Function Classes that Approximate the Bayes Risk

Ingo Steinwart, Don Hush, and Clint Scovel

CCS-3, Los Alamos National Laboratory, Los Alamos NM 87545, USA
{ingo,dhush,jcs}@lanl.gov

Abstract. Many learning algorithms approximately minimize a risk functional over a predefined function class. In order to establish consistency for such algorithms it is therefore necessary to know whether this function class approximates the Bayes risk. In this work we present necessary and sufficient conditions for the latter. We then apply these results to reproducing kernel Hilbert spaces used in support vector machines (SVMs). Finally, we briefly discuss universal consistency of SVMs for non-compact input domains.

1 Introduction

Many learning problems such as classification and regression are characterized by a loss function $L : X \times Y \times \mathbb{R} \to [0, \infty]$ and an unknown distribution P on $X \times Y$. Having a sample set $T \in (X \times Y)^n$ drawn in an i.i.d. fashion from P the learning goal is then to find a measurable function $f : X \to \mathbb{R}$ whose *L-risk*

$$\mathcal{R}_{L,\mathrm{P}}(f) := \int_{X \times Y} L\big(x, y, f(x)\big) d\mathrm{P}(x, y)$$

is close to the *Bayes L-risk*, i.e. the smallest possible risk

$$\mathcal{R}^*_{L,\mathrm{P}} := \inf\big\{\mathcal{R}_{L,\mathrm{P}}(f) \,\big|\, f : X \to \mathbb{R} \text{ measurable}\big\}\,.$$

In order to find such a function many learning methods minimize a (modified) empirical risk over a predefined function class F. Examples of such learning methods include empirical risk minimization, SVMs, (regularized) boosting, some neural networks, and certain decision trees.

In a first step of the consistency analysis of such a learning method one typically shows that the algorithm produces with high probability a function f_T whose risk $\mathcal{R}_{L,\mathrm{P}}(f_T)$ is close to the F-optimal L-risk $\mathcal{R}^*_{L,\mathrm{P},F} := \inf_{f \in F} \mathcal{R}_{L,\mathrm{P}}(f)$. In order to show that $\mathcal{R}_{L,\mathrm{P}}(f_T)$ is close to the Bayes risk $\mathcal{R}^*_{L,\mathrm{P}}$ it thus remains to prove that the function class F is (L, P)-*rich*, i.e. that it satisfies

$$\mathcal{R}^*_{L,\mathrm{P},F} = \mathcal{R}^*_{L,\mathrm{P}}\,. \tag{1}$$

In this work we provide both necessary and sufficient conditions for (1) for a variety of loss functions and distributions. Moreover, we apply these general conditions to reproducing kernel Hilbert spaces (RKHSs) used in SVMs. In particular,

for universal kernels introduced in [1] we establish (L, P)-richness for essentially all reasonable continuous loss functions. Furthermore, for kernels acting on discrete spaces X we establish the first sufficient conditions for (L, P)-richness, and we also show that the Gaussian RBF kernels are rich for $X := \mathbb{R}^d$. Finally, we use these results to discuss consistency of SVMs over non-compact input domains.

The rest of this work is organized as follows. In Section 2 we introduce basic notions for loss functions and provide various examples of losses satisfying these notions. In Section 3 we then present our main results on richness and apply them to RKHSs. Finally, the proofs of all results are gathered in Section 4.

2 Preliminaries: Losses and their Risks

In the following X is always a measurable space if not mentioned otherwise and $Y \subset \mathbb{R}$ is always a closed subset. Moreover, $\mathcal{L}_0(X)$ denotes the set of all measurable functions $f : X \to \mathbb{R}$, and $L_p(\mu)$ stands for the standard space of p-integrable functions with respect to the measure μ on X.

Let us now introduce the fundamental definitions of this work:

Definition 1. *A function $L : X \times Y \times \mathbb{R} \to [0, \infty]$ is called a* loss function *if it is measurable. In this case L is called:*

i) convex *if $L(x, y, \,.\,) : \mathbb{R} \to [0, \infty]$ is convex for all $x \in X$, $y \in Y$.*
ii) continuous *if $L(x, y, \,.\,) : \mathbb{R} \to [0, \infty]$ is continuous for all $x \in X$, $y \in Y$.*

It is obvious that the risk of a convex loss is convex on $\mathcal{L}_0(X)$. However, in general the risk of a continuous loss is not continuous. In order to ensure this continuity (cf. Lemma 2) we need the following definition:

Definition 2. *We call a loss function $L : X \times Y \times \mathbb{R} \to [0, \infty]$ a* Nemitski loss function *if there exist a measurable function $b : X \times Y \to [0, \infty)$ and an increasing function $h : [0, \infty) \to [0, \infty)$ with*

$$L(x, y, t) \leq b(x, y) + h\big(|t|\big)\,, \qquad (x, y, t) \in X \times Y \times \mathbb{R}. \tag{2}$$

Furthermore, we say that L is a Nemitski loss of order $p \in (0, \infty)$*, if there exists a constant $c > 0$ with $h(t) = c\, t^p$ for all $t \geq 0$. Finally, if P is a distribution on $X \times Y$ with $b \in L_1(\mathrm{P})$ we say that L is a* P-integrable Nemitski loss.

Note that P-integrable Nemitski loss functions L satisfy $\mathcal{R}_{L,\mathrm{P}}(f) < \infty$ for all $f \in L_\infty(\mathrm{P}_X)$, and consequently we also have $\mathcal{R}_{L,\mathrm{P}}(0) < \infty$ and $\mathcal{R}^*_{L,\mathrm{P}} < \infty$.

Let us now present some examples of loss functions that satisfy the above definitions. We begin with the class of locally Lipschitz continuous loss functions:

Example 1. A loss $L : X \times Y \times \mathbb{R} \to [0, \infty)$ is called *locally Lipschitz continuous* if

$$|L|_{a,1} := \sup_{\substack{t,t' \in [-a,a] \\ t \neq t'}} \sup_{\substack{x \in X \\ y \in Y}} \frac{\big|L(x, y, t) - L(x, y, t')\big|}{|t - t'|} \,<\, \infty\,, \qquad a > 0. \tag{3}$$

Moreover, L is called *Lipschitz continuous* if $|L|_1 := \sup_{a>0} |L|_{a,1} < \infty$.

Note that if $Y \subset \mathbb{R}$ is a *finite* subset and $L : Y \times \mathbb{R} \to [0,\infty)$ is a convex loss function then L is a locally Lipschitz continuous loss function. Moreover, a locally Lipschitz continuous loss function L is a Nemitski loss since (3) yields

$$L(x,y,t) \le L(x,y,0) + |L|_{|t|,1}|t|\,, \qquad (x,y,t) \in X \times Y \times \mathbb{R}. \tag{4}$$

In particular, a locally Lipschitz continuous loss L is a P-integrable Nemitski loss if and only if $\mathcal{R}_{L,\mathrm{P}}(0) < \infty$. Moreover, if L is Lipschitz continuous then L is a Nemitski loss of order 1. $\lhd$

The following two examples present some commonly used types of loss functions that satisfy the above definitions:

Example 2. A loss $L : Y \times \mathbb{R} \to [0,\infty)$ of the form $L(y,t) = \varphi(yt)$ for a suitable function $\varphi : \mathbb{R} \to \mathbb{R}$ and all $y \in Y := \{-1,1\}$ and $t \in \mathbb{R}$, is called *margin-based.* Recall that margin-based losses such as the (squared) hinge loss, the AdaBoost loss, the logistic loss and the least squares loss are used in many classification algorithms. Obviously, L is convex, continuous, or (locally) Lipschitz continuous if and only if φ is. In addition, convexity of L implies local Lipschitz continuity of L. Moreover, L is always a P-integrable Nemitski loss since we have

$$L(y,t) \le \max\{\varphi(-t), \varphi(t)\}$$

for all $y \in Y$ and all $t \in \mathbb{R}$. From this we can also easily derive a characterization for L being a P-integrable Nemitski loss of order p. $\lhd$

Example 3. A loss $L : Y \times \mathbb{R} \to [0,\infty)$ of the form $L(y,t) = \psi(y-t)$ for a suitable function $\psi : \mathbb{R} \to \mathbb{R}$ and all $y \in Y := \mathbb{R}$ and $t \in \mathbb{R}$, is called *distance-based.* Distance-based losses such as the least squares loss, Huber's insensitive loss, the logistic loss, or the ϵ-insensitive loss are usually used for regression. It is easy to see that L is convex, continuous, or Lipschitz continuous if and only if ψ is. Let us say that L is of *upper growth* $p \in [1,\infty)$ if there is a $c > 0$ with

$$\psi(r) \le c\left(|r|^p + 1\right), \qquad r \in \mathbb{R}.$$

Then it is obvious that L is of upper growth type 1 if it is Lipschitz continuous, and if L is convex the converse implication also holds. In addition, a distance-based loss function of upper growth type $p \in [1,\infty)$ is a Nemitski loss of order p, and if the distribution P satisfies $\mathbb{E}_{(x,y)\sim \mathrm{P}}|y|^p < \infty$ it is also P-integrable. $\lhd$

3 Main Results

In this section we present our main results establishing (L,P)-richness, i.e. Equation (1). Let us begin with some *sufficient* conditions:

Theorem 1. *Let $L : X \times Y \times \mathbb{R} \to [0,\infty]$ be a continuous loss function,* P *be a distribution on $X \times Y$ such that L is a* P*-integrable Nemitski loss, and $F \subset L_\infty(\mathrm{P}_X)$. Furthermore, assume that for all $g \in L_\infty(\mathrm{P}_X)$ there exists a sequence $(f_n) \subset F$ with $\sup_{n\geq 1} \|f_n\|_\infty < \infty$ and*

$$\lim_{n\to\infty} f_n(x) \;=\; g(x) \tag{5}$$

for P_X-almost all $x \in X$. Then F is (L,P)-rich.

Note that the assumptions on F in Theorem 1 are satisfied if and only if for all $g \in L_\infty(\mathrm{P}_X)$ there exists a sequence $(f_n) \subset F$ with $\sup_{n\geq 1} \|f_n\|_\infty < \infty$ and $f_n \to g$ in probability P_X.

Let us now provide an interesting example of a set F satisfying the assumption of Theorem 1. To this end let X be a compact topological Hausdorff space, $k : X \times X \to \mathbb{R}$ be a continuous kernel and H be its associated RKHS. Following [1] we say that H (or k) is universal if H is dense in the space $C(X)$ of continuous functions $f : X \to \mathbb{R}$, equipped with the $\|.\|_\infty$-norm. For examples of such kernels we refer to [1]. Now the following result applies Theorem 1 to universal kernels:

Corollary 1. *Let X be a compact metric space, $L : X \times Y \times \mathbb{R} \to [0,\infty]$ be a continuous loss function, and* P *be a distribution on $X \times Y$ such that L is a* P*-integrable Nemitski loss. Then every universal RKHS over X is (L,P)-rich.*

At first glance it seems disappointing that Corollary 1 only holds for compact *metric* spaces. However, the following theorem shows that these spaces are the only ones which permit universal kernels:

Theorem 2. *For a compact topological Hausdorff space X the following statements are equivalent:*

i) There exists a universal kernel k on X.
ii) X is metrizable, i.e. there exists a metric generating the topology.

Most common losses are of some order $p \in [1,\infty)$. For such losses we now present a weaker sufficient condition for richness than that of Theorem 1:

Theorem 3. *Let $L : X \times Y \times \mathbb{R} \to [0,\infty]$ be a continuous loss function and* P *be a distribution on $X \times Y$ such L is a* P*-integrable Nemitski loss of order $p \in [1,\infty)$. Then every dense subset $F \subset L_p(\mathrm{P}_X)$ is (L,P)-rich.*

Note that if F is a dense subset of $L_p(\mathrm{P}_X)$ for some $p \in [1,\infty)$ then it is also a dense subset of $L_q(\mathrm{P}_X)$ for all q with $1 \leq q \leq p$. Consequently, the denseness assumption in Theorem 3 becomes stronger for loss functions of higher order.

The condition that F is dense in $L_p(\mu)$ can often be guaranteed by means from functional analysis. In the following we will illustrate this for RKHSs. To this end let $k : X \times X \to \mathbb{R}$ be a measurable kernel and H be its associated RKHS. Given a measure μ on X and a real number $p \in [1,\infty)$ we define

$$\|k\|_{L_p(\mu)} := \Big(\int_X k^{\frac{p}{2}}(x,x)d\mu(x)\Big)^{\frac{1}{p}}.$$

It is elementary to check that the inclusion $I : H \to L_p(\mu)$ is well-defined and continuous if $\|k\|_{L_p(\mu)} < \infty$. Furthermore, its adjoint operator I' is the integral operator $T_k : L_{p'}(\mu) \to H$ defined by

$$T_k g(x) := \int_X k(x, x') g(x') d\mu(x') \,, \qquad g \in L_{p'}(\mu),\, x \in X, \tag{6}$$

where p' is defined by $\frac{1}{p} + \frac{1}{p'} = 1$. With these preparations we can now characterize when H is dense in $L_p(\mu)$:

Proposition 1. *Let X be a measurable space, μ be a measure on X, k be a measurable kernel on X with RKHS H such that $\|k\|_{L_p(\mu)} < \infty$ for some $p \in [1, \infty)$. Then the following statements are equivalent:*

i) H is dense in $L_p(\mu)$.
ii) The integral operator $T_k : L_{p'}(\mu) \to H$ defined by (6) is injective.

The next result provides injectivity for the integral operators of the Gaussian RBF kernels k_σ defined by $k_\sigma(x, x') := \exp(-\|x - x'\|_2^2/\sigma^2)$, $x, x' \in \mathbb{R}^d$, $\sigma > 0$.

Theorem 4. *Let μ be a finite measure on $\mathbb{R}^d$ and H_σ be the RKHS of k_σ. Then $T_{k_\sigma} : L_{p'}(\mu) \to H_\sigma$ is injective for all $p \in (1, \infty)$ and all $\sigma > 0$.*

Combining Theorem 4 with a stability argument in the sense of [2], it is not hard to show that an SVM with e.g. the hinge loss and a Gaussian RBF kernel is classification consistent for all distributions P on $\mathbb{R}^d \times \{-1, 1\}$. This extends the known consistency result [3, 4] from bounded to unbounded input domains.

If P_X is absolutely continuous with respect to some measure μ then Proposition 1 yields the following sufficient condition for H being dense in $L_p(\mathrm{P}_X)$:

Corollary 2. *Let X be a measurable space, μ be a measure on X, and k be a measurable kernel on X with RKHS H and $\|k\|_{L_p(\mu)} < \infty$ for some $p \in [1, \infty)$. Assume that $T_k : L_{p'}(\mu) \to H$ is injective. Then H is dense in $L_q(h\mu)$ for all $q \in [1, p]$ and all measurable $h : X \to [0, \infty)$ with $h \in L_s(\mu)$, where $s := \frac{p}{p-q}$.*

Let us now investigate denseness properties of RKHSs over discrete spaces X. To this end let us write $\ell_p(X) := L_p(\nu)$, where $p \in [1, \infty]$ and ν is the counting measure on X. Note that these spaces obviously satisfy $\ell_p(X) \subset \ell_q(X)$ for $p \le q$ which is used in the proof of the following corollary:

Corollary 3. *Let X be a countable set and k be a kernel on X with $\|k\|_{\ell_p(X)} < \infty$ for some $p \in [1, \infty)$. If k satisfies*

$$\sum_{x, x' \in X} k(x, x') f(x) f(x') > 0 \tag{7}$$

for all $f \in \ell_{p'}(X)$ with $f \neq 0$ then the RKHS of k is dense in $L_q(\mu)$ for all $q \in [1, \infty)$ and all distributions μ on X.

Note that the sharpest case $p = q = \infty$ is excluded in the above corollary. The reason for this is that the dual of $\ell_\infty(X)$ is *not* $\ell_1(X)$. However, if instead we consider the *pre*-dual of $\ell_1(X)$, namely the Banach space

$$c_0(X) := \left\{ f : X \to \mathbb{R} \,\middle|\, \forall \varepsilon > 0 \, \exists \text{ finite } A \subset X \text{ such that } \forall x \in X \setminus A : |f(x)| \leq \varepsilon \right\}$$

which is equipped with the usual $\|.\|_\infty$-norm, then we obtain:

Theorem 5. *Let X be a countable set and k be a bounded kernel on X that satisfies $k(.,x) \in c_0(X)$ for all $x \in X$, and (7) for all $f \in \ell_1(X)$ with $f \neq 0$. Then the RKHS of k is (L,P)-rich for all distribution P on $X \times Y$ and all continuous, P-integrable Nemitski losses $L : X \times Y \times \mathbb{R} \to [0,\infty)$.*

It is not hard to see that there exist non-trivial kernels satisfying the assumptions of Theorem 3 and Theorem 5. Using a stability argument we hence see that for all countable X there exist non-trivial, universally consistent SVMs.

Let us now present some *necessary* conditions for (L,P)-richness. To this end we first recall some concepts from [5]: let $L : Y \times \mathbb{R} \to [0,\infty]$ be a loss function and Q be a distribution on Y. We define the *inner L-risk* of Q by

$$\mathcal{C}_{L,\mathrm{Q}}(t) := \int_Y L(y,t)\, d\mathrm{Q}(y)\,, \qquad\qquad t \in \mathbb{R}\,.$$

Furthermore, the *minimal inner L-risk* is denoted by $\mathcal{C}^*_{L,\mathrm{Q}} := \inf_{t\in\mathbb{R}} \mathcal{C}_{L,\mathrm{Q}}(t)$, and the corresponding set of (non-trivial) minimizers is defined by

$$\mathcal{M}_{L,\mathrm{Q}}(0^+) := \left\{ t \in \mathbb{R} : \mathcal{C}_{L,\mathrm{Q}}(t) = \mathcal{C}^*_{L,\mathrm{Q}} \right\}$$

if $\mathcal{C}^*_{L,\mathrm{Q}} < \infty$, and by $\mathcal{M}_{L,\mathrm{Q}}(0^+) := \emptyset$ otherwise. Finally, we need the *self-calibration function* which is defined by

$$\delta_{\max,L}(\varepsilon,\mathrm{Q}) := \inf\left\{ \mathcal{C}_{L,\mathrm{Q}}(t) - \mathcal{C}^*_{L,\mathrm{Q}} : t \in \mathbb{R} \text{ with } |t-t'| \geq \varepsilon \text{ for all } t' \in \mathcal{M}_{L,\mathrm{Q}}(0^+) \right\}$$

for $\varepsilon \in [0,\infty)$ and Q with $\mathcal{C}^*_{L,\mathrm{Q}} < \infty$. Note that in [5] this function is denoted by $\delta_{\max,\check{L},L}(\varepsilon,\mathrm{Q})$. Moreover, in [5] it was shown that $\mathcal{M}_{L,\mathrm{Q}}(0^+) = \{t^*_\mathrm{Q}\}$ implies

$$\delta_{\max,L}(|t - t^*_{L,\mathrm{Q}}|,\mathrm{Q}) \leq \mathcal{C}_{L,\mathrm{Q}}(t) - \mathcal{C}^*_{L,\mathrm{Q}}\,, \qquad\qquad t \in \mathbb{R},$$

i.e. the self-calibration function quantifies how well an approximate minimizer of $\mathcal{C}_{L,\mathrm{Q}}(.)$ approximates the exact minimizer t^*_Q. Finally, given a set of distributions $\mathcal{Q}$ on Y we say that a distribution P on $X \times Y$ is of type $\mathcal{Q}$ if its conditional probabilities satisfy $\mathrm{P}(.|x) \in \mathcal{Q}$ for P_X-almost all $x \in X$.

Now we can formulate our first necessary condition for (L,P)-richness:

Theorem 6. *Let $L : Y \times \mathbb{R} \to [0,\infty)$ be a convex loss function such that there exists two distributions Q_1, Q_2 on Y, and two real numbers $t^*_1 \neq t^*_2$ with $\mathcal{M}_{L,\mathrm{Q}_1}(0^+) = \{t^*_1\}$ and $\mathcal{M}_{L,\mathrm{Q}_2}(0^+) = \{t^*_2\}$. Furthermore, let X be a measurable space and μ be a distribution on X. Assume that $F \subset L_\infty(\mu)$ is a (L,P)-rich linear subspace for all $\{\mathrm{Q}_1,\mathrm{Q}_2\}$-type distributions P on $X \times Y$ with $\mathrm{P}_X = \mu$. Then for all $g \in L_\infty(\mu)$ there exists a sequence $(f_n) \subset F$ with*

$$\lim_{n\to\infty} f_n(x) = g(x) \qquad\qquad \text{for } \mu\text{-almost all } x \in X.$$

Corollary 4. *Let X, L and Q_1, Q_2 be as in Theorem 6. Furthermore, let k be a measurable kernel on X whose RKHS H is (L,P)-rich for all $\{\mathrm{Q}_1,\mathrm{Q}_2\}$-type distributions P on $X \times Y$. Then k is strictly positive definite.*

There is an gap between the sufficient condition of Theorem 1 and the necessary condition of Theorem 6. Our next goal is to close this gap for certain Nemitski losses of order p. To this end we first present the following necessary condition which uses an additional assumption on the self-calibration function:

Theorem 7. *Let $L : Y \times \mathbb{R} \to [0,\infty)$ be a convex loss function such that there exists two distributions Q_1, Q_2 on Y, and two real numbers $t_1^* \neq t_2^*$ with $\mathcal{M}_{L,\mathrm{Q}_1}(0^+) = \{t_1^*\}$ and $\mathcal{M}_{L,\mathrm{Q}_2}(0^+) = \{t_2^*\}$. In addition assume that there exist constants $B > 0$ and $p > 0$ with*

$$\delta_{\max,L}(\varepsilon, \mathrm{Q}_i) \geq B\,\varepsilon^p\,, \qquad\qquad \varepsilon > 0,\ i = 1,2.$$

Furthermore, let X be a measurable space, μ be a distribution on X, and $F \subset L_p(\mu)$ be a (L,P)-rich linear subspace for all $\{\mathrm{Q}_1,\mathrm{Q}_2\}$-type distributions P on $X \times Y$ with $\mathrm{P}_X = \mu$. Then F is dense in $L_p(\mu)$.

By combining Theorem 3 with Theorem 7 we now obtain the following characterization of (L,P)-richness:

Theorem 8. *Let $L : Y \times \mathbb{R} \to [0,\infty)$ be a convex Nemitski loss of order $p \in [1,\infty)$, i.e. we have (2) with $b : Y \to [0,\infty)$. Furthermore, let Q_1, Q_2 be distributions on Y with $b \in L_1(\mathrm{Q}_1) \cap L_1(\mathrm{Q}_2)$, and t_1^*, $t_2^* \in \mathbb{R}$ be real numbers with $\mathcal{M}_{L,\mathrm{Q}_1}(0^+) = \{t_1^*\}$, $\mathcal{M}_{L,\mathrm{Q}_2}(0^+) = \{t_2^*\}$, and $t_1^* \neq t_2^*$. In addition, assume that there exists a constant $B > 0$ such that their self-calibration functions satisfy*

$$\delta_{\max,L}(\varepsilon, \mathrm{Q}_i) \geq B\,\varepsilon^p\,, \qquad\qquad \varepsilon > 0,\ i = 1,2.$$

Furthermore, let X be a measurable space, μ be a distribution on X, and $F \subset L_p(\mu)$ be a subspace. Then the following statements are equivalent:

i) F is (L,P)-rich for all distributions P on $X \times Y$ with $\mathrm{P}_X = \mu$ for which L is a P-integrable Nemitski loss of order p.
ii) F is (L,P)-rich for all $\{\mathrm{Q}_1,\mathrm{Q}_2\}$-type distributions P on $X \times Y$ with $\mathrm{P}_X = \mu$.
iii) F is dense in $L_p(\mu)$.

The following examples illustrate that many important loss functions satisfy the assumptions of Theorem 8:

Example 4. For $p \geq 1$ let L be the loss defined by $L(y,t) := |y-t|^p$, $y,t \in \mathbb{R}$. Furthermore, let $\mathrm{Q}_1 := \delta_{\{y_1\}}$, $\mathrm{Q}_1 := \delta_{\{y_2\}}$ be two Dirac distributions on $\mathbb{R}$ with $y_1 \neq y_2$. Then L, Q_1, and Q_2 satisfy the assumptions of Theorem 8 for p.

In order to see this, we first observe with Example 3 that L is a Nemitski loss of order p. Furthermore, for the Dirac measure $\mathrm{Q} := \delta_{\{y_0\}}$ at some $y_0 \in \mathbb{R}$ we have $\mathcal{C}_{L,\mathrm{Q}}(t) = |t-y_0|^p$, $t \in \mathbb{R}$. Consequently, we have $\mathcal{C}_{L,\mathrm{Q}}^* = 0$ and $\mathcal{M}_{L,\mathrm{Q}}(0^+) = \{y_0\}$. With these equalities it is easy to check that the self-calibration function of L is

$$\delta_{\max,L}(\varepsilon, \mathrm{Q}) = \varepsilon^p\,, \qquad\qquad \varepsilon \geq 0. \qquad\qquad \triangleleft$$

The above example shows that F being a dense subspace of $L_p(\mathrm{P}_X)$ characterizes the (L,P)-richness. Moreover, it also shows that restricting the class of distributions to noise-free distributions P, i.e. to distributions with $\mathrm{P}(.|x) = \delta_{\{f(x)\}}$ for measurable functions $f : X \to \mathbb{R}$, does not change this characterization. In other words, if we do not want to impose further assumptions on P then F being dense in $L_p(\mathrm{P}_X)$ is the condition we should look for.

The following example provides a similar characterization for the ϵ-insensitive loss function used in the standard SVM formulation for regression:

Example 5. Let $\epsilon > 0$ and L be the ϵ-insensitive loss defined by $L(y,t) := \max\{0, |y-t| - \epsilon\}$, $y, t \in \mathbb{R}$. Furthermore, for $y_1, y_2 \in \mathbb{R}$ with $y_1 \neq y_2$ we define $\mathrm{Q}_i := \frac{1}{2}\delta_{\{y_i-\epsilon\}} + \frac{1}{2}\delta_{\{y_i+\epsilon\}}$, $i = 1, 2$. Then L, Q_1, and Q_2 satisfy the assumptions of Theorem 8 for $p = 1$.

In order to see this, we first observe that L is a Nemitski loss of order 1. Let us define $\psi(r) := \max\{0, |r| - \epsilon\}$, $r \in \mathbb{R}$. For Q_i, $i = 1, 2$ we then have

$$2\mathcal{C}_{L,\mathrm{Q}_i}(t) = \psi(y_i - \epsilon - t) + \psi(y_i + \epsilon - t)\,, \qquad t \in \mathbb{R},$$

and thus we have $\mathcal{C}_{L,\mathrm{Q}_i}(y_i) = 0 \leq \mathcal{C}_{L,\mathrm{Q}_i}(t)$ for all $t \in \mathbb{R}$. For $t \geq 0$ this yields

$$\mathcal{C}_{L,\mathrm{Q}_i}(y_i \pm t) - \mathcal{C}^*_{L,\mathrm{Q}_i} = \frac{1}{2}\psi(\epsilon + t) + \frac{1}{2}\psi(\epsilon - t) \geq \frac{1}{2}\psi(\epsilon + t) = \frac{t}{2}\,,$$

and hence we find both $\mathcal{M}_{L,\mathrm{Q}_i}(0^+) = \{y_i\}$ and $\delta_{\max,L}(\varepsilon, \mathrm{Q}) = \frac{\varepsilon}{2}$ for $\varepsilon \geq 0$. $\lhd$

Our last example provides a characterization of richness for the hinge loss used in the standard SVM formulation for classification:

Example 6. Let L be the hinge loss defined by $L(y,t) := \max\{0, 1 - yt\}$, $y \in Y := \{-1, 1\}$, $t \in \mathbb{R}$. Furthermore, let Q_1, Q_2 be distributions on Y with $\eta_1 := \mathrm{Q}_1(\{1\}) \in (0, 1/2)$ and $\eta_2 := \mathrm{Q}_2(\{1\}) \in (1/2, 1)$. Then L, Q_1, and Q_2 satisfy the assumptions of Theorem 8 for $p = 1$.

In order to see this we first observe that L is Lipschitz continuous and hence a Nemitski loss of order 1. Moreover, it is well-known that $\mathcal{M}_{L,\eta}(0^+) = \{\mathrm{sign}(2\eta - 1)\}$ for $\eta \neq 0, \frac{1}{2}, 1$, and in addition, for such η an elementary calculation shows

$$\delta_{\max,L}(\varepsilon, \eta) = \varepsilon \, \min\bigl\{\eta, 1 - \eta, 2\eta - 1\bigr\}\,, \qquad \varepsilon \geq 0. \qquad \lhd$$

Note that unlike the distributions in Example 4 the distributions in Example 6 are noisy. This is due to the fact that only noise makes the hinge loss minimizer unique. Moreover, note that using e.g. the least squares loss for a classification problem requires $L_2(\mu)$-denseness which in general is a strictly stronger condition than the $L_1(\mu)$-denseness required for the hinge loss or the logistic loss. This is remarkable since the target functions for the former two losses are bounded, whereas in general the target function for the logistic loss is not even integrable.

Obviously, Condition (7) implies that k is strictly positive definite, and we have already seen in Corollary 1 that this property is *necessary* for richness. Our last result now shows that in general it is *not sufficient*:

Theorem 9. *There exists a bounded strictly positive definite kernel k on $X := \mathbb{N}_0$ with $k(.,x) \in c_0(X)$ for all $x \in X$, such that for all measures μ on X with $\mu(\{x\}) > 0$, $x \in X$, the RKHS H of k is not dense in $L_1(\mu)$.*

4 Proofs

Lemma 1. *Let $L : X \times Y \times \mathbb{R} \to [0,\infty]$ be a continuous loss, P be a distribution on $X \times Y$, and $(f_n) \subset \mathcal{L}_0(X)$ be a sequence that converges to an $f \in \mathcal{L}_0(X)$ in probability P_X. Then we have $\mathcal{R}_{L,\mathrm{P}}(f) \le \liminf_{n\to\infty} \mathcal{R}_{L,\mathrm{P}}(f_n)$.*

Proof. Since (f_n) converges in probability, there exists a subsequence (f_{n_k}) of (f_n) with

$$\lim_{k\to\infty} \mathcal{R}_{L,\mathrm{P}}(f_{n_k}) = \liminf_{n\to\infty} \mathcal{R}_{L,\mathrm{P}}(f_n)$$

and $f_{n_k}(x) \to f(x)$ for P_X-almost all $x \in X$. By the continuity of L we then have $L(x,y,f_{n_k}(x)) \to L(x,y,f(x))$ almost surely and hence Fatou's lemma gives

$$\begin{aligned}\mathcal{R}_{L,\mathrm{P}}(f) = \int_{X\times Y} \lim_{k\to\infty} L\big(x,y,f_{n_k}(x)\big)\, d\mathrm{P}(x,y) &\le \liminf_{k\to\infty} \int_{X\times Y} L\big(x,y,f_{n_k}(x)\big)\, d\mathrm{P}(x,y) \\ &= \liminf_{n\to\infty} \mathcal{R}_{L,\mathrm{P}}(f_n)\,. \qquad \square\end{aligned}$$

Lemma 2. *Let P be a distribution on $X \times Y$ and $L : X \times Y \times \mathbb{R} \to [0,\infty]$ be a continuous, P-integrable Nemitski loss function. Then we have:*

i) Let $f_n \in \mathcal{L}_0(X)$, $n \ge 1$, be functions with $B := \sup_{n\ge1} \|f_n\|_\infty < \infty$. If the sequence (f_n) converges P_X-almost surely to an $f \in \mathcal{L}_0(X)$ then we have

$$\lim_{n\to\infty} \mathcal{R}_{L,\mathrm{P}}(f_n) = \mathcal{R}_{L,\mathrm{P}}(f)\,.$$

ii) The map $\mathcal{R}_{L,\mathrm{P}} : L_\infty(\mathrm{P}_X) \to [0,\infty)$ is continuous.

iii) If L is of order $p \in [1,\infty)$ then $\mathcal{R}_{L,\mathrm{P}} : L_p(\mathrm{P}_X) \to [0,\infty)$ is continuous.

Proof. i). Obviously, f is a bounded and measurable function with $\|f\|_\infty \le B$. Furthermore, the continuity of L shows $L(x,y,f_{n_k}(x)) \to L(x,y,f(x))$ almost surely. In addition, for P-almost all $(x,y) \in X \times Y$ and all $n \ge 1$ we have

$$\big|L\big(x,y,f_n(x)\big) - L\big(x,y,f(x)\big)\big| \le 2b(x,y) + 2h(B)\,.$$

Since the function on the right hand side is P-integrable, we then obtain the assertion from Lebesgue's convergence theorem and

$$\big|\mathcal{R}_{L,\mathrm{P}}(f_n) - \mathcal{R}_{L,\mathrm{P}}(f)\big| \le \int_X \Big|L\big(x,y,f_n(x)\big) - L\big(x,y,f(x)\big)\Big|\, d\mathrm{P}(x,y)\,.$$

ii). This is a direct consequence of Condition (2) and *i).*

iii). Since L is a P-integrable Nemitski loss of order p we find $\mathcal{R}_{L,\mathrm{P}}(f) < \infty$ for all $f \in L_p(\mathrm{P}_X)$. Now let $(f_n) \subset L_p(\mathrm{P}_X)$ be a sequence converging to some $f \in L_p(\mathrm{P}_X)$. Lemma 1 then yields $\mathcal{R}_{L,\mathrm{P}}(f) \le \liminf_{n\to\infty} \mathcal{R}_{L,\mathrm{P}}(f_n)$. Moreover, $\tilde{L}(x,y,t) := b(x,y) + c|t|^p - L(x,y,t)$ is also a continuous loss, and hence we have

$$\begin{aligned}\|b\|_{L_1(\mathrm{P})} + c\|f\|_p^p - \mathcal{R}_{L,\mathrm{P}}(f) = \mathcal{R}_{\tilde{L},\mathrm{P}}(f) &\le \liminf_{n\to\infty} \mathcal{R}_{\tilde{L},\mathrm{P}}(f_n) \\ &= \|b\|_{L_1(\mathrm{P})} + \liminf_{n\to\infty} c\|f_n\|_p^p - \mathcal{R}_{L,\mathrm{P}}(f_n)\end{aligned}$$

by Lemma 1. Using the fact that $\|.\|_p^p$ is a continuous function on $L_p(\mathrm{P}_X)$, we thus obtain $\limsup_{n\to\infty} \mathcal{R}_{L,\mathrm{P}}(f_n) \le \mathcal{R}_{L,\mathrm{P}}(f)$. $\square$

Lemma 3. *Let $L : X \times Y \times \mathbb{R} \to [0,\infty]$ be a loss, and* P *be a distribution on $X \times Y$ such that L is a* P*-integrable Nemitski loss. Then $L_\infty(\mathrm{P}_X)$ is (L,P)-rich.*

Proof. Let us fix a measurable function $f : X \to \mathbb{R}$ with $\mathcal{R}_{L,\mathrm{P}}(f) < \infty$. Then the functions $f_n := \mathbf{1}_{\{|f|\leq n\}} f$, $n \geq 1$, are bounded, and an easy calculation shows

$$\begin{aligned}\bigl|\mathcal{R}_{L,\mathrm{P}}(f_n) - \mathcal{R}_{L,\mathrm{P}}(f)\bigr| &\leq \int_{\{|f|>n\}\times Y} \bigl|L(x,y,0) - L(x,y,f(x))\bigr|\, d\mathrm{P}(x,y) \\ &\leq \int_{\{|f|>n\}\times Y} b(x,y) + h(0) + L\bigl(x,y,f(x)\bigr)\, d\mathrm{P}(x,y)\end{aligned}$$

for all $n \geq 1$. In addition, the integrand in the last integral is integrable since $\mathcal{R}_{L,\mathrm{P}}(f) < \infty$ and $b \in L_1(\mathrm{P})$, and consequently Lebesgue's theorem yields $\mathcal{R}_{L,\mathrm{P}}(f_n) \to \mathcal{R}_{L,\mathrm{P}}(f)$ for $n \to \infty$. From this we easily get the assertion. □

Proof (of Theorem 1). By Lemma 3 we know $\mathcal{R}^*_{L,\mathrm{P},L_\infty(\mathrm{P}_X)} = \mathcal{R}^*_{L,\mathrm{P}}$, and since $F \subset L_\infty(\mathrm{P}_X)$ we also have $\mathcal{R}^*_{L,\mathrm{P},F} \geq \mathcal{R}^*_{L,\mathrm{P},L_\infty(\mathrm{P}_X)}$. In order to show the converse inequality we fix a $g \in L_\infty(\mathrm{P}_X)$. Let $(f_n) \subset F$ be a sequence of functions according to the assumptions of the theorem. Lemma 2 then yields $\mathcal{R}_{L,\mathrm{P}}(f_n) \to \mathcal{R}_{L,\mathrm{P}}(g)$, and hence we easily find $\mathcal{R}^*_{L,\mathrm{P},F} \leq \mathcal{R}^*_{L,\mathrm{P},L_\infty(\mathrm{P}_X)}$. □

Proof (of Corollary 1). Let us fix a $g \in L_\infty(\mathrm{P}_X)$. Then there exists a sequence $(g_n) \subset C(X)$ with $\|g_n\|_\infty \leq \|g\|_\infty$ for all $n \geq 1$ and $g_n(x) \to g(x)$ for P_X-almost all $x \in X$. Moreover, the universality of H gives functions $f_n \in H$ with $\|f_n - g_n\|_\infty \leq 1/n$ for all $n \geq 1$. Since this yields both $\|f_n\|_\infty \leq 1 + \|g\|_\infty$, $n \geq 1$, and $f_n(x) \to g(x)$ almost surely, we obtain the assertion by Theorem 1. □

Proof (of Theorem 2). By [6, Thm. 3.2.11 and Cor. 3.3.2] we know that X is completely regular and hence [7, Thm. V.6.6.] shows that X is metrizable if and only if $C(X)$ is separable.
i) $\Rightarrow$ *ii)*. Let H be the RKHS of k and $\Phi : X \to H$ be the canonical feature map. Then Φ is continuous and thus $\Phi(X)$ is compact. Since H is obviously a metric space we hence see that $\Phi(X)$ is separable, and consequently so is $H = \overline{\operatorname{span}\Phi(X)}$. Since H is dense in $C(X)$ we then obtain that $C(X)$ is separable.
ii) $\Rightarrow$ *i)*. Since our preliminary consideration shows that $C(X)$ is separable there exists a dense subset $\{f_n : n \in \mathbb{N}\}$ of $C(X)$. For $n \in \mathbb{N}$ we define $\Phi_n := 2^{-n}\|f_n\|_\infty^{-1} f_n$ if $f_n \neq 0$ and $\Phi_n := 0$ otherwise. Then it is easy to see that $\Phi(x) := (\Phi_n(x))_n$ satisfies $\Phi(x) \in \ell_2$ for all $x \in X$ and hence

$$k(x,x') := \bigl\langle \Phi(x), \Phi(x') \bigr\rangle_{\ell_2}, \qquad x, x' \in X,$$

defines a kernel on X with feature map $\Phi : X \to \ell_2$. Let us now fix an $f \in C(X)$ and an $\varepsilon > 0$. Then there exists an integer n with $\|f_n - f\|_\infty \leq \varepsilon$. We define $w := 2^n \|f_n\|_\infty e_n$, where (e_n) is the canonical orthonormal basis (ONB) of ℓ_2. This gives $\langle w, \Phi(x)\rangle = f_n(x)$ for all $x \in X$, and since $H := \bigl\{\langle v, \Phi(.)\rangle : v \in \ell_2\bigr\}$ is the RKHS of k we obtain the universality of k. □

Proof (of Theorem 3). Since $L_\infty(\mathrm{P}_X) \subset L_p(\mathrm{P}_X)$ we have $\mathcal{R}^*_{L,\mathrm{P},L_p(\mathrm{P}_X)} = \mathcal{R}^*_{L,\mathrm{P}}$ by Lemma 3. Now the assertion easily follows from the denseness of F in $L_p(\mathrm{P}_X)$ and the continuity of $\mathcal{R}_{L,\mathrm{P}} : L_p(\mathrm{P}_X) \to [0,\infty)$ established in Lemma 2. □

Proof (of Proposition 1). It is easy to check that T_k is the adjoint operator of the inclusion map $I : H \to L_p(\mu)$. Recalling that I has dense image if and only if its adjoint is injective (see e.g. [8, Satz III.4.5]) we then obtain the assertion. □

Proof (of Theorem 4). Let us fix an $f \in L_{p'}(\mu)$ with $T_{k_\sigma} f = 0$. For $t > 0$ we write $g_t(x,x') := (4\pi t)^{-d/2} k_{2\sqrt{t}}(x,x')$, and define

$$u(x,t) \;:=\; \int_{\mathbb{R}^d} g_t(x,x') f(x') d\mu(x'), \qquad\qquad x \in \mathbb{R}^d,\, t > 0.$$

Differentiation then shows that u satisfies the heat equation $\partial_t u = \partial_{xx} u$ and since $T_{k_\sigma} f = 0$ implies $u(x, \frac{\sigma^2}{4}) = 0$ for all $x \in \mathbb{R}^d$ the unique continuation theorem of Itô and Yamabe [9] implies that $u(x,t) = 0$ for all $x \in \mathbb{R}^d$ and $t > 0$. Now let $h : \mathbb{R}^d \to \mathbb{R}$ be a continuous function with compact support. Then we obviously have $\|h\|_\infty < \infty$, $h \in L_p(\mu)$, and

$$0 = \int_{\mathbb{R}^d} h(x) u(x,t) dx = \int_{\mathbb{R}^d} h(x) \biggl(\int_{\mathbb{R}^d} g_t(x,x') f(x') d\mu(x') \biggr) dx, \quad t > 0. \tag{8}$$

Since μ is finite it follows that $f \in L_1(\mu)$ and that $h(x) g_t(x,x') f(x')$ is integrable with respect to the product of μ and the Lebesgue measure on $\mathbb{R}^d$. Let us define

$$h_t(x') \;:=\; \int_{\mathbb{R}^d} g_t(x,x') h(x) dx, \qquad\qquad x' \in \mathbb{R}^d.$$

For $t > 0$ Fubini's theorem and (8) then yields

$$0 = \int_{\mathbb{R}^d} f(x') \biggl(\int_{\mathbb{R}^d} g_t(x,x') h(x) dx \biggr) d\mu(x') = \int_{\mathbb{R}^d} f(x') h_t(x') d\mu(x')\,. \tag{9}$$

Now fix an $x \in \mathbb{R}^d$ and an $\varepsilon > 0$. Then there exists a $\delta > 0$ such that for all $x' \in \mathbb{R}^d$ with $|x' - x| \le \delta$ we have $|h(x') - h(x)| \le \epsilon$. Since $\int_{\mathbb{R}^d} g_t(x,x') dx = 1, x' \in \mathbb{R}^d$ we hence obtain

$$h_t(x) - h(x) = \int\limits_{|x'-x| \le \delta} \bigl(h(x') - h(x) \bigr) g_t(x,x') dx' + \int\limits_{|x'-x| > \delta} \bigl(h(x') - h(x) \bigr) g_t(x,x') dx'.$$

The absolute value of the first term is bounded by ϵ and the absolute value of the second term can be made less than ϵ by choosing t small enough. Therefore we conclude that $\lim_{t\to 0} h_t(x) = h(x)$, $x \in \mathbb{R}^d$. Moreover, we have $|h_t(x)| \le \|h\|_\infty < \infty$, and hence the dominated convergence theorem and (9) yield

$$0 = \int_{\mathbb{R}^d} f(x') h(x') d\mu(x') = \langle f, h \rangle_{L_{p'}(\mu), L_p(\mu)}.$$

Since it follows from [10, Thm. 29.12] and [10, Thm. 29.14] that the continuous functions with compact support are dense in $L_p(\mu)$, we conclude $f = 0$. □

Proof (of Corollary 2). Let us fix an $f \in L_{q'}(h\mu)$. Then we have $f|h|^{\frac{1}{q'}} \in L_{q'}(\mu)$ and for r defined by $\frac{1}{q'} + \frac{1}{r} = \frac{1}{p'}$ Hölder's inequality and $\frac{r}{q} = s$ thus yield

$$\|fh\|_{L_{p'}(\mu)} \;=\; \big\| f|h|^{\frac{1}{q'}} |h|^{\frac{1}{q}} \big\|_{L_{p'}(\mu)} \;\leq\; \big\| f|h|^{\frac{1}{q'}} \big\|_{L_{q'}(\mu)} \big\| |h|^{\frac{1}{q}} \big\|_{L_r(\mu)} \;<\; \infty\,.$$

Moreover, if $f \neq 0$ in $L_{q'}(h\mu)$ we have $\mu\{fh \neq 0\} > 0$ and hence we obtain

$$0 \neq T_k(fh) = \int_X f(x)h(x)k(.,x)\,d\mu(x) = \int_X f(x)k(.,x)\,d(h\mu)(x)\,.$$

Since the latter integral describes the integral operator $L_{q'}(h\mu) \to H$ we then obtain the assertion by Proposition 1. □

Proof (of Corollary 3). Let us fix an $f \in \ell_{p'}(X)$ with $f \neq 0$. Then we have $T_k f \in H \subset \ell_p(X)$ and hence we obtain

$$\langle T_k f, f\rangle_{\ell_p(X),\ell_{p'}(X)} = \sum_{x,x'\in X} k(x,x')f(x)f(x') > 0\,.$$

This shows that $T_k : \ell_{p'}(X) \to H$ is injective. Now let μ be a distribution on X and ν be the counting measure on X. Then there exists a function $h \in \ell_1(X)$ with $\mu = h\nu$. Since for $q \in [1,p]$ we have $s := \frac{p}{p-q} \geq 1$ we then find $h \in \ell_s(X)$ and hence we obtain the assertion by applying Corollary 2. In addition, for $q > p$ we have $\|k\|_{\ell_q(X)} \leq \|k\|_{\ell_p(X)} < \infty$ and $\ell_{q'}(X) \subset \ell_{p'}(X)$, and consequently, this case follows from the already shown case $q = p$. □

Proof (of Theorem 5). The completeness of $c_0(X)$ and $k(.,x) \in c_0(X)$, $x \in X$, implies that the inclusion $I : H \to c_0(X)$ is well-defined. In addition, k is bounded and thus I is continuous. Moreover, a simple calculation shows that its adjoint operator is the integral operator $T_k : \ell_1(X) \to H$ which is injective by (7). Consequently, H is dense in $c_0(X)$, and by Lemma 2 we hence find $\mathcal{R}^*_{L,\mathrm{P},H} = \mathcal{R}^*_{L,\mathrm{P},c_0(X)}$. Therefore it remains to show that $c_0(X)$ is (L,P)-rich. To this end let ν be the counting measure on X and $h : X \to [0,1]$ be the map that satisfies $\mathrm{P}_X = h\nu$. In addition recall that we have $\mathcal{R}_{L,\mathrm{P}}(0) < \infty$ since L is a P-integrable Nemitski loss. Given an $\varepsilon > 0$ there hence exists a finite set $A \subset X$ with

$$\sum_{x\in X\setminus A} h(x) \int_Y L(x,y,0)d\mathrm{P}(y|x) \leq \varepsilon\,.$$

In addition, there exists a $g : X \to \mathbb{R}$ with $\mathcal{R}_{L,\mathrm{P}}(g) \leq \mathcal{R}^*_{L,\mathrm{P}} + \varepsilon$. Let us define $f := \mathbf{1}_A g$. Then we have $f \in c_0(X)$ and

$$\begin{aligned}\mathcal{R}_{L,\mathrm{P}}(f) &= \sum_{x\in A} h(x) \int_Y L\big(x,y,g(x)\big)d\mathrm{P}(y|x) + \sum_{x\in X\setminus A} h(x) \int_Y L(x,y,0)d\mathrm{P}(y|x) \\ &\leq \mathcal{R}_{L,\mathrm{P}}(g) + \varepsilon\,.\end{aligned}$$

□

Lemma 4. *Let μ be a distribution on X. Assume that we have a subspace $F \subset L_\infty(\mu)$ such that for all measurable $A \subset X$ there exists a sequence $(f_n) \subset F$ with $\lim_{n\to\infty} f_n(x) = \mathbf{1}_A(x)$ for μ-almost all $x \in X$. Then for all $g \in L_\infty(\mu)$ there exists a sequence $(f_n) \subset F$ with $\lim_{n\to\infty} f_n(x) = g(x)$ for μ-almost all $x \in X$.*

Proof. We first observe that for step functions $g \in L_\infty(\mu)$ the assertion immediately follows from the fact that F is a vector space. Let us now fix an arbitrary $g \in L_\infty(\mu)$. For $n \geq 1$ there then exists a step function $g_n \in L_\infty(\mu)$ with $\|g_n - g\|_\infty \leq 1/n$. Moreover, for this g_n there exists a sequence $(f_{m,n})_{m\geq 1} \subset F$ with $\lim_{m\to\infty} f_{m,n}(x) = g_n(x)$ for μ-almost all $x \in X$. By Egoroff's theorem we then find a measurable subset $A_n \subset X$ with $\mu(X\setminus A_n) \leq 1/n$ and

$$\lim_{m\to\infty} \|(f_{m,n} - g_n)_{|A_n}\|_\infty = 0\,.$$

Consequently, there is an index $m_n \geq 1$ with $\|(f_{m_n,n} - g_n)_{|A_n}\|_\infty \leq 1/n$. By putting all estimates together we now obtain

$$\mu\big(\{x \in X : |f_{m_n,n}(x) - g(x)| \leq 2/n\}\big) \geq\ 1 - 1/n\,, \qquad\qquad n \geq 1.$$

This shows that $(f_{m_n,n})_{n\geq 1}$ converges to g in probability μ, and consequently there exists a subsequence of it that converges to g almost surely. □

Proof (of Theorem 6). Let $\mathcal{A}$ be the σ-algebra of X. We fix an $A_1 \in \mathcal{A}$, and write $A_2 := \emptyset$. Let us define two distributions P_1 and P_2 on $X \times Y$ by

$$\mathrm{P}_i(\,.\,|x) := \begin{cases} \mathrm{Q}_1 & \text{if } x \in A_i \\ \mathrm{Q}_2 & \text{if } x \in X\setminus A_i \end{cases}$$

and $(\mathrm{P}_i)_X := \mu$ for $i = 1, 2$. Our assumptions on Q_1 and Q_2 guarantee $\mathcal{C}^*_{L,\mathrm{Q}_1} < \infty$ and $\mathcal{C}^*_{L,\mathrm{Q}_2} < \infty$, and hence we find $\mathcal{R}^*_{L,\mathrm{P}_i} < \infty$ for $i = 1, 2$. Moreover, every function minimizing $\mathcal{R}_{L,\mathrm{P}_i}$ has μ-almost surely the form

$$f^*_{L,\mathrm{P}_i} := t^*_1 \mathbf{1}_{A_i} + t^*_2 \mathbf{1}_{X\setminus A_i}\,, \qquad\qquad i = 1, 2.$$

Now, our assumptions yield $\mathcal{R}^*_{L,\mathrm{P}_i,F} = \mathcal{R}^*_{L,\mathrm{P}_i}$, $i = 1, 2$, and hence there are sequences $(f^{(1)}_n) \subset F$ and $(f^{(2)}_n) \subset F$ with $\lim_{n\to\infty} \mathcal{R}_{L,\mathrm{P}_i}(f^{(i)}_n) = \mathcal{R}^*_{L,\mathrm{P}_i}$ for $i = 1, 2$. By Remark 2.39 in [5] we then have

$$\lim_{n\to\infty} f^{(i)}_n = f^*_{L,\mathrm{P}_i}\,, \qquad\qquad i = 1, 2, \tag{10}$$

in probability $\hat{\mu}$, where $\hat{\mu}$ is the extension of μ to the μ-completed σ-algebra of $\mathcal{A}$. Now observe that all functions in (10) are $\mathcal{A}$-measurable, and hence (10) actually holds in probability μ. Consequently, there exist subsequences $(f^{(1)}_{n_j})$ and $(f^{(2)}_{n_j})$ for which (10) holds μ-almost surely. For

$$f_j := \frac{1}{t^*_1 - t^*_2}\big(f^{(1)}_{n_j} - f^{(2)}_{n_j}\big)\,, \qquad\qquad j \geq 1,$$

we then have $f_j \in F$, and in addition our construction yields

$$\lim_{j\to\infty} f_j = \frac{1}{t_1^* - t_2^*}\bigl(f^*_{L,\mathrm{P}_1} - f^*_{L,\mathrm{P}_2}\bigr) = \frac{1}{t_1^* - t_2^*}\bigl(t_1^*\mathbf{1}_{A_1} + t_2^*\mathbf{1}_{X\setminus A_1} - t_2^*\mathbf{1}_X\bigr) = \mathbf{1}_{A_1}$$

μ-almost surely. By Lemma 4 we thus obtain the assertion. □

Proof (of Corollary 4). Let $x_1, \ldots, x_n \in X$ be mutually different points and μ be the associated empirical distribution. Obviously, it suffices to show that the kernel matrix $K := (k(x_i, x_j))$ has full rank. Let us assume the converse, i.e. that there exists an $y \in \mathbb{R}^n$ with $K\alpha \neq y$ for all $\alpha \in \mathbb{R}^n$. Since $K\mathbb{R}^n$ is closed there exists an $\varepsilon > 0$ with $\|K\alpha - y\|_\infty \geq \varepsilon$ for all $\alpha \in \mathbb{R}^n$. Moreover, by decomposing H into $\operatorname{span}\{k(.,x_i) : i = 1,\ldots,n\}$ and its orthogonal complement we see that for every $f \in H$ there is an $\alpha \in \mathbb{R}^n$ with

$$f(x_j) = \sum_{i=1}^n \alpha_i k(x_j, x_i)\,, \qquad\qquad j = 1,\ldots,n,$$

and hence for all $f \in H$ there is an index $j \in \{1,\ldots,n\}$ with $|f(x_j) - y_j| > \varepsilon$. On the other hand, Theorem 6 gives a sequence $(f_n) \subset H$ with $f_n(x_i) \to y_i$ for all $i \in \{1,\ldots,n\}$. Since $\{1,\ldots,n\}$ is finite we then easily find a contradiction. □

Lemma 5. *Let μ be a distribution on X and $p > 0$. Assume that $F \subset L_p(\mu)$ is a linear subspace such that for all measurable $A \subset X$ there exists a sequence $(f_n) \subset F$ with $\lim_{n\to\infty} \|f_n - \mathbf{1}_A\|_{L_p(\mu)} = 0$. Then F is dense in $L_p(\mu)$.*

Proof. If $g \in L_p(\mu)$ is a measurable step function there obviously exists a sequence $(f_n) \subset F$ with $\lim_{n\to\infty} \|f_n - g\|_p = 0$. Moreover, if $g \in L_p(\mu)$ is bounded and n is an integer there exists a measurable step function g_n with $\|g_n - g\|_\infty \leq 1/n$. In addition, we have just see that there exists an $f_n \in F$ with $\|f_n - g_n\|_p \leq 1/n$, and hence we find $\lim_{n\to\infty} \|f_n - g\|_p = 0$. Finally, for general $g \in L_p(\mu)$ we then find an approximating sequence by first approximating g with the bounded measurable functions $g_n := \mathbf{1}_{|g|\leq n} g$, $n \geq 1$, and then approximating these g_n with suitable functions $f_n \in F$. □

Proof (of Theorem 7). Following the argument used in the proof of Theorem 6 we may assume without loss of generality that X is a complete measurable space. Let us now fix a measurable $A_1 \subset X$, and write $A_2 := \emptyset$. Furthermore, we define the distributions P_i, the functions f^*_{L,P_i}, and the approximating sequences $(f_n^{(i)}) \subset F$, $i = 1, 2$, as in the proof of Theorem 6. Then $\lim_{n\to\infty} \mathcal{R}_{L,\mathrm{P}_i}(f_n^{(i)}) = \mathcal{R}^*_{L,\mathrm{P}_i}$, $i = 1, 2$, together with Remark 2.41 in [5] yields

$$\lim_{n\to\infty} \|f_n^{(i)} - f^*_{L,\mathrm{P}_i}\|_{L_p(\mu)} = 0\,.$$

For $f_n := \frac{1}{t_1^* - t_2^*}\bigl(f_n^{(1)} - f_n^{(2)}\bigr)$, $n \geq 1$, we then obtain $\lim_{n\to\infty} \|f_n - \mathbf{1}_{A_1}\|_{L_p(\mu)} = 0$, and hence we obtain the assertion by Lemma 5. □

Proof (of Theorem 9). Let us write $p_n := \mu(\{n\})$, $n \in \mathbb{N}_0$. Moreover, let $(b_i)_{i\geq 1}$ be a strictly positive sequence with $\|(b_i)\|_2 = 1$ and $(b_i) \in \ell_1$. Furthermore, let (e_n) be the canonical ONB of ℓ_2. We write $\Phi(0) := (b_i)$ and $\Phi(n) := e_n$, $n \geq 1$. Then we have $\Phi(n) \in \ell_2$ for all $n \in \mathbb{N}_0$ and hence

$$k(n,m) := \bigl\langle \Phi(n), \Phi(m) \bigr\rangle_{\ell_2}\,, \qquad\qquad n,m \geq 0,$$

defines a kernel. Obviously, $\{\Phi(n) : n \geq 1\}$ is (algebraically) linearly independent and from this it is easy to conclude that k is strictly positive definite. Moreover, an easy calculation shows $k(0,0) = 1$, $k(n,m) = \delta_{n,m}$, and $k(n,0) = b_n$ for $n,m \geq 1$. Since $b_n \to 0$ we hence find $k(.,n) \in c_0(X)$ for all $n \in \mathbb{N}_0$. Let us define $f : \mathbb{N}_0 \to \mathbb{R}$ by $f(0) := 1$ and $f(n) := -\frac{b_n}{p_n} p_0$ for $n \geq 1$. Then we have $\|f\|_{L_1(\mu)} = p_0 + p_0 \|(b_i)\|_{\ell_1} < \infty$, and a simple calculation shows

$$T_k f(0) = k(0,0) f(0) p_0 + \sum_{n=1}^{\infty} k(0,n) f(n) p_n = p_0 - p_0 \sum_{n=1}^{\infty} b_n^2 = 0\,.$$

Moreover, for $m \geq 1$ our construction yields

$$T_k f(m) = k(m,0) f(0) p_0 + \sum_{n=1}^{\infty} k(m,n) f(n) p_n = b_m f(0) p_0 - f(m) p_m = 0\,. \quad \square$$

References

1. Steinwart, I.: On the influence of the kernel on the consistency of support vector machines. J. Mach. Learn. Res. **2** (2001) 67–93
2. Bousquet, O., Elisseeff, A.: Stability and generalization. J. Mach. Learn. Res. **2** (2002) 499–526
3. Steinwart, I.: Consistency of support vector machines and other regularized kernel machines. IEEE Trans. Inform. Theory **51** (2005) 128–142
4. Zhang, T.: Statistical behaviour and consistency of classification methods based on convex risk minimization. Ann. Statist. **32** (2004) 56–134
5. Steinwart, I.: How to compare loss functions and their risks. Technical Report (2005) `http://www.c3.lanl.gov/~ingo/pubs.shtml`.
6. Schurle, A.: Topics in Topology. Elsevier North Holland (1979)
7. Conway, J.B.: A Course in Functional Analysis. 2nd edn. Springer (1990)
8. Werner, D.: Funktionalanalysis. Springer, Berlin (1995)
9. Itô, S., Yamabe, H.: A unique continuation theorem for solutions of a parabolic differential equation. J. Math. Soc. Japan **10** (1958) 314–321
10. Bauer, H.: Measure and Integration Theory. De Gruyter, Berlin (2001)