Kernel Gradient Descent Algorithm for Information Theoretic Learning

Ting Hu
School of Mathematics and Statistics, Wuhan University,
Wuhan 430072, China.
Email: tinghu@whu.edu.cn

Qiang Wu
Department of Mathematical Sciences, Middle Tennessee State University,
Murfreesboro, TN 37132, USA.
Email: qwu@mtsu.edu

Ding-Xuan Zhou
School of Data Science, Department of Mathematics,
Liu Bie Ju Center for Mathematical Sciences, City University of Hong Kong,
Kowloon, Hong Kong.
Email: mazhou@cityu.edu.hk

Abstract

Information theoretic learning is a learning paradigm that uses concepts of entropies and divergences from information theory. A variety of signal processing and machine learning methods fall into this framework. Minimum error entropy principle is a typical one amongst them. In this paper, we study a kernel version of minimum error entropy methods that can be used to find nonlinear structures in the data. We show that the kernel minimum error entropy can be implemented by kernel based gradient descent algorithms with or without regularization. Convergence rates for both algorithms are deduced.

1 Introduction

Information theoretical learning (ITL) refers to a framework of learning methods that use concepts of entropies and divergences from information theory to substitute the conventional statistical descriptors of variances and covariances. It becomes an important research topic in signal processing and machine learning as many algorithms have been developed within this framework and many applications domains have been discovered. In the literature, the study of ITL has mostly focused on linear models. Kernel based ITL was introduced to the minimum error entropy principle in [19,24] to deal with nonlinear models. The purpose of this paper is to study a kernel based gradient descent algorithm for ITL with a focus on the kernel minimum error entropy for regression.

Minimum error entropy (MEE) might be the most important principle in the ITL framework. In the context of regression analysis, it serves as an important alternative to the classical least square method and has attracted continuous attention for more than a decade since it was introduced. The least squares method relies only on the variance of the error, so it falls into the second-order statistics and its optimality depends heavily on the assumption of Gaussianity. In contrast, entropy is a function defined on the error probability density function and all moments of the error are constrained when entropy is minimized. Thus, MEE belongs to a type of high order methods and has the robustness to deal with non-Gaussian models or heavy outliers. Because of its robustness and ability to deal with non-Gaussian impulse noises, MEE methods have been successfully applied in a variety of applications including signal processing, regression analysis, feature selection, and data clustering. There are a vast of literatures exploring their application domains and devoting on their computational and mathematical properties; see [5–7,12–16,18,19,26–28,33] and the references therein.

In regression analysis, MEE is motivated by minimizing the information of the prediction error and thus maintaining the useful information by the predictor as much as possible. Assume $X \in \mathbb{R}^n$ is a vector of explanatory variables, $Y \in \mathbb{R}$ is the response variable, and they are linked by

$$Y = f^*(X) + \epsilon, \qquad \mathbf{E}(\epsilon|X) = 0,$$

where f^* is a target function and ϵ is the noise. When a function f is used as a predictor, the error variable is $E = Y - f(X)$. Let p_E denote its probability density function. Then Shannon's entropy is defined by

$$H_S(f) = -\mathbf{E}[\log(p_E)]$$

and Rényi's entropy of order $\alpha > 1$ is

$$H_\alpha(f) = -\frac{1}{1-\alpha} \log\left(\mathbf{E}[p_E^{\alpha-1}]\right).$$

Recall the traditional least squares method minimizes the mean squared error, which is the second moment of error variable E. It is optimal for Gaussian noise but suboptimal for general non-Gaussian noise. Error entropy is a functional defined on the probability density of the error variable and takes information of all moments into account. It is an efficient measure to estimate the learning ability of the predictor when non-Gaussian or impulse noise is involved. Thus MEE may work well in these situations.

In the literature, most MEE methods have been designed for linear models. They are usually implemented by gradient descent algorithms. The convergence has been studied in [7, 20]. But these methods may not be applied to analysis of data with nonlinear structures. There is a necessity to develop MEE methods for nonlinear models so that the algorithms can deal with nonlinearity in the data and simultaneously preserve the advantages of robustness and ability to deal with non-Gaussian noises. As reproducing kernel Hilbert spaces are effective tools to represent nonlinear features via feature mappings, we can use the so-called "kernel trick" to extend MEE to nonlinear settings. MEE methods in reproducing kernel Hilbert spaces were introduced in [19, 24] and the consistency of a regularization scheme was investigated in [19]. Its implementation, however, has not been addressed. Recall that there are usually two classes of implementation algorithms for kernel methods. One is to write the kernel method as a finite dimensional optimization problem by using a representer theorem. Solving the optimization problem could be very challenging for large scale data. The other is to use the gradient descent or stochastic gradient descent in Hilbert spaces. They have the advantages of being implementable in the big data setting.

The purpose of this paper is to study the convergence of gradient descent algorithms when they are used to implement kernel MEE methods. Note that the convergence of gradient descent algorithms for linear MEE [7, 20] highly depends on the fact that the covariance matrix of the explanatory variables is either invertible or has a smallest positive eigenvalue. In kernel MEE, the reproducing kernel Hilbert space may be infinite dimensional. The eigenvalues of the kernel covariance operator decay to zero. So the kernel covariance operator is neither invertible nor has its positive eigenvalues lower bounded away from zero. Compared to traditional kernel learning methods such as regularization kernel networks and support vector machines where the loss functions are convex, the loss function of MEE is non-convex, which makes the analysis of MEE methods essentially difficult. Therefore, the convergence analysis of gradient descent algorithms for kernel MEE is not an easy extension of that for linear MEE or traditional kernel methods, though those studies may provide useful insights and techniques.

The rest of the paper is organized as follows. In Section 2 we describe two gradient descent algorithms for kernel MEE and state our convergence results. In the first algorithm, regularization is adopted to control the computational complexity. The second algorithm does not involve regularization explicitly but adopts an early stopping to play the role of regularization. The proofs of the convergence results are given in Section 3 and in Section 4 for the two algorithms respectively.

2 Algorithms and main results

In this section we first describe the MEE kernel gradient descent algorithms and state their convergence rates. Then we compare our results with those in the literature.

2.1 kernel gradient descent algorithms and their convergence

Throughout this paper, we assume the sample space of X is a compact subset $\mathcal{X} \subset \mathbb{R}^n$ and the sample space of Y is a bounded subset $\mathcal{Y} \subset \mathbb{R}$. They are also called the input space and output space, respectively. Let ρ denote a joint probability measure on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, $\rho_{\mathcal{X}}$ be the marginal distribution of ρ on $\mathcal{X}$ and $\rho(\cdot|x)$ the conditional distribution of ρ for given $x \in \mathcal{X}$. In the supervised learning setting, ρ is assumed to be unknown and the goal of regression analysis is to estimate the regression function

$$f^*(x) = \mathbf{E}[Y|X = x] = \int_{\mathcal{Y}} y d\rho(y|x)$$

from a sample $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ of m observations which are drawn independently according to ρ.

Most MEE methods in the literature have focused on the use of Renyi's quadratic entropy (i.e. $\alpha = 2$) for its simplicity. Given the sample $\mathbf{z}$, Renyi's quadratic entropy can be estimated empirically as follows. For a hypothetical predictor f, let $e_i = y_i - f(x_i)$. Since $\{e_i\}_{i=1}^m$ is a sample of $E = Y - f(X)$, a kernel density estimator can be used to estimate the probability density function of E as

$$\hat{p}_E(e) = \frac{1}{mh} \sum_{j=1}^{m} G\left(\frac{(e - e_j)^2}{2h^2}\right),$$

where the function G defined on $[0, \infty]$ is a *windowing function* and $h > 0$ is a *scaling parameter*. A usual choice of the windowing function is $G(u) =$

$\frac{1}{\sqrt{2\pi}}\exp(-u)$ which leads to the Gaussian kernel density estimator. The *empirical Renyi's quadratic entropy* can then be estimated by

$$\begin{aligned}\hat{H}_2(f) &= -\log\left(\frac{1}{m}\sum_{i=1}^{m}\hat{p}_E(e_i)\right)\\ &= -\log\left\{\frac{1}{m^2h}\sum_{i=1}^{m}\sum_{j=1}^{m}G\Big(\frac{[(y_i-f(x_i))-(y_j-f(x_j))]^2}{2h^2}\Big)\right\}.\end{aligned}$$

The MEE method learns a function by minimizing the empirical Renyi's quadratic entropy. Notice that the log function is monotone and does not affect the minimizer, the MEE method can be implemented by minimizing the empirical risk

$$\mathcal{R}_{\mathbf{z}}(f) = -\frac{h^2}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}G\left(\frac{[(y_i-f(x_i))-(y_j-f(x_j))]^2}{2h^2}\right).$$

The kernel MEE method minimizes $\mathcal{R}_{\mathbf{z}}(f)$ in a reproducing kernel Hilbert space. Recall that $K:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$ is a Mercer kernel if it is continuous, symmetric and positive semidefinite. The reproducing kernel Hilbert space $\mathcal{H}_K$ is the completion of the linear span of the function set $\{K_x = K(x,\cdot) : x\in\mathcal{X}\}$ with the inner product induced by $\langle K_x, K_y\rangle_K = K(x,y)$. The reproducing property is given by $f(x) = \langle f, K_x\rangle_K$ and implies $\|f\|_\infty \leq \sup_{x\in\mathcal{X}}\sqrt{K(x,x)}\|f\|_K$. The regularized MEE method in the RKHS $\mathcal{H}_K$ is defined by

$$f_{\mathbf{z},\lambda} = \arg\min_{f\in\mathcal{H}_K}\left\{\mathcal{R}_{\mathbf{z}}(f) + \frac{\lambda}{2}\|f\|_K^2\right\}, \tag{1}$$

where $\lambda > 0$ is a regularization parameter. In [19] it is proved that the regularized kernel MEE algorithm is consistent when the scaling parameter h is chosen large enough. For small h, the consistency of MEE algorithms is a more complicated issue and has been discussed in [15]. In this paper we will study the convergence of gradient descent implementation of the regularized kernel MEE method (1), which is defined with the initial function $f_1 = 0$, and an updating rule

$$f_{t+1} = f_t - \eta_t\left(\nabla\mathcal{R}_{\mathbf{z}}(f_t) + \lambda f_t\right),\ t = 1,2,\cdots, \tag{2}$$

where $0<\lambda<1$, $\{\eta_t\}$ is the sequence of step sizes,

$$\nabla\mathcal{R}_{\mathbf{z}}(f_t) = \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}G'\Big(\frac{\xi_t^2(i,j)}{2h^2}\Big)\xi_t(i,j)(K_{x_i}-K_{x_j})$$

is the functional gradient of $\mathcal{R}_{\mathbf{z}}$ at f_t and $\xi_t(i,j) = (y_i - f_t(x_i)) - (y_j - f_t(x_j))$.

It is observed in the literature that early stopping plays a role of regularization and thus explicit regularization is not necessary to guarantee the convergence of gradient descent [34]. Regularization is also shown to be unnecessary in stochastic gradient descent algorithms by certain sacrifice of learning rates [35, 37]. In this paper, we also investigate the unregularized gradient descent algorithm for kernel MEE, which starts with $\tilde{f}_1 = 0$ and adopts the updating rule

$$\tilde{f}_{t+1} = \tilde{f}_t - \eta_t \nabla \mathcal{R}_{\mathbf{z}}(\tilde{f}_t), \ t = 1, 2, \cdots . \tag{3}$$

The unregularized algorithm has the advantage of no need to validate the regularization parameter.

In regression, we usually measure the learning performance via the $L^2_{\rho_{\mathcal{X}}}$ distance or equivalently, the excess mean squared error of the learned function. In MEE algorithms, however, since the empirical risk $\mathcal{R}_{\mathbf{z}}(f)$ is invariant to constant shift, the best we can expect is to have the learned function plus an appropriate constant shift to approximate the regression function f^* well. A good measure, in this case, is $\mathbf{var}[f_t(X) - f^*(X)]$, the variance of the random variance $f_t(X) - f^*(X)$ with respect to X, because small variance guarantees the existence of good approximation to f^* by a constant adjustment. How to choose the constant has been studied empirically in [13] and theoretically in [15, 18] and is omitted in this paper.

Next we state our main results. We need the following assumptions. Without loss of generality, we assume that $\sup_{x\in\mathcal{X}} \sqrt{K(x,x)} = 1$ and the response variable Y is uniformly bounded by 1. Also, the windowing function G is assumed to be differentiable and satisfy $G'_+(0) = -1$, $G'(u) < 0$ for $u > 0$, $C_G := \sup_{u\in[0,\infty)} |G'(u)| < \infty$, and there exist constants $p > 0$ and $c_p > 0$ such that

$$|G'(u) - G'_+(0)| \le c_p u^p, \ \text{for } u > 0. \tag{4}$$

These conditions can be satisfied by a variety of kernel density estimators. For instance, when Gaussian kernel density estimator is used, up to a constant multiplication which does affect the minimizer, we can use $G(u) = \exp(-u)$ so that the above assumptions hold with $C_G = 1$, $p = 1$ and $c_p = 1$.

Following [18,19] we use the pairwise squared loss to measure the approximation error of MEE in this paper. Define for each $f \in L^2_{\rho_{\mathcal{X}}}$ the pairwise squared risk as

$$\mathcal{E}(f) = \mathbf{E}\left[\Big((Y - f(X)) - \big(Y' - f(X')\big) \Big)^2\right]$$

where (X, Y) and (X', Y') are independent and identically distributed random pairs. The approximation error $\mathcal{D}(\lambda)$ is defined by

$$\begin{aligned}\mathcal{D}(\lambda) &= \arg \min_{f \in \mathcal{H}_K} \left\{\mathcal{E}(f) - \mathcal{E}(f^*) + \lambda \|f\|_K^2\right\} \\ &= \mathcal{E}(f_\lambda) - \mathcal{E}(f^*) + \lambda \|f_\lambda\|_K^2,\end{aligned}$$

where the regularization function f_λ is an minimizer of the regularized pairwise squared risk $\mathcal{E}(f) + \lambda \|f\|_K^2$ over the RKHS $\mathcal{H}_K$. Recall that the pairwise square loss has been used in least square ranking problems; see e.g. [9, 40] and references therein. The pairwise feature of MEE algorithm makes it appropriate to use $\mathcal{D}(\lambda)$ to characterize the approximation error of MEE.

Throughout this paper, we shall assume that for some constant $\mathcal{D}_0 \geq 1$ and $0 < \beta \leq 1$,

$$\mathcal{D}(\lambda) \leq \mathcal{D}_0 \lambda^\beta, \quad \forall\, \lambda > 0. \tag{5}$$

Note that

$$\mathcal{E}(f) - \mathcal{E}(f^*) = 2\mathbf{var}[f(X) - f^*(X)] \leq 2\|f - f^*\|^2_{L^2_{\rho_\mathcal{X}}} \tag{6}$$

and as a result

$$\mathcal{D}(\lambda) \leq \min_{f \in \mathcal{H}_K} \left\{2\|f - f^*\|^2_{L^2_{\rho_X}} + \lambda \|f\|_K^2\right\}. \tag{7}$$

The K-functional on right hand side was widely used in the learning theory literature as a measure of approximation ability of reproducing kernel Hilbert spaces and has been well studied. By (7), we see that that the the assumption (5) always holds with $\beta = 0$ since the right hand side of (7) is obviously bounded by the constant $2\|f^*\|_{L^2_{\rho_X}}$. If the target function f^* lies in $\mathcal{H}_K$ then (5) holds with $\beta = 1$. If $\mathcal{H}_K$ is dense in $C(\mathcal{X})$, the space of bounded continuous functions on $\mathcal{X}$, then the right hand side of (7) converges to 0 as $\lambda \to 0$. Thus, the assumption of a decay rate in (5) is natural and can be related to the definition of interpolation spaces. See e.g. [29, 30, 32, 34] and the references therein for more details.

Theorem 1. *Let $\{f_t\}_{t \geq 1}$ be defined by (2). Assume* (5) *holds for some* $0 < \beta \leq 1$. *Let $\eta_t = \eta t^{-\theta}$ with $\eta \leq \frac{1}{\lambda + 2}$ and $0 \leq \theta < 1$. If $\lambda \leq 1$ and $h \geq \frac{12(6c_p)^{1/2p}}{\sqrt{2}\lambda^{1+1/2p}}$, then for any $0 < \delta < 1$, with confidence at least $1 - \delta$, we* *have*

$$\mathbf{var}[f_{T+1}(X)-f^*(X)] \leq C\Bigg(\exp\Big(\frac{-2\eta_1\lambda T^{1-\theta}}{1-\theta}\Big)\lambda^{\beta-1} + \frac{\log\frac{8}{\delta}}{m\lambda^{3-\beta}} + \frac{1}{\lambda^{4p+4}h^{4p}} + \lambda^{\beta}\Bigg)$$

where C is a constant independent of m, T, h, λ or δ. As a result, if $\lambda \sim m^{-\frac{1}{3}}$, $T \geq (\frac{\log m}{6\eta m^{1/3}})^{\frac{1}{1-\theta}}$, and $h \geq m^{\frac{1}{3}(p+1+\beta/4)}$, then

$$\mathbf{var}[f_{T+1}(X)-f^*(X)] = O(m^{-\frac{\beta}{3}}).$$

The error bound in Theorem 1 decays exponentially fast in terms of the number T of iteration steps when λ and h are fixed. This indicates that, with regularization, increasing the number of iterations will never hurt the learning performance, though $T = O((\frac{\log m}{6\eta_1 m^{1/3}})^{\frac{1}{1-\theta}})$ is sufficient. Note the learning rate $O(m^{-\frac{\beta}{3}})$ is capacity independent. It matches the learning rate obtained in [19] under the worst capacity assumptions.

Next we turn to the unregularized method. To state our result, we need to measure the capacity of the hypothesis space $\mathcal{H}_K$. In learning theory, many capacity measures have been used, for instance, the VC-dimension, covering numbers, Rademacher complexity, and eigenvalues decay. In this paper, we will use the uniform covering number.

Definition 2. *For a subset S of $C(\mathcal{X})$ and $\varepsilon > 0$, the covering number $\mathcal{N}(S, \varepsilon)$ is the minimal integer $\ell \in \mathbb{N}$ such that there exist ℓ balls with radius ε covering S.*

For any $R > 0$, denote $B_R = \{f \in \mathcal{H}_K : \|f\|_K \leq R\}$ be the ball of radius R in $\mathcal{H}_K$. It can be embedded into $C(\mathcal{X})$ and let $\mathcal{N}(B_R, \varepsilon)$ denote its covering number in $C(\mathcal{X})$. We assume that for some $q > 0$ and $c_q > 0$, the covering number of B_1 satisfies

$$\log \mathcal{N}(B_1, \varepsilon) \leq c_q \varepsilon^{-q}, \qquad \forall \varepsilon > 0. \tag{8}$$

Note the uniform covering numbers of reproducing kernel Hilbert space has been studied in [41, 42]. The smaller q, the more stringent the capacity assumption is. In particular, when $\mathcal{X}$ is a compact subset of $\mathbb{R}^n$ and has a piecewise smooth boundary, if $K \in C^\alpha(\mathcal{X} \times \mathcal{X})$ then the condition (8) holds true with $q = \frac{2n}{\alpha}$. If the kernel $K \in C^\infty(\mathcal{X} \times \mathcal{X})$, then (8) is satisfied for an arbitrarily small $q > 0$. Note further that (8) holds with $q \leq 2$ for all Mercer kernels and therefore an assumption of (8) with $q = 2$ is equivalent to no capacity assumption and the convergence results become capacity independent.

Theorem 3. *Assume* (8) *and* (5). *Let* $\{\tilde{f}_t\}_{t\geq 1}$ *be defined by the algorithm* (3). *If* $\eta_t = \eta t^{-\theta}$ *with* $\frac{1}{2} \leq \theta < 1$ *and*

$$\eta \leq \min\left\{\frac{1}{2}, \frac{(1-\theta)2^{-\theta}}{64C_G^2 + 2C_G}\right\}, \tag{9}$$

then for any $0 < \delta < 1$, *with confidence at least* $1 - \delta$,

$$\mathcal{E}(\tilde{f}_T) - \mathcal{E}(f^*) \leq \tilde{C} \max\left\{\frac{1}{T^{\beta(1-\theta)}}, \frac{T^{1-\theta}}{(m-1)^{\frac{1}{1+q}}}, \frac{T^{(p+2)(1-\theta)}}{h^{2p}}\right\} \log\frac{4}{\delta}$$

where $\tilde{C}$ *is a constant independent of* m, T, h, *or* δ. *Choosing* $T \sim (m-1)^{\frac{1}{(1+q)(1+\beta)(1-\theta)}}$ *and* $h \geq (m-1)^{\frac{p+2+\beta}{2(1+q)(1+\beta)}}$, *we have*

$$\mathbf{var}[\tilde{f}_T(X) - f^*(X)] = O\left((m-1)^{-\frac{\beta}{(1+\beta)(1+q)}}\right).$$

Let us compare Theorem 3 and Theorem 1. We first notice that the error bound in Theorem 3 is not a deceasing function of the number of iteration steps. To achieve convergence for the unregularized gradient descent algorithm, early stopping is required and iterating too many steps may hurt the learning performance. This is a price paid for computational instability without regularization. Secondly, if we let $q \to 2$, we obtain the capacity independent result for the the unregularized algorithm as $O(m^{-\frac{\beta}{3(1+\beta)}})$ which is worse than the rate in Theorem 1 for the regularized algorithm. Recall that the unregularized algorithm does not need to validate the regularization parameter. Theorem 3 indicates that this computational advantage requires a sacrifice of convergence rates.

2.2 Comparisons with the literature and discussions

We compare our results with those in the literature and provide some remarks before moving to the proofs of our main theorems.

The regularized least squares method in reproducing kernel Hilbert spaces has been extensively studied in the literature. Under the assumption (5) on the approximation error decay (or an analogous source condition), the capacity condition (8) (or the near-equivalent assumption on the effective dimension of $\mathcal{H}_K$), and some other mild conditions, the regularized least square method has been proved to reach the minimax optimal rate $O(m^{-\frac{2\beta}{2\beta+q}})$ [3, 23, 31] and the optimal capacity independent rate is $O(m^{-\frac{\beta}{\beta+1}})$ [2, 39]. When the gradient descent or stochastic gradient algorithms are used to implement the method, efforts have been made to purse fast convergence rates; see e.g., [22, 25, 34, 35]. While the optimal capacity independent rate can be proved, the minimax optimal rates have not been verified. To our best knowledge the best result so far is the rate $O(m^{-\frac{2\beta}{(2+q)(1+\beta)}})$ obtained in [22]. The gap is usually attributed to lack of analysis tools but not an inherent feature of the gradient descent algorithms.

For pairwise learning, interactions between observations make the analysis more complicated. The capacity independent rate $O(m^{-\frac{\beta}{\beta+2}}\log^2 m)$ has been proved for unregularized stochastic gradient descent algorithm with the pairwise square loss and $O(m^{-\frac{1}{3}})$ for a general convex Lipschitz pairwise classification loss if the target function lies in $\mathcal{H}_K$ [37, 38]. Both are worse than their counterparts for pointwise learning.

When moving to MEE, the nonconvexity of the loss function added more complication to the analysis. Consequently, both the capacity independent rate $O(m^{-\frac{\beta}{3}})$ in Theorem 1 for regularized gradient descent algorithm and the capacity dependent rate $O(m^{-\frac{\beta}{(1+\beta)(1+q)}})$ in Theorem 3 for the unregularized one are suboptimal. As no empirical evidence shows MEE has worse performance than least squares method, we argue that the gap is not due to the inherent feature of these algorithms. A more plausible interpretation is that the pairwise non-convex loss caused essential difficulty for the analysis of kernel MEE. Developing new techniques to overcome or circumvent this difficulty would be an interesting problem. It is worth mentioning that, when this paper is revised, we proved in [21] that minimax optimal rates can be achieved by kernel gradient MEE if the function $f^*(x)-f^*(\tilde{x})$ defined on the product space $\mathcal{X}^2$ lies in $\mathcal{H}_{\widetilde{K}}$ and the pairwise kernel defined on $\mathcal{X}^2\times\mathcal{X}^2$ by

$$\widetilde{K}((x,\tilde{x}),(u,\tilde{u})) = K(x,u) + K(\tilde{x},\tilde{u}) - K(x,\tilde{u}) - K(\tilde{x},u)$$

satisfies some appropriate conditions. But how to derive sharper error bounds and faster convergence rates when f^* is not in $\mathcal{H}_K$ and the assumptions are made directly on the kernel K is still an open problem.

We also remark that the theoretical choices of the number T of iteration steps and the bandwidth parameter h in Theorem 3 are able to help us understand convergence properties of the algorithms. But they cannot be used in practice because of their dependence on the unknown parameters q, β, and θ. Deriving data-driven choice for T and h for practical use is important. In the literature of gradient descent algorithms for least squares kernel regression, several data-driven approaches have been proposed for the early stopping rule, see, for instance, the hold-out rule [2, 4, 34], the Rademacher complexity based rule in [25], and the balancing principle in [11]. It is interesting to investigate whether these rules can be adapted to the kernel gradient descent algorithm for MEE.

As for the data-driven choice of bandwidth parameter h, to our best knowledge, the study is very sparse. Our theory indicates it should depend on the sample size and be chosen large enough to guarantee the convergence. However, empirical simulations showed that the learning performance is not very sensitive to h and successful applications of MEE with h varying from 0.01 to 10 had been reported in te literature. It seems a moderate choice can lead sufficiently good results in most scenarios, though tuning it via cross validation or other strategies may be necessary for the best performance. It is still an open problem for future research.

3 Convergence of gradient descent with regularization

In this section we prove Theorem 1. To this end, we first prove several useful lemmas.

For g, $h \in \mathcal{H}_K$, let $g \otimes h$ denote the rank-one tensor product operator defined by $(g \otimes h)f = \langle f, h\rangle_K g$. It has Hilbert-Schmidt norm $\|g \otimes h\|_{HS} = \|g\|_K \|h\|_K$. If $h = g$, it is easy to check that $g \otimes g$ is a symmetric positive operator. Define the operator T_{XX} on $\mathcal{H}_K$ by

$$T_{XX} = \mathbf{E}\left[(K_X - K_{X'}) \otimes (K_X - K_{X'})\right]$$

and its empirical version by

$$\hat{T}_{XX} = \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} (K_{x_i} - K_{x_j}) \otimes (K_{x_i} - K_{x_j}).$$

By the reproducing property, for any $f \in \mathcal{H}_K$, we have

$$T_{XX} f = \mathbf{E}\left[(f(X) - f(X'))(K_X - K_{X'})\right]$$

and

$$\hat{T}_{XX} f = \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} (f(x_i) - f(x_j))(K_{x_i} - K_{x_j}).$$

Lemma 4. *Both T_{XX} and $\hat{T}_{XX}$ are symmetric and positive operators with their operator norms bounded by* 2*.*

Proof. It is easy to check that

$$T_{XX} = 2\Big(\mathbf{E}[K_X \otimes K_X] - (\mathbf{E}[K_X]) \otimes (\mathbf{E}[K_X])\Big). \tag{10}$$

Thus, T_{XX} is twice of the covariance operator of the Hilbert space valued random variable K_X. As a result, it is symmetric and positive. Moreover,

$$\begin{aligned} \|T_{XX}\| &\leq 2\, \|\mathbf{E}[K_X \otimes K_X]\|_{HS} \leq 2 \sup_{x \in \mathcal{X}} \|K_x \otimes K_x\|_{HS} \\ &= 2 \sup_{x \in \mathcal{X}} K(x, x) = 2. \end{aligned}$$

For $\hat{T}_{XX}$, we see

$$\hat{T}_{XX} = 2 \left[\frac{1}{m} \sum_{i=1}^{m} K_{x_i} \otimes K_{x_i} - \Big(\frac{1}{m} \sum_{i=1}^{m} K_{x_i} \Big) \otimes \Big(\frac{1}{m} \sum_{i=1}^{m} K_{x_i} \Big) \right] \tag{11}$$

is twice of the sample covariance operator of K_X. So it is symmetric, positive, and

$$\|\hat{T}_{XX}\| \leq \frac{2}{m} \sum_{i=1}^{m} \|K_{x_i} \otimes K_{x_i}\|_{HS} \leq \frac{2}{m} \sum_{i=1}^{m} K(x_i, x_i) \leq 2.$$

This finishes the proof. □

Next we define

$$T_{XY} = \mathbf{E}\left[(Y - Y')(K_X - K_{X'}) \right]$$

and its empirical version

$$\hat{T}_{XY} = \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} (y_i - y_j)(K_{x_i} - K_{x_j}).$$

Lemma 5. *We have $\|T_{XY}\|_K \leq 4$ and $\|\hat{T}_{XY}\|_K \leq 4$.*

Proof. We can verify that

$$T_{XY} = 2\Big(\mathbf{E}[YK_X] - \mathbf{E}[Y]\mathbf{E}[K_X]\Big) \tag{12}$$

and

$$\hat{T}_{XY} = 2\left[\frac{1}{m}\sum_{i=1}^{m} y_i K_{x_i} - \left(\frac{1}{m}\sum_{i=1}^{m} y_i\right)\left(\frac{1}{m}\sum_{i=1}^{m} K_{x_i}\right)\right]. \tag{13}$$

By $|Y| \leq 1$ and $\|K_x\|_K = \sqrt{K(x,x)} \leq 1$, we easily conclude the desired bounds. □

The following lemma was proved in [20].

Lemma 6. *Let $\mathcal{H}$ be a Hilbert space and ξ be a random variable with values in $\mathcal{H}$. Assume that $\|\xi\| \leq M$ almost surely. Let $\{\xi_1, \xi_2, \dots, \xi_m\}$ be a sample of m independent observations for ξ. Then, any $0 < \delta < 1$, we have with confidence $1-\delta$,*

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

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

By the above lemma, we have the following estimates.

Lemma 7. *For any $0 < \delta < 1$, with probability at least $1-\delta$, we have*

$$\|T_{XX} - \hat{T}_{XX}\| \leq 12\sqrt{\tau} \tag{14}$$

and

$$\|T_{XY} - \hat{T}_{XY}\| \leq 12\sqrt{\tau} \tag{15}$$

simultaneously, with $\tau = \frac{\log\frac{8}{\delta}}{m}$.

Proof. Applying Lemma 6 to the $\mathcal{H}_K$ valued random variable $\xi = K_X$, we obtain with probability at least $1 - \frac{\delta}{4}$,

$$\left\|\frac{1}{m}\sum_{i=1}^{m} K_{x_i} - \mathbf{E}[K_X]\right\| \leq \frac{1}{2}\left(\tau + \sqrt{8\tau + \tau^2}\right). \tag{16}$$

Applying Lemma 6 to the Hilbert-Schmidt operator valued random variable $\xi = K_X \otimes K_X$, we obtain with probability at least $1 - \frac{\delta}{4}$,

$$\begin{aligned}
&\left\| \frac{1}{m} \sum_{i=1}^{m} K_{x_i} \otimes K_{x_i} - \mathbf{E}[K_X \otimes K_X] \right\| \\
\leq &\left\| \frac{1}{m} \sum_{i=1}^{m} K_{x_i} \otimes K_{x_i} - \mathbf{E}[K_X \otimes K_X] \right\|_{HS} \\
\leq &\frac{1}{2} \left(\tau + \sqrt{8\tau + \tau^2} \right). \qquad (17)
\end{aligned}$$

Applying Lemma 6 to the real valued random variable $\xi = Y$, we obtain with probability at least $1 - \frac{\delta}{4}$,

$$\left| \frac{1}{m} \sum_{i=1}^{m} y_i - \mathbf{E}[Y] \right| \leq \frac{1}{2} \left(\tau + \sqrt{8\tau + \tau^2} \right). \qquad (18)$$

Applying Lemma 6 to the $\mathcal{H}_K$ valued random variable $\xi = YK_X$, we obtain with probability at least $1 - \frac{\delta}{4}$,

$$\left\| \frac{1}{m} \sum_{i=1}^{m} y_i K_{x_i} - \mathbf{E}[YK_X] \right\|_K \leq \frac{1}{2} \left(\tau + \sqrt{8\tau + \tau^2} \right). \qquad (19)$$

So, with probability at least $1 - \delta$, estimates (16)-(19) hold simultaneously.

By the facts (10), (11) and using (16), (17), we obtain

$$\begin{aligned}
&\|T_{XX} - \hat{T}_{XX}\| \\
\leq &2 \left\| \frac{1}{m} \sum_{i=1}^{m} K_{x_i} \otimes K_{x_i} - \mathbf{E}[K_X \otimes K_X] \right\| \\
&+ 2 \left\| (\mathbf{E}[K_X]) \otimes (\mathbf{E}[K_X]) - \Big(\frac{1}{m} \sum_{i=1}^{m} K_{x_i} \Big) \otimes \Big(\frac{1}{m} \sum_{i=1}^{m} K_{x_i} \Big) \right\| \\
\leq &\left(\tau + \sqrt{8\tau + \tau^2} \right) + 2 \left\| \mathbf{E}[K_X] + \Big(\frac{1}{m} \sum_{i=1}^{m} K_{x_i} \Big) \right\|_K \left\| \mathbf{E}[K_X] - \frac{1}{m} \sum_{i=1}^{m} K_{x_i} \right\|_K \\
\leq &3 \Big(\tau + \sqrt{8\tau + \tau^2} \Big) \leq 12\sqrt{\tau}.
\end{aligned}$$

Similarly, we can verify (15) by the facts (12), (13) and using (19), (16), (18). □

The following lemma characterizes f_λ.

Lemma 8. *We have* $(\lambda I + T_{XX})f_\lambda = T_{XY}$.

Proof. The functional derivative of $\mathcal{E}(f) + \lambda\|f\|_K^2$ with respect to $f \in \mathcal{H}_K$ is $-2T_{XY} + 2T_{XX}f + 2\lambda f$. Since f_λ is the minimizer of $\mathcal{E}(f) + \lambda\|f\|_K^2$, we see $-2T_{XY} + 2T_{XX}f_\lambda + 2\lambda f_\lambda = 0$. This implies that $(\lambda I + T_{XX})f_\lambda = T_{XY}$. □

We will also need the following two useful lemmas which have been proved in [36].

Lemma 9. *For* $v \in (0,1]$ *and* $\theta \in [0,1]$,

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

Lemma 10. *For any* $0 < t < T$ *and* $0 \leq \theta < 1$, *there holds*

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

To prove Theorem 1, we first establish a uniform bound for the solution path $\{f_t\}_{t\geq 1}$.

Lemma 11. *Assume* $\lambda \leq 1$, $\eta_t \leq \frac{1}{\lambda+2}$ *and*

$$h \geq \frac{12(6c_p)^{1/2p}}{\sqrt{2}\lambda^{1+1/2p}}. \tag{20}$$

We have

$$\|f_t\|_K \leq \frac{5}{\lambda}, \qquad \forall\ t \in \mathbb{N}. \tag{21}$$

Proof. We prove the bound (21) by induction on $t \in \mathbb{N}$. The case $t = 1$ is trivial since $f_1 = 0$ by definition. Suppose that $\|f_t\|_K \leq \frac{5}{\lambda}$. Consider the case f_{t+1}. By $G'_+(0) = -1$ and the definitions of $\hat{T}_{XX}$ and $\hat{T}_{XY}$, we can write

$$\begin{aligned} f_{t+1} &= f_t - \eta_t(\hat{T}_{XX}f_t + \lambda f_t - \hat{T}_{XY} + E_t) \\ &= \Big[(1-\eta_t\lambda)I - \eta_t\hat{T}_{XX}\Big] f_t + \eta_t\hat{T}_{XY} - \eta_t E_t, \end{aligned} \tag{22}$$

where I is the identity operator and

$$E_t = \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} \left[G' \left(\frac{\xi_t^2(i,j)}{2h^2} \right) - G'_+(0) \right] \xi_t(i,j)(K_{x_i} - K_{x_j}).$$

By Lemma 4, $\hat{T}_{XX}$ is positive and has operator norm bounded by 2. Thus $(1 - \eta_t \lambda)I - \eta_t \hat{T}_{XX}$ is positive for $\eta_t \leq \frac{1}{2+\lambda}$. Moreover,

$$1 - \eta_t \lambda - 2\eta_t \leq \|(1 - \eta_t \lambda)I - \eta_t \hat{T}_{XX}\| \leq 1 - \eta_t \lambda. \tag{23}$$

By Lemma 5, we have $\|\hat{T}_{XY}\|_K \leq 4$.

By $|y_i| \leq 1$, the induction hypothesis $\|f_t\| \leq \frac{5}{\lambda}$, and the restriction $\lambda \leq 1$ on the regularization parameter, we have $|\xi_t(i,j)| \leq 2+2\|f_t\|_K \leq 2+\frac{10}{\lambda} \leq \frac{12}{\lambda}$ for all (i,j) pairs. By the assumption (4),

$$\begin{aligned} \|E_t\|_K &\leq \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} \left| G' \left(\frac{\xi_t^2(i,j)}{2h^2} \right) - G'_+(0) \right| |\xi_t(i,j)| \|K_{x_i} - K_{x_j}\|_K. \\ &\leq \frac{2c_p}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} \frac{|\xi_t(i,j)|^{2p+1}}{2^p h^{2p}} \leq \frac{12^{2p+1} c_p}{2^{p-1} \lambda^{2p+1} h^{2p}}. \end{aligned} \tag{24}$$

Combining the above estimates, we obtain

$$\begin{aligned} \|f_{t+1}\|_K &\leq \left\| (1 - \lambda \eta_t)I - \eta_t \hat{T}_{XX} \right\| \|f_t\|_K + \eta_t \|\hat{T}_{XY}\|_K + \eta_t \|E_t\|_K \\ &\leq (1 - \lambda \eta_t)\|f_t\|_K + 4\eta_t + \eta_t \|E_t\|_K \\ &\leq (1 - \lambda \eta_t)\frac{5}{\lambda} + 4\eta_t + \eta_t \frac{12^{2p+1} c_p}{2^{p-1} \lambda^{2p+1} h^{2p}} \\ &= \frac{5}{\lambda} - \eta_t \left(1 - \frac{12^{2p+1} c_p}{2^{p-1} \lambda^{2p+1} h^{2p}} \right). \end{aligned}$$

The condition (20) on h ensures that

$$1 - \frac{12^{2p+1} c_p}{2^{p-1} \lambda^{2p+1} h^{2p}} \geq 0.$$

Therefore, we have $\|f_{t+1}\|_K \leq \frac{5}{\lambda}$. By induction, the proof is complete. □

The following lemma measures how the error changes by each step of updating.

Lemma 12. *Let* $0 < \lambda \leq 1$, $\eta_t \leq \frac{1}{2+\lambda}$ *and* h *satisfy* (20). *If* (14) *and* (15) *hold simultaneously, we have*

$$\begin{aligned}\|f_{t+1} - f_\lambda\|_K \leq & (1 - \lambda\eta_t)\|f_t - f_\lambda\|_K \\ & + \eta_t \left(12(\|f_\lambda\|_K + 1)\sqrt{\tau} + \frac{12^{2p+1} c_p}{2^{p-1}\lambda^{2p+1}h^{2p}}\right)\end{aligned}$$

where $\tau = \frac{\log \frac{8}{\delta}}{m}$.

Proof. By (22) and the fact $(T_{XX} + \lambda I)f_\lambda = T_{XY}$ from Lemma 8, we have

$$\begin{aligned}f_{t+1} - f_\lambda =& [(1 - \eta_t\lambda)I - \eta_t\hat{T}_{XX}](f_t - f_\lambda) \\ & + \eta_t(T_{XX} - \hat{T}_{XX})f_\lambda + \eta_t(\hat{T}_{XY} - T_{XY}) - \eta_t E_t.\end{aligned}$$

By (23) we have

$$\begin{aligned}\|f_{t+1} - f_\lambda\|_K \leq & (1 - \eta_t\lambda)\|f_t - f_\lambda\|_K \\ & + \eta_t\|T_{XX} - \hat{T}_{XX}\|\|f_\lambda\|_K \\ & + \eta_t\|T_{XY} - \hat{T}_{XY}\|_K + \eta_t\|E_t\|_K.\end{aligned}$$

This together with (24) and the assumptions (14), (15) gives the desired conclusion. □

With the above lemma, we can prove the following bound of the estimation error.

Lemma 13. *Under the same conditions as Lemma 12, if the step sizes are chosen as* $\eta_t = \eta t^{-\theta}$ *for some* $0 < \eta < \frac{1}{\lambda+2}$ *and* $0 \leq \theta < 1$, *we have*

$$\begin{aligned}\|f_{T+1} - f_\lambda\|_K \leq C_1 \Bigg(& \exp\left(\frac{-\eta_1 \lambda T^{1-\theta}}{1-\theta}\right)\lambda^{\frac{\beta-1}{2}} \\ & + \lambda^{\frac{\beta-3}{2}}\sqrt{\frac{\log\frac{8}{\delta}}{m}} + \frac{1}{\lambda^{2p+2}h^{2p}}\Bigg)\end{aligned} \quad (25)$$

where C_1 *is a constant independent of* $m, T, h, \lambda,$ *or* δ.

Proof. Applying Lemma 12 iteratively for $t = 1, \cdots, T$, we obtain

$$\begin{aligned}\|f_{T+1} - f_\lambda\|_K \leq & \prod_{t=1}^{T}(1 - \lambda\eta_t)\|f_\lambda\|_K \\ & + \left(12(\|f_\lambda\|_K + 1)\sqrt{\tau} + \frac{12^{2p+1}c_p}{2^{p-1}\lambda^{2p+1}h^{2p}}\right)\sum_{t=1}^{T}\prod_{j=t+1}^{T}(1 - \lambda\eta_j)\eta_t.\end{aligned}$$

Since $\eta_t = \eta t^{-\theta}$ and $\lambda \leq 1$, by Lemma 10, we have

$$\begin{aligned}\prod_{t=1}^{T}(1-\lambda\eta_t) \leq \exp\left(-\lambda\sum_{t=1}^{T}\eta_t\right) &\leq \exp\Big(\frac{\eta\lambda(1-T^{1-\theta})}{1-\theta}\Big)\\ &\leq \exp\Big(\frac{\eta}{1-\theta}\Big)\exp\Big(-\frac{\eta\lambda T^{1-\theta}}{1-\theta}\Big).\end{aligned}$$

Lemma 9 yields

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

Noting the bound $\|f_\lambda\|_K \leq \sqrt{\mathcal{D}(\lambda)/\lambda} \leq \sqrt{\mathcal{D}_0}\lambda^{\frac{\beta-1}{2}}$, we have

$$\begin{aligned}\|f_{T+1}-f_\lambda\|_K \leq &\exp\Big(\frac{\eta}{1-\theta}\Big)\exp\Big(-\frac{\eta\lambda T^{1-\theta}}{1-\theta}\Big)\sqrt{\mathcal{D}_0}\lambda^{\frac{\beta-1}{2}}\\ &+\frac{3}{\lambda}\left(12(\sqrt{\mathcal{D}_0}\lambda^{\frac{\beta-1}{2}}+1)\sqrt{\tau}+\frac{12^{2p+1}c_p}{2^{p-1}\lambda^{2p+1}h^{2p}}\right).\end{aligned}$$

Taking the constant $C_1 = \max\{\sqrt{\mathcal{D}_0}\exp(\frac{\eta}{1-\theta}), 36(\sqrt{\mathcal{D}_0}+1), \frac{3(12)^{2p+1}c_p}{2^{p-1}}\}$, we get the desired conclusion (25). □

Now we can prove our first main theorem.

Proof of Theorem 1. Note that

$$\begin{aligned}\mathbf{var}[f_{T+1}-f^*] &\leq 2\mathbf{var}[f_{T+1}-f_\lambda]+2\mathbf{var}[f_\lambda-f^*]\\ &\leq 2\|f_{T+1}-f_\lambda\|_K^2+\mathcal{D}_0\lambda^{\beta}.\end{aligned}$$

By Lemma 7 and Lemma 13, Theorem 1 holds with constant $C = \max\{6C_1^2, \mathcal{D}_0\}$. □

4 Convergence of unregularized gradient descent algorithm

In this section, we prove the convergence of the unregularized gradient descent algorithm for kernel MEE. To this end, we need to deal with U-statistics. Let g be a symmetric function defined on $\mathcal{Z}\times\mathcal{Z}$ and $\{z_i\}_{i=1}^m$ be a sample of m independent observations drawn from a probability distribution ρ on $\mathcal{Z}$. The U-statistic induced by g is defined to be

$$\mathcal{U}(g) = \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{\substack{j=1 \\ j \neq i}}^{m} g(z_i, z_j).$$

We define a variant formula of $\mathcal{U}(g)$ by

$$\bar{\mathcal{U}}(g) = \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} g(z_i, z_j).$$

Recall the following Hoeffding inequality for U-statistics [17].

Lemma 14. *If g is symmetric, $\|g\|_\infty \leq B$, and $\mathbf{var}[g] = \sigma^2$, then for any $\varepsilon > 0$,*

$$\Pr\left\{ \left| \mathbf{E}[g] - \mathcal{U}(g) \right| > \varepsilon \right\} \leq 2 \exp\left\{ -\frac{(m-1)\varepsilon^2}{4(\sigma^2 + \frac{2}{3}B\varepsilon)} \right\}.$$

Immediately, we have ratio probability inequalities for a single random variable in Lemma 15 and for a set of functions in Lemma 16 below, respectively. The proofs can be easily given by using the techniques and following the processes in [1,8]. We omit the details.

Lemma 15. *Suppose that a symmetric function g on $\mathcal{Z} \times \mathcal{Z}$ satisfies $\mathbf{E}[g] > 0$ and $\|g\|_\infty \leq B$. If $\mathbf{E}[g^2] \leq c\mathbf{E}[g]$, then for any $\varepsilon > 0$ and $0 < \alpha \leq 1$, there holds*

$$\Pr\left\{ \frac{|\mathbf{E}[g] - \mathcal{U}(g)|}{\sqrt{\mathbf{E}[g] + \varepsilon}} > \alpha\sqrt{\varepsilon} \right\} \leq 2 \exp\left\{ -\frac{(m-1)\alpha^2\varepsilon}{4(c + \frac{2}{3}B)} \right\}.$$

Lemma 16. *Let $\mathcal{G}$ be a set of symmetric functions on $\mathcal{Z} \times \mathcal{Z}$ such that for all $g \in \mathcal{G}$, $\mathbf{E}[g] \geq 0$, $\|g\|_\infty \leq B$, and $\mathbf{E}[g^2] \leq c\mathbf{E}[g]$, then for every $\varepsilon > 0$ and $0 < \alpha \leq 1$, we have*

$$\Pr\left\{ \sup_{g \in \mathcal{G}} \frac{|\mathbf{E}[g] - \mathcal{U}(g)|}{\sqrt{\mathbf{E}[g] + \varepsilon}} > 4\alpha\sqrt{\varepsilon} \right\} \leq 2\mathcal{N}(\mathcal{G}, \alpha\varepsilon) \exp\left\{ -\frac{(m-1)\alpha^2\varepsilon}{4(c + \frac{2}{3}B)} \right\}.$$

The following two lemmas are easy corollaries of Lemma 15 and Lemma 16, respectively.

Lemma 17. *Suppose that a symmetric function g on $\mathcal{Z}\times\mathcal{Z}$ satisfies $\mathbf{E}[g] \geq 0$ and $\|g\|_\infty \leq B$. If $\mathbf{E}[g^2] \leq c\mathbf{E}[g]$, then for any $0 < \delta < 1$,*

$$\left|\mathbf{E}[g] - \bar{\mathcal{U}}(g)\right| \leq \frac{1}{2}\mathbf{E}[g] + \frac{4(c + \frac{2}{3}B)\log(\frac{2}{\delta}) + 2B}{m-1}$$

with probability at least $1 - \delta$.

Proof. Applying Lemma 15 with $\alpha = 1$, we obtain that, for any $0 < \delta < 1$,

$$\left|\mathbf{E}[g] - \mathcal{U}(g)\right| \leq \frac{1}{2}\mathbf{E}[g] + \frac{4(c + \frac{2}{3}B)\log(\frac{2}{\delta})}{m-1}$$

with probability at least $1 - \delta$. Noting that

$$\left|\mathcal{U}(g) - \bar{\mathcal{U}}(g)\right| \leq \left|\frac{1}{m^2(m-1)}\sum_{i=1}^{m}\sum_{j=1}^{m} g(z_i, z_j)\right| + \left|\frac{1}{m^2}\sum_{i=1}^{m} g(z_i, z_i)\right| \leq \frac{2B}{m-1},$$

we obtain the desired estimate. □

Lemma 18. *Let $\mathcal{G}$ be set of symmetric functions on $\mathcal{Z} \times \mathcal{Z}$ such that for each $g \in \mathcal{G}$, $\mathbf{E}[g] > 0$, $\|g\|_\infty \leq B$, and $\mathbf{E}[g^2] \leq c\mathbf{E}[g]$. If $\log\left(\mathcal{N}(\mathcal{G}, \varepsilon)\right) \leq a\varepsilon^{-q}$ for some $a > 0$ and $q > 0$, then for any $0 < \delta < 1$,*

$$\left|\mathbf{E}[g] - \bar{\mathcal{U}}(g)\right| \leq \frac{1}{2}\mathbf{E}[g] + 72\max\left\{\frac{(c + \frac{2}{3}B)\log(\frac{2}{\delta})}{m-1}, \left(\frac{(c + \frac{2}{3}B)a}{m-1}\right)^{\frac{1}{1+q}}\right\}.$$

for all $g \in \mathcal{G}$ with probability at least $1 - \delta$.

Proof. Applying Lemma 16 with $\alpha = 1$ and using the assumption on the covering number of $\mathcal{G}$, we obtain that, for any $0 < \delta < 1$,

$$|\mathbf{E}[g] - \mathcal{U}(g)| \leq 4\sqrt{\mathbf{E}[g] + \varepsilon^*}\sqrt{\varepsilon^*} \leq \frac{1}{2}\mathbf{E}[g] + \frac{17}{2}\varepsilon^*, \ \forall\, g \in \mathcal{G}, \tag{26}$$

holds with probability at least $1 - \delta$, where ε^* is a positive solution to the equation

$$a\varepsilon^{-q} - \frac{(m-1)\varepsilon}{4(c + \frac{2}{3}B)} = \log(\tfrac{\delta}{2}).$$

Note the equation is equivalent to

$$\varepsilon^{1+q} - \frac{4(c + \frac{2}{3}B)\log(\frac{2}{\delta})}{m-1}\varepsilon^q - \frac{4(c + \frac{2}{3}B)a}{m-1} = 0.$$

By [10, Lemma 7], the equation has a unique positive solution and

$$\varepsilon^* \leq \max\left\{ \frac{8(c+\frac{2}{3}B)\log(\frac{2}{\delta})}{m-1}, \left(\frac{8(c+\frac{2}{3}B)a}{m-1}\right)^{\frac{1}{1+q}} \right\}.$$

By (26) and the fact that $|\mathcal{U}(g) - \bar{\mathcal{U}}(g)| \leq \frac{2B}{m-1}$ for all $g \in \mathcal{G}$, we obtain the desired estimate. □

In the sequel let the empirical pairwise squared error on a sample $\mathbf{z}$ be defined by

$$\mathcal{E}_{\mathbf{z}}(f) = \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} \Big((y_i - f(x_i)) - (y_j - f(x_j)) \Big)^2.$$

By the relation (6), it suffices to bound the the excess error $\mathcal{E}(\tilde{f}_T) - \mathcal{E}(f^*)$. To this end we use the following error decomposition.

Lemma 19. *For any* $\lambda > 0$, *we have*

$$\mathcal{E}(\tilde{f}_T) - \mathcal{E}(f^*) \leq \mathcal{Q}_1 + \mathcal{Q}_2 + \mathcal{Q}_3 + \mathcal{D}(\lambda),$$

where

$$\begin{aligned}
\mathcal{Q}_1 &= \mathcal{E}_{\mathbf{z}}(\tilde{f}_T) - \mathcal{E}_{\mathbf{z}}(f_\lambda), \\
\mathcal{Q}_3 &= \Big(\mathcal{E}(\tilde{f}_T) - \mathcal{E}(f^*)\Big) - \Big(\mathcal{E}_{\mathbf{z}}(\tilde{f}_T) - \mathcal{E}_{\mathbf{z}}(f^*)\Big), \\
\mathcal{Q}_2 &= \Big(\mathcal{E}_{\mathbf{z}}(f_\lambda) - \mathcal{E}_{\mathbf{z}}(f^*)\Big) - \Big(\mathcal{E}(f_\lambda) - \mathcal{E}(f^*)\Big).
\end{aligned}$$

We first estimate $\mathcal{Q}_3$.

Proposition 20. *For any* $\lambda > 0$, *with confidence at least* $1 - \frac{\delta}{2}$,

$$\mathcal{Q}_3 \leq \frac{2^9 \mathcal{D}(\lambda)\log(\frac{4}{\delta})}{(m-1)\lambda} + \frac{1}{2}\mathcal{D}(\lambda).$$

Proof. Consider the function

$$g(z,z') = \Big((y - f_\lambda(x)) - \big(y' - f_\lambda(x')\big) \Big)^2 - \Big((y - f^*(x)) - \big(y' - f^*(x')\big) \Big)^2$$

defined on $\mathcal{Z} \times \mathcal{Z}$. Note that $\mathbf{E}[g] = \mathcal{E}(f_\lambda) - \mathcal{E}(f^*) \geq 0$ and

$$\begin{aligned}
\|g\|_\infty &\leq (6 + 2\|f_\lambda\|_\infty)|f_\lambda(x') - f_\lambda(x) - f^*(x') + f^*(x)| \\
&\leq (6 + 2\|f_\lambda\|_K)^2 \leq 64\mathcal{D}(\lambda)/\lambda.
\end{aligned}$$

It follows that

$$\begin{aligned}
&\mathbf{E}[g^2] \leq \mathbf{E}[\|g\|_\infty^2] \\
\leq&(6+2\|f_\lambda\|_\infty)^2 \mathbf{E}\left[(f_\lambda(x')-f_\lambda(x)-f^*(x')+f^*(x))^2\right] \\
=&(6+2\|f_\lambda\|_K)^2 2\mathbf{var}[f_\lambda(X)-f^*(X)] \\
=&(6+2\|f_\lambda\|_K)^2(\mathcal{E}(f_\lambda)-\mathcal{E}(f^*)) \\
\leq&64(\mathcal{D}(\lambda)/\lambda)\mathbf{E}[g].
\end{aligned}$$

By Lemma 17 with $c = B = 64\mathcal{D}(\lambda)/\lambda$, we obtain

$$\begin{aligned}
\mathcal{Q}_3 &\leq \frac{1}{2}\left(\mathcal{E}(f_\lambda)-\mathcal{E}(f^*)\right) + \frac{2^9\mathcal{D}(\lambda)\log(\frac{4}{\delta})}{(m-1)\lambda} \\
&\leq \frac{1}{2}\mathcal{D}(\lambda) + \frac{2^9\mathcal{D}(\lambda)\log(\frac{4}{\delta})}{(m-1)\lambda}.
\end{aligned}$$

with probability at least $1-\frac{\delta}{2}$. □

In order to bound $\mathcal{Q}_1$ and $\mathcal{Q}_2$ we need the following bound on the sequence $\{\tilde{f}_t\}$. In the sequel we will denote $\tilde{\xi}_t(i,j) = (y_i-\tilde{f}_t(x_i))-(y_j-\tilde{f}_t(x_j))$ and $\xi_f(i,j) = (y_i - f(x_i)) - (y_j - f(x_j))$.

Lemma 21. *Define $\{\tilde{f}_t\}_{t\geq 1}$ by (3). Let $\eta_t = \eta t^{-\theta}$ with $\frac{1}{2} \leq \theta < 1$ and η satisfying (9). Then for $t = 1, \cdots, T$,*

$$\|\tilde{f}_t\|_K \leq t^{\frac{1-\theta}{2}}. \tag{27}$$

Proof. We prove (27) by induction. Suppose that $\|\tilde{f}_t\|_K \leq t^{\frac{1-\theta}{2}}$. Denote

$$\tilde{H}_t = \frac{1}{m^2}\sum_{i=1}^m\sum_{j=1}^m G'\left(\frac{\tilde{\xi}_t^2(i,j)}{2h^2}\right)\tilde{\xi}_t(i,j)(K_{x_i}-K_{x_j}).$$

Then $\tilde{f}_{t+1} = \tilde{f}_t - \eta_t H_t$ and we can write

$$\begin{aligned}
&\|\tilde{f}_{t+1}\|_K^2 = \|\tilde{f}_t\|_K^2 - 2\eta_t\langle\tilde{f}_t,\tilde{H}_t\rangle_K + \eta_t^2\|\tilde{H}_t\|_K^2 \\
&= \|\tilde{f}_t\|_K^2 + \eta_t^2\|\tilde{H}_t\|_K^2 + \frac{2\eta_t}{m^2}\sum_{i=1}^m\sum_{j=1}^m G'\left(\frac{\tilde{\xi}_t^2(i,j)}{2h^2}\right)\tilde{\xi}_t(i,j)(\tilde{f}_t(x_j)-\tilde{f}_t(x_i)).
\end{aligned} \tag{28}$$

It is easy to check that

$$\begin{aligned}\|\tilde{H}_t\|_K &\le \frac{2}{m^2}\sum_{i=1}^m\sum_{j=1}^m \left|G'\left(\frac{\tilde{\xi}_t^2(i,j)}{2h^2}\right)\right| |\tilde{\xi}_t(i,j)| \\ &\le 2C_G(2+2\|\tilde{f}_t\|_K) \le 8C_G t^{\frac{1-\theta}{2}}. \qquad (29)\end{aligned}$$

Recall that $G'(u) < 0$ for $u > 0$. For each pair (i,j), we have

$$\begin{aligned}&G'\left(\frac{\tilde{\xi}_t^2(i,j)}{2h^2}\right)\tilde{\xi}_t(i,j)(\tilde{f}_t(x_j)-\tilde{f}_t(x_i)) \\ =&G'\left(\frac{\tilde{\xi}_t^2(i,j)}{2h^2}\right)\tilde{\xi}_t(i,j)[\tilde{\xi}_t(i,j)+y_j-y_i] \\ =&G'\left(\frac{\tilde{\xi}_t^2(i,j)}{2h^2}\right)\left[\left(\tilde{\xi}_t(i,j)+\frac{y_j-y_i}{2}\right)^2-\frac{(y_i-y_j)^2}{4}\right] \\ \le&-G'\left(\frac{\tilde{\xi}_t^2(i,j)}{2h^2}\right)\frac{(y_i-y_j)^2}{4} \le C_G.\end{aligned}$$

Plugging this estimate and (29) into (28) we obtain

$$\begin{aligned}\|\tilde{f}_{t+1}\|_K^2 &\le \|\tilde{f}_t\|_K^2 + 64\eta_t^2 C_G^2 t^{1-\theta} + 2\eta_t C_G \\ &\le t^{1-\theta} + 64\eta^2 C_G^2 t^{1-3\theta} + 2\eta C_G t^{-\theta}.\end{aligned}$$

By the Taylor expansion, we see that $(t+1)^{1-\theta} - t^{1-\theta} = (1-\theta)(t+w)^{-\theta}$ for some $w \in (0,1)$. This implies $(t+1)^{1-\theta} - t^{1-\theta} \ge (1-\theta)2^{-\theta}t^{-\theta}$. Thus, by the condition (9) on η and $\theta \ge \frac{1}{2}$,

$$\begin{aligned}\|\tilde{f}_{t+1}\|_K^2 &\le (t+1)^{1-\theta} + 64\eta^2 C_G^2 t^{1-3\theta} + 2\eta C_G t^{-\theta} - (1-\theta)2^{-\theta}t^{-\theta} \\ &= (t+1)^{1-\theta} + t^{-\theta}[64\eta^2 C_G^2 t^{1-2\theta} + 2\eta C_G - (1-\theta)2^{-\theta}] \\ &\le (t+1)^{1-\theta} + 64\eta^2 C_G^2 + 2\eta C_G - (1-\theta)2^{-\theta} \\ &\le (t+1)^{1-\theta} + (64C_G^2 + 2C_G)\eta - (1-\theta)2^{-\theta} \\ &\le (t+1)^{1-\theta}.\end{aligned}$$

This proves our statement (27). □

The estimation of $\mathcal{Q}_2$ will need the following result.

Proposition 22. *For any $R \geq 1$ and $f \in B_R$, with confidence at least $1-\frac{\delta}{2}$,*

$$\begin{aligned}\mathcal{E}(f)-\mathcal{E}(f^*))-(\mathcal{E}_{\mathbf{z}}(f)-\mathcal{E}_{\mathbf{z}}(f^*)) \leq & \frac{1}{2}(\mathcal{E}(f)-\mathcal{E}(f^*)) \\ & + 2^{13}R^2 \max\left\{\frac{\log(\frac{4}{\delta})}{m-1}, \left(\frac{c_q}{m-1}\right)^{\frac{1}{q+1}}\right\}.\end{aligned}$$

Proof. For each $f \in \mathcal{H}_K$, denote

$$\begin{aligned}g_f(z,z') = & \Big((y-f(x))-(y'-f(x'))\Big)^2 \\ & -\Big((y-f^*(x))-(y'-f^*(x'))\Big)^2.\end{aligned}$$

Consider the function set $\mathcal{G} = \{g_f(z,z') : f \in B_R\}$. For each $g \in \mathcal{G}$, by the same techniques used in the proof of Proposition 20, we can verify that $\mathbf{E}[g] \geq 0$, $\|g\|_\infty \leq 64R^2$ and $\mathbf{E}[g^2] \leq 64R^2\mathbf{E}[g]$. Also, it is easy to check that $\mathcal{N}(\mathcal{G},\varepsilon) \leq \mathcal{N}(B_1, \frac{\varepsilon}{16R^2})$. Thus $\log(\mathcal{N}(\mathcal{G},\varepsilon)) \leq c_q(16R^2)^q\varepsilon^{-q}$. Applying Lemma 18 with $a = c_q(16R^2)^q$, $c = B = 64R^2$ we obtain the desired estimate. □

The estimation for $\mathcal{Q}_1$ needs the following result.

Proposition 23. *Let $\eta_t = \eta t^{-\theta}$ with $\theta > \frac{1}{2}$ and η satisfying the restriction (9). For any fixed $f \in B_R$, we have*

$$\mathcal{E}_{\mathbf{z}}(\tilde{f}_T)-\mathcal{E}_{\mathbf{z}}(f) \leq \frac{2R^2}{\eta}T^{\theta-1} + \frac{2c'_{p,\theta}}{\eta}\Big(T^{(p+2)(1-\theta)} + RT^{(p+\frac{1}{2})(1-\theta)}\Big)h^{-2p}, \tag{30}$$

where $c'_{p,\theta}$ is a constant independent of R, h, T and will be explicitly given in the proof.

Proof. By the elementary equality $u^2 = (u')^2 + 2u'(u-u') + (u-u')^2$ for $u, u' \in \mathbb{R}$, we have

$$\begin{aligned}\tilde{\xi}_{t+1}^2(i,j) = & \tilde{\xi}_t^2(i,j) + 2\tilde{\xi}_t(i,j)\Big[(\tilde{f}_{t+1}(x_j)-\tilde{f}_t(x_j))-(\tilde{f}_{t+1}(x_i)-\tilde{f}_t(x_i))\Big] \\ & + \Big[(\tilde{f}_{t+1}(x_j)-\tilde{f}_t(x_j))-(\tilde{f}_{t+1}(x_i)-\tilde{f}_t(x_i))\Big]^2 \\ = & \tilde{\xi}_t^2(i,j) + 2\tilde{\xi}_t(i,j)\Big\langle \tilde{f}_{t+1}-\tilde{f}_t, K_{x_j}-K_{x_i}\Big\rangle_K \\ & + \Big\langle \tilde{f}_{t+1}-\tilde{f}_t, K_{x_j}-K_{x_i}\Big\rangle_K^2\end{aligned}$$

for each (i, j) pair. Summing up with $i, j = 1, \cdots, m$, we have

$$\mathcal{E}_{\mathbf{z}}(\tilde{f}_{t+1}) \leq \mathcal{E}_{\mathbf{z}}(\tilde{f}_t) + 2\left\langle \tilde{f}_{t+1} - \tilde{f}_t, \tilde{W}_t \right\rangle_K + 2\|\tilde{f}_{t+1} - \tilde{f}_t\|_K^2,$$

where we used the notation

$$\tilde{W}_t = \frac{1}{m^2}\sum_{i=1}^m \sum_{j=1}^m \tilde{\xi}_t(i,j)(K_{x_j} - K(x_i))$$

and the estimate

$$\frac{1}{m^2}\sum_{i=1}^m \sum_{j=1}^m \left\langle \tilde{f}_{t+1} - \tilde{f}_t, K_{x_j} - K_{x_i} \right\rangle_K^2 = \left\langle \tilde{f}_{t+1} - \tilde{f}_t, \hat{T}_{XX}(\tilde{f}_{t+1} - \tilde{f}_t) \right\rangle_K$$
$$\leq 2\|\tilde{f}_{t+1} - \tilde{f}_t\|_K^2.$$

Note that $\tilde{f}_{t+1} - \tilde{f}_t = -\eta_t \tilde{H}_t$ and let $\tilde{E}_t = \tilde{H}_t - \tilde{W}_t$. Then

$$\begin{aligned}\mathcal{E}_{\mathbf{z}}(\tilde{f}_{t+1}) &\leq \mathcal{E}_{\mathbf{z}}(\tilde{f}_t) - 2(1-\eta_t)\eta_t\|\tilde{H}_t\|_K^2 + 2\eta_t \left\langle \tilde{H}_t, \tilde{E}_t \right\rangle_K \\ &\leq \mathcal{E}_{\mathbf{z}}(\tilde{f}_t) - \eta_t\|\tilde{H}_t\|_K^2 + 2\eta_t\|\tilde{H}_t\|_K\|\tilde{E}_t\|_K,\end{aligned}$$

where we used the fact that $\eta \leq \frac{1}{2}$ implies that $1 - \eta_t = 1 - \eta t^{-\theta} \geq \frac{1}{2}$ for all $t \geq 1$. Similar to the estimation (24), but using (27) this time, we obtain $|\tilde{\xi}_t(i,j)| \leq 2 + 2t^{\frac{1-\theta}{2}} \leq 4t^{\frac{1-\theta}{2}}$ and

$$\|\tilde{E}_t\|_K \leq 2^{3p+3} c_p t^{\frac{(1-\theta)(2p+1)}{2}} h^{-2p}.$$

This together with (29) implies

$$\mathcal{E}_{\mathbf{z}}(\tilde{f}_{t+1}) \leq \mathcal{E}_{\mathbf{z}}(\tilde{f}_t) - \eta_t\|\tilde{H}_t\|_K^2 + 2^{6+3p} c_p C_G t^{(p+1)-(p+2)\theta} h^{-2p}. \tag{31}$$

By the elementary inequality $u^2 \leq (u')^2 + 2u(u - u')$ for $u, u' \in \mathbb{R}$, we obtain

$$\tilde{\xi}_t^2(i,j) = \tilde{\xi}_f^2(i,j) + 2\tilde{\xi}_t(i,j)\left\langle \tilde{f}_t - f, K_{x_j} - K_{x_i} \right\rangle_K,$$

which implies $\mathcal{E}_{\mathbf{z}}(\tilde{f}_t) \leq \mathcal{E}_{\mathbf{z}}(f) + 2\langle \tilde{f}_t - f, \tilde{W}_t\rangle_K$. Then it is easy to check that

$$\begin{aligned}\mathcal{E}_{\mathbf{z}}(\tilde{f}_t) &\leq \mathcal{E}_{\mathbf{z}}(f) + 2\left\langle \tilde{f}_t - f, \tilde{H}_t \right\rangle_K - 2\left\langle \tilde{f}_t - f, \tilde{E}_t \right\rangle_K \\ &\leq \mathcal{E}_{\mathbf{z}}(f) + 2\left\langle \tilde{f}_t - f, \tilde{H}_t \right\rangle_K + 2(\|\tilde{f}_t\|_K + \|f\|_K)\|\tilde{E}_t\|_K \\ &\leq \mathcal{E}_{\mathbf{z}}(f) + 2\left\langle \tilde{f}_t - f, \tilde{H}_t \right\rangle_K + 2^{3p+5} c_p (t^{(p+1)(1-\theta)} \\ &+ R t^{(p+\frac{1}{2})(1-\theta)}) h^{-2p}.\end{aligned}$$

Plugging it into (31), we get

$$\begin{aligned}\mathcal{E}_{\mathbf{z}}(\tilde{f}_{t+1}) &\leq \mathcal{E}_{\mathbf{z}}(f) + \frac{1}{\eta_t}\left(2\eta_t \left\langle \tilde{f}_t - f, H_t \right\rangle_K - \eta_t^2 \|H_t\|_K^2\right) + \Lambda_t(h, R) \\ &= \mathcal{E}_{\mathbf{z}}(f) + \frac{1}{\eta_t}\left(\|\tilde{f}_t - f\|_K^2 - \|\tilde{f}_{t+1} - f\|_K^2\right) + \Lambda_t(h, R)\end{aligned}$$

where

$$\Lambda_t(h, R) = 2^{7+3p} c_p C_G (t^{(p+1)(1-\theta)} + R t^{(p+\frac{1}{2})(1-\theta)}) h^{-2p}.$$

Thus,

$$\eta_t \left(\mathcal{E}_{\mathbf{z}}(\tilde{f}_{t+1}) - \mathcal{E}_{\mathbf{z}}(f)\right) \leq \|\tilde{f}_t - f\|_K^2 - \|\tilde{f}_{t+1} - f\|_K^2 + \eta_t \Lambda_t(h, R).$$

Summing over $t = 1, \cdots, T-1$ with $\tilde{f}_1 = 0$

$$\begin{aligned}&\sum_{t=1}^{T-1} \eta_t \left(\mathcal{E}_{\mathbf{z}}(\tilde{f}_{t+1}) - \mathcal{E}_{\mathbf{z}}(f)\right) \\ \leq &\|\tilde{f}_1 - f\|_K^2 - \|\tilde{f}_T - f\|_K^2 + \sum_{t=1}^{T-1} \eta_t \Lambda_t(h, R) \leq \|f\|_K^2 + \sum_{t=1}^{T-1} \eta_t \Lambda_t(h, R) \\ \leq &R^2 + \sum_{t=1}^{T-1} \eta_t \Lambda_t(h, R).\end{aligned}$$

By (31), it is obvious that $\mathcal{E}_{\mathbf{z}}(\tilde{f}_{t+1}) \leq \mathcal{E}_{\mathbf{z}}(\tilde{f}_t) + 2^{6+3p} c_p C_G t^{(p+1)-(p+2)\theta} h^{-2p}$. Thus, for all $t \leq T$,

$$\begin{aligned}\mathcal{E}_{\mathbf{z}}(\tilde{f}_T) &\leq \mathcal{E}_{\mathbf{z}}(\tilde{f}_t) + 2^{6+3p} c_p C_G h^{-2p} \sum_{k=t}^{T-1} k^{(p+1)-(p+2)\theta} \\ &\leq \mathcal{E}_{\mathbf{z}}(\tilde{f}_t) + \frac{2^{6+3p} c_p C_G}{(p+2)(1-\theta)} h^{-2p} T^{(p+2)(1-\theta)},\end{aligned}$$

where we have used the fact $(p+2)\theta - (p+1) < 1$ and the elementary inequality

$$\sum_{k=1}^{T-1} k^{-s} \leq \frac{T^{1-s}}{1-s}, \qquad \forall\ s < 1. \tag{32}$$

It follows that

$$\begin{aligned}&\Big(\mathcal{E}_{\mathbf{z}}(\tilde{f}_T)-\mathcal{E}_{\mathbf{z}}(f)\Big)\sum_{t=1}^{T-1}\eta_t\\ \leq&\sum_{t=1}^{T-1}\eta_t\left(\mathcal{E}_{\mathbf{z}}(\tilde{f}_t)-\mathcal{E}_{\mathbf{z}}(f)+\frac{2^{6+3p}c_pC_Gh^{-2p}}{(p+2)(1-\theta)}T^{(p+2)(1-\theta)}\right)\\ \leq& R^2+\sum_{t=1}^{T-1}\eta_t\Lambda_t(h,R)+\frac{2^{6+3p}c_pC_Gh^{-2p}}{(p+2)(1-\theta)}T^{(p+2)(1-\theta)}\sum_{t=1}^{T-1}\eta_t. \qquad (33)\end{aligned}$$

By (32) again we have

$$\sum_{t=1}^{T-1}\eta_t\Lambda_t(h,R)\leq c_{p,\theta}\eta\big(T^{(p+2)(1-\theta)}+RT^{(p+\frac{3}{2})(1-\theta)}\big)h^{-2p}$$

where $c_{p,\theta}=2^{6+3p}c_pC_G(\frac{1}{(p+2)(1-\theta)}+\frac{1}{(p+\frac{3}{2})(1-\theta)})$. Note that

$$\sum_{t=1}^{T-1}\eta_t=\eta\sum_{t=1}^{T-1}t^{-\theta}\geq\frac{\eta}{2}T^{1-\theta}.$$

Plugging these two estimate into (33), we obtain the desired conclusion (30) with $c'_{p,\theta}=c_{p,\theta}+\frac{2^{6+3p}c_pC_G}{(p+2)(1-\theta)}$. □

Now we can prove the main theorems for the unregularized gradient descent algorithm.

Proof of Theorem 3. Fix $\lambda>0$ which will be chosen later. Note that $\|f_\lambda\|_K\leq\frac{\mathcal{D}(\lambda)}{\lambda}\leq\sqrt{\mathcal{D}_0}\lambda^{\frac{\beta-1}{2}}$. Applying Proposition 23 with $f=f_\lambda$, we obtain

$$\mathcal{Q}_1\leq\frac{2\mathcal{D}_0\lambda^{\beta-1}}{\eta}T^{\theta-1}+\frac{2c'_{p,\theta}}{\eta}\left(T^{(p+2)(1-\theta)}+\sqrt{\mathcal{D}_0}\lambda^{\frac{\beta-1}{2}}T^{(p+\frac{1}{2})(1-\theta)}\right)h^{-2p}.$$

Applying Proposition 22 with $f=\tilde{f}_T$ and $R=T^{\frac{1-\theta}{2}}$, we know that with confidence at least $1-\frac{\delta}{2}$,

$$\mathcal{Q}_2\leq\frac{1}{2}\mathcal{E}(\tilde{f}_T)-\mathcal{E}(f^*)+2^{13}T^{1-\theta}\max\left\{\frac{\log(\frac{4}{\delta})}{m-1},\left(\frac{c_q}{m-1}\right)^{\frac{1}{q+1}}\right\}$$

These two estimates together with the estimate of $\mathcal{Q}_3$ in Proposition 20 imply that, with confidence at least $1-\delta$,

$$\begin{aligned}\mathcal{E}(\tilde{f}_T)-\mathcal{E}(f^*) \leq\ & \frac{4\mathcal{D}_0\lambda^{\beta-1}}{\eta}T^{\theta-1} \\ & + \frac{4c'_{p,\theta}}{\eta}\left(T^{(p+2)(1-\theta)}+\sqrt{\mathcal{D}_0}\lambda^{\frac{\beta-1}{2}}T^{(p+\frac{1}{2})(1-\theta)}\right)h^{-2p} \\ & + 2^{14}T^{1-\theta}\max\left\{\frac{\log(\frac{4}{\delta})}{m-1}, \left(\frac{c_q}{m-1}\right)^{\frac{1}{q+1}}\right\} \\ & + \frac{2^{10}\mathcal{D}_0\lambda^{\beta-1}\log(\frac{4}{\delta})}{(m-1)}+3\mathcal{D}_0\lambda^{\beta}.\end{aligned}$$

Choosing $\lambda = T^{-(1-\theta)}$, with confidence at least $1-\delta$, we have

$$\mathcal{E}(\tilde{f}_T)-\mathcal{E}(f^*) \leq \tilde{C}\max\left\{\frac{1}{T^{\beta(1-\theta)}}, \frac{T^{1-\theta}}{(m-1)^{\frac{1}{1+q}}}, \frac{T^{(p+2)(1-\theta)}}{h^{2p}}\right\}\log\frac{4}{\delta}$$

where $\tilde{C} = \max\left\{\frac{4\mathcal{D}_0}{\eta_1} + \frac{4c'_{p,\theta}(1+\sqrt{\mathcal{D}_0})}{\eta_1} + (2^{10}+3)\mathcal{D}_0 + 2^{14}\max\{1, (c_q)^{\frac{1}{q+1}}\}\right\}$. This proves the theorem. □

Acknowledgement

The work is partially supported by National Natural Science Foundation of China [Grant No. 12071356] and the Research Grants Council of Hong Kong [Project No. CityU 11306318].

References

[1] M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 2009.

[2] F. Bauer, S. Pereverzev, and L. Rosasco. On regularization algorithms in learning theory. *Journal of complexity*, 23(1):52–72, 2007.

[3] A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

[4] A. Caponnetto and Y. Yao. Cross-validation based adaptation for regularization operators in learning theory. *Analysis and Applications*, 8(02):161–183, 2010.

[5] B. Chen and J. C. Principe. Stochastic gradient algorithm under (h, ϕ)-entropy criterion. *Circuits, Systems, and Signal Processing*, 26(6):941–960, 2007.

[6] B. Chen, P. Zhu, and J. C. Principe. Survival information potential: a new criterion for adaptive system training. *IEEE Transactions on Signal Processing*, 60(3):1184–1194, 2012.

[7] B. Chen, Y. Zhu, and J. Hu. Mean-square convergence analysis of adaline training with minimum error entropy criterion. *IEEE Transactions on Neural Networks*, 21(7):1168–1179, 2010.

[8] D.-R. Chen, Q. Wu, Y. Ying, and D.-X. Zhou. Support vector machine soft margin classifiers: error analysis. *Journal of Machine Learning Research*, 5:1143–1175, 2004.

[9] H. Chen. The convergence rate of a regularized ranking algorithm. *Journal of Approximation Theory*, 164(12):1513–1519, 2012.

[10] F. Cucker and S. Smale. Best choices for regularization parameters in learning theory: on the bias-variance problem. *Foundations of Computational Mathematics*, 2(4):413–428, 2002.

[11] E. De Vito, S. Pereverzyev, and L. Rosasco. Adaptive kernel methods using the balancing principle. *Foundations of Computational Mathematics*, 10(4):455–479, 2010.

[12] D. Erdogmus, K. Hild II, and J. C. Principe. Blind source separation using Rényi's α-marginal entropies. *Neurocomputing*, 49:25–38, 2002.

[13] D. Erdogmus and J. C. Principe. Comparison of entropy and mean square error criteria in adaptive system training using higher order statistics. In *Proceedings of the Intl. Conf. on ICA and Signal Separation*, pages 75–90. Berlin: Springer-Verlag, 2000.

[14] D. Erdogmus and J. C. Principe. Convergence properties and data efficiency of the minimum error entropy criterion in adaline training. *IEEE Transactions on Signal Processing*, 51:1966–1978, 2003.

[15] J. Fan, T. Hu, Q. Wu, and D.-X. Zhou. Consistency analysis of an empirical minimum error entropy algorithm. *Applied and Computational Harmonic Analysis*, 41:164–189, 2016.

[16] E. Gokcay and J. C. Principe. Information theoretic clustering. *IEEE Transactions on Pattern Analysis and Machine Learning*, 24:2:158–171, 2002.

[17] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

[18] T. Hu, J. Fan, Q. Wu, and D.-X. Zhou. Learning theory approach to a minimum error entropy criterion. *Journal of Machine Learning Research*, 14:377–397, 2013.

[19] T. Hu, J. Fan, Q. Wu, and D.-X. Zhou. Regularization schemes for minimum error entropy principle. *Analysis and Applications*, 13(4):437–455, 2015.

[20] T. Hu, Q. Wu, and D.-X. Zhou. Convergence of gradient descent method for minimum error entropy principle in linear regression. *IEEE Transactions on Signal Processing*, 64(24):6571–6579, 2016.

[21] T. Hu, Q. Wu, and D.-X. Zhou. Distributed kernel gradient descent algorithm for minimum error entropy principle. *Applied and Computational Harmonic Analysis*, 49(1):229–256, 2020.

[22] J. Lin, L. Rosasco, and D.-X. Zhou. Iterative regularization for learning with convex loss functions. *Journal of Machine Learning Research*, 17:1–38, 2016.

[23] J. Lin, A. Rudi, L. Rosasco, and V. Cevher. Optimal rates for spectral algorithms with least-squares regression over hilbert spaces. *Applied and Computational Harmonic Analysis*, 48(3):868–890, 2020.

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

[25] G. Raskutti, M. J. Wainwright, and B. Yu. Early stopping and non-parametric regression: an optimal data-dependent stopping rule. *Journal of Machine Learning Research*, 15(1):335–366, 2014.

[26] P. Shen and C. Li. Minimum total error entropy method for parameter estimation. *IEEE Transactions on Signal Processing*, 63(15):4079–4090, 2015.

[27] L. M. Silva, J. Marques de Sá, and L. A. Alexandre. Neural network classification using Shannon's entropy. In *Proceedings of the European Symposium on Artificial Neural Networks*, pages 217–222. Bruges: d-side, 2005.

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

[29] S. Smale and D.-X. Zhou. Estimating the approximation error in learning theory. *Analysis and Applications*, 1:17–41, 2003.

[30] I. Steinwart and A. Christmann. *Support Vector Machines.* Springer, 2008.

[31] I. Steinwart, D. Hush, and C. Scovel. Optimal rates for regularized least squares regression. In S. Dasgupta and A. Klivans, editors, *Proceedings of the 22nd Annual Conference on Learning Theory*, pages 79–93, 2009.

[32] Q. Wu, Y. Ying, and D.-X. Zhou. Learning rates of least-square regularized regression. *Foundations of Computational Mathematics*, 6(2):171–192, 2006.

[33] Z. Wu, S. Peng, W. Ma, B. Chen, and J. C. Principe. Minimum error entropy algorithms with sparsity penalty constraints. *Entropy*, 17(5):3419–3437, 2015.

[34] Y. Yao, L. Rosasco, and A. Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

[35] Y. Ying and M. Pontil. Online gradient descent learning algorithms. *Foundations of Computational Mathematics*, 8(5):561–596, 2008.

[36] Y. Ying and D.-X. Zhou. Online regularized classification algorithms. *IEEE Transactions on Information Theory*, 52(11):4775–4788, 2006.

[37] Y. Ying and D.-X. Zhou. Unregularized online learning algorithms with general loss functions. *Applied and Computational Harmonic Analysis*, 2015.

[38] Y. Ying and D.-X. Zhou. Online pairwise learning algorithms. *Neural Computation*, 28(4):743–777, 2016.

[39] T. Zhang. Leave-one-out bounds for kernel methods. *Neural computation*, 15(6):1397–1437, 2003.

[40] Y. Zhao, J. Fan, and L. Shi. Learning rates for regularized least squares ranking algorithm. *Analysis and Applications*, 15(06):815–836, 2017.

[41] D.-X. Zhou. The covering number in learning theory. *Journal of Complexity*, 18(3):739–767, 2002.

[42] D.-X. Zhou. Capacity of reproducing kernel spaces in learning theory. *IEEE Transactions on Information Theory*, 49(7):1743–1752, 2003.