Convergence of Gradient Descent for Minimum Error Entropy Principle in Linear Regression

Ting Hu, Qiang Wu, and Ding-Xuan Zhou

***Abstract*—We study the convergence of minimum error entropy (MEE) algorithms when they are implemented by a gradient descent. This method has been used in practical applications for more than one decade, but there has been no consistency or rigorous error analysis. This paper gives the first rigorous proof for the convergence of the gradient descent method for MEE in a linear regression setting. The mean square error is proved to decay exponentially fast in terms of the iteration steps and of order $O(\frac{1}{m})$ in terms of the sample size m. The mean square convergence is guaranteed when the step size is chosen appropriately and the scaling parameter is large enough.**

***Index Terms*—Minimum error entropy, error information, gradient descent method, error analysis, global convergence.**

I. Introduction

Regression analysis plays important roles in many fields of science and engineering. The traditional least square method is the mostly used algorithm for regression in practice. However, it is suboptimal when the system noise is not normally distributed. Variant approaches have been proposed to deal with data with outliers or heavy-tailed distributions. Minimum error entropy (MEE) criterion is one of them. It is motivated by the idea of minimizing the information as measured by entropy in the prediction error. The estimated model is expected to preserve information as much as possible and thus improves the predictive performance. Unlike the traditional least square method which relies only on the variance of the prediction error, the error entropy takes all higher order moments into account and is thus advantageous when MEE is used to handle non-Guassian and heavy tailed error distributions [1], [2]. As non-Gaussian noise is ubiquitous in real world applications, the superiority of MEE has been evidenced in a variety of applications, which include adaptive filtering, clustering, classification, feature selection, and blind source separation [3]–[8].

Manuscript received November 12, 2015; revised April 15, 2016 and June 29, 2016; accepted August 29, 2016. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Dmitry M. Malioutov. The work was supported in part by the National Natural Science Foundation of China under Grants 11671307, 11501078, 11671171, 11461161006, and 11471292, in part by the U.S. Department of Agriculture National Institute of Food and Agriculture under Grant 2016-70001-24636, and in part by the Research Grants Council of Hong Kong (Project no. CityU 11303915).

T. Hu is with the School of Mathematics and Statistics, Wuhan University, Wuhan 430072, China (e-mail: tinghu@whu.edu.cn).

Q. Wu is with the Department of Mathematical Sciences and the Computational Science Ph.D. Program, Middle Tennessee State University, Murfreesboro, TN 37132 USA (e-mail: qwu@mtsu.edu).

D.-X. Zhou is with the Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong (e-mail: mazhou@cityu.edu.hk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2016.2612169

Let X be a multivariate random variable with values in a compact subset of $\mathbb{R}^n$ and Y a real valued response variable. The purpose of regression analysis is to study the quantitative relationship between X and Y. This usually leads to estimating the regression function $f_*(\mathbf{x}) = \mathbf{E}(Y|X=\mathbf{x})$ from a sample of m observations $\mathbf{z} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ drawn independently and identically. As most statistical and machine learning algorithms for regression analysis have focused on the use of convex losses such as the squared loss in the least square method and the insensitive loss in support vector regression, approximation powers of learning algorithms with convex losses have been well studied in the literature; see e.g. [9]–[11] and the references therein. The MEE algorithms, however, use the error entropy as the loss function which is not convex. It brings essential difficulties to the analysis. Although the MEE algorithms have been verified effective in many empirical studies, the study on its computational and mathematical properties is lagged a little bit behind.

The MEE approach was introduced in [1]. It aims to minimize the information contained in the error and maximize the information captured by the estimated model. Given an estimator f of the regression function, define the error variable as $E = Y - f(X)$. One can measure the error information by Renyi's entropy or Shannon's entropy. In this paper we consider the second order Renyi's entropy

$$H(E) = -\log \mathbf{E}(p_{_E}) = -\log \int p_{_E}^2(e)de$$

where $p_{_E}$ denotes the probability density function of E. For the given sample $\mathbf{z}$, define $e_i = y_i - f(\mathbf{x}_i)$. Then $p_{_E}$ can be estimated by Parzen windowing [12] which, given a kernel function $K : \mathbb{R} \to [0, \infty)$ and a scaling parameter $h > 0$, defines a kernel density estimator by

$$\hat{p}_{_E}(e) = \frac{1}{m}\sum_{i=1}^m K_h(e - e_i) = \frac{1}{mh}\sum_{i=1}^m K\left(\frac{e - e_i}{h}\right).$$

A usual choice is the Gaussian kernel density estimator where $K(u) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{u^2}{2})$ and $K_h(u) = \frac{1}{\sqrt{2\pi}h}\exp(-\frac{u^2}{2h^2})$. The empirical error information is

$$\hat{H}(f) = -\log\left\{\frac{1}{m^2 h}\sum_{i=1}^m\sum_{j=1}^m K\left(\frac{e_i - e_j}{h}\right)\right\}.$$

The MEE algorithm searches for an estimator that minimizes $\hat{H}$ over a hypothesis space.

The structure of the empirical entropy $\hat{H}$ exhibits that the scaling parameter h plays an important role in the MEE algorithm design. The value of h is adjusted for different learning tasks in MEE algorithms and the corresponding learning effects are presented in a series of numerical simulations; see e.g. [6], [7]. Mathematically, the predictive performance of MEE algorithms was analyzed in [13]–[16]. The convergence of MEE algorithms can be guaranteed only for homoscedastic model if the scaling parameter h is chosen small. The scaling parameter h should be chosen large enough to guarantee the algorithms to be asymptotically consistent for more general models. This coincides with the empirical studies in the literature.

From a computational perspective, the loss function is close to the squared loss by weighing less on the high order statistics of the error when h is large. Thus, using a relatively large scaling parameter reduces the risk that MEE algorithms suffer from being stuck in local minima. MEE algorithms are usually implemented by gradient descent or stochastic gradient descent [1], [17]–[19]. However, because the optimization problem arising from MEE is non-convex, the convergence of the gradient descent method is not unconditionally guaranteed. A mean squared convergence result is proved in [20] which, however, only guarantees the solution of the stochastic gradient descent method converges to a local minima but not necessarily the global minima. In this paper, our purpose is to derive conditions and stopping criteria for the gradient descent method to achieve global convergence.

We focus on linear regression models in this paper. Assume

$$y = \mathbf{w}_*^\top \mathbf{x} + \epsilon, \quad \mathbf{E}[\epsilon|\mathbf{x}] = 0$$

for some $\mathbf{w}_* \in \mathbb{R}^n$, where ϵ is a mean zero noise random variable. The regression function takes the form $f_*(\mathbf{x}) = \mathbf{w}_*^\top \mathbf{x}$ and the target of regression analysis is to estimate $\mathbf{w}_*$ from the sample. For an estimator $\hat{\mathbf{w}}$, the goodness could be measured by the squared error $\|\hat{\mathbf{w}} - \mathbf{w}_*\|^2$.

The MEE estimator $\hat{\mathbf{w}}$ is defined as

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w} \in \mathbb{R}^n} \hat{H}(\mathbf{w})$$

where, given $e_i = y_i - \mathbf{w}^\top \mathbf{x}_i$,

$$\hat{H}(\mathbf{w}) = -\log \left\{ \frac{1}{m^2 h} \sum_{i=1}^m \sum_{j=1}^m K\left(\frac{e_i - e_j}{h}\right) \right\}.$$

As the logarithmic function is monotone and does not affect the minimization process, we remove it and consider the transformed empirical error information

$$R(\mathbf{w}) = -\frac{h^2}{m^2} \sum_{i=1}^m \sum_{j=1}^m K\left(\frac{(y_i - \mathbf{w}^\top \mathbf{x}_i) - (y_j - \mathbf{w}^\top \mathbf{x}_j)}{h}\right).$$

It is obvious the MEE estimator can also be obtained by

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w} \in \mathbb{R}^n} R(\mathbf{w}). \tag{1}$$

When K is differentiable, the gradient descent algorithm for MEE starts with $\hat{\mathbf{w}}_0 = 0$ and updates the estimator by

$$\hat{\mathbf{w}}_t = \hat{\mathbf{w}}_{t-1} - \eta_t \nabla R(\hat{\mathbf{w}}_{t-1})$$

in the t-th step, where ∇ is the gradient operator and $\eta_t > 0$ is the step size. When this method is used to solve the MEE estimator (1), the first question might be the convergence of $\hat{\mathbf{w}}_t$ to $\hat{\mathbf{w}}$ as the number of iterations becomes large. However, we would consider the problem in an alternative way. Recall the ultimate goal is to learn the true regression coefficients vector $\mathbf{w}_*$. On one hand, if $\hat{\mathbf{w}}_t$ provide good estimates of $\mathbf{w}_*$, the convergence of $\hat{\mathbf{w}}_t$ to $\hat{\mathbf{w}}$ does not matter much. On the other hand, notice that

$$\|\hat{\mathbf{w}}_t - \mathbf{w}_*\| \le \|\hat{\mathbf{w}}_t - \hat{\mathbf{w}}\| + \|\hat{\mathbf{w}} - \mathbf{w}_*\|.$$

Even if $\hat{\mathbf{w}}_t$ does converge to $\hat{\mathbf{w}}$, it does not make much sense to iterate the gradient descent steps till convergence because the second term on the right will dominate the error. Instead, the algorithm should be stopped earlier when the performance of the estimate does not improve.

In order to state our main results we need some assumptions. Firstly, we assume both X and Y are uniformly bounded by a constant M. Also, the covariance matrix V_{XX} of X is non-degenerate, that is, all the eigenvalues of V_{XX} are positive. In particular, we denote by $\lambda_{\max}$ and $\lambda_{\min}$ the largest and the smallest eigenvalues of V_{XX}, respectively.

To simplify our presentation and notations in the proofs, we focus on symmetric kernels and define $\Psi : [0, \infty) \to [0, \infty)$ as $\Psi(u) = K(\sqrt{2u})$ or equivalently, $\Psi(\frac{u^2}{2}) = K(u)$. With this notation, the empirical error can be rewritten as

$$R(\mathbf{w}) = -\frac{h^2}{m^2} \sum_{i=1}^m \sum_{j=1}^m \Psi\left(\frac{[(y_i - \mathbf{w}^\top \mathbf{x}_i) - (y_j - \mathbf{w}^\top \mathbf{x}_j)]^2}{2h^2}\right).$$

Assume Ψ is decreasing and differentiable, $c_0 = -\Psi'_+(0) > 0$, and for some $p > 0$,

$$|\Psi'(u) - \Psi'_+(0)| \le c_p u^p, \ \forall\, u > 0. \tag{2}$$

When the Gaussian kernel is used, it is easy to verify that $\Psi(u) = \frac{1}{\sqrt{2\pi}} \exp(-u)$. We have $c_0 = \frac{1}{\sqrt{2\pi}}$ and (2) holds with $p = 1$ and $c_p = \frac{1}{\sqrt{2\pi}}$.

Our first result, Theorem 1 below, shows that $\hat{\mathbf{w}}_t$ is uniformly bounded with large probability.

Theorem 1: If $0 < \eta_t \le \frac{1}{2c_0\lambda_{\max}}$ for all $t \in \mathbb{N}$ and $h \ge \left(\frac{2^{5p+4} c_p M^{6p+2}}{c_0 \lambda_{\min}^{2p+1}}\right)^{1/2p}$, then for any $0 < \delta < 1$, we have

$$\|\hat{\mathbf{w}}_t\| \le \frac{3M^2}{\lambda_{\min}} \qquad \text{for all } t \in \mathbb{N}$$

with probability $1 - \delta$ provided that $m \ge \frac{900 M^4 \log(8/\delta)}{\lambda_{\min}^2}$.

Because any bounded closed set in $\mathbb{R}^n$ is compact, Theorem 1 guarantees that a subsequence of $\{\hat{\mathbf{w}}_t\}$ converges to some point. To ensure the accumulation point is the solution $\mathbf{w}_*$ as we expected, the step size and the scaling parameters should be selected appropriately.

Theorem 2: Let $\eta_t = \eta t^{-\theta}$ for some $0 \le \theta < 1$ and $0 < \eta \le \frac{\lambda_{\min}}{12 c_0 \lambda_{\max}^2}$. Let $h \ge \left(\frac{2^{5p+4} c_p M^{6p+2}}{c_0 \lambda_{\min}^{2p+1}}\right)^{1/2p}$. For any $0 < \delta < 1$, we have

$$\|\hat{\mathbf{w}}_T - \mathbf{w}_*\|^2 \le C' \left\{ \exp\left(-\frac{\eta c_0 \lambda_{\min} T^{1-\theta}}{1-\theta}\right) + \frac{1}{h^{4p}} + \frac{\log(8/\delta)}{m} \right\}$$

with probability $1 - \delta$ provided that $m \ge \frac{900 M^4 \log(8/\delta)}{\lambda_{\min}^2}$. Here the constant C' is independent of m, h, or δ, and will be given explicitly in the proof.

Theorem 2 indicates that, under appropriate choices of the parameters, $\hat{\mathbf{w}}_t$ converges to $\mathbf{w}_*$ exponentially fast in terms of the number of iterations and is of order $O(\frac{1}{m})$ in terms of the sample size. In particular, the convergence holds with a fixed step size $\eta_t = \eta$ provided that η is small enough. In practice, given a set of observations, the sample size m is fixed. The number of iteration steps $T = O(\log m)$ is usually sufficient to achieve the best possible learning performance.

II. Preliminaries

We first give several basic facts associated to the linear regression model. Throughout this section we denote $\mu_X = \mathbf{E}(X)$ and $\mu_Y = \mathbf{E}(Y)$.

Lemma 3: The covariance matrix V_{XX} satisfies $\lambda_{\max} = \|V_{XX}\| \le M^2$.

Proof: Note that $V_{XX} = \mathbf{E}(XX^\top) - \mu_X \mu_X^\top$. Since X is bounded by M, we have $\|\mathbf{E}(XX^\top)\| \le M^2$. Since both $\mathbf{E}(XX^\top)$ and $\mu_X\mu_X^\top$ are positive semidefinite, we have

$$\|V_{XX}\| \le \|\mathbf{E}(XX^\top)\| \le M^2.$$

This proves the conclusion. ■

Lemma 4: Let V_{XY} denote the covariance vector between X and Y. We have $V_{XX}\mathbf{w}_* = V_{XY}$ and $\|\mathbf{w}_*\| \le \frac{2M^2}{\lambda_{\min}}$.

Proof: By the model assumption we have $\mu_Y = \mu_X^\top \mathbf{w}_*$. Therefore, $y - \mu_Y = (\mathbf{x} - \mu_X)^\top \mathbf{w}_* + \epsilon$ and

$$(y - \mu_Y)(\mathbf{x} - \mu_X) = (\mathbf{x} - \mu_X)(\mathbf{x} - \mu_X)^\top \mathbf{w}_* + \epsilon(\mathbf{x} - \mu_X).$$

Taking expectation both sides and noting the fact $\mathbf{E}(\epsilon|\mathbf{x}) = 0$, we obtain $V_{XY} = V_{XX}\mathbf{w}_*$.

Since both X and Y are bounded by M, we have

$$\|V_{XY}\| = \|\mathbf{E}(XY) - \mu_X\mu_Y\| \le 2M^2.$$

Thus

$$\|\mathbf{w}_*\| = \|V_{XX}^{-1}V_{XY}\| \le \frac{2M^2}{\lambda_{\min}}.$$

This finishes the proof. ■

In our analysis, we need to deal with matrix and vector valued functions. For this purpose we need probability inequalities for Hilbert space valued random variables. The following one can be found in [21].

Lemma 5: Let $\mathcal{H}$ be a Hilbert space and $\{\xi_i\}_{i=1}^m$ be m independent random variables with values in $\mathcal{H}$. Suppose that for each i, $\|\xi_i\| \le M$ almost surely. Denote $\sigma^2 = \sum_{i=1}^m \mathbf{E}(\|\xi_i\|^2)$. Then, for any $\varepsilon > 0$,

$$\Pr\left\{\left\|\frac{1}{m}\sum_{i=1}^m [\xi_i - \mathbf{E}(\xi_i)]\right\| \ge \varepsilon\right\} \le 2\exp\left\{-\frac{m\varepsilon}{2M}\log\left(1 + \frac{mM\varepsilon}{\sigma^2}\right)\right\}.$$

By this lemma, we can prove the following inequality.

Lemma 6: Let $\mathcal{H}$ be a Hilbert space and ξ be a random variable with values in $\mathcal{H}$. Assume that $\|\xi\| \le M$ almost surely. Let $\{\xi_1, \xi_2, \ldots, \xi_m\}$ be a sample of m independent observations for ξ. Then, for any $\varepsilon > 0$,

$$\Pr\left\{\left\|\frac{1}{m}\sum_{i=1}^m \xi_i - \mathbf{E}(\xi)\right\| \ge \varepsilon\right\} \le 2\exp\left\{-\frac{m\varepsilon^2}{2M^2 + M\varepsilon}\right\}. \tag{3}$$

Proof: Since $\|\xi\| \le M$ almost surely, we have

$$\sigma^2 = \sum_{i=1}^m \mathbf{E}(\|\xi_i\|^2) = m\mathbf{E}(\|\xi\|^2) \le mM^2.$$

Applying Lemma 5 we obtain

$$\Pr\left\{\left\|\frac{1}{m}\sum_{i=1}^m \xi_i - \mathbf{E}(\xi)\right\| \ge \varepsilon\right\} \le 2\exp\left\{-\frac{m\varepsilon}{2M}\log\left(1 + \frac{\varepsilon}{M}\right)\right\}. \tag{4}$$

By the elementary inequality $\log(1+t) > \frac{2t}{2+t}$ for $t > 0$, we have

$$\frac{\varepsilon}{M}\log\left(1 + \frac{\varepsilon}{M}\right) \ge \frac{2\varepsilon^2}{2M^2 + M\varepsilon}.$$

Plugging this into (4) gives (3). ■

Lemma 7: Let $\mathcal{H}$ be a Hilbert space and ξ be a random variable with values in $\mathcal{H}$. Assume that $\|\xi\| \le M$ almost surely. Let $\{\xi_1, \xi_2, \ldots, \xi_m\}$ be a sample of m independent observations for ξ. Then, for any $0 < \tilde{\delta} < 1$, we have with confidence $1 - \tilde{\delta}$,

$$\left\|\frac{1}{m}\sum_{i=1}^m \xi_i - \mathbf{E}(\xi)\right\| \le \frac{1}{2}M\left(\tau + \sqrt{8\tau + \tau^2}\right)$$

where $\tau = \frac{\log(2/\tilde{\delta})}{m}$.

Using this lemma we can prove the following estimate.

Lemma 8: For any $0 < \delta < 1$, with probability at least $1 - \delta$, we have

$$\left\|\frac{1}{m^2}\sum_{i=1}^m\sum_{j=1}^m (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top - 2V_{XX}\right\| \le 10M^2\sqrt{\tau} \tag{5}$$

and

$$\left\|\frac{1}{m^2}\sum_{i=1}^m\sum_{j=1}^m (y_i - y_j)(\mathbf{x}_i - \mathbf{x}_j) - 2V_{XY}\right\| \le 12M^2\sqrt{\tau} \tag{6}$$

simultaneously, where $\tau = \frac{\log(8/\delta)}{m}$.

Proof: Let $\bar{\mathbf{x}} = \frac{1}{m}\sum_{i=1}^m \mathbf{x}_i$ and $\bar{y} = \frac{1}{m}\sum_{i=1}^m y_i$ be the corresponding sample means of X and Y.

Applying Lemma 7 to $\xi = X$ with $\tilde{\delta} = \frac{\delta}{4}$, we obtain

$$\left\|\frac{1}{m}\sum_{i=1}^m \mathbf{x}_i - \mu_X\right\| \le \frac{1}{2}M\left(\tau + \sqrt{8\tau + \tau^2}\right) \tag{7}$$

with probability at least $1 - \frac{\delta}{4}$. Applying Lemma 7 to $\xi = Y$ with $\tilde{\delta} = \frac{\delta}{4}$, we obtain

$$\left|\frac{1}{m}\sum_{i=1}^m y_i - \mu_Y\right| \le \frac{1}{2}M\left(\tau + \sqrt{8\tau + \tau^2}\right) \tag{8}$$

with probability at least $1-\frac{\delta}{4}$. Recall that all $n\times n$ matrices form a Hilbert space under the Frobenius norm. Consider the matrix valued random variable $\xi = XX^\top$ which satisfies $\|\xi\|_F = \|X\|^2 \le M^2$. Applying Lemma 7 with $\tilde{\delta} = \frac{\delta}{4}$, we obtain

$$\left\| \frac{1}{m}\sum_{i=1}^m \mathbf{x}_i\mathbf{x}_i^\top - \mathbf{E}(XX^\top)\right\|_F \le \frac{1}{2}M^2\left(\tau + \sqrt{8\tau+\tau^2}\right)$$

with probability at least $1-\frac{\delta}{4}$. Since the operator norm is bounded by the Frobenius norm, we have

$$\left\| \frac{1}{m}\sum_{i=1}^m \mathbf{x}_i\mathbf{x}_i^\top - \mathbf{E}(XX^\top)\right\| \le \frac{1}{2}M^2\left(\tau + \sqrt{8\tau+\tau^2}\right) \quad (9)$$

with probability at least $1-\frac{\delta}{4}$. Applying Lemma 7 to $\xi = XY$ with $\tilde{\delta}=\frac{\delta}{4}$, we obtain

$$\left\| \frac{1}{m}\sum_{i=1}^m \mathbf{x}_i y_i - \mathbf{E}(XY)\right\| \le \frac{1}{2}M^2\left(\tau + \sqrt{8\tau+\tau^2}\right) \quad (10)$$

with probability at least $1-\frac{\delta}{4}$. Thus, (7)–(10) hold simultaneously with probability at least $1-\delta$. (We have used the fact that for a sequence of k events $A_1, A_2, \ldots, A_k$, $\Pr(\bigcap_{i=1}^k A_i) = \Pr((\bigcup_{i=1}^k A_i^{\mathsf{c}})^{\mathsf{c}}) \ge 1 - \sum_{i=1}^k \Pr(A_i^{\mathsf{c}})$.) What is left is to verify (5) and (6) from (7)–(10).

Let us first prove (5). Note that

$$\frac{1}{m^2}\sum_{i=1}^m\sum_{j=1}^m (\mathbf{x}_i-\mathbf{x}_j)(\mathbf{x}_i-\mathbf{x}_j)^\top = \frac{2}{m}\sum_{i=1}^m \mathbf{x}_i\mathbf{x}_i^\top - 2\bar{\mathbf{x}}\bar{\mathbf{x}}^\top.$$

Both terms on the right hand side are positive semidefinite matrices and their norms are no greater than $2M^2$. Thus,

$$\left\|\frac{1}{m^2}\sum_{i=1}^m\sum_{j=1}^m (\mathbf{x}_i-\mathbf{x}_j)(\mathbf{x}_i-\mathbf{x}_j)^\top\right\| \le 2M^2.$$

This, together with Lemma 3, implies

$$\left\|\frac{1}{m^2}\sum_{i=1}^m\sum_{j=1}^m (\mathbf{x}_i-\mathbf{x}_j)(\mathbf{x}_i-\mathbf{x}_j)^\top - 2V_{XX}\right\| \le 2M^2.$$

So (5) holds almost surely if $\tau > \frac{1}{25}$. When $\tau \le \frac{1}{25}$, by (7) and (9), we obtain

$$\begin{aligned}
&\left\|\frac{1}{m^2}\sum_{i=1}^m\sum_{j=1}^m (\mathbf{x}_i-\mathbf{x}_j)(\mathbf{x}_i-\mathbf{x}_j)^\top - 2V_{XX}\right\| \\
&\le 2\left\|\frac{1}{m}\sum_{i=1}^m \mathbf{x}_i\mathbf{x}_i^\top - \mathbf{E}(XX^\top)\right\| + 2\left\|\bar{\mathbf{x}}\bar{\mathbf{x}}^\top - \mu_X\mu_X^\top\right\| \\
&\le M^2\left(\tau+\sqrt{8\tau+\tau^2}\right) + 4M\|\bar{\mathbf{x}}-\mu_X\| \\
&\le 3M^2\left(\tau+\sqrt{8\tau+\tau^2}\right) \\
&\le 3M^2\sqrt{\tau}\left(\sqrt{\frac{1}{25}} + \sqrt{8+\frac{1}{25}}\right) \\
&\le 10M^2\sqrt{\tau}.
\end{aligned}$$

This proves (5).

Now we turn to (6). The proof is quite similar. First note that the left hand side is bounded by $8M^2$ almost surely. So the inequality is always true when $\tau > 1$. When $\tau \le 1$, we need the fact that

$$\frac{1}{m^2}\sum_{i=1}^m\sum_{j=1}^m (y_i-y_j)(\mathbf{x}_i-\mathbf{x}_j) = \frac{2}{m}\sum_{i=1}^m y_i\mathbf{x}_i - 2\bar{\mathbf{x}}\bar{y}.$$

By (7), (8) and (10), we obtain

$$\begin{aligned}
&\left\|\frac{1}{m^2}\sum_{i=1}^m\sum_{j=1}^m (y_i-y_j)(\mathbf{x}_i-\mathbf{x}_j) - 2V_{XY}\right\| \\
&\le 2\left\|\frac{1}{m}\sum_{i=1}^m y_i\mathbf{x}_i - \mathbf{E}(XY)\right\| \\
&\quad + 2M\Big(\|\bar{\mathbf{x}}-\mu_X\| + |\bar{y}-\mu_Y|\Big) \\
&\le 3M^2\left(\tau+\sqrt{8\tau+\tau^2}\right) \\
&\le 12M^2\sqrt{\tau}.
\end{aligned}$$

We finish the proof. ■

According to Lemma 8, we will adopt the notations

$$\hat{V}_{XX} = \frac{1}{2m^2}\sum_{i=1}^m\sum_{j=1}^m (\mathbf{x}_i-\mathbf{x}_j)(\mathbf{x}_i-\mathbf{x}_j)^\top$$

and

$$\hat{V}_{XY} = \frac{1}{2m^2}\sum_{i=1}^m\sum_{j=1}^m (y_i-y_j)(\mathbf{x}_i-\mathbf{x}_j)$$

because they provide sample estimates of the covariance matrix V_{XX} and the covariance vector V_{XY}, respectively.

III. Uniform Bound for the Solution Path

In this section, we prove Theorem 1 which states that $\hat{\mathbf{w}}_t$ are uniformly bounded with large probability.

To simplify our presentation, we adopt the notation

$$\zeta_t(i,j) = (y_i - \hat{\mathbf{w}}_t^\top\mathbf{x}_i) - (y_j - \hat{\mathbf{w}}_t^\top\mathbf{x}_j)$$

for each $t\in\mathbb{N}$ in the sequel. The following proposition gives conditions for the solution $\hat{\mathbf{w}}_t$ to be uniformly bounded.

Proposition 9: Let $0 < \eta_t \le \frac{1}{2c_0\lambda_{\max}}$ for all $t \ge 1$ and h is chosen such that

$$h \ge \left(\frac{2^{5p+4}c_pM^{6p+2}}{c_0\lambda_{\min}^{2p+1}}\right)^{1/2p}. \quad (11)$$

If the sample $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$ satisfies

$$\|\hat{V}_{XX} - V_{XX}\| \le \frac{1}{6}\lambda_{\min}, \quad (12)$$

then

$$\|\hat{\mathbf{w}}_t\| \le \frac{3M^2}{\lambda_{\min}}.$$

Proof: By the definition of $\hat{V}_{XX}$ and $\hat{V}_{XY}$ and the fact $\Psi'_+(0) = -c_0$, we can write

$$\begin{aligned}
&\nabla R(\hat{\mathbf{w}}_{t-1}) \\
&= \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\Psi'\left(\frac{\zeta_{t-1}^2(i,j)}{2h^2}\right)\zeta_{t-1}(i,j)(\mathbf{x}_i - \mathbf{x}_j) \\
&= \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\left[\Psi'\left(\frac{\zeta_{t-1}^2(i,j)}{2h^2}\right) - \Psi'_+(0)\right]\zeta_{t-1}(i,j)(\mathbf{x}_i - \mathbf{x}_j) \\
&\quad - 2c_0\hat{V}_{XY} + 2c_0\left(\hat{V}_{XX} - V_{XX}\right)\hat{\mathbf{w}}_{t-1} + 2c_0 V_{XX}\hat{\mathbf{w}}_{t-1} \\
&:= Q_1 + Q_2 + Q_3 + 2c_0 V_{XX}\hat{\mathbf{w}}_{t-1}.
\end{aligned}$$

We prove the conclusion by induction. First it is obvious $\|\hat{\mathbf{w}}_0\| = 0 \le \frac{3M^2}{\lambda_{\min}}$. Assume $\|\hat{\mathbf{w}}_{t-1}\| \le \frac{3M^2}{\lambda_{\min}}$. We need to prove $\|\hat{\mathbf{w}}_t\| \le \frac{3M^2}{\lambda_{\min}}$.

By the definition of $\hat{\mathbf{w}}_t$, we have

$$\begin{aligned}
\hat{\mathbf{w}}_t &= \hat{\mathbf{w}}_{t-1} - \eta_t \nabla R(\hat{\mathbf{w}}_{t-1}) \\
&= (I - 2\eta_t c_0 V_{XX})\hat{\mathbf{w}}_{t-1} - \eta_t(Q_1 + Q_2 + Q_3).
\end{aligned}$$

Since $\eta_t \le \frac{1}{2c_0\lambda_{\max}}$, the matrix $I - 2\eta_t c_0 V_{XX}$ is positive semidefinite. We have

$$\|(I - 2\eta_t c_0 V_{XX})\hat{\mathbf{w}}_{t-1}\| \le (1 - 2\eta_t c_0 \lambda_{\min})\frac{3M^2}{\lambda_{\min}}.$$

Since X and Y are bounded by M almost surely and $\|\hat{\mathbf{w}}_{t-1}\| \le \frac{3M^2}{\lambda_{\min}}$, we have

$$\begin{aligned}
\|\zeta_{t-1}(i,j)\| &\le 2M(1 + \|\hat{\mathbf{w}}_{t-1}\|) \\
&\le 2M\left(1 + \frac{3M^2}{\lambda_{\min}}\right) \le \frac{8M^3}{\lambda_{\min}},
\end{aligned}$$

where we have used the fact $\lambda_{\min} \le \lambda_{\max} \le M^2$. This together with the Lipschitz assumption on Ψ' gives

$$\begin{aligned}
\|Q_1\| &\le \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} c_p\left(\frac{|\zeta_{t-1}(i,j)|^2}{2h^2}\right)^p |\zeta_{t-1}(i,j)|(2M) \\
&\le 2^{5p+4}c_p M^{6p+4}\lambda_{\min}^{-2p-1}h^{-2p}.
\end{aligned}$$

Under the condition (11), we have $\|Q_1\| \le c_0 M^2$. It is easy to verify $\|Q_2\| \le 4c_0 M^2$. As for Q_3, under the condition (12), we have $\|Q_3\| \le c_0 M^2$. Therefore, we have

$$\|\hat{\mathbf{w}}_t\| \le (1 - 2\eta_t c_0 \lambda_{\min})\frac{3M^2}{\lambda_{\min}} + 6\eta_t c_0 M^2 \le \frac{3M^2}{\lambda_{\min}}.$$

This finishes the proof. ∎

Now Theorem 1 can be proved by combining Proposition 9 and Lemma 8.

Proof of Theorem 1: By Lemma 8,

$$\|\hat{V}_{XX} - V_{XX}\| \le 5M^2\sqrt{\frac{\log(8/\delta)}{m}}$$

with probability $1-\delta$. Thus, when $m \ge \frac{900M^4\log(8/\delta)}{\lambda_{\min}^2}$, the condition (12) holds with probability at least $1-\delta$. By Proposition 9, we obtain the desired conclusion. ∎

IV. One Step Error Analysis

In this section we show that the estimation error decreases after each iteration step, which plays an essential role for the proof of Theorem 2.

Proposition 10: Let $0 < \eta_t \le \frac{\lambda_{\min}}{12c_0\lambda_{\max}^2}$ for all $t \ge 1$ and $h \ge \left(\frac{2^{5p+4}c_p M^{6p+4}}{c_0\lambda_{\min}^{2p+1}}\right)^{1/2p}$. If the sample $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$ satisfies

$$\|\hat{V}_{XX} - V_{XX}\| \le 5M^2\sqrt{\frac{\log(8/\delta)}{m}} \le \frac{1}{6}\lambda_{\min} \tag{13}$$

and

$$\|\hat{V}_{XY} - V_{XY}\| \le 6M^2\sqrt{\frac{\log(8/\delta)}{m}}, \tag{14}$$

then

$$\begin{aligned}
\|\hat{\mathbf{w}}_t - \mathbf{w}_*\|^2 &\le (1 - \eta_t c_0\lambda_{\min})\|\hat{\mathbf{w}}_{t-1} - \mathbf{w}_*\|^2 \\
&\quad + \eta_t C\left(h^{-4p} + \frac{\log(8/\delta)}{m}\right)
\end{aligned}$$

for some constant C independent of $m,\ \delta$, or h.

Proof: By the definition of $\hat{\mathbf{w}}_t$, we have

$$\begin{aligned}
\|\hat{\mathbf{w}}_t - \mathbf{w}_*\|^2 = \|\hat{\mathbf{w}}_{t-1} - \mathbf{w}_*\|^2 \\
- 2\eta_t(\hat{\mathbf{w}}_{t-1} - \mathbf{w}_*)^\top \nabla R(\hat{\mathbf{w}}_{t-1}) \\
+ \eta_t^2\|\nabla R(\hat{\mathbf{w}}_{t-1})\|^2.
\end{aligned} \tag{15}$$

The key to prove Proposition 10 is to estimate $\nabla R(\hat{\mathbf{w}}_{t-1})$ appropriately. For this purpose we write

$$\begin{aligned}
&\nabla R(\hat{\mathbf{w}}_{t-1}) \\
&= \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\Psi'\left(\tfrac{\zeta_{t-1}^2(i,j)}{2h^2}\right)\zeta_{t-1}(i,j)(\mathbf{x}_i - \mathbf{x}_j) \\
&= \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\left[\Psi'\left(\tfrac{\zeta_{t-1}^2(i,j)}{2h^2}\right) - \Psi'_+(0)\right]\zeta_{t-1}(i,j)(\mathbf{x}_i - \mathbf{x}_j) \\
&\quad - \frac{c_0}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\zeta_{t-1}(i,j)(\mathbf{x}_i - \mathbf{x}_j) \\
&= \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\left[\Psi'\left(\tfrac{\zeta_{t-1}^2(i,j)}{2h^2}\right) - \Psi'_+(0)\right]\zeta_{t-1}(i,j)(\mathbf{x}_i - \mathbf{x}_j) \\
&\quad - 2c_0\left\{\left(\hat{V}_{XY} - V_{XY}\right) - \left(\hat{V}_{XX} - V_{XX}\right)\hat{\mathbf{w}}_{t-1}\right\} \\
&\quad + 2c_0 V_{XX}(\hat{\mathbf{w}}_{t-1} - \mathbf{w}_*) \\
&:= D_1 + D_2 + D_3,
\end{aligned}$$

where we have used the fact $V_{XY} = V_{XX}\mathbf{w}_*$ obtained in Lemma 4.

Note that all the conditions for Proposition 9 hold. So we have the bound $\|\hat{\mathbf{w}}_t\| \le \frac{3M^2}{\lambda_{\min}}$ for all t. For D_1, by the Lipschitz condition of Ψ' and the bound for $\hat{\mathbf{w}}_{t-1}$, as we have shown in the proof of Proposition 9, we have

$$\|D_1\| \le 2^{5p+4}c_p M^{6p+4}\lambda_{\min}^{-2p-1}h^{-2p}. \tag{16}$$

For D_2, by (13), (14) and the bound for $\hat{\mathbf{w}}_{t-1}$, we have

$$\begin{aligned}\|D_2\| &\le 2c_0\left(\|\hat{V}_{XY}-V_{XY}\|+\|\hat{V}_{XX}-V_{XX}\|\|\hat{\mathbf{w}}_{t-1}\|\right)\\ &\le 2c_0\left(6M^2+\frac{15M^4}{\lambda_{\min}}\right)\sqrt{\frac{\log(8/\delta)}{m}}\\ &\le 42c_0M^4\lambda_{\min}^{-1}\sqrt{\frac{\log(8/\delta)}{m}}. \qquad (17)\end{aligned}$$

Now we can estimate the second term on the right of (15). For notational simplicity let

$$\tilde{C}=\max\{2^{5p+4}c_pM^{6p+4}\lambda_{\min}^{-2p-1},\ 42c_0M^4\lambda_{\min}^{-1}\}.$$

By (16) and the elementary inequality $ab \le \frac{a^2}{2}+\frac{b^2}{2}$, we have

$$\begin{aligned}&\left|(\hat{\mathbf{w}}_{t-1}-\mathbf{w}_*)^\top D_1\right|\\ &\le \frac{c_0\lambda_{\min}}{2}\|\hat{\mathbf{w}}_{t-1}-\mathbf{w}_*\|^2+\frac{1}{2c_0\lambda_{\min}}\|D_1\|^2.\\ &\le \frac{c_0\lambda_{\min}}{2}\|\hat{\mathbf{w}}_{t-1}-\mathbf{w}_*\|^2+\frac{\tilde{C}^2}{2c_0\lambda_{\min}}h^{-4p}.\end{aligned}$$

Similarly, by (17), we have

$$\begin{aligned}&\left|(\hat{\mathbf{w}}_{t-1}-\mathbf{w}_*)^\top D_2\right|\\ &\le \frac{c_0\lambda_{\min}}{2}\|\hat{\mathbf{w}}_{t-1}-\mathbf{w}_*\|^2+\frac{1}{2c_0\lambda_{\min}}\|D_2\|^2\\ &\le \frac{c_0\lambda_{\min}}{2}\|\hat{\mathbf{w}}_{t-1}-\mathbf{w}_*\|^2+\frac{\tilde{C}^2}{2c_0\lambda_{\min}}\frac{\log(8/\delta)}{m}.\end{aligned}$$

These together with the fact that

$$\begin{aligned}(\hat{\mathbf{w}}_{t-1}-\mathbf{w}_*)^\top D_3 &= 2c_0(\hat{\mathbf{w}}_{t-1}-\mathbf{w}_*)^\top V_{XX}(\hat{\mathbf{w}}_{t-1}-\mathbf{w}_*)\\ &\ge 2c_0\lambda_{\min}\|\hat{\mathbf{w}}_{t-1}-\mathbf{w}_*\|^2\end{aligned}$$

enable us to obtain

$$\begin{aligned}&-2\eta_t(\hat{\mathbf{w}}_{t-1}-\mathbf{w}_*)^\top\nabla R(\hat{\mathbf{w}}_{t-1})\\ &\le -2\eta_tc_0\lambda_{\min}\|\hat{\mathbf{w}}_{t-1}-\mathbf{w}_*\|^2\\ &\quad+\frac{\eta_t\tilde{C}^2}{c_0\lambda_{\min}}\left(h^{-4p}+\frac{\log(8/\delta)}{m}\right). \qquad (18)\end{aligned}$$

We turn to estimate the last term on the right hand side of (15). We need the trivial bound

$$\|D_3\|\le 2c_0\lambda_{\max}\|\hat{\mathbf{w}}_{t-1}-\mathbf{w}_*\|. \qquad (19)$$

Combining the estimates in (16), (17), and (19), we have

$$\begin{aligned}&\eta_t^2\|\nabla R(\hat{\mathbf{w}}_{t-1})\|^2\\ &\le 3\eta_t^2(\|D_1\|^2+\|D_2\|^2+\|D_3\|^2)\\ &\le 12\eta_t^2c_0^2\lambda_{\max}^2\|\hat{\mathbf{w}}_{t-1}-\mathbf{w}_*\|^2\\ &\quad+3\eta_t^2\tilde{C}^2\left(h^{-4p}+\frac{\log(8/\delta)}{m}\right)\\ &\le \eta_tc_0\lambda_{\min}\|\hat{\mathbf{w}}_{t-1}-\mathbf{w}_*\|^2\\ &\quad+\frac{1}{4c_0\lambda_{\min}}\eta_t\tilde{C}^2\left(h^{-4p}+\frac{\log(8/\delta)}{m}\right), \qquad (20)\end{aligned}$$

where we used the assumption $\eta_t \le \frac{\lambda_{\min}}{12c_0\lambda_{\max}^2}\le\frac{1}{12c_0\lambda_{\min}}$.

Let $C=\frac{5\tilde{C}^2}{4c_0\lambda_{\min}}$. Plugging the estimates in (18) and (20) into (15), we obtain the desired conclusion. ■

V. Error Bounds and Convergence Rates

To prove Theorem 2, we need two lemmas from [22].

Lemma 11: For $v\in(0,1]$ and $\theta\in[0,1]$,

$$\sum_{t=1}^{T}\frac{1}{t^\theta}\prod_{j=t+1}^{T}\left(1-\frac{v}{j^\theta}\right)\le\frac{3}{v}.$$

Lemma 12: For any $0\le\ell<T$ and $0<\theta<1$, there holds

$$\sum_{t=\ell+1}^{T}t^{-\theta}\ge\frac{1}{1-\theta}\left[(T+1)^{1-\theta}-(\ell+1)^{1-\theta}\right].$$

Proof of Theorem 2: For a sample satisfying the conditions (13) and (14), Proposition 10 states that

$$\begin{aligned}\|\hat{\mathbf{w}}_t-\mathbf{w}_*\|^2 &\le (1-\eta_tc_0\lambda_{\min})\|\hat{\mathbf{w}}_{t-1}-\mathbf{w}_*\|^2\\ &\quad+\eta_tC\left(h^{-4p}+\frac{\log(8/\delta)}{m}\right)\end{aligned}$$

for all t. Applying this estimate iteratively we obtain

$$\begin{aligned}\|\hat{\mathbf{w}}_T-\mathbf{w}_*\|^2 &\le \|\mathbf{w}_*\|^2\prod_{t=1}^{T}(1-\eta_tc_0\lambda_{\min})\\ &\quad+C\left(h^{-4p}+\frac{\log(8/\delta)}{m}\right)\sum_{t=1}^{T}\prod_{j=t+1}^{T}(1-\eta_jc_0\lambda_{\min})\eta_t.\end{aligned}$$

Since $\eta_t=\eta t^{-\theta}$, by the elementary inequality $1-u\le\exp(-u)$ and Lemma 12 with $\ell=0$, we have

$$\begin{aligned}\prod_{t=1}^{T}(1-\eta_tc_0\lambda_{\min}) &\le \exp\left(-c_0\lambda_{\min}\sum_{t=1}^{T}\eta_t\right)\\ &\le \exp\left(\frac{\eta c_0\lambda_{\min}\left(1-(T+1)^{1-\theta}\right)}{1-\theta}\right)\\ &\le \exp\left(\frac{\eta c_0\lambda_{\min}\left(1-T^{1-\theta}\right)}{1-\theta}\right).\end{aligned}$$

Lemma 11 with $v = \eta c_0 \lambda_{\min}$ yields

$$\sum_{t=1}^{T} \prod_{j=t+1}^{T} (1 - \eta_j c_0 \lambda_{\min}) \eta_t$$

$$= \eta \sum_{t=1}^{T} \frac{1}{t^\theta} \prod_{j=t+1}^{T} \left(1 - \frac{\eta c_0 \lambda_{\min}}{j^\theta}\right) \leq \frac{3}{c_0 \lambda_{\min}}.$$

Therefore,

$$\|\hat{\mathbf{w}}_T - \mathbf{w}_*\|^2 \leq \|\mathbf{w}_*\|^2 \exp\left(\frac{\eta c_0 \lambda_{\min}(1 - T^{1-\theta})}{1-\theta}\right) + \frac{3C}{c_0 \lambda_{\min}} \left(h^{-4p} + \frac{\log(8/\delta)}{m}\right)$$

$$\leq C' \left\{\exp\left(-\tfrac{\eta c_0 \lambda_{\min} T^{1-\theta}}{1-\theta}\right) + h^{-4p} + \frac{\log(8/\delta)}{m}\right\},$$

where $C' = \|\mathbf{w}_*\|^2 \exp\left(\frac{\eta c_0 \lambda_{\min}}{1-\theta}\right) + \frac{3C}{c_0 \lambda_{\min}}$. The proof of Theorem 2 is completed after noticing that the conditions (13) and (14) hold with probability at least $1 - \delta$, as are guaranteed by Lemma 8. ■

VI. Simulations

In this section we study the empirical performance of the gradient descent method for MEE by simulations and compare it with our theoretical analysis. On one hand we expect the theoretical analysis provides some guidance to the empirical implementation. On the other hand, since the theoretical analysis is based on upper bounds which might be far from tight, it is important to understand the gap between the theory and empirical applications.

In the simulation, let $\mathbf{x} \in \mathbb{R}^{10}$ and the model be defined by $Y = \mathbf{w}_*^\top \mathbf{x} + \epsilon$ with $\mathbf{w}_* = [1\, -1\; 1\, -1\; 1\, -1\; 1\, -1\; 1\, -1]^\top$ and $\mathbf{x} \sim N(0, I_{10})$. We consider two types of noise. The first type is the Gaussian noise $\epsilon \sim N(0, c\mathbf{w}_*^\top \mathbf{x})$ for each given $\mathbf{x}$. The second type is the generalized Gaussian noise with a probability density function $f(\epsilon) \propto \exp(-|c\epsilon|^{0.3})$. It is a typical heavy tailed distribution and has been employed in [2] to explore the effectiveness of a minimum total error entropy algorithm. For both noise models we select the constant c so that the signal to noise ratio equal to one. As indicated by Theorem 2, a small constant step size can be used to guarantee convergence and the scaling parameter should be large enough. In our simulations we have chosen $\eta_t = \sqrt{0.005\pi}$ (so that it satisfies the condition in Theorem 2) and $h = 10$. We let the sample size m vary from 50 to 500. The simulation results based on 100 repeated experiments were reported in Figs. 1 and 2 for the two noise models, respectively.

In Figs. 1(a) and 2(a) we report the change of the mean squared error as the number of iterations increases. In Figs. 1(b) and 2(b) we compare the mean squared error with iteration to convergence and the mean squared error with optimal iteration (i.e., the number of iterations that leads to minimal mean squared error). In Figs. 1(c) and 2(c) we compare the number of iterations to convergence and the optimal number of iterations. Similar results are seen regardless of the noise types. All these results indicate that the optimal solution can be achieved by early stopping the gradient descent iteration while further increasing the number of iterations may hurt the learning performance. The performance degradation is notable in a small sample setting while negligible in a large sample setting. Therefore, early stopping is not only sufficient but also necessary when the sample size is small. An interesting observation is that the number of iterations to convergence and the optimal number of iterations tend to coincide when the sample size is large. A plausible explanation is that when the sample size is large, the empirical risk approximates the expected risk well and thus $\hat{\mathbf{w}}$ approximates $\mathbf{w}_*$ well. So the optimal solution does require $\hat{\mathbf{w}}_t$ to converge to $\hat{\mathbf{w}}$. We observed that the number of iterations to convergence decreases as the sample size increases. Although it

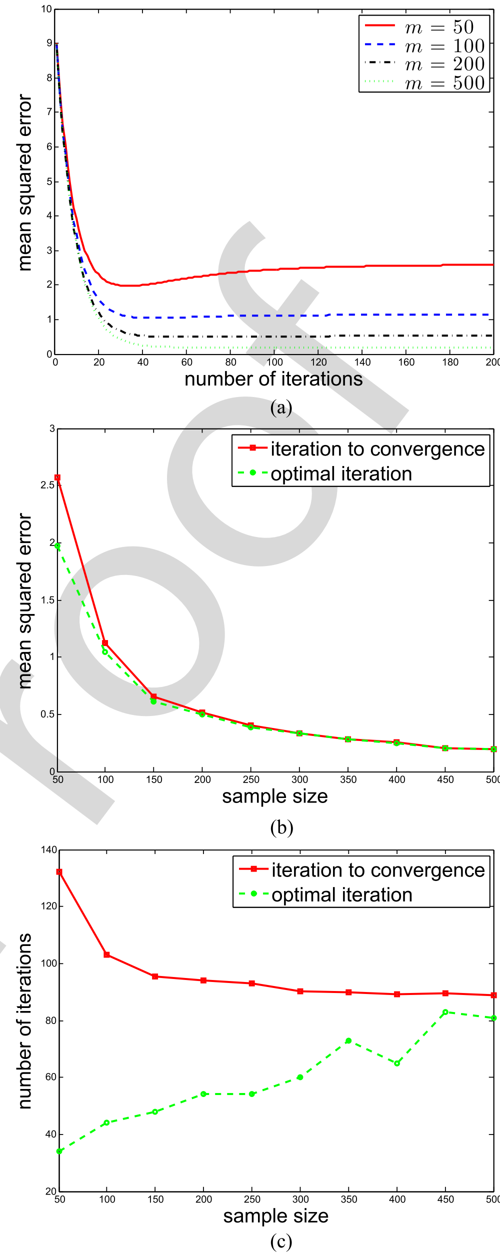


Fig. 1. Simulation results for Gaussian noise.

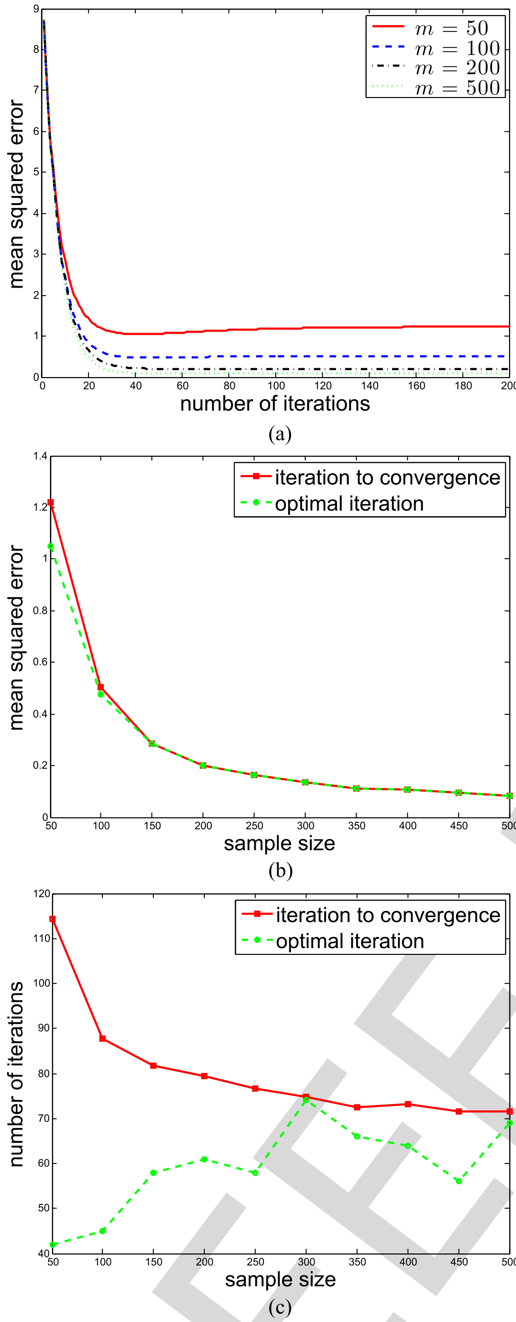


Fig. 2. Simulation results for generalized Gaussian noise.

does not contradict our analysis, it does seem surprising. We do not have an explanation at the moment and would leave it for future study.

Recall that the upper bound in Theorem 2 implies the sufficiency of early stopping for an optimal solution in a large sample size setting. Moreover, although it is hard to verify the optimal number of iterations is of order $O(\log m)$, it does increase very slowly according Figs. 1(c) and 2(c).

Theorem 2 also provides useful insight on the choice of the step size. The upper bound of η can be estimated from the sample. Choosing η around half of the upper bound usually works well in practice. However, there is a gap between the theoretical analysis and empirical applications regarding the choice of the scaling parameter h. The theoretical lower bound on h is too restrictive. In practice it is found that a moderately large h is good and a very large h is not necessary.

VII. Discussions

To derive our results, we have assumed that the covariance matrix V_{XX} of the input variable X is non-degenerate. This condition, however, may not be true in many situations. A very typical model is the classical linear regression model where an intercept is included:

$$Y = \beta_0 + \beta_1^\top Z + \epsilon \tag{21}$$

with $\beta_0 \in \mathbb{R}$, $\beta_1 \in \mathbb{R}^n$, and Z a vector valued random variable containing n explanatory variables. In this case, $\mathbf{w}_* = [\beta_0 \ \beta_1^\top]^\top$ and $X = [1 \ Z^\top]^\top$. Note that

$$V_{XX} = \begin{pmatrix} 0 & 0 \\ 0 & V_{ZZ} \end{pmatrix}.$$

So it is always degenerate.

When V_{XX} is degenerate, we cannot prove the convergence of $\hat{\mathbf{w}}_t$ to $\mathbf{w}_*$. Instead, we need to consider their projections onto the principal component space. Let U denote the subspace of $\mathbb{R}^n$ spanned by the principal components associated to the positive eigenvalues and P_U the projection onto U. Let $\lambda_{\min}$ denote the smallest positive eigenvalue. We can prove

$$\begin{aligned}\lambda_{\min}\|P_U(\mathbf{w}-\mathbf{w}_*)\|^2 &\le (\mathbf{w}-\mathbf{w}_*)^\top V_{XX}(\mathbf{w}-\mathbf{w}_*)\\ &\le \lambda_{\max}\|P_U(\mathbf{w}-\mathbf{w}_*)\|^2.\end{aligned}$$

By this relationship and the techniques developed in this paper and [23], we can prove the convergence of $P_U(\hat{\mathbf{w}}_t)$ to $P_U(\mathbf{w}_*)$. This guarantees the variance of $\hat{\mathbf{w}}_t^\top X$ converges to the variance $\mathbf{w}_*^\top X$ and thus $\hat{\mathbf{w}}_t^\top X$ plus an appropriate intercept provides good predictive performance. In the model (21), if V_{ZZ} is positive definite, we see $\hat{\mathbf{w}}_t$ estimates the slope coefficients β_1.

As for the implementation of the algorithm, we remark that if V_{XX} is non-degenerate, the initial point is not necessarily chosen as $\hat{\mathbf{w}}_0 = 0$. The convergence holds true for any starting point. If V_{XX} is degenerate, the convergence of $\hat{\mathbf{w}}_t$ to $\mathbf{w}_*$ can be proved if the starting value is in the principal components space U. Actually, since $\mathbf{x}_i - \mathbf{x}_j$ is in U, all $\hat{\mathbf{w}}_t$ are in U. Thus, $P_U(\hat{\mathbf{w}}_t) = \hat{\mathbf{w}}_t$ and the convergence of $\hat{\mathbf{w}}_t$ to $P_U(\mathbf{w}_*)$ is exactly the convergence of $P_U(\hat{\mathbf{w}}_t)$ to $P_U(\mathbf{w}_*)$. However, if the starting point has a nonzero components normal to U, it will never diminish during the iteration process.

We have focused on linear regression models in this paper. Note that the MEE principle can be extended to nonlinear regression by the kernel trick [14], [17]. Regularization theory plays an important role to overcome the overfitting problem in this case. It would be interesting to study the use of gradient descent for the kernel MEE method in the future.

Acknowledgment

The authors thank the anonymous reviewers for their valuable comments.

REFERENCES

[1] D. Erdogmus and J. C. Principe, "Comparison of entropy and mean square error criteria in adaptive system training using higher order statistics," in *Proc. 2nd Int. Workshop Independent Compon. Anal. Blind Signal Separation*, 2000, pp. 75–90.

[2] P. Shen and C. Li, "Minimum total error entropy method for parameter estimation," *IEEE Trans. Signal Process.*, vol. 63, no. 15, pp. 4079–4090, Aug. 2015.

[3] D. Erdogmus, K. Hild II, and J. C. Principe, "Blind source separation using Rényi's α-marginal entropies," *Neurocomput.*, vol. 49, pp. 25–38, 2002.

[4] D. Erdogmus and J. C. Principe, "Convergence properties and data efficiency of the minimum error entropy criterion in adaline training," *IEEE Trans. Signal Process.*, vol. 51, no. 7, pp. 1966–1978, Jul. 2003.

[5] E. Gokcay and J. C. Principe, "Information theoretic clustering," *IEEE Trans. Pattern Anal. Mach. Learn.*, vol. 24, no. 2, pp. 158–171, Feb. 2002.

[6] L. M. Silva, J. Marques de Sá, and L. A. Alexandre, "Neural network classification using Shannon's entropy," in *Proc. Eur. Symp. Artif. Neural Netw.*, 2005, pp. 217–222.

[7] L. M. Silva, J. Marques de Sá, and L. A. Alexandre, "The MEE principle in data classification: A perceptron-based analysis," *Neural Comput.*, vol. 22, pp. 2698–2728, 2010.

[8] B. Chen, P. Zhu, and J. C. Principe, "Survival information potential: A new criterion for adaptive system training," *IEEE Trans. Signal Process.*, vol. 60, no. 3, pp. 1184–1194, Mar. 2012.

[9] F. Cucker and D. X. Zhou, *Learning Theory: An Approximation Theory Viewpoint*. Cambridge, U.K.: Cambridge Univ. Press, 2007.

[10] I. Steinwart, D. Hush, and C. Scovel, "Optimal rates for regularized least squares regression," in S. Dasgupta and A. Klivans, eds., *Proc. 22nd Annu. Conf. Learn. Theory*, 2009, pp. 79–93.

[11] Y. Yao, L. Rosasco, and A. Caponnetto, "On early stopping in gradient descent learning," *Constructive Approx.*, vol. 26, no. 2, pp. 289–315, 2007.

[12] E. Parzen, "On the estimation of a probability density function and the mode," *Ann. Math. Statist.*, vol. 33, pp. 1049–1051, 1962.

[13] T. Hu, J. Fan, Q. Wu, and D.-X. Zhou, "Learning theory approach to a minimum error entropy criterion," *J. Mach. Learn. Res.*, vol. 14, pp. 377–397, 2013.

[14] T. Hu, J. Fan, Q. Wu, and D.-X. Zhou, "Regularization schemes for minimum error entropy principle," *Anal. Appl.*, vol. 13, no. 4, pp. 437–455, 2015.

[15] J. Fan, T. Hu, Q. Wu, and D.-X. Zhou, "Consistency analysis of an empirical minimum error entropy algorithm," *Appl. Comput. Harmonic Anal.*, vol. 41, pp. 164–189, 2016.

[16] B. Chen and J. C. Principe, "Some further results on the minimum error entropy estimation," *Entropy*, vol. 14, no. 5, pp. 966–977, 2012.

[17] J. C. Principe, *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*. New York, NY, USA: Springer-Verlag, 2010.

[18] B. Chen and J. C. Principe, "Stochastic gradient algorithm under (h, ϕ)-entropy criterion," *Circuits, Syst., Signal Process.*, vol. 26, no. 6, pp. 941–960, 2007.

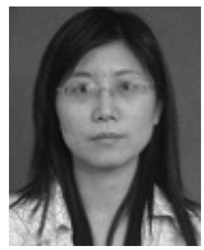
[19] Z. Wu, S. Peng, W. Ma, B. Chen, and J. C. Principe, "Minimum error entropy algorithms with sparsity penalty constraints," *Entropy*, vol. 17, no. 5, pp. 3419–3437, 2015.

[20] B. Chen, Y. Zhu, and J. Hu, "Mean-square convergence analysis of adaline training with minimum error entropy criterion," *IEEE Trans. Neural Netw.*, vol. 21, no. 7, pp. 1168–1179, Jul. 2010.

[21] S. Smale and D. X. Zhou, "Learning theory estimates via integral operators and their approximations," *Construct. Approx.*, vol. 26, pp. 153–172, 2007.

[22] Y. Ying and D.-X. Zhou, "Online regularized classification algorithms," *IEEE Trans. Inf. Theory*, vol. 52, no. 11, pp. 4775–4788, Nov. 2006.

[23] J. Lin and D.-X. Zhou, "Learning theory of randomized Kaczmarz algorithm," *J. Mach. Learn. Res.*, vol. 16, pp. 3341–3365, 2015.



Ting Hu received the B.S. degree in computational mathematics in 2004 and the Ph.D. degree in mathematics in 2009, both from Wuhan University, Hubei, China. From 2009 to 2011, she was a Research Fellow at the City University of Hong Kong, Hong Kong, China. She is currently an Associate Professor in the School of Mathematics and Statistics, Wuhan University, Hubei, China. Her research interests include learning theory, machine learning, wavelet analysis, and applications.

Qiang Wu received the Ph.D. degree in mathematics from the City University of Hong Kong, Hong Kong, China, in 2005. He was with Duke University, Michigan State University, and the University of Liverpool before joinedMiddle Tennessee State University in 2011. He is currently an Associate Professor of mathematics and actuarial science in the Department of Mathematical Sciences and a Faculty Member of the Computational Science Ph.D. program. His research interests include computational harmonic analysis, statistical learning theory, high-dimensional data mining, and their applications.

Ding-Xuan Zhou received the B.Sc. and Ph.D. degrees in mathematics from Zhejiang University, Hangzhou, China, in 1988 and 1991, respectively. He joined City University of Hong Kong as a Research Assistant Professor in 1996, and is currently the Chair Professor in the Department of Mathematics. His research interests include learning theory, data science, wavelet analysis, and approximation theory. He has published more than 100 research papers and is serving on the editorial board of more than ten international journals, and is an Editor-in-Chief of the journal *Analysis and Applications*. He received a Research Fund for Distinguished Young Scholars from the National Science Foundation of China in 2005 and a Humboldt Research Fellowship in 1993, and was rated by Thomson Reuters highly-cited researcher in 2014 and 2015. He has co-organised more than 20 international conferences and conducted more than 20 research grants.

Convergence of Gradient Descent for Minimum Error Entropy Principle in Linear Regression

Ting Hu, Qiang Wu, and Ding-Xuan Zhou

Abstract—**We study the convergence of minimum error entropy (MEE) algorithms when they are implemented by a gradient descent. This method has been used in practical applications for more than one decade, but there has been no consistency or rigorous error analysis. This paper gives the first rigorous proof for the convergence of the gradient descent method for MEE in a linear regression setting. The mean square error is proved to decay exponentially fast in terms of the iteration steps and of order $O(\frac{1}{m})$ in terms of the sample size m. The mean square convergence is guaranteed when the step size is chosen appropriately and the scaling parameter is large enough.**

Index Terms—**Minimum error entropy, error information, gradient descent method, error analysis, global convergence.**

I. Introduction

REGRESSION analysis plays important roles in many fields of science and engineering. The traditional least square method is the mostly used algorithm for regression in practice. However, it is suboptimal when the system noise is not normally distributed. Variant approaches have been proposed to deal with data with outliers or heavy-tailed distributions. Minimum error entropy (MEE) criterion is one of them. It is motivated by the idea of minimizing the information as measured by entropy in the prediction error. The estimated model is expected to preserve information as much as possible and thus improves the predictive performance. Unlike the traditional least square method which relies only on the variance of the prediction error, the error entropy takes all higher order moments into account and is thus advantageous when MEE is used to handle non-Guassian and heavy tailed error distributions [1], [2]. As non-Gaussian noise is ubiquitous in real world applications, the superiority of MEE has been evidenced in a variety of applications, which include adaptive filtering, clustering, classification, feature selection, and blind source separation [3]–[8].

Manuscript received November 12, 2015; revised April 15, 2016 and June 29, 2016; accepted August 29, 2016. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Dmitry M. Malioutov. The work was supported in part by the National Natural Science Foundation of China under Grants 11671307, 11501078, 11671171, 11461161006, and 11471292, in part by the U.S. Department of Agriculture National Institute of Food and Agriculture under Grant 2016-70001-24636, and in part by the Research Grants Council of Hong Kong (Project no. CityU 11303915).

T. Hu is with the School of Mathematics and Statistics, Wuhan University, Wuhan 430072, China (e-mail: tinghu@whu.edu.cn).

Q. Wu is with the Department of Mathematical Sciences and the Computational Science Ph.D. Program, Middle Tennessee State University, Murfreesboro, TN 37132 USA (e-mail: qwu@mtsu.edu).

D.-X. Zhou is with the Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong (e-mail: mazhou@cityu.edu.hk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2016.2612169

Let X be a multivariate random variable with values in a compact subset of $\mathbb{R}^n$ and Y a real valued response variable. The purpose of regression analysis is to study the quantitative relationship between X and Y. This usually leads to estimating the regression function $f_*(\mathbf{x}) = \mathbf{E}(Y|X = \mathbf{x})$ from a sample of m observations $\mathbf{z} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ drawn independently and identically. As most statistical and machine learning algorithms for regression analysis have focused on the use of convex losses such as the squared loss in the least square method and the insensitive loss in support vector regression, approximation powers of learning algorithms with convex losses have been well studied in the literature; see e.g. [9]–[11] and the references therein. The MEE algorithms, however, use the error entropy as the loss function which is not convex. It brings essential difficulties to the analysis. Although the MEE algorithms have been verified effective in many empirical studies, the study on its computational and mathematical properties is lagged a little bit behind.

The MEE approach was introduced in [1]. It aims to minimize the information contained in the error and maximize the information captured by the estimated model. Given an estimator f of the regression function, define the error variable as $E = Y - f(X)$. One can measure the error information by Renyi's entropy or Shannon's entropy. In this paper we consider the second order Renyi's entropy

$$H(E) = -\log \mathbf{E}(p_{_E}) = -\log \int p_{_E}^2(e)de$$

where $p_{_E}$ denotes the probability density function of E. For the given sample $\mathbf{z}$, define $e_i = y_i - f(\mathbf{x}_i)$. Then $p_{_E}$ can be estimated by Parzen windowing [12] which, given a kernel function $K : \mathbb{R} \to [0, \infty)$ and a scaling parameter $h > 0$, defines a kernel density estimator by

$$\hat{p}_{_E}(e) = \frac{1}{m}\sum_{i=1}^m K_h(e - e_i) = \frac{1}{mh}\sum_{i=1}^m K\left(\frac{e - e_i}{h}\right).$$

A usual choice is the Gaussian kernel density estimator where $K(u) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{u^2}{2})$ and $K_h(u) = \frac{1}{\sqrt{2\pi}h}\exp(-\frac{u^2}{2h^2})$. The empirical error information is

$$\hat{H}(f) = -\log\left\{\frac{1}{m^2h}\sum_{i=1}^m\sum_{j=1}^m K\left(\frac{e_i - e_j}{h}\right)\right\}.$$

The MEE algorithm searches for an estimator that minimizes $\hat{H}$ over a hypothesis space.

The structure of the empirical entropy $\hat{H}$ exhibits that the scaling parameter h plays an important role in the MEE algorithm design. The value of h is adjusted for different learning tasks in MEE algorithms and the corresponding learning effects are presented in a series of numerical simulations; see e.g. [6], [7]. Mathematically, the predictive performance of MEE algorithms was analyzed in [13]–[16]. The convergence of MEE algorithms can be guaranteed only for homoscedastic model if the scaling parameter h is chosen small. The scaling parameter h should be chosen large enough to guarantee the algorithms to be asymptotically consistent for more general models. This coincides with the empirical studies in the literature.

From a computational perspective, the loss function is close to the squared loss by weighing less on the high order statistics of the error when h is large. Thus, using a relatively large scaling parameter reduces the risk that MEE algorithms suffer from being stuck in local minima. MEE algorithms are usually implemented by gradient descent or stochastic gradient descent [1], [17]–[19]. However, because the optimization problem arising from MEE is non-convex, the convergence of the gradient descent method is not unconditionally guaranteed. A mean squared convergence result is proved in [20] which, however, only guarantees the solution of the stochastic gradient descent method converges to a local minima but not necessarily the global minima. In this paper, our purpose is to derive conditions and stopping criteria for the gradient descent method to achieve global convergence.

We focus on linear regression models in this paper. Assume

$$y = \mathbf{w}_*^\top \mathbf{x} + \epsilon, \quad \mathbf{E}[\epsilon|\mathbf{x}] = 0$$

for some $\mathbf{w}_* \in \mathbb{R}^n$, where ϵ is a mean zero noise random variable. The regression function takes the form $f_*(\mathbf{x}) = \mathbf{w}_*^\top \mathbf{x}$ and the target of regression analysis is to estimate $\mathbf{w}_*$ from the sample. For an estimator $\hat{\mathbf{w}}$, the goodness could be measured by the squared error $\|\hat{\mathbf{w}} - \mathbf{w}_*\|^2$.

The MEE estimator $\hat{\mathbf{w}}$ is defined as

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^n} \hat{H}(\mathbf{w})$$

where, given $e_i = y_i - \mathbf{w}^\top \mathbf{x}_i$,

$$\hat{H}(\mathbf{w}) = -\log\left\{\frac{1}{m^2 h}\sum_{i=1}^m\sum_{j=1}^m K\left(\frac{e_i - e_j}{h}\right)\right\}.$$

As the logarithmic function is monotone and does not affect the minimization process, we remove it and consider the transformed empirical error information

$$R(\mathbf{w}) = -\frac{h^2}{m^2}\sum_{i=1}^m\sum_{j=1}^m K\left(\frac{(y_i - \mathbf{w}^\top\mathbf{x}_i) - (y_j - \mathbf{w}^\top\mathbf{x}_j)}{h}\right).$$

It is obvious the MEE estimator can also be obtained by

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^n} R(\mathbf{w}). \tag{1}$$

When K is differentiable, the gradient descent algorithm for MEE starts with $\hat{\mathbf{w}}_0 = 0$ and updates the estimator by

$$\hat{\mathbf{w}}_t = \hat{\mathbf{w}}_{t-1} - \eta_t \nabla R(\hat{\mathbf{w}}_{t-1})$$

in the t-th step, where ∇ is the gradient operator and $\eta_t > 0$ is the step size. When this method is used to solve the MEE estimator (1), the first question might be the convergence of $\hat{\mathbf{w}}_t$ to $\hat{\mathbf{w}}$ as the number of iterations becomes large. However, we would consider the problem in an alternative way. Recall the ultimate goal is to learn the true regression coefficients vector $\mathbf{w}_*$. On one hand, if $\hat{\mathbf{w}}_t$ provide good estimates of $\mathbf{w}_*$, the convergence of $\hat{\mathbf{w}}_t$ to $\hat{\mathbf{w}}$ does not matter much. On the other hand, notice that

$$\|\hat{\mathbf{w}}_t - \mathbf{w}_*\| \le \|\hat{\mathbf{w}}_t - \hat{\mathbf{w}}\| + \|\hat{\mathbf{w}} - \mathbf{w}_*\|.$$

Even if $\hat{\mathbf{w}}_t$ does converge to $\hat{\mathbf{w}}$, it does not make much sense to iterate the gradient descent steps till convergence because the second term on the right will dominate the error. Instead, the algorithm should be stopped earlier when the performance of the estimate does not improve.

In order to state our main results we need some assumptions. Firstly, we assume both X and Y are uniformly bounded by a constant M. Also, the covariance matrix V_{XX} of X is nondegenerate, that is, all the eigenvalues of V_{XX} are positive. In particular, we denote by $\lambda_{\max}$ and $\lambda_{\min}$ the largest and the smallest eigenvalues of V_{XX}, respectively.

To simplify our presentation and notations in the proofs, we focus on symmetric kernels and define $\Psi : [0,\infty) \to [0,\infty)$ as $\Psi(u) = K(\sqrt{2u})$ or equivalently, $\Psi(\frac{u^2}{2}) = K(u)$. With this notation, the empirical error can be rewritten as

$$R(\mathbf{w}) = -\frac{h^2}{m^2}\sum_{i=1}^m\sum_{j=1}^m \Psi\left(\frac{[(y_i - \mathbf{w}^\top\mathbf{x}_i) - (y_j - \mathbf{w}^\top\mathbf{x}_j)]^2}{2h^2}\right).$$

Assume Ψ is decreasing and differentiable, $c_0 = -\Psi'_+(0) > 0$, and for some $p > 0$,

$$|\Psi'(u) - \Psi'_+(0)| \le c_p u^p, \ \forall\, u > 0. \tag{2}$$

When the Gaussian kernel is used, it is easy to verify that $\Psi(u) = \frac{1}{\sqrt{2\pi}}\exp(-u)$. We have $c_0 = \frac{1}{\sqrt{2\pi}}$ and (2) holds with $p = 1$ and $c_p = \frac{1}{\sqrt{2\pi}}$.

Our first result, Theorem 1 below, shows that $\hat{\mathbf{w}}_t$ is uniformly bounded with large probability.

Theorem 1: If $0 < \eta_t \le \frac{1}{2c_0\lambda_{\max}}$ for all $t \in \mathbb{N}$ and $h \ge \left(\frac{2^{5p+4}c_p M^{6p+2}}{c_0\lambda_{\min}^{2p+1}}\right)^{1/2p}$, then for any $0 < \delta < 1$, we have

$$\|\hat{\mathbf{w}}_t\| \le \frac{3M^2}{\lambda_{\min}} \qquad \text{for all } t \in \mathbb{N}$$

with probability $1 - \delta$ provided that $m \ge \frac{900M^4\log(8/\delta)}{\lambda_{\min}^2}$.

Because any bounded closed set in $\mathbb{R}^n$ is compact, Theorem 1 guarantees that a subsequence of $\{\hat{\mathbf{w}}_t\}$ converges to some point. To ensure the accumulation point is the solution $\mathbf{w}_*$ as we expected, the step size and the scaling parameters should be selected appropriately.

Theorem 2: Let $\eta_t = \eta t^{-\theta}$ for some $0 \le \theta < 1$ and $0 < \eta \le \frac{\lambda_{\min}}{12c_0\lambda_{\max}^2}$. Let $h \ge \left(\frac{2^{5p+4}c_p M^{6p+2}}{c_0\lambda_{\min}^{2p+1}}\right)^{1/2p}$. For any $0 < \delta < 1$, we have

$$\|\hat{\mathbf{w}}_T - \mathbf{w}_*\|^2 \le C'\left\{\exp\left(-\frac{\eta c_0\lambda_{\min}T^{1-\theta}}{1-\theta}\right) + \frac{1}{h^{4p}} + \frac{\log(8/\delta)}{m}\right\}$$

with probability $1 - \delta$ provided that $m \ge \frac{900M^4\log(8/\delta)}{\lambda_{\min}^2}$. Here the constant C' is independent of m, h, or δ, and will be given explicitly in the proof.

Theorem 2 indicates that, under appropriate choices of the parameters, $\hat{\mathbf{w}}_t$ converges to $\mathbf{w}_*$ exponentially fast in terms of the number of iterations and is of order $O(\frac{1}{m})$ in terms of the sample size. In particular, the convergence holds with a fixed step size $\eta_t = \eta$ provided that η is small enough. In practice, given a set of observations, the sample size m is fixed. The number of iteration steps $T = O(\log m)$ is usually sufficient to achieve the best possible learning performance.

II. Preliminaries

We first give several basic facts associated to the linear regression model. Throughout this section we denote $\mu_X = \mathbf{E}(X)$ and $\mu_Y = \mathbf{E}(Y)$.

Lemma 3: The covariance matrix V_{XX} satisfies $\lambda_{\max} = \|V_{XX}\| \le M^2$.

Proof: Note that $V_{XX} = \mathbf{E}(XX^\top) - \mu_X \mu_X^\top$. Since X is bounded by M, we have $\|\mathbf{E}(XX^\top)\| \le M^2$. Since both $\mathbf{E}(XX^\top)$ and $\mu_X \mu_X^\top$ are positive semidefinite, we have

$$\|V_{XX}\| \le \|\mathbf{E}(XX^\top)\| \le M^2.$$

This proves the conclusion. ■

Lemma 4: Let V_{XY} denote the covariance vector between X and Y. We have $V_{XX}\mathbf{w}_* = V_{XY}$ and $\|\mathbf{w}_*\| \le \frac{2M^2}{\lambda_{\min}}$.

Proof: By the model assumption we have $\mu_Y = \mu_X^\top \mathbf{w}_*$. Therefore, $y - \mu_Y = (\mathbf{x} - \mu_X)^\top \mathbf{w}_* + \epsilon$ and

$$(y - \mu_Y)(\mathbf{x} - \mu_X) = (\mathbf{x} - \mu_X)(\mathbf{x} - \mu_X)^\top \mathbf{w}_* + \epsilon(\mathbf{x} - \mu_X).$$

Taking expectation both sides and noting the fact $\mathbf{E}(\epsilon|\mathbf{x}) = 0$, we obtain $V_{XY} = V_{XX}\mathbf{w}_*$.

Since both X and Y are bounded by M, we have

$$\|V_{XY}\| = \|\mathbf{E}(XY) - \mu_X \mu_Y\| \le 2M^2.$$

Thus

$$\|\mathbf{w}_*\| = \|V_{XX}^{-1} V_{XY}\| \le \frac{2M^2}{\lambda_{\min}}.$$

This finishes the proof. ■

In our analysis, we need to deal with matrix and vector valued functions. For this purpose we need probability inequalities for Hilbert space valued random variables. The following one can be found in [21].

Lemma 5: Let $\mathcal{H}$ be a Hilbert space and $\{\xi_i\}_{i=1}^m$ be m independent random variables with values in $\mathcal{H}$. Suppose that for each i, $\|\xi_i\| \le M$ almost surely. Denote $\sigma^2 = \sum_{i=1}^m \mathbf{E}(\|\xi_i\|^2)$. Then, for any $\varepsilon > 0$,

$$\Pr\left\{\left\|\frac{1}{m}\sum_{i=1}^m [\xi_i - \mathbf{E}(\xi_i)]\right\| \ge \varepsilon\right\} \le\ 2\exp\left\{-\frac{m\varepsilon}{2M}\log\left(1 + \frac{mM\varepsilon}{\sigma^2}\right)\right\}.$$

By this lemma, we can prove the following inequality.

Lemma 6: Let $\mathcal{H}$ be a Hilbert space and ξ be a random variable with values in $\mathcal{H}$. Assume that $\|\xi\| \le M$ almost surely. Let $\{\xi_1, \xi_2, \ldots, \xi_m\}$ be a sample of m independent observations for ξ. Then, for any $\varepsilon > 0$,

$$\Pr\left\{\left\|\frac{1}{m}\sum_{i=1}^m \xi_i - \mathbf{E}(\xi)\right\| \ge \varepsilon\right\} \le 2\exp\left\{-\frac{m\varepsilon^2}{2M^2 + M\varepsilon}\right\}. \tag{3}$$

Proof: Since $\|\xi\| \le M$ almost surely, we have

$$\sigma^2 = \sum_{i=1}^m \mathbf{E}(\|\xi_i\|^2) = m\mathbf{E}(\|\xi\|^2) \le mM^2.$$

Applying Lemma 5 we obtain

$$\Pr\left\{\left\|\frac{1}{m}\sum_{i=1}^m \xi_i - \mathbf{E}(\xi)\right\| \ge \varepsilon\right\} \le\ 2\exp\left\{-\frac{m\varepsilon}{2M}\log\left(1 + \frac{\varepsilon}{M}\right)\right\}. \tag{4}$$

By the elementary inequality $\log(1+t) > \frac{2t}{2+t}$ for $t > 0$, we have

$$\frac{\varepsilon}{M}\log\left(1 + \frac{\varepsilon}{M}\right) \ge \frac{2\varepsilon^2}{2M^2 + M\varepsilon}.$$

Plugging this into (4) gives (3). ■

Lemma 7: Let $\mathcal{H}$ be a Hilbert space and ξ be a random variable with values in $\mathcal{H}$. Assume that $\|\xi\| \le M$ almost surely. Let $\{\xi_1, \xi_2, \ldots, \xi_m\}$ be a sample of m independent observations for ξ. Then, for any $0 < \tilde{\delta} < 1$, we have with confidence $1 - \tilde{\delta}$,

$$\left\|\frac{1}{m}\sum_{i=1}^m \xi_i - \mathbf{E}(\xi)\right\| \le \frac{1}{2}M\left(\tau + \sqrt{8\tau + \tau^2}\right)$$

where $\tau = \frac{\log(2/\tilde{\delta})}{m}$.

Using this lemma we can prove the following estimate.

Lemma 8: For any $0 < \delta < 1$, with probability at least $1 - \delta$, we have

$$\left\|\frac{1}{m^2}\sum_{i=1}^m\sum_{j=1}^m (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top - 2V_{XX}\right\| \le 10M^2\sqrt{\tau} \tag{5}$$

and

$$\left\|\frac{1}{m^2}\sum_{i=1}^m\sum_{j=1}^m (y_i - y_j)(\mathbf{x}_i - \mathbf{x}_j) - 2V_{XY}\right\| \le 12M^2\sqrt{\tau} \tag{6}$$

simultaneously, where $\tau = \frac{\log(8/\delta)}{m}$.

Proof: Let $\bar{\mathbf{x}} = \frac{1}{m}\sum_{i=1}^m \mathbf{x}_i$ and $\bar{y} = \frac{1}{m}\sum_{i=1}^m y_i$ be the corresponding sample means of X and Y.

Applying Lemma 7 to $\xi = X$ with $\tilde{\delta} = \frac{\delta}{4}$, we obtain

$$\left\|\frac{1}{m}\sum_{i=1}^m \mathbf{x}_i - \mu_X\right\| \le \frac{1}{2}M\left(\tau + \sqrt{8\tau + \tau^2}\right) \tag{7}$$

with probability at least $1 - \frac{\delta}{4}$. Applying Lemma 7 to $\xi = Y$ with $\tilde{\delta} = \frac{\delta}{4}$, we obtain

$$\left|\frac{1}{m}\sum_{i=1}^m y_i - \mu_Y\right| \le \frac{1}{2}M\left(\tau + \sqrt{8\tau + \tau^2}\right) \tag{8}$$

with probability at least $1 - \frac{\delta}{4}$. Recall that all $n \times n$ matrices form a Hilbert space under the Frobenius norm. Consider the matrix valued random variable $\xi = XX^\top$ which satisfies $\|\xi\|_F = \|X\|^2 \le M^2$. Applying Lemma 7 with $\tilde{\delta} = \frac{\delta}{4}$, we obtain

$$\left\| \frac{1}{m} \sum_{i=1}^{m} \mathbf{x}_i \mathbf{x}_i^\top - \mathbf{E}(XX^\top) \right\|_F \le \frac{1}{2} M^2 \left(\tau + \sqrt{8\tau + \tau^2} \right)$$

with probability at least $1 - \frac{\delta}{4}$. Since the operator norm is bounded by the Frobenius norm, we have

$$\left\| \frac{1}{m} \sum_{i=1}^{m} \mathbf{x}_i \mathbf{x}_i^\top - \mathbf{E}(XX^\top) \right\| \le \frac{1}{2} M^2 \left(\tau + \sqrt{8\tau + \tau^2} \right) \quad (9)$$

with probability at least $1 - \frac{\delta}{4}$. Applying Lemma 7 to $\xi = XY$ with $\tilde{\delta} = \frac{\delta}{4}$, we obtain

$$\left\| \frac{1}{m} \sum_{i=1}^{m} \mathbf{x}_i y_i - \mathbf{E}(XY) \right\| \le \frac{1}{2} M^2 \left(\tau + \sqrt{8\tau + \tau^2} \right) \quad (10)$$

with probability at least $1 - \frac{\delta}{4}$. Thus, (7)–(10) hold simultaneously with probability at least $1 - \delta$. (We have used the fact that for a sequence of k events $A_1, A_2, \ldots, A_k$, $\Pr(\bigcap_{i=1}^k A_i) = \Pr((\bigcup_{i=1}^k A_i^{\complement})^{\complement}) \ge 1 - \sum_{i=1}^k \Pr(A_i^{\complement})$.) What is left is to verify (5) and (6) from (7)–(10).

Let us first prove (5). Note that

$$\frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top = \frac{2}{m} \sum_{i=1}^{m} \mathbf{x}_i \mathbf{x}_i^\top - 2\bar{\mathbf{x}}\bar{\mathbf{x}}^\top.$$

Both terms on the right hand side are positive semidefinite matrices and their norms are no greater than $2M^2$. Thus,

$$\left\| \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top \right\| \le 2M^2.$$

This, together with Lemma 3, implies

$$\left\| \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top - 2V_{XX} \right\| \le 2M^2.$$

So (5) holds almost surely if $\tau > \frac{1}{25}$. When $\tau \le \frac{1}{25}$, by (7) and (9), we obtain

$$\begin{aligned}
&\left\| \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top - 2V_{XX} \right\| \\
&\le 2 \left\| \frac{1}{m} \sum_{i=1}^{m} \mathbf{x}_i \mathbf{x}_i^\top - \mathbf{E}(XX^\top) \right\| + 2 \left\| \bar{\mathbf{x}}\bar{\mathbf{x}}^\top - \mu_X \mu_X^\top \right\| \\
&\le M^2 \left(\tau + \sqrt{8\tau + \tau^2} \right) + 4M \|\bar{\mathbf{x}} - \mu_X\| \\
&\le 3M^2 \left(\tau + \sqrt{8\tau + \tau^2} \right) \\
&\le 3M^2 \sqrt{\tau} \left(\sqrt{\frac{1}{25}} + \sqrt{8 + \frac{1}{25}} \right) \\
&\le 10 M^2 \sqrt{\tau}.
\end{aligned}$$

This proves (5).

Now we turn to (6). The proof is quite similar. First note that the left hand side is bounded by $8M^2$ almost surely. So the inequality is always true when $\tau > 1$. When $\tau \le 1$, we need the fact that

$$\frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} (y_i - y_j)(\mathbf{x}_i - \mathbf{x}_j) = \frac{2}{m} \sum_{i=1}^{m} y_i \mathbf{x}_i - 2\bar{\mathbf{x}}\bar{y}.$$

By (7), (8) and (10), we obtain

$$\begin{aligned}
&\left\| \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} (y_i - y_j)(\mathbf{x}_i - \mathbf{x}_j) - 2V_{XY} \right\| \\
&\le 2 \left\| \frac{1}{m} \sum_{i=1}^{m} y_i \mathbf{x}_i - \mathbf{E}(XY) \right\| \\
&\quad + 2M \Big(\|\bar{\mathbf{x}} - \mu_X\| + |\bar{y} - \mu_Y| \Big) \\
&\le 3M^2 \left(\tau + \sqrt{8\tau + \tau^2} \right) \\
&\le 12 M^2 \sqrt{\tau}.
\end{aligned}$$

We finish the proof. ■

According to Lemma 8, we will adopt the notations

$$\hat{V}_{XX} = \frac{1}{2m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top$$

and

$$\hat{V}_{XY} = \frac{1}{2m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} (y_i - y_j)(\mathbf{x}_i - \mathbf{x}_j)$$

because they provide sample estimates of the covariance matrix V_{XX} and the covariance vector V_{XY}, respectively.

III. Uniform Bound for the Solution Path

In this section, we prove Theorem 1 which states that $\hat{\mathbf{w}}_t$ are uniformly bounded with large probability.

To simplify our presentation, we adopt the notation

$$\zeta_t(i,j) = (y_i - \hat{\mathbf{w}}_t^\top \mathbf{x}_i) - (y_j - \hat{\mathbf{w}}_t^\top \mathbf{x}_j)$$

for each $t \in \mathbb{N}$ in the sequel. The following proposition gives conditions for the solution $\hat{\mathbf{w}}_t$ to be uniformly bounded.

Proposition 9: Let $0 < \eta_t \le \frac{1}{2c_0 \lambda_{\max}}$ for all $t \ge 1$ and h is chosen such that

$$h \ge \left(\frac{2^{5p+4} c_p M^{6p+2}}{c_0 \lambda_{\min}^{2p+1}} \right)^{1/2p}. \quad (11)$$

If the sample $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$ satisfies

$$\|\hat{V}_{XX} - V_{XX}\| \le \frac{1}{6} \lambda_{\min}, \quad (12)$$

then

$$\|\hat{\mathbf{w}}_t\| \le \frac{3M^2}{\lambda_{\min}}.$$

Proof: By the definition of $\hat{V}_{XX}$ and $\hat{V}_{XY}$ and the fact $\Psi'_+(0) = -c_0$, we can write

$$\begin{aligned}
&\nabla R(\hat{\mathbf{w}}_{t-1}) \\
&= \frac{1}{m^2}\sum_{i=1}^m\sum_{j=1}^m \Psi'\left(\frac{\zeta_{t-1}^2(i,j)}{2h^2}\right)\zeta_{t-1}(i,j)(\mathbf{x}_i - \mathbf{x}_j) \\
&= \frac{1}{m^2}\sum_{i=1}^m\sum_{j=1}^m \left[\Psi'\left(\frac{\zeta_{t-1}^2(i,j)}{2h^2}\right) - \Psi'_+(0)\right]\zeta_{t-1}(i,j)(\mathbf{x}_i - \mathbf{x}_j) \\
&\quad - 2c_0\hat{V}_{XY} + 2c_0\left(\hat{V}_{XX} - V_{XX}\right)\hat{\mathbf{w}}_{t-1} + 2c_0 V_{XX}\hat{\mathbf{w}}_{t-1} \\
&:= Q_1 + Q_2 + Q_3 + 2c_0 V_{XX}\hat{\mathbf{w}}_{t-1}.
\end{aligned}$$

We prove the conclusion by induction. First it is obvious $\|\hat{\mathbf{w}}_0\| = 0 \le \frac{3M^2}{\lambda_{\min}}$. Assume $\|\hat{\mathbf{w}}_{t-1}\| \le \frac{3M^2}{\lambda_{\min}}$. We need to prove $\|\hat{\mathbf{w}}_t\| \le \frac{3M^2}{\lambda_{\min}}$.

By the definition of $\hat{\mathbf{w}}_t$, we have

$$\begin{aligned}
\hat{\mathbf{w}}_t &= \hat{\mathbf{w}}_{t-1} - \eta_t\nabla R(\hat{\mathbf{w}}_{t-1}) \\
&= (I - 2\eta_t c_0 V_{XX})\hat{\mathbf{w}}_{t-1} - \eta_t(Q_1 + Q_2 + Q_3).
\end{aligned}$$

Since $\eta_t \le \frac{1}{2c_0\lambda_{\max}}$, the matrix $I - 2\eta_t c_0 V_{XX}$ is positive semidefinite. We have

$$\|(I - 2\eta_t c_0 V_{XX})\hat{\mathbf{w}}_{t-1}\| \le (1 - 2\eta_t c_0\lambda_{\min})\frac{3M^2}{\lambda_{\min}}.$$

Since X and Y are bounded by M almost surely and $\|\hat{\mathbf{w}}_{t-1}\| \le \frac{3M^2}{\lambda_{\min}}$, we have

$$\begin{aligned}
\|\zeta_{t-1}(i,j)\| &\le 2M(1 + \|\hat{\mathbf{w}}_{t-1}\|) \\
&\le 2M\left(1 + \frac{3M^2}{\lambda_{\min}}\right) \le \frac{8M^3}{\lambda_{\min}},
\end{aligned}$$

where we have used the fact $\lambda_{\min} \le \lambda_{\max} \le M^2$. This together with the Lipschitz assumption on Ψ' gives

$$\begin{aligned}
\|Q_1\| &\le \frac{1}{m^2}\sum_{i=1}^m\sum_{j=1}^m c_p\left(\frac{|\zeta_{t-1}(i,j)|^2}{2h^2}\right)^p |\zeta_{t-1}(i,j)|(2M) \\
&\le 2^{5p+4}c_p M^{6p+4}\lambda_{\min}^{-2p-1}h^{-2p}.
\end{aligned}$$

Under the condition (11), we have $\|Q_1\| \le c_0M^2$. It is easy to verify $\|Q_2\| \le 4c_0M^2$. As for Q_3, under the condition (12), we have $\|Q_3\| \le c_0M^2$. Therefore, we have

$$\|\hat{\mathbf{w}}_t\| \le (1 - 2\eta_t c_0\lambda_{\min})\frac{3M^2}{\lambda_{\min}} + 6\eta_t c_0 M^2 \le \frac{3M^2}{\lambda_{\min}}.$$

This finishes the proof. ■

Now Theorem 1 can be proved by combining Proposition 9 and Lemma 8.

Proof of Theorem 1: By Lemma 8,

$$\|\hat{V}_{XX} - V_{XX}\| \le 5M^2\sqrt{\frac{\log(8/\delta)}{m}}$$

with probability $1-\delta$. Thus, when $m \ge \frac{900M^4\log(8/\delta)}{\lambda_{\min}^2}$, the condition (12) holds with probability at least $1-\delta$. By Proposition 9, we obtain the desired conclusion. ■

IV. One Step Error Analysis

In this section we show that the estimation error decreases after each iteration step, which plays an essential role for the proof of Theorem 2.

Proposition 10: Let $0 < \eta_t \le \frac{\lambda_{\min}}{12c_0\lambda_{\max}^2}$ for all $t \ge 1$ and $h \ge \left(\frac{2^{5p+4}c_pM^{6p+4}}{c_0\lambda_{\min}^{2p+1}}\right)^{1/2p}$. If the sample $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$ satisfies

$$\|\hat{V}_{XX} - V_{XX}\| \le 5M^2\sqrt{\frac{\log(8/\delta)}{m}} \le \frac{1}{6}\lambda_{\min} \tag{13}$$

and

$$\|\hat{V}_{XY} - V_{XY}\| \le 6M^2\sqrt{\frac{\log(8/\delta)}{m}}, \tag{14}$$

then

$$\begin{aligned}
\|\hat{\mathbf{w}}_t - \mathbf{w}_*\|^2 &\le (1 - \eta_t c_0\lambda_{\min})\|\hat{\mathbf{w}}_{t-1} - \mathbf{w}_*\|^2 \\
&\quad + \eta_t C\left(h^{-4p} + \frac{\log(8/\delta)}{m}\right)
\end{aligned}$$

for some constant C independent of m, δ, or h.

Proof: By the definition of $\hat{\mathbf{w}}_t$, we have

$$\begin{aligned}
\|\hat{\mathbf{w}}_t - \mathbf{w}_*\|^2 &= \|\hat{\mathbf{w}}_{t-1} - \mathbf{w}_*\|^2 \\
&\quad - 2\eta_t(\hat{\mathbf{w}}_{t-1} - \mathbf{w}_*)^\top\nabla R(\hat{\mathbf{w}}_{t-1}) \\
&\quad + \eta_t^2\|\nabla R(\hat{\mathbf{w}}_{t-1})\|^2.
\end{aligned} \tag{15}$$

The key to prove Proposition 10 is to estimate $\nabla R(\hat{\mathbf{w}}_{t-1})$ appropriately. For this purpose we write

$$\begin{aligned}
&\nabla R(\hat{\mathbf{w}}_{t-1}) \\
&= \frac{1}{m^2}\sum_{i=1}^m\sum_{j=1}^m \Psi'\left(\frac{\zeta_{t-1}^2(i,j)}{2h^2}\right)\zeta_{t-1}(i,j)(\mathbf{x}_i - \mathbf{x}_j) \\
&= \frac{1}{m^2}\sum_{i=1}^m\sum_{j=1}^m \left[\Psi'\left(\frac{\zeta_{t-1}^2(i,j)}{2h^2}\right) - \Psi'_+(0)\right]\zeta_{t-1}(i,j)(\mathbf{x}_i - \mathbf{x}_j) \\
&\quad - \frac{c_0}{m^2}\sum_{i=1}^m\sum_{j=1}^m \zeta_{t-1}(i,j)(\mathbf{x}_i - \mathbf{x}_j) \\
&= \frac{1}{m^2}\sum_{i=1}^m\sum_{j=1}^m \left[\Psi'\left(\frac{\zeta_{t-1}^2(i,j)}{2h^2}\right) - \Psi'_+(0)\right]\zeta_{t-1}(i,j)(\mathbf{x}_i - \mathbf{x}_j) \\
&\quad - 2c_0\left\{\left(\hat{V}_{XY} - V_{XY}\right) - \left(\hat{V}_{XX} - V_{XX}\right)\hat{\mathbf{w}}_{t-1}\right\} \\
&\quad + 2c_0 V_{XX}(\hat{\mathbf{w}}_{t-1} - \mathbf{w}_*) \\
&:= D_1 + D_2 + D_3,
\end{aligned}$$

where we have used the fact $V_{XY} = V_{XX}\mathbf{w}_*$ obtained in Lemma 4.

Note that all the conditions for Proposition 9 hold. So we have the bound $\|\hat{\mathbf{w}}_t\| \le \frac{3M^2}{\lambda_{\min}}$ for all t. For D_1, by the Lipschitz condition of Ψ' and the bound for $\hat{\mathbf{w}}_{t-1}$, as we have shown in the proof of Proposition 9, we have

$$\|D_1\| \le 2^{5p+4}c_pM^{6p+4}\lambda_{\min}^{-2p-1}h^{-2p}. \tag{16}$$

For D_2, by (13), (14) and the bound for $\hat{\mathbf{w}}_{t-1}$, we have

$$\begin{aligned}\|D_2\| &\leq 2c_0\left(\|\hat{V}_{XY}-V_{XY}\|+\|\hat{V}_{XX}-V_{XX}\|\|\hat{\mathbf{w}}_{t-1}\|\right)\\ &\leq 2c_0\left(6M^2+\frac{15M^4}{\lambda_{\min}}\right)\sqrt{\frac{\log(8/\delta)}{m}}\\ &\leq 42c_0M^4\lambda_{\min}^{-1}\sqrt{\frac{\log(8/\delta)}{m}}. \qquad (17)\end{aligned}$$

Now we can estimate the second term on the right of (15). For notational simplicity let

$$\tilde{C}=\max\{2^{5p+4}c_pM^{6p+4}\lambda_{\min}^{-2p-1},\ 42c_0M^4\lambda_{\min}^{-1}\}.$$

By (16) and the elementary inequality $ab\leq\frac{a^2}{2}+\frac{b^2}{2}$, we have

$$\begin{aligned}&\left|(\hat{\mathbf{w}}_{t-1}-\mathbf{w}_*)^\top D_1\right|\\ &\leq\frac{c_0\lambda_{\min}}{2}\|\hat{\mathbf{w}}_{t-1}-\mathbf{w}_*\|^2+\frac{1}{2c_0\lambda_{\min}}\|D_1\|^2.\\ &\leq\frac{c_0\lambda_{\min}}{2}\|\hat{\mathbf{w}}_{t-1}-\mathbf{w}_*\|^2+\frac{\tilde{C}^2}{2c_0\lambda_{\min}}h^{-4p}.\end{aligned}$$

Similarly, by (17), we have

$$\begin{aligned}&\left|(\hat{\mathbf{w}}_{t-1}-\mathbf{w}_*)^\top D_2\right|\\ &\leq\frac{c_0\lambda_{\min}}{2}\|\hat{\mathbf{w}}_{t-1}-\mathbf{w}_*\|^2+\frac{1}{2c_0\lambda_{\min}}\|D_2\|^2\\ &\leq\frac{c_0\lambda_{\min}}{2}\|\hat{\mathbf{w}}_{t-1}-\mathbf{w}_*\|^2+\frac{\tilde{C}^2}{2c_0\lambda_{\min}}\frac{\log(8/\delta)}{m}.\end{aligned}$$

These together with the fact that

$$\begin{aligned}(\hat{\mathbf{w}}_{t-1}-\mathbf{w}_*)^\top D_3&=2c_0(\hat{\mathbf{w}}_{t-1}-\mathbf{w}_*)^\top V_{XX}(\hat{\mathbf{w}}_{t-1}-\mathbf{w}_*)\\ &\geq 2c_0\lambda_{\min}\|\hat{\mathbf{w}}_{t-1}-\mathbf{w}_*\|^2\end{aligned}$$

enable us to obtain

$$\begin{aligned}&-2\eta_t(\hat{\mathbf{w}}_{t-1}-\mathbf{w}_*)^\top\nabla R(\hat{\mathbf{w}}_{t-1})\\ &\leq -2\eta_tc_0\lambda_{\min}\|\hat{\mathbf{w}}_{t-1}-\mathbf{w}_*\|^2\\ &\quad+\frac{\eta_t\tilde{C}^2}{c_0\lambda_{\min}}\left(h^{-4p}+\frac{\log(8/\delta)}{m}\right). \qquad (18)\end{aligned}$$

We turn to estimate the last term on the right hand side of (15). We need the trivial bound

$$\|D_3\|\leq 2c_0\lambda_{\max}\|\hat{\mathbf{w}}_{t-1}-\mathbf{w}_*\|. \qquad (19)$$

Combining the estimates in (16), (17), and (19), we have

$$\begin{aligned}&\eta_t^2\|\nabla R(\hat{\mathbf{w}}_{t-1})\|^2\\ &\leq 3\eta_t^2(\|D_1\|^2+\|D_2\|^2+\|D_3\|^2)\\ &\leq 12\eta_t^2c_0^2\lambda_{\max}^2\|\hat{\mathbf{w}}_{t-1}-\mathbf{w}_*\|^2\\ &\quad+3\eta_t^2\tilde{C}^2\left(h^{-4p}+\frac{\log(8/\delta)}{m}\right)\\ &\leq\eta_tc_0\lambda_{\min}\|\hat{\mathbf{w}}_{t-1}-\mathbf{w}_*\|^2\\ &\quad+\frac{1}{4c_0\lambda_{\min}}\eta_t\tilde{C}^2\left(h^{-4p}+\frac{\log(8/\delta)}{m}\right), \qquad (20)\end{aligned}$$

where we used the assumption $\eta_t\leq\frac{\lambda_{\min}}{12c_0\lambda_{\max}^2}\leq\frac{1}{12c_0\lambda_{\min}}$.

Let $C=\frac{5\tilde{C}^2}{4c_0\lambda_{\min}}$. Plugging the estimates in (18) and (20) into (15), we obtain the desired conclusion. ■

V. Error Bounds and Convergence Rates

To prove Theorem 2, we need two lemmas from [22].

Lemma 11: For $v\in(0,1]$ and $\theta\in[0,1]$,

$$\sum_{t=1}^T\frac{1}{t^\theta}\prod_{j=t+1}^T\left(1-\frac{v}{j^\theta}\right)\leq\frac{3}{v}.$$

Lemma 12: For any $0\leq\ell<T$ and $0<\theta<1$, there holds

$$\sum_{t=\ell+1}^T t^{-\theta}\geq\frac{1}{1-\theta}\left[(T+1)^{1-\theta}-(\ell+1)^{1-\theta}\right].$$

Proof of Theorem 2: For a sample satisfying the conditions (13) and (14), Proposition 10 states that

$$\begin{aligned}\|\hat{\mathbf{w}}_t-\mathbf{w}_*\|^2&\leq(1-\eta_tc_0\lambda_{\min})\|\hat{\mathbf{w}}_{t-1}-\mathbf{w}_*\|^2\\ &\quad+\eta_tC\left(h^{-4p}+\frac{\log(8/\delta)}{m}\right)\end{aligned}$$

for all t. Applying this estimate iteratively we obtain

$$\begin{aligned}\|\hat{\mathbf{w}}_T-\mathbf{w}_*\|^2&\leq\|\mathbf{w}_*\|^2\prod_{t=1}^T(1-\eta_tc_0\lambda_{\min})\\ &\quad+C\left(h^{-4p}+\frac{\log(8/\delta)}{m}\right)\sum_{t=1}^T\prod_{j=t+1}^T(1-\eta_jc_0\lambda_{\min})\eta_t.\end{aligned}$$

Since $\eta_t=\eta t^{-\theta}$, by the elementary inequality $1-u\leq\exp(-u)$ and Lemma 12 with $\ell=0$, we have

$$\begin{aligned}\prod_{t=1}^T(1-\eta_tc_0\lambda_{\min})&\leq\exp\left(-c_0\lambda_{\min}\sum_{t=1}^T\eta_t\right)\\ &\leq\exp\left(\frac{\eta c_0\lambda_{\min}\left(1-(T+1)^{1-\theta}\right)}{1-\theta}\right)\\ &\leq\exp\left(\frac{\eta c_0\lambda_{\min}\left(1-T^{1-\theta}\right)}{1-\theta}\right).\end{aligned}$$

Lemma 11 with $v = \eta c_0 \lambda_{\min}$ yields

$$\sum_{t=1}^{T} \prod_{j=t+1}^{T} (1 - \eta_j c_0 \lambda_{\min}) \eta_t$$

$$= \eta \sum_{t=1}^{T} \frac{1}{t^\theta} \prod_{j=t+1}^{T} \left(1 - \frac{\eta c_0 \lambda_{\min}}{j^\theta}\right) \leq \frac{3}{c_0 \lambda_{\min}}.$$

Therefore,

$$\begin{aligned}\|\hat{\mathbf{w}}_T - \mathbf{w}_*\|^2 &\leq \|\mathbf{w}_*\|^2 \exp\left(\frac{\eta c_0 \lambda_{\min}(1 - T^{1-\theta})}{1-\theta}\right) \\ &\quad + \frac{3C}{c_0 \lambda_{\min}} \left(h^{-4p} + \frac{\log(8/\delta)}{m}\right) \\ &\leq C' \left\{\exp\left(-\frac{\eta c_0 \lambda_{\min} T^{1-\theta}}{1-\theta}\right) + h^{-4p} + \frac{\log(8/\delta)}{m}\right\},\end{aligned}$$

where $C' = \|\mathbf{w}_*\|^2 \exp\left(\frac{\eta c_0 \lambda_{\min}}{1-\theta}\right) + \frac{3C}{c_0 \lambda_{\min}}$. The proof of Theorem 2 is completed after noticing that the conditions (13) and (14) hold with probability at least $1 - \delta$, as are guaranteed by Lemma 8. ■

VI. Simulations

In this section we study the empirical performance of the gradient descent method for MEE by simulations and compare it with our theoretical analysis. On one hand we expect the theoretical analysis provides some guidance to the empirical implementation. On the other hand, since the theoretical analysis is based on upper bounds which might be far from tight, it is important to understand the gap between the theory and empirical applications.

In the simulation, let $\mathbf{x} \in \mathbb{R}^{10}$ and the model be defined by $Y = \mathbf{w}_*^\top \mathbf{x} + \epsilon$ with $\mathbf{w}_* = [1\, -1\, 1\, -1\, 1\, -1\, 1\, -1\, 1\, -1]^\top$ and $\mathbf{x} \sim N(0, I_{10})$. We consider two types of noise. The first type is the Gaussian noise $\epsilon \sim N(0, c\mathbf{w}_*^\top \mathbf{x})$ for each given $\mathbf{x}$. The second type is the generalized Gaussian noise with a probability density function $f(\epsilon) \propto \exp(-|c\epsilon|^{0.3})$. It is a typical heavy tailed distribution and has been employed in [2] to explore the effectiveness of a minimum total error entropy algorithm. For both noise models we select the constant c so that the signal to noise ratio equal to one. As indicated by Theorem 2, a small constant step size can be used to guarantee convergence and the scaling parameter should be large enough. In our simulations we have chosen $\eta_t = \sqrt{0.005\pi}$ (so that it satisfies the condition in Theorem 2) and $h = 10$. We let the sample size m vary from 50 to 500. The simulation results based on 100 repeated experiments were reported in Figs. 1 and 2 for the two noise models, respectively.

In Figs. 1(a) and 2(a) we report the change of the mean squared error as the number of iterations increases. In Figs. 1(b) and 2(b) we compare the mean squared error with iteration to convergence and the mean squared error with optimal iteration (i.e., the number of iterations that leads to minimal mean squared error). In Figs. 1(c) and 2(c) we compare the number of iterations to convergence and the optimal number of iterations. Similar results are seen regardless of the noise types. All these results indicate that the optimal solution can be achieved by early stopping the gradient descent iteration while further increasing the number of iterations may hurt the learning performance. The performance degradation is notable in a small sample setting while negligible in a large sample setting. Therefore, early stopping is not only sufficient but also necessary when the sample size is small. An interesting observation is that the number of iterations to convergence and the optimal number of iterations tend to coincide when the sample size is large. A plausible explanation is that when the sample size is large, the empirical risk approximates the expected risk well and thus $\hat{\mathbf{w}}$ approximates $\mathbf{w}_*$ well. So the optimal solution does require $\hat{\mathbf{w}}_t$ to converge to $\hat{\mathbf{w}}$. We observed that the number of iterations to convergence decreases as the sample size increases. Although it

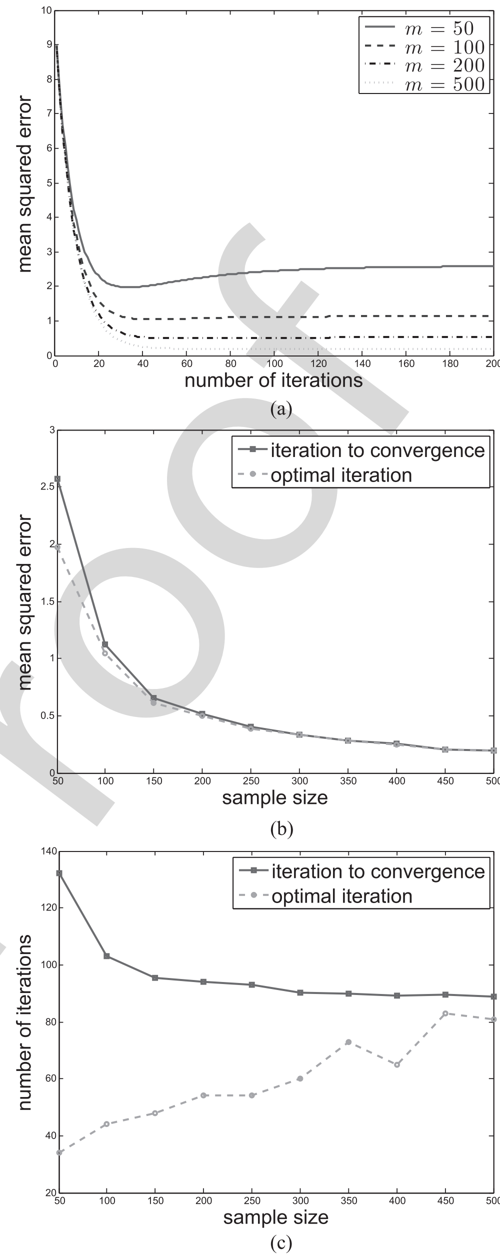


Fig. 1. Simulation results for Gaussian noise.

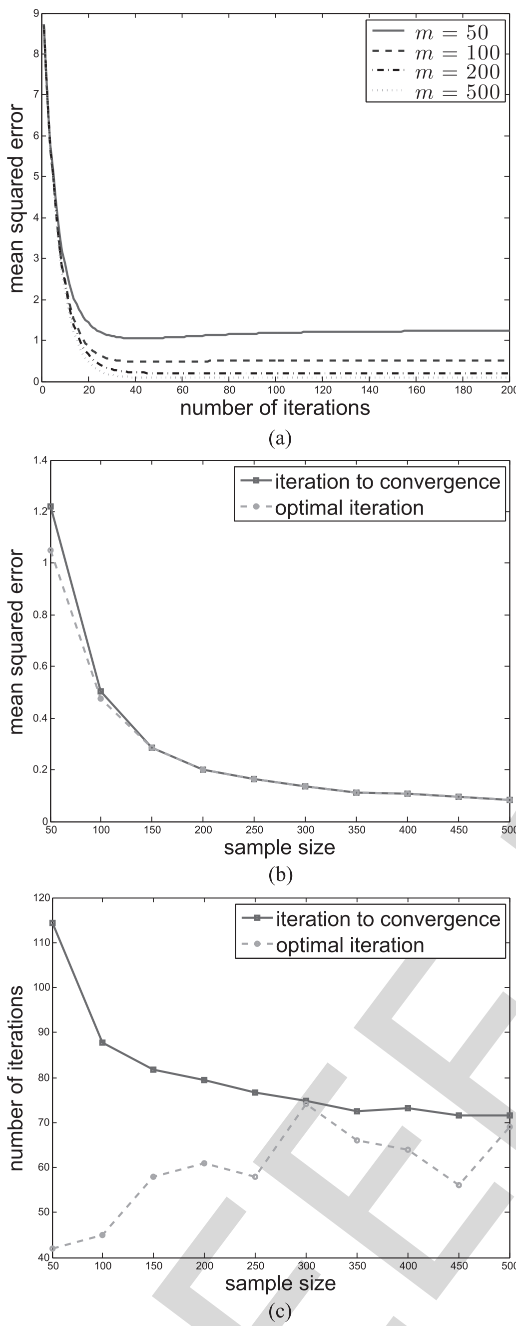


Fig. 2. Simulation results for generalized Gaussian noise.

does not contradict our analysis, it does seem surprising. We do not have an explanation at the moment and would leave it for future study.

Recall that the upper bound in Theorem 2 implies the sufficiency of early stopping for an optimal solution in a large sample size setting. Moreover, although it is hard to verify the optimal number of iterations is of order $O(\log m)$, it does increase very slowly according Figs. 1(c) and 2(c).

Theorem 2 also provides useful insight on the choice of the step size. The upper bound of η can be estimated from the sample. Choosing η around half of the upper bound usually works well in practice. However, there is a gap between the theoretical analysis and empirical applications regarding the choice of the scaling parameter h. The theoretical lower bound on h is too restrictive. In practice it is found that a moderately large h is good and a very large h is not necessary.

VII. Discussions

To derive our results, we have assumed that the covariance matrix V_{XX} of the input variable X is non-degenerate. This condition, however, may not be true in many situations. A very typical model is the classical linear regression model where an intercept is included:

$$Y = \beta_0 + \beta_1^\top Z + \epsilon \tag{21}$$

with $\beta_0 \in \mathbb{R}$, $\beta_1 \in \mathbb{R}^n$, and Z a vector valued random variable containing n explanatory variables. In this case, $\mathbf{w}_* = [\beta_0\ \beta_1^\top]^\top$ and $X = [1\ Z^\top]^\top$. Note that

$$V_{XX} = \begin{pmatrix} 0 & 0 \\ 0 & V_{ZZ} \end{pmatrix}.$$

So it is always degenerate.

When V_{XX} is degenerate, we cannot prove the convergence of $\hat{\mathbf{w}}_t$ to $\mathbf{w}_*$. Instead, we need to consider their projections onto the principal component space. Let U denote the subspace of $\mathbb{R}^n$ spanned by the principal components associated to the positive eigenvalues and P_U the projection onto U. Let $\lambda_{\min}$ denote the smallest positive eigenvalue. We can prove

$$\begin{aligned}\lambda_{\min}\|P_U(\mathbf{w} - \mathbf{w}_*)\|^2 &\le (\mathbf{w} - \mathbf{w}_*)^\top V_{XX}(\mathbf{w} - \mathbf{w}_*) \\ &\le \lambda_{\max}\|P_U(\mathbf{w} - \mathbf{w}_*)\|^2.\end{aligned}$$

By this relationship and the techniques developed in this paper and [23], we can prove the convergence of $P_U(\hat{\mathbf{w}}_t)$ to $P_U(\mathbf{w}_*)$. This guarantees the variance of $\hat{\mathbf{w}}_t^\top X$ converges to the variance $\mathbf{w}_*^\top X$ and thus $\hat{\mathbf{w}}_t^\top X$ plus an appropriate intercept provides good predictive performance. In the model (21), if V_{ZZ} is positive definite, we see $\hat{\mathbf{w}}_t$ estimates the slope coefficients β_1.

As for the implementation of the algorithm, we remark that if V_{XX} is non-degenerate, the initial point is not necessarily chosen as $\hat{\mathbf{w}}_0 = 0$. The convergence holds true for any starting point. If V_{XX} is degenerate, the convergence of $\hat{\mathbf{w}}_t$ to $\mathbf{w}_*$ can be proved if the starting value is in the principal components space U. Actually, since $\mathbf{x}_i - \mathbf{x}_j$ is in U, all $\hat{\mathbf{w}}_t$ are in U. Thus, $P_U(\hat{\mathbf{w}}_t) = \hat{\mathbf{w}}_t$ and the convergence of $\hat{\mathbf{w}}_t$ to $P_U(\mathbf{w}_*)$ is exactly the convergence of $P_U(\hat{\mathbf{w}}_t)$ to $P_U(\mathbf{w}_*)$. However, if the starting point has a nonzero components normal to U, it will never diminish during the iteration process.

We have focused on linear regression models in this paper. Note that the MEE principle can be extended to nonlinear regression by the kernel trick [14], [17]. Regularization theory plays an important role to overcome the overfitting problem in this case. It would be interesting to study the use of gradient descent for the kernel MEE method in the future.

Acknowledgment

The authors thank the anonymous reviewers for their valuable comments.

REFERENCES

[1] D. Erdogmus and J. C. Principe, "Comparison of entropy and mean square error criteria in adaptive system training using higher order statistics," in *Proc. 2nd Int. Workshop Independent Compon. Anal. Blind Signal Separation*, 2000, pp. 75–90.

[2] P. Shen and C. Li, "Minimum total error entropy method for parameter estimation," *IEEE Trans. Signal Process.*, vol. 63, no. 15, pp. 4079–4090, Aug. 2015.

[3] D. Erdogmus, K. Hild II, and J. C. Principe, "Blind source separation using Rényi's α-marginal entropies," *Neurocomput.*, vol. 49, pp. 25–38, 2002.

[4] D. Erdogmus and J. C. Principe, "Convergence properties and data efficiency of the minimum error entropy criterion in adaline training," *IEEE Trans. Signal Process.*, vol. 51, no. 7, pp. 1966–1978, Jul. 2003.

[5] E. Gokcay and J. C. Principe, "Information theoretic clustering," *IEEE Trans. Pattern Anal. Mach. Learn.*, vol. 24, no. 2, pp. 158–171, Feb. 2002.

[6] L. M. Silva, J. Marques de Sá, and L. A. Alexandre, "Neural network classification using Shannon's entropy," in *Proc. Eur. Symp. Artif. Neural Netw.*, 2005, pp. 217–222.

[7] L. M. Silva, J. Marques de Sá, and L. A. Alexandre, "The MEE principle in data classification: A perceptron-based analysis," *Neural Comput.*, vol. 22, pp. 2698–2728, 2010.

[8] B. Chen, P. Zhu, and J. C. Principe, "Survival information potential: A new criterion for adaptive system training," *IEEE Trans. Signal Process.*, vol. 60, no. 3, pp. 1184–1194, Mar. 2012.

[9] F. Cucker and D. X. Zhou, *Learning Theory: An Approximation Theory Viewpoint*. Cambridge, U.K.: Cambridge Univ. Press, 2007.

[10] I. Steinwart, D. Hush, and C. Scovel, "Optimal rates for regularized least squares regression," in S. Dasgupta and A. Klivans, eds., *Proc. 22nd Annu. Conf. Learn. Theory*, 2009, pp. 79–93.

[11] Y. Yao, L. Rosasco, and A. Caponnetto, "On early stopping in gradient descent learning," *Constructive Approx.*, vol. 26, no. 2, pp. 289–315, 2007.

[12] E. Parzen, "On the estimation of a probability density function and the mode," *Ann. Math. Statist.*, vol. 33, pp. 1049–1051, 1962.

[13] T. Hu, J. Fan, Q. Wu, and D.-X. Zhou, "Learning theory approach to a minimum error entropy criterion," *J. Mach. Learn. Res.*, vol. 14, pp. 377–397, 2013.

[14] T. Hu, J. Fan, Q. Wu, and D.-X. Zhou, "Regularization schemes for minimum error entropy principle," *Anal. Appl.*, vol. 13, no. 4, pp. 437–455, 2015.

[15] J. Fan, T. Hu, Q. Wu, and D.-X. Zhou, "Consistency analysis of an empirical minimum error entropy algorithm," *Appl. Comput. Harmonic Anal.*, vol. 41, pp. 164–189, 2016.

[16] B. Chen and J. C. Principe, "Some further results on the minimum error entropy estimation," *Entropy*, vol. 14, no. 5, pp. 966–977, 2012.

[17] J. C. Principe, *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*. New York, NY, USA: Springer-Verlag, 2010.

[18] B. Chen and J. C. Principe, "Stochastic gradient algorithm under (h, ϕ)-entropy criterion," *Circuits, Syst., Signal Process.*, vol. 26, no. 6, pp. 941–960, 2007.

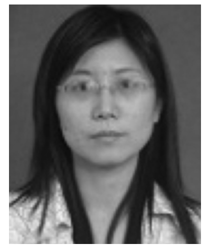
[19] Z. Wu, S. Peng, W. Ma, B. Chen, and J. C. Principe, "Minimum error entropy algorithms with sparsity penalty constraints," *Entropy*, vol. 17, no. 5, pp. 3419–3437, 2015.

[20] B. Chen, Y. Zhu, and J. Hu, "Mean-square convergence analysis of adaline training with minimum error entropy criterion," *IEEE Trans. Neural Netw.*, vol. 21, no. 7, pp. 1168–1179, Jul. 2010.

[21] S. Smale and D. X. Zhou, "Learning theory estimates via integral operators and their approximations," *Construct. Approx.*, vol. 26, pp. 153–172, 2007.

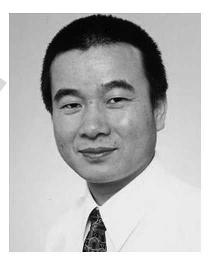
[22] Y. Ying and D.-X. Zhou, "Online regularized classification algorithms," *IEEE Trans. Inf. Theory*, vol. 52, no. 11, pp. 4775–4788, Nov. 2006.

[23] J. Lin and D.-X. Zhou, "Learning theory of randomized Kaczmarz algorithm," *J. Mach. Learn. Res.*, vol. 16, pp. 3341–3365, 2015.



Ting Hu received the B.S. degree in computational mathematics in 2004 and the Ph.D. degree in mathematics in 2009, both from Wuhan University, Hubei, China. From 2009 to 2011, she was a Research Fellow at the City University of Hong Kong, Hong Kong, China. She is currently an Associate Professor in the School of Mathematics and Statistics, Wuhan University, Hubei, China. Her research interests include learning theory, machine learning, wavelet analysis, and applications.

Qiang Wu received the Ph.D. degree in mathematics from the City University of Hong Kong, Hong Kong, China, in 2005. He was with Duke University, Michigan State University, and the University of Liverpool before joinedMiddle Tennessee State University in 2011. He is currently an Associate Professor of mathematics and actuarial science in the Department of Mathematical Sciences and a Faculty Member of the Computational Science Ph.D. program. His research interests include computational harmonic analysis, statistical learning theory, high-dimensional data mining, and their applications.



Ding-Xuan Zhou received the B.Sc. and Ph.D. degrees in mathematics from Zhejiang University, Hangzhou, China, in 1988 and 1991, respectively. He joined City University of Hong Kong as a Research Assistant Professor in 1996, and is currently the Chair Professor in the Department of Mathematics. His research interests include learning theory, data science, wavelet analysis, and approximation theory. He has published more than 100 research papers and is serving on the editorial board of more than ten international journals, and is an Editor-in-Chief of the journal *Analysis and Applications*. He received a Research Fund for Distinguished Young Scholars from the National Science Foundation of China in 2005 and a Humboldt Research Fellowship in 1993, and was rated by Thomson Reuters highly-cited researcher in 2014 and 2015. He has co-organised more than 20 international conferences and conducted more than 20 research grants.