entropy

MDPI

Article

Online Gradient Descent for Kernel-Based Maximum Correntropy Criterion

Baobin Wang [1] and Ting Hu [2,*]

[1] School of Mathematics and Statistics, South-Central University for Nationalities, Wuhan 430074, China
[2] School of Mathematics and Statistics, Wuhan University, Wuhan 430072, China
* Correspondence: tinghu@whu.edu.cn

Received: 15 May 2019; Accepted: 24 June 2019; Published: 29 June 2019

Abstract: In the framework of statistical learning, we study the online gradient descent algorithm generated by the correntropy-induced losses in Reproducing kernel Hilbert spaces (RKHS). As a generalized correlation measurement, correntropy has been widely applied in practice, owing to its prominent merits on robustness. Although the online gradient descent method is an efficient way to deal with the maximum correntropy criterion (MCC) in non-parameter estimation, there has been no consistency in analysis or rigorous error bounds. We provide a theoretical understanding of the online algorithm for MCC, and show that, with a suitable chosen scaling parameter, its convergence rate can be min–max optimal (up to a logarithmic factor) in the regression analysis. Our results show that the scaling parameter plays an essential role in both robustness and consistency.

Keywords: correntropy; maximum correntropy criterion; online algorithm; robustness; reproducing kernel Hilbert spaces

1. Introduction

Regression analysis is an important problem in many fields of science. The traditional least squares method may be the most used algorithm for regression in practice. However, it only relies on the mean squared error and belongs to second-order statistics, whose optimality depends heavily on the assumption of Gaussian noise. Thus, it usually performs poorly when the noise is not normally distributed. Alterative approaches have been proposed to deal with outliers or heavy-tailed distributions. A generalized correlation function named correntropy [1] is introduced as a substitute for the least squares loss, and the maximum correntropy criterion (MCC) [2–5] is used to improve robustness in situations of non-Gaussian and heavy-tailed error distributions. Recently, MCC has been succeeded in many real applications, e.g., wind power forecasting and pattern recognition [6,7].

In the standard framework of statistical learning, let $X \in \mathcal{R}^n$ be an explanatory variable with values taken in a compact metric space $(\mathcal{X}, d)$, Y be a real response variable with $Y \in \mathcal{Y} \subset \mathcal{R}$. Here we investigate the application of MCC in the following regression model

$$Y = f_\rho(X) + \epsilon, \quad \mathbb{E}(\epsilon | X = x) = 0,$$

where ϵ is the noise and $f_\rho(x)$ is the regression function, defined as the conditional mean $\mathbb{E}(Y|X = x)$ at each $x \in \mathcal{X}$. The purpose of regression is to estimate the unknown target function f_ρ according to the sample $\mathbf{z} = \{z_i = (x_i, y_i)\}_{i=1}^T$, which is drawn independently from the underlying unknown probability distribution ρ on $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$. For a hypothesis function $f : \mathcal{X} \to \mathcal{Y}$, with the scaling parameter $\sigma > 0$, the correntropy between $f(X)$ and Y is defined by $V_\sigma(f) := \mathbb{E}G\left(\frac{(f(X)-Y)^2}{2\sigma^2}\right)$ where $G(u)$ is the Exponential function $\exp\{-u\}, u \in \mathbb{R}$. For the given sample $\mathbf{z}$, the empirical form of V_σ is

$\hat{V}_\sigma(f) := \frac{1}{T}\sum_{i=1}^{T} G\left(\frac{(f(x_i)-y_i)^2}{2\sigma^2}\right)$. When applied to regression problems, MCC intends to maximize the empirical correntropy $\hat{V}_\sigma$ over a certain underlying hypothesis space $\mathcal{H}$, that is

$$f_{\mathbf{z},\mathcal{H}} := \arg\max_{f\in\mathcal{H}} \hat{V}_\sigma(f). \tag{1}$$

MCC in regression problems has shown its efficiency for cases when the noises are non-Gaussian, and also with large outliers, see [8–10]. It also has drawn much attention in the signal processing, machine learning and optimization communities [2,5,11–14].

Let $K : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ be a *Mercer kernel*, i.e., a continuous, symmetric and positive semi-definite function. We say that K is a positive semi-definite, if for any finite set $\{u_1, \cdots, u_m\} \subset \mathcal{X}$ and $m \in \mathbb{N}$, the matrix $\left(K(u_i, u_j)\right)_{i,j=1}^{m}$ is positive semi-definite. An RKHS $(\mathcal{H}_K, \|\cdot\|_K)$ associated with the Mercer kernel K is defined as the completion of the linear span of the functions set $\{K_x := K(x,\cdot), x \in \mathcal{X}\}$. It has the reproducing property

$$f(x) = \langle f, K_x\rangle_K \tag{2}$$

for any $f \in \mathcal{H}_K$ and $x \in \mathcal{X}$. Since $\mathcal{X}$ is compact, the RKHS $\mathcal{H}_K$ is contained in $C(\mathcal{X})$, the space of continuous functions on $\mathcal{X}$ with the norm $\|f\|_\infty := \sup_{x\in\mathcal{X}} |f(x)|$. Moreover, if $\mathcal{X}$ is a Euclidean ball in $\mathbb{R}^n$ with some $\alpha > \frac{n}{2}$, then the Sobolev space $H^\alpha(\mathcal{X})$ is an RKHS. For more families of RKHS in statistical learning, one can refer to [15]. Denote $\kappa := \sup_{x\in\mathcal{X}} \sqrt{K(x,x)}$, then, by the reproducing property (2), there holds

$$\|f\|_\infty \le \kappa\|f\|_K, \ \textit{for any } f \in \mathcal{H}_K. \tag{3}$$

Denote $\ell_\sigma : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ as the correntropy induced regression loss, given by

$$\ell_\sigma(u,v) := \sigma^2\left(1 - G\left(\frac{(u-v)^2}{2\sigma^2}\right)\right) = \sigma^2\left(1 - \exp\left\{-\frac{(u-v)^2}{2\sigma^2}\right\}\right).$$

Associated with this regression loss ℓ_σ and the RKHS $\mathcal{H}_K$, MCC for regression (1) in the context of learning theory is reformulated as

$$f_{\mathbf{z}} := \arg\min_{f\in\mathcal{H}_K} \frac{1}{T}\sum_{i=1}^{T} \ell_\sigma(f(x_i), y_i). \tag{4}$$

Notice that ℓ_σ is not convex, MCC algorithms are usually implemented by various gradient descent methods [14,16,17]. In this paper, we take the online gradient descent method as follows to solve the above optimization scheme (4) since it is scalable to large datasets and applicable to situations where the samples are presented in sequence.

Definition 1. *Given the sample* $\mathbf{z} = \{z_i = (x_i, y_i)\}_{i=1}^{T}$, *the online gradient descent method for MCC is defined by* $f_1 = 0$, *and*

$$f_{t+1} = f_t - \eta\ell'_\sigma(f_t(x_t), y_t)K_{x_t}, \quad t \in \mathbb{N}, \tag{5}$$

where $\eta > 0$ *is the step size and* ℓ'_σ *denotes the derivative of* ℓ_σ *with respect to the first variable.*

In the literature, most MCC algorithms have been implemented for linear models and cannot be applied to analysis of data with nonlinear structures. Kernel methods provide efficient non-parametric learning algorithms for dealing with nonlinear features. So, RKHS are used in this work as hypothesis spaces in the design of learning algorithms.

An online algorithm for MCC has been used in practical applications for more than one decade, but there still is a lack of the theoretical guarantee or strict analysis for its asymptotical convergence. Because the optimization problem arising from MCC is not convex, the global optimization convergence of the online algorithm (5) for MCC is not unconditionally guaranteed. This also makes the theoretical analysis for MCC essentially difficult. In fact, vast numerical studies show that MCC can lead robust estimators while keeping convenient convergence properties. Thus, our goal is to fill the gap between the theoretical analysis and the optimization process so that the output function of the online algorithm (5) can converge to a global minima while the existing work can not ensure the global optimization of this output. To this end, we study the approximation ability of f_{T+1} generated by (5) at the T-iteration to the regression function f_ρ. We derive the explicit error rate for (5) with suitable choice of step sizes, which is competitive with those in the regression analysis. In this work, we show that the scaling parameter σ plays an important role in providing robustness and a fast convergence rate.

2. Preliminaries and Main Results

We begin with some preliminaries and notations. Throughout the paper, we assume that the unknown distribution ρ on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ can be decomposed into the marginal distribution $\rho_{\mathcal{X}}$ on $\mathcal{X}$ and the conditional distribution $\rho(\cdot|x)$ at each $x \in \mathcal{X}$. We also require that $|Y| < M$ almost surely for some $M > 1$. In the regression analysis, the approximation power of f_{T+1} by (5) is usually measured in terms of the mean squared error in $L^2_{\rho_{\mathcal{X}}}$-metric $\|f_{T+1} - f_\rho\|_\rho$, that is defined as $\|\cdot\|_\rho = \|\cdot\|_{L^2_{\rho_{\mathcal{X}}}} := \left(\int_{\mathcal{X}} |\cdot|^2 d\rho_{\mathcal{X}}\right)^{\frac{1}{2}}$.

To present our main result for the error bound of $f_{T+1} - f_\rho$, the assumption on the target function f_ρ will be given as below. Define an integral operator $L_K : L^2_{\rho_{\mathcal{X}}} \longrightarrow L^2_{\rho_{\mathcal{X}}}$ associated with the kernel K by

$$L_K(f) := \int_{\mathcal{X}} f(x) K_x d\rho_{\mathcal{X}}, \quad f \in L^2_{\rho_{\mathcal{X}}}.$$

By the reproducing property (2) of $\mathcal{H}_K$, for any $f \in \mathcal{H}_K$, it can be expressed as

$$L_K(f) = \int_{\mathcal{X}} \langle f, K_x \rangle_K K_x d\rho_{\mathcal{X}}. \tag{6}$$

Since K is a Mercer kernel, L_K is compact and positive. Denote L_K^r as the r-th power of L_K, then it is well defined for any $r > 0$ by the spectral theorem. Let $\{\lambda_i\}_{i \geq 1}$ be the eigenvalues of L_K, arranged in decreasing order. The corresponding eigenfunctions $\{\phi_i\}_{i \geq 1}$ form an orthonormal basis of $L^2_{\rho_{\mathcal{X}}}$ space. Hence, the regularity space $L_K^r(L^2_{\rho_{\mathcal{X}}})$ is expressed as [18]

$$L_K^r(L^2_{\rho_{\mathcal{X}}}) := \left\{ f = \sum_{i=1}^{\infty} \lambda_j^r a_i \phi_i : \|L_K^{-r} f\|_\rho = \sum_{i=1}^{\infty} a_i^2 < \infty \right\}.$$

It implies that for any $r_1 > r_2 > 0$, there holds $L_K^{r_1}(L^2_{\rho_{\mathcal{X}}}) \subset L_K^{r_2}(L^2_{\rho_{\mathcal{X}}})$. In particular, we know that $L_K^r(L^2_{\rho_{\mathcal{X}}}) \subseteq \mathcal{H}_K$ for any $r \geq \frac{1}{2}$ and $L_K^{\frac{1}{2}}(L^2_{\rho_{\mathcal{X}}}) = \mathcal{H}_K$ satisfying

$$\|f\|_K = \|L_K^{-\frac{1}{2}} f\|_\rho, \quad \forall f \in \mathcal{H}_K. \tag{7}$$

Throughout the paper, the *regularity assumption* holds for f_ρ, i.e.,

$$f_\rho = L_K^r(g), \quad \text{for some } r > 0 \text{ and } g \in L^2_{\rho_{\mathcal{X}^2}}, \tag{8}$$

and $\|L^{-r} f_\rho\|_\rho = \|g\|_\rho$.

This assumption is called the source condition [19] in inverse problems and it characterizes the smoothness of the target function f_ρ. Obviously, the larger the parameter r is, the higher the regularity of f_ρ is. The general source conditions considered in inverse problems usually take the form of

$$f_\rho = \psi(L_K)h, \ for\ some\ h \in \mathcal{H}_K \tag{9}$$

where ψ is non-decreasing and $\psi(0) = 0$, called the index function. It is clear that when $r > \frac{1}{2}$, The above assumption is a special case of (9) with $\psi(L_K) = L_K^{r-\frac{1}{2}}$ and $h = L_K^{\frac{1}{2}}g$. It should be pointed that our analysis in this work also can applied to more general cases by taking source conditions (9).

We are now in a position to state our convergence rate for (5) in $L^2_{\rho_\mathcal{X}}$-space as well as in $\mathcal{H}_K$ by choosing the step size $\eta := \eta(T)$. For brevity, let $\kappa = 1$ without losing generality and denote the expectation $\mathbb{E}_{z_1,\cdots,z_t}$ as $\mathbb{E}_t$ for each $t \in \mathbb{N}$.

Theorem 1. *Define* $\{f_t\}_{t=1}^{T+1}$ *by (5). Suppose that the assumption (8) holds for* $r > 0$. *Take* $\eta = T^{-\frac{2r}{2r+1}}$ *and* $T > \left(24\left((1/2e)^{1/2}+1\right)^2\log(T)\right)^{\frac{2r+1}{2r}}$, *then*

$$\mathbb{E}_T\left[\|f_{T+1} - f_\rho\|_\rho^2\right] \le C\max\left\{T^{-\frac{2r}{2r+1}}\log(T), T^{\frac{5}{2r+1}}\sigma^{-4}\right\} \tag{10}$$

and if $r > \frac{1}{2}$,

$$\mathbb{E}_T\left[\|f_{T+1} - f_\rho\|_K^2\right] \le C'\max\left\{T^{-\frac{2r-1}{2r+1}}, T^{\frac{5}{2r+1}}\sigma^{-4}\right\} \tag{11}$$

where the constants C, C' *are independent of* T, σ, *and will be given in the proof.*

Remark 1. *Besides the error* $\|f_{T+1} - f_\rho\|_\rho$, *the error bound (11) in* $\mathcal{H}_K$*-norm is also given if* $r > \frac{1}{2}$, *i.e.,* $f_\rho \in \mathcal{H}_K$. *By (3), it leads the pointwise convergence of* f_{T+1} *to* f_ρ *since for each* $u \in \mathcal{X}$, $|f_{T+1}(u) - f_\rho(u)| \le \|f_{T+1} - f_\rho\|_K$. *Compared with the global error* $\|\|f_{T+1} - f_\rho\|_\rho$, *the error rate in* $\mathcal{H}_K$ *characterizes the local performance of (5) and is much stronger. Furthermore [18], when the kernel* K *lies in* $C^\alpha(\mathcal{X}\times\mathcal{X})$ *for some* $\alpha > 0$, *its associated RKHS* $\mathcal{H}_K$ *can be embedded into* $C^{\alpha/2}(\mathcal{X})$, *whose partial derivative up to order* $\alpha/2$ *are continuous with* $\|f\|_{C^{\alpha/2}(\mathcal{X})} = \sum_{|s|\le\frac{\alpha}{2}}\|D^{\frac{\alpha}{2}}f\|_\infty$. *So, the convergence in* $\mathcal{H}_K$ *implies that* f_{T+1} *will converge to* f_ρ *in* $C^{\frac{\alpha}{2}}$, *that ensures the convergence of the derivatives of* f_{T+1} *to those of* f_ρ.

Remark 2. *It has been proved in [20] that the min–max optimal rate for regression problems is of order* $O\left(T^{-\frac{2r}{2r+s}}\right)$ *when there exists constants* $C_s > 0$, $0 < s \le 1$ *such that the following effective dimension condition holds, i.e.,*

$$Trace((L_K + \lambda I)^{-1}L_K) \le C_s\lambda^{-s}, for\ any\ \lambda > 0,$$

where $Trace(\cdot)$ *denotes the trace of the operator. This condition measures the complexity [15,20,21] of* $\mathcal{H}_K$ *with respect to the marginal distribution* $\rho_\mathcal{X}$. *It is always satisfied with* $s = 1$ *by taking the constant* $C_s = Trace(L_K)$. *Hence, the min–max optimal rate for capacity-independent cases is of order* $O\left(T^{-\frac{2r}{2r+1}}\right)$ *by taking a universal parameter* $s = 1$.

When $\sigma \ge T^{\frac{2r+5}{4(2r+1)}}$, *we see that our convergence rate in* $L^2_{\rho_\mathcal{X}}$*-norm is of order* $O\left(T^{-\frac{2r}{2r+1}}\log(T)\right)$. *Thus, it is nearly optimal in the capacity-independent sense that up to a logarithmic factor, it matches the min–max optimal rate above. We also find that the convergence rates (10) and (11) keep decreasing as the regularity parameter* r *increases. Hence, the online algorithm (5) does not suffer from the saturation phenomenon existing in Tikhonov regularization schemes [22], where the error rate of the estimators will not improve if* r *is out of the range* $(0, 1]$. *This again shows the advantage of the online algorithm (5).*

Remark 3. *Recent paper [2] investigated the approximation ability of the empirical scheme (4) over general hypothesis spaces $\mathcal{H}$. This work shows that with a complexity parameter $0 < \beta \leq 2$, their error rate is of order $O(T^{-\frac{2}{2+\beta}})$ if the scaling parameter $\sigma = T^{\frac{1}{2+\beta}}$. To be fair, do not take the capacity of $\mathcal{H}$ into consideration by taking $\beta = 2$. Then, their order reduces to $O(T^{-\frac{1}{2}})$, which is far from capacity-independent optimality and inferior to our rates.*

In the work [17], iterative regularization techniques (alternatively called early stopping) are taken to solve the optimization problems associated with general robust losses including the correntropy induced loss ℓ_σ, where the whole sample $\mathbf{z}$ are presented at each iteration. In their analysis, under the polynomial decay of the eigenvalues $\{\lambda_i\}$, that is, there exists some constants $C_b > 0$ and $b \geq 1$ such that

$$\lambda_i \leq c_b i^{-b}, \quad \forall i \geq 1,$$

the obtained rate is $O(T^{-\frac{2br}{2br+1}})$ if $r \geq \frac{1}{2}$, else, it is $O(T^{-\frac{2br}{b+1}})$. This decay is also a measurement for the complexity of $\mathcal{H}_K$, please refer to [21]. Recall that the compactness of $\mathcal{X}$ implies that $\sum_i \lambda_i < \infty$ and $\lambda_i \leq ci^{-1}$ for some $c > 0$. So, their rate for capacity-independent cases is $O(T^{-\frac{2r}{2r+1}})$ if $r \geq \frac{1}{2}$, else, it is $O(T^{-r})$. We can see that our results in (10) are superior in the case $0 < r < \frac{1}{2}$. It shows in theory that the online algorithm (5) for MCC can achieve better approximation rate when f_ρ is not in $\mathcal{H}_K$.

Remark 4. *It is easy to check that the roots of the second derivative of ℓ_σ is $\pm\sigma$, i.e., when $|f(x) - y| < \sigma$, this loss is convex and behaves as the least squares loss; when $|f(x) - y| \geq \sigma$, the loss function becomes concave and rapidly tends to be flat as the value of $|f(x) - y|$ goes to infinity. It implies that ℓ_σ satisfies the redescending property, and with a suitable chosen scaling parameter σ, ℓ_σ can reject gross outliers while keeping a prediction accuracy. In Theorem 1, we observe that σ should be large enough to guarantee the nice convergence, which coincides with the work in [2]. They also pointed that too small σ may prevent the estimator to converge to f_ρ. In a recent paper [23], correntropy with small σ is interpreted as modal regression. According to the above discussions and empirical studies [2,14,17], we conclude that the value of σ would determine the learning target and a moderate σ may be more appropriate for balancing the convergence and robustness in practice.*

Based on the above remarks, we see that the convergence rate of online kernel-based MCC is comparable to that of the least squares that has appeared in the literature [24]. Meanwhile, MCC's redescending property will produce robustness to various outliers including sub-Gaussain, Student's t-distribution, and Cauchy distribution. These all shows the superiority of MCC in a variety of applications, such as clustering, classification and feature selection [14]. At the end of this section, we would like to point out that although our work is carried out under the boundness condition of $\mathcal{Y}$, it can be extended to more general situations such as the moment conditions [20].

3. Proofs of Main Result

In this section, we prove our main results in Theorem 1. First, we derive the uniform bound for the iteration sequence $\{f_t\}_{t=1}^{T+1}$ by (5).

Lemma 1. *Define $\{f_t\}_{t=1}^{T+1}$ by (5). If $0 < \eta \leq 1$, then*

$$\|f_t\|_K \leq M\eta^{\frac{1}{2}}(t-1)^{\frac{1}{2}}, \quad t \in \mathbb{N}. \tag{12}$$

Proof. We prove (12) by induction. It is trivial that (12) holds for $t = 1$. Suppose (12) holds for $t \geq 2$. Notice that $\ell'_\sigma(f_t(x_t), y_t) = G\left(-\frac{(f_t(x_t)-y_t)^2}{2\sigma^2}\right)[f_t(x_t) - y_t]$. Write (12) as $f_{t+1} = f_t - \eta H_t$ where $H_t = G\left(-\frac{(f_t(x_t)-y_t)^2}{2\sigma^2}\right)[f_t(x_t) - y_t]K_{x_t}$. Then by (2),

$$\begin{aligned}\|f_{t+1}\|_K^2 &= \|f_t\|_K^2 - 2\eta\langle f_t, H_t\rangle_K + \eta^2\|H_t\|_K^2 \\ &= \|f_t\|_K^2 - 2\eta G\left(-\frac{(f_t(x_t)-y_t)^2}{2\sigma^2}\right)[f_t(x_t)-y_t]\, f_t(x_t) + \eta^2\|H_t\|_K^2\end{aligned}$$

and

$$\begin{aligned}\|H_t\|_K^2 &= G\left(-\frac{(f_t(x_t)-y_t)^2}{\sigma^2}\right)[f_t(x_t)-y_t]^2\, K_{(x_t,x_t)} \\ &\leq G\left(-\frac{(f_t(x_t)-y_t)^2}{\sigma^2}\right)[f_t(x_t)-y_t]^2.\end{aligned}$$

Then, we have

$$\|f_{t+1}\|_K^2 \leq \|f_t\|_K^2 + \eta\left\{\eta G\left(-\frac{(f_t(x_t)-y_t)^2}{2\sigma^2}\right)[f_t(x_t)-y_t]^2 - 2(f_t(x_t)-y_t)f_t(x_t)\right\} G\left(-\frac{(f_t(x_t)-y_t)^2}{2\sigma^2}\right).$$

For the part of the above inequality, we have

$$\begin{aligned}&\eta G\left(-\frac{(f_t(x_t)-y_t)^2}{2\sigma^2}\right)[f_t(x_t)-y_t]^2 - 2(f_t(x_t)-y_t)f_t(x_t) \\ &= \left(\eta G\left(-\frac{(f_t(x_t)-y_t)^2}{2\sigma^2}\right)-2\right)\left((f_t(x_t)-y_t) - \frac{y_t}{\eta G\left(-\frac{(f_t(x_t)-y_t)^2}{2\sigma^2}\right)-2}\right)^2 \\ &+ \frac{y_t^2}{2-\eta G\left(-\frac{(f_t(x_t)-y_t)^2}{2\sigma^2}\right)}.\end{aligned}$$

Since $\eta \leq 1$, it follows that $\eta G\left(-\frac{(f_t(x_t)-y_t)^2}{2\sigma^2}\right) - 2 < 0$ and $2 - \eta G\left(-\frac{(f_t(x_t)-y_t)^2}{2\sigma^2}\right) \geq 1$. Recall that $|y| \leq M$ for all $y \in \mathcal{Y}$, then

$$\eta G\left(-\frac{(f_t(x_t)-y_t)^2}{2\sigma^2}\right)[f_t(x_t)-y_t]^2 - 2(f_t(x_t)-y_t)f_t(x_t) \leq \frac{y_t^2}{2-\eta G\left(-\frac{(f_t(x_t)-y_t)^2}{2\sigma^2}\right)} \leq M^2.$$

Based on the above analysis,

$$\|f_{t+1}\|_K^2 \leq \|f_t\|_K^2 + \eta M^2 G\left(-\frac{(f_t(x_t)-y_t)^2}{2\sigma^2}\right) \leq \|f_t\|_K^2 + \eta M^2 \leq M^2\eta(t-1) + \eta M^2 = M^2\eta t.$$

Then the proof is completed. □

Next, we will establish a proposition which is crucial to prove the convergence rates in Theorem 1. It is closely related to the generalization error of f_t. Define the *generalization error* $\mathcal{E}(f)$ for any measurable function $f : \mathcal{X} \to \mathbb{R}$ by

$$\mathcal{E}(f) = \int_{\mathcal{Z}} (f(x)-y)^2 d\rho.$$

The regression function f_ρ that we want to learn or approximate is a minimizer of $\mathcal{E}(f)$, that is

$$f_\rho = \arg\min\{\mathcal{E}(f) : f \textit{ is a measurable function from } \mathcal{X} \textit{ to } \mathcal{Y}\}.$$

A simple computation yields the relation for $f : \mathcal{X} \to \mathbb{R}$

$$\|f - f_\rho\|_\rho^2 = \mathcal{E}(f) - \mathcal{E}(f_\rho). \tag{13}$$

For brevity, set the operator $\pi_k^t(L_K) := \prod_{j=k}^t (I - \eta L_K)$ for $k, t \in \mathbb{N}$ and $\pi_{t+1}^t(L_K) := I$.

Proposition 1. *Define* $\{f_t\}_{t=1}^{T+1}$ *by (5). If the step size* $0 < \eta < 1$, *then we have*

$$\mathbb{E}_T\left[\|f_{T+1} - f_\rho\|_\rho^2\right] \leq 2\left\|\pi_1^T(L_K) f_\rho\right\|_\rho^2 + 2\eta^2 \sum_{t=1}^T \frac{2\left((1/2e)^{1/2}+1\right)^2}{1+\eta(T-t)} \mathbb{E}_{t-1}\left[\mathcal{E}(f_t)\right] + 2\mathbb{E}_T\left[\left\|\eta \sum_{t=1}^T \pi_{t+1}^T(L_K)\Delta_t\right\|_\rho^2\right], \tag{14}$$

furthermore, if $f_\rho \in \mathcal{H}_K$,

$$\mathbb{E}_T\left[\|f_{T+1} - f_\rho\|_K^2\right] \leq 2\left\|\pi_1^T(L_K) f_\rho\right\|_K^2 + 2\eta^2 \sum_{t=1}^T \mathbb{E}_{t-1}\left[\mathcal{E}(f_t)\right] + 2\mathbb{E}_T\left[\left\|\eta \sum_{t=1}^T \pi_{t+1}^T(L_K)\Delta_t\right\|_K^2\right], \tag{15}$$

where Δ_t *is defined in the proof.*

Proof. Denote

$$\begin{aligned}\Delta_t &= \left[G(0) - G\left(-\frac{(f_t(x_t)-y_t)^2}{2\sigma^2}\right)\right][f_t(x_t) - y_t] K_{x_t} \\ &= \left[1 - G\left(-\frac{(f_t(x_t)-y_t)^2}{2\sigma^2}\right)\right][f_t(x_t) - y_t] K_{x_t}\end{aligned} \tag{16}$$

and define a random variable $\xi(f_t, z_t) := L_K(f_t - f_\rho) - (f_t(x_t) - y_t)K_{x_t}$.

By (5), we have that for any $t \in \mathbb{N}$,

$$f_{t+1} - f_\rho = f_t - f_\rho - \eta\left[f_t(x_t) - y_t\right] K_{x_t} + \eta\Delta_t = (I - \eta L_K)(f_t - f_\rho) + \eta\xi(f_t, z_t) + \eta\Delta_t.$$

Applying the above equality iteratively from $t = T$ to $t = 1$, we get that by $f_1 = 0$,

$$f_{T+1} - f_\rho = -\pi_1^T(L_K) f_\rho + \eta \sum_{t=1}^T \pi_{t+1}^T(L_K)\xi(f_t, z_t) + \eta \sum_{t=1}^T \pi_{t+1}^T(L_K)\Delta_t. \tag{17}$$

It follows from the elementary inequality that $\|g_1 + g_2\|_\rho^2 \leq 2\|g_1\|_\rho^2 + 2\|g_2\|_\rho^2$ for any $g_1, g_2 \in L^2_{\rho_\mathcal{X}}$, that

$$\mathbb{E}_T\left[\|f_{T+1} - f_\rho\|_\rho^2\right] \leq 2\mathbb{E}_T\left[\left\|-\pi_1^T(L_K) f_\rho + \eta \sum_{t=1}^T \pi_{t+1}^T(L_K)\xi(f_t, z_t)\right\|_\rho^2\right] + 2\mathbb{E}_T\left[\left\|\eta \sum_{t=1}^T \pi_{t+1}^T(L_K)\Delta_t\right\|_\rho^2\right]. \tag{18}$$

To prove (14), we consider the part of the first term on the right-hand side of (18)

$$\begin{aligned}\mathbb{E}_T\left[\left\|-\pi_1^T(L_K) f_\rho + \eta \sum_{t=1}^T \pi_{t+1}^T(L_K)\xi(f_t, z_t)\right\|_\rho^2\right] &= \left\|\pi_1^T(L_K) f_\rho\right\|_\rho^2 \\ + \mathbb{E}_T\left[\left\|\eta \sum_{t=1}^T \pi_{t+1}^T(L_K)\xi(f_t, z_t)\right\|_\rho^2\right] &- 2\mathbb{E}_T\left[\left\langle \pi_1^T(L_K) f_\rho, \eta \sum_{t=1}^T \pi_{t+1}^T(L_K)\xi(f_t, z_t)\right\rangle_\rho\right].\end{aligned} \tag{19}$$

Observe that f_t is only dependent on $\{z_1, \cdots, z_{t-1}\}$, not on z_t. Thus, by the fact that $\int_{\mathcal{X}} y d\rho = f_\rho(x)$, we have

$$\mathbb{E}_{z_t}\left[\xi(f_t, z_t)\right] = 0, \quad t = 1, \cdots, T. \tag{20}$$

We consider the second term on the right-hand side of (19). It can be rewritten as

$$\mathbb{E}_T\left[\left\|\eta \sum_{t=1}^{T} \pi_{t+1}^T(L_K)\xi(f_t, z_t)\right\|_\rho^2\right] = \eta^2 \sum_{t=1}^{T}\sum_{l=1}^{T} \mathbb{E}_T\left\langle \pi_{t+1}^T(L_K)\xi(f_t, z_t), \pi_{l+1}^T(L_K)\xi(f_l, z_l)\right\rangle_\rho.$$

When $t < l \leq T$, by (20),

$$\begin{aligned}&\mathbb{E}_T\left\langle \pi_{t+1}^T(L_K)\xi(f_t, z_t), \pi_{l+1}^T(L_K)\xi(f_l, z_l)\right\rangle_\rho = \mathbb{E}_{l-1}\mathbb{E}_{z_l}\left\langle \pi_{t+1}^T(L_K)\xi(f_t, z_t), \pi_{l+1}^T(L_K)\xi(f_l, z_l)\right\rangle_\rho \\ &= \mathbb{E}_{l-1}\left\langle \pi_{l+1}^T(L_K)\pi_{t+1}^T(L_K)\xi(f_t, z_t), \mathbb{E}_{z_l}\xi(f_l, z_l)\right\rangle_\rho = 0.\end{aligned}$$

Obviously, the above equality holds for $l < t \leq T$. So, with (7), we get

$$\begin{aligned}&\eta^2 \sum_{t=1}^{T}\sum_{l=1}^{T} \mathbb{E}_T\left\langle \pi_{t+1}^T(L_K)\xi(f_t, z_t), \pi_{l+1}^T(L_K)\xi(f_l, z_l)\right\rangle_\rho \\ &= \eta^2 \sum_{t=1}^{T} \mathbb{E}_t\left[\left\|\pi_{t+1}^T(L_K)\xi(f_t, z_t)\right\|_\rho^2\right] \leq \eta^2 \sum_{t=1}^{T}\left\|\pi_{t+1}^T(L_K)L_K^{\frac{1}{2}}\right\|^2 \mathbb{E}_t\left[\left\|L_K^{-\frac{1}{2}}\xi(f_t, z_t)\right\|_\rho^2\right] \\ &= \eta^2 \sum_{t=1}^{T}\left\|\pi_{t+1}^T(L_K)L_K^{\frac{1}{2}}\right\|^2 \mathbb{E}_t\left[\|\xi(f_t, z_t)\|_K^2\right].\end{aligned}$$

To bound $\mathbb{E}_t\left[\|\xi(f_t, z_t)\|_K^2\right]$, we have

$$\begin{aligned}&\mathbb{E}_t\left[\|\xi(f_t, z_t)\|_K^2\right] = \mathbb{E}_t\left[\|(f_t(x_t) - y_t)K_{x_t}\|_K^2\right] - \|\mathbb{E}_t\left[(f_t(x_t) - y_t)K_{x_t}\right]\|_K^2 \\ &\leq \mathbb{E}_{t-1}\mathbb{E}_{z_t}\left[\|(f_t(x_t) - y_t)K_{x_t}\|_K^2\right] \leq \mathbb{E}_{t-1}\mathbb{E}_{z_t}\left[(f_t(x_t) - y_t)^2\right] = \mathbb{E}_{t-1}\left[\mathcal{E}(f_t)\right]\end{aligned}$$

where the last inequality is derived from (3). Applying Lemma A1 with $\beta = \frac{1}{2}$, $l = t+1$ and $k = T$, we have

$$\begin{aligned}&\sum_{t=1}^{T}\left\|\pi_{t+1}^T(L_K)L_K^{\frac{1}{2}}\right\|^2 = \sum_{t=1}^{T-1}\left\|\pi_{t+1}^T(L_K)L_K^{\frac{1}{2}}\right\|^2 + \left\|\pi_{T+1}^T(L_K)L_K^{\frac{1}{2}}\right\|^2 \\ &\leq \sum_{t=1}^{T-1}\frac{2\left((1/2e)^{1/2}+1\right)^2}{1+\eta(T-t)} + 1 \leq \sum_{t=1}^{T}\frac{2\left((1/2e)^{1/2}+1\right)^2}{1+\eta(T-t)}.\end{aligned}$$

Based on the above analysis, we have

$$\mathbb{E}_T\left[\left\|\eta \sum_{t=1}^{T} \pi_{t+1}^T(L_K)\xi(f_t, z_t)\right\|_\rho^2\right] \leq \eta^2 \sum_{t=1}^{T}\frac{2\left((1/2e)^{1/2}+1\right)^2}{1+\eta(T-t)}\mathbb{E}_{t-1}\left[\mathcal{E}(f_t)\right]. \tag{21}$$

Now, we estimate the last term on the right-hand side of (19). Using (20) again, we have

$$\begin{aligned}&\mathbb{E}_T\left[\left\langle \pi_1^T(L_K)f_\rho, \eta\sum_{t=1}^T \pi_{t+1}^T(L_K)\xi(f_t,z_t)\right\rangle_\rho\right]\\&=\left\langle \pi_1^T(L_K)f_\rho, \eta\sum_{t=1}^T \pi_{t+1}^T(L_K)\mathbb{E}_{t-1}\mathbb{E}_{z_t}\left[\xi(f_t,z_t)\right]\right\rangle_\rho=0.\end{aligned}\tag{22}$$

Plugging (21) and (22) into (19), we get

$$\begin{aligned}&\mathbb{E}_T\left[\left\|-\pi_1^T(L_K)f_\rho+\eta\sum_{t=1}^T\pi_{t+1}^T(L_K)\xi(f_t,z_t)\right\|_\rho^2\right]=\left\|\pi_1^T(L_K)f_\rho\right\|_\rho^2\\&+\eta^2\sum_{t=1}^T\frac{2\left((1/2e)^{1/2}+1\right)^2}{1+\eta(T-t)}\mathbb{E}_{t-1}\left[\mathcal{E}(f_t)\right].\end{aligned}\tag{23}$$

This together with (18) yields the desired conclusion (14).

Now we turn to bound $f_{T+1}-f_\rho$ in $\mathcal{H}_K$-norm. By (17) again, we have

$$\mathbb{E}_T\left[\|f_{T+1}-f_\rho\|_K^2\right]\le 2\mathbb{E}_T\left[\left\|-\pi_1^T(L_K)f_\rho+\eta\sum_{t=1}^T\pi_{t+1}^T(L_K)\xi(f_t,z_t)\right\|_K^2\right]+2\mathbb{E}_T\left[\left\|\eta\sum_{t=1}^T\pi_{t+1}^T(L_K)\Delta_t\right\|_K^2\right].$$

Following the similar procedure in estimating (14), we also get

$$\mathbb{E}_T\left[\|f_{T+1}-f_\rho\|_K^2\right]\le 2\left\|\pi_1^T(L_K)f_\rho\right\|_K^2+2\eta^2\sum_{t=1}^T\left\|\pi_{t+1}^T(L_K)\right\|^2\mathbb{E}_{t-1}\left[\mathcal{E}(f_t)\right]+2\mathbb{E}_T\left[\left\|\eta\sum_{t=1}^T\pi_{t+1}^T(L_K)\Delta_t\right\|_K^2\right].$$

Noticing that $\left\|\pi_{t+1}^T(L_K)\right\|\le 1$, then the bound (15) is obtained. □

Based on the error bounds of $f_{T+1}-f_\rho$ in Proposition 1, we need to estimate the generalization error $\mathcal{E}(f_t)$.

Lemma 2. *Define* $\{f_t\}_{t=1}^{T+1}$ *by (5). If*

$$0<\eta\le\min\left\{1,\frac{1}{8}\left((1/2e)^{1/2}+1\right)^{-2}(\log(et)+1)^{-1}\right\},\tag{24}$$

then for $t\ge 2$,

$$\mathbb{E}_{t-1}\left[\mathcal{E}(f_t)\right]\le 2\mathcal{E}(f_\rho)+4\|f_\rho\|_\rho^2+64\eta^2\sigma^{-4}(t-1)^2\left(\sup_{1\le k\le t-1}\{\|f_k\|_K,M\}\right)^6.\tag{25}$$

Proof. We shall prove (25) by induction. Obviously, (25) holds for $t=2$. Suppose (25) holds for $t\ge 2$. Applying (14) with $T=t$, then

$$\begin{aligned}\mathbb{E}_t\left[\|f_{t+1}-f_\rho\|_\rho^2\right]&\le 2\left\|\pi_1^t(L_K)f_\rho\right\|_\rho^2+2\eta^2\sum_{k=1}^t\frac{2\left((1/2e)^{1/2}+1\right)^2}{1+\eta(t-k)}\mathbb{E}_{k-1}\left[\mathcal{E}(f_k)\right]+2\mathbb{E}_t\left[\left\|\eta\sum_{k=1}^t\pi_{k+1}^t(L_K)\Delta_k\right\|_\rho^2\right]\\&\le 2\left\|\pi_1^t(L_K)f_\rho\right\|_\rho^2+2\eta^2\sum_{k=1}^t\frac{2\left((1/2e)^{1/2}+1\right)^2}{1+\eta(t-k)}\mathbb{E}_{k-1}\left[\mathcal{E}(f_k)\right]+2\eta^2\left(\sum_{k=1}^t\|\pi_{k+1}^t(L_K)\|\|\Delta_k\|_\infty\right)^2.\end{aligned}\tag{26}$$

Since the Gaussian G is Lipschitz continuous, we have that for each $1 \le k \le T$,

$$\begin{aligned}\|\Delta_k\|_\infty &\le \left\| \left[G(0) - G\left(-\frac{(f_k(x_k)-y_t)^2}{2\sigma^2} \right) \right] [f_k(x_k)-y_k] K_{x_k} \right\|_\infty \\ &\le \left| \left[G(0) - G\left(-\frac{(f_k(x_k)-y_k)^2}{2\sigma^2} \right) \right] [f_t(x_k)-y_k] \right| \|K_{x_k}\|_K \\ &\le \frac{(f_k(x_k)-y_k)^2}{2\sigma^2} |f_k(x_k)-y_k| \le \frac{(\|f_k\|_\infty + M)^3}{2\sigma^2} \le \frac{(\|f_k\|_K + M)^3}{2\sigma^2}\end{aligned}$$

where the last inequality is derived from (3).

Notice that by $0 < \eta \le 1$, there holds $\|\pi_k^t(L_K)\| \le \prod_{l=k}^t \|I - \eta L_K\| \le 1$ for each $1 \le k \le t \le T$. Then the last term on the right-hand side of (26) is bounded as

$$2\eta^2 \left(\sum_{k=1}^t \|\pi_{k+1}^t(L_K)\| \|\Delta_k\|_\infty \right)^2 \le 32\eta^2\sigma^{-4}t^2 \left(\sup_{1\le k\le t} \{\|f_k\|_K, M\} \right)^6. \tag{27}$$

For the first term $2\left\|\pi_1^t(L_K)f_\rho\right\|_\rho^2$, it is easy to get that $2\left\|\pi_1^t(L_K)f_\rho\right\|_\rho^2 \le 2\|f_\rho\|_\rho^2$.

Putting the above estimates into (26) and using the relation (13) with $f = f_{t+1}$, we have

$$\begin{aligned}&\mathbb{E}_t\left[\mathcal{E}(f_{t+1})\right] = \mathbb{E}_t\left[\|f_{t+1}-f_\rho\|_\rho^2\right] + \mathcal{E}(f_\rho) \\ &\le \mathcal{E}(f_\rho) + 2\|f_\rho\|_\rho^2 + 32\eta^2\sigma^{-4}t^2 \left(\sup_{1\le k\le t} \{\|f_k\|_K, M\} \right)^6 + 2\eta^2 \sum_{k=1}^t \frac{2\left((1/2e)^{1/2}+1\right)^2}{1+\eta(t-k)} \mathbb{E}_{k-1}\left[\mathcal{E}(f_k)\right] \\ &\le \mathcal{E}(f_\rho) + 2\|f_\rho\|_\rho^2 + 32\eta^2\sigma^{-4}t^2 \left(\sup_{1\le k\le t} \{\|f_k\|_K, M\} \right)^6 \\ &+ 2\eta^2 \sum_{k=1}^t \frac{2\left((1/2e)^{1/2}+1\right)^2}{1+\eta(t-k)} \left(2\mathcal{E}(f_\rho) + 4\|f_\rho\|_\rho^2 + 64\eta^2\sigma^{-4}(t-1)^2 \left(\sup_{1\le k\le t-1} \{\|f_k\|_K, M\} \right)^6 \right). \end{aligned} \tag{28}$$

By the restriction (24) of η and Lemma A3, we know that

$$2\eta^2 \sum_{k=1}^t \frac{2\left((1/2e)^{1/2}+1\right)^2}{1+\eta(t-k)} \le 4\eta \left((1/2e)^{1/2}+1\right)^2 (\log(et)+1) \le \frac{1}{2}.$$

Plugging it into (28), we have

$$\begin{aligned}&\mathbb{E}_t\left[\mathcal{E}(f_{t+1})\right] \le \mathcal{E}(f_\rho) + 2\|f_\rho\|_\rho^2 + 32\eta^2\sigma^{-4}t^2 \left(\sup_{1\le k\le t} \{\|f_k\|_K, M\} \right)^6 \\ &+ \frac{1}{2}\left(2\mathcal{E}(f_\rho) + 4\|f_\rho\|_\rho^2 + 64\eta^2\sigma^{-4}(t-1)^2 \left(\sup_{1\le k\le t-1} \{\|f_k\|_K, M\} \right)^6 \right) \\ &\le 2\mathcal{E}(f_\rho) + 4\|f_\rho\|_\rho^2 + 64\eta^2\sigma^{-4}t^2 \left(\sup_{1\le k\le t} \{\|f_k\|_K, M\} \right)^6.\end{aligned}$$

Then the proof is completed. □

With these preliminaries in place, we shall prove our main results.

Proof of Theorem 1. We shall prove Theorem 1 by Proposition 1. First, we will use (14) to estimate the error rate for (5) in $L^2_{\rho_X}$-space. For the first term on the right-hand side of (14), applying Lemma A2 with $f = f_\rho$ and $\eta = T^{-\frac{2r}{2r+1}}$, we have that

$$\left\|\pi_1^T(L_K)f_\rho\right\|_\rho^2 \le 4\left((r/e)^r+1\right)^2\|L_K^{-r}f_\rho\|_\rho^2 T^{-\frac{2r}{2r+1}} = 4\left((r/e)^r+1\right)^2\|g\|_\rho^2 T^{-\frac{2r}{2r+1}}.$$

For the second term on the right-hand side of (14), the choice of η and T in Theorem 1 implies that the restriction (24) holds. Then we can put the bound (12) into (25) and get that for $t \ge 2$

$$\begin{aligned}
&\mathbb{E}_{t-1}\left[\mathcal{E}(f_t)\right] \le 2\mathcal{E}(f_\rho) + 4\|f_\rho\|_\rho^2 + 64M^6\sigma^{-4}\eta^5(t-1)^5\\
&\le \left(2\mathcal{E}(f_\rho) + 4\|f_\rho\|_\rho^2 + 64M^6\right)\left(1+\sigma^{-4}\eta^5(t-1)^5\right)\\
&\le \left(2\mathcal{E}(f_\rho) + 4\|f_\rho\|_\rho^2 + 64M^6\right)\left(1+\sigma^{-4}\eta^5T^5\right) := c_{M,\rho}(1+\sigma^{-4}T^{\frac{5}{2r+1}}).
\end{aligned}$$

This together with Lemma A3 yields that

$$\begin{aligned}
\eta^2\sum_{t=1}^T \frac{2\left((1/2e)^{1/2}+1\right)^2}{1+\eta(T-t)}\mathbb{E}_{t-1}\left[\mathcal{E}(f_t)\right] &\le 2\left((1/2e)^{1/2}+1\right)^2 c_{M,\rho}\eta(\log(eT)+1)(1+\sigma^{-4}T^{\frac{5}{2r+1}})\\
&\le 4\left((1/2e)^{1/2}+1\right)^2 c_{M,\rho}\log(T)(T^{-\frac{2r}{2r+1}}+\sigma^{-4}T^{\frac{5-2r}{2r+1}}).
\end{aligned}$$

Finally, we bound the last term on the right-hand side of (14). Notice that

$$\left\|\eta\sum_{t=1}^T \pi_{t+1}^T(L_K)\Delta_t\right\|_\rho \le \eta\sum_{t=1}^T\|\pi_{t+1}^t(L_K)\|\|\Delta_t\|_\infty.$$

Then, using the estimate (27) and the bound (12) of $\{f_t\}$, we have

$$\begin{aligned}
\mathbb{E}_T\left[\left\|\eta\sum_{t=1}^T \pi_{t+1}^T(L_K)\Delta_t\right\|_\rho^2\right] &\le \eta^2\left(\sum_{t=1}^T\|\pi_{t+1}^T(L_K)\|\|\Delta_t\|_\infty\right)^2\\
&\le 16\eta^2\sigma^{-4}t^2\left(\sup_{1\le t\le T}\{\|f_t\|_K, M\}\right)^6 \le 16M^6\sigma^{-4}T^{\frac{5}{2r+1}}.
\end{aligned}$$

Based on the above analysis, the conclusion (10) is obtained by taking

$$C = 8\left((r/e)^r+1\right)^2\|g\|_\rho^2 + 16\left((1/2e)^{1/2}+1\right)^2 c_{M,\rho} + 32M^6.$$

Similarity, we can get the conclusion (11) by taking

$$C' = 8\left(((2r-1)/2e)^{r-\frac{1}{2}}+1\right)^2\|g\|_\rho^2 + 8c_{M,\rho} + 32M^6.$$

□

Author Contributions: B.W. conceived of the presented idea. T.H. developed the theory and performed the computations. All authors discussed the results and contributed to the final manuscript.

Funding: The work described in this paper is partially supported by National Natural Science Foundation of China [Nos. 11671307 and 11571078], Natural Science Foundation of Hubei Province in China [No. 2017CFB523] and the Fundamental Research Funds for the Central Universities, South-Central University for Nationalities [No. CZY18033].

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Useful Lemmas

The following two lemmas are slightly modified forms of Lemma 3, Lemma 7 in [24], respectively.

Lemma A1. *Let $\beta > 0$ and $0 < \eta \leq 1$. Then for any $t \in [l, k]$, there holds*

$$\|\pi_l^k(L_K)L_K^\beta\|^2 \leq \frac{2\left((\beta/e)^\beta + 1\right)^2}{1 + \eta^{2\beta}(k-l+1)^{2\beta}}.$$

Lemma A2. *If $f \in L_K^r(L^2_{\rho_X})$ for some $r > 0$, then*

$$\|\pi_1^T(L_K)f\|_\rho \leq 2\left((r/e)^r + 1\right)\|L_K^{-r}f\|_\rho \eta^{-r}T^{-r}.$$

In addition, if $r > \frac{1}{2}$, then

$$\|\pi_1^T(L_K)f\|_K \leq 2\left(((2r-1)/2e)^{r-\frac{1}{2}} + 1\right)\|L_K^{-r}f\|_\rho \eta^{-r+\frac{1}{2}}T^{-r+\frac{1}{2}}.$$

Lemma A3. *For any $0 < \eta \leq 1$, there holds for $t \geq 2$,*

$$\sum_{k=1}^{t} \frac{1}{1+\eta(t-k)} \leq \eta^{-1}(\log(et) + 1).$$

Proof. By the elementary inequality $\sum_{k=1}^t k^{-1} \leq \log e(t+1)$, we know that for $t \geq 2$,

$$\begin{aligned}\sum_{k=1}^{t} \frac{1}{1+\eta(t-k)} &= \sum_{k=1}^{t-1} \frac{1}{1+\eta(t-k)} + 1 \leq \eta^{-1}\sum_{k=1}^{t-1}(t-k)^{-1} + 1 = \eta^{-1}\sum_{k=1}^{t-1}\frac{1}{k} + 1 \\ &\leq \eta^{-1}\log(et) + 1 \leq \eta^{-1}(\log(et) + 1).\end{aligned}$$

Then the proof is completed. □

References

1. Santamaria, I.; Pokharel, P.P.; Principe, J.C. Generalized correlation function: Definition, properties, and application to blind equalization. *IEEE Trans. Signal Process.* **2006**, *54*, 2187–2197. [CrossRef]
2. Feng, Y.L.; Huang, X.L.; Shi, L.; Yang, Y.N.; Suykens, J.A.K. Learning with the Maximum Correntropy Criterion Induced Losses for Regression. *J. Mach. Learn. Res.* **2015**, *16*, 993–1034.
3. He, R.; Zheng, W.S.; Hu, B.G. Maximum Correntropy Criterion for Robust Face Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 1561–1576. [PubMed]
4. Liu, W.F.; Pokharel, P.P.; Principe, J.C. Correntropy: Properties and applications in non-Gaussian signal processing. *IEEE Trans. Signal Process.* **2007**, *55*, 5286–5298. [CrossRef]
5. Principe, J.C. Renyi's Entropy and Kernel Perspectives. In *Information Theoretic Learning*; Springer: New York, NY, USA, 2010.
6. He, R.; Zheng, W.S.; Hu, B.G.; Kong, X.W. A regularized correntropy framework for robust pattern recognition. *Neural Comput.* **2011**, *23*, 2074–2100. [CrossRef]
7. Bessa, R.J.; Miranda, V.; Gama, J. Entropy and correntropy against minimum square error in offline and online three-day ahead wind power forecasting. *IEEE Trans. Power Syst.* **2009**, *24*, 1657–1666. [CrossRef]
8. He, R.; Hu, B.G.; Zheng, W.S.; Kong, X.W. Robust Principal Component Analysis Based on Maximum Correntropy Criterion. *IEEE Trans. Image Process.* **2011**, *20*, 1485–1494. [PubMed]
9. Chen, B.; Xing, L.; Liang, J.; Zheng, N.; Principe, J.C. Steady-State Mean-Square Error Analysis for Adaptive Filtering under the Maximum Correntropy Criterion. *IEEE Signal Process. Lett.* **2014**, *21*, 880–883.

10. Wu, Z.; Peng, S.; Chen, B.; Zhao, H. Robust Hammerstein Adaptive Filtering under Maximum Correntropy Criterion. *Entropy* **2015**, *17*, 7149–7166. [CrossRef]
11. Liu, W.; Pokharel, P.P.; Principe, J.C. Error Entropy, Correntropy and M-Estimation. In Proceedings of the 16th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing, Arlington, VA, USA, 6–8 September 2006.
12. Syed, M.N.; Pardalos, P.M.; Principe, J.C. Invexity of the minimum error entropy criterion. *IEEE Signal Process. Lett.* **2013**, *20*, 1159–1162. [CrossRef]
13. Syed, M.N.; Pardalos, P.M.; Principe, J.C. On the optimization properties of the correntropic loss function in data analysis. *Optim. Lett.* **2014**, *8*, 823–839. [CrossRef]
14. Marques de Sá, J.P.; Silva, L.M.A.; Santos, J.M.F.; Alexandre, L.A. *Minimum Error Entropy Classification*; Studies in Computational Intelligence; Springer: Berlin/Heidelberg, Germany, 2013.
15. Cucker, F.; Zhou, D.X. *Learning Theory: An Approximation Theory Viewpoint*; Cambridge University Press: Cambridge, UK, 2007.
16. Singh, A.; Pokharel, R.; Principe, J.C. The C-loss function for pattern classification. *Pattern Recognit.* **2014**, *47*, 441–453. [CrossRef]
17. Guo, Z.C.; Hu, T.; Shi, L. Gradient descent for robust kernel-based regression. *Inverse Prob.* **2018**, *34*. [CrossRef]
18. Smale, F.; Zhou, D.X. Learning theory estimates via integral operators and their approximations. *Constr. Approx.* **2007**, *26*, 153–172. [CrossRef]
19. Lu, S.; Pereverzev, S.V. *Regularization Theory for Ill-Posed Problems: Selected Topics*; Walter de Gruyter: Berlin, Germany, 2013.
20. Caponnetto, A.; Vito, E.D. Optimal rates for the regularized least-squares algorithm. *Found. Comput. Math.* **2007**, *7*, 331–368. [CrossRef]
21. Steinwart, I.; Christmann, A. *Support Vector Machines*; Springer: New York, NY, USA, 2008.
22. Bauer, F.; Pereverzev, S.V.; Rosasco, L. On regularization algorithms in learning theory. *J. Complexity* **2007**, *23*, 52–72. [CrossRef]
23. Feng, Y.L.; Fan, J.; Suykens, J.A. A statistical learning approach to modal regression. *arXiv* **2017**, arXiv:1702.05960.
24. Ying, Y.; Pontil, M. Online gradient descent learning algorithms. *Found. Comput. Math.* **2008**, *8*, 561–596. [CrossRef]