mathematics

MDPI

Article

Huber Regression Analysis with a Semi-Supervised Method

Yue Wang [1], Baobin Wang [1], Chaoquan Peng [1], Xuefeng Li [1] and Hong Yin [2,*]

[1] School of Mathematics and Statistics, South-Central Minzu University, Wuhan 430074, China
[2] School of Mathematics, Renmin University of China, Beijing 100872, China
* Correspondence: yinhong@ruc.edu.cn

Abstract: In this paper, we study the regularized Huber regression algorithm in a reproducing kernel Hilbert space (RKHS), which is applicable to both fully supervised and semi-supervised learning schemes. Our focus in the work is two-fold: first, we provide the convergence properties of the algorithm with fully supervised data. We establish optimal convergence rates in the minimax sense when the regression function lies in RKHSs. Second, we improve the learning performance of the Huber regression algorithm by a semi-supervised method. We show that, with sufficient unlabeled data, the minimax optimal rates can be retained if the regression function is out of RKHSs.

Keywords: robust regression; Huber loss function; reproducing kernel Hilbert space; semi-supervised data

MSC: 62J02

1. Introduction

The ordinary least squares (OLS) is an important statistical tool applied in regression analysis. However, OLS does not perform well when the data are contaminated by the occurrence of outliers or heavy-tailed noise. Thus, OLS is suboptimal in the robust regression analysis and a variety of robust loss functions have been developed that are not so easily affected by noises. Among them, Huber loss function is usually a popular choice in the fields of statistics, machine learning and optimization since it is less sensitive to outliers and can address the issue of heavy-tailed errors effectively. Huber regression was initiated by Peter Huber in his seminal work [1,2]. Statistical bounds and convergence properties for Huber estimation and inference have been further investigated in the subsequent works. See, e.g., [3–9].

Semi-supervised learning has been gaining increased attention as an active research area in the fields of science and engineering. The original idea of semi-supervised method can date back to self-learning in the context of classification [10] and then is well developed in decision-directed learning, co-training in text classification, and manifold learning [11–13]. Most existing research on Huber regression work is in the supervised framework. Unlabeled data had been deemed useless and thus thrown away in the design of algorithms. Recently, it has been shown in vast literature that utilizing the additional information in unlabeled data can effectively improve the learning performance of algorithms. See, e.g., [14–18]. In this paper, we focus on the Huber regression algorithm performance with unlabeled data. By the semi-supervised method, we find that optimal learning rates are available if sufficient unlabeled data are added in the Huber regression analysis.

In the standard framework of statistical learning, we let the explanatory variable X take values in a compact domain $\mathcal{X}$ in a Euclidean space, and the response variable Y takes values in the output space $\mathcal{Y} \subset \mathbb{R}$. This work investigates the application of the Huber loss that is linked to the following regression model:

$$Y = f^*(X) + \epsilon,$$

Citation: Wang, Y.; Wang, B.; Peng, C.; Li, X.; Yin, H. Huber Regression Analysis with a Semi-Supervised Method. *Mathematics* **2022**, *10*, 3734. https://doi.org/10.3390/math10203734

Academic Editor: Kai-Tai Fang

Received: 27 August 2022
Accepted: 8 October 2022
Published: 11 October 2022

where f^* is the regression function and ϵ is the noise in the regression model. Let ρ be a Borel probability measure on the product space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Let $\rho_{\mathcal{X}}$ and $\rho(y|x)$ and denote the marginal distribution of ρ on $\mathcal{X}$, and the conditional distribution on $\mathcal{Y}$ given $x \in \mathcal{X}$, respectively. In the supervised learning setting, ρ is assumed to be unknown and the purpose of regression is to estimate $f^*(X)$ according to a sample $D = \{(x_i, y_i)\}_{i=1}^N$ drawn independently from ρ, where N is the sample size, the cardinality of D. The Huber loss function $\ell_\sigma(\cdot)$ is defined as

$$\ell_\sigma(u) = \begin{cases} u^2, & \text{if } |u| \le \sigma, \\ 2\sigma|u| - \sigma^2, & \text{if } |u| > \sigma, \end{cases}$$

where $\sigma > 0$ is a robustification parameter. Given the prediction function $f : \mathcal{X} \to \mathcal{Y}$, Huber regression searches for a good approximation of $f^*(X)$ by minimizing the empirical prediction error with the Huber loss

$$\mathcal{E}_D(f) := \frac{1}{N} \sum_{i=1}^{N} \ell_\sigma(y_i - f(x_i)) \tag{1}$$

over a suitable hypothesis space.

In this work, we study the kernel based Huber regression algorithm and the minimization of (1) performs in a reproducing kernel Hilbert space (RKHS) [19]. Recall that $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a Mercer kernel if it is continuous, symmetric, and positive semidefinite. The RKHS $\mathcal{H}_K$ is the completion of the linear span of the function set $\{K_x = K(x, \cdot), x \in \mathcal{X}\}$ with the inner product induced by $\langle K_x, K_y \rangle_K = K(x, y)$. The reproducing property is given by $f(x) = \langle f, K_x \rangle_K$. Note that, by Cauchy–Schwarz inequality and [19],

$$\|f\|_\infty = \sup_{x \in \mathcal{X}} |\langle f, K_x \rangle_K| \le \sup_{x \in \mathcal{X}} \|f\|_K \|K_x\|_K = \sup_{x \in \mathcal{X}} \sqrt{K(x, x)} \|f\|_K.$$

To avoid overfitting, the regularized Huber regression algorithm in the RKHS $\mathcal{H}_K$ is given as

$$f_{D,\lambda} = \arg \min_{f \in \mathcal{H}_K} \left\{ \mathcal{E}_D(f) + \lambda \|f\|_K^2 \right\}, \tag{2}$$

where $\lambda > 0$ is a regularization parameter.

In this paper, we derive the explicit learning rate of Algorithm (2) in the supervised learning, which is comparable to the minimax optimal rate of OLS. By a semi-supervised method, we show that utilizing unlabeled data can conquer the bottleneck that optimal learning rates for algorithm (2) are only achievable when f^* lies in $\mathcal{H}_K$.

2. Assumptions and Main Results

To present our main results, we introduce some necessary assumptions. In this section, we study the convergence of $f_{D,\lambda}$ to f^* in the square integrable space $(L^2_{\rho_{\mathcal{X}}}, \|\cdot\|_\rho)$.

Below, we elaborate on three important assumptions to carry out the analysis. The first assumption (3) is about the regularity of the regression function f^*. Define the integral operator $L_K : L^2_{\rho_{\mathcal{X}}} \to L^2_{\rho_{\mathcal{X}}}$ associated with the kernel K by

$$L_K f := \int_{\mathcal{X}} f(x) K_x d\rho_{\mathcal{X}}(x), \quad \forall f \in L^2_{\rho_{\mathcal{X}}}.$$

Since K is a Mercer kernel on the compact domain $\mathcal{X}$, L_K is compact and positive. Thus, L_K^r as the r-th power of L_K for $r > 0$ is well defined [20]. Our error bounds are stated in terms of the regularity of f^*, given by

$$f^* = L_K^r(h), \quad \text{for some } r > 0 \text{ and } h \in L^2_{\rho_{\mathcal{X}}}. \tag{3}$$

The condition (3) characterizes the regularity of f^* and is directly related to the smoothness of f^* when $\mathcal{H}_K$ is a Sobolev space. If (3) holds with $r \geq \frac{1}{2}$, f^* lies in the space $\mathcal{H}_K$ [21].

The second assumption (4) is about the capacity of $\mathcal{H}_K$, measured by the *effective dimension* [22–24]

$$\mathcal{N}(\lambda) = \mathsf{Trace}((L_K + \lambda I)^{-1} L_K), \quad \text{for } \lambda > 0,$$

where I is the identity operator on $\mathcal{H}_K$. In this paper, we assume that

$$\mathcal{N}(\lambda) \leq C\lambda^{-s} \quad \text{for some } C > 0, 0 < s \leq 1. \tag{4}$$

This condition measures the complexity of $\mathcal{H}_K$ with respect to the marginal distribution $\rho_{\mathcal{X}}$. It is typical in the analysis of the performances of kernel methods' estimators. It is always satisfied with $s = 1$ by taking the constant $C = \mathsf{Trace}(L_K)$. When $\mathcal{H}_K$ is a Sobolev space $W^{\alpha}(\mathcal{X}), \mathcal{X} \subset \mathbb{R}^n$ with all derivatives of an order up to $\alpha > \frac{n}{2}$, then (4) is satisfied with $s = \frac{n}{2\alpha}$ [25]. When $0 < s < 1$, (4) is weaker than the eigenvalue decaying assumption in the literature [17,23].

The third assumption is about the conditional probability distribution $\rho(y|x)$ on the output space $\mathcal{Y}$. We assume that the output variable Y satisfies the *moment condition* when there exist two positive numbers $t, M > 0$ such that, for any integer $q \geq 2$,

$$\mathbb{E}(|Y|^q|X) \leq \frac{1}{2} q! t^2 M^{q-2}. \tag{5}$$

The assumption (5) covers many common distributions, for example, Gaussian, sub-Gaussian, and the distributions with compact support [26].

Now, we are ready to present the main results of this paper. Without loss of generality, we assume $\sup_{x \in \mathcal{X}} K(x, x) = 1$.

2.1. Convergence in the Supervised Learning

The following error estimate for Algorithm (2) is the first result of this section, which presents the convergence of Huber regression with fully supervised data and will be proved in Section 3.

Theorem 1. *Define $f_{D,\lambda}$ by Algorithm (2) with the fully supervised data set $D = \{(x_i, y_i)\}_{i=1}^N$. Suppose that (3) holds for some $r > 0$, (4) and (5). If*

$$\lambda = \begin{cases} N^{-\frac{1}{1+s}}, & \text{for } 0 < r < \frac{1}{2}, \\ N^{-\frac{1}{s+2\min\{1,r\}}}, & \text{for } r \geq \frac{1}{2}, \end{cases} \tag{6}$$

then, for any $0 < \delta < 1$, with probability $1 - \delta$,

$$\|f_{D,\lambda} - f^*\|_\rho \leq C_1 \max\left\{\lambda^{\min\{r,1\}}, \sigma^{-1}\lambda^{-\frac{3}{2}} (\log N)^4\right\} \left(\log \frac{8}{\delta}\right)^4, \tag{7}$$

where C_1 is a constant independent of N, δ, or σ.

The above theorem shows that the parameter σ in the Huber loss ℓ_σ balances the robustness of Algorithm (2) and its convergence rates. We can see that, when the Huber loss function is employed in nonparametric regression problems, the enhancement of robustness occurs with the sacrifice of the convergence rate of Algorithm (2). Thus, what one needs to do is to find a trade-off. It is then direct to obtain the following corollary that provides the explicit learning rates for (2) with a suitable choice of σ.

Corollary 1. *Under the same conditions of Theorem 1, if* $\sigma \geq \left[\lambda^{-r-\frac{3}{2}}(\log N)^4\right]$, *then with probability at least* $1-\delta$,

$$\|f_{D,\lambda} - f^*\|_\rho = \begin{cases} \mathcal{O}\left(N^{-\frac{r}{s+1}}\left(\log\frac{8}{\delta}\right)^4\right), & \text{for } 0 < r < \frac{1}{2}, \\ \mathcal{O}\left(N^{-\min\left\{\frac{r}{s+2r},\frac{1}{s+2}\right\}}\left(\log\frac{8}{\delta}\right)^4\right), & \text{for } r \geq \frac{1}{2}. \end{cases} \tag{8}$$

Remark 1. *The above corollary tells us that, when* $\frac{1}{2} \leq r \leq 1$, *Algorithm (2) achieves the error rate* $\mathcal{O}\left(N^{-\frac{r}{2r+s}}\right)$, *which coincides with the minimax lower bound proved in [23,25], and is optimal. We also notice that the convergence rate can not improve when* $r > 1$. *It is referred to as the saturation phenomenon, which has been found in a vast amount of literature [20,22,25].*

2.2. Convergence in the Semi-Supervised Learning

Although optimal convergence rates of the Algorithm (2) were deduced when f^* lies in $\mathcal{H}_K$ ($r \geq \frac{1}{2}$) in the previous subsection, the error rate for the case $0 < r < \frac{1}{2}$ needs improvements. In this subsection, we study the influence of unlabeled data on the convergence of (2) by using semi-supervised data.

Let an unlabeled data set $\tilde{D}(x) = \{\tilde{x}_i\}_{i=1}^{\tilde{N}}$ be drawn independently according to the marginal distribution $\rho_{\mathcal{X}}$, where $\tilde{N}$ is the cardinality of $\tilde{D}(x)$. With the fully supervised data set $D = \{(x_i, y_i)\}_{i=1}^{N}$, we then introduce the supervised data set associated with Huber regression problems as $D^* = \{(x_i^*, y_i^*)\}_{i=1}^{N+\tilde{N}}$, given by

$$(x_i^*, y_i^*) = \begin{cases} (x_i, \frac{N+\tilde{N}}{N} y_i), & \text{for } 1 \leq i \leq N, \\ (\tilde{x}_{i-N}, 0), & \text{for } N+1 \leq i \leq N+\tilde{N}. \end{cases} \tag{9}$$

By replacing D with D^* in Algorithm (2), we then obtain the output function $f_{D^*,\lambda}$ with semi-supervised data D^*. The enhanced convergence results are as follows.

Theorem 2. *Suppose (3), (4) and (5) hold for* $0 < r \leq 1$, $r + s \geq \frac{1}{2}$, *and* $\tilde{N} \geq \max\{N^{\frac{s+1}{2r+s}} - N + 1, 1\}$. *If* $\lambda = N^{-\frac{1}{2r+s}}$, *then, with a probability at least* $1-\delta$,

$$\|f_{D^*,\lambda} - f_\rho\|_\rho \leq C_2 \max\left\{N^{-\frac{r}{2r+s}}, \frac{\Delta_{N,\tilde{N},\lambda}(\log N)^4}{\sqrt{\lambda}\sigma}\right\} \log\left(\frac{8}{\delta}\right)^4, \tag{10}$$

where

$$\Delta_{N,\tilde{N},\lambda} = \frac{N+\tilde{N}}{\lambda N} + \left(\frac{N+\tilde{N}}{N}\right)^2$$

and C_2 *is a constant independent of* $N, \tilde{N}, \sigma$, *or* δ.

Based on the theorem above, we can obtain the improved convergence rate as follows.

Corollary 2. *Under the same conditions of Theorem 2, if*

$$\sigma \geq N^{\frac{2r+1}{2(2r+s)}} \Delta_{N,\tilde{N},\lambda}(\log N)^4, \tag{11}$$

then, with probability $1-\delta$,

$$\|f_{D^*,\lambda} - f_\rho\|_\rho = \mathcal{O}\left(N^{-\frac{r}{2r+s}}\left(\log\frac{8}{\delta}\right)^4\right). \tag{12}$$

Remark 2. *Corollary 1 shows that, provided no unlabeled data are involved, the minimax optimal convergence rate for (2) is obtained only in the situation* $r > \frac{1}{2}$. *When* $0 < r \leq \frac{1}{2}$, *the rate reduces to* $\mathcal{O}\left(N^{-\frac{r}{s+1}}\right)$. *It implies that the regression function* f^* *is assumed to belong to* $\mathcal{H}_K$ *for achieving the optimal rate, which is difficult to verify in practice. In contrast, Corollary 2 tells us that, with sufficient unlabeled data* $\tilde{D}(x)$ *engaged in Algorithm (2), the minimax optimal rate* $\mathcal{O}\left(N^{-\frac{r}{2r+s}}\right)$ *is retained for* $0 < r \leq 1$. *This removes the strict regularity condition on* f^*.

3. Proofs

Now, we are in a position of proving results stated in Section 2.

3.1. Useful Estimates

First, we will estimate the bound of $f_{D,\lambda}$ defined by (2). In the sequel, for notational simplicity, let $z = (x, y)$ and define the empirical operator $L_{K,D} : \mathcal{H}_K \to \mathcal{H}_K$ by

$$L_{K,D} := \frac{1}{N}\sum_{i=1}^{N}\langle \cdot, K_{x_i}\rangle_K K_{x_i},\ z_i = (x_i, y_i) \in D,$$

so, for any $f \in \mathcal{H}_K$, $L_{K,D}f = \frac{1}{N}\sum_{i=1}^{N} f(x_i)K_{x_i}$. Then, we have the following representation for $f_{D,\lambda}$.

Lemma 1. *Define* $f_{D,\lambda}$ *by (2). Then, it satisfies*

$$f_{D,\lambda} = (L_{K,D} + \lambda I)^{-1}\hat{f}_{\rho,D} + (L_{K,D} + \lambda I)^{-1}W_{D,\lambda} \tag{13}$$

where

$$\hat{f}_{\rho,D} = \frac{1}{N}\sum_{i=1}^{N} y_i K_{x_i},\ z_i = (x_i, y_i) \in D$$

and

$$W_{D,\lambda} = \frac{1}{N}\sum_{i=1}^{N}\left[G'_+\left(\frac{(f_{D,\lambda}(x_i) - y_i)^2}{\sigma^2}\right) - G'_+(0)\right](f_{D,\lambda}(x_i) - y_i)K_{x_i}$$

with

$$G(s) = \begin{cases} s, & \text{if } 0 \leq s \leq 1, \\ 2s^{\frac{1}{2}} - 1, & \text{if } s \geq 1. \end{cases}$$

Proof. Note that $\ell_\sigma(u) = \sigma^2 G\left(\frac{u^2}{\sigma^2}\right)$. Since $f_{D,\lambda}$ is the minimizer of Algorithm (2), we take the gradient of the regularized functional on $\mathcal{H}_K$ to give

$$\frac{1}{N}\sum_{i=1}^{N} G'_+\left(\frac{(f_{D,\lambda}(x_i) - y_i)^2}{\sigma^2}\right)(f_{D,\lambda}(x_i) - y_i)K_{x_i} + \lambda f_{D,\lambda} = 0.$$

With the fact $G'_+(0) = 1$, it yields

$$\frac{1}{N}\sum_{i=1}^{N}(f_{D,\lambda}(x_i) - y_i)K_{x_i} + \lambda f_{D,\lambda} - W_{D,\lambda} = 0,$$

which is $(L_{K,D} + \lambda I)f_{D,\lambda} - \hat{f}_{\rho,D} - W_{D,\lambda} = 0$.

The proof is complete. □

Based on the above lemma, we can obtain the bound of $f_{D,\lambda}$.

Lemma 2. *Under the moment condition (5), with a probability at least* $1-\delta$*, there holds*

$$\|f_{D,\lambda}\|_K \leq (4M+5t)\lambda^{-\frac{1}{2}} \log \frac{N}{\delta}. \tag{14}$$

Proof. Under the moment condition (5), it has been proven in [27] that, with a probability of at least $1-\delta$, there holds

$$\max\{|y| : \text{ there exists an } x \in \mathcal{X}, \text{such that } (x,y) \in D\} \leq (4M+5t) \log \frac{N}{\delta}. \tag{15}$$

By the definition of $f_{D,\lambda}$, we have that $\mathcal{E}_D(f_{D,\lambda}) + \lambda \|f_{D,\lambda}\|_K^2 \leq \mathcal{E}_D(0)$. Thus,

$$\lambda \|f_{D,\lambda}\|_K^2 \leq \mathcal{E}_D(0) \leq \frac{1}{N}\sum_{i=1}^{N} \ell_\sigma(y_i) \leq \frac{1}{N}\sum_{i=1}^{N} y_i^2 \leq \max_{(x,y)\in D} |y|^2.$$

It follows that

$$\|f_{D,\lambda}\|_K \leq \lambda^{-\frac{1}{2}} \max_{(x,y)\in D} |y|. \tag{16}$$

This together with (15) yields the desired conclusion. □

Furthermore, we see that

$$\begin{aligned} \|W_{D,\lambda}\|_K &\leq \sigma^{-1}\frac{1}{N}\sum_{i=1}^{N}(\|f_{D,\lambda}\|_K + |y_i|)^2 \leq 2\sigma^{-1}\frac{1}{N}\sum_{i=1}^{N}\Big(\|f_{D,\lambda}\|_K^2 + |y_i|^2\Big) \\ &\leq 2\sigma^{-1}\left(\|f_{D,\lambda}\|_K^2 + \max_{(x,y)\in D}|y|^2\right). \end{aligned} \tag{17}$$

This in combination with the bounds (15) and (16) provides that, with probability at least $1-\delta$,

$$\|W_{D,\lambda}\|_K \leq 2(4M+5t)^2\Big(\lambda^{-1}+1\Big)\sigma^{-1}\left(\log\frac{N}{\delta}\right)^2. \tag{18}$$

3.2. Error Decomposition

To derive the explicit convergence rate of Algorithm (2), we introduce the *regularization function* f_λ in $\mathcal{H}_K$, defined by

$$f_\lambda := \arg\min_{f\in\mathcal{H}_K} \mathcal{E}_{\mathsf{ls}}(f) + \lambda\|f\|_K^2$$

where $\mathcal{E}_{\mathsf{ls}}(f) = \int_{\mathcal{Z}}(f(x)-y)^2 d\rho$ is the expected risk associated with the least squares loss. It is direct to verify that

$$f_\lambda = (L_K + \lambda I)^{-1} L_K f^*, \tag{19}$$

so $f_\lambda - f^* = -\lambda(L_K+\lambda I)^{-1} f^*$. By the work in [20], we know that under the regularity assumption (3) with $r > 0$,

$$\|f_\lambda - f^*\|_\rho \leq \begin{cases} \|h\|_\rho \lambda^r, & \text{when } 0 < r \leq 1, \\ \|h\|_\rho \lambda, & \text{when } r > 1, \end{cases} \tag{20}$$

and

$$\|f_\lambda\|_K \leq \begin{cases} \|h\|_\rho \lambda^{r-\frac{1}{2}}, & \text{when } 0 < r < 1/2, \\ \|h\|_\rho, & \text{when } r \geq 1/2. \end{cases} \tag{21}$$

Now, we state two error decompositions for $f_{D,\lambda} - f_\lambda$. By (19), we have

$$-L_{K,D} f_\lambda - \lambda f_\lambda = -L_{K,D} f_\lambda + L_K f_\lambda - L_K f^*.$$

It implies

$$-f_\lambda = (L_{K,D} + \lambda I)^{-1}[(L_K - L_{K,D}) f_\lambda - L_K f^*], \tag{22}$$

which leads the decomposition by (13),

$$\begin{aligned} f_{D,\lambda} - f_\lambda =& (L_{K,D} + \lambda I)^{-1}(L_K - L_{K,D}) f_\lambda + (L_{K,D} + \lambda I)^{-1}\left(\hat{f}_{\rho,D} - L_K f^*\right) \\ &+ (L_{K,D} + \lambda I)^{-1} W_{D,\lambda}. \end{aligned} \tag{23}$$

In the sequel, we denote

$$\begin{aligned} \mathcal{B}_{D,\lambda} &= \|(L_{K,D} + \lambda I)^{-1}(L_K + \lambda I)\|, \\ \mathcal{C}_{D,\lambda} &= \|(L_K + \lambda I)^{-\frac{1}{2}}(L_K - L_{K,D})\|, \\ \mathcal{G}_{D,\lambda} &= \|(L_K + \lambda I)^{-\frac{1}{2}}(\hat{f}_{\rho,D} - L_K f^*)\|_K. \end{aligned}$$

Noting that, for any $f \in \mathcal{H}_K$,

$$\max\{\|f\|_\rho, \sqrt{\lambda}\|f\|_K\} \leq \|(L_K + \lambda I)^{\frac{1}{2}} f\|_K \tag{24}$$

by the fact $\|f\|_\rho = \|L_K^{\frac{1}{2}} f\|_K$ [21], one obtains a bound for the *sample error* $\|f_{D,\lambda} - f_\lambda\|_\rho$ by the decomposition (23) above.

Proposition 1. *Define $f_{D,\lambda}$ by (2). Then, there holds*

$$\|f_{D,\lambda} - f_\lambda\|_\rho \leq \mathcal{B}_{D,\lambda}\mathcal{C}_{D,\lambda}\|f_\lambda\|_K + \mathcal{B}_{D,\lambda}\mathcal{G}_{D,\lambda} + \lambda^{-\frac{1}{2}}\mathcal{B}_{D,\lambda}\|W_{D,\lambda}\|_K. \tag{25}$$

Proof. Let I_1, I_2, and I_3 denote the three terms on the right-hand side of (23), respectively. Consider the $\mathcal{H}_K$ norm of

$$(L_K + \lambda I)^{1/2}(f_{D,\lambda} - f_\lambda) = (L_K + \lambda I)^{1/2}(I_1 + I_2 + I_3).$$

Then,

$$\begin{aligned} &\|(L_K + \lambda I)^{1/2} I_1\|_K \\ \leq& \|(L_K + \lambda I)^{1/2}(L_{K,D} + \lambda I)^{-1/2}\| \|(L_{K,D} + \lambda I)^{-1/2}(L_K + \lambda I)^{1/2}\| \\ &\times \|(L_K + \lambda I)^{-1/2}(L_K - L_{K,D})\| \|f_\lambda\|_K \\ \leq& \mathcal{B}_{D,\lambda}\mathcal{C}_{D,\lambda}\|f_\lambda\|_K. \end{aligned}$$

Similarly,

$$\begin{aligned} &\|(L_K + \lambda I)^{1/2} I_2\|_K \\ \leq& \|(L_K + \lambda I)^{1/2}(L_{K,D} + \lambda I)^{-1}(L_K + \lambda I)^{1/2}\| \|(L_K + \lambda I)^{-1/2}(\hat{f}_{\rho,D} - L_K f^*)\|_K \\ \leq& \mathcal{B}_{D,\lambda}\mathcal{G}_{D,\lambda}, \end{aligned}$$

and

$$\begin{aligned}&\|(L_K+\lambda I)^{1/2}I_3\|_K\\ \le&\|(L_K+\lambda I)^{1/2}(L_{K,D}+\lambda I)^{-1}(L_K+\lambda I)^{1/2}\|\frac{1}{\sqrt{\lambda}}\|W_{D,\lambda}\|_K\\ \le&\lambda^{-1/2}\mathcal{B}_{D,\lambda}\|W_{D,\lambda}\|_K.\end{aligned}$$

With the above bounds, we use (24) to obtain the statement.
The proof is finished. □

3.3. Deriving Main Results

To prove our main results, we need to bound the quantities $\mathcal{B}_{D,\lambda}, \mathcal{C}_{D,\lambda}, \mathcal{G}_{D,\lambda}$ by the following probability estimates.

Lemma 3. *With a confidence of at least* $1-\delta$, *there holds*

$$\mathcal{B}_{D,\lambda} \le 2\left(\frac{2\mathcal{A}_{D,\lambda}\log\frac{2}{\delta}}{\sqrt{\lambda}}\right)^2+2, \qquad \mathcal{C}_{D,\lambda} \le 2\mathcal{A}_{D,\lambda}\log\frac{2}{\delta}, \text{ and}$$
$$\mathcal{G}_{D,\lambda} \le 4(M+t)\mathcal{A}_{D,\lambda}\log\frac{2}{\delta}$$

where $\mathcal{A}_{D,\lambda} = \frac{1}{N\sqrt{\lambda}} + \frac{\sqrt{\mathcal{N}(\lambda)}}{\sqrt{N}}$.

These inequalities are well studied in the literature and can be found in [17,18].

Proof of Theorem 1. We can decompose $\|f_{D,\lambda}-f^*\|_\rho$ as the *sample error* $\|f_{D,\lambda}-f_\lambda\|_\rho$ and the *approximation error* $\|f_\lambda-f^*\|_\rho$. As stated in (20), $\|f_\lambda-f^*\|_\rho \le \lambda^r\|h\|_\rho$ for $0<r\le 1$. Thus, we just estimate $\|f_{D,\lambda}-f_\lambda\|_\rho$ by Proposition 1.

By Lemma 3 and the bound (18), with probability at least $1-4\delta$, the following bounds hold simultaneously:

$$\mathcal{B}_{D,\lambda}\mathcal{C}_{D,\lambda}\|f_\lambda\|_K \le 4\left[\left(\frac{2\mathcal{A}_{D,\lambda}}{\sqrt{\lambda}}\right)^2+2\right]\mathcal{A}_{D,\lambda}\left(\log\frac{2}{\delta}\right)^3\|f_\lambda\|_K,$$
$$\mathcal{B}_{D,\lambda}\mathcal{G}_{D,\lambda} \le 8(M+t)\left[\left(\frac{2\mathcal{A}_{D,\lambda}}{\sqrt{\lambda}}\right)^2+2\right]\mathcal{A}_{D,\lambda}\left(\log\frac{2}{\delta}\right)^3,$$

and

$$\lambda^{-\frac{1}{2}}\mathcal{B}_{D,\lambda}\|W_{D,\lambda}\|_K \le 8(4M+5t)^2\left[\left(\frac{2\mathcal{A}_{D,\lambda}}{\sqrt{\lambda}}\right)^2+2\right]\lambda^{-\frac{3}{2}}\sigma^{-1}\left(\log\frac{N}{\delta}\right)^4.$$

Scaling 4δ to δ, by (20) and the estimates above, we have with confidence at least $1-\delta$

$$\begin{aligned}&\|f_{D,\lambda}-f^*\|_\rho \le \|f_{D,\lambda}-f_\lambda\|_\rho+\|f_\lambda-f^*\|_\rho\\ &\le 24(4M+5t)^2\left[\left(\frac{2\mathcal{A}_{D,\lambda}}{\sqrt{\lambda}}\right)^2+2\right]\\ &\left[(\mathcal{A}_{D,\lambda}+\mathcal{A}_{D,\lambda}\|f_\lambda\|_K)\left(\log\frac{8}{\delta}\right)^3+\lambda^{-\frac{3}{2}}\sigma^{-1}\left(\log\frac{4N}{\delta}\right)^4\right]+\|h\|_\rho\lambda^r. \qquad (26)\end{aligned}$$

By (4),

$$\mathcal{A}_{D,\lambda} = \frac{1}{N\sqrt{\lambda}} + \frac{\sqrt{\mathcal{N}(\lambda)}}{\sqrt{N}} \leq \frac{1}{N\sqrt{\lambda}} + \sqrt{\frac{C\lambda^{-s}}{N}} \leq (\sqrt{C}+1)\frac{\lambda^{-\frac{s}{2}}}{\sqrt{N}}\left(\frac{\lambda^{\frac{s-1}{2}}}{\sqrt{N}}+1\right).$$

The choice (6) of λ results in the fact that

$$\frac{\lambda^{-\frac{s}{2}}}{\sqrt{N}} \leq \begin{cases} \lambda^{-\frac{s}{2}+\frac{1+s}{2}} = \lambda^{\frac{1}{2}}, & \text{when } 0 < r < \frac{1}{2}, \\ \lambda^{-\frac{s}{2}+\frac{s}{2}+\min\{1,r\}} = \lambda^{\min\{1,r\}}, & \text{when } r \geq \frac{1}{2}, \end{cases} \tag{27}$$

$$\frac{\lambda^{\frac{s-1}{2}}}{\sqrt{N}} \leq \left\{ \begin{array}{ll} \lambda^{\frac{s-1}{2}+\frac{1+s}{2}}, & \text{when } 0 < r < \frac{1}{2} \\ \lambda^{\frac{s-1}{2}+\frac{s}{2}+\min\{1,r\}}, & \text{when } r \geq \frac{1}{2} \end{array} \right\} \leq \lambda^{s} \leq 1, \tag{28}$$

and

$$\frac{\lambda^{-s-1}}{N} \leq \left\{ \begin{array}{ll} \lambda^{-s-1}\lambda^{s+1} = 1, & \text{when } 0 < r < \frac{1}{2} \\ \lambda^{-1-s}\lambda^{s+2\min\{r,1\}} = \lambda^{\min\{2r-1,1\}}, & \text{when } r \geq \frac{1}{2} \end{array} \right\} \leq 1. \tag{29}$$

Collecting the above estimates,

$$\left(\frac{\mathcal{A}_{D,\lambda}}{\sqrt{\lambda}}\right)^2 \leq (\sqrt{C}+1)^2\frac{\lambda^{-s-1}}{N}\left(\frac{\lambda^{\frac{s-1}{2}}}{\sqrt{N}}+1\right)^2 \leq 4(\sqrt{C}+1)^2 \tag{30}$$

and

$$\mathcal{A}_{D,\lambda} \leq \begin{cases} \lambda^{\frac{1}{2}}, & \text{when } 0 < r < \frac{1}{2}, \\ \lambda^{\min\{1,r\}}, & \text{when } r \geq \frac{1}{2}. \end{cases} \tag{31}$$

Putting (21), (30) and (31) into (26), we can get (7) with

$$C_1 = 96(4M+5t)^2[(\sqrt{C}+1)^2+1](2+\|h\|_\rho).$$

The proof is complete. □

Proof of Theorem 2. Similar as the proof of (25), there holds

$$\|f_{D^*,\lambda} - f_\lambda\|_\rho \leq \mathcal{B}_{D^*,\lambda}\mathcal{C}_{D^*,\lambda}\|f_\lambda\|_K + \mathcal{B}_{D^*,\lambda}\mathcal{G}_{D^*,\lambda} + \lambda^{-\frac{1}{2}}\mathcal{B}_{D^*,\lambda}\|W_{D^*,\lambda}\|_K. \tag{32}$$

Note that, by (9),

$$\hat{f}_{\rho,D^*} = \frac{1}{N+\tilde{N}}\sum_{i=1}^{N+\tilde{N}} y_i^* K_{x_i^*} = \frac{1}{N+\tilde{N}}\sum_{i=1}^{N}\frac{N+\tilde{N}}{N} y_i K_{x_i} = \frac{1}{N}\sum_{i=1}^{N} y_i K_{x_i} = \hat{f}_{\rho,D}.$$

It means $\mathcal{G}_{D^*,\lambda} = \mathcal{G}_{D,\lambda}$. Furthermore, similar to (16), we have

$$\|f_{D^*,\lambda}\|_K^2 \leq \frac{N+\tilde{N}}{N\lambda}\max_{(x,y)\in D}|y|^2.$$

In addition, by (17),

$$\begin{aligned}\|W_{D^*,\lambda}\|_K &\leq \sigma^{-1}\frac{1}{N+\tilde{N}}\sum_{i=1}^{N+\tilde{N}}\left(\|f_{D^*,\lambda}\|_K + |y_i|\right)^2 \leq 2\sigma^{-1}\frac{1}{N+\tilde{N}}\sum_{i=1}^{N+\tilde{N}}\left(\|f_{D^*,\lambda}\|_K^2 + |y_i^*|^2\right) \\ &\leq 2\sigma^{-1}\left(\|f_{D^*,\lambda}\|_K^2 + \left(\frac{N+\tilde{N}}{N}\right)^2\max_{(x,y)\in D}|y|^2\right).\end{aligned}$$

Then, by (15), with confidence at least $1-\delta$,

$$\|W_{D^*,\lambda}\|_K \le 2(4M+5t)^2\left(\frac{N+\tilde{N}}{N\lambda}+\left(\frac{N+\tilde{N}}{N}\right)^2\right)\sigma^{-1}\left(\log\frac{N}{\delta}\right)^2.$$

This together with Lemma 3 yields that, with a confidence of at least $1-4\delta$,

$$\mathcal{B}_{D^*,\lambda}\mathcal{C}_{D^*,\lambda}\|f_\lambda\|_K \le 4\left[\left(\frac{2\mathcal{A}_{D^*,\lambda}}{\sqrt{\lambda}}\right)^2+2\right]\mathcal{A}_{D^*,\lambda}\left(\log\frac{2}{\delta}\right)^3\|f_\lambda\|_K,$$

$$\mathcal{B}_{D^*,\lambda}\mathcal{G}_{D^*,\lambda} = \mathcal{B}_{D^*,\lambda}\mathcal{G}_{D,\lambda} \le 8(M+t)\left[\left(\frac{2\mathcal{A}_{D^*,\lambda}}{\sqrt{\lambda}}\right)^2+2\right]\mathcal{A}_{D,\lambda}\left(\log\frac{2}{\delta}\right)^3,$$

and

$$\lambda^{-\frac{1}{2}}\mathcal{B}_{D^*,\lambda}\|W_{D^*,\lambda}\|_K \le 8(4M+5t)^2\left[\left(\frac{2\mathcal{A}_{D^*,\lambda}}{\sqrt{\lambda}}\right)^2+2\right]\Delta_{N,\tilde{N},\lambda}\lambda^{-\frac{1}{2}}\sigma^{-1}\left(\log\frac{N}{\delta}\right)^4.$$

Scaling 4δ to δ, by (32) and (20), then, with a confidence at least $1-\delta$,

$$\begin{aligned}&\|f_{D^*,\lambda}-f^*\|_\rho \le \|f_{D^*,\lambda}-f_\lambda\|_\rho+\|f_\lambda-f^*\|_\rho\\ &\le 24(4M+5t)^2\left[\left(\frac{2\mathcal{A}_{D^*,\lambda}}{\sqrt{\lambda}}\right)^2+2\right]\\ &\left[(\mathcal{A}_{D,\lambda}+\mathcal{A}_{D^*,\lambda}\|f_\lambda\|_K)\left(\log\frac{8}{\delta}\right)^3+\Delta_{N,\tilde{N},\lambda}\lambda^{-\frac{1}{2}}\sigma^{-1}\left(\log\frac{4N}{\delta}\right)^4\right]+\|h\|_\rho\lambda^r. \qquad (33)\end{aligned}$$

Thus, to prove Theorem 2, we need the estimates as follows:
Since $r+s>\frac{1}{2}$ and $\lambda = N^{-\frac{1}{2r+s}}$,

$$\frac{\lambda^{\frac{s-1}{2}}}{\sqrt{N}} = N^{\frac{1-2(r+s)}{2(2r+s)}} \le 1, \quad \frac{\lambda^{-\frac{s}{2}}}{\sqrt{N}} = \lambda^r.$$

Then, by (4),

$$\mathcal{A}_{D,\lambda} = \frac{1}{N\sqrt{\lambda}}+\frac{\sqrt{\mathcal{N}(\lambda)}}{\sqrt{N}} \le \frac{1}{N\sqrt{\lambda}}+\sqrt{\frac{C\lambda^{-s}}{N}} \le (\sqrt{C}+1)\frac{\lambda^{-\frac{s}{2}}}{\sqrt{N}}\left(\frac{\lambda^{\frac{s-1}{2}}}{\sqrt{N}}+1\right) \le (\sqrt{C}+1)\lambda^r$$

and

$$\begin{aligned}\mathcal{A}_{D^*,\lambda} &= \frac{1}{(N+\tilde{N})\sqrt{\lambda}}+\frac{\sqrt{\mathcal{N}(\lambda)}}{\sqrt{N+\tilde{N}}} \le \frac{1}{(N+\tilde{N})\sqrt{\lambda}}+\sqrt{\frac{C\lambda^{-s}}{(N+\tilde{N})}}\\ &\le (\sqrt{C}+1)\frac{\lambda^{-\frac{s}{2}}}{\sqrt{N+\tilde{N}}}\left(\frac{\lambda^{\frac{s-1}{2}}}{\sqrt{N+\tilde{N}}}+1\right) \le 2(\sqrt{C}+1)\frac{\lambda^{-\frac{s}{2}}}{\sqrt{N+\tilde{N}}}.\end{aligned}$$

Thus,

$$\left(\frac{\mathcal{A}_{D^*,\lambda}}{\sqrt{\lambda}}\right)^2+1 \le 4(\sqrt{C}+1)^2\frac{\lambda^{-s-1}}{N+\tilde{N}}+1 \le 4(\sqrt{C}+1)^2+1.$$

Furthermore, by (21),

$$\mathcal{A}_{D^*,\lambda}\|f_\lambda\|_K \le \begin{cases}2(\sqrt{C}+1)\|h\|_\rho\frac{\lambda^{-\frac{s}{2}+r-\frac{1}{2}}}{\sqrt{N+\tilde{N}}}, & \text{when } 0<r<1/2,\\ 2(\sqrt{C}+1)\|h\|_\rho\frac{\lambda^{-\frac{s}{2}}}{\sqrt{N+\tilde{N}}}, & \text{when } r\ge 1/2.\end{cases}$$

By the restriction $\tilde{N} \geq \max\{N^{\frac{s+1}{2r+s}} - N + 1, 1\}$, we conclude that

$$\mathcal{A}_{D^*,\lambda}\|f_\lambda\|_K \leq 2(\sqrt{C}+1)\|h\|_\rho \lambda^r, \text{ for } r > 0.$$

Putting the estimates above into (33) yields that the desired conclusion (10) with

$$C_2 = 96(4M+5t)^2[(\sqrt{C}+1)^2 + (\sqrt{C}+1) + 1](2+\|h\|_\rho).$$

The proof is finished. □

4. Numerical Simulation

In this part, we carry out simulations to verify our theoretical statements. We employ the mean squared error of a testing set for the comparison. We generate $N = 500$ labeled data $\{x_i, y_i\}_{i=1}^{500}$ by the regression model $y_i = f^*(x_i) + \epsilon$, where $f^*(x) = x(1-x)$, and the random inputs x_i's are independently drawn according to the Normal distribution $\mathcal{N}(0,1)$, and ϵ is the independent Gaussian noise $\mathcal{N}(0, 0.005)$. We also generate $\tilde{N} = 200$ unlabeled data $\{\tilde{x}_i\}_{i=1}^{200}$ with $\tilde{x}_i$'s drawn independently according to the uniform distribution on $[0,1]$. We choose the Gaussian kernel $K(x,u) = \exp\{-|x-u|^2/2\}$, $h = 5$ and regularization parameter $\lambda = 0.7$. Algorithm 1 shows the mean squared error of Algorithm (1) with the training data set $D = \{x_i, y_i\}_{i=1}^{500}$. Algorithm 2 shows the mean squared error of Algorithm (2) with the semi-supervised data set D^* by (9). Algorithm (2)' s error is obviously smaller than Algorithm (1) if 20 unlabeled data are added into the training data. When we add more unlabeled data from 20 to 200, Algorithm (2)' s curve decreases continuously. These experimental results coincide with our theoretical analysis through the following Figure 1.

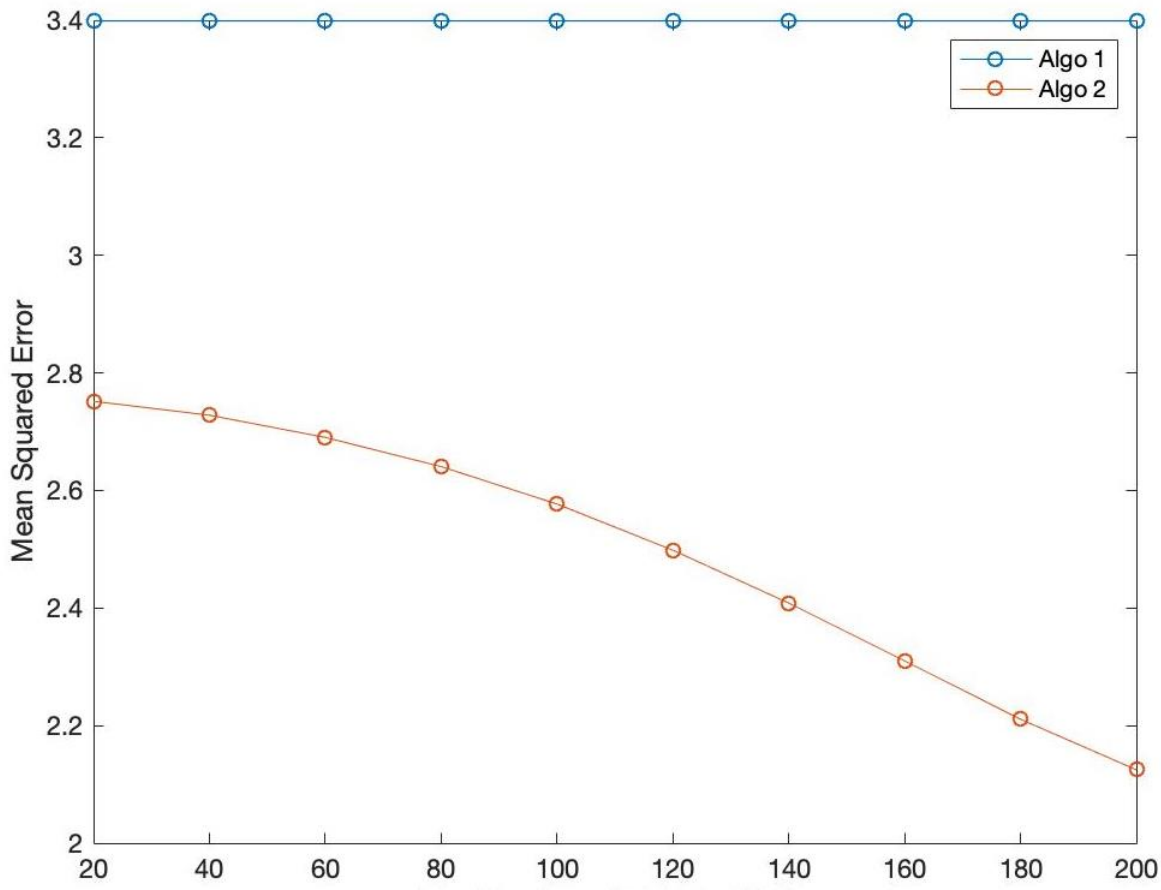

Figure 1. The number of unlabeled data.

5. Discussion

Unlabeled data are ubiquitous in a variety of fields including signal processing, privacy concerns, feature selection, and data clustering. For the applications of Huber regression that have robustness, we adopted a semi-supervised learning method to our regularized Huber regression algorithm. We derived the explicit learning rate of algorithm (2) in the supervised learning, which was comparable to the minimax optimal rate of OLS. By a semi-supervised method, we showed that an inflation of unlabeled data could improve learning performance for Huber regression analysis. It suggested that using the additional information of unlabeled data could extend the application of Huber regression.

Author Contributions: Conceptualization, Y.W.; Funding acquisition, C.P.; Methodology, B.W. and X.L.; Project administration, B.W.; Resources, X.L.; Supervision, C.P. and H.Y.; Writing—original draft, Y.W.; Writing—review & editing, H.Y. All authors have read and agreed to the published version of the manuscript.

Funding: The work described in this paper is partially supported by the National Natural Science Foundation of China (Project 12071356).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Huber, P.J. Robust Estimation of a Location Parameter. *Ann. Math. Stat.* **1964**, *35*, 73–101. [CrossRef]
2. Huber, P.J. Robust Regression: Asymptotics, Conjectures and Monte Carlo. *Ann. Stat.* **1973**, *1*, 799–821.
3. Christmann, A.; Steinwart, I. Consistency and robustness of kernel based regression. *Bernoulli* **2007**, *13*, 799–819. [CrossRef]
4. Fan, J.; Li, Q.; Wang, Y. Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *J. R. Stat. Soc.* **2017**, *79*, 247–265. [CrossRef]
5. Feng, Y.; Wu, Q. A statistical learning assessment of Huber regression. *J. Approx. Theory* **2022**, *273*, 105660. [CrossRef]
6. Loh, P.L. Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *Statistics* **2015**, *45*, 866–896. [CrossRef]
7. Rao, B. Asymptotic behavior of M-estimators for the linear model with dependent errors. *Bull. Inst. Math. Acad. Sin.* **1981**, *9*, 367–375.
8. Sun, Q.; Zhou, W.; Fan, J. Adaptive Huber Regression. *J. Am. Stat. Assoc.* **2017**, *115*, 254–265. [CrossRef]
9. Wang, Z.; Liu, H.; Zhang, T. Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *Ann. Stat.* **2013**, *42*, 2164–2201. [CrossRef] [PubMed]
10. Chapelle, O.; Schölkopf, B.; Zien, A. *Semi-Supervised Learning*; MIT Press: Cambridge, MA, USA, 2006.
11. Belkin, M.; Niyogi, P. Semi-Supervised Learning on Riemannian Manifolds. *Mach. Learn.* **2004**, *56*, 209–239. [CrossRef]
12. Blum, A.; Mitchell, T. Combining Labeled and Unlabeled Data with Co-Training. In Proceedings of the 11th Annual Conference on Computational Learning Theory, Madison, WI, USA, 24–26 July 1998.
13. Wang, J.; Jebara, T.; Chang, S.F. Semi-Supervised Learning Using Greedy Max-Cut. *J. Mach. Learn. Res.* **2013**, *14*, 771–800.
14. Andrea Caponnetto, Y.Y. Cross-validation based adaptation for regularization operators in learning theory. *Anal. Appl.* **2010**, *8*, 161–183. [CrossRef]
15. Guo, X.; Hu, T.; Wu, Q. Distributed Minimum Error Entropy Algorithms. *J. Mach. Learn. Res.* **2020**, *21*, 1–31.
16. Hu, T.; Fan, J.; Xiang, D.H. Convergence Analysis of Distributed Multi-Penalty Regularized Pairwise Learning. *Anal. Appl.* **2019**, *18*, 109–127. [CrossRef]
17. Lin, S.B.; Guo, X.; Zhou, D.X. Distributed Learning with Regularized Least Squares. *J. Mach. Learn. Res.* **2016**, *18*, 3202–3232.
18. Lin, S.B.; Zhou, D.X. Distributed Kernel-Based Gradient Descent Algorithms. *Constr. Approx.* **2018**, *47*, 249–276. [CrossRef]
19. Aronszajn, N. Theory of reproducing kernels. *Trans. Am. Math. Soc.* **1950**, *686*, 337–404. [CrossRef]
20. Smale, S.; Zhou, D.X. Learning Theory Estimates via Integral Operators and Their Approximations. *Constr. Approx.* **2007**, *26*, 153–172. [CrossRef]
21. Cucker, F.; Ding, X.Z. *Learning Theory: An Approximation Theory Viewpoint*; Cambridge University Press: Cambridge, MA, USA, 2007.
22. Bauer, F.; Pereverzev, S.; Rosasco, L. On regularization algorithms in learning theory. *J. Complex.* **2007**, *23*, 52–72. [CrossRef]
23. Caponnetto, A.; Vito, E.D. Optimal Rates for the Regularized Least-Squares Algorithm. *Found. Comput. Math.* **2007**, *7*, 331–368. [CrossRef]
24. Tong, Z. Effective Dimension and Generalization of Kernel Learning. In Proceedings of the Advances in Neural Information Processing Systems 15, NIPS 2002, Vancouver, BC, Canada, 9–14 December 2002.
25. Neeman, M.J. Regularization in kernel learning. *Ann. Stat.* **2010**, *38*, 526–565.
26. Raskutti, G.; Wainwright, M.J.; Yu, B. Early stopping and non-parametric regression. *J. Mach. Learn. Res.* **2014**, *15*, 335–366.
27. Wang, C.; Hu, T. Online minimum error entropy algorithm with unbounded sampling. *Anal. Appl.* **2019**, *17*, 293–322. [CrossRef]