On the Robustness of Regularized Pairwise Learning Methods Based on Kernels†

Andreas Christmann[1] and **Ding-Xuan Zhou**[2]
[1] University of Bayreuth, Germany
[2] City University of Hong Kong, China

Date: June 21, 2016

Abstract

Regularized empirical risk minimization including support vector machines plays an important role in machine learning theory. In this paper regularized pairwise learning (RPL) methods based on kernels will be investigated. One example is regularized minimization of the error entropy loss which has recently attracted quite some interest from the viewpoint of consistency and learning rates. This paper shows that such RPL methods and also their empirical bootstrap have additionally good statistical robustness properties, if the loss function and the kernel are chosen appropriately. We treat two cases of particular interest: (i) a bounded and non-convex loss function and (ii) an unbounded convex loss function satisfying a certain Lipschitz type condition.

Key words and phrases. Machine learning, pairwise loss function, regularized risk, robustness.

1 Introduction

Regularized empirical risk minimization based on kernels has attracted a lot of interest during the last decades in statistical machine learning. To fix ideas, let $D_n = ((x_1, y_1), \ldots, (x_n, y_n))$ be a given data set, where the value x_i denotes the input value and y_i denotes the output value of the i-the data point. Let L be a loss function which is typically of the form $L(x, y, f(x))$, where $f(x)$ denotes the predicted value for y, when x is observed, and the real-valued function f is unknown. Many regularized learning methods are then defined as minimizers of the optimization problem

$$\inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} L(x_i, y_i, f(x_i)) + \operatorname{pen}(\lambda, f), \tag{1.1}$$

where the set $\mathcal{F}$ consists of real-valued functions f, $\lambda > 0$ is a regularization constant, and $\operatorname{pen}(\lambda, f) \geq 0$ is some regularization term to avoid overfitting for the case, that $\mathcal{F}$ is rich. One example is that $\mathcal{F}$ is a reproducing kernel Hilbert space H and $\operatorname{pen}(\lambda, f) = \lambda \|f\|_H^2$, see e.g. Vapnik (1995, 1998), Poggio and Girosi (1998), Wahba (1999), Schölkopf and Smola (2002), Cucker and Zhou (2007), Steinwart and Christmann (2008) and the references cited there. Regularized empirical risk minimization based on kernels has also been investigated for additive models. We refer to Christmann and Hable (2012) for results on consistency and robustness and to Christmann and Zhou (2015) for fast learning rates.

†Corresponding author: A. Christmann, Email: andreas.christmann@uni-bayreuth.de
The work by A. Christmann described in this paper is partially supported by a grant of the Deutsche Forschungsgesellschaft [Project No. CH/291/2-1]. The work by D.-X. Zhou described in this paper is supported partially by a grant from the NSFC/RGC Joint Research Scheme [RGC Project No. N_CityU120/14 and NSFC Project No. 11461161006].

In recent years there is quite some interest in related learning methods where a *pairwise loss function* is used, which yields optimization problems like

$$\inf_{f\in H} \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} L(x_i, y_i, x_j, y_j, f(x_i), f(x_j)) + \lambda\|f\|_H^2 \tag{1.2}$$

or asymptotically equivalent versions of it. In other words, the estimator for f is defined as the minimizer of the sum of a V-statistic of degree 2 and the regularizing term $\lambda\|f\|_H^2$, see e.g. Serfling (1980). An example of this class of learning methods occurs when one is interested in minimizing Renyi's entropy of order 2, see e.g. Hu *et al.* (2013), Fan *et al.* (2016), and Ying and Zhou (2015) for consistency and fast learning rates. Another example arises from ranking algorithms, see e.g. Clemencon *et al.* (2008) and Agarwal and Niyogi (2009). Other examples include gradient learning, and metric and similarity learning, see e.g. Mukherjee and Zhou (2006), Xing *et al.* (2002), and Cao *et al.* (2015). However, much less theory is currently known for such regularized learning methods given by (1.2) based on a pairwise loss function than for the more classical problem (1.1) using a standard loss function. This is true in particular for statistical robustness aspects. Statistical robustness is one important facet of a statistical method, especially if the data quality is only moderate or unknown, which is often the case in the so-called big data situation.

The main goal of this paper is to show that such regularized learning methods given by (1.2) have nice statistical robustness properties if a bounded and continuous kernel is used in combination with a convex, smooth, and separately Lipschitz continuous (see Definition 2.5) pairwise loss function. We also establish a representer theorem for such regularized pairwise learning methods, because we need it for our proofs, but the representer theorem may also be helpful to further research.

The rest of the paper has the following structure. In Section 2, we define pairwise loss functions, their corresponding risks, derive some basic properties of pairwise loss functions and their risks, and give some examples. In Section 3 we define regularized pairwise learning (RPL) methods treated in this paper and derive results on existence and uniqueness. We will show that shifted loss functions (defined in (3.9)) are useful to define RPL methods on the set of *all* probability measures without making moment assumptions. This is of course desirable, because the probability measure chosen by nature to generate the data is completely unknown in machine learning theory. Section 4 contains a representer theorem for RPL methods, which is our first main result, see Theorem 4.3. This result is interesting in its own right, but we use it as a tool to prove our statistical robustness results in Section 5. For a bounded kernel in combination with a bounded, but not necessarily convex pairwise loss function, we show that RPL methods have a bounded maxbias, see Theorem 5.1. For a bounded continuous kernel in combination with a convex pairwise loss function, which is separately Lipschitz continuous in the sense of Definition 2.5, we can formulate the two other main results of this paper: Theorem 5.3 shows that the RPL operator has a bounded Gâteaux derivative and hence a bounded influence function, see Corollary 5.4, and Theorem 5.5 shows that RPL methods and even their empirical bootstrap approximations are qualitatively robust, if some *non-stochastic* conditions are satisfied. Hence these statistical robustness properties of RPL methods hold for all probability measures provided that weak conditions on the input space, on the output space, on the kernel, and on the loss function are fulfilled. We like to emphasize that all these assumptions are non-stochastic and can therefore easily be checked by the user. All proofs and some technical results are given in the Appendix.

2 Pairwise Loss Functions and Basic Properties

If not otherwise mentioned, we will assume the following setup.

Assumption 2.1. *Let $\mathcal{X}$ be a complete separable metric space and $\mathcal{Y} \subset \mathbb{R}$ be closed. Let (X, Y) and (X_i, Y_i), $i \in \mathbb{N}$, be independent and identically distributed pairs of random quantities with values in*

$\mathcal{X} \times \mathcal{Y}$. *We denote the joint distribution of* (X_i, Y_i) *by* $\mathrm{P} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$, *where* $\mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ *is the set of all Borel probability measures on the Borel* σ*-algebra* $\mathcal{B}_{\mathcal{X}\times\mathcal{Y}}$.

As usual we will denote the realisations of (X_i, Y_i) by (x_i, y_i). For a given data set $D_n = \big((x_1, y_1), \ldots, (x_n, y_n)\big) \in (\mathcal{X} \times \mathcal{Y})^n$ we denote the empirical distribution by $\mathrm{D}_n = \frac{1}{n}\sum_{i=1}^n \delta_{(x_i,y_i)}$. Furthermore, we write $\mathbb{P} := \mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{(X_i,Y_i)}$. Hence $\mathbb{P}_n(\omega) = \mathrm{D}_n$ for the realisations $(X_i(\omega), Y_i(\omega)) = (x_i, y_i)$, $i = 1, \ldots, n$.

Leading examples are of course $\mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \{-1, +1\}$ for binary classification and $\mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \mathbb{R}$ for regression, where d is some positive integer.

For $\mathrm{P} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$, we denote the marginal distribution of X by $\mathrm{P}_\mathcal{X}$ and the conditional probability of Y given $X = x$ by $\mathrm{P}(y|x)$. The n-fold product measure of P is denoted by $\mathrm{P} \otimes \ldots \otimes \mathrm{P}$ or simply by P^n.

The classical definition of a loss function in the machine learning literature is a measurable function from $\mathcal{X} \times \mathcal{Y} \times \mathbb{R}$ to $[0, \infty)$ and one goal is to minimize the expected loss plus some regularization term over a hypothesis space, which is often a reproducing kernel Hilbert space (RKHS), say H, defined implicitly by a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, see e.g. Vapnik (1995, 1998), Poggio and Girosi (1998), Wahba (1999), Schölkopf and Smola (2002), Cucker and Zhou (2007), Steinwart and Christmann (2008) and the references cited there.

Here we consider the case that minimizing a regularized risk, where the loss function for pairwise learning has six instead of three arguments. I.e., we wish to find a measurable function $f : \mathcal{X} \to \mathbb{R}$ such that for some non-negative loss function L the value $L(x, y, \tilde{x}, \tilde{y}, f(x), f(\tilde{x}))$ is small, if the pair $(f(x), f(\tilde{x}))$ is a good prediction for the pair $(y, \tilde{y})$. The close connection to V-statistics and U-statistics (both of degree 2) is obvious, see e.g. Serfling (1980, p. 172-174) and Koroljuk and Borovskich (1994).

Definition 2.2. *Let* $(\mathcal{X}, \mathcal{A})$ *be a measurable space and* $Y \subset \mathbb{R}$ *be closed. Then a function*

$$L : (\mathcal{X} \times \mathcal{Y})^2 \times \mathbb{R}^2 \to [0, \infty) \tag{2.1}$$

is called a ***pairwise loss function****, or simply a* ***pairwise loss****, if it is measurable. A pairwise loss* L *is* ***represented by*** ρ*, if* $\rho : \mathbb{R} \to [0, \infty)$ *is a measurable function and, for all* $(x, y) \in \mathcal{X} \times \mathcal{Y}$*, for all* $(\tilde{x}, \tilde{y}) \in \mathcal{X} \times \mathcal{Y}$*, and for all* $t, \tilde{t} \in \mathbb{R}$*,*

$$L(x, y, \tilde{x}, \tilde{y}, t, \tilde{t}) := \rho\big((y - t) - (\tilde{y} - \tilde{t})\big). \tag{2.2}$$

In the following, we will interpret $L(x, y, \tilde{x}, \tilde{y}, f(x), f(\tilde{x}))$ as the *loss* when we predict $(f(x), f(\tilde{x}))$ if $(x, \tilde{x})$ is observed, but the true outcome is $(y, \tilde{y})$. The smaller the value $L(x, y, \tilde{x}, \tilde{y}, f(x), f(\tilde{x}))$ is, the better $(f(x), f(\tilde{x}))$ predicts $(y, \tilde{y})$ by means of L. Hence constant loss functions, such as $L := 0$, are rather meaningless for our purposes, since they do not distinguish between good and bad predictions. Therefore, we only consider non-constant pairwise loss functions.

Let us now recall from the introduction that our major goal is to have a small *average* loss for future unseen observations (x, y). This leads to the following definition.

Definition 2.3. *Let* L *be a pairwise loss function and* $\mathrm{P} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$.

(*i*) *The* L***-risk*** *for a measurable function* $f : \mathcal{X} \to \mathbb{R}$, *i.e.* $f \in \mathcal{L}_0(\mathcal{X})$, *is defined by*

$$\mathcal{R}_{L,\mathrm{P}}(f) := \int_{(\mathcal{X}\times\mathcal{Y})^2} L\big(x, y, \tilde{x}, \tilde{y}, f(x), f(\tilde{x})\big)\, d\mathrm{P}^2(x, y, \tilde{x}, \tilde{y}). \tag{2.3}$$

(*ii*) *The minimal* L*-risk*

$$\mathcal{R}^*_{L,\mathrm{P}} := \inf\big\{\mathcal{R}_{L,\mathrm{P}}(f)\,;\, f : (\mathcal{X}, \mathcal{A}) \to (\mathbb{R}, \mathcal{B}) \text{ measurable}\big\} \tag{2.4}$$

*is called the **Bayes risk** with respect to* P *and* L*. In addition, a measurable function* $f^*_{L,\mathrm{P}} : \mathcal{X} \to \mathbb{R}$ *with* $\mathcal{R}_{L,\mathrm{P}}(f^*_{L,\mathrm{P}}) = \mathcal{R}^*_{L,\mathrm{P}}$ *is called a **Bayes decision function**.*

If f is measurable, then the function $(x, \tilde{x}, y, \tilde{y}) \mapsto L(x, y, \tilde{x}, \tilde{y}, f(x), f(\tilde{x}))$ is measurable and $\mathcal{R}_{L,\mathrm{P}}(f) \in [0, \infty]$.

If $\mathcal{X}$ is a Polish space and $\mathcal{Y} \subset \mathbb{R}$ is closed, then $\mathcal{X} \times \mathcal{Y}$ is a Polish space. Hence we can split up P into the regular conditional probability $\mathrm{P}(dy|x)$ and the marginal distribution $\mathrm{P}_{\mathcal{X}}$, *cf.* Dudley (2002, Section 10.2). If we combine this with the Tonelli-Fubini theorem, we can write $\mathcal{R}_{L,\mathrm{P}}(f)$ as

$$\int_{\mathcal{X}} \int_{\mathcal{X}} \int_{\mathcal{Y}} \int_{\mathcal{Y}} L\big(x, y, \tilde{x}, \tilde{y}, f(x), f(\tilde{x})\big)\, \mathrm{P}(dy|x)\, \mathrm{P}(d\tilde{y}|\tilde{x})\, \mathrm{P}_{\mathcal{X}}(dx)\, \mathrm{P}_{\mathcal{X}}(d\tilde{x}). \tag{2.5}$$

For a given sequence $D := D_n := ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$, we denote by $\mathrm{D} := \mathrm{D}_n := \frac{1}{n}\sum_{i=1}^n \delta_{(x_i, y_i)}$ the empirical measure associated to the data set D. The risk of a function $f : \mathcal{X} \to \mathbb{R}$ with respect to D is called the **empirical L-risk**

$$\mathcal{R}_{L,\mathrm{D}}(f) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n L\big(x_i, x_j, y_i, y_j, f(x_i), f(x_j)\big)\,. \tag{2.6}$$

Analogously to (2.6) one can also define modifications in which one only adds terms in (2.6) over all pairs (i, j) with $i \neq j$ or over all pairs (i, j) with $i \leq j$. Here we will not investigate these modifications.

We will now introduce some useful properties of pairwise loss functions and their risks in a similar way as for classical loss functions. The first step is of course measurability.

Lemma 2.4 (Measurability of risks)**.** *Let L be a pairwise loss and $\mathcal{F} \subset \mathcal{L}_0(\mathcal{X})$ be a subset that is equipped with a complete and separable metric d and its corresponding Borel σ-algebra. Assume that the metric d dominates the pointwise convergence, i.e., $\lim_{n\to\infty} d(f_n, f) = 0$ implies $\lim_{n\to\infty} f_n(x) = f(x)$ for all $x \in \mathcal{X}$ and for all $f, f_n \in \mathcal{F}$. Then the evaluation map $\mathcal{F} \times \mathcal{X} \to \mathbb{R}$ defined by $(f, x) \mapsto f(x)$ is measurable, and consequently the map $(x, y, \tilde{x}, \tilde{y}, f) \mapsto L(x, y, \tilde{x}, \tilde{y}, f(x), f(\tilde{x}))$ defined on $(\mathcal{X} \times \mathcal{Y})^2 \times \mathcal{F}$ and the map $(x, y, \tilde{x}, \tilde{y}, f, f) \mapsto L(x, y, \tilde{x}, \tilde{y}, f(x), f(\tilde{x}))$ defined on $(\mathcal{X} \times \mathcal{Y})^2 \times \mathcal{F}^2$ are also measurable. Finally, given $\mathrm{P} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$, the risk functional $\mathcal{R}_{L,\mathrm{P}} : \mathcal{F} \to [0, \infty]$ is measurable.*

Obviously, the metric defined by the supremum norm $\|\cdot\|_\infty$ dominates the pointwise convergence for every $\mathcal{F} \subset C(\mathcal{X}) \cap \mathcal{L}_\infty(\mathcal{X})$. It is well-known that the metric of reproducing kernel Hilbert spaces (RKHSs) also dominates the pointwise convergence.

Definition 2.5. *A pairwise loss L is called*

(*i*) ***(strictly) convex**, **continuous**, or **differentiable**, if*

$$L(x, y, \tilde{x}, \tilde{y}, \cdot\,, \cdot\,) : \mathbb{R}^2 \to [0, \infty)$$

is (strictly) convex, continuous, or (total) differentiable for all $(x, y, \tilde{x}, \tilde{y}) \in (\mathcal{X} \times \mathcal{Y})^2$, respectively.

(*ii*) ***locally separately Lipschitz continuous**, if for all $b \geq 0$ there exists a constant $c_b \geq 0$ such that, for all $t, \tilde{t}, t', \tilde{t}' \in [-b, b]$, we have*

$$\sup_{\substack{x,\tilde{x}\in\mathcal{X}\\ y,\tilde{y}\in\mathcal{Y}}} \big|L(x, y, \tilde{x}, \tilde{y}, t, \tilde{t}) - L(x, y, \tilde{x}, \tilde{y}, t', \tilde{t}')\big| \leq c_b \left(|t - t'| + |\tilde{t} - \tilde{t}'|\right). \tag{2.7}$$

Moreover, for $b \geq 0$, the smallest such constant c_b is denoted by $|L|_{b,1}$. Furthermore, L is called ***separately Lipschitz continuous***[1]*, if there exists a constant $|L|_1 \in [0,\infty)$ such that, for all $t, \tilde{t}, t', \tilde{t}' \in \mathbb{R}$, (2.7) is satisfied, if we replace c_b by $|L|_1$.*

If L is differentiable, we denote by $DL(x, y, \tilde{x}, \tilde{y}, t, \tilde{t})$ the (total) derivative of $L(x, y, \tilde{x}, \tilde{y}, \cdot, \cdot)$ at $(t, \tilde{t}) \in \mathbb{R}^2$. If $L(x, y, \tilde{x}, \tilde{y}, \cdot, \cdot)$ is differentiable with respect to the 5th or 6t argument, we denote the corresponding partial derivative by $D_5L(x, y, \tilde{x}, \tilde{y}, \cdot, \cdot)$ and $D_6L(x, y, \tilde{x}, \tilde{y}, \cdot, \cdot)$, respectively.

Theorem 5.3 and Theorem 5.5 show that separate Lipschitz continuity and bounded derivatives are key properties of pairwise loss functions to achieve a RPL method with good robustness properties.

In the following we often need that the risk functional is convex to achieve uniqueness of the estimator. This can easily be achieved by the following result.

Lemma 2.6 (Convexity of risks)**.** *Let L be a (strictly) convex pairwise loss and $\mathrm{P} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$. Then $\mathcal{R}_{L,\mathrm{P}} : \mathcal{L}_0(\mathcal{X}) \to [0,\infty]$ is (strictly) convex.*

We also need some additional relationships between a pairwise loss function and its risk. Such relationships are of course well-known for standard loss functions, see e.g. Steinwart and Christmann (2008).

Lemma 2.7 (Lipschitz continuity of risks)**.** *Let $\mathrm{P} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ and L be a locally separately Lipschitz continuous pairwise loss. Then for all $B \geq 0$ and all $f, g \in L_\infty(\mathrm{P}_\mathcal{X})$ with $\|f\|_\infty \leq B$ and $\|g\|_\infty \leq B$, we have*

$$\bigl|\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{P}}(g)\bigr| \leq 2\,|L|_{B,1} \cdot \|f - g\|_{L_1(\mathrm{P}_\mathcal{X})}\,.$$

Furthermore, the risk functional $\mathcal{R}_{L,\mathrm{P}} : L_\infty(\mathrm{P}_\mathcal{X}) \to [0,\infty)$ is well-defined and continuous.

In general, we can not expect that the risk of a differentiable loss function is differentiable. In this paper we are mainly interested in convex, separately Lipschitz continuous and differentiable pairwise loss functions, for which all partial derivatives (up to order one or two) are continuous and uniformly bounded. Such loss functions yield several desirable statistical robustness properties of the learning methods, as we aim to show. However, we conjecture that similar results can be shown for certain integrable Nemitski losses, see e.g. Steinwart and Christmann (2008, Lem. 2.21) for a result for standard loss functions from $X \times Y \times \mathbb{R} \to [0,\infty)$.

Lemma 2.8 (Differentiability of risks)**.** *Let $\mathrm{P} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ and L be a differentiable pairwise loss such that, for all $x, \tilde{x} \in \mathcal{X}$ and all $y, \tilde{y} \in \mathcal{Y}$, the partial derivatives $D_iL(x, y, \tilde{x}, \tilde{y}, \cdot, \cdot)$, $i \in \{5, 6\}$, are continuous and uniformly bounded by some constant $c_L \in [0,\infty)$. Then the risk functional $\mathcal{R}_{L,\mathrm{P}} : L_\infty(\mathrm{P}_\mathcal{X}) \to [0,\infty)$ is Fréchet differentiable and its derivative at $f \in L_\infty(\mathrm{P}_\mathcal{X})$ is the bounded linear operator $\mathcal{R}'_{L,\mathrm{P}}(f) : L_\infty(\mathrm{P}_\mathcal{X}) \to \mathbb{R}$, where $\mathcal{R}'_{L,\mathrm{P}}(f)g$ equals*

$$\int_{(\mathcal{X}\times\mathcal{Y})^2} D_5L(x, y, \tilde{x}, \tilde{y}, f(x), f(\tilde{x}))g(x) + D_6L(x, y, \tilde{x}, \tilde{y}, f(x), f(\tilde{x}))g(\tilde{x})d\mathrm{P}^2(x, y, \tilde{x}, \tilde{y})\,.$$

Let us now consider a few examples of pairwise loss functions.

(i) Our leading example for the *non-convex case* is the **minimum error entropy (MEE) loss.** Fix $h \in (0,\infty)$. Define the pairwise loss

$$L_{MEE}(x, y, \tilde{x}, \tilde{y}, t, \tilde{t}) := \rho_{MEE}(u) := 1 - \exp\bigl(-u^2/(2h^2)\bigr), \tag{2.8}$$

where $u = (y - t) - (\tilde{y} - \tilde{t}) \in \mathbb{R}$, see e.g. Hu *et al.* (2013), Fan *et al.* (2016), and Feng *et al.* (2015). Some easy calculations show that the first two derivatives ρ' and ρ'' are continuous and bounded. However, ρ_{MEE} is not convex and therefore L_{MEE} is not a convex pairwise loss.

[1] We mention that Rio (2013) used the related term "separately 1-Lipschitz" in a different context.

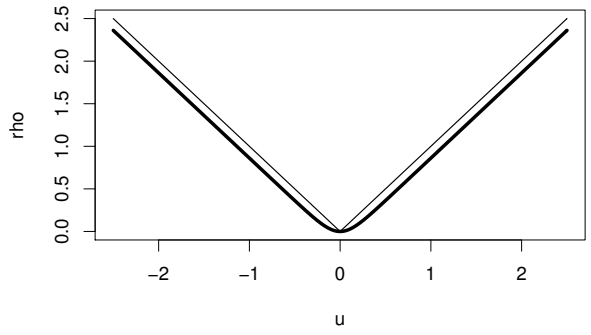

Figure 1: Comparison of ρ_a with $a = 0.01$ (thick curve) with the absolute value (thin curve).

(ii) Our leading example for the *convex case* is the **logistic pairwise loss.** Fix some $a \in (0, \infty)$, e.g. $a = 0.01$ or a equals the rounding precision of the observations. Denote the cumulative distribution function of the logistic distribution by $\Lambda(r) := 1/[1 + \exp(-r)]$, $r \in \mathbb{R}$. Define the pairwise logistic loss

$$L_a(x, y, \tilde{x}, \tilde{y}, t, \tilde{t}) := \rho_a(u) := u - 2a \log\bigl(2\Lambda(u/a)\bigr), \tag{2.9}$$

where $u = (y - t) - (\tilde{y} - \tilde{t}) \in \mathbb{R}$, see also Figure 1. Some easy calculations show that ρ_a is Lipschitz continuous with Lipschitz constant 1 and ρ_a' and ρ_a'' are continuous and bounded. Furthermore, L_a is a convex, continuous, differentiable, and separately Lipschitz continuous pairwise loss with $|L_a|_1 = 1$ and

$$\sup_{(x,y,\tilde{x},\tilde{y}) \in (\mathcal{X} \times \mathcal{Y})^2} |D_i L_a(x, y, \tilde{x}, \tilde{y}, \cdot, \cdot)| \;\leq\; 1 \tag{2.10}$$

$$\sup_{(x,y,\tilde{x},\tilde{y}) \in (\mathcal{X} \times \mathcal{Y})^2} |D_i D_j L_a(x, y, \tilde{x}, \tilde{y}, \cdot, \cdot)| \;\leq\; \frac{1}{2a}\,, \tag{2.11}$$

where $i, j \in \{5, 6\}$. All partial derivatives of L_a up to order two (w.r.t. the last two arguments) are continuous and bounded. Therefore, Lemma 2.8 yields that the risk functional $\mathcal{R}_{L,\mathrm{P}} : L_\infty(\mathrm{P}_\mathcal{X}) \to [0, \infty)$ is Fréchet differentiable and its derivative at $f \in L_\infty(\mathrm{P}_\mathcal{X})$ is the bounded linear operator $\mathcal{R}'_{L,\mathrm{P}}(f) : L_\infty(\mathrm{P}_\mathcal{X}) \to \mathbb{R}$, where $\mathcal{R}'_{L,\mathrm{P}}(f)g$ is given by

$$\int_{(\mathcal{X} \times \mathcal{Y})^2} \left(1 - 2\Lambda\Bigl(\frac{(y - f(x)) - (y - f(\tilde{x}))}{\varepsilon}\Bigr)\right) \cdot \bigl(g(x) - g(\tilde{x})\bigr)\, d\mathrm{P}^2(x, y, \tilde{x}, \tilde{y})\,,$$

where $g \in L_\infty(\mathrm{P}_\mathcal{X})$. Since $1 - 2\Lambda(r) \in (-1, +1)$ for all $r \in \mathbb{R}$, we immediately obtain

$$|\mathcal{R}'_{L,\mathrm{P}}(f)g| \leq 2\|g\|_\infty\,, \qquad g \in L_\infty(\mathrm{P}_\mathcal{X}).$$

(iii) The **squared pairwise loss** L_{LS} is represented by ρ_{LS}, where $\rho_{LS}(u) = u^2$, $u \in \mathbb{R}$. Obviously, ρ_{LS} is only locally Lipschitz continuous and ρ'_{LS} and ρ''_{LS} are continuous. However, ρ'_{LS} is unbounded. Hence L_{LS} is a convex, continuous, and differentiable pairwise loss, but

$$\sup_{x,\tilde{x} \in \mathcal{X},\ y,\tilde{y} \in \mathcal{Y},\ t,\tilde{t} \in \mathbb{R}} |D_i L_{LS}(x, y, \tilde{x}, \tilde{y}, t, \tilde{t})| = \infty\,, \quad i \in \{5, 6\}, \tag{2.12}$$

if $\mathcal{Y} = \mathbb{R}$. This is in contrast to the separately Lipschitz continuous pairwise loss L_a, as the previous example showed.

(iv) **Ranking loss**. Many ranking algorithms can be induced by a pairwise loss of the form $L(x,y,\tilde{x},\tilde{y},t,\tilde{t}) := \ell(t-\tilde{t},y-\tilde{y})$ with a bivariate function $\ell : \mathbb{R}^2 \to [0,\infty)$. See e.g. Agarwal and Niyogi (2009), the **hinge ranking loss** by $L(x,y,\tilde{x},\tilde{y},t,\tilde{t}) = \max\{0,v\}$ and the **least squares ranking loss** by $L(x,y,\tilde{x},\tilde{y},t,\tilde{t}) = v^2$, where $v := |y-\tilde{y}| - (t-\tilde{t})\operatorname{sign}(y-\tilde{y})$. The hinge ranking loss is not differentiable and the least squares ranking loss is not separately Lipschitz continuous in the sense of Definition 2.5. In contrast, the **logistic ranking loss**, which we define by $L(x,y,\tilde{x},\tilde{y},t,\tilde{t}) = \rho_a(v)$ by using the ρ_a function from (2.9) for some $a > 0$, is a separately Lipschitz continuous, differentiable pairwise loss function with bounded first and second order partial derivatives w.r.t. the last two arguments.

(v) **Similarity loss**. Some distance metric or similarity learning algorithms for $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} := \{-1,+1\}$ can be induced by a pairwise loss of the form $L(x,y,\tilde{x},\tilde{y},t,\tilde{t}) = \ell((x-\tilde{x})^T A(x-\tilde{x}), r(y,\tilde{y}))$ or $L(x,y,\tilde{x},\tilde{y},t,\tilde{t}) = \ell(x^T A\tilde{x}, r(y,\tilde{y}))$ with a positive semidefinite symmetric matrix $A \in \mathbb{R}^{d\times d}$ and $\ell : \mathbb{R}^2 \to [0,\infty)$, $r : \mathbb{R}^2 \to \mathbb{R}$. We refer to Cao *et al.* (2015) for the **hinge similarity loss** given by $L(x,y,\tilde{x},\tilde{y},t,\tilde{t}) := \max\{0,1-w\}$, where $w := y\tilde{y}x^{\mathsf{T}}A\tilde{x}$. We can define a smoothed version as the **logistic similarity loss** by $L(x,y,\tilde{x},\tilde{y},t,\tilde{t}) := \ln\bigl(1+\exp(-w)\bigr)$.

3 Regularized Pairwise Learning Methods

Definition 3.1. *Let L be a pairwise loss, H be the RKHS of a measurable kernel on $\mathcal{X}$, and $\lambda > 0$. For $f \in H$, define the regularized risk by $\mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(f) = \mathcal{R}_{L,\mathrm{P}}(f) + \lambda\|f\|_H^2$. A function $f_{L,\mathrm{P},\lambda} \in H$ which satisfies*

$$\mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(f_{L,\mathrm{P},\lambda}) = \inf_{f\in H} \mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(f) \tag{3.1}$$

*is called a **regularized pairwise learning (RLP) method**.*

If $f_{L,\mathrm{P},\lambda}$ exists, we have

$$\lambda\|f_{L,\mathrm{P},\lambda}\|_H^2 \le \mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(f_{L,\mathrm{P},\lambda}) \le \mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(0) = \mathcal{R}_{L,\mathrm{P}}(0)\,, \tag{3.2}$$

or in other words

$$\|f_{L,\mathrm{P},\lambda}\|_H \le \sqrt{\frac{\mathcal{R}_{L,\mathrm{P}}(0)}{\lambda}}\,. \tag{3.3}$$

Let us now investigate under which assumptions there exists an $f_{L,\mathrm{P},\lambda} \in H$ and when it is unique.

Assumption 3.2. *Let $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ be a **continuous and bounded kernel** with reproducing kernel Hilbert space H and define $\|k\|_\infty := \sup_{x\in\mathcal{X}} \sqrt{k(x,x)} \in (0,\infty)$. Denote the canonical feature map by $\Phi(x) := k(\cdot,x)$, $x \in \mathcal{X}$.*

It is well-known that if k is a continuous kernel defined on a Polish space, then Φ is continuous, too. This assumption on the kernel is fulfilled e.g., if $\mathcal{X} = \mathbb{R}^d$ and k is a Gaussian RBF-kernel $k(x,x') := \exp(-\|x-x'\|_2^2/\gamma)$, an Abel RBF-kernel $k(x,x') := \exp(-\|x-x'\|_1/\gamma)$, where $\gamma > 0$, or a compactly supported kernel, see e.g. Wu (1995) and Wendland (1995).

The following facts, which we need later on, are well-known for any bounded kernel k on $\mathcal{X}$ with RKHS H, all $f \in H$, and all $x \in \mathcal{X}$:

$$\begin{aligned}
\langle f, \Phi(x)\rangle_H &= f(x)\,, && (3.4)\\
\|f\|_\infty &\le \|k\|_\infty \cdot \|f\|_H\,, && (3.5)\\
\|\Phi(x)\|_H &= \sqrt{k(x,x)} \;\le\; \|k\|_\infty\,, && (3.6)\\
\|\Phi(x)\|_\infty &\le \|k\|_\infty \cdot \|\Phi(x)\|_H \;\le\; \|k\|_\infty^2\,. && (3.7)
\end{aligned}$$

Theorem 3.3 (Existence)**.** *If L is a separately Lipschitz continuous pairwise loss function, $\mathrm{P} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$, $\mathcal{R}_{L,\mathrm{P}}(f_0) < \infty$ for some $f_0 \in H$, and H is an RKHS with bounded measurable kernel k on $\mathcal{X}$, then a minimizer $f_{L,\mathrm{P},\lambda} \in H$ exists for any $\lambda > 0$.*

Lemma 3.4 (Uniqueness)**.** *Let $\mathrm{P} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$, L be a convex pairwise loss with $\mathcal{R}_{L,\mathrm{P}}(f_0) < \infty$ for* some *$f_0 \in H$, and H be the RKHS of a measurable kernel over $\mathcal{X}$. Then for all $\lambda > 0$ there exists at most one $f_{L,\mathrm{P},\lambda}$.*

Theorem 3.5 (Existence)**.** *Let $\mathrm{P} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$, L be a convex, locally separately Lipschitz continuous pairwise loss function with $\mathcal{R}_{L,\mathrm{P}}(0) < \infty$, and H be the RKHS of a bounded measurable kernel over $\mathcal{X}$. Then, for all $\lambda > 0$, there exists $f_{L,\mathrm{P},\lambda}$.*

Corollary 3.6 (Existence and Uniqueness)**.** *Let $\mathrm{P} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$, L be a convex, separately Lipschitz continuous pairwise loss function with $\mathcal{R}_{L,\mathrm{P}}(0) < \infty$, and H be the RKHS of a bounded measurable kernel over $\mathcal{X}$. Then, for all $\lambda > 0$, there exists a uniquely defined $f_{L,\mathrm{P},\lambda} \in H$ and*

$$\|f_{L,\mathrm{P},\lambda}\|_H \leq \left(\mathcal{R}_{L,\mathrm{P}}(0)/\lambda\right)^{1/2}. \tag{3.8}$$

Obviously, we would like to get rid of the moment assumption $\mathcal{R}_{L,\mathrm{P}}(0) < \infty$, because otherwise we can not define $f_{L,\mathrm{P},\lambda}$ on $\mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ for arbitrary input and output spaces. Although an assumption like $\mathcal{R}_{L,\mathrm{P}}(0) < \infty$ is quite common in machine learning, we like to emphasize that the validity of this condition can in general *not* be checked, because P is completely unknown. E.g., let $\mathcal{Y} = \mathbb{R}$ and $L = L_{LS}$. Assume that

$$\mathcal{R}_{L_{LS},\mathrm{P}}(0) := \int_{(\mathcal{X}\times\mathcal{Y})^2} (y - \tilde{y})^2 \, d\mathrm{P}^2(x, y, \tilde{x}, \tilde{y}) < \infty.$$

It is easy to see that, for any $\varepsilon \in (0, 1)$, there exists a mixture distribution $(1-\varepsilon)\mathrm{P} + \varepsilon\tilde{\mathrm{P}}$ such that

$$\mathcal{R}_{L_{LS},(1-\varepsilon)\mathrm{P}+\varepsilon\tilde{\mathrm{P}}}(0) = \infty,$$

e.g., let the conditional distribution $\tilde{\mathrm{P}}(Y|X = x)$ of Y given $X = x$ be a Cauchy distribution. If ε is small enough, say $\varepsilon = 10^{-10}$ or $\varepsilon = 10^{-100}$, we will never be able to check whether the data were generated by P or by $(1-\varepsilon)\mathrm{P} + \varepsilon\tilde{\mathrm{P}}$.

The idea to shift the loss function by an appropriate function which is independent of the last argument(s) of L is useful and it was already used e.g. by Huber (1967) for M-estimators and by Christmann *et al.* (2009) for support vector machines based on a general loss function $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ and on a general kernel.

Let L be a pairwise loss function and define the corresponding **shifted pairwise loss function** (or simply the shifted version of L) by

$$L^\star : (\mathcal{X} \times \mathcal{Y})^2 \times \mathbb{R}^2 \to \mathbb{R}, \tag{3.9}$$

$$L^\star(x, y, \tilde{x}, \tilde{y}, t, \tilde{t}) := L(x, y, \tilde{x}, \tilde{y}, t, \tilde{t}) - L(x, y, \tilde{x}, \tilde{y}, 0, 0). \tag{3.10}$$

We adopt the definitions of continuity, (locally) separately Lipschitz continuity, and differentiability of $L^\star$ from the same definitions for L, i.e. these properties are meant to be valid for the last two arguments, when the first four arguments are arbitrary but fixed. In the same manner we define the $L^\star$-risk, the regularized $L^\star$-risk, and the RPL method based on $L^\star$ by

$$\mathcal{R}_{L^\star,\mathrm{P}}(f) := \mathbb{E}_{\mathrm{P}^2} L^\star(X, Y, \tilde{X}, \tilde{Y}, f(X), f(\tilde{X})) \tag{3.11}$$

$$\mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f) := \mathcal{R}_{L^\star,\mathrm{P}}(f) + \lambda\|f\|_H^2 \tag{3.12}$$

$$f_{L^\star,\mathrm{P},\lambda} := \arg\inf_{f\in H} \mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f), \tag{3.13}$$

respectively. There exists a strong connection between L and $L^\star$ in terms of convexity and separate Lipschitz continuity and also for the corresponding risks, see Lemma 7.8 to Lemma 7.11 in the Appendix.

Of course, shifting the loss function L to $L^\star$ changes the objective function, but the *minimizers* of $\mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(\cdot)$ and $\mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(\cdot)$ coincide for those $\mathrm{P} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ for which $\mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(\cdot)$ has a minimizer in H. I.e., we have

$$f_{L^\star,\mathrm{P},\lambda} = f_{L,\mathrm{P},\lambda}, \qquad \text{if } f_{L,\mathrm{P},\lambda} \in H \text{ exists.} \tag{3.14}$$

Furthermore, (3.14) is valid for all empirical distributions D based on a data set consisting of n data points (x_i, y_i), $1 \le i \le n$, because $f_{L,\mathrm{D},\lambda}$ exists and is unique since $\mathcal{R}_{L,\mathrm{D}}(0) < \infty$.

Let us now show that shifting a pairwise loss function indeed helps to get rid of the moment assumption $\mathcal{R}_{L,\mathrm{P}}(0) < \infty$ which was essential for Lemma 3.4 and Theorem 3.5. Assume that L is a separately Lipschitz continuous pairwise loss. Then we obtain, for all $f \in L_1(\mathrm{P}_\mathcal{X})$,

$$\begin{aligned}
\mathcal{R}_{L^\star,\mathrm{P}}(f) &= \mathbb{E}_{\mathrm{P}^2}\big(L(X,Y,\tilde{X},\tilde{Y},f(X),f(\tilde{X})) - L(X,Y,\tilde{X},\tilde{Y},0,0)\big) \\
&\le \int_{(\mathcal{X}\times\mathcal{Y})^2} |L(x,y,\tilde{x},\tilde{y},f(x),f(\tilde{x})) - L(x,y,\tilde{x},\tilde{y},0,0)|\, d\mathrm{P}^2(x,y,\tilde{x},\tilde{y}) \\
&\le |L|_1 \int_{\mathcal{X}^2} \big(|f(x)-0| + |f(\tilde{x})-0|\big)\, d\mathrm{P}^2_\mathcal{X}(x,\tilde{x}) \\
&\le 2\,|L|_1 \|f\|_{L_1(\mathrm{P}_\mathcal{X})} \;<\; \infty\,,
\end{aligned} \tag{3.15}$$

without assuming $\mathcal{R}_{L,\mathrm{P}}(0) < \infty$. The assumption $f \in L_1(\mathrm{P}_\mathcal{X})$ can easily be satisfied by choosing a *bounded* kernel k, because then all $f \in H$ are bounded due to $\|f\|_\infty \le \|k\|_\infty \|f\|_H$, see (3.5). Therefore, taking (3.14) into account, the use of a shifted loss function just enlarges the set of probability measures where the minimizer of the regularized risk is well-defined.

Theorem 3.7 (Uniqueness of $f_{L^\star,\mathrm{P},\lambda}$)**.** *Let L be a convex pairwise loss, H be the RKHS of a measurable kernel over $\mathcal{X}$, and $\mathrm{P} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$. Assume that (i) $\mathcal{R}_{L^\star,\mathrm{P}}(f_0) < \infty$ for* some *$f_0 \in H$ and $\mathcal{R}_{L^\star,\mathrm{P}}(f) > -\infty$ for* all *$f \in H$ or (ii) L is separately Lipschitz continuous and $f \in L_1(\mathrm{P}_\mathcal{X})$ for all $f \in H$. Then, for all $\lambda > 0$, there exists at most one decision function $f_{L^\star,\mathrm{P},\lambda}$.*

Theorem 3.8 (Existence and Uniqueness of $f_{L^\star,\mathrm{P},\lambda}$)**.** *Let L be a convex, separately Lipschitz continuous pairwise loss, H be the RKHS of a bounded measurable kernel k, and $\mathrm{P} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$. Then, for all $\lambda > 0$, there exists a unique decision function $f_{L^\star,\mathrm{P},\lambda}$.*

4 Representer Theorem for RPL Methods

In this section we establish a representer theorem for a general probability measure P. This result is interesting in its own right, but also useful to prove several statistical robustness properties of RPL methods.

Assumption 4.1. *Let L be a* ***separately Lipschitz-continuous***, ***differentiable*** *pairwise loss function for which all partial derivatives up to order 2 with respect to the last two arguments are* ***continuous*** *and* ***uniformly bounded*** *in the sense that there exist constants $c_{L,1} \in (0,\infty)$ and $c_{L,2} \in (0,\infty)$ with*

$$\sup_{x,\tilde{x}\in\mathcal{X},\ y,\tilde{y}\in\mathcal{Y}} |D_i L(x,y,\tilde{x},\tilde{y},\cdot,\cdot)| \;\le\; c_{L,1}\,, \qquad i \in \{5,6\}, \tag{4.1}$$

$$\sup_{x,\tilde{x}\in\mathcal{X},\ y,\tilde{y}\in\mathcal{Y}} |D_i D_j L(x,y,\tilde{x},\tilde{y},\cdot,\cdot)| \;\le\; c_{L,2}\,, \qquad i,j \in \{5,6\}. \tag{4.2}$$

Additionally, assume that

$$L(x,y,x,y,t,t) \equiv 0, \qquad \forall\ (x,y,t) \in \mathcal{X} \times \mathcal{Y} \times \mathbb{R}. \tag{4.3}$$

Of course, L_{LS} does not satisfy (4.1), whereas e.g. L_a fulfills Assumption 4.1. The assumption (4.3) is quite plausible and is satisfied for many loss functions of practical use, e.g. for L_{MEE}, L_a, L_{LS}, and hinge, least squares, and logistic ranking loss. If a pairwise loss L is represented by ρ, then (4.3) is satisfied if $\rho(0) = 0$. A ranking loss with $\ell(0,0) = 0$ also satisfies (4.3).

Assumption 4.2. *Let L be a convex pairwise loss function.*

We will reconsider these assumptions at the end of Section 5 and it will become clear that these assumptions on L and k are very plausible to guarantee the existence of a *bounded* Gâteaux derivative of the map $\mathrm{P} \mapsto f_{L^\star,\mathrm{P},\lambda}$.

As usual we will denote Bochner integrals of an H-valued function g with respect to some Borel measure μ by $\int g\, d\mu$, we refer to Denkowski *et al.* (2003, p. 365ff). If μ is a probability measure, we denote the Bochner integral occasionally by $\mathbb{E}_\mu[g]$.

Theorem 4.3 (**Representer theorem**). *Let the Assumptions 2.1, 3.2, and 4.1 be valid. Then we have, for all $\mathrm{P} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ and all $\lambda \in (0,\infty)$:*

(*i*) *If $f_{L^\star,\mathrm{P},\lambda} \in H$ is any fixed minimizer of $\min_{f\in H}\left(\mathcal{R}_{L^\star,\mathrm{P}}(f) + \lambda\|f\|_H^2\right)$, then:*

$$f_{L^\star,\mathrm{P},\lambda} = -\frac{1}{2\lambda}\mathbb{E}_{\mathrm{P}^2}\left[h_{5,\mathrm{P}}(X,Y,\tilde{X},\tilde{Y})\Phi(X) + h_{6,\mathrm{P}}(X,Y,\tilde{X},\tilde{Y})\Phi(\tilde{X})\right], \tag{4.4}$$

where $h_{5,\mathrm{P}}$ and $h_{6,\mathrm{P}}$ denote the partial derivatives

$$h_{5,\mathrm{P}}(X,Y,\tilde{X},\tilde{Y}) := D_5 L\big(X,Y,\tilde{X},\tilde{Y},f_{L^\star,\mathrm{P},\lambda}(X),f_{L^\star,\mathrm{P},\lambda}(\tilde{X})\big) \tag{4.5}$$

$$h_{6,\mathrm{P}}(X,Y,\tilde{X},\tilde{Y}) := D_6 L\big(X,Y,\tilde{X},\tilde{Y},f_{L^\star,\mathrm{P},\lambda}(X),f_{L^\star,\mathrm{P},\lambda}(\tilde{X})\big). \tag{4.6}$$

(*ii*) *(Convex Case.) If additionally Assumption 4.2 is valid (and hence $f_{L^\star,\mathrm{P},\lambda}$ uniquely exists by Theorem 3.8), then $f_{L^\star,\mathrm{P},\lambda}$ has the representation (4.4) and we have additionally, for all $\mathrm{Q} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$:*

$$\begin{aligned} &\|f_{L^\star,\mathrm{P},\lambda} - f_{L^\star,\mathrm{Q},\lambda}\|_H \\ \leq\ & \frac{1}{\lambda}\Big\|\mathbb{E}_{\mathrm{P}^2}\left[h_{5,\mathrm{P}}(X,Y,\tilde{X},\tilde{Y})\Phi(X) + h_{6,\mathrm{P}}(X,Y,\tilde{X},\tilde{Y})\Phi(\tilde{X})\right] \\ &\quad - \mathbb{E}_{\mathrm{Q}^2}\left[h_{5,\mathrm{P}}(X,Y,\tilde{X},\tilde{Y})\Phi(X) + h_{6,\mathrm{P}}(X,Y,\tilde{X},\tilde{Y})\Phi(\tilde{X})\right]\Big\|_H. \end{aligned} \tag{4.7}$$

5 Robustness of RPL Methods

In this section we show that an RPL method has several desirable statistical robustness properties, if the pairwise loss function L and the kernel k fulfill some weak assumptions. As these assumptions are independent of P and independent of the data set, these assumptions can be checked in advance. We start with the case of bounded pairwise loss functions. The case of convex pairwise loss functions is investigated in Subsection 5.2.

5.1 Case 1: Non-convex and Bounded Pairwise Loss

The minimizer $f_{L^\star,\mathrm{P},\lambda}$ typically exists, but it is unfortunately in general *not* uniquely defined for *non-convex* pairwise loss functions. However we show in this subsection, that RPL-methods based on a non-convex and bounded pairwise loss function yields a statistically robust approximation of the regularized risk under weak conditions. More precisely, we show that the regularized *risk* functional has a small bias in neighborhoods defined by the norm of total variation, if L is a

bounded, but in general *non-convex* pairwise loss function. This is also valid, if we consider the classical contamination "neighborhoods", see e.g. Huber (1981, p.11). This result indicates that we can expect a bounded influence function for the regularized *risk operator* for non-convex pairwise loss functions under appropriate conditions, provided the influence function exists.

Our leading example for this subsection is the minimum entropy loss L_{MEE}, see (2.8).

In this subsection, let L be a *bounded* pairwise loss function, i.e. we assume

$$L(x, y, \tilde{x}, \tilde{y}, t, \tilde{t}) \in [0, c] \qquad \forall\, (x, y, \tilde{x}, \tilde{y}, t, \tilde{t}) \in (\mathcal{X} \times \mathcal{Y})^2 \times \mathbb{R}^2$$

for some constant $c \in (0, \infty)$. Hence the risk $\mathcal{R}_{L,\mathrm{P}}(f) \in [0, c]$ for all $\mathrm{P} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ and there is no need to consider shifted loss functions.

The norm of total variation of two probability measures $\mathrm{P}, \mathrm{Q} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ is defined by

$$d_{TV}(\mathrm{P}, \mathrm{Q}) := \sup_{A \in \mathcal{B}(\mathcal{X} \times \mathcal{Y})} |\mathrm{P}(A) - \mathrm{Q}(A)| = \frac{1}{2} \sup_h \left| \int h \, d\mathrm{P} - \int h \, d\mathrm{Q} \right|,$$

where the supremum is with respect to all $h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ with $\|h\|_\infty \leq 1$. It is well-known that $d_{TV}(\mathrm{P}, \mathrm{Q}) \in [0, 1]$ for all $\mathrm{P}, \mathrm{Q} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$.

Define the function

$$\begin{aligned} &R^{reg} : \mathcal{M}_1(\mathcal{X} \times \mathcal{Y}) \to [0, \infty], \\ &R^{reg}(\mathrm{P}) := \mathcal{R}_{L,\mathrm{P}}(f_{L,\mathrm{P},\lambda}) = \inf_{f \in H} \mathcal{R}_{L,\mathrm{P}}(f) + \lambda \|f\|_H^2. \end{aligned} \tag{5.1}$$

Recall that the *maximum bias* of R^{reg} is defined by

$$b_1(\varepsilon; \mathrm{P}) := \sup_{\mathrm{Q} \in N(\varepsilon; \mathrm{P})} |R^{reg}(\mathrm{Q}) - R^{reg}(\mathrm{P})|, \quad \varepsilon \in (0, 1), \tag{5.2}$$

where $N(\varepsilon; \mathrm{P})$ denotes an ε-neighborhood of P, see Huber (1981, p.11, (4.5)). Common examples are the *total variation neighborhood*

$$N_{TV}(\varepsilon; \mathrm{P}) := \{\mathrm{Q} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y}); d_{TV}(\mathrm{Q}, \mathrm{P}) \leq \varepsilon\}$$

and the so-called *contamination "neighborhood"*

$$N_{con}(\varepsilon; \mathrm{P}) = \{\mathrm{P}_\varepsilon := (1 - \varepsilon)\mathrm{P} + \varepsilon \bar{\mathrm{P}};\ \bar{\mathrm{P}} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})\}.$$

One desirable property from the viewpoint of robust statistics is a bounded maximum bias for sufficiently large positive values of ε. If two statistical methods have a bounded maximum bias, the one with the smaller maximum bias is considered to be more robust.

Theorem 5.1 (**Bounds for the bias**). *Let $\varepsilon \in (0, 1)$ and $\mathrm{P}, \mathrm{Q} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$. Let L be a bounded pairwise loss function satisfying $L \leq c \in (0, \infty)$. Consider the regularised risk functional R^{reg} defined in (5.1).*

(*i*) *Then*

$$|R^{reg}(\mathrm{Q}) - R^{reg}(\mathrm{P})| \leq c\, d_{TV}(\mathrm{Q}^2, \mathrm{P}^2) \leq 2\, c\, d_{TV}(\mathrm{Q}, \mathrm{P}) \tag{5.3}$$

and an upper bound for the maximum bias over total variation neighbourhoods is given by

$$b_1(\varepsilon; \mathrm{P}) \leq 2c\varepsilon,$$

uniformly for all P.

(*ii*) *If* $\mathrm{P}_\varepsilon = (1-\varepsilon)\mathrm{P} + \varepsilon\bar{\mathrm{P}}$ *for some* $\bar{\mathrm{P}} \in \mathcal{M}_1(\mathcal{X}\times\mathcal{Y})$*, then*

$$|R^{reg}(\mathrm{P}_\varepsilon) - R^{reg}(\mathrm{P})| \leq 2\,c\,d_{TV}(\bar{\mathrm{P}},\mathrm{P})\cdot\varepsilon(1+\varepsilon) \tag{5.4}$$

and the maximum bias over contamination "neighborhoods" satisfies

$$b_1(\varepsilon;\mathrm{P}) \leq 2c\varepsilon(1+\varepsilon) \leq 4c\varepsilon,$$

uniformly for all P *and* $\bar{\mathrm{P}}$*.*

Hence even an upper bound for the maximum bias in total variation or contamination neighborhoods increases at most linearly in ε, uniformly for all P and all $\bar{\mathrm{P}}$. An obvious consequence of the second part of Theorem 5.1 is, that the limit

$$\lim_{\varepsilon\to 0}\frac{R^{reg}(\mathrm{P}_\varepsilon) - R^{reg}(\mathrm{P})}{\varepsilon} \tag{5.5}$$

is bounded by $2\,c\,d_{TV}(\bar{\mathrm{P}},\mathrm{P}) \leq 2c$, provided the limit exists. If we specialize P_ε to $\mathrm{P}_\varepsilon = (1-\varepsilon)\mathrm{P} + \varepsilon\delta_{(x,y)}$ for some $(x,y)\in\mathcal{X}\times\mathcal{Y}$, we obtain immediately from (5.5), that R^{reg} has a *uniformly bounded* influence function in the sense of Hampel (1968, 1974), whenever the influence function exists.

Let us now consider an interesting special case of the previous theorem. Define the discrete probability measures $\mathrm{P} := \frac{1}{n}\sum_{i=1}^n \delta_{(x_i,y_i)}$ and $\mathrm{P}_{n-1} := (n-1)^{-1}\sum_{i=1}^{n-1}\delta_{(x_i,y_i)}$ for given data sets with n and $n-1$ data points $(x_i,y_i)\in\mathcal{X}\times\mathcal{Y}$, respectively. Let $\bar{\mathrm{P}}$ be the Dirac measure $\delta_{(x_0,y_0)}$ at some point $(x_0,y_0)\in\mathcal{X}\times\mathcal{Y}$ and let $\varepsilon := \frac{1}{n}$. Then we obtain

$$\frac{R^{reg}(\mathrm{P}_\varepsilon) - R^{reg}(\mathrm{P})}{\varepsilon} = \frac{R^{reg}\big((1-\frac{1}{n})\mathrm{P}_{n-1} + \frac{1}{n}\delta_{(x_0,y_0)}\big) - R^{reg}(\mathrm{P}_n)}{\frac{1}{n}}.$$

This ratio is the so-called *sensitivity curve* at the point (x_0,y_0), see Tukey (1977) or Hampel *et al.* (1986, p. 93), and is usually denoted by

$$SC_n((x_0,y_0);R^{reg},\mathrm{P}_{n-1})$$

in the literature on robust statistics. It measures the influence which an additional single data point (x_0,y_0) has on the statistical method S, if the original data set contains $n-1$ data points. The influence function can under appropriate assumptions be considered as a finite-sample version of the influence function, see Hampel *et al.* (1986, p. 94). A similar version of the sensitivity curve exists, if we replace one data point from an original data set with n data points. An immediate consequence of part (ii) in Theorem 5.1 is, that the sensitivity curve $SC_n((x_0,y_0);R^{reg},\mathrm{P}_{n-1})$ is *uniformly bounded* by $2c(1+\frac{1}{n})$ for all data sets and any additional data point (x_0,y_0), no matter where (x_0,y_0) is located in $\mathcal{X}\times\mathcal{Y}$. If we are interested in the slightly more general problem how to obtain an upper bound for the influence of ℓ additional data points, we just define $\varepsilon := \frac{\ell}{n}$, $\ell\in\{1,\ldots,\lfloor\frac{n}{2}\rfloor\}$, and use again part (ii) in Theorem 5.1 to obtain a uniform upper bound.

Example 5.2. *Theorem 5.1 is applicable for the non-convex minimum entropy loss* L_{MEE} *represented by* $\rho_{MEE}(u) := 1-\exp\big(-u^2/(2h^2)\big)\in[0,1)$*,* $u\in\mathbb{R}$*, where* $h\in(0,\infty)$*. A division by* ε *shows that the absolute value of these difference quotients are bounded by* 2 *or* $2(1+\varepsilon)$*, respectively, which is an immediate consequence of Theorem 5.1. Let us now additionally assume for the case of contamination "neighborhoods" in Theorem 5.1*(ii)*, that the limit* $\lim_{\varepsilon\searrow 0}\frac{R^{reg}(\mathrm{P}_\varepsilon)-R^{reg}(\mathrm{P})}{\varepsilon}$ *exists, where* $\mathrm{P}_\varepsilon := (1-\varepsilon)\mathrm{P}+\varepsilon\delta_z$ *and* $\delta_{(x,y)}$ *denotes the Dirac measure in* $(x,y)\in\mathcal{X}\times\mathcal{Y}$*. Then this limit equals the influence function of* R^{reg} *at* P *and its absolute value is bounded by* $\lim_{\varepsilon\searrow 0}\frac{2\varepsilon(1+\varepsilon)}{\varepsilon} = 2$*. A bounded influence function is of course highly desirable from a robustness point of view.*

5.2 Case 2: Convex Pairwise Loss

Our most important special case for this subsection is the logistic pairwise loss function L_a, see (2.9).

An immediate consequence of the second part of our representer theorem, see (4.7), is the inequality

$$\|f_{L^\star,\mathrm{P},\lambda} - f_{L^\star,\mathrm{Q},\lambda}\|_H \leq \frac{4}{\lambda} c_{L,1} \|k\|_\infty^2 < \infty, \tag{5.6}$$

which is valid for all $\mathrm{P}, \mathrm{Q} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ and all $\lambda \in (0, \infty)$ if the Assumptions 2.1, 3.2, 4.1, and 4.2 are valid.

The goal of this subsection however is to show that the **RPL operator**

$$S : \mathcal{M}_1(\mathcal{X} \times \mathcal{Y}) \to H, \qquad S(\mathrm{P}) = f_{L^\star,\mathrm{P},\lambda} \tag{5.7}$$

has two additional desirable robustness properties, if some weak and *non-stochastic* conditions on $\mathcal{X}$, $\mathcal{Y}$, L, and k are satisfied:

(i) The following Theorem 5.3 shows that S has a *bounded* Gâteaux derivative for any probability measure P and hence a *bounded* influence function in the sense of Hampel (1968, 1974), see also Hampel *et al.* (1986).

(ii) Theorem 5.5 given below shows that the sequence of RPL estimators $(f_{L^\star,\mathbb{P}_n,\lambda})_{n\in\mathbb{N}}$ are qualitatively robust, which is a kind of equicontinuity described later in more detail. If additionally $\mathcal{X} \times \mathcal{Y}$ is a compact metric space, then even the empirical bootstrap approximations are qualitatively robust.

Please note, that the following results of this subsection are all formulated for $f_{L^\star,\mathrm{P},\lambda}$ and not for $f_{L,\mathrm{P},\lambda}$, because the latter is in general not well-defined for *all* $\mathrm{P} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$, as was explained in Section 3. Please recall the obvious equalities $D_i L(x, y, \tilde{x}, \tilde{y}, t, \tilde{t}) = D_i L^\star(x, y, \tilde{x}, \tilde{y}, t, \tilde{t})$ and $D_i D_j L(x, y, \tilde{x}, \tilde{y}, t, \tilde{t}) = D_i D_j L^\star(x, y, \tilde{x}, \tilde{y}, t, \tilde{t})$ for $i, j \in \{5, 6\}$.

Theorem 5.3 (**Bounded Gâteaux derivative**). *Let the Assumptions 2.1, 3.2, 4.1, and 4.2 be satisfied. Denote the shifted version of L by $L^\star$. Then, for all Borel probability measures $\mathrm{P}, \mathrm{Q} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$, the RPL operator*

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

has a bounded Gâteaux derivative $S'_G(\mathrm{P})$ at P and

$$S'_G(\mathrm{P})(\mathrm{Q}) = -M(\mathrm{P})^{-1} T(\mathrm{Q}; \mathrm{P})\,. \tag{5.8}$$

Here

$$\begin{aligned}
T(\mathrm{Q};\mathrm{P}) \;=\; & -2\,\mathbb{E}_{\mathrm{P}\otimes\mathrm{P}}\Big[D_5 L\big(X, Y, \tilde{X}, \tilde{Y}, f_{L^\star,\mathrm{P},\lambda}(X), f_{L^\star,\mathrm{P},\lambda}(\tilde{X})\big)\Phi(X) \\
& \qquad + D_6 L\big(X, Y, \tilde{X}, \tilde{Y}, f_{L^\star,\mathrm{P},\lambda}(X), f_{L^\star,\mathrm{P},\lambda}(\tilde{X})\big)\Phi(\tilde{X})\Big] \\
& + \;\mathbb{E}_{\mathrm{P}\otimes\mathrm{Q}}\Big[D_5 L\big(X, Y, \tilde{X}, \tilde{Y}, f_{L^\star,\mathrm{P},\lambda}(X), f_{L^\star,\mathrm{P},\lambda}(\tilde{X})\big)\Phi(X) \\
& \qquad + D_6 L\big(X, Y, \tilde{X}, \tilde{Y}, f_{L^\star,\mathrm{P},\lambda}(X), f_{L^\star,\mathrm{P},\lambda}(\tilde{X})\big)\Phi(\tilde{X})\Big] \\
& + \;\mathbb{E}_{\mathrm{Q}\otimes\mathrm{P}}\Big[D_5 L\big(X, Y, \tilde{X}, \tilde{Y}, f_{L^\star,\mathrm{P},\lambda}(X), f_{L^\star,\mathrm{P},\lambda}(\tilde{X})\big)\Phi(X) \\
& \qquad + D_6 L\big(X, Y, \tilde{X}, \tilde{Y}, f_{L^\star,\mathrm{P},\lambda}(X), f_{L^\star,\mathrm{P},\lambda}(\tilde{X})\big)\Phi(\tilde{X})\Big]
\end{aligned}$$

equals the gradient of the regularized risk and

$$\begin{aligned} M(\mathrm{P}) \;=\;& 2\lambda \operatorname{id}_H \\ &+ \mathbb{E}_{\mathrm{P}\otimes\mathrm{P}}\Big[D_5 D_5 L\big(X,Y,\tilde{X},\tilde{Y}, f_{L^\star,\mathrm{P},\lambda}(X), f_{L^\star,\mathrm{P},\lambda}(\tilde{X})\big)\Phi(X)\otimes\Phi(X) \\ &+ \quad D_6 D_5 L\big(X,Y,\tilde{X},\tilde{Y}, f_{L^\star,\mathrm{P},\lambda}(X), f_{L^\star,\mathrm{P},\lambda}(\tilde{X})\big)\Phi(X)\otimes\Phi(\tilde{X}) \\ &+ \quad D_5 D_6 L\big(X,Y,\tilde{X},\tilde{Y}, f_{L^\star,\mathrm{P},\lambda}(X), f_{L^\star,\mathrm{P},\lambda}(\tilde{X})\big)\Phi(\tilde{X})\otimes\Phi(X) \\ &+ \quad D_6 D_6 L\big(X,Y,\tilde{X},\tilde{Y}, f_{L^\star,\mathrm{P},\lambda}(X), f_{L^\star,\mathrm{P},\lambda}(\tilde{X})\big)\Phi(\tilde{X})\otimes\Phi(\tilde{X})\Big] \end{aligned}$$

equals the Hessian of the regularized risk.

Please note, that the operator $M(\mathrm{P})$ and the first integral of $T(\mathrm{Q};\mathrm{P})$ only depend on P. Only the second and the third integral in the formula of $T(\mathrm{Q};\mathrm{P})$ depend on Q and describe how the RPL operator S changes, if the probability measure equals the mixture $(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q}$ instead of P. Of course, we have $S'_G(\mathrm{P})(\mathrm{P}) = 0 \in H$.

The influence function is an important approach in robust statistics and was proposed by Hampel (1968, 1974); we also refer to the classical textbook by Hampel *et al.* (1986). The influence function is strongly related to Gâteaux differentiation of the operator S in direction of the Dirac measure $\mathrm{Q} := \delta_{(x_0,y_0)}$, where $(x_0,y_0) \in \mathcal{X}\times\mathcal{Y}$, i.e.

$$\mathrm{IF}((x_0,y_0);S,\mathrm{P}) = S'_G(\mathrm{P})(\delta_{(x_0,y_0)}).$$

However, the influence function does not need to be a continuous linear operator, see Hampel *et al.* (1986, Def. 1, p. 84), which is in contrast to Gâteaux derivatives.

The influence function has the interpretation that it measures the influence of an (infinitesimal) small amout of contamination of the original measure P in the direction of a Dirac measure located in the point (x_0,y_0) on the theoretical quantity $S(\mathrm{P})$ of interest. Hence, it is desirable that a statistical method has a *bounded* influence function. If different methods have a bounded influence function, the one with the lower bound is considered to be more robust within this approach.

Corollary 5.4 (**Bounded influence function**). *Let the Assumptions 2.1, 3.2, 4.1, and 4.2 be satisfied. Denote the shifted version of L by $L^\star$. Then, for all $\mathrm{P} \in \mathcal{M}_1(\mathcal{X}\times\mathcal{Y})$, for all $(x_0,y_0) \in \mathcal{X}\times\mathcal{Y}$, and for all $\lambda \in (0,\infty)$, the influence function of $S : \mathcal{M}_1(\mathcal{X}\times\mathcal{Y}) \to H$ defined by $S(\mathrm{P}) := f_{L^\star,\mathrm{P},\lambda}$ is bounded. It holds*

$$\mathrm{IF}((x_0,y_0);S,\mathrm{P}) = -M(\mathrm{P})^{-1} T(\delta_{(x_0,y_0)};\mathrm{P}), \tag{5.9}$$

where $T(\delta_{(x_0,y_0)};\mathrm{P})$ and $M(\mathrm{P})$ are given by Theorem 5.3. Here $T(\delta_{(x_0,y_0)};\mathrm{P})$ simplifies to

$$\begin{aligned} &-2\,\mathbb{E}_{\mathrm{P}\otimes\mathrm{P}}\Big[D_5 L\big(X,Y,\tilde{X},\tilde{Y}, f_{L^\star,\mathrm{P},\lambda}(X), f_{L^\star,\mathrm{P},\lambda}(\tilde{X})\big)\Phi(X) \\ &\qquad + D_6 L\big(X,Y,\tilde{X},\tilde{Y}, f_{L^\star,\mathrm{P},\lambda}(X), f_{L^\star,\mathrm{P},\lambda}(\tilde{X})\big)\Phi(\tilde{X})\Big] \\ &+ \;\mathbb{E}_{\mathrm{P}}\Big[D_5 L\big(X,Y,x_0,y_0, f_{L^\star,\mathrm{P},\lambda}(X), f_{L^\star,\mathrm{P},\lambda}(x_0)\big)\Phi(X) \\ &\qquad + D_6 L\big(X,Y,x_0,y_0, f_{L^\star,\mathrm{P},\lambda}(X), f_{L^\star,\mathrm{P},\lambda}(x_0)\big)\Phi(x_0) \\ &\qquad + D_5 L\big(x_0,y_0,X,Y, f_{L^\star,\mathrm{P},\lambda}(x_0), f_{L^\star,\mathrm{P},\lambda}(X)\big)\Phi(x_0) \\ &\qquad + D_6 L\big(x_0,y_0,X,Y, f_{L^\star,\mathrm{P},\lambda}(x_0), f_{L^\star,\mathrm{P},\lambda}(X)\big)\Phi(X)\Big]. \end{aligned}$$

The pairwise loss function L_a fulfills the Assumptions 4.1 and 4.2 with $c_{L_a,1} = 1$ and $c_{L_a,2} = \frac{1}{2a}$ for any $a \in (0,\infty)$, see (2.10) and (2.11). Hence, Theorem 5.3 and Corollary 5.4 are applicable for L_a, if used in combination with a bounded and continuous kernel, e.g. a Gaussian RBF kernel.

Let us now reconsider the assumptions on L and k in Theorem 5.3. Due to

$$S'_G(\mathrm{P})(\mathrm{Q}) = -M(\mathrm{P})^{-1}T(\mathrm{Q};\mathrm{P})$$

and the specific form of $T(\mathrm{Q};\mathrm{P})$ and $M(\mathrm{P})$, we see that the boundedness of the Gâteaux derivative stems from the fact that L is separately Lipschitz continuous *and* k is bounded. One of these properties will in general not be enough to guarantee the boundedness of the Gâteaux derivative in unbounded spaces $\mathcal{X}$ and $\mathcal{Y}$. Let us give one simple example. If $\mathcal{Y}$ is unbounded, e.g. $\mathcal{Y} = \mathbb{R}$, we do not expect a bounded influence function for $f_{L^\star,\mathrm{P},\lambda}$, if the squared loss L_{LS} is used, because the supremum of the absolute values of the partial derivatives are unbounded in this case, as follows from (2.12). Please note that this is no contradiction to Theorem 5.3, because L_{LS} is clearly not separately Lipschitz continuous and $f_{L^\star_{LS},\mathrm{P},\lambda}$ is in general not even defined on the set of *all* Borel probability measures $\mathcal{M}_1(\mathcal{X}\times\mathcal{Y})$, if $\mathcal{Y}$ is unbounded.

In this sense, Theorem 5.3 and its corollary are in good agreement with results obtained by Christmann *et al.* (2009) for the case of support vector machines based on a general loss function and on a general kernel.

Besides the maximum bias over neighbourhoods and a bounded influence function, qualitative robustness is another key notion in robust statistics. Qualitative robustness was proposed by Hampel (1968, 1971) and generalized to more abstract spaces by Cuevas (1988), see also Cuevas and Romo (1993) for qualitative robustness of the empirical bootstrap. Define

$$\mathbb{P}_n := \frac{1}{n}\sum_{i=1}^{n}\delta_{(X_i,Y_i)}.$$

Now we show that the sequence of estimators

$$(S_n)_{n\in\mathbb{N}}, \quad S_n := f_{L^\star,\mathbb{P}_n,\lambda},$$

is qualitatively robust for all probability measures and any fixed regularization parameter $\lambda \in (0,\infty)$. We will also give an analogous qualitative robustness result for the empirical bootstrap approximations.

According to Hampel (1968) and Cuevas (1988) a sequence of estimators $(S_n)_{n\in\mathbb{N}}$ is called *qualitatively robust* at a probability measure P if and only if

$$\forall\,\varepsilon>0\ \ \exists\,\delta>0\ :\quad \Big[d_*(\mathrm{Q},\mathrm{P})\,<\,\delta \quad\Longrightarrow\quad d_*\big(\mathscr{L}_\mathrm{Q}(S_n),\mathscr{L}_\mathrm{P}(S_n)\big)\,<\,\varepsilon\ \ \forall\, n\in\mathbb{N}\Big]\,. \tag{5.10}$$

Here $\mathscr{L}_\mathrm{Q}(S_n)$ and $\mathscr{L}_\mathrm{P}(S_n)$ denote the image measures of Q and P by S_n, if all pairs (X_i,Y_i) are independent and identically distributed with $(X_i,Y_i)\sim\mathrm{P}$ or $(X_i,Y_i)\sim\mathrm{Q}$, respectively. Another common notation for $\mathscr{L}_\mathrm{P}(S_n)$ is $S_n(\mathrm{P}^n)$. Originally, Hampel (1971) used for d_* the Prohorov metric d_Pro, but one can also use the bounded Lipschitz metric d_BL defined by

$$d_\mathrm{BL}(\mathrm{P},\mathrm{Q}) := \sup\left\{\left|\int g\,d\mathrm{P}-\int g\,d\mathrm{Q}\right|;\|g\|_\mathrm{BL}\le 1\right\}\,,\quad \mathrm{P},\mathrm{Q}\in\mathcal{M}_1(\mathcal{X}\times\mathcal{Y})$$

in separable metric spaces, where $\|g\|_L := \sup_{x_1\neq x_2}|g(x_1)-g(x_2)|/d(x_1,x_2)$ and $\|g\|_\mathrm{BL} := \|g\|_L + \|g\|_\infty$, see Dudley (2002, Chapter 11.2). The reason for this is that, for any *separable* metric space – and in our case $\mathcal{X}\times\mathcal{Y}$ is separable –, both d_Pro and d_BL metrize the weak convergence for sequences of probability measures, i.e.

$$\mathrm{P}_n \rightsquigarrow \mathrm{P} \quad\Longleftrightarrow\quad d_\mathrm{Pro}(\mathrm{P}_n,\mathrm{P})\to 0 \quad\Longleftrightarrow\quad d_\mathrm{BL}(\mathrm{P}_n,\mathrm{P})\to 0, \tag{5.11}$$

we refer to Dudley (2002, Thm. 11.3.3, p. 395) for details. Hence, qualitative robustness as defined in (5.10) is a kind of equicontinuity concerning the weak convergence of the image measures of S_n with respect to n.

The finite sample distribution of RPL estimators is in general unknown. One method to obtain approximations of this finite sample distribution is the empirical bootstrap proposed by Efron (1979, 1982). As the next theorem also contains a qualitative robustness result of empirical bootstrap approximations, we need some more notation. Define $\mathbb{P}_n := \frac{1}{n}\sum_{i=1}^n \delta_{(X_i,Y_i)}$, where all pairs (X_i, Y_i) are independent and identically distributed with $(X_i, Y_i) \sim \mathrm{P}$ [abbreviation: $(X_i, Y_i) \overset{i.i.d.}{\sim} \mathrm{P}$]. Furthermore, we denote the distribution of the H-valued RPL estimator $f_{L^\star,\mathbb{P}_n,\lambda}$, by

$$\mathscr{L}_n(S;\mathrm{P}), \qquad n \in \mathbb{N}, \tag{5.12}$$

where $S : \mathcal{M}_1(\mathcal{X}\times\mathcal{Y}) \to H$ with $S(\mathrm{P}) = f_{L^\star,\mathrm{P},\lambda}$. As P is unknown but fixed, $\mathscr{L}_n(S;\mathrm{P})$ is a fixed but unknown probability measure of an H-valued random function. In the same manner we denote the distribution of the H-valued RPL estimator $f_{L^\star,\mathbb{P}_n,\lambda}$, when all pairs $(X_i^{(b)}, Y_i^{(b)}) \overset{i.i.d.}{\sim} \mathbb{P}_n := \frac{1}{n}\sum_{i=1}^n \delta_{(X_i,Y_i)}$, where $(X_i, Y_i) \overset{i.i.d.}{\sim} \mathrm{P}$, by

$$\mathscr{L}_n(S;\mathbb{P}_n), \quad n \in \mathbb{N}. \tag{5.13}$$

We mention that $\mathscr{L}_n(S;\mathbb{P}_n)$ is a probability kernel, see e.g. Kallenberg (2002, p. 106), because $\mathscr{L}_n(S;\mathbb{P}_n)$ denotes a probability measure, but it can also be considered as a random function in an abstract sense as it depends on the random probability measure $\mathbb{P}_n$.

We can now state our result on the qualitative robustness of regularized pairwise learning methods.

Theorem 5.5 (**Qualitative robustness**). *Let the Assumptions 2.1, 3.2, 4.1, and 4.2 be valid. Then, for all $\lambda \in (0,\infty)$, we have:*

(*i*) *The sequence of RPL estimators $(S_n)_{n\in\mathbb{N}}$, where $S_n := f_{L^\star,\mathbb{P}_n,\lambda}$, is qualitatively robust for all Borel probability measures $\mathrm{P} \in \mathcal{M}_1(\mathcal{X}\times\mathcal{Y})$.*

(*ii*) *If the metric space $\mathcal{X}\times\mathcal{Y}$ is additionally compact, then the sequence $\mathscr{L}_n(S;\mathbb{P}_n)$, $n \in \mathbb{N}$, of empirical bootstrap approximations of $\mathscr{L}_n(S;\mathrm{P})$ is qualitatively robust for all Borel probability measures $\mathrm{P} \in \mathcal{M}_1(\mathcal{X}\times\mathcal{Y})$.*

The proof of Theorem 5.5 is based on the following two results which are interesting in their own.

Theorem 5.6 (**Continuity of the operator**). *Let the Assumptions 2.1, 3.2, 4.1, and 4.2 be valid. Then, for all Borel probability measures $\mathrm{P} \in \mathcal{M}_1(\mathcal{X}\times\mathcal{Y})$ and for all $\lambda \in (0,\infty)$, we have:*

(*i*) *The operator $S : \mathcal{M}_1(\mathcal{X}\times\mathcal{Y}) \to H$, where $S(\mathrm{P}) = f_{L^\star,\mathrm{P},\lambda}$, is continuous with respect to the weak topology on $\mathcal{M}_1(\mathcal{X}\times\mathcal{Y})$ and the norm topology on H.*

(*ii*) *The operator $\tilde{S} : \mathcal{M}_1(\mathcal{X}\times\mathcal{Y}) \to \mathcal{C}_b(\mathcal{X})$, where $S(\mathrm{P}) = f_{L^\star,\mathrm{P},\lambda}$, is continuous with respect to the weak topology on $\mathcal{M}_1(\mathcal{X}\times\mathcal{Y})$ and the norm topology on $\mathcal{C}_b(\mathcal{X})$.*

Corollary 5.7 (**Continuity of the estimator**). *Let the Assumptions 2.1, 3.2, 4.1, and 4.2 be valid. For any data set $D_n \in (\mathcal{X}\times\mathcal{Y})^n$ denote the corresponding empirical measure by $\mathrm{D}_n := \frac{1}{n}\sum_{i=1}^n \delta_{(x_i,y_i)}$. Then, for every $\lambda \in (0,\infty)$ and every $n \in \mathbb{N}$, the mapping*

$$S_n : \Big((\mathcal{X}\times\mathcal{Y})^n, d_{(\mathcal{X}\times\mathcal{Y})^n}\Big) \to \big(H, d_H\big), \quad S_n(D_n) = f_{L^\star,\mathrm{D}_n,\lambda}, \tag{5.14}$$

is continuous.

It is known that support vector machines (SVMs) are qualitatively robust for fixed values of $\lambda \in (0,\infty)$. However they *can not* be qualitatively robust for the usual null-sequences λ_n needed to obtain universal consistency, because universal consistency and qualitatively robustness are under very mild conditions concurrent goals, see Hable and Christmann (2013) for a discussion. This is not only true for SVMs. Is is known that SVMs can satisfy somewhat weaker property called finite-sample qualitative robustness, see Hable and Christmann (2011). It is an open problem, whether a similar result is true for RPL methods, and we will not address this question here.

6 Discussion

In this paper we proved several desirable statistical robustness properties (upper bounds for the maxbias, a bounded influence function, and qualitative robustness) for a broad class of regularized pairwise learning methods based on kernels. Such kernel methods are used in the fields of information theoretic learning, ranking, gradient learning, and metric and similarity learning. In particular, our work complements to some respect earlier work on consistency and learning rates for minimum error entropy principles by Hu *et al.* (2013), Fan *et al.* (2016), Hu *et al.* (2015), for ranking algorithms by Agarwal and Niyogi (2009), for metric and similarity learning problems by Cao *et al.* (2015), and for gradient learning methods by Mukherjee and Zhou (2006).

The following aspects are beyond the scope of this paper and remain open for further work. *(i)* We did not address the question of an influence function of regularized pairwise learning methods, if a bounded but *non-convex* pairwise loss function is used. The main problem seems to be that in this case the function $f_{L^\star,\mathrm{P},\lambda}$ is in general not unique. *(ii)* We did not add numerical comparisons because it is known from Principe (2010) that for minimum error entropy principles such methods can be computed in an efficient gradient descent manner. The convergence of the gradient descent algorithm for the minimum error entropy principle is recently proved rigorously in a linear regression setting by Hu *et al.* (2016). *(iii)* It is obvious that the results developed here for pairwise learning can in principle be established also to higher order, if one uses U- or V-statistics of degree $\ell > 2$. E.g., if $\ell = 3$, one can consider loss functions with 9 instead of 6 arguments which yields instead of (1.2) the optimization problem

$$\inf_{f\in H} \frac{1}{n^3} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{m=1}^{n} L(x_i, y_i, x_j, y_j, x_m, y_m, f(x_i), f(x_j), f(x_m)) + \lambda \|f\|_H^2 \,. \tag{6.1}$$

We conjecture that the numerical effort to solve such problems will strongly increase with ℓ.

7 Appendix

7.1 Appendix A: Some Tools

To improve the readability of the paper, we list some known results which are used in our proofs. The following theorem provides a criterion for the existence of a global minimizer. We refer to Ekeland and Turnbull (1983, Prop. 6, p. 75) for the following result.

Theorem 7.1 (Existence of minimizers)**.** *Let E be a reflexive Banach space and $f : E \to \mathbb{R} \cup \{\infty\}$ be a convex and lower semi-continuous map. If there exists an $M > 0$ such that $\{x \in E : f(x) \leq M\}$ is non-empty and bounded, then f has a global minimum, i.e., there exists an $x_0 \in E$ with $f(x_0) \leq f(x)$ for all $x \in E$. Moreover, if f is strictly convex, then x_0 is the only element minimizing f.*

Definition 7.2 (Derivatives, see e.g. Denkowski *et al.* (2003, p. 518f))**.** *Let E and F be normed spaces, $U \subset E$ and $V \subset F$ be open sets, and $G : U \to V$ be a map. We say that G is Gâteaux differentiable at $x_0 \in U$ if there exists a bounded linear operator $A : E \to F$ such that*

$$\lim_{\substack{t\to 0 \\ t\neq 0}} \frac{\|G(x_0 + tx) - G(x_0) - tAx\|_F}{t} = 0 \,, \qquad x \in E.$$

In this case, A is called the derivative of G at x_0, and since A is uniquely determined, we write $G'(x_0) := \frac{\partial G}{\partial E}(x_0) := A$. Moreover, we say that G is Fréchet differentiable at x_0 if A actually satisfies

$$\lim_{\substack{x\to 0 \\ x\neq 0}} \frac{\|G(x_0 + x) - G(x_0) - Ax\|_F}{\|x\|_E} = 0 \,.$$

Furthermore, we say that G is (Gâteaux, Fréchet) differentiable if it is (Gâteaux, Fréchet) differentiable at every $x_0 \in U$. Finally, G is said to be continuously differentiable if it is Fréchet differentiable and the derivative $G' : U \to \mathcal{L}(E, F)$ is continuous.

Theorem 7.3 (Partial Fréchet differentiability, see e.g. Akerkar (1999, Theorem 2.6, p. 37))**.** *Let E_1, E_2, and F be Banach spaces, $U_1 \subset E_1$ and $U_2 \subset E_2$ be open subsets, and $G : U_1 \times U_2 \to F$ be a continuous map. Then G is continuously differentiable if and only if G is partially Fréchet differentiable and the partial derivatives $\frac{\partial G}{\partial E_1}$ and $\frac{\partial G}{\partial E_2}$ are continuous. In this case, the derivative of G at $(x_1, x_2) \in U_1 \times U_2$ is given by*

$$G'(x_1, x_2)(y_1, y_2) = \frac{\partial G}{\partial E_1}(x_1, x_2)y_1 + \frac{\partial G}{\partial E_2}(x_1, x_2)y_2\,, \qquad (y_1, y_2) \in E_1 \times E_2\,.$$

The proof of Theorem 5.3 heavily relies on an implicit function theorem in Banach spaces. Recall the following simplified version of this theorem, see Akerkar (1999, Thm. 4.1, Cor. 4.2). Here and throughout this appendix B_E denotes the open unit ball of a Banach space E.

Theorem 7.4 (Implicit function theorem)**.** *Let E, F be Banach spaces and $G : E \times F \to F$ be a continuously differentiable map. Suppose that we have $(x_0, y_0) \in E \times F$ such that $G(x_0, y_0) = 0$ and $\frac{\partial G}{\partial F}(x_0, y_0)$ is invertible. Then there exists a $\delta > 0$ and a continuously differentiable map $f : x_0 + \delta B_E \to y_0 + \delta B_F$ such that for all $x \in x_0 + \delta B_E$, $y \in y_0 + \delta B_F$ we have: $G(x, y) = 0$ if and only if $y = f(x)$. Moreover, the derivative of f is given by*

$$f'(x) = -\left(\frac{\partial G}{\partial F}\big(x, f(x)\big)\right)^{-1} \frac{\partial G}{\partial E}\big(x, f(x)\big)\,. \tag{7.1}$$

Definition 7.5 (Bochner integral)**.** *Let E be a Banach space and $(\Omega, \mathcal{A}, \mu)$ be a σ-finite measure space. An E-valued measurable function $f : \Omega \to E$ is called Bochner μ-integrable if there exists a sequence $(f_n)_{n\in\mathbb{N}}$ of E-valued measurable step functions $f_n : \Omega \to E$ such that $\lim_{n\to\infty} \int_\Omega \|f_n - f\|_E \, d\mu = 0$. In this case, the limit $\int_\Omega f \, d\mu := \lim_{n\to\infty} \int_\Omega f_n \, d\mu$ exists and is called the Bochner integral of f. Finally, if μ is a probability measure, we sometimes write $\mathbb{E}_\mu[f]$ for this integral.*

Theorem 7.6 (Dominated convergence theorem, see e.g. Denkowski *et al.* (2003, Thm. 3.10.12, p. 367))**.** *Let E be a Banach space, $(\Omega, \mathcal{A}, \mu)$ be a finite measure space, and $(f_n)_{n\in\mathbb{N}}$ be a sequence of Bochner μ-integrable functions $f_n : \Omega \to E$. If $\lim_{n\to\infty} \mu\{\|f_n - f\| \geq \varepsilon\} = 0$ for every $\varepsilon > 0$ and if there exists a μ-integrable function $g : \Omega \to \mathbb{R}$ with $\|f_n(\omega)\| \leq g(\omega)$ μ-almost everywhere for all $n \in \mathbb{N}$, then f is Bochner μ-integrable and $\lim_{n\to\infty} \int_\Omega f_n \, d\mu = \int_\Omega f \, d\mu$.*

7.2 Appendix B: Proofs

To shorten the notation, we occasionally use the abbreviations $z := (x, y)$, $\tilde{z} := (\tilde{x}, \tilde{y})$, $Z := (X, Y)$, $\tilde{Z} := (\tilde{X}, \tilde{Y})$, $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$, and $D_i L(z, \tilde{z}, t, \tilde{t}) := D_i L(x, y, \tilde{x}, \tilde{y}, t, \tilde{t})$, $D_i D_j L(z, \tilde{z}, t, \tilde{t}) := D_i D_j L(x, y, \tilde{x}, \tilde{y}, t, \tilde{t})$ for $i \in \{5, 6\}$ etc.

The proofs for the results given in Section 2 and Section 3 are similar to corresponding results for "classical" loss functions of the form $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ used by support vector machines and related kernel based methods, see e.g. Steinwart and Christmann (2008).

Proof of Lemma 2.4. Since d dominates the pointwise convergence, we see that, for fixed $x \in \mathcal{X}$, the $\mathbb{R}$-valued map $f \mapsto f(x)$ defined on $\mathcal{F}$ is continuous with respect to d. Furthermore, $\mathcal{F} \subset \mathcal{L}_0(\mathcal{X})$ implies that, for fixed $f \in \mathcal{F}$, the $\mathbb{R}$-valued map $x \mapsto f(x)$ defined on $\mathcal{X}$ is measurable. By a well-known result from Carathéodory, see e.g. Castaing and Valadier (1977, p. 70), we then obtain the first assertion. Since this implies that the maps $(x, y, \tilde{x}, \tilde{y}, f) \mapsto (x, y, \tilde{x}, \tilde{y}, f(x), f(\tilde{x}))$ and $(x, y, \tilde{x}, \tilde{y}, f, f) \mapsto (x, y, \tilde{x}, \tilde{y}, f(x), f(\tilde{x}))$ are measurable, we obtain the second assertion. The third assertion now follows from the measurability statement in Tonelli-Fubini's theorem, see e.g. Dudley (2002, p. 137). □

Proof of Lemma 2.6. Let $c \in [0,1]$, $f, g \in \mathcal{L}_0(\mathcal{X})$, and assume that L is a convex pairwise loss. We immediately obtain, for all $(x, y, \tilde{x}, \tilde{y}) \in (\mathcal{X} \times \mathcal{Y})^2$,

$$\begin{aligned} & L\bigl(x, y, \tilde{x}, \tilde{y}, cf(x) + (1-c)g(x), cf(\tilde{x}) + (1-c)g(\tilde{x})\bigr) \\ \leq \; & cL\bigl(x, y, \tilde{x}, \tilde{y}, f(x), f(\tilde{x})\bigr) + (1-c)L\bigl(x, y, \tilde{x}, \tilde{y}, g(x), g(\tilde{x})\bigr). \end{aligned}$$

The linearity of integrals yields the assertion $\mathcal{R}_{L,\mathrm{P}}(cf + (1-c)g) \leq c\mathcal{R}_{L,\mathrm{P}}(f) + (1-c)\mathcal{R}_{L,\mathrm{P}}(g)$. The case of strict convexity can be shown in an analogous manner. □

Proof of Lemma 2.7. As L is a locally separately Lipschitz continuous pairwise loss, we have

$$\begin{aligned} & |\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{P}}(g)| \\ \leq \; & \int_{(\mathcal{X}\times\mathcal{Y})^2} \bigl|L(x, y, \tilde{x}, \tilde{y}, f(x), f(\tilde{x})) - L(x, y, \tilde{x}, \tilde{y}, g(x), g(\tilde{x}))\bigr| \, d\mathrm{P}^2(x, y, \tilde{x}, \tilde{y}) \\ \leq \; & |L|_{B,1} \int_{\mathcal{X}^2} |f(x) - g(x)| + |f(\tilde{x}) - g(\tilde{x})| \, d\mathrm{P}_{\mathcal{X}}^2(x, \tilde{x}) \\ \leq \; & 2\,|L|_{B,1} \cdot \|f - g\|_{L_1(\mathrm{P}_{\mathcal{X}})}\,, \end{aligned}$$

which gives the assertion. □

Proof of Lemma 2.8. As we consider L as a function of its last two arguments, while the first four arguments are held fixed, we define $L_{z,\tilde{z}}(t, \tilde{t}) := L(z, \tilde{z}, t, \tilde{t})$, where $z := (x, y)$ and $\tilde{z} = (\tilde{x}, \tilde{y})$. We first observe that all derivatives $(DL_{z,\tilde{z}})(t, \tilde{t})$ are measurable since we assumed continuous partial derivatives. Now let $f \in L_\infty(\mathrm{P}_{\mathcal{X}})$ and $(f_n)_{n\in\mathbb{N}} \subset L_\infty(\mathrm{P}_{\mathcal{X}})$ be a sequence with $f_n \neq 0$, $n \geq 1$, and $\lim_{n\to\infty} \|f_n\|_\infty = 0$. Without loss of generality, we additionally assume for later use that $\|f_n\|_\infty \leq 1$ for all $n \geq 1$. For $z = (x, y), \tilde{z} = (\tilde{x}, \tilde{y}) \in \mathcal{Z}$ and $n \geq 1$, we now define

$$\begin{aligned} G_n(z, \tilde{z}) \;:=\; & \sqrt{2} \cdot \Bigl|L_{z,\tilde{z}}\bigl(f(x) + f_n(x), f(\tilde{x}) + f_n(\tilde{x})\bigr) - L_{z,\tilde{z}}\bigl(f(x), f(\tilde{x})\bigr) \\ & \qquad - \bigl\langle (DL_{z,\tilde{z}})(f(x), f(\tilde{x})), \bigl(f_n(x), f_n(\tilde{x})\bigr)\bigr\rangle\Bigr| \Big/ \|(f_n(x), f_n(\tilde{x}))\|_2\,, \end{aligned}$$

if $\|(f_n(x), f_n(\tilde{x}))\|_2 \neq 0$, and $G_n(z, \tilde{z}) := 0$ otherwise. We obtain

$$\begin{aligned} & \left|\frac{\mathcal{R}_{L,\mathrm{P}}(f + f_n) - \mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}'_{L,\mathrm{P}}(f)f_n}{\|f_n\|_\infty}\right| \\ \leq \; & \int_{(\mathcal{X}\times\mathcal{Y})^2} \frac{1}{\|f_n\|_\infty} \cdot \Bigl|L_{z,\tilde{z}}\bigl(f(x) + f_n(x), f(\tilde{x}) + f_n(\tilde{x})\bigr) - L_{z,\tilde{z}}\bigl(f(x), f(\tilde{x})\bigr) \\ & \qquad\qquad - \bigl\langle (DL_{z,\tilde{z}})(f(x), f(\tilde{x})), \bigl(f_n(x), f_n(\tilde{x})\bigr)\bigr\rangle\Bigr| \, d\mathrm{P}^2(x, y, \tilde{x}, \tilde{y}) \\ \leq \; & \int_{(\mathcal{X}\times\mathcal{Y})^2} \frac{\sqrt{2}}{\|(f_n(x), f_n(\tilde{x}))\|_2} \cdot \qquad (7.2) \\ & \quad \Bigl|L_{z,\tilde{z}}\bigl(f(x) + f_n(x), f(\tilde{x}) + f_n(\tilde{x})\bigr) - L_{z,\tilde{z}}\bigl(f(x), f(\tilde{x})\bigr) \\ & \qquad - \bigl\langle (DL_{z,\tilde{z}})(f(x), f(\tilde{x})), \bigl(f_n(x), f_n(\tilde{x})\bigr)\bigr\rangle\Bigr| \, d\mathrm{P}^2(x, y, \tilde{x}, \tilde{y}) \\ = \; & \int_{\mathcal{Z}^2} G_n(z, \tilde{z}) \, d\mathrm{P}^2(z, \tilde{z}) \qquad (7.3) \end{aligned}$$

for all $n \geq 1$, where the well-known relationship $\|v\|_2 \leq \sqrt{2}\,\|v\|_\infty$ for $v \in \mathbb{R}^2$, was used in (7.2). Furthermore, for $z, \tilde{z} \in \mathcal{Z}$, the definition of G_n and the definition of the (total) derivative $DL_{z,\tilde{z}}$ obviously yield

$$\lim_{n\to\infty} G_n(z, \tilde{z}) = 0. \qquad (7.4)$$

Denote the gradient of $L_{z,\tilde{z}}$ by $\nabla L_{z,\tilde{z}}$. For $z, \tilde{z} \in \mathcal{Z}$ and $n \geq 1$ with $f_n(x) \neq 0$, the mean value theorem for functions from $\mathbb{R}^2$ to $\mathbb{R}$ shows that there exists some $g_n(z, \tilde{z})$ with

$$\begin{aligned} & L_{z,\tilde{z}}\big(f(x) + f_n(x), f(\tilde{x}) + f_n(\tilde{x})\big) - L_{z,\tilde{z}}\big(f(x), f(\tilde{x})\big) \qquad (7.5) \\ = \; & \Big\langle (\nabla L_{z,\tilde{z}})\big((1 - g_n(z,\tilde{z}))f(x) + g_n(z,\tilde{z})f_n(x), \\ & \qquad (1 - g_n(z,\tilde{z}))f(\tilde{x}) + g_n(z,\tilde{z})f_n(\tilde{x})\big), \;\; \big(f_n(x), f_n(\tilde{x})\big)\Big\rangle . \end{aligned}$$

As L has by assumption uniformly bounded partial derivatives $D_5 L$ and $D_6 L$, there exists a constant $c_L \in [0, \infty)$ such that

$$\sup_{x,\tilde{x} \in \mathcal{X},\; y,\tilde{y} \in \mathcal{Y},\; t,\tilde{t} \in \mathbb{R}^2} \big| D_i L(x, y, \tilde{x}, \tilde{y}, t, \tilde{t}) \big| \leq c_L , \qquad i \in \{5, 6\}.$$

If we combine this with the equality in (7.5), we obtain

$$\begin{aligned} & \Big| L_{z,\tilde{z}}\big(f(x) + f_n(x), f(\tilde{x}) + f_n(\tilde{x})\big) - L_{z,\tilde{z}}\big(f(x), f(\tilde{x})\big) \Big| \qquad (7.6) \\ \leq \; & 2c_L\big(|f_n(x)| + |f_n(\tilde{x})|\big) \;\leq\; 4c_L \|f_n\|_\infty \;\longrightarrow 0, \quad n \to \infty . \end{aligned}$$

Combining (7.3), (7.5), and (7.6), we get $\int_{Z^2} G_n(z, \tilde{z}) \, d\mathrm{P}^2(z, \tilde{z}) < \infty$. Hence, G_n is a non-negative convergent dominating function and the assertion follows from Lebesgue's theorem. □

The following result is helpful for the proof of Theorem 3.3.

Lemma 7.7. *Let $r \in (0, \infty)$. If $f_0 \in H$ and if the sequence $(f_j)_{j=1}^\infty \subset B(f_0, r) := \{f \in H : \|f - f_0\|_H \leq r\}$, then there exists a subsequence $(f_{j_\ell})_{\ell=1}^\infty$ and $f^* \in B(f_0, r)$ such that $\|f^*\|_H \leq \underline{\lim}_{\ell \to \infty} \|f_{j_\ell}\|_H$ and*

$$\lim_{\ell \to \infty} f_{j_\ell}(x) = f^*(x), \qquad \forall x \in \mathcal{X}. \qquad (7.7)$$

Proof. Since the closed ball $B(f_0, r)$ of the Hilbert space H is weakly compact, there exists a subsequence $(f_{j_\ell})_{\ell=1}^\infty$ weakly converging to some $f \in B(f_0, r)$. That is,

$$\lim_{\ell \to \infty} \langle f_{j_\ell}, f \rangle_H = \langle f^*, f \rangle_H, \qquad \forall f \in H. \qquad (7.8)$$

Let $f = f^*$ in (7.8). Then

$$\|f^*\|_H^2 = \langle f^*, f^* \rangle_H = \lim_{\ell \to \infty} \langle f_{j_\ell}, f^* \rangle_H \leq \underline{\lim}_{\ell \to \infty} \|f_{j_\ell}\|_H \|f^*\|_H$$

which implies $\|f^*\|_H \leq \underline{\lim}_{\ell \to \infty} \|f_{j_\ell}\|_H$.

Let $f = \Phi(x)$, $x \in \mathcal{X}$, in (7.8). Then the reproducing property of the kernel yields

$$\lim_{\ell \to \infty} f_{j_\ell}(x) = \lim_{\ell \to \infty} \langle f_{j_\ell}, \Phi(x) \rangle_H = \langle f^*, \Phi(x) \rangle_H = f^*(x).$$

This is true for any $x \in \mathcal{X}$. So (7.7) is verified. □

Proof of Theorem 3.3. For every $\ell \in \mathbb{N}$, we take a function $f_\ell \in H$ such that

$$\mathcal{R}_{L,\mathrm{P}}(f_\ell) + \lambda \|f_\ell\|_H^2 \leq \inf_{f \in H} \mathcal{R}_{L,\mathrm{P}}(f) + \lambda \|f\|_H^2 + \frac{1}{\ell}. \qquad (7.9)$$

Taking $f = f_0$, we find that

$$\lambda \|f_\ell\|_H^2 \leq \mathcal{R}_{L,\mathrm{P}}(f_0) + \lambda \|f_0\|_H^2 + 1$$

and

$$f_\ell \in B(0,r), \text{ with } r = \frac{\mathcal{R}_{L,\mathrm{P}}(f_0) + \lambda\|f_0\|_H^2 + 1}{\lambda}.$$

Now we apply Lemma 7.7. We know that there exists a subsequence $(f_{j_\ell})_{\ell=1}^\infty$ and $f^* \in B(0,r)$ such that $\|f^*\|_H \leq \underline{\lim}_{\ell\to\infty}\|f_{j_\ell}\|_H$ and (7.7) is valid.

Consider $\mathcal{R}_{L,\mathrm{P}}(f_\ell)$. By the Lipschitz continuity of L, the integrated function is bounded by

$$L\left(x,y,\tilde{x},\tilde{y},f_\ell(x),f_\ell(\tilde{x})\right) \leq L\left(x,y,\tilde{x},\tilde{y},f_0(x),f_0(\tilde{x})\right) + 4|L|_1\|k\|_\infty r.$$

The upper bound is integrable with respect to P^2. Also, for any $(x,y,\tilde{x},\tilde{y}) \in (\mathcal{X}\times\mathcal{Y})^2$, by the continuity of L and (7.7), we have

$$\lim_{\ell\to\infty} L\left(x,y,\tilde{x},\tilde{y},f_\ell(x),f_\ell(\tilde{x})\right) = L\left(x,y,\tilde{x},\tilde{y},f^*(x),f^*(\tilde{x})\right).$$

So by the Lebesgue Dominated Theorem, we have

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

Then we take $\underline{\lim}$ on both sides of (7.9) and find from $\|f^*\|_H \leq \underline{\lim}_{\ell\to\infty}\|f_{j_\ell}\|_H$ that

$$\mathcal{R}_{L,\mathrm{P}}(f^*) + \lambda\|f^*\|_H \leq \inf_{f\in H} \mathcal{R}_{L,\mathrm{P}}(f) + \lambda\|f\|_H^2.$$

It means that f^* is a minimizer $f_{L,\mathrm{P},\lambda}$. This proves our statement. □

Proof of Lemma 3.4. Let us assume that the map $f \mapsto \mathcal{R}_{L,\mathrm{P}}(f) + \lambda\|f\|_H^2$ has two minimizers $f_1, f_2 \in H$ with $f_1 \neq f_2$. Recall that the parallelogram law $\|x+x'\|^2 + \|x-x'\|^2 = 2\|x\|^2 + 2\|x'\|^2$ is valid for all points x and x' in a Hilbert space, *cf.* Denkowski *et al.* (2003, Thm. 3.7.7, p. 310). Therefore, we have $\|\frac{1}{2}(f_1+f_2)\|_H^2 < \frac{1}{2}\|f_1\|_H^2 + \frac{1}{2}\|f_2\|_H^2$. As L is a convex pairwise loss, the map $f \mapsto \mathcal{R}_{L,\mathrm{P}}(f)$ is convex by Lemma 2.6. This together with $\mathcal{R}_{L,\mathrm{P}}(f_1) + \lambda\|f_1\|_H^2 = \mathcal{R}_{L,\mathrm{P}}(f_2) + \lambda\|f_2\|_H^2$ then shows for $f^* := \frac{1}{2}(f_1+f_2)$ that

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

i.e., f_1 is not a minimizer of $f \mapsto \mathcal{R}_{L,\mathrm{P}}(f) + \lambda\|f\|_H^2$. Consequently, the assumption that there are two minimizers is false. □

Proof of Theorem 3.5. Since the kernel k of H is measurable, H consists of measurable functions, see e.g. Steinwart and Christmann (2008, Lem. 4.24). Moreover, k is bounded and thus $\mathrm{id} : H \to L_\infty(\mathrm{P}_\mathcal{X})$ is continuous, see e.g. Steinwart and Christmann (2008, Lem. 4.23). In addition, we have $L(z,\tilde{z},t,\tilde{t}) < \infty$ for all $(z,\tilde{z},t,\tilde{t}) \in \mathcal{Z}^2 \times \mathbb{R}^2$. Recall that every convex function $g : \mathbb{R}^2 \to \mathbb{R}\cup\{\infty\}$, which is not identically equal to $+\infty$, is continuous on the interior of its effective domain $\operatorname{Dom} g := \{t \in \mathbb{R}; g(t) < \infty\}$, see e.g. Ekeland and Témam (1999, Cor. 2.3 on p. 12). Hence L is a continuous pairwise loss by the convexity of L. Therefore, Lemma 2.7 shows that $\mathcal{R}_{L,\mathrm{P}} : L_\infty(\mathrm{P}_\mathcal{X}) \to \mathbb{R}$ is continuous, and hence $\mathcal{R}_{L,\mathrm{P}} : H \to \mathbb{R}$ is continuous. In addition, Lemma 2.6 provides the convexity of this map. Furthermore, $f \mapsto \lambda\|f\|_H^2$ is also convex and continuous, which yields the continuity and the convexity of the map $f \mapsto \mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(f)$. Now consider the set $A := \left\{f \in H : \mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(f) \leq \mathcal{R}_{L,\mathrm{P}}(0)\right\}$. We obviously have $0 \in A$. In addition, $f \in A$ implies $\lambda\|f\|_H^2 \leq \mathcal{R}_{L,\mathrm{P}}(0)$, and hence $A \subset (\mathcal{R}_{L,\mathrm{P}}(0)/\lambda)^{1/2}\bar{B}_H$, where $\bar{B}_H$ denotes the closed unit ball of H. Hence A is a non-empty and bounded subset of H and thus Theorem 7.1 gives the existence of a minimizer $f_{L,\mathrm{P},\lambda}$. □

Proof of Corollary 3.6. As L is a separately Lipschitz continuous pairwise loss, we have, for all $(z, \tilde{z}, t, \tilde{t}) \in \mathcal{Z}^2 \times \mathbb{R}^2$,

$$\begin{aligned} L(z, \tilde{z}, t, \tilde{t}) &= L(z, \tilde{z}, 0, 0) + L(z, \tilde{z}, t, \tilde{t}) - L(z, \tilde{z}, 0, 0) \\ &\leq L(z, \tilde{z}, 0, 0) + 2|L|_1 \left(|t| + |\tilde{t}|\right). \end{aligned}$$

The assumption $\mathcal{R}_{L,\mathrm{P}}(0) < \infty$ yields $\mathcal{R}_{L,\mathrm{P}}(f) \leq \mathcal{R}_{L,\mathrm{P}}(0) + 4|L|_1 \|f\|_{L_1(\mathrm{P}_X)} < \infty$. As L is a convex pairwise loss, Lemma 3.4 and Theorem 3.5 yield the existence and the uniqueness of $f_{L,\mathrm{P},\lambda}$ and (3.8) equals (3.3). □

Lemma 7.8. *Let L be a pairwise loss. Then the following statements are valid.*

(*i*) *If L is (strictly) convex, then $L^\star$ is (strictly) convex.*

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

Proof. Follows immediately from the definition of $L^\star$. □

Lemma 7.9. *Let L be a pairwise loss and $L^\star$ its shifted version. Then the following assertions are valid.*

(*i*) $\inf_{f \in \mathcal{L}_0(X)} L^\star(x, y, \tilde{x}, \tilde{y}, f(x), f(\tilde{x})) \leq 0$.

(*ii*) *If L is a separately Lipschitz continuous pairwise loss, then for all $f \in H$:*

$$-2|L|_1 \mathbb{E}_{\mathrm{P}_X}|f(X)| \leq \mathcal{R}_{L^\star,\mathrm{P}}(f) \leq 2|L|_1 \mathbb{E}_{\mathrm{P}_X}|f(X)|, \tag{7.10}$$

$$-2|L|_1 \mathbb{E}_{\mathrm{P}_X}|f(X)| + \lambda\|f\|_H^2 \;\leq\; \mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f) \;\leq\; 2|L|_1 \mathbb{E}_{\mathrm{P}_X}|f(X)| + \lambda\|f\|_H^2. \tag{7.11}$$

(*iii*) $\inf_{f \in H} \mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f) \leq 0$ *and* $\inf_{f \in H} \mathcal{R}_{L^\star,\mathrm{P}}(f) \leq 0$.

(*iv*) *Let L be a separately Lipschitz continuous pairwise loss and assume that $f_{L^\star,\mathrm{P},\lambda}$ exists. Then we have*

$$\begin{aligned} \lambda\|f_{L^\star,\mathrm{P},\lambda}\|_H^2 &\leq -\mathcal{R}_{L^\star,\mathrm{P}}(f_{L^\star,\mathrm{P},\lambda}) \;\leq\; \mathcal{R}_{L,\mathrm{P}}(0), \\ 0 &\leq -\mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f_{L^\star,\mathrm{P},\lambda}) \;\leq\; \mathcal{R}_{L,\mathrm{P}}(0), \\ \lambda\|f_{L^\star,\mathrm{P},\lambda}\|_H^2 &\leq \min\bigl\{|L|_1 \mathbb{E}_{\mathrm{P}_X}|f_{L^\star,\mathrm{P},\lambda}(X)|, \mathcal{R}_{L,\mathrm{P}}(0)\bigr\}. \end{aligned} \tag{7.12}$$

If the kernel k is additionally bounded, then

$$\|f_{L^\star,\mathrm{P},\lambda}\|_\infty \leq \lambda^{-1}|L|_1\|k\|_\infty^2 < \infty, \tag{7.13}$$

$$|\mathcal{R}_{L^\star,\mathrm{P}}(f_{L^\star,\mathrm{P},\lambda})| \leq \lambda^{-1}|L|_1^2\|k\|_\infty^2 < \infty. \tag{7.14}$$

(*v*) *If the partial derivatives D_iL and D_iD_jL of L exist for all $(x, y, \tilde{x}, \tilde{y}) \in (\mathcal{X} \times \mathcal{Y})^2$ and all $i, j \in \{5, 6\}$, then, for all $(t, \tilde{t}) \in \mathbb{R}^2$,*

$$D_iL^\star(x, y, \tilde{x}, \tilde{y}, t, \tilde{t}) = D_iL(x, y, \tilde{x}, \tilde{y}, t, \tilde{t}), \tag{7.15}$$

$$D_iD_jL^\star(x, y, \tilde{x}, \tilde{y}, t, \tilde{t}) = D_iD_jL(x, y, \tilde{x}, \tilde{y}, t, \tilde{t}). \tag{7.16}$$

Proof. (i) Obviously, we have

$$\inf_{f \in \mathcal{L}_0(X)} L^\star(x, y, \tilde{x}, \tilde{y}, f(x), f(\tilde{x})) \leq L^\star(x, y, \tilde{x}, \tilde{y}, 0, 0) = 0.$$

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

$$\begin{aligned} |\mathcal{R}_{L^\star,\mathrm{P}}(f)| &= |\mathbb{E}_{\mathrm{P}^2} L(X,Y,\tilde{X},\tilde{Y},f(X),f(\tilde{X})) - L(X,Y,\tilde{X},\tilde{Y},0,0)| \\ &\leq \mathbb{E}_{\mathrm{P}^2} |L(X,Y,\tilde{X},\tilde{Y},f(X),f(\tilde{X})) - L(X,Y,\tilde{X},\tilde{Y},0,0)| \\ &\leq |L|_1 \, \mathbb{E}_{\mathrm{P}^2}\big(|f(X)| + |f(\tilde{X})|\big) \\ &\leq 2|L|_1 \, \mathbb{E}_{\mathrm{P}_\mathcal{X}} |f(X)|, \end{aligned}$$

which proves (7.10). Equation (7.11) follows from $\mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f) = \mathcal{R}_{L^\star,\mathrm{P}}(f) + \lambda\|f\|_H^2$.
(iii) As $0 \in H$, we obtain $\inf_{f\in H} \mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f) \leq \mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(0) = 0$ and the same reasoning holds for $\inf_{f\in H} \mathcal{R}_{L^\star,\mathrm{P}}(f)$.
(iv) Due to *(iii)* we have $\mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f_{L^\star,\mathrm{P},\lambda}) \leq 0$. As $L \geq 0$ we obtain

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

Using similar arguments as above, we obtain

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

Furthermore, we obtain

$$-2|L|_1 \, \mathbb{E}_{\mathrm{P}_\mathcal{X}} |f_{L^\star,\mathrm{P},\lambda}(X)| + \lambda\|f_{L^\star,\mathrm{P},\lambda}\|_H^2 \leq \mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f_{L^\star,\mathrm{P},\lambda}) \leq \mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(0) = 0.$$

This yields (7.12). Using (3.5), (3.7), and (7.12), we obtain for $f_{L^\star,\mathrm{P},\lambda} \neq 0$ that

$$\begin{aligned} \|f_{L^\star,\mathrm{P},\lambda}\|_\infty &\leq \|k\|_\infty \|f_{L^\star,\mathrm{P},\lambda}\|_H \\ &\leq \|k\|_\infty \sqrt{(2/\lambda)\,|L|_1 \, \mathbb{E}_{\mathrm{P}_\mathcal{X}} |f_{L^\star,\mathrm{P},\lambda}(X)|} \\ &\leq \|k\|_\infty \sqrt{(2/\lambda)\,|L|_1 \|f_{L^\star,\mathrm{P},\lambda}\|_\infty} < \infty\,. \end{aligned}$$

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

(v) This follows immediately from the definition of $L^\star$, because we just subtract a term, which is constant w.r.t. the last two arguments of $L^\star$. □

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

Lemma 7.10. *Let L be a separately Lipschitz continuous pairwise loss function and $f \in L_1(\mathrm{P}_\mathcal{X})$. Then $\mathcal{R}_{L^\star,\mathrm{P}}(f) \notin \{-\infty,+\infty\}$. Moreover, we have $\mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f) > -\infty$ for all $f \in L_1(\mathrm{P}_\mathcal{X}) \cap H$.*

Proof. The inequality $|\mathcal{R}_{L^\star,\mathrm{P}}(f)| \leq 2\,|L|_1 \mathbb{E}_{\mathrm{P}_\mathcal{X}} |f(X)| < \infty$ for $f \in L_1(\mathrm{P}_\mathcal{X})$ follows from (7.10). Then (7.11) yields $\mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f) \geq -|L|_1 \mathbb{E}_{\mathrm{P}_\mathcal{X}} |f(X)| + \lambda\|f\|_H^2 > -\infty$. □

Lemma 7.11 (Convexity of $L^\star$-risks)**.** *Let L be a (strictly) convex loss. Then $\mathcal{R}_{L^\star,\mathrm{P}} : H \to [-\infty,\infty]$ is (strictly) convex and $\mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda} : H \to [-\infty,\infty]$ is strictly convex.*

Proof. Lemma 7.8 yields that $L^\star$ is (strictly) convex. Trivially $\mathcal{R}_{L^\star,\mathrm{P}}$ is also (strictly) convex. Further $f \mapsto \lambda\|f\|_H^2$ is strictly convex, and hence the map $f \mapsto \mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f) = \mathcal{R}_{L^\star,\mathrm{P}}(f) + \lambda\|f\|_H^2$ is strictly convex. □

Proof of Theorem 3.7. The proof of part (i) is almost identical to the proof of Lemma 3.4. We only have to use Lemma 7.11 instead of Lemma 2.6. *(ii)* This condition implies that $|\mathcal{R}_{L^\star,\mathrm{P}}(f)| < \infty$, see Lemma 7.10, and the assertion follows from *(i)*. □

Proof of Theorem 3.8. As the proof is very similar to the proof of Theorem 3.5, we omit it. We also refer to the proof of Theorem 6 by Christmann *et al.* (2009) for details. The uniqueness of $f_{L^\star,\mathrm{P},\lambda}$ follows immediately from Theorem 3.7(ii), because the boundedness of k guarantees $\|f\|_\infty \leq \|k\|_\infty \|f\|_H$, see (3.5). □

Before we can prove Theorem 5.3, we need the following results.

Lemma 7.12. *Let the Assumptions 2.1, 3.2, and 4.1 be satisfied. Let $f_{L^\star,\mathrm{P},\lambda} \in H$ be any fixed minimizer of $\min_{f\in H}\big(\mathcal{R}_{L^\star,\mathrm{P}}(f) + \lambda\|f\|_H^2\big)$. Then we have, for any $g \in H$,*

$$\mathbb{E}_{\mathrm{P}^2}\Big[D_5L\big(X,Y,\tilde{X},\tilde{Y},f_{L^\star,\mathrm{P},\lambda}(X),f_{L^\star,\mathrm{P},\lambda}(\tilde{X})\big)g(X) + D_6L\big(X,Y,\tilde{X},\tilde{Y},f_{L^\star,\mathrm{P},\lambda}(X),f_{L^\star,\mathrm{P},\lambda}(\tilde{X})\big)g(\tilde{X})\Big] + 2\lambda\langle f_{L^\star,\mathrm{P},\lambda}, g\rangle_H = 0.$$

Proof. As $L^\star$ and λ are fixed, we write $f_\mathrm{P} := f_{L^\star,\mathrm{P},\lambda}$ to shorten the notation in the proof. Let $g \in H$. We define the continuous function

$$\tilde{G} : [-1,1] \to \mathbb{R}, \quad \tilde{G}(t) = \mathcal{R}_{L^\star,\mathrm{P}}(f_\mathrm{P} + tg) + \lambda\|f_\mathrm{P} + tg\|_H^2. \tag{7.17}$$

Recall that the partial derivatives of L and $L^\star$ w.r.t. to the last two arguments are identical because L and $L^\star$ differ only by the term $L(x,y,\tilde{x},\tilde{y},0,0)$. Observe that for $t \neq 0$,

$$\begin{aligned}\frac{\tilde{G}(t) - \tilde{G}(0)}{t} &= \int_{(\mathcal{X}\times\mathcal{Y})^2} \frac{1}{t}\Big(L\big(x,y,\tilde{x},\tilde{y},f_\mathrm{P}(x)+tg(x),f_\mathrm{P}(\tilde{x})+tg(\tilde{x})\big) \\ &\qquad - L\big(x,y,\tilde{x},\tilde{y},f_\mathrm{P}(x),f_\mathrm{P}(\tilde{x})\big)\Big)d\mathrm{P}^2(x,y,\tilde{x},\tilde{y}) \\ &\quad + 2\lambda\langle f_\mathrm{P}, g\rangle_H + t\|g\|_H^2.\end{aligned}$$

By the separate Lipschitz continuity of L, the absolute value of the above integrand is bounded by

$$|L|_1\left(|g(x)| + |g(\tilde{x})|\right) \leq 2|L|_1\|g\|_\infty < \infty\,.$$

Also, for any $(x,y),(\tilde{x},\tilde{y}) \in \mathcal{X}\times\mathcal{Y}$, we have

$$\begin{aligned}&\lim_{t\to 0}\frac{1}{t}\Big(L\big(x,y,\tilde{x},\tilde{y},f_\mathrm{P}(x)+tg(x),f_\mathrm{P}(\tilde{x})+tg(\tilde{x})\big) - L\big(x,y,\tilde{x},\tilde{y},f_\mathrm{P}(x),f_\mathrm{P}(\tilde{x})\big)\Big) \\ &\qquad = D_5L\big(x,y,\tilde{x},\tilde{y},f_\mathrm{P}(x),f_\mathrm{P}(\tilde{x})\big)g(x) + D_6L\big(x,y,\tilde{x},\tilde{y},f_\mathrm{P}(x),f_\mathrm{P}(\tilde{x})\big)g(\tilde{x}).\end{aligned}$$

An application of Lebesgue's dominated convergence theorem yields

$$\begin{aligned}\lim_{t\to 0}\frac{\tilde{G}(t) - \tilde{G}(0)}{t} &= \int_{(\mathcal{X}\times\mathcal{Y})^2}\Big(D_5L\big(x,y,\tilde{x},\tilde{y},f_\mathrm{P}(x),f_\mathrm{P}(\tilde{x})\big)g(x) \\ &\qquad + D_6L\big(x,y,\tilde{x},\tilde{y},f_\mathrm{P}(x),f_\mathrm{P}(\tilde{x})\big)g(\tilde{x})\Big)d\mathrm{P}^2(x,y,\tilde{x},\tilde{y}) + 2\lambda\langle f_\mathrm{P}, g\rangle_H.\end{aligned}$$

Since $\tilde{G}(t) - \tilde{G}(0) \geq 0$ for any t by definition of $\tilde{G}$, we know that the term on the right hand of the above equality is greater or equal to 0. This inequality is also true for the function $-g$. So the desired identity follows. □

Definition 7.13. *We define the local modulus of continuity for the second order derivatives of a pairwise loss function L with respect to the last two variables as*

$$\omega(h)_r := \sup\Big\{ \Big| D_iD_jL\big(x,y,\tilde{x},\tilde{y},f,\tilde{f}\big) - D_iD_jL\big(x,y,\tilde{x},\tilde{y},g,\tilde{g}\big)\Big| : \ x,\tilde{x}\in\mathcal{X}, \\ y,\tilde{y}\in\mathcal{Y}, f,\tilde{f},g,\tilde{g}\in[-r,r], |f-g|\le h, |\tilde{f}-\tilde{g}|\le h, i,j\in\{5,6\}\Big\} \tag{7.18}$$

where $h, r > 0$.

If the sets $\mathcal{X}$ and $\mathcal{Y}$ are bounded, the continuity of the second order derivatives of L implies that $\lim_{h\to 0}\omega(h)_r = 0$ uniformly with respect to $r > 0$.

Let $\mathrm{P},\mathrm{Q}\in\mathcal{M}((\mathcal{X}\times\mathcal{Y})^2)$ and $\varepsilon\in\mathbb{R}$. Define the signed measure $\mathrm{P}_\varepsilon := (1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q}$. Note that P_ε is a probability distribution if $\varepsilon\in[0,1]$.

The key property we need to prove Theorem 5.3 is formulated in the following result.

Theorem 7.14. *Let L satisfy the Assumptions 2.1, 3.2, 4.1, and 4.2. If $\lim_{h\to 0}\omega(h)_r = 0$ for any fixed $r > 0$, then the function $G : \mathbb{R}\times H\to H$ defined by*

$$\begin{aligned} G(\varepsilon,f) &= 2\lambda f + \mathbb{E}_{\mathrm{P}_\varepsilon^2}\Big[D_5L(X,Y,\tilde{X},\tilde{Y},f(X),f(\tilde{X}))\Phi(X) \\ &\qquad + D_6L(X,Y,\tilde{X},\tilde{Y},f(X),f(\tilde{X}))\Phi(\tilde{X})\Big] \end{aligned}$$

is continuously differentiable. Moreover, $\frac{\partial G}{\partial H}(0,f)$ is invertible for any $f\in H$.

Proof. By Theorem 7.3, we only need to show that the partial derivatives $\frac{\partial G}{\partial\varepsilon}$ and $\frac{\partial G}{\partial H}$ are continuous.

To shorten the notation in the proof, we denote the random functions $D_5L\big(X,Y,\tilde{X},\tilde{Y},f(X),f(\tilde{X})\big)$ by D_5L_f and $D_6L\big(X,Y,\tilde{X},\tilde{Y},f(X),f(\tilde{X})\big)$ by D_6L_f, respectively. Analogously we denote second order partial derivatives of L by $D_iD_jL_f$, where $i,j\in\{5,6\}$.

Note that for $\varepsilon\in\mathbb{R}$ and $f\in H$,

$$\begin{aligned} \frac{\partial G}{\partial\varepsilon}(\varepsilon,f) &= -2(1-\varepsilon)\mathbb{E}_{\mathrm{P}\otimes\mathrm{P}}\big(D_5L_f\Phi(X)+D_6L_f\Phi(\tilde{X})\big) \\ &\quad +(1-2\varepsilon)\mathbb{E}_{\mathrm{P}\otimes\mathrm{Q}}\big(D_5L_f\Phi(X)+D_6L_f\Phi(\tilde{X})\big) \\ &\quad +(1-2\varepsilon)\mathbb{E}_{\mathrm{Q}\otimes\mathrm{P}}\big(D_5L_f\Phi(X)+D_6L_f\Phi(\tilde{X})\big) \\ &\quad +2\varepsilon\mathbb{E}_{\mathrm{Q}\otimes\mathrm{Q}}\big(D_5L_f\Phi(X)+D_6L_f\Phi(\tilde{X})\big). \end{aligned} \tag{7.19}$$

Then for $\varepsilon,\tilde{\varepsilon}\in\mathbb{R}$ and $f,\tilde{f}\in H$, we have

$$\begin{aligned} \frac{\partial G}{\partial\varepsilon}(\varepsilon,f) - \frac{\partial G}{\partial\varepsilon}(\tilde{\varepsilon},\tilde{f}) &= \left(\frac{\partial G}{\partial\varepsilon}(\varepsilon,f) - \frac{\partial G}{\partial\varepsilon}(\varepsilon,\tilde{f})\right) + \left(\frac{\partial G}{\partial\varepsilon}(\varepsilon,\tilde{f}) - \frac{\partial G}{\partial\varepsilon}(\tilde{\varepsilon},\tilde{f})\right) \\ &=: \partial G_1 + \partial G_2. \end{aligned}$$

Here

$$\begin{aligned} \partial G_1 &= -2(1-\varepsilon)\mathbb{E}_{\mathrm{P}\otimes\mathrm{P}}\Big[\big(D_5L_f - D_5L_{\tilde{f}}\big)\Phi(X) + \big(D_6L_f - D_6L_{\tilde{f}}\big)\Phi(\tilde{X})\Big] \\ &\quad +(1-2\varepsilon)\mathbb{E}_{\mathrm{P}\otimes\mathrm{Q}}\Big[\big(D_5L_f - D_5L_{\tilde{f}}\big)\Phi(X) + \big(D_6L_f - D_6L_{\tilde{f}}\big)\Phi(\tilde{X})\Big] \\ &\quad +(1-2\varepsilon)\mathbb{E}_{\mathrm{Q}\otimes\mathrm{P}}\Big[\big(D_5L_f - D_5L_{\tilde{f}}\big)\Phi(X) + \big(D_6L_f - D_6L_{\tilde{f}}\big)\Phi(\tilde{X})\Big] \\ &\quad +2\varepsilon\mathbb{E}_{\mathrm{Q}\otimes\mathrm{Q}}\Big[\big(D_5L_f - D_5L_{\tilde{f}}\big)\Phi(X) + \big(D_6L_f - D_6L_{\tilde{f}}\big)\Phi(\tilde{X})\Big]. \end{aligned}$$

Applying (3.6) and (4.2) yields

$$\|\partial G_1\|_H \leq (2|1-\varepsilon| + 2|1-2\varepsilon| + 2|\varepsilon|) \cdot 2 \cdot \left(\|k\|_\infty \cdot 2 \cdot c_{L,2} \|k\|_\infty \|f - \tilde{f}\|_H \right).$$

Moreover,

$$\begin{aligned} \partial G_2 &= 2(\varepsilon - \tilde{\varepsilon}) \mathbb{E}_{\mathrm{P}\otimes\mathrm{P}} \left[D_5 L_{\tilde{f}} \Phi(X) + D_6 L_{\tilde{f}} \Phi(\tilde{X}) \right] \\ &\quad + 2(\tilde{\varepsilon} - \varepsilon) \mathbb{E}_{\mathrm{P}\otimes\mathrm{Q}} \left[D_5 L_{\tilde{f}} \Phi(X) + D_6 L_{\tilde{f}} \Phi(\tilde{X}) \right] \\ &\quad + 2(\tilde{\varepsilon} - \varepsilon) \mathbb{E}_{\mathrm{Q}\otimes\mathrm{P}} \left[D_5 L_{\tilde{f}} \Phi(X) + D_6 L_{\tilde{f}} \Phi(\tilde{X}) \right] \\ &\quad + 2(\varepsilon - \tilde{\varepsilon}) \mathbb{E}_{\mathrm{Q}\otimes\mathrm{Q}} \left[D_5 L_{\tilde{f}} \Phi(X) + D_6 L_{\tilde{f}} \Phi(\tilde{X}) \right]. \end{aligned}$$

Hence by (4.1) we have

$$\|\partial G_2\|_H \leq 8 \cdot 2 \cdot |\varepsilon - \tilde{\varepsilon}| c_{L,1} \|k\|_\infty.$$

Thus

$$\left\| \frac{\partial G}{\partial \varepsilon}(\varepsilon, f) - \frac{\partial G}{\partial \varepsilon}(\tilde{\varepsilon}, \tilde{f}) \right\|_H \leq (4 + 8|\varepsilon|)\, 4\|k\|_\infty^2 c_{L,2} \|f - \tilde{f}\|_H + 16 c_{L,1} \|k\|_\infty |\varepsilon - \tilde{\varepsilon}|.$$

This proves the continuity of the partial derivative $\frac{\partial G}{\partial \varepsilon}$.

The other partial derivative $\frac{\partial G}{\partial H}$ can be expressed with $\varepsilon \in \mathbb{R}$ and $f \in H$ as

$$\begin{aligned} \frac{\partial G}{\partial H}(\varepsilon, f) &= 2\lambda id_H + \mathbb{E}_{\mathrm{P}_\varepsilon \otimes \mathrm{P}_\varepsilon} \Big[D_5 D_5 L_f \Phi(X) \otimes \Phi(X) + D_6 D_5 L_f \Phi(X) \otimes \Phi(\tilde{X}) \\ &\quad + D_5 D_6 L_f \Phi(\tilde{X}) \otimes \Phi(X) + D_6 D_6 L_f \Phi(\tilde{X}) \otimes \Phi(\tilde{X}) \Big]. \end{aligned}$$

To prove its continuity, we first observe, for any $\tilde{f} \in H$,

$$\begin{aligned} &\frac{\partial G}{\partial H}(\varepsilon, f) - \frac{\partial G}{\partial H}(\varepsilon, \tilde{f}) \\ = \; &\mathbb{E}_{\mathrm{P}_\varepsilon \otimes \mathrm{P}_\varepsilon} \Big[D_5 \left(D_5 L_f - D_5 L_{\tilde{f}} \right) \Phi(X) \otimes \Phi(X) + D_6 \left(D_5 L_f - D_5 L_{\tilde{f}} \right) \Phi(X) \otimes \Phi(\tilde{X}) \\ &\quad + D_5 \left(D_6 L_f - D_6 L_{\tilde{f}} \right) \Phi(\tilde{X}) \otimes \Phi(X) + D_6 \left(D_6 L_f - D_6 L_{\tilde{f}} \right) \Phi(\tilde{X}) \otimes \Phi(\tilde{X}) \Big]. \end{aligned}$$

By the definition of the local modulus of continuity for the second order derivatives of L, see Definition 7.13, and the bound $\|\Phi(\tilde{X}) \otimes \Phi(\tilde{X})\|_{\mathcal{L}(H,H)} \leq \|k\|_\infty^2$, if $f, \tilde{f} \in \{g \in H : \|g\|_H \leq r\}$, we obtain the bound

$$\left\| \frac{\partial G}{\partial H}(\varepsilon, f) - \frac{\partial G}{\partial H}(\varepsilon, \tilde{f}) \right\|_{\mathcal{L}(H,H)} \leq 4\|k\|_\infty^2 \omega(\|k\|_\infty \|f - \tilde{f}\|_H)_{r\|k\|_\infty}.$$

The second difference we need to consider is the following sum of four terms, where the integrands

are the same but the factors and the probability measures differ:

$$\begin{aligned}
&\frac{\partial G}{\partial H}(\varepsilon,\tilde f)-\frac{\partial G}{\partial H}(\tilde\varepsilon,\tilde f)\\
=\;&(\tilde\varepsilon-\varepsilon)(2-\tilde\varepsilon-\varepsilon)\mathbb{E}_{\mathrm{P}\otimes\mathrm{P}}\Big[D_5D_5L_f\Phi(X)\otimes\Phi(X)+D_6D_5L_f\Phi(X)\otimes\Phi(\tilde X)\\
&\quad+D_5D_6L_f\Phi(\tilde X)\otimes\Phi(X)+D_6D_6L_f\Phi(\tilde X)\otimes\Phi(\tilde X)\Big]\\
&+(\varepsilon-\tilde\varepsilon)(1-\tilde\varepsilon-\varepsilon)\mathbb{E}_{\mathrm{P}\otimes\mathrm{Q}}\Big[D_5D_5L_f\Phi(X)\otimes\Phi(X)+D_6D_5L_f\Phi(X)\otimes\Phi(\tilde X)\\
&\quad+D_5D_6L_f\Phi(\tilde X)\otimes\Phi(X)+D_6D_6L_f\Phi(\tilde X)\otimes\Phi(\tilde X)\Big]\\
&+(\varepsilon-\tilde\varepsilon)(1-\tilde\varepsilon-\varepsilon)\mathbb{E}_{\mathrm{Q}\otimes\mathrm{P}}\Big[D_5D_5L_f\Phi(X)\otimes\Phi(X)+D_6D_5L_f\Phi(X)\otimes\Phi(\tilde X)\\
&\quad+D_5D_6L_f\Phi(\tilde X)\otimes\Phi(X)+D_6D_6L_f\Phi(\tilde X)\otimes\Phi(\tilde X)\Big]\\
&+(\varepsilon-\tilde\varepsilon)(\tilde\varepsilon+\varepsilon)\mathbb{E}_{\mathrm{Q}\otimes\mathrm{Q}}\Big[D_5D_5L_f\Phi(X)\otimes\Phi(X)+D_6D_5L_f\Phi(X)\otimes\Phi(\tilde X)\\
&\quad+D_5D_6L_f\Phi(\tilde X)\otimes\Phi(X)+D_6D_6L_f\Phi(\tilde X)\otimes\Phi(\tilde X)\Big].
\end{aligned}$$

Let $\varepsilon,\tilde\varepsilon\in[-1,+1]$. Then assumption (4.2) and the bound $\|\Phi(\tilde X)\otimes\Phi(\tilde X)\|_{\mathcal{L}(H,H)}\le\|k\|_\infty^2$ yield the following inequality for the norm of this difference:

$$\left\|\frac{\partial G}{\partial H}(\varepsilon,\tilde f)-\frac{\partial G}{\partial H}(\tilde\varepsilon,\tilde f)\right\|_{\mathcal{L}(H,H)}\le 4c_{L,2}\|k\|_\infty^2|\varepsilon-\tilde\varepsilon|\,(4+4|\varepsilon|+4|\tilde\varepsilon|)\,.$$

Thus we have

$$\begin{aligned}
&\left\|\frac{\partial G}{\partial H}(\varepsilon,f)-\frac{\partial G}{\partial H}(\tilde\varepsilon,\tilde f)\right\|_{\mathcal{L}(H,H)}\\
\le\;&4\|k\|_\infty^2\omega(\|k\|_\infty\|f-\tilde f\|_H)_{r\|k\|_\infty}+4c_{L,2}\|k\|_\infty^2|\varepsilon-\tilde\varepsilon|\,(4+4|\varepsilon|+4|\tilde\varepsilon|)\,.
\end{aligned}$$

Then the continuity of the partial derivative $\frac{\partial G}{\partial H}$ follows. This proves the continuous differentiability of G.

Let $f\in H$. Consider the linear operator $\frac{\partial G}{\partial H}(0,f)$. It can be expressed as

$$\begin{aligned}
&\frac{\partial G}{\partial H}(0,f)\\
=\;&2\lambda id_H+\mathbb{E}_{\mathrm{P}\otimes\mathrm{P}}\Big[D_5D_5L_f\Phi(X)\otimes\Phi(X)+D_6D_5L_f\Phi(X)\otimes\Phi(\tilde X)\\
&\quad+D_5D_6L_f\Phi(\tilde X)\otimes\Phi(X)+D_6D_6L_f\Phi(\tilde X)\otimes\Phi(\tilde X)\Big].
\end{aligned}\tag{7.20}$$

It is important to note that L is a twice continuously differentiable pairwise loss function by Assumption 4.1 which implies that

$$D_6D_5L_f=D_5D_6L_f\,,\qquad f\in H.$$

Hence for any $g,\tilde g\in H$, the following holds

$$\begin{aligned}
&\Big\langle\frac{\partial G}{\partial H}(0,f)(g),\tilde g\Big\rangle_H\\
=\;&2\lambda\langle g,\tilde g\rangle_H+\mathbb{E}_{\mathrm{P}\otimes\mathrm{P}}\Big[D_5D_5L_f\,g(X)\tilde g(X)\\
&+D_6D_5L_f\Big(g(X)\tilde g(\tilde X)+g(\tilde X)g(\tilde X)\Big)+D_6D_6L_f\,g(\tilde X)\tilde g(\tilde X)\Big].
\end{aligned}$$

So the linear operator $\frac{\partial G}{\partial H}(\varepsilon, \tilde{f})(0, f)$ is symmetric. Hence its spectrum lies in the closed interval $[a, b]$ where

$$a := \inf_{\|g\|_H=1} \left\langle \frac{\partial G}{\partial H}(0, f)(g), g \right\rangle_H, \quad b := \sup_{\|g\|_H=1} \left\langle \frac{\partial G}{\partial H}(0, f)(g), g \right\rangle_H.$$

Now, L is a *convex* pairwise loss function due to Assumption 4.2. Therefore, we obtain

$$\left\langle \frac{\partial G}{\partial H}(0, f)(g), g \right\rangle_H \geq 2\lambda \|g\|_H^2, \qquad g \in H.$$

Hence, $a \geq 2\lambda > 0$. This shows that the operator $\frac{\partial G}{\partial H}(\varepsilon, \tilde{f})(0, f)$ is invertible. □

We are now ready for the

Proof of Theorem 4.3. We will first prove part *(i)* by using Lemma 7.12. Let $\mathrm{P} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$. Fix some $\xi \in \mathcal{X}$ and define $g_\xi := \Phi(\xi) = k(\cdot, \xi) \in H$. By the reproducing property (3.4) of the kernel k, we have

$$\langle f_{L^\star,\mathrm{P},\lambda}, g_\xi \rangle_H = \langle f_{L^\star,\mathrm{P},\lambda}, \Phi(\xi) \rangle_H = f_{L^\star,\mathrm{P},\lambda}(\xi). \tag{7.21}$$

Obviously, we also have $g_\xi(x) = \Phi(\xi)(x) = k(x, \xi)$ and $g_\xi(\tilde{x}) = \Phi(\xi)(\tilde{x}) = k(\tilde{x}, \xi)$ for all $x, \tilde{x} \in \mathcal{X}$. Note that the partial derivatives of L and of $L^\star$ with respect of the last two arguments are identical, because L and its shifted version $L^\star$ only differ by the term $L(x, y, \tilde{x}, \tilde{y}, 0, 0)$ which is independent of $f \in H$. Therefore, Lemma 7.12 yields for the function $g_\xi \in H$ the equality

$$\begin{aligned} 0 &= \mathbb{E}_{\mathrm{P}^2}\Big[D_5 L\big(X, Y, \tilde{X}, \tilde{Y}, f_{L^\star,\mathrm{P},\lambda}(X), f_{L^\star,\mathrm{P},\lambda}(\tilde{X})\big) g_\xi(X) \\ &\qquad + D_6 L\big(X, Y, \tilde{X}, \tilde{Y}, f_{L^\star,\mathrm{P},\lambda}(X), f_{L^\star,\mathrm{P},\lambda}(\tilde{X})\big) g_\xi(\tilde{X}) \Big] + 2\lambda \langle f_{L^\star,\mathrm{P},\lambda}, g_\xi \rangle_H \\ &= \mathbb{E}_{\mathrm{P}^2}\big[h_{5,\mathrm{P}}(X, Y, \tilde{X}, \tilde{Y}) k(X, \xi) + h_{6,\mathrm{P}}(X, Y, \tilde{X}, \tilde{Y}) k(\tilde{X}, \xi) \big] + 2\lambda f_{L^\star,\mathrm{P},\lambda}(\xi), \end{aligned}$$

where we used in the last equality the definition of $h_{5,\mathrm{P}}$ and $h_{6,\mathrm{P}}$ from (4.5) and (4.6), respectively, and (7.21). From this we easily conclude that, for *all* $\xi \in \mathcal{X}$,

$$f_{L^\star,\mathrm{P},\lambda}(\xi) = -\frac{1}{2\lambda} \mathbb{E}_{\mathrm{P}^2}\big[h_{5,\mathrm{P}}(X, Y, \tilde{X}, \tilde{Y}) k(X, \xi) + h_{6,\mathrm{P}}(X, Y, \tilde{X}, \tilde{Y}) k(\tilde{X}, \xi) \big]$$

which gives the assertion of part *(i)*.

Let us now prove part *(ii)*. As $L^\star$ and $\lambda \in (0, \infty)$ are fixed, we use the abbreviations $f_\mathrm{P} := f_{L^\star,\mathrm{P},\lambda}$ and $f_\mathrm{Q} := f_{L^\star,\mathrm{Q},\lambda}$ in the proof. The inequality is trivial, if $f_\mathrm{P} = f_\mathrm{Q}$. Hence let us assume that $f_\mathrm{P} \neq f_\mathrm{Q}$. Recall the following well-known inequality from convex analysis. If $g : \mathbb{R}^2 \to \mathbb{R}$ is a convex and totally differentiable function with continuous second partial derivatives, then

$$g(\tilde{t}) - g(t) \geq \langle \nabla g(t), \tilde{t} - t \rangle_{\mathbb{R}^2} \qquad \forall\, t, \tilde{t} \in \mathbb{R}^2,$$

where $\nabla g(t)$ is the gradient of g at t, see Rockafellar (1970, Thm. 25.1, p. 242) for a more general result using sub-gradients. To apply this result, we define, for any fixed $(x, y, \tilde{x}, \tilde{y}) \in (\mathcal{X} \times \mathcal{Y})^2$, the function $g : \mathbb{R}^2 \to \mathbb{R}$, where $g(t_1, t_2) := L^\star(x, y, \tilde{x}, \tilde{y}, t_1, t_2)$ and $t := (t_1, t_2) \in \mathbb{R}^2$. For any $\tilde{t} := (\tilde{t}_1, \tilde{t}_2) \in \mathbb{R}^2$ we thus obtain

$$\begin{aligned} & L^\star(x, y, \tilde{x}, \tilde{y}, \tilde{t}_1, \tilde{t}_2) - L^\star(x, y, \tilde{x}, \tilde{y}, t_1, t_2) \\ \geq\ & D_5 L^\star(x, y, \tilde{x}, \tilde{y}, t_1, t_2)(\tilde{t}_1 - t_1) + D_6 L^\star(x, y, \tilde{x}, \tilde{y}, t_1, t_2)(\tilde{t}_2 - t_2)\,. \end{aligned} \tag{7.22}$$

If we specialize $(t_1, t_2) := (f_{\mathrm{P}}(x), f_{\mathrm{P}}(\tilde{x}))$ and $(\tilde{t}_1, \tilde{t}_2) := (f_{\mathrm{Q}}(x), f_{\mathrm{Q}}(\tilde{x}))$, we obtain from (7.22) the inequality

$$\begin{aligned}
& L^\star(x, y, \tilde{x}, \tilde{y}, f_{\mathrm{Q}}(x), f_{\mathrm{Q}}(\tilde{x})) - L^\star(x, y, \tilde{x}, \tilde{y}, f_{\mathrm{P}}(x), f_{\mathrm{P}}(\tilde{x})) \\
\geq\ & D_5 L^\star(x, y, \tilde{x}, \tilde{y}, f_{\mathrm{P}}(x), f_{\mathrm{P}}(\tilde{x}))(f_{\mathrm{Q}}(x) - f_{\mathrm{P}}(x)) \\
& + D_6\, L^\star(x, y, \tilde{x}, \tilde{y}, f_{\mathrm{P}}(x), f_{\mathrm{P}}(\tilde{x}))(f_{\mathrm{Q}}(\tilde{x}) - f_{\mathrm{P}}(\tilde{x})) \\
=\ & D_5 L(x, y, \tilde{x}, \tilde{y}, f_{\mathrm{P}}(x), f_{\mathrm{P}}(\tilde{x}))(f_{\mathrm{Q}}(x) - f_{\mathrm{P}}(x)) \\
& + D_6 L(x, y, \tilde{x}, \tilde{y}, f_{\mathrm{P}}(x), f_{\mathrm{P}}(\tilde{x}))(f_{\mathrm{Q}}(\tilde{x}) - f_{\mathrm{P}}(\tilde{x})),
\end{aligned}$$

where we used in the last step that L and $L^\star$ only differ by a term which does not depend on the last two arguments. By calculating the corresponding Bochner integral with respect to the product measure Q^2, it follows from the reproducing property (3.4) of k that

$$\begin{aligned}
& \mathcal{R}_{L^\star,\mathrm{Q}}(f_{\mathrm{Q}}) - \mathcal{R}_{L^\star,\mathrm{Q}}(f_{\mathrm{P}}) \qquad (7.23) \\
\geq\ & \Big\langle f_{\mathrm{Q}} - f_{\mathrm{P}}\ ,\ \mathbb{E}_{\mathrm{Q}^2}\Big[D_5 L\big(X, Y, \tilde{X}, \tilde{Y}, f_{\mathrm{P}}(X), f_{\mathrm{P}}(\tilde{X})\big)\Phi(X) \\
& \qquad + D_6 L\big(X, Y, \tilde{X}, \tilde{Y}, f_{\mathrm{P}}(X), f_{\mathrm{P}}(\tilde{X})\big)\Phi(\tilde{X})\Big]\Big\rangle_H \\
=\ & \Big\langle f_{\mathrm{Q}} - f_{\mathrm{P}}\ ,\ \mathbb{E}_{\mathrm{Q}^2}\big[h_{5,\mathrm{P}}(X, Y, \tilde{X}, \tilde{Y})\Phi(X) + h_{6,\mathrm{P}}(X, Y, \tilde{X}, \tilde{Y})\Phi(\tilde{X})\big]\Big\rangle_H , \qquad (7.24)
\end{aligned}$$

where we used in the last step only the definition of $h_{5,\mathrm{P}}$ and $h_{6,\mathrm{P}}$ given in (4.5) and (4.6), respectively. Moreover, an easy calculation shows

$$2\lambda\langle f_{\mathrm{Q}} - f_{\mathrm{P}}, f_{\mathrm{P}}\rangle_H + \lambda\|f_{\mathrm{P}} - f_{\mathrm{Q}}\|_H^2 = \lambda\|f_{\mathrm{Q}}\|_H^2 - \lambda\|f_{\mathrm{P}}\|_H^2 . \qquad (7.25)$$

We thus obtain

$$\begin{aligned}
& \Big\langle f_{\mathrm{Q}} - f_{\mathrm{P}}, \mathbb{E}_{\mathrm{Q}^2}\big[h_{5,\mathrm{P}}(X, Y, \tilde{X}, \tilde{Y})\Phi(X) + h_{6,\mathrm{P}}(X, Y, \tilde{X}, \tilde{Y})\Phi(\tilde{X})\big] + 2\lambda f_{\mathrm{P}}\Big\rangle_H \\
& + \lambda\|f_{\mathrm{P}} - f_{\mathrm{Q}}\|_H^2 \qquad (7.26) \\
\overset{(7.24),(7.23)}{\leq}\ & \mathcal{R}_{L^\star,\mathrm{Q}}(f_{\mathrm{Q}}) - \mathcal{R}_{L^\star,\mathrm{Q}}(f_{\mathrm{P}}) + 2\lambda\langle f_{\mathrm{Q}} - f_{\mathrm{P}}, f_{\mathrm{P}}\rangle_H + \lambda\|f_{\mathrm{P}} - f_{\mathrm{Q}}\|_H^2 \\
\overset{(7.25)}{=}\ & \mathcal{R}_{L^\star,\mathrm{Q}}(f_{\mathrm{Q}}) - \mathcal{R}_{L^\star,\mathrm{Q}}(f_{\mathrm{P}}) + \lambda\|f_{\mathrm{Q}}\|_H^2 - \lambda\|f_{\mathrm{P}}\|_H^2 \\
=\ & \mathcal{R}^{reg}_{L^\star,\mathrm{Q},\lambda}(f_{\mathrm{Q}}) - \mathcal{R}^{reg}_{L^\star,\mathrm{Q},\lambda}(f_{\mathrm{P}}) \quad \leq 0, \qquad (7.27)
\end{aligned}$$

where the term on the left hand side of (7.27) is less than or equal to zero, because the regularized risk with respect to Q is minimized for f_{Q}. Recall that we have

$$2\lambda f_{\mathrm{P}} = -\mathbb{E}_{\mathrm{P}^2}\Big[h_{5,\mathrm{P}}(X, Y, \tilde{X}, \tilde{Y})\Phi(X) + h_{6,\mathrm{P}}(X, Y, \tilde{X}, \tilde{Y})\Phi(\tilde{X})\Big] \qquad (7.28)$$

due to (4.4) in the first part of the representer theorem. If we combine these two facts with the Cauchy-Schwarz inequality we obtain from (7.26) that

$$\begin{aligned}
& \lambda\|f_{\mathrm{P}} - f_{\mathrm{Q}}\|_H^2 \\
\leq\ & -\Big\langle f_{\mathrm{Q}} - f_{\mathrm{P}}\ ,\ \mathbb{E}_{\mathrm{Q}^2}\big[h_{5,\mathrm{P}}(X, Y, \tilde{X}, \tilde{Y})\Phi(X) + h_{6,\mathrm{P}}(X, Y, \tilde{X}, \tilde{Y})\Phi(\tilde{X})\big] + 2\lambda f_{\mathrm{P}}\Big\rangle_H \\
=\ & \Big\langle f_{\mathrm{P}} - f_{\mathrm{Q}}\ ,\ \mathbb{E}_{\mathrm{Q}^2}\big[h_{5,\mathrm{P}}(X, Y, \tilde{X}, \tilde{Y})\Phi(X) + h_{6,\mathrm{P}}(X, Y, \tilde{X}, \tilde{Y})\Phi(\tilde{X})\big] + 2\lambda f_{\mathrm{P}}\Big\rangle_H \\
\overset{(7.28)}{=}\ & \Big\langle f_{\mathrm{P}} - f_{\mathrm{Q}}\ ,\ \mathbb{E}_{\mathrm{Q}^2}\big[h_{5,\mathrm{P}}(X, Y, \tilde{X}, \tilde{Y})\Phi(X) + h_{6,\mathrm{P}}(X, Y, \tilde{X}, \tilde{Y})\Phi(\tilde{X})\big] \\
& \qquad - \mathbb{E}_{\mathrm{P}^2}\big[h_{5,\mathrm{P}}(X, Y, \tilde{X}, \tilde{Y})\Phi(X) + h_{6,\mathrm{P}}(X, Y, \tilde{X}, \tilde{Y})\Phi(\tilde{X})\big]\Big\rangle_H \\
\overset{C.-S.}{\leq}\ & \|f_{\mathrm{P}} - f_{\mathrm{Q}}\|_H \cdot \Big\|\mathbb{E}_{\mathrm{Q}^2}\big[h_{5,\mathrm{P}}(X, Y, \tilde{X}, \tilde{Y})\Phi(X) + h_{6,\mathrm{P}}(X, Y, \tilde{X}, \tilde{Y})\Phi(\tilde{X})\big] \\
& \qquad - \mathbb{E}_{\mathrm{P}^2}\big[h_{5,\mathrm{P}}(X, Y, \tilde{X}, \tilde{Y})\Phi(X) + h_{6,\mathrm{P}}(X, Y, \tilde{X}, \tilde{Y})\Phi(\tilde{X})\big]\Big\|_H .
\end{aligned}$$

After multiplication with $1/\big(\lambda\|f_{\mathrm{P}} - f_{\mathrm{Q}}\|_H\big)$, which is allowed since $f_{\mathrm{P}} \neq f_{\mathrm{Q}}$, we immediately obtain the assertion. □

Proof of Theorem 5.1. The proof only needs some elementary arguments. To shorten the notation, we write $\mathrm{P}_\varepsilon := (1-\varepsilon)\mathrm{P} + \varepsilon\bar{\mathrm{P}}$, $f_{\mathrm{P}} := f_{L^\star,\mathrm{P},\lambda}$, $f_{\mathrm{Q}} := f_{L^\star,\mathrm{Q},\lambda}$, $f_{\mathrm{P}_\varepsilon} := f_{L^\star,\mathrm{P}_\varepsilon,\lambda}$, and $L_f := L(X,Y,\tilde{X},\tilde{Y},f(X),f(\tilde{X}))$, $f \in H$. Denote $f_0 := 0 \in H$. As $L \in [0,c]$ and $\lambda\|f_0\|_H^2 = 0$, we immediately obtain, for all $\mathrm{P} \in \mathcal{M}_1(\mathcal{Z})$,

$$0 \leq R^{reg}(\mathrm{P}) = \inf_{f\in H}\Big(\int L_f\,d\mathrm{P}^2 + \lambda\|f\|_H^2\Big) \leq \int L_{f_0}\,d\mathrm{P}^2 + \lambda\|f_0\|_H^2 \leq c.$$

Hence, there is no need to consider shifted loss functions.

Let us start with part *(i)*. As $f_{\mathrm{Q}} \in H$ and $L \in [0,c]$, we have $R^{reg}(\mathrm{Q}) = \int L_{f_{\mathrm{Q}}}\,d\mathrm{Q}^2 + \lambda\|f_{\mathrm{Q}}\|_H^2$ and $R^{reg}(\mathrm{P}) \leq \int L_{f_{\mathrm{Q}}}\,d\mathrm{P}^2 + \lambda\|f_{\mathrm{Q}}\|_H^2$. Therefore,

$$\begin{aligned} R^{reg}(\mathrm{Q}) - R^{reg}(\mathrm{P}) &\geq \int L_{f_{\mathrm{Q}}}\,d\mathrm{Q}^2 + \lambda\|f_{\mathrm{Q}}\|_H^2 - \int L_{f_{\mathrm{Q}}}\,d\mathrm{P}^2 - \lambda\|f_{\mathrm{Q}}\|_H^2 \\ &= \int L_{f_{\mathrm{Q}}}\,d\mathrm{Q}^2 - \int L_{f_{\mathrm{Q}}}\,d\mathrm{P}^2 \\ &\geq -c\,d_{TV}(\mathrm{Q}^2,\mathrm{P}^2) \geq -2c\,d_{TV}(\mathrm{Q},\mathrm{P}), \end{aligned}$$

where we used in the last inequality that

$$d_{TV}(\mathrm{Q},\mathrm{P}) \leq d_{TV}(\mathrm{Q}^2,\mathrm{P}^2) \leq 2d_{TV}(\mathrm{Q},\mathrm{P}), \quad \mathrm{P},\mathrm{Q} \in \mathcal{M}_1(\mathcal{X}\times\mathcal{Y}), \tag{7.29}$$

see Hoeffding and Wolfowitz (1958, p.709, (4.4) and (4.5)). Analogously, from $f_{\mathrm{P}} \in H$ and $L(x,y,\tilde{x},\tilde{y},t,\tilde{t}) \in [0,c]$, we conclude $R^{reg}(\mathrm{Q}) \leq \int L_{f_{\mathrm{P}}}\,d\mathrm{Q}^2 + \lambda\|f_{\mathrm{P}}\|_H^2$ and $R^{reg}(\mathrm{P}) = \int L_{f_{\mathrm{P}}}\,d\mathrm{P}^2 + \lambda\|f_{\mathrm{P}}\|_H^2$, which yields

$$\begin{aligned} R^{reg}(\mathrm{Q}) - R^{reg}(\mathrm{P}) &\leq \int L_{f_{\mathrm{P}}}\,d\mathrm{Q}^2 + \lambda\|f_{\mathrm{P}}\|_H^2 - \int L_{f_{\mathrm{P}}}\,d\mathrm{P}^2 - \lambda\|f_{\mathrm{P}}\|_H^2 \\ &= \int L_{f_{\mathrm{P}}}\,d\mathrm{Q}^2 - \int L_{f_{\mathrm{P}}}\,d\mathrm{P}^2 \\ &\leq c\,d_{TV}(\mathrm{Q}^2,\mathrm{P}^2) \overset{(7.29)}{\leq} 2c\,d_{TV}(\mathrm{Q},\mathrm{P}). \end{aligned}$$

If we combine both inequalities, we obtain the assertion from part *(i)*.

To part *(ii)*. As $f_{\mathrm{P}_\varepsilon} \in H$ and $L(x,y,\tilde{x},\tilde{y},t,\tilde{t}) \in [0,c]$, we have $R^{reg}(\mathrm{P}_\varepsilon) = \int L_{f_{\mathrm{P}_\varepsilon}}\,d\mathrm{P}_\varepsilon^2 + \lambda\|f_{\mathrm{P}_\varepsilon}\|_H^2$

and $R^{reg}(\mathrm{P}) \le \int L_{f_{\mathrm{P}_\varepsilon}}\, d\mathrm{P}^2 + \lambda \|f_{\mathrm{P}_\varepsilon}\|_H^2$. Therefore,

$$
\begin{aligned}
& R^{reg}(\mathrm{P}_\varepsilon) - R^{reg}(\mathrm{P}) \\
\geq\ & \int L_{f_{\mathrm{P}_\varepsilon}}\, d\mathrm{P}_\varepsilon^2 - \int L_{f_{\mathrm{P}_\varepsilon}}\, d\mathrm{P}^2 \\
=\ & \int\int L_{f_{\mathrm{P}_\varepsilon}}\, d((1-\varepsilon)\mathrm{P} + \varepsilon\bar{\mathrm{P}})\, d((1-\varepsilon)\mathrm{P} + \varepsilon\bar{\mathrm{P}}) - \int\int L_{f_{\mathrm{P}_\varepsilon}}\, d\mathrm{P}\, d\mathrm{P} \\
=\ & \big((1-\varepsilon)^2 - 1\big) \int\int L_{f_{\mathrm{P}_\varepsilon}}\, d\mathrm{P}\, d\mathrm{P} + \varepsilon(1-\varepsilon) \int\int L_{f_{\mathrm{P}_\varepsilon}}\, d\mathrm{P}\, d\bar{\mathrm{P}} \\
& + \varepsilon(1-\varepsilon) \int\int L_{f_{\mathrm{P}_\varepsilon}}\, d\bar{\mathrm{P}}\, d\mathrm{P} + \varepsilon^2 \int\int L_{f_{\mathrm{P}_\varepsilon}}\, d\bar{\mathrm{P}}\, d\bar{\mathrm{P}} \\
=\ & \varepsilon \left(\int\int L_{f_{\mathrm{P}_\varepsilon}}\, d\mathrm{P}\, d\bar{\mathrm{P}} + \int\int L_{f_{\mathrm{P}_\varepsilon}}\, d\bar{\mathrm{P}}\, d\mathrm{P} - 2 \int\int L_{f_{\mathrm{P}_\varepsilon}}\, d\mathrm{P}\, d\mathrm{P} \right) \\
& + \varepsilon^2 \left(\int\int L_{f_{\mathrm{P}_\varepsilon}}\, d\mathrm{P}\, d\mathrm{P} + \int\int L_{f_{\mathrm{P}_\varepsilon}}\, d\bar{\mathrm{P}}\, d\bar{\mathrm{P}} - \int\int L_{f_{\mathrm{P}_\varepsilon}}\, d\bar{\mathrm{P}}\, d\mathrm{P} - \int\int L_{f_{\mathrm{P}_\varepsilon}}\, d\mathrm{P}\, d\bar{\mathrm{P}} \right) \\
=\ & \varepsilon \left(\int \left[\int L_{f_{\mathrm{P}_\varepsilon}}\, d\mathrm{P} \right] d\big(\bar{\mathrm{P}} - \mathrm{P}\big) + \int \left[\int L_{f_{\mathrm{P}_\varepsilon}}\, d\big(\bar{\mathrm{P}} - \mathrm{P}\big) \right] d\mathrm{P} \right) \\
& + \varepsilon^2 \left(\int \left[\int L_{f_{\mathrm{P}_\varepsilon}}\, d\mathrm{P} \right] d\big(\mathrm{P} - \bar{\mathrm{P}}\big) + \int \left[\int L_{f_{\mathrm{P}_\varepsilon}}\, d\bar{\mathrm{P}} \right] d\big(\bar{\mathrm{P}} - \mathrm{P}\big) \right) \\
\overset{(*)}{\geq}\ & -2c\, d_{TV}(\bar{\mathrm{P}}, \mathrm{P})\varepsilon - 2c\, d_{TV}(\bar{\mathrm{P}}, \mathrm{P})\varepsilon^2 \\
=\ & -2c\, d_{TV}(\bar{\mathrm{P}}, \mathrm{P})\varepsilon(1+\varepsilon),
\end{aligned}
$$

where we used in (*) that $\int L_{f_\mathrm{Q}}\, d\mathrm{Q} \le c$ for all $\mathrm{Q} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$. As $f_\mathrm{P} \in H$ and $L(x, y, \tilde{x}, \tilde{y}, t, \tilde{t}) \in [0, c]$, we have $R^{reg}(\mathrm{P}_\varepsilon) \le \int L_{f_\mathrm{P}}\, d\mathrm{P}_\varepsilon^2 + \lambda \|f_\mathrm{P}\|_H^2$ and $R^{reg}(\mathrm{P}) = \int L_{f_\mathrm{P}}\, d\mathrm{P}^2 + \lambda \|f_\mathrm{P}\|_H^2$. Hence, we obtain with the same argumentation as given above that

$$
R^{reg}(\mathrm{P}_\varepsilon) - R^{reg}(\mathrm{P}) \ \leq\ 2c\, d_{TV}(\bar{\mathrm{P}}, \mathrm{P})\varepsilon(1+\varepsilon).
$$

The combination of both inequalities yields the assertion. □

Proof of Theorem 5.3. Partial derivatives of L with respect to the fifth or sixth argument are denoted by $D_5 L$ or $D_6 L$, respectively. In the same manner we denote partial derivatives of L of order two by $D_i D_j L$, where $i, j \in \{5, 6\}$. Recall that due to (7.15), $D_i L^\star = D_i L$ and $D_i D_j L^\star = D_i D_j L$ for $i, j \in \{5, 6\}$.

Fix $\mathrm{Q} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ and $\lambda \in (0, \infty)$. Define $\mathrm{P}_\varepsilon := (1-\varepsilon)\mathrm{P} + \varepsilon\mathrm{Q}$ and denote the product measure $\mathrm{P}_\varepsilon \otimes \mathrm{P}_\varepsilon$ by P_ε^2.

The function $G : \mathbb{R} \times H \to H$ defined by

$$
\begin{aligned}
G(\varepsilon, f) \ :=\ & 2\lambda f + \mathbb{E}_{\mathrm{P}_\varepsilon^2} \Big[D_5 L(X, Y, \tilde{X}, \tilde{Y}, f(X), f(\tilde{X}))\Phi(X) \\
& \qquad + D_6 L(X, Y, \tilde{X}, \tilde{Y}, f(X), f(\tilde{X}))\Phi(\tilde{X}) \Big],
\end{aligned}
$$

where $\varepsilon \in \mathbb{R}$ and $f \in H$, plays a key role in the proof. Since k is bounded by Assumption 3.2, we have $\|f\|_\infty < \infty$ for all $f \in H$, see (3.5). Additionally the partial derivatives $D_5 L$ and $D_6 L$ are continuous and uniformly bounded by Assumption 4.1. It follows from $\| \int f\, d\mathrm{P} \|_H \le \int \|f\|_H\, d\mathrm{P}$

and (3.7) that, for all $\varepsilon \in \mathbb{R}$ and all $f \in H$,

$$\begin{aligned} & \|G(\varepsilon, f)\|_H \\ \leq \; & 2\lambda \|f\|_H \\ & + \int_{(\mathcal{X}\times\mathcal{Y})^2} \Big(\big(|D_5 L(x,y,\tilde{x},\tilde{y}, f(x), f(\tilde{x})| + |D_6 L(x,y,\tilde{x},\tilde{y}, f(x), f(\tilde{x}))| \big) \cdot \\ & \qquad \cdot \sup_{x\in\mathcal{X}} \|\Phi(x)\|_H \Big) \, d\mathrm{P}_\varepsilon^2(x,y,\tilde{x},\tilde{y}) \\ \overset{(3.6)}{\leq} \; & 2\lambda \|f\|_H + 2c_{L,1} \cdot \|k\|_\infty < \infty \,. \end{aligned}$$

Therefore, the map G is well-defined and bounded with respect to the H-norm. Due to (3.5) we have

$$\|G(\varepsilon, f)\|_\infty \leq 2\big(\lambda \|f\|_H + c_{L,1} \cdot \|k\|_\infty\big) \|k\|_\infty \; < \; \infty \,.$$

Note that for $\varepsilon \notin [0,1]$ the H-valued Bochner integral is with respect to a signed measure. Hence Lemma 2.8 yields, for all $\varepsilon \in [0,1]$, that

$$G(\varepsilon, f) = \frac{\partial(\mathcal{R}_{L^\star,\mathrm{P}_\varepsilon}(\cdot) + \lambda\|\cdot\|_H^2)}{\partial H}(f) \,. \tag{7.30}$$

As the map $f \mapsto \mathcal{R}_{L^\star,\mathrm{P}_\varepsilon}(f) + \lambda\|f\|_H^2$ is continuous and strictly convex for all $\varepsilon \in [0,1]$ due to Lemma 2.6, equation (7.30) shows that we have $G(\varepsilon, f) = 0$ if and only if $f = f_{L^\star,\mathrm{P}_\varepsilon,\lambda}$ for such ε. Our aim is to show the existence of a differentiable function $\varepsilon \mapsto f_\varepsilon$ defined on a small interval $(-\delta, \delta)$ for some $\delta > 0$ that satisfies $G(\varepsilon, f_\varepsilon) = 0$ for all $\varepsilon \in (-\delta, \delta)$. Once we have shown the existence of this function we immediately obtain

$$S'_G(\mathrm{P})(\mathrm{Q}) = \frac{\partial f_\varepsilon}{\partial \varepsilon}(0) \,. \tag{7.31}$$

For the existence of this map $\varepsilon \mapsto f_\varepsilon$ we have to check by the implicit function theorem (*cf.* Theorem 7.4) that G is *continuously differentiable* and that $\frac{\partial G}{\partial H}(0, f_{\mathrm{P},\lambda})$ is *invertible*. However, these properties of G were shown in Theorem 7.14. Hence we can apply Theorem 7.4 on implicit functions to see that the map $\varepsilon \mapsto f_\varepsilon$ is differentiable on a small non-empty interval $(-\delta, \delta)$. Therefore, we obtain

$$\begin{aligned} S'_G(\mathrm{P})(\mathrm{Q}) \quad & \overset{(7.31)}{=} \quad \frac{\partial f_\varepsilon}{\partial \varepsilon}(0) \\ & \overset{(7.1)}{=} \quad - \left(\frac{\partial G}{\partial H}(0, f_{L^\star,\mathrm{P},\lambda}) \right)^{-1} \circ \frac{\partial G}{\partial \varepsilon}(0, f_{L^\star,\mathrm{P},\lambda}) \\ & \overset{(7.19),\,(7.20)}{=} \quad -M(\mathrm{P})^{-1} T(\mathrm{Q};\mathrm{P}) \,, \end{aligned}$$

which yields the assertion. □

Proof of Corollary 5.4. The proof follows immediately by specifying Q to the Dirac-measure $\delta_{(x_0,y_0)}$ in Theorem 5.3. □

Proof of Theorem 5.6. We will first prove part *(i)*. Let $\mathrm{P} \in \mathcal{M}_1(\mathcal{X}\times\mathcal{Y})$ be fixed. As $L^\star$ and λ are fixed, we use again the shorter notations

$$h_{i,\mathrm{P}}(X, Y, \tilde{X}, \tilde{Y}) := D_i L^\star\big(X, Y, \tilde{X}, \tilde{Y}, f_{L^\star,\mathrm{P},\lambda}(X), f_{L^\star,\mathrm{P},\lambda}(\tilde{X})\big), \quad i \in \{5, 6\},$$

in the proof, see (4.5) and (4.6).

Let $\mathrm{P}_n \in \mathcal{M}_1(\mathcal{X}\times\mathcal{Y})$, $n \in \mathbb{N}$, be a weakly convergent sequence with limit P, i.e. $\mathrm{P}_n \rightsquigarrow \mathrm{P}$. We know from (5.11) that $\mathrm{P}_n \rightsquigarrow \mathrm{P}$ is equivalent to $d_{\mathrm{BL}}(\mathrm{P}_n, \mathrm{P}) \to 0$, where d_{BL} denotes the bounded

Lipschitz metric, because $\mathcal{X} \times \mathcal{Y}$ is separable by Assumption 2.1. Hence the metric space $(\mathcal{X} \times \mathcal{Y})^2$ is separable, too. The separability guarantees that

$$\mathrm{P}_n \rightsquigarrow \mathrm{P} \qquad \Longleftrightarrow \qquad \mathrm{P}_n^2 \rightsquigarrow \mathrm{P}^2 \qquad (n \to \infty), \tag{7.32}$$

see Billingsley (1999, Thm. 2.8 (ii), p. 23). Note that $\mathrm{P}_n^2 \rightsquigarrow \mathrm{P}^2$ guarantees by definition the convergence $\int g \, d\mathrm{P}_n^2 \to \int g \, d\mathrm{P}^2$ for all continuous and bounded *real-valued* functions $g : (\mathcal{X} \times \mathcal{Y})^2 \to \mathbb{R}$. However, we will need a corresponding convergence result of Bochner integrals where the integrand is a special H*-valued* function.

The second part of Theorem 4.3 (representer theorem) yields

$$\|S(\mathrm{P}_n) - S(\mathrm{P})\|_H := \|f_{L^\star,\mathrm{P}_n,\lambda} - f_{L^\star,\mathrm{P},\lambda}\|_H \tag{7.33}$$

$$\overset{(4.7)}{\leq} \frac{1}{\lambda} \Big\| \int \big[h_{5,\mathrm{P}}(X,Y,\tilde{X},\tilde{Y})\Phi(X) + h_{6,\mathrm{P}}(X,Y,\tilde{X},\tilde{Y})\Phi(\tilde{X})\big] \, d\mathrm{P}_n^2 \tag{7.34}$$

$$- \int \big[h_{5,\mathrm{P}}(X,Y,\tilde{X},\tilde{Y})\Phi(X) + h_{6,\mathrm{P}}(X,Y,\tilde{X},\tilde{Y})\Phi(\tilde{X})\big] \, d\mathrm{P}^2 \Big\|_H ,$$

where $\Phi(X) = k(\cdot, X)$ and $\Phi(\tilde{X}) = k(\cdot, \tilde{X})$. As k is continuous and bounded by Assumption 3.2, the canonical feature map Φ is continuous and bounded, see e.g. Steinwart and Christmann (2008, Lemma 4.23, Lemma 4.29). Furthermore, because the shifted loss function $L^\star$ is by Assumption 4.1 twice continuously differentiable and the partial derivatives are uniformly bounded, it follows that, for every fixed $\mathrm{P} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ and every fixed $\lambda \in (0, \infty)$, the function

$$\begin{aligned} &\Psi_\mathrm{P} : \big((\mathcal{X} \times \mathcal{Y})^2, d_{(\mathcal{X}\times\mathcal{Y})^2}\big) \to (H, d_H), \\ &\Psi_\mathrm{P}(x,y,\tilde{x},\tilde{y}) := h_{5,\mathrm{P}}(x,y,\tilde{x},\tilde{y})\Phi(x) + h_{6,\mathrm{P}}(x,y,\tilde{x},\tilde{y})\Phi(\tilde{x}) \end{aligned} \tag{7.35}$$

is *continuous and bounded*, where $d_H(\cdot,\cdot) := \|\cdot - \cdot\|_H$. We mention that the H-valued function Ψ_P does not depend on P_n. As Ψ_P is continuous and bounded, we obtain from Bourbaki (2004, p. III.40) the following convergence result for Bochner integrals:

$$\mathrm{P}_n^2 \rightsquigarrow \mathrm{P}^2 \qquad \Longrightarrow \qquad \lim_{n\to\infty} \int \Psi_\mathrm{P} \, d\mathrm{P}_n^2 = \int \Psi_\mathrm{P} \, d\mathrm{P}^2 , \tag{7.36}$$

see also Hable and Christmann (2011, Thm. A.1, p. 1000). Combining (7.32)–(7.36), we obtain that $\mathrm{P}_n \rightsquigarrow \mathrm{P}$, which is equivalent to $d_\mathrm{BL}(\mathrm{P}_n, \mathrm{P}) \to 0$ by (5.11), implies $\|S(\mathrm{P}_n) - S(\mathrm{P})\|_H \to 0$, which is the assertion of part *(i)*.

The proof of the second part follows immediately from part *(i)* and the fact that the inclusion $\mathrm{id} : H \to \mathcal{C}_b(\mathcal{X})$ is continuous and bounded, see e.g. Steinwart and Christmann (2008, Lemma 4.28). □

Proof of Corollary 5.7. Let $(D_{n,m})_{m\in\mathbb{N}}$ be a sequence in $(\mathcal{X} \times \mathcal{Y})^n$ which converges to some $D_{n,0} \in (\mathcal{X} \times \mathcal{Y})^n$, if $m \to \infty$. Then, the corresponding sequence of empirical measures weakly converges, i.e. $\mathrm{D}_{n,m} \rightsquigarrow \mathrm{D}_{n,0}$, if $m \to \infty$. Therefore, the assertion follows from Theorem 5.6 and $f_{L^\star,\mathrm{D}_n,\lambda} = S_n(D_n)$. □

Proof of Theorem 5.5. Fix $\lambda \in (0, \infty)$. We will first prove part *(i)*. For any $D_n \in (\mathcal{X} \times \mathcal{Y})^n$ denote its empirical measure by $\mathrm{D}_n := \frac{1}{n}\sum_{i=1}^n \delta_{(x_i,y_i)}$. According to Corollary 5.7, the functions

$$S_n : \big((\mathcal{X} \times \mathcal{Y})^n, d_{(\mathcal{X}\times\mathcal{Y})^n}\big) \to (H, d_H), \quad S_n(D_n) = f_{L^\star,\mathrm{D}_n,\lambda}$$

are continuous and therefore measurable with respect to the corresponding Borel-σ-algebras for every $n \in \mathbb{N}$. The mapping

$$S : \big(\mathcal{M}_1(\mathcal{X} \times \mathcal{Y}), d_\mathrm{BL}\big) \to (H, d_H), \quad S(\mathrm{P}) = f_{L^\star,\mathrm{P},\lambda}, \tag{7.37}$$

is a continuous operator due to Theorem 5.6. Furthermore,

$$S_n(D_n) = S(\mathrm{D}_n) \qquad \forall\ D_n \in (\mathcal{X} \times \mathcal{Y})^n \ \ \forall\ n \in \mathbb{N}.$$

As $\mathcal{X}$ is a separable metric space and the kernel k is continuous, the RKHS H of k is separable, see e.g. Steinwart and Christmann (2008, Lemma 4.33, p. 130). Hence (H, d_H) is a complete and separable metric space.

Therefore, the sequence of RPL estimators $(f_{L^\star,\mathbb{P}_n,\lambda})_{n\in\mathbb{N}}$, where $\mathbb{P}_n := \frac{1}{n}\sum_{i=1}^n \delta_{(X_i,Y_i)}$, is qualitatively robust for all Borel probability measures $\mathrm{P} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ according to Cuevas (1988, Thm. 2), which states: If $(S_n)_{n\in\mathbb{N}}$ is any sequence of estimators which can be represented via a *continuous* operator S, which maps each probability measure P to a value in a *complete and separable metric space* and satisfies (in our notation) $S_n(D_n) = S(\mathrm{D}_n)$, is qualitatively robust for all P. Hence the assertion of part *(i)* is shown.

Let us now prove part *(ii)*. It follows from the first part of Theorem 5.6, that the operator S defined in (7.37) is continuous for all $\mathrm{P} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$. Hence all conditions of Assumption 16.3 in Christmann *et al.* (2013) are satisfied, because $\mathscr{Z} := \mathcal{X} \times \mathcal{Y}$ is a *compact* metric space by assumption of Theorem 5.5(ii) and $\mathscr{W} := H$ is a complete and separable metric space due to the continuity of k by Assumption 3.2, e.g. Steinwart and Christmann (2008, Lemma 4.33, p. 130). Hence, Corollary 16.1 by Christmann *et al.* (2013) is applicable and immediately yields the assertion. We like to note that the compactness of the metric space $\mathscr{Z}$ was used in the proof of the above mentioned Corollary 16.1 to show that the continuous operator S is even uniformly continuous for all $\mathrm{P} \in \mathcal{M}_1(\mathscr{Z})$. □

References

Agarwal, S. and Niyogi, P. (2009). Generalization bounds for ranking algorithms via algorithmic stability. *J. Mach. Learn. Res.*, **10**, 441–474.

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

Billingsley, P. (1999). *Convergence of Probability Measures*. John Wiley & Sons, New York, 2nd edition.

Bourbaki, N. (2004). *Integration I. (Translated from the 1959, 1965, and 1967 French originals by Sterling K. Berberian. Chapters 1–6)*. Springer, Berlin.

Cao, Q., Guo, Z. C., and Ying, T. (2015). Generalization bounds for metric and similarity learning. *to appear in: Machine Learning, Online First. DOI 10.1007/s10994-015-5499-7*.

Castaing, C. and Valadier, M. (1977). *Convex Analysis and Measurable Multifunctions*. Springer, Berlin.

Christmann, A. and Hable, R. (2012). Consistency of support vector machines using additive kernels for additive models. *Comput. Statist. Data Anal.*, **56**, 854–873.

Christmann, A. and Steinwart, I. (2004). On robust properties of convex risk minimization methods for pattern recognition. *J. Mach. Learn. Res.*, **5**, 1007–1034.

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

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**, 311–327.

Christmann, A., Salibían-Barrera, M., and Aelst, S. V. (2013). Qualitative robustness of bootstrap approximations for kernel based methods. In C. Becker, R. Fried, and S. Kuhnt, editors, *Robustness and Complex Data Structures*, pages 277–293. Springer, Heidelberg.

Christmann, A. and Zhou, D. X. (2015). Learning rates for the risk of kernel-based quantile regression estimators in additive models. *Anal. Appl.*, **14**, 449–477.

Clemencon, S., Lugosi, G., and Vayatis, N. (2008). Ranking and empirical minimization of U-statistics. *Ann. Statist.*, **36**, 844–874.

Cucker, F. and Zhou, D. X. (2007). *Learning Theory: An Approximation Theory Viewpoint*. Cambridge University Press, Cambridge.

Cuevas, A. (1988). Qualitative robustness in abstract inference. *J. Statist. Plann. Inference*, **18**, 277–289.

Cuevas, A. and Romo, R. (1993). On robustness properties of bootstrap approximations. *J. Statist. Plann. Inference*, **37**, 181–191.

Denkowski, Z., Migórski, S., and Papageorgiou, N. (2003). *An Introduction to Nonlinear Analysis: Theory*. Kluwer Academic Publishers, Boston.

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

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, **7**, 1–26.

Efron, B. (1982). *The Jackknife, the Bootstrap, and Other Resampling Plans*, volume 38. CBMS Monograph, Society for Industrial and Applied Mathematics, Philadelphia.

Ekeland, I. and Témam, R. (1999). *Convex Analysis and Variational Problems*. SIAM, Philadelphia.

Ekeland, I. and Turnbull, T. (1983). *Infinite-Dimensional Optimization and Convexity*. University of Chicago Press, Chicago.

Fan, J., Hu, T., Wu, Q., and Zhou, D. X. (2016). Consistency analysis of an empirical minimum error entropy algorithm. *Appl. Comput. Harmonic Anal.*, **41**, 164–189.

Feng, Y., Huang, X., Shi, L., Yang, Y., and Suykens, J. (2015). Learning with the maximum correntropy criterion induced losses for regression. *J. Mach. Learn. Res.*, **16**, 993–1034.

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

Hable, R. and Christmann, A. (2013). Robustness versus consistency in ill-posed classification and regression problems. In A. Giusti, G. Ritter, and M. Vichi, editors, *Classification and Data Mining*, pages 27–35. Springer, Berlin.

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

Hampel, F. R. (1971). A general qualitative definition of robustness. *Ann. Math. Statist.*, **42**, 1887–1896.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.*, **69**, 383–393.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust statistics: The Approach Based on Influence Functions*. John Wiley & Sons, New York.

Hoeffding, W. and Wolfowitz, J. (1958). Distinguishability of sets of distributions. the case of independent and identically distributed chance variables. *Ann. Math. Statist.*, **29**, 700–718.

Hu, T., Fan, J., Wu, Q., and Zhou, D. X. (2013). Learning theory approach to minimum error entropy criterion. *J. Mach. Learn. Res.*, **14**, 377–397.

Hu, T., Fan, J., Wu, Q., and Zhou, D. X. (2015). Regularization schemes for minimum error entropy principle. *Anal. Appl.*, **13**, 437–455.

Hu, T., Wu, Q., and Zhou, D. X. (2016). Convergence of Gradient Descent for Minimum Error Entropy Principle in Linear Regression. *IEEE Transactions on Signal Processing*. Accepted for publication subject to a minor revision.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proc. 5th Berkeley Symp.*, **1**, 221–233.

Huber, P. J. (1981). *Robust Statistics*. John Wiley & Sons, New York.

Kallenberg, O. (2002). *Foundations of Modern Probability, 2nd edition*. Springer, New York.

Koroljuk, V. and Borovskich, Y. (1994). *Theory of U-Statistics*. Springer, Dordrecht.

Mukherjee, S. and Zhou, D. X. (2006). Learning coordinate covariances via gradients. *J. Mach. Learn. Res.*, **7**, 519–549.

Poggio, T. and Girosi, F. (1998). A sparse representation for function approximation. *Neural Comput.*, **10**, 1445–1454.

Principe, J. (2010). *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*. Springer, New York.

Rio, E. (2013). On McDiarmid's concentration inequality. *Electron. Commun. Probab.*, **44**, 1–011.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton University Press, Princeton, NJ.

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

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, New York.

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

Tukey, J. W. (1977). *Exploratory Data Analysis. [preliminary edition 1970-1971]*. Addison-Wesley, Reading, MA.

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

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

Wahba, G. (1999). Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV. In B. Schölkopf, C. J. C. Burges, and A. Smola, editors, *Advances in Kernel Methods–Support Vector Learning*, pages 69–88. MIT Press, Cambridge, MA.

Wendland, H. (1995). Piecewise polynomial, positive definite and compactly supported radial basis functions of minimal degree. *Adv. Comput. Math.*, **4**, 389–396.

Wu, Z. (1995). Compactly supported positive definite radial functions. *Adv. Comput. Math.*, **4**, 283–292.

Xing, E., Ng, A., Jordan, M., and Russell, S. (2002). Distance metric learning, with application to clustering with side-information. *Advances in Neural Information Processing Systems*, **15**, 505–512.

Ying, Y. and Zhou, D. X. (2015). Unregularized online learning algorithms with general loss functions. *To appear in: Appl. Comput. Harmonic Anal., Online first:* doi: 10. 1016/ j. acha. 2015. 08. 007.