Statistical Robustness of Empirical Risks in Machine Learning

Shaoyan Guo,† Huifu Xu‡ and Liwei Zhang ‡

May 19, 2020

Abstract

This paper studies convergence of empirical risks in reproducing kernel Hilbert spaces (RKHS). A conventional assumption in the existing research is that empirical training data do not contain any noise but this may not be satisfied in some practical circumstances. Consequently the existing convergence results do not provide a guarantee as to whether empirical risks based on empirical data are reliable or not when the data contain some noise. In this paper, we fill out the gap in a few steps. First, we derive moderate sufficient conditions under which the expected risk changes stably (continuously) against small perturbation of the probability distribution of the underlying random variables and demonstrate how the cost function and kernel affect the stability. Second, we examine the difference between laws of the statistical estimators of the expected optimal loss based on pure data and contaminated data using Prokhorov metric and Kantorovich metric and derive some qualitative and quantitative statistical robustness results. Third, we identify appropriate metrics under which the statistical estimators are uniformly asymptotically consistent. These results provide theoretical grounding for analysing asymptotic convergence and examining reliability of the statistical estimators in a number of well-known machine learning models.

Keywords. Empirical risks, stability analysis, qualitative statistical robustness, quantitative statistical robustness, uniform consistency

1 Introduction

A key element of supervised learning is to find a function which optimally fits to a training set of input-output data and validate its performance with new data. Classical regression models and classification models are typical examples. However, with rapid development of social and economic activities and computer technology, data size increases at an exponential rate. This in turn requires much more powerful optimization models to understand the behavior of complex

---

∗School of Mathematical Sciences, Dalian University of Technology, Dalian, 116024, China. Email: syguo@dlut.edu.cn.
†Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong. Email: hfxu@se.cuhk.edu.hk
‡School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, China. Email: lwzhang@dlut.edu.cn
systems with uncertainties on high dimensional parameter spaces and efficient computational algorithms to solve them. Empirical risk minimization (ERM) is one of them. The essence of ERM models is to use various approximation methods such as sample average approximation (SAA) and stochastic approximation to approximate the expected value of a random function with sampled data. Regularization is often needed since these problems are usually ill-conditioned. Convergence analysis of SAA is well documented in the literature of stochastic optimization, see for instance Ruszczyński and Shapiro [26] and references therein.

In the context of machine learning, the focus is not only on the convergence of statistical estimators to their true counterparts as sample size increases, but also on scalability of the learning algorithms because the size of machine learning problems are often very large under some circumstances [29]. For instance, Norkin and Keyzer [22] consider a general nonparametric regression in finite dimensional RKHS and derive nonasymptotic bounds on the minimization error, exponential bounds on the tail distribution of errors, and sufficient conditions for uniform convergence of kernel estimators to the true (normal) solution with probability one. In the regularized empirical least squares risk minimization, the convergence of estimators can be referred to [9, 10, 23, 30]. Caponnetto and Vito [6] develop a theoretical analysis of the performance of the regularized least-square algorithm in the regression setting when the output space is a general Hilbert space. They use the concept of effective dimension to choose the regularization parameter as a function of the number of samples and derive optimal convergence rates over a suitable class of priors defined by the considered kernel. More recently, Davis and Drusvyatskiy [11] consider a stochastic optimization problem of minimizing population risk, where the loss defining the risk is assumed to be weakly convex. They establish dimension-dependent rates on subgradient estimation in full generality and dimension-independent rates when the loss is a generalized linear model. We refer readers to monograph [10] for the ML models in infinite dimensional spaces for a comprehensive overview.

The problem of characterizing learnability is the most basic question of statistical learning theory. For the case of supervised classification and regression, the learnability is equivalent to uniform convergence of the empirical risk to the expected risk [1, 2]. For the general learning setting, Shalev-Shwartz et al. [29] establish that the stability is the key necessary and sufficient condition for learnability. The existing literature on stability in learning uses many different stability measures. Much of them consider the effect on the optimal value when there exist small changes to the sample such as replacing, adding or removing one instant from the sample, see the review paper [29] for more detail. A conventional assumption in the above stability is that all the instants used in the sample are independent and identically distributed (i.i.d.) and are drawn from the true probability distribution, but this may not be satisfied in some practical circumstances, which means that the empirical training data may contain some noise. Consequently the existing convergence results do not provide a guarantee as to whether empirical risks and kernel learning estimators obtained from solving the ERM models is reliable when the empirical data contain some noise. In this paper, we investigate the issue for learning algorithms on a RKHS from statistical robustness perspective [8, 19] in three main steps.

First, we carry out stability analysis on the optimal expected risk of a generic expected loss minimization problem with respect to perturbation of the probability distribution of the underlying random data. This kind of analysis is well known in stochastic programming (see
and references therein) but not known in machine learning as far as we are concerned. The main challenge in the latter is that the decision variable is often a functional (a function of the underlying random data). In the case when the support of the random data is unbounded, the tail of the probability distribution of the random variables, the tail of the kernel and the tail of the cost function interact and have a joint effect on the stability of the optimal expected risk. We derive moderate sufficient conditions under which the expected risk changes stably (continuously) against small perturbation of the probability distribution and demonstrate how the cost function, the kernel and the random data interactively affect the stability.

Second, we investigate the quality of empirical risk by examining the difference between laws of the statistical estimators of the expected risk based on pure data and contaminated data using metrics on probability measures/distributions. This kind of approach stems from statistics and is applied to risk management where empirical data are used to estimate risk measures of some random losses by Cont et al. [8], Krätschmer et al. [19, 20] and many others. Here we extend the research to machine learning as we believe the approach can be effectively used to look into the interactions between model errors and data errors from statistical point of view, and we do so in both qualitative and quantitative manners.

Third, we discuss convergence of empirical risk which has a vast literature in machine learning. Our focus in this paper is on a generic expected loss minimization model in an infinite dimensional RKHS which requires us to take a particular caution on the tails of the kernel and the cost function when they are both unbounded. We also look into the uniform convergence of the statistical estimator with respect to a set of empirical distributions generated near the true one and identify appropriate metrics under which the statistical estimators are uniformly asymptotically consistent. A combination of all of these results provides some new theoretical grounding for analysing asymptotic convergence and examining reliability of the statistical estimators in a number of well-known machine learning models.

The rest of the paper are organized as follows. Section 2 sets up the background of the model and statistical robustness, Section 3 presents stability of the expected risk against perturbation of the probability distribution, Section 4 details qualitative and quantitative analysis of statistical robustness and Section 5 gives uniform consistency analysis, Section 6 points out some future research.

2 Problem statement

Let $X$ be the input space and $Y$ the output space. The relation between an input $x \in X$ and an output $y \in Y$ is described by a probability distribution $P(x, y)$. Let $Z$ denote the product space $X \times Y$. For each input $x \in X$, output $y \in Y$ and $z = (x, y)$, let $c(z, f(x))$ denote the loss caused by the use of $f$ as a model for the unknown process producing $y$ from $x$ and $\mathbb{E}_P[c(z, f(x))]$ the statistical average of the losses. If $P$ is known, then the problem of learning is down to find an optimal model such that the average loss is minimized, i.e.,

$$\min_{f \in \mathcal{F}} R(f) := \mathbb{E}_P[c(z, f(x))] = \int_Z c(z, f(x))P(dz),$$

(2.1)
where $\mathcal{F}$ is some functional class to be specified. Let $\vartheta(P)$ denote the optimal value and $\mathcal{F}^*(P)$ the set of optimal solutions in (2.1). By indicating their dependence on $P$, we will investigate the effect of a perturbation of $P$ in forthcoming discussions. Without loss of generality, we assume throughout the paper that $c(z, f(x))$ takes non-negative value. In practice, $\mathcal{F}$, $Z$ and $c(\cdot, \cdot)$ are known to learners. Here we list a few examples [29].

- **Regression.** Let $Z = X \times Y$ where $X$ and $Y$ are bounded subsets of $\mathbb{R}^n$ and $\mathbb{R}$ respectively, let $\mathcal{F}$ be a set of functions $f : \mathbb{R}^n \to \mathbb{R}$ and $c(z, f(x)) = L(f(x) - y)$, where $L(\cdot)$ is a loss function. Specific interesting cases include squared loss function $L(t) = \frac{1}{2}t^2$, $\epsilon$-insensitive loss function $L(t) = \max\{0, |t| - \epsilon\}$ with $\epsilon > 0$, hinge loss function $L(t) = \max\{0, 1 - t\}$ and log-loss function $L(t) = \log(1 + e^{-t})$ in various regression and support vector machine models, see [25].

- **Binary Classification.** Let $Z = X \times \{0, 1\}$ and $\mathcal{F}$ be a set of functions $f : X \to \{0, 1\}$, let $c(z, f(x)) = 1_{f(x)\neq y}$. Here $c(\cdot, \cdot)$ is a 0–1 loss function, measuring whether $f$ misclassifies the pair $(x, y)$.

- **Density estimation.** Let $Z$ be a subset of $\mathbb{R}^n$ and $\mathcal{F}$ be a set of bounded probability densities on $Z$, let $c(z, f(x)) = -\log(f(z))$. Here $c(\cdot, \cdot)$ is simply the negative log-likelihood of an instance $z$ according to the hypothesis model density $f$.

### 2.1 Reproducing kernel Hilbert space

The nature of functions $f$ in (2.1) needs to be specified. Let $\mathcal{H}$ denote a class of functions $f : X \to Y$. $\mathcal{H}$ is called hypotheses space if $f$ is restricted to $\mathcal{H}$. This is because the choice of $\mathcal{H}$ is based on hypotheses of the structure of these functions.

**Definition 2.1** Let $\mathcal{H}(X)$ be a Hilbert space of functions with inner product $\langle \cdot, \cdot \rangle$ and $k(\cdot, \cdot) : X \times X \to \mathbb{R}$ be a kernel, that is, there is a feature map $\Phi : X \to \mathcal{H}$ such that $k(x, x) = \langle \Phi(x), \Phi(x) \rangle$. $\mathcal{H}(X)$ is said to be a reproducing kernel Hilbert space (RKHS for short) if there is a kernel function $k(\cdot, \cdot) : X \times X \to \mathbb{R}$ such that: (a) $k(\cdot, x) \in \mathcal{H}(X)$ for all $x \in X$ and (b) $f(x) = \langle f, k(\cdot, x) \rangle$ for all $f \in \mathcal{H}(X)$ and $x \in X$. The corresponding scalar product and norm are denoted by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|_k$ respectively.

A kernel: $k : X \times X \to \mathbb{R}$ is said to be symmetric if $k(x, t) = k(t, x)$ for each $x, t \in X$, positive definite symmetric (PDS) if for any $x_1, \cdots, x_m \in X$ the matrix $[k(x_i, x_j)]_{ij} \in \mathbb{R}^{m \times m}$ is symmetric positive semidefinite (SPSD). A kernel $k$ is called Mercer kernel if it is continuous, symmetric and positive semidefinite.

Examples of Mercer kernels abound. Here we list some of them.

- **Polynomial kernel:** $k(x_1, x_2) = (\gamma \langle x_1, x_2 \rangle + 1)^d$, $\forall x_1, x_2 \in \mathbb{R}^N$, where $\gamma > 0$ is a constant, $d \in \mathbb{N}$ and $\mathbb{N}$ denotes the set of positive integers.

- **Gussian kernel:** $k(x_1, x_2) = e^{-\gamma \|x_1 - x_2\|^2}$, $\forall x_1, x_2 \in \mathbb{R}^N$, where $\gamma > 0$ is a constant.
• **Sigmoid kernel**: \( k(x_1, x_2) = \tanh(a(x_1, x_2) + b), \forall x_1, x_2 \in \mathbb{R}^N, \) where \( a, b > 0 \) are constants, \( \tanh(t) = \frac{e^t - e^{-t}}{e^t + e^{-t}} \) is the hyperbolic tangent function.

Let \( k : X \times X \to \mathbb{R} \) be a positive definite symmetric kernel (Mercer kernel). Then there exists a Hilbert space \( \mathcal{H}_k(X) \) and a mapping \( \Phi : X \to \mathcal{H}_k(X) \) such that

\[
k(x, x') = \langle \Phi(x), \Phi(x') \rangle, \forall x, x' \in X.
\]

Moreover \( \mathcal{H}_k(X) \) has the reproducing property, see \([15, \text{Theorem } 5.2]\). If we let

\[
\mathcal{F} = \left\{ \sum_{i=1}^{n} \alpha_i k(x_i, \cdot) : n \in \mathbb{N}, \alpha_i \in \mathbb{R}, x_i \in X \right\}
\]

with the inner product

\[
\left\langle \sum_{i=1}^{n} \alpha_i k(x_i, \cdot), \sum_{j=1}^{n} \beta_j k(x_j, \cdot) \right\rangle = \sum_{i,j=1}^{n} \alpha_i \beta_j k(x_i, x_j),
\]

then \( \mathcal{F} \) can be completed into the RKHS, see \([3]\). Algorithms working with kernels usually perform minimization of a cost function on a ball of the associated RKHS of the form

\[
\mathcal{F}_\sigma = \left\{ \sum_{j=1}^{N} a_j k(x_j, \cdot) : N \in \mathbb{N}, \sum_{i,j=1}^{N} a_i a_j k(x_i, x_j) \leq \sigma^2, x_1, \cdots, x_N \in X \right\}. \quad (2.2)
\]

Throughout the paper, we assume that a positive definite symmetric kernel \( k(\cdot, \cdot) \) is given and \( \mathcal{H}_k \) is the RKHS associated with \( k \). The functional class \( \mathcal{F} \) in (2.1) and (2.3) is a subset of \( \mathcal{H}_k \).

### 2.2 Sample average approximation

In practice, the true probability distribution \( P \) is unknown, but it is possible to obtain an independent and identically distributed (i.i.d.) samples \( \{z^i = (x^i, y^i)\}_{i=1}^{N} \) generated by \( P \), which is known as training data. Given the sample, the goal of machine learning is to find a function \( f : X \to Y \) such that \( f \) solves

\[
\min_{f \in \mathcal{F}} \mathbb{E}_{P_N}[c(z, f(x))] := \frac{1}{N} \sum_{i=1}^{N} c(z^i, f(x^i)), \quad (2.3)
\]

where

\[
P_N(\cdot) := \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{z^i}(\cdot) \quad (2.4)
\]

denotes the empirical probability measure/distribution and \( \mathbb{1}_{z^i}(\cdot) \) denotes the Dirac measure at \( z^i \). Let \( \vartheta(P_N) \) denote the optimal value (empirical risk), \( R_{P_N}(f) \) the objective function, and \( \mathcal{F}_{P_N}^* \) the set of optimal solutions of the sample average approximation problem (2.3). Let
\(f_N(P_N) \in \mathcal{F}^*_P\) denote an optimal solution of \((2.3)\). Then \(f_N(P_N)\) is called the estimator and the framework generating \(f_N(P_N)\) is called learning algorithm. Notice that from sampling point of view, we may write \(\hat{\vartheta}_N(z^1, \ldots, z^N)\) and \(\hat{f}_N(z^1, \ldots, z^N)\) for \(\vartheta(P_N)\) and \(f_N(P_N)\) respectively to indicate their dependence on the samples.

From computationally perspective, problem \((2.3)\) is often ill-conditioned. The issue can be addressed by adopting a simple Tikhonov regularization approach:

\[
\vartheta(P_N, \lambda_N) = \min_{f \in \mathcal{F}} R^\lambda_P(f) := E_{P_N}[c(z, f(x))] + \lambda_N \|f\|_F^2, \tag{2.5}
\]

where \(\lambda_N > 0\) is a regularization parameter. In general \(\lambda_N\) is driven to \(0\) but the choice of the value may affect the rate of convergence. A number of papers have been devoted to this, see for instance Brehney and Huang [4] for logistic regression models in a finite dimensional space, Cucker and Smale [3] and Caponnetto and Vito [1] for regularized least squares models in infinite dimensional RKHS. Let \(\mathcal{F}^*_{P_N, \lambda_N}\) denote the set of optimal solutions in \((2.5)\) and \(f_N(P_N, \lambda_N) \in \mathcal{F}^*_{P_N, \lambda_N}\) an optimal solution. In the case that \(c(z, f(x))\) is convex in \(f\) for almost every \(z\), \(\mathcal{F}^*_{P_N, \lambda_N}\) is a singleton. By virtue of the representer theorem (see Kimeldorf and Wahba [18], [27, Theorem 4.2]), problem \((2.5)\) has a solution which takes the following form \(f_N^{\lambda_N}(x) = \sum_{j=1}^N \alpha_j k(x_j, x)\) and by the reproducing property \((2.2)\), \(\|f_N^{\lambda_N}\|_F^2 = \langle f_N^{\lambda_N}, f_N^{\lambda_N} \rangle = \sum_{i,j=1}^N \alpha_i \alpha_j k(x_i, x_j)\). In that case the feasible set may be written as \((2.2)\). As we commented earlier, here we may also write \(\hat{\vartheta}_N(z^1, \ldots, z^N, \lambda_N)\) and \(\hat{f}_N(z^1, \ldots, z^N, \lambda_N)\) for \(\vartheta(P_N, \lambda_N)\) and \(f_N(P_N, \lambda_N)\) respectively to indicate their dependence on the samples.

### 2.3 Contamination of the training data

The current research of machine learning is mostly focused on the case that sample data are generated by the true probability distribution \(P\) which means that they do not contain any noise. This assumption may not be satisfied in some data-driven problems. Let \(\bar{z}^1, \ldots, \bar{z}^N\) denote the perceived data which may contain noise and

\[
Q_N(\cdot) := \frac{1}{N} \sum_{i=1}^N \mathbb{1}_{\bar{z}^i}(\cdot) \tag{2.6}
\]

be the respective empirical distribution. Instead of solving \((2.5)\), we solve, in practice,

\[
\min_{f \in \mathcal{F}} E_{Q_N}[c(z, f(x))] + \lambda_N \|f\|_F^2. \tag{2.7}
\]

Let \(R^\lambda_{Q_N}(f), \vartheta(Q_N, \lambda_N)\) and \(f_N(Q_N, \lambda_N)\) denote respectively the objective function, the optimal value and the optimal solution of problem \((2.7)\). We are then concerned with the quality of the learning model estimator \(f_N(Q_N, \lambda_N)\) and the associated empirical risk \(\vartheta(Q_N, \lambda_N)\).

There are two ways to proceed the research. One is to look into convergence of the statistical quantities as the sample size \(N\) increases and the regularization parameter \(\lambda_N\) goes to zero. Assume without loss of generality that the samples are independent and identically generated. By law of large numbers, \(Q_N\) converges to some probability distribution \(Q\) and subsequently

\[
f_N(Q_N, \lambda_N) \to f(Q) \quad \text{and} \quad \vartheta(Q_N, \lambda_N) \to \vartheta(Q). \tag{2.8}
\]
On the other hand, if we regard $Q$ as a perturbation of the true unknown probability distribution $P$, then we need to investigate whether

$$f(Q) \to f(P) \quad \text{and} \quad \vartheta(Q) \to \vartheta(P)$$

as $Q$ approaches $P$. The former is known as asymptotic convergence/consistency and the latter is known as stability in the literature of stochastic programming [25]. However, if we want to establish

$$f_N(Q_N, \lambda_N) \to f(P) \quad \text{and} \quad \vartheta(Q_N, \lambda_N) \to \vartheta(P),$$

then we require not only (2.9) but also (2.8) to hold uniformly for all $Q$ near $P$. This will be more demanding than the currently established convergence results.

The other is to examine the discrepancy between $f_N(Q_N, \lambda_N)$ and $f_N(P_N, \lambda_N)$ ($\vartheta(Q_N, \lambda_N)$ and $\vartheta(P_N, \lambda_N)$) via law of these estimators. The latter should be understood as estimators when the noise in the samples is detached (an ideal case). This kind of research is in alignment with qualitative robustness in the literature of robust statistics and risk measurement, see [8, 13, 19, 20] and references therein. We will give a formal definition in Section 4.

In both steps leading towards statistical robustness of $\vartheta(\cdot)$, we will need to restrict the perturbation of the probability measure from $P$ to the space of $\phi$-topology of weak convergence instead of usual weak convergence.

### 2.4 $\phi$-weak topology

We recall some basic concepts and results about weak topology which are needed for the analysis. The materials are mainly extracted from [7], we refer readers to [7, Chapter 2] and references therein for a more comprehensive discussion on the subject.

**Definition 2.2** Let $\phi : Z \to [0, \infty)$ be a continuous function and

$$\mathcal{M}_Z^\phi := \left\{ P \in \mathcal{P}(Z) : \int_Z \phi(z)P(dz) < \infty \right\},$$

where $\mathcal{P}(Z)$ is the set of all probability measures on the measurable space $(Z, \mathcal{B}(Z))$ with Borel sigma algebra $\mathcal{B}(Z)$ of $Z$.

$\mathcal{M}_Z^\phi$ defines a subset of probability measures in $\mathcal{P}(Z)$ which satisfies the generalized moment condition of $\phi$.

**Definition 2.3 (\(\phi\)-weak topology)** Let $\phi : Z \to [0, \infty)$ be a gauge function, that is, $\phi \geq 1$ holds outside a compact set. Define $C_Z^\phi$ the linear space of all continuous functions $h : Z \to \mathbb{R}$ for which there exists a positive constant $c$ such that

$$h(z) \leq c(\phi(z) + 1), \forall z \in Z.$$
The $\phi$-weak topology, denoted by $\tau_\phi$, is the coarsest topology on $M^\phi_Z$ for which the mapping $g_h : M^\phi_Z \to \mathbb{R}$ defined by
$$g_h(P) := \int_Z h(z)P(dz), \ h \in C^\phi_Z$$
is continuous. A sequence $\{P_t\} \subset M^\phi_Z$ is said to converge $\phi$-weakly to $P \in M^\phi_Z$ written $P_t \overset{\phi}{\to} P$ if it converges with respect to (w.r.t.) $\tau_\phi$.

From the definition, we can see immediately that $\phi$-weak convergence implies weak convergence under usual topology of weak convergence. We denote the latter by $P_t \overset{w}{\to} P$. Moreover, it follows by [7, Corollary 2.66] that the $\phi$-weak topology on $M^\phi_Z$ is generated by the metric $d_\phi : M^\phi_Z \times M^\phi_Z \to \mathbb{R}$ defined by
$$d_\phi(P', P'') := d_{\text{Prok}}(P', P'') + \left| \int_Z \phi dP' - \int_Z \phi dP'' \right|,$$
where $d_{\text{Prok}} : \mathcal{P}(Z) \times \mathcal{P}(Z) \to \mathbb{R}_+$ is the Prokhorov metric defined as follows:
$$d_{\text{Prok}}(P', P'') := \inf \{ \epsilon > 0 : P'(A) \leq P''(A^c) + \epsilon \text{ for all } A \in \mathcal{B}(Z) \},$$
where $A' := A + B_r(0)$ denotes the Minkowski sum of $A$ and the open ball centred at 0 (w.r.t. the norm in $Z$). When $\phi \equiv 1$, the second term in (2.11) disappears and consequently $d_\phi(P', P'') = d_{\text{Prok}}(P', P'')$. In that case, the $\phi$-weak topology reduces to the usual topology of weak convergence (defined through bounded continuous functions). Equivalence between the two topologies may be established over a set which satisfies some uniform integration conditions, see [7, Lemma 2.66] and the reference therein.

**Definition 2.4 (Fortet-Mourier metric) Let**
$$\mathcal{F}_p(Z) := \{ \psi : Z \to \mathbb{R} : |\psi(z) - \psi(\tilde{z})| \leq c_p(z, \tilde{z})\|z - \tilde{z}\|, \forall z, \tilde{z} \in Z \},$$
where $\| \cdot \|$ denotes some norm on $Z$ and $c_p(z, \tilde{z}) := \max \{1, \|z\|, \|\tilde{z}\| \}^{p-1}$ for all $z, \tilde{z} \in Z$ and $p \geq 1$ describes the growth of the local Lipschitz constants. The $p$-th order Fortet-Mourier metric over $\mathcal{P}(Z)$ is defined by
$$\zeta_p(P, Q) := \sup_{\psi \in \mathcal{F}_p(Z)} \left| \int_Z \psi(z)P(dz) - \int_Z \psi(z)Q(dz) \right|.$$

Fortet-Mourier metric is well-known in stochastic programming. The unique feature of the metric is that it is induced by a class of locally Lipschitz continuous functions with specified modulus and rate of growth. In the case when $p = 1$, it reduces to Kontorovich metric. We refer readers to see Römisch [25] for a comprehensive overview of the topic. From the definition, we can see that
$$\zeta_p(P, Q) \leq \mathbb{E}_{P \times Q}[c_p(z, \tilde{z})\|z - \tilde{z}\|],$$
where $P \times Q$ denotes the joint probability distribution of $z$ and $\tilde{z}$. In the case when $P$ and $Q$ are empirical distributions generated by i.i.d. sample, we have
$$\mathbb{E}_{P \times Q}[c_p(z, \tilde{z})\|z - \tilde{z}\|] = \frac{1}{N^2} \sum_{i,j=1}^N c_p(z^i, \tilde{z}^j)\|z^i - \tilde{z}^j\|.$$

The latter may be used to give an estimate of $\zeta_p(P, Q)$ if we are able to obtain the i.i.d. samples in practice.

## 3 Stability analysis

In this section, we investigate how the model risk of problem (2.1) is affected by a small perturbation of the probability measure $P$. This kind of research is well known in the literature of stochastic programming [25] but not in machine learning as far as we are concerned. We proceed with some technical assumptions which stipulate the properties of the cost function and the kernel.

**Assumption 3.1**

(a) For any compact subset $Z_0$ of $Z$, let $X_0$ be its orthogonal projection on $X$. The set of functions $\{k(\cdot, x) : x \in X_0\}$ are equi-continuous on $X_0$, i.e., for any $\epsilon > 0$, there exists a constant $\eta > 0$ such that

$$\|k(\cdot, x') - k(\cdot, x)\|_k < \epsilon, \forall x, x' \in X_0 : \|x' - x\| < \eta,$$

where $\| \cdot \|$ is some norm on $X$.

(b) There is a positive constant $\beta$ such that $\|f\|_k \leq \beta$ for all $f \in F$.

**Remark 3.1**

To see how Assumption 3.1 (a) can be possibly satisfied, we recall the notion of calmness of kernel introduced by Shafieezadeh-Abadeh et al. [28, Assumption 25]. The kernel function $k$ is said to be calm from above, if there exists a concave smooth growth function $g : \mathbb{R}^+ \to \mathbb{R}^+$ with $g(0) = 0$ and $g'(t) \geq 1$ for all $t \in \mathbb{R}^+$ such that

$$\sqrt{k(x', x') - 2k(x, x') + k(x, x)} \leq g(\|x - x'\|), \forall x, x \in X.$$

Under the calmness condition, there exists $\eta > 0$ such that

$$\|k(\cdot, x') - k(\cdot, x)\|_k = \sqrt{(k(\cdot, x') - k(\cdot, x), k(\cdot, x') - k(\cdot, x))}$$

$$= \sqrt{k(x', x') - k(x, x') + k(x, x) - k(x, x')}$$

$$\leq g(\|x - x'\|) < \epsilon$$

for all $x, x'$ with $\|x - x'\| \leq \eta$. The last inequality is due to the fact that the growth function $g$ is continuous with $g(0) = 0$, thus for any $\epsilon > 0$, there exists a positive constant $\eta > 0$ such that $|g(t) - g(0)| = |g(t)| < \epsilon$ for $|t| < \eta$. The calmness condition is non-restrictive, which can be satisfied in the following cases for $X = \mathbb{R}^n$, see [28, Example 1].

- **Linear kernel**: for $k(x_1, x_2) = \langle x_1, x_2 \rangle$, $g(t) = t$.
- **Gaussian kernel**: for $k(x_1, x_2) = e^{-\gamma \|x_1 - x_2\|^2_2}$, $g(t) = \max\{\sqrt{2\gamma}, 1\}t$.
- **Laplacian kernel**: for $k(x_1, x_2) = e^{-\gamma \|x_1 - x_2\|_1}$, $g(t) = \sqrt{2\gamma t \sqrt{n}}$ if $0 \leq t \leq \gamma \sqrt{n}/2$ and $g(t) = t + \gamma \sqrt{n}/2$ otherwise.
• **Polynomial kernel:** the kernel \( k(x_1, x_2) = (\gamma(x_1, x_2) + 1)^d \) with \( \gamma > 0 \) and \( d \in \mathbb{N} \) fails to satisfy the calmness condition if \( X \) is unbounded and \( d > 1 \), in which case \( \sqrt{k(x_1, x_1) - 2k(x_1, x_2) + k(x_2, x_2)} \) grows superlinearly. If \( X \subset \{x \in \mathbb{R}^n : \|x\|_2 \leq R\} \) for some \( R > 0 \), however, the polynomial kernel is calm with respect to the growth function

\[
g(t) = \begin{cases} 
\max\{\frac{1}{\pi R} \sqrt{2(\gamma R^2 + 1)^d}, 1\}t & \text{d is even,} \\
\max\{\frac{1}{\pi R} \sqrt{2(\gamma R^2 + 1)^d - 2(1 - \gamma R^2)^d}, 1\}t & \text{d is odd.}
\end{cases}
\]  

Assumption 3.2 (b) may be guaranteed by restricting the set of feasible solutions to lie within a ball, see [22, Assumption D].

**Assumption 3.2** The cost function \( c(\cdot, \cdot) \) satisfies the following properties.

(a) There is a gauge function \( \phi(\cdot) \) such that

\[
c(z, f(x)) \leq \phi(z), \forall z \in Z \text{ and } f \in F,
\]

where \( \phi(z) \to \infty \) as \( \|z\| \to \infty \).

(b) \( c(z, y) : Z \times Y \to \mathbb{R} \) is continuous.

**Remark 3.2** Condition (a) is known as a growth condition where \( \phi(z) \) controls the growth of the cost function as \( \|z\| \) goes to infinity. It is trivially satisfied when \( Z \) is compact. Our focus here is on the case that \( Z \) is unbounded. Obviously \( \phi \) depends on the concrete structure of \( c(\cdot, \cdot) \). Consider for example \( c(z, f(x)) = \frac{1}{2}\|y - f(x)\|^2 \). Then

\[
c(z, f(x)) \leq \|y\|^2 + \|f(x)\|^2 = \|y\|^2 + |\langle f, k(\cdot, x) \rangle|^2 \\
\leq \|y\|^2 + \|f\|_k^2 \|k(\cdot, x)\|_k^2.
\]

Moreover, under Assumption 3.1 (b), i.e., \( \|f\|_k \leq \beta \), we can work out an explicit form of \( \phi \) for some specific kernels.

- If \( k \) is a Linear kernel, then \( \|k(\cdot, x)\|_k^2 = |k(x, x)| = \|x\|^2 \) and \( \phi(z) := \|y\|^2 + \beta^2 \|x\|^2 \);
- If \( k \) is a Gaussian or Laplacian kernel, then \( \|k(\cdot, x)\|_k^2 = 0 \) and \( \phi(z) := \|y\|^2 \);
- If \( k \) is a Polynomial kernel, then \( \|k(\cdot, x)\|_k^2 = (\gamma \|x\|^2 + 1)^d \) and

\[
\phi(z) := \|y\|^2 + \beta^2 (\gamma \|x\|^2 + 1)^d.
\]  

(3.15)

From the examples above, we can see that \( \phi \) captures not only the growth of the cost function \( c(\cdot, \cdot) \) but also the kernel. The growth rate of \( \phi \) at the tail in turn affects the topology of weak convergence to be used in the stability analysis in the next theorem.

**Theorem 3.1** Under Assumptions 3.1 and 3.2, the following holds for any \( p \geq 1 \),

\[
\lim_{\varrho(P') \to \varrho P} \vartheta(P') = \vartheta(P).
\]  

(3.16)
Proof. Since \((\mathcal{M}^{\phi}_{Z}, \tau_{\phi})\) is a Polish space, it suffices to show that (3.16) holds for any sequence \(\{P_l\} \subset \mathcal{M}^{\phi}_{Z}\) with \(P_l \overset{\phi}{\rightarrow} P \in \mathcal{M}^{\phi}_{Z}\). First, \(P_l \overset{\phi}{\rightarrow} P\) implies that \(P_l \overset{w}{\rightarrow} P\) and

\[
\lim_{l \to \infty} \int_{Z} \phi^{p}(z) P_l(dz) = \int_{Z} \phi(z)^{p} P(dz).
\]

Moreover, by [24, Theorem 3.1], equi-continuous of \(k\). Thus, by [7, Lemma 2.61], for any \(\epsilon > 0\), there exists a positive constant \(M > 1\) such that

\[
\int_{Z} \phi^{p}(z) 1_{(M, \infty)}(\phi^{p}(z)) P(dz) < \epsilon \tag{3.17}
\]

and

\[
\sup_{l \in \mathbb{N}} \int_{Z} \phi^{p}(z) 1_{(M, \infty)}(\phi^{p}(z)) P_l(dz) < \epsilon, \tag{3.18}
\]

where \(1_{(M, \infty)}(t) = 1\) if \(t \in (M, \infty)\) otherwise 0. Since \(\phi\) is coercive, i.e., \(\phi^{p}(z) \to \infty\) as \(\|z\| \to \infty\), then exists a compact continuity set \(Z_{M} \subset Z\) of \(P\) such that \(Z \setminus Z_{M} \subset \{z \in Z : \phi(z) > M\}\). Here the continuity set means that \(P(\partial Z_{M}) = 0\) where \(\partial Z_{M}\) denotes the boundary of \(Z_{M}\).

Let

\[
\mathcal{G} := \{g : g(z) := c(z, f(x)) \text{ for } f \in F\}
\]

and

\[
\mathcal{G}_{M} := \{g_{M} : Z_{M} \to \mathbb{R}| g_{M}(z) := g(z) \text{ for } z \in Z_{M}, g \in \mathcal{G}\}.
\]

It follows from Assumption 3.2(a) that for each \(g_{M} \in \mathcal{G}_{M}\) and \(z \in Z_{M}\), \(|g_{M}(z)| \leq \sup_{z \in Z_{M}} \phi(z) < \infty\), which implies that \(\mathcal{G}_{M}\) is uniformly bounded.

Next, we prove that \(\mathcal{G}_{M}\) is equi-continuous over \(Z_{M}\). By the reproducing property of the kernel \(k(\cdot, \cdot)\), i.e., \(f(x) = \langle f, k(\cdot, x) \rangle\) for every \(f \in F\), we have

\[
|f(x') - f(x)| = |\langle f, k(\cdot, x') \rangle - \langle f, k(\cdot, x) \rangle| \leq \|f\|_{k} \|k(\cdot, x') - k(\cdot, x)\|_{k}
\]

\[
\leq \beta \|k(\cdot, x') - k(\cdot, x)\|_{k}. \tag{3.19}
\]

The equicontinuity of \(k(\cdot, x)\) over \(X_{M}\) (under Assumption 3.1(a)) ensures the equicontinuity of \(\mathcal{F}\) over \(X_{M}\). Moreover, under Assumption 3.2(b), \(\mathcal{G}_{M}\) is also equi-continuous because \(c(\cdot, \cdot)\) is uniformly continuous over any compact set.

Let \(Q_{l}, Q\) be measures on \(Z_{M}\) defined by \(Q_{l}(A) = P_{l}(A)\) and \(Q(A) = P(A)\) respectively. Since \(Z_{M}\) is a continuity set of \(P\), then \(P_{l} \overset{w}{\rightarrow} P\) imply \(Q_{l} \overset{w}{\rightarrow} Q\). Since \(\mathcal{G}_{M}\) is uniformly bounded and equi-continuous, by [24, Theorem 3.1],

\[
\lim_{l \to \infty} \sup_{g_{M} \in \mathcal{G}_{M}} \left| \int_{Z_{M}} g_{M}(z) Q_{l}(dz) - \int_{Z_{M}} g_{M}(z) Q(dz) \right| = 0. \tag{3.20}
\]

On the other hand, under the growth condition (3.14), (3.17) and (3.18) imply

\[
\int_{Z \setminus Z_{M}} |g(z)| P(dz) \leq \int_{Z \setminus Z_{M}} \phi^{p}(z) P(dz) \leq \int_{Z} \phi^{p}(z) 1_{(M, \infty)}(\phi^{p}(z)) P(dz) < \epsilon \tag{3.21}
\]
and
\[
\sup_{l\in\mathbb{N}} \int_{Z\setminus Z_M} |g(z)|P_l(dz) \leq \sup_{l\in\mathbb{N}} \int_{Z\setminus Z_M} \phi^p(z)P_l(dz) \leq \sup_{l\in\mathbb{N}} \int_Z \phi^p(z)1_{(M,\infty)}(\phi^p(z))P_l(dz) < \epsilon.
\]

(3.22)

Together with (3.20), we have
\[
|\vartheta(P_l) - \vartheta(P)| \leq \sup_{f\in\mathcal{F}} \left| \int_Z c(z, f(x))P_l(dz) - \int_Z c(z, f(x))P(dz) \right|
= \sup_{g\in\mathcal{G}} \left| \int_Z g(z)P_l(dz) - \int_Z g(z)P(dz) \right|
\leq \sup_{g\in\mathcal{G}} \left| \int_{Z_M} g(z)P_l(dz) - \int_{Z_M} g(z)P(dz) \right| + \int_{Z\setminus Z_M} |g(z)|P_l(dz)
\leq \sup_{g_M\in\mathcal{G}_M} \left| \int_{Z_M} g_M(z)Q_l(dz) - \int_{Z_M} g_M(z)Q(dz) \right| + 2\epsilon < 3\epsilon
\]

for sufficiently large \(l\). The proof is complete.

The theorem tells us that \(\vartheta(Q)\) is close to \(\vartheta(P)\) when \(Q\) is perturbed from \(P\) under the \(\tau_{\phi^p}\)-weak topology for any fixed \(p \geq 1\). Since the empirical probability measure \(P_N \in \mathcal{M}_Z^{\phi^p}\), we have
\[
\lim_{N\to\infty} \vartheta(P_N) = \vartheta(P)
\]
almost surely. The topological structure of set \(\mathcal{M}_Z^{\phi^p}\) affects the stability of \(\vartheta(\cdot)\): a larger \(\mathcal{M}_Z^{\phi^p}\) means that \(\vartheta(\cdot)\) remains stable w.r.t. a greater freedom of perturbation from \(P\). In the case when \(Z\) is a compact set, \(\mathcal{M}_Z^{\phi^p} = \mathcal{P}(Z)\), which means \(\vartheta(\cdot)\) remains stable for any perturbation of the probability measure from \(P\) locally. The tail behaviour of \(c(z, f(x))\) affects the structure of \(\mathcal{M}_Z^{\phi^p} = \mathcal{P}(Z)\), we explain this through next example.

**Example 3.1** Consider the least squares regression model with Polynomial kernel. By (3.15)
\[
\mathcal{M}_Z^{\phi} = \left\{ P \in \mathcal{P}(Z) : \int_Z \left[ \|y\|^2 + \beta^2(\gamma\|x\|^2 + 1)^d \right] P(dz) < \infty \right\}
= \left\{ P \in \mathcal{P}(Z) : \int_Z \|y\|^2 P(dz) < \infty, \int_Z \|x\|^{2d} P(dz) < \infty \right\}.
\]

We can see from the formulation above that a larger \(d\) requires a thinner tail of \(P\) and hence a smaller set of \(\mathcal{M}_Z^{\phi}\), consequently the stability result is valid for a smaller class of probability distributions.

*In the case of Gaussian kernel or Laplacian kernel,*
\[
\mathcal{M}_Z^{\phi} = \left\{ P \in \mathcal{P}(Z) : \int_Z \|y\|^2 P(dz) < \infty \right\},
\]
which is the set of probability measures with finite second order moment of \(y\).
Finally, we note that our stability result should be distinguished from those in [29] where stability is used to examine the difference of the costs resulting from kernel learning estimators based on different samples.

4 Statistical robustness

We now move on to discuss statistical robustness of the machine learning model [2,5]. To ease the exposition, let $Z^\otimes N$ denote the Cartesian product $Z \otimes \cdots \otimes Z$ and $\mathcal{B}(Z)^\otimes N$ its Borel sigma algebra. Let $P^\otimes N$ denote the probability measure on the measurable space $(Z^\otimes N, \mathcal{B}(Z)^\otimes N)$ with marginal $P$ and $Q^\otimes N$ with marginal $Q$. We will consider statistical estimators mapping from $(Z^\otimes N, \mathcal{B}(Z)^\otimes N)$ to $\mathbb{R}$ and examine their convergence under $Q^\otimes N$ and $P^\otimes N$.

4.1 Qualitative robustness

We begin by a formal definition of statistical estimator $T(\cdot, \lambda)$ parameterized by $\lambda$, where $T(\cdot, \lambda)$ maps from a subset of $\mathcal{M} \subset \mathcal{P}(Z)$ to $\mathbb{R}$. To ease the exposition, we write $z^N$ for $(z^1, \cdots, z^N)$ and $\hat{T}_N(z^N, \lambda_N)$ for $T(P_N, \lambda_N)$ for fixed sample size $N$. The following definition is based on Krätschmer et al. [19, Definition 2.11].

**Definition 4.1 (Statistical robustness)** Let $\mathcal{M} \subset \mathcal{P}(Z)$ be a set of probability measures and $d_{\phi}$ be defined as in [2,11] for some gauge function $\phi : Z \to \mathbb{R}$, let $\{\lambda_N\}$ be a sequence of parameters. A parameterized statistical estimator $T(\cdot, \lambda_N)$ is said to be robust on $\mathcal{M}$ with respect to $d_{\phi}$ and $d_{\text{Prok}}$ if for all $P \in \mathcal{M}$ and $\epsilon > 0$, there exist $\delta > 0$ and $N_0 \in \mathbb{N}$ such that for all $N \geq N_0$

$$Q \in \mathcal{M}, \ d_{\phi}(P, Q) \leq \delta \implies d_{\text{Prok}} \left( P^\otimes N \circ \hat{T}_N(\cdot, \lambda_N)^{-1}, Q^\otimes N \circ \hat{T}_N(\cdot, \lambda_N)^{-1} \right) \leq \epsilon.$$

In this definition, $P^\otimes N \circ \hat{T}_N(\cdot, \lambda_N)^{-1}$ and $Q^\otimes N \circ \hat{T}_N(\cdot, \lambda_N)^{-1}$ are two probability distributions of random variable $\hat{T}_N(\cdot, \lambda_N)$ mapping from probability spaces $(Z^\otimes N, \mathcal{B}(Z)^\otimes N, P^\otimes N)$ and $(Z^\otimes N, \mathcal{B}(Z)^\otimes N, Q^\otimes N)$ respectively to $\mathbb{R}$, and the Prokhorov metric is used to measure the difference of the two distributions (also known as laws in the literature [8, 19]). The statistical robustness requires the difference under the Prokhorov metric to be small when the difference between $P$ and $Q$ is small under $d_{\phi}$. The definition relies heavily on the adoption of the two metrics. In Cont et al. [8], the authors use Lévy metric for both. Krätschmer et al. [19] argue that the Levy metric underestimates the impact of the tail distributions of $P$ and $Q$ and subsequently propose to use $d_{\phi}$ to replace the Lévy metric. Since the former is tighter than the later, it means the perturbation under $d_{\phi}$ is more restrictive and hence enables one to examine finer difference between the laws of the statistical estimators.

Statistical robustness is also called qualitative robustness in this paper in that there is no explicit quantitative relationship between $\epsilon$ and $\delta$. To establish the statistical robustness, we need the following Uniform Glivenko-Cantelli property.
Definition 4.2 (Uniform Glivenko-Cantelli property) Let $\phi$ be a gauge function and $d_{\phi}$ be defined as in (2.11). Let $M$ be a subset of $M^\phi_Z$, the metric space $(M, d_{\phi})$ is said to have Uniform Glivenko-Cantelli (UGC) property if for every $\epsilon > 0$ and $\delta > 0$, there exists $N_0 \in \mathbb{N}$ such that for all $P \in M$

$$P^\otimes N \left[ z^N : d_{\phi}(P, P_N) \geq \delta \right] \leq \epsilon, \forall P \in M$$

(4.24)

for all $N \geq N_0$.

Recall that $P_N$ is constructed through i.i.d. samples generated by random variable $z$ with probability distribution $P$. The UGC property requires that for all $P \in M$, their empirical probability measures converge to their true counterparts uniformly as the sample size goes to infinity. The convergence is under $d_{\phi}$ which means not only the weak convergence but also convergence of the $\phi$ moments, the latter captures the tails of $P$.

Theorem 4.1 (Statistical robustness) Let $\{P_N\}$ be a sequence of empirical probability measures defined by (2.4) and $M^\phi_{Z,\kappa}$ be the class of all $P \in \mathcal{P}(Z)$ such that

$$\int_Z \phi(z)^p P(dz) \leq \kappa,$$

(4.25)

for $\kappa \geq 0$ and $p > 1$. Let Assumptions $3.A$ and $3.B$ hold, $\lambda_N \to 0$ as $N \to \infty$. Then for any $\epsilon > 0$, there exist positive numbers $\delta > 0$ and $N_0 \in \mathbb{N}$ such that when $Q \in M \subset M^\phi_{Z,\kappa}$, $d_{\phi}(P, Q) \leq \delta$, we have

$$d_{\text{Prok}}\left(P^\otimes N \circ \hat{\phi}_N(\cdot, \lambda_N)^{-1}, Q^\otimes N \circ \hat{\phi}_N(\cdot, \lambda_N)^{-1}\right) \leq \epsilon$$

(4.26)

for all $N \geq N_0$ and $\lambda_N \leq \frac{6\beta}{\delta \beta}$, where $\hat{\phi}_N(z^N, \lambda_N) = \hat{\phi}(P_N, \lambda_N)$ denotes the optimal value of problem (2.7).

Proof. By triangle inequality

$$d_{\text{Prok}}\left(P^\otimes N \circ \hat{\phi}_N(\cdot, \lambda_N)^{-1}, Q^\otimes N \circ \hat{\phi}_N(\cdot, \lambda_N)^{-1}\right) \leq d_{\text{Prok}}\left(P^\otimes N \circ \hat{\phi}_N(\cdot, \lambda_N)^{-1}, 1_{\inf_{f \in \mathcal{F}} R_p(f)}\right) + d_{\text{Prok}}\left(1_{\inf_{f \in \mathcal{F}} R_p(f)}, 1_{\inf_{f \in \mathcal{F}} R_Q(f)}\right) + d_{\text{Prok}}\left(1_{\inf_{f \in \mathcal{F}} R_Q(f)}, Q^\otimes N \circ \hat{\phi}_N(\cdot, \lambda_N)^{-1}\right),$$

where $1_a$ denotes the Dirac measure at $a \in \mathbb{R}$. By Theorem 3.1 for the given $\epsilon$ there exists a constant $\delta_0 > 0$ such that

$$d_{\text{Prok}}\left(1_{\inf_{f \in \mathcal{F}} R_p(f)}, 1_{\inf_{f \in \mathcal{F}} R_Q(f)}\right) \leq \frac{\epsilon}{3}, \forall Q \in M \subset M^\phi_{Z,\kappa} \text{ with } d_{\phi^p}(P, Q) \leq \delta_0.$$ 

So we are left to show that

$$d_{\text{Prok}}\left(P^\otimes N \circ \hat{\phi}_N(\cdot, \lambda_N)^{-1}, 1_{\inf_{f \in \mathcal{F}} R_p(f)}\right) \leq \frac{\epsilon}{3}$$

(4.27)
and

\[ \text{d}_{\mathcal{P}_{\text{rok}}} \left( 1_{\inf_{f \in \mathcal{F}} R_Q(f)}, Q^{N} \circ \hat{\vartheta}_{N} \cdot, \lambda_{N} \right)^{-1} \leq \frac{\epsilon}{3} \] (4.28)

for \( N \) sufficiently large. By Strassen’s theorem [16, (4.27) and (4.28)] are implied respectively by

\[ P^{N} \left[ z^{N} : \hat{\vartheta}_{N}(z^{N}, \lambda_{N}) - \inf_{f \in \mathcal{F}} R_{P}(f) \geq \frac{\epsilon}{3} \right] \leq \frac{\epsilon}{3} \] (4.29)

and

\[ Q^{N} \left[ \tilde{z}^{N} : \hat{\vartheta}_{N}(\tilde{z}^{N}, \lambda_{N}) - \inf_{f \in \mathcal{F}} R_{Q}(f) \geq \frac{\epsilon}{3} \right] \leq \frac{\epsilon}{3}. \] (4.30)

Using the definition of the optimal values, (4.29) and (4.30) can be rewritten respectively as

\[ P^{N} \left[ z^{N} : \inf_{f \in \mathcal{F}} E_{P_{N}} \{c(z, f(x)) + \lambda_{N} \| f \|_{k}^{2} \} - \inf_{f \in \mathcal{F}} R_{P}(f) \geq \frac{\epsilon}{3} \right] \leq \frac{\epsilon}{3} \] (4.31)

and

\[ Q^{N} \left[ \tilde{z}^{N} : \inf_{f \in \mathcal{F}} E_{Q_{N}} \{c(z, f(x)) + \lambda_{N} \| f \|_{k}^{2} \} - \inf_{f \in \mathcal{F}} R_{Q}(f) \geq \frac{\epsilon}{3} \right] \leq \frac{\epsilon}{3}. \] (4.32)

Note that we may set \( N_{0} \in \mathbb{N} \) sufficiently large such that \( \lambda_{N} \leq \frac{\epsilon}{6 \beta} \) for all \( N \geq N_{0} \). Consequently the two inequalities above are implied by

\[ P^{N} \left[ z^{N} : \inf_{f \in \mathcal{F}} R_{P_{N}}(f) - \inf_{f \in \mathcal{F}} R_{P}(f) \geq \frac{\epsilon}{6} \right] \leq \frac{\epsilon}{3} \] (4.33)

and

\[ Q^{N} \left[ \tilde{z}^{N} : \inf_{f \in \mathcal{F}} R_{Q_{N}}(f) - \inf_{f \in \mathcal{F}} R_{Q}(f) \geq \frac{\epsilon}{6} \right] \leq \frac{\epsilon}{3}. \] (4.34)

or equivalently

\[ P^{N} \left[ z^{N} : \hat{\vartheta}_{N}(z^{N}) - \vartheta(P) \geq \frac{\epsilon}{6} \right] \leq \frac{\epsilon}{3} \] (4.35)

and

\[ Q^{N} \left[ \tilde{z}^{N} : \hat{\vartheta}_{N}(\tilde{z}^{N}) - \vartheta(Q) \geq \frac{\epsilon}{6} \right] \leq \frac{\epsilon}{3}. \] (4.36)

By Theorem 3.1 there exists a constant \( \delta > 0 \) such that when \( d_{\phi}(P', P) < 2\delta, |\vartheta(P') - \vartheta(P)| < \frac{\epsilon}{12} \). On the other hand, it follows by [20 Corollary 3.5] that \( (\mathcal{M}_{Z, \phi, d_{\phi}}) \) has the UGC property which implies that

\[ Q^{N} \left[ d_{\phi}(Q_{N}, Q) \geq \delta \right] \leq \frac{\epsilon}{3} \] (4.37)
for all $Q \in M_{\phi, Z, \kappa}$ including $Q = P$. This shows \((4.35)\) when $N_0$ is chosen sufficiently large. To show \((4.36)\), let $\dd(Q, P) \leq \delta$. Then

\begin{align*}
\frac{\epsilon}{3} &\geq Q \circ N \left[ \tilde{z}^N : \dd(Q_N, Q) \geq \delta \right] \\
&\geq Q \circ N \left[ \tilde{z}^N : \dd(Q_N, P) \geq \delta + \dd(Q, P) \right] \\
&\geq Q \circ N \left[ \tilde{z}^N : \dd(Q_N, P) \geq 2\delta \right] \\
&\geq Q \circ N \left[ \tilde{z}^N : |\vartheta(Q_N) - \vartheta(P)| \geq \frac{\epsilon}{12} \right] \\
&\geq Q \circ N \left[ \tilde{z}^N : |\vartheta(Q_N) - \vartheta(Q)| \geq |\vartheta(P) - \vartheta(Q)| + \frac{\epsilon}{12} \right] \\
&\geq Q \circ N \left[ \tilde{z}^N : |\vartheta(Q_N) - \vartheta(Q)| \geq \frac{\epsilon}{6} \right], \forall Q \in M_{\phi, Z, \kappa}.
\end{align*}

The conclusion follows.

We make a few comments about the conditions and results of this theorem.

First, the set $M_{\phi, Z, \kappa}$ differs from $M_{\phi, Z}$ in that the former imposes a bound for the moment value uniformly for all $P \in M_{\phi, Z, \kappa}$ whereas the latter does not have such uniformity. This is because we need the UGC property of $(M_{\phi, Z, \kappa}, \dd)$ in order for us to apply \([20, \text{Corollary 3.5}]\). For example, in the least squares regression model with polynomial kernel, we have

\begin{align*}
M_{\phi, Z, \kappa} &= \left\{ P \in \mathcal{P}(Z) : \int_Z [\|y\|^2 + \beta^2(\gamma \|x\|^2 + 1)^d] P(dz) < \kappa \right\}.
\end{align*}

In the case of Gaussian kernel or Laplacian kernel,

\begin{align*}
M_{\phi, Z, \kappa} &= \left\{ P \in \mathcal{P}(Z) : \int_Z \|y\|^{2p} P(dz) < \kappa \right\}.
\end{align*}

Second, by \((4.36)\), we can obtain for any $\epsilon > 0$, there exist constants $\delta > 0$ and $N_0 \in \mathbb{N}$ such that

\begin{align*}
Q \in M, \dd(P, Q) \leq \delta &\implies Q \circ N \left[ \tilde{z}^N : |\vartheta(Q_N) - \vartheta(Q)| \geq \frac{\epsilon}{6} \right] \leq \frac{\epsilon}{3}
\end{align*}

for $N \geq N_0$. This implies uniform convergence of $\vartheta(Q_N)$ to $\vartheta(Q)$ for all $Q$ near $P$ as opposed to pointwise convergence (for each fixed $Q$) in stochastic programming. The uniformity does not come out for free: it restricts both $P$ and $Q$ to the $\phi$-weak topological space of probability measures.

Third, in practice, since $P$ is unknown, it is difficult to identify $\delta$ for a specified $\epsilon$. The usefulness of \((4.26)\) should be understood as that it provides a theoretical guarantee: if the training data are generated by some probability distribution $Q$ which is close to the true distribution $P$, and $Q$ satisfies moment condition \((4.25)\) (which may be examined through empirical data), then the optimal value obtained with the perceived data is close to the one with real data. There are potentially two ways to move forward the research. One is to derive quantitative statistical robustness under some additional conditions in which case the relationship between $\epsilon$ and $\delta$ may be explicitly established, we will come back to this in the next subsection. The other is to use...
the training data to construct an ambiguity set of probability distributions and use the latter to develop a model which is robust both in preference and in brief. This will effectively create a robust mechanism to mitigate the risk arising from noise in perceived data. We leave this for future research.

4.2 Quantitative robustness

In the previous section, there is no explicit relationship between $\epsilon$ and $\delta$ in the qualitative robustness result. In this section, we address the issue under the following additional conditions.

**Assumption 4.1** The cost function $c(z, f(x))$ satisfies the following property:

$$|c(z, f(x)) - c(z', f(x'))| \leq c_p(z, z') \|z - z'\|, \forall z, z' \in Z, f \in F,$$

where $c_p(z, z') := \max\{1, \|z\|, \|z'\|\}^{p-1}$ and $p \geq 1$ is a fixed positive number.

To see how the assumption may be satisfied, we consider the case that $c(z, f(x))$ is locally Lipschitz continuous with modulus being bounded by $L(z)$, then

$$|c(z, f(x)) - c(z', f(x'))| \leq \max\{L(z), L(z')\}(\|z - z'\| + |f(x) - f(x')|), \forall z, z' \in Z.$$

Under Assumption 3.1 (b) and the calmness condition in Remark 3.1

$$|f(x) - f(x')| = |\langle f(k(\cdot, x)) - f(k(\cdot, x'))\rangle| \leq \beta \|k(\cdot, x) - k(\cdot, x')\| \leq \beta g(\|x - x'\|).$$

Consequently we have

$$|c(z, f(x)) - c(z', f(x'))| \leq \max\{L(z), L(z')\}(\|z - z'\| + \beta g(\|x - x'\|)), \forall z, z' \in Z.$$

(4.40)

In Example 4.1 we will explain in detail how $L(\cdot)$ may be figured out and in a combination with specific form of function $g(\cdot)$, inequality (4.40) leads to inequality (4.39) for some specific cost functions and kernel functions in regression models.

We now return to our discussion on the quantitative description of the discrepancy between $P^N \circ \hat{\vartheta}_N(\cdot, \lambda_N)^{-1}$ and $Q^N \circ \hat{\vartheta}_N(\cdot, \lambda_N)^{-1}$. Our idea is to use Kantorovich metric to measure the difference, i.e., $\text{dl}_{K,1}\left(P^N \circ \hat{\vartheta}_N(\cdot, \lambda_N)^{-1}, Q^N \circ \hat{\vartheta}_N(\cdot, \lambda_N)^{-1}\right)$, which can be converted to the estimate of the difference between $P^N$ and $Q^N$ under some metric by the $\zeta$-metric of $P$ and $Q$. The next technical result prepares for such a conversion.

**Lemma 4.1** Let $\hat{z} := (z^1, \cdots, z^N) \in Z^N$ and

$$\Psi := \left\{\psi : Z^N \to \mathbb{R} : |\psi(\hat{z}) - \psi(\bar{z})| \leq \frac{1}{N} \sum_{j=1}^{N} c_p(z^j, \bar{z}^j)\|z^j - \bar{z}^j\|\right\}.$$

Let $\text{dl}_\psi(P^N, Q^N) = \sup_{\psi \in \Psi} \left|\int_Z \psi(z)P^N(dz) - \int_Z \psi(z)Q^N(dz)\right|$. Then

$$\text{dl}_\psi(P^N, Q^N) \leq \zeta_p(P, Q).$$
The proof is established in [31 Lemma 4.1] which is an extension of [13 Lemma 1] (which is presented when \( p = 1 \)). Here we include a proof for self-containedness. Let \( \tilde{z} := \{z^1, \ldots, z^j\} \) and \( \bar{z}^j := \{z^1, \ldots, z^{j-1}, z^{j+1}, \ldots, z^N\} \). For any \( P_1, \ldots, P_N \in \mathcal{P}(Z) \) and any \( j \in \{1, \ldots, N\} \), denote
\[
P_{-j}(d\tilde{z}^j) := P_1(dz^1) \cdots P_{j-1}(dz^{j-1}) P_{j+1}(dz^{j+1}) \cdots P_N(dz^N)
\]
and \( h_{\bar{z}^j}(z^j) := \int_{Z^\otimes(N-1)} \psi(\tilde{z}^j, z^j) P_{-j}(d\bar{z}^j) \). Then
\[
|h_{\bar{z}^j}(\tilde{z}^j) - h_{\bar{z}^j}(\hat{z}^j)| \leq \int_{Z^\otimes(N-1)} |\psi(\tilde{z}^j, z^j) - \psi(\hat{z}^j, z^j)| P_{-j}(d\bar{z}^j)
\]
\[
\leq \int_{Z^\otimes(N-1)} \frac{1}{N} c_p(\tilde{z}^j, z^j) \|z^j - \hat{z}^j\| P_{-j}(d\bar{z}^j)
\]
\[
\leq \frac{1}{N} c_p(\tilde{z}^j, z^j) \|z^j - \hat{z}^j\|.
\]
Let \( \mathcal{W} \) denote the set of functions \( h_{\bar{z}^j}(z^j) \) generated by \( \psi \in \Psi \). By the definition of \( \mathcal{d}_\psi \) and the \( p \)-th order Fortet-Mourier metric,
\[
\mathcal{d}_\psi(P_{-j} \times \hat{P}_j, P_{-j} \times \hat{P}_j) = \sup_{\psi \in \Psi} \left| \int_{Z} \int_{Z^{\otimes(N-1)}} \psi(\tilde{z}^j, z^j) P_{-j}(d\bar{z}^j) \hat{P}_j(dz^j) \right|
\]
\[
- \int_{Z} \int_{Z^{\otimes(N-1)}} \psi(\tilde{z}^j, z^j) P_{-j}(d\bar{z}^j) \hat{P}_j(dz^j) \left| \int_{Z} \int_{Z^{\otimes(N-1)}} h_{\bar{z}^j}(z^j) \hat{P}_j(dz^j) - \int_{Z} h_{\bar{z}^j}(z^j) \hat{P}_j(dz^j) \right|
\]
\[
\leq \frac{1}{N} \zeta_p(\hat{P}_j, \hat{P}_j), \quad (4.41)
\]
where the inequality is due to \( Nh_{\bar{z}^j}(z^j) \in \mathcal{F}_p(Z) \) and the definition of \( \zeta_p(P, Q) \). Finally, by the triangle inequality of the pseudo-metric, we have
\[
\mathcal{d}_\psi(P^{\otimes N}, Q^{\otimes N}) \leq \mathcal{d}_\psi(P^{\otimes N}, P^{\otimes(N-1)} \times Q) + \mathcal{d}_\psi(P^{\otimes(N-1)} \times Q, P^{\otimes(N-2)} \times Q^{\otimes 2}) + \cdots + \mathcal{d}_\psi(P \times Q^{\otimes(N-1)}, Q^{\otimes N}) \leq \frac{1}{N} \zeta_p(P, Q) \times N = \zeta_p(P, Q).
\]
The proof is complete. \( \Box \)

With Lemma [4.1] we are ready to state our main result.

**Theorem 4.2 (Quantitative statistical robustness)** Let \( \phi(z) \) be defined as in Assumption 3.3 and \( \mathcal{M}_Z^\phi = \{P' \in \mathcal{P}(Z) : \int_{Z} \phi(z) P'(dz) < \infty \} \). Under Assumptions 3.1 (b), 3.2 (a) and 4.1,
\[
\mathcal{d}_{K,1} \left( P^{\otimes N} \circ \hat{\nu}_{N}(\cdot, \lambda_N)^{-1}, Q^{\otimes N} \circ \hat{\nu}_{N}(\cdot, \lambda_N)^{-1} \right) \leq \zeta_p(P, Q)
\]
for any \( N \in \mathbb{N} \) and any \( P, Q \in \mathcal{M}_Z^\phi \), where \( p \) is defined as in Assumption 4.1. In the case when \( p = 1 \),
\[
\mathcal{d}_{K,1} \left( P^{\otimes N} \circ \hat{\nu}_{N}(\cdot, \lambda_N)^{-1}, Q^{\otimes N} \circ \hat{\nu}_{N}(\cdot, \lambda_N)^{-1} \right) \leq \mathcal{d}_{K,Z}(P, Q).
\]

18
Proof. By definition

\[ \text{d}_K,1 \left( P^{\otimes N} \circ \hat{\theta}_N(\cdot, \lambda_N)^{-1}, Q^{\otimes N} \circ \hat{\theta}_N(\cdot, \lambda_N)^{-1} \right) \]

\[ = \sup_{g \in \mathcal{G}} \left| \int_{\mathbb{R}} g(t) P^{\otimes N} \circ \hat{\theta}_N(\cdot, \lambda_N)^{-1}(dt) - \int_{\mathbb{R}} g(t) Q^{\otimes N} \circ \hat{\theta}_N(\cdot, \lambda_N)^{-1}(dt) \right| \]

\[ = \sup_{g \in \mathcal{G}} \left| \int_{Z^{\otimes N}} g(\hat{\theta}_N(z^N, \lambda_N)) P^{\otimes N}(dz^N) - \int_{Z^{\otimes N}} g(\hat{\theta}_N(z^N, \lambda_N)) Q^{\otimes N}(dz^N) \right|, \]

where we write \( z^N \) for \( (z^1, \ldots, z^N) \) and \( \hat{\theta}(z^N, \lambda_N) \) for \( \hat{\theta}_N \) to indicate its dependence on \( z^1, \ldots, z^N \).

To see the well-definiteness of the pseudo-metric, we note that for each \( g \in \mathcal{G} \),

\[ |g(\hat{\theta}_N(z^N, \lambda_N))| \leq |g(\hat{\theta}_N(z_0^N, \lambda_N))| + |\hat{\theta}_N(z^N, \lambda_N) - \hat{\theta}_N(z_0^N, \lambda_N)|, \]

where \( z_0^N \in Z^{\otimes N} \) is fixed. By the definition of \( \hat{\theta}(z^N, \lambda_N) \), we have

\[ |\hat{\theta}_N(z^N, \lambda_N)| = \left| \min_{f \in \mathcal{F}} \frac{1}{N} \sum_{j=1}^{N} \left( c(z^j, f(x^j)) + \lambda_N \|f\|_k^2 \right) \right| \leq \frac{1}{N} \sum_{j=1}^{N} \phi(z^j) + \lambda_N \beta^2. \]

Thus

\[ \int_{Z^{\otimes N}} |\hat{\theta}_N(z^N, \lambda_N)|^{\otimes N}(dz^N) \leq \int_{Z^{\otimes N}} \frac{1}{N} \sum_{j=1}^{N} \phi(z^j) P^{\otimes N}(dz^N) + \lambda_N \beta^2 \]

\[ = \int_{Z} \phi(z) P(dz) + \lambda_N \beta^2 < \infty, \forall P \in \mathcal{M}_Z^\phi, \]

where the equality holds due to the fact that \( z^1, \ldots, z^N \) are i.i.d.. The same inequality can be established for \( \int_{Z^{\otimes N}} |\hat{\theta}_N(z_0^N, \lambda_N)|^{\otimes N}(dz^N) \). Combining (4.45) and (4.46), we deduce that

\[ \int_{Z^{\otimes N}} g(\hat{\theta}_N(z^N, \lambda_N))^{\otimes N}(dz^N) < \infty, \forall P \in \mathcal{M}_Z^\phi. \]

The same argument can be made on \( \int_{Z^{\otimes N}} g(\hat{\theta}_N(z^N, \lambda_N)) Q^{\otimes N}(dz^N) \) for \( Q \in \mathcal{M}_Z^\phi \).

Next, we show (4.42). We do so by applying Lemma 4.11 to the right hand side of (4.44). To this end, we need to verify the condition of the lemma. Define \( \psi : Z^{\otimes N} \to \mathbb{R} \) by \( \psi(z^N) := g(\hat{\nu}(z^N, \lambda_N)) \). Since \( g \) is Lipschitz continuous with modulus bounded by 1, we have

\[ \left| \psi(\hat{\nu}^N) - \psi(\hat{\nu}^N) \right| = \left| g(\hat{\theta}_N(z^N, \lambda_N)) - g(\hat{\theta}_N(z^N, \lambda_N)) \right| \]

\[ \leq \left| \hat{\theta}_N(z^N, \lambda_N) - \hat{\theta}_N(z^N, \lambda_N) \right| \]

\[ = \left| \min_{f \in \mathcal{F}} \frac{1}{N} \sum_{j=1}^{N} \left( c(z^j, f(x^j)) + \lambda_N \|f\|_k^2 \right) - \min_{f \in \mathcal{F}} \frac{1}{N} \sum_{j=1}^{N} \left( c(z^j, f(x^j)) + \lambda_N \|f\|_k^2 \right) \right| \]

\[ \leq \frac{1}{N} \sum_{j=1}^{N} \sup_{f \in \mathcal{F}} \left| c(z^j, f(x^j)) - c(\hat{z}^j, f(\hat{x}^j)) \right| \]

\[ \leq \frac{1}{N} \sum_{j=1}^{N} c_p(\hat{z}^j, \hat{x}^j) \|\hat{z}^j - \hat{x}^j\|, \]

19
which means that $\psi$ is in the set of functions $\Psi$ in Lemma 4.1. The rest follows from application of the lemma to (4.44). ■

The strength of Theorem 4.2 lies in the fact that it gives rise to an explicit quantitative relationship between $d_{K,1}\left(P^{\otimes N} \circ \hat{\nu}_N(\cdot, \lambda_N)^{-1}, Q^{\otimes N} \circ \hat{\nu}_N(\cdot, \lambda_N)^{-1}\right)$ and $\zeta(P, Q)$. This is benefited partially from use of the dual representation of the Kantorovich metric in the quantification of the former and partially from use of the Fortet-Mourier metric for quantification of the latter. As noted immediately after Definition 2.3, $\zeta(P, Q)$ may be estimated via sample data, which means the error bound established in (4.42) is practically obtainable and this is a significant step forward from the qualitative robustness result. Note also that both $d_{\phi}$ and $\zeta_{\phi}$ capture (restrict) the tail behaviour of $P$ but there is no explicit relationship between the two metrics as far as we are concerned: the former provides weak convergence of each measurable function dominated by $\phi$ whereas the latter requires uniform convergence of a class of locally Lipschitz continuous functions with specified rate of growth. Finally, we note that the error bound does not depend on the regularization parameters because from the proof we can see that the regularization terms are cancelled. It does not mean that the parameter has no effect on the statistical performance of the empirical risk, rather it means the error bound does not capture such effect.

The next example illustrates how the theorem works in some concrete regression models.

**Example 4.1** Consider the least squares regression model, where $c(z, f(x)) = \frac{1}{2}|y - f(x)|^2$. We have

$$|c(z, f(x)) - c(z', f(x'))| = \frac{1}{2}||y - f(x)|^2 - |y' - f(x')|^2|$$

$$\leq \frac{1}{2} \left(|y| + |f(x)| + |y'| + |f(x')|(\|y - y\| + |f(x) - f(x')|)\right).$$

Under Assumption 3.1 (b) and the calmness condition in Remark 3.1,

$$|f(x)| \leq \|f\|_k k(x, \cdot)\|_k \leq \beta \|k(x, \cdot)\|_k = \beta \sqrt{k(x, x)}, \forall f \in \mathcal{F}$$

and

$$|f(x) - f(x')| = |\langle f, k(\cdot, x) \rangle - \langle f, k(\cdot, x') \rangle| \leq \beta \|k(\cdot, x) - k(\cdot, x')\|_k \leq \beta g(||x - x'||).$$

Let $\eta(z) := |y| + \beta \sqrt{k(x, x)}$. Then,

$$|c(z, f(x)) - c(z', f(x'))| \leq \max \{\eta(z), \eta(z')\} (|y - y'| + \beta g(||x - x'||)).$$

- In the case of linear kernel, $\eta(z) = |y| + \beta \|x\| \leq \max\{1, \beta\}\|z\|$, $g(t) = t$, and

$$|c(z, f(x)) - c(z', f(x'))| \leq \max\{1, \beta\} \max\{1, \|z\|, \|z'\|\} \|z - z'\|.$$

By Theorem 4.2, $d_{K,1}\left(P^{\otimes N} \circ \hat{\nu}_N^{-1}, Q^{\otimes N} \circ \hat{\nu}_N^{-1}\right) \leq \max\{1, \beta\}^2 \zeta(P, Q)$ for all $N \in \mathbb{N}$ and any $P, Q \in \mathcal{M}^2_Z$, where $\phi(z) = ||y||^2 + \beta^2 ||x||^2$.

- In the case of Gaussian kernel, $\eta(z) = |y| \leq ||z||$, $g(t) = \max\{\sqrt{t}, 1\}t$, and

$$|c(z, f(x)) - c(z', f(x'))| \leq \max\{\sqrt{2}, 1\} \max\{1, \|z\|, \|z'\|\} \|z - z'\|.$$

20
By Theorem 4.2, \( \|k(\cdot, x) - k(\cdot, x')\|_k \leq \max\{\frac{1}{2R}\sqrt{2(\gamma R^2 + 1)^d}, 1\}|x - x'|, \) if \( d \) is even, and \( \max\{\frac{1}{2R}\sqrt{2(\gamma R^2 + 1)^d} - 2(1 - \gamma R^2)^d, 1\}|x - x'|, \) if \( d \) is odd.

\[
\begin{aligned}
\leq & \left\{ \begin{array}{ll}
\max\{\frac{1}{2R}\sqrt{2(\gamma R^2 + 1)^d}, 1\}|x - x'|, & \text{if } d \text{ is even}, \\
\max\{\frac{1}{2R}\sqrt{2(\gamma R^2 + 1)^d} - 2(1 - \gamma R^2)^d, 1\}|x - x'|, & \text{if } d \text{ is odd},
\end{array} \right.
\leq \left\{ \begin{array}{ll}
\max\{\frac{1}{2R}\sqrt{2(\gamma R^2 + 1)^d}, 1\}|x - x'|, & \text{if } d \text{ is even}, \\
\max\{\frac{1}{2R}\sqrt{2(\gamma R^2 + 1)^d} - 2(1 - \gamma R^2)^d, 1\}|x - x'|, & \text{if } d \text{ is odd}.
\end{array} \right.
\end{aligned}
\]

The last inequality is due to the fact that \( a - b \leq \max\{2a, -2b\} \) for any two numbers \( a, b \) where \( a > 0 \) and \( b \) could be either negative or positive. Let

\[
A_1 := (1 + \beta(\gamma + 1)^{d/2}) \max\left\{ \beta\left(\frac{1 + \gamma}{2}\right)^{d/2}, \beta, 1 \right\},
\]

\[
A_2 := (1 + \beta(\gamma + 1)^{d/2}) \max\left\{ 2\beta\left(\frac{1 + \gamma}{2}\right)^{d/2}, 4\beta, 4\beta \gamma, 1 \right\}.
\]

Then

\[
|c(z, f(x)) - c(z', f(x'))| \leq \left\{ \begin{array}{ll}
A_1 \max\{1, \|z\|, \|z'\|\}^{2d-1} \|z - z'\|, & \text{if } d \text{ is even}, \\
A_2 \max\{1, \|z\|, \|z'\|\}^{3d-1} \|z - z'\|, & \text{if } d \text{ is odd}.
\end{array} \right.
\]

By Theorem 4.2,

\[
\|P^{\otimes N} \circ \hat{\vartheta}_N^{-1}, Q^{\otimes N} \circ \hat{\vartheta}_N^{-1}\|_{K, 1} \leq \left\{ \begin{array}{ll}
A_1 \zeta_{2d}(P, Q), & \text{if } d \text{ is even}, \\
A_2 \zeta_{3d+1}(P, Q), & \text{if } d \text{ is odd},
\end{array} \right.
\]

for all \( N \in \mathbb{N} \) and any \( P, Q \in \mathcal{M}_2^0 \), where \( \phi(z) = \|y\|^2 + \beta^2(\gamma \|x\|^2 + 1)^d \).

We can derive similar results for the regression models with \( \epsilon \)-insensitive loss function \( c(z, f(x)) = \max\{0, |y - f(x)| - \epsilon\} \), hinge loss \( c(z, f(x)) = \max\{0, 1 - (y - f(x))\} \), and log-loss function \( c(z, f(x)) = \log(1 + e^{-(y-f(x))}) \) respectively, we omit the details.

**Remark 4.1** It might be interesting to study the discrepancy between \( f^{\lambda N}_N(P_N) \) and \( f^{\lambda N}_N(Q_N) \). To this end, we assume that \( c(z, f(x)) \) is strong convex in \( f \) for almost all \( z \). In such a case, \( R(f) = \mathbb{E}_P[c(z, f(x))] \) is also strongly convex and so is \( R(f) + \lambda \|f\|_k \), which implies that problem (2.3) and the regularized problem (2.4) have a unique solution. Moreover, the strong convexity implies that problem (2.3) satisfies second order growth condition at \( f^{\lambda N}_N(P_N) \), that is, there exists a positive constant \( \alpha \) such that

\[
R^{\lambda N}_{P_N}(f) - \vartheta(P_N, \lambda_N) \geq \alpha \|f - f^{\lambda N}_{N}(P_N)\|^2_k, \forall f \in \mathcal{F}.
\]
By virtue of \cite[Lemma 3.8]{21}, we can use the inequality to obtain
\[
\|f_N^{\lambda N}(P_N) - f_N^{\lambda N}(Q_N)\|_k \leq \sqrt{\frac{3}{\alpha} \sup_{f \in \mathcal{F}} |\mathbb{E}_{P_N}[c(z, f(x))] - \mathbb{E}_{Q_N}[c(z, f(x))]|} \\
\leq \sqrt{\frac{3}{\alpha} \mathbb{E}_{P_N \times Q_N}[c_p(\hat{z}, \tilde{z})]\|\hat{z} - \tilde{z}\|}.
\]
Since \(\mathbb{E}_{P_N \times Q_N}[c_p(\hat{z}, \tilde{z})]\|\hat{z} - \tilde{z}\| - \mathbb{E}_{P \times Q}[c_p(\hat{z}, \tilde{z})]\|\hat{z} - \tilde{z}\| \to 0\) as \(Q \to P\) and \(N\) goes to infinity, then
\[
\|f_N^{\lambda N}(P_N) - f_N^{\lambda N}(Q_N)\| \to 0.
\]
However, we are unable to establish the kind of estimation in \cite[(4.42)]{4} for the optimal solutions because of the non-linearity of the bound
\[
\sqrt{\frac{3}{\alpha} \sup_{f \in \mathcal{F}} |\mathbb{E}_{P_N}[c(z, f(x))] - \mathbb{E}_{Q_N}[c(z, f(x))]|}
\]
for \(\|f_N^{\lambda N}(P_N) - f_N^{\lambda N}(Q_N)\|_k\) in terms of the difference of the function values.

5 Uniform consistency

In this section, we move on to investigate convergence of \(\vartheta(P_N, \lambda_N)\) to \(\vartheta(P)\) as \(N \to \infty\) and \(\lambda_N \to 0\). We proceed the investigation in two steps: first pointwise convergence, i.e., for each fixed \(P \in \mathcal{P}(Z)\) and then uniform convergence for all \(P\) over a subset \(\mathcal{M}\) of \(\mathcal{P}(Z)\). To this end, we introduce the following assumption on the cost function.

**Assumption 5.1** There exist a measurable function \(r(\cdot) : Z \to \mathbb{R}_+\) and a constant \(\nu \in (0, 1]\) such that
\[
|c(z, f(x)) - c(z, g(x))| \leq r(z)\|f - g\|_{\infty}, \forall f, g \in \mathcal{F}, z \in Z. \quad (5.47)
\]

The assumption requires \(c(z, \cdot)\) to be Hölder continuous over \(\mathcal{F}\) uniformly for \(z \in Z\). It should be distinguished from Assumption \cite[4.1]{1} which requires \(c(z, f(x))\) to be locally Lipschitz continuous in \(z\) for all \(f \in \mathcal{F}\). The assumption is satisfied by all of the loss functions in regression models that we listed at the beginning of Section 2.

**Theorem 5.1 (Consistency of \(\vartheta(P_N, \lambda_N)\))** Let Assumptions \cite[4.1, 4.2 and 5.1]{1} hold. Then for any \(\delta > 0\), there exist positive constants \(\epsilon < \delta/6\), \(\alpha(\epsilon, \delta)\) and \(\gamma(\epsilon, \delta)\), independent of \(N\) and a positive number \(N_0\) such that
\[
P^\otimes N \left( \sup_{f \in \mathcal{F}} |\mathbb{E}_{P_N}[c(z, f(x))] + \lambda_N\|f\|_k^2 - \mathbb{E}_P[c(z, f(x))]| \geq \delta \right) \leq \alpha(\epsilon, \delta)e^{-N\gamma(\epsilon, \delta)} \quad (5.48)
\]
when \(N \geq N_0\) and \(\lambda_N \leq \epsilon/\beta^2\) and hence
\[
P^\otimes N (|\vartheta(P_N, \lambda_N) - \vartheta(P)| \geq \delta) \leq \alpha(\epsilon, \delta)e^{-N\gamma(\epsilon, \delta)} \quad (5.49)
\]
and
\[ P^\otimes N \left( \| \mathbb{E}_P[c(z, f^N_{\lambda N}(x))] - \vartheta(P) \| \geq \delta \right) \leq 2\alpha(\epsilon, \delta) e^{-N\gamma(\epsilon, \delta)}, \tag{5.50} \]

where \( f^N_{\lambda N} \in F^*_{N, \lambda N}. \)

In the literature of machine learning, consistency analysis refers to (5.50) whereas in stochastic programming, it refers to (5.49). The consistency analysis is mostly focused on the case when \( Z \) is a compact set, we refer readers to Norkin and Keyzer \[22\] which provides an excellent overview about this. Caponnetto and Vito \[6\] is one of a few exceptions which studies convergence of the empirical risk of a regularized least-square problem in a reproducing kernel Hilbert space with unbounded feasible set. Under some moderate conditions, they derive optimal choice of the regularization parameter and optimal rate of convergence of the empirical risk over a class of priors defined by a uniformly bounded kernel. Our focus here is slightly different: while we are also aiming to derive exponential rate of convergence, we concentrate more on how to overcome the complexities and challenges arising from a generic form of the cost function and an unbounded kernel. For instance, the exponential rate of convergence in (5.48) holds uniformly for all \( f \in F \). This kind of result may not hold in general, see a counter example in \[29\]. Here we manage to establish the uniform convergence by showing equi-continuity of the class of functions in \( F \) under Assumption \[3.1\] and their uniform boundedness over a compact subset of \( Z \).

**Proof of Theorem 5.1** Observe that inequality (5.48) implies
\[ P^\otimes N \left( \| \mathbb{E}_P[c(z, f^N_{\lambda N}(x))] + \lambda_N \| f^N_{\lambda N} \|_k^2 \right) \leq \mathbb{E}_P[c(z, f^N_{\lambda N}(x))] \geq \delta \) \leq \alpha(\epsilon, \delta) e^{-N\gamma(\epsilon, \delta)}, \tag{5.51} \]

and a combination of (5.51) and (5.49) yields (5.50). Thus it suffices to prove (5.48) and (5.49). Since \( P \in \mathcal{M}_Z^\phi \), then for any \( \epsilon > 0 \), there exist a constant \( r > 0 \) such that
\[ \int_Z \phi(z) 1_{(r, \infty)}(\phi(z)) P(dz) \leq \epsilon. \]

Moreover, by the large deviation theory, there exist positive numbers \( C_0 \) and \( \gamma_0 \) such that
\[ P^\otimes N \left( \int_Z \phi(z) 1_{(r, \infty)}(\phi(z)) P_N(dz) \geq 2\epsilon \right) \leq C_0 e^{-\gamma_0 N}. \]

Under the coercive condition on \( \phi \) in Assumption \[3.2\] (a), there exists a compact set \( Z_\epsilon = (X_\epsilon, Y_\epsilon) \subset Z \) such that \( \{ z \in Z : \phi(z) \leq r \} \subset Z_\epsilon \). Thus
\[ \sup_{f \in F} \int_{Z \setminus Z_\epsilon} |c(z, f(x))| P(dz) \leq \int_{Z \setminus Z_\epsilon} \phi(z) P(dz) \leq \int_{\{ z \in Z : \phi(z) > r \}} \phi(z) P(dz) \leq \epsilon \tag{5.52} \]

and
\[ P^\otimes N \left( \sup_{f \in F} \int_{Z \setminus Z_\epsilon} |c(z, f(x))| P_N(dz) \geq 2\epsilon \right) \leq P^\otimes N \left( \int_{Z \setminus Z_\epsilon} \phi(z) P_N(dz) \geq 2\epsilon \right) \tag{5.53} \]
\[ \leq P^\otimes N \left( \int_{\{ z \in Z : \phi(z) > r \}} \phi(z) P_N(dz) \geq 2\epsilon \right) \leq C_0 e^{-\gamma_0 N}. \]
By Assumption 3.1, there exists $\eta > 0$ such that for any $x, x' \in X_\epsilon$ satisfying $\|x - x'\| < \eta$, we have
\[
|f(x') - f(x)| = |\langle f, k(\cdot, x') \rangle - \langle f, k(\cdot, x) \rangle| \leq \|f\|_k \|k(\cdot, x') - k(\cdot, x)\|_k
\leq \beta \|k(\cdot, x') - k(\cdot, x)\|_k \leq \beta \epsilon,
\]
which implies $\mathcal{F}$ is equi-continuous when it is restricted to $X_\epsilon$.

Let $\Delta_\epsilon := \sup_{x \in X_\epsilon} \|k(\cdot, x)\|_k$. Then for any $f \in \mathcal{F}$,
\[
\sup_{x \in X_\epsilon} |f(x)| = \sup_{x \in X_\epsilon} |\langle f, k(\cdot, x) \rangle| \leq \|f\|_k \sup_{x \in X_\epsilon} \|k(\cdot, x)\|_k \leq \beta \Delta_\epsilon,
\]
which implies that $\mathcal{F}$ is uniformly bounded when it is restricted to $X_\epsilon$. Let $\bar{r} := \max\{|r(z)| : z \in Z_\epsilon\}$ and $\bar{\epsilon} := (\epsilon/\bar{r})^{1/\nu}$. By Ascoli-Arzela Theorem [5], there exists an $\bar{\epsilon}$-net of $\mathcal{F}_K := \{f_1, \ldots, f_K\} \subset \mathcal{F}$ such that $\mathcal{F} = \bigcup_{k=1}^K \mathcal{F}_k$, where $\mathcal{F}_k := \{f \in \mathcal{F} : \sup_{x \in X_\epsilon} |f(x) - f_k(x)| \leq \bar{\epsilon}\}$ for $k = 1, \ldots, K$. Therefore,
\[
|\vartheta(P_N, \lambda_N) - \vartheta(P)| = \sup_{f \in \mathcal{F}} \left\{ \mathbb{E}_P[c(z, f(x))] + \lambda_N \|f\|_k^2 \right\} - \sup_{f \in \mathcal{F}} \mathbb{E}_P[c(z, f(x))]
\leq \sup_{f \in \mathcal{F}} \mathbb{E}_P[c(z, f(x))1_{Z_\epsilon}(z)] - \sup_{f \in \mathcal{F}} \mathbb{E}_P[c(z, f(x))]1_{Z_\epsilon}(z)] + \lambda_N \beta^2
+ \sup_{f \in \mathcal{F}} \int_{Z \setminus Z_\epsilon} |c(z, f(x))|P_N(dz) + \sup_{f \in \mathcal{F}} \int_{Z \setminus Z_\epsilon} |c(z, f(x))|P(dz)
\leq \sup_{k \in K} \sup_{f \in \mathcal{F}_k^c} \mathbb{E}_P[c(z, f_k(x))1_{Z_\epsilon}(z)] - \mathbb{E}_p[c(z, f_k(x))1_{Z_\epsilon}(z)]
+ \mathbb{E}_p[c(z, f_k(x))1_{Z_\epsilon}(z)] + 2\epsilon
\leq \sup_{k \in \{1, \ldots, K\}} \sup_{f \in \mathcal{F}_k^c} |\mathbb{E}_P[c(z, f_k(x))1_{Z_\epsilon}(z)] - \mathbb{E}_P[c(z, f_k(x))1_{Z_\epsilon}(z)]| + 4\epsilon
+ \sup_{f \in \mathcal{F}} \int_{Z \setminus Z_\epsilon} |c(z, f(x))|P_N(dz),
\]
where the first inequality holds due to $\|f\|_k \leq \beta$, and the last inequality holds because under Assumption 5.1 we have
\[
\mathbb{E}_P[c(z, f(x))1_{Z_\epsilon}(z)] - \mathbb{E}_P[c(z, f_k(x))1_{Z_\epsilon}(z)] \leq \mathbb{E}_P[r(z)\|f - f_k\|_k^\nu 1_{Z_\epsilon}(z)] \leq \bar{r}\epsilon^\nu = \epsilon
\]
and
\[
\mathbb{E}_P[c(z, f(x))1_{Z_\epsilon}(z)] - \mathbb{E}_P[c(z, f_k(x))1_{Z_\epsilon}(z)] \leq \mathbb{E}_P[r(z)\|f - f_k\|_k^\nu 1_{Z_\epsilon}(z)] \leq \bar{r}\epsilon^\nu = \epsilon.
\]
It follows from the classical Cramér’s large deviation theorem [12] that for each \( k \) there exist positive constants \( C(\epsilon, \delta, f_k) \) and \( \gamma(\epsilon, \delta, f_k) \) such that

\[
P^N(\|E_P[c(z, f_k(x))I_{Z_k}(z)]| - E_P[c(z, f_k(x))I_{Z_k}(z)]\| \geq \delta - 6\epsilon) \leq C(\epsilon, \delta, f_k)e^{-N\gamma(\epsilon, \delta, f_k)}.
\]

Hence, we have

\[
P^N\left(\sup_{f \in H}|E_P[c(z, f(x))| - E_P[c(z, f(x))I_{Z_k}(z)]| \geq \delta - 6\epsilon\right)
\]

\[
\leq P^N\left(\sup_{k \in \{1, \ldots, K\}}\|E_P[c(z, f_k(x))I_{Z_k}(z)]| - E_P[c(z, f_k(x))I_{Z_k}(z)]\| \geq \delta - 6\epsilon\right)
\]

\[
+ P^N\left(\sup_{f \in F} \int_{Z \setminus Z_k} |c(z, f(x))| P_N(dz) \geq 2\epsilon\right)
\]

\[
\leq \sum_{k \in \{1, \ldots, K\}} P^N(\|E_P[c(z, f_k(x))I_{Z_k}(z)]| - E_P[c(z, f_k(x))I_{Z_k}(z)]\| \geq \delta - 6\epsilon) + C_0 e^{-\gamma_0 N}
\]

\[
\leq \sum_{k \in \{1, \ldots, K\}} C(\epsilon, \delta, f_k)e^{-N\gamma(\epsilon, \delta, f_k)} + C_0 e^{-\gamma_0 N},
\]

which implies (5.48).

Next we study uniform convergence of the regularized empirical risk with respect to a class of empirical probability distributions as the sample size increases. In practice, we may be able to obtain empirical data but often do not know the true probability distribution generating the data. Our next result states that the empirical risk converges to its true counterpart uniformly for all empirical data to be used in the machine learning model.

**Theorem 5.2 (Uniform consistency of \( \vartheta(P_N, \lambda_N) \))** Let Assumptions 3.1, 3.2 and 5.1 hold. Let

\[
\mathcal{M}_\kappa^{\phi_p} := \left\{ P \in \mathcal{P}(Z) : \int_Z \phi(z)^p P(dz) < \kappa \right\}
\]

for some fixed \( p > 1 \) and \( \mathcal{M} \) be a compact subset of \( \mathcal{M}_\kappa^{\phi_p} \). Then for every \( \epsilon > 0 \) and \( \delta > 0 \), there exists \( N_0 \) such that

\[
\sup_{P \in \mathcal{M}} P^N(\|\vartheta(P_N, \lambda_N) - \vartheta(P)\| \geq \delta) \leq \epsilon,
\]

(5.54)

when \( \lambda_N \leq \delta/4\beta^2 \) and \( N \geq N_0 \).

The uniform convergence (5.54) is closely related to learnability in statistical learning theory which is defined as the uniform convergence of \( R(f_N(P)) \) to \( \vartheta(P) \) for all empirical probability distributions drawn from \( \mathcal{P}(Z) \), where \( R(\cdot) \) is defined as in (2.1), see [29, Definition 1]. Here we are looking into the convergence for all \( P_N \) whose true counterpart is drawn \( \mathcal{M} \). This applies to the case that there is some incomplete information about the nature of \( P \).
Proof of Theorem 5.2. We first show that (5.54) holds for each $P \in \mathcal{M} \subset \mathcal{M}_s^{\phi_p}$. For fixed $\bar{P}$, by the continuity of $\vartheta(\cdot)$ at $\bar{P}$ in Theorem 5.1 for any $\delta > 0$, there exists a positive constant $\eta > 0$ such that

$$|\vartheta(Q) - \vartheta(\bar{P})| < \delta/2,$$

for each $Q$ satisfying $d_{\phi}(Q, \bar{P}) < \eta$. It follows by [20, Corollary 3.5] that $(\mathcal{M}_s^{\phi_p}, dl_{\phi})$ has the UGC property for all $p > 1$ and $\kappa > 0$, that is, for any $\epsilon, \eta > 0$, there exists $N_0 \in \mathbb{N}$ such that for all $N \geq N_0$

$$P^{\otimes N} [dl_{\phi}(P_N, P) \geq \eta] \leq \epsilon, \forall P \in \mathcal{M}_s^{\phi_p}.$$

Thus, for any $\epsilon > 0$ and $\delta > 0$, there exists $N_0$ such that for all $N \geq N_0$

$$\bar{P}^{\otimes N} \left[|\vartheta(\bar{P}_N) - \vartheta(\bar{P})| \geq \delta/2\right] \leq P^{\otimes N} \left[dl_{\phi}(\bar{P}_N, \bar{P}) \geq \eta\right] \leq \epsilon.$$

Since

$$|\vartheta(\bar{P}_N, \lambda_N) - \vartheta(\bar{P})| = |\inf_{f \in \mathcal{F}} \{\mathbb{E}_P[c(z, f(x))] + \lambda_N \|f\|_k^2\} - \inf_{f \in \mathcal{F}} \mathbb{E}_{P_N}[c(z, f(x))]|$$

$$\leq |\inf_{f \in \mathcal{F}} \mathbb{E}_P[c(z, f(x))] - \inf_{f \in \mathcal{F}} \mathbb{E}_{P_N}[c(z, f(x))]| + \sup_{f \in \mathcal{F}} \lambda_N \|f\|_k^2$$

$$= |\vartheta(\bar{P}_N) - \vartheta(\bar{P})| + \lambda_N \beta^2,$$

then

$$\bar{P}^{\otimes N} \left[|\vartheta(\bar{P}_N, \lambda_N) - \vartheta(\bar{P})| \geq \delta\right] \leq P^{\otimes N} \left[|\vartheta(\bar{P}_N) - \vartheta(\bar{P})| \geq \delta/2\right] \leq \epsilon$$

when $\lambda_N \leq \delta/4\beta^2$. Therefore, (5.54) holds when $P$ is fixed at $\bar{P}$.

Now we show (5.54) holds for all $P \in \mathcal{M}$. Assume for the sake of a contradiction that there exist some positive numbers $\epsilon_0$ and $\delta_0$ such that for any $s \in \mathbb{N}$, there exist $s' > s$, $P_{s'} \in \mathcal{M}$ and some $N_{s'} \geq s$ such that

$$P_{s'}^{\otimes N_{s'}} \left[|\vartheta(P_{N_{s'}}, \lambda_{N_{s'}}) - \vartheta(P_{s'})| \geq \delta_0\right] > \epsilon_0. \quad (5.55)$$

Let $s$ increase. Then we obtain a sequence of $\{P_{s'}\}$ which satisfies (5.55). Since $\mathcal{M}$ is compact under the $\phi$-weak topology, then $\{P_{s'}\}$ has a converging subsequence. Assume without loss of generality that $P_{s'} \xrightarrow{\phi} P_s \in \mathcal{M}$. Since $\vartheta(\cdot)$ is continuous at $P_s$, then there exists $\eta > 0$ such that $|\vartheta(Q) - \vartheta(P_s)| < \delta_0/4$ for $P$ satisfying $dl_{\phi}(Q, P_s) < \eta$ and then

$$|\vartheta(Q, \lambda') - \vartheta(P_s)| \leq |\vartheta(Q) - \vartheta(P_s)| + \lambda'\beta^2 < \delta_0/2$$

for $\lambda' \leq \delta_0/4\beta^2$. By $P_{s'} \xrightarrow{\phi} P_s$, there exists $s'_0$ such that $dl_{\phi}(P_{s'}, P_s) < \eta/2$ for $s' \geq s'_0$, and then $|\vartheta(P_{s'}, \lambda_{s'}) - \vartheta(P_s)| < \delta_0/2$ for $\lambda_{s'} \leq \delta_0/4\beta^2$. On the other hand, by the UGC property

$$P_{s}^{\otimes N_s} (dl_{\phi}(P_{N_{s'}}, P_s) \geq \eta) \leq P_{s}^{\otimes N_s} (dl_{\phi}(P_{N_{s'}}, P_{s'}) + dl_{\phi}(P_{s'}, P_s) \geq \eta)$$

$$= P_{s}^{\otimes N_s} (dl_{\phi}(P_{N_{s'}}, P_{s'}) \geq \eta - dl_{\phi}(P_{s'}, P_s) \geq \eta/2) \leq P_{s}^{\otimes N_s} (dl_{\phi}(P_{N_{s'}}, P_{s'}) \geq \eta/2) \leq \epsilon_0$$

26
for sufficiently large $N_{s'}$. Therefore,

$$P_{s'}^{\otimes N_{s'}} \left[ |\vartheta(P_{N_{s'}}, \lambda_{N_{s'}}) - \vartheta(P_{s})| \geq \delta_0/2 \right] \leq \epsilon_0,$$

and

$$P_{s'}^{\otimes N_{s'}} \left[ |\vartheta(P_{N_{s'}}, \lambda_{N_{s'}}) - \vartheta(P_{s'})| \geq \delta_0 \right]$$

$$\leq P_{s'}^{\otimes N_{s'}} \left[ |\vartheta(P_{N_{s'}}, \lambda_{N_{s'}}) - \vartheta(P_{s'})| + |\vartheta(P_{s'}, \lambda_{s'}) - \vartheta(P_{s})| \geq \delta_0 \right]$$

$$\leq P_{s'}^{\otimes N_{s'}} \left[ |\vartheta(P_{N_{s'}}, \lambda_{N_{s'}}) - \vartheta(P_{s})| \geq \delta_0/2 \right] \leq \epsilon_0,$$

which leads to a contradiction with (5.55) as desired.

\section{Concluding remarks}

In this paper, we present some theoretical analysis about statistical robustness of empirical risk in machine learning. Our focus is on empirical risk but it might be interesting to extend the discussion to kernel learning estimators. Moreover, our analysis in statistical robustness and uniform consistency does not capture the effect of the optimal choice of the regularization parameter in learning process, but we envisage the effect exists and will be helpful to quantify it. Finally, it might be interesting to carry out some numerical experiments to examine the statistical robustness of the empirical risk. We leave all these for future research as they require much more intensive work.

\section*{References}

[1] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler, Scale-sensitive dimensions, uniform convergence, and learnability, \textit{Journal of the ACM}, 44(4):615-631, 1997.

[2] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth, Learnability and the Vapnik-Chervonenkis dimension, \textit{Journal of the Association for Computing Machinery}, 36(4):929-965, 1989.

[3] S. Boucheron, O. Bousquet and G. Lugosi, Theory of classification: a survey of some recent advances, \textit{ESAIM: Probability and Statistics}, 9:323-375, 2005.

[4] P. Breheny and J. Huang, Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors, \textit{Statistics and Computing}, 25:173-187, 2015.

[5] R. F. Brown, \textit{A Topological Introduction to Nonlinear Analysis}, Spring, New York, 2004.

[6] A. Caponnetto and E. De Vito, Optimal rates for the regularized least-squares algorithm, \textit{Foundations of Computational Mathematics}, 331-368, 2007.

[7] M. Claus, \textit{Advancing Stability Analysis of Mean-Risk Stochastic Programs: Bilevel and Two-Stage Models}, PhD Thesis, Universität Dusburg-Essen, 2016.
[8] R. Cont, R. Deguest and G. Scandolo, Robustness and sensitivity analysis of risk measurement procedures, *Quantitative Finance*, 10:593-606, 2010.

[9] F. Cucker and S. Smale, Best choices for regularization parameters in learning theory: on the bias-variance problem, *Foundations of Computational Mathematics*, 2: 413-428, 2002.

[10] F. Cucker and D.X. Zhou, *Learning Theory: An Approximation Theory Viewpoint*, Cambridge University Press, 2007.

[11] D. Davis and D. Drusvyatskiy, Graphical Convergence of Subgradients in Nonconvex Optimization and Learning, [arXiv:1810.07590](https://arxiv.org/abs/1810.07590).

[12] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, Springer-Verlag, New York, 1998.

[13] S. Guo and H. Xu, Statistical robustness in utility preference robust optimization models, preprint, School of Mathematical Sciences, University of Southampton, 2018.

[14] F. R. Hampel, A general statistical definition of robustness, *The Annals of Mathematical Statistics*, 42:1887-1896, 1971.

[15] M. Mohri, A. Rostamizadeh and A. Talwalkar, *Foundations of Machine Learning*, The MIT Press, 2012.

[16] P. J. Huber, *Robust Statistics*, 3rd Edition, John Wiley & Sons, New York, 1981.

[17] P. J. Huber and E. M. Ronchetti, *Robust Statistics*, 2nd Edition, John Wiley & Sons, New Jersey, 2009.

[18] G. S. Kimeldorf and G. Wahba, A correspondence between Bayesian estimation on stochastic processes and smoothing by splines, *The Annals of Mathematical Statistics*, 45:495-502, 1970.

[19] V. Krätschmer, A. Schied and H. Zähle, Comparative and statistical robustness for law-invariant risk measures, *Finance and Stochastics*, 18:271-295, 2014.

[20] V. Krätschmer, A. Schied and H. Zähle, Qualitative and infinitesimal robustness of tail-dependent statistical functionals, *Journal of Multivariate Analysis*, 103:35-47, 2012.

[21] Y. Liu and H. Xu, Stability analysis of stochastic programs with second order dominance constraints, Mathematical Programming, 142:435-460, 2013.

[22] V. Norkin and M. Keyzer, On convergence of kernal learning estimators, *SIAM Journal on Optimization*, 20:1205-1223, 2009.

[23] T. Poggio and S. Smale, The mathematics of learning: dealing with data, Notices of the AMS, 2003.

[24] R. Ranga Rao, Relations between weak and uniform convergence of measures with applications, *The Annals of Mathematical Statistics*, 33:659-680, 1962.
[25] W. Römisch, Stability of stochastic programming problems. In: Ruszczyński, A., Shapiro, A. (eds.) Stochastic Programming, Handbooks in Operations Research and Management Science, volume 10, chapter 8. Elsevier, Amsterdam, 2003.

[26] A. Ruszcynski and A. Shapiro, Stochastic Programming Models, in A. Ruszcynski and A.Shapiro, editors, Stochastic Programming, Handbooks in OR & MS, Vol. 10, North-Holland Publishing Company, Amsterdam, 2003.

[27] B. Schölkopf and A. J. Smola, Learning with Kernels. Support Vector Machines, Regularization, Optimization and Beyong, MIT Press, Cambridge, MA, 2002.

[28] S. Shafieezadeh-Abadeh, D. Kuhn and P. Esfahani, Regularization via mass transportation, Journal of Machine Learning Research, 20:1-68, 2019.

[29] S. Shalev-Shwartz, O. Shamir, N. Srebro and K. Sridharan, Learnability, stability and uniform convergence, Journal of Machine Learning Research, 11:2635-2670, 2010.

[30] S. Smale and Y. Yao, Online learning algorithms, Foundations of Computational Mathematics, 6:145-170, 2006.

[31] W. Wang, H. Xu and T. Ma, Quantitative Statistical Robustness for Tail-Dependent Law Invariant Risk Measures, submitted, 2020.