Generalization ability of extreme learning machine with uniformly ergodic Markov chains

Peipei Yuan, Hong Chen, Yicong Zhou, Xiaoyan Deng, Bin Zou

1. Introduction

Extreme learning machine (ELM) can be considered as a single-hidden layer feedforward neural networks (FNNs), where the output weights can be adjusted while the input weights and the threshold of hidden layer are fixed randomly [6,9]. This idea of training FNNs is different from the traditional neural network theories and is related with the discussions in [13,14]. Because only the Moore–Penrose generalized inverse is necessary to be calculated, the original ELM and its variations have shown the computational feasibility in the various applications, see, e.g., [2,4,5,11,23]. With the rapid development of the ELM-based applications, there are some theoretical works for its universal consistency in [25] and generalization ability in [10,19,2]. In particular, the generalization bounds of ELM are established in [10], which demonstrate that ELM can achieve the same learning rates as FNNs under mild conditions. Therefore, it is important to further investigate the generalization ability of ELM with dependent samples.

Recently, the Markov chain samples have attracted increasing attention in statistical learning theory. In [17], the learning rate is estimated for the online algorithm with the Markov chains. For the uniformly ergodic Markov chains (u.e.M.c) samples. The upper bound of the misclassiﬁcation error is estimated for the ELM classiﬁcation showing that the satisfactory learning rate can be achieved even for the dependent samples. Empirical evaluations on real-word datasets are provided to compare the predictive performance of ELM with independent and Markov sampling.

2. Preliminaries

Let $X \in \mathbb{R}^d$ be the input space and $Y = \{-1, 1\}$. The training samples $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m \in \mathbb{Z}^m$ are drawn from a probability distribution $p$ on $Z = X \times Y$. Given $\mathbf{z}$, the main goal of the classification algorithm is searching a predictor $f_\mathbf{z}: X \to \{-1, 1\}$ such that
the misclassification rate is as low as possible. In learning theory, the misclassification risk is defined as
\[ \mathcal{R}(f) = \int_{\mathcal{X}} |y - f(x)| \, dp \]
and the Bayes risk is denoted by
\[ \mathcal{R}^* = \min_{f \in \mathcal{F}} \mathcal{R}(f). \]
For the regression function \( f_\rho = \int_{\mathcal{Y}} y \, dp(y|x) \), we know that \( \mathcal{R}^* = \mathcal{R}(f_\rho) \), where \( f_\rho = \text{sign}(f_\rho) \), and \( \text{sign}(t) = 1 \) if \( t \geq 0 \) and \( \text{sign}(t) = -1 \) otherwise. The performance of a classifier is measured by the excess risk \( \mathcal{R}(f) - \mathcal{R}(f_\rho) \). Since the inductor loss \( l \) is nonconvex and noncontinuous, we usually use the convex loss to replace it.

In original ELM, the least square loss \( \frac{1}{2} \| f - y \|^2 \) is used. Denote \( A = (a_1, a_2, \ldots, a_n)^T \in \mathbb{R}^n \), in which \( a_i \) is generated independently and identically according to a uniform distribution \( \mu \) on \([0, 1]^d\). In ELM, the hypothesis space is defined as
\[ \mathcal{M}_n = \left\{ f(x, \alpha, \beta) = \sum_{i=1}^n \beta_i \phi(a_i, x) : x \in \mathcal{X}, \beta = (\beta_1, \ldots, \beta_n)^T \in \mathbb{R}^n \right\}, \tag{1} \]
where \( \phi : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R} \) is an activation function. The activation functions include the sigmoid function, Gaussian function, hyperbolic tangent function, multiquadric function and Fourier series. This paper focuses on bounding the excess risk \( \mathcal{E}(f) - \mathcal{E}(f_\rho) \) to measure the generalization ability of ELM. The current analysis is based on the u.e.M.c samples different from the previous works in [10,19].

From Proposition 1, we recall some preliminary definition and properties of the u.e.M.c [12,18,22]. Let \((Z, \mathcal{F}, \mu)\) be a measurable space. We call \( \{Z_t\}_{t=1}^\infty \) is a Markov chain, if the sequence \( \{Z_t\}_{t=1}^\infty \) is randomly generated and its transition probability measure satisfies
\[ P^t(Z_t) = \text{Prob}[Z_{t+1} \in A|Z_t = z]. \tag{3} \]
Starting from the initial state \( z \) at time \( 0 \), the probability that the state \( z_{k+1} \) will belong to set \( A \) after \( k \) steps, is denoted by \( P^k(Z_t) \). Hence, if \( k = 1 \), we have \( P^1(Z_t) = \text{Prob}[Z_{t+1} \in A|Z_t = z] \), which is independent of the values of \( Z(f < t) \). For the given probabilities \( p_1 \) and \( p_2 \), the total variation distance is defined as
\[ \| p_1 - p_2 \|_{TV} = \sup_{A \in \mathcal{F}} |p_1(A) - p_2(A)|. \] (4)
Definition 1. A Markov chain \( \{Z_t\}_{t=1}^\infty \) is said to be uniformly ergodic if
\[ P^t(z_1 - \pi(z)) \|_{TV} \leq y^t, \tag{4} \]
for some \( 0 < y < \infty \) and \( 0 < \tau < 1 \). Here \( k \geq 1, k \in \mathbb{N} \) and \( \pi(z) \) is the stationary distribution of \( \{Z_t\}_{t=1}^\infty \).

From [12], we know that the transition probability \( P^k(Z_t) \) of the u.e.M.c satisfies the Dooblin condition as below.

**Proposition 1.** Let \( \{Z_t\}_{t=1}^\infty \) be a Markov chain with the transition probability measure \( P^k(Z_t) \) and let \( \mu \) be a nonnegative measure with nonzero mass \( \mu_0 \). Assume that, for some integer \( t \) and all measurable sets \( A, P^t(A) \leq \mu(A), \forall z \in \mathcal{F} \). Then, we have
\[ \| P^t(z_1 - P^t(z_2)) \|_{TV} \leq 2(1 - \mu_0)^t, \quad \forall k \in \mathbb{N}, \ z, z' \in Z. \tag{5} \]

3. Generalization bound
To evaluate the generalization ability of ELM, we should estimate the approximation between \( f_{\rho, \lambda} \) and \( f_{\rho} \). That is to say, we should estimate the excess convex risk \( \mathcal{E}(f_{\rho, \lambda}) - \mathcal{E}(f_{\rho}) \). Propreposition 2. For any \( z \in \mathbb{R}^m \) and \( f_{\rho, \lambda} \) defined in (2), there holds
\[ \mathcal{E}(f_{\rho, \lambda}) - \mathcal{E}(f_{\rho}) \leq S_1 + S_2, \tag{6} \]
where
\[ S_1 = \mathcal{E}(f_{\rho, \lambda}) - \mathcal{E}(f_{\rho}) - (\mathcal{E}(f_{\rho, \lambda}) - \mathcal{E}(f_{\rho})) \]
and
\[ S_2 = \mathcal{E}(f_{\rho, \lambda}) - \mathcal{E}(f_{\rho}) + \lambda \| f_{\rho, \lambda} \|_{L_2}^2. \]

Definition 2. For a subset \( \mathcal{G} \) of a metric space and any \( \epsilon > 0 \), the covering number \( N(\mathcal{G}, \epsilon) \) is defined to be the smallest integer \( l \in \mathbb{N} \) such that there exist \( l \) disks with radius \( \epsilon \) and centers in \( \mathcal{G} \) covering \( \mathcal{G} \).

For any given \( R > 0 \), we define a class of functions:
\[ B_R = \{ f \in \mathcal{M}_n : \| f \|_{L_2}^2 \leq R^2 \}. \]
The covering number of \( B_R \) is estimated in [3].

Lemma 1. For any \( R > 0, \epsilon > 0 \), there holds
\[ \log N(B_R, \epsilon) \leq n \log \left( \frac{4R}{\epsilon} \right). \tag{7} \]
random samples (see [18]). In order to estimate the generalization bound, we introduce the following lemma established in [22].

**Lemma 2.** Let \( G \) be a countable class of bounded measurable functions and let \( z = \{z_i\}_{i=1}^m \) be a set of u.e.M.c samples. For some \( C > 0 \), there exists \( 0 \leq g(z) \leq C \) for all \( g \in G, z \in Z \). Then, for any \( \epsilon > 0 \), we have

\[
\Pr_{z \sim \mathcal{Z}} \left( \frac{1}{m} \sum_{i=1}^m g(z_i) - E(g) \geq \epsilon \right) \leq 2 \exp \left( \frac{-m \epsilon^2}{56C \|f\|^2 E(g)} \right).
\]

As shown in [22], the following lemma can be deduced by Lemma 2. For completeness, we present its proof in Appendix.

**Lemma 3.** Let \( G \) be a countable class of bounded measurable functions and let \( z = \{z_i\}_{i=1}^m \) be a set of u.e.M.c samples. For all \( g \in G, z \in Z \), assume that \( 0 \leq g(z) \leq C \) for some \( C > 0 \). Then, for any \( \epsilon > 0 \), there holds that

\[
\Pr_{z \sim \mathcal{Z}} \left( \sup_{g \in G} \frac{1}{m} \sum_{i=1}^m g(z_i) - E(g) \geq \epsilon \right) \leq N(G, \epsilon) \exp \left( \frac{-m \epsilon^2}{56C \|f\|^2} \right).
\]

Now, we present the main result on the excess risk \( \mathcal{E}(f_{\mathcal{X}}) - \mathcal{E}(f) \).

**Theorem 1.** Assume that \( z = \{z_i\}_{i=1}^m \) is a set of u.e.M.c samples and \( \|\phi\| \leq \kappa \). For any \( m \geq 1 \), we have

\[
P_{\mathbf{P}} \mathcal{E}_{\mathbf{P}}(\|f_{\mathcal{X}} - f_{\phi}\|_2^2) \leq C_n \log (m/\delta) \left( \frac{2}{m} \right)^{\frac{\mathbf{E}}{\mathbf{P}} \phi \phi} + 2E_{\mathbf{P}} \mathcal{E}_{\mathbf{P}} \left( \sum_{f \in \mathbf{P}} \frac{E_{\mathbf{P}}(\mathcal{E}(f_{\mathcal{X}}) - \mathcal{E}(f_{\phi})))}{\mathcal{E}(f_{\phi})} + \mathcal{E}_{\mathbf{P}} \mathcal{E}_{\mathbf{P}}(\kappa f_{\phi}) + \lambda \|f_{\mathcal{X}}\|_2^2 \right) \leq \mathcal{N}(\mathbf{P}, \delta) \exp \left( \frac{-m \epsilon^2}{56C \|f\|^2} \right).
\]

**Proof.** From the Proposition 2, we conclude that

\[
\mathcal{E}_{\mathbf{P}}(\|f_{\mathcal{X}} - f_{\phi}\|_2^2) \leq \mathcal{E}_{\mathbf{P}}(\mathcal{E}(f_{\mathcal{X}}) - \mathcal{E}(f_{\phi})) + \mathcal{E}_{\mathbf{P}}(\kappa f_{\phi}) + \lambda \|f_{\mathcal{X}}\|_2^2.
\]

Firstly, we estimate \( S_1 \). Set

\[ S_1 = \{(y - f_{\phi}(x))^2 : f \in \mathcal{B}_R \}, \]

for any \( g \in S_1 \), there exists \( f \in \mathcal{B}_R \) such that

\[ g(z) = (y - f_{\phi}(x))^2 - (y - f_{\phi}(x))^2. \]

We can observe that

\[ E_{\mathbf{P}}(g) = \mathcal{E}(f) - \mathcal{E}(f_{\phi}) \geq 0. \]

and

\[ \frac{1}{m} \sum_{i=1}^m g(z_i) = \mathcal{E}(f) - \mathcal{E}(f_{\phi}). \]

Since \( \|\phi\| \leq \kappa \), from Cauchy–Schwarz inequality, we have

\[ |f(x)| = \left| \sum_{\alpha=1}^n \beta_\alpha \phi_\alpha x_\alpha \right| \leq \sqrt{\sum_{\alpha=1}^n \beta_\alpha^2 \sum_{\alpha=1}^n \phi_\alpha^2} \leq \kappa R. \]

Then we deduce that

\[ |g(z)| = |(y - f_{\phi}(x))^2 - (y - f_{\phi}(x))^2| \leq (\kappa R + 1)(\kappa R + 3). \]

By Lemma 3, we have that

\[
\Pr_{z \sim \mathcal{Z}} \left( \sup_{f \in \mathcal{B}_R} \frac{\mathcal{E}(f) - \mathcal{E}(f_{\phi}) - \mathcal{E}(f_{\phi})}{\mathcal{E}(f) - \mathcal{E}(f_{\phi}) + \epsilon} \geq 4 \sqrt{\epsilon} \right)
\]

\[
= \Pr_{z \sim \mathcal{Z}} \left( \sup_{f \in \mathcal{B}_R} \frac{\mathcal{E}(g(z)) - \frac{1}{m} \sum_{i=1}^m g(z_i)}{\mathcal{E}(g(z)) + \epsilon} \geq 4 \sqrt{\epsilon} \right)
\]

\[
\leq \mathcal{N}(\mathcal{G}_R, \epsilon) \exp \left( \frac{-m \epsilon^2}{56C \|f\|^2} \right). \tag{9}
\]

For any \( g_1, g_2 \in \mathcal{G}_R, z \in Z \), there exists that

\[ |g_1(z) - g_2(z)| \leq 2(\kappa R + 1)\|f_1 - f_2\|_\infty. \]

Therefore, for any \( \epsilon > 0 \), an \( \epsilon/(2(\kappa R + 1)) \)-covering of \( \mathcal{G}_R \) can provide \( \epsilon \)-covering of \( \mathcal{G}_R \). Accordingly,

\[
\mathcal{N}(\mathcal{G}_R, \epsilon) \leq \mathcal{N} \left( \kappa R + 1, \epsilon \right).
\]

From Lemma 1, we have

\[ \log \mathcal{N}(\mathcal{G}_R, \epsilon) \leq n \log \left( \frac{8R(\kappa R + 1)}{\epsilon} \right). \]

Then, (9) tells us that

\[
\Pr_{z \sim \mathcal{Z}} \left( \sup_{f \in \mathcal{B}_R} \frac{\mathcal{E}(g(z)) - \frac{1}{m} \sum_{i=1}^m g(z_i)}{\mathcal{E}(g(z)) + \epsilon} \geq 4 \sqrt{\epsilon} \right)
\]

\[
\leq \mathcal{N}(\mathcal{G}_R, \epsilon) \exp \left( \frac{-m \epsilon^2}{56C \|f\|^2} \right).
\]

Since

\[ \sqrt{\epsilon} \left( \mathcal{E}(f) - \mathcal{E}(f_{\phi}) + \epsilon \right) \leq \frac{1}{2} \left( \mathcal{E}(f) - \mathcal{E}(f_{\phi}) \right), \]

there exists that

\[
\sup_{f \in \mathcal{B}_R} \left( \mathcal{E}(f) - \mathcal{E}(f_{\phi}) \right) \leq \frac{1}{2} \left( \mathcal{E}(f) - \mathcal{E}(f_{\phi}) \right) + \epsilon.
\]

with confidence at least

\[ 1 - \exp \left( n \log \left( \frac{16R(\kappa R + 1)}{\epsilon} \right) \right). \]

Hence,

\[
\Pr_{z \sim \mathcal{Z}} \left( \sup_{f \in \mathcal{B}_R} \left( \mathcal{E}(f) - \mathcal{E}(f_{\phi}) \right) \leq \epsilon \right) \geq 1 - \exp \left( n \log \left( \frac{16R(\kappa R + 1)}{\epsilon} \right) \right) \leq \frac{1}{2} \left( \mathcal{E}(f) - \mathcal{E}(f_{\phi}) \right),
\]

then we have

\[
\mathcal{E}(f) - \mathcal{E}(f_{\phi}) \leq \mathcal{K} + 2S_2. \tag{11}
\]
For any $t \geq 16(k + \sqrt{\lambda})/m$, we conclude that
\[
E_p^m(K) = \int_0^t \text{Prob}_{x,z}^m \left\{ \mathcal{E}(f) - \mathcal{E}(f_p) - 2(\mathcal{E}(f) - \mathcal{E}(f_p)) \leq \epsilon \right\} \, de \\
\leq t + \epsilon m \left\{ \lambda m e^{-\frac{\lambda m}{e^{12(k + \sqrt{\lambda})}} (\sqrt{\lambda} + 3\sqrt{\lambda}) \|f\|^2} \right\} \\
\leq t + \epsilon m \left\{ \lambda m e^{-\frac{\lambda m}{e^{12(k + \sqrt{\lambda})}} (\sqrt{\lambda} + 3\sqrt{\lambda}) \|f\|^2} \right\} \\
\leq t + \epsilon m \left\{ \lambda m e^{-\frac{\lambda m}{e^{12(k + \sqrt{\lambda})}} (\sqrt{\lambda} + 3\sqrt{\lambda}) \|f\|^2} \right\} \\
\leq t + \beta^{-n} \text{exp} \left\{ -\frac{\lambda m t}{112(k + \sqrt{\lambda}) (3\sqrt{\lambda}) \|f\|^2} \right\} m^n t.
\]

Setting $t = 112n(k + \sqrt{\lambda}) (\sqrt{\lambda} + 3\sqrt{\lambda}) \|f\|^2 \log m/m$, we have
\[
E_p^m(K) \leq 2t = \frac{224n(k + \sqrt{\lambda}) (\sqrt{\lambda} + 3\sqrt{\lambda}) \|f\|^2 \log(m/\lambda)}{\lambda m}.
\]

Now, we give the upper bound of $E_p^m(S_2)$:
\[
E_p^m(S_2) = E_p^m \left\{ \frac{1}{m} \sum_{i=1}^m (y_i - f(X_i))^2 - \frac{1}{m} \sum_{i=1}^m (y_i - f_p(X_i))^2 + \lambda \|f\|^2 \right\} \\
= E_p^m \left\{ \inf_{f \in \mathcal{M}_p} \left\{ \frac{1}{m} \sum_{i=1}^m (y_i - f(X_i))^2 - \frac{1}{m} \sum_{i=1}^m (y_i - f_p(X_i))^2 + \lambda \|f\|^2 \right\} \right\} \\
\leq \inf_{f \in \mathcal{M}_p} \left\{ \left( \int_X (f(X) - f_p(X))^2 \, dp + \lambda \|f\|^2 \right) \right\}.
\]

The desired result follows by combining the inequations (11) and (12).

From Theorem 1, we know that the excess convex risk $E_pE_p^m(\mathcal{E}(f_{Z}) - \mathcal{E}(f_p))$ depends on the sample number $m$, the net number $n$, the regularized parameter $\lambda$, and the hypothesis space $\mathcal{M}_p$. The generalization bound for ELM with the u.e.M.c samples is consistent with the result in [10] for the i.i.d samples.

**Corollary 1.** Under the condition and notations in Theorem 1, we have
\[
E_p^m(\mathcal{R}(f_{Z}) - \mathcal{R}(f_p)) \\
\leq \sqrt{\frac{Cn \log (m/\lambda)^2}{m}} + 2E_p^m \inf_{f \in \mathcal{M}_p} \left( \int_X (f(X) - f_p(X))^2 \, dp + \lambda \|f\|^2 \right).
\]

The learning rate $O(\sqrt{n \log m/m})$ can be achieved when the optimal $\lambda$ is selected and $\inf_{f \in \mathcal{M}_p} \left( \int_X (f(X) - f_p(X))^2 \, dp + \lambda \|f\|^2 \right)$

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|}
\hline
Datasets & Attributes & Training size & Testing size \\
\hline
Waveform & 21 & 2500 & 2500 \\
Abalone & 7 & 3133 & 1044 \\
Magic & 10 & 9510 & 9510 \\
Letter & 16 & 15,000 & 5000 \\
Shuttle & 9 & 10,000 & 4500 \\
\hline
\end{tabular}
\caption{Specifications of datasets.}
\end{table}

4. Empirical evaluations

To better verify the theoretical analysis of ELM with the Markov chains, we evaluate its performance on some datasets. Here, we generate the u.e.M.c samples by the sampling algorithm in Table 1. In fact, this sampling algorithm has been used for learning algorithms in [27,21,22].

The UCI datasets are used to evaluate ELM and their characteristics are summarized in Table 2. The experiment can be divided into three steps: firstly, the training set $Z$ with $m$ samples is

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|}
\hline
Datasets & Attributes & Training size & Testing size \\
\hline
Waveform & 21 & 2500 & 2500 \\
Abalone & 7 & 3133 & 1044 \\
Magic & 10 & 9510 & 9510 \\
Letter & 16 & 15,000 & 5000 \\
Shuttle & 9 & 10,000 & 4500 \\
\hline
\end{tabular}
\caption{Misclassification rate (MR) for 1000 training samples.}
\end{table}

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|}
\hline
Datasets & MR(i.i.d.) & MR(Markov) \\
\hline
Waveform & 0.1410 ± 0.0076 & 0.1364 ± 0.0062 \\
Abalone & 0.2068 ± 0.0041 & 0.2087 ± 0.0030 \\
Magic & 0.1956 ± 0.0069 & 0.1946 ± 0.0053 \\
Letter & 0.1791 ± 0.0078 & 0.1779 ± 0.0072 \\
Shuttle & 0.0141 ± 0.0193 & 0.0114 ± 0.0028 \\
\hline
\end{tabular}
\caption{Misclassification rate (MR) for 1500 training samples.}
\end{table}
generated by the Markov sampling algorithm in [21,22]; secondly, we consider the combination of the square function and Gaussian function as the activation function of ELM [15]; finally, we train ELM on $Z$ and evaluate its performance on the test set.

We conduct the experiment for 50 times and the average misclassification rates are presented in Tables 3 and 4. The results tell us that ELM with Markov sampling can provide the competitive prediction according to the misclassification rates and standard deviations. We also evaluate the ELM with Markov sampling for different numbers of training samples in Fig. 1, which shown that the misclassification rate will decrease with the increasing training samples. This empirical result is consistent with the theoretical analysis in Theorem 1.

In order to better understand the efficiency of ELM, we also present several experiments to compare the standard deviations with the independent and Markov sampling methods. Figs. 2, 3, 4, and 5 report these experimental results for different training numbers on Waveform, Abalone, Magic, and Shuttle datasets. From these figures, we can find that the standard deviations of ELM with Markov sampling are usually smaller than ELM with i.i.d samples. That is to say, the Markov sampling usually can improve the stability of ELM with i.i.d samples.

![Fig. 1. Misclassification rates for Waveform, Abalone, Magic, and Letter with different training samples.](image1)

![Fig. 2. Standard deviations for Waveform with $m = 1000$ (left) and $m = 1500$ (right).](image2)

![Fig. 3. Standard deviations for Abalone with $m = 1000$ (left) and $m = 1500$ (right).](image3)
5. Conclusion

In this paper, we have investigated the generalization ability of ELM with the Markov chain samples. The generalization bound of ELM has been established and some empirical evaluations have been provided. In particular, the learning rate derived here is the same as the previous work for the i.i.d samples. Along the line of the present work, some subjects deserve to further study, e.g., the generalization bounds of semi-supervised and unsupervised ELMs in [4] and the generalization analysis for the sparse ELMs in [1].

Acknowledgments

This work was supported in part by the Fundamental Research Funds for the Central Universities (Program no. 2014PY025, 2662015PY046), the National Natural Science Foundation of China (61370002, 61403132), the Macau Science and Technology Development Fund (FDCT) under Grant 106/2013/A3, and by the Research Committee at University of Macau under Grants MYRG2014-00003-FST, MRG017/ZYC/2014/FST, MYRG113(Y1-L3)-FST12-ZYC and MRG001/ZYC/2013/FST.

Appendix

Proof of Lemma 3. Firstly, denote \( J = \mathcal{N}(\mathcal{G}, \epsilon) \) and consider \( |D_j| = 1 \) as a cover of \( \mathcal{G} \). Here, the balls \( D_j = \{ g \in \mathcal{G} : \| g - g_j \| \leq \epsilon \} \). Then, for any \( g \in \mathcal{G} \), there is \( g_j \) such that \( \| g - g_j \| \leq \epsilon \). Therefore, we have

\[
|E(g) - E(g_j)| \leq \| g - g_j \| \leq \epsilon
\]

and

\[
\frac{1}{m} \sum_{i=1}^{m} g(z_i) - \frac{1}{m} \sum_{i=1}^{m} g_j(z_i) \leq \| g - g_j \| \leq \epsilon.
\]

Then, there exists

\[
|E(g) - E(g_j)|/\sqrt{E(g) + \epsilon} \leq \epsilon
\]

and

\[
\frac{1}{m} \sum_{i=1}^{m} g(z_i) - \frac{1}{m} \sum_{i=1}^{m} g_j(z_i) \sqrt{E(g) + \epsilon} \leq \epsilon.
\]

Fig. 4. Standard deviations for Magic with \( m = 1000 \) (left) and \( m = 1500 \) (right).

Fig. 5. Standard deviations for Shuttle with \( m = 1000 \) (left) and \( m = 1500 \) (right).
Secondly, from Lemma 2, for any $\epsilon > 0$, we have
\[
\mathbb{P}\left( E(g) - E(f) \geq \epsilon \right) \leq \exp\left( -\frac{m\epsilon^2}{\max_{i,j} (\|g_i - g_j\|^2)} \right).
\] (13)

Then for any $g_i \in G$, we have
\[
\mathbb{P}\left( E(g_i) - E(f) \geq \epsilon \right) \leq \exp\left( -\frac{m\epsilon^2}{\max_{i,j} (\|g_i - g_j\|^2)} \right).
\]

Lastly, for $E(g_i) - E(g) \leq \epsilon$, we have $\sqrt{E(g_i) + \epsilon} \leq 2\sqrt{E(g) + \epsilon}$.

Therefore, there holds
\[
\mathbb{P}\left( \sup_{g \in G} E(g) - E(f) \geq \epsilon \right) \leq \sum_{i=1}^{n} \mathbb{P}\left( E(g_i) - E(f) \geq \epsilon \right) \leq n(\epsilon, 2\sqrt{E(g) + \epsilon}) \exp\left( -\frac{m\epsilon^2}{\max_{i,j} (\|g_i - g_j\|^2)} \right).
\]

References


