Journal of Machine Learning Research 24 (2023) 1-38 Submitted 9/20; Revised 4/23; Published 4/23

Statistical Robustness of Empirical Risks in Machine Learning

Shaoyan Guo SYGUO@DLUT.EDU.CN
School of Mathematical Sciences
Dalian University of Technology
Dalian, 116024, China

Huifu Xu HFXU@SE.CUHK.EDU.HK
Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong

Liwei Zhang LWZHANG@DLUT.EDU.CN
School of Mathematical Sciences
Dalian University of Technology
Dalian, 116024, China

Editor: Francis Bach

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 are generated by the unknown true probability distribution but this may not be satisfied in some practical circumstances. Consequently the existing convergence results may not provide a guarantee as to whether the empirical risks are reliable or not when the data are potentially corrupted (generated by a distribution perturbed from the true). In this paper, we fill out the gap from robust statistics perspective (Krätschmer, Schied and Zähle (2012); Krätschmer, Schied and Zähle (2014); Guo and Xu (2020)). First, we derive moderate sufficient conditions under which the expected risk changes stably (continuously) against small perturbation of the probability distributions 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 asymptotic qualitative and non-asymptotic 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 regression models.

Keywords: Empirical risks, stability analysis, asymptotic qualitative statistical robustness, non-asymptotic 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 test 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 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, Rusczyński and Shapiro (2003) 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 (Shalev-Shwartz et al., 2010). For instance, Norkin and Keyzer (2009) consider a general nonparametric regression in 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 Cucker and Smale (2002a); Cucker and Zhou (2007); Poggio and Smale (2003); Smale and Yao (2006). Caponnetto and De Vito (2007) 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 of distribution probabilities encoding our knowledge on the relation between input and output data. More recently, Davis and Drusvyatskiy (2018) 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 (Cucker and Zhou, 2007) for the machine learning 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 (Alon et al., 1997; Blumer et al., 1989). For the general learning setting, Shalev-Shwartz et al. (2010) and Shalev-Shwartz and Ben-David (2014) 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 (Shalev-Shwartz et al., 2010) for more detail. A conventional assumption in the above stability is that all of the instants used in the sample are independent and identically distributed (i.i.d.) and are drawn from the true probability distribution, indeed, many classical procedures of machine learning such as LASSO heavily rely on the i.i.d. data with sub-Gaussian behaviour (Tibshirani, 1996). However,

this may not be satisfied in some practical circumstances. As noted in Balasubramanian and Yuan (2016), data from real-world experiments oftentimes tend to be corrupted with outliers and/or exhibiting heavy tails. In finance, heavy-tailed processes are routinely used, and in biology or medical experiments, datasets are regularly subject to some corruption by outliers, see Lecué and Lerasle (2020). 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 are reliable when the empirical data contain some noise. This issue is investigated by Steinwart and Christmann (2008) from robust statistics point of view, that is, how data perturbation may affect the learning models, see Chapter 10 of the book. Robust statistics stems from Tukey (1960, 1962) and Hampel (1968, 1971) and has been popularized by many others particularly the monographs by Huber (1981); Huber and Ronchetti (2009). A well known approach in robust statistics is to examine how the distribution of a statistical estimator is affected by the distribution of the underlying random variables generating the data under the Prokhorov metric, see Cont, Deguest and Scandolo (2010); Krätschmer, Schied and Zähle (2012); Krätschmer, Schied and Zähle (2014); Krätschmer, Schied and Zähle (2017). Another approach is to quantify the sensitivity of a statistical estimator with respect to (w.r.t.) perturbation of a single data point known as an outlier using a so-called influence function. Steinwart and Christmann (2008) discuss in detail how the second approach can be effectively used to analyse impact of data perturbation on learning models, see Chapter 10 of the book.

In a more recent development, Lecué and Lerasle (2020) propose a new robust machine learning approach where the estimators for robust machine learning are based on the median-of-means (MOM) of the estimators of the mean of real valued random variables and demonstrate that these estimators achieve optimal rates of convergence under minimal assumptions on the dataset. Moreover, by studying the breakdown number of outliers that a dataset can contain without deteriorating the estimation properties of a given estimator, they demonstrate that the breakdown number of the estimator is of the order of number of observations times the rate of convergence, and beyond the breakdown point, the rate of convergence achieved by the estimator is the number of outliers divided by the number of observations.

In this paper, we complement the existing research of statistical robustness in machine learning from a different perspective: instead of focusing on the impact of outliers in a dataset, we consider generic data perturbation and its impact on the empirical risks. The rational behind this consideration is that in data driven problems, we might only know the data are polluted but lack of specific information to identify a borderline between good data and bad data. In that case we have to treat all of the data are potentially bad and investigate the extent of data perturbation by which the resulting statistical estimators remain stable. This requires us to take a topological approach to analyse the data structure and we do so by taking the cutting edge results on qualitative statistical robustness by Krätschmer, Schied and Zähle (2012); Krätschmer, Schied and Zähle (2014). The research is carried in three main steps.

First, we carry out stability analysis on the optimal expected risk of a generic expected loss minimization problem w.r.t. perturbation of the probability distribution of the underlying random data. This kind of analysis is well known in stochastic programming (see Römisch (2003) 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 (Hampel, 1971; Huber, 1981; Huber and Ronchetti, 2009) and is recently applied to risk management, where empirical data are used to estimate risk measures of some random losses by Cont, Deguest and Scandolo (2010), Krätschmer, Schied and Zähle (2012); Krätschmer, Schied and Zähle (2014) and optimization by Guo and Xu (2020); Jiang and Li (2022); Xu and Zhang (2022). 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 estimators w.r.t. 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 regression 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 $\mathbb{E}_P[c(z, f(x))] := \int_Z c(z, f(x))P(dz)$ 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.,

$$\inf_{f \in \mathcal{F}} R(f) := \mathbb{E}_P[c(z, f(x))], \tag{1}$$

where $\mathcal{F}$ is some functional class to be specified. Let $\vartheta(P)$ denote the optimal value and $S^*(P)$ the set of optimal solutions of (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 as our focus will be mainly on regression models, which means $0 \leq \vartheta(P) < +\infty$ so long as there exists $f \in \mathcal{F}$ such that $R(f) < +\infty$. Existence of an optimal solution requires more conditions, we will come back to this in the next subsection. In practice, $\mathcal{F}$, Z and $c(\cdot, \cdot)$ are known to learners. Here we list a few examples (Shalev-Shwartz et al., 2010).

- **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 $L(t) = \frac{1}{2}t^2$, ϵ-insensitive loss function $L(t) = \max\{0, |t| - \epsilon\}$ with $\epsilon > 0$, hinge loss function $L(t) = \max\{0, 1 - t\}$, log-loss function $L(t) = \log(1 + e^{-t})$, Huber loss function $L_\alpha(t) = t^2/2$ for $|t| \leq \alpha$ and $\alpha|t| - \alpha^2/2$ otherwise where α is some positive constant, p-th power absolute loss function $L(t) = |t|^p$ for $p > 0$ in various regression and support vector machine models, see Shafieezadeh-Abadeh et al. (2019).

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

In the rest of paper, our focus will be on the regression models.

2.1 Reproducing kernel Hilbert space

The nature of functions f in (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 1 *Let $\mathcal{H}(X)$ be a Hilbert space of functions with inner product $\langle \cdot, \cdot \rangle$ and $k : X \times X \to \mathbb{R}$ 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 $k : X \times X \to \mathbb{R}$ 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 norm is denoted by $\| \cdot \|_k$.*

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 semidefinite if for any finite set $\{x_1, \cdots, x_m\} \subset X$, the $m \times m$ matrix $k[x]$ whose (i, j) entry is $k(x_i, x_j)$ is positive semidefinite. 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.

- **Gaussian kernel**: $k(x_1, x_2) = e^{-\gamma \|x_1 - x_2\|_2^2}, \forall x_1, x_2 \in \mathbb{R}^n$, where $\gamma > 0$ is a constant.

- **Laplacian kernel**: $k(x_1, x_2) = e^{-\gamma\|x_1 - x_2\|_1}, \forall x_1, x_2 \in \mathbb{R}^n$, where $\gamma > 0$ is a constant.
- **Sigmoid kernel**: $k(x_1, x_2) = \tanh\left(a\langle x_1, x_2\rangle + b\right), \forall x_1, x_2 \in \mathbb{R}^n$, 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 Mercer kernel. Then there exist 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 Theorem 5.2 in Mohri et al. (2012). 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 Boucheron et al. (2005). Throughout the paper, we assume that a Mercer kernel $k(\cdot, \cdot)$ is given and $\mathcal{H}_k$ is the RKHS associated with k. The functional class $\mathcal{F}$ in (1) is a subset of $\mathcal{H}_k$ and Z is a Polish space.

Before concluding this subsection, we come back to address our earlier question as to when problem (1) has an optimal solution. The next theorem addresses this.

Theorem 1 *Assume: (a) there exists a constant α such that the lower level set $\{f \in \mathcal{F} : R(f) \le \alpha\}$ is nonempty and bounded, (b) $c(z, y)$ is convex in y for each z and c is continuous over $Z \times Y$, (c) there is a function ϕ such that*

$$c(z, f(x)) \le \phi(z), \forall z \in Z \text{ and } f \in \mathcal{F}.$$

Then problem (1) has an optimal solution when $\int_Z \phi(z) P(dz) < \infty$.

The existence result is perhaps known, for instance, Theorem 5.2 in Steinwart and Christmann (2008) shows existence of an optimal solution for a similar learning model with Nemitski loss function. Here we include a proof as the setting is slightly different.

Proof. We first show that $R(f)$ is continuous in f, where $R(f)$ is defined as in (1). For each $x \in X$,

$$|\tilde{f}(x) - f(x)| = |\langle \tilde{f} - f, k(\cdot, x)\rangle| \le \|\tilde{f} - f\|_k \sqrt{k(x, x)}.$$

It follows from the continuity of $c(x, \cdot)$ that for each $z \in Z$,

$$|c(z, \tilde{f}(x) - c(z, f(x))| \to 0 \text{ as } \|\tilde{f} - f\|_k \to 0.$$

Since $R(f) \leq \int_Z \phi(z)P(dz) < \infty$ for all $f \in \mathcal{F}$, by the Lebesgue dominated convergence theorem,

$$\begin{aligned}\lim_{\|\tilde{f}-f\|_k\to 0} |R(\tilde{f}) - R(f)| &= \lim_{\|\tilde{f}-f\|_k\to 0} \left| \int_Z c(z, \tilde{f}(x))P(dz) - \int_Z c(z, f(x))P(dz) \right| \\ &= \left| \int_Z \lim_{\|\tilde{f}-f\|_k\to 0} (c(z, \tilde{f}(x)) - c(z, f(x)))P(dz) \right| \\ &= 0,\end{aligned}$$

which shows continuity of R in f as desired. Moreover, since $c(z, y)$ is convex in y, $R(f)$ is also convex. Together with condition (a), we conclude by virtue of Proposition 6 on page 75 of Ekeland and Turnbull (1983) that R attains minimum in $\mathcal{F}$. ■

Condition (a) is known as inf-compactness condition which is widely used for securing existence of an optimal solution in the literature of continuous optimization, see e.g. Rockafellar and Wets (1998). It is satisfied when either $\mathcal{F}$ is bounded and/or $c(z, \cdot)$ is coercive for almost every fixed z. Condition (c) is a kind of growth condition to be used for securing the well-definedness of $R(f)$. To ease the discussion, we assume in the rest of the paper that $\mathcal{F}$ is bounded, that is, there exists a positive number β such that $\|f\|_k \leq \beta$ for all $f \in \mathcal{F}$, see e.g. Norkin and Keyzer (2009).

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.) sample $\{z^i = (x^i, y^i)\}_{i=1}^N$ 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

$$\inf_{f\in\mathcal{F}} \mathbb{E}_{P_N}[c(z, f(x))] := \frac{1}{N}\sum_{i=1}^N c(z^i, f(x^i)), \tag{2}$$

where

$$P_N(\cdot) := \frac{1}{N}\sum_{i=1}^N \mathbb{1}_{z^i}(\cdot) \tag{3}$$

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 $S^*_{P_N}$ the set of optimal solutions of the sample average approximation problem (2). Let $f_N(P_N) \in S^*_{P_N}$ denote an optimal solution of (2). Then $f_N(P_N)$ is called an estimator and the framework generating $f_N(P_N)$ is called a learning algorithm. Notice that from sampling point of view, we may write $\hat{\vartheta}_N(z^1, \cdots, z^N)$ and $\hat{f}_N(z^1, \cdots, z^N)$ for $\vartheta(P_N)$ and $f_N(P_N)$ respectively to indicate their dependence on the sample.

From computational perspective, problem (2) is often ill-conditioned. The issue can be addressed by adopting a simple Tikhonov regularization approach:

$$\vartheta(P_N, \lambda_N) = \inf_{f\in\mathcal{F}} R^{\lambda_N}_{P_N}(f) := \mathbb{E}_{P_N}[c(z, f(x))] + \lambda_N\|f\|_k^2, \tag{4}$$

where $\lambda_N > 0$ is a regularization parameter. Note that problem (4) is well-defined even when $\mathcal{F}$ is unbounded since the objective is coercive for each fixed N. Let $S^*_{P_N,\lambda_N}$ denote the set of optimal solutions of problem (4) and $f_N(P_N,\lambda_N) \in S^*_{P_N,\lambda_N}$ an optimal solution. Under conditions in Theorem 1, we can show that a unique optimal solution exists. By virtue of the representer theorem (see Kimeldorf and Wahba (1970), Schölkopf and Smola (2002)), problem (4) has a solution which takes the form $f_N^{\lambda_N}(x) = \sum_{j=1}^N \alpha_j k(x_j, x)$ and by the reproducing property (Norkin and Keyzer, 2009), $\|f_N^{\lambda_N}\|_k^2 = \langle f_N^{\lambda_N}, f_N^{\lambda_N}\rangle = \sum_{i,j=1}^N \alpha_i\alpha_j k(x_i, x_j)$. As we commented earlier, here we may write $\hat{\vartheta}_N(z^1,\cdots,z^N,\lambda_N)$ and $\hat{f}_N(z^1,\cdots,z^N,\lambda_N)$ for $\vartheta(P_N,\lambda_N)$ and $f_N(P_N,\lambda_N)$ respectively to indicate their dependence on the sample.

In general λ_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, Breheny and Huang (2015) for logistic regression models in a finite dimensional space, Cucker and Smale (2002a) and Caponnetto and De Vito (2007) for regularized least squares models in a infinite dimensional RKHS.

Note also that under some special circumstances, the regularization may be interpreted as a result of robust formulation or distributionally robust formulation of problem (2). For instance, Xu et al. (2009) consider the case where the input data are potentially contaminated and show that a robust version of the model is equivalent to a regularized regression model. Chen and Paschalidis (2018) consider a linear regression model where the true probability distribution of input-output data is unknown but it is possible to use empirical data to construct a Wasserstein ball of probability distributions, the optimal solution is based on the worst probability distribution from the ball. Under these circumstance, the authors demonstrate that the distributionally robust regression model is equivalent to a regularized regression model where the regularization parameter is the radius of the Wasserstein ball. Shafieezadeh-Abadeh et al. (2019) extend the result to a nonlinear regression model, see Theorem 28 in the paper.

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. As discussed in the introduction, this assumption may not be satisfied in practice. Let $\tilde{z}^1,\cdots,\tilde{z}^N$ denote the perceived data which are potentially contaminated and

$$Q_N(\cdot) := \frac{1}{N}\sum_{i=1}^N \mathbb{1}_{\tilde{z}^i}(\cdot) \tag{5}$$

be the respective empirical distribution. Instead of solving problem (4), we solve, in practice,

$$\inf_{f\in\mathcal{F}} \mathbb{E}_{Q_N}[c(z,f(x))] + \lambda_N\|f\|_k^2. \tag{6}$$

The coerciveness of the objective function ensures well-definedness of the problem. Let $R_{Q_N}^{\lambda_N}(f)$, $\vartheta(Q_N,\lambda_N)$ and $f_N(Q_N,\lambda_N)$ denote respectively the objective function, the optimal value and the optimal solution of problem (6). We are then concerned with the quality of the learning model estimator $f_N(Q_N,\lambda_N)$ and the associated empirical risk $\vartheta(Q_N,\lambda_N)$. Measurability of these quantities are guaranteed by Lemma 6.23 and Lemma A.3.18 in

Steinwart and Christmann (2008). Note that in this setup, we often assume that $\tilde{z}^1, \cdots, \tilde{z}^N$ are i.i.d.. This assumption differs from Lecué and Lerasle (2020) where the authors divide the data into two categories: the outliers and the informative data. The former are usually non-independent and not identically distributed. Here we follow the stream of new research on statistical robustness in risk measurement led by Cont, Deguest and Scandolo (2010); Krätschmer, Schied and Zähle (2012); Krätschmer, Schied and Zähle (2014) to assume that each data could be potentially contaminated and they are generated by some underlying probability distribution Q.

Note also that if we interpret the regularization model (4) as a result of distributionally robust formulation (see Theorem 28 of Shafieezadeh-Abadeh et al. (2019)), then model (6) may be interpreted as an equivalence of a distributionally robust optimization (DRO) model where the sample data used for constructing the nominal empirical distribution in the Wasserstein ball are potentially contaminated. In our view, the DRO models in Chen and Paschalidis (2018); Shafieezadeh-Abadeh et al. (2019) are not about contamination of training data, rather they are about incomplete information of the true probability distribution. The DRO model picks up the worst estimate of the true probability distribution rather than the worst perturbed data. This issue disappears when the sample size increases, but the data contamination issue persists. The analysis in this paper has a potential to address the issue in DRO models via (6).

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 λ_N goes to zero. Assume without loss of generality that the samples are i.i.d.. By law of large numbers, Q_N converges to some probability distribution Q almost surely (a.s. for short) as N goes to infinity and subsequently

$$f_N(Q_N, \lambda_N) \to f(Q) \quad \text{and} \quad \vartheta(Q_N, \lambda_N) \to \vartheta(Q), \text{ a.s..} \tag{7}$$

On the other hand, if we regard Q as a perturbation of the true unknown probability distribution P, then we need to investigate whether

$$f(Q) \to f(P) \quad \text{and} \quad \vartheta(Q) \to \vartheta(P) \tag{8}$$

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 (Römisch, 2003). However, if we want to establish

$$f_N(Q_N, \lambda_N) \to f(P) \quad \text{and} \quad \vartheta(Q_N, \lambda_N) \to \vartheta(P), \text{ a.s.}, \tag{9}$$

then we require not only (8) but also (7) 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 Cont, Deguest and Scandolo (2010); Guo and Xu (2020); Krätschmer, Schied and Zähle (2014); Krätschmer, Schied and Zähle (2012) 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 P in the space with ϕ-weak topology instead of usual weak convergence. This is primarily because we need to capture interactions between the tails of the cost function and the kernel and the tail of the probability distribution of z.

2.4 ϕ-weak topology

We recall some basic concepts and results about weak topology which are needed for the analysis. The materials are mainly extracted from Claus (2016), we refer readers to Chapter 2 in Claus (2016) and references therein for a more comprehensive discussion on the subject.

Definition 2 *Let $\phi : Z \to [0, \infty)$ be a continuous function and*

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

where $\mathscr{P}(Z)$ 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 $\mathscr{P}(Z)$ which satisfies the generalized moment condition of ϕ.

Definition 3 (ϕ-weak topology) *Let $\phi : Z \to [0, \infty)$ be a gauge function, that is, $\phi \geq 1$ holds outside a compact set. Define $\mathcal{C}_Z^\phi$ the linear space of all continuous functions $h : Z \to \mathbb{R}$ such that for each $h \in \mathcal{C}_Z^\phi$, there exists a positive constant C_h such that*

$$|h(z)| \leq C_h(\phi(z) + 1), \forall z \in Z.$$

The ϕ-weak topology, denoted by τ_ϕ, is the coarsest topology on $\mathcal{M}_Z^\phi$ for which the mapping $g_h : \mathcal{M}_Z^\phi \to \mathbb{R}$ defined by

$$g_h(P) := \int_Z h(z) P(dz), \ h \in \mathcal{C}_Z^\phi$$

is continuous. A sequence $\{P_l\} \subset \mathcal{M}_Z^\phi$ is said to converge ϕ-weakly to $P \in \mathcal{M}_Z^\phi$ written $P_l \xrightarrow{\phi} P$ if it converges w.r.t. τ_ϕ.

From the definition, we can see immediately that ϕ-weak convergence implies weak convergence under usual topology of weak convergence (defined through bounded continuous functions). We denote the latter by $P_l \xrightarrow{w} P$. Moreover, it follows by Corollary 2.62 in Claus (2016) that the ϕ-weak topology on $\mathcal{M}_Z^\phi$ is generated by the metric $\mathsf{dl}_\phi : \mathcal{M}_Z^\phi \times \mathcal{M}_Z^\phi \to \mathbb{R}$ defined by

$$\mathsf{dl}_\phi(P', P'') := \mathsf{dl}_{\text{Prok}}(P', P'') + \left| \int_Z \phi(z) dP' - \int_Z \phi(z) dP'' \right| \text{ for } P', P'' \in \mathcal{M}_Z^\phi, \tag{10}$$

where $\mathsf{dl}_{\text{Prok}} : \mathscr{P}(Z) \times \mathscr{P}(Z) \to \mathbb{R}_+$ is the Prokhorov metric defined as follows:

$$\mathsf{dl}_{\text{Prok}}(P', P'') := \inf\{\epsilon > 0 : P'(A) \leq P''(A^\epsilon) + \epsilon \text{ for all } A \in \mathcal{B}(Z)\}, \tag{11}$$

where $A^\epsilon := A + B_\epsilon(0)$ 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 (10) disappears and consequently $d_\phi(P', P'') = d_{\text{Prok}}(P', P'')$. In that case, the ϕ-weak topology reduces to the usual topology of weak convergence. Equivalence between the two topologies may be established over a set which satisfies some uniform integration conditions, see Lemma 2.66 in Claus (2016) and the reference therein.

Definition 4 (Fortet-Mourier metric) *Let*

$$\mathcal{F}_p(Z) := \{\psi : Z \to \mathbb{R} : |\psi(z) - \psi(\tilde{z})| \leq L_p(z, \tilde{z})\|z - \tilde{z}\|, \forall z, \tilde{z} \in Z\}, \tag{12}$$

where $\|\cdot\|$ denotes some norm on Z and $L_p(z, \tilde{z}) := \max\{1, \|z\|, \|\tilde{z}\|\}^{p-1}$ 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 $\mathscr{P}(Z)$ is defined by

$$\zeta_p(P, Q) := \sup_{\psi \in \mathcal{F}_p(Z)} \left| \int_Z \psi(z) P(dz) - \int_Z \psi(z) Q(dz) \right|. \tag{13}$$

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 Kantorovich metric

$$\mathsf{dl}_{K,Z}(P, Q) := \sup_{\psi \in \mathcal{F}_1(Z)} \left| \int_Z \psi(z) P(dz) - \int_Z \psi(z) Q(dz) \right|. \tag{14}$$

If $Z = \mathbb{R}$, $\mathsf{dl}_{K,\mathbb{R}}(P, Q)$ is denoted by $\mathsf{dl}_{K,1}(P, Q)$ for simplicity. We refer readers to see Römisch (2003) for a comprehensive overview of the topic. From the definition, we can see that

$$\zeta_p(P, Q) \leq \mathbb{E}_{P \times Q}[L_p(z, \tilde{z})\|z - \tilde{z}\|],$$

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

$$\mathbb{E}_{P \times Q}[L_p(z, \tilde{z})\|z - \tilde{z}\|] = \frac{1}{N^2} \sum_{i,j=1}^{N} L_p(z^i, \tilde{z}^j)\|z^i - \tilde{z}^j\|.$$

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

3. Stability analysis

In this section, we investigate how the true risk of problem (1) is affected by a small perturbation of the probability distribution P. This kind of research is well known in the literature of stochastic programming (Römisch, 2003) 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 1 *For any compact subset Z_0 of Z, let X_0 be its orthogonal projection on X. The set of functions $\{k(\cdot, x) : x \in X_0\}$ is uniformly continuous over 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.

Remark 1 *To see how Assumption 1 can be possibly satisfied, we recall the notion of calmness of kernel introduced by Assumption 25 in Shafieezadeh-Abadeh et al. (2019). The kernel 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, for any $\epsilon > 0$, there exists $\eta > 0$ such that

$$\begin{array}{rcl} \|k(\cdot, x') - k(\cdot, x)\|_k & = & \sqrt{\langle k(\cdot, x') - k(\cdot, x), k(\cdot, x') - k(\cdot, x)\rangle} \\ & = & \sqrt{k(x', x') - 2k(x, x') + k(x, x)} \\ & \leq & g(\|x - x'\|) < \epsilon \end{array}$$

for all x, x' with $\|x - x'\| < \eta$. The last inequality is due to the fact that the growth function g is continuous with $g(0) = 0$. The calmness condition is non-restrictive, which can be satisfied in the following cases for $X = \mathbb{R}^n$, see Example 1 in Shafieezadeh-Abadeh et al. (2019).

- *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.*
- *Polynominal kernel: The kernel $k(x_1, x_2) = (\gamma\langle x_1, x_2\rangle + 1)^d$ 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 w.r.t. the growth function*

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

Assumption 2 *The cost function $c(\cdot, \cdot)$ satisfies the following properties.*

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

$$c(z, f(x)) \leq \phi(z), \forall z \in Z \text{ and } f \in \mathcal{F}, \tag{15}$$

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

(b) $c : Z \times Y \to \mathbb{R}$ *is uniformly continuous over any compact subset of* $Z \times Y$.

Remark 2 *Inequality (15) 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* ϕ *depends on the concrete structure of* $c(\cdot,\cdot)$. *Consider for example* $c(z, f(x)) = \frac{1}{2}\|y - f(x)\|^2$. *Then*

$$\begin{aligned} c(z, f(x)) &\leq \|y\|^2 + \|f(x)\|^2 = \|y\|^2 + |\langle f, k(\cdot, x)\rangle|^2 \\ &\leq \|y\|^2 + \|f\|_k^2 \|k(\cdot, x)\|_k^2. \end{aligned}$$

Moreover, if there esists a positive number β *such that* $\|f\|_k \leq \beta$, *then we can work out an explicit form of* ϕ *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 kernel or Laplacian kernel, then* $\|k(\cdot, x)\|_k^2 = 0$ *and* $\phi(z) := \|y\|^2$.
- *If* k *is a Polynominal kernel, then* $\|k(\cdot, x)\|_k^2 = (\gamma\|x\|^2 + 1)^d$ *and*

$$\phi(z) := \|y\|^2 + \beta^2(\gamma\|x\|^2 + 1)^d. \tag{16}$$

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

Theorem 2 *Assume that* $\mathcal{F}$ *is bounded with* $\|f\|_k \leq \beta$ *for all* $f \in \mathcal{F}$ *and Assumptions 1 and 2 hold. Then*

$$\lim_{P' \xrightarrow{\phi} P} \vartheta(P') = \vartheta(P). \tag{17}$$

Proof. Since $(\mathcal{M}_Z^\phi, \tau_\phi)$ is a Polish space by Theorem 2.59 Claus (2016), it suffices to show that (17) holds for any sequence $\{P_l\} \subset \mathcal{M}_Z^\phi$ with $P_l \xrightarrow{\phi} P \in \mathcal{M}_Z^\phi$. First, $P_l \xrightarrow{\phi} P$ implies that $P_l \xrightarrow{w} P$ and

$$\lim_{l\to\infty} \int_Z \phi(z) P_l(dz) = \int_Z \phi(z) P(dz).$$

Moreover, by Lemma 2.61 in Claus (2016), for any $\epsilon > 0$, there exists a positive constant $M_0 > 1$ such that

$$\int_Z \phi(z) \mathbf{1}_{(M_0,\infty)}(\phi(z)) P(dz) < \epsilon \tag{18}$$

and

$$\sup_{l\in\mathbb{N}} \int_Z \phi(z) \mathbf{1}_{(M_0,\infty)}(\phi(z)) P_l(dz) < \epsilon, \tag{19}$$

where $\mathbf{1}_{(M_0,\infty)}(t) = 1$ if $t \in (M_0, \infty)$ and 0 otherwise. Let $M > M_0$ and $Z_M := \mathrm{cl}\{z \in Z : \phi(z) < M\}$. Since ϕ is continuous, $\partial Z_M \subset \{z \in Z : \phi(z) = M\}$, where ∂Z_M denotes the

boundary of Z_M. This means Z_M is a P-continuity set except for countably many M [1]. Thus we can choose some M such that $P(\partial Z_M) = 0$ and $Z \backslash Z_M \subset \{z \in Z : \phi(z) \geq M\} \subset \{z \in Z : \phi(z) > M_0\}$. Moreover since ϕ is coercive, i.e., $\phi(z) \to \infty$ as $\|z\| \to \infty$, then Z_M is a compact set. Let $\mathscr{G} := \{g : g(z) := c(z, f(x)) \text{ for } f \in \mathcal{F}\}$ and

$$\mathscr{G}_M := \{g_M : Z_M \to \mathbb{R} | g_M(z) := g(z) \text{ for } z \in Z_M, g \in \mathscr{G}\}.$$

It follows from Assumption 2 (a) that for each $g_M \in \mathscr{G}_M$ and $z \in Z_M$, $|g_M(z)| \leq \sup_{z \in Z_M} \phi(z) < \infty$, which implies that $\mathscr{G}_M$ is uniformly bounded.

Next, we prove that $\mathscr{G}_M$ is equi-continuous over Z_M. Since $\mathcal{F}$ is bounded, problem (1) is equivalent to

$$\min_{\|f\|_k \leq \beta} R(f). \tag{20}$$

By the reproducing property of the kernel $k(\cdot, \cdot)$, i.e., $f(x) = \langle f, k(\cdot, x) \rangle$ for every $f \in \mathcal{F}$, we have

$$\begin{aligned} |f(x') - f(x)| &= |\langle f, k(\cdot, x') \rangle - \langle f, k(\cdot, x) \rangle| \leq \|f\|_k \|k(\cdot, x') - k(\cdot, x)\|_k \\ &\leq \beta \|k(\cdot, x') - k(\cdot, x)\|_k. \end{aligned} \tag{21}$$

The uniform continuity of $k(\cdot, x)$ over Z_M (under Assumption 1) ensures the equicontinuity of $\mathcal{F}$ over Z_M. Moreover, under Assumption 2(b), $\mathscr{G}_M$ is also equicontinuous 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 $Q(A) = P(A)$ for $A \in \mathcal{B}(Z_M)$ respectively. Since Z_M is a continuity set of P, then $P_l \xrightarrow{w} P$ imply $Q_l \xrightarrow{w} Q$. Since $\mathscr{G}_M$ is uniformly bounded and equi-continuous, by Theorem 3.1 in Rao (1962),

$$\lim_{l \to \infty} \sup_{g_M \in \mathscr{G}_M} \left| \int_{Z_M} g_M(z) Q_l(dz) - \int_{Z_M} g_M(z) Q(dz) \right| = 0. \tag{22}$$

On the other hand, under the growth condition (15), (18) and (19) imply

$$\begin{aligned} \sup_{g \in \mathscr{G}} \int_{Z \backslash Z_M} |g(z)| P(dz) &\leq \int_{Z \backslash Z_M} \phi(z) P(dz) \\ &\leq \int_Z \phi(z) \mathbf{1}_{(M_0, \infty)}(\phi(z)) P(dz) < \epsilon \end{aligned} \tag{23}$$

and

$$\begin{aligned} \sup_{g \in \mathscr{G}} \sup_{l \in \mathbb{N}} \int_{Z \backslash Z_M} |g(z)| P_l(dz) &\leq \sup_{l \in \mathbb{N}} \int_{Z \backslash Z_M} \phi(z) P_l(dz) \\ &\leq \sup_{l \in \mathbb{N}} \int_Z \phi(z) \mathbf{1}_{(M_0, \infty)}(\phi(z)) P_l(dz) < \epsilon. \end{aligned} \tag{24}$$

1. A set S is said to be a P-continuity set if $P(\partial S) = 0$, see Section 2 in Chapter 1 in Billingsley (1999) and the proof of Theorem 2.1 in Billingsley (1999).

Together with (22), we have

$$\begin{aligned}
|\vartheta(P_l) - \vartheta(P)| &\leq \sup_{f\in\mathcal{F}} \left| \int_Z c(z, f(x)) P_l(dz) - \int_Z c(z, f(x)) P(dz) \right| \\
&= \sup_{g\in\mathscr{G}} \left| \int_Z g(z) P_l(dz) - \int_Z g(z) P(dz) \right| \\
&\leq \sup_{g\in\mathscr{G}} \left| \int_{Z_M} g(z) P_l(dz) - \int_{Z_M} g(z) P(dz) \right| \\
&\quad + \int_{Z\setminus Z_M} |g(z)| P(dz) + \int_{Z\setminus Z_M} |g(z)| P_l(dz) \\
&\leq \sup_{g_M\in\mathscr{G}_M} \left| \int_{Z_M} g_M(z) Q_l(dz) - \int_{Z_M} g_M(z) Q(dz) \right| + 2\epsilon < 3\epsilon
\end{aligned}$$

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 ϕ-weak topology. Observe that the empirical probability measure $P_N \in \mathcal{M}_Z^\phi$. Moreover, by Theorem 11.4.1 in Dudley (2004), P_N converges to P a.s., and since $\mathbb{E}_P[\phi] < \infty$, it follows by the strong law of large numbers (see e.g. Theorem 8.3.5 in Dudley (2004)), $\mathbb{E}_{P_N}[\phi] \to \mathbb{E}_P[\phi]$ a.s.. Together, we conclude by definition that $P_N \xrightarrow{\phi} P$ a.s. as $N \to \infty$. Consequently we have

$$\lim_{N\to\infty} \vartheta(P_N) = \vartheta(P), a.s.. \tag{25}$$

The topological structure of $\mathcal{M}_Z^\phi$ affects the stability of $\vartheta(\cdot)$: a larger $\mathcal{M}_Z^\phi$ 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, $\mathcal{M}_Z^\phi = \mathscr{P}(Z)$, 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 $\mathcal{M}_Z^\phi$, we explain this through next example.

Example 1 *Consider the least squares regression model with Polynomial kernel. By (16)*

$$\begin{aligned}
\mathcal{M}_Z^\phi &= \left\{ P \in \mathscr{P}(Z) : \int_Z \left[\|y\|^2 + \beta^2(\gamma\|x\|^2 + 1)^d \right] P(dz) < \infty \right\} \\
&= \left\{ P \in \mathscr{P}(Z) : \int_Z \|y\|^2 P(dz) < \infty, \int_Z \|x\|^{2d} P(dz) < \infty \right\}.
\end{aligned}$$

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

In the case of Gaussian kernel or Laplacian kernel,

$$\mathcal{M}_Z^\phi = \left\{ P \in \mathscr{P}(Z) : \int_Z \|y\|^2 P(dz) < \infty \right\},$$

which is the set of probability measures with finite second order moment of y. In all these cases, $\mathcal{M}_Z^\phi$ consists of sub-Gaussian distributions on $\mathscr{P}(Z)$.

It might be interesting to ask whether we will be able to show the continuity of $\vartheta(\cdot)$ by using the well-known maximum theorem in parametric programming. The answer is yes but it requires similar conditions and we will not be able to simplify the proof. Let us explain why.

A key step to use the maximum theorem is to show that $\mathbb{E}_P[c(z, f(x))]$ is jointly continuous in (f, P). Let $\tilde{f}, f \in \mathcal{F}$ and $\tilde{P}, P \in \mathcal{M}_Z^{\phi}$. Observe that

$$\begin{aligned} & |\mathbb{E}_{\tilde{P}}[c(z, \tilde{f}(x))] - \mathbb{E}_P[c(z, f(x)]| \\ \leq \; & |\mathbb{E}_{\tilde{P}}[c(z, \tilde{f}(x))] - \mathbb{E}_{\tilde{P}}[c(z, f(x)]| + |\mathbb{E}_{\tilde{P}}[c(z, f(x))] - \mathbb{E}_P[c(z, f(x)]|. \end{aligned}$$

Assumption 2 (a) ensures

$$\lim_{\tilde{P} \xrightarrow{\phi} P} |\mathbb{E}_{\tilde{P}}[c(z, f(x))] - \mathbb{E}_P[c(z, f(x)]| = 0. \tag{26}$$

On the other hand, under Assumption 2 (a),

$$|\mathbb{E}_{\tilde{P}}[c(z, \tilde{f}(x))] - \mathbb{E}_{\tilde{P}}[c(z, f(x)]| \leq 2\mathbb{E}_{\tilde{P}}[\phi(z)] < \infty.$$

By the Lebesgue dominated convergence theorem,

$$\lim_{\tilde{f} \to f} |\mathbb{E}_{\tilde{P}}[c(z, \tilde{f}(x))] - \mathbb{E}_{\tilde{P}}[c(z, f(x)]| \;=\; |\mathbb{E}_{\tilde{P}}[\lim_{\tilde{f} \to f}(c(z, \tilde{f}(x)) - c(z, f(x)))]| = 0. \tag{27}$$

However, (27) holds only for fixed $\tilde{P}$ but we need the equality hold uniformly for all $\tilde{P}$ close to P, which in turn requires some delicate handling of the tails as in the proof of Theorem 2. We leave interested readers for an exercise.

Finally, we note that our stability result in Theorem 2 should be distinguished from those in Shalev-Shwartz et al. (2010) where stability is used to examine the difference of the costs resulting from kernel learning estimators based on different samples. It should also be differentiated from classical stability results in stochastic programming under pseudo-metric:

$$\mathsf{dl}_{\mathscr{G}}(Q, P) := \sup_{\psi \in \mathscr{G}} \left| \int_Z \psi(z) Q(dz) - \int_Z \psi(z) P(dz) \right|, \tag{28}$$

where $\mathscr{G}$ is a class of measurable functions mapping from Z to $\mathbb{R}$ (see e.g. Römisch and Schultz (1991); Römisch (2003)). To see this, let $\{Q_l\} \subset \mathscr{P}(Z)$ be a sequence of probability measures converging to P under the metric, i.e.,

$$\mathsf{dl}_{\mathscr{G}}(Q_l, P) = \sup_{\psi \in \mathscr{G}} \left| \int_Z \psi(z) Q_l(dz) - \int_Z \psi(z) P(dz) \right| \to 0. \tag{29}$$

The convergence above does not indicate under which topology Q_l converges to P. Conversely the convergence of Q_l to P under ϕ-weak topology may not guarantee the uniform convergence (29). However, if $\mathscr{G}$ is equicontinuous and $|\psi(z)| \leq \phi(z)$ for all $\psi \in \mathscr{G}$ and $z \in Z$, then the convergence of Q_l to P under ϕ-weak topology implies $\mathsf{dl}_{\mathscr{G}}(Q_l, P) \to 0$, see Theorem 3.2 of Rao (1962). In a particular case that $\mathscr{G}$ comprises all Lipschitz continuous functions with modulus 1, $\mathsf{dl}_{\mathscr{G}}(Q, P)$ reduces to the well-known Kantorovich metric. Since the Kantorovich metric generates the ϕ-weak topology with $\phi(z) = \|z\|$ in the finite dimensional space (see Proposition 2.63 of Claus (2016)), then the two convergences are equivalent.

4. Statistical robustness

We now move on to discuss statistical robustness of the machine learning model (4). 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 $\left(Z^{\otimes N}, \mathcal{B}(Z)^{\otimes N}\right)$ with marginal P and $Q^{\otimes N}$ with marginal Q. We will consider statistical estimators mapping from $\left(Z^{\otimes N}, \mathcal{B}(Z)^{\otimes N}\right)$ to $\mathbb{R}$ and examine their convergence.

4.1 Qualitative robustness

We begin by a formal definition of statistical estimator $T(\cdot, \lambda)$ parameterized by λ, where $T(\cdot, \lambda)$ maps from a subset of $\mathcal{M} \subset \mathscr{P}(Z)$ to $\mathbb{R}$. To ease the exposition, we write $\vec{z}^N$ for $(z^1, \cdots, z^N)$ and $\hat{T}_N(\vec{z}^N, \lambda_N)$ for $T(P_N, \lambda_N)$ for fixed sample size N. The following definition is based on Definition 2.11 in Krätschmer, Schied and Zähle (2014).

Definition 5 (Statistical robustness) *Let $\mathcal{M} \subset \mathscr{P}(Z)$ be a set of probability measures and dl_ϕ be defined as in (10) 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 dl_ϕ and $\mathsf{dl}_{\text{Prok}}$ if for all $P \in \mathcal{M}$ and $\epsilon > 0$, there exist $\delta > 0$ and $N_0 \in \mathbb{N}$ such that*

$$Q \in \mathcal{M}, \mathsf{dl}_\phi(P, Q) \leq \delta \Longrightarrow \mathsf{dl}_{\text{Prok}}\left(P^{\otimes N} \circ \hat{T}_N(\cdot, \lambda_N)^{-1}, Q^{\otimes N} \circ \hat{T}_N(\cdot, \lambda_N)^{-1}\right) \leq \epsilon \text{ for } N \geq N_0.$$

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 $\left(Z^{\otimes N}, \mathcal{B}(Z)^{\otimes N}, P^{\otimes N}\right)$ and $\left(Z^{\otimes N}, \mathcal{B}(Z)^{\otimes N}, Q^{\otimes N}\right)$ respectively to $\mathbb{R}$, and the Prokohorov metric is used to measure the difference of the two distributions (also known as laws in the literature (Cont, Deguest and Scandolo, 2010; Krätschmer, Schied and Zähle , 2014)). The statistical robustness requires the difference of statistical estimators under the Prokhorov metric to be small when the difference between P and Q is small under the metric dl_ϕ. The definition relies heavily on the adoption of the two metrics. In Cont, Deguest and Scandolo (2010), the authors use Lévy metric for both. Krätschmer, Schied and Zähle (2014) argue that the Levy metric underestimates the impact of the tail distributions of P and Q and subsequently propose to use dl_ϕ to replace the Lévy metric. Since the former is tighter than the later, it means the perturbation under dl_ϕ 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 ϵ and δ. To establish the statistical robustness, we need the following Uniform Glivenko-Cantelli property.

Definition 6 (Uniform Glivenko-Cantelli property) *Let ϕ be a gauge function and dl_ϕ be defined as in (10). Let $\mathcal{M}$ be a subset of $\mathcal{M}_Z^\phi$. The metric space $(\mathcal{M}, \mathsf{dl}_\phi)$ 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*

$$P^{\otimes N}\left(\vec{z}^N : \mathsf{dl}_\phi(P, P_N) \geq \delta\right) \leq \epsilon \text{ for all } P \in \mathcal{M}, N \geq N_0. \tag{30}$$

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

Theorem 3 (Statistical robustness) *Let*

$$\mathcal{M}_{Z,\kappa}^{\phi^p} := \{P \in \mathscr{P}(Z) : \int_Z \phi(z)^p P(dz) \leq \kappa\}, \tag{31}$$

where $\kappa > 0$ and $p > 1$ are some positive constants. Assume: (a) $\vartheta(P)$ is well-defined for every $P \in \mathcal{M}_{Z,\kappa}^{\phi^p}$, (b) the conditions in Theorem 2 are satisfied, (c) $\lambda_N \to 0$ as $N \to \infty$. Then for any $\epsilon > 0$ and fixed $P \in \mathcal{M}_{Z,\kappa}^{\phi^p}$, there exist positive numbers $\delta > 0$ and $N_0 \in \mathbb{N}$ such that

$$Q \in \mathcal{M}_{Z,\kappa}^{\phi^p}, \mathsf{dl}_\phi(P,Q) \leq \delta \Longrightarrow \mathsf{dl}_{\text{Prok}}\left(P^{\otimes N} \circ \hat{\vartheta}_N(\cdot,\lambda_N)^{-1}, Q^{\otimes N} \circ \hat{\vartheta}_N(\cdot,\lambda_N)^{-1}\right) \leq \epsilon \tag{32}$$

for $N \geq N_0$ and $\lambda_N \leq \frac{\epsilon}{6\beta^2}$, where $\hat{\vartheta}_N(\vec{z}^N,\lambda_N) := \vartheta(P_N,\lambda_N)$ denotes the optimal value of problem (4), P_N is a sequence of empirical probability measures defined by (3) and β is a positive constant.

Proof. The results follow straightforwardly from Theorem 2 and Theorem 2.4 in Krätschmer, Schied and Zähle (2012). We include a proof for self-containing. By triangle inequality

$$\begin{aligned} & \mathsf{dl}_{\text{Prok}}\left(P^{\otimes N} \circ \hat{\vartheta}_N(\cdot,\lambda_N)^{-1}, Q^{\otimes N} \circ \hat{\vartheta}_N(\cdot,\lambda_N)^{-1}\right) \\ \leq\ & \mathsf{dl}_{\text{Prok}}\left(P^{\otimes N} \circ \hat{\vartheta}_N(\cdot,\lambda_N)^{-1}, \mathbb{1}_{\inf_{f\in\mathcal{F}} R_P(f)}\right) + \mathsf{dl}_{\text{Prok}}\left(\mathbb{1}_{\inf_{f\in\mathcal{F}} R_P(f)}, \mathbb{1}_{\inf_{f\in\mathcal{F}} R_Q(f)}\right) \\ & + \mathsf{dl}_{\text{Prok}}\left(\mathbb{1}_{\inf_{f\in\mathcal{F}} R_Q(f)}, Q^{\otimes N} \circ \hat{\vartheta}_N(\cdot,\lambda_N)^{-1}\right), \end{aligned}$$

where $\mathbb{1}_a$ denotes the Dirac measure at $a \in \mathbb{R}$. Under condition (b), for the given ϵ there exists a constant $\delta_0 > 0$ such that

$$\mathsf{dl}_{\text{Prok}}\left(\mathbb{1}_{\inf_{f\in\mathcal{F}} R_P(f)}, \mathbb{1}_{\inf_{f\in\mathcal{F}} R_Q(f)}\right) = \mathsf{dl}_{\text{Prok}}\left(\mathbb{1}_{\vartheta(P)}, \mathbb{1}_{\vartheta(Q)}\right) = |\vartheta(P) - \vartheta(Q)| \leq \frac{\epsilon}{3}$$

for all $Q \in \mathcal{M}_{Z,\kappa}^{\phi^p}$ with $\mathsf{dl}_\phi(P,Q) \leq \delta_0$. So we are left to show that

$$\mathsf{dl}_{\text{Prok}}\left(P^{\otimes N} \circ \hat{\vartheta}_N(\cdot,\lambda_N)^{-1}, \mathbb{1}_{\inf_{f\in\mathcal{F}} R_P(f)}\right) \leq \frac{\epsilon}{3} \tag{33}$$

and

$$\mathsf{dl}_{\text{Prok}}\left(\mathbb{1}_{\inf_{f\in\mathcal{F}} R_Q(f)}, Q^{\otimes N} \circ \hat{\vartheta}_N(\cdot,\lambda_N)^{-1}\right) \leq \frac{\epsilon}{3} \tag{34}$$

for N sufficiently large. By Strassen's theorem (Huber, 1981), (33) and (34) are implied respectively by

$$P^{\otimes N}\left(\vec{z}^N : \left|\hat{\vartheta}_N(\vec{z}^N,\lambda_N) - \inf_{f\in\mathcal{F}} R_P(f)\right| \geq \frac{\epsilon}{3}\right) \leq \frac{\epsilon}{3} \tag{35}$$

and

$$Q^{\otimes N}\left(\tilde{\vec{z}}^N : \left|\hat{\vartheta}_N(\tilde{\vec{z}}^N, \lambda_N) - \inf_{f\in\mathcal{F}} R_Q(f)\right| \geq \frac{\epsilon}{3}\right) \leq \frac{\epsilon}{3}. \tag{36}$$

Using the definition of the optimal values, (35) and (36) can be rewritten respectively as

$$P^{\otimes N}\left(\vec{z}^N : \left|\inf_{f\in\mathcal{F}} \mathbb{E}_{P_N}\{c(z, f(x)) + \lambda_N \|f\|_k^2\} - \inf_{f\in\mathcal{F}} R_P(f)\right| \geq \frac{\epsilon}{3}\right) \leq \frac{\epsilon}{3},$$
$$Q^{\otimes N}\left(\tilde{\vec{z}}^N : \left|\inf_{f\in\mathcal{F}} \mathbb{E}_{Q_N}\{c(z, f(x)) + \lambda_N \|f\|_k^2\} - \inf_{f\in\mathcal{F}} R_Q(f)\right| \geq \frac{\epsilon}{3}\right) \leq \frac{\epsilon}{3}.$$

Note that we may set $N_0 \in \mathbb{N}$ sufficiently large such that $\lambda_N \leq \frac{\epsilon}{6\beta^2}$ for all $N \geq N_0$. Consequently the two inequalities above are implied by

$$P^{\otimes N}\left(\vec{z}^N : \left|\inf_{f\in\mathcal{F}} R_{P_N}(f) - \inf_{f\in\mathcal{F}} R_P(f)\right| \geq \frac{\epsilon}{6}\right) \leq \frac{\epsilon}{3}$$

and

$$Q^{\otimes N}\left(\tilde{\vec{z}}^N : \left|\inf_{f\in\mathcal{F}} R_{Q_N}(f) - \inf_{f\in\mathcal{F}} R_Q(f)\right| \geq \frac{\epsilon}{6}\right) \leq \frac{\epsilon}{3},$$

or equivalently

$$P^{\otimes N}\left(\vec{z}^N : \left|\hat{\vartheta}_N(\vec{z}^N) - \vartheta(P)\right| \geq \frac{\epsilon}{6}\right) \leq \frac{\epsilon}{3} \tag{37}$$

and

$$Q^{\otimes N}\left(\tilde{\vec{z}}^N : \left|\hat{\vartheta}_N(\tilde{\vec{z}}^N) - \vartheta(Q)\right| \geq \frac{\epsilon}{6}\right) \leq \frac{\epsilon}{3} \tag{38}$$

for all $Q \in \mathcal{M}_{Z,\kappa}^{\phi^p}$ and $\mathsf{dl}_\phi(P, Q)$ sufficiently small. By the continuity of ϑ, there exists a constant $\delta > 0$ such that when $\mathsf{dl}_\phi(P', P) < 2\delta$, $|\vartheta(P') - \vartheta(P)| < \frac{\epsilon}{12}$. On the other hand, it follows by Corollary 3.5 in Krätschmer, Schied and Zähle (2012) that $(\mathcal{M}_{Z,\kappa}^{\phi^p}, \mathsf{dl}_\phi)$ has the UGC property which implies that

$$Q^{\otimes N}\left(\mathsf{dl}_\phi(Q_N, Q) \geq \delta\right) \leq \frac{\epsilon}{3} \tag{39}$$

for all $Q \in \mathcal{M}_{Z,\kappa}^{\phi^p}$ including $Q = P$. This shows (37) when N_0 is chosen sufficiently large. Next, we show (38) which is more challenging than (37) because it requires the inequality to hold uniformly for all Q close to P. Let $\mathsf{dl}_\phi(Q, P) \leq \delta$. By (39)

$$\begin{aligned}
\frac{\epsilon}{3} &\geq Q^{\otimes N}\left(\tilde{\vec{z}}^N : \mathsf{dl}_\phi(Q_N, Q) \geq \delta\right) \qquad (40)\\
&\geq Q^{\otimes N}\left(\tilde{\vec{z}}^N : \mathsf{dl}_\phi(Q_N, P) \geq \delta + \mathsf{dl}_\phi(Q, P)\right)\\
&\geq Q^{\otimes N}\left(\tilde{\vec{z}}^N : \mathsf{dl}_\phi(Q_N, P) \geq 2\delta\right)\\
&\geq Q^{\otimes N}\left(\tilde{\vec{z}}^N : |\vartheta(Q_N) - \vartheta(P)| \geq \frac{\epsilon}{12}\right)\\
&\geq Q^{\otimes N}\left(\tilde{\vec{z}}^N : |\vartheta(Q_N) - \vartheta(Q)| \geq |\vartheta(P) - \vartheta(Q)| + \frac{\epsilon}{12}\right)\\
&\geq Q^{\otimes N}\left(\tilde{\vec{z}}^N : |\vartheta(Q_N) - \vartheta(Q)| \geq \frac{\epsilon}{6}\right)\\
&= Q^{\otimes N}\left(\tilde{\vec{z}}^N : \left|\hat{\vartheta}_N(\tilde{\vec{z}}^N) - \vartheta(Q)\right| \geq \frac{\epsilon}{6}\right), \forall Q \in \mathcal{M}_{Z,\kappa}^{\phi^p}.
\end{aligned}$$

The conclusion follows. ■

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

First, the conclusions of Theorem 3 hold for the case that $\lambda_N \equiv 0$ for all N, which means that the empirical risk obtained from solving the unregularized problem (2) is statistically robust under the same conditions.

Second, the set $\mathcal{M}_{Z,\kappa}^{\phi^p}$ differs from $\mathcal{M}_Z^{\phi^p}$ in that the former imposes a bound for the moment value uniformly for all $P \in \mathcal{M}_{Z,\kappa}^{\phi^p}$ whereas the latter does not have such uniformity. This is because we need the UGC property of $(\mathcal{M}_{Z,\kappa}^{\phi^p}, \mathsf{dl}_\phi)$ in order for us to apply Corollary 3.5 in Krätschmer, Schied and Zähle (2012). For example, in the least squares regression model with polynomial kernel, we have

$$\mathcal{M}_{Z,\kappa}^{\phi^p} = \left\{ P \in \mathscr{P}(Z) : \int_Z \left[\|y\|^2 + \beta^2(\gamma\|x\|^2 + 1)^d \right]^p P(dz) < \kappa \right\}.$$

In the case of Gaussian kernel or Laplacian kernel,

$$\mathcal{M}_{Z,\kappa}^{\phi^p} = \left\{ P \in \mathscr{P}(Z) : \int_Z \|y\|^{2p} P(dz) < \kappa \right\}.$$

Third, by (38), we can obtain for any $\epsilon > 0$, there exist constants $\delta > 0$ and $N_0 \in \mathbb{N}$ such that

$$Q \in \mathcal{M}, \mathsf{dl}_\phi(P, Q) \le \delta \Longrightarrow Q^{\otimes N} \left(\tilde{z}^N : |\vartheta(Q) - \vartheta(Q_N)| \ge \frac{\epsilon}{6} \right) \le \frac{\epsilon}{3}$$

for $N \ge N_0$. 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 ϕ-weak topological space of probability measures.

Fourth, in practice, since P is unknown, it is difficult to identify δ for a specified ϵ. The usefulness of (32) should be understood as that it provides a theoretical guarantee: if the training data is generated by some probability distribution Q which is close to the true distribution P [2], and Q satisfies moment condition (31) (which may be examined by through empirical data, i.e., $\int_Z \phi(z)^p Q_N(dz) \le \kappa$), then the optimal value obtained with the perceived data is close to the one with pure data. An effective way to address the difficulty is to derive quantitative statistical robustness under some additional conditions in which case the relationship between ϵ and δ may be explicitly established, we will come back to this in the next subsection.

Fifth, Theorem 3 does not tell us how N_0 depends on ϵ and δ. The dependence may be derived under some special circumstances when $\phi(\cdot) = \|\cdot\|^q$ for some positive number q and Z is a finite dimensional space. In the rest of this subsection, we discuss this.

Observe first that from Remark 2, $\phi(\cdot)$ can be chosen as $\|\cdot\|^q$ for some kernels. For this particular form of ϕ, we show next how the threshold value N_0 in Theorem 3 behaves as a function of ϵ, δ and the dimension of z. To this end, we need the following two intermediate results.

2. In practice, we may draw M groups of samples with size N and use them to construct histograms, if all of the histograms are close to each other, then we may predict that P_N is in the vicinity of Q_N and hence P is in the vicinity of Q.

Lemma 1 *(Claus, 2016, Proposition 2.63) For any $q \geq 1$, let $\mathcal{M}^q_{\mathbb{R}^n} := \{P \in \mathscr{P}(\mathbb{R}^n) : \int_{\mathbb{R}^n} \|z\|^q P(dz) < \infty\}$. The $\|\cdot\|^q$-weak topology on $\mathcal{M}^q_{\mathbb{R}^n}$ is generated by the Wasserstein metric $\mathsf{dl}_{W,q} : \mathcal{M}^q_{\mathbb{R}^n} \times \mathcal{M}^q_{\mathbb{R}^n} \to \mathbb{R}$ of order q:*

$$\mathsf{dl}_{W,q}(P,Q) := \inf_{\pi} \left\{ \left(\int_{\mathbb{R}^n \times \mathbb{R}^n} \|\xi - \tilde{\xi}\|^q \pi(\mathrm{d}\xi, \mathrm{d}\tilde{\xi}) \right)^{\frac{1}{q}} \right\},$$

where π is among all probability measures over $\mathbb{R}^n \times \mathbb{R}^n$ with marginals P and Q.

Lemma 2 *Let $P \in \mathcal{M}^q_{\mathbb{R}^n}$ with $1 \leq q < n/2$. Assume that there exist $\alpha > 2$ and $\kappa > 0$ such that $\int_{\mathbb{R}^n} \|\xi\|^{\alpha q} P(d\xi) \leq \kappa$. Then for all $N \geq 1$ and $\delta > 0$*

$$P^N(\mathsf{dl}_{W,q}(P, P_N) \geq \delta) \leq a(N,\delta)\mathbf{1}_{(-\infty,1]}(\delta) + b(N,\delta), \tag{41}$$

where $a(N,\delta) := c_1 \exp(-c_2 N \delta^n)$ and $b(N,\delta) := c_1 N (N\delta^q)^{-(\alpha q - \eta)/q}$ for any $\eta \in (0, \alpha q)$. The positive constants c_1 and c_2 depend on q, n, α, κ and η.

Proof. The result follows directly from Theorem 2 in Fournier and Guilline (2015). ■

Note that by setting $\eta := \frac{\alpha q}{2}$ and the right hand side of (41) to ϵ, we can obtain

$$P^N(\mathsf{dl}_{W,q}(P, P_N) \geq \delta) \leq \epsilon, \tag{42}$$

for all $N \geq N_0$, where

$$N_0 := \max \left\{ \frac{\ln(2c_1/\epsilon)}{c_2 \delta^n}, \left(\frac{\epsilon}{2c_1} \delta^{\alpha q/2} \right)^{1-\alpha/2} \right\}.$$

The dependence of the constants c_1 and c_2 on q, n α and κ implies that N_0 depends on δ, ϵ, q, n, α and κ. This shows inequality (42) holds for all P satisfying $\int_{\mathbb{R}^n} \|\xi\|^{\alpha q} P(d\xi) \leq \kappa$ and hence the space $(\mathcal{M}^{\alpha q}_{\mathbb{R}^n,\kappa}, \mathsf{dl}_{W,q})$ has the UGC property, the next proposition addresses this.

Proposition 1 *Let $\mathcal{M}^{\alpha q}_{\mathbb{R}^n,\kappa} := \{P \in \mathscr{P}(\mathbb{R}^n) : \int_{\mathbb{R}^n} \|\xi\|^{\alpha q} P(d\xi) \leq \kappa\}$ with $1 \leq q < n/2$, $\kappa > 0$ and $\alpha > 2$. Then the space $(\mathcal{M}^{\alpha q}_{\mathbb{R}^n,\kappa}, \mathsf{dl}_{W,q})$ has the UGC property, that is, for every $\epsilon > 0$ and $\delta > 0$,*

$$P^N(\mathsf{dl}_{W,q}(P, P_N) \geq \delta) \leq \epsilon, \forall P \in \mathcal{M}^{\alpha q}_{\mathbb{R}^n,\kappa} \tag{43}$$

for $N \geq N_0 := \max \left\{ \frac{\ln(2c_1/\epsilon)}{c_2\delta^n}, \left(\frac{\epsilon}{2c_1} \delta^{\alpha q/2} \right)^{1-\alpha/2} \right\}$, where the positive constants c_1 and c_2 depend on q, n α and κ.

Theorem 4 (Statistical robustness for $\phi(\cdot) := \|\cdot\|^q$) *Consider the case that $Z = \mathbb{R}^n$. Let $\mathcal{M}^{\alpha q}_{\mathbb{R}^n,\kappa}$ be defined as in Proposition 1 with $\kappa > 0$ and $\alpha > 2$, and $\lambda_N \equiv 0$. Assume: (a) the conditions in Theorem 2 are satisfied, (b) the function ϕ in Assumption 2 (a) take the*

form of $\|\cdot\|^q$ *with* $1 \le q < n/2$. *Then for any* $\epsilon > 0$ *and* $P \in \mathcal{M}^{\alpha q}_{\mathbb{R}^n,\kappa}$, *there exists positive number* $\delta > 0$ *such that for all* $Q \in \mathcal{M}^{\alpha q}_{\mathbb{R}^k,\kappa}$ *with* $\mathsf{dl}_{W,q}(P,Q) \le \delta$,

$$\mathsf{dl}_{\mathrm{Prok}}\left(P^{\otimes N} \circ \hat{\vartheta}_N(\cdot,\lambda_N)^{-1}, Q^{\otimes N} \circ \hat{\vartheta}_N(\cdot,\lambda_N)^{-1}\right) \le \epsilon \tag{44}$$

for all $N \ge N_0 := \max\left\{\frac{\ln(6c_1/\epsilon)}{c_2\delta^n}, \left(\frac{\epsilon}{6c_1}\delta^{\alpha q/2}\right)^{1-\alpha/2}\right\}$, *where the positive constants* c_1 *and* c_2 *depend on* q, n α *and* κ.

Proof. By the proof of Theorem 3, it suffices to find δ and N_0 such that

$$|\vartheta(P) - \vartheta(Q)| \le \frac{\epsilon}{3}, \tag{45}$$

$$P^{\otimes N}\left(\vec{z}^N : \left|\hat{\vartheta}_N(\vec{z}^N) - \vartheta(P)\right| \ge \frac{\epsilon}{3}\right) \le \frac{\epsilon}{3}, \tag{46}$$

$$Q^{\otimes N}\left(\tilde{\vec{z}}^N : \left|\hat{\vartheta}_N(\tilde{\vec{z}}^N) - \vartheta(Q)\right| \ge \frac{\epsilon}{3}\right) \le \frac{\epsilon}{3}. \tag{47}$$

By Theorem 2 and Lemma 1, there exists a constant $\delta > 0$ such that $|\vartheta(P') - \vartheta(P)| < \frac{\epsilon}{6}$ when $\mathsf{dl}_{W,q}(P',P) < 2\delta$. On the other hand, it follows by Proposition 1 that $(\mathcal{M}^{\alpha q}_{\mathbb{R}^n,\kappa}, \mathsf{dl}_{W,q})$ has the UGC property, which implies that

$$Q^{\otimes N}\left(\vec{z}^N : \mathsf{dl}_{W,q}(Q_N, Q) \ge \delta\right) \le \frac{\epsilon}{3} \tag{48}$$

for all $Q \in \mathcal{M}^{\alpha q}_{\mathbb{R}^n,\kappa}$ including $Q = P$ when $N \ge N_0 := \max\left\{\frac{\ln(6c_1/\epsilon)}{c_2\delta^n}, \left(\frac{\epsilon}{6c_1}\delta^{\alpha q/2}\right)^{1-\alpha/2}\right\}$. Then

$$P^{\otimes N}\left(\vec{z}^N : \left|\hat{\vartheta}_N(\vec{z}^N) - \vartheta(P)\right| \ge \frac{\epsilon}{3}\right) \le P^{\otimes N}\left(\mathsf{dl}_{W,q}(P_N, P) \ge \delta\right) \le \frac{\epsilon}{3},$$

which means (46) holds. Next, we show (47) is guaranteed by (48). Let $Q \in \mathcal{M}^{\alpha q}_{\mathbb{R}^n,\kappa}$ with $\mathsf{dl}_{W,q}(Q,P) \le \delta$, Following a similar argument to that of (40), we can obtain (with $\mathsf{dl}_\phi = \mathsf{dl}_{W,q}$)

$$\frac{\epsilon}{3} \;\ge\; Q^{\otimes N}\left(\tilde{\vec{z}}^N : \mathsf{dl}_{W,q}(Q_N, Q) \ge \delta\right) \ge Q^{\otimes N}\left(\tilde{\vec{z}}^N : \left|\hat{\vartheta}_N(\tilde{\vec{z}}^N) - \vartheta(Q)\right| \ge \frac{\epsilon}{3}\right).$$

Since the inequality holds uniformly for all $Q \in \mathcal{M}^{\alpha q}_{\mathbb{R}^n,\kappa}$ with $\mathsf{dl}_{W,q}(Q,P) \le \delta$, the conclusion follows. ■

Note that in Theorem 4, N_0 depends on the dimension of z and hence suffers from curse of dimensionality when the dimension is large. To address the issue, we may strengthen the condition on $c(z, f(x))$ by requiring it to be locally Lipschitz continuous with specified growth of the Lipschitz modulus as z goes to infinity. We will address the issue in the next subsection.

4.2 Quantitative robustness

In the previous section, there is no explicit relationship between ϵ and δ in the qualitative robustness result. In this section, we address the issue under the following additional condition.

Assumption 3 *The cost function $c(z, f(x))$ satisfies the following property:*

$$|c(z, f(x)) - c(z', f(x'))| \leq L_p(z, z')\|z - z'\|, \forall z, z' \in Z, f \in \mathcal{F}, \tag{49}$$

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

To see how the assumption may be satisfied, we consider the case that $c : Z \times Y \to \mathbb{R}$ satisfies

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

When $\|f\|_k \leq \beta$ for some positive number β (see (20)), the calmness condition in Remark 1 implies

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

Consequently, we have

$$|c(z, f(x)) - c(z', f(x'))| \leq \max\{L(z), L(z')\}(\|z - z'\| + \beta g(\|x - x'\|)) \tag{50}$$

for all $z, z' \in Z$ and $f \in \mathcal{F}$. In Example 2, we will explain in detail how $L(\cdot)$ may be figured out and in a combination with specific form of function $g(\cdot)$, inequality (50) leads to inequality (49) for some specific cost functions and kernels in regression models.

We now return to our discussion on the quantitative description of the discrepancy between $P^{\otimes N} \circ \hat{\vartheta}_N(\cdot, \lambda_N)^{-1}$ and $Q^{\otimes N} \circ \hat{\vartheta}_N(\cdot, \lambda_N)^{-1}$. Our idea is to use Kantorovich metric to measure the difference, i.e.,

$$\mathsf{dl}_{K,1}\left(P^{\otimes N} \circ \hat{\vartheta}_N(\cdot, \lambda_N)^{-1}, Q^{\otimes N} \circ \hat{\vartheta}_N(\cdot, \lambda_N)^{-1}\right),$$

which is bounded by the difference of P and Q under ζ_p metric. The next technical result prepares for such a conversion.

Lemma 1 *For $\vec{z} := (z^1, \cdots, z^N) \in Z^{\otimes N}$ and $L_p(\cdot, \cdot)$ being defined in Assumption 3, let*

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

and

$$\mathsf{dl}_\Psi(P^{\otimes N}, Q^{\otimes N}) := \sup_{\psi\in\Psi}\left|\int_{Z^{\otimes N}} \psi(\vec{z})P^{\otimes N}(d\vec{z}) - \int_{Z^{\otimes N}} \psi(\vec{z})Q^{\otimes N}(d\vec{z})\right|.$$

Then

$$\mathsf{dl}_\Psi(P^{\otimes N}, Q^{\otimes N}) \leq \zeta_p(P, Q) < +\infty, \forall P, Q \in \mathcal{P}_p(Z),$$

where $\zeta_p(P, Q)$ is defined in Definition 4 and

$$\mathcal{P}_p(Z) := \left\{P \in \mathscr{P}(Z) : \int_Z \|z\|^p P(dz) < +\infty\right\}. \tag{51}$$

Proof. The result is established in Lemma 4.1 in Wang et al. (2020) which is an extension of Lemma 1 in Guo and Xu (2020) (which is presented when $p = 1$). Here we include a proof for self-containedness. Let $\vec{z}^j := \{z^1, \cdots, z^j\}$ and $\vec{z}^{-j} := \{z^1, \cdots, z^{j-1}, z^{j+1}, \cdots, z^N\}$ with $z^1, \cdots, z^N \in Z$. For any $P^1, \cdots, P^N \in \mathscr{P}(Z)$ and any $j \in \{1, \cdots, N\}$, denote

$$P^{-j}(d\vec{z}^{-j}) := P^1(dz^1) \cdots P^{j-1}(dz^{j-1}) P^{j+1}(dz^{j+1}) \cdots P^N(dz^N)$$

and $h_{\vec{z}^{-j}}(z^j) := \int_{Z^{\otimes(N-1)}} \psi(\vec{z}^{-j}, z^j) P^{-j}(d\vec{z}^{-j})$. Then

$$\begin{aligned} |h_{\vec{z}^{-j}}(\tilde{z}^j) - h_{\vec{z}^{-j}}(\hat{z}^j)| &\leq \int_{Z^{\otimes(N-1)}} \left|\psi(\vec{z}^{-j}, \tilde{z}^j) - \psi(\vec{z}^{-j}, \hat{z}^j)\right| P^{-j}(d\vec{z}^{-j}) \\ &\leq \int_{Z^{\otimes(N-1)}} \frac{1}{N} L_p(\tilde{z}^j, \hat{z}^j) \|\tilde{z}^j - \hat{z}^j\| P^{-j}(d\vec{z}^{-j}) \\ &\leq \frac{1}{N} L_p(\tilde{z}^j, \hat{z}^j) \|\tilde{z}^j - \hat{z}^j\|. \end{aligned}$$

Let $\mathcal{W}$ denote the set of functions $h_{\vec{z}^{-j}}(z^j)$ generated by $\psi \in \Psi$. By the definition of dl_Ψ and the p-th order Fortet-Mourier metric,

$$\begin{aligned} \mathsf{dl}_\Psi(P^{-j} \times \tilde{P}^j, P^{-j} \times \hat{P}^j) &= \sup_{\psi \in \Psi} \left| \int_Z \int_{Z^{\otimes(N-1)}} \psi(\vec{z}^{-j}, z^j) P^{-j}(d\vec{z}^{-j}) \tilde{P}^j(dz^j) \right. \\ &\quad \left. - \int_Z \int_{Z^{\otimes(N-1)}} \psi(\vec{z}^{-j}, z^j) P^{-j}(d\vec{z}^{-j}) \hat{P}^j(dz^j) \right| \\ &= \sup_{h_{\vec{z}^{-j}} \in \mathcal{W}} \left| \int_Z h_{\vec{z}^{-j}}(z^j) \tilde{P}^j(dz^j) - \int_Z h_{\vec{z}^{-j}}(z^j) \hat{P}^j(dz^j) \right| \\ &\leq \frac{1}{N} \zeta_p(\tilde{P}^j, \hat{P}^j), \end{aligned} \tag{52}$$

where the inequality is due to $N h_{\vec{z}^{-j}}(z^j) \in \mathcal{F}_p(Z)$ and the definition of $\zeta_p(P, Q)$. Finally, by the triangle inequality of the pseudo-metric, we have

$$\begin{aligned} \mathsf{dl}_\Psi\left(P^{\otimes N}, Q^{\otimes N}\right) &\leq \mathsf{dl}_\Psi\left(P^{\otimes N}, P^{\otimes(N-1)} \times Q\right) + \mathsf{dl}_\Psi\left(P^{\otimes(N-1)} \times Q, P^{\otimes(N-2)} \times Q^{\otimes 2}\right) \\ &\quad + \cdots + \mathsf{dl}_\Psi\left(P \times Q^{\otimes(N-1)}, Q^{\otimes N}\right) \\ &\leq \frac{1}{N} \zeta_p(P, Q) \times N = \zeta_p(P, Q). \end{aligned}$$

The proof is complete. ∎

With Lemma 1, we are ready to state our main result.

Theorem 5 (Quantitative statistical robustness) *Let $\mathcal{P}_p(Z)$ be defined as in (51). Under Assumption 3,*

$$\mathsf{dl}_{K,1}\left(P^{\otimes N} \circ \hat{\vartheta}_N(\cdot, \lambda_N)^{-1}, Q^{\otimes N} \circ \hat{\vartheta}_N(\cdot, \lambda_N)^{-1}\right) \leq \zeta_p(P, Q) \tag{53}$$

for any $N \in \mathbb{N}$ and any $P, Q \in \mathcal{P}_p(Z)$, where p is defined as in Assumption 3. In the case when $p = 1$,

$$\mathsf{dl}_{K,1}\left(P^{\otimes N} \circ \hat{\vartheta}_N(\cdot, \lambda_N)^{-1}, Q^{\otimes N} \circ \hat{\vartheta}_N(\cdot, \lambda_N)^{-1}\right) \leq \mathsf{dl}_{K,Z}(P, Q) \tag{54}$$

and

$$\mathsf{dl}_{\mathrm{Prok}}\left(P^{\otimes N}\circ\hat{\vartheta}_N(\cdot,\lambda_N)^{-1},Q^{\otimes N}\circ\hat{\vartheta}_N(\cdot,\lambda_N)^{-1}\right)\leq\sqrt{\mathsf{dl}_{K,Z}(P,Q)}. \tag{55}$$

Proof. Inequality (54) follows from inequality (53) whereas inequality (55) follows from inequality (54) and Corollary 2.18 in Huber and Ronchetti (2009). Thus it suffices to show (53). By definition

$$\begin{aligned}
&\mathsf{dl}_{K,1}\left(P^{\otimes N}\circ\hat{\vartheta}_N(\cdot,\lambda_N)^{-1},Q^{\otimes N}\circ\hat{\vartheta}_N(\cdot,\lambda_N)^{-1}\right) \qquad (56)\\
=\;&\sup_{g\in\mathscr{G}}\left|\int_{\mathbb{R}}g(t)P^{\otimes N}\circ\hat{\vartheta}_N(\cdot,\lambda_N)^{-1}(dt)-\int_{\mathbb{R}}g(t)Q^{\otimes N}\circ\hat{\vartheta}_N(\cdot,\lambda_N)^{-1}(dt)\right|\\
=\;&\sup_{g\in\mathscr{G}}\left|\int_{Z^{\otimes N}}g(\hat{\vartheta}_N(\vec{z}^N,\lambda_N))P^{\otimes N}(d\vec{z}^N)-\int_{Z^{\otimes N}}g(\hat{\vartheta}_N(\vec{z}^N,\lambda_N))Q^{\otimes N}(d\vec{z}^N)\right|,
\end{aligned}$$

where $\mathscr{G}$ denotes the set of all Lipschitz continuous functions with modulus bounded by 1 and we write $\vec{z}^N$ for $(z^1,\cdots,z^N)$ and $\hat{\vartheta}_N(\vec{z}^N,\lambda_N)$ for $\hat{\vartheta}_N$ to indicate its dependence on $z^1,\cdots,z^N$. To see the well-definiteness of the pseudo-metric, we note that for each $g\in\mathscr{G}$,

$$|g(\hat{\vartheta}_N(\vec{z}^N,\lambda_N))|\leq|g(\hat{\vartheta}_N(\vec{z}_0^N,\lambda_N))|+|\hat{\vartheta}_N(\vec{z}^N,\lambda_N)-\hat{\vartheta}_N(\vec{z}_0^N,\lambda_N)|, \tag{57}$$

where $\vec{z}_0^N\in Z^{\otimes N}$ is fixed. By the definition of $\hat{\vartheta}_N(\vec{z}^N,\lambda_N)$, we have

$$\begin{aligned}
&|\hat{\vartheta}_N(\vec{z}^N,\lambda_N)-\hat{\vartheta}_N(\vec{z}_0^N,\lambda_N)|\\
=\;&\left|\min_{f\in\mathcal{F}}\;\frac{1}{N}\sum_{j=1}^N\left(c(z^j,f(x^j))+\lambda_N\|f\|_k^2\right)-\min_{f\in\mathcal{F}}\;\frac{1}{N}\sum_{j=1}^N\left(c(z_0^j,f(x_0^j))+\lambda_N\|f\|_k^2\right)\right|\\
\leq\;&\frac{1}{N}\sum_{j=1}^N\sup_{f\in\mathcal{F}}|c(z^j,f(x^j))-c(z_0^j,f(x_0^j))|\\
\leq\;&\frac{1}{N}\sum_{j=1}^N L_p(z^j,z_0^j)\|z^j-z_0^j\|. \qquad (58)
\end{aligned}$$

Combining (57) and (58), we deduce that

$$\int_{Z^{\otimes N}}g(\hat{\vartheta}_N(\vec{z}^N,\lambda_N))P^{\otimes N}(d\vec{z}^N)<\infty,\forall P\in\mathcal{P}_p(Z).$$

The same argument can be made on $\int_{Z^{\otimes N}}g(\hat{\vartheta}_N(\vec{z}^N,\lambda_N))Q^{\otimes N}(d\vec{z}^N)$ for $Q\in\mathcal{P}_p(Z)$.

Next, we show (53). We do so by applying Lemma 1 to the right hand side of (56). To this end, we need to verify the condition of the lemma. Define $\psi:Z^{\otimes N}\to\mathbb{R}$ by $\psi(\vec{z}^N):=g(\hat{\vartheta}_N(\vec{z}^N,\lambda_N))$. Since g is Lipschitz continuous with modulus bounded by 1, by (58) we have

$$\begin{aligned}
\left|\psi(\tilde{\vec{z}}^N)-\psi(\hat{\vec{z}}^N)\right|&=|g(\hat{\vartheta}_N(\tilde{\vec{z}}^N,\lambda_N))-g(\hat{\vartheta}_N(\hat{\vec{z}}^N,\lambda_N))|\\
&\leq|\hat{\vartheta}_N(\tilde{\vec{z}}^N,\lambda_N)-\hat{\vartheta}_N(\hat{\vec{z}}^N,\lambda_N)|\\
&\leq\frac{1}{N}\sum_{j=1}^N L_p(\tilde{z}^j,\hat{z}^j)\|\tilde{z}^j-\hat{z}^j\|,
\end{aligned}$$

which means that ψ is in the set Ψ in Lemma 1. The rest follows from application of the lemma to (56). ■

Theorem 5 strengthens the qualitative statistical robustness results in several aspects.

1. It gives rise to an explicit quantitative relationship between $\mathsf{dl}_{K,1}(P^{\otimes N}\circ\hat{\vartheta}_N(\cdot,\lambda_N)^{-1}, Q^{\otimes N}\circ\hat{\vartheta}_N(\cdot,\lambda_N)^{-1})$ and $\zeta_p(P,Q)$. 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 Fortet-Mourier metric for quantification of the latter. As noted immediately after Definition 4, $\zeta_p(P,Q)$ may be estimated via sample data, which means the error bound established in (53) is practically obtainable and this is a significant step forward from the qualitative robustness result.

2. 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. It also raises the prospect of application of the quantitative statistical robustness results to problems with non-Hilbertian regularizations (Unser (2019)).

3. Inequalities (53) and (54) hold for all $N \in \mathbb{N}$ which means the quantitative statistical results are independent of the sample size. This effectively addresses curse of dimensionality suffered by the qualitative statistical robustness results as indicated in Theorem 4.

4. As we can see from the proof of Theorem 5, the quantitative statistical robustness results are established without relying on the stability result established in Theorem 2. Of course, under the strengthened conditions as stated in Assumption 3, we can obtain by the definition of the Fortet-Mourier metric that

$$|\vartheta(Q) - \vartheta(P)| \leq \zeta_p(Q,P).$$

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

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

$$\begin{aligned} |c(z,f(x)) - c(z',f(x'))| &= \frac{1}{2}\left||y-f(x)|^2 - |y'-f(x')|^2\right| \\ &\leq \frac{1}{2}\left(|y|+|f(x)|+|y'|+|f(x')|\right)\left(|y-y'|+|f(x)-f(x')|\right). \end{aligned}$$

In the case when there exists a positive constant β *such that* $\|f\|_k \leq \beta$ *and the calmness condition in Remark 1 holds,*

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

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,*

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

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

$$|c(z, f(x)) - c(z', f(x'))| \le (1+\beta)^2 \max\{1, \|z\|, \|z'\|\}\,\|z - z'\|.$$

  By Theorem 5, $\mathsf{dl}_{K,1}\left(P^{\otimes N} \circ \hat{\vartheta}_N^{-1}, Q^{\otimes N} \circ \hat{\vartheta}_N^{-1}\right) \le (1+\beta)^2 \zeta_2(P,Q)$ *for all* $N \in \mathbb{N}$ *and any* $P, Q \in \mathcal{P}_2(\mathbb{R}^{n+1})$.

- *In the case of Gaussian kernel,* $\eta(z) = |y| \le \|z\|$, $g(t) = \max\{\sqrt{2\gamma}, 1\}t$, *and*

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

  By Theorem 5, $\mathsf{dl}_{K,1}\left(P^{\otimes N} \circ \hat{\vartheta}_N^{-1}, Q^{\otimes N} \circ \hat{\vartheta}_N^{-1}\right) \le \max\{\sqrt{2\gamma}, 1\}\zeta_2(P,Q)$ *for all* $N \in \mathbb{N}$ *and any* $P, Q \in \mathcal{P}_2(\mathbb{R}^{n+1})$.

- *In the case of polynomial kernel,*

$$\eta(z) = |y| + \beta\sqrt{(\gamma\|x\|^2 + 1)^d} \le (1 + \beta(\gamma+1)^{d/2}) \max\{1, \|z\|\}^d.$$

  For fixed z *and* z', *let* $R := \max\{1, \|z\|, \|z'\|\}$. *Then by Remark 1,*

$$\begin{aligned} & \|k(\cdot, x) - k(\cdot, x')\|_k \\ \le & \begin{cases} \max\{\frac{1}{2R}\sqrt{2(\gamma R^2 + 1)^d}, 1\}\|x - x'\|, & \text{if } d \text{ is even,} \\ \max\{\frac{1}{2R}\sqrt{2(\gamma R^2 + 1)^d - 2(1 - \gamma R^2)^d}, 1\}\|x - x'\|, & \text{if } d \text{ is odd,} \end{cases} \\ \le & \max\{(1+\gamma)^{d/2}, 1\} \max\{1, \|z\|, \|z'\|\}^{d-1}\|z - z'\|. \end{aligned}$$

  Let $A := (1 + \beta(\gamma+1)^{d/2}) \max\{2\beta(1+\gamma)^{d/2}, 2\beta, 2\}$. *Then*

$$|c(z, f(x)) - c(z', f(x'))| \le A \max\{1, \|z\|, \|z'\|\}^{2d-1}\,\|z - z'\|.$$

  By Theorem 5, $\mathsf{dl}_{K,1}\left(P^{\otimes N} \circ \hat{\vartheta}_N^{-1}, Q^{\otimes N} \circ \hat{\vartheta}_N^{-1}\right) \le A\zeta_{2d}(P,Q)$ *for all* $N \in \mathbb{N}$ *and any* $P, Q \in \mathcal{P}_{2d}(\mathbb{R}^{n+1})$.

We can derive similar results for the regression models with ϵ*-insensitive loss function, hinge loss, log-loss function, Huber's loss function and* $p-$*th power absolute loss function, we leave readers for exercises.*

Remark 3 *It might be interesting to study the discrepancy between* $f_N^{\lambda_N}(P_N)$ *and* $f_N^{\lambda_N}(Q_N)$. *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 (1) and the regularized problem (4) have a unique solution. Moreover, the strong convexity implies that problem (4) satisfies the second order growth condition at* $f_N^{\lambda_N}(P_N)$, *that is, there exists a positive constant* α *such that*

$$R_{P_N}^{\lambda_N}(f) - \vartheta(P_N, \lambda_N) \ge \alpha\|f - f_N^{\lambda_N}(P_N)\|_k^2, \forall f \in \mathcal{F}.$$

By virtue of Lemma 3.8 in Liu and Xu (2013), under Assumption 3, we can use the inequality to obtain

$$\begin{aligned}\|f_N^{\lambda_N}(P_N) - f_N^{\lambda_N}(Q_N)\|_k &\leq \sqrt{\frac{3}{\alpha}\sup_{f\in\mathcal{F}}|\mathbb{E}_{P_N}[c(z,f(x))] - \mathbb{E}_{Q_N}[c(z,f(x))]|} \\ &\leq \sqrt{\frac{3}{\alpha}\zeta_p(P_N,Q_N)}.\end{aligned}$$

If $\zeta_p(P_N,Q_N)\to 0$, then $\|f_N^{\lambda_N}(P_N) - f_N^{\lambda_N}(Q_N)\|_k \to 0$. However, we are unable to establish the kind of estimation in (53) 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 for fixed $P\in\mathscr{P}(Z)$ and then uniform convergence for all P over a subset $\mathcal{M}$ of $\mathscr{P}(Z)$. To this end, we introduce the following assumption on the cost function.

Assumption 4 *There exist a continuous function $r: Z\to\mathbb{R}_+$ and a constant $\nu\in(0,1]$ such that for any compact subset $\hat{Z}\subset Z$*

$$|c(z,f(x)) - c(z,g(x))| \leq r(z)\|f-g\|_{\hat{Z},\infty}^{\nu}, \forall f,g\in\mathcal{F}, z\in\hat{Z}, \tag{59}$$

where $\|f-g\|_{\hat{Z},\infty} := \sup_{z=(x,y)\in\hat{Z}}|f(x)-g(x)|$.

The assumption requires $c(z,\cdot)$ to be Hölder continuous over $\mathcal{F}$ uniformly for $z\in\hat{Z}$. It should be distinguished from Assumption 3 which requires $c(z,f(x))$ to be locally Lipschitz continuous in z for all $f\in\mathcal{F}$. In a particular case when there exists a positive constant such that $\|f\|_k\leq\beta$, this assumption is satisfied by all of the loss functions in regression models that we list at the beginning of Section 2.

Theorem 6 (Consistency of $\vartheta(P_N,\lambda_N)$) *Assume: (a) the conditions in Theorem 2 are satisfied, (b) Assumption 4 holds, (c) $M(t) := \mathbb{E}_P[e^{t\phi(z)}]$ is finite valued for all t in a neighborhood of zero and $\mathbb{E}_P[r(z)]<\infty$ for $P\in\mathcal{M}_Z^{\phi}$. 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*

$$P^{\otimes N}\left(\sup_{f\in\mathcal{F}}|\mathbb{E}_{P_N}[c(z,f(x))] + \lambda_N\|f\|_k^2 - \mathbb{E}_P[c(z,f(x))| \geq \delta\right) \leq \alpha(\epsilon,\delta)e^{-N\gamma(\epsilon,\delta)}, \tag{60}$$

when $N \geq N_0$ *and* $\lambda_N \leq \epsilon/\beta^2$ *for some positive constant* β *and hence*

$$P^{\otimes N}\left(|\vartheta(P_N, \lambda_N) - \vartheta(P)| \geq \delta\right) \leq \alpha(\epsilon, \delta) e^{-N\gamma(\epsilon,\delta)} \tag{61}$$

and

$$P^{\otimes N}\left(|\mathbb{E}_P[c(z, f_N^{\lambda_N}(x))] - \vartheta(P)| \geq 2\delta\right) \leq 2\alpha(\epsilon, \delta) e^{-N\gamma(\epsilon,\delta)}, \tag{62}$$

where $f_N^{\lambda_N} \in S^*_{P_N,\lambda_N}$.

In the literature of machine learning, consistency analysis refers to (62) whereas in stochastic programming, it refers to (61). The consistency analysis is mostly focused on the case when Z is a compact set. Norkin and Keyzer (2009) comment that compactness of Z or Y is commonly accepted in the statistical learning literature, where it allows us to apply exponential concentration measure inequalities for bounded random variables as developed by Bernstein, McDiarmid, and Hoeffding; see, for example, Cucker and Smale (2002a,b), Bousquet and Elisseeff (2002), Bartlett and Mendelson (2002), Schölkopf and Smola (2002), Poggio and Smale (2003), De Vito et al. (2005), Boucheron et al. (2005), Takeuchi et al. (2006), and Cucker and Zhou (2007).

Caponnetto and De Vito (2007) 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 underlying data generating distributions (priors) defined by a uniformly bounded kernel. In the case when X is a complete measurable space, $Y = \mathbb{R}$ and k is a bounded kernel, Steinwart and Christmann (2008) show that $\mathbb{E}_P[c(z, f_N^{\lambda_N}(x))]$ converges to $\vartheta(P)$ in distribution as $N \to \infty$, see Theorem 9.1 in Steinwart and Christmann (2008).

Note also that in machine learning, the constants $\alpha(\epsilon, \delta)$ and $\gamma(\epsilon, \delta)$ are often optimized. 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 (60) holds uniformly for all $f \in \mathcal{F}$. This kind of result may not hold in general, see a counter example in Shalev-Shwartz et al. (2010). Here we manage to establish the uniform convergence by showing equi-continuity of the class of functions in $\mathcal{F}$ and their uniform boundedness over a compact subset of Z under some moderate conditions.

Proof of Theorem 6. Observe that inequality (60) implies

$$\begin{aligned} & P^{\otimes N}\left(|\mathbb{E}_{P_N}[c(z, f_N^{\lambda_N}(x))] + \lambda_N \|f_N^{\lambda_N}\|_k^2 - \mathbb{E}_P[c(z, f_N^{\lambda_N}(x))| \geq \delta\right) \\ \leq \quad & \alpha(\epsilon, \delta) e^{-N\gamma(\epsilon,\delta)}, \end{aligned} \tag{63}$$

and a combination of (63) and (61) yields (62). Thus it suffices to prove (60) and (61). Since $P \in \mathcal{M}_Z^\phi$, then for any $\epsilon > 0$, there exists a constant $r > 0$ such that

$$\int_Z \phi(z) \mathbf{1}_{(r,\infty)}(\phi(z)) P(dz) \leq \epsilon.$$

Moreover, by Cramér's large deviation theory (Dembo and Zeitouni, 1998), there exist positive numbers C_0 and γ_0 such that

$$P^{\otimes N}\left(\int_Z \phi(z)\mathbf{1}_{(r,\infty)}(\phi(z))P_N(dz) \geq 2\epsilon\right) \leq C_0 e^{-\gamma_0 N}.$$

Under the coercive condition on ϕ in Assumption 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

$$\sup_{f\in\mathcal{F}} \int_{Z\setminus Z_\epsilon} |c(z, f(x))|P(dz) \leq \int_{Z\setminus Z_\epsilon} \phi(z)P(dz) \leq \int_{\{z\in Z:\phi(z)>r\}} \phi(z)P(dz) \leq \epsilon \tag{64}$$

and

$$\begin{aligned} & P^{\otimes N}\left(\sup_{f\in\mathcal{F}} \int_{Z\setminus Z_\epsilon} |c(z, f(x))|P_N(dz) \geq 2\epsilon\right) \leq P^{\otimes N}\left(\int_{Z\setminus Z_\epsilon} \phi(z)P_N(dz) \geq 2\epsilon\right) \\ \leq \quad & P^{\otimes N}\left(\int_{\{z\in Z:\phi(z)>r\}} \phi(z)P_N(dz) \geq 2\epsilon\right) \leq C_0 e^{-\gamma_0 N}. \end{aligned} \tag{65}$$

Moreover, under Assumption 1, there exists $\eta > 0$ such that for any $x, x' \in X_\epsilon$ satisfying $\|x - x'\| < \eta$,

$$\begin{aligned} |f(x') - f(x)| &= |\langle f, k(\cdot, x')\rangle - \langle f, k(\cdot, x)\rangle| \leq \|f\|_k \|k(\cdot, x') - k(\cdot, x)\|_k \\ &\leq \beta\|k(\cdot, x') - k(\cdot, x)\|_k \leq \beta\epsilon, \end{aligned}$$

which implies that $\mathcal{F}$ is equi-continuous when it is restricted to X_ϵ.

Let $\Delta_\epsilon := \sup_{x\in X_\epsilon} \|k(\cdot, x)\|_k$. Then for any $f \in \mathcal{F}$,

$$\sup_{x\in X_\epsilon} |f(x)| = \sup_{x\in X_\epsilon} |\langle f, k(\cdot, x)| \leq \|f\|_k \sup_{x\in X_\epsilon} \|k(\cdot, x)\|_k \leq \beta\Delta_\epsilon,$$

which implies that $\mathcal{F}$ is uniformly bounded when it is restricted to X_ϵ. Let $\bar{r} := \max\{|r(z)| : z \in Z_\epsilon\}$ where $r(z)$ is defined in Assumption 4, and $\bar{\epsilon} := (\epsilon/\bar{r})^{1/\nu}$. By Ascoli-Arzela Theorem (Brown, 2004), there exists an $\bar{\epsilon}$-net of $\mathcal{F}_K := \{f_1, \ldots, f_K\} \subset \mathcal{F}$ such that $\mathcal{F} = \bigcup_{k=1}^{K} \mathcal{F}_k^{\bar{\epsilon}}$, where

$\mathcal{F}_k^{\bar{\epsilon}} := \{f \in \mathcal{F} : \sup_{x\in X_\epsilon} |f(x) - f_k(x)| \le \bar{\epsilon}\}$ for $k = 1, \ldots, K$. Therefore,

$$
\begin{aligned}
& |\vartheta(P_N, \lambda_N) - \vartheta(P)| \\
= & \left| \sup_{f\in\mathcal{F}} \{\mathbb{E}_{P_N}[c(z, f(x))] + \lambda_N \|f\|_k^2\} - \sup_{f\in\mathcal{F}} \mathbb{E}_P[c(z, f(x))] \right| \\
\le & \left| \sup_{f\in\mathcal{F}} \mathbb{E}_{P_N}[c(z, f(x))\mathbf{1}_{Z_\epsilon}(z)] - \sup_{f\in\mathcal{F}} \mathbb{E}_P[c(z, f(x))\mathbf{1}_{Z_\epsilon}(z)] \right| + \epsilon \\
& + \sup_{f\in\mathcal{F}} \int_{Z\setminus Z_\epsilon} |c(z, f(x))| P_N(dz) + \sup_{f\in\mathcal{F}} \int_{Z\setminus Z_\epsilon} |c(z, f(x))| P(dz) \\
\le & \left| \sup_{k\in K} \sup_{f\in\mathcal{F}_k^{\bar{\epsilon}}} \mathbb{E}_{P_N}[c(z, f(x))\mathbf{1}_{Z_\epsilon}(z)] - \sup_{k\in K} \sup_{f\in\mathcal{F}_k^{\bar{\epsilon}}} \mathbb{E}_P[c(z, f(x))\mathbf{1}_{Z_\epsilon}(z)] \right| + 2\epsilon \\
& + \sup_{f\in\mathcal{F}} \int_{Z\setminus Z_\epsilon} |c(z, f(x))| P_N(dz) \\
\le & \sup_{k\in\{1,\ldots,K\}} \sup_{f\in\mathcal{F}_k^{\bar{\epsilon}}} |\mathbb{E}_{P_N}[c(z, f(x))\mathbf{1}_{Z_\epsilon}(z) - c(z, f_k(x))\mathbf{1}_{Z_\epsilon}(z) + c(z, f_k(x))\mathbf{1}_{Z_\epsilon}(z)] \\
& - \mathbb{E}_P[c(z, f(x))\mathbf{1}_{Z_\epsilon}(z) - c(z, f_k(x))\mathbf{1}_{Z_\epsilon}(z) + c(z, f_k(x))\mathbf{1}_{Z_\epsilon}(z)]| + 2\epsilon \\
& + \sup_{f\in\mathcal{F}} \int_{Z\setminus Z_\epsilon} |c(z, f(x))| P_N(dz) \\
\le & \sup_{k\in\{1,\ldots,K\}} |\mathbb{E}_{P_N}[c(z, f_k(x))\mathbf{1}_{Z_\epsilon}(z)] - \mathbb{E}_P[c(z, f_k(x))\mathbf{1}_{Z_\epsilon}(z)]| + 4\epsilon \\
& + \sup_{f\in\mathcal{F}} \int_{Z\setminus Z_\epsilon} |c(z, f(x))| P_N(dz),
\end{aligned}
$$

where the first inequality holds due to $\|f\|_k \le \beta$ and $\lambda_N \le \epsilon/\beta^2$ for $N \ge N_0$, and the last inequality holds because under Assumption 4 we have

$$
\begin{aligned}
\mathbb{E}_P[c(z, f(x))\mathbf{1}_{Z_\epsilon}(z) - c(z, f_k(x))\mathbf{1}_{Z_\epsilon}(z)] &\le \mathbb{E}_P[r(z)\|f - f_k\|_{Z_{\epsilon,\infty}}^{\nu}\mathbf{1}_{Z_\epsilon}(z)] \\
&\le \bar{r}\bar{\epsilon}^{\nu} = \epsilon
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbb{E}_{P_N}[c(z, f(x))\mathbf{1}_{Z_\epsilon}(z) - c(z, f_k(x))\mathbf{1}_{Z_\epsilon}(z)] &\le \mathbb{E}_{P_N}[r(z)\|f - f_k\|_{Z_{\epsilon,\infty}}^{\nu}\mathbf{1}_{Z_\epsilon}(z)] \\
&\le \bar{r}\bar{\epsilon}^{\nu} = \epsilon.
\end{aligned}
$$

It follows from by the classical Cramér's large deviation theorem that for each k there exist positive constants $C(\epsilon, \delta, f_k)$ and $\gamma(\epsilon, \delta, f_k)$ such that

$$
P^{\otimes N}\left(|\mathbb{E}_{P_N}[c(z, f_k(x))\mathbf{1}_{Z_\epsilon}(z)] - \mathbb{E}_P[c(z, f_k(x))\mathbf{1}_{Z_\epsilon}(z)]| \ge \delta - 6\epsilon\right) \le C(\epsilon, \delta, f_k)e^{-N\gamma(\epsilon,\delta,f_k)}.
$$

Hence, we have

$$\begin{aligned}
& P^{\otimes N}\left(\sup_{f\in\mathcal{F}}|\mathbb{E}_{P_N}[c(z,f(x))]+\lambda_N\|f\|_k^2-\mathbb{E}_{P_N}[c(z,f(x))]|\geq\delta\right)\\
\leq\ & P^{\otimes N}\left(\sup_{k\in\{1,\ldots,K\}}|\mathbb{E}_{P_N}[c(z,f_k(x))\mathbf{1}_{Z_\epsilon}(z)]-\mathbb{E}_P[c(z,f_k(x))\mathbf{1}_{Z_\epsilon}(z)]|\geq\delta-6\epsilon\right)\\
& +P^{\otimes N}\left(\sup_{f\in\mathcal{F}}\int_{Z\setminus Z_\epsilon}|c(z,f(x))|P_N(dz)\geq 2\epsilon\right)\\
\leq\ & \sum_{k\in\{1,\ldots,K\}}P^{\otimes N}\left(|\mathbb{E}_{P_N}[c(z,f_k(x))\mathbf{1}_{Z_\epsilon}(z)]-\mathbb{E}_P[c(z,f_k(x))\mathbf{1}_{Z_\epsilon}(z)]|\geq\delta-6\epsilon\right)\\
& +C_0e^{-\gamma_0 N}\\
\leq\ & \sum_{k\in\{1,\ldots,K\}}C(\epsilon,\delta,f_k)e^{-N\gamma(\epsilon,\delta,f_k)}+C_0e^{-\gamma_0 N},
\end{aligned}$$

which implies (60). Finally, (61) follows from (60) due to the fact that

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

The proof is complete. ■

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 7 (Uniform consistency of $\vartheta(P_N,\lambda_N)$) *Let*

$$\mathcal{M}_{Z,\kappa}^{\phi^p}=\left\{P\in\mathscr{P}(Z):\int_Z\phi(z)^pP(dz)\leq\kappa\right\}$$

for some fixed $p>1$ and $\kappa>0$. Assume: (a) the conditions in Theorem 2 are satisfied, (b) Assumption 4 hold, (c) $\mathcal{M}$ is a weakly compact (i.e., tight [3] and closed under the ϕ-weak topology) subset of $\mathcal{M}_{Z,\kappa}^{\phi^p}$. Then for every $\epsilon>0$ and $\delta>0$, there exists N_0 such that

$$\sup_{P\in\mathcal{M}}P^{\otimes N}\left(|\vartheta(P_N,\lambda_N)-\vartheta(P)|\geq\delta\right)\leq\epsilon, \tag{66}$$

when $\lambda_N\leq\delta/(4\beta^2)$, where β is is some positive constant and $N\geq N_0$.

The uniform convergence (66) is closely related to learnability in statistical learning theory which is defined as the uniform convergence of $R(f_N(P_N))$ to $\vartheta(P)$ for all empirical probability distributions drawn from $\mathscr{P}(Z)$, where $R(\cdot)$ is defined as in (1), see Definition

3. $\mathcal{M}$ is tight under the ϕ-weak topology if for any $\epsilon>0$, there exists a compact set $K\subset Z$ such that $\sup_{P\in\mathcal{M}}\int_{Z\setminus K}\phi(t)P(dt)\leq\epsilon$.

1 in Shalev-Shwartz et al. (2010). 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 7. We first show that (66) holds for each $P \in \mathcal{M} \subset \mathcal{M}_{Z,\kappa}^{\phi^p}$. For fixed $\bar{P}$, by the continuity of $\vartheta(\cdot)$ at $\bar{P}$, 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 Corollary 3.5 in Krätschmer, Schied and Zähle (2012) that $(\mathcal{M}_{Z,\kappa}^{\phi^p}, \mathsf{dl}_\phi)$ 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 $N \geq N_0$

$$P^{\otimes N}\left(\mathsf{dl}_\phi(P_N, P) \geq \eta\right) \leq \epsilon, \forall P \in \mathcal{M}_{Z,\kappa}^{\phi^p}.$$

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

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

Consequently, we have

$$\begin{aligned}
\vartheta(\bar{P}_N, \lambda_N) - \vartheta(\bar{P}) &= \inf_{f\in\mathcal{F}}\{\mathbb{E}_{\bar{P}_N}[c(z, f(x))] + \lambda_N\|f\|_k^2\} - \inf_{f\in\mathcal{F}}\mathbb{E}_{\bar{P}}[c(z, f(x))] \\
&= \inf_{f\in\mathcal{F}}\{\mathbb{E}_{\bar{P}_N}[c(z, f(x))] + \lambda_N\|f\|_k^2\} - \inf_{f\in\mathcal{F}}\mathbb{E}_{\bar{P}}[c(z, f(x))] \\
&\leq \inf_{f\in\mathcal{F}}\mathbb{E}_{\bar{P}_N}[c(z, f(x))] + \sup_{f\in\mathcal{F}}\lambda_N\|f\|_k^2 - \inf_{f\in\mathcal{F}}\mathbb{E}_{\bar{P}}[c(z, f(x))] \\
&\leq |\inf_{f\in\mathcal{F}}\mathbb{E}_{\bar{P}_N}[c(z, f(x))] - \inf_{f\in\mathcal{F}}\mathbb{E}_{\bar{P}}[c(z, f(x))]| + \sup_{f\in\mathcal{F}}\lambda_N\|f\|_k^2 \\
&= |\vartheta(\bar{P}_N) - \vartheta(\bar{P})| + \lambda_N\beta^2.
\end{aligned}$$

Likewise, we can show that

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

Thus

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

when $\lambda_N \leq \delta/(4\beta^2)$. Therefore, (66) holds when P is fixed at $\bar{P}$.

Now we show (66) holds for all $P \in \mathcal{M}$. Assume for the sake of a contradiction that there exist some positive numbers ϵ_0 and δ_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

$$P_{s'}^{\otimes N_{s'}}\left(|\vartheta(P_{N_{s'}}, \lambda_{N_{s'}}) - \vartheta(P_{s'})| \geq \delta_0\right) > \epsilon_0. \tag{67}$$

Let s increase. Then we obtain a sequence of $\{P_{s'}\}$ which satisfies (67). Since $\mathcal{M}$ is weakly compact under the ϕ-weak topology, then $\{P_{s'}\}$ has a converging subsequence. Assume

without loss of generality that $P_{s'} \xrightarrow{\phi} P_* \in \mathcal{M}$. Since $\vartheta(\cdot)$ is continuous at P_*, then there exists $\eta > 0$ such that $|\vartheta(Q) - \vartheta(P_*)| < \delta_0/4$ for Q satisfying $\mathsf{dl}_\phi(Q, P_*) < \eta$ and then

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

for $\lambda' \leq \delta_0/(4\beta^2)$. By $P_{s'} \xrightarrow{\phi} P_*$, there exists s'_0 such that $\mathsf{dl}_\phi(P_{s'}, P_*) < \eta/2$ and $\lambda_{s'} \leq \delta_0/(4\beta^2)$ for $s' \geq s'_0$, and then $|\vartheta(P_{s'}) - \vartheta(P_*)| < \delta_0/4$ and $|\vartheta(P_{s'}, \lambda_{s'}) - \vartheta(P_*)| < \delta_0/2$. On the other hand, by the UGC property, there exists a constant $N_0 > 0$ such that

$$\begin{aligned} P_{s'}^{\otimes N_{s'}}(\mathsf{dl}_\phi(P_{N_{s'}}, P_*) \geq \eta) &\leq P_{s'}^{\otimes N_{s'}}(\mathsf{dl}_\phi(P_{N_{s'}}, P_{s'}) + \mathsf{dl}_\phi(P_{s'}, P_*) \geq \eta) \\ &= P_{s'}^{\otimes N_{s'}}(\mathsf{dl}_\phi(P_{N_{s'}}, P_{s'}) \geq \eta - \mathsf{dl}_\phi(P_{s'}, P_*)) \\ &\leq P_{s'}^{\otimes N_{s'}}(\mathsf{dl}_\phi(P_{N_{s'}}, P_{s'}) \geq \eta/2) \leq \epsilon_0 \end{aligned}$$

for $N_{s'} \geq N_0$ and $s' \geq s'_0$. Therefore,

$$P_{s'}^{\otimes N_{s'}}\left(|\vartheta(P_{N_{s'}}, \lambda_{N_{s'}}) - \vartheta(P_*)| \geq \delta_0/2\right) \leq P_{s'}^{\otimes N_{s'}}(\mathsf{dl}_\phi(P_{N_{s'}}, P_*) \geq \eta) \leq \epsilon_0,$$

and

$$\begin{aligned} & P_{s'}^{\otimes N_{s'}}\left(|\vartheta(P_{N_{s'}}, \lambda_{N_{s'}}) - \vartheta(P_{s'})| \geq \delta_0\right) \\ \leq\ & P_{s'}^{\otimes N_{s'}}\left(|\vartheta(P_{N_{s'}}, \lambda_{N_{s'}}) - \vartheta(P_*)| + |\vartheta(P_{s'}) - \vartheta(P_*)| \geq \delta_0\right) \\ \leq\ & P_{s'}^{\otimes N_{s'}}\left(|\vartheta(P_{N_{s'}}, \lambda_{N_{s'}}) - \vartheta(P_*)| \geq 3\delta_0/4\right) \leq \epsilon_0, \end{aligned}$$

which leads to a contradiction with (67) as desired. ■

6. Concluding remarks

In this paper, we present some theoretical analysis about statistical robustness of empirical risk in machine learning and we do so for the regularized problem (6). All of our results hold when the regularization parameter is set zero, which means that they are applicable to non-regularized problem (2). There are a number of issues remain to be explored.

First, 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. Furthermore, it might be interesting to carry out some numerical experiments to examine the statistical robustness of the empirical risk. Second, there is a prospect to extend the statistical robustness results established in this paper to deep learning model based on the recent representer theorem (Unser, 2019, Theorem 4). To see this, we note that our main statistical robustness results, Theorem 3 and Theorem 5 do not rely on the structure of RKHS directly, rather they depend on the continuity of the optimal value of $\vartheta(P)$ w.r.t. P and $\vartheta(P_N, \lambda_N)$ w.r.t. P_N (samples). The main stability result, Theorem 2, depends on the structure of RKHS but we envisage that a similar result can be established outside framework of RKHS when f is confined to a class of piecewise linear functions as specified in Theorem 4 of Unser (2019). Of course, all these will have to be carried out under a

completely different regularized learning framework in a Banach space equipped with total variation distance, see (15) in Unser (2019). We leave all these for future research as they require much more intensive work.

Acknowledgments

The authors would like to thank two referees for valuable comments and Francis Bach for effective handling of the review. The first author would like to acknowledge support from the National Key R&D Program of China (No.2022YFA1004000) and the Natural Science Foundation of China (No.12271077). The second author would like to acknowledge support from RGC grant (No.14204821). The third author would like to acknowledge support from the Natural Science Foundation of China (No.11971089).

References

N. Alon, S. Ben-David, N. Cesa-Bianchi and D. Haussler, Scale-sensitive dimensions, uniform convergence, and learnability, *Journal of the ACM*, 44(1997): 615-631.

K. Balasubramanian and M. Yuan, Discussion of estimating structured high-dimensional covariance and precision matrices: optimal rates and adaptive estimation, *Electron. J. Stat.*, 10(2016): 71-73.

P. L. Bartlett and S. Mendelson, Rademacher and Gaussian complexities: Risk bounds and structural results, *J. Mach. Learn. Res.*, 3(2002): 463-482. .

P. Billingsley, *Convergence of Probability Measures*, Wiley, New York, 1999.

A. Blumer, A. Ehrenfeucht, D. Haussler and M. Warmuth, Learnability and the Vapnik-Chervonenkis dimension, *J. Assoc. Comp. Mach.*, 36(1989): 929-965.

S. Boucheron, O. Bousquet and G. Lugosi, Theory of classification: a survey of some recent advances, *ESAIM: Prob. Statist.*, 9(2005): 323-375.

O. Bousquet and A. Elisseeff, Stability and generalization, *J. Mach. Learn. Res.*, 2(2002): 499-526.

P. Breheny and J. Huang, Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors, *Statist. Comput.*, 25(2015): 173-187.

R. F. Brown, *A Topological Introduction to Nonlinear Analysis*, Spring, New York, 2004.

A. Caponnetto and E. De Vito, Optimal rates for the regularized least-squares algorithm, *Found. Comput. Math.*, (2007): 331-368.

R. Chen and I. C. Paschalidis, A robust learning approach for regression models based on distributionally robust optimization, *J. Mach. Learn. Res.*, 19(2018): 517-564.

M. Claus, *Advancing Stability Analysis of Mean-Risk Stochastic Programs: Bilevel and Two-Stage Models*, PhD Thesis, Universität Dusburg-Essen, 2016.

R. Cont, R. Deguest and G. Scandolo, Robustness and sensitivity analysis of risk measurement procedures, *Quant. Finan.*, 10(2010): 593-606.

F. Cucker and S. Smale, Best choices for regularization parameters in learning theory: on the bias-variance problem, *Found. Comput. Math.*, 2(2002): 413-428.

F. Cucker and S. Smale, On the mathematical foundations of learning, *Bull. Amer. Math. Soc. (N.S.),* 39(2002): 1-49.

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

D. Davis and D. Drusvyatskiy, Graphical convergence of subgradients in nonconvex optimization and learning, *Math. Oper. Res*, 47(2021): 209-231.

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

E. De Vito, A. Caponnetto and L. Rosasco, Model selection for regularized least-squares algorithm in learning theory, *Found. Comput. Math.*, 5(2005): 59-85.

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

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

N. Fournier and A. Guilline, On the rate of convergence in Wasserstein distance of the empirical measure, *Probab. Theory Rel.*, 162(2015): 707-738.

S. Guo and H. Xu, Statistical robustness in utility preference robust optimization models, *Math. Program.*, 190(2021): 679-720.

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

F. R. Hampel. *Contribution to the theory of robust estimation.* Ph. D. Thesis, University of California, Berkeley, 1968.

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

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

J. Jiang and S. Li, Statistical robustness of two-stage stochastic variational inequalities, *Optim. Lett.*, 16(2022): 2591-2605.

G. S. Kimeldorf and G. Wahba, A correspondence between Bayesian estimation on stochastic processes and smoothing by splines, *Ann. Math. Statist.*, 45(1970): 495-502.

V. Krätschmer, A. Schied and H. Zähle, Comparative and statistical robustness for law-invariant risk measures, *Finance Stoch.*, 18(2014): 271-295.

V. Krätschmer, A. Schied and H. Zähle, Qualitative and infinitesimal robustness of tail-dependent statistical functionals, *J. Multi. Anal.*, 103(2012): 35-47.

V. Krätschmer, A. Schied and H. Zähle, Domains of weak continuity of statistical functionals with a view toward robust statistics, *J. Multi. Anal.*. 158(2017): 1-19.

G. Lecué and M. Lerasle, Robust machine learning by median-of-means: theory and practice, *Ann. Stat.*, 48(2020): 906-931.

Y. Liu and H. Xu, Stability analysis of stochastic programs with second order dominance constraints, *Math. Program.*, 142(2013): 435-460.

M. Mohri, A. Rostamizadeh and A. Talwalkar, *Foundation of Machine Learning*, MIT Press, Cambridge, 2012.

V. Norkin and M. Keyzer, On convergence of kernel learning estimators, *SIAM J. Optim.*, 20(2009): 1205-1223.

T. Poggio and S. Smale, The mathematics of learning: Dealing with data, *Notices Amer. Math. Soc.*, 50(2003): 537-544.

R. Ranga Rao, Relations between weak and uniform convergence of measures with applications, *Ann. Math. Statist.*, 33(1962): 659-680.

R. T. Rockafellar and J-B Wets, *Variational Analysis*, Springer, New York, 1998.

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.

W. Römisch and R. Schultz, Stability analysis for stochastic programs, *Ann. Oper. Res.*, 30(1991): 241-266.

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

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

S. Shafieezadeh-Abadeh, D. Kuhn and P. Esfahani, Regularization via mass transportation, *J. Mach. Learn. Res.*, 20(2019): 1-68.

S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*, Cambridge University Press, Cambridge, 2014.

S. Shalev-Shwartz, O. Shamir, N. Srebro and K. Sridharan, Learnability, stability and uniform convergence, *J. Mach. Learn. Res.*, 11(2010): 2635-2670.

S. Smale and Y. Yao, Online learning algorithms, *Found. Comput. Math.*, 6(2006): 145–170.

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

I. Takeuchi, Q. Le, T. Sears, and A. Smola, Nonparametric quantile estimation, *J. Mach. Learn. Res.*, 7(2006): 1231-1264.

R. Tibshirani, Regression shrinkage and selection via the lasso, *J. Roy. Statist. Soc. Ser. B*, 58(1996): 267-288.

J. W. Tukey, A survey of sampling from contaminated distributions, *Contri. Prob. Statist.*, 2(1960): 448–485.

J. W. Tukey, The future of data analysis, *Ann. Math. Statist.*, 33(1962): 1-67.

M. Unser, A representer theorem for deep neural networks, *J. Mach. Learn. Res.*, 20(2019): 1-30.

W. Wang, H. Xu and T. Ma, Quantitative statistical robustness for tail-dependent law invariant risk measures, *Quant. Finance*, 21(2021): 1669-1685.

H. Xu, C. Caramanis and S. Mannor, Robustness and regularization of support vector machines, *J. Mach. Learn. Res.*, 10(2009): 1485-1510.

H. Xu and S. Zhang, Quantitative Statistical Robustness in Distributionally Robust Optimization Models, *Pac. J. Optim.*, 19(2023): 335-361.