Neurocomputing 289 (2018) 101–118

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

# Total stability of kernel methods☆

Andreas Christmann[a], Daohong Xiang[b,c,*], Ding-Xuan Zhou[d]

[a] University of Bayreuth, Germany
[b] Zhejiang Normal University, China
[c] University of Bayreuth, Germany
[d] City University of Hong Kong, China

ARTICLE INFO

*Article history:*
Received 22 September 2017
Revised 30 November 2017
Accepted 1 February 2018
Available online 9 February 2018

Communicated by Dr Yiming Ying

*MSC:*
68Q32
62G35
68T05
68T10
62M20

*Keywords:*
Machine learning
Stability
Robustness
Kernel
Regularization

ABSTRACT

Regularized empirical risk minimization using kernels and their corresponding reproducing kernel Hilbert spaces (RKHSs) plays an important role in machine learning. However, the actually used kernel often depends on one or on a few hyperparameters or the kernel is even data dependent in a much more complicated manner. Examples are Gaussian RBF kernels, kernel learning, and hierarchical Gaussian kernels which were recently proposed for deep learning. Therefore, the actually used kernel is often computed by a grid search or in an iterative manner and can often only be considered as an approximation to the "ideal" or "optimal" kernel.

The paper gives conditions under which classical kernel based methods based on a convex Lipschitz loss function and on a bounded and smooth kernel are stable, if the probability measure P, the regularization parameter $\lambda$, and the kernel $K$ may slightly change in a *simultaneous* manner. Similar results are also given for pairwise learning. Therefore, the topic of this paper is somewhat more general than in classical robust statistics, where usually only the influence of small perturbations of the probability measure P on the estimated function is considered.



## 1. Introduction

Regularized empirical risk minimization using the kernel approach including support vector machines (SVMs) based on a general convex loss function and regularized pairwise learning (RPL) methods plays a very important role in machine learning. Such kernel methods have been widely investigated from the points of view of universal consistency, learning rates, and statistical robustness, see e.g. [12,13,28,31,34,35], and the references cited in these books. In short words, universal consistency describes the property that the statistical method or the algorithm converges to the asymptotical optimal value of interest (i.e. the Bayes risk or the Bayes decision function) for *all* probability measures P, if the sample size $n$ converges to infinity and if the regularization parameter $\lambda_n$ converges in an appropriate manner to 0. Unfortunately, it turns out by the so-called no-free-lunch theorem shown by Devroye [15] that universally consistent methods can in general not have a *uniform rate* of convergence for *all* P. However, there is a vast literature that regularized empirical risk minimization based on kernels yields optimal guaranteed rates of convergence on *large subsets* of the set $\mathcal{M}_1$ of all probability measures, see e.g. [5,12,30,32,39], and the references cited therein. Results on the statistical robustness or on various notations of stability have shown that under weak conditions on the loss function $L$ and on the kernel $K$ or its RKHS $H$, many regularized empirical risk minimization methods including general SVMs and RPL methods are stable with respect to small changes in the probability measure P or w.r.t. small changes of the data set, see e.g. [4,7–9,11,20,21,25,26] and the references cited therein. Such kernel methods can often be represented by operators which are continuous or differentiable (in the sense of Gâteaux or Hadamard) with respect to *all* probability measures P.

☆ The work by A. Christmann described in this paper is partially supported by two grants of the Deutsche Forschungsgesellschaft [Project No. CH/291/2-1 and CH/291/3-1]. The work by D. H. Xiang described in this paper is supported by the National Natural Science Foundation of China under Grant 11471292 and the Alexander von Humboldt Foundation of Germany. The work by D.-X. Zhou described in this paper is supported partially by the Research Grants Council of Hong Kong under project # CityU 11304114.

* Corresponding author at: Department of Mathematics, Zhejiang Normal University, Jinhua 321004, China.
*E-mail address:* daohongxiang@zjnu.cn (D. Xiang).

https://doi.org/10.1016/j.neucom.2018.02.009
0925-2312/

The aim of the present paper is to take a step further: we establish some total stability results which show that many regularized empirical risk minimization methods based on kernels are even stable, if the full triple (P, $\lambda$, $K$) consisting of the – of course completely unknown – underlying probability measure P, the regularization parameter $\lambda$, and the kernel $K$ (or its RKHS $H$) changes slightly. Our main results are Theorem 2.7, Corollary 2.9, and Theorem 2.10 for classical loss functions and Theorem 3.3, Corollary 3.4, and Theorem 3.5 for pairwise learning. In particular, we establish results like

$$\|f_{\mathrm{P}_1,\lambda_1,K_1} - f_{\mathrm{P}_2,\lambda_2,K_2}\|_\infty = \mathcal{O}\big(\|\mathrm{P}_1 - \mathrm{P}_2\|_{tv}\big) + \mathcal{O}\big(|\lambda_1 - \lambda_2|\big) + \mathcal{O}\big(\|K_1 - K_2\|_\infty\big), \tag{1.1}$$

where $f_{\mathrm{P}_j,\lambda_j,K_j}$ denotes the regularized empirical risk minimization method for the triple $(\mathrm{P}_j, \lambda_j, K_j)$, $j \in \{1, 2\}$, and $\|\mathrm{P}_1 - \mathrm{P}_2\|_{tv}$ denotes the norm of total variation between the two probability measures. We explicitly give the constants in (1.1), although the constants may not be optimal.

The rest of the paper has the following structure. Section 2 yields results for general SVM-type methods based on a classical loss function $L(x, y, f(x))$. Section 3 yields similar results for pairwise learning based on functions of the form $L(x, \tilde{x}, y, \tilde{y}, f(x), f(\tilde{x}))$. Section 4 gives some examples of practical importance. Gaussian RBF kernels and the recently introduced hierarchical Gaussian RBF kernels for deep learning, see [33], are covered by our results. Section 5 contains a short discussion. All proofs are given in the appendix. As this is a theoretical paper, we omit numerical examples.

## 2. Results for SVMs

In this section we show that many kernel based methods like SVMs have nice total stability properties if simultaneously the distribution P, the regularization parameter $\lambda$ and the kernel $K$ slightly change.

**Assumption 2.1.** Let $\mathcal{X}$ be a complete separable metric space and $\mathcal{Y} \subset \mathbb{R}$ be closed. Let $(X, Y)$ and $(X_i, Y_i)$, $i \in \mathbb{N}$, be independent and identically distributed pairs of random quantities with values in $\mathcal{X} \times \mathcal{Y}$. We denote the joint distribution of $(X_i, Y_i)$ by $\mathrm{P} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$, where $\mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ is the set of all Borel probability measures on the Borel $\sigma$-algebra $\mathcal{B}_{\mathcal{X}\times\mathcal{Y}}$.

Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a continuous, symmetric and positive semidefinite function, i.e., for any finite set of distinct points $\{x_1, \ldots, x_n\} \subset \mathcal{X}$, the kernel matrix $(K(x_i, x_j))_{i,j=1}^n$ is positive semidefinite. Such a function is called a *Mercel kernel*. The *reproducing kernel Hilbert space (RKHS) H* associated with the kernel $K$ is defined in [1] to be the completion of the linear span of the set of functions $\{K(\cdot, x) : x \in \mathcal{X}\}$ with the inner product $\langle \cdot, \cdot \rangle_H$ given by $\langle \Phi(x), \Phi(y) \rangle_H = K(x, y)$, where $\Phi(x) := K(\cdot, x)$ denotes the canonical feature map of $K$, $x \in \mathcal{X}$. RKHSs are interesting, because they satisfy the reproducing property

$$\langle \Phi(x), f \rangle_H = f(x), \quad x \in \mathcal{X}, f \in H. \tag{2.1}$$

**Assumption 2.2.** Let $K, K_1, K_2 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be continuous and bounded kernels with reproducing kernel Hilbert space $H$, $H_1$, $H_2$, respectively. Define $\|K\|_\infty := \sup_{x\in\mathcal{X}} \sqrt{K(x, x)} \in (0, \infty)$, $\|K_j\|_\infty := \sup_{x\in\mathcal{X}} \sqrt{K_j(x, x)} \in (0, \infty)$ for $j \in \{1, 2\}$, and denote $\kappa = \max\{\|K_1\|_\infty, \|K_2\|_\infty\}$. Denote the corresponding canonical feature maps by $\Phi_j(x)$, $j \in \{1, 2\}$.

A function $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ is called a loss function if $L$ is measurable with respect to all Borel probability measures. Because constant loss functions are not useful for applications, we will always assume that $L$ is not a constant function.

A loss function $L(x, y, t)$ is usually represented by a *margin-based* loss function $\tilde{L}(yt)$ for classification and represented by a *distance-based* loss function $\tilde{L}(y - t)$ for regression if $\tilde{L} : \mathbb{R} \to [0, \infty)$ is a measurable function. For example, the hinge loss $L_{\text{hinge}}(x, y, t) = \max\{0, 1 - yt\}$ and the logistic loss $L_{\text{c-logist}}(x, y, t) = \ln(1 + \exp(-yt))$ for classification, the $\epsilon$-insensitive loss $L_{\epsilon\text{-insens}}(x, y, t) = \max\{0, |y - t| - \epsilon\}$ for some $\epsilon > 0$, the Huber's loss $L_{\alpha\text{-Huber}}(x, y, t) = \begin{cases} 0.5(y-t)^2 & \text{if } |y-t| \le \alpha \\ \alpha|y-t| - 0.5\alpha^2 & \text{if } |y-t| > \alpha \end{cases}$ for some $\alpha > 0$ and the logistic loss $L_{\text{r-logist}}(x, y, t) = -\ln \frac{4\exp(y-t)}{(1+\exp(y-t))^2}$ for regression, the pinball loss $L_{\tau\text{-pin}}(x, y, t) = \begin{cases} (\tau - 1)(y - t) & \text{if } |y - t| < 0 \\ \tau(y - t) & \text{if } |y - t| \ge 0 \end{cases}$ for some $\tau > 0$ for quantile regression. We refer to [2,13,28,29,31,34,35,41] for details and more examples of kernels.

**Definition 2.3.** The loss function $L$ is called Lipschitz continuous, if there exists a constant $|L|_1 < \infty$ such that

$$|L(x, y, t_1) - L(x, y, t_2)| \le |L|_1 |t_1 - t_2| \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}, t_1, t_2 \in \mathbb{R}. \tag{2.2}$$

**Assumption 2.4.** Let $L$ be a convex with respect to the last argument and Lipschitz continuous loss function with Lipschitz constant $|L|_1 \in (0, \infty)$.

**Assumption 2.5.** For all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, let $L(x, y, \cdot)$ be differentiable and its derivative be Lipschitz continuous with Lipschitz constant $|L'|_1 \in (0, \infty)$.

The moment condition $\mathbb{E}_\mathrm{P} L(X, Y, 0) < \infty$ excludes heavy-tailed distributions such as the Cauchy distribution and many other stable distributions used in financial or actuarial problems. We avoid the moment condition by shifting the loss with by the term $L(x, y, 0)$. This trick is well-known in the literature on robust statistics, see, e.g., [9,10,23].

Denote the shifted loss function of $L$ by

$$L^\star(x, y, t) := L(x, y, t) - L(x, y, 0), \ (x, y, t) \in \mathcal{X} \times \mathcal{Y} \times \mathbb{R}.$$

The shifted loss function $L^\star$ still shares the properties of $L$ specified in Assumptions 2.4 and2.5, see [10, Proposition 2], in particular, if $L$ is convex, differentiable, and Lipschitz continuous with Lipschitz constant $|L|_1$ with respect to the third argument, then $L^\star$ inherits convexity, differentiability and Lipschitz continuity from $L$ with identical Lipschitz constant $|L^\star|_1 = |L|_1$. Additionally, if the derivative $L'$ satisfies Lipschitz continuity with Lipschitz constant $|L'|_1$, so does $(L^\star)'$ with the identical Lipschitz constant $|(L^\star)'|_1 = |L'|_1$.

The SVM associated with $L^\star$ can be defined to solve a minimization problem as follows

$$f_{\mathrm{P},\lambda,K} := \arg\min_{f\in H} \big(\mathbb{E}_\mathrm{P} L^\star(X, Y, f(X)) + \lambda \|f\|_H^2\big), \tag{2.3}$$

where $\mathrm{P} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$, $H$ is the RKHS of a kernel $K$, and $\lambda > 0$ is a regularization parameter to avoid overfitting.

Although the shifted loss function $L^\star$ changes the objective function of SVMs, the minimizers defined by $L^\star$ and $L$ respectively are the same for all $\mathrm{P} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ and in particular for all empirical distributions D based on a data set consisting of $n$ data points $(x_i, y_i)$, $1 \le i \le n$, if the minimizer of an SVM in terms of $L$ instead of $L^\star$ exists.

Our first main result states that the kernel based estimator $f_{\mathrm{P},\lambda,K}$ defined by (2.3) only changes slightly if the regularization parameter wiggles a little bit. Ye and Zhou [40, Theorem 1] proved the assertion of the following result for margin-based loss functions for classification. Here we show it holds true for more general loss functions.