LEARNING RATES FOR THE KERNEL REGULARIZED REGRESSION WITH A DIFFERENTIABLE STRONGLY CONVEX LOSS

BAOHUAI SHENG*, HUANXIANG LIU* AND HUIMIN WANG

Department of Applied Statistics, Shaoxing University
Shaoxing, 312000, China

Abstract. We consider learning rates of kernel regularized regression (KRR) based on reproducing kernel Hilbert spaces (RKHSs) and differentiable strongly convex losses and provide some new strongly convex losses. We first show the robustness with the maximum mean discrepancy (MMD) and the Hutchinson metric respectively, and, along this line, bound the learning rate of the KRR. We first provide a capacity dependent learning rate and then give the learning rates for four concrete strongly convex losses respectively. In particular, we provide the learning rates when the hypothesis RKHS’s logarithmic complexity exponent is arbitrarily small as well as sufficiently large.

1. Introduction. It is known that convex analysis is an effective tool in dealing with convex optimization problems. The learning problem which comes from many disciplines such as Psychology, Animal Behavior, Economic Decision Makings, Engineering, Computer Science and the study of human thought process is a kind of nonlinear optimization problem whose study needs knowledge of convex analysis, mathematical statistics, functional analysis and measure theory, and therefore has been investigated by many researchers ([25, 77]). The KRR model is one of the most important schemes in learning theory.

Let $X$ be a compact subset of $q$-dimensional Euclidean space $R^q$, $Y$ be a non-empty closed subset contained in $[-M, M]$ for a given $M > 0$. Then it is known that the regression learning with Support Vector Machine (SVM) is to find a function $f(x)$ between the input $x \in X$ and the output $y \in Y$ from a hypothesis space such that its values correspond to the condition mean of $y$ or closely related quantities. The function can be modelled by a positive probability distribution (measure) $\rho(x, y) = \rho(y|x) \rho_X(x)$ on $Z := X \times Y$, where $\rho(y|x)$ is the conditional probability of $y$ for a given $x$ and $\rho_X(x)$ is the marginal probability of $x$.

Generally speaking, the distribution $\rho$ is unknown and what one can know is a sample $z := \{z_i\}_{i=1}^m \in Z^m$ independently drawn (i.i.d.) according to $\rho$. Given a sample $z$, the regression problem aims at finding a function $f_z : X \rightarrow R$ such that

2010 Mathematics Subject Classification. Primary: 90C25, 68Q32, 68T40; Secondary: 41A25.

Key words and phrases. Kernel regularized regression, learning rate, differentiable strongly convex loss, conjugate loss, $K-$functional, maximum mean discrepancy (MMD), Hutchinson metric, reproducing kernel Hilbert space.

This work is supported by the National Natural Science Foundation of China under Grants (No. 61877039, 11501375) and the Natural Science Foundation of Zhejiang Province under Grant (No. LQ14A010005).

* Corresponding authors.
it is a good approximation of $y$ when a new input $x$ is given. In most cases, the function $f_z(x)$ is found from a reproducing kernel Hilbert space (RKHS) ([25, 27]).

Let $K : X \times X \to \mathbb{R}$ be continuous, symmetric and positive semi-definite, i.e., for any finite set of distinct points $t = \{x_1, x_2, \ldots, x_l\} \subset X$, the matrix $(K(x_i, x_j))_{i,j=1}^l$ is positive semi-definite. Such functions are called Mercer kernels.

The RKHS $H_K$ associating with the Mercer kernel $K$ is defined ([3]) to be the closure of the linear span of the set of functions $\{K_x : = K(x, \cdot) : x \in X\}$ with the inner product $\langle \cdot, \cdot \rangle_K$ satisfying the reproducing property

$$f(x) = \langle f, K_x \rangle_K, \quad f \in H_K, x \in X.$$ (1.1)

and having the embedding inequality

$$\|f\|_\infty \leq k \|f\|_K, \quad f \in H_K,$$ (1.2)

where $k = \sup_{x,y \in X} \sqrt{K(x,y)}$. We assume in this paper that $K$ is a bounded kernel on $X \times X$, i.e., $0 < k < +\infty$.

We call a given non-negative even convex function $V$ a regression loss function if it is defined on $\mathbb{R}$ and satisfies $V(0) = \min_{t \in \mathbb{R}} V(t)$. The Tikhonov regularization KRR learning algorithm based on an RKHS $H_K$ and a regression loss $V$ is defined to be the minimizer of the following algorithm ([34, 40])

$$f^V_{z,\lambda} := f_{z,\lambda,V,H_K} = \arg \min_{f \in H_K} \left[ \frac{1}{m} \sum_{i=1}^{m} V\left(y_i - f(x_i)\right) + \lambda \|f\|_K^2 \right],$$ (1.3)

where $0 < \lambda < 1$ is a given regularization parameter which is commonly used to overcome the ill-posedness.

It is known that the most ideal regression function $f^V_\rho(x)$ is

$$f^V_\rho := \arg \min_{f \in H_K} \mathcal{E}_{\rho,V}(f),$$ (1.4)

where

$$\mathcal{E}_{\rho,V}(f) = \int_Z V(y - f(x)) \, d\rho = E_\rho[V(\cdot - f(\cdot))].$$

and the minimum is taken over all measurable functions with respect to $\rho$. It is known that if $V(t) = t^2$, then ([25, 26])

$$f_\rho(x) = E(y|x) = \int_Y y \, d\rho(y|x) \quad \text{and} \quad f_\rho := \arg \min_{f \in L^2(\rho_X)} \mathcal{E}_{\rho}(f),$$

where

$$\mathcal{E}_{\rho}(f) = \int_Z (y - f(x))^2 \, d\rho.$$ 

For the needs of the statements we define the integral risk minimization scheme by

$$f^V_{\rho,\lambda} := \arg \min_{f \in H_K} \left[ \mathcal{E}_{\rho,V}(f) + \lambda \|f\|_K^2 \right].$$ (1.5)

In particular, we define

$$f_{\rho,\lambda} := \arg \min_{f \in H_K} \left[ \mathcal{E}_{\rho}(f) + \lambda \|f\|_K^2 \right].$$

Define an empirical expectation $\rho_z$ on $Z := X \times Y$ as (the empirical measure)

$$E_{\rho_z}[f(x,y)] = \int_Z f(x,y) \, d\rho_z = \frac{1}{m} \sum_{i=1}^{m} f(x_i, y_i).$$

for any bounded $\rho$-measurable function $f$ on $Z = X \times Y$. Then $f_{z,\lambda}^V = f_{\rho,\gamma}^V$ and by the Representer Theorem (Chapter 5 of [72] and [30]) we know $f_{z,\lambda}$ has the form

$$f_{z,\lambda}^V(x) = \sum_{j=1}^m \alpha_j K(x, x_j), \quad x \in X, \quad (1.6)$$

for a real coefficient vector $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_m)^\top$.

The concept of stability is often used to describe the dependence of the optimizer on the parameters qualitatively ([12]), for example, for the optimization problem $V(u) = \arg\min_{x \in X} \varphi(x, u)$, $u \in D$, where $u$ is a parameter varying in a given domain $D$.

To discuss the continuity of $V$ about $u$ on $D$ belongs to the scope of stability. By robustness we mean to describe the dependence of $V$ on the parameter $u$ quantitatively.

In the field of kernel regularized learning, the robustness of $f_{\rho,\lambda}^V$ on the distributions $\rho$ is systematically studied by I. Steinwart and A. Christmann. They show that if $V$ is a differentiable convex function, then (Theorem 5.9 of [72])

$$\left\| f_{\rho,\lambda}^V - f_{\rho,\gamma}^V \right\|_{\mathcal{K}} \leq \frac{1}{\lambda} \left( \int_Z V'(y - f_{\rho,\lambda}^V(x)) K_x(\cdot) d\rho - \int_Z V'(y - f_{\rho,\gamma}^V(x)) K_x(\cdot) d\gamma \right), \quad (1.7)$$

where $\gamma$ is an another probability distribution on $Z$, $f_{\rho,\lambda}^V$ is the corresponding optimization solution defined in the same way as that of $f_{\rho,\lambda}^V$ in (1.5) ([18, 19, 21, 22, 32, 39, 23, 28, 45, 69]).

Let $\mathcal{P}_Z$ be the set of all probability measures on $Z$, $d(\cdot, \cdot)$ be a metric on $\mathcal{P}_Z$. Then we expect to form a robust method so that $d(\rho, \gamma)$ can be used to describe the right side of (1.7) quantitatively.

Let $\mu \in \mathcal{P}_Z$ be a probability measures on a Banach space $E$. We define the norm of total variation as

$$\|\mu\|_{\mathcal{P}_Z} = \sup \left\{ \sum_{i=1}^n |\mu(E_i)| : E_1, E_2, \cdots, E_n \text{ is a partition of } E \right\}.$$

Theorem 10. 15 of [72] shows that

$$\lim_{\varepsilon \to 0} \left\| f_{(1-\varepsilon)\rho + \varepsilon\gamma,\lambda}^V - f_{\rho,\lambda}^V \right\|_{\mathcal{K}} \leq C(V, \lambda)\|\gamma - \rho\|_{\mathcal{P}_Z}, \quad (1.8)$$

where $C(V, \lambda) > 0$ is a constant depending upon $V$ and $\lambda$, see also[17]. If $V$ is a Lipschitz loss, then (Theorem 10.27 of [72] or [18, 19])

$$\left\| f_{(1-\varepsilon)\rho + \varepsilon\gamma,\lambda}^V - f_{\rho,\lambda}^V \right\|_{\mathcal{K}} \leq C_{\gamma, \lambda}\varepsilon, \quad (1.9)$$

where $C_{\gamma, \lambda} = \lambda^{-1}\|K\|_{\infty} |V|_1 \|\gamma - \rho\|_{\mathcal{P}_Z}$ and $|V|_1 = \max_{x \in R} |V'(x)|$.

When $V$ is a Lipschitz loss, [22] provides a bound for $\|f_{\rho,\lambda}^V - f_{\rho,\gamma}^V\|_{\mathcal{K}}$. i.e.,

$$\|f_{\gamma,\lambda}^V - f_{\rho,\lambda}^V\|_{\infty} \leq C_{\gamma,\lambda}\|\gamma - \rho\|_{\mathcal{P}_Z}, \quad (1.10)$$

where

$$\|\gamma - \rho\|_{\mathcal{P}_Z} = \frac{1}{2} \sup_{h \in \mathcal{K}} \left| \int_Z h d\rho - \int_Z h d\gamma \right|$$

and $C_{\gamma,\lambda}$ is a constant depending upon $\lambda$ and $V$. Similar estimates are given in [24] for kernel regularized pairwise regression. The Lipschitz assumption is a strong condition since even $V(t) = t^2$ is not a Lipschitz function on $R$. An interesting problem is if the assumption that $V(t)$ is a Lipschitz loss can be removed or weaken.
Then what can we say about the estimate? This is the first motivation for writing this paper.

Another research topic in learning theory is to bound the error between $f^V_{\gamma,\lambda}$ and $f^V_\lambda$ in probability. Many mathematicians have devoted to this field and many approaches have been developed, see e.g., the convex analysis approach ([19, 72]), the capacity approach ([15, 25, 26, 27, 78]), the integral operator approach ([44, 53, 74]), the spectral learning approach ([2, 6, 29]) and the optimization approach ([5, 86]). To investigate the optimal learning rates for algorithm (1.3), many researchers choose the quantile loss ([21, 23, 70, 71, 84]) since it has the Lipschitz property. It is known that the least square loss $V(t) = t^2$ has many nice properties, the learning approach based on it has been studied by many researches ([9, 10, 14, 35, 33, 36, 42, 43, 59, 73, 80, 83]). When $V$ is a general convex loss, the error has also been studied in many papers ([19, 20, 46, 75, 76]).

There are three issues that deserve our attention. (i) It is known that for $V(t) = t^2$ the errors are described by norm $\| \cdot \|_{L^2(\rho_X)}$ due to the famous comparison equality ([25])
\[
\mathcal{E}_\rho(g) - \mathcal{E}_\rho(f_\rho) = \| g - f_\rho \|_{L^2(\rho_X)}^2, \quad g \in L^2(\rho_X).
\] (1.11)

Since many convex losses have no such comparison equalities, many papers describe the error in expectation: $\mathcal{E}_{\rho,V}(f) - \mathcal{E}_{\rho,V}(f^V_\rho)$. Recently, some mathematicians have paid their attention to build the comparison inequality of the type
\[
C\| g - f^V_\rho \|_{L^2(\rho_X)}^2 \leq \mathcal{E}_{\rho,V}(g) - \mathcal{E}_{\rho,V}(f^V_\rho), \quad g \in L^2(\rho_X).
\] (1.12)

For example, some comparison inequalities for the pinball loss are established under certain realistic assumptions on $\rho$ ([35, 70, 84, 85]). It will be of interest if we can establish certain comparison inequality similar to (1.12) for the strongly convex losses. This is the second motivation for writing this paper. (ii) Although the error bound for a convex loss has been estimated by many papers, we find that the typical examples of losses are in fact not many as we hope. It is known that the losses satisfying the Lipschitz condition may provide nice learning rates (the pinball loss in [21]). Are there other losses which do not satisfy the Lipschitz property but can attain the same learning rate as that of $V(t) = t^2$? It is known that strongly convex losses have many advantages. In particular, the Bregman distance has been used for clustering ([4]). Whether or not the kernel regularized regression associating them can attain the same learning rate as that of $V(t) = t^2$. This is the third motivation for writing this paper. (iii) We notice that, to provide the learning rate for a general convex loss, the variance-expectation bound condition
\[
\sigma^2\mathcal{E}(y - f(x)) - \mathcal{E}(y - f^V_\rho(x))^2 \leq C_\alpha (\mathcal{E}_{\rho,V}(f) - \mathcal{E}_{\rho,V}(f^V_\rho))^\alpha,
\] (1.13)

where $|f(x)| \leq M$, $|f^V_\rho(x)| \leq M$, is needed for the loss ([75, 76]). This assumption always holds if $V$ is a strongly convex loss (P.338 of [90]). Then (1.13) is not a necessary assumption. To provide learning rates without the variance-expected bound condition (1.13) is an interesting topic. This forms the fourth motivation for writing this paper.

The contributions of this manuscript are three folds. First we provide two kinds of explicit upper bounds for the error $\| f^V_{\gamma,\lambda} - f^V_\lambda \|_{L^2(\rho_X)}$ in MMDIPM (maximum mean discrepancy integral probability metric) and the Hutchinson metric respectively when $V$ is a differentiable convex loss and its conjugate loss $V^*$ is a strongly convex loss. Second, we provide three new differentiable strongly convex losses. Some of them can attain the same learning rate as that of $V(t) = t^2$ in the setting of the
present paper. Third, we find that if \( V \) is a convex loss and its conjugate loss \( V^* \) is a strongly convex loss, \( V' \) is an important quantity in describing the learning rate, which may lead to new metrics, e.g., the metrics \( \ln^2 \| \frac{x}{\epsilon} \| \) and \( \sqrt{\frac{|x|}{1+|x|}} \) on \( R \).

In a word, we find in this manuscript that the strong convexity of the conjugate loss \( V^* \) of a convex loss \( V \) has more contributions to the learning rate than that of the strong convexity of \( V \) itself.

The paper is organized as follows. In Section 2 we shall provide some notions and concepts of convex analysis. Three new convex losses and their conjugate losses will be provided. In Section 3 we shall provide some bounds for metric rates for four concrete differentiable convex losses. In Section 5 we shall give some discussions for the obtained results. Some lemmas used in proving the main results will be provided. In Section 6 we shall provide some bounds for metric

\[
\int_Z \left| V'(y - f^V_{\rho,\lambda}(x)) - V'(y - f^V_{\gamma,\lambda}(x)) \right|^2 d\rho \tag{1.14}
\]

with MMDIPM and the Hutchinson metric respectively. The upper bounds are provided explicitly with the help of the \( K \)-functional defined in learning theory.

In Section 4 we shall show the learning rates for some strongly convex losses. We first give a general estimate for metric

\[
\int_Z \left| V'(y - f^V_{\rho,\lambda}(x)) - V'(y - f^V_{\rho}(x)) \right|^2 d\rho \tag{1.15}
\]

in case that \( V \) is a differentiable convex loss and its conjugate loss \( V^* \) is a strongly convex loss with convex modulus \( \frac{1}{A} \) and \( c > 0 \) and then provide some learning rates for four concrete differentiable convex losses. In Section 5 we shall give some discussions for the obtained results. Some lemmas used in proving the main results are given in Subsection 6.1. All the proofs are given in Subsection 6.2.

Throughout the paper, we assume that \( \rho \) has finite second moment, i.e., \( \int_Z y^2 d\rho < +\infty \). We say \( A = O(B) \) if there exists a constant \( C > 0 \) such that \( A \leq CB \). We say \( A \sim B \) if both \( A = O(B) \) and \( B = O(A) \). For any \( a = (a_1, a_2, \ldots, a_q)^\top, b = (b_1, b_2, \ldots, b_q)^\top \in R^q \), we define in this paper

\[
\|a\|_2^2 = \sum_{i=1}^q |a_i|^2 = a^\top a, \quad (a, b)_2 = \sum_{i=1}^q a_i b_i = a^\top b.
\]

2. Notations and concepts. Let \( \mathcal{P}_X \) be the set of marginal probability distributions on \( X \). Then we can define on \( \mathcal{P}_X \) a metric with characteristic kernels. We call a bounded measurable positive definite kernel \( K \) a characteristic kernel if by \( \rho_X, \gamma_X \in \mathcal{P}_X \) and \( \| \int_X K_x(\cdot) d\rho_X - \int_X K_x(\cdot) d\gamma_X \|_K = 0 \) we have \( \rho_X = \gamma_X \), which is not only equivalent to the fact that the kernel mean embedding

\[\mu_{\rho_X} : \rho_X \to \int_X K_x(\cdot) d\rho_X\]

is injective, but also is equivalent to the universality of the kernel ([47, 49, 67, 68]), where we say a Mercer kernel \( K_x(y) = K(x, y) \) is universal if for any given subset \( W \subset X \), any positive number \( \varepsilon \) and any function \( f \in C(W) \) there is a function \( g \) having the form of (1.6), i.e., \( g(x) = \sum \alpha_i K_x(x_i), \ x_i \in W, \alpha_i \in R \) such that \( \|f - g\|_{C(W)} < \varepsilon \). One can find such kind of kernels from [55, 57, 60].

Define the maximum mean discrepancy (MMD) integral probability metric (IPM) between \( \rho_X \) and \( \gamma_X \) as

\[d_K(\rho_X, \gamma_X) = \sup_{\|f\|_K \leq 1} \left| \int_X f d\rho_X - \int_X f d\gamma_X \right|\]
Then, by Section 3.5 of [48] or Theorem 1 of [68] we know
\[ d_K(\rho x, \gamma x) = ||\mu_{\rho x} - \mu_{\gamma x}||_K. \]
To get the bounds of \([f^x_{\mu,\lambda} - f^x_{\gamma,\lambda}]\) with \(d_K(\rho x, \gamma x)\) is a topic deserves us to consider.

Let \(B\) be a metric space with metric \(d(x, y)\). We denote by \(C(B)\) the set of all the continuous functions \(f : B \to R\). We say \(f \in C(B)\) has Lipschitz property of \(\alpha\) \((0 < \alpha \leq 1)\) if there is a constant \(C(f) > 0\) such that
\[ |f(x) - f(y)| \leq C(f)(d(x, y))^\alpha, \quad \forall x, y \in B. \]
We call \(C(f)\) a Lipschitz constant. In particular we denote
\[ \text{Lip}_{\alpha} = \{f \in C(B) : |f(x) - f(y)| \leq (d(x, y))^\alpha\}. \]

Define the Hutchinson metric as
\[ D_H(\rho, \gamma) = \sup \left\{ \int_B f d\rho - \int_B f d\gamma : f \in \text{Lip}_1 \right\}, \quad \rho, \gamma \in \mathcal{P}_B. \]

**Proposition 2.1** ([54]). \(D_H\) is a metric on \(\mathcal{P}_B\) and, moreover, \((\mathcal{P}_B, D_H)\) is a complete metric space.

We call a function \(f : X \to R\) a strongly convex function on \(X\) if for any \(x, x' \in X\) and any \(\lambda \in [0, 1]\) there holds
\[ f(\lambda x + (1 - \lambda)x') \leq \lambda f(x) + (1 - \lambda)f(x') - \frac{c\lambda(1 - \lambda)}{2}||x - x'||_2^2, \]
where \(c > 0\) is the convex modulus. If \(c = 0\), we call it a convex function.

By Proposition 10.6 and Proposition 17.10 of [7] we know if \(f\) is a differential convex loss on \(X\), then it is a strongly convex function with convex modulus \(c\) if and only if for any \(x, x' \in X\)
\[ f(x') - f(x) \geq \langle \nabla f(x), x' - x \rangle_2 + \frac{c}{2}||x' - x||_2^2. \tag{2.1} \]

For a function \(f : R^q \to R \cup \{+\infty\}\) not identically \(+\infty\) the conjugate function \(f^*\) is defined by
\[ f^*(y) = \sup \{ \langle y, x \rangle_2 - f(x) : x \in \text{dom} f \}, \quad y \in R^q, \]
where \(\text{dom} f = \{x \in R^q : -\infty < f(x) < +\infty\}\) and \(\langle ., . \rangle_2\) is the inner product in \(R^q\).

It is known that \(f\) is convex and lower semicontinuous if, and only if \((f^*)^* = f\).

By the \(N\)-function theory of Orlicz space (Chapter 1 of [52]) we know that if \(V : R \to R\) is an even convex function, then its conjugate function \(V^* : R \to R\) is
\[ V^*(v) = \sup_{u \geq 0} (u|v| - V(u)), \quad v \in R. \tag{2.2} \]

From now on we shall denote by \(V^*\) the conjugate function of a function \(V\).

It is known that for \(M(u) = \frac{|u|^p}{p}, p > 1\), we have \(M^*(u) = \frac{|u|^{p'}}{p'}, p' = \frac{p}{p-1}\).

We show in Lemma 6.9 that for the convex loss
\[ V_1(t) = |t| \ln \left| t + \sqrt{t^2 + 4} - \sqrt{t^2 + 4} + 2 \right|, \quad t \in R, \tag{2.3} \]
the conjugate loss is \(V_1^*(t) = e^{|t|} - e^{-|t|} - 2\), it is a strongly convex loss on \(R\) with convex modulus 2.

By the Example 11 in Chapter 1 of [52] we know the convex loss
\[ V_2(v) = (1 + |v|) \ln(1 + |v|) - |v|, \quad v \in R, \tag{2.4} \]
has a conjugate loss $V_2^*(u) = e^{|u|} - |u| - 1$, it is a strongly convex loss on $R$ with convex modulus 1.

The convex loss ((66))

$$V_3(u) = \sqrt{1 + u^2} - 1, \quad u \in R,$$

whose conjugate function

$$V_3^*(u) = \begin{cases} 1 - \sqrt{1 - u^2}, & |u| < 1; \\ +\infty, & |u| \geq 1, \end{cases}$$

is a strongly convex loss with convex modulus 1 (see Lemma 6.10).

By Theorem 4.2.1 of [41] we known if $f^*$ is a differentiable strong convex with convex modulus $\frac{1}{c}$ on $R^q$, then $\nabla f$ is Lipschitzian with constant $c$ on $R^q$:

$$\|\nabla f(x') - \nabla f(x)\|_2 \leq c \|x' - x\|_2, \quad \forall x, x' \in X. \quad (2.6)$$

for all $(x', x) \in R^q \times R^q$. It follows, in this case, that (from the page 120 of [41])

$$f(x') \geq f(x) + \langle \nabla f(x), x' - x \rangle + \frac{1}{2c} \|\nabla f(x') - \nabla f(x)\|_2^2, \quad \forall x, x' \in X, \quad (2.7)$$

and (Lemma 1.30 of [50])

$$f(x') \leq f(x) + \langle \nabla f(x), x' - x \rangle + \frac{1}{2c} \|x' - x\|_2^2, \quad \forall x, x' \in X. \quad (2.8)$$

Inequality (2.7) is more useful than inequality (2.1) in setting of this paper. It will lead to a new metric (1.15) and a comparison inequality (5.6).

It is known that if $V$ is a differentiable even convex function on $(-\infty, +\infty)$, then $V'$ is an odd and increasing function on $(-\infty, +\infty)$ and $|V'(t)| = V'(t)$ for $t \geq 0$.

The $K$-functional

$$D(f^\rho, \lambda) = \inf_{g \in H_K} \left[ \mathcal{E}_{\rho, V}(g) - \mathcal{E}_{\rho, V}(f^\rho) + \lambda \|g\|_K^2 \right]$$

is a key quantity in learning theory ([16, 27, 30, 55, 58]). It shows the approximation ability of $H_K$ with respect to $L^2(\rho_X)$ and is influenced by the loss $V(t)$. In particular when $V(t) = t^2$ it becomes

$$D(f^\rho, \lambda) = \inf_{g \in H_K} \left[ \|g - f^\rho\|_{L^2(\rho_X)}^2 + \lambda \|g\|_K^2 \right].$$

To show an explicit learning rate for the algorithm (1.3), one often use an assumption ([63, 61, 64, 73, 75, 76]):

**Assumption.** There are $0 < \beta \leq 1$ and $c_\beta > e$ such that

$$D(f^\rho, \lambda) \leq c_\beta \lambda^\beta, \quad \forall \lambda > 0. \quad (2.9)$$

Since $\lambda \to 0^+$ we know $\frac{D(f^\rho, \lambda)}{\lambda^2}$ is unbounded unless $\beta = 1$, we assume in this paper, without loss of generality, that

$$\frac{D(f^\rho, \lambda)}{\lambda^2} \geq M^2,$$

where $M$ is the given constant in the symbol $Y$ and $Z$. 

3. The robustness. We provide some robustness estimates for \( f_{\rho,\lambda}^V \) in \( d_K(\rho, \gamma) \) and \( D_H(\rho, \gamma) \).

**Proposition 3.1.** Let \( f_{\rho,\lambda}^V \) and \( f_{\gamma,\lambda}^V \) be the solutions of algorithm (1.5) with respect to \( \rho \) and \( \gamma \) respectively. If \( V \) is a differentiable convex loss and its conjugate loss \( V^* \) is a differentiable strongly convex loss with convex modulus \( \frac{1}{\epsilon} \) and \( V' \) is continuous on \( R \), then

\[
\int_Z \left| V'(y - f_{\rho,\lambda}^V(x)) - V'(y - f_{\gamma,\lambda}^V(x)) \right|^2 \, d\rho \\
\leq 2c \frac{\sup_{\|h\|_K \leq 1} \left| \int_Z V'(y - f_{\rho,\lambda}^V(x))h(x) \, d\rho - \int_Z V'(y - f_{\gamma,\lambda}^V(x))h(x) \, d\gamma \right|}{\lambda} \\
\times \sup_{\|h\|_K \leq 1} \left| \int_Z V'(y - f_{\rho,\lambda}^V(x))h(x) \, d\rho - \int_Z V'(y - f_{\rho,\lambda}^V(x))h(x) \, d\gamma \right|. \tag{3.1}
\]

By (3.1) we know \( V' \) is a key factor in quantitatively describing the robustness.

To show the robustness of \( f_{\rho,\lambda}^V \) about \( \rho \) in metric \( D_H(\rho, \gamma) \), we borrow the concept of separating kernels ([31]). We call a Mercer kernel \( K : X \times X \rightarrow R \) a separating kernel if for all \( x, y \in X \) with \( x \neq y \) we have \( K_x \neq K_y \). By the Proposition 1 of [31] we know if \( K \) is a separating kernel, then

\[
\|K_x - K_y\|_K = \sqrt{K(x, x) + K(y, y) - 2K(x, y)} \tag{3.2}
\]

is a distance on \( X \). Therefore for \( z = (x, y), z' = (x', y') \in Z \) we may define on \( X \) the distance

\[
d_H(z, z')_K = |y - y'| + \|K_x - K_y\|_K.
\]

We use this distance to define the Hutchinson metric \( D_H(\rho, \gamma) \) on \( \mathcal{P}_Z \) and use it to describe the robustness for \( f_{\rho,\lambda}^V \).

**Theorem 3.1.** Let \( K : X \times X \rightarrow R \) be a separating kernel, \( f_{\rho,\lambda}^V \) and \( f_{\gamma,\lambda}^V \) be the solutions of algorithm (1.5) with respect to \( \rho \) and \( \gamma \) respectively. If \( V \) is a differentiable convex loss and its conjugate loss \( V^* \) is a differentiable strongly convex loss with convex modulus \( \frac{1}{\epsilon} \) and \( V' \) is continuous on \( R \), then

\[
\int_Z \left| V'(y - f_{\rho,\lambda}^V(x)) - V'(y - f_{\gamma,\lambda}^V(x)) \right|^2 \, d\rho \\
\leq \frac{8c^3 \beta' \beta^2}{\lambda} D_H^2(\rho, \gamma), \tag{3.3}
\]

where

\[
\beta' = \max \left\{ \sqrt{\frac{D(f_{\rho}^V, \lambda)}{\lambda}}, \left. V'(2k \sqrt{\frac{D(f_{\rho}^V, \lambda)}{\lambda}}) \right\}
\]

and

\[
\beta = \max \left\{ \sqrt{\frac{D(f_{\gamma}^V, \lambda)}{\lambda}}, \left. V'(2k \sqrt{\frac{D(f_{\gamma}^V, \lambda)}{\lambda}}) \right\}
\]

Moreover, if \( K \) is universal and \( V' \) is a bounded function with \( |V'(t)| \leq M' \), then

\[
\int_Z \left| V'(y - f_{\rho,\lambda}^V(x)) - V'(y - f_{\gamma,\lambda}^V(x)) \right|^2 \, d\rho \\
\leq \frac{2c (M')^2}{\lambda} \times d_K^2(\rho_X, \gamma_X), \tag{3.4}
\]

We give several estimates under the condition that \( K : X \times X \rightarrow R \) is a separating kernel.
where $f$ defined to be the minimal positive integer number $l$

We borrow the concept of covering number.

4. Learning rates. To give a capacity dependent learning rate for algorithm (1.3), we borrow the concept of covering number.

Let $S$ be a metric space and $\eta > 0$. Then, the covering number $\mathcal{N}(S, \eta)$ is defined to be the minimal positive integer number $l$ such that there exist $l$ disks in $S$ with radius $\eta$ covering $S$.

We say a compact subset $E$ in a metric space $(\mathcal{H}, ||\cdot||_\mathcal{H})$ has logarithmic complexity exponent $s \geq 0$ if there is a constant $c_s > 0$ such that the closed ball of radius $r$ centered at origin, i.e., $\mathcal{B}_r = \{ f \in E : ||f||_\mathcal{H} \leq r \}$ satisfies

$$\log \mathcal{N}(\mathcal{B}_r, \eta) \leq c_s \left( \frac{r}{\eta} \right)^s, \quad \forall \eta > 0.$$  \hspace{1cm} (4.1)

Let $X \subset R^q$ be a compact set, for a positive integer $s$ we denote by $C^{(s)}(X)$ the space of functions on $X$ that are $s$-times differentiable and whose $s$-th partial
derivatives $D^\alpha$ are continuous, i.e., $\|D^\alpha f\|_{C^\alpha(X)} = \max_{|\alpha| \leq s} \|D^\alpha f\|_{C(X)} < +\infty$, where for $\alpha \in N_0^q$ we define

$$D^\alpha f = \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \ldots \frac{\partial^{\alpha_q}}{\partial x_q^{\alpha_q}} f, \quad |\alpha| = \alpha_1 + \cdots + \alpha_q.$$

It is known by Theorem 5.1 of [27] that if $K \in C(\alpha)(X \times X)$, then $\mathcal{H}_K$ has the logarithmic complexity with exponent $\frac{2q}{s}$.

It is proved in [87] that for a $C^\infty(X \times X)$ kernel (such as Gaussians) (4.1) is valid for an arbitrarily small $s > 0$.

Let $C([a,b]^q, B, c)$ denote the class of real-valued convex functions defined on $[a,b]^q$ that are uniformly bounded in absolute value by $B$ and uniformly Lipschitz with constant $c$. Then by [37] or Theorem 6 of [13] we know

$$\log \mathcal{N}[C([a,b]^q, B, c), \eta] \sim \eta^{\frac{2}{q}}. \tag{4.2}$$

Also, by Theorem 5.8 of [27] we know if $X$ is a closed subset of $R^q$ with piecewise smooth boundary, and $K : X \times X \to R$ is a Mercer kernel. Let $\alpha > 0$ such that $K \in Lip^* [\alpha, C(X \times X)]$, where $0 < \alpha \leq 1$,

$$Lip^* [\alpha, C(X \times X)] = \{ f \in Lip(\alpha) : \| f \|_{Lip(\alpha)} = |f|_{Lip(\alpha)} + \| f \|_{C(X)} < +\infty \},$$

and $|f|_{Lip(\alpha)} = \sup_{x \neq y \in X} \frac{|f(x) - f(y)|}{(\alpha(x, y))^{\frac{1}{p}}}$.

Then,

$$\log \mathcal{N}(B_r, \eta) \leq c\left(\frac{r}{\eta}\right)^{\frac{2q}{s}}, \quad \forall \eta > 0. \tag{4.3}$$

(4.2) and (4.3) show that the exponent $s$ depends upon the dimension $q$. It increases to $+\infty$ if $\alpha \to 0^+$ or $q \to +\infty$. These facts encourage us to consider the learning rate for the cases of $K \in C^\infty(X \times X)$ and $K \in Lip^* [\alpha, C(X \times X)]$ with $\alpha \to 0^+$.

We first give a general learning rate for algorithm (1.3).

**Proposition 4.1.** Let $f_{z, \lambda}^V$ and $f_{\rho}^V$ be the solutions for algorithms (1.3) and (1.4) respectively. If $V$ is a differentiable convex loss and its conjugate loss $V^*$ is a differentiable strongly convex loss with convex modulus $\frac{1}{c}$ and $V^*$ is continuous on $R$. Assume $\mathcal{H}_K$ has the logarithmic complexity with exponent $s > 0$, i.e., (4.1) holds for $\mathcal{B}^{\alpha, k}_K = \{ f \in \mathcal{H}_K : \| f \|_{\mathcal{H}_K} \leq r \}$. Then for any $0 < \delta < 1$, with confidence $1-\delta$, holds

$$\int_Z |V'(y - f_{\rho}^V(x)) - V'(y - f_{z, \lambda}^V(x))|^2 d\rho \leq \frac{32kc(16k^2c_\alpha)^{\frac{2}{q+1}} (4k^2c)^{\frac{2}{q+1}} (V(M))^{\frac{2}{q+1}}}{\lambda^{\frac{q}{q+1}} m^{\frac{q}{q+1}} \log 2} \times V'(2k \sqrt{\frac{D(f_{\rho}^V, \lambda)}{\lambda}}) \left[ V'(2k \sqrt{\frac{V(M)}{\lambda}}) \right]^{\frac{2}{q+1}} \left( \log \frac{4}{\delta} \right)^{\frac{2}{q+1}} + 8cD(f_{\rho}^V, \lambda) \tag{4.4}$$

if

$$V'(2k \sqrt{\frac{V(M)}{\lambda}}) \leq kc(c_\alpha)^{\frac{1}{q}} V(M)^{\frac{1}{q}} \left( \frac{m}{\lambda} \right)^{\frac{1}{q}}. \tag{4.5}$$

In particular, if $\int_Z |V'(y - f_{z, \lambda}^V(x))|^2 d\rho \leq \sigma^2$, then there is a constant $C = C(s, k, M, c_\alpha, c, V) > 0$ such that

$$\int_Z |V'(y - f_{\rho}^V(x)) - V'(y - f_{z, \lambda}^V(x))|^2 d\rho$$
Theorem 4.2. Under the conditions of Proposition 4.1, if

\[ \frac{a}{b} \]

then we have a constant \( C = C(s, \beta, k, M, C_s) > 0 \) such that

\[ \| f_{z, \lambda} - f_\rho \|_{L^2(\rho_X)}^2 \leq C \left[ \left( \frac{\log \frac{\rho}{\lambda}}{\lambda} \right)^{\frac{3}{2}} \sqrt{\frac{D(f_\rho, \lambda)}{\lambda m \sqrt{\rho}}} + D(f_\rho, \lambda) \right] \]

(4.7)

if \( \int_Z y^2 d\rho < +\infty \) and

\[ \frac{D(f_\rho, \lambda)}{\lambda} \leq m. \]

Moreover, we give the following theorems.

Theorem 4.1. Under the conditions of Proposition 4.1, if \( V(t) = \frac{t^2}{2} \), then we have a constant \( C = C(s, \beta, k, M, C_s) > 0 \) such that

\[ \| f_{z, \lambda} - f_\rho \|_{L^2(\rho_X)}^2 \leq C \left[ \left( \frac{\log \frac{\rho}{\lambda}}{\lambda} \right)^{\frac{3}{2}} \sqrt{\frac{D(f_\rho, \lambda)}{\lambda m \sqrt{\rho}}} + D(f_\rho, \lambda) \right] \]

(4.8)

if \( \int_Z y^2 d\rho < +\infty \). Let \( K \in C^\infty(X \times X) \), and \( \lambda = m^{-\frac{4}{6+3+2\beta(2+\gamma)}} \). Then under the assumption (2.9)

\[ \| f_{z, \lambda} - f_\rho \|_{L^2(\rho_X)}^2 = O \left[ \frac{\left( \log \frac{\rho}{\lambda} \right)^{\frac{3}{2}}}{\frac{\lambda}{m^{\frac{1}{6+3+2\beta(2+\gamma)}}}} + m^{-\frac{4\beta}{6+3+2\beta(2+\gamma)}} \right]. \]

(4.9)

Let \( K \in Lip^*(\alpha, C(X \times X)) \) for \( \alpha \to 0^+ \) and \( \lambda = m^{-\frac{4\beta}{6+3+2\beta(2+\gamma)}} \). Then under the assumption (2.9)

\[ \| f_{z, \lambda} - f_\rho \|_{L^2(\rho_X)}^2 = O \left[ \frac{\left( \log \frac{\rho}{\lambda} \right)^{\frac{3}{2}}}{\frac{\lambda}{m^{\frac{1}{6+3+2\beta(2+\gamma)}}}} + m^{-\frac{4\beta}{6+3+2\beta(2+\gamma)}} \right]. \]

(4.10)

Theorem 4.2. Under the conditions of Proposition 4.1, if \( V = V_1 \) and (2.9) holds for \( f_\rho^{V_1} \), then we have a constant \( C_1 = C_1(s, \beta, k, M, C_s) > 0 \) such that

\[ \int_Z \ln^2 \left( \frac{|y - f_{z, \lambda}^{V_1}(x)| + \sqrt{|y - f_{z, \lambda}^{V_1}(x)|^2 + 4}}{|y - f_\rho^{V_1}(x)| + \sqrt{|y - f_\rho^{V_1}(x)|^2 + 4}} \right) d\rho \]

\[ \leq \begin{cases} 
C_1 \left[ \frac{\left( \log \frac{\rho}{\lambda} \right)^{\frac{3}{2}}}{\frac{\lambda}{m^{\frac{1}{6+3+2\beta(2+\gamma)}}}} + \lambda^\beta \right], & \beta = 1; \\
C_1 \left[ \frac{\left( \log \frac{\rho}{\lambda} \right)^{\frac{3}{2}}}{\frac{\lambda}{m^{\frac{1}{6+3+2\beta(2+\gamma)}}}} + \lambda^\beta \right], & 0 < \beta < 1.
\end{cases} \]

(4.11)

Let \( K \in C^\infty(X \times X) \) and \( \lambda = m^{-\frac{4\beta}{6+3+2\beta(2+\gamma)}} \). Then

\[ \int_Z \ln^2 \left( \frac{|y - f_{z, \lambda}^{V_1}(x)| + \sqrt{|y - f_{z, \lambda}^{V_1}(x)|^2 + 4}}{|y - f_\rho^{V_1}(x)| + \sqrt{|y - f_\rho^{V_1}(x)|^2 + 4}} \right) d\rho \]
Under the conditions of Proposition 4.1, if
\[ f \in \mathcal{L}^1, \]\nTheorem 4.3.

Let
\[ \int_X \ln^2 \left( \frac{1 + |y - f^V_{Z,\lambda}(x)|}{1 + |y - f^V_{\rho}(x)|} \right) d\rho \]
\[ = \begin{cases} 
O \left( \frac{1}{m^\frac{1}{2+s}} \right), & \beta = 1; \\
O \left( \frac{1}{m^\frac{1}{2+s}} \right), & 0 < \beta < 1.
\end{cases} \] (4.12)

Moreover, let \( K \in \mathcal{L}^1 \) and \( f \) be as stated in Proposition 4.1. Then
\[ \int_X \ln^2 \left( \frac{1 + |y - f^V_{Z,\lambda}(x)|}{1 + |y - f^V_{\rho}(x)|} \right) d\rho = \begin{cases} 
O \left( \frac{1}{m^\frac{1}{2+s}} \right), & \beta = 1; \\
O \left( \frac{1}{m^\frac{1}{2+s}} \right), & 0 < \beta < 1.
\end{cases} \] (4.13)

Also the term \( \frac{1}{2+s} \) is useful since for a given small \( s > 0 \) and \( m \to +\infty \) we have
\[ \int_X \ln^2 \left( \frac{1 + |y - f^V_{Z,\lambda}(x)|}{1 + |y - f^V_{\rho}(x)|} \right) d\rho = \begin{cases} 
O \left( \frac{1}{m^\frac{1}{2+s}} \right), & \beta = 1; \\
O \left( \frac{1}{m^\frac{1}{2+s}} \right), & 0 < \beta < 1.
\end{cases} \] (4.14)

Theorem 4.3. Under the conditions of Proposition 4.1, if \( V = V_2 \) and (2.9) holds for \( f^V_{Z,\lambda} \), then we have
\[ \int_X \ln^2 \left( \frac{1 + |y - f^V_{Z,\lambda}(x)|}{1 + |y - f^V_{\rho}(x)|} \right) d\rho = \begin{cases} 
O \left( \frac{1}{m^\frac{1}{2+s}} \right), & \beta = 1; \\
O \left( \frac{1}{m^\frac{1}{2+s}} \right), & 0 < \beta < 1.
\end{cases} \] (4.15)

Moreover, let \( K \in \mathcal{L}^1 \) and \( f \) be as stated in Proposition 4.1. Then
\[ \int_X \ln^2 \left( \frac{1 + |y - f^V_{Z,\lambda}(x)|}{1 + |y - f^V_{\rho}(x)|} \right) d\rho = \begin{cases} 
O \left( \frac{1}{m^\frac{1}{2+s}} \right), & \beta = 1; \\
O \left( \frac{1}{m^\frac{1}{2+s}} \right), & 0 < \beta < 1.
\end{cases} \] (4.16)

The same results as that of the right-sides of (4.12)-(4.13) also hold for \( V_2 \):
Let \( K \in \mathcal{L}^1 \) and \( f \) be as stated in Proposition 4.1. Then
\[ \int_X \ln^2 \left( \frac{1 + |y - f^V_{Z,\lambda}(x)|}{1 + |y - f^V_{\rho}(x)|} \right) d\rho = \begin{cases} 
O \left( \frac{1}{m^\frac{1}{2+s}} \right), & \beta = 1; \\
O \left( \frac{1}{m^\frac{1}{2+s}} \right), & 0 < \beta < 1.
\end{cases} \] (4.17)

Moreover, let \( K \in \mathcal{L}^1 \) and \( f \) be as stated in Proposition 4.1. Then
\[ \int_X \ln^2 \left( \frac{1 + |y - f^V_{Z,\lambda}(x)|}{1 + |y - f^V_{\rho}(x)|} \right) d\rho = \begin{cases} 
O \left( \frac{1}{m^\frac{1}{2+s}} \right), & \beta = 1; \\
O \left( \frac{1}{m^\frac{1}{2+s}} \right), & 0 < \beta < 1.
\end{cases} \] (4.18)
\[
\begin{align*}
&= \begin{cases} 
O \left[ \left( \log \frac{4}{\delta} \right)^{\frac{2}{3}} \frac{(\ln m) \frac{2 \alpha}{\beta}}{m^{\frac{\alpha}{\beta} - 1}} \right], & \beta = 1; \\
O \left[ \left( \log \frac{4}{\delta} \right)^{\frac{2}{3}} \frac{(\ln m) \frac{2 \alpha}{\beta}}{m^{\frac{\alpha}{\beta} - 1}} \right], & 0 < \beta < 1.
\end{cases} 
\end{align*}
\] (4.16)

Theorem 4.4. Under the conditions of Proposition 4.1, if \( V = V_3 \) and (2.9) holds for \( f_\rho^{V_3} \), then we have
\[
\int_{X} \left( \frac{|y - f_{z,\lambda}^{V_3}(x)|}{\sqrt{1 + |y - f_{z,\lambda}^{V_3}(x)|^2}} - \frac{|y - f_{\rho}^{V_3}(x)|}{\sqrt{1 + |y - f_{\rho}^{V_3}(x)|^2}} \right)^2 d\rho = \begin{cases} 
O \left[ \left( \log \frac{4}{\delta} \right)^{\frac{2}{3}} \frac{(\ln m) \frac{2 \alpha}{\beta}}{m^{\frac{\alpha}{\beta} - 1}} \right], & \beta = 1; \\
O \left[ \left( \log \frac{4}{\delta} \right)^{\frac{2}{3}} \frac{(\ln m) \frac{2 \alpha}{\beta}}{m^{\frac{\alpha}{\beta} - 1}} \right], & 0 < \beta < 1.
\end{cases}
\] (4.17)

The similar results as that of the right-sides of (4.12)-(4.13) also hold for \( V_3(t) \).

It can be seen that the loss \( V_3 \) provide faster learning rates than that of the least square loss in the setting of this paper.

5. Further discussions. We now give some further analysis for the above results.

- **On the robustness.** In this paper, by robustness we mean the quantitative description of the dependence of \( f_{\rho,\lambda}^{V_3} \) on \( \mathcal{P}_Z \), which is different from the robustness in the kernel based robust regression ([38, 79]), where the robustness is the description of how the parameters in the windowing loss influences the learning rates.

- **Comparison the learning rates with the related literature.** For \( V(t) = t^2 \) Theorem 2 of [82] shows that if \( K \) satisfies (4.1) and (2.9) holds, then
\[
\|f_{z,\lambda} - f_{\rho}\|_{L^2(\rho,x)} = O \left[ \left( \log \frac{4}{\delta} \right)^{\frac{2}{3}} \frac{(\ln m) \frac{2 \alpha}{\beta}}{m^{\frac{\alpha}{\beta} - 1}} \right], \quad 0 < \varepsilon < \frac{\beta}{s + 1}. \] (5.1)

It is easy to see that (4.10) is nicer than (5.1) for sufficiently large \( s \). Also, if \( K \) satisfies (4.1) and (2.9) holds, then [83] shows that
\[
\|f_{z,\lambda} - f_{\rho}\|_{L^2(\rho,x)} = O \left[ \left( \frac{1}{m} \right)^{\frac{s \alpha}{s + \beta}} \log \frac{2}{\delta} \right]. \] (5.2)

Let \( W_2^s(X) \) be the Sobolev space defined in Appendix. Let \( f_{\rho} \in W_2^s(X) \). Then [33] shows a learning rate
\[
\left( \frac{1}{m} \right)^{\frac{s \alpha}{s + \beta}} \zeta, \quad \text{for all } \zeta > 0.
\] (5.3)

(5.3) is improved by [35] and shows that for \( f_{\rho} \in B_2^{s,\infty} \) (whose definition is in the appendix) a learning rate
\[
(\log m)^{q+1} \left( \frac{1}{m} \right)^{\frac{s \alpha}{s + \beta}}.
\] (5.4)

Then the right side of (4.9) is faster than (5.4) for sufficiently small \( s > 0 \) and \( q > 6 \).

The right side of (4.10) is faster than (5.2) for sufficiently large \( s \).

Above comparisons show that the error estimates in this paper have certain advantages over some in the literature. These advantages arise from the advantages of strong convexity. Because of the strong convexity of \( V^* \), the learning rate is contributed by both \( E \) and \( G \) (see the proof of Theorem 4.1) and \( V' \) has the Lipschitz property.
On the comparison inequality. By (iv) of Lemma 6.1 we know if $V$ is a strongly convex loss with convex modulus $c > 0$, then
\[
\frac{c}{2} \| f - f_{\rho}^V \|_{L^2(\rho_X)}^2 \leq \mathcal{E}_{\rho,V}(f) - \mathcal{E}_{\rho,V}(f_{\rho}^V), \quad f \in L^2(\rho_X). \tag{5.5}
\]
Also, if $\star$ is a differentiable convex loss and its conjugate loss $V^*$ is a differentiable strongly convex loss with convex modulus $1/c$, then we have the following comparison inequality with respect to $V^*$
\[
\mathcal{E}_{\rho,V}(f) - \mathcal{E}_{\rho,V}(f_{\rho}^V)
\geq \frac{1}{2c} \int_Z |V^*(y - f(x)) - V^*(y - f_{\rho}^V(x))|^2 \, d\rho, \quad f, f_{\rho}^V(x) \in L^2(\rho_X). \tag{5.6}
\]

Although both inequalities (5.5) and (5.6) are comparison theorems for a strongly convex loss, they are essential different. Inequality (5.6) is more practical than (5.5) when $V^*$ is a differentiable strongly convex loss, usually $V^*$ will satisfy Lipschitz condition and has high decay, however, $(V^*)'$ itself often has a high increasing rate, e.g. $V^*(t) = e^{t^2}$, the kernel regularized algorithm associating with it usually has very slow convergence rate ([56, 62]).

On the variance-expectation bound condition (1.13). We point out here that the variance-expectation bound condition (1.13) is satisfied by many convex losses (chapter 10.4 of [27]), in particular, when $V(t) = t^2$ we have
\[
E\{(y - f(x))^2 - (y - f_{\rho}^V(x))^2\} \leq 16M^2 (\mathcal{E}_{\rho}(f) - \mathcal{E}_{\rho}(f_{\rho}^V)). \tag{5.7}
\]

The variance-expectation bound condition (1.13) is an outcome of the capacity approach due to the error decomposition. In the approaches of the convex approach and the integral operator approach, it won’t be necessary. It can be applied to a large class of convex losses, but for some concrete losses it cannot provide the optimal learning rate. For example, if $V$ satisfies (1.13), (4.1) and (2.9), then[75] shows that
\[
\mathcal{E}_{\rho}(f_{\rho}^V) - \mathcal{E}_{\rho}(f_{\rho}^V) \leq c_2 \log(\frac{3}{\delta}) \left( \frac{1}{m} \right)^{\min\left\{ \frac{2\alpha}{1+\gamma}, \frac{2\beta}{2\gamma+\alpha+\beta+2\gamma} \right\}}. \tag{5.8}
\]

If $V(t) = t^2$, then (5.8) becomes (since $\alpha = 1$)
\[
\mathcal{E}_{\rho}(f_{\rho}^V) - \mathcal{E}_{\rho}(f_{\rho}^V) = O \left[ \log(\frac{3}{\delta}) \left( \frac{1}{m} \right)^{\frac{2\beta}{2\gamma+\alpha+\beta+2\gamma}} \right]. \tag{5.9}
\]

The right side of (5.9) is nice for very small $s > 0$, but is very weak for sufficiently large $s$ and is not as nice as that of (4.10).

On the $K$–functional. By the inequality (6.7) in Lemma 6.1 we have
\[
D(f_{\rho}^V, \lambda) \leq \inf_{g \in \mathcal{K}} \left[ \| g - f_{\rho}^V \|_{L^2(\rho_X)}^2 + \lambda \| g \|_{\mathcal{K}}^2 \right], \quad f_{\rho}^V \in L^2(\rho_X). \tag{5.10}
\]

The decay rate for the right side of (5.10) may be described by a modulus of smoothness ([58]).

On the characteristic kernels. It is known that the characteristic kernel theory is based on the universality of Mercer kernels. The universality of the deep convolution kernels (neural networks) has been studied by many mathematics ([51, 81, 88, 89]). The kernel regularized learning approach for multi-layer convolution kernel has been investigated by [8, 11] et al. It will be an interesting topic if one can extend the characteristic kernel theory to the deep convolution kernels (neural networks).
6. **Proofs.** To prove the theorems in Section 3 and Section 4, we need some lemmas.

6.1. **Lemmas.**

**Lemma 6.1.** Let $V$ be a differentiable convex loss and $V'$ be continuous on $R$, let $f_{\rho,\lambda}^V$ be the optimal solution of scheme (1.5). Then

(i) There holds
\[ \|f_{\rho,\lambda}^V\|_{\infty} \leq k\|f_{\rho,\lambda}^V\|_{K} \leq k \sqrt{\frac{D(f_{\rho}^V,\lambda)}{\lambda}}. \]  

(ii) There hold
\[ \nabla f_{\rho}E_{\rho,V}(f)(\cdot) = -\int_{V} V'(y - f(\cdot))d\rho(y), \quad f \in L^2(\rho_X) \]  

and
\[ \nabla f_{\rho,V}(f)(\cdot) = -\int_{V} V'(y - f(x))K_x(\cdot)d\rho, \quad f \in H_K. \]  

(iii) There holds
\[ f_{\rho,\lambda}^V(\cdot) = \frac{1}{2\lambda} \int_{V} V'(y - f_{\rho,\lambda}^V(x))K_x(\cdot)d\rho. \]  

(iv) If $V$ is a differentiable strongly convex loss with convex modulus $c > 0$, then
\[ E_{\rho,V}(f) - E_{\rho,V}(f_{\rho}^V) \geq \frac{c}{2}\|f - f_{\rho}^V\|_{L^2(\rho_X)}^2, \quad f, f_{\rho}^V \in L^2(\rho_X). \]  

(v) If $V$ is a differentiable convex loss and its conjugate loss $V^*$ is a differentiable strongly convex loss with convex modulus $\frac{1}{2}$, then
\[ E_{\rho,V}(f) - E_{\rho,V}(f_{\rho}^V) \geq \frac{1}{2c} \int_{X} \|f - f_{\rho}^V(x)\|_{2}^2 d\rho_X, \quad f, f_{\rho}^V(x) \in L^2(\rho_X). \]  

Proof. (i) By inequality (1.2) and the definition of $f_{\rho,\lambda}^V$ we have
\[ \|f_{\rho,\lambda}^V\|_{\infty} \leq k\|f_{\rho,\lambda}^V\|_{K} \leq k \sqrt{\frac{E_{\rho,V}(f_{\rho,\lambda}^V) - E_{\rho,V}(f_{\rho}^V) + \lambda\|f_{\rho,\lambda}^V\|_{K}^2}{\lambda}} = k \sqrt{\frac{D(f_{\rho}^V,\lambda)}{\lambda}}. \]  

(ii) Since $V$ is a differentiable convex loss we have by the (ii) of Proposition 7.1 in Appendix that \[ V'(t)(t-s) \leq V(t) - V(s) \leq V'(s)(t-s), \quad t, s \in R. \]  

It follows
\[ (-tg(x))V'[y - (f_{\rho,\lambda}^V(x) + tg(x))] \geq V[y - (f_{\rho,\lambda}^V(x) + tg(x))] - V[y - f_{\rho,\lambda}^V(x)] \]
\[ \geq V'[y - f_{\rho,\lambda}^V(x)](-tg(x)). \]

Then
\[ -\lim_{t \to 0} \int_{\mathbb{R}} V'[y - (f_{\rho,\lambda}^V(x) + tg(x))]g(x) d\rho \geq \lim_{t \to 0} \frac{E_{\rho,V}(f + tg) - E_{\rho,V}(f)}{t}. \]
Since
\[\lim_{t \to 0} \int_Z V'[y - (f_{r,\lambda}^V(x) + tg(x))]g(x)d\rho = \int_Z \lim_{t \to 0} V'[y - (f_{r,\lambda}^V(x) + tg(x))]g(x)d\rho \]
\[= \int_Z V'[y - f_{r,\lambda}^V(x)]g(x)d\rho. \tag{6.9}\]

We have by (6.9) and (6.8) that
\[\lim_{t \to 0} \frac{\mathcal{E}_{\rho,V}(f + tg) - \mathcal{E}_{\rho,V}(f)}{t} = -\int_Z V'[y - f_{r,\lambda}^V(x)]g(x)d\rho. \tag{6.10}\]

When \(g \in L^2(\rho X)\) we have
\[\int_Z V'[y - f_{r,\lambda}^V(x)]g(x)d\rho = \int_X \left( \int_Y V'[y - f_{r,\lambda}^V(x)]d\rho(y|x) \right)g(x)d\rho_X \]
\[= \langle \int_Y V'[y - f_{r,\lambda}^V(\cdot)]d\rho(y|\cdot), g \rangle_{L^2(\rho X)}. \tag{6.11}\]

By the definition of Gâteaux derivative we have (6.2). When \(g \in \mathcal{H}_K\) we have \(g(x) = \langle g, K(x) \rangle_K\). Hence
\[\int_Z V'[y - f_{r,\lambda}^V(x)]g(x)d\rho = \int_Z V'[y - f_{r,\lambda}^V(x)]\langle g, K(x) \rangle_K d\rho \]
\[= \langle g, \int_Z V'[y - f_{r,\lambda}^V(x)]K(x)d\rho \rangle_K. \tag{6.12}\]

where we have used the (6.17) in the Lemma 6.3. By the definition of Gâteaux derivative we have (6.3).

(iii) By the definition of \(f_{r,\lambda}^V\) and (i) of Proposition 7.1 in Appendix we have
\[0 = \nabla f(\mathcal{E}_{\rho,V}(f) + \lambda \|f\|^2_{K})|_{f = f_{r,\lambda}^V} = -\int_Z V'[y - f_{r,\lambda}^V(x)]K(x)d\rho + 2\lambda f_{r,\lambda}^V(\cdot). \]
So (6.4) holds.

(iv) By the definition of \(f_{\rho}^V\) we have
\[0 = \nabla f \mathcal{E}_{\rho,V}(f)|_{f = f_{\rho}^V} = -\int_Y V'[y - f_{\rho}^V(\cdot)]d\rho(y|\cdot). \tag{6.13}\]

On the other hand, by the property of Gâteaux derivative we have for \(z = (x,y) \in Z\) that
\[V(y - f(x)) - V(y - f_{\rho}^V(x)) \geq -V'(y - f_{\rho}^V(x))(f(x) - f_{\rho}^V(x)) + \frac{c}{2} |f(x) - f_{\rho}^V(x)|^2. \]
It follows
\[\mathcal{E}_{\rho,V}(f) - \mathcal{E}_{\rho,V}(f_{\rho}^V) \]
\[\geq -\int_Z V'(y - f_{\rho}^V(x))(f(x) - f_{\rho}^V(x))d\rho + \frac{c}{2} \int_X |f(x) - f_{\rho}^V(x)|^2d\rho_X \]
\[= -\int_X \left( \int_Y V'(y - f_{\rho}^V(x))d\rho(y|x) \right)(f(x) - f_{\rho}^V(x))d\rho_X + \frac{c}{2} \|f - f_{\rho}^V\|^2_{L^2(\rho X)} \]
\[= \frac{c}{2} \|f - f_{\rho}^V\|^2_{L^2(\rho X)}, \]
where in the last derivation we have used (6.13).
Lemma 6.2. Let \( \rho, \gamma \) be distributions on \( Z = X \times Y \). \( f_{\rho, \lambda}^V \) and \( f_{\gamma, \lambda}^V \) are the solutions of scheme (1.5) for \( \rho \) and \( \gamma \) respectively. If \( V \) is a differentiable convex loss and its conjugate loss \( V^* \) is a differentiable strongly convex loss with convex modulus \( \frac{1}{c} \), then

\[
\int_Z \left| V'(y - f_{\gamma, \lambda}^V(x)) - V'(y - f_{\rho, \lambda}^V(x)) \right|^2 d\rho 
\leq 2c \left\| f_{\rho, \lambda}^V - f_{\gamma, \lambda}^V \right\|_K.
\]

and (1.7) holds as well.

(1.7) and (6.15) show the robustness of the solutions according to the distributions.

Proof. Since \( V^* \) is a strongly convex function with convex modulus \( \frac{1}{c} \), we have by (2.7) that

\[
V'(y - f_{\rho, \lambda}^V(x)) - V'(y - f_{\gamma, \lambda}^V(x)) 
\geq -V'(y - f_{\rho, \lambda}^V(x))(f_{\rho, \lambda}^V(x) - f_{\gamma, \lambda}^V(x)) + \frac{1}{2c} \left| V'(y - f_{\rho, \lambda}^V(x)) - V'(y - f_{\gamma, \lambda}^V(x)) \right|^2.
\]

Therefore

\[
\int_Z V'(y - f_{\rho, \lambda}^V(x)) d\rho - \int_Z V'(y - f_{\gamma, \lambda}^V(x)) d\rho 
\geq \left\langle f_{\rho, \lambda}^V - f_{\gamma, \lambda}^V, - \int_Z V'(y - f_{\gamma, \lambda}^V(x)) K_x(\cdot) d\rho \right\rangle_K 
+ \frac{1}{2c} \int_Z \left| V'(y - f_{\rho, \lambda}^V(x)) - V'(y - f_{\gamma, \lambda}^V(x)) \right|^2 d\rho.
\]
Therefore, by (iii) of Proposition 7.1 in Appendix we have
\[ \| f_{p,\lambda}^V - f_{\gamma,\lambda}^V \|_K^2 = (f_{p,\lambda}^V - f_{\gamma,\lambda}^V, 2 f_{p,\lambda}^V - f_{\gamma,\lambda}^V) + \| f_{p,\lambda}^V - f_{\gamma,\lambda}^V \|_K \]

Therefore,
\[
(f_{p,\lambda}^V - f_{\gamma,\lambda}^V, - \int_Z V'(y - f_{\gamma,\lambda}^V(x)) K_x(\cdot) d\rho) + 2\lambda(f_{p,\lambda}^V - f_{\gamma,\lambda}^V, f_{p,\lambda}^V - f_{\gamma,\lambda}^V) + \lambda\| f_{p,\lambda}^V - f_{\gamma,\lambda}^V \|_K^2 + \frac{1}{2c} \int_Z |V'(y - f_{p,\lambda}^V(x)) - V'(y - f_{\gamma,\lambda}^V(x))|^2 d\rho
\]
\[\geq \left( f_{p,\lambda}^V - f_{\gamma,\lambda}^V, \int_Z V'(y - f_{\gamma,\lambda}^V(x)) K_x(\cdot) d\gamma - \int_Z V'(y - f_{\gamma,\lambda}^V(x)) K_x(\cdot) d\gamma \right)
\]
\[\geq \left( f_{p,\lambda}^V - f_{\gamma,\lambda}^V, \int_Z V'(y - f_{\gamma,\lambda}^V(x)) K_x(\cdot) d\gamma - \int_Z V'(y - f_{\gamma,\lambda}^V(x)) K_x(\cdot) d\gamma \right)_K \]
\[+ \frac{1}{2c} \int_Z |V'(y - f_{p,\lambda}^V(x)) - V'(y - f_{\gamma,\lambda}^V(x))|^2 d\rho + \lambda\| f_{p,\lambda}^V - f_{\gamma,\lambda}^V \|_K^2,\]
where in the last deduce we have used (6.4).

The definitions of \( f_{p,\lambda}^V \) and \( f_{\gamma,\lambda}^V \) lead to
\[ 0 \geq (\mathcal{E}_{p,V}(f_{p,\lambda}^V) + \lambda\| f_{p,\lambda}^V \|_K^2) - (\mathcal{E}_{p,V}(f_{\gamma,\lambda}^V) + \lambda\| f_{\gamma,\lambda}^V \|_K^2). \]

It follows
\[ \lambda\| f_{p,\lambda}^V - f_{\gamma,\lambda}^V \|_K^2 + \frac{1}{2c} \int_Z |V'(y - f(x)) - V'(y - f_{\gamma,\lambda}^V(x))|^2 d\rho \]
\[\leq \left( f_{p,\lambda}^V - f_{\gamma,\lambda}^V, \int_Z V'(y - f_{\gamma,\lambda}^V(x)) K_x(\cdot) d\gamma - \int_Z V'(y - f_{\gamma,\lambda}^V(x)) K_x(\cdot) d\gamma \right)_K \]
\[\leq \left( f_{p,\lambda}^V - f_{\gamma,\lambda}^V, \int_Z V'(y - f_{\gamma,\lambda}^V(x)) K_x(\cdot) d\gamma - \int_Z V'(y - f_{\gamma,\lambda}^V(x)) K_x(\cdot) d\gamma \right)_K \]
\[\times \| f_{p,\lambda}^V - f_{\gamma,\lambda}^V \|_K, \quad (6.16)\]
where in the last derivation we have used Cauchy’s inequality. We thus have (1.7) and (6.15) by (6.16).

**Lemma 6.3.** Let \( K \) be a \( d\rho_X \)-measurable and bounded kernel on \( X \times X, \mathcal{H}_K \) be its reproducing kernel Hilbert space, \( M \) be a bounded function on \( X \). Then for any \( f \in \mathcal{H}_K \) we have
\[ \int_X M(x) f(x) d\rho_X = \int_X M(x) \langle f, K_x \rangle d\rho_X = \langle f, \int_X M(x) K_x(\cdot) d\rho_X \rangle_K. \quad (6.17) \]

**Proof.** (6.17) is a generalization of the Lemma 26 of [67]. Define a linear functional by
\[ T_{\rho_X}(f) = \int_X M(u) f(u) d\rho_X(u). \]
Then
\[ \| T_{\rho_X}(f) \| = \sup_{f \in \mathcal{H}_K} \frac{|T_{\rho_X}(f)|}{\|f\|_K} \leq \sup_{f \in \mathcal{H}_K} \frac{\int_X M(x) \langle f, K_x \rangle d\rho_X}{\|f\|_K} \leq \sup_{x \in X} |M(x)| \sqrt{\sup_{x \in X} K(x,x)} \leq \| M \|_{\infty} k < +\infty. \]
Lemma 6.4. Let \( b \) be a bounded function on \( Z \). Then

\[
\left\| \int_Z b(x, y)K_x(\cdot)d\rho - \int_Z b(x, y)K_x(\cdot)d\gamma \right\|_K \leq d^b_K(\rho, \gamma),
\]

where

\[
d^b_K(\rho, \gamma) = \sup_{\|h\|_K \leq 1} \left| \int_Z b(x, y)h(x)d\rho - \int_Z b(x, y)h(x)d\gamma \right|.
\]

Proof. Since \( \mathcal{H}_K \) is a Hilbert space, we have \( \|f\|_K = \sup_{\|g\|_K \leq 1} \langle f, g \rangle_K \). It follows

\[
\left\| \int_Z b(x, y)K_x(\cdot)d\rho - \int_Z b(x, y)K_x(\cdot)d\gamma \right\|_K
= \sup_{\|h\|_K \leq 1} \left| \int_X \left( \int_Y b(x, y)K_x(\cdot)d\rho(y|x) \right) d\rho_X - \int_X \left( \int_Y b(x, y)K_x(\cdot)d\gamma(y|x) \right) d\gamma_X \right|
= \sup_{\|h\|_K \leq 1} \left| \int_X \left( \int_Y b(x, y)K_x(\cdot)d\rho(y|x) \right) K_x(\cdot)d\rho_X - \int_X \left( \int_Y b(x, y)K_x(\cdot)d\gamma(y|x) \right) K_x(\cdot)d\gamma_X \right|
= \sup_{\|h\|_K \leq 1} \left| \int_X \left( \int_Y b(x, y)h(x)d\rho(y|x) \right) d\rho_X - \int_X \left( \int_Y b(x, y)h(x)d\gamma(y|x) \right) d\gamma_X \right|.
\]

Replacing \( M \) in (6.17) by \( \int_Y b(x, y)d\rho(y|x) \) and \( \int_Y b(x, y)d\gamma(y|x) \) respectively we have

\[
\int_X \left( \int_Y b(x, y)d\rho(y|x) \right) h(x)d\rho_X = \left\langle h, \int_X \left( \int_Y b(x, y)d\rho(y|x) \right) K_x(\cdot)d\rho_X \right\rangle_K,
\]

and

\[
\int_X \left( \int_Y b(x, y)d\gamma(y|x) \right) h(x)d\rho_X = \left\langle h, \int_X \left( \int_Y b(x, y)d\gamma(y|x) \right) K_x(\cdot)d\rho_X \right\rangle_K.
\]

Therefore

\[
\left\| \int_Z b(x, y)K_x(\cdot)d\rho - \int_Z b(x, y)K_x(\cdot)d\gamma \right\|_K
= \sup_{\|h\|_K \leq 1} \left| \int_X \left( \int_Y b(x, y)d\rho(y|x) \right) h(x)d\rho_X - \int_X \left( \int_Y b(x, y)d\gamma(y|x) \right) h(x)d\Gamma \right|
= \sup_{\|h\|_K \leq 1} \left| \int_Z b(x, y)h(x)d\rho - \int_Z b(x, y)h(x)d\gamma \right|.
\]

(6.18) thus holds. \( \square \)
Lemma 6.5 (Proposition 3.13 of [27]). Let $\mathcal{F}$ be a family of functions from a probability space $Z$ to $R$ and $d(\cdot, \cdot)$ a metric on $\mathcal{F}$. Let $\mathcal{U} \subset Z$ be of full measure and constants $B, L > 0$ such that

(i) $|\xi(z)| \leq B$ for all $\xi \in \mathcal{F}$ and all $z \in \mathcal{U}$, and

(ii) $|L_z(\xi_1) - L_z(\xi_2)| \leq L d(\xi_1, \xi_2)$ for all $\xi_1, \xi_2 \in \mathcal{F}$ and all $Z = (z_i)_{i=1}^m \in \mathcal{U}^m$, where

$$L_z(\xi) = \int_Z \xi(z) - \frac{1}{m} \sum_{i=1}^m \xi(z_i).$$

Then for all $\epsilon > 0$,

$$\text{Prob}_{Z \in \mathcal{Z}^m}\left\{ \sup_{\xi \in \mathcal{F}} |L_z(\xi)| \leq \epsilon \right\} \geq 1 - N(\mathcal{F}, \frac{\epsilon}{2L}) \times 2 \exp\left( - \frac{m \epsilon^2}{8B^2} \right). \quad (6.19)$$

Lemma 6.6. Let $\mathcal{F}$ be the family of functions defined as in Lemma 6.5. Let $V$ be a convex loss on $R$ with $c$-Lipschitz continuous. For the class $\mathcal{V}_F$ consisting of loss functions $\mathcal{V}_F = \{V(y - \eta(z)) : \eta \in \mathcal{F}, |y| \leq B\}$, there holds

$$\text{Prob}_{Z \in \mathcal{Z}^m}\left\{ \sup_{\xi \in \mathcal{V}_F} |L_z(\xi)| \leq \epsilon \right\} \geq 1 - N(\mathcal{V}_F, \frac{\epsilon}{2Lc}) \times 2 \exp\left( - \frac{m \epsilon^2}{8 V^2(2B)} \right). \quad (6.20)$$

Proof. Since $|V(y - \eta(z))| \leq V(2B)$, we have by (6.19) that

$$\text{Prob}_{Z \in \mathcal{Z}^m}\left\{ \sup_{\xi \in \mathcal{V}_F} |L_z(\xi)| \leq \epsilon \right\} \geq 1 - N(\mathcal{V}_F, \frac{\epsilon}{2Lc}) \times 2 \exp\left( - \frac{m \epsilon^2}{8 V^2(2B)} \right). \quad (6.21)$$

By (2.6) we know for $\eta', \eta \in \mathcal{F}$ there holds

$$|V(y - \eta'(z)) - V(y - \eta(z))| \leq c|\eta'(z) - \eta(z)|,$$

Therefore

$$N(\mathcal{V}_F, \frac{\epsilon}{2Lc}) \leq N(\mathcal{F}, \frac{\epsilon}{2Lc}). \quad (6.22)$$

Collecting (6.22) and (6.21) we have (6.20). \hfill \Box

Lemma 6.7 ([26]). Let $c_1 > 0, c_2 > 0$ and $u > t > 0$. Then the equation

$$x^u - c_1 x^t - c_2 = 0$$

has a unique positive zero $x^b$. In addition, there holds

$$x^b \leq \max \left\{ \left( 2c_1 \right)^{\frac{1}{u-t}}, \left( 2c_2 \right)^{\frac{1}{t}} \right\}. \quad (6.23)$$

Lemma 6.8. Let $a, b$ and $c$ be nonnegative functions of random variables $z \in Z^m$. They satisfy almost surely

$$a(z) \leq b(z)c(z). \quad (6.24)$$

Let $\varepsilon_i, \delta_i < 1$ be real numbers and $0 < \varepsilon_i, 0 < \delta_i < 1, (i = 1, 2)$. If $P\{z \in Z^m : b(z) \leq \varepsilon_1\} \geq 1 - \delta_1$ and $P\{z \in Z^m : c(z) \leq \varepsilon_2\} \geq 1 - \delta_2$, then,

$$P\{z \in Z^m : a(z) \leq \varepsilon_1 \varepsilon_2\} \geq 1 - \delta_1 - \delta_2. \quad (6.25)$$

Proof. Define events $B = \{z \in Z^m : b(z) \leq \varepsilon_1\}, C = \{z \in Z^m : c(z) \leq \varepsilon_2\}$ and $A = \{z \in Z^m : a(z) \leq \varepsilon_1 \varepsilon_2\}$. Then, by (6.24) we have $BC \subset A$, therefore, $P(BC) \leq P(A)$. Since $P(A) \geq P(B) + P(C) - P(B \cup C) \geq P(B) + P(C) - 1$, we have

$$P(A) \geq (1 - \delta_1) + (1 - \delta_2) - 1 = 1 - \delta_1 - \delta_2. \quad \Box$$
Lemma 6.9. The function
\[ V_1(t) = |t| \ln \left| \frac{|t| + \sqrt{t^2 + 4}}{2} - \sqrt{t^2 + 4} + 2 \right|, \quad t \in R \]
is an even convex loss on R. \( V_1' \) is a Lipschitz function with Lipschitz constant \( \frac{1}{2} \), whose conjugate loss is
\[ V_1^*(u) = e^{\|u\|} + e^{-\|u\|} - 2, \quad u \in R, \]
where \( \|u\| \) is an even convex loss on \( R \). Assume that there are two positive constants \( \sigma \) and \( \sigma' \). \( \Box \)

Lemma 6.10. The function
\[ V_3(u) = \sqrt{1 + u^2} - 1, \quad u \in R \]
is an even convex loss on R. \( V_3' \) is Lipschitz with Lipschitz constant 2, whose conjugate function is
\[ V_3^*(u) = \begin{cases} 1 - \sqrt{1 - u^2}, & |u| < 1; \\ +\infty, & |u| \geq 1, \end{cases} \]
which is a strongly convex loss with convex modulus 1.

Proof. By the definition of conjugate function we have
\[ V_3(t) = \max_{u \geq 0} \{ |t|u - (e^u + e^{-u} - 2) \} = \max_{u \geq 0} \Omega_t(u), \quad t \in R, \]
where \( \Omega_t(u) = |t|u - (e^u + e^{-u} - 2) \). If \( \frac{d\Omega_t(u)}{u} = 0 \), then \( |t| - (e^u - e^{-u}) = 0 \), which gives the stagnation point \( u_0 = \ln \frac{|t| + \sqrt{t^2 + 4}}{2} \). Therefore,
\[ V_3(t) = \Omega_t(u_0) = |t|u_0 - (e^{u_0} + e^{-u_0} - 2) = |t| \ln \left| \frac{|t| + \sqrt{t^2 + 4}}{2} - \sqrt{t^2 + 4} + 2 \right|, \quad t \in R. \]
Simple computations yield \( V_3'(t) = \ln \frac{|t| + \sqrt{t^2 + 4}}{2} \) and \( V_3''(t) = \frac{1}{\sqrt{t^2 + 4}} \leq \frac{1}{2} \). So \( V_3'(t) \) is a \( \frac{1}{2} \)-Lipschitz function. Also by \( V_3^{**}(u) = e^u + e^{-u} \geq 2 \) and Taylor formula we know \( V_3^*(u) \) is a strongly convex function with convex modulus 2. \( \Box \)

Lemma 6.11. Let \( \xi \) be a random variable taking values in a real separable Hilbert space \( H \) on a probability space \( (\Omega, \mathcal{F}, P) \). Assume that there are two positive constants \( L \) and \( \sigma \) such that
\[ ||\xi||_H \leq L \quad \text{and} \quad E(||\xi||_H^2) \leq \sigma^2. \]
Then, for all $n \geq 1$ and $0 < \eta < 1$, it holds, with confidence $1 - \eta$, that
\[
\left\| \frac{1}{n} \sum_{i=1}^{n} \xi(\omega_i) - E(\xi) \right\|_H \leq \frac{2L}{n} \log \frac{2}{\eta} + \sqrt{\frac{2\sigma^2 \log \frac{2}{\eta}}{n}}.
\] (6.26)
In particular, we have
\[
\left\| \frac{1}{n} \sum_{i=1}^{n} \xi(\omega_i) - E(\xi) \right\|_H \leq \frac{4L}{\sqrt{n}} \log \frac{2}{\eta}.
\] (6.27)

Proof. See [65] or [31].

6.2. Proof of the results.

Proof of Proposition 3.1. By (1.7) and (6.15) we have
\[
\int_Z \left| V'(y - f^V_{\gamma,\lambda}(x)) - V'(y - f^V_{\gamma,\lambda}(x)) \right|^2 \, d\rho
\]
\[
= \frac{1}{\lambda} \left\| \int_Z \left| V'(y - f^V_{\gamma,\lambda}(x)) - V'(y - f^V_{\gamma,\lambda}(x)) \right|^2 \, d\rho \right\|_{L^1} \times \left\| f^V_{\gamma,\lambda} - f^V_{\gamma,\lambda} \right\|_K
\]
\[
\leq \frac{2c}{\lambda} \left\| \int_Z V'(y - f^V_{\gamma,\lambda}(x)) K_x(\cdot) d\rho - \int_Z V'(y - f^V_{\gamma,\lambda}(x)) K_x(\cdot) d\gamma \right\|_K
\]
\[
\leq \frac{2c}{\lambda} \left\| \int_Z V'(y - f^V_{\gamma,\lambda}(x)) K_x(\cdot) d\rho - \int_Z V'(y - f^V_{\gamma,\lambda}(x)) K_x(\cdot) d\gamma \right\|_K.
\] (6.28)

By Lemma 6.4 we have
\[
\left\| \int_Z V'(y - f^V_{\gamma,\lambda}(x)) K_x(\cdot) d\rho - \int_Z V'(y - f^V_{\gamma,\lambda}(x)) K_x(\cdot) d\gamma \right\|_K
\]
\[
\leq \sup_{|h| \leq 1} \left| \int_Z V'(y - f^V_{\gamma,\lambda}(x)) h(x) d\rho - \int_Z V'(y - f^V_{\gamma,\lambda}(x)) h(x) d\gamma \right|.
\] (6.29)

and
\[
\left\| \int_Z V'(y - f^V_{\gamma,\lambda}(x)) K_x(\cdot) d\rho - \int_Z V'(y - f^V_{\gamma,\lambda}(x)) K_x(\cdot) d\gamma \right\|_K
\]
\[
\leq \sup_{|h| \leq 1} \left| \int_Z V'(y - f^V_{\gamma,\lambda}(x)) h(x) d\rho - \int_Z V'(y - f^V_{\gamma,\lambda}(x)) h(x) d\gamma \right|.
\] (6.30)

Collecting (6.28), (6.29) and (6.30) we have (3.1).}

Proof of Theorem 3.1 i.e. Proof of (3.3). Since $V'$ is a Lipschitz function with a Lipschitz constant $c > 0$, we have for any $(x, y), (x', y') \in Z$ and a given $h \in H_K$ satisfying $\|h\|_K \leq 1$ that
\[
\left| V'(y - f^V_{\gamma,\lambda}(x)) h(x) - V'(y' - f^V_{\gamma,\lambda}(x')) h(x') \right|
\]
\[
\leq \left| \left( V'(y - f^V_{\gamma,\lambda}(x)) - V'(y' - f^V_{\gamma,\lambda}(x')) \right) h(x) \right|
\]
\[
+ \left| V'(y - f^V_{\gamma,\lambda}(x')) (h(x) - h(x')) \right|
\]
\[
\leq c \left( |y - y'| + |f^V_{\gamma,\lambda}(x) - f^V_{\gamma,\lambda}(x')| \right) + \left| V'(y - f^V_{\gamma,\lambda}(x')) | h(x) - h(x') | \right|
\]
\[
\leq c \left( |y - y'| + \| f^V_{\gamma,\lambda} K_x - K_{x'} \|_K \right) + V' \left( 2k \sqrt{\frac{D(f^V_{\gamma,\lambda}, \lambda)}{\lambda}} \right) | h, K_x - K_{x'} |_K.
\]
Collecting \((6.33), (6.32)\) and \((3.1)\) we have \((3.3)\).

In the same method, we have

\[
\text{Replacing } f, \gamma, \lambda, \text{ and } \kappa \text{ where we have used } (6.1) \text{ and the inequality } \langle f, g \rangle_K \leq \|f\|_K \|g\|_K, \ f, g \in \mathcal{H}_K.
\]

It follows
\[
\frac{V'(y - f_{\gamma, \lambda}^V(x)) h(x)}{2c\beta} \leq d_H(z, z')_K \tag{6.31}
\]
and \(\frac{V'(y - f_{\gamma, \lambda}^V(x)) h(x)}{2c\beta} \in \text{Lip}1\). By the definition of \(D_H(\rho, \gamma)\) we have

\[
\sup_{\|h\|_K \leq 1} \left| \int Z V'(y - f_{\gamma, \lambda}^V(x)) h(x) \, d\rho - \int Z V'(y - f_{\gamma, \lambda}^V(x)) h(x) \, d\gamma \right| \leq 2c\beta \sup_{\|h\|_K \leq 1} \left| \int Z \frac{V'(y - f_{\gamma, \lambda}^V(x)) h(x)}{2c\beta} \, d\rho - \int Z \frac{V'(y - f_{\gamma, \lambda}^V(x)) h(x)}{2c\beta} \, d\gamma \right| .
\]

\[
\leq 2c\beta \sup_{\|h\|_K \leq 1} \left| \int Z f \, d\rho - \int Z f \, d\gamma : f \in \text{Lip}1 \right| = 2c\beta D_H(\rho, \gamma). \tag{6.32}
\]

In the same method, we have

\[
\sup_{\|h\|_K \leq 1} \left| \int Z V'(y - f_{\rho, \lambda}^V(x)) h(x) \, d\rho - \int Z V'(y - f_{\rho, \lambda}^V(x)) h(x) \, d\gamma \right| \leq 2c\beta' \sup_{\|h\|_K \leq 1} \left| \int Z \frac{V'(y - f_{\rho, \lambda}^V(x)) h(x)}{2c\beta'} \, d\rho - \int Z \frac{V'(y - f_{\rho, \lambda}^V(x)) h(x)}{2c\beta'} \, d\gamma \right| .
\]

\[
\leq 2c\beta' \sup_{\|h\|_K \leq 1} \left| \int Z f \, d\rho - \int Z f \, d\gamma : f \in \text{Lip}1 \right| = 2c\beta' D_H(\rho, \gamma). \tag{6.33}
\]

Collecting \((6.33), (6.32)\) and \((3.1)\) we have \((3.3)\).

\[\square\]

**Proof of (3.4).** Replacing \(\text{sgn}(V'(y - f_{\gamma, \lambda}^V(x)))\) \(\text{sgn}(h(x))h(x)\) with \(h(x)\), we have

\[
\sup_{\|h\|_K \leq 1} \left| \int Z V'(y - f_{\gamma, \lambda}^V(x)) h(x) \, d\rho - \int Z V'(y - f_{\gamma, \lambda}^V(x)) h(x) \, d\gamma \right| = \sup_{\|h\|_K \leq 1} \left| \int Z V'(y - f_{\gamma, \lambda}^V(x)) |h(x)| \, dp - \int Z V'(y - f_{\gamma, \lambda}^V(x)) |h(x)| \, d\gamma \right| = D. \tag{6.34}
\]

Since \(|V'(t)| \leq M'\) we have

\[
D = M' \sup_{\|h\|_K \leq 1} \left| \int Z \frac{|V'(y - f_{\gamma, \lambda}^V(x))|}{M'} \times |h(x)| \, dp - \int Z \frac{|V'(y - f_{\gamma, \lambda}^V(x))|}{M'} \times |h(x)| \, d\gamma \right| \leq M' \sup_{\|h\|_K \leq 1} \left| \int Z |h(x)| \, dp - \int Z |h(x)| \, d\gamma \right| = M' \, d_K(\rho_X, \gamma_X). \tag{6.35}
\]
where we have used the fact that
\[
\int_Z |h(x)| \, d\rho = \int_X \left( \int_Y \rho(y|x) \right) |h(x)| \, d\rho_X = \int_X |h(x)| \, d\rho_X.
\]
By the same way, replacing \(sgn(V'(y - f_{\rho,\lambda}^V(x)))sgn(h(x))h(x)\) with \(h(x)\), we have
\[
\sup_{\|h\|_\kappa \leq 1} \left| \int_Z V'(y - f_{\rho,\lambda}^V(x))h(x) \, d\rho - \int_Z V'(y - f_{\rho,\lambda}^V(x))h(x) \, d\gamma \right|
\leq M' \sup_{\|h\|_\kappa \leq 1} \left| \int_Z |h(x)| \, d\rho - \int_Z |h(x)| \, d\gamma \right| = M'd_K(\rho_X, \gamma_X). \tag{6.36}
\]
Collecting (6.34), (6.35), (6.36) and (3.1) we have (3.4). \(\Box\)

**Proof of (3.5).** In this case \(f_{\rho}^V = f_{\rho}\). By the comparison equality (1.11) we have
\[
D(f_{\rho}^V, \lambda) = D(f_{\rho}, \lambda) = \inf_{g \in \mathbb{H}_\kappa} \left[ \|g - f_{\rho}\|_{L^2(\rho_X)}^2 + \lambda \|g\|_K^2 \right],
\]
which, the inequality (1.2), \(|V_2^V(t)| = 2t\) for \(t \geq 0\) together with (3.3) give (3.5). \(\Box\)

**Proof of (3.6).** Taking \(V_1^V(t) = \ln \frac{\rho + \rho_{\rho,\lambda}^V}{\rho - \rho_{\rho,\lambda}^V}\) in (6.13), we know \(f_{\rho}^V_{\rho_1}\) satisfies (3.7). Also, taking \(V_1^V(t)\) in (3.3) we have (3.6). \(\Box\)

**Proof of (3.8).** Taking \(V_2^V(t) = \ln(1 + |t|)\) in (6.13), we know \(f_{\rho}^V_{\rho_2}\) satisfies (3.9). Taking \(V_2^V(t)\) in (3.3) we have (3.8). \(\Box\)

**Proof of (3.10).** Taking \(V_3^V(t) = \frac{|t|}{\sqrt{1 + t^2}}\) in (6.13), we know \(f_{\rho}^V_{\rho_3}\) satisfies (3.11). Taking \(V_3^V(t)\) in (3.4) we have (3.10). \(\Box\)

**Proof of Proposition 4.1.** By the inequality
\[
|a - b|^2 \leq 4|a - c|^2 + 4|c - b|^2, \quad a, b, c \in \mathbb{R},
\]
we have
\[
\int_Z \left| V'(y - f_{z,\lambda}^V(x)) - V'(y - f_{\rho}^V(x)) \right|^2 \, d\rho
\leq 4 \int_Z \left| V'(y - f_{z,\lambda}^V(x)) - V'(y - f_{\rho,\lambda}^V(x)) \right|^2 \, d\rho
+ 4 \int_Z \left| V'(y - f_{\rho,\lambda}^V(x)) - V'(y - f_{\rho}^V(x)) \right|^2 \, d\rho. \tag{6.37}
\]
Collecting inequality (6.15)(taking \(\gamma = \rho_{\rho,\lambda}\)) and inequality (1.7) we have
\[
\int_Z \left| V'(y - f_{z,\lambda}^V(x)) - V'(y - f_{\rho,\lambda}^V(x)) \right|^2 \, d\rho
= \int_Z \left| V'(y - f_{z,\lambda}^V(x)) - V'(y - f_{\rho,\lambda}^V(x)) \right|^2 \, d\rho
\leq \frac{2c}{\lambda} E G, \tag{6.38}
\]
where
\[
E = \sup_{\|h\|_\kappa \leq 1} \left| \int_Z V'(y - f_{z,\lambda}^V(x))h(x) \, d\rho - \frac{1}{m} \sum_{i=1}^m V'(y_i - f_{z,\lambda}^V(x_i))h(x_i) \right|
\]
and
\[
G = \left\| \int_Z V' \left( y - f_{p,\lambda}^{V,(z)}(x) \right) K_z(\cdot) \, d\rho - \frac{1}{m} \sum_{i=1}^{m} V' \left( y_i - f_{p,\lambda}^{V,(z)}(x_i) \right) K_{x_i}(\cdot) \right\|_K.
\]

By (6.37) and (6.38) we have
\[
\begin{align*}
&\int_Z \left| V'(y - f_{p,\lambda}^{V,(z)}(x)) - V'(y - f_{p}^{V}(x)) \right|^2 \, d\rho \\
\leq &\int_Z \left| V'(y - f_{p,\lambda}^{V,(z)}(x)) - V'(y - f_{p,\lambda}^{V,(z)}(x)) \right|^2 \, d\rho + 8c \left( \varepsilon_{p,V}(f_{p,\lambda}^{V}) - \varepsilon_{p,V}(f_{p}^{V}) + \lambda \left\| f_{p,\lambda}^{V} \right\|_K^2 \right) \\
\leq &\frac{8c}{\lambda} E_G + 8c D(f_{p}^{V}, \lambda),
\end{align*}
\]

(6.39)

Take
\[
L_z(\xi) = \int_Z \xi(x, y) \, d\rho - \frac{1}{m} \sum_{i=1}^{m} \xi(x_i, y_i), \quad \xi \in C(Z), z = \{(x_i, y_i)\}_{i=1}^{m}.
\]

Then for \( \xi_1, \xi_2 \in C(Z) \)
\[
|L_z(\xi_1) - L_z(\xi_2)|
\]
\[
\leq \left| \left( \int_Z \xi_1(x, y) \, d\rho - \frac{1}{m} \sum_{i=1}^{m} \xi_1(x_i, y_i) \right) \right| - \left( \int_Z \xi_2(x, y) \, d\rho - \frac{1}{m} \sum_{i=1}^{m} \xi_2(x_i, y_i) \right) | \\
\leq \left| \int_Z \xi_1(x, y) \, d\rho - \int_Z \xi_2(x, y) \, d\rho \right| + \left| \frac{1}{m} \sum_{i=1}^{m} \xi_1(x_i, y_i) - \frac{1}{m} \sum_{i=1}^{m} \xi_2(x_i, y_i) \right| \\
\leq 2 \| \xi_1 - \xi_2 \|_{C(Z)}
\]

and
\[
\left| V'(y - f_{p,\lambda}^{V,(z)}(x)) h(x) \right| \leq k \left( 2k \sqrt{\frac{V(M)}{\lambda}} \right) \| h \|_K \leq k \left( 2k \sqrt{\frac{V(M)}{\lambda}} \right).
\]

Define
\[
\mathcal{F} = \{ \eta_\varepsilon(x) \in \mathcal{H}_K : \| \eta_\varepsilon \|_K \leq \sqrt{\frac{V(M)}{\lambda}} \} \text{ and } V'_\varepsilon = \{ V'(y - \eta_\varepsilon(x)) : \eta_\varepsilon(x) \in \mathcal{F} \}.
\]

By Lemma 6.5 we have
\[
\Prob_{Z \in Z^m} \{ |E| \leq \varepsilon \} \geq 1 - \mathcal{N} \left( \frac{\mathcal{F}}{4k}, \frac{\epsilon}{4k} \right) 2 \exp \left( - \frac{m \epsilon^2}{8 k^2 \left( V'(2k \sqrt{\frac{V(M)}{\lambda}}) \right)^2} \right).
\]

(6.40)

Since \( V' \) is a differentiable strongly convex loss with convex modulus \( \frac{1}{\varepsilon} \), we have by (2.6) and Lemma 6.6 that
\[
\Prob_{Z \in Z^m} \{ |E| \leq \varepsilon \} \geq 1 - \mathcal{N} \left( \frac{\mathcal{F}}{4kc}, \frac{\varepsilon}{4kc} \right) 2 \exp \left( - \frac{m \epsilon^2}{8 k^2 \left( V'(2k \sqrt{\frac{V(M)}{\lambda}}) \right)^2} \right).
\]

(6.41)

Since (4.1) we have
\[
\mathcal{N} \left( \frac{\mathcal{F}}{4kc}, \frac{\varepsilon}{4kc} \right) \leq \exp \left\{ c_s \left( \frac{k \sqrt{\frac{V(M)}{\lambda}}}{\varepsilon} \right)^s \right\} = \exp \left\{ c_s \left( \frac{4k^2 c \sqrt{\frac{V(M)}{\lambda}}}{\varepsilon} \right)^s \right\}.
\]
It follows that
\[ \text{Prob}\{E \leq \varepsilon \} \geq 1 - 2 \exp\left\{ -c_s\left(4k^2c\sqrt{\frac{V(M)}{\lambda}}\right)s \right\} - \frac{m \varepsilon^2}{8k^2\left(V'(2k\sqrt{\frac{V(M)}{\lambda}})\right)^2} \}

Take
\[ 2 \exp\left\{ -c_s\left(4k^2c\sqrt{\frac{V(M)}{\lambda}}\right)s \right\} - \frac{m \varepsilon^2}{8k^2\left(V'(2k\sqrt{\frac{V(M)}{\lambda}})\right)^2} = \delta. \]

Then
\[ c_s\left(4k^2c\sqrt{\frac{V(M)}{\lambda}}\right)s - \frac{m \varepsilon^2}{8k^2\left(V'(2k\sqrt{\frac{V(M)}{\lambda}})\right)^2} = \log \frac{\delta}{2}. \]

It follows
\[ \varepsilon^{2+s} - \varepsilon^s \times \frac{8k^2\left(V'(2k\sqrt{\frac{V(M)}{\lambda}})\right)^2}{m} \log \frac{2}{\delta} \]
\[ - \frac{8k^2c_s(4k^2c\sqrt{\frac{V(M)}{\lambda}})^s\left(V'(2k\sqrt{\frac{V(M)}{\lambda}})\right)^2}{m} = 0. \]

By Lemma 6.7 we have
\[ \varepsilon \leq \max \left\{ \frac{4kV'(2k\sqrt{\frac{V(M)}{\lambda}})}{\sqrt{\log 2}}, \right. \]
\[ \frac{(16k^2c_s)^{\frac{1}{2}}\pi^{\frac{1}{2}}\left(4k^2c\sqrt{\frac{V(M)}{\lambda}}\right)^{\frac{1}{2}}\pi^{\frac{1}{2}}\left(V'(2k\sqrt{\frac{V(M)}{\lambda}})\right)^{\frac{1}{2}}}{\sqrt{2^{s\sqrt{\log 2}} \sqrt{m}}}, \right\} \]
\[ (16k^2c_s)^{\frac{1}{2}}\pi^{\frac{1}{2}}\left(4k^2c\sqrt{\frac{V(M)}{\lambda}}\right)^{\frac{1}{2}}\pi^{\frac{1}{2}}\left(V'(2k\sqrt{\frac{V(M)}{\lambda}})\right)^{\frac{1}{2}} \sqrt{\log \frac{2}{\delta}} \}
\[ (6.42) \]

if (4.5) holds, where we have used the fact \(1 \leq \frac{\log 2}{\log 2}\). Therefore, with confidence \(1 - \delta\), holds
\[ E \leq \frac{(16k^2c_s)^{\frac{1}{2}}\pi^{\frac{1}{2}}\left(4k^2c\sqrt{\frac{V(M)}{\lambda}}\right)^{\frac{1}{2}}\pi^{\frac{1}{2}}\left(V'(2k\sqrt{\frac{V(M)}{\lambda}})\right)^{\frac{1}{2}}}{\sqrt{2^{s\sqrt{\log 2}} \sqrt{m}}} \sqrt{\log \frac{2}{\delta}}. \]
\[ (6.43) \]

We now bound (6.39). Since \(\xi(x, y, \cdot) = V'(y - f_{\rho,\lambda}^V(x))K(x, \cdot)\) satisfies
\[ \|\xi(x, y, \cdot)\|_K = \left|V'(y - f_{\rho,\lambda}^V(x))\right| \sqrt{K(x, x)} \leq kV'(2k\sqrt{\frac{D(f_{\rho,\lambda}^V, \lambda)}{\lambda}}), \]
we have by (6.27) that, with confidence \(1 - \delta\), holds
\[ G \leq \frac{4k}{\sqrt{m}}V'(2k\sqrt{\frac{D(f_{\rho,\lambda}^V, \lambda)}{\lambda}}) \log \frac{2}{\delta}. \]
\[ (6.44) \]
Combining (6.44) with (6.39) together with (6.43) and Lemma 6.8 we have
\[
\int_Z |V'(y - f^V_{\lambda}(x)) - V'(y - f^V_{\rho,\lambda}(x))|^2 \, d\rho \\
\leq \frac{32ke(16k^2c_s)(4k^2c)\pi^2}{\sqrt{\log 2} \lambda \frac{1}{2} \frac{1}{\pi^2} m ^{\frac{1}{2}+\gamma}} \\
\times \left( V'(2k\sqrt{\frac{V(M)}{\lambda}}) \right)^{\frac{2}{\gamma}} V'(2k\sqrt{\frac{D(f^V_{\lambda})}{\lambda}}) \left( \log \frac{4}{\delta} \right)^{\frac{1}{2}} + 8cD(f^V_{\lambda},\lambda).
\] (6.45)

(4.4) thus holds. \qed

We now show (4.6). Since \( \int_Z |V'(y - f^V_{\rho,\lambda}(x))|^2 \, d\rho \leq \sigma^2 \) we have by (6.26) that
\[
G \leq \frac{2k}{m} V'(2k\sqrt{\frac{D(f^V_{\lambda})}{\lambda}}) \log \frac{2}{\delta} + \sqrt{\frac{2\sigma^2 \log 2}{m}} \\
\leq \frac{2}{\sqrt{m}} \left( \int V'(2k\sqrt{\frac{D(f^V_{\lambda})}{\lambda}}) + \sigma \right) \log \frac{2}{\delta}.
\] (6.46)

Combining (6.46), (6.43) and (6.39) we have by Lemma 6.8 the inequality (4.6).

Proof of (4.7). In this case, we have \( V'(t) = t \) and \( f^V_{\rho} = f_{\rho} \) and
\[
\int_Z |V'(y - f_{\rho,\lambda}(x))|^2 \, d\rho = \int_Z (y - f_{\rho,\lambda}(x))^2 \, d\rho \\
\leq E_{\rho}(f_{\rho,\lambda}) + \lambda\|f_{\rho,\lambda}\|_K^2 = \inf_{f \in H_K} (E_{\rho}(f) + \lambda\|f\|_K^2) \\
\leq E_{\rho}(0) = \int_Z y^2 \, d\rho < +\infty.
\] (6.47)

Also,
\[
V'(2k\sqrt{\frac{V(M)}{\lambda}}) = 2k\sqrt{\frac{V(M)}{\lambda}}, V'(2k\sqrt{\frac{D(f_{\rho,\lambda})}{\lambda}}) = 2k\sqrt{\frac{D(f_{\rho,\lambda})}{\lambda}}.
\] (6.48)

Putting (6.47) and (6.48) into (4.6), we have (4.7). \qed

Proof of (4.11). In this case, we have \( V'_1(s) = \ln \frac{s + \sqrt{s^2 + 4}}{2} \). If \( \beta = 1 \), then
\[
V'_1(2k\sqrt{\frac{D(f_{\rho,\lambda})}{\lambda}}) = \ln \frac{2k^2O(1) + \sqrt{4kO(1) + 4}}{2} = O(1)
\] (6.49)

if \( 0 < \beta < 1 \) we have a constant \( c_\beta > 0 \) such that
\[
V'_1(2k\sqrt{\frac{D(f_{\rho,\lambda})}{\lambda}}) \leq \ln \frac{2k^2c_\beta\lambda^{\beta-1} + \sqrt{4k^2c_\beta\lambda^{\beta-1} + 4}}{2} \sim \ln \frac{1}{\lambda}.
\] (6.50)

where we have used the fact \( \ln \left( \frac{1}{\lambda} \right)^{\frac{\beta-2}{2}} \right) = \frac{\beta-2}{2} \ln \frac{1}{\lambda} \). And also
\[
V'_1(2k\sqrt{\frac{V_1(M)}{\lambda}}) = \ln \frac{2k^2V_1(M) + \sqrt{4k^2 \times V_1(M) + 4}}{2} \sim \ln \frac{1}{\lambda}.
\] (6.51)

Putting (6.49), (6.50) and (6.51) into (4.4), we have (4.11). \qed

In the same way we can show (4.14).
Proof of (4.17). In this case, we have \(|V'_3(t)| = \sqrt{\lambda t} \leq 1\). It follows that
\[
\int_Z |V'_3(y - f_{\rho, \lambda}^3(x))|^2 \rho \leq 1
\]  
(6.52)
and
\[
V'_3(2k \sqrt{\frac{V_3(M)}{\lambda}}) \leq 1, \quad V'_3(2k \sqrt{\frac{D(f_{\rho, \lambda}^3, \lambda)}{\lambda}}) \leq 1.
\]  
(6.53)
Putting (6.52) and (6.53) into (4.6), we have (4.17).

7. Appendix: some notations and auxiliary results. For the needs of proofs we give here some notations and auxiliary results.

7.1. Some properties of convex functions. To give the feature description for the optimal solution, we need the concept of Gâteaux derivative.

Let \((\mathcal{H}, \| \cdot \|_\mathcal{H})\) be a Hilbert space, \(F : \mathcal{H} \to \mathbb{R} \cup \{\pm \infty\}\) be a real function. We say \(F\) is Gâteaux differentiable at \(f \in \mathcal{H}\) if there is an \(\xi \in \mathcal{H}\) such that for any \(g \in \mathcal{H}\) there holds
\[
\lim_{t \to 0} \frac{F(f + tg) - F(f)}{t} = \langle g, \xi \rangle_\mathcal{H}
\]
and write \(\nabla F(f) = \xi\) as the Gâteaux derivative of \(F(f)\) at \(f\).

The following Proposition 7.1 can be found from Proposition 17.4, Proposition 17.10 and Proposition 17.12 in [7].

**Proposition 7.1.** Let \((\mathcal{H}, \| \cdot \|_\mathcal{H})\) be a Hilbert space and \(F : \mathcal{H} \to \mathbb{R} \cup \{\pm \infty\}\) be a function defined on \(\mathcal{H}\). Then

(i) if \(F\) is a convex function, then \(F\) attains minimal value at \(f_0\) if and only if \(\nabla F(f_0) = 0\).

(ii) if \(F : \mathcal{H} \to \mathbb{R} \cup \{\pm \infty\}\) is a Gâteaux differentiable function, then, \(F\) is a convex function on \(\mathcal{H}\) if and only if for any \(f, g \in \mathcal{H}\) there holds
\[
F(g) - F(f) \geq \left\langle g - f, \nabla F(f) \right\rangle_\mathcal{H}.
\]

(iii) For any \(a, b \in H\) there holds
\[
\|a\|_\mathcal{H}^2 - \|b\|_\mathcal{H}^2 = \langle a - b, 2b \rangle_\mathcal{H} + \|a - b\|_\mathcal{H}^2.
\]

7.2. Sobolev spaces and Besov spaces. Recall (see Definition 3.1 and 3.2 in the Chapter 3 of [1]) that for any integer \(k \geq 0, 1 \leq p \leq +\infty\) and a subset \(X \subset \mathbb{R}^d\) with nonempty interiors, the Sobolev space \(W^k_p(X)\) of order \(k\) is defined by
\[
W^k_p(X) := \{f \in L^p(X) : D^{(\alpha)}f \in L^p(X) \text{ exists for all } \alpha \in \mathbb{N}_0^q \text{ with } |\alpha| \leq k\}
\]
with the norm
\[
\|f\|_{W^k_p(X)} = \left\{ \begin{array}{ll}
\left( \sum_{|\alpha| \leq k} \|D^{(\alpha)}f\|_{L^p(X)}^p \right)^{\frac{1}{p}}, & 1 \leq p < +\infty,\\
\max_{|\alpha| \leq k} \|D^{(\alpha)}f\|_{C(X)}, & p = +\infty,
\end{array} \right.
\]
where \(D^{(\alpha)}\) is the \(\alpha\)-weak partial derivative for multi-index \(\alpha = (\alpha_1, \ldots, \alpha_q) \in \mathbb{N}_0^q\) of modulus \(|\alpha| = |\alpha_1| + \cdots + |\alpha_q|\).

For \(0 < p, \theta \leq +\infty, \alpha > 0, \gamma = |\alpha| + 1\), the Besov space \(B^\alpha_{p, \theta}(X)\) based on moduli of smoothness for \(X \subset \mathbb{R}^d\) is defined by
\[
B^\alpha_{p, \theta}(X) = \{f \in L^p(X) : |f|_{B^\alpha_{p, \theta}(X)} < +\infty\},
\]
where the semi-norm $|f|_{B^0_{\gamma,\sigma}(X)}$ is given by

$$|f|_{B^0_{\gamma,\sigma}(X)} = \left\{ \left( \int_0^{+\infty} \left( t^{-\frac{\gamma}{2}} \omega_{\gamma,L^p(X)}(f,t) \right)^{\theta} \frac{dt}{t} \right)^{\frac{1}{\theta}}, \quad 1 \leq \theta < +\infty, 1 \leq p < +\infty, \theta = +\infty, \right. \right.$$ 

and for $\gamma \in \mathbb{N}$, the $\gamma$-th modulus of smoothness of $f$ is defined by

$$\omega_{\gamma,L^p(X)}(f,t) = \sup_{\|h\|_2 \leq t} \|\Delta_h^\gamma(f,\cdot)\|_{L^p(X)},$$

where $\| \cdot \|_2$ denotes the norm in $\mathbb{R}^d$ and $\Delta_h^\gamma(f,\cdot)$ is the $\gamma$-th difference ([33]).

### 7.3. The modulus of smoothness associating with RKHSs.

Assume the Mercer kernel has the expansion form

$$K(x,y) = \sum_{l=0}^{+\infty} \lambda_l \phi_l(x) \phi_l(y), \quad x \in X, y \in X,$$

where $\{\phi_l(x)\}_{l=0}^{+\infty}$ is a normalized orthogonal basis in $L^2(\rho_X)$. Then for the $K$-functional

$$K(f, t)_{\mathcal{H}_K(\rho_X)} = \inf_{g \in \mathcal{H}_K(\rho_X)} (\|g - f\|_{L^2(\rho_X)} + t\|g\|_{\mathcal{H}_K(\rho_X)}),$$

we show in [58] an equivalent relation

$$K(f, t)_{\mathcal{H}_K(\rho_X)} \sim \omega_{\mathcal{H}_K(\mu)}(f, t)_{2,\mu}, \quad f \in L^2(\mu), \quad t > 0,$$

where

$$\omega_{\mathcal{H}_K(\rho_X)}(f,t)_{L^2(\mu)} = \| (T_K(t) - I)f \|_{L^2(\rho_X)}, \quad f \in L^2(\mu), \quad t > 0.$$

and

$$T_K(t)f = \sum_{k=0}^{+\infty} e^{-\frac{\sqrt{\gamma}}{\mu} t} a_k(f) \phi_k(x), \quad a_k(f) = \int_X f(y) \phi_k(y) d\rho_X(y), \quad t > 0.$$

Define a semi-norm $|f|_{B^2_{\mathcal{H}_K(\rho_X),\theta}(X)}$ by

$$|f|_{B^2_{\mathcal{H}_K(\rho_X),\theta}(X)} = \left\{ \left( \int_0^{+\infty} \left( t^{-\frac{\theta}{2}} \omega_{\mathcal{H}_K(\rho_X)}(f,t)_{L^2(\mu)} \right)^{\theta} \frac{dt}{t} \right)^{\frac{1}{\theta}}, \quad 1 \leq \theta < +\infty, \theta = +\infty, \right. \right.$$ 

and define a Besov space by

$$B^2_{\mathcal{H}_K(\rho_X),\theta}(X) = \{ f \in L^2(\rho_X) : |f|_{B^2_{\mathcal{H}_K(\rho_X),\theta}(X)} < +\infty \}.$$

Then $f^V_\rho$ satisfy (2.9) is equivalent to $f^V_\rho \in B^2_{\mathcal{H}_K(\rho_X),+\infty}(X)$.

**Acknowledgments.** The authors would like to take this opportunity to express their deeply grateful to the anonymous reviewers whose constructive suggestions and encouraging comments help the authors to improve the main results of this manuscript.
REFERENCES

[1] R. A. Adams and J. J. F. Fournier, *Sobolev Space*, 2nd edition, Academic Press, 2009.

[2] A. Argyriou, C. A. Micchelli and M. Pontil, On spectral learning, *J. Mach. Learn. Res.*, 11 (2010), 935–953.

[3] N. Aronszajn, Theory of reproducing kernels, *Trans. Amer. Math. Soc.*, 68 (1950), 337–404.

[4] A. Banerjee, S. Merugu, I. S. Dhillon and J. Ghosh, Clustering with Bregman divergences, *J. Mach. Learn. Res.*, 6 (2005), 1705–1749.

[5] P. L. Bartlett, M. I. Jordan and J. D. McAuliffe, Convex, classification, and risk bounds, *J. Amer. Statist. Assoc.*, 101 (2006), 138–156.

[6] F. Bauer, S. Pereverzey and L. Rosasco, On regularization algorithms in learning theory, *J. Complexity*, 23 (2007), 52–72.

[7] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer-Verlag, New York, 2010.

[8] A. Bietti and J. Mairal, Group invariance, stability to deformations, and complexity of deep convolutional representations, *J. Mach. Learn. Res.*, 20 (2019), 1–49.

[9] G. Blanchard and N. Krämer, Optimal learning rates for kernel conjugate gradient regression, preprint, arXiv:1009.5839.

[10] G. Blanchard and N. Müller, Optimal rates for regularized of statistical inverse learning problems, *Found. Math. Comput.*, 18 (2018), 971–1013.

[11] B. Bohn and C. Rieger, A representer theorem for deep kernel learning, *J. Mach. Learn. Res.*, 20 (2019), 1–32.

[12] J. F. Bonnans and A. Shapiro, *Perturbation Analysis of Optimization Problem*, Springer-Verlag, New York, 2000.

[13] E. M. Bronshtein, $\varepsilon$-entropy of convex sets and functions, *Siberian Math. J.*, 17 (1976), 393–398.

[14] A. Caponnetto and E. De Vito, Optimal rates for the regularized least-squares algorithm, *Found. Comput. Math.*, 7 (2007), 331–368.

[15] X. Y. Chang, Z. B. Xu, B. Zou and H. Zhang, Generalization bounds of regularization algorithms derived simultaneously through hypothesis space complexity, algorithmic stability and data quality, *Int. J. Wavelets Multiresolut. Inform. Process.*, 9 (2011), 549–570.

[16] D. R. Chen, Q. Wu, Y. M. Ying and D. X. Zhou, Support vector machine soft margin classifiers: error analysis, *J. Mach. Learn. Res.*, 5 (2004), 1143–1175.

[17] A. Christmann and A. V. Messem, Bouligand derivatives and robustness of support vector machines for regression, *J. Mach. Learn. Res.*, 9 (2008), 915–936.

[18] A. Christmann, A. V. Messem and I. Steinwart, On consistency and robustness properties of support vector machines for heavy-tailed distributions, *Statist. Interface*, 2 (2009), 311–327.

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

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

[21] A. Christmann and I. Steinwart, Consistency of kernel based quantile regression, *Appl. Stoch. Model. Bus. and Industr.*, 24 (2008), 171–183.

[22] A. Christmann, D. H. Xiang and D. X. Zhou, Total stability of kernel methods, *Neurocomputing*, 289 (2018), 101–118.

[23] A. Christmann and D. X. Zhou, Learning rates for the risk of kernel-based quantile regression estimators in additive models, *Anal. Appl.*, 14 (2016), 449–477.

[24] A. Christmann and D. X. Zhou, On the robustness of regularized pairwise learning methods based on kernels, *J. Complexity*, 37 (2016), 1–33.

[25] F. Cucker and S. Smale, On the mathematical foundations of learning theory, *Bull. Amer. Math. Soc.*, 39 (2001), 1–49.

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

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

[28] M. Debruyne, A. Christmann, M. Huber and J. A. K. Suykens, Robustness of reweighted least squares kernel based regression, *J. Multi. Anal.*, 101 (2010), 447–463.

[29] E. De Vito, A. Caponnetto and L. Rosasco, Model selection for regularized least-squares algorithm in learning theory, *Found. Comput. Math.*, 5 (2005), 59–85.
[30] E. De Vito, L. Rosasco, A. Caponnetto, M. Piana and A. Verri, Some properties of regularized kernel methods, *J. Mach. Learn. Res.*, 5 (2004), 1363–1390.
[31] E. De Vito, L. Rosasco and A. Toigo, Learning sets with separating kernels, *Appl. Comput. Harmon. Anal.*, 37 (2014), 185–217.
[32] F. Dumpert, Universal consistency and robustness of localized support vector machines, *Neurocomputing*, 315 (2018), 96–106.
[33] M. Ebets and I. Steinwart, Optimal regression rates for SVMs using Gaussian kernels, *Electron. J. Statist.*, 6 (2012), 2627–2668.
[34] T. Evgeniou, M. Pontil and T. Poggio, Regularization networks and support vector machines, *Adv. Comput. Math.*, 13 (2000), 1–50.
[35] M. Farooq and I. Steinwart, Learning rates for kernel-based expectile regression, *Mach. Learn.*, 108 (2019), 203–227.
[36] F. Dumpert, Universal consistency and robustness of localized support vector machines, *Neurocomputing*, 315 (2018), 96–106.
[37] M. Ebets and I. Steinwart, Optimal regression rates for SVMs using Gaussian kernels, *Electron. J. Statist.*, 6 (2012), 2627–2668.
[38] T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer-Verlag, Berlin, 2001.
[39] J. B. Hiriart-urruty and C. Lemaréchal, *Fundamentals of Convex Analysis*, Springer-Verlag, Berlin, 2001.
[40] Z. C. Guo, T. Hu and L. Shi, Gradient descent for robust kernel-based regression, *Inverse Problem*, 34 (2018), Art. 065009, 29 pp.
[41] R. Hable and A. Christmann, Qualitative robustness of support vector machines, *J. Multi. Anal.*, 102 (2011), 993–1007.
[42] T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer-Verlag, Berlin, 2001.
[43] J. B. Hiriart-urruty and C. Lemaréchal, *Fundamentals of Convex Analysis*, Springer-Verlag, Berlin, 2001.
[44] A. V. Messem and A. Christmann, A review on consistency and robustness properties of support vector machines for heavy-tailed distributions, *Adv. Data Anal. Classif.*, 4 (2010), 199–220.
[45] C. A. Micchelli and M. Pontil, Learning the kernel function via regularization, *J. Mach. Learn. Res.*, 6 (2005), 1099–1125.
[46] C. A. Micchelli, Y. S. Xu and H. Z. Zhang, Universal kernels, *J. Mach. Learn. Res.*, 7 (2006), 2651–2667.
[47] K. Muandet, K. Fukumizu, B. K. Sriperumbudur and B. Schölkopf, Kernel mean embedding of distributions: a review and beyonds, *Found. Trends Mach. Learn.*, 10 (2017), 1–14.
[48] Y. Nishiyama, Characteristic kernels and infinitely divisible distributions, *J. Mach. Learn. Res.*, 17 (2016), 1–28.
[49] J. Peypouquet, *Convex Optimization in Normed Spaces: Theory, Methods and Examples*, Springer-Verlag, 2015.
[50] Z. Qin, F. X. Yu, C. C. Liu and X. Chen, How convolutional neural networks see the world: a survey of convolution neural network visualization methods, *Math. Found. Comput.*, 1 (2018), 149–180.
[51] M. M. Rao and Z. D. Ren, *Applications of Orlicz Spaces*, Marcel Dekker, Inc., New York, 2002.
[52] L. Rosasco and M. Belkin, On learning with integral operators, *J. Mach. Learn. Res.*, 11 (2010), 905–934.
[53] Z. Sha and H. J. Ruan, *Fractal and Fitting (in Chinese)*, Zhejiang University Press, Hangzhou, 2005, 82–85.
[54] B. H. Sheng, On approximation by reproducing kernel spaces in weighted $L^p$–spaces, *J. Syst. Sci. Complex.*, 20 (2007), 623–638.
[55] B. H. Sheng, L. Q. Duan and P. X. Ye, Strong convex loss can increase the learning rates of online learning, *J. Comput.*, 9 (2014), 1606–1611.
[57] B. H. Sheng and H. T. Li, On approximation by spherical zonal translation networks based on Bochner Riesz means, *J. Syst. Sci. Complex.*, 18 (2005), 361–374.
[58] B. H. Sheng and J. L. Wang, On the K-functional in learning theory, *Anal. Appl.*, 18 (2020), 423–446.
[59] B. H. Sheng, J. L. Wang and D. H. Xiang, Error analysis on Hérmite learning with gradient data, *Chin. Ann. Math. Ser. B.*, 39 (2018), 705–720.
[60] B. H. Sheng, J. L. Wang and S. P. Zhou, A way of constructing spherical zonal translation network operators with linear bounded operators, *Taiwan. J. Math.*, 12 (2008), 77–92.
[61] B. H. Sheng and D. H. Xiang, The convergence rate for a K-functional in learning theory, *J. Inequal. Appl.*, (2010), Art. 249507, 18 pp.
[62] B. H. Sheng and D. H. Xiang, The learning rate of $l_2$-coefficient regularized classification with strong loss, *Acta Math. Sin. English Ser.*, 29 (2013), 2397–2408.
[63] B. H. Sheng, P. X. Ye and J. L. Wang, Learning rates for least square regressions with coefficient regularization, *Acta Math. Sin.*, 28 (2012), 2205–2212.
[64] S. Smale and D. X. Zhou, Estimating the approximation error in learning theory, *Anal. Appl.*, 1 (2003), 17–41.
[65] S. Smale and D. X. Zhou, Learning theory estimates via integral operators and their applications, *Constr. Approx.*, 26 (2007), 153–172.
[66] B. H. Sheng and H. C. Zhu, The convergence rate of semi-supervised regression with quadratic loss, *Appl. Math. Comput.*, 321 (2018), 11–24.
[67] B. K. Sriperumbudur, K. Fukumizu and G. R. G. Lanckriet, Universality, characteristic kernels and RKHS embedding of measures, *J. Mach. Learn. Res.*, 9 (2010), 773–780.
[68] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf and G. R. G. Lanckriet, Hilbert space embeddings and metrics on probability measures, *J. Mach. Learn. Res.*, 11 (2010), 1517–1561.
[69] I. Steinwart, Sparseness of support vector machines, *J. Mach. Learn. Res.*, 4 (2003), 1071–1105.
[70] I. Steinwart and A. Christmann, How SVMs can estimate quantiles and the median, *Adv. Neural Inform. Process. Syst.*, 20 (2008), 305–312.
[71] I. Steinwart and A. Christmann, Estimating conditional quantiles with the help of the pinball loss, *Bernoulli*, 17 (2011), 211–225.
[72] I. Steinwart and A. Christmann, *Support Vector Machines*, Springer-Verlag, New York, 2008.
[73] I. Steinwart, D. Hush and C. Scovel, Optimal rates for regularized least square regression, in *Proceedings of the 22nd Annual Conference on Learning Theory*, (2009), 79–93.
[74] H. W. Sun and Q. Wu, Application of integral operator for regularized least-square regression, *Math. Comput. Model.*, 49 (2009), 276–285.
[75] H. Z. Tong, D. R. Chen and L. Z. Peng, Analysis of support vector machines, *Found. Comput. Math.*, 9 (2009), 243–257.
[76] H. Z. Tong, D. R. Chen and F. H. Yang, Fast learning rates for regularized regression algorithms (in Chinese), *Sci. Sin. Math.*, 42 (2012), 1251–1262.
[77] B. H. Sheng, Convergence rate of SVM for kernel-based robust regression, *Int. J. Wavelets Multiresolut. Inform. Process.*, 17 (2019), Art. 1950004, 21 pp.
[78] C. Wang and J. Cai, Convergence analysis of coefficient-based regularization under moment incremental condition, *Int. J. Wavelets Multiresolut. Inform. Process.*, 12 (2014), Art. 1450008, 19 pp.
[79] B. H. Sheng, Convergence rate of SVM for kernel-based robust regression, *Int. J. Wavelets Multiresolut. Inform. Process.*, 17 (2019), Art. 1950004, 21 pp.
[80] C. Wang and Z. C. Guo, ERM learning with unbounded sampling, *Acta Math. Sin. English Ser.*, 28 (2012), 97–104.
[81] X. Z. Wang, Y. X. Zhao and F. Pourpanah, Recent advances in deep learning, *Int. J. Math. Learn. Cyber.*, 11 (2020), 747–750.
[82] C. Wang and D. X. Zhou, Optimal learning rates for least squares regularized regression with unbounded sampling, *J. Complexity*, 27 (2011), 55–67.
[83] Q. Wu, Y. M. Ying and D. X. Zhou, Learning rates of least-square regularized regression, *Found. Comput. Math.*, 6 (2006), 171–192.
[84] D. H. Xiang, Conditional quantiles with varying Gaussians, *Adv. Comput. Math.*, 38 (2013), 723–735.
[85] D. H. Xiang, ERM scheme for quantile regression, *Abstr. Appl. Anal.*, (2013), Art. 148490, 6 pp.
[86] T. Zhang, Statistical behavior and consistency of classification methods based on convex risk minimization, *Ann. Statist.*, **32** (2004), 56–134.

[87] D. X. Zhou, Capacity of reproducing kernel spaces in learning theory, *IEEE Trans. Inform. Theory*, **49** (2003), 1743–1752.

[88] D. X. Zhou, Deep distributed convolutional neural networks: universality, *Anal. Appl.*, **16** (2018), 895–919.

[89] D. X. Zhou, Universality of deep convolutional neural networks, *Appl. Comput. Harmonic Anal.*, **48** (2020), 787–794.

[90] D. X. Zhou and K. Jetter, Approximation with polynomial kernels and SVM, *Adv. Comput. Math.*, **25** (2006), 323–344.

Received June 2019; revised March 2020.

E-mail address: bhsheng@usx.edu.cn
E-mail address: lhxiang@usx.edu.cn
E-mail address: glanne@usx.edu.cn