Journal of Machine Learning Research 25 (2024) 1-33 Submitted 3/23; Revised 1/24; Published 7/24

On the Connection between Lp- and Risk Consistency and its Implications on Regularized Kernel Methods

Hannes Köhler HANNES.KOEHLER@UNI-BAYREUTH.DE
Department of Mathematics
University of Bayreuth
95440 Bayreuth, Germany

Editor: Florence d'Alché-Buc

Abstract

As a predictor's quality is often assessed by means of its risk, it is natural to regard risk consistency as a desirable property of learning methods, and many such methods have indeed been shown to be risk consistent. The first aim of this paper is to establish the close connection between risk consistency and L_p-consistency for a considerably wider class of loss functions than has been done before. The attempt to transfer this connection to shifted loss functions surprisingly reveals that this shift does not reduce the assumptions needed on the underlying probability measure to the same extent as it does for many other results. The results are applied to regularized kernel methods such as support vector machines.
Keywords: machine learning, consistency, regression, kernel methods, support vector machines

1. Introduction

The goal of non-parametric statistical machine learning is to predict an output random variable Y based on an input random variable X with (almost) no prior knowledge about the distribution P of (X, Y) on some space $\mathcal{X} \times \mathcal{Y}$, all information about P typically stemming from a data set $D_n := ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$ consisting of independent and identically distributed (i.i.d.) observations sampled from P. More specifically, one aims at finding a measurable function $f\colon \mathcal{X} \to \mathcal{Y}$ which captures certain characteristics of the conditional distribution $\mathrm{P}(\cdot \,|\, X)$, like its conditional mean function or conditional quantile function.

Such learning tasks can often be formalized by aiming at finding a measurable function that minimizes the *L-risk* (or just *risk*)

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

for a suitable *loss function*, which is a measurable function $L\colon \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$. Here, $L(x, y, f(x))$ quantifies the cost of the prediction $f(x)$ if the observed true output belonging to x is y. Hence, the choice of L controls how different deviations between y and $f(x)$ are penalized and specifies the exact goal of the prediction, and the risk assesses the quality of the whole predictor f with respect to the whole distribution P. For example, the two aforementioned goals of finding the conditional means (least squares regression) and conditional quantiles (quantile regression) can be approached by using the least squares loss and the pinball loss respectively, as it is known that the according risks are minimized by the respective target functions one aims to estimate.

To this end, we define the *Bayes risk* $\mathcal{R}^*_{L,\mathrm{P}}$ as usual as the smallest possible risk, that is,

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

and call a measurable function $f^*_{L,\mathrm{P}}$ achieving $\mathcal{R}_{L,\mathrm{P}}(f^*_{L,\mathrm{P}}) = \mathcal{R}^*_{L,\mathrm{P}}$ a *Bayes function.* Further assume that a learning method yields the predictor f_n based on the data set D_n, $n \in \mathbb{N}$.

Because of the risk assessing a predictor's quality, a desirable property for the learning method is *risk consistency*, i.e. that

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

in a suitable sense (which is in probability or almost surely in most results). As this is a very natural type of consistency to consider and the classic one aimed at in statistical learning theory (cf. Vapnik, 1995), results on risk consistency exist for many such learning methods, see for example Steinwart (2005) (regularized kernel methods for classification), Zhang and Yu (2005) (boosting), Christmann and Steinwart (2007) (regularized kernel methods for regression; see also Section 4), Biau et al. (2008) (averaging classifiers such as random forests), Lin et al. (2022) (deep convolutional neural networks).

We are however also interested in taking a look at a different type of consistency, namely L_p-*consistency*, i.e. that

$$\left|\left|f_n - f^*_{L,\mathrm{P}}\right|\right|_{L_p(\mathrm{P}_X)} \to 0\,, \qquad n \to \infty\,,$$

in a suitable sense for some $p \in [1,\infty)$. This type of consistency is of interest for us because it compares the estimator f_n and the Bayes function $f^*_{L,\mathrm{P}}$ directly, weighted only based on the marginal distribution P_X, instead of additionally depending on the loss function and the conditional distribution of Y. Additionally, it is a quite powerful type of consistency that for example also directly implies weak consistency.

We show in Section 3 that L_p- and risk consistency are actually equivalent under rather mild assumptions. Note that the results from that section are non-probabilistic and therefore yield a very strong connection between the two examined types of convergence, stating that $\mathcal{R}_{L,\mathrm{P}}(f_n) \to \mathcal{R}^*_{L,\mathrm{P}}$ holds true if and only if $||f_n - f^*_{L,\mathrm{P}}||_{L_p(\mathrm{P}_X)} \to 0$ (under the assumptions required by the results) even in an analytical sense. In this equivalence, the more surprising direction certainly is risk consistency implying L_p-consistency as the latter tackles the generally more demanding task of system identification instead of only system imitation, as it is described by Cherkassky and Mulier (2007, Section 2.1.1), see also Györfi et al. (2002, Section 1.4) for the classification case. Whereas this implication had already been established for certain special loss functions (which we briefly recap in Section 3.1), we considerably generalize it to a large class of loss functions including those as special cases. Additionally, we examine whether it is possible to transfer these results to risks that are based on shifted loss functions—which are useful for working with heavy-tailed distributions—and stumble upon some difficulties when trying to do this in all generality, which is somewhat surprising considering that many other results can be transferred to shifted loss functions quite seamlessly. Before successfully transferring our results by imposing some assumptions on the underlying distribution, we therefore also derive some interesting negative results. Lastly, in Section 4, the derived connection is applied to regularized kernel methods, in which the predictors are defined as minimizers of regularized risks. Because of this definition, it is natural to examine their risk consistency and this has already been well investigated in the past, but there did not exist any general results on their L_p-consistency so far. Here, these two types of consistency are for the most part examined in the sense of convergence in probability as this is the one that we are able to derive under the mild assumptions imposed in that section. It should however be noted that the results from Section 3 on the connection between L_p- and risk consistency can also be applied to other learning methods for which one of the two types of consistency is known, and that their non-probabilistic nature means that they can also be used to even obtain either of the two types of consistency in the almost sure sense if the other one is known to hold true in that sense as well, see also Remark 16 and Corollary 19.

We wish to emphasize that our goal is not to derive learning rates for any learning method (like for example for the regularized kernel methods from Section 4). Instead, we aim at deriving results on consistency under minimal assumptions on the underlying probability distribution—much weaker assumptions than those needed for deriving learning rates—and such that the results are applicable to general learning methods in a general setting.

2. Prerequisites

Before presenting our results, we first need to state some additional prerequisites: As mentioned in the introduction, we aim at estimating certain properties of the unknown conditional distribution $\mathrm{P}(\cdot\,|\,X)$ such as the conditional mean or conditional quantiles. This conditional distribution $\mathrm{P}(\cdot\,|\,X)$ uniquely exists, and P can therefore be split into a marginal distribution P_X on $\mathcal{X}$ and this conditional distribution, whenever $\mathcal{Y}$ is a Polish space (cf. Dudley, 2004, Theorems 10.2.1 and 10.2.2), for example if $\mathcal{Y} \subseteq \mathbb{R}$ closed (cf. Bauer, 2001, p. 157). Hence, by choosing $\mathcal{Y}$ in such a way, we are guaranteed to always be able to perform this factorization of P, which leads us to one part of the following standard and rather general assumption which we assume to hold true throughout this paper.

Assumption 1 *Let $\mathcal{X}$ be a complete separable metric space and let $\mathcal{Y} \subseteq \mathbb{R}$ be closed. Let $\mathcal{X}$ and $\mathcal{Y}$ be equipped with their respective Borel σ-algebras $\mathcal{B}_\mathcal{X}$ and $\mathcal{B}_\mathcal{Y}$. Let $\mathrm{P} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$, where $\mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ denotes the set of all Borel probability measures on the measurable space $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}_{\mathcal{X}\times\mathcal{Y}})$.*

We are mainly interested in continuous and in convex loss functions, by which we mean continuity respectively convexity of L in its third argument. Furthermore, the loss functions will be assumed to additionally be distance-based. Distance-based losses are a special type of loss functions which are typically used in regression tasks, and which are defined in the following way:

Definition 2 *A loss function $L\colon \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ is called distance-based if there exists a representing function $\psi\colon \mathbb{R} \to [0, \infty)$ satisfying $\psi(0) = 0$ and $L(x, y, t) = \psi(y - t)$ for all*

$(x, y, t) \in \mathcal{X} \times \mathcal{Y} \times \mathbb{R}$. *If* $\psi(r) = \psi(-r)$ *for all* $r \in \mathbb{R}$, *then* L *is called symmetric.*
Let $p \in (0, \infty)$. *A distance-based loss* $L \colon \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ *with representing function* ψ *is of*

(i) upper growth type p *if there is a constant* $c > 0$ *such that* $\psi(r) \leq c\,(|r|^p + 1)$ *for all* $r \in \mathbb{R}$.

(ii) lower growth type p *if there is a constant* $c > 0$ *such that* $\psi(r) \geq c\,|r|^p - 1$ *for all* $r \in \mathbb{R}$.

(iii) growth type p *if* L *is of both upper and lower growth type* p.

Since the first argument does not matter in distance-based loss functions, we often ignore it and write $L \colon \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ and $L(y, t)$ instead.

Distance-based losses are typically used in regression tasks, but some of them, like the least squares loss, are also popular choices for classification tasks, see for example Györfi et al. (2002, Section 1.4). As an example of a distance-based loss, the mentioned least squares loss is of growth type 2 whereas many other common loss functions for regression tasks, like the pinball loss, Huber loss or ε-insensitive loss, are of growth type 1. We will later see that this sometimes leads to slightly more restrictive conditions regarding P when using the least squares loss.

More specifically, if p denotes the growth type of the loss function that is applied, some results from the subsequent sections require that the *averaged* p*-th moment* of P is finite. This averaged p-th moment is defined as

$$|\mathrm{P}|_p := \left(\int_{\mathcal{X} \times \mathcal{Y}} |y|^p \, \mathrm{dP}(x, y) \right)^{1/p} = \left(\int_{\mathcal{X}} \int_{\mathcal{Y}} |y|^p \, \mathrm{dP}(y \,|\, x) \, \mathrm{dP}_X(x) \right)^{1/p} .$$

3. Connection between Lp- and Risk Consistency

In Section 3.1, we show that L_p- and risk consistency are equivalent under certain conditions. Section 3.2 contains the rather surprising result that some of these results can not be transferred to risks that are based on shifted loss functions in the generality we would have hoped for, but we also introduce some additional conditions under which it is possible to transfer the results after all.

Remark 3 *We will often write "the Bayes function", implying there exists exactly one such measurable function minimizing* $\mathcal{R}_{L,\mathrm{P}}$. *This does not always hold true and is not necessary for risk consistency (neither existence nor uniqueness). We however assume that the Bayes function indeed exists and is* P_X*-almost-surely (a.s.) unique whenever we investigate the difference between some predictor and the Bayes function directly (e.g. in the results on* L_p*-consistency) instead of the difference between the according risks.*

3.1 Connection between Lp- and risk consistency for regular loss functions

So far, there are no general results on L_p-consistency following from risk consistency, but only results regarding special loss functions: For the least squares loss, it has been known for many years that a function's excess risk, i.e. the difference between its risk and the Bayes risk, corresponds to the squared $L_2(\mathrm{P}_X)$-norm of its deviation from the Bayes function, and risk consistency therefore implies L_2-consistency, cf. Cucker and Smale (2001, Proposition 1) or Cherkassky and Mulier (2007, pp. 26–28). Recently, this L_2-difference between a function and the Bayes function has also been bounded by the excess risk—by means of so-called comparison or self-calibration inequalities—in case of the asymmetric least squares loss by Farooq and Steinwart (2019) and in case of more general strongly convex loss functions under additional assumptions by Sheng et al. (2020). Additionally, Hable and Christmann (2014) showed that L_1-consistency follows from risk consistency in case of the pinball loss, and Steinwart and Christmann (2011); Xiang et al. (2012) derived self-calibration inequalities for this loss under additional assumptions. Tong and Ng (2019) did so for the ε-insensitive loss.

The following theorem generalizes the aforementioned special cases to general convex, distance-based loss functions:

Theorem 4 *Let $L\colon \mathcal{Y} \times \mathbb{R} \to [0,\infty)$ be a convex, distance-based loss function of lower growth type $p \in [1,\infty)$. Assume that $f^*_{L,\mathrm{P}}$ is P_X-a.s. unique, $f^*_{L,\mathrm{P}} \in L_p(\mathrm{P}_X)$ and $\mathcal{R}^*_{L,\mathrm{P}} < \infty$. Then, for every sequence $(f_n)_{n\in\mathbb{N}} \subseteq L_p(\mathrm{P}_X)$, we have*

$$\lim_{n\to\infty} \mathcal{R}_{L,\mathrm{P}}(f_n) = \mathcal{R}^*_{L,\mathrm{P}} \qquad \Rightarrow \qquad \lim_{n\to\infty} ||f_n - f^*_{L,\mathrm{P}}||_{L_p(\mathrm{P}_X)} = 0\,.$$

Remark 5 *If L is of growth type p instead of only being of lower growth type p, the conditions $f^*_{L,\mathrm{P}} \in L_p(\mathrm{P}_X)$ and $\mathcal{R}^*_{L,\mathrm{P}} < \infty$ in Theorem 4 can also be replaced by the perhaps more intuitive and in this case equivalent moment condition $|\mathrm{P}|_p < \infty$. This equivalence can easily be obtained from parts (i) and (iii) of Steinwart and Christmann (2008, Lemma 2.38) by noting that $\mathcal{R}^*_{L,\mathrm{P}} \leq \mathcal{R}_{L,\mathrm{P}}(0)$, with 0 denoting the zero function, always holds true by definition of the Bayes risk.*

Notably, Theorem 4 strengthens Steinwart and Christmann (2008, Corollary 3.62), which stated that risk consistency implies weak consistency.

As mentioned in the introduction, the opposite direction—risk consistency following from L_p-consistency—is generally the easier one. We formally state this implication in the subsequent Theorem 6. Hence, this theorem can be seen as the counterpart of Theorem 4, even though the conditions of the two theorems differ in some details. Notably, the function f^*, which the sequence is converging to, does not necessarily need to be the Bayes function $f^*_{L,\mathrm{P}}$ here:

Theorem 6 *Let $L\colon \mathcal{Y} \times \mathbb{R} \to [0,\infty)$ be a continuous, distance-based loss function of upper growth type $p \in [1,\infty)$. Assume that $|\mathrm{P}|_p < \infty$. Then, for every sequence $(f_n)_{n\in\mathbb{N}} \subseteq L_p(\mathrm{P}_X)$ and every function $f^* \in L_p(\mathrm{P}_X)$, we have*

$$\lim_{n\to\infty} ||f_n - f^*||_{L_p(\mathrm{P}_X)} = 0 \qquad \Rightarrow \qquad \lim_{n\to\infty} \mathcal{R}_{L,\mathrm{P}}(f_n) = \mathcal{R}_{L,\mathrm{P}}(f^*)\,.$$

3.2 Connection between Lp- and risk consistency for shifted loss functions

When looking at Theorem 4, it is obvious that the assumptions $f^*_{L,\mathrm{P}} \in L_p(\mathrm{P}_X)$ and $\mathcal{R}^*_{L,\mathrm{P}} < \infty$ are indeed necessary for the theorem's conclusion and that one cannot hope to derive L_p-

from risk consistency without them. Because these assumptions are equivalent to $|\mathrm{P}|_p < \infty$ if L is of growth type p (cf. Remark 5), this however excludes heavy-tailed distributions such as the Cauchy distribution—even for $p = 1$. Analogously, Theorem 6 also requires $|\mathrm{P}|_p < \infty$ and can therefore not be applied to such heavy-tailed distributions.

To circumvent this problem, we now try to transfer the results from Section 3.1 to *shifted loss functions*, which have been applied in robust statistics for a long time, see for example Huber (1967) or Huber and Ronchetti (2009, Chapter 3), and which can be defined in a very easy way: Given a loss function $L\colon \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$, the associated shifted loss function is

$$\begin{aligned} L^\star\colon\ & \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to \mathbb{R}\,, \\ & (x, y, t) \mapsto L(x, y, t) - L(x, y, 0)\,, \end{aligned}$$

which can be used to estimate the same quantities as the original loss function since the shift is fixed independently of t. Risks can be defined in the same way as for regular loss functions.

Remark 7 *By Steinwart and Christmann (2008, Lemma 2.34), a convex and distance-based loss function of upper growth type 1 is always Lipschitz continuous. We call a loss function L Lipschitz continuous if it is Lipschitz continuous with respect to its last argument, that is, if*

$$|L(x, y, t) - L(x, y, t')| \le |L|_1 \cdot |t - t'| \qquad \forall\, (x, y) \in \mathcal{X} \times \mathcal{Y}\,,\ t, t' \in \mathbb{R}\,,$$

for some constant $|L|_1 \ge 0$ which is called the Lipschitz constant of L.

With Remark 7 in mind, the risk with respect to the shifted version of a convex and distance-based loss function of upper growth type 1 can be bounded by

$$|\mathcal{R}_{L^\star,\mathrm{P}}(f)| \le \int_{\mathcal{X}\times\mathcal{Y}} |L(y, f(x)) - L(y, 0)|\,\mathrm{dP}(x, y) \le |L|_1 \int_{\mathcal{X}} |f(x)|\,\mathrm{dP}_X(x)\,. \tag{1}$$

Hence, even if $|\mathrm{P}|_1 = \infty$, this risk is finite for all $f \in L_1(\mathrm{P}_X)$. Using the shifted loss therefore seems like a promising approach for extending the applicability of the results from Section 3.1 to heavy-tailed distributions and getting rid of the moment condition $|\mathrm{P}|_1 < \infty$ in the case of having a convex loss function of growth type 1. Indeed, Christmann et al. (2009) showed that the moment condition can in this case be eliminated from many results regarding regular loss functions by transferring them to shifted loss functions.

When looking at the proof of Theorem 4, it is however easy to see that (9) does not hold true for shifted loss functions and the proof can thus not be transferred to the situation of this section. The following negative result shows that this is indeed not a failing of the specific proof we used, but that L_1-consistency does, somewhat surprisingly, actually not follow from $L^\star$-risk consistency in the generality one would have hoped for:

Proposition 8 *Let $\mathcal{Y} = \mathbb{R}$. Let $L\colon \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a convex, distance-based and symmetric loss function of growth type 1, and let $L^\star$ be its shifted version. Then, even if $f^*_{L^\star,\mathrm{P}}$ is P_X-a.s. unique with $f^*_{L^\star,\mathrm{P}} \in L_1(\mathrm{P}_X)$, a sequence $(f_n)_{n\in\mathbb{N}} \subseteq L_1(\mathrm{P}_X)$ of functions satisfying*

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

*does in general **not** imply*

$$\lim_{n\to\infty} \left|\left|f_n - f^*_{L^\star,\mathrm{P}}\right|\right|_{L_1(\mathrm{P}_X)} = 0$$

without any additional assumptions besides Assumption 1 being imposed.

Note that in the situation of Proposition 8, risk consistency does also not imply L_p-consistency for any $p > 1$ since L_p-consistency for $p > 1$ would imply L_1-consistency.

We now take a special look at the *τ-pinball loss* (or just *pinball loss*)

$$\begin{aligned} L_{\tau\text{-pin}} \colon\ & \mathcal{Y} \times \mathbb{R} \to [0,\infty)\,, \\ & (y,t) \mapsto \begin{cases} (1-\tau)\cdot(t-y) & \text{, if } y < t\,, \\ \tau \cdot (y-t) & \text{, if } y \geq t\,, \end{cases} \end{aligned} \tag{2}$$

$\tau \in (0,1)$, which is convex and distance-based with growth type 1, but not symmetric for $\tau \neq 0.5$. As mentioned in the introduction, the pinball loss can be used for quantile regression, i.e. for estimating the conditional quantiles

$$\begin{aligned} F^*_{\tau,\mathrm{P}} \colon\ & \mathcal{X} \to 2^{\mathbb{R}}\,, \\ & x \mapsto \{t^* \mid \mathrm{P}((-\infty,t^*]|x) \geq \tau \text{ and } \mathrm{P}([t^*,\infty)|x) \geq 1-\tau\}\,, \end{aligned}$$

see also Koenker and Bassett (1978); Koenker and Hallock (2001); Takeuchi et al. (2006); Steinwart and Christmann (2011).

If one assumes these conditional quantiles $F^*_{\tau,\mathrm{P}}(x)$ to P_X-a.s. be singletons, it is possible to denote them by the P_X-a.s. unique quantile function $f^*_{\tau,\mathrm{P}} \colon \mathcal{X} \to \mathbb{R}$ defined by $\{f^*_{\tau,\mathrm{P}}(x)\} = F^*_{\tau,\mathrm{P}}(x)$ for all $x \in \mathcal{X}$. Recall that this $f^*_{\tau,\mathrm{P}}$ is the up to P_X-zero sets only measurable function satisfying

$$\mathcal{R}_{L_{\tau\text{-pin}},\mathrm{P}}(f^*_{\tau,\mathrm{P}}) = \mathcal{R}^*_{L_{\tau\text{-pin}},\mathrm{P}} \tag{3}$$

if $\mathcal{R}^*_{L_{\tau\text{-pin}},\mathrm{P}}$ is finite, and similarly, that $f^*_{\tau,\mathrm{P}}$ satisfies

$$\mathcal{R}_{L^\star_{\tau\text{-pin}},\mathrm{P}}(f^*_{\tau,\mathrm{P}}) = \mathcal{R}^*_{L^\star_{\tau\text{-pin}},\mathrm{P}} \tag{4}$$

and is the up to P_X-zero sets only measurable function doing so if $\mathcal{R}^*_{L^\star_{\tau\text{-pin}},\mathrm{P}}$ is finite. This ties our assumption of the conditional quantiles P_X-a.s. being singletons to Remark 3 about the required P_X-a.s. uniqueness of the Bayes function and yields $f^*_{L^\star_{\tau\text{-pin}},\mathrm{P}} \equiv f^*_{\tau,\mathrm{P}}$ P_X-a.s.

As non-symmetric loss functions are not covered by Proposition 8 and as the pinball loss is the probably most popular among these, we specifically investigate this loss function's behavior and obtain the following analogous result to Proposition 8:

Proposition 9 *Let $\mathcal{Y} = \mathbb{R}$. Let $\tau \in (0,1)$ and let $L^\star_{\tau\text{-pin}}$ be the shifted version of the τ-pinball loss.*[1] *Then, even if $f^*_{\tau,\mathrm{P}}$ is P_X-a.s. unique with $f^*_{\tau,\mathrm{P}} \in L_1(\mathrm{P}_X)$, a sequence $(f_n)_{n\in\mathbb{N}} \subseteq$*

1. It can easily be seen that this shifted pinball loss function is, for $\tau \in (0,1)$,

$$\begin{aligned} L^\star_{\tau\text{-pin}} \colon\ & \mathcal{Y} \times \mathbb{R} \to \mathbb{R} \\ & (y,t) \mapsto L_{\tau\text{-pin}}(y,t) - L_{\tau\text{-pin}}(y,0) = \begin{cases} (1-\tau)\cdot t & \text{, if } y < \min\{0,t\}\,, \\ (1-\tau)\cdot t - y & \text{, if } 0 \leq y < t\,, \\ y - \tau \cdot t & \text{, if } t \leq y < 0\,, \\ -\tau \cdot t & \text{, if } y \geq \max\{0,t\}\,. \end{cases} \end{aligned}$$

$L_1(\mathrm{P}_X)$ *of functions satisfying*

$$\lim_{n\to\infty} \mathcal{R}_{L^\star_{\tau\text{-}pin},\mathrm{P}}(f_n) = \mathcal{R}^*_{L^\star_{\tau\text{-}pin},\mathrm{P}}$$

does in general ***not*** *imply*

$$\lim_{n\to\infty} \left|\left|f_n - f^*_{\tau,\mathrm{P}}\right|\right|_{L_1(\mathrm{P}_X)} = 0$$

without any additional assumptions besides Assumption 1 being imposed.

As the preceding results allow for arbitrary sequences of functions in $L_1(\mathrm{P}_X)$, we might still hope to deduce L_1-consistency following from $L^\star$-risk consistency by restricting ourselves to smaller function spaces with more structure like Sobolev spaces. However, the subsequent corollary shows that Proposition 8 and Proposition 9 can even be strengthened to sequences of functions from Sobolev spaces. Here, we assume that $\mathcal{X} \subseteq \mathbb{R}^d$ open for some $d \in \mathbb{N}$, and we denote by $W^{m,q}(\mathcal{X})$ the Sobolev space consisting of all functions from $L_q(\mathcal{X})$ whose weak derivatives up to order m are also in $L_q(\mathcal{X})$, cf. Adams and Fournier (2003, Definition 3.2). Here, as usual, $L_q(\mathcal{X})$ denotes the L_q-space with respect to the Lebesgue measure on $\mathcal{X}$.

Corollary 10 *Let $d \in \mathbb{N}$, $\mathcal{X} \subseteq \mathbb{R}^d$ open, and $\mathcal{Y} = \mathbb{R}$. Let $L\colon \mathcal{Y} \times \mathbb{R} \to [0,\infty)$ be a convex, distance-based and symmetric loss function of growth type 1, or the τ-pinball loss for some $\tau \in (0,1)$. Let $L^\star$ be its shifted version. Let $m \in \mathbb{N}$ and $1 \le q \le \infty$. Then, even if $f^*_{L^\star,\mathrm{P}}$ is P_X-a.s. unique with $f^*_{L^\star,\mathrm{P}} \in L_1(\mathrm{P}_X)$, a sequence $(f_n)_{n\in\mathbb{N}} \subseteq W^{m,q}(\mathcal{X}) \cap L_1(\mathrm{P}_X)$ of functions satisfying*

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

does in general ***not*** *imply*

$$\lim_{n\to\infty} \left|\left|f_n - f^*_{L^\star,\mathrm{P}}\right|\right|_{L_1(\mathrm{P}_X)} = 0$$

without any additional assumptions besides Assumption 1 being imposed.

The preceding results show that it is not possible to get rid of the moment condition from Theorem 4 (cf. Remark 5) just by transferring it to shifted loss functions. It might, however, still be possible to circumvent this moment condition by instead imposing some different and less restrictive conditions. For the pinball loss from (2), i.e. for doing quantile regression, we are indeed able to derive such an alternative and in many cases less restrictive condition regarding P. To be more specific, the conditional distribution $\mathrm{P}(\cdot\,|\,X)$ is, in some sense, not allowed to be too heteroscedastic and it has to be continuous in the conditional quantiles $f^*_{\tau,\mathrm{P}}(x)$, $x \in \mathcal{X}$:

Theorem 11 *Let $\tau \in (0,1)$ and $L^\star_{\tau\text{-}pin}$ be the shifted version of the τ-pinball loss. Assume that $f^*_{\tau,\mathrm{P}}$ is P_X-a.s. unique, $f^*_{\tau,\mathrm{P}} \in L_1(\mathrm{P}_X)$, and P additionally satisfies at least one of the following conditions:*

(i) $|\mathrm{P}|_1 < \infty$.

(ii) *There exist* $c_1, c_2 > 0$ *such that*

$$\mathrm{P}\Big((f^*_{\tau,\mathrm{P}}(X) - c_1, f^*_{\tau,\mathrm{P}}(X)) \,\Big|\, X \Big) \geq c_2 \quad \textit{and} \quad \mathrm{P}\Big((f^*_{\tau,\mathrm{P}}(X), f^*_{\tau,\mathrm{P}}(X) + c_1) \,\Big|\, X \Big) \geq c_2 \tag{5}$$

P_X*-a.s., as well as*

$$\mathrm{P}(f^*_{\tau,\mathrm{P}}(X) \,|\, X) = 0 \tag{6}$$

P_X*-a.s.*

Then, for every sequence $(f_n)_{n\in\mathbb{N}} \subseteq L_1(\mathrm{P}_X)$*, we have*

$$\lim_{n\to\infty} \mathcal{R}_{L^\star_{\tau\text{-}pin},\mathrm{P}}(f_n) = \mathcal{R}^*_{L^\star_{\tau\text{-}pin},\mathrm{P}} \qquad \Rightarrow \qquad \lim_{n\to\infty} ||f_n - f^*_{\tau,\mathrm{P}}||_{L_1(\mathrm{P}_X)} = 0\,.$$

Even though it was not possible to get rid of the moment condition (i) without imposing the new condition (ii), this still substantially expands the theorem's applicability since there are many cases in which (ii) (whose first part is visualized in Figure 1) is satisfied even though (i) is not:

Example 1 *Assume that* $\tau \in (0,1)$ *and that we have an underlying homoscedastic regression model like*

$$Y = f(X) + \varepsilon\,,$$

where $f\colon \mathcal{X} \to \mathcal{Y}$ *is an arbitrary measurable function and* ε *is a continuous random variable whose distribution does not depend on the value of* X*. Whenever* ε *has a unique* τ*-quantile* $q_\tau \in \mathbb{R}$*, (ii) from Theorem 11 holds true with* $f^*_{\tau,\mathrm{P}} = f + q_\tau$*. For example,* ε *can follow a Cauchy distribution with location and scale parameters which are fixed independently of the value of* X*. In this case, the moment condition (i) does not hold true, but Theorem 11 does still yield* L_1*-consistency following from risk consistency.*

Example 2 *The independence of* ε *from* X *in Example 1 is not even strictly necessary. Assume the more general heteroscedastic model*

$$Y = f(X) + \varepsilon_X\,,$$

where the distribution of ε_X *is now allowed to depend on the value* x *of* X*. If, for example, there exist* $C > 0$ *and* $c_1 > 0$ *such that* ε_x *has a unique* τ*-quantile* $q_{x,\tau} \in \mathbb{R}$ *and Lebesgue density greater than* C *on* $(q_{x,\tau} - c_1, q_{x,\tau} + c_1)$ *for* P_X*-almost all* $x \in \mathcal{X}$*, condition (ii) from Theorem 11 is still satisfied.*

For example, this situation is on hand if $\mathcal{X} = \mathbb{R}^d$ *for some* $d \in \mathbb{N}$*,* $\mathcal{Y} = \mathbb{R}$*, and* ε_x *follows a Cauchy distribution with location parameter* $\cos(||x||_2)$ *and scale parameter* $2 + \sin(||x||_2)$ *for all* $x \in \mathcal{X}$*. More generally, the same also holds true for different choices of location and scale parameters, as long as they are bounded from above and from below (in the case of the scale parameter we mean bounded away from zero by bounded from below).*

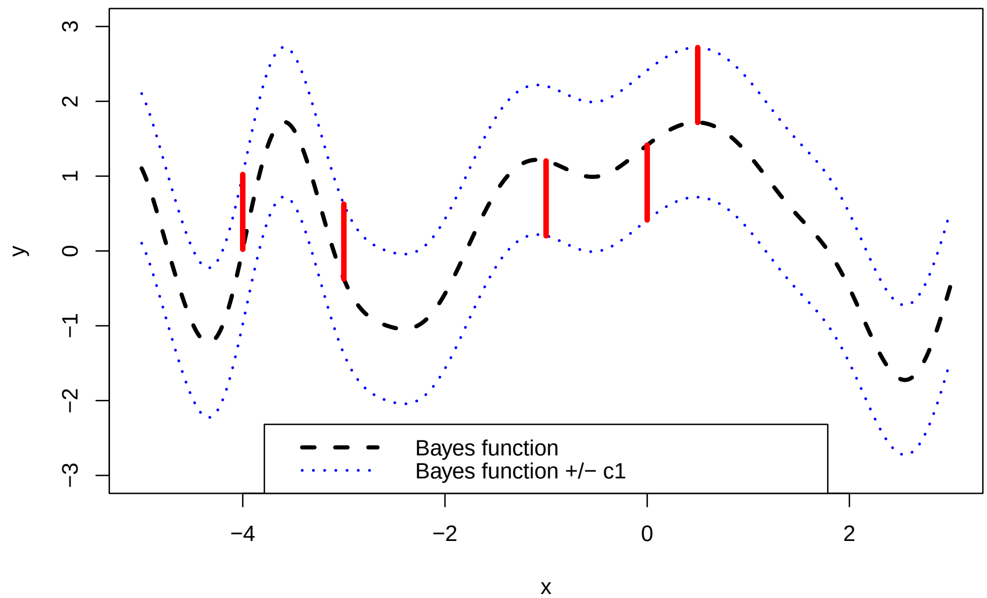

Figure 1: Visualization of (5). Each vertical slice between $f^*_{\tau,\mathrm{P}} - c_1$ and $f^*_{\tau,\mathrm{P}}$ as well as between $f^*_{\tau,\mathrm{P}}$ and $f^*_{\tau,\mathrm{P}} + c_1$ needs to have a conditional probability (given x) of at least c_2. The solid vertical lines depict some examples of such slices whose conditional probability needs to be at least c_2.

We saw that L_1-consistency can not be obtained from risk consistency without imposing some different, albeit in some sense weaker, condition regarding P in exchange for omitting the moment condition. It is, however, indeed possible to just omit the moment condition in the reverse statement (Theorem 6) when transferring this to shifted loss functions in the case of having a convex loss function of upper growth type 1, which again hints at this direction being the easier one as it was mentioned in the introduction.

Theorem 12 *Let $L\colon \mathcal{Y} \to \mathbb{R}$ be a convex, distance-based loss function of upper growth type 1, and let $L^\star$ be its shifted version. Then, for every sequence $(f_n)_{n\in\mathbb{N}} \subseteq L_1(\mathrm{P}_X)$ and every function $f^* \in L_1(\mathrm{P}_X)$, we have*

$$\lim_{n\to\infty} ||f_n - f^*||_{L_1(\mathrm{P}_X)} = 0 \qquad \Rightarrow \qquad \lim_{n\to\infty} \mathcal{R}_{L^\star,\mathrm{P}}(f_n) = \mathcal{R}_{L^\star,\mathrm{P}}(f^*)\,.$$

4. Consistency of Regularized Kernel Methods

After having derived general results regarding the connection between L_p- and risk consistency in Section 3, we would like to apply these results to special predictors now. More specifically, we investigate kernel-based regularized risk minimizers, which we also call *support vector machines (SVMs)*. We are thus using the term SVM in a broad sense, allowing not only for the hinge loss (as the expression SVM is used in some works) but rather for arbitrary loss functions including the distance-based losses used in Section 3.

We first give a formal definition and some further mathematical prerequisites regarding SVMs as well as a short recap of some of their known properties in Section 4.1. In Section 4.2, we then first use our results from Section 3.1 to derive a result on their L_p-consistency, where no general result existed so far, and then derive a new result on their risk consistency, which in some part slightly weakens the conditions from existing results on risk consistency. Finally, we examine SVMs based on shifted loss functions in Section 4.3.

4.1 Prerequisites regarding regularized kernel methods

As the true distribution P is usually unknown in practice, one has to make do with the information available about P, i.e. the data set $D_n := ((x_1, y_1), \dots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$ mentioned in the introduction and consisting of i.i.d. observations sampled from P, instead of minimizing $\mathcal{R}_{L,\mathrm{P}}$ directly. This is approached by using the empirical distribution

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

corresponding to D_n, with $\delta_{(x_i,y_i)}$ denoting the Dirac measure in (x_i, y_i), and defining the *empirical risk* $\mathcal{R}_{L,\mathrm{D}_n}$ analogously to $\mathcal{R}_{L,\mathrm{P}}$, which results in

$$\mathcal{R}_{L,\mathrm{D}_n}(f) := \mathbb{E}_{\mathrm{D}_n}\left[L(X, Y, f(X))\right] = \frac{1}{n} \sum_{i=1}^{n} L(x_i, y_i, f(x_i)) .$$

Because just minimizing $\mathcal{R}_{L,\mathrm{D}_n}$ constitutes an ill-posed problem and usually results in some extent of overfitting, a regularization term has to be added. This leads to the definition of SVMs as minimizers of the regularized risk. More specifically, the *empirical SVM* is defined as

$$f_{L,\mathrm{D}_n,\lambda} := \arg \inf_{f \in H} \mathcal{R}_{L,\mathrm{D}_n}(f) + \lambda ||f||_H^2 , \tag{7}$$

and the *theoretical SVM* analogously as

$$f_{L,\mathrm{P},\lambda} := \arg \inf_{f \in H} \mathcal{R}_{L,\mathrm{P}}(f) + \lambda ||f||_H^2 . \tag{8}$$

In both definitions, $\lambda > 0$ is a regularization parameter which controls the amount of regularization and H is the *reproducing kernel Hilbert space* (RKHS) of a measurable *kernel* on $\mathcal{X}$, i.e. a symmetric and positive definite function $k \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, cf. Aronszajn (1950); Berlinet and Thomas-Agnan (2004); Saitoh and Sawano (2016) among others for a thorough introduction to this topic. We are often be interested in bounded kernels for which we define $||k||_\infty := \sup_{x \in \mathcal{X}} \sqrt{k(x, x)}$. Additionally, we define the *canonical feature map* $\Phi \colon \mathcal{X} \to H$ by $\Phi(x) := k(\cdot, x)$.

SVMs have been widely investigated and have been shown to possess many desirable properties including existence, uniqueness, risk consistency, statistical robustness, and the existence of representation theorems under rather mild assumptions. See for example Vapnik (1995, 1998); Schölkopf and Smola (2002); Cucker and Zhou (2007); Steinwart and

Christmann (2008); Van Messem (2020) for a detailed introduction. More recent results regarding statistical robustness and stability in general have for example been derived by Hable and Christmann (2011); Sheng et al. (2020); Eckstein et al. (2023); Köhler and Christmann (2022). Results on learning rates (which have to make more restrictive assumptions regarding P because of the no-free-lunch-theorem, cf. Devroye, 1982) can for example be found in Caponnetto and De Vito (2007); Steinwart et al. (2009); Eberts and Steinwart (2013); Hang and Steinwart (2017); Fischer and Steinwart (2020).

4.2 Consistency of regularized kernel methods based on regular loss functions

Whereas SVMs based on distance-based losses are known to be risk consistent under mild assumptions (cf. Christmann and Steinwart, 2007, Theorem 12), there are no general results on their L_p-consistency so far, but instead only corollaries for special loss functions based on the results mentioned at the beginning of Section 3.1.

Since the conditions required by Christmann and Steinwart (2007, Theorem 12) also imply the validity of Theorem 4, L_p-consistency of such SVMs would now directly follow under these conditions. However, by some more thorough investigations, we are even able to slightly relax the conditions on the sequence $(\lambda_n)_{n\in\mathbb{N}}$ of regularization parameters, namely only requiring it to satisfy $\lambda_n^{p^*} n \to \infty$ (as $n \to \infty$) for $p^* = \max\{p+1, p(p+1)/2\}$ instead of for $p^* = \max\{2p, p^2\}$, which is required by Christmann and Steinwart (2007, Theorem 12).

Theorem 13 *Let $L\colon \mathcal{Y} \times \mathbb{R} \to [0,\infty)$ be a convex, distance-based loss function of growth type $p \in [1,\infty)$. Let $H \subseteq L_p(\mathrm{P}_X)$ dense and separable be the RKHS of a bounded and measurable kernel k. Assume that $f^*_{L,\mathrm{P}}$ is P_X-a.s. unique and $|\mathrm{P}|_p < \infty$. Define $p^* := \max\{p+1, p(p+1)/2\}$. If the sequence $(\lambda_n)_{n\in\mathbb{N}}$ satisfies $\lambda_n > 0$ for all $n \in \mathbb{N}$ as well as $\lambda_n \to 0$ and $\lambda_n^{p^*} n \to \infty$ for $n \to \infty$, then*

$$\lim_{n\to\infty} ||f_{L,\mathrm{D}_n,\lambda_n} - f^*_{L,\mathrm{P}}||_{L_p(\mathrm{P}_X)} = 0 \qquad \textit{in probability } \mathrm{P}^\infty.$$

The mentioned slight relaxation of the conditions on $(\lambda_n)_{n\in\mathbb{N}}$ means that Theorem 13 (as well as Corollary 15) also guarantees consistency for sequences of regularization parameters, for which Christmann and Steinwart (2007, Theorem 12) did not guarantee this, as can be seen from the following example:

Example 3 *The popular least squares loss is of growth type $p = 2$. Hence in this case $\max\{p+1, p(p+1)/2\} = 3 < 4 = \max\{2p, p^2\}$. Thus, Theorem 13 yields L_p-consistency (and Corollary 15 will yield risk consistency) of SVMs using the least squares loss under the condition that $\lambda_n^3 n \to \infty$ as $n \to \infty$, which is for example satisfied if $\lambda_n \propto n^{-1/4}$. On the other hand, Christmann and Steinwart (2007, Theorem 12) guarantees risk consistency of such SVMs only if $\lambda_n^4 n \to \infty$ as $n \to \infty$, which is not satisfied for $\lambda_n \propto n^{-1/4}$. Thus, our new results allow for slightly faster convergence of the regularization parameter to 0 and one therefore becomes more flexible in choosing the regularization parameters while still being guaranteed consistency.*

It should be noted that such a relaxation takes place whenever $p > 1$ holds true. $p = 1$ is the only case, in which $\max\{p+1, p(p+1)/2\} = \max\{2p, p^2\}$.

Remark 14 *The conditions on H in Theorem 13 can be difficult to check directly. However, if $\mathcal{X}$ is separable, the separability of H immediately follows whenever k is continuous (cf. Berlinet and Thomas-Agnan, 2004, Corollary 4) and it suffices to verify this continuity instead. For example, the commonly used Gaussian RBF kernel (among many other kernels) satisfies this continuity, and since additionally its RKHS is dense in $L_p(\mathrm{P}_X)$ (cf. Steinwart and Christmann, 2008, Theorem 4.63), the RKHS satisfies both conditions from Theorem 13.*

As we successfully slightly reduced the conditions regarding $(\lambda_n)_{n\in\mathbb{N}}$ compared to the referenced result on risk consistency, we can now transfer this slight relaxation back from L_p-consistency to risk consistency by using Theorem 6:

Corollary 15 *Let $L\colon \mathcal{Y}\times\mathbb{R}\to[0,\infty)$ be a convex, distance-based loss function of growth type $p\in[1,\infty)$. Let $H\subseteq L_p(\mathrm{P}_X)$ dense and separable be the RKHS of a bounded and measurable kernel k. Assume that $f^*_{L,\mathrm{P}}$ is P_X-a.s. unique and $|\mathrm{P}|_p<\infty$. Define $p^*:=\max\{p+1,p(p+1)/2\}$. If the sequence $(\lambda_n)_{n\in\mathbb{N}}$ satisfies $\lambda_n>0$ for all $n\in\mathbb{N}$ as well as $\lambda_n\to 0$ and $\lambda_n^{p^*}n\to\infty$ for $n\to\infty$, then*

$$\lim_{n\to\infty}\mathcal{R}_{L,\mathrm{P}}(f_{L,\mathrm{D}_n,\lambda_n})=\mathcal{R}^*_{L,\mathrm{P}}\qquad \textit{in probability } \mathrm{P}^\infty.$$

Alas, the slight relaxation of the mentioned condition regarding the regularization parameters also comes along with an additional condition compared to Christmann and Steinwart (2007, Theorem 12): Corollary 15 requires $f^*_{L,\mathrm{P}}$ to be P_X-a.s. unique. Thus, Corollary 15 pays for the slight relaxation in one condition by introducing this new additional condition and should therefore not be seen as a replacement of Theorem 12 from Christmann and Steinwart (2007) but as an addition instead.

Remark 16 *In the special case of using the pinball loss, Christmann and Steinwart (2008, Theorem 5) derived risk consistency of SVMs even in the almost sure sense under slightly stricter conditions regarding $(\lambda_n)_{n\in\mathbb{N}}$. Because of the non-probabilistic nature of the results from Section 3, Theorem 4 could in this special case be used to immediately obtain L_1-consistency of such SVMs in the almost sure sense as well.*

4.3 Consistency of regularized kernel methods based on shifted loss functions

SVMs based on shifted loss functions can be defined analogously as in the non-shifted case in (7) and (8). Christmann et al. (2009) proved that SVMs using Lipschitz continuous shifted loss functions inherit many of the desirable properties from their non-shifted counterparts, even without requiring the moment condition. These results include existence, uniqueness, representation and statistical robustness as well as risk consistency. Furthermore, they showed that $f_{L^\star,\mathrm{P},\lambda}=f_{L,\mathrm{P},\lambda}$ whenever $f_{L,\mathrm{P},\lambda}$ uniquely exists.

The natural hope that Theorem 13 can be transferred to the shifted case similarly, thus also ridding it of the moment condition, might have already decreased because of the negative results from Section 3.2. As SVMs are always contained in some RKHS H, one might however still hope that counterexamples like the ones from these results' proofs can not occur in such RKHSs because of the additional structure they possess compared to

$L_1(\mathrm{P}_X)$.[2] Alas, Sobolev spaces like the ones considered in Corollary 10 are also RKHSs if one chooses a suiting kernel like for example the ones found in Wu (1995); Wendland (2005), which are classical examples of kernels with compact support. Hence, we obtain the following:

Corollary 17 *Let $d \in \mathbb{N}$, $\mathcal{X} \subseteq \mathbb{R}^d$, and $\mathcal{Y} = \mathbb{R}$. Let $L\colon \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a convex, distance-based and symmetric loss function of growth type 1, or the τ-pinball loss for some $\tau \in (0, 1)$. Let $L^\star$ be its shifted version. Then, even if H is the RKHS of a bounded and measurable kernel k and $f^*_{L^\star,\mathrm{P}}$ is P_X-a.s. unique with $f^*_{L^\star,\mathrm{P}} \in L_1(\mathrm{P}_X)$, a sequence $(f_n)_{n\in\mathbb{N}} \subseteq H$ of functions satisfying*

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

*does in general **not** imply*

$$\lim_{n\to\infty} \left|\left|f_n - f^*_{L^\star,\mathrm{P}}\right|\right|_{L_1(\mathrm{P}_X)} = 0$$

without any additional assumptions besides Assumption 1 being imposed.

As the (probably) most commonly used RKHSs for computing SVMs are those of the *Gaussian RBF kernels* k_γ, $\gamma \in (0, \infty)$, defined by

$$k_\gamma(x, x') := \exp\left(-\frac{||x - x'||_2^2}{\gamma^2}\right) \qquad \forall\, x, x' \in \mathcal{X}\,,$$

we also want to take a special look at these. After proving in Corollary 17 that RKHSs, in which L_1-consistency does not follow from risk consistency, do in fact exist, we see in the subsequent Corollary 18 that this phenomenon can not only occur for kernels whose RKHS is a Sobolev space but also for that of the Gaussian RBF kernel.

Corollary 18 *Let $d \in \mathbb{N}$, $\mathcal{X} \subseteq \mathbb{R}^d$, and $\mathcal{Y} = \mathbb{R}$. Let $L\colon \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a convex, distance-based and symmetric loss function of growth type 1, or the τ-pinball loss for some $\tau \in (0, 1)$. Let $L^\star$ be its shifted version. Let $\gamma \in (0, \infty)$ and H_γ be the RKHS of the Gaussian RBF kernel k_γ. Then, even if $f^*_{L^\star,\mathrm{P}}$ is P_X-a.s. unique with $f^*_{L^\star,\mathrm{P}} \in L_1(\mathrm{P}_X)$, a sequence $(f_n)_{n\in\mathbb{N}} \subseteq H_\gamma$ of functions satisfying*

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

*does in general **not** imply*

$$\lim_{n\to\infty} \left|\left|f_n - f^*_{L^\star,\mathrm{P}}\right|\right|_{L_1(\mathrm{P}_X)} = 0$$

without any additional assumptions besides Assumption 1 being imposed.

2. The associated kernel k being bounded and measurable implies that all $f \in H$ are bounded and measurable as well, and hence that $H \subseteq L_1(\mathrm{P}_X)$, cf. Steinwart and Christmann (2008, Lemma 4.23 and 4.24).

The previous results show that L_1-consistency of SVMs using shifted loss functions does in general not follow from their risk consistency, with the latter being known from Christmann et al. (2009, Theorem 8). Note that it might still be possible for such SVMs to be L_1-consistent for different reasons though.

At least in the special case of the shifted pinball loss, we found some alternative conditions to replace—and in many situations weaken—the moment condition from Theorem 13. With this, we can now at least deduce L_1-consistency of SVMs using this shifted pinball loss without needing to impose the moment condition:

Corollary 19 *Let $\tau \in (0,1)$ and $L^\star_{\tau\text{-}pin}$ be the shifted τ-pinball loss. Let $H \subseteq L_1(\mathrm{P}_X)$ dense and separable be the RKHS of a bounded and measurable kernel k. Assume that $f^*_{\tau,\mathrm{P}}$ is P_X-a.s. unique, $f^*_{\tau,\mathrm{P}} \in L_1(\mathrm{P}_X)$ and P additionally satisfies at least one of the additional conditions (i) and (ii) from Theorem 11. If the sequence $(\lambda_n)_{n\in\mathbb{N}}$ satisfies $\lambda_n > 0$ for all $n \in \mathbb{N}$ as well as $\lambda_n \to 0$ and $\lambda_n^2 n \to \infty$ for $n \to \infty$, then*

$$\lim_{n\to\infty} ||f_{L^\star_{\tau\text{-}pin},\mathrm{D}_n,\lambda_n} - f^*_{\tau,\mathrm{P}}||_{L_1(\mathrm{P}_X)} = 0 \qquad \textit{in probability } \mathrm{P}^\infty.$$

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

Remark 20 *It would be possible to use Corollary 19 to derive a result on risk consistency of SVMs which are based on the shifted pinball loss, similarly to what we did in the non-shifted case in Section 4.2, where we used Theorem 13 to derive Corollary 15. In the latter result, we however only achieved an actual improvement (over already existing results) regarding the conditions on the regularization parameters if the loss function is of growth type $p > 1$, cf. Example 3. Similarly, a result on risk consistency which is based on Corollary 19 would offer no benefit over Theorem 8 from Christmann et al. (2009) because of the pinball loss being of growth type 1.*

5. Discussion

This paper considerably generalized existing results regarding the close relationship between L_p- and risk consistency by deriving results which are applicable to a wide range of loss functions. We additionally tried to eliminate the moment condition from the results connecting L_p- and risk consistency by switching to shifted loss functions. Somewhat surprisingly, this only worked for one of the two directions (risk consistency following from L_p-consistency), but in general not for the reverse. We proved that it is indeed not possible to infer L_p-consistency from risk consistency if neither some standard moment condition nor some suitable alternative condition holds true.

In case of using the shifted pinball loss, which can be used for quantile regression, we derived such an alternative condition, which is in many cases considerably weaker than the moment condition, thus still gaining some benefit from switching to shifted loss functions. It remains to be seen whether similar alternative conditions can also be derived for different loss functions or whether it might even be possible to derive a general alternative condition applicable to a wider array of loss functions.

Lastly, we applied our results to regularized kernel methods. By doing so, we proved their L_p-consistency in considerably greater generality than it had been known so far, and we slightly reduced a condition from results on their risk consistency from the literature.

Acknowledgments

I would like to thank Andreas Christmann for helpful discussions on this topic.

Appendix A. Proofs

In this appendix, we prove our results, split based on which section they were stated in.

A.1 Proofs for Section 3.1

Proof of Theorem 4 Let $g_n \colon \mathcal{X} \times \mathcal{Y} \to [0,\infty), (x,y) \mapsto L(y, f_n(x))$ for $n \in \mathbb{N}$, and $g^* \colon \mathcal{X} \times \mathcal{Y} \to [0,\infty), (x,y) \mapsto L(y, f^*_{L,\mathrm{P}}(x))$. According to Steinwart and Christmann (2008, Corollary 3.62)—where it is easy to see that we do not need the assumption of the sets $\mathcal{M}_{L,\mathrm{P}(\cdot\,|\,x),x}$ being singletons since we already know that $f^*_{L,\mathrm{P}}$ P_X-a.s. uniquely exists—, we have $f_n \xrightarrow{\mathrm{P}_X} f^*_{L,\mathrm{P}}$. Thus, because of the continuous mapping theorem and the continuity of L, we also have $g_n \xrightarrow{\mathrm{P}} g^*$. Since

$$
\begin{aligned}
\lim_{n\to\infty} \int |g_n| \,\mathrm{dP} = \lim_{n\to\infty} \int g_n \,\mathrm{dP} &= \lim_{n\to\infty} \mathcal{R}_{L,\mathrm{P}}(f_n) \\
&= \mathcal{R}_{L,\mathrm{P}}(f^*_{L,\mathrm{P}}) = \int g^* \,\mathrm{dP} = \int |g^*| \,\mathrm{dP}\,, \qquad (9)
\end{aligned}
$$

the sequence $(|g_n|)_{n\in\mathbb{N}}$ is thus equi-integrable according to Bauer (2001, Theorem 21.7). That theorem can be applied because $\mathcal{R}_{L,\mathrm{P}}(f^*_{L,\mathrm{P}}) < \infty$, and hence $\mathcal{R}_{L,\mathrm{P}}(f_n) < \infty$ for n sufficiently large because of (9), and therefore $g^* \in L_1(\mathrm{P}_X)$ and $g_n \in L_1(\mathrm{P}_X)$ for n sufficiently large.

Because of L being of lower growth type p, there now exists a constant $c > 0$ such that

$$
\begin{aligned}
|f_n(x) - f^*_{L,\mathrm{P}}(x)|^p &\le \max\left\{(2|y - f_n(x)|)^p\,,\, (2|y - f^*_{L,\mathrm{P}}(x)|)^p\right\} \\
&\le 2^p \cdot \max\left\{c^{-1}\big(L(y, f_n(x)) + 1\big)\,,\, c^{-1}\big(L(y, f^*_{L,\mathrm{P}}(x)) + 1\big)\right\} \\
&= \frac{2^p}{c} \cdot \big(\max\{g_n(x,y), g^*(x,y)\} + 1\big) \\
&\le \frac{2^p}{c} \cdot \big(g_n(x,y) + g^*(x,y) + 1\big) \qquad \forall\,(x,y,n) \in \mathcal{X} \times \mathcal{Y} \times \mathbb{N}\,, \qquad (10)
\end{aligned}
$$

since g_n, $n \in \mathbb{N}$, and g^* are non-negative.

As $(|g_n|)_{n\in\mathbb{N}}$ is equi-integrable, and $g^* \in L_1(\mathrm{P}_X)$ and hence also equi-integrable (cf. Bauer, 2001, part 2 of the example on p. 122), every summand occurring on the right hand side of (10) is equi-integrable (as a sequence in n). By employing the example on p. 121 of Bauer (2001) as well as Corollary 21.3 from the same book, we hence obtain equi-integrability of the whole right hand side (as a sequence in n).

Thus, the sequence $(|f_n - f^*_{L,\mathrm{P}}|^p)_{n\in\mathbb{N}}$ is equi-integrable as well and L_p-convergence of f_n to $f^*_{L,\mathrm{P}}$, follows from Bauer (2001, Theorem 21.7). ■

Proof of Theorem 6 Since $||f_n - f^*||_{L_p(\mathrm{P}_X)} \to 0$, we also have $f_n \xrightarrow{\mathrm{P}_X} f^*$, and Bauer (2001, Theorem 21.7) yields equi-integrability of the sequence $(|f_n|^p)_{n\in\mathbb{N}}$. Let $g_n \colon \mathcal{X}\times\mathcal{Y} \to [0,\infty), (x,y) \mapsto L(y, f_n(x))$ for $n \in \mathbb{N}$, and $g^* \colon \mathcal{X}\times\mathcal{Y} \to [0,\infty), (x,y) \mapsto L(y, f^*(x))$. Because of L being of upper growth type p, there then exists a $c > 0$ such that

$$\begin{aligned} |g_n(x,y)| = g_n(x,y) = L(y, f_n(x)) &\leq c \cdot (|y - f_n(x)|^p + 1) \\ &\leq c \cdot (2^p \cdot (|y|^p + |f_n(x)|^p) + 1) \end{aligned} \tag{11}$$

for all $(x,y,n) \in \mathcal{X}\times\mathcal{Y}\times\mathbb{N}$.

Since every summand on the right hand side of (11) is equi-integrable (because $|\mathrm{P}|_p < \infty$), the whole right hand side is equi-integrable as well (as a sequence in n) by the example on p. 121 of Bauer (2001) and Corollary 21.3 from the same book. Hence, the sequence $(|g_n|)_{n\in\mathbb{N}}$ is equi-integrable as well.

Additionally, $g_n \xrightarrow{\mathrm{P}} g^*$ because of $f_n \xrightarrow{\mathrm{P}_X} f^*$ and the continuous mapping theorem in combination with the continuity of L, and thus, Bauer (2001, Theorem 21.7) yields

$$\lim_{n\to\infty} \mathcal{R}_{L,\mathrm{P}}(f_n) = \lim_{n\to\infty} \int g_n \,\mathrm{dP} = \lim_{n\to\infty} \int |g_n| \,\mathrm{dP} = \int |g^*| \,\mathrm{dP} = \int g^* \,\mathrm{dP} = \mathcal{R}_{L,\mathrm{P}}(f^*)\,.$$

■

A.2 Proofs for Section 3.2

Before proving Proposition 8, we first need the following auxiliary lemma:

Lemma 21 *Let $L \colon \mathcal{X}\times\mathcal{Y}\times\mathbb{R} \to [0,\infty)$ be a convex and Lipschitz continuous loss function, and let $L^\star$ be its shifted version. If there exists a measurable function $f \colon \mathcal{X} \to \mathbb{R}$ satisfying $\mathcal{R}_{L^\star,\mathrm{P}}(f) = -\infty$, there also exists a measurable function $g \colon \mathcal{X} \to \mathbb{R}$ satisfying $\mathrm{P}_X(g \neq 0) > 0$ and $\mathcal{R}_{L^\star,\mathrm{P}}(g) \in (-\infty, 0]$.*

Proof If we denote the inner risk by

$$\mathcal{C}_{L^\star,\mathrm{P}(\cdot\,|\,x)} \colon \mathbb{R} \to \mathbb{R} \cup \{-\infty, +\infty\}\,,\ t \mapsto \int_{\mathcal{Y}} L^\star(x,y,t)\,\mathrm{dP}(y\,|\,x)\,,$$

we have

$$\begin{aligned} \mathcal{R}_{L^\star,\mathrm{P}}(f) &= \int L^\star(x,y,f(x))\,\mathrm{dP}(x,y) = \int \mathcal{C}_{L^\star,\mathrm{P}(\cdot\,|\,x)}(f(x))\,\mathrm{dP}_X(x) \\ &= \int \mathcal{C}^+_{L^\star,\mathrm{P}(\cdot\,|\,x)}(f(x))\,\mathrm{dP}_X(x) - \int \mathcal{C}^-_{L^\star,\mathrm{P}(\cdot\,|\,x)}(f(x))\,\mathrm{dP}_X(x) = -\infty\,, \end{aligned}$$

with $\mathcal{C}^+_{L^\star,\mathrm{P}(\cdot\,|\,x)} := \max\{\mathcal{C}_{L^\star,\mathrm{P}(\cdot\,|\,x)}\,, 0\}$ and $\mathcal{C}^-_{L^\star,\mathrm{P}(\cdot\,|\,x)} := \max\{-\mathcal{C}_{L^\star,\mathrm{P}(\cdot\,|\,x)}\,, 0\}$ denoting the positive and the negative part of $\mathcal{C}_{L^\star,\mathrm{P}(\cdot\,|\,x)}$ respectively. From the definition of the integral, we hence obtain

$$\int \mathcal{C}^-_{L^\star,\mathrm{P}(\cdot\,|\,x)}(f(x))\,\mathrm{dP}_X(x) = \infty \tag{12}$$

and therefore the existence of $c \in (0,\infty)$ and $A \subseteq \mathcal{X}$ measurable such that $\mathrm{P}_X(A) > 0$ and $\mathcal{C}^-_{L^\star,\mathrm{P}(\cdot\,|\,x)}(f(x)) \geq c$ for all $x \in A$.

We further know that $|L|_1 > 0$ because it is clear from the definition of Lipschitz continuous loss functions (cf. Remark 7) that $|L|_1 = 0$ would imply $L(x,y,f(x)) = L(x,y,0)$ for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$ and hence $\mathcal{R}_{L^\star,\mathrm{P}}(f) = 0$, which contradicts our assumptions. Therefore, (12) directly implies that $|f(x)| \geq \frac{c}{|L|_1} > 0$ for all $x \in A$ because otherwise

$$\begin{aligned}\mathcal{C}^-_{L^\star,\mathrm{P}(\cdot\,|\,x)}(f(x)) &= \left(\int L^\star(x,y,f(x))\,\mathrm{dP}_X(x)\right)^- \leq \int \big|L^\star(x,y,f(x))\big|\,\mathrm{dP}_X(x)\\ &= \int \big|L(x,y,f(x)) - L(x,y,0)\big|\,\mathrm{dP}_X(x) \leq |L|_1 \cdot |f(x)| < c\,,\end{aligned}$$

which would form a contradiction to x coming from A.

Define

$$g(x) := \begin{cases} 0 & \text{, if } x \notin A\,,\\ \frac{c}{|L|_1}\cdot \mathrm{sign}(f(x)) & \text{, if } x \in A\,.\end{cases}$$

Then, $\mathrm{P}_X(g \neq 0) > 0$ and

$$\mathcal{R}_{L^\star,\mathrm{P}}(g) = \int_A \mathcal{C}_{L^\star,\mathrm{P}(\cdot\,|\,x)}(g(x))\,\mathrm{dP}_X(x) + \underbrace{\int_{\mathcal{X}\setminus A} \mathcal{C}_{L^\star,\mathrm{P}(\cdot\,|\,x)}(g(x))\,\mathrm{dP}_X(x)}_{=0}\,. \tag{13}$$

All that remains to investigate is the first integral on the right hand side. For all $x \in A$, we know that

$$\big|\mathcal{C}_{L^\star,\mathrm{P}(\cdot\,|\,x)}(g(x))\big| \leq \int |L(x,y,g(x)) - L(x,y,0)|\,\mathrm{dP}(y\,|\,x) \leq |L|_1 \cdot |g(x)| = c$$

and

$$\mathcal{C}_{L^\star,\mathrm{P}(\cdot\,|\,x)}(g(x)) \leq \max\left\{\mathcal{C}_{L^\star,\mathrm{P}(\cdot\,|\,x)}(0), \mathcal{C}_{L^\star,\mathrm{P}(\cdot\,|\,x)}(f(x))\right\} = \mathcal{C}_{L^\star,\mathrm{P}(\cdot\,|\,x)}(0) = 0$$

because $g(x)$ lies between 0 and $f(x)$, $\mathcal{C}_{L^\star,\mathrm{P}(\cdot\,|\,x)}(f(x)) < 0$ by definition of A, and $\mathcal{C}_{L^\star,\mathrm{P}(\cdot\,|\,x)}$ is convex (which follows from L being convex).

Plugging this into the right hand side of (13) yields $\mathcal{R}_{L^\star,\mathrm{P}}(g) \in [-c, 0]$ and hence the assertion. ■

Proof of Proposition 8 We prove the statement by providing a counterexample.

Because of L being of lower growth type 1,

$$c_0 := \sup\{r \in [0,\infty) \,|\, \psi(r) = 0\}$$

is finite, where ψ denotes the representing function belonging to L, as introduced in Definition 2. Because of L being convex, distance-based, and symmetric, we have

$$L(y,t) = \psi(y-t) = 0 \qquad \Leftrightarrow \qquad y-t \in [-c_0, c_0]\,. \tag{14}$$

Assume without loss of generality that $c_0 \leq \frac{1}{2}$ (else just scale the subsequent example accordingly).

Choose $\mathcal{X} := (0,1)$, $\mathrm{P}_X := \mathcal{U}(0,1)$ and

$$\mathrm{P}(\cdot\,|\,X=x) := x \cdot \mathcal{U}(-1,1) + \frac{1-x}{2} \cdot \left(\delta_{-a_x} + \delta_{a_x}\right) \qquad \forall\, x \in \mathcal{X}\,, \tag{15}$$

where $\mathcal{U}(a,b)$ denotes the uniform distribution on (a,b), δ_z denotes the Dirac distribution in $z \in \mathbb{R}$ and $a_x > 1$ is a constant depending on x (and on L) that we will specify later on.[3] Further define

$$f_n \colon \mathcal{X} \to \mathbb{R}\,, \qquad x \mapsto \begin{cases} n & , \text{ if } x \in \left(0, \frac{1}{n}\right)\,, \\ 0 & , \text{ else}\,, \end{cases} \tag{16}$$

for $n \in \mathbb{N}$. As f_n is bounded for all $n \in \mathbb{N}$, we obviously have $(f_n)_{n\in\mathbb{N}} \subseteq L_1(\mathrm{P}_X)$. We now show that this example also possesses the remaining properties mentioned in the proposition, which consists of three main steps:

First, we show that $f^*_{L^\star,\mathrm{P}}$ is P_X-a.s. unique, more specifically $f^*_{L^\star,\mathrm{P}} \equiv 0$ P_X-a.s., and $f^*_{L^\star,\mathrm{P}} \in L_1(\mathrm{P}_X)$:

Choose $f^* \equiv 0$. We show that $\mathcal{R}_{L^\star,\mathrm{P}}(f^*) < \mathcal{R}_{L^\star,\mathrm{P}}(f)$ for all measurable $f \colon \mathcal{X} \to \mathbb{R}$ satisfying $\mathrm{P}_X(f \neq 0) > 0$. As $\mathcal{R}_{L^\star,\mathrm{P}}(f^*) = 0$, the case $\mathcal{R}_{L^\star,\mathrm{P}}(f) = \infty$ is trivial. Furthermore, if there was an f satisfying $\mathcal{R}_{L^\star,\mathrm{P}}(f) = -\infty$ and thus contradicting our claim, there would by Lemma 21 (which is applicable by Remark 7) also exist a measurable g with $\mathrm{P}_X(g \neq 0) > 0$ and $-\infty < \mathcal{R}_{L^\star,\mathrm{P}}(g) \leq 0 = \mathcal{R}_{L^\star,\mathrm{P}}(f^*)$, which would also contradict our claim. Hence, we can without loss of generality assume that $\mathcal{R}_{L^\star,\mathrm{P}}(f) \in \mathbb{R}$.

Since $f^* \equiv 0$, we have, for each $x \in \mathcal{X}$ and $y \geq 0$,

$$\begin{aligned} L^\star\left(-y, f^*(x)\right) + L^\star\left(y, f^*(x)\right) &= 2 \cdot L^\star(y,0) \\ &= 2 \cdot L^\star\left(y,\, \frac{1}{2} \cdot (-f(x)) + \frac{1}{2} \cdot f(x)\right) \\ &\leq L^\star\left(y, -f(x)\right) + L^\star\left(y, f(x)\right) \\ &= L^\star\left(-y, f(x)\right) + L^\star\left(y, f(x)\right) \end{aligned} \tag{17}$$

because of L being distance-based, symmetric and convex.

Furthermore, by the definition of f, there exists $\varepsilon := (\varepsilon_1, \varepsilon_2)$ with $\varepsilon_1, \varepsilon_2 > 0$ such that

3. For the sake of strictly adhering to the completeness assumption from Assumption 1, we can also choose $\mathcal{X}$ as $[0,1]$ or $\mathbb{R}$, and $\mathrm{P}(\cdot\,|\,X=x)$ as an arbitrary probability measure for $x \notin (0,1)$ without changing anything else.

$\mathrm{P}_X(\mathcal{X}_\varepsilon) > 0$, where $\mathcal{X}_\varepsilon := \{x \in \mathcal{X} : |f(x)| \geq \varepsilon_1 \text{ and } x \geq \varepsilon_2\}$. Now, specifically look at $x \in \mathcal{X}_\varepsilon$ and $y \in [c_0, c_0 + \min\{\frac{1}{2}, \frac{|f(x)|}{4}\}] \subseteq [0,1]$. First, only consider such x that satisfy $f(x) > 0$. We then obtain that

$$|-y - f(x)| = y + f(x) \geq c_0 + f(x) \quad \text{and} \quad |\pm y - f^*(x)| = y \leq c_0 + \frac{f(x)}{4}\,, \tag{18}$$

and hence

$$L(-y, f(x)) \geq 4 \cdot L(-y, f^*(x)) = 2 \cdot \Big(L\left(y, f^*(x)\right) + L\left(-y, f^*(x)\right) \Big)$$

because of (14) and the convexity, symmetry and distance-basedness of L. Thus,

$$\begin{aligned}
&\Big(L^\star\left(-y, f(x)\right) + L^\star\left(y, f(x)\right) \Big) - \Big(L^\star\left(-y, f^*(x)\right) + L^\star\left(y, f^*(x)\right) \Big) \\
&= \Big(L\left(-y, f(x)\right) + L\left(y, f(x)\right) \Big) - \Big(L\left(-y, f^*(x)\right) + L\left(y, f^*(x)\right) \Big) \\
&\geq \frac{1}{2} \cdot L\left(-y, f(x)\right) = \frac{1}{2} \cdot \psi(|-y - f(x)|) \geq \frac{1}{2} \cdot \psi(c_0 + f(x))\,,
\end{aligned}$$

where, in the last step, we again applied the convexity and symmetry of L, as well as (18). By interchanging the roles of y and $-y$ in the preceding paragraph, we obtain an analogous inequality for the case that $f(x) < 0$. Combining these two cases yields that

$$\begin{aligned}
&\Big(L^\star\left(-y, f(x)\right) + L^\star\left(y, f(x)\right) \Big) - \Big(L^\star\left(-y, f^*(x)\right) + L^\star\left(y, f^*(x)\right) \Big) \\
&\geq \frac{1}{2} \cdot \psi(c_0 + |f(x)|)
\end{aligned} \tag{19}$$

for all $x \in \mathcal{X}_\varepsilon$ and $y \in [c_0, c_0 + \min\{\frac{1}{2}, \frac{|f(x)|}{4}\}] \subseteq [0,1]$.
Because $\mathcal{R}_{L^\star,\mathrm{P}}(f^*) = 0 \in \mathbb{R}$ by the definition of f^* and $\mathcal{R}_{L^\star,\mathrm{P}}(f) \in \mathbb{R}$ by assumption, our considerations yield

$$\begin{aligned}
&\mathcal{R}_{L^\star,\mathrm{P}}(f) - \mathcal{R}_{L^\star,\mathrm{P}}(f^*) \\
&= \int_{\mathcal{X}} \int_{\mathcal{Y}} L^\star\left(y, f(x)\right) - L^\star\left(y, f^*(x)\right) \,\mathrm{dP}(y\,|\,x)\,\mathrm{dP}_X(x) \\
&= \int_{\mathcal{X}} \int_{[0,\infty)} \Big(L^\star\left(-y, f(x)\right) + L^\star\left(y, f(x)\right) \Big) \\
&\qquad\qquad\qquad - \Big(L^\star\left(-y, f^*(x)\right) + L^\star\left(y, f^*(x)\right) \Big) \,\mathrm{dP}(y\,|\,x)\,\mathrm{dP}_X(x) \\
&\overset{(17),(19)}{\geq} \int_{\mathcal{X}_\varepsilon} \int_{[c_0, c_0 + \min\{\frac{1}{2}, \frac{|f(x)|}{4}\}]} \frac{1}{2} \cdot \psi(c_0 + |f(x)|) \,\mathrm{dP}(y\,|\,x)\,\mathrm{dP}_X(x) \\
&= \int_{\mathcal{X}_\varepsilon} \frac{x}{2} \cdot \min\left\{\frac{1}{2}, \frac{|f(x)|}{4}\right\} \cdot \frac{1}{2} \cdot \psi(c_0 + |f(x)|) \,\mathrm{dP}_X(x) \\
&\geq \mathrm{P}_X(\mathcal{X}_\varepsilon) \cdot \frac{\varepsilon_2}{2} \cdot \min\left\{\frac{1}{2}, \frac{\varepsilon_1}{4}\right\} \cdot \frac{1}{2} \cdot \psi(c_0 + \varepsilon_1) \\
&\overset{(14)}{>} 0\,.
\end{aligned}$$

In the second step, we multiplied the integrand by 2 for $y = 0$, which does not change the value of the integral since $\mathrm{P}(Y = 0 \,|\, X = x) = 0$ for all $x \in \mathcal{X}$. In the final steps, we additionally applied that $\mathrm{P}(\cdot \,|\, X = x)$ has Lebesgue density $\frac{x}{2}$ on $[c_0, c_0 + \min\{\frac{1}{2}, \frac{|f(x)|}{4}\}] \subseteq [0,1]$, respectively the definition of $\mathcal{X}_\varepsilon$.
Hence, $f^*_{L^\star,\mathrm{P}} \equiv 0$ P_X-a.s. and thus also $f^*_{L^\star,\mathrm{P}} \in L_1(\mathrm{P}_X)$.

Next, we show that $\lim_{n\to\infty} \mathcal{R}_{L^\star,\mathrm{P}}(f_n) = \mathcal{R}^*_{L^\star,\mathrm{P}}$:
Recall the definition of f_n, $n \in \mathbb{N}$, from (16). For all $n \in \mathbb{N}$, we have $f^*_{L^\star,\mathrm{P}}, f_n \in L_1(\mathrm{P}_X)$ and therefore $\mathcal{R}^*_{L^\star,\mathrm{P}} = \mathcal{R}_{L^\star,\mathrm{P}}(f^*_{L^\star,\mathrm{P}}) \in \mathbb{R}$ and $\mathcal{R}_{L^\star,\mathrm{P}}(f_n) \in \mathbb{R}$ by (1). Hence, we can write

$$\begin{aligned}
&\mathcal{R}_{L^\star,\mathrm{P}}(f_n) - \mathcal{R}^*_{L^\star,\mathrm{P}} \\
&= \int_{\mathcal{X}} \int_{\mathcal{Y}} L^\star\left(y, f_n(x)\right) - L^\star\left(y, f^*_{L^\star,\mathrm{P}}(x)\right) \,\mathrm{d}\mathrm{P}(y\,|\,x)\,\mathrm{d}\mathrm{P}_X(x) \\
&= \int_{\mathcal{X}} \int_{\mathcal{Y}} L\left(y, f_n(x)\right) - L\left(y, f^*_{L^\star,\mathrm{P}}(x)\right) \,\mathrm{d}\mathrm{P}(y\,|\,x)\,\mathrm{d}\mathrm{P}_X(x) \\
&= \int_0^{1/n} \int_{-1}^{1} \frac{x}{2} \cdot \big(L(y,n) - L(y,0)\big)\,\mathrm{d}y\,\mathrm{d}x \\
&\quad + \int_0^{1/n} \frac{1-x}{2} \cdot \Big(\big(L(-a_x, n) + L(a_x, n)\big) - \big(L(-a_x, 0) + L(a_x, 0)\big)\Big)\,\mathrm{d}x\,,
\end{aligned} \tag{20}$$

where we applied the definition of f_n, $f^*_{L^\star,\mathrm{P}}$, and P in the last step. We will now analyze the two integrals on the right hand side separately and show that they both converge to 0 as $n \to \infty$, starting with the first one:

$$\begin{aligned}
&\left|\int_0^{1/n} \int_{-1}^{1} \frac{x}{2} \cdot \big(L(y,n) - L(y,0)\big)\,\mathrm{d}y\,\mathrm{d}x\right| \\
&\leq \int_0^{1/n} \int_{-1}^{1} \frac{x}{2} \cdot |L|_1 \cdot |n - 0|\,\mathrm{d}y\,\mathrm{d}x = \frac{|L|_1}{2n} \xrightarrow{n\to\infty} 0
\end{aligned}$$

with L being Lipschitz continuous by Remark 7.
As for the second integral on the right hand side of (20):
We take a look at the subdifferential $\partial\psi$ (cf. Phelps, 1993, Definition 1.9) of the representing function ψ of L. Because of the symmetry of L, we will without loss of generality only investigate $\partial\psi(r)$ for $r \in [0,\infty)$. Define

$$z(r) := \sup \partial\psi(r) \in [0,\infty) \qquad \forall\, r \in [0,\infty)\,,$$

where $z(r) < \infty$ will follow from (21) and $z(r) \geq 0$ follows from L being monotonically increasing on $[0,\infty)$ because of it being distance-based and convex. Furthermore, let c_L be the constant from the definition of the upper growth type 1 of L, that is

$$\psi(r) \leq c_L \cdot (|r| + 1) \qquad \forall\, r \in \mathbb{R}\,.$$

Assume there was an $r_0 \in [0,\infty)$ such that $z(r_0) > c_L$. Then, by the definition of the subdifferential, we would obtain

$$c_L \cdot (r+1) \geq \psi(r) \geq \psi(r_0) + z(r_0) \cdot (r - r_0) \qquad \forall\, r \in [0,\infty)$$

and hence

$$r \leq \frac{\psi(r_0) - z(r_0)r_0 - c_L}{c_L - z(r_0)} \qquad \forall\, r \in [0, \infty)\,,$$

which is a contradiction because the right hand side is a constant in $\mathbb{R}$ that is independent of r. Hence, z is bounded by c_L. Because of L additionally being monotonically increasing on $[0, \infty)$, we obtain that

$$\tilde{c}_L := \lim_{r\to\infty} z(r) = \sup_{r\in[0,\infty)} z(r) \leq c_L \tag{21}$$

exists.
We can therefore, for each $x \in (0, 1)$, choose $r_x \in [0, \infty)$ such that

$$0 \leq \tilde{c}_L - z(r_x) \leq x \tag{22}$$

and

$$\psi(r_x) + z(r_x)\cdot(r - r_x) \leq \psi(r) \leq \psi(r_x) + \tilde{c}_L \cdot (r - r_x) \qquad \forall\, r \in [r_x, \infty)\,. \tag{23}$$

Now choose a_x in the definition of $\mathrm{P}(\cdot\,|\,X = x)$ in (15) as $a_x := r_x + \frac{1}{x}$ for all $x \in (0, 1)$. Please note that $a_x > 1$ for all $x \in (0, 1)$. We obtain

$$\begin{aligned}
&L(-a_x, n) + L(a_x, n)\\
&= \psi\left(|-a_x - n|\right) + \psi\left(|a_x - n|\right)\\
&= \psi\left(r_x + \frac{1}{x} + n\right) + \psi\left(r_x + \frac{1}{x} - n\right)\\
&\in \left[2\cdot\psi(r_x) + z(r_x)\cdot\left(\frac{1}{x} + n + \frac{1}{x} - n\right)\,,\, 2\cdot\psi(r_x) + \tilde{c}_L\cdot\left(\frac{1}{x} + n + \frac{1}{x} - n\right)\right]\\
&= \left[2\cdot\left(\psi(r_x) + \frac{z(r_x)}{x}\right)\,,\, 2\cdot\left(\psi(r_x) + \frac{\tilde{c}_L}{x}\right)\right] \qquad \forall\, n \in \mathbb{N}\,,\, x \in \left(0, \frac{1}{n}\right),
\end{aligned}$$

where we applied the symmetry of L as well as (23) combined with the fact that $\frac{1}{x} + n \geq 0$ and $\frac{1}{x} - n \geq 0$. Analogously, we obtain

$$\begin{aligned}
&L(-a_x, 0) + L(a_x, 0)\\
&= 2\cdot\psi\left(r_x + \frac{1}{x}\right)\\
&\in \left[2\cdot\left(\psi(r_x) + \frac{z(r_x)}{x}\right)\,,\, 2\cdot\left(\psi(r_x) + \frac{\tilde{c}_L}{x}\right)\right] \qquad \forall\, x \in \left(0, \frac{1}{n}\right).
\end{aligned}$$

Plugging these results into the second integral on the right hand side of (20) finally yields

$$\begin{aligned}
&\left|\int_0^{1/n} \frac{1-x}{2} \cdot \Big(\big(L(-a_x, n) + L(a_x, n)\big) - \big(L(-a_x, 0) + L(a_x, 0)\big)\Big)\, \mathrm{d}x\right| \\
&\le \int_0^{1/n} \frac{1-x}{2} \cdot \left(2 \cdot \left(\psi(r_x) + \frac{\tilde{c}_L}{x}\right) - 2 \cdot \left(\psi(r_x) + \frac{z(r_x)}{x}\right)\right)\, \mathrm{d}x \\
&= \int_0^{1/n} \frac{1-x}{2} \cdot \frac{2}{x} \cdot (\tilde{c}_L - z(r_x))\, \mathrm{d}x \\
&\overset{(22)}{\le} \int_0^{1/n} (1-x)\, \mathrm{d}x = \frac{1}{n} - \frac{1}{2n^2} \xrightarrow{n\to\infty} 0\,,
\end{aligned}$$

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

Finally and as a last step, we have to show that $\lim_{n\to\infty} \left|\left|f_n - f^*_{L^\star,\mathrm{P}}\right|\right|_{L_1(\mathrm{P}_X)} \ne 0$:

$$\lim_{n\to\infty} \left|\left|f_n - f^*_{L^\star,\mathrm{P}}\right|\right|_{L_1(\mathrm{P}_X)} = \lim_{n\to\infty} \int_0^{1/n} |n-0|\, \mathrm{d}x = \lim_{n\to\infty} 1 \ne 0\,.$$

■

Proof of Proposition 9 Similarly to Proposition 8, we prove the statement by providing a counterexample:

Choose $\mathcal{X} := (0,1)$, $\mathcal{Y} := \mathbb{R}$, $\mathrm{P}_X := \mathcal{U}(0,1)$, and

$$\begin{aligned}
\mathrm{P}(\cdot\,|\,X=x) = x \cdot \Big(\tau \cdot \mathcal{U}((-1,0)) + (1-\tau)\cdot \mathcal{U}((0,1))\Big) \\
+ (1-x)\cdot \Big(\tau \cdot \delta_{-1/x} + (1-\tau)\cdot \delta_{1/x}\Big) \qquad \forall\, x \in \mathcal{X}\,,
\end{aligned}$$

where $\mathcal{U}((a,b))$ denotes the uniform distribution on (a,b) and δ_z denotes the Dirac distribution in $z \in \mathbb{R}$.[4] From this definition, we immediately obtain that $f^*_{\tau,\mathrm{P}} \equiv 0 \in L_1(\mathrm{P}_X)$.

Further define

$$f_n\colon \mathcal{X} \to \mathbb{R}\,, \qquad x \mapsto \begin{cases} n & \text{, if } x \in \left(0, \frac{1}{n}\right)\,, \\ 0 & \text{, else}\,, \end{cases}$$

for all $n \in \mathbb{N}$. As f_n is bounded for all $n \in \mathbb{N}$, we obviously have $(f_n)_{n\in\mathbb{N}} \subseteq L_1(\mathrm{P}_X)$.

Because of the occurring risks both being finite, cf. (1), and $\mathcal{R}^*_{L^\star_{\tau\text{-pin}},\mathrm{P}} = \mathcal{R}_{L^\star_{\tau\text{-pin}},\mathrm{P}}(f^*_{\tau,\mathrm{P}})$, cf. (3), we can for all $n \in \mathbb{N}$ write

$$\begin{aligned}
&\mathcal{R}_{L^\star_{\tau\text{-pin}},\mathrm{P}}(f_n) - \mathcal{R}^*_{L^\star_{\tau\text{-pin}},\mathrm{P}} \\
&= \int_{(0,1)} \int_{\mathbb{R}} L^\star_{\tau\text{-pin}}(y, f_n(x)) - L^\star_{\tau\text{-pin}}(y, f^*_{\tau,\mathrm{P}}(x))\, \mathrm{d}\mathrm{P}(y\,|\,x)\, \mathrm{d}\mathrm{P}_X(x)\,. \qquad (24)
\end{aligned}$$

4. For the sake of strictly adhering to the completeness assumption from Assumption 1, we can also choose $\mathcal{X}$ as $[0,1]$ or $\mathbb{R}$, and $\mathrm{P}(\cdot\,|\,X=x)$ as an arbitrary probability measure for $x \notin (0,1)$ without changing anything else.

For P_X-almost all $x \in \mathcal{X}$, we can now further analyze the inner integral, applying that $f_n(x) \geq f^*_{\tau,\mathrm{P}}(x)$, by

$$
\begin{aligned}
&\int_{\mathbb{R}} L^\star_{\tau\text{-pin}}(y, f_n(x)) - L^\star_{\tau\text{-pin}}(y, f^*_{\tau,\mathrm{P}}(x)) \,\mathrm{dP}(y\,|\,x) \\
&= \int_{\mathbb{R}} L_{\tau\text{-pin}}(y, f_n(x)) - L_{\tau\text{-pin}}(y, f^*_{\tau,\mathrm{P}}(x)) \,\mathrm{dP}(y\,|\,x) \\
&= \int_{\left(-\infty, f^*_{\tau,\mathrm{P}}(x)\right)} (1-\tau) \cdot \left(f_n(x) - f^*_{\tau,\mathrm{P}}(x)\right) \,\mathrm{dP}(y\,|\,x) \\
&\quad + \int_{\left[f^*_{\tau,\mathrm{P}}(x), f_n(x)\right)} (-\tau) \cdot \left(f_n(x) - f^*_{\tau,\mathrm{P}}(x)\right) + (f_n(x) - y) \,\mathrm{dP}(y\,|\,x) \\
&\quad + \int_{[f_n(x),\infty)} (-\tau) \cdot \left(f_n(x) - f^*_{\tau,\mathrm{P}}(x)\right) \,\mathrm{dP}(y\,|\,x) \\
&= \int_{\left[f^*_{\tau,\mathrm{P}}(x), f_n(x)\right)} (f_n(x) - y) \,\mathrm{dP}(y\,|\,x)\,. \qquad (25)
\end{aligned}
$$

In the last step, we employed that, for P_X-almost all $x \in \mathcal{X}$, we know from the definition of P that $\mathrm{P}(\{f^*_{\tau,\mathrm{P}}(x)\}\,|\,x) = 0$ and therefore $\mathrm{P}((-\infty, f^*_{\tau,\mathrm{P}}(x))\,|\,x) = \tau$ and $\mathrm{P}([f^*_{\tau,\mathrm{P}}(x), \infty)\,|\,x) = 1-\tau$ by the definition of $f^*_{\tau,\mathrm{P}}$.

Plugging (25) and the definition of f_n and $f^*_{\tau,\mathrm{P}}$ into (24), we obtain

$$
\begin{aligned}
\mathcal{R}_{L^\star_{\tau\text{-pin}},\mathrm{P}}(f_n) - \mathcal{R}^*_{L^\star_{\tau\text{-pin}},\mathrm{P}} &= \int_{\left(0,\frac{1}{n}\right)} \int_{[0,n)} (n-y) \,\mathrm{dP}(y\,|\,x)\,\mathrm{dP}_X(x) \\
&= \int_0^{\frac{1}{n}} \int_0^1 (n-y) \cdot x \cdot (1-\tau)\,\mathrm{d}y\,\mathrm{d}x \\
&= (1-\tau) \cdot \frac{2n-1}{4n^2} \to 0\,, \qquad n \to \infty\,.
\end{aligned}
$$

On the other hand,

$$
\left|\left|f_n - f^*_{\tau,\mathrm{P}}\right|\right|_{L_1(\mathrm{P}_X)} = \int_0^{\frac{1}{n}} |n-0| \,\mathrm{d}x = 1 \not\to 0\,, \qquad n \to \infty\,,
$$

which completes the proof. ∎

Proof of Corollary 10 The assertion follows directly from the proof of Proposition 8 respectively Proposition 9 by changing the functions f_n, $n \in \mathbb{N}$, to

$$
f_n \colon \mathcal{X} \to \mathbb{R}\,, \qquad x \mapsto \begin{cases} n \cdot (1-nx)^m & \text{, if } x \in \left(0, \frac{1}{n}\right)\,, \\ 0 & \text{, else}\,. \end{cases}
$$

Since, for all $n \in \mathbb{N}$, f_n is bounded and m times weakly differentiable, we obtain $(f_n)_{n\in\mathbb{N}} \subseteq W^{m,\infty}(\mathcal{X}) \cap L_1(\mathrm{P}_X) \subseteq W^{m,q}(\mathcal{X}) \cap L_1(\mathrm{P}_X)$.[5]

5. If $\mathcal{X}$ is not chosen as $(0,1)$ but instead as $[0,1]$ or $\mathbb{R}$ in the proofs of Proposition 8 and Proposition 9, it is obviously possible to extend the functions f_n, $n \in \mathbb{N}$, in such a way that they are still in $W^{m,\infty}(\mathcal{X}) \cap L_1(\mathrm{P}_X)$.

If we denote the functions from the mentioned proofs by g_n, $n \in \mathbb{N}$, we have $f^*_{L^\star,\mathrm{P}}(x) \le f_n(x) \le g_n(x)$ for P_X-almost all $x \in \mathcal{X}$ because $f^*_{L^\star,\mathrm{P}} = 0$ P_X-a.s. (with $f^*_{L^\star,\mathrm{P}} = f^*_{\tau,\mathrm{P}}$ P_X-a.s. in the situation of $L^\star = L^\star_{\tau\text{-pin}}$ by the considerations prior to Proposition 9). It is easy to see that the convexity of L and the definition of $f^*_{L^\star,\mathrm{P}}$ as a minimizer of $\mathcal{R}_{L^\star,\mathrm{P}}$ therefore implies $\mathcal{R}_{L^\star,\mathrm{P}}(f_n) - \mathcal{R}^*_{L^\star,\mathrm{P}} \le \mathcal{R}_{L^\star,\mathrm{P}}(g_n) - \mathcal{R}^*_{L^\star,\mathrm{P}}$, which then yields $\lim_{n\to\infty} \mathcal{R}_{L^\star,\mathrm{P}}(f_n) = \mathcal{R}^*_{L^\star,\mathrm{P}}$.

At the same time, we obtain

$$\left|\left|f_n - f^*_{L^\star,\mathrm{P}}\right|\right|_{L_1(\mathrm{P}_X)} = \int_0^{1/n} |n \cdot (1-nx)^m - 0| \,\mathrm{d}x = \frac{1}{m+1} \not\to 0\,, \qquad n \to \infty\,,$$

which completes the proof. ■

Proof of Theorem 11 By (1), both $\mathcal{R}_{L^\star_{\tau\text{-pin}},\mathrm{P}}(f_n)$, $n \in \mathbb{N}$, and $\mathcal{R}_{L^\star_{\tau\text{-pin}},\mathrm{P}}(f^*_{\tau,\mathrm{P}})$ are finite.

If condition (i) is satisfied, we further obtain as in Remark 5 that $\mathcal{R}_{L_{\tau\text{-pin}},\mathrm{P}}(0)$ and $\mathcal{R}_{L_{\tau\text{-pin}},\mathrm{P}}(f_n)$, for $n \in \mathbb{N}$, are finite, and therefore also $\mathcal{R}^*_{L_{\tau\text{-pin}},\mathrm{P}}$. As $\mathcal{R}^*_{L_{\tau\text{-pin}},\mathrm{P}} = \mathcal{R}_{L_{\tau\text{-pin}},\mathrm{P}}(f^*_{\tau,\mathrm{P}})$ and $\mathcal{R}^*_{L^\star_{\tau\text{-pin}},\mathrm{P}} = \mathcal{R}_{L^\star_{\tau\text{-pin}},\mathrm{P}}(f^*_{\tau,\mathrm{P}})$ by (3) and (4), we hence obtain

$$\mathcal{R}_{L_{\tau\text{-pin}},\mathrm{P}}(f_n) = \mathcal{R}_{L^\star_{\tau\text{-pin}},\mathrm{P}}(f_n) + \mathcal{R}_{L_{\tau\text{-pin}},\mathrm{P}}(0) \qquad \forall\, n \in \mathbb{N}$$

and

$$\mathcal{R}^*_{L_{\tau\text{-pin}},\mathrm{P}} = \mathcal{R}_{L_{\tau\text{-pin}},\mathrm{P}}(f^*_{\tau,\mathrm{P}}) = \mathcal{R}_{L^\star_{\tau\text{-pin}},\mathrm{P}}(f^*_{\tau,\mathrm{P}}) + \mathcal{R}_{L_{\tau\text{-pin}},\mathrm{P}}(0) = \mathcal{R}^*_{L^\star_{\tau\text{-pin}},\mathrm{P}} + \mathcal{R}_{L_{\tau\text{-pin}},\mathrm{P}}(0)\,.$$

Theorem 4 and Remark 5 then yield the assertion because of $L_{\tau\text{-pin}}$ being of growth type 1. Thus, it is only left to show that condition (ii) yields the assertion as well:

Because of the finiteness of $\mathcal{R}_{L^\star_{\tau\text{-pin}},\mathrm{P}}(f_n)$, $n \in \mathbb{N}$, and $\mathcal{R}_{L^\star_{\tau\text{-pin}},\mathrm{P}}(f^*_{\tau,\mathrm{P}})$, the assumed risk consistency implies that the P-integral of $L^\star_{\tau\text{-pin}}(y, f_n(x)) - L^\star_{\tau\text{-pin}}(y, f^*_{\tau,\mathrm{P}}(x))$ converges to 0 as $n \to \infty$. We will now begin by fixing an $x \in \mathcal{X}$ and further analyzing the inner integral with respect to $\mathrm{P}(\cdot\,|\,x)$:

First, we look at the case that $f_n(x) \ge f^*_{\tau,\mathrm{P}}(x)$. In this case, repeating the considerations from (25), where we can apply (6) in the last step, yields for P_X-almost all such x that

$$\begin{aligned}
&\int_{\mathcal{Y}} L^\star_{\tau\text{-pin}}(y, f_n(x)) - L^\star_{\tau\text{-pin}}(y, f^*_{\tau,\mathrm{P}}(x)) \,\mathrm{dP}(y|x) \\
&\overset{(25)}{=} \int_{\left[f^*_{\tau,\mathrm{P}}(x), f_n(x)\right)} (f_n(x) - y) \,\mathrm{dP}(y|x) \\
&\ge \int_{\left[f^*_{\tau,\mathrm{P}}(x), \frac{f_n(x)+f^*_{\tau,\mathrm{P}}(x)}{2}\right)} (f_n(x) - y) \,\mathrm{dP}(y|x) \\
&\ge \left(f_n(x) - \frac{f_n(x)+f^*_{\tau,\mathrm{P}}(x)}{2}\right) \cdot \mathrm{P}\left(\left(f^*_{\tau,\mathrm{P}}(x), \frac{f_n(x)+f^*_{\tau,\mathrm{P}}(x)}{2}\right) \,\middle|\, x\right) \\
&= \frac{f_n(x) - f^*_{\tau,\mathrm{P}}(x)}{2} \cdot \mathrm{P}\left(\left(f^*_{\tau,\mathrm{P}}(x), \frac{f_n(x)+f^*_{\tau,\mathrm{P}}(x)}{2}\right) \,\middle|\, x\right)\,.
\end{aligned}$$

If on the other hand $f_n(x) < f^*_{\tau,\mathrm{P}}(x)$, we analogously obtain for P_X-almost all such x:

$$\begin{aligned}&\int_{\mathcal{Y}} L^\star_{\tau\text{-pin}}(y, f_n(x)) - L^\star_{\tau\text{-pin}}(y, f^*_{\tau,\mathrm{P}}(x))\,\mathrm{dP}(y|x)\\ &\geq \frac{f^*_{\tau,\mathrm{P}}(x) - f_n(x)}{2} \cdot \mathrm{P}\left(\left(\frac{f_n(x) + f^*_{\tau,\mathrm{P}}(x)}{2}, f^*_{\tau,\mathrm{P}}(x)\right)\middle|\, x\right).\end{aligned}$$

In summary,

$$\int_{\mathcal{Y}} L^\star_{\tau\text{-pin}}(y, f_n(x)) - L^\star_{\tau\text{-pin}}(y, f^*_{\tau,\mathrm{P}}(x))\,\mathrm{dP}(y|X) \geq \frac{|f_n(X) - f^*_{\tau,\mathrm{P}}(X)|}{2} \cdot \mathrm{P}\left(J_{X,n}|\, X\right) \tag{26}$$

P_X-a.s., where $J_{x,n} := \left(\min\left\{f^*_{\tau,\mathrm{P}}(x), \frac{f_n(x)+f^*_{\tau,\mathrm{P}}(x)}{2}\right\}, \max\left\{f^*_{\tau,\mathrm{P}}(x), \frac{f_n(x)+f^*_{\tau,\mathrm{P}}(x)}{2}\right\}\right)$ for all $x \in \mathcal{X}$.

Additionally, Christmann et al. (2009, Corollary 31) yields $f_n \xrightarrow{\mathrm{P}_X} f^*_{\tau,\mathrm{P}}$, i.e.

$$\lim_{n\to\infty} \mathrm{P}_X(|f_n(X) - f^*_{\tau,\mathrm{P}}(X)| > \varepsilon) = 0 \qquad \forall \varepsilon > 0\,. \tag{27}$$

Now, let $\varepsilon > 0$ be an arbitrary positive number (without loss of generality $\varepsilon < 2c_1$). $\mathcal{X}$ can be partitioned as $\mathcal{X} = \dot{\bigcup}_{i=1}^3 \mathcal{X}_{i,\varepsilon}$, where

$$\begin{aligned}\mathcal{X}_{1,\varepsilon} &:= \left\{x \in \mathcal{X} : |f_n(x) - f^*_{\tau,\mathrm{P}}(x)| \leq \varepsilon\right\},\\ \mathcal{X}_{2,\varepsilon} &:= \left\{x \in \mathcal{X} : \varepsilon < |f_n(x) - f^*_{\tau,\mathrm{P}}(x)| \leq 2 \cdot c_1\right\},\\ \mathcal{X}_{3,\varepsilon} &:= \mathcal{X}_3 := \left\{x \in \mathcal{X} : |f_n(x) - f^*_{\tau,\mathrm{P}}(x)| > 2 \cdot c_1\right\},\end{aligned}$$

such that

$$||f_n - f^*_{\tau,\mathrm{P}}||_{L_1(\mathrm{P}_X)} = \sum_{i=1}^3 \int_{\mathcal{X}_{i,\varepsilon}} |f_n(x) - f^*_{\tau,\mathrm{P}}(x)|\,\mathrm{dP}_X(x)\,. \tag{28}$$

The three summands can now be analyzed separately:

$$\begin{aligned}&\int_{\mathcal{X}_{1,\varepsilon}} |f_n(x) - f^*_{\tau,\mathrm{P}}(x)|\,\mathrm{dP}_X(x) \leq \varepsilon,\\ &\int_{\mathcal{X}_{2,\varepsilon}} |f_n(x) - f^*_{\tau,\mathrm{P}}(x)|\,\mathrm{dP}_X(x) \leq 2 \cdot c_1 \cdot \mathrm{P}_X(\mathcal{X}_{2,\varepsilon}) \xrightarrow{(27)} 0\,, \qquad n \to \infty\,,\end{aligned}$$

and

$$\begin{aligned}&\int_{\mathcal{X}_{3,\varepsilon}} |f_n(x) - f^*_{\tau,\mathrm{P}}(x)|\,\mathrm{dP}_X(x)\\ &= \int_{\mathcal{X}_3} \left(\frac{|f_n(x) - f^*_{\tau,\mathrm{P}}(x)|}{2} \cdot \mathrm{P}(J_{x,n}|x)\right) \cdot \frac{2}{\mathrm{P}(J_{x,n}|x)}\,\mathrm{dP}_X(x)\\ &\overset{(5),(26)}{\leq} \frac{2}{c_2} \cdot \int_{\mathcal{X}_3}\int_{\mathcal{Y}} L^\star_{\tau\text{-pin}}(y, f_n(x)) - L^\star_{\tau\text{-pin}}(y, f^*_{\tau,\mathrm{P}}(x))\,\mathrm{dP}(y|x)\,\mathrm{dP}_X(x)\\ &\to 0\,, \qquad n \to \infty\,,\end{aligned}$$

with the last convergence holding true because

$$\int_{\mathcal{X}}\int_{\mathcal{Y}} L^{\star}_{\tau\text{-pin}}(y, f_n(x)) - L^{\star}_{\tau\text{-pin}}(y, f^*_{\tau,\mathrm{P}}(x))\,\mathrm{dP}(y|x)\,\mathrm{dP}_X(x) \to 0\,, \qquad n\to\infty\,,$$

by assumption and

$$\int_{\mathcal{Y}} L^{\star}_{\tau\text{-pin}}(y, f_n(x)) - L^{\star}_{\tau\text{-pin}}(y, f^*_{\tau,\mathrm{P}}(x))\,\mathrm{dP}(y|X) \geq 0$$

P_X-a.s. by (26).

Plugging these results into (28) yields the assertion. ■

Proof of Theorem 12 We know from (1) that all risks appearing in this result are finite. L additionally being Lipschitz continuous (cf. Remark 7) yields

$$\begin{aligned}
|\mathcal{R}_{L^\star,\mathrm{P}}(f_n) - \mathcal{R}_{L^\star,\mathrm{P}}(f^*)| &\leq \int |L^\star(y, f_n(x)) - L^\star(y, f^*(x))|\,\mathrm{dP}(x,y)\\
&= \int |L(y, f_n(x)) - L(y, f^*(x))|\,\mathrm{dP}(x,y)\\
&\leq |L|_1 \cdot \int |f_n(x) - f^*(x)|\,\mathrm{dP}(x,y)\\
&= |L|_1 \cdot ||f_n - f^*||_{L_1(\mathrm{P}_X)} \to 0 \qquad n\to\infty\,.
\end{aligned}$$

■

A.3 Proofs for Section 4.2

Proof of Theorem 13 We can split up the difference, which we have to investigate, as

$$\begin{aligned}
\left|\left|f_{L,\mathrm{D}_n,\lambda_n} - f^*_{L,\mathrm{P}}\right|\right|_{L_p(\mathrm{P}_X)} &\leq ||f_{L,\mathrm{D}_n,\lambda_n} - f_{L,\mathrm{P},\lambda_n}||_{L_p(\mathrm{P}_X)} + \left|\left|f_{L,\mathrm{P},\lambda_n} - f^*_{L,\mathrm{P}}\right|\right|_{L_p(\mathrm{P}_X)}\\
&\leq ||k||_\infty\, ||f_{L,\mathrm{D}_n,\lambda_n} - f_{L,\mathrm{P},\lambda_n}||_H + \left|\left|f_{L,\mathrm{P},\lambda_n} - f^*_{L,\mathrm{P}}\right|\right|_{L_p(\mathrm{P}_X)} \qquad (29)
\end{aligned}$$

by Steinwart and Christmann (2008, Lemma 4.23). We will now examine the two summands on the right hand side separately, starting with the first one:

First, note that applying Steinwart and Christmann (2008, Lemma 4.23, equation (5.4) and Lemma 2.38(i)) yields

$$||f_{L,\mathrm{P},\lambda_n}||_\infty \leq ||k||_\infty \cdot ||f_{L,\mathrm{P},\lambda_n}||_H \leq ||k||_\infty \cdot \mathcal{R}_{L,\mathrm{P}}(0)^{1/2} \cdot \lambda_n^{-1/2} \leq c_{p,L,\mathrm{P},k} \cdot \lambda_n^{-1/2} \qquad (30)$$

for all $n \in \mathbb{N}$, with $c_{p,L,\mathrm{P},k} \in (0,\infty)$ denoting a constant depending only on p, L, P and k, but not on λ_n.

We know from Steinwart and Christmann (2008, Corollary 5.11) that there exist functions $h_n\colon \mathcal{X}\times\mathcal{Y}\to\mathbb{R}$, $n\in\mathbb{N}$, such that

$$||f_{L,\mathrm{D}_n,\lambda_n} - f_{L,\mathrm{P},\lambda_n}||_H \leq \frac{1}{\lambda_n} \cdot ||\mathbb{E}_{\mathrm{D}_n}[h_n\Phi] - \mathbb{E}_{\mathrm{P}}[h_n\Phi]||_H \qquad \forall\, n\in\mathbb{N}\,, \qquad (31)$$

and, for $s := p/(p-1)$,

$$\begin{aligned} \|h_n\|_{L_s(\mathrm{P})} &\leq 8^p \cdot c_L \cdot \Big(1 + |\mathrm{P}|_p^{p-1} + \|f_{L,\mathrm{P},\lambda_n}\|_\infty^{p-1}\Big) \\ &\leq 8^p \cdot c_L \cdot \Big(1 + |\mathrm{P}|_p^{p-1} + c_{p,L,\mathrm{P},k}^{p-1} \cdot \lambda_n^{-(p-1)/2}\Big) \\ &\leq \tilde{c}_{p,L,\mathrm{P},k} \cdot \lambda_n^{-(p-1)/2} \qquad\qquad \forall\, n \in \mathbb{N}\,, \end{aligned} \tag{32}$$

where we employed (30) in the second and the boundedness of $(\lambda_n)_{n\in\mathbb{N}}$ in the third step, and where $c_L \in (0,\infty)$ and $\tilde{c}_{p,L,\mathrm{P},k} \in (0,\infty)$ denote constants depending only on L respectively p, L, P and k.

Now, we can apply Steinwart and Christmann (2008, Lemma 9.2) with $q := p/(p-1)$ if $p > 1$ and $q := 2$ if $p = 1$, which leads to $q^* := \min\{1/2, 1 - 1/q\} = \min\{1/2, 1/p\} = (p+1)/(2p^*)$, to the functions $h_n\Phi$, $n \in \mathbb{N}$: First of all, with the help of (32) we obtain

$$\|h_n\Phi\|_q := \big(\mathbb{E}_{\mathrm{P}}\big[\|h_n\Phi\|_H^q\big]\big)^{1/q} \leq \|k\|_\infty \cdot \|h_n\|_{L_q(\mathrm{P})} \leq \|k\|_\infty \cdot \tilde{c}_{p,L,\mathrm{P},k} \cdot \lambda_n^{-(p-1)/2} < \infty$$

for all $n \in \mathbb{N}$. We employed that, for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$,

$$\|h_n(x,y)\Phi(x)\|_H^q = |h_n(x,y)|^q \cdot \|\Phi(x)\|_H^q = |h_n(x,y)|^q \cdot k(x,x)^{q/2} \leq |h_n(x,y)|^q \cdot \|k\|_\infty^q$$

by the reproducing property (cf. for example Schölkopf and Smola, 2002, Definition 2.9). Hence, we obtain for all $\varepsilon > 0$, by combining this Lemma 9.2 with (31),

$$\begin{aligned} &\mathrm{P}^n\big(D_n \in (\mathcal{X}\times\mathcal{Y})^n : \|f_{L,\mathrm{D}_n,\lambda_n} - f_{L,\mathrm{P},\lambda_n}\|_H \geq \varepsilon\big) \\ &\leq \mathrm{P}^n\left(D_n \in (\mathcal{X}\times\mathcal{Y})^n : \|\mathbb{E}_{\mathrm{D}_n}[h_n\Phi] - \mathbb{E}_{\mathrm{P}}[h_n\Phi]\|_H \geq \lambda_n \cdot \varepsilon\right) \\ &\leq c_q \cdot \left(\frac{\|h_n\Phi\|_q}{\lambda_n \varepsilon n^{q^*}}\right)^q \leq \hat{c}_{p,L,\mathrm{P},k} \cdot \left(\frac{1}{\lambda_n^{(p+1)/2}\varepsilon n^{q^*}}\right)^q \to 0\,, \qquad n \to \infty\,, \end{aligned}$$

with $c_q \in (0,\infty)$ and $\hat{c}_{p,L,\mathrm{P},k} \in (0,\infty)$ denoting constants depending only on q (that is, only on p) respectively p, L, P and k, and with the convergence in the last step holding true because

$$\lambda_n^{(p+1)/2} n^{q^*} = \Big(\lambda_n^{(p+1)/(2q^*)} n\Big)^{q^*} = \Big(\lambda_n^{p^*} n\Big)^{q^*} \to \infty\,, \qquad n \to \infty\,,$$

by the assumptions on $(\lambda_n)_{n\in\mathbb{N}}$. Thus, the first summand on the right hand side of (29) converges to 0 in probability as $n \to \infty$.

Now, we can turn our attention to the second summand: First of all, Steinwart and Christmann (2008, Lemma 2.38(i)) yields that L is a P-integrable Nemitski loss of order p. Hence, we know from Steinwart and Christmann (2008, Theorem 5.31) that

$$\mathcal{R}_{L,\mathrm{P},H}^* := \inf_{f\in H} \mathcal{R}_{L,\mathrm{P}}(f) = \mathcal{R}_{L,\mathrm{P}}^*\,,$$

and Steinwart and Christmann (2008, Lemma 5.15) (with $\mathcal{R}_{L,\mathrm{P},H}^* = \mathcal{R}_{L,\mathrm{P}}^* < \infty$ by Remark 5) then yields

$$\lim_{n\to\infty} \lambda_n \|f_{L,\mathrm{P},\lambda_n}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_{L,\mathrm{P},\lambda_n}) - \mathcal{R}_{L,\mathrm{P}}^* = 0$$

because $\lambda_n \to 0$ as $n \to \infty$. Since $\lambda_n \, ||f_{L,\mathrm{P},\lambda_n}||_H^2$ is non-negative and $\mathcal{R}_{L,\mathrm{P}}(f_{L,\mathrm{P},\lambda_n}) \geq \mathcal{R}^*_{L,\mathrm{P}}$ by the definition of $\mathcal{R}^*_{L,\mathrm{P}}$, we obtain

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

Hence, Theorem 4, whose conditions are satisfied because of the considerations from Remark 5, yields convergence to 0 (as $n \to \infty$) of the second summand on the right hand side of (29), which completes the proof. ■

Proof of Corollary 15 The assertion follows directly from Theorem 13 and Theorem 6. ■

A.4 Proofs for Section 4.3

Proof of Corollary 17 There exist different kernels whose RKHS is $W^{2,2}(\mathcal{X})$. Examples of such kernels can be found in Wu (1995), Berlinet and Thomas-Agnan (2004, Chapter 7), Saitoh and Sawano (2016, Theorem 1.11) among others. For this proof, we will however use the kernel $k_{1,1}$ defined by $k_{1,1}(x, x') := \phi_{1,1}(||x - x'||_2)$ with $\phi_{1,1}$ as in Wendland (2005, Definition 9.11), that is $\phi_{1,1}(r) \propto (1 - r)^3_+(3r + 1)$ (cf. Wendland, 2005, Table 9.1). By Wendland (2005, Theorem 10.35), the RKHS of $k_{1,1}$ is indeed $W^{2,2}(\mathcal{X})$. Additionally, $k_{1,1}$ is bounded by $\phi_{1,1}(0) < \infty$ and because of its continuity also measurable. Applying Corollary 10 yields the assertion. ■

Proof of Corollary 18 Denote, for some $m \in \mathbb{N}$, the functions from the proof of Corollary 10 by g_n, $n \in \mathbb{N}$. Because $(g_n)_{n\in\mathbb{N}} \subseteq L_1(\mathrm{P}_X)$, there exists by Steinwart and Christmann (2008, Theorem 4.63) a sequence $(f_n)_{n\in\mathbb{N}} \subseteq H_\gamma$ such that

$$||f_n - g_n||_\infty \leq \frac{1}{n}$$

for all $n \in \mathbb{N}$.

Since both f_n and g_n are bounded, we obtain from (1) that, for all $n \in \mathbb{N}$, $\mathcal{R}_{L^\star,\mathrm{P}}(f_n) \in \mathbb{R}$ and $\mathcal{R}_{L^\star,\mathrm{P}}(g_n) \in \mathbb{R}$. Hence,

$$\begin{aligned}
&|\mathcal{R}_{L^\star,\mathrm{P}}(f_n) - \mathcal{R}_{L^\star,\mathrm{P}}(g_n)| \leq \int_{\mathcal{X}\times\mathcal{Y}} |L^\star(y, f_n(x)) - L^\star(y, g_n(x))| \, \mathrm{dP}(x, y) \\
&= \int_{\mathcal{X}\times\mathcal{Y}} |L(y, f_n(x)) - L(y, g_n(x))| \, \mathrm{dP}(x, y) \leq |L|_1 \cdot \int_{\mathcal{X}\times\mathcal{Y}} |f_n(x) - g_n(x)| \, \mathrm{dP}(x, y) \\
&\leq |L|_1 \cdot \frac{1}{n} \to 0 \,, \qquad n \to \infty \,.
\end{aligned}$$

with L being Lipschitz continuous by Remark 7. The risk consistency of $(g_n)_{n\in\mathbb{N}}$ shown in the proof of Corollary 10 then yields risk consistency of $(f_n)_{n\in\mathbb{N}}$.

On the other hand,

$$\lim_{n\to\infty} ||f_n - g_n||_{L_1(\mathrm{P}_X)} = \lim_{n\to\infty} \int_{\mathcal{X}} |f_n(x) - g_n(x)| \, \mathrm{dP}_X(x) \leq \lim_{n\to\infty} \frac{1}{n} = 0$$

combined with

$$\lim_{n\to\infty} \big|\big|g_n - f^*_{L^\star,\mathrm{P}}\big|\big|_{L_1(\mathrm{P}_X)} = \frac{1}{m+1}\,,$$

which is known from the proof of Corollary 10, yields

$$\lim_{n\to\infty} \big|\big|f_n - f^*_{L^\star,\mathrm{P}}\big|\big|_{L_1(\mathrm{P}_X)} \geq \lim_{n\to\infty} \Big(\big|\big|g_n - f^*_{L^\star,\mathrm{P}}\big|\big|_{L_1(\mathrm{P}_X)} - ||f_n - g_n||_{L_1(\mathrm{P}_X)}\Big) = \frac{1}{m+1}$$

and thus $(f_n)_{n\in\mathbb{N}}$ not being L_1-consistent. ■

Proof of Corollary 19 Christmann et al. (2009, Theorem 8) yields

$$\lim_{n\to\infty} \mathcal{R}_{L^\star,\mathrm{P}}(f_{L^\star_{\tau\text{-pin}},\mathrm{D}_n,\lambda_n}) = \mathcal{R}_{L^\star,\mathrm{P}}(f^*_{\tau,\mathrm{P}})$$

in probability P^∞ respectively even P^∞-almost surely. The assertion follows directly from Theorem 11. ■

References

Adams, R. A. and Fournier, J. J. F. (2003). *Sobolev Spaces*. Pure and Applied Mathematics. Elsevier.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404.

Bauer, H. (2001). *Measure and Integration Theory*. de Gruyter Studies in Mathematics. de Gruyter, Berlin, Boston.

Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer Science+Business Media, New York.

Biau, G., Devroye, L., and Lugosi, G. (2008). Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 9:2015–2033.

Caponnetto, A. and De Vito, E. (2007). Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368.

Cherkassky, V. and Mulier, F. (2007). *Learning from Data*. Wiley, Hoboken, New Jersey, 2nd edition.

Christmann, A. and Steinwart, I. (2007). Consistency and robustness of kernel-based regression in convex risk minimization. *Bernoulli*, 13(3):799–819.

Christmann, A. and Steinwart, I. (2008). Consistency of kernel-based quantile regression. *Applied Stochastic Models in Business and Industry*, 24:171–183.

Christmann, A., Van Messem, A., and Steinwart, I. (2009). On consistency and robustness properties of support vector machines for heavy-tailed distributions. *Statistics and Its Interface*, 2(3):311–327.

Cucker, F. and Smale, S. (2001). On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(1):1–49.

Cucker, F. and Zhou, D.-X. (2007). *Learning Theory: An Approximation Theory Viewpoint*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, Cambridge.

Devroye, L. (1982). Any discrimination rule can have an arbitrarily bad probability of error for finite sample size. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 4(2):154–157.

Dudley, R. M. (2004). *Real Analysis and Probability*. Cambridge University Press, Cambridge.

Eberts, M. and Steinwart, I. (2013). Optimal regression rates for SVMs using Gaussian kernels. *Electronic Journal of Statistics*, 7:1–42.

Eckstein, S., Iske, A., and Trabs, M. (2023). Dimensionality reduction and Wasserstein stability for kernel regression. *Journal of Machine Learning Research*, 24:1–35.

Farooq, M. and Steinwart, I. (2019). Learning rates for kernel-based expectile regression. *Machine Learning*, 108:203–227.

Fischer, S. and Steinwart, I. (2020). Sobolev norm learning rates for regularized least-squares algorithms. *Journal of Machine Learning Research*, 21:1–38.

Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer Series in Statistics. Springer, New York.

Hable, R. and Christmann, A. (2011). On qualitative robustness of support vector machines. *Journal of Multivariate Analysis*, 102:993–1007.

Hable, R. and Christmann, A. (2014). Estimation of scale functions to model heteroscedasticity by regularised kernel-based quantile methods. *Journal of Nonparametric Statistics*, 26(2):219–239.

Hang, H. and Steinwart, I. (2017). A Bernstein-type inequality for some mixing processes and dynamical systems with an application to learning. *The Annals of Statistics*, 45(2):708–743.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In Le Cam, L. M. and Neyman, J., editors, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 221–233, Berkeley.

Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, New Jersey, 2nd edition.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46(1):33–50.

Koenker, R. and Hallock, K. F. (2001). Quantile regression. *Journal of Economic Perspectives*, 15(4):143–156.

Köhler, H. and Christmann, A. (2022). Total stability of SVMs and localized SVMs. *Journal of Machine Learning Research*, 23(100):1–41.

Lin, S.-B., Wang, K., Wang, Y., and Zhou, D.-X. (2022). Universal consistency of deep convolutional neural networks. *IEEE Transactions on Information Theory*, 68(7):4610–4617.

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

Saitoh, S. and Sawano, Y. (2016). *Theory of Reproducing Kernels and Applications*, volume 44 of *Developments in Mathematics*. Springer Science+Business Media, Singapore.

Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels.* Adaptive Computation and Machine Learning. MIT Press, Cambridge, Massachusetts.

Sheng, B., Liu, H., and Wang, H. (2020). Learning rates for the kernel regularized regression with a differentiable strongly convex loss. *Communications on Pure & Applied Analysis*, 19(8):3973–4005.

Steinwart, I. (2005). Consistency of support vector machines and other regularized kernel classifiers. *IEEE Transactions on Information Theory*, 51(1):128–142.

Steinwart, I. and Christmann, A. (2008). *Support Vector Machines.* Information Science and Statistics. Springer, New York.

Steinwart, I. and Christmann, A. (2011). Estimating conditional quantiles with the help of the pinball loss. *Bernoulli*, 17(1):211–225.

Steinwart, I., Hush, D., and Scovel, C. (2009). Optimal rates for regularized least squares regression. In *Proceedings of the 22nd Annual Conference on Learning Theory*, pages 79–93.

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

Tong, H. and Ng, M. K. (2019). Calibration of ε-insensitive loss in support vector machines regression. *Journal of the Franklin Institute*, 356:2111–2129.

Van Messem, A. (2020). Support vector machines: A robust prediction method with applications in bioinformatics. In Rao, A. S. R. S. and Rao, C. R., editors, *Handbook of Statistics*, volume 43, pages 391–466. Elsevier.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory.* Springer, New York.

Vapnik, V. N. (1998). *Statistical Learning Theory.* Adaptive and Learning Systems for Signal Processing, Communications and Control. Wiley, New York.

Wendland, H. (2005). *Scattered Data Approximation.* Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, Cambridge.

Wu, Z. (1995). Compactly supported positive definite radial functions. *Advances in Computational Mathematics*, 4:283–292.

Xiang, D.-H., Hu, T., and Zhou, D.-X. (2012). Approximation analysis of learning algorithms for support vector regression and quantile regression. *Journal of Applied Mathematics*, 2012.

Zhang, T. and Yu, B. (2005). Boosting with early stopping: Convergence and consistency. *The Annals of Statistics*, 33(4):1538–1579.