Journal of Machine Learning Research 16 (2015) 993-1034 Submitted 12/13; Revised 8/14; Published 5/15

Learning with the Maximum Correntropy Criterion Induced Losses for Regression

Yunlong Feng YUNLONG.FENG@ESAT.KULEUVEN.BE
Xiaolin Huang HUANGXL06@MAILS.TSINGHUA.EDU.CN
Department of Electrical Engineering, ESAT-STADIUS, KU Leuven
Kasteelpark Arenberg 10, Leuven, B-3001, Belgium

Lei Shi LEISHI@FUDAN.EDU.CN
Shanghai Key Laboratory for Contemporary Applied Mathematics
School of Mathematical Sciences, Fudan University, Shanghai, 200433, P.R. China

Yuning Yang YUNING.YANG@ESAT.KULEUVEN.BE
Johan A. K. Suykens JOHAN.SUYKENS@ESAT.KULEUVEN.BE
Department of Electrical Engineering, ESAT-STADIUS, KU Leuven
Kasteelpark Arenberg 10, Leuven, B-3001, Belgium

Editor: Saharon Rosset

Abstract

Within the statistical learning framework, this paper studies the regression model associated with the correntropy induced losses. The correntropy, as a similarity measure, has been frequently employed in signal processing and pattern recognition. Motivated by its empirical successes, this paper aims at presenting some theoretical understanding towards the maximum correntropy criterion in regression problems. Our focus in this paper is two-fold: first, we are concerned with the connections between the regression model associated with the correntropy induced loss and the least squares regression model. Second, we study its convergence property. A learning theory analysis which is centered around the above two aspects is conducted. From our analysis, we see that the scale parameter in the loss function balances the convergence rates of the regression model and its robustness. We then make some efforts to sketch a general view on robust loss functions when being applied into the learning for regression problems. Numerical experiments are also implemented to verify the effectiveness of the model.

Keywords: correntropy, the maximum correntropy criterion, robust regression, robust loss function, least squares regression, statistical learning theory

1. Introduction and Motivation

Recently, a generalized correlation function named correntropy (see Santamaría et al., 2006) has drawn much attention in the signal processing and machine learning community (see Liu et al., 2007; Gunduz and Príncipe, 2009; He et al., 2011a,b). Formally speaking, correntropy is a generalized similarity measure between two scalar random variables U and V, which is defined by $\mathcal{V}_\sigma(U,V) = \mathbb{E}\mathcal{K}_\sigma(U,V)$. Here $\mathcal{K}_\sigma$ is a Gaussian kernel given by $\mathcal{K}_\sigma(u,v) = \exp\left\{-(u-v)^2/\sigma^2\right\}$ with the scale parameter $\sigma > 0$, (u,v) being a realization of (U,V).

In this paper, we are interested in the application of the similarity measure $\mathcal{V}_\sigma$ in regression problems, namely, the maximum correntropy criterion for regression. Therefore, we first assume that the data generation model is given as

$$Y = f^\star(X) + \epsilon, \quad \mathbb{E}(\epsilon \,|\, X = x) = 0. \tag{1}$$

In model (1), X is the explanatory variable that takes values in a separable metric space $\mathcal{X}$ and $Y \in \mathcal{Y} = \mathbb{R}$ stands for the response variable. The main purpose of the regression problem is to estimate f^* from a set of observations generated by (1). The underlying unknown probability distribution on $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$ is denoted as ρ.

Under the regression model (1), probably the most widely employed methodology for quantifying the regression efficiency is the mean squared error. This is the classical tool that minimizes the variance of ϵ and belongs to the second-order statistics. The drawback of the second-order statistics is that its optimality depends heavily on the assumption of Gaussianity. However, in many real-life applications, data may be contaminated by non-Gaussian noise or outliers. This motivates the introduction of the maximum correntropy criterion into the regression problems.

Given a set of i.i.d observations $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$, for any $f : \mathcal{X} \to \mathcal{Y}$, the empirical estimator of the correntropy between $f(X)$ and Y is given as

$$\widehat{\mathcal{V}}_{\sigma,\mathbf{z}}(f) = \frac{1}{m} \sum_{i=1}^{m} \mathcal{K}_\sigma(y_i, f(x_i)).$$

The maximum correntropy criterion for regression models the output function via maximizing the empirical estimator of $\mathcal{V}_\sigma$ as follows

$$f_{\mathbf{z}} = \arg\max_{f \in \mathcal{H}} \widehat{\mathcal{V}}_{\sigma,\mathbf{z}}(f),$$

where $\mathcal{H}$ is a certain underlying hypothesis space. The maximum correntropy criterion in regression problems has shown its efficiency for cases when the noises are non-Gaussian, and also with large outliers (see Santamaría et al., 2006; Liu et al., 2007; Príncipe, 2010; Wang et al., 2013). It has also succeeded in many real-world applications, e.g., wind power forecasting (see Bessa et al., 2009) and pattern recognition (see He et al., 2011b).

In this paper, we attempt to present a theoretical understanding on the maximum correntropy criterion for regression (MCCR) within the statistical learning framework. To this end, we first generalize the idea of the maximum correntropy criterion in regression problems using the following supervised regression loss:

Definition 1 *The correntropy induced regression loss $\ell_\sigma : \mathbb{R} \times \mathbb{R} \to [0, +\infty)$ is defined as*

$$\ell_\sigma(y, t) = \sigma^2 \left(1 - e^{-\frac{(y-t)^2}{\sigma^2}}\right), \; y \in \mathcal{Y}, \, t \in \mathbb{R},$$

with $\sigma > 0$ being a scale parameter.

Figure 1 plots the correntropy induced loss function ℓ_σ (the ℓ_σ loss for short in what follows) with different choices of σ. Associated with this regression loss, the MCCR model

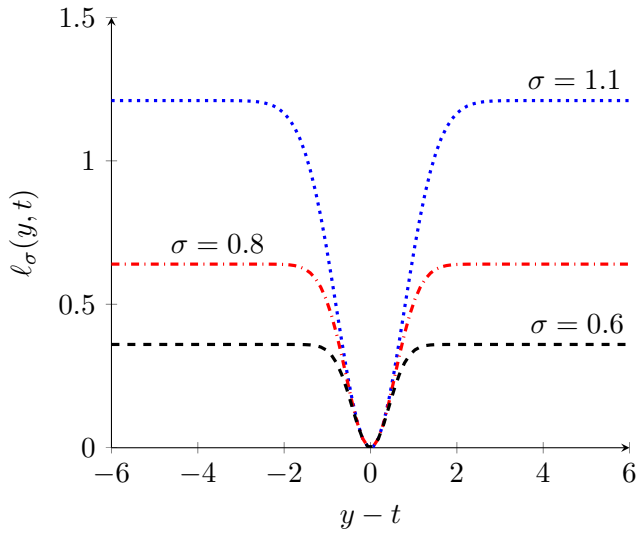

Figure 1: Plots of $\ell_\sigma(y,t) = \sigma^2(1 - e^{-(y-t)^2/\sigma^2})$ with respect to $y - t$ for different σ values: $\sigma = 0.6$ (the dashed curve), $\sigma = 0.8$ (the dotted-dashed curve), and $\sigma = 1.1$ (the dotted curve).

that we will investigate is the following empirical risk minimization (ERM) scheme

$$f_{\mathbf{z}} = \arg\min_{f\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} \ell_\sigma(y_i, f(x_i)), \tag{2}$$

where, throughout, the hypothesis space $\mathcal{H}$ is assumed to be a compact subset of $C(\mathcal{X})$. Here $C(\mathcal{X})$ is the Banach space of continuous functions on $\mathcal{X}$ with the norm $\|f\|_\infty = \sup_{x\in\mathcal{X}} |f(x)|$. Note that the compactness of $\mathcal{H}$ ensures the existence of the empirical target function $f_{\mathbf{z}}$.

We remark that the ℓ_σ loss is in fact a variant of the Welsch function, which was originally introduced in robust statistics (see Holland and Welsch, 1977; Dennis and Welsch, 1978). Consequently, the estimator from the MCCR model (2) is essentially a non-parametric M-estimator. For linear regression models, the robustness and the consistency properties of the M-estimator induced by the ℓ_σ loss have been investigated in Wang et al. (2013). In Santamaría et al. (2006) and Liu et al. (2007), an information-theoretical interpretation of the ℓ_σ loss by viewing it as a correlation measurement is provided.

However, existing theoretical results on understanding the ℓ_σ loss and the MCCR model are still very limited, the barriers of which lie in their non-convexity properties. From Taylor's expansion, it is easy to see that there holds $\ell_\sigma(t) \approx t^2$ for sufficiently large σ. Therefore, in some existing empirical studies, the ℓ_σ loss has been roughly taken as the least squares loss when σ is large enough. However, our studies in this paper suggest that this is in general not the case. On the other hand, the consistency property and the convergence rates of the MCCR model are yet unknown, which are the central focuses of the statistical learning research. In view of the above considerations, in this paper, our main concerns are the following two aspects:

- We are concerned with the connections between the ℓ_σ loss and the least squares loss when they are employed in regression problems. Therefore, we will study the relations between the MCCR model (2) and the ERM-based least squares regression (LSR) model.

- We are concerned with the approximation ability of the output function $f_{\mathbf{z}}$ modeled by (2). More concretely, we aim at carrying out a learning theory analysis to bound the difference between $f_{\mathbf{z}}$ and $f^\star$.

It should be mentioned that our study on the MCCR model (2) is inspired by Hu et al. (2013), which presented comprehensive and thorough studies on the minimum error entropy criterion from a learning theory viewpoint. According to Hu et al. (2013), a specific form of the minimum error entropy criterion for regression (MEECR) can be stated as

$$\widetilde{f}_{\mathbf{z}} = \arg\min_{f\in\mathcal{H}} \left\{ -\frac{\sigma^2}{m(m-1)} \sum_{i=1}^{m} \sum_{j\neq i} G\left\{ \frac{[(y_i - f(x_i)) - (y_j - f(x_j))]^2}{2\sigma^2} \right\} \right\},$$

where $G(\cdot)$ is a window function and can be chosen as $G(t) = \exp(-t)$. Hu et al. (2013, 2014) presented the first results concerning the regression consistency and convergence rates of the above MEECR model and its regularized variant when σ becomes large. Concerning the two regression models, we can see that MEECR models the empirical target function $\widetilde{f}_{\mathbf{z}}$ via a pairwise empirical risk minimization scheme while the MCCR model learns in a point-wise fashion. More discussions on the two different learning schemes will be provided in Section 2.

The rest of this paper is organized as follows. In Section 2, results on the convergence rates of the MCCR model (2) in different situations are provided. Discussions and comparisons with related studies will be also presented. Section 3 concerns connections between the two regression models: MCCR and LSR, which are explored from three aspects. Section 4 is dedicated to analyzing the MCCR model and giving proofs of theoretical results stated in Section 2. Discussions on the role that the scale parameter σ in the ℓ_σ loss plays is given in Section 5. Section 6 makes some efforts in sketching a general view of learning with robust regression losses. Numerical experiments are implemented in Section 7. We end this paper with concluding remarks in Section 8.

2. Theoretical Results on Convergence Rates and Discussions

In this section, we provide theoretical results on the convergence rates of the MCCR model (2). Explicitly, denoting $\rho_{\mathcal{X}}$ as the marginal distribution of ρ on $\mathcal{X}$, we are going to bound $\|f_{\mathbf{z}} - f_\rho\|^2_{\mathcal{L}^2_{\rho_{\mathcal{X}}}}$, where f_ρ is defined as

$$f_\rho(x) = \int_{\mathcal{Y}} y d\rho(y|x), \ \ x \in \mathcal{X},$$

and is assumed to satisfy that $f_\rho \in L^\infty_{\rho_X}$ throughout this paper. Due to the zero-mean noise assumption in the data generation model (1), almost surely there holds $f_\rho = f^\star$. To analyze

the convergence of the model, we need to introduce the following target function in $\mathcal{H}$

$$f_{\mathcal{H}} = \arg\min_{f\in\mathcal{H}} \int_{\mathcal{Z}} (f(x)-y)^2 d\rho.$$

In addition, the convergence rates that we are going to present are obtained by controlling the complexity of the hypothesis space $\mathcal{H}$. Therefore, we need the following definitions and assumptions to state our main results.

2.1 Definitions and Assumptions

Definition 2 (Covering Number) *The covering number of the hypothesis space $\mathcal{H}$, which is denoted as $\mathcal{N}(\mathcal{H},\eta)$ with the radius $\eta>0$, is defined as*

$$\mathcal{N}(\mathcal{H},\eta) := \inf\left\{ l\geq 1 : \text{there exist } f_1,\ldots,f_l\in\mathcal{H}, \text{such that } \mathcal{H}\subset\bigcup_{i=1}^{l} B(f_i,\eta)\right\},$$

where $B(f,\eta)=\{g\in\mathcal{H}:\|f-g\|_\infty\leq\eta\}$ denotes the closed ball in $C(\mathcal{X})$ with center $f\in\mathcal{H}$ and radius η.

Definition 3 (ℓ^2-Empirical Covering Number) *Let $\mathbf{x}=\{x_1,x_2,\ldots,x_n\}\subset\mathcal{X}^n$. The ℓ^2-empirical covering number of the hypothesis space $\mathcal{H}$, which is denoted as $\mathcal{N}_2(\mathcal{H},\eta)$ with radius $\eta>0$, is defined by*

$$\mathcal{N}_2(\mathcal{H},\eta) := \sup_{n\in\mathbb{N}}\sup_{\mathbf{x}\in\mathcal{X}^n}\inf\Big\{\ell\in\mathbb{N}:\exists\{f_i\}_{i=1}^{\ell}\subset\mathcal{H} \text{ such that for each } f\in\mathcal{H}, \text{there exists some } i\in\{1,2,\ldots,\ell\} \text{ with } \frac{1}{n}\sum_{j=1}^{n}|f(x_j)-f_i(x_j)|^2\leq\eta^2\Big\}.$$

Assumption 1 (Complexity Assumption I) *There exist positive constants p and $c_{I,p}$ such that*

$$\log\mathcal{N}(\mathcal{H},\eta)\leq c_{I,p}\eta^{-p},\ \forall\ \eta>0.$$

Assumption 2 (Complexity Assumption II) *There exist positive constants s and $c_{II,s}$ with $0<s<2$, such that*

$$\log\mathcal{N}_2(\mathcal{H},\eta)\leq c_{II,s}\eta^{-s},\forall\ \eta>0.$$

In learning theory, the covering number is frequently used to measure the capacity of the hypothesis spaces (see Anthony and Bartlett, 1999; Zhou, 2002). As explained in Zhou (2002), the Complexity Assumption I is typical in the statistical learning theory literature. For instance, it holds when $\mathcal{H}$ is chosen as a ball of reproducing kernel Hilbert spaces induced by Sobolev smooth kernels. The ℓ^2-empirical covering number is another data-dependent complexity measurement and usually leads to sharper convergence rates. Several examples of hypothesis spaces satisfying the Complexity Assumption II can be found in Guo and Zhou (2013).

Assumption 3 (Moment Assumption) *Assume that the tail behavior of the response variable Y satisfies $\int_{\mathcal{Z}} y^4 d\rho < \infty$.*

We will give some discussions on the above Moment Assumption in Subsection 2.3. In our study, the Moment Assumption will be employed to analyze the convergence of the MCCR model. For some specific situations of the regression model (1), in our study we will also restrict ourselves to the noise that satisfies the following Noise Assumption.

Assumption 4 (Noise Assumption) *The density function of the noise variable ϵ for any given $X = x$, which is denoted as $p_{\epsilon|X=x}$, is symmetric and uniformly bounded by the interval $[-M_0, M_0]$ with $M_0 > 0$.*

2.2 Theoretical Results on Convergence Rates

We are now ready to state our main results on the convergence rates of the MCCR model (2). Our first result considers a general case of the regression model (1), where the Moment Assumption is assumed to hold.

Theorem 4 *Assume that the Complexity Assumption I with $p > 0$ and the Moment Assumption hold. Let $f_{\mathbf{z}}$ be produced by (2). For any $0 < \delta < 1$, with confidence $1 - \delta$, there holds*

$$\|f_{\mathbf{z}} - f_\rho\|^2_{\mathcal{L}^2_{\rho_{\mathcal{X}}}} \leq 3\,\|f_{\mathcal{H}} - f_\rho\|^2_{\mathcal{L}^2_{\rho_{\mathcal{X}}}} + C_{\mathcal{H},\rho} \log(2/\delta) \left(\sigma^{-2} + \sigma m^{-1/(1+p)}\right),$$

where $C_{\mathcal{H},\rho}$ is a positive constant independent of m, σ or δ and will be given explicitly in the proof.

Discussions on the convergence rates established in Theorem 4 are postponed to Subsection 2.3. Here we remark that the moment condition in the Moment Assumption which is used in Theorem 4 can be relaxed to a weaker moment condition, i.e., $\int_{\mathcal{Z}} |y|^\ell d\rho < \infty$ with $\ell > 2$, where meaningful convergence rates can be still derived. Meanwhile, when the condition in the Moment Assumption is further strengthened, refined convergence rates can be derived. For instance, when $|y| \leq M$ almost surely for some $M > 0$, we can get the following improved convergence rates:

Theorem 5 *Assume that the Complexity Assumption II with $0 < s < 2$ holds, and $|y| \leq M$ almost surely for some $M > 0$. Let $f_\rho \in \mathcal{H}$ and $f_{\mathbf{z}}$ be produced by (2) with $\sigma = m^{1/(2+s)}$. For any $0 < \delta < 1$, with confidence $1 - \delta$, there holds*

$$\|f_{\mathbf{z}} - f_\rho\|^2_{\mathcal{L}^2_{\rho_{\mathcal{X}}}} \leq C'_{\mathcal{H},\rho} \log(2/\delta) m^{-\frac{2}{2+s}},$$

where $C'_{\mathcal{H},\rho}$ is a positive constant independent of m, σ or δ and will be given explicitly in the proof.

From Theorem 4 and Theorem 5, we can see that meaningful convergence rates can be obtained when σ is properly chosen, e.g., $\sigma = \mathcal{O}(m^\alpha)$ with some $\alpha > 0$. That is, σ has to grow in accordance with the sample size m to ensure non-trivial convergence rates. In view of this, it is natural to ask whether one can also get consistency properties or even convergence rates for the MCCR model (2) when σ is fixed. Under certain conditions, we give a positive answer in the following theorem.

Theorem 6 *Assume that the Complexity Assumption II with $0 < s < 2$ and the Noise Assumption hold. Let $f_\rho \in \mathcal{H}$, $f_{\mathbf{z}}$ be produced by (2) with σ being fixed and $\sigma > \sigma_{\mathcal{H},\rho}$ where*

$$\sigma_{\mathcal{H},\rho} = \sqrt{2}\Big(M_0 + \|f_\rho\|_\infty + \sup_{f\in\mathcal{H}} \|f\|_\infty\Big).$$

For any $0 < \delta < 1$, with confidence $1-\delta$, there holds

$$\|f_{\mathbf{z}} - f_\rho\|^2_{L^2_{\rho_\mathcal{X}}} \le C_{\mathcal{H},\sigma,\rho} \log(1/\delta) m^{-\frac{2}{2+s}},$$

where $C_{\mathcal{H},\sigma,\rho}$ is a positive constant independent of m or δ and will be given explicitly in the proof.

Proofs of the above theorems will be given in Subsection 4.3.

2.3 Discussions and Comparisons

We now give some discussions on the obtained convergence rates, the Moment Assumption and also comparisons with related studies.

2.3.1 Convergence Rates

As shown in Theorem 4, under the Moment Assumption, the convergence rates of the MCCR model depend on the choice of σ and the regularity of f_ρ. In the case when $f_\rho \in \mathcal{H}$ and $\sigma = \mathcal{O}(m^{1/(2+2p)})$, the convergence rate of $\mathcal{O}(m^{-2/(3+3p)})$ can be obtained. We then show in Theorem 5 and Theorem 6 that under the boundedness assumption on Y, or with the Noise Assumption, refined convergence rates of $\mathcal{O}(m^{-2/(2+s)})$ can be derived. Note that when s tends to zero which corresponds to the case where functions in $\mathcal{H}$ are smooth enough, convergence rates established in Theorem 5 and Theorem 6 tend to $\mathcal{O}(m^{-1})$, which are considered as the optimal rates in learning theory according to the law of large numbers (see Caponnetto and De Vito, 2007; Steinwart et al., 2009; Mendelson and Neeman, 2010; Wang and Zhou, 2011). The established convergence rates indicate the feasibility of applying the ℓ_σ loss in regression problems.

2.3.2 Moment Assumption and Related Studies on Robustness

Note that convergence rates in Theorem 4 are obtained under the Moment Assumption, which restricts the tail behavior of Y. In fact, as commented in Christmann and Steinwart (2007), tail properties of Y are frequently used in linear regression as well as nonparametric regression problems. For instance, tail behaviors of Y are usually employed to study the robustness and the consistency properties of M-estimators in linear regression problems, see e.g., Hampel et al. (1986); Davies (1993); Audibert and Catoni (2011) and many others. In the statistical learning literature, some recent studies have also confined the tail properties of Y to explore the robustness of the kernel-based regression schemes, see e.g., Christmann and Steinwart (2007); Christmann and Messem (2008); Steinwart and Christmann (2008); De Brabanter et al. (2009); Debruyne et al. (2010).

Note also that in the statistical learning literature there are many existing studies on the robust regression problem. For instance, Suykens et al. (2002a,b) presented a weighted

least squares method to pursue a robust approximation to the regression function. Debruyne et al. (2008) addressed the model selection problem in kernel-based robust regression. Some efforts have been made in Steinwart and Christmann (2008) to understand generalization abilities of regression schemes associated with convex robust loss functions, e.g., Huber's loss, which are also conducted by restricting the tail behavior of Y. As shown in Steinwart and Christmann (2008), under certain conditions, empirical estimators learned from the ERM schemes associated with certain convex robust loss functions can generalize. However, this does not directly indicate the regression consistency property of the empirical estimators, e.g., the convergence from the empirical estimator to the regression function with respect to the $\mathcal{L}^2_{\rho_\mathcal{X}}$-distance. On the other hand, as far as we can see, few studies can be found in the statistical learning literature towards understanding regression schemes associated with nonconvex robust loss functions, which are frequently employed in robust statistics (see Huber, 1981; Hampel et al., 1986).

2.3.3 Comparisons with Related Studies

As mentioned earlier, our study is motivated by recent work towards understanding the minimum error entropy criterion in regression problems (see Hu et al., 2013). Observing that when being applied to regression problems, both of the two models aim at modeling an empirical estimator that approximates the regression function f_ρ. Therefore, we can give comparisons on the convergence rates of the two models. Under the same assumptions on the tail behavior of Y and the Complexity Assumption I, when $f_\rho \in \mathcal{H}$, the convergence rates established in Hu et al. (2013) are of the type $\mathcal{O}(m^{-2/(3+3p)})$, which are presented with respect to the variance of $\widetilde{f}_\mathbf{z}(X) - f_\rho(X)$ due to the mean insensitive property of the MEECR model. In addition, when Y is bounded, under the Complexity Assumption I, Hu et al. (2013) reported convergence rates of the type $\mathcal{O}(m^{-1/(1+p)})$. In view of the convergence rates reported in Theorem 4 and Theorem 5, we conclude that the convergence rates of the two regression models are comparable. This is a nice property of the MCCR model considering that it has a lower computational complexity.

3. Connections between MCCR and LSR

As aforementioned, it is not suggested to roughly treat the ℓ_σ loss as the least squares loss in regression problems even if σ is sufficiently large. This section is dedicated to explaining this issue and trying to explore the connections between the two different regression models: MCCR and LSR.

To this end, we first give some notations. For any measurable function $f : \mathcal{X} \to \mathcal{Y}$, the generalization error of f under the ℓ_σ loss and the least squares loss are defined, respectively, as

$$\mathcal{E}^\sigma(f) = \int_\mathcal{Z} \ell_\sigma(y, f(x)) d\rho(x, y), \text{ and } \mathcal{E}(f) = \int_\mathcal{Z} (y - f(x))^2 d\rho(x, y).$$

The corresponding target functions with respect to the hypothesis space $\mathcal{H}$ are given, respectively, by

$$f^\sigma_\mathcal{H} = \arg\min_{f\in\mathcal{H}} \mathcal{E}^\sigma(f), \text{ and } f_\mathcal{H} = \arg\min_{f\in\mathcal{H}} \mathcal{E}(f).$$

3.1 A Useful Lemma

We first give a lemma which bounds the deviation of the excess risks of f associated with the ℓ_σ loss and the least squares loss for any $f \in \mathcal{H}$. It will play an important role in our following analysis. In this context, the excess risk of f with respect to the ℓ_σ loss refers to the term $\mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f_\rho)$ while the excess risk of f with respect to the least squares loss refers to the term $\mathcal{E}(f) - \mathcal{E}(f_\rho)$.

Lemma 7 *Assume that the Moment Assumption holds. For any $f \in \mathcal{H}$, the deviation of the two excess risk terms can be bounded as follows*

$$\Big| \{\mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f_\rho)\} - \{\mathcal{E}(f) - \mathcal{E}(f_\rho)\} \Big| \leq \frac{c_{\mathcal{H},\rho}}{\sigma^2},$$

where $c_{\mathcal{H},\rho}$ is a positive constant given by

$$c_{\mathcal{H},\rho} = 8 \int_{\mathcal{Z}} y^4 d\rho + 4 \sup_{f \in \mathcal{H}} \|f\|_\infty^4 + 4\|f_\rho\|_\infty^4. \tag{3}$$

Proof Following the inequality $|1 - t - e^{-t}| \leq \frac{t^2}{2}$ for $t > 0$, one has

$$\left| 1 - \frac{(y - f(x))^2}{\sigma^2} - \exp\left\{ -\frac{(y - f(x))^2}{\sigma^2} \right\} \right| \leq \frac{(y - f(x))^4}{2\sigma^4}.$$

Simple computations show that

$$\left| \mathcal{E}^\sigma(f) - \int_{\mathcal{Z}} (y - f(x))^2 d\rho \right| \leq \frac{1}{2\sigma^2} \int_{\mathcal{Z}} (y - f(x))^4 d\rho. \tag{4}$$

Since $f_\rho \in L^\infty_{\rho_X}$, the same estimation process can be applied to f_ρ, which gives

$$\left| \mathcal{E}^\sigma(f_\rho) - \int_{\mathcal{Z}} (y - f_\rho(x))^2 d\rho \right| \leq \frac{1}{2\sigma^2} \int_{\mathcal{Z}} (y - f_\rho(x))^4 d\rho. \tag{5}$$

Combining estimates in (4) and (5), we come to the following inequality

$$\Big| \{\mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f_\rho)\} - \{\mathcal{E}(f) - \mathcal{E}(f_\rho)\} \Big| \leq \frac{1}{\sigma^2} \left(8 \int_{\mathcal{Z}} y^4 d\rho + 4\|f\|_\infty^4 + 4\|f_\rho\|_\infty^4 \right),$$

where the basic inequality $(a + b)^4 \leq 8a^4 + 8b^4$ for $a, b \in \mathbb{R}$ has been applied. By denoting

$$c_{\mathcal{H},\rho} = 8 \int_{\mathcal{Z}} y^4 d\rho + 4 \sup_{f \in \mathcal{H}} \|f\|_\infty^4 + 4\|f_\rho\|_\infty^4,$$

we complete the proof of Lemma 7. ■

3.2 An Equivalence Relation between MCCR and LSR

In this part, we proceed with exploring the connections between the two models: MCCR and LSR. We will show that, when σ is large enough, under certain conditions, there does exist an equivalence relation between the two regression models. By equivalence, we mean that the two regression models admit the same target function when working in the same hypothesis space, i.e., $f_{\mathcal{H}}^{\sigma} = f_{\mathcal{H}}$ in our study.

Theorem 8 *Suppose that the Noise Assumption holds. Under the condition that $f_\rho \in \mathcal{H}$ and $\sigma > \sigma_{\mathcal{H},\rho}$ with*

$$\sigma_{\mathcal{H},\rho} = \sqrt{2}\Big(M_0 + \|f_\rho\|_\infty + \sup_{f\in\mathcal{H}} \|f\|_\infty\Big),$$

almost surely we have

$$f_{\mathcal{H}}^{\sigma} = f_{\mathcal{H}}.$$

Proof Since $f_\rho \in \mathcal{H}$, it is immediate to see that almost surely we have $f_{\mathcal{H}} = f_\rho$. To finish the proof, it remains to show that there holds $f_{\mathcal{H}}^{\sigma} = f_\rho$. In fact, for any $f \in \mathcal{H}$, we know that

$$\mathcal{E}^{\sigma}(f) = \sigma^2 \int_{\mathcal{Z}} \left(1 - \exp\left\{-\frac{(y-f(x))^2}{\sigma^2}\right\}\right) d\rho(x,y) = \sigma^2 \int_{\mathcal{X}} F_x(f(x) - f_\rho(x)) d\rho_{\mathcal{X}}(x),$$

where

$$F_x(u) := 1 - \int_{-M_0}^{M_0} \exp\left\{-\frac{(t-u)^2}{\sigma^2}\right\} p_{\epsilon|X=x}(t)dt,\ x \in \mathcal{X}.$$

By taking the derivative of F with respect to u, we get

$$F_x'(u) = -2\int_{-M_0}^{M_0} \exp\left\{-\frac{(t-u)^2}{\sigma^2}\right\} \left(\frac{t-u}{\sigma^2}\right) p_{\epsilon|X=x}(t)dt,\ x \in \mathcal{X}.$$

According to the symmetry property of $p_{\epsilon|X=x}$, we know that $F_x'(0) = 0$. Moreover,

$$F_x''(u) = 2\int_{-M_0}^{M_0} \exp\left\{-\frac{(t-u)^2}{\sigma^2}\right\} \left(\frac{\sigma^2 - 2(t-u)^2}{\sigma^4}\right) p_{\epsilon|X=x}(t)dt,\ x \in \mathcal{X}.$$

Obviously, $F_x''(u) > 0$ for all $x \in \mathcal{X}$ when $\sigma > \sigma_{\mathcal{H},\rho}$. Consequently, $u = 0$ is the unique minimizer of $F_x(\cdot)$ for any $x \in \mathcal{X}$. The proof of Theorem 8 can be completed by recalling the definitions of $f_{\mathcal{H}}^{\sigma}$ and f_ρ. ■

Theorem 8 provides a situation where the equivalence relation between the two regression models holds. In the sense of Theorem 8, one can take the ℓ_σ loss as the least squares loss when σ is large enough. However, Theorem 8 also indicates that the equivalence relation holds when the Noise Assumption is valid, $f_\rho \in \mathcal{H}$ and σ is sufficiently large. Note that the condition $f_\rho \in \mathcal{H}$ imposes a regularity requirement on the regression function f_ρ while the

Noise Assumption asks for the boundedness and symmetry of the noise. In view of these, we conclude that one is not suggested to simply treat the ℓ_σ loss as the least squares loss even if σ is sufficiently large.

We remark that Theorem 8 merely provides a sufficient condition to ensure the existence of the equivalence relation between the two models. It would be meaningful to explore some other relaxed conditions to get a similar equivalence relation. However, we also remark that the non-convexity of the ℓ_σ loss makes it non-trivial since in this case there exists more than one local optimum of the MCCR model.

3.3 Comparisons on the Convergence Rates of MCCR and LSR

To further elucidate connections between the two regression models, in this part we move our attention to comparing the learning performance of their empirical estimators, i.e., the convergence rates of $\|f_{\mathbf{z}} - f_\rho\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}}$ and $\|f^{\rm ls}_{\mathbf{z}} - f_\rho\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}}$ where $f^{\rm ls}_{\mathbf{z}}$ is modeled by the following ERM scheme

$$f^{\rm ls}_{\mathbf{z}} = \arg\min_{f\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^m (f(x_i) - y_i)^2. \tag{6}$$

Noticing that due to the assumption that $\mathcal{H}$ is a compact subset of $C(\mathcal{X})$, (6) is in fact a constrained optimization model. When $\mathcal{H}$ is taken as a bounded subset of a certain reproducing kernel Hilbert space $\mathcal{H}_\mathcal{K}$, there exists an equivalence relation between the constrained optimization model (6) and the following unconstrained model

$$f^{\rm ls}_{\mathbf{z},\lambda} = \arg\min_{f\in\mathcal{H}_\mathcal{K}} \frac{1}{m}\sum_{i=1}^m (f(x_i) - y_i)^2 + \lambda\|f\|^2_\mathcal{K}, \tag{7}$$

where $\lambda > 0$ is a regularization parameter. Therefore, our comparison will be conducted between the MCCR model (2) and the regularized least squares regression model (7), which has been well understood in the statistical learning literature.

When Y is bounded, $f_\rho \in \mathcal{H}$ and the Complexity Assumption II with $0 < s < 2$ holds, the convergence rate of $\|f_{\mathbf{z}} - f_\rho\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}}$ established in Theorem 5 belongs to the type of $\mathcal{O}(m^{-2/(2+s)})$, which is the same as that of the regularized LSR (7) under the same conditions as revealed in Wu et al. (2006). In fact, when $\mathcal{H}$ is taken as a bounded subset of $\mathcal{H}_\mathcal{K}$ and the Mercer kernel $\mathcal{K}$ is sufficiently smooth, the constant s in the Complexity Assumption II can be arbitrarily small. As mentioned earlier, in this case, learning rates of the type $\mathcal{O}(m^{-1})$ can be derived which are regarded as the optimal learning rates in learning theory according to the law of large numbers.

On the other hand, due to the non-robustness of the least squares loss, almost all the existing convergence rates established for (7) are reported under the restriction that the response variable has a sub-Gaussian tail (see Wu et al., 2006; Caponnetto and De Vito, 2007; Steinwart et al., 2009; Mendelson and Neeman, 2010; Wang and Zhou, 2011). However, we see from Theorem 4 that for the MCCR model, convergence rates can be obtained under the Moment Assumption. This shows that the MCCR model can deal with non-Gaussian noise, which consequently distinguishes the two models in terms of conditions needed to establish meaningful convergence rates.

Before ending this section, let us briefly summarize the connections between MCCR and LSR as follows:

- For any given $f \in \mathcal{H}$, the difference between the excess risk of f with respect to the two regression models can be upper bounded by $\mathcal{O}(\sigma^{-2})$;

- Under certain conditions, we do see the existence of an equivalence relation between the two models, as commonly expected when σ is large enough. However, this equivalence relation might hold only under very specific conditions as suggested by our analysis;

- The MCCR model can deal with the heavy-tailed noise while the LSR model can only deal with sub-Gaussian noise. Moreover, when being restricted to cases with the bounded output or with the Gaussian noise, the performance of the two regression models are comparable. Therefore, in the above sense, we suggest that one can count on the MCCR model (2) to solve regression problems.

4. Deriving the Convergence Rates

This section presents detailed convergence analysis of the MCCR model (2) and proofs of theorems given in Section 2. The main difficulty in analyzing the model lies in the non-convexity of the loss function ℓ_σ, which disables usual techniques for analyzing convex learning models (see Cucker and Zhou, 2007; Steinwart and Christmann, 2008). We overcome this difficulty by introducing a novel error decomposition strategy with the help of Lemma 7. Analysis presented in this section is inspired by Cucker and Zhou (2007); Hu et al. (2013) and Fan et al..

4.1 Decomposing the Error into Bias-Variance Terms

The $\mathcal{L}^2_{\rho_\mathcal{X}}$-distance between the empirical target function $f_\mathbf{z}$ and the regression function f_ρ can be decomposed into the bias and the variance terms (see Vapnik, 1998; Cucker and Zhou, 2007; Steinwart and Christmann, 2008). Roughly speaking, the bias refers to the data-free error terms while the variance refers to the data-dependent error terms. The spirit of the learning theory approach to analyzing the convergence of learning models is trying to find a compromise between bias and variance by controlling the complexity of the hypothesis space. The following proposition offers a method for such compromise with respect to the MCCR model (2).

Proposition 9 *Assume that the Moment Assumption holds and let $f_\mathbf{z}$ be produced by (2). The $\mathcal{L}^2_{\rho_\mathcal{X}}$-distance between $f_\mathbf{z}$ and f_ρ can be decomposed as follows:*

$$\|f_\mathbf{z} - f_\rho\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}} \leq \mathcal{A}_{\mathcal{H},\sigma,\rho} + \mathcal{A}_{\mathcal{H},\rho} + \mathcal{S}_1(\mathbf{z}) + \mathcal{S}_2(\mathbf{z}),$$

where

$$\begin{aligned}
\mathcal{A}_{\mathcal{H},\sigma,\rho} &= 2c_{\mathcal{H},\rho}/\sigma^2, \\
\mathcal{A}_{\mathcal{H},\rho} &= \mathcal{E}(f_{\mathcal{H}}) - \mathcal{E}(f_\rho), \\
\mathcal{S}_1(\mathbf{z}) &= \{\mathcal{E}^\sigma_{\mathbf{z}}(f^\sigma_{\mathcal{H}}) - \mathcal{E}^\sigma_{\mathbf{z}}(f_\rho)\} - \{\mathcal{E}^\sigma(f^\sigma_{\mathcal{H}}) - \mathcal{E}^\sigma(f_\rho)\}, \\
\mathcal{S}_2(\mathbf{z}) &= \{\mathcal{E}^\sigma(f_{\mathbf{z}}) - \mathcal{E}^\sigma(f_\rho)\} - \{\mathcal{E}^\sigma_{\mathbf{z}}(f_{\mathbf{z}}) - \mathcal{E}^\sigma_{\mathbf{z}}(f_\rho)\}.
\end{aligned}$$

Proof Following from Lemma 7, with simple computations, we see that

$$\begin{aligned}
\|f_{\mathbf{z}} - f_\rho\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}} &\le \mathcal{E}^\sigma(f_{\mathbf{z}}) - \mathcal{E}^\sigma(f_\rho) + c_{\mathcal{H},\rho}/\sigma^2 \\
&\le \{\mathcal{E}^\sigma(f_{\mathbf{z}}) - \mathcal{E}^\sigma_{\mathbf{z}}(f_{\mathbf{z}})\} + \{\mathcal{E}^\sigma_{\mathbf{z}}(f_{\mathbf{z}}) - \mathcal{E}^\sigma_{\mathbf{z}}(f^\sigma_{\mathcal{H}})\} + \{\mathcal{E}^\sigma_{\mathbf{z}}(f^\sigma_{\mathcal{H}}) - \mathcal{E}^\sigma(f^\sigma_{\mathcal{H}})\}. \\
&\quad + \{\mathcal{E}^\sigma(f^\sigma_{\mathcal{H}}) - \mathcal{E}^\sigma(f_{\mathcal{H}})\} + \{\mathcal{E}^\sigma(f_{\mathcal{H}}) - \mathcal{E}^\sigma(f_\rho)\} + c_{\mathcal{H},\rho}/\sigma^2 \\
&\le \{\mathcal{E}^\sigma(f_{\mathbf{z}}) - \mathcal{E}^\sigma_{\mathbf{z}}(f_{\mathbf{z}})\} + \{\mathcal{E}^\sigma_{\mathbf{z}}(f_{\mathbf{z}}) - \mathcal{E}^\sigma_{\mathbf{z}}(f^\sigma_{\mathcal{H}})\} + \{\mathcal{E}^\sigma_{\mathbf{z}}(f^\sigma_{\mathcal{H}}) - \mathcal{E}^\sigma(f^\sigma_{\mathcal{H}})\}. \\
&\quad + \{\mathcal{E}^\sigma(f^\sigma_{\mathcal{H}}) - \mathcal{E}^\sigma(f_{\mathcal{H}})\} + \{\mathcal{E}(f_{\mathcal{H}}) - \mathcal{E}(f_\rho)\} + 2c_{\mathcal{H},\rho}/\sigma^2.
\end{aligned}$$

The definitions of $f_{\mathbf{z}}$, $f^\sigma_{\mathcal{H}}$ and $f_{\mathcal{H}}$ tell us that the second and the fourth terms of right-hand side of the last inequality are at most zero. By introducing intermediate terms $\mathcal{E}^\sigma_{\mathbf{z}}(f_\rho)$, $\mathcal{E}^\sigma(f_\rho)$ and corresponding notations, we finish the proof of Proposition 9. ■

As shown in Proposition 9, the $\mathcal{L}^2_{\rho_\mathcal{X}}$-distance between $f_{\mathbf{z}}$ and f_ρ are decomposed into four error terms: $\mathcal{A}_{\mathcal{H},\sigma,\rho}$, $\mathcal{A}_{\mathcal{H},\rho}$, $\mathcal{S}_1(\mathbf{z})$, and $\mathcal{S}_2(\mathbf{z})$. It is easy to see that the first two error terms are data-independent and correspond to the bias while the last two terms are data-dependent, which consequently are referred as the sample error (variance). The quantity $\mathcal{A}_{\mathcal{H},\rho}$ can be translated as the approximation ability of $f_{\mathcal{H}}$ to f_ρ, the estimation of which belongs to the topics of the approximation theory and has been well conducted. For instance, when $\mathcal{H}$ is chosen as a bounded subset of a certain reproducing kernel Hilbert space (RKHS), a comprehensive study on this term can be found in Smale and Zhou (2003). On the other hand, we remind that the bias term $\mathcal{A}_{\mathcal{H},\sigma,\rho}$ is introduced into the above error decomposition method, which not only depends on the hypothesis space $\mathcal{H}$ and the underlying probability distribution ρ, but also relies on the scale parameter σ. As explained later, this is caused by the introduction of the robustness into the regression model. This makes the decomposition strategy for the MCCR model different from those for convex regression models (see Cucker and Zhou, 2007; Steinwart and Christmann, 2008).

As a consequence of Proposition 9, to bound $\|f_{\mathbf{z}} - f_\rho\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}}$, it suffices to estimate the two sample error terms: $\mathcal{S}_1(\mathbf{z})$ and $\mathcal{S}_2(\mathbf{z})$, which will be tackled in the next subsection.

4.2 Concentration Estimates of Sample Error Terms

This part presents concentration estimates for the sample error terms $\mathcal{S}_1(\mathbf{z})$ and $\mathcal{S}_2(\mathbf{z})$ when the Moment Assumption is assumed. In learning theory, this is typically done by applying concentration inequalities to certain random variables that may be function-space valued.

In our study, for this purpose we introduce the following two random variables, $\xi_1(z)$ and $\xi_2(z)$ with $z \in \mathcal{Z}$, which are defined by

$$\xi_1(z) := -\sigma^2 \exp\left\{-(y - f^\sigma_{\mathcal{H}}(x))^2/\sigma^2\right\} + \sigma^2 \exp\left\{-(y - f_\rho(x))^2/\sigma^2\right\},$$

and

$$\xi_2(z) := -\sigma^2 \exp\left\{-(y - f_{\mathbf{z}}(x))^2/\sigma^2\right\} + \sigma^2 \exp\left\{-(y - f_\rho(x))^2/\sigma^2\right\}.$$

By applying the one-sided Bernstein's inequality in Lemma 12 to the random variable ξ_1, we can get the concentrated estimate for the sample error term $\mathcal{S}_1(\mathbf{z})$. However, the estimation of the sample error term $\mathcal{S}_2(\mathbf{z})$ requires us to apply concentration inequalities to the function-space valued random variable ξ_2 and consequently depends on the capacity of the hypothesis space $\mathcal{H}$. This is due to the fact that the random variable ξ_2 is dependent with $f_{\mathbf{z}}$ which varies in accordance with the sample $\mathbf{z}$.

Concentrated estimates for $\mathcal{S}_1(\mathbf{z})$ and $\mathcal{S}_2(\mathbf{z})$ are presented in the following two propositions, the proofs of which are given in Subsection 4.3.

Proposition 10 *Assume that the Moment Assumption holds. For any $0 < \delta < 1$, with confidence $1 - \delta/2$, there holds*

$$\mathcal{S}_1(\mathbf{z}) \leq \frac{1}{2}\left\|f_{\mathcal{H}} - f_\rho\right\|^2_{\mathcal{L}^2_{\rho_{\mathcal{X}}}} + C_{\mathcal{H},\rho,1}\left(\log\frac{2}{\delta}\right)\left(\frac{\sigma}{m} + \frac{1}{\sigma^2}\right),$$

where $C_{\mathcal{H},\rho,1}$ is a positive constant independent of m, σ or δ and will be given explicitly in the proof.

Proposition 11 *Assume that the Complexity Assumption I with $p > 0$ and the Moment Assumption hold. For any $0 < \delta < 1$, with confidence $1 - \delta/2$, there holds*

$$\mathcal{S}_2(\mathbf{z}) \leq \frac{1}{2}(\mathcal{S}_1(\mathbf{z}) + \mathcal{S}_2(\mathbf{z})) + \frac{1}{2}\left\|f_{\mathcal{H}} - f_\rho\right\|^2_{\mathcal{L}^2_{\rho_{\mathcal{X}}}} + C_{\mathcal{H},\rho,2}\left(\log\frac{2}{\delta}\right)\left\{\frac{1}{\sigma^2} + \frac{\sigma}{m^{\frac{1}{1+p}}}\right\},$$

where $C_{\mathcal{H},\rho,2}$ is a positive constant independent of m, σ or δ and will be given explicitly in the proof.

4.3 Proofs

4.3.1 Lemmas

We first list several lemmas that will be used in the proofs. Lemma 12 and Lemma 13 are one-sided Bernstein's concentration inequalities, which were introduced in Bernstein (1946) and can be found in many statistical learning textbooks, see e.g., Cucker and Zhou (2007); Steinwart and Christmann (2008). Lemma 14 was proved in Wu et al. (2007).

Lemma 12 *Let ξ be a random variable on a probability space $\mathcal{Z}$ with variance $\sigma^2_\star$ satisfying $|\xi - \mathbb{E}\xi| \leq M_\xi$ almost surely for some constant M_ξ and for all $z \in \mathcal{Z}$. Then*

$$Prob_{z\in\mathcal{Z}^m}\left\{\frac{1}{m}\sum_{i=1}^m \xi(z_i) - \mathbb{E}\xi \geq \varepsilon\right\} \leq \exp\left\{-\frac{m\varepsilon^2}{2(\sigma^2_\star + \frac{1}{3}M_\xi\varepsilon)}\right\}.$$

Lemma 13 *Let ξ be a random variable on a probability space $\mathcal{Z}$ with variance $\sigma_\star^2$ satisfying $|\xi - \mathbb{E}\xi| \leq M_\xi$ almost surely for some constant M_ξ and for all $z \in \mathcal{Z}$. Then for any $0 < \delta < 1$, with confidence $1-\delta$, we have*

$$\frac{1}{m}\sum_{i=1}^{m}\xi(z_i) - \mathbb{E}\xi \leq \frac{2M_\xi \log\frac{1}{\delta}}{3m} + \sqrt{\frac{2\sigma_\star^2 \log\frac{1}{\delta}}{m}}.$$

Lemma 14 *Let $\mathcal{F}$ be a class of measurable functions on $\mathcal{Z}$. Assume that there are constants $B, c > 0$ and $\theta \in [0,1]$ such that $\|f\|_\infty \leq B$ and $\mathbb{E}f^2 \leq c(\mathbb{E}f)^\theta$ for every $f \in \mathcal{F}$. If for some $a > 0$ and $s \in (0,2)$,*

$$\log \mathcal{N}_2(\mathcal{F}, \eta) \leq a\eta^{-s}, \qquad \forall\, \eta > 0,$$

then there exists a constant α_p depending only on p such that for any $t > 0$, with probability at least $1 - e^{-t}$, there holds

$$\mathbb{E}f - \frac{1}{m}\sum_{i=1}^{m} f(z_i) \leq \frac{1}{2}\gamma^{1-\theta}(\mathbb{E}f)^\theta + \alpha_p\gamma + 2\left(\frac{ct}{m}\right)^{\frac{1}{2-\theta}} + \frac{18Bt}{m}, \qquad \forall\, f \in \mathcal{F},$$

where

$$\gamma := \max\left\{c^{\frac{2-s}{4-2\theta+s\theta}}\left(\frac{a}{m}\right)^{\frac{2}{4-2\theta+s\theta}}, B^{\frac{2-s}{2+s}}\left(\frac{a}{m}\right)^{\frac{2}{2+s}}\right\}.$$

4.3.2 Proof of Proposition 10

Proof To bound the sample error term $\mathcal{S}_1(\mathbf{z})$, we apply the one-sided Bernstein's inequality in Lemma 13 to the random variable ξ_1 introduced in Subsection 4.2. To this end, we need to verify conditions in Lemma 13.

We first verify the boundedness condition. Recall that the random variable ξ_1 is defined as

$$\xi_1(z) := -\sigma^2 \exp\left\{-(y - f_{\mathcal{H}}^\sigma(x))^2/\sigma^2\right\} + \sigma^2 \exp\left\{-(y - f_\rho(x))^2/\sigma^2\right\},\ z \in \mathcal{Z}.$$

Introducing the auxiliary function $h(t) = \exp\{-t^2\}$ with $t \in \mathbb{R}$, it is easy to see that $\|h'\|_\infty = \sqrt{2/e}$. By taking $t_1 = (y - f_{\mathcal{H}}^\sigma(x))/\sigma$, $t_2 = (y - f_\rho(x))/\sigma$ and applying the mean value theorem to h, we see that

$$|\xi_1(z)| \leq \sqrt{2/e}\sigma|f_{\mathcal{H}}^\sigma(x) - f_\rho(x)| \leq \sqrt{2/e}\sigma\|f_{\mathcal{H}}^\sigma - f_\rho\|_\infty,\ z \in \mathcal{Z}.$$

Consequently,

$$|\xi_1 - \mathbb{E}\xi_1| \leq 2\|\xi_1\|_\infty \leq 2\sqrt{2/e}\sigma\|f_{\mathcal{H}}^\sigma - f_\rho\|_\infty \leq 2\sqrt{2/e}\sigma \sup_{f\in\mathcal{H}}\|f - f_\rho\|_\infty.$$

We are now in a position to bound the variance of the random variable ξ_1, which is denoted as $\mathrm{var}(\xi_1)$. Applying the mean value theorem to the auxiliary function $h_1(t) =$

$\exp(-t)$ at $t_1 = (y - f_{\mathcal{H}}^{\sigma}(x))^2/\sigma^2$, $t_2 = (y - f_\rho(x))^2/\sigma^2$ and recalling that $\|h_1'\|_\infty \leq 1$, we get

$$\begin{aligned}
\mathrm{var}(\xi_1) &= \mathbb{E}\xi_1^2 - (\mathbb{E}\xi_1)^2 \leq \mathbb{E}\xi_1^2 \\
&\leq \mathbb{E}\left((f_{\mathcal{H}}^{\sigma}(x) - f_\rho(x))^2 (2y - f_{\mathcal{H}}^{\sigma}(x) - f_\rho(x))^2\right) \\
&\leq \int_{\mathcal{Y}} \Big(12y^2 + 3\sup_{f\in\mathcal{H}} \|f\|_\infty^2 + 3\|f_\rho\|_\infty^2\Big) d\rho(y|x) \int_{\mathcal{X}} (f_{\mathcal{H}}^{\sigma}(x) - f_\rho(x))^2 d\rho_{\mathcal{X}}(x) \\
&= c_{\mathcal{H},\rho,0} \int_{\mathcal{X}} (f_{\mathcal{H}}^{\sigma}(x) - f_\rho(x))^2 d\rho_{\mathcal{X}}(x),
\end{aligned}$$

where the second inequality is from the elementary inequality $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$ for $a, b, c \in \mathbb{R}$ and the positive constant $c_{\mathcal{H},\rho,0}$ is denoted as

$$c_{\mathcal{H},\rho,0} = 12 \int_{\mathcal{Z}} y^2 d\rho + 3 \sup_{f\in\mathcal{H}} \|f\|_\infty^2 + 3\|f_\rho\|_\infty^2. \tag{8}$$

Now applying Lemma 13 to the random variable ξ_1, we see that for any $0 < \delta < 1$, with confidence $1 - \delta/2$, there holds

$$\mathcal{S}_1(\mathbf{z}) \leq \frac{4\sqrt{2/e}\sup_{f\in\mathcal{H}} \|f - f_\rho\|_\infty}{3} \frac{\sigma \log(2/\delta)}{m} + \sqrt{\frac{2c_{\mathcal{H},\rho,0} \log(2/\delta) \|f_{\mathcal{H}}^{\sigma} - f_\rho\|_{\mathcal{L}^2_{\rho_{\mathcal{X}}}}^2}{m}}. \tag{9}$$

The elementary inequality $\sqrt{ab} \leq (a+b)/2$ for $a, b \geq 0$ gives[1]

$$\sqrt{\frac{2c_{\mathcal{H},\rho,0} \log(2/\delta) \|f_{\mathcal{H}}^{\sigma} - f_\rho\|_{\mathcal{L}^2_{\rho_{\mathcal{X}}}}^2}{m}} \leq \frac{1}{2}\|f_{\mathcal{H}}^{\sigma} - f_\rho\|_{\mathcal{L}^2_{\rho_{\mathcal{X}}}}^2 + \frac{c_{\mathcal{H},\rho,0} \log(2/\delta)}{m}. \tag{10}$$

In addition, as a consequence of Lemma 7, we have

$$\begin{aligned}
\|f_{\mathcal{H}}^{\sigma} - f_\rho\|_{\mathcal{L}^2_{\rho_{\mathcal{X}}}}^2 &\leq \mathcal{E}^{\sigma}(f_{\mathcal{H}}^{\sigma}) - \mathcal{E}^{\sigma}(f_\rho) + c_{\mathcal{H},\rho}/\sigma^2 \\
&= \mathcal{E}^{\sigma}(f_{\mathcal{H}}^{\sigma}) - \mathcal{E}^{\sigma}(f_{\mathcal{H}}) + \mathcal{E}^{\sigma}(f_{\mathcal{H}}) - \mathcal{E}^{\sigma}(f_\rho) + c_{\mathcal{H},\rho}/\sigma^2 \\
&\leq \|f_{\mathcal{H}} - f_\rho\|_{\mathcal{L}^2_{\rho_{\mathcal{X}}}}^2 + 2c_{\mathcal{H},\rho}/\sigma^2,
\end{aligned} \tag{11}$$

where the last inequality is due to the fact that $f_{\mathcal{H}}^{\sigma}$ is the minimizer of the risk functional $\mathcal{E}^{\sigma}(\cdot)$ in $\mathcal{H}$.

Combining estimates in (9), (10), and (11), we come to the conclusion that for any $0 < \delta < 1$, with confidence $1 - \delta/2$, there holds

$$\mathcal{S}_1(\mathbf{z}) \leq \frac{1}{2}\|f_{\mathcal{H}} - f_\rho\|_{\mathcal{L}^2_{\rho_{\mathcal{X}}}}^2 + C_{\mathcal{H},\rho,1} \left(\log\frac{2}{\delta}\right)\left(\frac{\sigma}{m} + \frac{1}{\sigma^2}\right),$$

where $C_{\mathcal{H},\rho,1}$ is a positive constant independent of m, σ or δ and given by

$$C_{\mathcal{H},\rho,1} = (4/3)\sqrt{2/e}\sup_{f\in\mathcal{H}} \|f - f_\rho\|_\infty + 2c_{\mathcal{H},\rho} + c_{\mathcal{H},\rho,0}.$$

Thus we have completed the proof of Proposition 10. ■

1. Refined estimate can be derived here by applying Young's inequality $ab \leq \frac{ta^2}{2} + \frac{b^2}{2t}$ for $a, b \in \mathbb{R}$, $t > 0$. In our proof, we choose $t = 1$ for simplification.

4.3.3 Proof of Proposition 11

To prove Proposition 11, we first need to prove the following intermediate conclusion, which is in fact a concentrated estimate for function-space valued random variables.

Proposition 15 *Assume that the Moment Assumption holds. Let ε satisfy $\varepsilon \geq c_{\mathcal{H},\rho}/\sigma^2$. For any $0 < \delta < 1$, with confidence $1 - \delta/2$, there holds*

$$\begin{aligned}&\mathrm{Prob}_{\mathbf{z}\in\mathcal{Z}^m}\left\{\sup_{f\in\mathcal{H}}\frac{(\mathcal{E}^\sigma(f)-\mathcal{E}^\sigma(f_\rho))-(\mathcal{E}^\sigma_{\mathbf{z}}(f)-\mathcal{E}^\sigma_{\mathbf{z}}(f_\rho))}{\sqrt{\mathcal{E}^\sigma(f)-\mathcal{E}^\sigma(f_\rho)+2\varepsilon}}>4\sqrt{\varepsilon}\right\}\\&\leq\mathcal{N}\left(\mathcal{H},\frac{\varepsilon}{\sqrt{2/e}\sigma}\right)\exp\left\{-\frac{3m\varepsilon}{4\sqrt{2/e}\sup_{f\in\mathcal{H}}\|f-f_\rho\|_\infty\sigma+6c_{\mathcal{H},\rho,0}}\right\},\end{aligned}$$

where $c_{\mathcal{H},\rho}$ is given in (3) *and $c_{\mathcal{H},\rho,0}$ is given in* (8)*, both of which are positive constants independent of m, σ or δ.*

Proof To derive the desired estimate, we will apply the one-sided Bernstein's inequality in Lemma 13 to the function set $\mathcal{H}$ by taking its capacity into account.

For any $f \in \mathcal{H}$, we redefine the random variable $\xi_2(z)$ as follows

$$\xi_2(z) = -\sigma^2\exp\left\{-(y-f(x))^2/\sigma^2\right\}+\sigma^2\exp\left\{-(y-f_\rho(x))^2/\sigma^2\right\},\ z\in\mathcal{Z}.$$

Following from the proof of Proposition 10, we know that

$$\|\xi_2\|_\infty \leq \sqrt{2/e}\sigma\|f-f_\rho\|_\infty \text{ and } |\xi_2-\mathbb{E}\xi_2|\leq 2\sqrt{2/e}\sigma\sup_{f\in\mathcal{H}}\|f-f_\rho\|_\infty.$$

Meanwhile, we also know from the proof of Proposition 10 that

$$\mathbb{E}\xi_2^2 \leq c_{\mathcal{H},\rho,0}\|f-f_\rho\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}},$$

where the constant $c_{\mathcal{H},\rho,0}$ is given in (8).

Consider a function set $\{f_j\}_{j=1}^J \subset \mathcal{H}$ with $J = \mathcal{N}(\mathcal{H}, \varepsilon/(\sqrt{2/e}\sigma))$. The compactness of $\mathcal{H}$ ensures the existence and finiteness of J. Now we let

$$\mu = \sqrt{\mathcal{E}^\sigma(f_j)-\mathcal{E}^\sigma(f_\rho)+2\varepsilon},$$

and choose ε such that $\varepsilon \geq c_{\mathcal{H},\rho}/\sigma^2$. Applying the one-sided Bernstein's inequality in Lemma 12 to the following group of random variables

$$\xi_{2,j}(z) = -\sigma^2\exp\left\{-(y-f_j(x))^2/\sigma^2\right\}+\sigma^2\exp\left\{-(y-f_\rho(x))^2/\sigma^2\right\},\ j=1,\ldots,J,$$

we come to the following conclusion

$$\begin{aligned}
&\mathrm{Prob}_{\mathbf{z}\in\mathcal{Z}^m}\left\{\frac{(\mathcal{E}^\sigma(f_j)-\mathcal{E}^\sigma(f_\rho))-(\mathcal{E}^\sigma_{\mathbf{z}}(f_j)-\mathcal{E}^\sigma_{\mathbf{z}}(f_\rho))}{\sqrt{\mathcal{E}^\sigma(f_j)-\mathcal{E}^\sigma(f_\rho)+2\varepsilon}}>\sqrt{\varepsilon}\right\}\\
&\quad\le\exp\left\{-\frac{3m\varepsilon\mu^2}{4\sqrt{2/e}\|f_j-f_\rho\|_\infty\sqrt{\varepsilon}\mu\sigma+6c_{\mathcal{H},\rho,0}\|f_j-f_\rho\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}}}\right\}\\
&\quad\le\exp\left\{-\frac{3m\varepsilon\mu^2}{4\sqrt{2/e}\|f_j-f_\rho\|_\infty\sqrt{\varepsilon}\mu\sigma+6c_{\mathcal{H},\rho,0}\mu^2}\right\}\\
&\quad\le\exp\left\{-\frac{3m\varepsilon}{4\sqrt{2/e}\sup_{f\in\mathcal{H}}\|f-f_\rho\|_\infty\sigma+6c_{\mathcal{H},\rho,0}}\right\},
\end{aligned}$$

where the last two inequalities follow from the inequality in Lemma 7, the equation that $\mathcal{E}(f_j)-\mathcal{E}(f_\rho)=\|f_j-f_\rho\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}}$, the fact that $\varepsilon\ge c_{\mathcal{H},\rho}/\sigma^2$ and

$$\mu^2=\mathcal{E}^\sigma(f_j)-\mathcal{E}^\sigma(f_\rho)+2\varepsilon\ge\mathcal{E}^\sigma(f_j)-\mathcal{E}^\sigma(f_\rho)+c_{\mathcal{H},\rho}/\sigma^2+\varepsilon\ge\mathcal{E}(f_j)-\mathcal{E}(f_\rho)+\varepsilon\ge\varepsilon.$$

From the choice of f_j, we know that for each $f\in\mathcal{H}$, there exists some j such that $\|f-f_j\|_\infty\le\varepsilon/(\sqrt{2/e}\sigma)$. Therefore $|\mathcal{E}^\sigma(f)-\mathcal{E}^\sigma(f_j)|$ and $|\mathcal{E}^\sigma_{\mathbf{z}}(f)-\mathcal{E}^\sigma_{\mathbf{z}}(f_j)|$ can be both upper bounded by ε, which yields

$$\frac{|(\mathcal{E}^\sigma_{\mathbf{z}}(f)-\mathcal{E}^\sigma_{\mathbf{z}}(f_\rho))-(\mathcal{E}^\sigma_{\mathbf{z}}(f_j)-\mathcal{E}^\sigma_{\mathbf{z}}(f_\rho))|}{\sqrt{\mathcal{E}^\sigma(f)-\mathcal{E}^\sigma(f_\rho)+2\varepsilon}}\le\sqrt{\varepsilon}\tag{12}$$

and

$$\frac{|(\mathcal{E}^\sigma(f)-\mathcal{E}^\sigma(f_\rho))-(\mathcal{E}^\sigma(f_j)-\mathcal{E}^\sigma(f_\rho))|}{\sqrt{\mathcal{E}^\sigma(f)-\mathcal{E}^\sigma(f_\rho)+2\varepsilon}}\le\sqrt{\varepsilon}.\tag{13}$$

The latter inequality together with the fact that $\varepsilon\le\mathcal{E}^\sigma(f)-\mathcal{E}^\sigma(f_\rho)+2\varepsilon$ implies

$$\begin{aligned}
\mathcal{E}^\sigma(f_j)-\mathcal{E}^\sigma(f_\rho)+2\varepsilon&=(\mathcal{E}^\sigma(f_j)-\mathcal{E}^\sigma(f_\rho))-(\mathcal{E}^\sigma(f)-\mathcal{E}^\sigma(f_\rho))+\mathcal{E}^\sigma(f)-\mathcal{E}^\sigma(f_\rho)+2\varepsilon\\
&\le\sqrt{\varepsilon}\sqrt{(\mathcal{E}^\sigma(f)-\mathcal{E}^\sigma(f_\rho))+2\varepsilon}+\mathcal{E}^\sigma(f)-\mathcal{E}^\sigma(f_\rho)+2\varepsilon\\
&\le 2(\mathcal{E}^\sigma(f)-\mathcal{E}^\sigma(f_\rho)+2\varepsilon).
\end{aligned}\tag{14}$$

For any $f\in\mathcal{H}$, if the following inequality holds

$$\frac{(\mathcal{E}^\sigma(f)-\mathcal{E}^\sigma(f_\rho))-(\mathcal{E}^\sigma_{\mathbf{z}}(f)-\mathcal{E}^\sigma_{\mathbf{z}}(f_\rho))}{\sqrt{\mathcal{E}^\sigma(f)-\mathcal{E}^\sigma(f_\rho)+2\varepsilon}}>4\sqrt{\varepsilon},$$

then combining estimates in (12) and (13) we know that there holds

$$\frac{(\mathcal{E}^\sigma(f_j)-\mathcal{E}^\sigma(f_\rho))-(\mathcal{E}^\sigma_{\mathbf{z}}(f_j)-\mathcal{E}^\sigma_{\mathbf{z}}(f_\rho))}{\sqrt{\mathcal{E}^\sigma(f)-\mathcal{E}^\sigma(f_\rho)+2\varepsilon}}>2\sqrt{\varepsilon}.$$

This together with inequality (14) tells us that the following inequality holds

$$\frac{(\mathcal{E}^\sigma(f_j)-\mathcal{E}^\sigma(f_\rho))-(\mathcal{E}^\sigma_{\mathbf{z}}(f_j)-\mathcal{E}^\sigma_{\mathbf{z}}(f_\rho))}{\sqrt{\mathcal{E}^\sigma(f_j)-\mathcal{E}^\sigma(f_\rho)+2\varepsilon}}>\sqrt{\varepsilon}.$$

Consequently, based on the above estimates, we come to the following conclusion

$$\begin{aligned}&\mathrm{Prob}_{\mathbf{z}\in\mathcal{Z}^m}\left\{\sup_{f\in\mathcal{H}}\frac{(\mathcal{E}^\sigma(f)-\mathcal{E}^\sigma(f_\rho))-(\mathcal{E}^\sigma_{\mathbf{z}}(f)-\mathcal{E}^\sigma_{\mathbf{z}}(f_\rho))}{\sqrt{\mathcal{E}^\sigma(f)-\mathcal{E}^\sigma(f_\rho)+2\varepsilon}}>4\sqrt{\varepsilon}\right\}\\ &\le\sum_{j=1}^{J}\mathrm{Prob}_{\mathbf{z}\in\mathcal{Z}^m}\left\{\frac{(\mathcal{E}^\sigma(f_j)-\mathcal{E}^\sigma(f_\rho))-(\mathcal{E}^\sigma_{\mathbf{z}}(f_j)-\mathcal{E}^\sigma_{\mathbf{z}}(f_\rho))}{\sqrt{\mathcal{E}^\sigma(f_j)-\mathcal{E}^\sigma(f_\rho)+2\varepsilon}}>\sqrt{\varepsilon}\right\}\\ &\le\mathcal{N}\left(\mathcal{H},\frac{\varepsilon}{\sqrt{2/e}\sigma}\right)\exp\left\{-\frac{3m\varepsilon}{4\sqrt{2/e}\sup_{f\in\mathcal{H}}\|f-f_\rho\|_\infty\sigma+6c_{\mathcal{H},\rho,0}}\right\}.\end{aligned}$$

This completes the proof of Proposition 15. ■

Proof [Proof of Proposition 11] From the Complexity Assumption I, we know that

$$\mathcal{N}\left(\mathcal{H},\varepsilon/(\sqrt{2/e}\sigma)\right)\le\exp\left\{c_{I,p}(\sqrt{2/e})^p\sigma^p/\varepsilon^p\right\}.$$

This in connection with Proposition 15 yields

$$\begin{aligned}&\mathrm{Prob}_{\mathbf{z}\in\mathcal{Z}^m}\left\{\sup_{f\in\mathcal{H}}\frac{(\mathcal{E}^\sigma(f)-\mathcal{E}^\sigma(f_\rho))-(\mathcal{E}^\sigma_{\mathbf{z}}(f)-\mathcal{E}^\sigma_{\mathbf{z}}(f_\rho))}{\sqrt{\mathcal{E}^\sigma(f)-\mathcal{E}^\sigma(f_\rho)+2\varepsilon}}>4\sqrt{\varepsilon}\right\}\\ &\le\exp\left\{\frac{A_p\sigma^p}{\varepsilon^p}-\frac{m\varepsilon}{\sigma B_{\mathcal{H},\rho}+2c_{\mathcal{H},\rho,0}}\right\},\end{aligned}$$

where A_p and $B_{\mathcal{H},\rho}$ are positive constants given by

$$A_p=c_{I,p}(\sqrt{2/e})^p\ \text{and}\ \ B_{\mathcal{H},\rho}=4\sqrt{2/e}\sup_{f\in\mathcal{H}}\|f-f_\rho\|_\infty/3.$$

By setting

$$\exp\left\{\frac{A_p\sigma^p}{\varepsilon^p}-\frac{m\varepsilon}{\sigma B_{\mathcal{H},\rho}+2c_{\mathcal{H},\rho,0}}\right\}\le\frac{\delta}{2},$$

we obtain

$$\varepsilon^{p+1}-\frac{\log(2/\delta)\left(\sigma B_{\mathcal{H},\rho}+2c_{\mathcal{H},\rho,0}\right)}{m}\varepsilon^p-\frac{A_p\left(\sigma B_{\mathcal{H},\rho}+2c_{\mathcal{H},\rho,0}\right)\sigma^p}{m}\ge 0.$$

Lemma 7.2 in Cucker and Zhou (2007) tells us that the above inequality holds if

$$\varepsilon\ge\max\left\{\frac{c_{\mathcal{H},\rho}}{\sigma^2},\frac{2\log(2/\delta)\left(\sigma B_{\mathcal{H},\rho}+2c_{\mathcal{H},\rho,0}\right)}{m},\left(\frac{2A_p\left(\sigma B_{\mathcal{H},\rho}+2c_{\mathcal{H},\rho,0}\right)\sigma^p}{m}\right)^{1/(1+p)}\right\}.$$

In view of the above condition, we choose a sufficient large $\varepsilon_{\mathcal{H},\rho}$ as follows

$$\varepsilon_{\mathcal{H},\rho} = c_{\mathcal{H},\rho,1} \log(2/\delta)(\sigma^{-2} + \sigma m^{-1/(1+p)}),$$

where $c_{\mathcal{H},\rho,1}$ is a positive constant independent of m, σ or δ and given by

$$c_{\mathcal{H},\rho,1} = 2c_{\mathcal{H},\rho} + 2(A_p + 1)(B_{\mathcal{H},\rho} + 2c_{\mathcal{H},\rho,0}).$$

With the above choice of $\varepsilon_{\mathcal{H},\rho}$ and following the above discussions, we see that for any $0 < \delta < 1$, with confidence $1 - \delta/2$, there holds

$$\sup_{f\in\mathcal{H}} \left\{ ((\mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f_\rho)) - (\mathcal{E}^\sigma_{\mathbf{z}}(f) - \mathcal{E}^\sigma_{\mathbf{z}}(f_\rho))) \Big/ \sqrt{\mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f_\rho) + \varepsilon_{\mathcal{H},\rho}} \right\} \leq 4\sqrt{\varepsilon_{\mathcal{H},\rho}},$$

which yields

$$(\mathcal{E}^\sigma(f_{\mathbf{z}}) - \mathcal{E}^\sigma(f_\rho)) - (\mathcal{E}^\sigma_{\mathbf{z}}(f_{\mathbf{z}}) - \mathcal{E}^\sigma_{\mathbf{z}}(f_\rho)) \leq 4\sqrt{\varepsilon_{\mathcal{H},\rho}} \sqrt{\mathcal{E}^\sigma(f_{\mathbf{z}}) - \mathcal{E}^\sigma(f_\rho) + 2\varepsilon_{\mathcal{H},\rho}}.$$

Applying the basic inequality $\sqrt{ab} \leq (a+b)/2$ for $a, b \geq 0$, we know that for any $0 < \delta < 1$, with confidence $1 - \delta$, there holds[2]

$$\mathcal{S}_2(\mathbf{z}) = (\mathcal{E}^\sigma(f_{\mathbf{z}}) - \mathcal{E}^\sigma(f_\rho)) - (\mathcal{E}^\sigma_{\mathbf{z}}(f_{\mathbf{z}}) - \mathcal{E}^\sigma_{\mathbf{z}}(f_\rho)) \leq \frac{1}{2}(\mathcal{E}^\sigma(f_{\mathbf{z}}) - \mathcal{E}^\sigma(f_\rho)) + 9\varepsilon_{\mathcal{H},\rho}. \tag{15}$$

Proposition 9 tells us that

$$\begin{aligned} \mathcal{E}^\sigma(f_{\mathbf{z}}) - \mathcal{E}^\sigma(f_\rho) &= \mathcal{E}^\sigma(f_{\mathbf{z}}) - \mathcal{E}^\sigma(f^\sigma_{\mathcal{H}}) + \mathcal{E}^\sigma(f^\sigma_{\mathcal{H}}) - \mathcal{E}^\sigma(f_\rho) \\ &\leq \mathcal{S}_1(\mathbf{z}) + \mathcal{S}_2(\mathbf{z}) + \|f_{\mathcal{H}} - f_\rho\|^2_{\mathcal{L}^2_{\rho_{\mathcal{X}}}} + c_{\mathcal{H},\rho}/\sigma^2, \end{aligned} \tag{16}$$

where the above inequality is due to Lemma 7 and the observation that

$$\begin{aligned} \mathcal{E}^\sigma(f^\sigma_{\mathcal{H}}) - \mathcal{E}^\sigma(f_\rho) &= \mathcal{E}^\sigma(f^\sigma_{\mathcal{H}}) - \mathcal{E}^\sigma(f_{\mathcal{H}}) + \mathcal{E}^\sigma(f_{\mathcal{H}}) - \mathcal{E}^\sigma(f_\rho) \\ &\leq \mathcal{E}^\sigma(f_{\mathcal{H}}) - \mathcal{E}^\sigma(f_\rho) \\ &\leq \|f_{\mathcal{H}} - f_\rho\|^2_{\mathcal{L}^2_{\rho_{\mathcal{X}}}} + c_{\mathcal{H},\rho}/\sigma^2. \end{aligned}$$

Combining estimates in (15) and (16), we come to the conclusion that for any $0 < \delta < 1$, with confidence $1 - \delta/2$, there holds

$$\mathcal{S}_2(\mathbf{z}) \leq \frac{1}{2}(\mathcal{S}_1(\mathbf{z}) + \mathcal{S}_2(\mathbf{z})) + \frac{1}{2}\|f_{\mathcal{H}} - f_\rho\|^2_{\mathcal{L}^2_{\rho_{\mathcal{X}}}} + C_{\mathcal{H},\rho,2}\left(\log\frac{2}{\delta}\right)\left\{\frac{1}{\sigma^2} + \frac{\sigma}{m^{1/(1+p)}}\right\},$$

where $C_{\mathcal{H},\rho,2}$ is a positive constant independent of m, σ or δ and given by $C_{\mathcal{H},\rho,2} = 2c_{\mathcal{H},\rho} + 9c_{\mathcal{H},\rho,1}$. This completes the proof of Proposition 11. ■

2. Similarly, refined estimate can be also derived here by using Young's inequality $ab \leq \frac{ta^2}{2} + \frac{b^2}{2t}$ for $a, b \in \mathbb{R}$, $t > 0$. In our proof, again we choose $t = 1$ for simplification.

4.3.4 Proof of Theorem 4

Proof From Lemma 7 and Proposition 9, we know that

$$\|f_{\mathbf{z}} - f_\rho\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}} \leq \mathcal{S}_1(\mathbf{z}) + \mathcal{S}_2(\mathbf{z}) + \|f_\mathcal{H} - f_\rho\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}} + 2c_{\mathcal{H},\rho}/\sigma^2. \tag{17}$$

Combining estimates in Proposition 10 and Proposition 11 for the sample error terms $\mathcal{S}_1(\mathbf{z})$ and $\mathcal{S}_2(\mathbf{z})$, we know that for any $0 < \delta < 1$, with confidence $1 - \delta$, there holds

$$\mathcal{S}_1(\mathbf{z}) + \mathcal{S}_2(\mathbf{z}) \leq 2\|f_\mathcal{H} - f_\rho\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}} + (2C_{\mathcal{H},\rho,1} + 4C_{\mathcal{H},\rho,2}) \log(2/\delta) \{\sigma^{-2} + \sigma m^{-1/(1+p)}\}.$$

This in connection with the estimate in (17) tells us that for any $0 < \delta < 1$, with confidence $1 - \delta$, there holds

$$\|f_{\mathbf{z}} - f_\rho\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}} \leq 3\|f_\mathcal{H} - f_\rho\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}} + C_{\mathcal{H},\rho} \log(2/\delta) \{\sigma^{-2} + \sigma m^{-1/(1+p)}\},$$

where $C_{\mathcal{H},\rho} = 2C_{\mathcal{H},\rho,1} + 4C_{\mathcal{H},\rho,2} + 4c_{\mathcal{H},\rho}$. This completes the proof of Theorem 4. ■

4.3.5 Proof of Theorem 5

The proof of Theorem 5 can be similarly conducted as that of Theorem 4, since the error decomposition in Proposition 9 holds when Y is bounded. Therefore, we also need to bound the two sample error terms $\mathcal{S}_1(\mathbf{z})$ and $\mathcal{S}_1(\mathbf{z})$, respectively.

Proposition 16 *Assume that $|y| \leq M$ almost surely for some $M > 0$, and $f_\rho \in \mathcal{H}$. For any $0 < \delta < 1$, with confidence $1 - \delta/2$, there holds*

$$\mathcal{S}_1(\mathbf{z}) \leq C'_{\mathcal{H},\rho,1} \log(2/\delta)(\sigma^{-2} + m^{-1}),$$

where $C'_{\mathcal{H},\rho,1}$ is a positive constant that independent of m, σ or δ and will be given explicitly in the proof.

Proof We will finish the proof by following similar process as done for Proposition 10. We first introduce the random variable $\bar{\xi}_1(z)$ as follows

$$\bar{\xi}_1(z) = -\sigma^2 \exp\left\{-(y - f^\sigma_\mathcal{H}(x))^2/\sigma^2\right\} + \sigma^2 \exp\left\{-(y - f_\rho(x))^2/\sigma^2\right\}, \ z \in \mathcal{Z}.$$

It follows from the proof of Proposition 10 and the boundedness of Y that for any $z \in \mathcal{Z}$, there holds

$$\begin{aligned} |\bar{\xi}_1(z)| &\leq \left|(2y - f^\sigma_\mathcal{H}(x) - f_\rho(x))(f^\sigma_\mathcal{H}(x) - f_\rho(x))\right| \\ &\leq \Big(2M + \|f_\rho\|_\infty + \sup_{f\in\mathcal{H}} \|f\|_\infty\Big) \sup_{f\in\mathcal{H}} \|f - f_\rho\|_\infty. \end{aligned}$$

Consequently, the following estimate holds

$$\left|\bar{\xi}_1 - \mathbb{E}\bar{\xi}_1\right| \leq 2\Big(2M + \|f_\rho\|_\infty + \sup_{f\in\mathcal{H}} \|f\|_\infty\Big) \sup_{f\in\mathcal{H}} \|f - f_\rho\|_\infty := c'_{\mathcal{H},\rho,0}.$$

Denote the variance of the random variable $\bar{\xi}_1$ as $\mathrm{var}(\bar{\xi}_1)$. From the proof of Proposition 10 and the boundedness of Y, we have

$$\begin{aligned}\mathrm{var}(\bar{\xi}_1) &= \mathbb{E}\bar{\xi}_1^2 - (\mathbb{E}\bar{\xi}_1)^2 \\ &\leq \mathbb{E}\bar{\xi}_1^2 \leq \mathbb{E}\left((f_{\mathcal{H}}^{\sigma}(x) - f_\rho(x))^2(2y - f_{\mathcal{H}}^{\sigma}(x) - f_\rho(x))^2\right) \\ &\leq \left(12M^2 + 3\sup_{f\in\mathcal{H}}\|f\|_\infty^2 + 3\|f_\rho\|_\infty^2\right)\int_{\mathcal{X}}(f_{\mathcal{H}}^{\sigma}(x) - f_\rho(x))^2 d\rho_{\mathcal{X}}(x).\end{aligned}$$

Recalling the fact that $f_\rho \in \mathcal{H}$, as a consequence of Lemma 7, we obtain

$$\int_{\mathcal{X}}(f_{\mathcal{H}}^{\sigma}(x) - f_\rho(x))^2 d\rho_{\mathcal{X}}(x) \leq \int_{\mathcal{X}}(f_{\mathcal{H}}(x) - f_\rho(x))^2 d\rho_{\mathcal{X}}(x) + \frac{2c_{\mathcal{H},\rho}}{\sigma^2} = \frac{2c_{\mathcal{H},\rho}}{\sigma^2}.$$

Combining the above two estimates, we obtain the following upper bound for the variance of $\bar{\xi}_1$:

$$\mathrm{var}(\bar{\xi}_1) \leq c'_{\mathcal{H},\rho,1}/\sigma^2 \quad \text{with} \quad c'_{\mathcal{H},\rho,1} = 2c_{\mathcal{H},\rho}\Big(12M^2 + 3\sup_{f\in\mathcal{H}}\|f\|_\infty^2 + 3\|f_\rho\|_\infty^2\Big).$$

Applying the one-sided Bernstein's inequality in Lemma 13 to the random variable $\bar{\xi}_1$ and with simple computations, we come to the conclusion that for any $0 < \delta < 1$, with confidence $1 - \delta/2$, there holds

$$\mathcal{S}_1(\mathbf{z}) \leq C'_{\mathcal{H},\rho,1}\log(2/\delta)(\sigma^{-2} + m^{-1}),$$

where $C'_{\mathcal{H},\rho,1}$ is a positive constant independent of m, σ or δ and given by $C'_{\mathcal{H},\rho,1} = 2 + c'_{\mathcal{H},\rho,1}/2 + 2c'_{\mathcal{H},\rho,0}/3$. This completes the proof. ■

We now turn to bound the sample error term $\mathcal{S}_2(\mathbf{z})$ when Y is bounded.

Proposition 17 *Assume that the Complexity Assumption II with $0 < s < 2$ holds, $|y| \leq M$ almost surely for some $M > 0$. Let $f_\rho \in \mathcal{H}$ and $\sigma \geq 1$. For any $f \in \mathcal{H}$ and $0 < \delta < 1$, with confidence $1 - \delta/2$, there holds*

$$\{\mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f_\rho)\} - \{\mathcal{E}_{\mathbf{z}}^\sigma(f) - \mathcal{E}_{\mathbf{z}}^\sigma(f_\rho)\} \leq \frac{1}{2}\{\mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f_\rho)\} + C'_{\mathcal{H},\rho,2}\log(2/\delta)m^{-\frac{2}{2+s}},$$

where $C'_{\mathcal{H},\rho,2}$ is a positive constant independent of m, σ or δ and will be given explicitly in the proof.

Proof To prove the proposition, we apply Lemma 14 to the function set $\mathcal{F}_{\mathcal{H}}$, which is defined as

$$\mathcal{F}_{\mathcal{H}} = \left\{g \;\middle|\; g(z) = \ell_\sigma(y, f(x)) - \ell_\sigma(y, f_\rho(x)) + \frac{c_{\mathcal{H},\rho}}{\sigma^2}, f \in \mathcal{H}, z \in \mathcal{Z}\right\}.$$

According to the definition of $\mathcal{F}_{\mathcal{H}}$, for any $g \in \mathcal{F}_{\mathcal{H}}$, it can be explicitly expressed as

$$g(z) = -\sigma^2\exp\left\{-(y - f(x))^2/\sigma^2\right\} + \sigma^2\exp\left\{-(y - f_\rho(x))^2/\sigma^2\right\} + \frac{c_{\mathcal{H},\rho}}{\sigma^2},$$

with $z \in \mathcal{Z}$ and $f \in \mathcal{H}$. Recalling that $|y| \le M$ almost surely and $\sigma \ge 1$, simple computations show that

$$\|g\|_\infty \le \Big(2M + \|f_\rho\|_\infty + \sup_{f\in\mathcal{H}} \|f\|_\infty\Big) \sup_{f\in\mathcal{H}} \|f - f_\rho\|_\infty + c_{\mathcal{H},\rho}.$$

Applying the mean value theorem again as done in the proof of Proposition 10, we get

$$\begin{aligned}
&\left(-\sigma^2 \exp\left\{-(y-f(x))^2/\sigma^2\right\} + \sigma^2 \exp\left\{-(y-f_\rho(x))^2/\sigma^2\right\}\right)^2 \\
&\qquad \le \left((y-f(x))^2 - (y-f_\rho(x))^2\right)^2 \\
&\qquad \le \Big(2M + \sup_{f\in\mathcal{H}} \|f\|_\infty + \|f_\rho\|_\infty\Big)^2 (f(x) - f_\rho(x))^2,
\end{aligned}$$

where the last inequality is again due to the boundedness of Y. This in connection with Lemma 7 tells us that

$$\begin{aligned}
\mathbb{E}g^2 &= \int_{\mathcal{Z}} \left(-\sigma^2 \exp\left\{-\frac{(y-f(x))^2}{\sigma^2}\right\} + \sigma^2 \exp\left\{-\frac{(y-f_\rho(x))^2}{\sigma^2}\right\}\right)^2 d\rho \\
&\quad + \frac{2c_{\mathcal{H},\rho}}{\sigma^2} \int_{\mathcal{Z}} \left(-\sigma^2 \exp\left\{-\frac{(y-f(x))^2}{\sigma^2}\right\} + \sigma^2 \exp\left\{-\frac{(y-f_\rho(x))^2}{\sigma^2}\right\}\right) d\rho + \frac{c_{\mathcal{H},\rho}^2}{\sigma^4} \\
&\le \Big(2M + \sup_{f\in\mathcal{H}} \|f\|_\infty + \|f_\rho\|_\infty\Big)^2 (\mathcal{E}(f) - \mathcal{E}(f_\rho)) + \frac{2c_{\mathcal{H},\rho}}{\sigma^2} \left(\mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f_\rho) + \frac{c_{\mathcal{H},\rho}}{\sigma^2}\right) \\
&\le \Big(2M + \sup_{f\in\mathcal{H}} \|f\|_\infty + \|f_\rho\|_\infty\Big)^2 \left(\mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f_\rho) + \frac{c_{\mathcal{H},\rho}}{\sigma^2}\right) + \frac{2c_{\mathcal{H},\rho}}{\sigma^2} \left(\mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f_\rho) + \frac{c_{\mathcal{H},\rho}}{\sigma^2}\right) \\
&= \left(\Big(2M + \sup_{f\in\mathcal{H}} \|f\|_\infty + \|f_\rho\|_\infty\Big)^2 + \frac{2c_{\mathcal{H},\rho}}{\sigma^2}\right) \left(\mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f_\rho) + \frac{c_{\mathcal{H},\rho}}{\sigma^2}\right) \\
&= \left(\Big(2M + \sup_{f\in\mathcal{H}} \|f\|_\infty + \|f_\rho\|_\infty\Big)^2 + \frac{2c_{\mathcal{H},\rho}}{\sigma^2}\right) \mathbb{E}g \\
&\le \left(\Big(2M + \sup_{f\in\mathcal{H}} \|f\|_\infty + \|f_\rho\|_\infty\Big)^2 + 2c_{\mathcal{H},\rho}\right) \mathbb{E}g,
\end{aligned}$$

where the last inequality is due to the assumption that $\sigma \ge 1$.

For any g_1, $g_2 \in \mathcal{F}_\mathcal{H}$, there exist f_1, $f_2 \in \mathcal{H}$ such that

$$g_1(z) = -\sigma^2 \exp\left\{-(y-f_1(x))^2/\sigma^2\right\} + \sigma^2 \exp\left\{-(y-f_\rho(x))^2/\sigma^2\right\} + \frac{c_{\mathcal{H},\rho}}{\sigma^2}$$

and

$$g_2(z) = -\sigma^2 \exp\left\{-(y-f_2(x))^2/\sigma^2\right\} + \sigma^2 \exp\left\{-(y-f_\rho(x))^2/\sigma^2\right\} + \frac{c_{\mathcal{H},\rho}}{\sigma^2}.$$

Applying the mean value theorem and noticing the boundedness of Y, we have

$$|g_1(z) - g_2(z)| \le 2\Big(M + \sup_{f\in\mathcal{H}} \|f\|_\infty\Big)\|f_1 - f_2\|_\infty,\ z \in \mathcal{Z}.$$

Under the Complexity Assumption II with $0 < s < 2$, the following relation between the ℓ^2-empirical covering numbers of $\mathcal{F}_{\mathcal{H}}$ and $\mathcal{H}$ holds

$$\log \mathcal{N}_2(\mathcal{F}_{\mathcal{H}}, \eta) \leq \log \mathcal{N}_2\Big(\mathcal{H}, \eta \Big/ \Big(2M + 2\sup_{f\in\mathcal{H}} \|f\|_\infty\Big)\Big) \leq c_{II,s}\Big(\Big(2M + 2\sup_{f\in\mathcal{H}} \|f\|_\infty\Big)\Big/\eta\Big)^s.$$

For notation simplification, we denote

$$\begin{aligned}
c'_{\mathcal{H},\rho,2} &= \Big(2M + \sup_{f\in\mathcal{H}} \|f\|_\infty + \|f_\rho\|_\infty\Big)^2 + 2c_{\mathcal{H},\rho},\\
B'_{\mathcal{H},\rho} &= \Big(2M + \sup_{f\in\mathcal{H}} \|f\|_\infty + \|f_\rho\|_\infty\Big) \sup_{f\in\mathcal{H}} \|f - f_\rho\|_\infty + c_{\mathcal{H},\rho},\\
a_{\mathcal{H},s} &= c_{II,s}\Big(2M + 2\sup_{f\in\mathcal{H}} \|f\|_\infty\Big)^s.
\end{aligned}$$

Applying Lemma 14 to the function set $\mathcal{F}_{\mathcal{H}}$, with simple computations, we come to the conclusion that when $\sigma \geq 1$, for any $0 < \delta < 1$ with confidence $1 - \delta/2$, there holds

$$\{\mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f_\rho)\} - \{\mathcal{E}^\sigma_{\mathbf{z}}(f) - \mathcal{E}^\sigma_{\mathbf{z}}(f_\rho)\} \leq \frac{1}{2}\{\mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f_\rho)\} + C'_{\mathcal{H},\rho,2}\log(2/\delta)m^{-\frac{2}{2+s}},$$

where $C'_{\mathcal{H},\rho,2}$ is a positive constant independent of m, σ or δ and given by

$$C'_{\mathcal{H},\rho,2} = 18B'_{\mathcal{H},\rho} + 2c'_{\mathcal{H},\rho,2} + 2a_s a^{2/(2+s)}_{\mathcal{H},s}(c'_{\mathcal{H},\rho,2} + B'_{\mathcal{H},\rho})^{(2-s)/(2+s)},$$

and a_s is a positive constant depending only on s. This completes the proof of Proposition 17. ■

Proof [Proof of Theorem 5] Following from the estimate in inequality (11), and recalling that $f_\rho \in \mathcal{H}$, we have

$$\|f_{\mathbf{z}} - f_\rho\|^2_{\mathcal{L}^2_{\rho_{\mathcal{X}}}} \leq \mathcal{S}_1(\mathbf{z}) + \mathcal{S}_2(\mathbf{z}) + 2c_{\mathcal{H},\rho}/\sigma^2. \tag{18}$$

As a consequence of Proposition 17, we know that when $\sigma \geq 1$, for any $0 < \delta < 1$ with confidence $1 - \delta/2$, there holds

$$\{\mathcal{E}^\sigma(f_{\mathbf{z}}) - \mathcal{E}^\sigma(f_\rho)\} - \{\mathcal{E}^\sigma_{\mathbf{z}}(f_{\mathbf{z}}) - \mathcal{E}^\sigma_{\mathbf{z}}(f_\rho)\} \leq \frac{1}{2}\{\mathcal{E}^\sigma(f_{\mathbf{z}}) - \mathcal{E}^\sigma(f_\rho)\} + C'_{\mathcal{H},\rho,2}\log(2/\delta)m^{-\frac{2}{2+s}}.$$

The above inequality together with Lemma 7 yields

$$\begin{aligned}
\mathcal{S}_2(\mathbf{z}) &= \{\mathcal{E}^\sigma(f_{\mathbf{z}}) - \mathcal{E}^\sigma(f_\rho)\} - \{\mathcal{E}^\sigma_{\mathbf{z}}(f_{\mathbf{z}}) - \mathcal{E}^\sigma_{\mathbf{z}}(f_\rho)\}\\
&\leq \frac{1}{2}\|f_{\mathbf{z}} - f_\rho\|^2_{\mathcal{L}^2_{\rho_{\mathcal{X}}}} + \frac{c_{\mathcal{H},\rho}}{2\sigma^2} + C'_{\mathcal{H},\rho,2}\log(2/\delta)m^{-\frac{2}{2+s}}.
\end{aligned}$$

This in connection with the upper bound for the sample error term $\mathcal{S}_1(\mathbf{z})$ in Proposition 16 and inequality (18), with the choice $\sigma = m^{1/(2+s)}$, yields that for any $0 < \delta < 1$, with confidence $1 - \delta$, there holds

$$\|f_{\mathbf{z}} - f_\rho\|^2_{\mathcal{L}^2_{\rho_{\mathcal{X}}}} \leq C'_{\mathcal{H},\rho}\log(2/\delta)m^{-\frac{2}{2+s}},$$

where $C'_{\mathcal{H},\rho} = 2C'_{\mathcal{H},\rho,1} + C'_{\mathcal{H},\rho,2} + 3c_{\mathcal{H},\rho}$. This completes the proof of Theorem 5. ■

4.3.6 Proof of Theorem 6

To prove Theorem 6, we first prove the following conclusion.

Lemma 18 *Assume that the Noise Assumption holds, and $f_\rho \in \mathcal{H}$. Let σ be fixed and satisfy*

$$\sigma > \sigma_{\mathcal{H},\rho} = \sqrt{2}\Big(M_0 + \|f_\rho\|_\infty + \sup_{f\in\mathcal{H}}\|f\|_\infty\Big).$$

For any $f \in \mathcal{H}$, there exists a positive constant $c_{\mathcal{H},\sigma,\rho} \in (0,1)$, such that

$$c_{\mathcal{H},\sigma,\rho}\left\{\mathcal{E}(f) - \mathcal{E}(f_\rho)\right\} \le \mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f_\rho).$$

Proof Under the Noise Assumption, when $\sigma > \sigma_{\mathcal{H},\rho}$, Theorem 8 shows that for any $f \in \mathcal{H}$,

$$\mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f_{\mathcal{H}}^\sigma) = \mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f_{\mathcal{H}}) = \mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f_\rho).$$

For any $x \in \mathcal{X}$, again we denote $F_x(u) = 1 - \int_{-M_0}^{M_0} \exp\left\{-\frac{(t-u)^2}{\sigma^2}\right\} p_{\epsilon|X=x}(t)dt$, then

$$\begin{aligned}
\mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f_\rho) &= \sigma^2 \int_{\mathcal{X}} \left(F_x(f(x) - f_\rho(x)) - F_x(0)\right) d\rho_{\mathcal{X}}(x) \\
&= \sigma^2 \int_{\mathcal{X}} \left\{F_x'(0)(f(x) - f_\rho(x)) + \frac{F_x''(\xi_x)}{2}(f(x) - f_\rho(x))^2\right\} d\rho_{\mathcal{X}}(x) \\
&= \int_{\mathcal{X}} \frac{\sigma^2 F_x''(\xi_x)}{2}(f(x) - f_\rho(x))^2 d\rho_{\mathcal{X}}(x),
\end{aligned}$$

where the last equality follows from the fact that $F_x'(0) = 0$ and ξ_x falls between 0 and $f(x) - f_\rho(x)$ for any $x \in \mathcal{X}$. It is easy to see that when σ is fixed and $\sigma > \sigma_{\mathcal{H},\rho}$, we have

$$\begin{aligned}
F_x''(\xi_x) &= 2\int_{-M_0}^{M_0} \exp\left\{-\frac{(t-\xi_x)^2}{\sigma^2}\right\}\left(\frac{\sigma^2 - 2(t-\xi_x)^2}{\sigma^4}\right) p_{\epsilon|X=x}(t)dt \\
&\ge (2\sigma^2 - 2\sigma_{\mathcal{H},\rho}^2)/\sigma^4 \exp(-\sigma_{\mathcal{H},\rho}^2/\sigma^2), \text{ for any } x \in \mathcal{X},
\end{aligned}$$

where the last inequality is due to the following fact

$$|t - \xi_x| \le \sqrt{2}\sigma_{\mathcal{H},\rho}/2, \quad t \in [-M_0, M_0],\, x \in \mathcal{X}.$$

As a result, we come to the conclusion that

$$\mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f_\rho) \ge c_{\mathcal{H},\sigma,\rho}\left\{\mathcal{E}(f) - \mathcal{E}(f_\rho)\right\},$$

where $c_{\mathcal{H},\sigma,\rho} = (\sigma^2 - \sigma_{\mathcal{H},\rho}^2)/\sigma^2 \exp(-\sigma_{\mathcal{H},\rho}^2/\sigma^2)$. Noticing that $0 < c_{\mathcal{H},\sigma,\rho} < 1$, we have verified our assertion. ■

The proof of Theorem 6 is different from the proofs of Theorem 4 and Theorem 5. This is because when σ is fixed, σ^{-1} does not tend to zero and consequently we cannot get

meaningful convergence rates via the error decomposition in Proposition 9. However, from Lemma 18, we know that

$$\|f_{\mathbf{z}} - f_\rho\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}} \leq c^{-1}_{\mathcal{H},\sigma,\rho} \left\{\mathcal{E}^\sigma(f_{\mathbf{z}}) - \mathcal{E}^\sigma(f_\rho)\right\} = c^{-1}_{\mathcal{H},\sigma,\rho} \left(\mathcal{S}_1(\mathbf{z}) + \mathcal{S}_2(\mathbf{z})\right),$$

where the definitions of $\mathcal{S}_1(\mathbf{z})$ and $\mathcal{S}_2(\mathbf{z})$ are inherited from Proposition 9.

We notice that under the condition that the Noise Assumption holds, and $f_\rho \in \mathcal{H}$, when σ is fixed and satisfies

$$\sigma > \sigma_{\mathcal{H},\rho} = \sqrt{2}\Big(M_0 + \|f_\rho\|_\infty + \sup_{f\in\mathcal{H}} \|f\|_\infty\Big),$$

Theorem 8 tells us that almost surely $f^\sigma_\mathcal{H} = f_\rho$. In this situation, almost surely we have $\mathcal{S}_1(\mathbf{z}) = 0$. Therefore, to prove Theorem 6, it suffices to bound the sample error term $\mathcal{S}_2(\mathbf{z})$. This can be done by applying Lemma 14 to the function set

$$\mathcal{F}_\mathcal{H} = \left\{g \,\middle|\, g(z) = \ell_\sigma(y, f(x)) - \ell_\sigma(y, f_\rho(x)) : f \in \mathcal{H}, z \in \mathcal{Z}\right\}.$$

Proposition 19 *Assume that the Complexity Assumption II with $0 < s < 2$ and the Noise Assumption hold. Let $f_\rho \in \mathcal{H}$, σ be fixed and satisfy*

$$\sigma > \sigma_{\mathcal{H},\rho} = \sqrt{2}\Big(M_0 + \|f_\rho\|_\infty + \sup_{f\in\mathcal{H}} \|f\|_\infty\Big).$$

For any $f \in \mathcal{H}$ and $0 < \delta < 1$, with confidence $1 - \delta$, there holds

$$\left\{\mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f_\rho)\right\} - \left\{\mathcal{E}^\sigma_{\mathbf{z}}(f) - \mathcal{E}^\sigma_{\mathbf{z}}(f_\rho)\right\} \leq \frac{1}{2}\left\{\mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f_\rho)\right\} + C_{\mathcal{H},\sigma,\rho,1} \log(1/\delta) m^{-\frac{2}{2+s}},$$

where $C_{\mathcal{H},\sigma,\rho,1}$ is a positive constant independent of m or δ and will be given explicitly in the proof.

Proof For any $g \in \mathcal{F}_\mathcal{H}$, we know from the definition of $\mathcal{F}_\mathcal{H}$ that g can be expressed as

$$g(z) = -\sigma^2 \exp\left\{-(y - f(x))^2/\sigma^2\right\} + \sigma^2 \exp\left\{-(y - f_\rho(x))^2/\sigma^2\right\}, z \in \mathcal{Z},$$

for some $f \in \mathcal{H}$. Following from the proof of Proposition 10, we know that

$$\|g\|_\infty \leq \sqrt{2/e}\sigma \sup_{f\in\mathcal{H}} \|f - f_\rho\|_\infty := B_{\mathcal{H},\sigma,\rho}.$$

When the Noise Assumption holds, $f_\rho \in \mathcal{H}$, and $\sigma > \sigma_{\mathcal{H},\rho}$, we have

$$\begin{aligned}
\mathbb{E}g^2 &\leq \mathbb{E}\left((f(x) - f_\rho(x))^2(2y - f(x) - f_\rho(x))^2\right) \\
&\leq \int_\mathcal{Y} \Big(12y^2 + 3\sup_{f\in\mathcal{H}} \|f\|^2_\infty + 3\|f_\rho\|^2_\infty\Big) d\rho(y|x) \int_\mathcal{X} (f(x) - f_\rho(x))^2 d\rho_\mathcal{X}(x) \\
&\leq c^{-1}_{\mathcal{H},\sigma,\rho}\Big(12\int_\mathcal{Z} y^2 d\rho + 3\sup_{f\in\mathcal{H}} \|f\|^2_\infty + 3\|f_\rho\|^2_\infty\Big)\mathbb{E}g := c_{\mathcal{H},\sigma,\rho,1}\mathbb{E}g,
\end{aligned}$$

where the last inequality follows from Lemma 18. For any $g_1,\, g_2 \in \mathcal{F}_{\mathcal{H}}$, there exist $f_1,\, f_2 \in \mathcal{H}$ such that

$$g_1(z) = -\sigma^2 \exp\left\{-(y - f_1(x))^2/\sigma^2\right\} + \sigma^2 \exp\left\{-(y - f_\rho(x))^2/\sigma^2\right\}$$

and

$$g_2(z) = -\sigma^2 \exp\left\{-(y - f_2(x))^2/\sigma^2\right\} + \sigma^2 \exp\left\{-(y - f_\rho(x))^2/\sigma^2\right\}.$$

From the proof of Proposition 10, we know that $|g_1 - g_2| \leq \sqrt{2/e}\sigma\|f_1 - f_2\|_\infty$. This in connection with the Complexity Assumption II yields

$$\log \mathcal{N}_2(\mathcal{F}_{\mathcal{H}}, \eta) \leq \log \mathcal{N}_2\left(\mathcal{H}, \eta/(\sqrt{2/e}\sigma)\right) \leq c_{II,s}\left(\sqrt{2/e}\sigma/\eta\right)^s := a_{\sigma,s}\eta^{-s}.$$

Applying Lemma 14 to the function set $\mathcal{F}_{\mathcal{H}}$, with simple computations, we see that for any $0 < \delta < 1$ with confidence $1 - \delta$, there holds

$$\{\mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f_\rho)\} - \{\mathcal{E}^\sigma_{\mathbf{z}}(f) - \mathcal{E}^\sigma_{\mathbf{z}}(f_\rho)\} \leq \frac{1}{2}\{\mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f_\rho)\} + C_{\mathcal{H},\sigma,\rho,1}\log(1/\delta)m^{-2/(2+s)},$$

where $C_{\mathcal{H},\rho,\sigma,1}$ is a positive constant independent of m or δ and given by

$$C_{\mathcal{H},\rho,\sigma,1} = 18B_{\mathcal{H},\sigma,\rho} + 2c_{\mathcal{H},\sigma,\rho,1} + 2a'_s a^{2/(2+s)}_{\sigma,s}(c_{\mathcal{H},\sigma,\rho,1} + B_{\mathcal{H},\sigma,\rho})^{(2-s)/(2+s)},$$

and a'_s is a positive constant depending only on s. This completes the proof of Proposition 19. ■

Proof [Proof of Theorem 6] As a consequence of Proposition 19, we see that for any $0 < \delta < 1$, with confidence $1 - \delta$, there holds

$$\mathcal{S}_2(\mathbf{z}) \leq \frac{1}{2}\{\mathcal{E}^\sigma(f_{\mathbf{z}}) - \mathcal{E}^\sigma(f_\rho)\} + C_{\mathcal{H},\sigma,\rho,1}\log(1/\delta)m^{-2/(2+s)}.$$

Following from Lemma 18 and recalling that $\mathcal{S}_1(\mathbf{z}) = 0$, we come to the conclusion that for any $0 < \delta < 1$, with confidence $1 - \delta$, there holds

$$\|f_{\mathbf{z}} - f_\rho\|^2_{\mathcal{L}^2_{\rho_{\mathcal{X}}}} \leq c^{-1}_{\mathcal{H},\sigma,\rho}\{\mathcal{E}^\sigma(f_{\mathbf{z}}) - \mathcal{E}^\sigma(f_\rho)\} = c^{-1}_{\mathcal{H},\sigma,\rho}\mathcal{S}_2(\mathbf{z}) \leq 2c^{-1}_{\mathcal{H},\sigma,\rho}C_{\mathcal{H},\sigma,\rho,1}\log(1/\delta)m^{-2/(2+s)}.$$

By denoting $C_{\mathcal{H},\sigma,\rho} = 2c^{-1}_{\mathcal{H},\sigma,\rho}C_{\mathcal{H},\sigma,\rho,1}$, we complete the proof of Theorem 6. ■

5. Towards the Role that σ Plays

We now move our attention to discuss the scale parameter σ in the ℓ_σ loss by making some attempts to interpret the role that σ plays from a learning theory viewpoint.

The first observation on the parameter σ in the ℓ_σ loss is that it determines the robustness of the regression models. For linear regression models, this observation has been quantitatively described in terms of the influence function and finite-sample breakdown

point in Wang et al. (2013). For nonlinear regression models, similar observations on the robustness have been also empirically reported. For instance, the robustness of the regression models induced by the ℓ_σ losses can be enhanced with a decreasing value of σ. In fact, this is reasonable if we look at the ℓ_σ loss in which a smaller σ would limit the influence of the outliers in the response variable. In addition, in the learning theory literature, the robustness property of kernel-based regression models has been studied by considering the growth type of the loss function and investigating the existence and boundedness of the corresponding influence function (see Christmann and Steinwart, 2007; Steinwart and Christmann, 2008). From Chapter 2 in Steinwart and Christmann (2008), it is easy to check that the ℓ_σ loss is of upper growth type 1 due to its Lipschitz continuity property and consequently can be used to deal with unbounded Y. It would be also worthwhile to derive a quantitative description on the robustness of the MCCR model (2) in terms of the influence function as done in Christmann and Steinwart (2007) and Christmann and Messem (2008) for convex regression models. However, we remark that due to the non-convexity of the ℓ_σ loss, the deduction of the influence function of the MCCR model in $\mathcal{H}$ (which is possibly infinite dimensional) can be much involved and is worthy for further study.

On the other hand, we realize that in the robustness literature, the scale parameter not only controls the robustness property of the regression model associated with the ℓ_σ loss but also specifies its efficiency and plays a trade-off role. Considering the nonparametric setting in our study and given that our primary concern is the convergence rates of the MCCR model (2), we restrict ourselves to discussions of the influence of the scale parameter σ on the convergence rates. To this end, we recall the following relation from the error decomposition in Proposition 9:

$$\|f_{\mathbf{z}} - f_\rho\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}} \leq \left\{\mathcal{E}^\sigma(f_{\mathbf{z}}) - \mathcal{E}^\sigma(f^\sigma_\mathcal{H})\right\} + \|f_\mathcal{H} - f_\rho\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}} + \mathcal{A}_{\mathcal{H},\sigma,\rho}.$$

On the right-hand side of the above inequality, the first term is the excess risk of the empirical estimator modeled by the MCCR model, the convergence of which can be ensured by controlling the complexity of the hypothesis space $\mathcal{H}$ and confining the tail behavior of the response variable. The second term $\|f_\mathcal{H} - f_\rho\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}}$ represents the approximation error and is independent of the scale parameter σ. The influence of the scale parameter σ on the convergence rates can be revealed from the bias term $\mathcal{A}_{\mathcal{H},\sigma,\rho}$. According to Proposition 9, we know that $\mathcal{A}_{\mathcal{H},\sigma,\rho} = 2c_{\mathcal{H},\rho}/\sigma^2$. Therefore, a decreasing value of σ will lead to increasing bias and consequently yields slower convergence rates.

From the above discussions, we can see that the parameter σ in the ℓ_σ loss balances the robustness of the MCCR model (2) and its convergence rates. We will continue our discussion on the role that σ plays by trying to extend our preceding analysis for the ℓ_σ loss to other robust regression loss functions in the next section.

6. Generalization to Other Robust Loss Functions

In the preceding sections, motivated by the information-theoretic interpretation of the maximum correntropy criterion and its empirical successes in real-world applications, we generalize the idea of the maximum correntropy criterion in regression with the ℓ_σ loss. We then present a theoretical understanding towards the maximum correntropy criterion in regression by conducting a learning theory analysis for $\|f_{\mathbf{z}} - f_\rho\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}}$. We conclude that one can

rely on the ℓ_σ loss to solve regression problems with non-Gaussian as well as Gaussian noise. However, one may argue that from a regression viewpoint, the ℓ_σ loss is merely a special case of robust loss functions arise in robust statistics. In view of this, in this section we try to generalize our previous analysis to other robust loss functions and see what happens when a robust loss function is applied into the learning for regression scenarios.

The robust loss functions refer to those used to obtain robust M-estimators in linear regression models. As mentioned earlier, the MCCR model can be viewed as a nonparametric M-estimator. Therefore, we first give a glimpse of the robust M-estimation methods in linear regression models to distinguish them from the robust nonparametric M-estimator we investigate in this paper. In linear regression models, it is assumed that the observations $\mathbf{z}$ are drawn i.i.d from $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ with $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}$. In this setting, the regression function $f^\star(x) := x^T\theta^\star$, where $\theta^\star \in \Theta := \mathbb{R}^d$ is unknown and one of the main tasks in linear regression problem is to estimate the regression parameter $\theta^\star$. A common approach to obtaining a robust estimator $\hat{\theta}$ for $\theta^\star$ is to solve the following optimization problem

$$\hat{\theta} = \arg\min_{\theta\in\mathbb{R}^d} \sum_{i=1}^m \phi\left(\frac{y_i - x_i^T\theta}{\sigma}\right), \tag{19}$$

where $\sigma > 0$ is the scale parameter and ϕ is a robust loss function that downweights large residual errors. In fact, by using the above robust loss function ϕ, concerning the nonlinear regression model (1), one can also propose the following robust nonparametric ERM-based regression scheme

$$\hat{f}_{\mathbf{z}} = \arg\min_{f\in\mathcal{H}} \sum_{i=1}^m \phi\left(\frac{y_i - f(x_i)}{\sigma}\right). \tag{20}$$

Notice that (19) aims at estimating a vector in $\mathbb{R}^d$ while (20) is proposed to estimate a function in a function space $\mathcal{H}$ that can have an infinite dimension. This gives the main difference between the two models. Denoting $\phi_\sigma(t) := \phi(t/\sigma)$, besides the ℓ_σ loss investigated in this paper, several frequently employed robust loss functions include:

- Huber's loss: $\phi_\sigma(t) = t^2 I_{\{|t|\le\sigma\}} + (2\sigma|t| - \sigma^2) I_{\{|t|>\sigma\}}$;
- Cauchy loss: $\phi_\sigma(t) = \sigma^2 \log\left(1 + t^2/\sigma^2\right)$;
- Tukey's biweight loss: $\phi_\sigma(t) = (\sigma^2/6)(1 - (1 - (t/\sigma)^2)^3) I_{\{|t|\le\sigma\}} + (\sigma^2/6) I_{\{|t|>\sigma\}}$.

In the above loss functions, I_S is an indicator function which takes the value 1 if S is true and gets the value 0 otherwise.

Recall that our previous analysis on the ℓ_σ loss and the MCCR model (2) relies heavily on Lemma 7. From the proof of Lemma 7, we know that similar analysis can be also applied to other robust loss functions that are sufficiently smooth and satisfy certain conditions, e.g., the Cauchy loss given above. On the other hand, although our analysis cannot cover all the robust loss functions, following from our previous analyzing process, we can still get a general view on the robust loss functions and see what happens when a robust loss function is employed from a learning theory viewpoint.

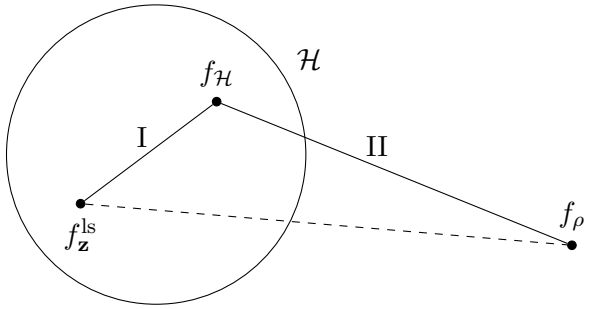

Figure 2: The statistical learning approach to bounding the $\mathcal{L}^2_{\rho_{\mathcal{X}}}$-distance between $f_{\mathbf{z}}$ and f_ρ for the ERM scheme (6), which is induced by the least squares loss.

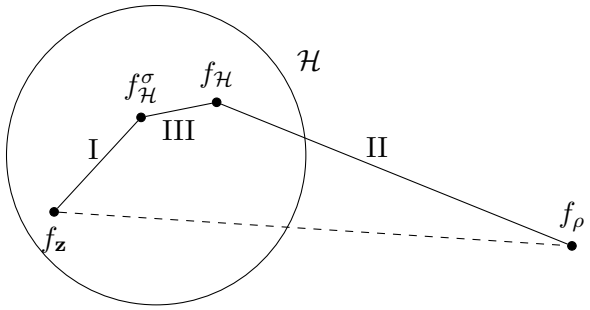

Figure 3: The statistical learning approach to bounding the $\mathcal{L}^2_{\rho_{\mathcal{X}}}$-distance between $f_{\mathbf{z}}$ and f_ρ for the ERM scheme induced by a robust loss function ϕ_σ.

To illustrate this, we first recall that to analyze the convergence of an ERM scheme associated with the least squares loss (e.g., the unconstrained regression model (6)), a typical statistical learning approach is proceeded as follows: instead of directly measuring the $\mathcal{L}^2_{\rho_\mathcal{X}}$-distance between $f^{\rm ls}_{\mathbf{z}}$ and f_ρ, one first introduces the projection of f_ρ in $\mathcal{H}$, i.e., $f_\mathcal{H}$. With the help of $f_\mathcal{H}$, one can decompose the distance into sample error and approximation error as follows:

$$\|f^{\rm ls}_{\mathbf{z}} - f_\rho\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}} \leq \|f^{\rm ls}_{\mathbf{z}} - f_\mathcal{H}\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}} + \|f_\mathcal{H} - f_\rho\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}}.$$

The idea of the above decomposition is depicted in Figure 2, where I represents the sample error $\|f^{\rm ls}_{\mathbf{z}} - f_\mathcal{H}\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}}$ while II gives the approximation error $\|f_\mathcal{H} - f_\rho\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}}$.

However, situations will be quite different if a robust regression loss ϕ_σ is employed. To explain this, we redefine $f^\sigma_\mathcal{H}$ as the target function of the regression model induced by a general robust loss ϕ_σ and $f_{\mathbf{z}}$ as the corresponding empirical target function, definitions of which are given as follows

$$f^\sigma_\mathcal{H} = \arg\min_{f\in\mathcal{H}} \int_\mathcal{Z} \phi_\sigma(y - f(x))d\rho \text{ and } f_{\mathbf{z}} = \arg\min_{f\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^m \phi_\sigma(y_i - f(x_i)).$$

The analysis in our study indicates that to analyze the convergence of a regression model induced by a robust loss function ϕ_σ, one may proceed via the following decomposition

$$\|f_{\mathbf{z}} - f_\rho\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}} \leq \|f_{\mathbf{z}} - f^\sigma_\mathcal{H}\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}} + \|f^\sigma_\mathcal{H} - f_\mathcal{H}\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}} + \|f_\mathcal{H} - f_\rho\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}}.$$

Figure 3 gives an intuitive description on the above decomposition. Similarly, in Figure 3, I represents the sample error term $\|f_{\mathbf{z}} - f^\sigma_\mathcal{H}\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}}$, II stands for the approximation error term $\|f_\mathcal{H} - f_\rho\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}}$ while III measures the $\mathcal{L}^2_{\rho_\mathcal{X}}$-distance between $f^\sigma_\mathcal{H}$ and $f_\mathcal{H}$. Notice that the bias term III is caused by the introduction of the scale parameter σ that delivers the robustness to the model. Due to the non-robustness of LSR and the fact that $f_\mathcal{H}$ is the target function of LSR, again we conclude that the smaller of the $\mathcal{L}^2_{\rho_\mathcal{X}}$-distance between $f^\sigma_\mathcal{H}$ and $f_\mathcal{H}$ is, the less robustness the regression model associated with the ϕ_σ loss possesses.

Taking the ℓ_σ loss for example, we know from our previous analysis that under very specific conditions the two points, $f^\sigma_\mathcal{H}$ and $f_\mathcal{H}$, meet and consequently the bias term III disappears. Technically speaking, a nice point of the ℓ_σ loss lies in that it is sufficiently smooth which makes it possible to bound the $\mathcal{L}^2_{\rho_\mathcal{X}}$-distance between $f^\sigma_\mathcal{H}$ and $f_\mathcal{H}$ explicitly. For instance, when the Moment Assumption holds and $f_\rho \in \mathcal{H}$, as a consequence of Lemma 7, we see that

$$\|f^\sigma_\mathcal{H} - f_\mathcal{H}\|^2_{\mathcal{L}^2_{\rho_\mathcal{X}}} \leq c_{\mathcal{H},\rho}/\sigma^2.$$

As mentioned in the previous section, the above estimate reveals that when the value of σ decreases, the upper bound of the bias term III increases.

Based on the above discussions, we conclude that when a robust loss function is employed in nonparametric regression problems, the enhancement of robustness is at the sacrifice of the convergence rate of the model and what one needs to do is to find a good compromise.

7. Numerical Experiments

Studies in this paper are motivated by empirical success of the MCCR model. However, for the sake of completeness, in this section, we carry out numerical experiments on synthetic and real data sets to show the effectiveness of the MCCR model (2).

7.1 Experimental Setup

Notice that the MCCR model (2) is a constrained optimization model since $\mathcal{H}$ is assumed to be a compact subset of $C(\mathcal{X})$. As mentioned previously, a typical choice of $\mathcal{H}$ is a bounded subset of a certain reproducing kernel Hilbert space $\mathcal{H}_K$ induced by some Mercer kernel K. However, to determine the diameter of this bounded subset in applications, prior information is usually required. In our experiments, instead of evaluating the optimization model (2), we focus on its unconstrained version

$$f_{\mathbf{z}} = \arg\min_{f\in\mathcal{H}_K} \frac{1}{m}\sum_{i=1}^{m} \ell_\sigma(y_i, f(x_i)) + \lambda\|f\|_K^2, \tag{21}$$

where λ is a positive regularization parameter.

The representer theorem ensures that we can search within the function set $\mathcal{H}_{K,\mathbf{z}}$ for the minimizer of the optimization model (21), where

$$\mathcal{H}_{K,\mathbf{z}} = \left\{ \sum_{i=1}^{m} \alpha_i K(x, x_i) + b,\, b\in\mathbb{R}, \alpha_i\in\mathbb{R},\, i = 1, \cdots, m \right\},$$

with b being an offset. In our experiments, we use the Gaussian kernel

$$K_h(x_i, x_j) = \exp\left(-\|x_i - x_j\|^2/h^2\right),$$

with the parameter h to be determined. To show the effectiveness of the MCCR model, we compare the empirical performance of (21) with other robust regression schemes, including robust regression models based on the Huber's loss and the least absolute deviation loss. These robust regression schemes are obtained by replacing the ℓ_σ loss in (21) with the Huber's loss and the least absolute deviation loss, respectively. Explicit definitions of the two loss functions are given as follows:

$$\phi_a^{\text{Huber}}(u,v) = \begin{cases} (u-v)^2, & \text{if } |u-v|\le a \\ 2a|u-v|, & \text{if } |u-v| > a \end{cases} \quad \text{and} \quad \phi^{\text{LAD}}(u,v) = |u-v|,\ u,v\in\mathbb{R}.$$

For notation simplifications, we denote the two robust regression models as Huber and LAD, respectively.

To solve (21), we apply the iteratively reweighted least squares method (IRLS). The basic procedure is to iteratively solve the weighted least squares problem and give weights according to the current solution. Due to the non-convexity of the MCCR model, solving (21) by using IRLS only guarantees a stationary point. In our experiment, we use the result of the least squares method as the starting point.

In our experiment, noise added to the toy examples is given as follows

$$\text{noise} := \tau_1\varepsilon_1 + \tau_2\varepsilon_2^p, \tag{22}$$

where ε_1 follows the standard Gaussian distribution and ε_2^p is an impulse noise (outliers) defined as

$$\text{Prob}(\varepsilon_2^p = t) = \begin{cases} 1-p, & t=0, \\ p/2, & t=1, \\ p/2, & t=-1. \end{cases}$$

τ_1 and τ_2 are introduced to set the variance of the Gaussian noise and the magnitude of the impulsive noise. In our experiment, we always set $p = 0.1$, i.e., 10% samples are contaminated by impulsive noise. In addition, in some of our experiments on synthetic data sets, we will also consider the noise ε_1 that is drawn from the Student's t-distribution with 3 degrees of freedom, and Cauchy distribution.

7.2 Example of the Noisy Sinc Function

We first choose the sinc function as the regression function. The one-dimension sinc function is given as

$$f(x) = \sin(\pi x)/(\pi x), \quad x \in [-4, 4], \tag{23}$$

which is frequently adopted to illustrate the regression models (see Vapnik, 1998; Suykens et al., 2002a,b; Schölkopf et al., 2000; Smola and Schölkopf, 2004).

In our experiment, we first draw a training set of size 100 from the sinc function (23) that are corrupted by the Gaussian noise. We then draw another training set with the same size corrupted by the Gaussian noise and the outliers. With each training set, the fitting results of the sinc function are plotted in Figure 4, in which the red dot-dashed curve is the one reconstructed by MCCR, the blue dashed curve represents the one from Huber while LAD gives the green dotted curve.

From Figure 4, one can see that all of the three models can fit the curve of the sinc function well when the data is only contaminated by the Gaussian noise. When the training data are also corrupted by outliers, all of the three robust regression models can still successfully reconstruct the curve. However, we can see that MCCR gives the best fitting results, especially at positions where data are corrupted by outliers.

7.3 Example of the Noisy Friedman's Benchmark Functions

Our second numerical experiment on toy examples considers multiple dimensional regression problems. We now use the Friedman's benchmark functions as our test functions, which were introduced in Friedman (1991) and have become widely employed models when studying regression problems (see Tipping, 2001; Brown et al., 2005; Debruyne et al., 2010).

The Friedman's benchmark functions are listed as follows:

- $f_1(x) = 10\sin(\pi x^1 x^2) + 20(x^3 - 0.5)^2 + 10x^4 + 5x^5$;
- $f_2(x) = \sqrt{(x^1)^2 + \left(x^2 x^3 - 1/(x^2 x^4)\right)^2}$;
- $f_3(x) = \arctan\left(1/x^1 \left(x^2 x^3 - 1/(x^2 x^4)\right)\right)$.

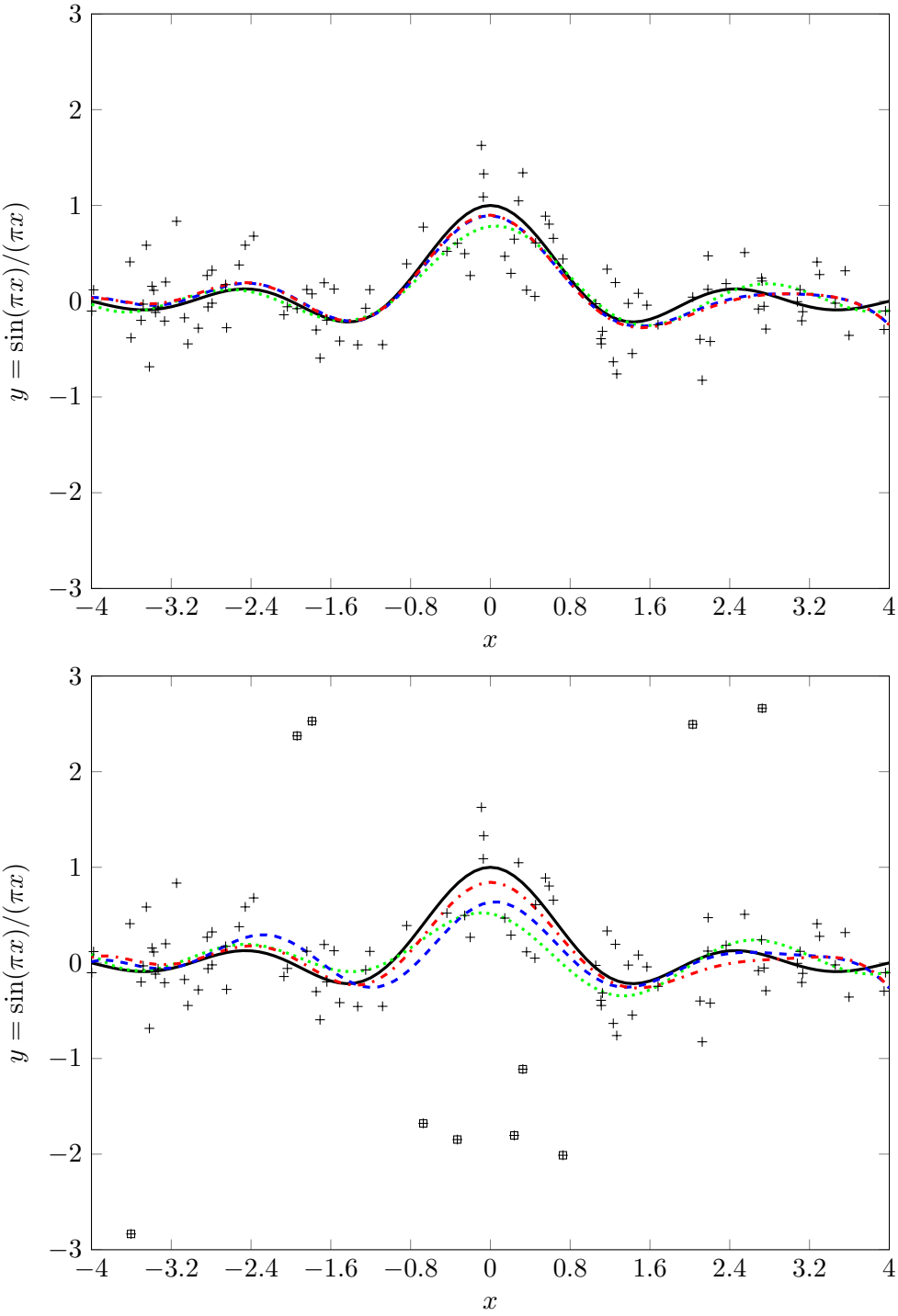

Figure 4: Sinc function (black solid curves) and the regression results (MCCR: red dot-dashed curve; Huber: blue dashed curve; LAD: green dotted curve). (top) The training data (crosses) are corrupted by Gaussian noise; (bottom) Some observed data are outliers (marked by squares).

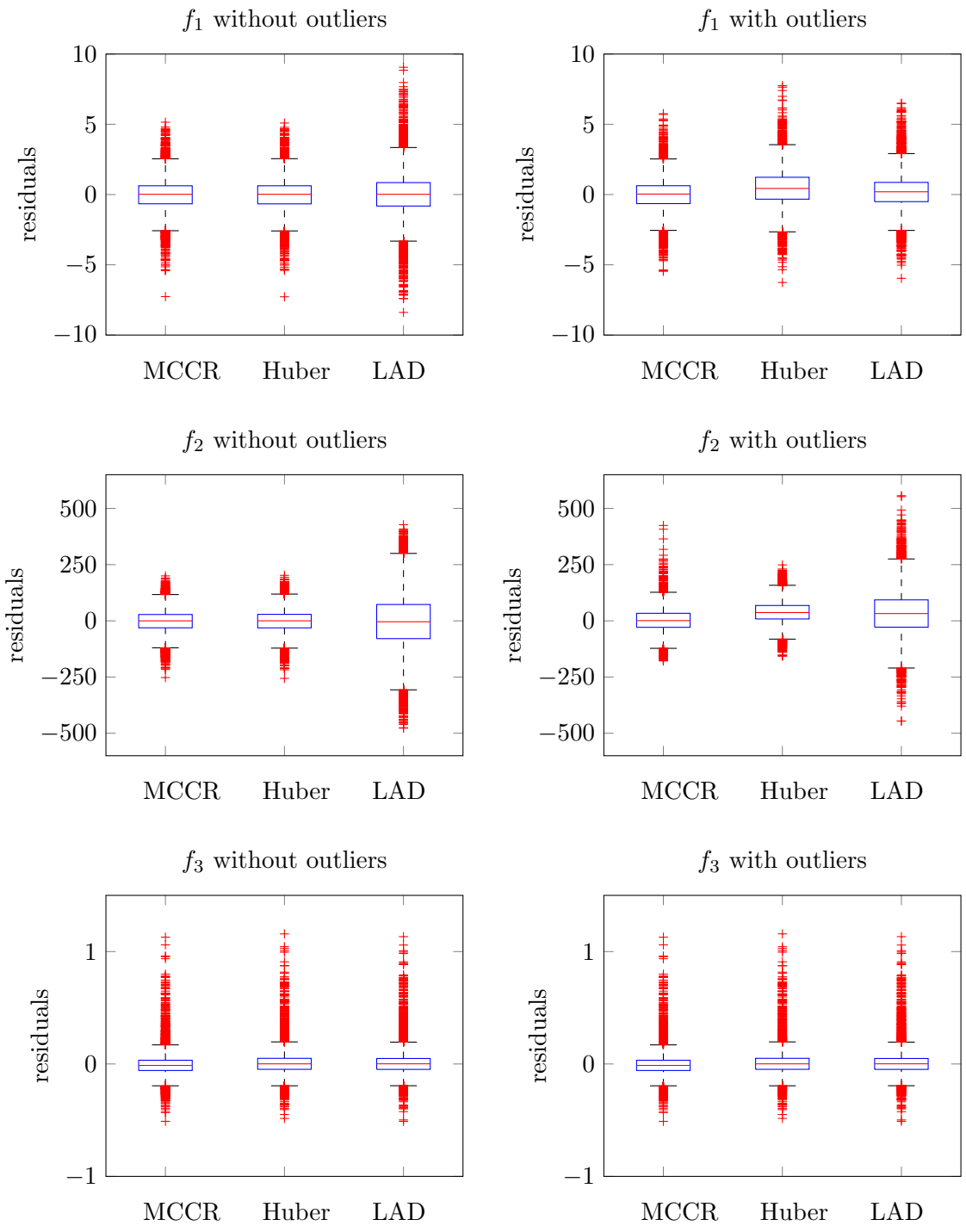

Figure 5: Box-plots of the residuals of Friedman's benchmark functions for the case of Gaussian noise. Each box-plot features a lower quartile (25 percentile) line, a median (50 percentile) line and an upper quartile (75 percentile) line for the residuals on test data.

For f_1, $x = (x^1, \ldots, x^{10})$ where each x^j, $j = 1, \ldots, 10$, is uniformly distributed in $[0, 1]$ and x^6, $\ldots$, x^{10} are noisy variables. For f_2 and f_3, $x = (x^1, x^2, x^3, x^4)$ and each is uniformly distributed in the following intervals: $x^1 \in [0, 100]$, $x^2 \in [40\pi, 560\pi]$, $x^3 \in [0, 1]$ and $x^4 \in [1, 11]$.

For each function, 1000 observations are randomly taken from corresponding domain for training and cross-validating. Another independent 1000 observations are also randomly drawn as the test set. Noise and outliers are then added according to (22). For f_1, we set $\tau_1 = 1$. For f_2 and f_3, τ_1 is set such that the ratio of the signal power to the power of ε_1 is 3. In the outlier-free cases, we set $\tau_2 = 0$. To observe the performance for the three models in the presence of outliers in the training data sets, we set $\tau_2 = \max_{x \in D} f(x) - \min_{x \in D} f(x)$, where D is the domain of each benchmark function. For each regression model, the width of the Gaussian kernel h, the regularization parameter λ and the scale parameter in the loss function (no scale parameter for the LAD loss) are all tuned via a 10-fold cross-validation under the mean squared error criterion. The residuals $\{y_i - f(x_i)\}_{i=1}^{1000}$ are recorded. For the case of Gaussian noise, we boxplot all the residuals in Figure 5. Each box-plot features a lower quartile (25 percentile) line, a median (50 percentile) line and an upper quartile (75 percentile) line.

In Table 1, we also report the relative sum of squared error (RSSE) on the test data set T, i.e.,

$$\mathrm{RSSE}(\hat{f}) = \sum_{x \in T} \Big(f(x) - \hat{f}(x) \Big)^2 \Big/ \sum_{x \in T} \big(f(x) - \bar{f}_T \big)^2 ,$$

where $\bar{f}_T$ is the mean value of $f(x)$ on T.

test function	noise	MCCR	Huber	LAD
f_1	Gaussian noise, no outliers	**0.048**	0.049	0.103
	Gaussian noise, outliers	**0.062**	0.073	0.157
f_2	Gaussian noise, no outliers	**0.020**	0.021	0.136
	Gaussian noise, outliers	**0.023**	0.032	0.156
f_3	Gaussian noise, no outliers	**0.091**	0.117	0.136
	Gaussian noise, outliers	**0.062**	0.073	0.157
f_1	Cauchy noise, no outliers	**0.042**	0.042	0.116
	Cauchy noise, outliers	**0.045**	0.049	0.089
f_2	Cauchy noise, no outliers	**0.005**	0.005	0.025
	Cauchy noise, outliers	**0.006**	0.006	0.021
f_3	Cauchy noise, no outliers	0.180	0.195	**0.177**
	Cauchy noise, outliers	0.219	**0.143**	0.154
f_1	Student noise, no outliers	**0.040**	0.040	0.101
	Student noise, outliers	**0.046**	0.075	0.092
f_2	Student noise, no outliers	**0.017**	0.017	0.129
	Student noise, outliers	**0.023**	0.024	0.123
f_3	Student noise, no outliers	**0.423**	0.429	0.430
	Student noise, outliers	0.471	0.544	**0.434**

Table 1: The relative sum of squared error on the test data

7.4 Evaluation on Real Data Sets

We also evaluate the three robust regression models on four real data sets downloaded from UCI repository of machine learning databases (see Bache and Lichman, 2013): Concrete Compressive Strength Data Set, Housing Data Set, Yacht Hydrodynamics Data Set and Airfoil Self-Noise Data Set.

For each data set, two third of the instances are used for training and the remaining are used for test. We repeat our experiment as done for the Friedman's benchmark functions for ten times. The residuals for the three robust regression models are displayed by box-plots in Figure 6, the accuracy of which are measured by RSSE. Experimental results on the RSSEs and the details of training data, including the size of features n and the size of instances m, are reported in Table 2.

data sets	n	m	MCCR	Huber	LAD
concrete	9	686	**0.061**	0.061	0.062
house	14	338	0.128	**0.126**	0.175
yacht-hydrodynamics	7	205	**0.022**	0.024	0.159
airfoil	6	1000	**0.184**	0.195	0.238

Table 2: The relative sum of squared error on real data

In the above numerical evaluations on toy examples and real data sets, our experiments show that when the data is only contaminated by Gaussian noise, a large sigma value in the MCCR model and a large a value in the regression model based on the Huber's criterion will be selected via cross-validation. However, for other noise and in the presence and absence of outliers, smaller values of the scale parameters in the two regression models will be selected. These coincide with our understandings on the robust regression models.

From the above experimental results, we can see the effectiveness of MCCR especially for the cases in the presence of impulsive noise.

8. Concluding Remarks

In this paper, we presented a statistical learning interpretation of the regression model associated with the correntropy induced regression loss. We investigated its connections with the least squares regression. We found that the correntropy induced loss could help for regression with non-Gaussian noise. Meanwhile, comparable performance could be obtained by applying this regression model when the noise is Gaussian. Convergence rates of the proposed model under various circumstances were derived explicitly. We showed that the scale parameter in the loss function balanced the convergence rates and the robustness of the model. We also made some efforts to extend our analysis to other robust loss functions and gave a general view on analyzing regression models induced by general robust loss functions. It is expected that our observations can shed some light towards future real-life applications.

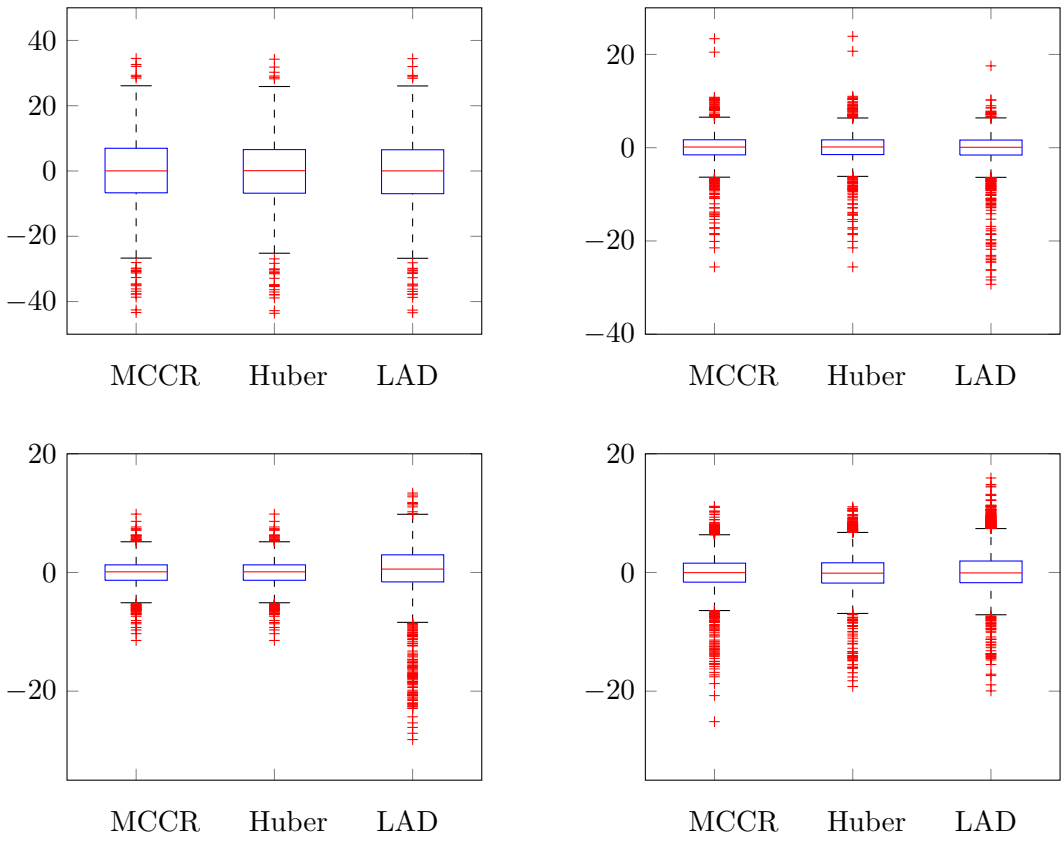

Figure 6: Box-plots of the residuals on four real data sets. Each box-plot features a lower quartile (25 percentile) line, a median (50 percentile) line and an upper quartile (75 percentile) line for the residuals on test data. (top left) concrete; (top right) Boston house; (bottom left) yacht hydrodynamics; (bottom right) airfoil.

Acknowledgments

The authors would like to thank the action editor and the reviewers for their insightful comments and constructive suggestions, which improved the quality of this paper. The authors would also like to thank Dr. Jun Fan from Department of Statistics, University of Wisconsin-Madison for helpful discussions.

EU: The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC AdG A-DATADRIVE-B (290923). This paper reflects only the authors' views, the Union is not liable for any use that may be made of the contained information. Research Council KUL: GOA/10/09 MaNet, CoE PFV/10/002 (OPTEC), BIL12/11T; PhD/Postdoc grants. Flemish Government: FWO: projects: G.0377.12 (Structured systems), G.088114N (Tensor based data similarity); PhD/Postdoc grants. IWT: projects: SBO POM (100031); PhD/Postdoc grants. iMinds Medical Information Technologies SBO 2014. Belgian Federal Science Policy Office: IUAP P7/19 (DYSCO, Dynamical systems, control and optimization, 2012-2017). L. Shi is supported by the National Natural Science Foundation of China Project No. 11201079), the Joint Research Fund by National Natural Science Foundation of China and Research Grants Council of Hong Kong (Project No. 11461161006 and Project No. CityU 104012) and the Fundamental Research Funds for the Central Universities of China (Project No. 20520133238, Project No. 20520131169). Johan Suykens is a professor at KU Leuven, Belgium.

References

Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge, 1999.

Jean-Yves Audibert and Olivier Catoni. Robust linear least squares regression. *Annals of Statistics*, 39(5):2766–2794, 2011.

Kevin Bache and Moshe Lichman. UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`.

Sergei N. Bernstein. *The Theory of Probabilities*. Gastehizdat Publishing House, Moscow, 1946.

Ricardo J. Bessa, Vladimiro Miranda, and João Gama. Entropy and correntropy against minimum square error in offline and online three-day ahead wind power forecasting. *Power Systems, IEEE Transactions on*, 24(4):1657–1666, 2009.

Gavin Brown, Jeremy L. Wyatt, and Peter Tiňo. Managing diversity in regression ensembles. *Journal of Machine Learning Research*, 6:1621–1650, 2005.

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

Andreas Christmann and Arnout Van Messem. Bouligand derivatives and robustness of support vector machines for regression. *Journal of Machine Learning Research*, 9:915–936, 2008.

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

Felipe Cucker and Ding-Xuan Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge University Press, Cambridge, 2007.

Patrick L. Davies. Aspects of robust linear regression. *Annals of Statistics*, 21(4):1843–1899, 1993.

Kris De Brabanter, Kristiaan Pelckmans, Jos De Brabanter, Michiel Debruyne, Johan A. K. Suykens, Mia Hubert, and Bart De Moor. Robustness of kernel based regression: a comparison of iterative weighting schemes. In *International Conference on Artificial Neural Networks–ICANN 2009*, pages 100–110. Springer, 2009.

Michiel Debruyne, Mia Hubert, and Johan A. K. Suykens. Model selection in kernel based regression using the influence function. *Journal of Machine Learning Research*, 9:2377–2400, 2008.

Michiel Debruyne, Andreas Christmann, Mia Hubert, and Johan A. K. Suykens. Robustness of reweighted least squares kernel based regression. *Journal of Multivariate Analysis*, 101 (2):447–463, 2010.

John E. Dennis and Roy E. Welsch. Techniques for nonlinear least squares and robust regression. *Communications in Statistics-Simulation and Computation*, 7(4):345–359, 1978.

Jun Fan, Ting Hu, Qiang Wu, and Ding-Xuan Zhou. Consistency analysis of an empirical minimum error entropy algorithm. *Applied and Computational Harmonic Analysis, in press*. doi: 10.1016/j.acha.2014.12.005.

Jerome H. Friedman. Multivariate adaptive regression splines. *Annals of Statistics*, 19(1): 1–141, 1991.

Aysegul Gunduz and José C. Príncipe. Correntropy as a novel measure for nonlinearity tests. *Signal Processing*, 89(1):14–23, 2009.

Zheng-Chu Guo and Ding-Xuan Zhou. Concentration estimates for learning with unbounded sampling. *Advances in Computational Mathematics*, 38(1):207–223, 2013.

Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. John Wiley & Sons, New York, 1986.

Ran He, Wei-Shi Zheng, and Bao-Gang Hu. Maximum correntropy criterion for robust face recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(8): 1561–1576, 2011a.

Ran He, Wei-Shi Zheng, Bao-Gang Hu, and Xiang-Wei Kong. A regularized correntropy framework for robust pattern recognition. *Neural Computation*, 23(8):2074–2100, 2011b.

Paul W. Holland and Roy E. Welsch. Robust regression using iteratively reweighted least-squares. *Communications in Statistics-Theory and Methods*, 6(9):813–827, 1977.

Ting Hu, Jun Fan, Qiang Wu, and Ding-Xuan Zhou. Learning theory approach to minimum error entropy criterion. *Journal of Machine Learning Research*, 14(1):377–397, 2013.

Ting Hu, Jun Fan, Qiang Wu, and Ding-Xuan Zhou. Regularization schemes for minimum error entropy principle. *Analysis and Applications*, 2014.

Peter J. Huber. *Robust Statistics*. John Wiley & Sons, New York, 1981.

Weifeng Liu, Puskal P. Pokharel, and José C. Príncipe. Correntropy: properties and applications in non-gaussian signal processing. *Signal Processing, IEEE Transactions on*, 55 (11):5286–5298, 2007.

Shahar Mendelson and Joseph Neeman. Regularization in kernel learning. *Annals of Statistics*, 38(1):526–565, 2010.

José C. Príncipe. *Information Theoretic Learning: Rényi's Entropy and Kernel Perspectives*. Springer, New York, 2010.

Ignacio Santamaría, Puskal P. Pokharel, and José C. Príncipe. Generalized correlation function: definition, properties, and application to blind equalization. *Signal Processing, IEEE Transactions on*, 54(6):2187–2197, 2006.

Bernhard Schölkopf, Alex J. Smola, Robert C. Williamson, and Peter L. Bartlett. New support vector algorithms. *Neural Computation*, 12(5):1207–1245, 2000.

Steve Smale and Ding-Xuan Zhou. Estimating the approximation error in learning theory. *Analysis and Applications*, 1(01):17–41, 2003.

Alex J. Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, 2004.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer, New York, 2008.

Ingo Steinwart, Don Hush, and Clint Scovel. Optimal rates for regularized least squares regression. In *Proceedings of the 22nd Conference on Learning Theory*, 2009.

Johan A. K. Suykens, Jos De Brabanter, Lukas Lukas, and Joos Vandewalle. Weighted least squares support vector machines: robustness and sparse approximation. *Neurocomputing*, 48(1):85–105, 2002a.

Johan A. K. Suykens, Tony Van Gestel, Jos De Brabanter, Bart De Moor, and Joos Vandewalle. *Least Squares Support Vector Machines*. World Scientific, Singapore, 2002b.

Michael E. Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001.

Vladimir Vapnik. *Statistical Learning Theory*. John Wiley & Sons, New York, 1998.

Cheng Wang and Ding-Xuan Zhou. Optimal learning rates for least squares regularized regression with unbounded sampling. *Journal of Complexity*, 27(1):55–67, 2011.

Xueqin Wang, Yunlu Jiang, Mian Huang, and Heping Zhang. Robust variable selection with exponential squared loss. *Journal of the American Statistical Association*, 108(502): 632–643, 2013.

Qiang Wu, Yiming Ying, and Ding-Xuan Zhou. Learning rates of least-square regularized regression. *Foundations of Computational Mathematics*, 6(2):171–192, 2006.

Qiang Wu, Yiming Ying, and Ding-Xuan Zhou. Multi-kernel regularized classifiers. *Journal of Complexity*, 23(1):108–134, 2007.

Ding-Xuan Zhou. The covering number in learning theory. *Journal of Complexity*, 18(3): 739–767, 2002.