Journal of Applied Mathematics and Physics, 2016, 4, 156-168
Published Online January 2016 in SciRes. http://www.scirp.org/journal/jamp
http://dx.doi.org/10.4236/jamp.2016.41019

Error Analysis of ERM Algorithm with Unbounded and Non-Identical Sampling*

Weilin Nie, Cheng Wang#

Department of Mathematics, Huizhou University, Huizhou, China
Email: niewl@hzu.edu.cn, #wangch@hzu.edu.cn

Received 9 November 2015; accepted 24 January 2016; published 27 January 2016

Abstract

A standard assumption in the literature of learning theory is the samples which are drawn independently from an identical distribution with a uniform bounded output. This excludes the common case with Gaussian distribution. In this paper we extend these assumptions to a general case. To be precise, samples are drawn from a sequence of unbounded and non-identical probability distributions. By drift error analysis and Bennett inequality for the unbounded random variables, we derive a satisfactory learning rate for the ERM algorithm.

Keywords

Learning Theory, ERM, Non-Identical, Unbounded Sampling, Covering Number

1. Introduction

In learning theory we study the problem of looking for a function or its approximation which reflects the relationship between the input and the output via samples. It can be considered as a mathematical analysis of artificial intelligence or machine learning. Since the exact distributions of the samples are usually unknown, we can only construct algorithms based on an empirical sample set. A typical setting of learning theory in mathematics can be like this: the input space X is a compact metric space, and the output space $Y \subset \mathbb{R}$ for regression. (When $Y = \{+1, -1\}$, it can be regarded as a binary classification problem.) Then $Z: = X \times Y$ is the whole sample space. We assume a distribution ρ on Z, which can be decomposed to two parts: marginal distribution ρ_X on X and conditional distribution $\rho(y|x)$ given some $x \in X$. This implies

*This work is supported by NSF of China (Grant No. 11326096, 11401247), Foundation for Distinguished Young Talents in Higher Education of Guangdong, China (No. 2013LYM 0089), Doctor Grants of Huizhou University (Grant No. C511.0206) and NSF of Guangdong Province in China (No. 2015A030313674).

#Corresponding author.

How to cite this paper: Nie, W.L. and Wang, C. (2016) Error Analysis of ERM Algorithm with Unbounded and Non-Identical Sampling. *Journal of Applied Mathematics and Physics*, **4**, 156-168. http://dx.doi.org/10.4236/jamp.2016.41019

$$\int_Z g(z)\mathrm{d}\rho = \int_X\int_Y g(x,y)\mathrm{d}\rho(y\mid x)\mathrm{d}\rho_X$$

for any integrable function $g(z)$ [1].

To evaluate the efficiency of a function $f: X \to Y$ we can choose the generalization error:

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

Here $\phi(f(x),y)$ is a loss function which measures the difference between the prediction $f(x)$ via *f* and the actual output *y*. It can be hinge loss in SVM (support vector machine) or pinball loss in quantile learning and etc.. In this paper we focus on the classical least square loss $\phi(f(x),y) = (f(x)-y)^2$ for simplicity. [2] shows that

$$\mathcal{E}(f) = \mathcal{E}(f_\rho) + \int_X (f(x)-f_\rho(x))^2 \mathrm{d}\rho_X. \tag{1}$$

From this we can see the regression function

$$f_\rho(x) = \int_Y y\mathrm{d}\rho(y\mid x)$$

is our goal minimizing the generalization error. The empirical risk minimization (ERM) algorithm aims to find a function which approximates the goal function f_ρ well. While ρ is always unknown beforehand, a sample set $\mathbf{z} = \{z_i\}_{i=1}^m = \{(x_i,y_i)\}_{i=1}^m \in Z^m$ is accessible. Then ERM algorithm can be described as

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

where function space $\mathcal{H}$ is the hypothesis space which will be chosen to be a compact subset of $C(X)$.

Then the error produced by ERM algorithm is $\mathcal{E}(f_{\mathbf{z}})$. We expect it is close to the optimal one $\mathcal{E}(f_\rho)$, which means the excess generalization error $\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho)$ should be small, while the sample size *m* tends to infinity.

Dependent sampling has considered in some literature such as [3] for concentration inequality and [4] [5] for learning. More recently, in [6] and [7], the authors studied learning with non-identical sampling and dependent sampling, and obtained satisfactory learning rates.

In this paper we concentrate on the non-identical setting that each sample z_i is drawn according to a different distribution $\rho^{(i)}$ on *Z*. And each $\rho^{(i)}$ can also be decomposed to marginal distribution $\rho_X^{(i)}$ and conditional distribution $\rho^{(i)}(y\mid x)$. Assume they are elements of $(C^s(X))^*$ and $(C^s(Y))^*$ respectively, where $C^s(X)$ and $C^s(Y)$ are Hölder spaces with $0 \le s \le 1$. Hölder spaces $C^s(X)$ is the set of continuous functions with finite norm

$$\|f\|_C^s(X) = \|f\|_C(X) + |f|_C^s(X),$$

where

$$|f|_C^s(X) = \sup_{x\neq x'} \frac{|f(x)-f(x')|}{|x-x'|^s}.$$

We assume a polynomial convergence condition for both sequences $\{\rho_X^{(i)}\}_{i\ge1}$ and $\{\rho^{(i)}(y\mid x)\}_{i\ge1}$, *i.e.*, there exist $b>0, C_b>0, \rho_X \in C^s(X)$ and $\rho(y\mid x) \in C^s(Y)$, such that

$$\left|\int_X f(x)\mathrm{d}\rho_X^{(i)} - \int_X f(x)\mathrm{d}\rho_X\right| \le C_b i^{-b}\|f\|_C^s(X), \quad \forall f\in C^s(X), i\in\mathbb{N}. \tag{2}$$

$$\left|\int_Y g(y)\mathrm{d}\rho^{(i)}(y\mid x) - \int_Y g(y)\mathrm{d}\rho(y\mid x)\right| \le C_b i^{-b}\|g\|_C^s(Y), \quad \forall g\in C^s(Y), i\in\mathbb{N}. \tag{3}$$

Power index b measures quantitatively differences between the non-identical setting and the i.i.d. case. The distributions are more similar as b is larger, and when $b=\infty$ it is indeed i.i.d. sampling, *i.e.* $\rho_X^{(i)}=\rho_X$ and $\rho^{(i)}(y|x)=\rho(y|x)$ for any $i\in\mathbb{N}$. The following example is taken from [8].

Example 1. *Let* $\{h^{(i)}\}$ *be a sequence of bounded functions on X such that* $\sup_{x\in X}\left|h^{(i)}(x)\right|\le C_i^{-b}$. *Then the sequence* $\{\rho_X^{(i)}\}_{i=1,2,\cdots}$ *defined by* $\mathrm{d}\rho_X^{(i)}=\mathrm{d}\rho_X+h^{(i)}(x)\mathrm{d}\rho_X$ *satisfies* (2) *for any* $0\le s\le 1$.

On the other hand, most literature assume the output space is uniformly bounded, that is, $|y|\le M$ for some positive constant M and almost surely with respect to ρ. A typical kernel dependent result for the least-squares regularization algorithm under this assumption is [9]. There the authors get a learning rate close to 1 under some capacity condition for the hypothesis space. However, the most common distribution-Gaussian distribution is not bounded. This requirement is from the bounded condition in Bernstein inequality and limits the application of algorithms. In [10]-[13], some unbounded conditions for the output space are discussed in different forms, which extends the classical bounded condition. Here we will follow the latter one which is more generalized and simple in expression, and this is the second novelty of this paper. We assume the moment incremental condition for the output space, an extension of that we proposed in [11]:

$$\mathbb{E}\left(|y|^\ell \mid x\right)=\int_Y |y|^\ell \,\mathrm{d}\rho(y\mid x)\le C\ell!M^\ell,\quad \forall 2\le\ell\in\mathbb{N},\tag{4}$$

and

$$\mathbb{E}^{(i)}\left(|y|^\ell \mid x\right)=\int_Y |y|^\ell \,\mathrm{d}\rho^{(i)}(y\mid x)\le C\ell!M^\ell,\quad \forall i\in\mathbb{N},2\le\ell\in\mathbb{N}.\tag{5}$$

We can see the Gaussian distribution satisfies this setting.

Example 2. *Let* $B>0$ *and* $B_0>0$. *If for each* $x\in X,\left|f_\rho(x)\right|\le B$ *and the condition distribution* $\rho(\cdot|x)$ *is a normal distribution with variance* σ_x^2 *bounded by* B_0, *then* (4) *is satisfied with* $M=\max\left\{\sqrt{2}B_0,B\right\}$ *and* $C=4$.

Next we need to introduce the covering number and interpolation space.

Definition 1. *The covering number* $\mathcal{N}(\mathcal{F},\eta)$ *for a subset* $\mathcal{F}$ *of* $C(X)$ *and* $\eta>0$ *is defined to be the minimal integer N such that there exist N balls with radius* η *covering* $\mathcal{F}$.

Let the hypothesis space $H\subseteq C(X)$, be a compact Banach space with inclusion $I:H\to C(X)$ bounded and compact. We follow the assumption [14] [15] that there exist some constants $r>0$ and $C_r>0$, such that the hypothesis space satisfies the capacity condition

$$\log\mathcal{N}(B_1,\eta)\le C_r\eta^{-r},\quad \forall\eta>0,\tag{6}$$

where $B_1=\left\{f\in H:\|f\|_H\le 1\right\}$. Capacity condition describes the amount of functions in the hypothesis space. The sample error will decrease but approximation error will increase when covering number of H is larger (or simply say H is larger). So how to choose an appropriate hypothesis space is the key problem of ERM algorithm. We will demonstrate this in our main theorem.

Definition 2. *The interpolation space* $\left(L^2_{\rho_X},H\right)_{\theta,\infty}$ *is a function space consists of* $f\in L^2_{\rho_X}$ *with norm*

$$\|f\|_{\theta,\infty}=\sup_{t>0}\frac{K(f,t)}{t^\theta}<+\infty,$$

where $K(f,t)$ is the K-functional defined as

$$K(f,t)=\inf_{g\in H}\left\{\|g-f\|_{L^2_{\rho_X}}+t\|g\|_H\right\},\quad t>0.$$

Interpolation space is used to characterize the position of the regression function, and it is related with the approximation error. Now we can state our main result as follow.

Theorem 1. *If* $H\subset C(X)$ *with bounded inclusion* $I:H\to C(X)$, *and satisfies* (6) *with* r, $C_r>0$,

$f_\rho \in \left(L^2_{\rho_X}, H\right)_{\theta,\infty}$ *for some* $0<\theta<1$, *the sample distribution satisfies* (2), (3) *for some* $b>0, C_b>0$ *and* $0<s<1$, (4) *and* (5). *For any* $1 \le p \in \mathbb{N}$, *choose the hypothesis space* $\mathcal{H}$ *to be the ball of H centered at* 0 *with radius* $R=\left(\gamma(m)\right)^{-1/\left(r+\frac{2}{1-\theta}\right)}$, *where* $\gamma(m)=\max\left\{\left(v_b(m)\right)^{(p-1)/p}, m^{-1/(1+r)}\right\}$ *and*

$$v_b(m)=\begin{cases} m^{-b}, & 0<b<1 \\ \dfrac{\log m}{m}, & b=1. \\ m^{-1}, & b>1 \end{cases}$$

Moreover, we assume all functions in *H* and f_ρ are Hölder continuous of order *s*, *i.e.*, there is a constant $C_s>0$, such that

$$\frac{\left|f(x)-f(y)\right|}{\left|x-y\right|^s} \le C_s, \quad \forall f \in H \cup \left\{f_\rho\right\}, x \ne y \in X. \tag{7}$$

Then for any $0<\delta<1$, with confidence at least $1-\delta$, we have

$$\mathcal{E}\left(f_{\mathbf{z}}\right)-\mathcal{E}\left(f_\rho\right) \le \tilde{C}\left(\gamma(m)\right)^{\frac{2\theta}{r-r\theta+2}} \log\frac{3}{\delta}.$$

Here $\tilde{C}$ is a constant independent with *m* and δ.

Remark 1. *In* [6], *the authors pointed out that if we choose the hypothesis space to be the reproducing kernel Hilbert space* (*RKHS*) $\mathcal{H}_K$ *on* $X \subset \mathbb{R}^n$, *and the kernel* $K \in C^2\left(X \times X\right)$, *then our assumption* (7) *will hold true. In particular, if the kernel is chosen to be Gaussian kernel* K_σ, *then* (7) *holds for any* $0<s\le 1$. [16] *discussed this in detail.*

In all, we extend the polynomial convergence condition on the conditional distribution sequense and accordingly, set the moment inremental condition for the sequence in the least squares ERM algorithm. By error decomposition, truncate technique and unbounded concentration inequality, we can finally obtain the total error bound Theorem 1.

Compared with the non-identical settings in [6] and [17], our setting is more general since the conditional distribution sequence $\left\{\rho^{(i)}\left(y \mid x\right)\right\}_{i\ge 1}$ is also a polynomially convergence sequence, but not identical as in their settings. This together with unbounded y lead to the main difficulty for the error analysis in this paper.

For the classical i.i.d. and bounded conditions, [9] indicates that $X \subset R^n$ and kernel $K \in C^\infty$ while $f_\rho \in \mathcal{H}_K$, the rate of least square regularization algorithm is $O_p\left(\left(1/m\right)^{1-\epsilon}\right)$ for any $\epsilon>0$. [17] shows that under some conditions on kernel, object function f_ρ, exponential convergence condition for distribution sequence and choose some special parameters, the optimal rate of online learning algorithm is close to $O_p\left(\left(1/m\right)^{1/4}\right)$. In [6], the best case occurs when $X \subset R^n$ and kernel $K \in C^\infty$. The rate of least square regularization algorithm can be close to $O_p\left(\left(1/m\right)^{2/3}\right)$. However, our result implicates that while $b>1$, θ tends to 1 and r, s tends to 0, since *p* can be any integer, the learning rate can be arbitrarily close to $O_p\left(1/m\right)$, which is the same as in i.i.d. case [9], and better than the former results with non-identical settings. With this result, we can extend the application of learning algorithm to more situations and still keep the best learning rate. The explicit expression of $\tilde{C}$ in the theorem can be found through the proof of the theorem below.

2. Error Decomposition

Our aim, the error $\mathcal{E}\left(f_{\mathbf{z}}\right)-\mathcal{E}\left(f_\rho\right)$ is hard to bound directly, we need a transitional function for analyzing. By the compactness of $\mathcal{H}$ and continuity of functional $\mathcal{E}$, we can denote

$$f_{\mathcal{H}} = \arg\min_{f\in\mathcal{H}} \mathcal{E}(f).$$

Then the generalization error can be written as

$$\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}(f_\rho)=\left\{\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}(f_{\mathcal{H}})\right\}+\left\{\mathcal{E}(f_{\mathcal{H}})-\mathcal{E}(f_\rho)\right\}.$$

The first term on the right hand side is the sample error, and the second term $A_{\mathcal{H}}=\mathcal{E}(f_{\mathcal{H}})-\mathcal{E}(f_\rho)$ is called approximation error which is independent with samples. [18] analyzed the approximation error by approximation theory. In the following we mainly study the sample error bound.

Now we break the sample error to some parts which can be bounded using truncate technique and unbounded concentration inequality. We refer the error decomposition $\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}(f_{\mathcal{H}})$ to [6]. Denote

$$\mathcal{E}_{\mathbf{z}}(f)=\frac{1}{m}\sum_{i=1}^{m}\left(f(x_i)-y_i\right)^2$$

$$\mathcal{E}_m(f)=\frac{1}{m}\sum_{i=1}^{m}\int_Z\left(f(x)-y\right)^2\mathrm{d}\rho^{(i)},$$

then $f_{\mathbf{z}}=\arg\min_{f\in\mathcal{H}}\mathcal{E}_{\mathbf{z}}(f)$ and we have

$$\begin{aligned}
&\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}(f_{\mathcal{H}})\\
&=\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})+\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})-\mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}})+\mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}})-\mathcal{E}(f_{\mathcal{H}})\le\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})+\mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}})-\mathcal{E}(f_{\mathcal{H}})\\
&=\left(\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}_m(f_{\mathbf{z}})\right)-\left(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})-\mathcal{E}_m(f_{\mathbf{z}})\right)+\left(\mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}})-\mathcal{E}_m(f_{\mathcal{H}})\right)-\left(\mathcal{E}(f_{\mathcal{H}})-\mathcal{E}_m(f_{\mathcal{H}})\right)\\
&=\left[\left(\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}(f_\rho)\right)-\left(\mathcal{E}_m(f_{\mathbf{z}})-\mathcal{E}_m(f_\rho)\right)\right]+\left[\left(\mathcal{E}_m(f_{\mathbf{z}})-\mathcal{E}_m(f_\rho)\right)-\left(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})-\mathcal{E}_{\mathbf{z}}(f_\rho)\right)\right]\\
&\quad+\left[\left(\mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}})-\mathcal{E}_{\mathbf{z}}(f_\rho)\right)-\left(\mathcal{E}_m(f_{\mathcal{H}})-\mathcal{E}_m(f_\rho)\right)\right]+\left[\left(\mathcal{E}_m(f_{\mathcal{H}})-\mathcal{E}_m(f_\rho)\right)-\left(\mathcal{E}(f_{\mathcal{H}})-\mathcal{E}(f_\rho)\right)\right].
\end{aligned}$$

In the following, we call the first and fourth brackets drift errors, and the left sample errors. We will bound the two types of errors respectively in the following sections, and finally obtain the total error bounds.

3. Drift Errors

Firstly we consider the drift error involving $f_{\mathcal{H}}$ in this section.To avoid handling two polynomial convergence sequences simultaneously, we break the drift errors to two parts. Meanwhile, a truncate technique is used to deal with the unbounded assumption. Since $\mathcal{H}$ is a subset of $C(X)$, functions in $\mathcal{H}$ is uniformly bounded. Then we have

Proposition 1. *Assume* $\|f_{\mathcal{H}}\|_C(X)\le B$ *for some* $B>0$, *for any* $1\le p\in\mathbb{N}$,

$$\begin{aligned}
\left(\mathcal{E}_m(f_{\mathcal{H}})-\mathcal{E}_m(f_\rho)\right)-\left(\mathcal{E}(f_{\mathcal{H}})-\mathcal{E}(f_\rho)\right)\le\left(\frac{w_b(m)}{m}\right)^{\frac{p-1}{p}}&\Big[(C_b+1)B^2+\left(2CM+3pC_bM+4\sqrt{2C}+4C_s\right)B\\
&+\left(4C^2+6CC_bp+8C\right)M^2+6\sqrt{2C}C_sM\Big].
\end{aligned}$$

Proof. From the definition of $\mathcal{E}_m$ and $\mathcal{E}$, we know that

$$\begin{aligned}
&\left(\mathcal{E}_m(f_{\mathcal{H}})-\mathcal{E}_m(f_\rho)\right)-\left(\mathcal{E}(f_{\mathcal{H}})-\mathcal{E}(f_\rho)\right)=\frac{1}{m}\sum_{i=1}^{m}\int_Z\left(f_{\mathcal{H}}^2(X)-f_\rho^2(X)-2\left(f_{\mathcal{H}}(X)-f_\rho(X)\right)y\right)\mathrm{d}\left(\rho^{(i)}-\rho\right)\\
&=\frac{1}{m}\sum_{i=1}^{m}\Big[\int_X\int_Y\left(f_{\mathcal{H}}^2(X)-f_\rho^2(X)-2\left(f_{\mathcal{H}}(X)-f_\rho(X)\right)y\right)\mathrm{d}\left(\rho^{(i)}(y\mid x)-\rho(y\mid x)\right)\mathrm{d}\rho_X^{(i)}\\
&\quad+\int_X\int_Y\left(f_{\mathcal{H}}^2(X)-f_\rho^2(X)-2\left(f_{\mathcal{H}}(X)-f_\rho(X)\right)y\right)\mathrm{d}\rho(y\mid x)\mathrm{d}\left(\rho_X^{(i)}-\rho_X\right)\Big].
\end{aligned}$$

Since $f_{\mathcal{H}}^2(X)\le B^2, f_\rho^2(X)=\left(\int_Y y\mathrm{d}\rho(y\mid x)\right)^2\le\int_Y y^2\mathrm{d}\rho(y\mid x)\le 2CM^2$, we can bound the first term inside the bracket as follow.

$$\begin{aligned}&\int_X\int_Y\left(f_{\mathcal{H}}^2(X)-f_\rho^2(X)-2\left(f_{\mathcal{H}}(X)-f_\rho(X)\right)y\right)\mathrm{d}\left(\rho^{(i)}(y\mid x)-\rho(y\mid x)\right)\mathrm{d}\rho_X^{(i)}\\ &\le\int_X\left[\left|f_{\mathcal{H}}^2(X)\int_Y\mathrm{d}\left(\rho^{(i)}(y\mid x)-\rho(y\mid x)\right)\right|+\left|f_\rho^2(X)\int_Y\mathrm{d}\left(\rho^{(i)}(y\mid x)-\rho(y\mid x)\right)\right|\right.\\ &\quad\left.+2\left|f_{\mathcal{H}}(X)-f_\rho(X)\right|\cdot\left|\int_Y y\mathrm{d}\left(\rho^{(i)}(y\mid x)-\rho(y\mid x)\right)\right|\right]\mathrm{d}\rho_X^{(i)}\\ &\le\left(B^2+2CM^2\right)C_b i^{-b}+2(B+CM)\int_X\left|\int_Y y\mathrm{d}\left(\rho^{(i)}(y\mid x)-\rho(y\mid x)\right)\right|\mathrm{d}\rho_X^{(i)}.\end{aligned}$$

But for any $K\ge 1$ and $1\le p\in\mathbb{N}$, there holds

$$\begin{aligned}\left|\int_Y y\mathrm{d}\left(\rho^{(i)}(y\mid x)-\rho(y\mid x)\right)\right|&\le\left|\int_{|y|>K}y\mathrm{d}\left(\rho^{(i)}(y\mid x)-\rho(y\mid x)\right)\right|+\left|\int_{|y|\le K}y\mathrm{d}\left(\rho^{(i)}(y\mid x)-\rho(y\mid x)\right)\right|\\ &\le\frac{1}{K^{p-1}}\int_{|y|>K}|y|^p\,\mathrm{d}\left(\rho^{(i)}(y\mid x)-\rho(y\mid x)\right)+\left\|y|_{|y|\le K}\right\|_{C^s(Y)}C_b i^{-b}\\ &\le\frac{2C_p!M^p}{K^{p-1}}+\left(K+\sup_{y\ne y'}\frac{|y-y'|}{|y-y'|^s}\right)C_b i^{-b}\le\frac{2C_p!M^p}{K^{p-1}}+3KC_b i^{-b}.\end{aligned}$$

From (3.12) in [6], we have

$$\sum_{i=1}^m i^{-b}\le w_b(m)=\begin{cases}\frac{1}{1-b}m^{1-b}, & 0<b<1,\\ 1+\log m, & b=1,\\ \frac{b}{b-1}, & b>1.\end{cases}$$

Then we can bound the sum of the first term as

$$\begin{aligned}&\frac{1}{m}\sum_{i=1}^m\int_X\int_Y\left(f_{\mathcal{H}}^2(X)-f_\rho^2(X)-2\left(f_{\mathcal{H}}(X)-f_\rho(X)\right)y\right)\mathrm{d}\left(\rho^{(i)}(y\mid x)-\rho(y\mid x)\right)\mathrm{d}\rho_X^{(i)}\\ &\le\left(B^2+2CM^2\right)C_b\frac{w_b(m)}{m}+2(B+CM)\left(\frac{2C_p!M^p}{K^{p-1}}+3KC_b\frac{w_b(m)}{m}\right).\end{aligned}$$

Choose K to be $\left(m/w_b(m)\right)^{1/p}pM$, we have

$$\begin{aligned}&\frac{1}{m}\sum_{i=1}^m\int_X\int_Y\left(f_{\mathcal{H}}^2(X)-f_\rho^2(X)-2\left(f_{\mathcal{H}}(X)-f_\rho(X)\right)y\right)\mathrm{d}\left(\rho^{(i)}(y\mid x)-\rho(y\mid x)\right)\mathrm{d}\rho_X^{(i)}\\ &\le\left(\frac{w_b(m)}{m}\right)^{\frac{p-1}{p}}\left[C_bB^2+\left(2C+3pC_p\right)MB+\left(4C^2+6CC_bp\right)M^2\right].\end{aligned}$$

For the second term, notice $\|fg\|_{C^s(X)}\le\|f\|_{C(X)}\|g\|_{C^s(X)}+\|f\|_{C^s(X)}\|g\|_{C(X)}$, and $\left\|f_\rho\right\|_{C^s(X)}\le\sqrt{2C}M+C_s$ so

$$\begin{aligned}&\int_X\int_Y\left(f_{\mathcal{H}}^2(X)-f_\rho^2(X)-2\left(f_{\mathcal{H}}(X)-f_\rho(X)\right)y\right)\mathrm{d}\rho(y\mid x)\mathrm{d}\left(\rho_X^{(i)}-\rho_X\right)\\ &\le\left|\int_X f_{\mathcal{H}}^2(X)\mathrm{d}\left(\rho_X^{(i)}-\rho_X\right)\right|+\left|\int_X f_\rho^2(X)\mathrm{d}\left(\rho_X^{(i)}-\rho_X\right)\right|+2\left|\int_X\left(f_{\mathcal{H}}(X)-f_\rho(X)\right)\int_Y y\mathrm{d}\rho(y\mid x)\mathrm{d}\left(\rho_X^{(i)}-\rho_X\right)\right|\\ &\le\left(B^2+\sup_{x\ne x'}\frac{\left|f_{\mathcal{H}}^2(X)-f_{\mathcal{H}}^2(x')\right|}{|x-x'|^s}\right)C_b i^{-b}+2\left\|f_\rho\right\|_{C^s(X)}\left\|f_\rho\right\|_{C(X)}C_b i^{-b}+2\left\|\left(f_{\mathcal{H}}(X)-f_\rho(X)\right)f_\rho(X)\right\|_{C^s(X)}C_b i^{-b}\\ &\le\left[B^2+2BC_s+2\left(\left\|f_{\mathcal{H}}-f_\rho\right\|_{C^s(X)}\left\|f_\rho\right\|_{C(X)}+\left\|f_{\mathcal{H}}-f_\rho\right\|_{C(X)}\left\|f_\rho\right\|_{C^s(X)}\right)\right]C_b i^{-b}\\ &\le\left[B^2+4\left(\sqrt{2C}+C_s\right)B+8CM^2+6\sqrt{2C}C_sM\right]C_b i^{-b}.\end{aligned}$$

Therefore

$$\frac{1}{m}\sum_{i=1}^{m}\int_X\int_Y\left(f_{\mathcal{H}}^2(X)-f_\rho^2(X)-2\left(f_{\mathcal{H}}(X)-f_\rho(X)\right)y\right)\mathrm{d}\rho(y\,|\,x)\,\mathrm{d}\left(\rho_X^{(i)}-\rho_X\right)$$
$$\le\frac{w_b(m)}{m}\left[B^2+4\left(\sqrt{2C}+C_s\right)B+8CM^2+6\sqrt{2C}C_sM\right].$$

Combining the two bounds, we have

$$\left(\mathcal{E}_m(f_{\mathcal{H}})-\mathcal{E}_m(f_\rho)\right)-\left(\mathcal{E}(f_{\mathcal{H}})-\mathcal{E}(f_\rho)\right)$$
$$\le\frac{w_b(m)}{m}^{\frac{p-1}{p}}\Big[(C_b+1)B^2+\left(2CM+3pC_bM+4\sqrt{2C}+4C_s\right)B$$
$$+\left(4C^2+6CC_bp+8C\right)M^2+6\sqrt{2C}C_sM\Big].$$

And this is indeed the proposition.

For the drift error involving $f_{\mathbf{z}}$, we have the same result since $f_{\mathbf{z}}\in\mathcal{H}$ as well, *i.e.*,

Proposition 2. *Assume* $\|f_{\mathbf{z}}\|_{C(X)}\le B$ *for some* $B>0$, *for any* $1\le p\in\mathbb{N}$, *we have*

$$\left(\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}(f_\rho)\right)-\left(\mathcal{E}_m(f_{\mathbf{z}})-\mathcal{E}_m(f_\rho)\right)$$
$$\le\left(\frac{w_b(m)}{m}\right)^{\frac{p-1}{p}}\Big[(C_b+1)B^2+\left(2CM+3pC_bM+4\sqrt{2C}+4C_s\right)B$$
$$+\left(4C^2+6CC_bp+8C\right)M^2+6\sqrt{2C}C_sM\Big].$$

4. Sample Error Estimate

We devote this section to the analysis of the sample errors. For the sample error term involving $f_{\mathcal{H}}$, we will use the Bennett inequality as in [11] and [19], which is initially introduced in [20]. Since two polynomial convergence conditions are posed on the marginal and conditional distribution sequences, we have to modify the Bennett inequality to fit our setting. Denote $\mathbb{E}g=\int_Z g(z)\,\mathrm{d}\rho$ and $\mathbb{E}_{\mathbf{z}}g=\frac{1}{m}\sum i=1^m g(z)$ for an integrable function *g*, the lemma can be stated as follow.

Lemma 1. *Assume* $\mathbb{E}|g-\mathbb{E}g|^\ell\le\frac{1}{2}\ell!M^{\ell-2}v$ *holds for* $\ell=2,3,\cdots$ *and some constants* $M,v>0$, *then we have*

$$\mathrm{Prob}_{\mathbf{z}\in Z^m}\left\{\mathbb{E}_{\mathbf{z}}g-\mathbb{E}g\ge\varepsilon\right\}\le\exp\left\{-\frac{m\varepsilon^2}{2(v+M\varepsilon)}\right\}.$$

For our non-identical setting, we can have a similar result from the same idea of proof. By denoting $\mathbb{E}^{(i)}g=\int_Z g(z)\,\mathrm{d}\rho^{(i)}$ and $\mathbb{E}_m g=\frac{1}{m}\sum i=1^m\int_Z g(z)\,\mathrm{d}\rho^{(i)}$, the following lemma holds.

Lemma 2. *Assume* $\mathbb{E}|g-\mathbb{E}g|^\ell\le\frac{1}{2}\ell!M^{\ell-2}v$ *and* $\mathbb{E}^{(i)}\left|g-\mathbb{E}^{(i)}g\right|^\ell\le\frac{1}{2}\ell!M^{\ell-2}v$ *for some constants* $M,v>0$ *and any* $i\in\mathbb{N},\ell=2,3,\cdots$, *then we have*

$$\mathrm{Prob}_{\mathbf{z}\in Z^m}\left\{\mathbb{E}_{\mathbf{z}}g-\mathbb{E}_m g\ge\varepsilon\right\}\le\exp\left\{-\frac{m\varepsilon^2}{2(v+M\varepsilon)}\right\}.$$

Now we can bound the sample error term $\mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}})-\mathcal{E}_m(f_{\mathcal{H}})$ by applying this lemma.

Proposition 3. *Under the moment incremental condition* (4), (5) *and notations above*, *with probability at least* $1-\delta/3$, *we have*

$$\left(\mathcal{E}_{\mathbf{z}}\left(f_{\mathcal{H}}\right)-\mathcal{E}_{\mathbf{z}}\left(f_{\rho}\right)\right)-\left(\mathcal{E}_{m}\left(f_{\mathcal{H}}\right)-\mathcal{E}_{m}\left(f_{\rho}\right)\right)\le\frac{2(C+2)M_B}{m}\log\frac{3}{\delta}+\frac{1}{2}A_{\mathcal{H}},$$

where $M_B=6\left(B+(C+3)M\right)^2$ and $A_{\mathcal{H}}$ is the approximation error.

Proof. Let

$g_{\mathcal{H}}(z)=\left(f_{\mathcal{H}}(x)-y\right)^2-\left(f_\rho(x)-y\right)^2=\left(f_{\mathcal{H}}(x)-f_\rho(x)\right)\left(f_{\mathcal{H}}(x)+f_\rho(x)-2y\right)$, then

$$\left(\mathcal{E}_{\mathbf{z}}\left(f_{\mathcal{H}}\right)-\mathcal{E}_{\mathbf{z}}\left(f_{\rho}\right)\right)-\left(\mathcal{E}_{m}\left(f_{\mathcal{H}}\right)-\mathcal{E}_{m}\left(f_{\rho}\right)\right)=\mathbb{E}_{\mathbf{z}}g-\mathbb{E}_{m}g.$$

Since $\left|f_\rho(x)\right|=\left|\int_Y y\,\mathrm{d}\rho(y\,|\,x)\right|\le\mathbb{E}\left(|y|\,|\,x\right)\le\sqrt{\mathbb{E}\left(|y|^2\,|\,x\right)}\le\sqrt{2C}M$, we have

$$\begin{aligned}
&\mathbb{E}\left|g_{\mathcal{H}}-\mathbb{E}g_{\mathcal{H}}\right|^{\ell}\le2^{\ell}\,\mathbb{E}\left(\left|g_{\mathcal{H}}\right|^{\ell}+\left|\mathbb{E}g_{\mathcal{H}}\right|^{\ell}\right)\le2^{\ell+1}\,\mathbb{E}\left|g_{\mathcal{H}}\right|^{\ell}\\
&=2^{\ell+1}\int_Z\left|f_{\mathcal{H}}(x)-f_\rho(x)\right|^{\ell}\cdot\left|f_{\mathcal{H}}(x)+f_\rho(x)-2y\right|^{\ell}\mathrm{d}\rho\\
&\le2^{\ell+1}\left(B+CM\right)^{\ell-2}\int_X\left|f_{\mathcal{H}}(x)-f_\rho(x)\right|^{2}\int_Y\left(B+\sqrt{2C}M+2|y|\right)^{\ell}\mathrm{d}\rho(y\,|\,x)\,\mathrm{d}\rho_X\\
&\le2^{\ell+1}\left(B+CM\right)^{\ell-2}\left(\mathcal{E}\left(f_{\mathcal{H}}\right)-\mathcal{E}\left(f_\rho\right)\right)\int_Y3^{\ell}\left(B^{\ell}+(2C)^{\frac{\ell}{2}}M^{\ell}+2^{\ell}|y|^{\ell}\right)\mathrm{d}\rho(y\,|\,x)\\
&\le2^{\ell+1}\left(B+CM\right)^{\ell-2}\cdot A_{\mathcal{H}}\cdot3^{\ell}\left(B^{\ell}+(2C)^{\frac{\ell}{2}}M^{\ell}+2^{\ell}C\ell!M^{\ell}\right)\\
&\le2\ell!\cdot6^{\ell}\left(B+(C+3)M\right)^{2\ell-2}(C+1)A_{\mathcal{H}}\le\frac{1}{2}\ell!M_B^{\ell-2}v_B
\end{aligned}$$

for any $\ell=1,2,\cdots$, where $_1$ and $v_B=4(C+1)M_BA_{\mathcal{H}}$. In the same way, we have the following bounds

$$\mathbb{E}^{(i)}\left|g_{\mathcal{H}}-\mathbb{E}^{(i)}g_{\mathcal{H}}\right|^{\ell}\le\frac{1}{2}\ell!M_B^{\ell-2}v_B,\quad i,\ell=1,2,\cdots$$

as well. Then from Lemma 2 above, we have

$$\mathrm{Prob}_{\mathbf{z}\in Z}\left\{\mathbb{E}_{\mathbf{z}}g_{\mathcal{H}}-\mathbb{E}_{m}g_{\mathcal{H}}\ge\varepsilon\right\}\le\exp\left\{-\frac{m\varepsilon^2}{2\left(v_B+M_B\varepsilon\right)}\right\}.$$

Set the right hand side to be $\delta/3$, we can solve that

$$\begin{aligned}
\mathbb{E}_{\mathbf{z}}g_{\mathcal{H}}-\mathbb{E}_{m}g_{\mathcal{H}}\le\varepsilon&=\frac{1}{m}\left(M_B\log\frac{3}{\delta}+\sqrt{M_B^2\log^2\frac{3}{\delta}+2mv_B\log\frac{3}{\delta}}\right)\\
&\le\frac{2M_B}{m}\log\frac{3}{\delta}+\sqrt{\frac{4(C+1)M_BA_{\mathcal{H}}}{m}\log\frac{3}{\delta}}\\
&\le\frac{2(C+2)M_B}{m}\log\frac{3}{\delta}+\frac{1}{2}A_{\mathcal{H}}.
\end{aligned}$$

Therefore with confidence at least $1-\delta/3$, there holds

$$\mathbb{E}_{\mathbf{z}}g_{\mathcal{H}}-\mathbb{E}_{m}g_{\mathcal{H}}\le\frac{2(C+2)M_B}{m}\log\frac{3}{\delta}+\frac{1}{2}A_{\mathcal{H}}.$$

This proves the proposition.

For the sample error term involving $f_{\mathbf{z}}$, analysis will be more involved since we need a concentration inequality for a set of functions. Firstly we have to introduce the ratio inequality [9].

Lemma 3. *Denote* $g(z)=\left(f(x)-y\right)^2-\left(f_\rho(x)-y\right)^2$ *for* $f\in\mathcal{H}$, *which satisfies* $\mathbb{E}\left|g-\mathbb{E}g\right|^{\ell}\le\frac{1}{2}\ell!M^{\ell-2}v$

and $\mathbb{E}^{(i)}\left|g-\mathbb{E}^{(i)}g\right|^{\ell}\le\frac{1}{2}\ell!M^{\ell-2}v$ *for some constants* $M,v>0$ *and* $i,\ell=2,3,\cdots$, *then we have*

$$\mathrm{Prob}_{\mathbf{z}\in Z^m}\left\{\mathbb{E}_z g-\mathbb{E}_m g\ge\sqrt{\varepsilon\left(\varepsilon+\mathbb{E}g\right)}\right\}\le\exp\left\{-\frac{m\varepsilon}{2M\left(4C+5\right)}\right\}.$$

Proof. Let ε to be $\sqrt{\varepsilon\left(\varepsilon+\mathcal{E}(f)-\mathcal{E}\left(f_\rho\right)\right)}$ in the Lemma 2, from the proof of the last proposition, we can conclude that

$$\begin{aligned}&\mathrm{Prob}_{\mathbf{z}\in Z^m}\left\{\mathbb{E}_{\mathbf{z}}g-\mathbb{E}_m g\ge\sqrt{\varepsilon\left(\varepsilon+\mathcal{E}(f)-\mathcal{E}\left(f_\rho\right)\right)}\right\}\\&\le\exp\left\{-\frac{m\varepsilon\left(\varepsilon+\mathcal{E}(f)-\mathcal{E}\left(f_\rho\right)\right)}{2\left[4(C+1)M_B\left(\mathcal{E}(f)-\mathcal{E}\left(f_\rho\right)\right)+M_B\left(\varepsilon+\mathcal{E}(f)-\mathcal{E}\left(f_\rho\right)\right)\right]}\right\}\\&\le\exp\left\{-\frac{m\varepsilon}{2M_B\left(4C+5\right)}\right\}.\end{aligned}$$

Note that $\mathcal{E}(f)-\mathcal{E}\left(f_\rho\right)=\mathbb{E}g$ and the lemma is proved.

Then we have the following result.

Lemma 4. *For a set of functions* $\left\{f_i\right\}_{i=1}^{N}\subset\mathcal{H}$ *with* $N\in\mathbb{N}$, *construct functions* $g_i(z)=-\left(\left(f_i(x)-y\right)^2-\left(f_\rho(x)-y\right)^2\right)$ *for* $i=1,2,\cdots,N$, *with confidence at least* $1-\delta/3$, *we have*

$$\mathbb{E}_{\mathbf{z}}g_i-\mathbb{E}_m g_i\le\frac{2M_B\left(4C+5\right)}{m}\log\frac{3N}{\delta}+\frac{1}{2}A_i,\quad\forall i=1,2,\cdots,N$$

where $A_i=\mathcal{E}\left(f_i\right)-\mathcal{E}\left(f_\rho\right)$ *for any* $i=1,2,\cdots,N.$

Proof. Since f_i is an element of $\mathcal{H}$, from Lemma 3 we have

$$\mathrm{Prob}_{\mathbf{z}\in Z^m}\left\{\frac{\mathbb{E}_{\mathbf{z}}g_i-\mathbb{E}_m g_i}{\sqrt{\varepsilon+\mathbb{E}g_i}}\ge\sqrt{\varepsilon}\right\}\le\exp\left\{-\frac{m\varepsilon}{2M_B\left(4C+5\right)}\right\},$$

then there holds

$$\mathrm{Prob}_{\mathbf{z}\in Z^m}\left\{\sup_{1\le i\le N}\frac{\mathbb{E}_{\mathbf{z}}g_i-\mathbb{E}_m g_i}{\sqrt{\varepsilon+\mathbb{E}g_i}}\ge\sqrt{\varepsilon}\right\}\le N\exp\left\{-\frac{m\varepsilon}{2M_B\left(4C+5\right)}\right\}.$$

Set the right hand side to be $\delta/3$ and we have with probability at least $1-\delta/3$,

$$\begin{aligned}\mathbb{E}_{\mathbf{z}}g_i-\mathbb{E}_m g_i&\le\sqrt{\varepsilon\left(\varepsilon+\mathbb{E}g_i\right)}\le\varepsilon+\frac{1}{2}\mathbb{E}g_i\\&\le\frac{2M_B\left(4C+5\right)}{m}\log\frac{3N}{\delta}+\frac{1}{2}A_i,\ i=1,2,\cdots,N.\end{aligned}$$

Here $A_i=\mathcal{E}\left(f_i\right)-\mathcal{E}\left(f_\rho\right)=\mathbb{E}g_i$. And this proves the lemma.

Now by a covering number argument we can bound the sample error term involving $f_{\mathbf{z}}$.

Proposition 4. *If* $\mathcal{H}=B_R(H):=\left\{f\in H:\|f\|_H\le R\right\}$ *for some* $R>0$, *where* H *satisfies the capacity condition, for any* $0<\delta<1$, *with confidence at least* $1-2\delta/3$, *there holds*

$$\begin{aligned}&\left(\mathcal{E}_m\left(f_{\mathbf{z}}\right)-\mathcal{E}_m\left(f_\rho\right)\right)-\left(\mathcal{E}_{\mathbf{z}}\left(f_{\mathbf{z}}\right)-\mathcal{E}_{\mathbf{z}}\left(f_\rho\right)\right)\\&\le m^{-\frac{1}{1+r}}\log\frac{3}{\delta}\left(5B+(13C+16)M+2M_B\left(4C+5\right)\left(C_rR^r+1\right)\right)+\frac{1}{2}\left(\mathcal{E}\left(f_{\mathbf{z}}\right)-\mathcal{E}\left(f_\rho\right)\right).\end{aligned}$$

Proof. Denote $N=\mathcal{N}(\mathcal{H},\eta)$ where η is to be determined, then we can find an η-net $\{f_i\}_{i=1}^{N}$ of $\mathcal{H}$, and there exist a function $f_j, j\in\{1,2,\cdots,N\}$, we have

$$\begin{aligned}
&\left(\mathcal{E}_m(f_{\mathbf{z}})-\mathcal{E}_m(f_\rho)\right)-\left(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})-\mathcal{E}_{\mathbf{z}}(f_\rho)\right)\\
&=\left(\mathcal{E}_m(f_{\mathbf{z}})-\mathcal{E}_m(f_j)\right)+\left(\mathcal{E}_m(f_j)-\mathcal{E}_m(f_\rho)\right)-\left(\mathcal{E}_{\mathbf{z}}(f_j)-\mathcal{E}_{\mathbf{z}}(f_\rho)\right)+\left(\mathcal{E}_{\mathbf{z}}(f_j)-\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})\right)\\
&=\left(\mathcal{E}_m(f_{\mathbf{z}})-\mathcal{E}_m(f_j)\right)+\left(\mathbb{E}_{\mathbf{z}}g_j-\mathbb{E}_m g_j\right)+\left(\mathcal{E}_{\mathbf{z}}(f_j)-\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})\right).
\end{aligned}$$

For the first term, since $\int_Y y\mathrm{d}\rho^{(i)}(y\mid x)\le\sqrt{2C}M$ for all $i=1,2,\cdots,m$, we have

$$\begin{aligned}
&\mathcal{E}_m(f_{\mathbf{z}})-\mathcal{E}_m(f_j)\\
&=\frac{1}{m}\sum_{i=1}^{m}\int_X\int_Y\left(f_{\mathbf{z}}(x)-f_j(x)\right)\left(f_{\mathbf{z}}(x)+f_j(x)-2y\right)\mathrm{d}\rho^{(i)}(y\mid x)\mathrm{d}\rho_X^{(i)}\\
&\le\frac{\eta}{m}\sum_{i=1}^{m}2\left(B+\sqrt{2C}M\right)=2\eta\left(B+\sqrt{2C}M\right).
\end{aligned}$$

And for the third term,

$$\begin{aligned}
\mathcal{E}_{\mathbf{z}}(f_j)-\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})&=\frac{1}{m}\sum_{i=1}^{m}\left(\left(f_j(x)-y\right)^2-\left(f_{\mathbf{z}}(x)-y\right)^2\right)\\
&=\frac{1}{m}\sum_{i=1}^{m}\left(f_j(x)-f_{\mathbf{z}}(x)\right)\left(f_j(x)+f_{\mathbf{z}}(x)-2y\right)\le\eta\left(2B+2\frac{1}{m}\sum_{i=1}^{m}|y|\right),
\end{aligned}$$

we need to bound

$$\frac{1}{m}\sum_{i=1}^{m}|y|=\mathbb{E}_z|y|=\left(\mathbb{E}_z|y|-\mathbb{E}|y|\right)+\mathbb{E}|y|.$$

Let $g(z)=|y|$ and then

$$\mathbb{E}|g-\mathbb{E}g|_\ell\le2^{\ell+1}\mathbb{E}|y|^\ell\le2C\ell!(2M)^\ell=\frac{1}{2}\ell!(2M)^{\ell-2}\left(16CM^2\right),\quad\ell=1,2,\cdots.$$

From Lemma 1 we have

$$\mathrm{Prob}_{\mathbf{z}\in Z^m}\left\{\mathbb{E}_{\mathbf{z}}g-\mathbb{E}g\ge\varepsilon\right\}\le\exp\left\{-\frac{m\varepsilon^2}{2\left(16CM^2+2M\varepsilon\right)}\right\}.$$

Set the right hand side to be $\delta/3$ and with confidence at least $1-\delta/3$ we have

$$\mathbb{E}_z g-\mathbb{E}g=\frac{1}{m}\sum_{i=1}^{m}|y|-\int_Z|y|\mathrm{d}\rho\le\varepsilon\le\frac{4M}{m}\log\frac{3}{\delta}+\sqrt{\frac{32CM^2}{m}\log\frac{3}{\delta}}\le4\left(1+\sqrt{2C}\right)\log\frac{3}{\delta}.$$

And this means,

$$\frac{1}{m}\sum_{i=1}^{m}|y|\le\int_Y|y|\mathrm{d}\rho(y\mid x)+4M\left(1+\sqrt{2C}\right)\log\frac{3}{\delta}\le M(5C+9)\log\frac{3}{\delta}$$

with probability at least $1-\delta/3$.

The second term can be bounded by 4 above. That is, with confidence at least $1-\delta/3$, we have

$$\mathbb{E}_{\mathbf{z}}g_j-\mathbb{E}g_j\le\frac{2M_B(4C+5)}{m}\log\frac{3N}{\delta}+\frac{1}{2}A_j.$$

Since $\log N=\log\mathcal{N}(\mathcal{H},\eta)=\log\mathcal{N}\left(B_1(H),\frac{\eta}{R}\right)\le C_r(R/\eta)^r$ by assumption, and

$$A_j = \mathcal{E}(f_j) - \mathcal{E}(f_\rho) = \mathcal{E}(f_j) - \mathcal{E}(f_{\mathbf{z}}) + \mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) \le \eta(2B + 2CM) + \left(\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho)\right),$$

combining the three parts above, we have the following bound with confidence at least $1 - 2\delta/3$,

$$\begin{aligned}
&\left(\mathcal{E}_m(f_{\mathbf{z}}) - \mathcal{E}_m(f_\rho)\right) - \left(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_\rho)\right) \\
&\le 2\eta\left(B + (5C+9)M \log\frac{3}{\delta}\right) + 2\eta\left(B + (C+1)M\right) + \frac{2M_B(4C+5)}{m}\log\frac{3N}{\delta} \\
&\quad + \eta(B + CM) + \frac{1}{2}\left(\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho)\right) \\
&\le \eta\left(5B + (13C+20)M \log\frac{3}{\delta}\right) + \frac{2M_B(4C+5)}{m}\log\frac{3N}{\delta} \\
&\quad + \frac{2C_r M_B(4C+5)R^r}{m}\eta^{-r} + \frac{1}{2}\left(\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho)\right).
\end{aligned}$$

By choosing $\eta = m^{-1/(1+r)}$ for balancing, we have

$$\begin{aligned}
&\left(\mathcal{E}_m(f_{\mathbf{z}}) - \mathcal{E}_m(f_\rho)\right) - \left(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_\rho)\right) \\
&\le m^{-\frac{1}{1+r}} \log\frac{3}{\delta}\left[5B + (13C+20)M + 2M_B(4C+5)\left(C_r R^r + 1\right)\right] + \frac{1}{2}\left(\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho)\right)
\end{aligned}$$

with confidence at least $1 - 2\delta/3$, this proves the proposition.

5. Approximation Error and Total Error

Combining the results above, we can derive the error bound for the generalization error $\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho)$.

Proposition 5. *Under the moment condition for the distribution of the sample and capacity condition for the hypothesis space* $\mathcal{H}$, *for any* $0 < \delta < 1$ *and* $1 \le p \in \mathbb{N}$, *with confidence at least* $1 - \delta$, *we have*

$$\begin{aligned}
&\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) = \mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_{\mathcal{H}}) + A_{\mathcal{H}} \\
&\le \left(\frac{w_b(m)}{m}\right)^{\frac{p-1}{p}} \left[2(C_b + 1)B^2 + 2\left(2CM + 3pC_bM + 4\sqrt{2C} + 4C_s\right)B\right. \\
&\quad \left. + 4\left(2C^2 + 3CC_bp + 4C\right)M^2 + 12\sqrt{2C}C_sM\right] \\
&\quad + m^{-\frac{1}{1+r}}\frac{3}{\delta}\cdot 2\left[5B + (13C+20)M + 2M_B(4C+5)\left(C_rR^r + 1\right)\right] \\
&\quad + \frac{4(C+2)M_B}{m}\frac{3}{\delta} + 3A_{\mathcal{H}},
\end{aligned}$$

where $M_B = 6\left(B + (C+3)M\right)^2$.

What is left to be determined in the proposition is the approximation error $A_{\mathcal{H}}$. By the choice of hypothesis space we can get our main result.

Proof of Theorem 1. Let

$$v_b(m) = \begin{cases} m^{-b}, & 0 < b < 1, \\ \dfrac{\log m}{m}, & b = 1, \\ m^{-1}, & b > 1, \end{cases}$$

and $\gamma(m) = \max\left\{\left(v_b(m)\right)^{(p-1)/p}, m^{-1/(1+r)}\right\}$, assume $B \ge 1$ without loss of generality, and $R \ge 1$, Proposition 5 indicates that

$$\mathcal{E}\left(f_{\mathbf{z}}\right)-\mathcal{E}\left(f_{\rho}\right)=\mathcal{E}\left(f_{\mathbf{z}}\right)-\mathcal{E}\left(f_{\mathcal{H}}\right)+A_{\mathcal{H}}\leq\gamma\left(m\right)\log\frac{3}{\delta}B^{2}R^{r}C_{b,s,r,p,M}+3A_{\mathcal{H}}$$

holds with confidence at least $1-\delta$ for any $1\leq p\in\mathbb{N}$, where $C_{b,s,r,p,M}$ is a constant independent on m or δ.

For the approximation error $A_{\mathcal{H}}$, we can bound it by Theorem 3.1 of [18]. Since the hypothesis space $\mathcal{H}=B_{R}\left(H\right)$, and $f_{\rho}\in\left(L^{2}_{\rho_X},H\right)_{0,\infty}$ with $0<\theta<1$, we have

$$A_{\mathcal{H}}=\int_{X}\left|f_{\mathcal{H}}\left(x\right)-f_{\rho}\left(x\right)\right|^{2}\mathrm{d}\rho_{X}\leq\left(\frac{1}{R}\right)^{\frac{2\theta}{1-\theta}}\left(\left\|f_{\rho}\right\|_{0,\infty}\right)^{\frac{2}{1-\theta}}.$$

The upper bound B is now chosen to be $\|I\|R$ since $\|f\|_{C}\left(X\right)\leq\|I\|\cdot\|f\|_{H}\leq\|I\|R$, then with confidence at least $1-\delta$,

$$\mathcal{E}\left(f_{\mathbf{z}}\right)-\mathcal{E}\left(f_{\rho}\right)\leq\gamma\left(m\right)\log\frac{3}{\delta}\|I\|^{2}R^{2+r}C_{b,s,r,p,M}+3R^{-\frac{2\theta}{1-\theta}}\left(\left\|f_{\rho}\right\|_{0,\infty}\right)^{\frac{2}{1-\theta}}.$$

By choosing

$$R=\left(\gamma\left(m\right)\right)^{-\frac{1}{r+\frac{2}{1-\theta}}},$$

we have

$$\mathcal{E}\left(f_{\mathbf{z}}\right)-\mathcal{E}\left(f_{\rho}\right)\leq\left(\gamma(m)\right)^{\frac{2\theta}{r-r\theta+2}}\log\frac{3}{\delta}\left[C_{b,s,r,p,M}\|I\|^{2}+3\left(\left\|f_{\rho}\right\|_{0,\infty}\right)^{\frac{2}{1-\theta}}\right]$$

holds with confidence at least $1-\delta$. Denote $\tilde{C}=C_{b,s,r,p,M,\theta,\mathcal{H}}=C_{b,s,r,p,M}\|I\|^{2}+3\left(\left\|f_{\rho}\right\|_{0,\infty}\right)^{\frac{2}{1-\theta}}$, then the theorem is obtained.

6. Summary and Future Work

We investigate the least squares ERM algorithm with non-identical and unbounded sample, *i.e.*, polynomial convergence for $\left\{\rho_{X}^{(i)}\right\}_{i\geq1}$ and $\left\{\rho^{(i)}\left(y\mid x\right)\right\}_{i\geq1}$ and moment inremental condition for the latter ones. Analogue error decomposition as classical analysis for least sqaures regularization [9] [11] is conducted. Truncate technique is introduced for handling unbounded setting, and Bennett concentration inequality is used for the sample error. By the above analysis we finally get the error bound and learning rate.

However, our work only considers the ERM algorithm. It is neccesary for us to extend this to the regularization algorithms which are more widely used in practice. A more recent relative reference can be found in [21]. Another interesting topic in future study is dependent sampling [7].

References

[1] Cucker, F. and Zhou, D.X. (2007) Learning Theory: An Approximation Theory Viewpoint. Cambridge University Press, Cambridge. http://dx.doi.org/10.1017/CBO9780511618796

[2] Cucker, F. and Smale, S. (2002) On the Mathematical Foundations of Learning. *Bulletin of the American Mathematical Society*, **39**, 1-49.

[3] Dehling, H., Mikosch, T. and Sorensen, M. (2002) Empirical Process Techniques for Dependent Data. Birkhauser Boston, Inc., Boston. http://dx.doi.org/10.1007/978-1-4612-0099-4

[4] Steinwart, I., Hush, D. and Scovel, C. (2009) Learning from Dependent Observations. *Journal of Multivariate Analysis*, **100**, 175-194.

[5] Steinwart, I. and Christmann, A. (2009) Fast Learning from Non-i.i.d. Observations. In: Bengio, Y., Schuurmans, D., Lafferty, J.D., Williams, C.K.I. and Culotta, A., Eds., *Advances in Neural Information Processing Systems* 22, Curran and Associates, Inc., Yellowknife, 1768-1776.

[6] Xiao, Q.W. and Pan, Z.W. (2010) Learning from Non-Identical Sampling for Classification. *Advances in Computational Mathematics*, **33**, 97-112.

[7] Pan, Z.W. and Xiao, Q.W. (2009) Least-Square Regularized Regression with Non-i.i.d. Sampling. *Journal of Statistical Planning and Inference*, **139**, 3579-3587.

[8] Hu, T. and Zhou, D.X. (2009) Online Learning with Samples Drawn from Non-identical Distributions. *Journal of Machine Learning Research*, **10**, 2873-2898.

[9] Wu, Q., Ying, Y. and Zhou, D.X. (2006) Learning Rates of Least-Square Regularized Regression. *Foundations of Computational Mathematics*, **6**, 171-192.

[10] Capponnetto, A. and De Vito, E. (2007) Optimal Rates for the Regularized Least Squares Algorithm. *Foundations of Computational Mathematics*, **7**, 331-368.

[11] Wang, C. and Zhou, D.X. (2011) Optimal Learning Rates for Least Squares Regularized Regression with Unbounded Sampling. *Journal of Complexity*, **27**, 55-67.

[12] Guo, Z.C. and Zhou, D.X. (2013) Concentration Estimates for Learning with Unbounded Sampling. *Advances in Computational Mathematics*, **38**, 207-223.

[13] He, F. (2014) Optimal Convergence Rates of High Order Parzen Windows with Unbounded Sampling. *Statistics & Probability Letters*, **92**, 26-32.

[14] Zhou, D.X. (2002) The Covering Number in Learning Theory. *Journal of Complexity*, **18**, 739-767.

[15] Zhou, D.X. (2003) Capacity of Reproducing Kernel Spaces in Learning Theory. *IEEE Transactions on Information Theory*, **49**, 1743-1752.

[16] Zhou, D.X. (2008) Derivative Reproducing Properties for Kernel Methods in Learning Theory. *Journal of Computational and Applied Mathematics*, **220**, 456-463.

[17] Smale, S. and Zhou, D.X. (2009) Online Learning with Markov Sampling. *Analysis and Applications*, **7**, 87-113.

[18] Smale, S. and Zhou, D.X. (2003) Estimating the Approximation Error in Learning Theory. *Analysis and Applications*, **1**, 17-41.

[19] Wang, C. and Guo, Z.C. (2012) ERM Learning with Unbounded Sampling. *Acta Mathematica Sinica*, *English Series*, **28**, 97-104.

[20] Bennett, G. (1962) Probability Inequalities for the Sum of Independent Random Variables. *Journal of the American Statistical Association*, **57**, 33-45.

[21] Cai, J. (2013) Coefficient-Based Regression with Non-Identical Unbounded Sampling. *Abstract and Applied Analysis*, **2013**, Article ID: 134727. http://dx.doi.org/10.1155/2013/134727