Statistics and Its Interface Volume 2 (2009) 311–327

On consistency and robustness properties of Support Vector Machines for heavy-tailed distributions*

Andreas Christmann†, Arnout Van Messem, and Ingo Steinwart

Support Vector Machines (SVMs) are known to be consistent and robust for classification and regression if they are based on a Lipschitz continuous loss function and on a bounded kernel with a dense and separable reproducing kernel Hilbert space. These facts are even true in the regression context for unbounded output spaces, if the target function f is integrable with respect to the marginal distribution of the input variable X and if the output variable Y has a finite first absolute moment. The latter assumption clearly excludes distributions with heavy tails, e.g., several stable distributions or some extreme value distributions which occur in financial or insurance projects. The main point of this paper is that we can enlarge the applicability of SVMs even to heavy-tailed distributions, which violate this moment condition. Results on existence, uniqueness, representation, consistency, and statistical robustness are given.

AMS 2000 subject classifications: Primary 62G05; secondary 62G35, 62G08, 68Q32, 62G20, 68T10, 62J02.

1. INTRODUCTION

The goal in non-parametric statistical machine learning, both for classification and for regression purposes, is to relate an $\mathcal{X}$-valued input random variable X to a $\mathcal{Y}$-valued output random variable Y, under the assumption that the joint distribution P of (X,Y) is (almost) completely unknown. Common choices of input and output spaces are $\mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \{-1,+1\}$ for classification and $\mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \mathbb{R}$ for regression. In order to model this relationship one typically assumes that one has a training data set $D_{train} = ((x_1,y_1),\ldots,(x_n,y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$ with observations from independent and identically distributed (i.i.d.) random variables (X_i,Y_i), $i = 1,\ldots,n$, which all have the same distribution P on $\mathcal{X} \times \mathcal{Y}$ equipped with the corresponding Borel σ-algebra. Informally, the aim is to build a predictor $f : \mathcal{X} \to \mathcal{Y}$ based on these observations such that $f(X)$ is a good approximation of Y.

*We would like to thank Ursula Gather and Xuming He for drawing our attention to the $L^\star$-trick.
†Corresponding author.

To formalize this aim we call a function $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0,\infty)$ a *loss function* (or just *loss*) if L is measurable. The loss function assesses the quality of a prediction $f(x)$ for an observed output value y by $L(x,y,f(x))$. We follow the convention that the smaller $L(x,y,f(x))$ is, the better the prediction is. We will further always assume that $L(x,y,y) = 0$ for all $y \in \mathcal{Y}$, because practitioners usually argue that the loss is zero, if the forecast $f(x)$ equals the observed value y.

The quality of a predictor f is measured by the expectation of the loss function, i.e., by the *L-risk*

$$\mathcal{R}_{L,\mathrm{P}}(f) := \mathbb{E}_{\mathrm{P}} L(X,Y,f(X)).$$

One tries to find a predictor whose risk is close to the minimal risk, i.e., close to the Bayes risk

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

One way to build a non-parametric predictor f is to use a support vector machine

$$f_{L,\mathrm{P},\lambda} := \arg\inf_{f \in \mathcal{H}} \mathcal{R}_{L,\mathrm{P}}(f) + \lambda \|f\|^2_{\mathcal{H}}\,, \tag{1}$$

where L is a loss function, $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) of a measurable *kernel* $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and $\lambda > 0$ is a regularization parameter to reduce the danger of overfitting, see e.g., Vapnik [1998] and Schölkopf and Smola [2002]. The *reproducing property* states, for all $f \in \mathcal{H}$ and all $x \in \mathcal{X}$,

$$f(x) = \langle f, \Phi(x) \rangle_{\mathcal{H}}.$$

A kernel k is called *bounded*, if

$$\|k\|_\infty := \sup\{\sqrt{k(x,x)} : x \in \mathcal{X}\} < \infty\,.$$

Using the reproducing property and $\|\Phi(x)\|_{\mathcal{H}} = \sqrt{k(x,x)}$, we obtain the well-known inequalities

$$\|f\|_\infty \le \|k\|_\infty \|f\|_{\mathcal{H}} \tag{2}$$

and

$$\|\Phi(x)\|_\infty \le \|k\|_\infty \|\Phi(x)\|_{\mathcal{H}} \le \|k\|^2_\infty \tag{3}$$

for $f \in \mathcal{H}$ and $x \in \mathcal{X}$. As an example of a bounded kernel we mention the popular Gaussian radial basis function (RBF) kernel defined by

$$k_{\mathrm{RBF}}(x, x') = \exp(-\gamma^{-2} \|x - x'\|^2), \quad x, x' \in \mathcal{X}, \tag{4}$$

where γ is a positive constant. Furthermore, it is *universal* in the sense of Steinwart [2001], that is, its RKHS is dense in $C(\mathcal{X})$ for all compact $\mathcal{X} \subset \mathbb{R}^d$. Finally, see Theorem 4.63 of Steinwart and Christmann [2008b], its RKHS is dense in $L_1(\mu)$ for all probability measures μ on $\mathbb{R}^d$.

Of course, the regularized risk

$$\mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(f) := \mathcal{R}_{L,\mathrm{P}}(f) + \lambda \|f\|^2_{\mathcal{H}}$$

is in general not computable, because P is unknown. However, the empirical distribution

$$\mathrm{D} = \frac{1}{n} \sum_{i=1}^{n} \delta_{(x_i, y_i)}$$

corresponding to the data set D can be used as an estimator of P. Here $\delta_{(x_i,y_i)}$ denotes the Dirac distribution in (x_i, y_i). If we replace P by D in (1), we obtain the regularized empirical risk $\mathcal{R}^{reg}_{L,\mathrm{D},\lambda}(f)$ and the empirical SVM $f_{L,\mathrm{D},\lambda}$.

SVMs based on a *convex* loss function have under weak assumptions at least the following four advantageous properties, which partially explain their success, see e.g., Vapnik [1998], Cristianini and Shawe-Taylor [2000], Schölkopf and Smola [2002], and Steinwart and Christmann [2008b] for details. *(i)* An SVM $f_{L,\mathrm{P},\lambda}$ exists and is the unique solution of a certain convex problem. *(ii)* SVMs are L-risk consistent, i.e., for suitable null-sequences (λ_n) with $\lambda_n > 0$ we have

$$\mathcal{R}_{L,\mathrm{P}}(f_{L,\mathrm{D},\lambda_n}) \to \mathcal{R}^*_{L,\mathrm{P}}, \qquad n \to \infty,$$

in probability. *(iii)* SVMs have good statistical robustness properties, if k is *bounded* in the sense of $\|k\|_\infty := \sup\{\sqrt{k(x,x)} : x \in \mathcal{X}\} < \infty$ and if L is *Lipschitz continuous* with respect to its third argument, i.e., there exists a constant $|L|_1 \in (0, \infty)$ such that, for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and all $t_1, t_2 \in \mathbb{R}$,

$$|L(x, y, t_1) - L(x, y, t_2)| \le |L|_1 \, |t_1 - t_2|. \tag{5}$$

In a nutshell, robustness implies that $f_{L,\mathrm{P},\lambda}$ only varies in a smooth and bounded manner if P changes slightly in the set $\mathcal{M}_1$ of all probability measures on $\mathcal{X} \times \mathcal{Y}$. *(iv)* There exist efficient numerical algorithms to determine $f_{L,\mathrm{D},\lambda}$ even for large and high-dimensional data sets D.

If $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ only depends on its last two arguments, i.e., if there exists a measurable function $\hat{L} : \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ such that $L(x, y, t) = \hat{L}(y, t)$ for all $(x, y, t) \in \mathcal{X} \times \mathcal{Y} \times \mathbb{R}$, then L is called a *supervised loss*. A loss function L is called a *Nemitski loss* if there exists a measurable function $b : \mathcal{X} \times \mathcal{Y} \to [0, \infty)$ and an increasing function $h : [0, \infty) \to [0, \infty)$ such that

$$L(x, y, t) \le b(x, y) + h(|t|), \quad (x, y, t) \in \mathcal{X} \times \mathcal{Y} \times \mathbb{R}.$$

If additionally $b \in \mathcal{L}_1(\mathrm{P})$, we say that L is a P-*integrable Nemitski loss*.

If not otherwise mentioned, we will restrict attention to Lipschitz continuous (w.r.t. the third argument) loss functions L for three reasons. *(i)* Many loss functions used in practice are Lipschitz continuous, e.g., the *hinge loss*

$$L(x, y, t) := \max\{0, 1 - yt\}$$

and the *logistic loss*

$$L(x, y, t) := \ln(1 + \exp(-yt)) \tag{6}$$

for classification; the ϵ-*insensitive loss*

$$L(x, y, t) := \max\{0, |y - t| - \epsilon\}$$

for some $\epsilon > 0$, *Huber's loss*

$$L(x, y, t) := \begin{cases} 0.5(y - t)^2 & \text{if } |y - t| \le \alpha \\ \alpha |y - t| - 0.5\alpha^2 & \text{if } |y - t| > \alpha \end{cases}$$

for some $\alpha > 0$, and the *logistic loss*

$$L(x, y, t) := -\ln \frac{4 \exp(y - t)}{(1 + \exp(y - t))^2} \tag{7}$$

for regression; and the *pinball loss*

$$L(x, y, t) := \begin{cases} (\tau - 1)(y - t), & \text{if } y - t < 0, \\ \tau(y - t), & \text{if } y - t \ge 0, \end{cases} \tag{8}$$

for some $\tau > 0$ for quantile regression. *(ii)* Lipschitz continuous loss functions are trivially Nemitski loss functions for *all* probability measures on $\mathcal{X} \times \mathcal{Y}$, because

$$\begin{aligned} L(x, y, t) &= L(x, y, 0) + L(x, y, t) - L(x, y, 0) \\ &\le b(x, y) + |L|_1 \, |t|, \end{aligned}$$

where $b(x, y) := L(x, y, 0)$ for $(x, y, t) \in \mathcal{X} \times \mathcal{Y} \times \mathbb{R}$ and $|L|_1 \in (0, \infty)$ denotes the Lipschitz constant of L. Furthermore, Lipschitz continuous L are P-integrable if $\mathcal{R}_{L,\mathrm{P}}(0)$ is finite. *(iii)* SVMs based on the combination of a Lipschitz continuous loss and a bounded kernel have good statistical robustness properties for classification and regression, see Christmann and Steinwart [2004, 2007] and Christmann and Van Messem [2008].

Let us assume that the probability measure P can be split up into the marginal distribution P_X on $\mathcal{X}$ and the

conditional probability $P(y|x)$ on $\mathcal{Y}$, which is possible if $\mathcal{Y} \subset \mathbb{R}$ is closed. Then we obtain for the L-risk the inequality

$$\begin{aligned}(9)\qquad \mathcal{R}_{L,P}(f) &= \mathbb{E}_P\big(L(X,Y,f(X)) - L(X,Y,Y)\big)\\ &\le |L|_1 \int_{\mathcal{X}}\int_{\mathcal{Y}} |f(x)-y|\, dP(y|x)\, dP_X(x)\\ &\le |L|_1 \int_{\mathcal{X}} |f(x)|\, dP_X(x)\\ &\quad + |L|_1 \int_{\mathcal{X}}\int_{\mathcal{Y}} |y| dP(y|x)\, dP_X(x),\end{aligned}$$

which is finite, if $f \in L_1(P_X)$ *and*

$$(10)\qquad \mathbb{E}_P|Y| = \int_{\mathcal{X}}\int_{\mathcal{Y}} |y|\, dP(y|x)\, dP_X(x) < \infty\,.$$

The latter condition excludes heavy-tailed distributions such as many stable distributions, including the Cauchy distribution, and many extreme value distributions which occur in financial or actuarial problems. The moment condition (10) is one of the assumptions made by Christmann and Steinwart [2007] and Steinwart and Christmann [2008b] for their consistency and robustness proofs of SVMs for an *unbounded* output set $\mathcal{Y}$.

The main point of this paper is to enlarge the applicability of SVMs even to heavy-tailed distributions, which violate the moment condition $\mathbb{E}_P|Y| < \infty$, by using a trick well-known in the literature on robust statistics, see e.g., Huber [1967]: we shift the loss $L(x,y,t)$ downwards by the amount of $L(x,y,0) \in [0,\infty)$. We will call the function $L^\star : \mathcal{X}\times\mathcal{Y}\times\mathbb{R} \to \mathbb{R}$ defined by

$$(11)\qquad L^\star(x,y,t) := L(x,y,t) - L(x,y,0)$$

the *shifted loss function* or the *shifted version of* L. We obtain, for all $f \in L_1(P_X)$,

$$\begin{aligned}(12)\qquad &\mathbb{E}_P L^\star(X,Y,f(X))\\ &= \mathbb{E}_P\big(L(X,Y,f(X)) - L(X,Y,0)\big)\\ &\le \int_{\mathcal{X}\times\mathcal{Y}} |L(x,y,f(x)) - L(x,y,0)|\, dP(x,y)\\ &\le |L|_1 \int_{\mathcal{X}} |f(x)|\, dP_X(x) < \infty,\end{aligned}$$

no matter whether the moment condition (10) is fulfilled. We will use this "$L^\star$-trick" to show that many important results on the SVM $f_{L,P,\lambda}$, such as existence, uniqueness, representation, consistency, and statistical robustness, can also be shown for

$$(13)\qquad f_{L^\star,P,\lambda} := \arg\inf_{f\in\mathcal{H}} \mathcal{R}_{L^\star,P}(f) + \lambda\|f\|_{\mathcal{H}}^2\,,$$

where

$$\mathcal{R}_{L^\star,P}(f) := \mathbb{E}_P L^\star(X,Y,f(X))$$

denotes the $L^\star$-risk of f. Moreover, we will show that

$$f_{L^\star,P,\lambda} = f_{L,P,\lambda}$$

if $f_{L,P,\lambda}$ exists. Hence, there is no need for new algorithms to compute $f_{L^\star,D,\lambda}$ because the empirical SVM $f_{L,D,\lambda}$ exists for all data sets D. The advantage of $f_{L^\star,P,\lambda}$ over $f_{L,P,\lambda}$ is that $f_{L^\star,P,\lambda}$ is still well-defined and useful for heavy-tailed conditional distributions $P(y|x)$, for which the first absolute moment $\int_{\mathcal{Y}} |y| dP(y|x)$ is infinite. In particular, our results will show that even in the case of heavy-tailed distributions, the forecasts $f_{L^\star,D,\lambda}(x) = f_{L,D,\lambda}(x)$ are consistent and robust, if the kernel is bounded and a Lipschitz continuous loss function such as, e.g., the pinball loss for quantile regression is used.

The paper is organized as follows. Section 2 gives some simple facts on $L^\star$ and on $\mathcal{R}_{L^\star,P}(f_{L^\star,P,\lambda})$ and their counterparts with respect to L. Section 3 contains our main results, i.e., existence, uniqueness, a representation theorem, risk consistency, and statistical robustness of SVMs based on $L^\star$. Section 4 contains a discussion. All proofs together with some general facts are given in the Appendix.

2. SHIFTED LOSS FUNCTIONS

In this section we will give some general facts on the function $L^\star$ which will be used to obtain our main results in the next section. Our general assumptions for the rest of the paper are summarized in

Assumption 1. Let $n \in \mathbb{N}$, $\mathcal{X}$ be a complete separable metric space (e.g., a closed $\mathcal{X} \subset \mathbb{R}^d$), $\mathcal{Y} \subset \mathbb{R}$ be a non-empty and closed set, and P be a probability distribution on $\mathcal{X}\times\mathcal{Y}$ enclipped with its Borel σ-algebra. Since $\mathcal{Y}$ is closed, P can be split up into the marginal distribution P_X on $\mathcal{X}$ and the conditional probability $P(y|x)$ on $\mathcal{Y}$. Let $L : \mathcal{X}\times\mathcal{Y}\times\mathbb{R} \to [0,\infty)$ be a loss function and define its **shifted loss function** $L^\star : \mathcal{X}\times\mathcal{Y}\times\mathbb{R} \to \mathbb{R}$ by

$$L^\star(x,y,t) := L(x,y,t) - L(x,y,0).$$

We say that L (or $L^\star$) is convex, Lipschitz continuous, continuous or differentiable, if L (or $L^\star$) has this property with respect to its third argument. If not otherwise mentioned, $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ is a measurable kernel with reproducing kernel Hilbert space $\mathcal{H}$ of measurable functions $f : \mathcal{X} \to \mathbb{R}$, and $\Phi : \mathcal{X} \to \mathcal{H}$ denotes the canonical feature map, i.e., $\Phi(x) := k(\cdot,x)$ for $x \in \mathcal{X}$.

Obviously, $L^\star < \infty$. As shown in the introduction, we obtain by (9) that the L-risk $\mathbb{E}_P L(X,Y,f(X))$ is finite, if $f \in L_1(P_X)$ and $\mathbb{E}_P|Y| < \infty$. On the other hand, (12) shows us that $\mathbb{E}_P L^\star(X,Y,f(X))$ is finite, if $f \in L_1(P_X)$ no matter whether $\mathbb{E}_P|Y| < \infty$ is finite or infinite. Therefore, by using the $L^\star$-trick, we can enlarge the applicability of SVMs by relaxing the finiteness of the risk.

The following result gives a relationship between L and $L^\star$ in terms of convexity and Lipschitz continuity.

Proposition 2. *Let L be a loss function. Then the following statements are valid.*

i) $L^\star$ *is (strictly) convex, if* L *is (strictly) convex.*
ii) $L^\star$ *is Lipschitz continuous, if* L *is Lipschitz continuous. Furthermore, both Lipschitz constants are equal, i.e.,* $|L|_1 = |L^\star|_1$.

It follows from Proposition 2 and the strict convexity of the mapping $f \mapsto \lambda \|f\|_{\mathcal{H}}^2$, $f \in \mathcal{H}$, that $L^\star(x,y,\cdot) + \lambda \|\cdot\|_{\mathcal{H}}^2$ is a strictly convex function if L is convex.

Proposition 3. *The following assertions are valid.*

i) $\inf_{t\in\mathbb{R}} L^\star(x,y,t) \le 0$.
ii) If L *is a Lipschitz continuous loss, then for all* $f \in \mathcal{H}$*:*

$$-|L|_1 \mathbb{E}_{\mathrm{P}_X}|f(X)| \le \mathcal{R}_{L^\star,\mathrm{P}}(f) \le |L|_1 \mathbb{E}_{\mathrm{P}_X}|f(X)|, \tag{14}$$

$$\begin{aligned} -|L|_1 \mathbb{E}_{\mathrm{P}_X}|f(X)| + \lambda \|f\|_{\mathcal{H}}^2 \\ \le \mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f) \le |L|_1 \mathbb{E}_{\mathrm{P}_X}|f(X)| + \lambda \|f\|_{\mathcal{H}}^2 . \end{aligned} \tag{15}$$

iii) $\inf_{f\in\mathcal{H}} \mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f) \le 0$ *and* $\inf_{f\in\mathcal{H}} \mathcal{R}_{L^\star,\mathrm{P}}(f) \le 0$.
iv) Let L *be a Lipschitz continuous loss and assume that* $f_{L^\star,\mathrm{P},\lambda}$ *exists. Then we have*

$$\begin{aligned} \lambda \|f_{L^\star,\mathrm{P},\lambda}\|_{\mathcal{H}}^2 \le -\mathcal{R}_{L^\star,\mathrm{P}}(f_{L^\star,\mathrm{P},\lambda}) \le \mathcal{R}_{L,\mathrm{P}}(0), \\ 0 \le -\mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f_{L^\star,\mathrm{P},\lambda}) \le \mathcal{R}_{L,\mathrm{P}}(0), \end{aligned}$$

$$\lambda \|f_{L^\star,\mathrm{P},\lambda}\|_{\mathcal{H}}^2 \le \min\bigl\{|L|_1 \mathbb{E}_{\mathrm{P}_X}|f_{L^\star,\mathrm{P},\lambda}(X)|, \mathcal{R}_{L,\mathrm{P}}(0)\bigr\}. \tag{16}$$

If the kernel k *is additionally bounded, then*

$$\|f_{L^\star,\mathrm{P},\lambda}\|_\infty \le \lambda^{-1}|L|_1 \|k\|_\infty^2 < \infty, \tag{17}$$

$$|\mathcal{R}_{L^\star,\mathrm{P}}(f_{L^\star,\mathrm{P},\lambda})| \le \lambda^{-1}|L|_1^2 \|k\|_\infty^2 < \infty. \tag{18}$$

v) If the partial Fréchet- and Bouligand-derivatives[1] *of* L *and* $L^\star$ *exist for* $(x,y) \in \mathcal{X}\times\mathcal{Y}$*, then*

$$\nabla_3^F L^\star(x,y,t) = \nabla_3^F L(x,y,t), \quad \forall\, t \in \mathbb{R}, \tag{19}$$

$$\nabla_3^B L^\star(x,y,t) = \nabla_3^B L(x,y,t), \quad \forall\, t \in \mathbb{R}. \tag{20}$$

The following proposition ensures that the optimization problem to determine $f_{L^\star,\mathrm{P},\lambda}$ is well-posed.

Proposition 4. *Let L be a Lipschitz continuous loss and $f \in L_1(\mathrm{P}_X)$. Then $\mathcal{R}_{L^\star,\mathrm{P}}(f) \notin \{-\infty,+\infty\}$. Moreover, we have $\mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f) > -\infty$ for all $f \in L_1(\mathrm{P}_X) \cap \mathcal{H}$.*

3. MAIN RESULTS

This section contains our main results on the SVM $f_{L^\star,\mathrm{P},\lambda}$, namely existence, uniqueness, representation theorem, consistency, and statistical robustness.

[1]See Appendix A.1.3.

Theorem 5 (Uniqueness of SVM)**.** *Let L be a convex loss function. Assume that (i) $\mathcal{R}_{L^\star,\mathrm{P}}(f) < \infty$ for some $f \in \mathcal{H}$ and $\mathcal{R}_{L^\star,\mathrm{P}}(f) > -\infty$ for all $f \in \mathcal{H}$ or (ii) L is Lipschitz continuous and $f \in L_1(\mathrm{P}_X)$ for all $f \in \mathcal{H}$. Then for all $\lambda > 0$ there exists at most one SVM solution $f_{L^\star,\mathrm{P},\lambda}$.*

Theorem 6 (Existence of SVM)**.** *Let L be a Lipschitz continuous and convex loss function and let $\mathcal{H}$ be the RKHS of a bounded measurable kernel k. Then for all $\lambda > 0$ there exists an SVM solution $f_{L^\star,\mathrm{P},\lambda}$.*

The application of the $L^\star$-trick is superfluous if $\mathcal{R}_{L,\mathrm{P}}(0) < \infty$, because in this case we obtain

$$\begin{aligned} &\mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f_{L^\star,\mathrm{P},\lambda}) \\ &\quad= \inf_{f\in\mathcal{H}} \mathbb{E}_{\mathrm{P}}\bigl(L(X,Y,f(X)) - L(X,Y,0)\bigr) + \lambda\|f\|_{\mathcal{H}}^2 \\ &\quad= \inf_{f\in\mathcal{H}} \bigl(\mathbb{E}_{\mathrm{P}} L(X,Y,f(X)) + \lambda\|f\|_{\mathcal{H}}^2\bigr) - \mathbb{E}_{\mathrm{P}} L(X,Y,0) \\ &\quad= \mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(f_{L,\mathrm{P},\lambda}) - \mathcal{R}_{L,\mathrm{P}}(0) \end{aligned}$$

and $\mathcal{R}_{L,\mathrm{P}}(0)$ is finite and independent of f. Hence, $f_{L^\star,\mathrm{P},\lambda} = f_{L,\mathrm{P},\lambda}$ if $\mathcal{R}_{L,\mathrm{P}}(0) < \infty$.

A loss function $L : \mathcal{X}\times\mathcal{Y}\times\mathbb{R} \to [0,\infty)$ is called *distance-based*, if there exists a representing function $\psi : \mathbb{R} \to \mathbb{R}$ with $L(x,y,t) = \psi(y-t)$ for all $(x,y,t) \in \mathcal{X}\times\mathcal{Y}\times\mathbb{R}$ and $\psi(0) = 0$. Such loss functions are often used in regression. If L is a distance-based loss, $L^\star$ does not necessarily share this property.[2]

The following result gives a useful representation of $f_{L^\star,\mathrm{P},\lambda}$ and shows that the mapping $\mathrm{P} \mapsto f_{L^\star,\mathrm{P},\lambda}$ behaves similar to a Lipschitz continuous function. The subdifferential of $L^\star$ is denoted by $\partial L^\star$, see Definition 20.

Theorem 7 (Representer theorem)**.** *Let L be a convex and Lipschitz continuous loss function, k be a bounded and measurable kernel with separable RKHS $\mathcal{H}$. Then, for all $\lambda > 0$, there exists an $h \in \mathcal{L}_\infty(\mathrm{P})$ such that*

$$h(x,y) \in \partial L^\star(x,y,f_{L^\star,\mathrm{P},\lambda}(x)) \ \forall\,(x,y), \tag{21}$$

$$f_{L^\star,\mathrm{P},\lambda} = -(2\lambda)^{-1}\mathbb{E}_{\mathrm{P}}(h\Phi), \tag{22}$$

$$\|h\|_\infty \le |L|_1, \tag{23}$$

$$\bigl\|f_{L^\star,\mathrm{P},\lambda} - f_{L^\star,\bar{\mathrm{P}},\lambda}\bigr\|_{\mathcal{H}} \le \lambda^{-1}\bigl\|\mathbb{E}_{\mathrm{P}}(h\Phi) - \mathbb{E}_{\bar{\mathrm{P}}}(h\Phi)\bigr\|_{\mathcal{H}}, \tag{24}$$

for all distributions $\bar{\mathrm{P}}$ on $\mathcal{X}\times\mathcal{Y}$. If L is additionally distance-based, we obtain for (21) that

$$h(x,y) \in -\partial\psi(y - f_{L^\star,\mathrm{P},\lambda}(x)) \ \forall\,(x,y). \tag{25}$$

The next result shows that the $L^\star$-risk of the SVM $f_{L^\star,\mathrm{D},\lambda_n}$ stochastically converges for $n \to \infty$ to the smallest possible risk, i.e., to the Bayes risk. This is somewhat astonishing at first glance because $f_{L^\star,\mathrm{D},\lambda_n}$ is evaluated

[2]For the least squares loss $L(x,y,t) = (y-t)^2$ we obtain $L^\star(x,y,t) = (y-t)^2 + (y-0)^2 = t(t-2y)$ which clearly cannot be written as a function in $y-t$ only.

by minimizing a *regularized empirical risk* over the RKHS $\mathcal{H}$, whereas the Bayes risk is defined as the minimal *non-regularized* risk over the broader set of *all* measurable functions $f : \mathcal{X} \to \mathbb{R}$.

Theorem 8 (Risk consistency)**.** *Let L be a convex, Lipschitz continuous loss function, $L^\star$ its shifted version, and $\mathcal{H}$ be a separable RKHS of a bounded measurable kernel k such that $\mathcal{H}$ is dense in $L_1(\mu)$ for all distributions μ on $\mathcal{X}$. Let (λ_n) be a sequence of strictly positive numbers with $\lambda_n \to 0$.*

i) If $\lambda_n^2 n \to \infty$, then, for all $\mathrm{P} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$,

$$\mathcal{R}_{L^\star,\mathrm{P}}(f_{L^\star,\mathrm{D},\lambda_n}) \to \mathcal{R}^*_{L^\star,\mathrm{P}}, \quad n \to \infty, \tag{26}$$

in probability P^∞ for all $|D| = n$.

ii) If $\lambda_n^{2+\delta} n \to \infty$ for some $\delta > 0$, then the convergence in (26) holds even P^∞-almost surely.

In general, it is unclear whether the convergence of the risks in (26) implies the convergence of $f_{L^\star,\mathrm{D},\lambda_n}$ to a minimizer $f^*_{L^\star,\mathrm{P}}$ of the Bayes risk $\mathcal{R}^*_{L^\star,\mathrm{P}}$. However, Theorem 9 will show such a convergence for the important special case of nonparametric quantile regression. Estimation of conditional quantiles instead of estimation of conditional means is especially interesting for heavy-tailed distributions that often have no finite moments. It is known that the pinball loss function defined in (8) can be used to estimate the conditional τ-quantiles, $\tau \in (0,1)$,

$$f^*_{\tau,\mathrm{P}}(x) := \big\{t^* \in \mathbb{R} : \ \mathrm{P}\big((-\infty, t^*] \,|x\big) \geq \tau \text{ and } \mathrm{P}\big([t^*, \infty) \,|x\big) \geq 1 - \tau\big\},$$

$x \in \mathcal{X}$, see Koenker [2005] and Takeuchi *et al.* [2006]. For some recent result on SVMs based on this loss function we refer to Christmann and Steinwart [2008] and Steinwart and Christmann [2008a]. The pinball loss function is convex and Lipschitz continuous, but asymmetric for $\tau \neq \frac{1}{2}$. Before we formulate the next result, we define

$$d_0(f,g) := \mathbb{E}_{\mathrm{P}_X} \min\{1, |f(X) - g(X)|\},$$

where $f, g : \mathcal{X} \to \mathbb{R}$ are arbitrary measurable functions. It is known that d_0 is a translation invariant metric describing the convergence in probability.

Theorem 9 (Consistency)**.** *For $\tau \in (0,1)$, let L be the τ-pinball loss and $L^\star$ its shifted version. Moreover, let P be a distribution on $\mathcal{X} \times \mathbb{R}$ whose conditional τ-quantile $f^*_{\tau,\mathrm{P}} : \mathcal{X} \to \mathbb{R}$ is P_X-almost surely unique. Under the assumptions of Theorem 8, we then have*

$$d_0(f_{L^\star,\mathrm{D},\lambda_n}, f^*_{\tau,\mathrm{P}}) \to 0, \qquad n \to \infty,$$

where the convergence is either in probability P^∞ or P^∞-almost surely, depending on whether assumption (i) or (ii) on the null-sequence (λ_n) is taken from Theorem 8.

Let us now consider robustness properties of SVMs. Define the function

$$T : \mathcal{M}_1(\mathcal{X} \times \mathcal{Y}) \to \mathcal{H}, \quad T(\mathrm{P}) := f_{L^\star,\mathrm{P},\lambda}.$$

In robust statistics we are often interested in smooth and bounded functions T, because this will give us stable regularized risks within small neighbourhoods of P. If an appropriately chosen derivative of $T(\mathrm{P})$ is bounded, then we expect the value of $T(\mathrm{Q})$ to be close to the value of $T(\mathrm{P})$ for distributions Q in a small neighbourhood of P.

One general approach to robustness [Hampel, 1968, 1974] is the one based on influence functions which are related to Gâteaux-derivatives. Let $\mathcal{M}_1$ be the set of distributions on some measurable space $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$ and let $\mathcal{H}$ be a reproducing kernel Hilbert space. The *influence function* (IF) of $T : \mathcal{M}_1 \to \mathcal{H}$ at $z \in \mathcal{Z}$ for a distribution P is defined as

$$\mathrm{IF}(z; T, \mathrm{P}) = \lim_{\varepsilon \downarrow 0} \frac{T((1-\varepsilon)\mathrm{P} + \varepsilon\delta_z) - T(\mathrm{P})}{\varepsilon}, \tag{27}$$

if the limit exists. Within this approach, a statistical method $T(\mathrm{P})$ is robust if it has a bounded influence function. The influence function is neither supposed to be linear nor continuous. If the influence functions exists for all points $z \in \mathcal{Z}$ and if it is continuous and linear, then the IF is a special Gâteaux-derivative.

Theorem 10 (Influence function)**.** *Let $\mathcal{X}$ be a complete separable metric space and $\mathcal{H}$ be a RKHS of a bounded continuous kernel k. Let L be a convex, Lipschitz continuous loss function with continuous partial Fréchet-derivatives $\nabla_3^F L(x,y,\cdot)$ and $\nabla_{3,3}^F L(x,y,\cdot)$ which are bounded by*

$$\kappa_1 := \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \big\|\nabla_3^F L(x,y,\cdot)\big\|_\infty \in (0,\infty), \tag{28}$$
$$\kappa_2 := \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \big\|\nabla_{3,3}^F L(x,y,\cdot)\big\|_\infty < \infty.$$

Then, for all probability measures P on $\mathcal{X} \times \mathcal{Y}$ and for all $z := (x,y) \in \mathcal{X} \times \mathcal{Y}$, the influence function $\mathrm{IF}(z; T, \mathrm{P})$ of $T(\mathrm{P}) := f_{L^\star,\mathrm{P},\lambda}$ exists, is bounded, and equals

$$\mathbb{E}_\mathrm{P} \nabla_3^F L^\star\big(X, Y, f_{L^\star,\mathrm{P},\lambda}(X)\big) S^{-1} \Phi(X) - \nabla_3^F L^\star\big(x, y, f_{L^\star,\mathrm{P},\lambda}(x)\big) S^{-1} \Phi(x), \tag{29}$$

where $S : \mathcal{H} \to \mathcal{H}$ is the Hessian of the regularized risk and is given by

$$S(\cdot) := 2\lambda \,\mathrm{id}_\mathcal{H}(\cdot) + \mathbb{E}_\mathrm{P} \nabla_{3,3}^F L^\star(X, Y, f_{L^\star,\mathrm{P},\lambda}(X)) \langle \Phi(X), \cdot \rangle \Phi(X). \tag{30}$$

The Lipschitz continuity of L already guarantees $\kappa_1 < \infty$. Some calculations for the logistic loss functions defined in (6) and (7) give $(\kappa_1, \kappa_2) = (1, \frac{1}{4})$ for classification and $(\kappa_1, \kappa_2) = (1, \frac{1}{2})$ for regression.

Remark 11. (i) Note that only the second term of $\mathrm{IF}(z;T,\mathrm{P})$ in (29) depends on z, where the contamination of P occurs. (ii) All assumptions of Theorem 10 can be verified *without* knowledge of P, which is not true for Steinwart and Christmann [2008b, Thm. 10.18]. It is easy to check that the assumptions of Theorem 10 on L are fulfilled, e.g., for the logistic loss functions for classification and for regression defined in (6) and (7). The Gaussian RBF kernel defined in (4) is bounded and continuous.

The next result shows that the $\mathcal{H}$-norm of the difference $f_{L^\star,(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q},\lambda} - f_{L^\star,\mathrm{P},\lambda}$ increases in $\varepsilon \in (0,1)$ at most linearly. We denote the norm of total variation of a signed measure μ by $\|\mu\|_{\mathcal{M}}$.

Theorem 12 (Bounds for bias)**.** *Let L be a convex and Lipschitz continuous loss function and let $\mathcal{H}$ be a separable RKHS of a bounded and measurable kernel k. Then, for all $\lambda > 0$, all $\varepsilon \in [0,1]$, and all probability measures* P *and* Q *on $\mathcal{X} \times \mathcal{Y}$, we have*

$$\left\| f_{L^\star,(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q},\lambda} - f_{L^\star,\mathrm{P},\lambda} \right\|_{\mathcal{H}} \le c_{\mathrm{P},\mathrm{Q}}\, \varepsilon, \tag{31}$$

where

$$c_{\mathrm{P},\mathrm{Q}} = \lambda^{-1} \|k\|_\infty |L|_1 \|\mathrm{P} - \mathrm{Q}\|_{\mathcal{M}}.$$

Let $\mathrm{Q} = \delta_z$ be the Dirac measure in $z = (x,y) \in \mathcal{X} \times \mathcal{Y}$. If the influence function of $T(\mathrm{P}) = f_{L^\star,\mathrm{P},\lambda}$ exists, then

$$\|\mathrm{IF}(z;T,\mathrm{P})\|_{\mathcal{H}} \le c_{\mathrm{P},\delta_z}.$$

The Bouligand influence function (BIF) was introduced by Christmann and Van Messem [2008] to investigate robustness properties of SVMs based on *non-Fréchet-differentiable* loss functions, such as, e.g., the ϵ-insensitive loss or the pinball loss. The BIF of the map $T\colon \mathcal{M}_1(\mathcal{X}\times\mathcal{Y}) \to \mathcal{H}$ for a distribution P in the direction of a distribution $\mathrm{Q} \ne \mathrm{P}$ is the special Bouligand-derivative[3] (if it exists)

$$\lim_{\varepsilon \downarrow 0} \frac{\left\| T\big((1-\varepsilon)\mathrm{P} + \varepsilon\mathrm{Q}\big) - T(\mathrm{P}) - \mathrm{BIF}(\mathrm{Q};T,\mathrm{P}) \right\|_{\mathcal{H}}}{\varepsilon} = 0. \tag{32}$$

The BIF has the interpretation that it measures the impact of an infinitesimal small amount of contamination of the original distribution P in the direction of Q on the quantity of interest $T(\mathrm{P})$. It is thus desirable that the function T has a *bounded* BIF.

Theorem 13 (Bouligand influence function)**.** *Let $\mathcal{X}$ be a complete separable normed linear space*[4] *and $\mathcal{H}$ be a RKHS of a bounded, continuous kernel k. Let L be a convex, Lipschitz continuous loss function with Lipschitz constant $|L|_1 \in (0,\infty)$. Let the partial Bouligand-derivatives $\nabla_3^B L(x,y,\cdot)$ and $\nabla_{3,3}^B L(x,y,\cdot)$ be measurable and bounded by*

$$\begin{aligned} \kappa_1 &:= \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \left\| \nabla_3^B L(x,y,\cdot) \right\|_\infty \in (0,\infty), \\ \kappa_2 &:= \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \left\| \nabla_{3,3}^B L(x,y,\cdot) \right\|_\infty < \infty. \end{aligned} \tag{33}$$

Let P *and* $\mathrm{Q} \ne \mathrm{P}$ *be probability measures on $\mathcal{X}\times\mathcal{Y}$, $\delta_1 > 0$, $\delta_2 > 0$,*

$$\mathcal{N}_{\delta_1}(f_{L^\star,\mathrm{P},\lambda}) := \{ f \in \mathcal{H} : \|f - f_{L^\star,\mathrm{P},\lambda}\|_{\mathcal{H}} < \delta_1 \},$$

and $\lambda > \frac{\kappa_2}{2}\|k\|_\infty^3$. Define $G : (-\delta_2,\delta_2) \times \mathcal{N}_{\delta_1}(f_{L^\star,\mathrm{P},\lambda}) \to \mathcal{H}$,

$$G(\varepsilon, f) := 2\lambda f + \mathbb{E}_{(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q}} \nabla_3^B L^\star(X,Y,f(X)) \cdot \Phi(X), \tag{34}$$

and assume that $\nabla_2^B G(0, f_{L^\star,\mathrm{P},\lambda})$ is strong. Then the Bouligand influence function $\mathrm{BIF}(\mathrm{Q};T,\mathrm{P})$ *of $T(\mathrm{P}) := f_{L^\star,\mathrm{P},\lambda}$ exists, is bounded, and equals*

$$\begin{aligned} & S^{-1}\big(\mathbb{E}_{\mathrm{P}} \nabla_3^B L^\star(X,Y,f_{L^\star,\mathrm{P},\lambda}(X)) \cdot \Phi(X)\big) \\ & -S^{-1}\big(\mathbb{E}_{\mathrm{Q}} \nabla_3^B L^\star(X,Y,f_{L^\star,\mathrm{P},\lambda}(X)) \cdot \Phi(X)\big), \end{aligned} \tag{35}$$

where $S := \nabla_2^B G(0, f_{L^\star,\mathrm{P},\lambda}) : \mathcal{H} \to \mathcal{H}$ is given by

$$\begin{aligned} S(\cdot) &= 2\lambda\, \mathrm{id}_{\mathcal{H}}(\cdot) \\ &\quad + \mathbb{E}_{\mathrm{P}} \nabla_{3,3}^B L^\star(X,Y,f_{L^\star,\mathrm{P},\lambda}(X)) \cdot \langle \Phi(X), \cdot \rangle_{\mathcal{H}} \Phi(X). \end{aligned}$$

Note that the Bouligand influence function of the SVM only depends on Q via the second term in (35). We have $(\kappa_1,\kappa_2) = (1,0)$ for the ϵ-insensitive loss and $(\kappa_1,\kappa_2) = (\max\{1-\tau,\tau\},0)$ for the pinball loss, see Christmann and Van Messem [2008].

4. DISCUSSION

Support vector machines play an important role in statistical machine learning and are successfully applied even to complex high-dimensional data sets. From a nonparametric point of view, we do not know in supervised machine learning whether the moment condition $\mathbb{E}_{\mathrm{P}}|Y| < \infty$ is fulfilled. However, some recent results on consistency and statistical robustness properties of SVMs for unbounded output spaces were derived under the assumption that this absolute moment is finite which excludes distributions with heavy tails such as many stable distributions, including the Cauchy distribution, and many extreme value distributions which occur in financial or actuarial problems.

The main goal of this paper was therefore to enlarge the applicability of support vector machines to situations where the output space $\mathcal{Y}$ is unbounded, e.g., $\mathcal{Y} = \mathbb{R}$ or $\mathcal{Y} = [0,\infty)$, *without* the above mentioned moment condition. We showed that SVMs can still be used in a satisfactory manner. Results on existence, uniqueness, representation, consistency, and statistical robustness were derived. There is no need to

[3] See Appendix A.1.3.

[4] E.g., $\mathcal{X} \subset \mathbb{R}^d$ closed. By definition of the Bouligand-derivative, $\mathcal{X}$ has to be a normed linear space.

establish new algorithms to compute the SVM based on the shifted loss function.

Finally, let us briefly comment some topics which were not treated in this paper. *(i)* We decided to consider only non-negative loss functions L (but the shifted loss function $L^\star$ can have negative values), because almost all loss functions used in practice are non-negative and no results on SVMs seem available for loss functions with negative values. *(ii)* It may be possible to derive results similar to ours for convex, *locally Lipschitzian* loss functions, including the least squares loss, but Lipschitz continuous loss functions can offer better robustness properties, see Christmann and Steinwart [2004, 2007] and Steinwart and Christmann [2008b]. *(iii)* From a robustness point of view, *bounded* and *non-convex* loss functions may also be of interest. We have not considered such loss functions for two reasons. Firstly, existence, uniqueness, consistency, and availability of efficient numerical algorithms are widely accepted as necessary properties which SVMs should have to avoid numerically intractable problems for large and high-dimensional data sets, say for $n > 10^5$ and $d > 100$, see e.g. Vapnik [1998] or Schölkopf and Smola [2002]. All these properties can be achieved if the risk is convex which is true for convex loss functions. Secondly, there are currently—to our best knowledge—no general results on SVMs available which guarantee that the risk remains convex although the loss function is non-convex and bounded. However, the convexity of the risk plays a key role in the proofs of the existence and uniqueness of SVMs. From our point of view, such results would be a prerequisite for an investigation of shifted versions of bounded and non-convex loss functions, but this is beyond the scope of this paper.

APPENDIX: MATHEMATICAL FACTS AND PROOFS

A.1 Mathematical prerequisites

A.1.1 Some definitions and properties

Let E and F be normed spaces and $S : E \to F$ a linear operator. We will denote the *closed unit ball* by $B_E := \{x \in E : \|x\|_E \leq 1\}$. The *convex hull* $\operatorname{co} A$ of $A \subset E$ is the smallest convex set containing A. The space of all bounded (linear) operators mapping from E to F is written as $\mathcal{L}(E, F)$. If $S \in \mathcal{L}(E, F)$ satisfies $\|Sx\|_F = \|x\|_E$ for all $x \in E$, then S is called an isometric embedding. Obviously, S is injective in this case. If, in addition, S is also surjective, then S is called an *isometric isomorphism* and E and F are said to be *isometrically isomorphic*. An $S \in \mathcal{L}(E, F)$ is called *compact* if $\overline{SB_E}$ is a compact subset in F. A special case of linear operators are the bounded linear functionals, i.e., the elements of the *dual space* $E' := \mathcal{L}(E, \mathbb{R})$. Note that, due to the completeness of $\mathbb{R}$, dual spaces are always Banach spaces. For $x \in E$ and $x' \in E'$, the evaluation of x' at x is often written as a *dual pairing*, i.e., $\langle x', x\rangle_{E',E} := x'(x)$. The smallest topology on E' for which the maps $x' \mapsto \langle x', x\rangle_{E',E}$ are continuous on E' for all $x \in E$ is called the *weak* topology*. For $S \in \mathcal{L}(E, F)$, the *adjoint operator* $S' : F' \to E'$ is defined by $\langle S'y', x\rangle_{E',E} := \langle y', Sx\rangle_{F',F}$ for all $x \in E$ and $y' \in F'$.

Given a measurable space $(\mathcal{X}, \mathcal{A})$, $\mathcal{L}_0(\mathcal{X})$ denotes the set of all real-valued measurable functions f on $\mathcal{X}$ and $\mathcal{L}_\infty(\mathcal{X})$ the set of all bounded measurable functions, i.e., $\mathcal{L}_\infty(\mathcal{X}) := \{f \in \mathcal{L}_0(\mathcal{X}) : \|f\|_\infty < \infty\}$. Let us now assume we have a measure μ on $\mathcal{A}$. For $p \in (0, \infty)$ and $f \in \mathcal{L}_0(\mathcal{X})$ we write $\|f\|_{\mathcal{L}_p(\mu)} := (\int_{\mathcal{X}} |f|^p d\mu)^{1/p}$. To treat the case $p = \infty$, we call $N \in \mathcal{A}$ a local μ-zero set if $\mu(N \cap A) = 0$ for all $A \in \mathcal{A}$ with $\mu(A) < \infty$. Then $\|f\|_{\mathcal{L}_\infty(\mu)} := \inf\{a \geq 0 : \{x \in \mathcal{X} : |f(x)| > a\}$ is a local μ-zero set$\}$. In both cases the *set of p-integrable functions* $\mathcal{L}_p(\mu) := \{f \in \mathcal{L}_0(\mathcal{X}) : \|f\|_{\mathcal{L}_p(\mu)} < \infty\}$ is a vector space of functions, and for $p \in [1, \infty]$ all properties of a norm on $\mathcal{L}_p(\mu)$ are followed by the mapping $\|\cdot\|_{\mathcal{L}_p(\mu)}$. As usual, we call $f, f' \in \mathcal{L}_p(\mu)$ equivalent, written $f \sim f'$, if $\|f - f'\|_{\mathcal{L}_p(\mu)} = 0$. In other words, $f \sim f'$ if and only if $f(x) = f'(x)$ for μ-almost all $x \in \mathcal{X}$. The set of equivalence classes $L_p(\mu) := \{[f]_\sim : f \in \mathcal{L}_p(\mu)\}$, where $[f]_\sim := \{f' \in \mathcal{L}_p(\mu) : f \sim f'\}$, is a vector space and $\|[f]_\sim\|_{L_p(\mu)} := \|f\|_{\mathcal{L}_p(\mu)}$ is a complete norm on $L_p(\mu)$ for $p \in [1, \infty]$, i.e., $(L_p(\mu), \|\cdot\|_{L_p(\mu)})$ is a Banach space. It is common practice to identify the Lebesgue spaces $\mathcal{L}_p(\mu)$ and $L_p(\mu)$ and hence we often abbreviate both $\|\cdot\|_{\mathcal{L}_p(\mu)}$ and $\|\cdot\|_{L_p(\mu)}$ as $\|\cdot\|_p$. In addition, we usually write $\mathcal{L}_p(\mathcal{X}) := \mathcal{L}_p(\mu)$ and $L_p(\mathcal{X}) := L_p(\mu)$ if $\mathcal{X} \subset \mathbb{R}^d$ and μ is the Lebesgue measure on $\mathcal{X}$. For μ the counting measure on $\mathcal{X}$, we write $\ell_p(\mathcal{X})$ instead of $\mathcal{L}_p(\mu)$.

Lemma 14 (Parallellogram identity)**.** *Let $(\mathcal{H}, \langle\cdot,\cdot\rangle)$ be a Hilbert space. Then, for all $f, g \in \mathcal{H}$, we have*

$$4\langle f, g\rangle = \|f + g\|_{\mathcal{H}}^2 - \|f - g\|_{\mathcal{H}}^2 ,$$
$$\|f + g\|_{\mathcal{H}}^2 + \|f - g\|_{\mathcal{H}}^2 = 2\|f\|_{\mathcal{H}}^2 + 2\|g\|_{\mathcal{H}}^2 .$$

We refer to Cheney [2001] for the following result.

Theorem 15 (Fredholm alternative)**.** *Let E be a Banach space and let $S : E \to E$ be a compact operator. Then $\mathrm{id}_E + S$ is surjective if and only if it is injective.*

We refer to Werner [2002] for the following fact.

Theorem 16 (Fréchet-Riesz representation)**.** *Let $\mathcal{H}$ be a Hilbert space and $\mathcal{H}'$ its dual. Then the mapping $\iota : \mathcal{H} \to \mathcal{H}'$ defined by $\iota x := \langle\cdot, x\rangle$ for all $x \in \mathcal{H}$ is an isometric isomorphism.*

For Hoeffding's inequality, we refer to Yurinsky [1995].

Theorem 17 (Hoeffding's inequality in Hilbert spaces)**.** *Let $(\Omega, \mathcal{A}, \mathrm{P})$ be a probability space, $\mathcal{H}$ be a separable Hilbert space, and $B > 0$. Furthermore, let $\xi_1, \ldots, \xi_n : \Omega \to \mathcal{H}$ be*

independent $\mathcal{H}$-valued random variables satisfying $\|\xi_i\|_\infty \leq B$ for all $i = 1, \ldots, n$. Then, for all $\tau > 0$, we have

$$\mathrm{P}\Big(\big\|n^{-1}\sum_{i=1}^{n}(\xi_i - \mathbb{E}_{\mathrm{P}}\xi_i)\big\|_{\mathcal{H}} \geq B\sqrt{\frac{2\tau}{n}} + B\sqrt{\frac{1}{n}} + \frac{4B\tau}{3n}\Big) \leq e^{-\tau}.$$

A.1.2 Some facts on convexity and subdifferentials

The following result on the continuity of convex functions can be found, e.g., in Rockafellar and Wets [1998].

Lemma 18 (Continuity of convex functions)**.** *Let $f : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ be a convex function with domain* $\mathrm{Dom} f := \{t \in \mathbb{R} : f(t) < \infty\}$*. Then f is continuous at all $t \in \mathrm{Int}\,\mathrm{Dom} f$.*

The next result is a consequence of Ekeland and Turnbull [1983, Prop. II.4.6].

Proposition 19. *Let E be a Banach space and let $f : E \to \mathbb{R} \cup \{\infty\}$ be a convex function. If f is continuous and $\lim_{\|x\|_E \to \infty} f(x) = \infty$, then f has a minimizer. Moreover, if f is strictly convex, then f has a unique minimizer in E.*

Now we will state some important properties of the subdifferential of a convex function [see e.g., Phelps, 1993]. For the remainder of this subsection, E and F will denote $\mathbb{R}$-Banach spaces. Let us begin by recalling the definition of subdifferentials.

Definition 20. Let $f : E \to \mathbb{R} \cup \{\infty\}$ be a convex function, and $w \in E$ with $f(w) < \infty$. Then the *subdifferential* of f at w is defined by

$$\partial f(w) := \left\{w' \in E' : \langle w', v - w\rangle \leq f(v) - f(w) \text{ for all } v \in E\right\}.$$

The following proposition provides some elementary facts on the subdifferential, see Phelps [1993, Proposition 1.11].

Proposition 21. *Let $f : E \to \mathbb{R} \cup \{\infty\}$ be a convex function and $w \in E$ such that $f(w) < \infty$. If f is continuous at w, then the subdifferential $\partial f(w)$ is a non-empty, convex, and weak*-compact subset of E'. In addition, if $c \geq 0$ and $\delta > 0$ are constants satisfying $\big|f(v) - f(w)\big| \leq c\,\|v - w\|_E$, $v \in w + \delta B_E$, then we have $\|w'\|_E \leq c$ for all $w' \in \partial f(w)$.*

This next proposition shows the extent to which the known rules of calculus carry over to subdifferentials.

Proposition 22 (Subdifferential calculus)**.** *Let $f, g : E \to \mathbb{R} \cup \{\infty\}$ be convex functions, $\lambda \geq 0$, and $A : F \to E$ be a bounded linear operator. We then have:*

i) For all $w \in E$ with $f(x) < \infty$, we have $\partial(\lambda f)(w) = \lambda \partial f(w)$.

ii) If there exists a $w_0 \in E$ at which f is continuous, then, for all $w \in E$ satisfying both $f(w) < \infty$ and $g(w) < \infty$, we have $\partial(f + g)(w) = \partial f(w) + \partial g(w)$.

iii) If there exists a $v_0 \in F$ such that f is finite and continuous at Av_0, then, for all $v \in F$ satisfying $f(Av) < \infty$, we have $\partial(f \circ A)(v) = A'\partial f(Av)$, where $A' : E' \to F'$ denotes the adjoint operator of A.

iv) The function f has a global minimum at $w \in E$ if and only if $0 \in \partial f(w)$.

v) If f is finite and continuous at $w \in E$, then f is Gâteaux-differentiable at w if and only if $\partial f(w)$ is a singleton, and in this case we have $\partial f(w) = \{f'(w)\}$.

vi) If f is finite and continuous at all $w \in E$, then ∂f is a monotone operator*, i.e., for all $v, w \in E$ and $v' \in \partial f(v)$, $w' \in \partial f(w)$, we have $\langle v' - w', v - w\rangle \geq 0$.*

The following proposition shows how the subdifferential of a function defined by an integral can be computed.

Proposition 23. *Let $\tilde{L} : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$ be a measurable function which is both convex and Lipschitz continuous with respect to its third argument, P be a distribution on $\mathcal{X} \times \mathcal{Y}$, and $p \in [1, \infty)$. Assume that $R : L_p(\mathrm{P}) \to \mathbb{R} \cup \{\pm\infty\}$ defined by*

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

exists for all $f \in L_p(\mathrm{P})$ and define p' by $\frac{1}{p} + \frac{1}{p'} = 1$. If $|R(f)| < \infty$ for at least one $f \in L_p(\mathrm{P})$, then, for all $f \in L_p(\mathrm{P})$, we have

$$\partial R(f) = \big\{h \in L_{p'}(\mathrm{P}) : h(x, y) \in \partial\tilde{L}(x, y, f(x, y)) \\ \text{for } \mathrm{P}\text{-almost all } (x, y)\big\},$$

where $\partial\tilde{L}(x, y, t)$ denotes the subdifferential of $\tilde{L}(x, y, \cdot)$ at the point t.

Proof of Proposition 23. Since $\tilde{L}$ is measurable, Lipschitz continuous, and finite, it is a continuous function with respect to its third argument. Thus it is a normal convex integrand by Proposition 2C of Rockafellar [1976]. Then Corollary 3E of Rockafellar [1976] gives the assertion. □

A.1.3 Some facts on derivatives

We first recall the definitions of the Gâteaux- and Fréchet-derivative. Let E and F be normed spaces, $U \subset E$ and $V \subset F$ be open sets, and $f : U \to V$ be a function. We say that f is *Gâteaux-differentiable* at $x_0 \in U$ if there exists a bounded linear operator $\nabla^G f(x_0) \in \mathcal{L}(E, F)$ such that

$$\lim_{t \to 0,\, t \neq 0} \frac{\big\|f(x_0 + tx) - f(x_0) - t\nabla^G f(x_0)(x)\big\|_F}{t} = 0, \quad x \in E.$$

We say that f is *Fréchet-differentiable* at x_0 if there exists a bounded linear operator $\nabla^F f(x_0) \in \mathcal{L}(E, F)$ such that

$$\lim_{x \to 0,\, x \neq 0} \frac{\big\|f(x_0 + x) - f(x_0) - \nabla^F f(x_0)(x)\big\|_F}{\|x\|_E} = 0.$$

We call $\nabla^G f(x_0)$ the Gâteaux-derivative and $\nabla^F f(x_0)$ the Fréchet-derivative of f at x_0. The function f is called Gâteaux- (or Fréchet-) differentiable if f is Gâteaux- (or Fréchet-) differentiable for all $x_0 \in U$, respectively.

We also recall some facts on Bouligand-derivatives and strong approximation of functions, because these notions will be used to investigate robustness properties for SVMs for nonsmooth loss functions in Theorem 13. Let E_1, E_2, W, and Z be normed linear spaces, and let us consider neighbourhoods $\mathcal{N}(x_0)$ of x_0 in E_1, $\mathcal{N}(y_0)$ of y_0 in E_2, and $\mathcal{N}(w_0)$ of w_0 in W. Let F and G be functions from $\mathcal{N}(x_0) \times \mathcal{N}(y_0)$ to Z, h_1 and h_2 functions from $\mathcal{N}(w_0)$ to Z, f a function from $\mathcal{N}(x_0)$ to Z and g a function from $\mathcal{N}(y_0)$ to Z. A function f *approximates* F in x at (x_0, y_0), written as $f \sim_x F$ at (x_0, y_0), if

$$F(x, y_0) - f(x) = o(x - x_0).$$

Similarly, $g \sim_y F$ at (x_0, y_0) if $F(x_0, y) - g(y) = o(y - y_0)$. A function h_1 *strongly approximates* h_2 at w_0, written as $h_1 \approx h_2$ at w_0, if for each $\varepsilon > 0$ there exists a neighbourhood $\mathcal{N}(w_0)$ of w_0 such that whenever w and w' belong to $\mathcal{N}(w_0)$,

$$\big\|\big(h_1(w) - h_2(w)\big) - \big(h_1(w') - h_2(w')\big)\big\| \leq \varepsilon \|w - w'\|.$$

A function f *strongly approximates* F in x at (x_0, y_0), written as $f \approx_x F$ at (x_0, y_0), if for each $\varepsilon > 0$ there exist neighbourhoods $\mathcal{N}(x_0)$ of x_0 and $\mathcal{N}(y_0)$ of y_0 such that whenever x and x' belong to $\mathcal{N}(x_0)$ and y belongs to $\mathcal{N}(y_0)$ we have

$$\big\|\big(F(x, y) - f(x)\big) - \big(F(x', y) - f(x')\big)\big\| \leq \varepsilon \|x - x'\|.$$

Strong approximation amounts to requiring $h_1 - h_2$ to have a strong Fréchet-derivative of 0 at w_0, though neither h_1 nor h_2 is assumed to be differentiable in any sense. A similar definition is made for strong approximation in y. We define strong approximation for functions of several groups of variables, for example $G \approx_{(x,y)} F$ at (x_0, y_0), by replacing W by $E_1 \times E_2$ and making the obvious substitutions. Note that one has both $f \approx_x F$ and $g \approx_y F$ at (x_0, y_0) exactly if $f(x) + g(y) \approx_{(x,y)} F$ at (x_0, y_0).

Recall that a function $f : E_1 \to Z$ is called *positive homogeneous* if

$$f(\alpha x) = \alpha f(x) \quad \forall\, \alpha \geq 0, \quad \forall\, x \in E_1.$$

Following Robinson [1987] we can now define the *Bouligand-derivative*. Given a function f from an open subset U of a normed linear space E_1 into another normed linear space Z, we say that f is *Bouligand-differentiable* at a point $x_0 \in U$, if there exists a positive homogeneous function $\nabla^B f(x_0) : U \to Z$ such that $f(x_0 + h) = f(x_0) + \nabla^B f(x_0)(h) + o(h)$, which can be rewritten as

$$\lim_{h \to 0} \frac{\big\|f(x_0 + h) - f(x_0) - \nabla^B f(x_0)(h)\big\|_Z}{\|h\|_{E_1}} = 0.$$

Sometimes we use the abbreviations B-, F-, and G-derivatives. Let $F : E_1 \times E_2 \to Z$, and suppose that F has a partial B-derivative[5] $\nabla_1^B F(x_0, y_0)$ with respect to x at (x_0, y_0). We say $\nabla_1^B F(x_0, y_0)$ is *strong* if

$$F(x_0, y_0) + \nabla_1^B F(x_0, y_0)(x - x_0) \approx_x F \text{ at } (x_0, y_0).$$

We refer to Akerkar [1999] for the following implicit function theorem for F-derivatives and to Robinson [1991, Cor. 3.4] for a similar implicit function theorem for B-derivatives.

Theorem 24 (Implicit function theorem)**.** *Let E_1 and E_2 be Banach spaces, and let $G : E_1 \times E_2 \to E_2$ be a continuously Fréchet-differentiable function. Suppose that we have $(x_0, y_0) \in E_1 \times E_2$ such that $G(x_0, y_0) = 0$ and $\nabla_2^F G(x_0, y_0)$ is invertible. Then there exists a $\delta > 0$ and a continuously Fréchet-differentiable function $f : x_0 + \delta B_{E_1} \to y_0 + \delta B_{E_2}$ such that for all $x \in x_0 + \delta B_{E_1}$, $y \in y_0 + \delta B_{E_2}$ we have $G(x, y) = 0$ if and only if $y = f(x)$. Moreover, the Fréchet-derivative of f is given by*

$$\nabla^F f(x) = -\big(\nabla_2^F G(x, f(x))\big)^{-1} \nabla_1^F G(x, f(x)).$$

A.1.4 Properties of the risk and of RKHSs

We will need the following three results, see, e.g., Steinwart and Christmann [2008b, Lemma 2.19, 4.23, 4.24]. The first lemma relates the Lipschitz continuity of L to the Lipschitz continuity of its risk.

Lemma 25 (Lipschitz continuity of the risks)**.** *Let L be a Lipschitz continuous loss and P be a probability measure on $\mathcal{X} \times \mathcal{Y}$. Then we have, for all $f, g \in L_\infty(\mathrm{P}_X)$,*

$$|\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{P}}(g)| \leq |L|_1 \cdot \|f - g\|_{L_1(\mathrm{P}_X)}.$$

Lemma 26 (RKHSs of bounded kernels)**.** *Let $\mathcal{X}$ be a set and k be a kernel on $\mathcal{X}$ with RKHS $\mathcal{H}$. Then k is bounded if and only if every $f \in \mathcal{H}$ is bounded. Moreover, in this case the inclusion $\mathrm{id} : \mathcal{H} \to \ell_\infty(\mathcal{X})$ is continuous and we have $\|\mathrm{id} : \mathcal{H} \to \ell_\infty(\mathcal{X})\| = \|k\|_\infty$.*

Lemma 27 (RKHSs of measurable kernels)**.** *Let $\mathcal{X}$ be a measurable space and k be a kernel on $\mathcal{X}$ with RKHS $\mathcal{H}$. Then all $f \in \mathcal{H}$ are measurable if and only if $k(\cdot, x) : \mathcal{X} \to \mathbb{R}$ is measurable for all $x \in \mathcal{X}$.*

A.2 Proofs for Section 2

Proof of Proposition 2. Let L be a convex loss function and fix $(x, y) \in \mathcal{X} \times \mathcal{Y}$. For all $\alpha \in [0, 1]$ we get

$$\begin{aligned}
&L^\star(x, y, \alpha t_1 + (1 - \alpha) t_2) \\
&\quad = L(x, y, \alpha t_1 + (1 - \alpha) t_2) - L(x, y, 0) \\
&\quad \leq \alpha L(x, y, t_1) + (1 - \alpha) L(x, y, t_2) - (1 - \alpha + \alpha) L(x, y, 0) \\
&\quad = \alpha L^\star(x, y, t_1) + (1 - \alpha) L^\star(x, y, t_2), \quad t_1, t_2 \in \mathbb{R},
\end{aligned}$$

[5] Partial B-derivatives of f are denoted by $\nabla_1^B f$, $\nabla_2^B f$, $\nabla_{2,2}^B f := \nabla_2^B\big(\nabla_2^B f\big)$ etc.

which proves the convexity of $L^\star$. For a strict convex loss, the calculation is analogous. If L is a Lipschitz continuous loss, we immediately obtain that $|L^\star(x,y,t_1)-L^\star(x,y,t_2)| = |L(x,y,t_1)-L(x,y,t_2)| \le |L|_1\,|t_1-t_2|$, $t_1,t_2\in\mathbb{R}$, and hence $L^\star$ is Lipschitzian with $|L^\star|_1 = |L|_1$. □

Proof of Proposition 3. *(i)* Obviously, $\inf_{t\in\mathbb{R}} L^\star(x,y,t) \le L^\star(x,y,0) = 0$.

(ii) We have for all $f\in\mathcal{H}$ that

$$\begin{aligned}
&|\mathcal{R}_{L^\star,\mathrm{P}}(f)| \\
&\quad = |\mathbb{E}_\mathrm{P} L^\star(X,Y,f(X))| = |\mathbb{E}_\mathrm{P} L(X,Y,f(X)) - L(X,Y,0)| \\
&\quad \le \mathbb{E}_\mathrm{P}|L(X,Y,f(X)) - L(X,Y,0)| \le |L|_1\,\mathbb{E}_{\mathrm{P}_X}|f(X)|,
\end{aligned}$$

which proves (14). Equation (15) follows from $\mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f) = \mathcal{R}_{L^\star,\mathrm{P}}(f) + \lambda\,\|f\|^2_\mathcal{H}$.

(iii) As $0\in\mathcal{H}$, we obtain $\inf_{f\in\mathcal{H}} \mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f) \le \mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(0) = 0$ and the same reasoning holds for $\inf_{f\in\mathcal{H}} \mathcal{R}_{L^\star,\mathrm{P}}(f)$.

(iv) Due to *(iii)* we have $\mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f_{L^\star,\mathrm{P},\lambda}) \le 0$. As $L\ge 0$ we obtain

$$\begin{aligned}
\lambda\,\|f_{L^\star,\mathrm{P},\lambda}\|^2_\mathcal{H} &\le -\mathcal{R}_{L^\star,\mathrm{P}}(f_{L^\star,\mathrm{P},\lambda}) \\
&= \mathbb{E}_\mathrm{P}\big(L(X,Y,0) - L(X,Y,f_{L^\star,\mathrm{P},\lambda}(X))\big) \\
&\le \mathbb{E}_\mathrm{P} L(X,Y,0) = \mathcal{R}_{L,\mathrm{P}}(0).
\end{aligned}$$

Using similar arguments as above, we obtain

$$\begin{aligned}
0 &\le -\mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f_{L^\star,\mathrm{P},\lambda}) \\
&= \mathbb{E}_\mathrm{P}\big(L(X,Y,0) - L(X,Y,f_{L^\star,\mathrm{P},\lambda}(X))\big) - \lambda\,\|f_{L^\star,\mathrm{P},\lambda}\|^2_\mathcal{H} \\
&\le \mathbb{E}_\mathrm{P} L(X,Y,0) = \mathcal{R}_{L,\mathrm{P}}(0).
\end{aligned}$$

Furthermore, we obtain

$$\begin{aligned}
-|L|_1\,\mathbb{E}_{\mathrm{P}_X}|f_{L^\star,\mathrm{P},\lambda}(X)| + \lambda\,\|f_{L^\star,\mathrm{P},\lambda}\|^2_\mathcal{H} &\le \mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f_{L^\star,\mathrm{P},\lambda}) \\
&\le \mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(0) = 0.
\end{aligned}$$

This yields (16). Using (2), (3), and (16), we obtain for $f_{L^\star,\mathrm{P},\lambda}\neq 0$ that

$$\begin{aligned}
\|f_{L^\star,\mathrm{P},\lambda}\|_\infty &\le \|k\|_\infty\,\|f_{L^\star,\mathrm{P},\lambda}\|_\mathcal{H} \\
&\le \|k\|_\infty \sqrt{\lambda^{-1}|L|_1\mathbb{E}_{\mathrm{P}_X}|f_{L^\star,\mathrm{P},\lambda}(X)|} \\
&\le \|k\|_\infty \sqrt{\lambda^{-1}|L|_1\,\|f_{L^\star,\mathrm{P},\lambda}\|_\infty} < \infty.
\end{aligned}$$

Hence $\|f_{L^\star,\mathrm{P},\lambda}\|_\infty \le \|k\|^2_\infty\,\lambda^{-1}|L|_1$. The case $f_{L^\star,\mathrm{P},\lambda}=0$ is trivial.

(v) By definition of $L^\star$ and of the Fréchet-derivative we immediately obtain

$$\begin{aligned}
\nabla_3^F L^\star(x,y,t) &= \lim_{h\to 0,\,h\neq 0} \frac{L^\star(x,y,t+h) - L^\star(x,y,t)}{h} \\
&= \nabla_3^F L(x,y,t).
\end{aligned}$$

An analogous calculation is valid for the Bouligand-derivative because the term $L(x,y,0)$ cancels out and we obtain $\nabla_3^B L^\star(x,y,t) = \nabla_3^B L(x,y,t)$. □

Proof of Proposition 4. Using (14) we have $|\mathcal{R}_{L^\star,\mathrm{P}}(f)| \le |L|_1\mathbb{E}_{\mathrm{P}_X}|f(X)| < \infty$ for $f\in L_1(\mathrm{P}_X)$. Then (15) yields $\mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f) \ge -|L|_1\mathbb{E}_{\mathrm{P}_X}|f(X)| + \lambda\,\|f\|^2_\mathcal{H} > -\infty$. □

A.3 Proofs for Section 3

Lemma 28 (Convexity of risks)**.** *Let L be a (strictly) convex loss. Then $\mathcal{R}_{L^\star,\mathrm{P}} : \mathcal{H}\to[-\infty,\infty]$ is (strictly) convex and $\mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda} : \mathcal{H}\to[-\infty,\infty]$ is strictly convex.*

Proof of Lemma 28. Proposition 2 yields that $L^\star$ is (strictly) convex. Trivially $\mathcal{R}_{L^\star,\mathrm{P}}$ is also convex. Further $f\mapsto\lambda\,\|f\|^2_\mathcal{H}$ is strictly convex, and hence the mapping $f\mapsto\mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f) = \mathcal{R}_{L^\star,\mathrm{P}}(f) + \lambda\,\|f\|^2_\mathcal{H}$ is strictly convex. □

Proof of Theorem 5. Let us assume that the mapping $f\mapsto\lambda\,\|f\|^2_\mathcal{H} + \mathcal{R}_{L^\star,\mathrm{P}}(f)$ has two minimizers f_1 and $f_2\in\mathcal{H}$ with $f_1\neq f_2$. *(i)* By Lemma 14, we then find

$$\|(f_1+f_2)/2\|^2_\mathcal{H} < \|f_1\|^2_\mathcal{H}/2 + \|f_2\|^2_\mathcal{H}/2.$$

The convexity of $f\mapsto\mathcal{R}_{L^\star,\mathrm{P}}(f)$, see Lemma 28, and

$$\lambda\,\|f_1\|^2_\mathcal{H} + \mathcal{R}_{L^\star,\mathrm{P}}(f_1) = \lambda\,\|f_2\|^2_\mathcal{H} + \mathcal{R}_{L^\star,\mathrm{P}}(f_2)$$

then shows for $f^* := \frac{1}{2}(f_1+f_2)$ that

$$\lambda\,\|f^*\|^2_\mathcal{H} + \mathcal{R}_{L^\star,\mathrm{P}}(f^*) < \lambda\,\|f_1\|^2_\mathcal{H} + \mathcal{R}_{L^\star,\mathrm{P}}(f_1),$$

i.e., f_1 is *not* a minimizer of $f\mapsto\lambda\,\|f\|^2_\mathcal{H} + \mathcal{R}_{L^\star,\mathrm{P}}(f)$. Consequently, the assumption that there are two minimizers is false. *(ii)* This condition implies that $|\mathcal{R}_{L^\star,\mathrm{P}}(f)| < \infty$, see Proposition 4, and the assertion follows from *(i)*. □

Lemma 2.17 from Steinwart and Christmann [2008b] gives us a result on the continuity of risks, which we will adapt to our needs.

Lemma 29 (Continuity of risks)**.** *Let L be a Lipschitz continuous loss function. Then the following statements hold:*

i) Let $f_n : \mathcal{X}\to\mathbb{R}$, $n\ge 1$, be bounded, measurable functions for which there exists a constant $B>0$ with $\|f_n\|_\infty \le B$ for all $n\ge 1$. If the sequence (f_n) converges P_X-almost surely to a measurable function $f : \mathcal{X}\to\mathbb{R}$, then we have

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

ii) The mapping $\mathcal{R}_{L^\star,\mathrm{P}} : L_\infty(\mathrm{P}_X)\to\mathbb{R}$ is well-defined and continuous.

A consequence of this lemma is that the function $f\mapsto\mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f)$ is continuous, since both mappings $f\mapsto\mathcal{R}_{L^\star,\mathrm{P}}(f)$ and $f\mapsto\lambda\,\|f\|^2_\mathcal{H}$ are continuous.

Proof of Lemma 29. *(i)* Obviously, f is a bounded and measurable function with $\|f\|_\infty \le B$. Furthermore, the continuity of L shows

$$\begin{aligned}\lim_{n\to\infty} &|L^\star(x,y,f_n(x)) - L^\star(x,y,f(x))| \\ &= \lim_{n\to\infty} |L(x,y,f_n(x)) - L(x,y,f(x))| = 0\end{aligned}$$

for P-almost all $(x,y) \in \mathcal{X}\times\mathcal{Y}$. In addition, we have

$$\begin{aligned}&|L^\star(x,y,f_n(x)) - L^\star(x,y,f(x))| \\ &\quad\le |L|_1\,|f_n(x) - f(x)| \\ &\quad\le |L|_1(\|f_n\|_\infty + \|f\|_\infty) \\ &\quad\le 2B|L|_1 < \infty\end{aligned}$$

for all $(x,y) \in \mathcal{X}\times\mathcal{Y}$ and all $n \ge 1$. Since the constant function $2B|L|_1$ is P-integrable, Lebesgue's theorem of dominated convergence together with

$$\begin{aligned}&|\mathcal{R}_{L^\star,\mathrm{P}}(f_n) - \mathcal{R}_{L^\star,\mathrm{P}}(f)| \\ &\quad\le \int_{\mathcal{X}\times\mathcal{Y}} |L^\star(x,y,f_n(x)) - L^\star(x,y,f(x))|\,d\mathrm{P}(x,y)\end{aligned}$$

gives the assertion.

(ii) We know from Proposition 4 that $|\mathcal{R}_{L^\star,\mathrm{P}}(f)| < \infty$ for $f \in L_1(\mathrm{P}_X)$ and thus also for all $f \in L_\infty(\mathrm{P}_X)$. Moreover, the continuity is a direct consequence of *(i)*. □

Proof of Theorem 6. Since the kernel k of $\mathcal{H}$ is measurable, $\mathcal{H}$ consists of measurable functions by Lemma 27. Moreover, k is bounded, and thus Lemma 26 shows that $\mathrm{id} : \mathcal{H} \to L_\infty(\mathrm{P}_X)$ is continuous. In addition we have $L(x,y,t) \in [0,\infty)$, and hence $-\infty < L^\star(x,y,t) < \infty$ for all $(x,y,t) \in \mathcal{X}\times\mathcal{Y}\times\mathbb{R}$. Thus $L^\star$ is continuous by the convexity of $L^\star$ and Lemma 18. Therefore, Lemma 29 shows that $\mathcal{R}_{L^\star,\mathrm{P}} : L_\infty(\mathrm{P}_X) \to \mathbb{R}$ is continuous and hence $\mathcal{R}_{L^\star,\mathrm{P}} : \mathcal{H} \to \mathbb{R}$ is continuous since $\mathcal{H} \subset L_\infty(\mathrm{P}_X)$, see Lemma 26. In addition, Lemma 28 provides the convexity of this mapping. These lemmas also yield that $f \mapsto \lambda\|f\|_\mathcal{H}^2 + \mathcal{R}_{L^\star,\mathrm{P}}(f)$ is strictly convex and continuous. Proposition 19 shows that if $\mathcal{R}_{L^\star,\mathrm{P}}(f) + \lambda\|f\|_\mathcal{H}^2$ is convex and continuous and additionally $\mathcal{R}_{L^\star,\mathrm{P}}(f) + \lambda\|f\|_\mathcal{H}^2 \to \infty$ for $\|f\|_\mathcal{H} \to \infty$, then $\mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(\cdot)$ will have a minimizer. Therefore we need to show that this limit is infinite. By using (2) and (3) we obtain

$$\begin{aligned}&\mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f) \\ &\quad\ge -|L|_1\mathbb{E}_{\mathrm{P}_X}|f(X)| + \lambda\|f\|_\mathcal{H}^2 \\ &\quad\ge -|L|_1\|f\|_\infty + \lambda\|f\|_\mathcal{H}^2 \\ &\quad\ge -|L|_1\|k\|_\infty\|f\|_\mathcal{H} + \lambda\|f\|_\mathcal{H}^2 \;\to\; \infty \text{ for } \|f\|_\mathcal{H} \to \infty,\end{aligned}$$

as $|L|_1\|k\|_\infty \in [0,\infty)$ and $\lambda > 0$. □

Proof of Theorem 7. The existence and uniqueness of $f_{L^\star,\mathrm{P},\lambda}$ follow from the Theorems 5 and 6. As k is bounded, Proposition 3*(iv)* is applicable and (17) and (18) yield $\|f_{L^\star,\mathrm{P},\lambda}\|_\infty \le \lambda^{-1}|L|_1\|k\|_\infty^2 < \infty$ and $|\mathcal{R}_{L^\star,\mathrm{P}}(f_{L^\star,\mathrm{P},\lambda})| \le \lambda^{-1}|L|_1^2\|k\|_\infty^2 < \infty$. Further, the shifted loss function $L^\star$ is continuous because L and hence $L^\star$ are Lipschitz continuous. Moreover, $R : L_1(\mathrm{P}) \to \mathbb{R}$ defined by

$$R(f) := \int_{\mathcal{X}\times\mathcal{Y}} L^\star\big(x,y,f(x,y)\big)\,d\mathrm{P}(x,y), \quad f \in L_1(\mathrm{P}),$$

is well-defined and continuous. The first property follows by the definition of $L^\star$ and its Lipschitz continuity, because

$$|R(f)| \le |L|_1 \int_{\mathcal{X}\times\mathcal{Y}} |f(x,y)|\,d\mathrm{P}(x,y) < \infty, \quad f \in L_1(\mathrm{P}), \tag{36}$$

and hence R is well-defined. The continuity of R can be shown as follows. Fix $\delta > 0$ and let $f_1, f_2 \in L_1(\mathrm{P})$ with $\|f_1 - f_2\|_{L_1(\mathrm{P})} < \delta$. The Lipschitz continuity of $L^\star$ yields

$$\begin{aligned}&|R(f_1) - R(f_2)| \\ &\quad\le \int_{\mathcal{X}\times\mathcal{Y}} \big|L^\star(x,y,f_1(x,y)) - L^\star(x,y,f_2(x,y))\big|\,d\mathrm{P}(x,y) \\ &\quad\le |L|_1 \int_{\mathcal{X}\times\mathcal{Y}} |f_1(x,y) - f_2(x,y)|\,d\mathrm{P}(x,y) < \delta|L|_1,\end{aligned}$$

which shows the continuity of R. We can now apply Proposition 23 with $p = 1$ because (36) guarantees that $R(f)$ exists and is finite for all $f \in L_1(\mathrm{P})$. The subdifferential of R can thus be computed by[6]

$$\begin{aligned}\partial R(f) = \big\{h \in L_\infty(\mathrm{P}) : h(x,y) \in \partial L^\star(x,y,f(x,y)) \\ \text{for P-almost all } (x,y)\big\}.\end{aligned}$$

Now, we infer from Lemma 26 that the inclusion map $I : \mathcal{H} \to L_1(\mathrm{P})$ defined by $(If)(x,y) := f(x)$, $f \in \mathcal{H}$, $(x,y) \in \mathcal{X}\times\mathcal{Y}$, is a bounded linear operator. Moreover, for $h \in L_\infty(\mathrm{P})$ and $f \in \mathcal{H}$, the reproducing property yields

$$\begin{aligned}\langle h, If\rangle_{L_\infty(\mathrm{P}),L_1(\mathrm{P})} &= \mathbb{E}_\mathrm{P} hIf = \mathbb{E}_\mathrm{P} h\langle f,\Phi\rangle_\mathcal{H} \\ &= \langle f, \mathbb{E}_\mathrm{P} h\Phi\rangle_\mathcal{H} = \langle \iota\mathbb{E}_\mathrm{P} h\Phi, f\rangle_{\mathcal{H}',\mathcal{H}},\end{aligned}$$

where $\iota : \mathcal{H} \to \mathcal{H}'$ is the Fréchet-Riesz isomorphism described in Theorem 16. Consequently, the adjoint operator I' of I is given by $I'h = \iota\mathbb{E}_\mathrm{P} h\Phi$, $h \in L_\infty(\mathrm{P})$. Moreover, the $L^\star$-risk functional $\mathcal{R}_{L^\star,\mathrm{P}} : \mathcal{H} \to \mathbb{R}$ restricted to $\mathcal{H}$ satisfies $\mathcal{R}_{L^\star,\mathrm{P}} = R \circ I$, and hence the chain rule for subdifferentials (see Proposition 22) yields $\partial\mathcal{R}_{L^\star,\mathrm{P}}(f) = \partial(R\circ I)(f) = I'\partial R(If)$ for all $f \in \mathcal{H}$. Applying the formula for $\partial R(f)$ thus yields

$$\begin{aligned}\partial\mathcal{R}_{L^\star,\mathrm{P}}(f) = \big\{\iota\mathbb{E}_\mathrm{P} h\Phi : h \in L_\infty(\mathrm{P}) \text{ with} \\ h(x,y) \in \partial L^\star(x,y,f(x)) \text{ P-a.s.}\big\}\end{aligned}$$

[6] We have $h \in L_\infty(\mathrm{P})$ since there exists an isometric isomorphism between $(L_1(\mathrm{P}))'$ and $L_\infty(\mathrm{P})$, see, e.g., Werner [2002, Thm. II.2.4].

for all $f \in \mathcal{H}$. In addition, $f \mapsto \|f\|_{\mathcal{H}}^2$ is Fréchet-differentiable and its derivative at f is $2\iota f$ for all $f \in \mathcal{H}$. By picking suitable representations of $h \in L_\infty(\mathrm{P})$, Proposition 22 thus gives

$$\partial \mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f) = 2\lambda\iota f + \big\{\iota \mathbb{E}_\mathrm{P} h\Phi : h \in \mathcal{L}_\infty(\mathrm{P}) \text{ with } h(x,y) \in \partial L^\star(x,y,f(x)) \; \forall\, (x,y)\big\}$$

for all $f \in \mathcal{H}$. Now recall that $\mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(\,\cdot\,)$ has a minimum at $f_{L^\star,\mathrm{P},\lambda}$, and therefore we have $0 \in \partial \mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f_{L^\star,\mathrm{P},\lambda})$ by another application of Proposition 22. This together with the injectivity of ι yields the assertions (21) and (22).

Let us now show that (23) holds. Since k is a bounded kernel, we have by (17) and (18) that

$$\|f_{L^\star,\mathrm{P},\lambda}\|_\infty \le \lambda^{-1}|L|_1 \, \|k\|_\infty^2 := B_\lambda < \infty.$$

Now (21) and Proposition 21 with $\delta := 1$ yield, for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$,

$$|h(x,y)| \le \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \big|\partial L^\star(x,y,f_{L^\star,\mathrm{P},\lambda}(x))\big| \le |L|_1$$

and hence we have shown $h \in \mathcal{L}_\infty(\mathrm{P})$ and (23).

Let us now establish (24). To this end, observe that we have by (21) and the definition of the subdifferential

$$\begin{aligned} &h(x,y)\big(f_{L^\star,\bar{\mathrm{P}},\lambda}(x) - f_{L^\star,\mathrm{P},\lambda}(x)\big) \\ &\quad \le L^\star\big(x,y,f_{L^\star,\bar{\mathrm{P}},\lambda}(x)\big) - L^\star\big(x,y,f_{L^\star,\mathrm{P},\lambda}(x)\big) \end{aligned}$$

for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$. By integrating with respect to $\bar{\mathrm{P}}$, we hence obtain

$$\begin{aligned} (37) \qquad &\langle f_{L^\star,\bar{\mathrm{P}},\lambda} - f_{L^\star,\mathrm{P},\lambda}, \, \mathbb{E}_{\bar{\mathrm{P}}} h\Phi\rangle_{\mathcal{H}} \\ &\quad \le \; \mathcal{R}_{L^\star,\bar{\mathrm{P}}}(f_{L^\star,\bar{\mathrm{P}},\lambda}) - \mathcal{R}_{L^\star,\bar{\mathrm{P}}}(f_{L^\star,\mathrm{P},\lambda}). \end{aligned}$$

Moreover, an easy calculation shows

$$\begin{aligned} (38) \qquad &2\lambda \, \langle f_{L^\star,\bar{\mathrm{P}},\lambda} - f_{L^\star,\mathrm{P},\lambda}, \, f_{L^\star,\mathrm{P},\lambda}\rangle_{\mathcal{H}} \\ &\quad + \lambda \, \big\|f_{L^\star,\mathrm{P},\lambda} - f_{L^\star,\bar{\mathrm{P}},\lambda}\big\|_{\mathcal{H}}^2 \\ &\qquad = \lambda \, \big\|f_{L^\star,\bar{\mathrm{P}},\lambda}\big\|_{\mathcal{H}}^2 - \lambda \, \|f_{L^\star,\mathrm{P},\lambda}\|_{\mathcal{H}}^2 . \end{aligned}$$

By combining (37) and (38), we thus find

$$\begin{aligned} &\big\langle f_{L^\star,\bar{\mathrm{P}},\lambda} - f_{L^\star,\mathrm{P},\lambda}, \, \mathbb{E}_{\bar{\mathrm{P}}} h\Phi + 2\lambda f_{L^\star,\mathrm{P},\lambda}\big\rangle_{\mathcal{H}} \\ &\quad + \lambda \, \big\|f_{L^\star,\mathrm{P},\lambda} - f_{L^\star,\bar{\mathrm{P}},\lambda}\big\|_{\mathcal{H}}^2 \\ &\qquad \le \mathcal{R}^{reg}_{L^\star,\bar{\mathrm{P}},\lambda}(f_{L^\star,\bar{\mathrm{P}},\lambda}) - \mathcal{R}^{reg}_{L^\star,\bar{\mathrm{P}},\lambda}(f_{L^\star,\mathrm{P},\lambda}) \le 0, \end{aligned}$$

and consequently the representation $f_{L^\star,\mathrm{P},\lambda} = -\frac{1}{2\lambda}\,\mathbb{E}_\mathrm{P} h\Phi$ yields in combination with the Cauchy-Schwarz inequality that

$$\begin{aligned} &\lambda \, \big\|f_{L^\star,\mathrm{P},\lambda} - f_{L^\star,\bar{\mathrm{P}},\lambda}\big\|_{\mathcal{H}}^2 \\ &\quad \le \big\langle f_{L^\star,\mathrm{P},\lambda} - f_{L^\star,\bar{\mathrm{P}},\lambda}, \, \mathbb{E}_{\bar{\mathrm{P}}} h\Phi - \mathbb{E}_\mathrm{P} h\Phi\big\rangle_{\mathcal{H}} \\ &\quad \le \big\|f_{L^\star,\mathrm{P},\lambda} - f_{L^\star,\bar{\mathrm{P}},\lambda}\big\|_{\mathcal{H}} \cdot \|\mathbb{E}_{\bar{\mathrm{P}}} h\Phi - \mathbb{E}_\mathrm{P} h\Phi\|_{\mathcal{H}} . \end{aligned}$$

From this we easily obtain (24).

It remains to show (25) for the special case of a distance-based loss function. By the definition of the subdifferential we obtain for L and $L^\star$ that, for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$,

$$\begin{aligned} &\partial L^\star(x,y,t) \\ &\quad = \big\{t' \in \mathbb{R}' : \langle t', v - t\rangle \le L^\star(x,y,v) - L^\star(x,y,t) \; \forall\, v \in \mathbb{R}\big\} \\ &\quad = \big\{t' \in \mathbb{R}' : \langle t', v - t\rangle \le L(x,y,v) - L(x,y,t) \; \forall\, v \in \mathbb{R}\big\} \\ &\quad = \partial L(x,y,t), \quad t \in \mathbb{R}. \end{aligned}$$

Hence $\partial L(f) = \partial L^\star(f)$ for all measurable functions $f : \mathcal{X} \to \mathbb{R}$. If we combine this with Proposition 22, it follows, for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$, that $\partial L^\star(x,y,t) = \partial L(x,y,t) = -\partial\psi(y-t)$ for all $t \in \mathbb{R}$, and therefore (21) implies (25). □

Proof of Theorem 8. *(i)* To avoid handling too many constants, let us assume $\|k\|_\infty = 1$. This implies $\|f\|_\infty \le \|k\|_\infty \, \|f\|_{\mathcal{H}} \le \|f\|_{\mathcal{H}}$ for all $f \in \mathcal{H}$. Now we use the Lipschitz continuity of L (and thus also of $L^\star$), $|L|_1 < \infty$, and Lemma 25 to obtain, for all $g \in \mathcal{H}$,

$$(39) \qquad \big|\mathcal{R}_{L^\star,\mathrm{P}}(f_{L^\star,\mathrm{P},\lambda_n}) - \mathcal{R}_{L^\star,\mathrm{P}}(g)\big| \le |L|_1 \, \|f_{L^\star,\mathrm{P},\lambda_n} - g\|_{\mathcal{H}} .$$

For $n \in \mathbb{N}$ and $\lambda_n > 0$, we write $h_n := h_{L^\star,n} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ for the function h obtained by the representer theorem 7. Let $\Phi : \mathcal{X} \to \mathcal{H}$ be the canonical feature map. We have $f_{L^\star,\mathrm{P},\lambda_n} = -(2\lambda_n)^{-1}\mathbb{E}_\mathrm{P} h_n\Phi$, and for all distributions Q on $\mathcal{X} \times \mathcal{Y}$, we have

$$\|f_{L^\star,\mathrm{P},\lambda_n} - f_{L^\star,\mathrm{Q},\lambda_n}\|_{\mathcal{H}} \le \lambda_n^{-1} \, \|\mathbb{E}_\mathrm{P} h_n\Phi - \mathbb{E}_\mathrm{Q} h_n\Phi\|_{\mathcal{H}} .$$

Note that $\|h_n\|_\infty \le |L|_1$ due to (23). Moreover, let $\varepsilon \in (0,1)$ and D be a training set of n data points and corresponding empirical distribution D such that

$$(40) \qquad \|\mathbb{E}_\mathrm{P} h_n\Phi - \mathbb{E}_\mathrm{D} h_n\Phi\|_{\mathcal{H}} \le \frac{\lambda_n\varepsilon}{|L|_1}.$$

Then Theorem 7 gives $\|f_{L^\star,\mathrm{P},\lambda_n} - f_{L^\star,\mathrm{D}_n,\lambda_n}\|_{\mathcal{H}} \le \frac{\varepsilon}{|L|_1}$ and hence (39) yields

$$\begin{aligned} (41) \qquad &\big|\mathcal{R}_{L^\star,\mathrm{P}}(f_{L^\star,\mathrm{P},\lambda_n}) - \mathcal{R}_{L^\star,\mathrm{P}}(f_{L^\star,\mathrm{D},\lambda_n})\big| \\ &\quad \le |L|_1 \cdot \|f_{L^\star,\mathrm{P},\lambda_n} - f_{L^\star,\mathrm{D},\lambda_n}\|_{\mathcal{H}} \le \varepsilon. \end{aligned}$$

Let us now estimate the probability of D satisfying (40). To this end, we first observe that $\lambda_n n^{1/2} \to \infty$ implies that $\lambda_n\varepsilon \ge n^{-1/2}$ for all sufficiently large $n \in \mathbb{N}$. Moreover, Theorem 7 shows $\|h_n\|_\infty \le |L|_1$, and our assumption $\|k\|_\infty = 1$ thus yields $\|h_n\Phi\|_\infty \le |L|_1$. Consequently, Hoeffding's inequality in Hilbert spaces (see Theorem 17) yields for $B = 1$ and

$$\xi = \frac{3}{8}\,\frac{|L|_1^{-2}\varepsilon^2\lambda_n^2 n}{|L|_1^{-1}\varepsilon\lambda_n + 3}$$

the bound

$$\begin{aligned}
&\mathrm{P}^n\Big(D \in (\mathcal{X}\times\mathcal{Y})^n : \|\mathbb{E}_{\mathrm{P}} h_n\Phi - \mathbb{E}_D h_n\Phi\|_{\mathcal{H}} \le \frac{\lambda_n\varepsilon}{|L|_1}\Big)\\
&\quad\ge \mathrm{P}^n\Big(D \in (\mathcal{X}\times\mathcal{Y})^n : \|\mathbb{E}_{\mathrm{P}} h_n\Phi - \mathbb{E}_D h_n\Phi\|_{\mathcal{H}} \le\\
&\qquad\qquad (\sqrt{2\xi}+1)n^{-1/2} + \frac{4\xi}{3n}\Big)\\
&\quad\ge 1 - \exp\Big(-\frac{3}{8}\cdot\frac{\varepsilon^2\lambda_n^2 n/|L|_1^2}{\varepsilon\lambda_n/|L|_1 + 3}\Big)\\
&\quad= 1 - \exp\Big(-\frac{3}{8}\cdot\frac{\varepsilon^2\lambda_n^2 n}{(\varepsilon\lambda_n + 3|L|_1)|L|_1}\Big)
\end{aligned}$$

for all sufficiently large values of n. Now using $\lambda > 0$, $\lambda_n \to 0$ and $\lambda_n n^{1/2} \to \infty$, we find that the probability of sample sets D satisfying (40) converges to 1 if $|D| = n \to \infty$. As we have seen above, this implies that (41) holds true with probability tending to 1. Now, since $\lambda_n > 0$ and $\lambda_n \to 0$, $n \to \infty$, we additionally have $|\mathcal{R}_{L^\star,\mathrm{P}}(f_{L^\star\mathrm{P},\lambda_n}) - \mathcal{R}^*_{L^\star,\mathrm{P}}| \le \varepsilon$ for all sufficiently large n, and hence we obtain the assertion of $L^\star$-risk consistency of $f_{L^\star\mathrm{P},\lambda_n}$.

(ii) In order to show the second assertion, we define $\varepsilon_n := (\ln(n+1))^{-1/2}$ and

$$\delta_n := \mathcal{R}_{L^\star,\mathrm{P}}(f_{L^\star,\mathrm{P},\lambda_n}) - \mathcal{R}^*_{L^\star,\mathrm{P}} + \varepsilon_n, \quad n \in \mathbb{N}.$$

Moreover, for an infinite sample

$$D_\infty := ((x_1,y_1),(x_2,y_2),\ldots) \in (\mathcal{X}\times\mathcal{Y})^\infty,$$

we write $D_n := ((x_1,y_1),\ldots,(x_n,y_n))$. With these notations, we define, for $n \in \mathbb{N}$,

$$A_n := \big\{D_\infty \in (\mathcal{X}\times\mathcal{Y})^\infty : \mathcal{R}_{L^\star,\mathrm{P}}(f_{L^\star,\mathrm{D}_n,\lambda_n}) - \mathcal{R}^*_{L^\star,\mathrm{P}} > \delta_n\big\}.$$

Now, our estimates above together with $\lambda_n^{2+\delta} n \to \infty$ for some $\delta > 0$ yield

$$\sum_{n\in\mathbb{N}} \mathrm{P}^\infty(A_n) \;\le\; \sum_{n\in\mathbb{N}} \exp\left(-\frac{3}{8}\cdot\frac{\varepsilon_n^2\lambda_n^2 n}{(\varepsilon_n\lambda_n + 3|L|_1)|L|_1}\right) < \infty.$$

We obtain by the Borel-Cantelli lemma [see e.g., Dudley, 2002] that

$$\begin{aligned}
\mathrm{P}^\infty\big(\,\{\, D_\infty \in (\mathcal{X}\times\mathcal{Y})^\infty : \exists\, n_0\, \forall\, n \ge n_0 \text{ with}\\
\mathcal{R}_{L^\star,\mathrm{P}}(f_{L^\star,\mathrm{D}_n,\lambda_n}) - \mathcal{R}^*_{L^\star,\mathrm{P}} \le \delta_n\big\}\,\big) = 1.
\end{aligned}$$

The assertion follows because $\lambda_n \to 0$ implies $\delta_n \to 0$. □

Before we can prove Theorem 9, we need some prerequisites on self-calibrated loss functions and related results. Let $\tau \in (0,1)$, L be a pinball loss function, and $L^\star$ its shifted version. Hence, L is Lipschitz continuous, convex, and $L(x,y,t) = \psi(y-t) \to \infty$ for $|t| \to \infty$. Our goal is to extend the consistency results derived in Christmann and Steinwart [2008] to *all* distributions P on $\mathcal{X}\times\mathbb{R}$. To this end, we adopt the inner risk notation from Steinwart and Christmann [2008b, Chapter 3] by writing, for $t \in \mathbb{R}$,

$$\mathcal{C}_{L^\star,\mathrm{Q}}(t) := \int_{\mathbb{R}} L^\star(x,y,t)\,d\mathrm{Q}(y) = \int_{\mathbb{R}} \psi(y-t) - \psi(y)\,d\mathrm{Q}(y),$$

where Q is a distribution on $\mathbb{R}$ that will serve us as a template for the conditional distribution $\mathrm{P}(\,\cdot\,|x)$. Similarly, we write $\mathcal{C}^*_{L^\star,\mathrm{Q}} := \inf_{t\in\mathbb{R}} \mathcal{C}_{L^\star,\mathrm{Q}}(t)$ for the minimal inner $L^\star$-risk. Note that, like for the $L^\star$-risk, we have $|\mathcal{C}^*_{L^\star,\mathrm{Q}}| < \infty$. Finally, for $\varepsilon \in [0,\infty]$, we denote the set of *$\varepsilon$-approximate minimizers* by

$$\mathcal{M}_{L^\star,\mathrm{Q}}(\varepsilon) := \big\{t \in \mathbb{R} : \mathcal{C}_{L^\star,\mathrm{Q}}(t) - \mathcal{C}^*_{L^\star,\mathrm{Q}} < \varepsilon\big\}$$

and the set of *exact minimizers* by

$$\mathcal{M}_{L^\star,\mathrm{Q}}(0^+) := \bigcap_{\varepsilon>0} \mathcal{M}_{L^\star,\mathrm{Q}}(\varepsilon) = \big\{t \in \mathbb{R} : \mathcal{C}_{L^\star,\mathrm{Q}}(t) = \mathcal{C}^*_{L^\star,\mathrm{Q}}\big\}.$$

Since $|\mathcal{C}^*_{L^\star,\mathrm{Q}}| < \infty$ it is easy to verify that these notations coincide with those of Steinwart and Christmann [2008b, Chapter 3] modulo the fact that we now consider the shifted loss function $L^\star$ rather than L. The following proposition, which is an $L^\star$-analogue to Steinwart and Christmann [2008b, Prop. 3.9], computes the $L^\star$-excess risk and the set of exact minimizers.

Proposition 30. *For $\tau \in (0,1)$, let L be the τ-pinball loss and $L^\star$ its shifted version. Moreover, let* Q *be a distribution on $\mathbb{R}$ and t^* be a τ-quantile of* Q*, i.e., we have*

$$\mathrm{Q}\big((-\infty,t^*]\big) \ge \tau \quad \text{and} \quad \mathrm{Q}\big([t^*,\infty)\big) \ge 1-\tau.$$

Then there exist real numbers $q_+, q_- \ge 0$ such that $q_+ + q_- = \mathrm{Q}(\{t^*\})$ and

$$\mathcal{C}_{L^\star,\mathrm{Q}}(t^*+t) - \mathcal{C}^*_{L^\star,\mathrm{Q}} = tq_+ + \int_0^t \mathrm{Q}\big((t^*,t^*+s)\big)\,ds, \tag{42}$$

$$\mathcal{C}_{L^\star,\mathrm{Q}}(t^*-t) - \mathcal{C}^*_{L^\star,\mathrm{Q}} = tq_- + \int_0^t \mathrm{Q}\big((t^*-s,t^*)\big)\,ds, \tag{43}$$

for all $t \ge 0$. Moreover, we have

$$\begin{aligned}
\mathcal{M}_{L^\star,\mathrm{Q}}(0^+) = \{t^*\} \cup \big\{t > t^* : q_+ + \mathrm{Q}((t^*,t)) = 0\big\}\\
\cup \big\{t < t^* : q_- + \mathrm{Q}((-t,t^*)) = 0\big\}.
\end{aligned}$$

Proof of Proposition 30. Let us consider the distribution $\mathrm{Q}^{(t^*)}$ defined by $\mathrm{Q}^{(t^*)}(A) := \mathrm{Q}(t^*+A)$ for all measurable sets $A \subset \mathbb{R}$. Then it is not hard to see that 0 is a τ-quantile of $\mathrm{Q}^{(t^*)}$. Moreover, we obviously have $\mathcal{C}_{L^\star,\mathrm{Q}}(t^*+t) = \mathcal{C}_{L^\star,\mathrm{Q}^{(t^*)}}(t)$. Therefore, we may assume without loss of generality that $t^* = 0$. Then our assumptions together with $\mathrm{Q}((-\infty,0]) + \mathrm{Q}([0,\infty)) = 1 + \mathrm{Q}(\{0\})$ yield $\tau \le \mathrm{Q}((-\infty,0]) \le \tau + \mathrm{Q}(\{0\})$, i.e., there exists a $q_+ \in \mathbb{R}$ satisfying $0 \le q_+ \le \mathrm{Q}(\{0\})$ and

$$\mathrm{Q}((-\infty,0]) = \tau + q_+. \tag{44}$$

Let us now prove the first expression for the excess inner risks of $L^\star$. To this end, we first observe that, for $t \geq 0$, we have

$$\begin{aligned}
&\mathcal{C}_{L^\star,\mathrm{Q}}(t) \\
&\quad = (1-\tau)\int_{y<0}(t-y) + y\,d\mathrm{Q}(y) \\
&\qquad + \int_{0\leq y<t}(1-\tau)(t-y) - \tau y\,d\mathrm{Q}(y) + \tau\int_{y\geq t}(y-t) - y\,d\mathrm{Q}(y) \\
&\quad = (1-\tau)t\mathrm{Q}((-\infty,0)) + \int_{0\leq y<t}(1-\tau)t - y\,d\mathrm{Q}(y) - \tau t\int_{y\geq t} d\mathrm{Q}(y) \\
&\quad = (1-\tau)t\mathrm{Q}((-\infty,t)) - \int_{0\leq y<t} y\,d\mathrm{Q}(y) - \tau t\mathrm{Q}([t,\infty)) \\
&\quad = t\mathrm{Q}((-\infty,0)) - \tau t + t\mathrm{Q}([0,t)) - \int_{0\leq y<t} y\,d\mathrm{Q}(y).
\end{aligned}$$

Moreover, using a well-known relationship between expectations and tail bounds, see Bauer [2001, p. 141], we get

$$\begin{aligned}
t\,\mathrm{Q}([0,t)) - \int_{0\leq y<t} y\,d\mathrm{Q}(y) &= \int_0^t \mathrm{Q}([0,t))\,ds - \int_0^t \mathrm{Q}([s,t))\,ds \\
&= t\,\mathrm{Q}(\{0\}) + \int_0^t \mathrm{Q}((0,s))\,ds,
\end{aligned}$$

and since (44) implies

$$\mathrm{Q}((-\infty,0)) + \mathrm{Q}(\{0\}) = \mathrm{Q}((-\infty,0]) = \tau + q_+,$$

we thus obtain

$$\mathcal{C}_{L^\star,\mathrm{Q}}(t) = tq_+ + \int_0^t \mathrm{Q}\big((0,s)\big)\,ds.$$

Applying this equation to the pinball loss with parameter $1-\tau$ and the distribution $\bar{\mathrm{Q}}$ defined by $\bar{\mathrm{Q}}(A) := \mathrm{Q}(-A)$, $A \subset \mathbb{R}$ measurable, gives a real number $0 \leq q_- \leq \mathrm{Q}(\{0\})$ such that $\mathrm{Q}([0,\infty)) = 1-\tau+q_-$ and

$$\mathcal{C}_{L^\star,\mathrm{Q}}(-t) = tq_- + \int_0^t \mathrm{Q}\big((-s,0)\big)\,ds$$

for all $t \geq 0$. Consequently, $t^* = 0$ is a minimizer of $\mathcal{C}_{L^\star,\mathrm{Q}}(\,\cdot\,)$ and we have $\mathcal{C}^*_{L^\star,\mathrm{Q}} = \mathcal{C}_{L^\star,\mathrm{Q}}(0) = 0$. From this we conclude both (42) and (43). Moreover, combining $\mathrm{Q}([0,\infty)) = 1-\tau+q_-$ with (44), we find $q_+ + q_- = \mathrm{Q}(\{0\})$. Finally, the formula for the set of exact minimizers is an obvious consequence of (42) and (43). □

In order to investigate how well approximate $L^\star$-risk minimizers approximate the exact $L^\star$-risk minimizers, we further have to adopt the *self-calibration approach* of Steinwart and Christmann [2008b, Chapter 3]. Fortunately, the fact that we always have $|\mathcal{C}^*_{L^\star,\mathrm{Q}}| < \infty$ makes our considerations a little easier than those in Steinwart and Christmann [2008b, Chapter 3] for general loss functions. To further decrease the notational burden we assume in the following that the considered distribution Q on $\mathbb{R}$ has a *unique* τ-quantile, denoted by $t^*_{\tau,\mathrm{Q}}$ or simply t^* if no confusion can arise. Fortunately, this uniqueness assumption is by no means necessary, and we refer the interested reader to Steinwart and Christmann [2008b, Chapter 3] for a modification to this general situation.

With these preparations, the $L^\star$-generalization of the *self-calibration function* now reads as follows:

$$\delta_{\max}(\varepsilon,\mathrm{Q}) := \inf_{|t-t^*|\geq\varepsilon} \mathcal{C}_{L^\star,\mathrm{Q}}(t) - \mathcal{C}^*_{L^\star,\mathrm{Q}}, \quad \varepsilon > 0.$$

Note that, for $t \in \mathbb{R}$ and $\varepsilon := |t-t^*|$, we have

$$\delta_{\max}(|t-t^*|,\mathrm{Q}) = \delta_{\max}(\varepsilon,\mathrm{Q}) \leq \mathcal{C}_{L^\star,\mathrm{Q}}(t) - \mathcal{C}^*_{L^\star,\mathrm{Q}},$$

i.e., as for standard loss functions $\delta_{\max}(\varepsilon,\mathrm{Q})$ measures how well approximate $\mathcal{C}_{L^\star,\mathrm{Q}}(\,\cdot\,)$-minimizers approximate the exact minimizer t^*. Moreover, by Proposition 30 we conclude that, for all $\varepsilon > 0$, we have

$$\begin{aligned}
\delta_{\max}(\varepsilon,\mathrm{Q}) = \min\Big\{&\varepsilon q_+ + \int_0^\varepsilon \mathrm{Q}\big((t^*,t^*+s)\big)\,ds, \\
&\varepsilon q_- + \int_0^\varepsilon \mathrm{Q}\big((t^*-s,t^*)\big)\,ds\Big\} > 0,
\end{aligned}$$

where we used the assumption that t^* is the only τ-quantile, i.e., the only exact $\mathcal{C}_{L^\star,\mathrm{Q}}(\,\cdot\,)$-minimizer. Since the proofs of Theorem 3.61 and its Corollary 3.62 in Steinwart and Christmann [2008b] only consider excess inner risks and not the underlying loss function itself, a literal repetition of these proofs then yields the following result.

Corollary 31. *For $\tau \in (0,1)$, let L be the τ-pinball loss and $L^\star$ its shifted version. Moreover, let P be a distribution on $\mathcal{X}\times\mathbb{R}$ whose conditional τ-quantile $f^*_{\tau,\mathrm{P}} : \mathcal{X} \to \mathbb{R}$ is P_X-almost surely unique. Then, for all sequences (f_n) of measurable functions $f_n : \mathcal{X} \to \mathbb{R}$, the convergence*

$$\mathcal{R}_{L^\star,\mathrm{P}}(f_n) \to \mathcal{R}^*_{L^\star,\mathrm{P}}$$

implies

$$f_n \to f^*_{\tau,\mathrm{P}} \quad \text{in probability } \mathrm{P}_X.$$

Proof of Theorem 9. Due to the assumptions, Theorem 8 is applicable and hence $f_{L^\star,\mathrm{D},\lambda_n}$ satisfies $\mathcal{R}_{L^\star,\mathrm{P}}(f_{L^\star,\mathrm{D},\lambda_n}) \to \mathcal{R}^*_{L^\star,\mathrm{P}}$ in probability (or almost surely) for $n \to \infty$. The existence of a unique minimizer $f^*_{\tau,\mathrm{P}}$ is guaranteed by the assumptions of Theorem 9. Hence, Corollary 31 yields the assertion. □

Proof of Theorem 10. Let $z = (x,y) \in \mathcal{X}\times\mathcal{Y}$. The two key ingredients of our analysis are the function $G : \mathbb{R}\times\mathcal{H} \to \mathcal{H}$ defined by

$$G(\varepsilon,f) := 2\lambda f + \mathbb{E}_{(1-\varepsilon)\mathrm{P}+\varepsilon\delta_z} \nabla_3^F L^\star(X,Y,f(X))\Phi(X), \tag{45}$$

and the application of an implicit function theorem for Fréchet-derivatives. Let us first check that G is well-defined. Recall that every function $f \in \mathcal{H}$ is bounded because we assumed that $\mathcal{H}$ has a bounded kernel k. By using (19) and (28) we get $\mathbb{E}_{\mathrm{P}}|\nabla_3^F L^\star(X,Y,f(X))| \le \kappa_1 \in (0,\infty)$ for all $f \in \mathcal{H}$. As $\Phi(x) := k(x,\cdot) \in \mathcal{H}$ for all $x \in \mathcal{X}$, we obtain that $\Phi : \mathcal{X} \to \mathcal{H}$ is a bounded mapping. Therefore, the $\mathcal{H}$-valued (Bochner) integral used in the definition of G is well-defined for all $\varepsilon \in \mathbb{R}$ and all $f \in \mathcal{H}$. Note that for $\varepsilon \notin [0,1]$ the $\mathcal{H}$-valued integral in (45) is with respect to a signed measure. As in Christmann and Steinwart [2007] we obtain for $\varepsilon \in [0,1]$ the equation

(46)
$$G(\varepsilon, f) = \frac{\partial \mathcal{R}^{reg}_{L^\star,(1-\varepsilon)\mathrm{P}+\varepsilon\delta_z,\lambda}}{\partial \mathcal{H}}(f) = \nabla_3^F \mathcal{R}^{reg}_{L^\star,(1-\varepsilon)\mathrm{P}+\varepsilon\delta_z,\lambda}(f).$$

Given an $\varepsilon \in [0,1]$, the function $f \mapsto \mathcal{R}^{reg}_{L^\star,(1-\varepsilon)\mathrm{P}+\varepsilon\delta_z,\lambda}(f)$ is convex and continuous (see the proof of Theorem 6) and hence (46) shows that $G(\varepsilon, f) = 0$ if and only if $f = f_{L^\star,(1-\varepsilon)\mathrm{P}+\varepsilon\delta_z,\lambda}$. Our aim is to show the existence of a Fréchet-differentiable function $\varepsilon \mapsto f_\varepsilon$ defined on a small interval $(-\delta,\delta)$ for some $\delta > 0$ that satisfies $G(\varepsilon, f_\varepsilon) = 0$ for all $\varepsilon \in (-\delta,\delta)$. Once we have shown the existence of this function, we immediately obtain

$$\mathrm{IF}(z; T, \mathrm{P}) = \nabla^F f_\varepsilon(0).$$

For the existence of $\varepsilon \mapsto f_\varepsilon$ we have to check by Theorem 24 that G is continuously differentiable and that $\nabla_2^F G(0, f_{L^\star,\mathrm{P},\lambda})$ is invertible. Let us start with the first. By the definition of G and by using $\nabla_3^F L^\star(x,y,\cdot) = \nabla_3^F L(x,y,\cdot)$ for all $(x,y) \in \mathcal{X}\times\mathcal{Y}$, we get

(47)
$$\begin{aligned}
&\nabla_1^F G(\varepsilon, f)\\
&\quad = -\mathbb{E}_{\mathrm{P}} \nabla_3^F L^\star(X,Y,f(X))\Phi(X) + \nabla_3^F L^\star(x,y,f(x))\Phi(x)\\
&\quad = -\mathbb{E}_{\mathrm{P}} \nabla_3^F L(X,Y,f(X))\Phi(X) + \nabla_3^F L(x,y,f(x))\Phi(x).
\end{aligned}$$

A similar, but slightly more involved computation using (19) and (30) yields

(48)
$$\begin{aligned}
&\nabla_2^F G(\varepsilon, f)\\
&\qquad = \mathbb{E}_{(1-\varepsilon)\mathrm{P}+\varepsilon\delta_z} \nabla_{3,3}^F L(X,Y,f(X))\langle \Phi(X), \cdot\rangle \Phi(X)\\
&\qquad\quad + 2\lambda \mathrm{id}_{\mathcal{H}},
\end{aligned}$$

which equals S. To prove that $\nabla_1^F G$ is continuous, we fix $\varepsilon \in \mathbb{R}$ and a sequence $(f_n)_{n\in\mathbb{N}}$ such that $f_n \in \mathcal{H}$ for all $n \in \mathbb{N}$ and $\lim_{n\to\infty} f_n = f \in \mathcal{H}$. Since k is bounded, the sequence $(f_n)_{n\in\mathbb{N}}$ is uniformly bounded. By (28), we have, for all $(x,y,t) \in \mathcal{X}\times\mathcal{Y}\times\mathbb{R}$, that $|\nabla_3^F L(x,y,t)| \le \kappa_1 + |t|$. Hence $|\nabla_3^F L|$ is a P-integrable Nemitski loss function for *all* probability measures P, because we only have to choose the constant function $b(x,y) \equiv \kappa_1$ in the definition of a P-integrable Nemitski loss defined in the introduction. We can thus find a bounded, measurable function $g : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$ with $|\nabla_3^F L^\star(x,y,f_n(x))| \le |\nabla_3^F L^\star(x,y,g(y))|$ for all $n \in \mathbb{N}$ and all $(x,y) \in \mathcal{X}\times\mathcal{Y}$. For the function $v : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$ with $v(x,y) := L^\star(x,y,g(y))$, we hence obtain by the definition of $L^\star$ and by the Lipschitz continuity of L that

$$\begin{aligned}
&\int_{\mathcal{X}\times\mathcal{Y}} |v(X,Y)|\, d\mathrm{P}\\
&\qquad = \int_{\mathcal{X}\times\mathcal{Y}} |L(X,Y,g(Y)) - L(X,Y,0)|\, d\mathrm{P} \le |L|_1\, \|g\|_\infty
\end{aligned}$$

is finite for *all* $\mathrm{P} \in \mathcal{M}_1(\mathcal{X}\times\mathcal{Y})$. Thus, an application of the dominated convergence theorem for Bochner integrals, see Diestel and Uhl [1977, Thm. 3, p. 45], gives the continuity of $\nabla_1^F G$. Because the continuity of G and $\nabla_2^F G$ can be shown analogously, we obtain that G is continuously differentiable, see for example Akerkar [1999, Thm. 2.6].

To show that $\nabla_2^F G(0, f_{L^\star,\mathrm{P},\lambda})$ is invertible, it suffices by the Fredholm alternative (see Theorem 15) to show that $\nabla_2^F G(0, f_{L^\star,\mathrm{P},\lambda})$ is injective and that

$$Ag := \mathbb{E}_{\mathrm{P}} \nabla_{3,3}^F L^\star(X,Y,f_{L^\star,\mathrm{P},\lambda}(X)) g(X)\Phi(X), \quad g \in \mathcal{H},$$

defines a compact operator on $\mathcal{H}$. To show the compactness of the operator A, recall that $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{X}\times\mathcal{Y}$ are Polish spaces because $\mathcal{X}$ is a complete separable metric space and $\mathcal{Y} \subset \mathbb{R}$ is closed, see Dudley [2002]. Furthermore, Borel probability measures on Polish spaces are regular by Ulam's theorem, that is, they can be approximated from inside by compact sets. Hence, there exists a sequence of measurable compact subsets $\mathcal{X}_n \times \mathcal{Y}_n \subset \mathcal{X}\times\mathcal{Y}$ with $\mathrm{P}(\mathcal{X}_n\times\mathcal{Y}_n) \ge 1 - \frac{1}{n}$, $n \in \mathbb{N}$. Let us also define a sequence of operators $A_n : \mathcal{H} \to \mathcal{H}$, where $A_n g$ equals

$$\int_{\mathcal{X}_n}\int_{\mathcal{Y}_n} \nabla_{3,3}^F L^\star(x,y,f_{L^\star,\mathrm{P},\lambda}(x))\, \mathrm{P}(dy|x)\, g(x)\Phi(x)\, d\mathrm{P}_X(x)$$

for all $g \in \mathcal{H}$. Note that if $\mathcal{X}\times\mathcal{Y}$ is compact, we can choose $\mathcal{X}_n \times \mathcal{Y}_n := \mathcal{X}\times\mathcal{Y}$, which implies $A = A_n$. Let us now show that A_n, $n \ge 1$, is a compact operator. To this end we assume without loss of generality that $\|k\|_\infty \le 1$. Denote the closed unit ball in $\mathcal{H}$ by $B_{\mathcal{H}}$. For $g \in B_{\mathcal{H}}$ and $x \in \mathcal{X}$, we have due to the assumption (28) that

$$\begin{aligned}
h_g(x) &:= \int_{\mathcal{Y}_n} \nabla_{3,3}^F L^\star(x,y,f_{L^\star,\mathrm{P},\lambda}(x))\, |g(x)|\, \mathrm{P}(dy|x)\\
&\le \kappa_2 \|g\|_\infty =: h(x).
\end{aligned}$$

Therefore, we have $h \in L_1(\mathrm{P})$, which implies $h_g \in L_1(\mathrm{P})$ with $\|h_g\|_1 \le \|h\|_1 < \infty$ for all $g \in B_{\mathcal{H}}$. Consequently, $\mu_g := h_g \mathrm{P}_X$ and $\mu := h\mathrm{P}_X$ are finite measures. By Diestel and Uhl

[1977, Cor. 8, p. 48] we hence obtain

$$\begin{aligned} A_n g &:= \int_{\mathcal{X}_n} \operatorname{sign} g(x)\Phi(x)h_g(x)\,d\mathrm{P}_X(x) \\ &= \int_{\mathcal{X}_n} \operatorname{sign} g(x)\Phi(x)\,d\mu_g(x) \\ &\in \mu_g(\mathcal{X}_n)\,\overline{\operatorname{aco}\Phi(X_n)} \subset \mu(\mathcal{X}_n)\,\overline{\operatorname{aco}\Phi(X_n)}, \quad g \in \mathcal{H}, \end{aligned}$$

where $\operatorname{aco}\Phi(X_n)$ denotes the absolute convex hull of $\Phi(X_n)$, and the closure is with respect to $\|\cdot\|_{\mathcal{H}}$. The continuity of k yields the continuity of the canonical feature map Φ. Thus, $\Phi(X_n)$ is compact and hence so is the closure of $\operatorname{aco}\Phi(X_n)$. This shows that A_n is a compact operator.

To see that A is compact, it therefore suffices to show $\|A_n - A\| \to 0$ with respect to the operator norm for $n \to \infty$. Recalling that the convexity of $L^\star$ and the existence of its second derivative implies $\nabla^F_{3,3}L^\star(x,y,\cdot) \geq 0$ for all $(x,y) \in \mathcal{X}\times\mathcal{Y}$, it follows from (28) that

$$0 \leq \int \nabla^F_{3,3}L^\star(x,y,f_{L^\star,\mathrm{P},\lambda}(x))\,d\mathrm{P}(x,y) \leq \kappa_2,$$

which shows due to (19) that $\nabla^F_{3,3}L^\star(\cdot,\cdot,f_{L^\star,\mathrm{P},\lambda}(\cdot)) = \nabla^F_{3,3}L(\cdot,\cdot,f_{L^\star,\mathrm{P},\lambda}(\cdot)) \in L_\infty(\mathrm{P})$ for all $\mathrm{P} \in \mathcal{M}_1(\mathcal{X}\times\mathcal{Y})$. Now define $B := (\mathcal{X}\times\mathcal{Y})\backslash(\mathcal{X}_n\times\mathcal{Y}_n)$. Then the desired convergence follows from (2) and (3), $\mathrm{P}(\mathcal{X}_n\times\mathcal{Y}_n) \geq 1-\frac{1}{n}$, and

$$\begin{aligned} &\|A_n g - Ag\|_{\mathcal{H}} \\ &\quad\leq \int_B \nabla^F_{3,3}L^\star(x,y,f_{L^\star,\mathrm{P},\lambda}(x))\,|g(x)|\,\|\Phi(x)\|_{\mathcal{H}}\,d\mathrm{P}(x,y) \\ &\quad\leq \|g\|_\infty \|\Phi(x)\|_{\mathcal{H}} \int_B \nabla^F_{3,3}L^\star(x,y,f_{L^\star,\mathrm{P},\lambda}(x))\,d\mathrm{P}(x,y) \\ &\quad\leq \kappa_2\,\|g\|_{\mathcal{H}}\,\|k\|^3_\infty\,. \end{aligned}$$

Let us now show that $\nabla^F_2 G(0,f_{L^\star,\mathrm{P},\lambda}) = 2\lambda\mathrm{id}_{\mathcal{H}} + A$ is injective. To this end, let us choose $g \in \mathcal{H}\backslash\{0\}$. Then we find

$$\begin{aligned} &\langle(2\lambda\mathrm{id}_{\mathcal{H}}+A)g,(2\lambda\mathrm{id}_{\mathcal{H}}+A)g\rangle_{\mathcal{H}} \\ &\quad> 4\lambda\,\langle g,Ag\rangle_{\mathcal{H}} \\ &\quad= 4\lambda\,\mathbb{E}_{\mathrm{P}}\nabla^F_{3,3}L^\star(X,Y,f_{L^\star,\mathrm{P},\lambda}(X))g^2(X) \geq 0, \end{aligned}$$

which shows the injectivity. The implicit function Theorem 24 for Fréchet-derivatives guarantees that $\varepsilon \mapsto f_\varepsilon$ is differentiable on $(-\delta,\delta)$ if $\delta > 0$ is small enough. Furthermore, (47) and (48) yield, for all $z = (x,y) \in \mathcal{X}\times\mathcal{Y}$, that

$$\begin{aligned} \mathrm{IF}(z;T,\mathrm{P}) &= \nabla^F f_\varepsilon(0) \\ &= -S^{-1}\circ\nabla^F_1 G(0,f_{L^\star,\mathrm{P},\lambda}) \\ &= S^{-1}\big(\mathbb{E}_{\mathrm{P}}\big(\nabla^F_3 L^\star(X,Y,f_{L^\star,\mathrm{P},\lambda}(X))\Phi(X)\big)\big) \\ &\quad - \nabla^F_3 L^\star(x,y,f_{L^\star,\mathrm{P},\lambda}(x))S^{-1}\Phi(x), \end{aligned}$$

which yields the existence of the influence function and (29). The boundedness follows from (28) and (29). □

Proof of Theorem 12. Theorem 7 guarantees the existence of a bounded measurable function $h : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$ such that $\|h\|_\infty \leq |L|_1$ and

$$\left\|f_{L^\star,\mathrm{P},\lambda} - f_{L^\star,(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q},\lambda}\right\|_{\mathcal{H}} \leq \frac{\varepsilon}{\lambda}\left\|\mathbb{E}_{\mathrm{P}}h\Phi - \mathbb{E}_{\mathrm{Q}}h\Phi\right\|_{\mathcal{H}}.$$

From (24) we get

$$\begin{aligned} &\left\|f_{L^\star,\mathrm{P},\lambda} - f_{L^\star,(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q},\lambda}\right\|_{\mathcal{H}} \\ &\quad\leq \frac{\varepsilon}{\lambda}\left\|\mathbb{E}_{\mathrm{P}}h\Phi - \mathbb{E}_{\mathrm{Q}}h\Phi\right\|_{\mathcal{H}} \leq \frac{1}{\lambda}\|h\|_\infty\|k\|_\infty\|\mathrm{P}-\mathrm{Q}\|_{\mathcal{M}}\,\varepsilon, \end{aligned}$$

which gives the assertion. □

Proof of Theorem 13. By definition of $L^\star$ it follows from (20) that $\nabla^B_3 L(x,y,t) = \nabla^B_3 L^\star(x,y,t)$. Therefore,

$$\begin{aligned} G(\varepsilon,f) &:= 2\lambda f + \mathbb{E}_{(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q}}\nabla^B_3 L^\star(X,Y,f(X))\Phi(X) \\ &= 2\lambda f + \mathbb{E}_{(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q}}\nabla^B_3 L(X,Y,f(X))\Phi(X). \end{aligned}$$

Hence $G(\varepsilon,f)$ is the same as in Theorem 2 in Christmann and Van Messem [2008]. All conditions of Theorem 2 are fulfilled since we assumed that $\nabla^B_2 G(0,f_{L^\star,\mathrm{P},\lambda})$ is strong. Hence the proof of Theorem 13 is identical to the proof of Theorem 2 in Christmann and Van Messem [2008], which is based on an implicit function theorem for B-derivatives [Robinson, 1991], and the assertion follows. □

Received 1 April 2009

REFERENCES

Akerkar, R. (1999). *Nonlinear Functional Analysis*. Narosa Publishing House, New Delhi. MR1684584

Bauer, H. (2001). *Measure and Integration Theory*. De Gruyter, Berlin. MR1897176

Cheney, W. (2001). *Analysis for Applied Mathematics*. Springer, New York. MR1838468

Christmann, A. and Steinwart, I. (2004). On robust properties of convex risk minimization methods for pattern recognition. *Journal of Machine Learning Research* **5** 1007–1034. MR2248007

Christmann, A. and Steinwart, I. (2007). Consistency and robustness of kernel based regression in convex minimization. *Bernoulli* **13** 799–819. MR2348751

Christmann, A. and Steinwart, I. (2008). Consistency of kernel based quantile regression. *Appl. Stoch. Models Bus. Ind.* **24** 171–183. MR2406112

Christmann, A. and Van Messem, A. (2008). Bouligand Derivatives and Robustness of Support Vector Machines for Regression. *Journal of Machine Learning Research* **9** 623–644. MR2417258

Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge.

Diestel, J. and Uhl, J. J. (1977). *Vector Measures*. American Mathematical Society, Providence, RI. MR0453964

Dudley, R. M. (2002). *Real Analysis and Probability*. Cambridge University Press. MR1932358

Ekeland, I. and Turnbull, T. (1983). *Infinite-dimensional Optimization and Convexity*. Chicago Lectures in Mathematics. The University of Chicago Press. MR0769469

Hampel, F. R. (1968). Contributions to the theory of robust estimation. Unpublished Ph.D. thesis, Dept. of Statistics, University of California, Berkeley.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association* **69** 383–393. MR0362657

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the 5^{th} Berkeley Symposium* **1** 221–233. MR0216620

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press, New York. MR2268657

Phelps, R. R. (1993). *Convex Functions, Monotone Operators and Differentiability*. Lecture Notes in Math. 1364. Springer, Berlin. MR1238715

Robinson, S. M. (1987). Local structure of feasible sets in nonlinear programming, Part III: Stability and sensitivity. *Mathematical Programming Study* **30** 45–66. MR0874131

Robinson, S. M. (1991). An implicit-function theorem for a class of nonsmooth functions. *Mathematics of Operations Research* **16** 292–309. MR1106803

Rockafellar, R. T. (1976). Integral functionals, normal integrands and measurable selections. In *Nonlinear Operators and the Calculus of Variations*, volume 543 of *Lecture Notes in Mathematics*, pages 157–207. MR0512209

Rockafellar, R. T. and Wets, R. J. B. (1998). *Variational Analysis*. Springer, Berlin. MR1491362

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

Steinwart, I. (2001). On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research* **2** 67–93. MR1883281

Steinwart, I. and Christmann, A. (2008a). How SVMs can estimate quantiles and the median. In J. C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*. MIT Press, Cambridge, Massachusetts.

Steinwart, I. and Christmann, A. (2008b). *Support Vector Machines*. Springer, New York. MR2450103

Takeuchi, I., Le, Q. V., Sears, T. D., and Smola, A. J. (2006). Nonparametric quantile estimation. *Journal of Machine Learning Research* **7** 1231–1264. MR2274404

Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley & Sons, New York. MR1641250

Werner, D. (2002). *Funktionalanalysis, 4^{th} ed.* Springer, Berlin. MR1787146

Yurinsky, V. (1995). *Sums and Gaussian Vectors*. Lecture Notes in Math. 1617. Springer, Berlin. MR1442713

Andreas Christmann
University of Bayreuth
Department of Mathematics
D-95440 Bayreuth
Germany
E-mail address: andreas.christmann@uni-bayreuth.de

Arnout Van Messem
Vrije Universiteit Brussel
Department of Mathematics
B-1050 Brussels
Belgium
E-mail address: avmessem@vub.ac.be

Ingo Steinwart
Los Alamos National Laboratory
Information Sciences Group (CCS-3)
MS B256
Los Alamos, NM 87545
USA
E-mail address: ingo@lanl.gov