega,\cdot)\big\rangle\big| \leqslant \sqrt{k(\omega,\omega)}\big\|f_m^\lambda\big\|_k \leqslant K\big\|f_m^\lambda\big\|_k. \tag{13}$$

Hence,

$$\big\|f_m^\lambda\big\|_\infty^2 \leqslant K^2\big\|f_m^\lambda\big\|_k^2 \leqslant K^2C/\lambda,$$

where we note that inequalities (12) and (13) can be found in Bousquet and Elisseeff (2002).

Remark that under Assumptions A(ii), B(iii) $\|f_m^\lambda\|_k \leqslant C^*$ and hence by Proposition 4, $\|f_m^\lambda\|_\infty \leqslant KC^*$ holds.

Lemma 2 (bounds on the loss function at minimizers f^* and f_m^λ). *Under Assumptions A(i), B the following bounds hold true*:

$$\begin{aligned}
&\sup_{\omega\in\Omega} c\big(\omega, f^*(\omega)\big) \leqslant C + L\big\|f^*\big\|_\infty =: N,\\
&\sup_{\omega\in\Omega} c\big(\omega, f_m^\lambda(\omega)\big) \leqslant C + LK\sqrt{C/\lambda} =: M(\lambda),\\
&\lambda M(\lambda) \leqslant \Lambda M(\Lambda) =: \bar{M}(\Lambda), \quad \textit{for any } \lambda, \quad 0 < \lambda \leqslant \Lambda.
\end{aligned} \tag{14}$$

Proof. For N, by B(i) we have:

$$\sup_{\omega\in\Omega} c\big(\omega, f^*(\omega)\big) \leqslant \max_{\omega\in\Omega} c(\omega, 0) + L\,\|f^*\|_\infty \leqslant C + L\,\|f^*\|_\infty = N,$$

where $\|f^*\|_\infty = \sup_{\omega\in\Omega} |f^*(\omega)|$. For $M(\lambda)$, by B(i) and Lemma 1 we have:

$$\sup_{\omega\in\Omega} c\big(\omega, f_m^\lambda(\omega)\big) \leqslant \max_{\omega\in\Omega} \Big(c(\omega, 0) + L\big|f_m^\lambda(\omega)\big|\Big) \leqslant C + LK\sqrt{C/\lambda} = M(\lambda).$$

Hence, for any $\lambda \in (0, \Lambda]$,

$$\begin{aligned}
\lambda M(\lambda) &\leqslant \lambda\big(C + LK\sqrt{C/\lambda}\big) = \lambda C + LK\sqrt{\lambda C}\\
&\leqslant \big(\Lambda C + LK\sqrt{\Lambda C}\big) = \Lambda M(\Lambda) = \bar{M}(\Lambda).
\end{aligned}$$

Remark that although $M(\lambda)$ may be unbounded as $\lambda \to 0$ but it enters into probabilistic bound (18) below in the combination $\lambda M(\lambda)$, which is bounded by (14). This is a key observation for obtaining convergence results from concentration measure inequality (18) under $\lambda \to 0$.

Under Assumptions A(ii), B $\sup_{\omega\in\Omega} c(\omega, f_m^\lambda(\omega)) \leqslant C + LKC^*$.

Lemma 3 (basic inequalities). *For any $\lambda > 0$, m and solutions f^* of* (1)*, f_m^λ of* (2) *the following inequalities hold*:

$$R(f_m^\lambda) \leqslant R\,(f^*) + \Delta_m^\lambda - \Delta_m^* + \lambda\,\|f^*\|_k^2\,, \tag{15}$$

$$\begin{aligned}
\big\|f_m^\lambda\big\|_k^2 &\leqslant \frac{1}{\lambda}\Big[R(f^*) - R(f_m^\lambda) + \Delta_m^\lambda - \Delta_m^* + \lambda\|f^*\|_k^2\Big]\\
&\leqslant \frac{1}{\lambda}\big(\Delta_m^\lambda - \Delta_m^*\big) + \big\|f^*\big\|_k^2,
\end{aligned} \tag{16}$$

where $\Delta_m^\lambda = [R(f_m^\lambda) - R_m(f_m^\lambda)]$ and $\Delta_m^* = [R(f^*) - R_m(f^*)]$, decompose the error in prediction of risk;

$$R(f) = E_\omega c\big(\omega, f(\omega)\big), \quad R_m(f) = \frac{1}{m}\sum_{i=1}^m c\big(\omega^i, f(\omega^i)\big).$$

Proof. By optimality of f_m^λ,

$$R_m^\lambda(f_m^\lambda) = R_m(f_m^\lambda) + \lambda\|f_m^\lambda\|_k^2 \leqslant R_m^\lambda(f^*) = R_m(f^*) + \lambda\|f^*\|_k^2$$

and hence

$$\begin{aligned} R(f_m^\lambda) + \lambda\|f_m^\lambda\|_k^2 &\leqslant R(f^*) + \big[R(f_m^\lambda) - R_m(f_m^\lambda)\big] \\ &\quad + \big[R_m(f_m^\lambda) - R(f^*)\big] + \lambda\|f_m^\lambda\|_k^2 \\ &\leqslant R(f^*) + \big[R(f_m^\lambda) - R_m(f_m^\lambda)\big] \\ &\quad + \big[R_m(f^*) - R(f^*)\big] + \lambda\|f^*\|_k^2. \end{aligned} \tag{17}$$

Inequality (15) is obtained by dropping the regularization term on the left hand side of (17) and the second inequality (16) follows from (17), noting that $R(f^*) - R(f_m^\lambda) \leqslant 0$.

Next Lemmas 4 and 5 give exponential bounds on distributions of Δ_m^λ and Δ_m^*.

Lemma 4 (bound on $P_m\{|\Delta_m^\lambda| - \beta_m^\lambda > \varepsilon\}$). *In Assumptions A, B and bound $M(\lambda) \geqslant c(\omega^i, f(\omega^i))$, for any $\varepsilon > 0$ we have*:

$$P_m\big\{|\Delta_m^\lambda| - \beta_m^\lambda > \varepsilon\big\} \leqslant 2\exp\big(-\eta_m^\lambda \varepsilon^2\big), \tag{18}$$

where $\beta_m^\lambda = \frac{L^2K^2}{m\lambda}$ and $\eta_m^\lambda = \frac{2m\lambda^2}{(2L^2K^2+\lambda M(\lambda))^2}$; probability measure P_m is a product of m copies of the original data generation distribution P.

Proof. See Bousquet and Elisseeff (2002) and also Scholkopf and Smola (2002; Theorem 12.5, p. 363). In the last reference there is an additional multiplier 1/2 (not present in our formulation) in the regularization terms and concentration measure inequalities are applied to two side deviations (as in our case). The proof is based on establishing uniform stability of minimizer $f_m^\lambda \in F$ with respect to any single observation ω^i (Bousquet and Elisseeff, 2002; Theorem 22) and on application of the concentration measure exponential inequality (McDiarmid, 1989).

Lemma 5 (bound on $P_m\{|\Delta_m^*| \geqslant \varepsilon\}$). *Given a uniform upper bound N on $c(\omega, f^*(\omega))$, and supposing that random variables $c(\omega^i, f^*(\omega^i))$ are i.i.d., we have by Hoeffding's inequality, see Devroye et al. (1996; p. 122),*

$$P_m\big\{|\Delta_m^*| \geqslant \varepsilon\big\} \leqslant 2\exp\left\{-\frac{2m}{N^2}\varepsilon^2\right\}. \tag{19}$$

Next, we present bounds on the expectation of the absolute value of Δ_m^λ and Δ_m^*, to be used in subsequent proofs. Symbol E_m below denotes the expectation over probability measure P_m, i.e., expectation over i.i.d data sample $\{\omega^1, \ldots, \omega^m\}$.

Lemma 6 (bound on $E_m|\Delta_m^\lambda|$). *From Assumptions B follows*:

$$E_m\big|\Delta_m^\lambda\big| \leqslant \frac{1}{\lambda\sqrt{m}}\left(2\big(2L^2K^2+\lambda M(\lambda)\big)+\frac{L^2K^2}{\sqrt{m}}\right).$$

Proof. Define $F_m^\lambda(\varepsilon) = \max\{0, 1 - 2\exp\{-\eta_m^\lambda\varepsilon^2\}\}$. Obviously, by Lemma 4 we have $P_m\{|\Delta_m^\lambda| - \beta_m^\lambda < \varepsilon\} \geqslant F_m^\lambda(\varepsilon)$. Denote by $\bar\varepsilon$ a solution of the equation $2\exp\{-\eta_m^\lambda\bar\varepsilon^2\}= 1$ (hence $\bar\varepsilon \leqslant 1\Big/\sqrt{\eta_m^\lambda}$). Then,

$$\begin{aligned}
E_m\big|\Delta_m^\lambda\big| - \beta_m^\lambda &\leqslant \int\limits_0^{+\infty}\big(1-F_m^\lambda(\varepsilon)\big)\,\mathrm{d}\varepsilon = \bar\varepsilon + 2\int\limits_{\bar\varepsilon}^{+\infty}\exp\big\{-\eta_m^\lambda\varepsilon^2\big\}\mathrm{d}\varepsilon \\
&\leqslant \bar\varepsilon + 2\int\limits_0^{+\infty}\exp\big\{-\eta_m^\lambda\varepsilon^2\big\}\mathrm{d}\varepsilon \leqslant \bar\varepsilon + \sqrt{\pi}\frac{1}{\sqrt{\eta_m^\lambda}} \\
&\leqslant \big(1+\sqrt{\pi}\big)\frac{1}{\sqrt{\eta_m^\lambda}} \leqslant \frac{2(2L^2K^2+\lambda M(\lambda))}{\lambda\sqrt{m}}.
\end{aligned}$$

Accounting for $\beta_m^\lambda = \frac{L^2K^2}{m\lambda}$ completes the proof.

Lemma 7 (bound on $E_m|\Delta_m^*|$). *If* $\sup_{\omega\in\Omega} c(\omega, f^*(\omega)) \leqslant N$, *then* $E_m|\Delta_m^*| \leqslant 2N\sqrt{m}$.

Proof. Define $F_m^*(\varepsilon) = \max\{0, 1 - 2\exp\{-\eta_m^*\varepsilon^2\}\}$, $\eta_m^* = 2m/N^2$. By Lemma 5, $P_m\{|\Delta_m^*| < \varepsilon\} \geqslant F_m^*(\varepsilon)$. Find $\bar\varepsilon$ such that $2\exp\{-\eta_m^*\bar\varepsilon^2\} = 1$, then $\bar\varepsilon \leqslant 1/\sqrt{\eta_m^*}$. Therefore,

$$E_m\big|\Delta_m^*\big| \leqslant \int\limits_0^{+\infty}\big(1-F_m^*(\varepsilon)\big)\mathrm{d}\varepsilon \leqslant \big(1+\sqrt{\pi}\big)\frac{1}{\sqrt{\eta_m^*}} \leqslant \frac{1+\sqrt{\pi}}{\sqrt{2}}\frac{N}{\sqrt{m}} \leqslant \frac{2N}{\sqrt{m}}.$$

Lemmas 6 and 7 enable us to obtain a bound on the expected value of the empirical risk $R(f_m^\lambda)$.

Proof of Theorem 1 (bounds on the expected functional value). As from Lemma 3 it follows that

$$E_mR(f_m^\lambda) \leqslant R\big(f^*\big) + E_m\big|\Delta_m^*\big| + E_m\big|\Delta_m^\lambda\big| + \lambda\big\|f^*\big\|_k^2.$$

Substitution of $E_m\,|\Delta_m^*|$, $E_m\big|\Delta_m^\lambda\big|$, from Lemmas 6, 7 gives the estimates:

$$\begin{aligned}
E_mR\big(f_m^\lambda\big) \leqslant\ & R(f^*) + 2\frac{C + L\,\|f^*\|_\infty}{\sqrt{m}} \\
&+ \frac{1}{\lambda\sqrt{m}}\left(2\big(2L^2K^2+\lambda M(\lambda)\big)+\frac{L^2K^2}{\sqrt{m}}\right) + \lambda\big\|f^*\big\|_k^2
\end{aligned}$$

$$\begin{aligned}
&\leqslant R(f^*) + 2\frac{C + L\,\|f^*\|_\infty}{\sqrt{m}} \\
&\quad + \frac{4L^2K^2 + 2(\lambda C + LK\sqrt{\lambda C})}{\lambda\sqrt{m}} + \frac{L^2K^2}{\lambda m} + \lambda\,\|f^*\|_k^2 \\
&\leqslant R(f^*) + 2\frac{2C + L\,\|f^*\|_\infty}{\sqrt{m}} + \frac{LK(5LK + 2\sqrt{\lambda C})}{\lambda\sqrt{m}} + \lambda\,\|f^*\|_k^2 .
\end{aligned}$$

To obtain the estimate under Assumption A(ii) one has to replace in the last inequality term $\sqrt{\lambda C}$ by λC^*.

A2. Proof of Consistency of Kernel Learning Estimators

Proof of Theorem 2 (sufficient conditions of consistency for kernel minimizers). We start by proving convergence (a.s.) of the risk criterion $R(f_m^{\lambda(m)}) \to R(f^*)$. Next, although deterministic bounds on $\|f_m^\lambda\|_k$ in Lemma 1 go to infinity as $\lambda \to 0$, random quantities $\|f_m^\lambda\|_k$ under condition $\lim_{m\to\infty} m\lambda^2(m)/\ln m = \infty$ are bounded in probability by (16), (21), (22) and even appear a.s. converging to $\|f^*\|$ if $\lim_{m\to\infty} m\lambda^4(m)/\ln m = \infty$. And finally, we invoke the convexity of $R(\cdot)$ in Hilbert space H_k and the uniqueness of the normal minimizer f^* to prove consistency of the kernel minimizers.

Fix any minimizer f^* of (1). From Lemma 3, we have estimates (15), (16). By Lemmas 4 and 5 for distributions of Δ_m^λ, Δ_m^* we have exponential bounds (18), (19). Since $\lim_{m\to\infty} \lambda(m) = 0$, then $\lambda(m) \leqslant \Lambda < \infty$. For $\lambda \leqslant \Lambda$ by (14), we have $\lambda M(\lambda) \leqslant \bar{M}(\Lambda)$ and hence

$$P_m\big\{\big|\Delta_m^\lambda\big| - \beta_m^\lambda > \varepsilon\big\} \leqslant 2\exp\big(-\bar{\eta}_m^\lambda \varepsilon^2\big), \tag{20}$$

where $\bar{\eta}_m^\lambda = \frac{2m\lambda^2}{(2L^2K^2 + M(\Lambda))^2}$. Under condition $\lim_{m\to\infty} m\lambda^2(m)/\ln m = \infty$ it takes place $\lim_{m\to\infty} \beta_m^\lambda = 0$ and for any $\varepsilon > 0$ holds

$$\begin{aligned}
&\sum_{m=1}^{\infty} P_m\Big\{\big|\,|\Delta_m^{\lambda(m)}| - \beta_m^{\lambda(m)}\big| > \varepsilon\Big\} \leqslant 2\sum_{m=1}^{\infty} \exp\big(-\bar{\eta}_m^{\lambda(m)}\varepsilon^2\big) < \infty, \\
&\sum_{m=1}^{\infty} P_m\big\{\big|\Delta_m^*\big| \geqslant \varepsilon\big\} \leqslant 2\sum_{m=1}^{\infty} \exp\Big\{-\frac{2m}{N^2}\varepsilon^2\Big\} < \infty.
\end{aligned}$$

From here by a criterion of almost sure convergence it follows $\lim_{m\to\infty} |\Delta_m^{\lambda(m)}| = \lim_{m\to\infty} |\Delta_m^*| = 0$ with probability one, which together with assumption $\lim_{m\to\infty} \lambda(m) = 0$ and relation (15) proves the first statement of the theorem, $\lim_{m\to\infty} R(f_m^{\lambda(m)}) = R(f^*)$ a.s.

Let us prove the second statement of the theorem. Let f^* be now the normal minimizer of (1). For Δ_m^λ/λ and Δ_m^*/λ we have due to (19), (20) the following estimates of

distributions:

$$P_m\{|\Delta_m^\lambda/\lambda| - \beta_m^\lambda/\lambda > \varepsilon\} \leqslant 2\exp\big(-\bar{\eta}_m^\lambda \lambda^2 \varepsilon^2\big), \tag{21}$$

$$P_m\{|\Delta_m^*/\lambda| \geqslant \varepsilon\} \leqslant 2\exp\Big\{-\frac{2m\lambda^2}{N^2}\varepsilon^2\Big\}. \tag{22}$$

Since under condition $\lim_{m\to\infty} m\lambda^4(m)/\ln m = \infty$ both

$$\sum_{m=1}^{\infty} \exp\big(-\bar{\eta}_m^\lambda \lambda^2(m)\varepsilon^2\big) < \infty$$

and

$$\sum_{m=1}^{\infty} \exp\big\{-2\varepsilon^2 m\lambda^2(m)/N^2\big\} < \infty.$$

It follows from (21), (22), by the criterion of almost sure convergence, that $\Delta_m^{\lambda(m)}\Big/\lambda(m)$ and $\Delta_m^*/\lambda(m)$ converge to zero with probability one as $m \to \infty$; and that they are bounded with probability one. Then, since $\lim_{m\to\infty}\lambda(m) = 0$, it follows also that $\lim_{m\to\infty}|\Delta_m^{\lambda(m)}| = \lim_{m\to\infty}|\Delta_m^*| = 0$ with probability one.

From this follows, by inequalities (15), (16), that with probability one, we have:

$$\lim_{m\to+\infty} R\big(f_m^{\lambda(m)}\big) = R\big(f^*\big),$$

$$\Big(\limsup_{m\to\infty} \big\|f_m^{\lambda(m)}\big\|_k\Big)^2 = \limsup_{m\to\infty} \big\|f_m^{\lambda(m)}\big\|_k^2$$

$$\leqslant \limsup_{m\to\infty} \Big(\big|\Delta_m^{\lambda(m)}\big|\,/\,\lambda(m) + |\Delta_m^*|/\lambda(m) + \big\|f^*\big\|_k^2\Big) = \big\|f^*\big\|_k^2,$$

and thus $\limsup_{m\to\infty}\|f_m^{\lambda(m)}\|_k \leqslant \big\|f^*\big\|_k$. Hence, because of the convexity of functional $R(\cdot)$, sequence $\{f_m^{\lambda(m)}\}$ a.s. converges to the solution set F^* of (1) in weak topology of the space H_k, (Ekeland and Temam, 1976; Section 2.1). By properties of weak topology convergence in Hilbert spaces for the normal minimizer f^* holds $\liminf_{m\to\infty}\|f_m^{\lambda(m)}\|_k \geqslant \|f^*\|_k$ and, therefore, $\lim_{m\to\infty}\|f_m^{\lambda(m)}\|_k = \|f^*\|_k$. Since $\{f_m^{\lambda(m)}\}$ converges to f^* in weak topology and norms are convergent, i.e., $\lim_{m\to\infty}\|f_m^{\lambda(m)}\|_k = \|f^*\|_k$, it follows from properties of Hilbert spaces that sequence $\{f_m^{\lambda(m)}\}$ strongly converges (in norm) to f^* as $m \to \infty$, see Kolmogorov and Fomin (1981; Ch. IV, §3, Sec. 2). Finally in RKHS with bounded kernel (Assumption B(iii)) convergence in norm implies the uniform one, by Proposition 4.

Proof of Theorem 3 (strong uniform consistency of kernel minimizers). The first two statements of the theorem follow from Theorem 2 and the third follows from Theorem 1 accounting for the bound $\lambda(m) = \Lambda(\ln m)^\varepsilon/\sqrt[4]{m} \leqslant 2\Lambda$.

References

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of American Mathematical Society*, **68**, 337–404.

Bousquet, O., and A. Elisseeff (2002). Stability and generalization. *Journal of Machine Learning Research*, **2**, 499–526.

Brumen B., M.B. Juric, T. Welzer, I. Rozman, H. Jaakkola and A. Papadopoulos (2007). Assessment of classification models with small amounts of data. *Informatica*, **18**(3), 343–362.

Chancelier, J.-Ph., and SOWG (2006). Epi-convergence of stochastic optimization problems involving both random variables and measurability constraints approximations. *Rapport de recherche du CERMICS* 2006-323, 1–14 (`http://cermics.enpc.fr/reports`).

Cucker, F., and S. Smale (2001). On the mathematial foundations of learning. *Bull. Amer. Math. Soc.*, **89**(1), 1–49.

Cucker, F., and S. Smale (2002). Best choices for regularization parameters in learning theory: on the bias-variance problem. *Found. Comput. Math.*, **2**(4), 413–428.

De Vito, E., A. Caponnetto and L. Rosasco (2005). Model selection for regularized least-squares algorithm in learning theory. *Found. Comput. Math*., **5**(1), 59–85.

Devroye, L., L. Gyorfi and G. Lugosi (1996). *A Probabilistic Theory of Pattern Recognition.* Springer-Verlag, New York.

Ekeland, I., and R. Temam (1976). *Convex Analysis and Variational Problems*. North Holland & American Elsevier, Amsterdam, Oxford, New York.

Ermoliev, Y.M. (1976). *Methods of Stochastic Programming*. Nauka, Moscow (in Russian).

Ermoliev, Y.M., and G. Leonardi (1982). Some proposals for stochastic facility location models. *Mathematical Modelling*, **3**, 407–420.

Ermoliev, Y.M., and A.I. Yastremski (1979). *Stochastic Models and Methods in Economic Planning*. Nauka, Moscow (in Russian).

Gyorfi, L., M. Kohler, A. Krzyzak and H. Walk (2002). *A Distribution Free Theory of Nonparametric Regression.* Springer-Verlag, New York, Berlin, Heidelberg.

Keyzer, M.A. (2005). Rule-based and support vector (SV-) regression/classification algorithms for joint processing of census, map, survey and district data. Working Paper WP-05-01. Centre for World Food Studies, Amsterdam.

Kimeldorf, G.S., and G. Wahba (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Stat*., **45**, 495–502.

Koenker, R., and G.W. Bassett (1978). Regression quantiles. *Econometrica*, **46**, 33–50.

Kolmogorov, A.N., and S.V. Fomin (1981). *Elements of the Theory of Functions and Functional Analysis*. 5th ed. Nauka, Moscow (in Russian).

McDiarmid, C. (1989). On the method of bounded differences. *Surveys in Combinatorics.* Cambridge University Press.

Mukherjee, S., P. Niyogi, T. Poggio and R. Rifkin (2006). Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Adv. Comput. Math.*, **25**(1–3), 161–193.

Norkin, V.I., and M.A. Keyzer (2008). On convergence of kernel learning estimators. In L. Sakalauskas, O.W. Weber and E.K. Zavadskas (Eds.), *Proceedings of 20th EURO Mini Conference "Continuous Optimization and Knowledge-Based Technologies"* (*EUROPT-2008*), Institute of Mathematics and Informatics, Vilnius. pp. 306–310.

Poggio, T., and S. Smale (2003). The mathematics of learning: dealing with data. *Notices Amer. Math. Soc.*, **50**(5), 537–544.

Raik, E. (1972). On stochastic programming problems with decision rules. In *Proceedings of the Estonian Academy of Sciences. Physics. Mathematics*, Vol. 21. Izvestia AN ESSR. Physika. Math., pp. 258–263.

Rockafellar, R.T., and R.J.-B. Wets (1998). *Variational Analysis.* Springer-Verlag, Berlin.

Ruszczynski, A., and A. Shapiro (Eds.) (2003). Stochastic programming. *Handbooks in OR & MS*, **10**.

Scholkopf, B., and A.J. Smola (2002). *Learning with Kernels. Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press, Cambridge, MA.

Sergienko, I.V., A.M. Gupal and A.A. Vagis (2008). Bayesian approach, theory of empirical risk minimization. Comparative analysis. *Cybernetics and Systems Analysis*, **44**(6), 822–831 (translated from *Kibernetika i Sistemnyi Analiz*, 2008, **6**, 39–49).

Smale, S., and Y. Yao (2006). Online Learning Algorithms. *Foundations of Computational Mathematics*, **6**(2), 145–170.

Smale, S., and D.X. Zhou (2005). Shannon sampling II: Connections to learning theory. *Applied Computational*

Harmonic Analysis, **19**(3), 285–302.
Takeuchi, I., Q.V. Le, T. Sears and A.J. Smola (2006). Nonparametric quantile estimation. *Journal of Machine Learning Research*, **7**, 1231–1264.
Tihonov, A.N., and V.Y. Arsenin (1977). *Solution of Ill-Posed Problems.* WH Winston, Washington, DC.
Vapnik, V.N. (1998). *Statistical Learning Theory.* Wiley, New York.
Yastremskii, A.I. (1980). Optimality conditions in stochastic programming. *Cybernetics and Systems Analysis*, **16**(1), 154–158.
Yudin, D.B. (1979). *Problems and Methods of Stochastic Programming*. Sovetskoe radio, Moscow (in Russian).
Yudin, D.B. (1974). *Mathematical Methods of Control under Incomplete Information* (*Problems and Methods of Stochastic Programming*). Sovetskoe radio, Moscow (in Russian).

V.I. Norkin has graduated from Moscow Institute of Physics and Technology (1974), received candidate (1983) and doctor (1998) degrees from Glushkov Institute of Cybernetics, Kiev, Ukraine. Presently he is a leading researcher at this institute and a member of the editorial board of SIAM Journal on Optimization. His research interests include systems analysis, optimization theory, stochastic programming, nonsmooth analysis, risk theory, actuarial and financial mathemetics, mathematical economics, statistical learning theory.

M.A. Keyzer is a professor of Economics and Director of the Centre for World Food Studies of VU University of Amsterdam (SOW-VU). His research interests include applied statistics, mathematical economics, econometrics, general equilibrium modelling, applied optimization, agricultural economics, land use and food studies.

Apie stochastinį optimizavimą ir statistinį mokymą saviredukuojančių branduolių Hilberto erdvėje

Vladimir NORKIN, Michiel KEYZER

Straipsnyje nagrinėjamos stochastinio optimizavimo problemos Saviredukuojančių Branduolių Hilberto Erdvėje. Tikslo funkcija tokioje erdvėje paprastai yra tikėtinos reikšmės funkcionalas, priklausantis nuo sprendimo taisyklių (strategijų), t.y. funkcijos nuo stebimų atsitiktinių parametrų. Tokio pobūdžio uždaviniai kyla interaktyvių sprendimų ir statistinio mokymo teorijoje. Šitokie uždaviniai sprendžiami pasinaudojant imties vidurkio aproksimacija ir Tichonovo reguliarizacijos metodu. Pasinaudojus atvaizdžio teorema aproksimuojančios reguliarizuotos problemos sprendinys randamas kaip baigtinė tiesinė branduolių funkcijų kombinacija, nustatoma per baigtinį optimizavimą. Statistinėje literatūroje toks būdas vadinamas atraminių vektorių mašinomis. Branduolių sprendinių konvergavimas yra tiriamas palaipsniui mažinant iki nulio reguliarizavimo parametrą, kai stebėjimų kiekis didėja. Darbe išvestos tikėtinos rizikos skirtumo nuo aproksimuojančio sprendinio neasimptotiniai įverčiai. Taip pat nustatytos tolygaus branduolių sprendinio konvergavimo su tikimybe 1 pakankamos sąlygos kartu su reguliarizavimo parametro keitimo taisykle, kai imties tūris didėja. Atskiru atveju išnagrinėtas branduolių regresijos įverčių ir binarinio klasifikatoriaus konvergavimas, kai imties tūris be galo auga.