entropy

MDPI

Article

Gradient Learning under Tilted Empirical Risk Minimization

Liyuan Liu [1,†], Biqin Song [1,†], Zhibin Pan [1,2], Chuanwu Yang [3], Chi Xiao [4,*] and Weifu Li [1,2,*]

[1] College of Science, Huazhong Agricultural University, Wuhan 430062, China; liulymt@foxmail.com (L.L.); biqin.song@mail.hzau.edu.cn (B.S.); pzbhallow@mail.hzau.edu.cn (Z.P.)
[2] Hubei Key Laboratory of Applied Mathematics, Hubei University, Wuhan 430062, China
[3] School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China; chuanwuyang@hust.edu.cn
[4] Key Laboratory of Biomedical Engineering of Hainan Province, School of Biomedical Engineering, Hainan University, Haikou 570228, China
* Correspondence: xiaochi@hainanu.edu.cn (C.X.); liweifu@mail.hzau.edu.cn (W.L.)
† These authors contributed equally to this work.

Abstract: Gradient Learning (GL), aiming to estimate the gradient of target function, has attracted much attention in variable selection problems due to its mild structure requirements and wide applicability. Despite rapid progress, the majority of the existing GL works are based on the empirical risk minimization (ERM) principle, which may face the degraded performance under complex data environment, e.g., non-Gaussian noise. To alleviate this sensitiveness, we propose a new GL model with the help of the tilted ERM criterion, and establish its theoretical support from the function approximation viewpoint. Specifically, the operator approximation technique plays the crucial role in our analysis. To solve the proposed learning objective, a gradient descent method is proposed, and the convergence analysis is provided. Finally, simulated experimental results validate the effectiveness of our approach when the input variables are correlated.

Keywords: gradient learning; operator approximation; reproducing kernel Hilbert spaces; tilted empirical risk minimization

check for updates

Citation: Liu, L.; Song, B.; Pan, Z.; Yang, C.; Xiao, C.; Li, W. Gradient Learning under Tilted Empirical Risk Minimization. *Entropy* **2022**, *24*, 956. https://doi.org/10.3390/e24070956

Academic Editor: Sotiris Kotsiantis

Received: 16 June 2022
Accepted: 7 July 2022
Published: 9 July 2022

1. Introduction

Data-driven variable selection aims to select informative features related with the response in high-dimensional statistics and plays a critical role in many areas. For example, if the milk production of dairy cows can be predicted by the blood biochemical indexes, then the doctors are eager to know which indexes can drive the milk production because each of them is independently measured with additional burden. Therefore, an explainable and interpretable system to select the effective variables is critical to convince domain experts. Currently, the methodologies on variable selection methods can be roughly divided into three categories including linear models [1–3], nonlinear additive models [4–6], and partial linear models [7–9]. Although achieving promising performance in some applications, these methods mentioned above still suffer from two main limitations. Firstly, the target function of these methods is restricted on the assumption of specific structures. Secondly, these methods cannot revive how the coordinates vary with respect to each other. As an alternative, Mukherjee and Zhou [10] proposed the gradient learning (GL) model, which aims to learn the gradient functions and enjoys the model-free property.

Despite the empirical success [11–13], there are still some limitations of the GL model, such as high computational cost, lacking the sparsity in high-dimensional data and lacking the robustness to complex noises. To this end, several variants of the GL model have been devoted to developing alternatives for individual purposes. For example, Dong and Zhou [14] proposed a stochastic gradient descent algorithm for learning the gradient and demonstrated that the gradient estimated by the algorithm converges to the true gradient. Mukherjee et al. [15] provided an algorithm to reduce dimension on manifolds for high-dimensional data with few observations. They obtained generalization error bounds of the gradient estimates and revealed that the convergence rate depends on the intrinsic dimension of the manifold. Borkar et al. [16] combined ideas from Spall's Simultaneous Perturbation Stochastic Approximation with compressive sensing and proposed to learn the gradient with few function evaluations. Ye et al. [17] originally proposed a sparse GL model to further address the sparsity for high-dimensional variable selection of the estimated sparse gradients. He et al. [18] developed a three-step sparse GL method which allows for efficient computation, admits general predictor effects, and attains desirable asymptotic sparsistency. Following the research direction of robustness, Guinney et al. [19] provided a multi-task model which are efficient and robust for high-dimensional data. In addition, Feng et al. [20] provided a robust gradient learning (RGL) framework by introducing a robust regression loss function. Meanwhile, a simple computational algorithm based on gradient descent was provided, and the convergence of the proposed method is also analyzed.

Despite rapid progress, the GL model and its extensions mentioned above are established under the framework of empirical risk minimization (ERM). While enjoying the nice statistical properties, ERM usually performs poorly in situations where average performance is not an appropriate surrogate for the problem of interest [21]. Recently, a novel framework, named tilted empirical risk minimization (TERM), is proposed to flexibly address the deficiencies in ERM [21]. By using a new loss named t-tilted loss, it has been shown that TERM (1) can increase or decrease the influence of outliers, respectively, to enable fairness or robustness; (2) has variance reduction properties that can benefit generalization; and (3) can be viewed as a smooth approximation to a superquantile method. Considering these strength, we propose to investigate the GL under the framework of TERM. The main contributions of this paper can be summarized as follows:

- New learning objective. We propose to learn the gradient function under the framework of TERM. Specifically, the t-tilted loss is embedded into the GL model. To the best of our knowledge, it may be the first endeavor in this topic.
- Theoretical guarantees. For the new learning objective, we estimate the generalization bound by error decomposition and operator approximation technique, and further provide the theoretical consistency and the convergence rate. To be specific, the convergence rate can recover the result of traditional GL as t tends 0 [10].
- Efficient computation. A gradient descent method is provided to solve the proposed learning objective. By showing the smoothness and strongly convex of the learning objective, the convergence to the optimal solution is proved.

The rest of this paper is organized as follows: Section 2 proposes the GL with t-tilted loss (TGL) and states the main theoretical results on the asymptotic estimation. Section 3 provides the computational algorithm and its convergence analysis. Numerical experiments on synthetic data sets will be implemented in Section 4. Finally, Section 5 closes this paper with some conclusions.

2. Learning Objective

In this section, we introduce TGL and provide the main theoretical results on the asymptotic estimation.

2.1. Gradient Learning with t-Tilted Loss

Let X be a compact subset of $\mathbb{R}^n$ and $Y \in \mathbb{R}$. Assume that ρ is a probability measure on $Z := X \times Y$. It induces the marginal distribution ρ_X on X and conditional distributions $\rho(\cdot|x)$ at $x \in X$. Denote $L^2_{\rho_X}$ as the L^2 space with the metric $\|f\|_\rho = (\int_X |f(x)|^2 d\rho_X)^{1/2}$. In addition, the regression function $f_\rho : X \to Y$ associated with ρ is defined as

$$f_\rho(x) = \int_Y y d\rho(y|x), \quad x \in X.$$

For $x = (x^1, x^2, \ldots, x^n)^\top \in X$, the gradient of f_ρ is the vector of functions (if the partial derivatives exist)

$$\nabla f_\rho = \left(\frac{\partial f_\rho}{\partial x^1}, \frac{\partial f_\rho}{\partial x^2}, \ldots, \frac{\partial f_\rho}{\partial x^n} \right)^\top.$$

The relevance between the l-th coordinate and f_ρ can be evaluated via the norm of its partial derivative $\|\frac{\partial f_\rho}{\partial x^l}\|$, where a large value implies a large change in the function f_ρ with respect to a sensitive change in the l-th coordinate. This fact gives an intuitive motivation for the GL. In terms of Taylor series expansion, the following equation holds:

$$f_\rho(x) \approx f_\rho(\tilde{x}) + \nabla f_\rho(\tilde{x}) \cdot (x - \tilde{x}), \tag{1}$$

for $x \approx \tilde{x}$ and $x, \tilde{x} \in X$. Inspired by (1), we denote the weighted square loss of $\vec{f}$ as

$$V(\vec{f}, z, \tilde{z}) = \omega(x, \tilde{x}) \big(\tilde{y} - y + \vec{f}(\tilde{x})^\top (x - \tilde{x})\big)^2, \quad \vec{f} \in (L^2_{\rho_X})^n, \; z, \tilde{z} \in Z, \tag{2}$$

where the restriction $x \approx \tilde{x}$ will be enforced by weights $\omega(x, \tilde{x})$ given by $\frac{1}{s^{n+2}} e^{-|x - \tilde{x}|^2 / 2s^2}$ with a constant $0 < s \leq 1$, see, e.g., [10,11,19]. Then, the expected risk of $\vec{f}$ can be given by

$$\mathcal{E}(\vec{f}) = \int_Z \int_Z V(\vec{f}, z, \tilde{z}) d\rho(z) d\rho(\tilde{z}). \tag{3}$$

As mentioned in [21], the $\vec{f}$ defined in (3) usually performs poorly in situations where average performance is not an appropriate surrogate. Inspired from [21], for $t \in \mathbb{R}^{\backslash 0}$, we address the deficiencies by introducing the t-tilted loss and define the expected risk of $\vec{f}$ with t-tilted loss as

$$\mathcal{E}(\vec{f}, t) = \frac{1}{t} \log \int_Z \int_Z e^{tV(\vec{f}, z, \tilde{z})} d\rho(z) d\rho(\tilde{z}). \tag{4}$$

Remark 1. *Note that $t \in \mathbb{R}^{\backslash 0}$ is a real-valued hyperparameter, and it can encompass a family of objectives which can address the fairness ($t > 0$) or robustness ($t < 0$) by different choices. In particular, it recovers the expected risk (3) as $t \to 0$.*

On this basis, the GL with t-tilted loss is formulated as the following regularization scheme:

$$\vec{f}_{\lambda,t} = \arg\min_{\vec{f} \in \mathcal{H}_K^n} \{\mathcal{E}(\vec{f}, t) + \lambda \|\vec{f}\|_K^2\}, \tag{5}$$

where $\lambda > 0$ is a regularization parameter. Here, $K : X \times X \to \mathbb{R}$ is a Mercer kernel that is continuous, symmetric, and positive semidefinite [22,23] and $\mathcal{H}_K$ induced by K be an RKHS defined as the closure of the linear span of the set of functions $\{K_x := K(x, \cdot) : x \in X\}$ with the inner product $\langle \cdot, \cdot \rangle_K$ satisfying $\langle K_x, K_{\tilde{x}} \rangle_K = K(x, \tilde{x})$. The reproducing property takes the form $\langle K_x, f \rangle_K = f(x)$, $\forall x \in X$, $\forall f \in \mathcal{H}_K$. Then, we denote $\mathcal{H}_K^n$ as an n-fold RKHS with the inner product

$$\langle \vec{f}, \vec{h} \rangle_K = \sum_{l=1}^n \langle f^l, h^l \rangle_K, \quad \vec{f} = (f^1, f^2, \ldots, f^n)^\top, \; \vec{h} = (h^1, h^2, \ldots, h^n)^\top \in \mathcal{H}_K^n,$$

and norm $\|\vec{f}\|_K^2 = \langle \vec{f}, \vec{f} \rangle_K$.

2.2. Main Results

This subsection states our main theoretical results on the asymptotic estimation of $\|\vec{f}_{\lambda,t} - \nabla f_\rho\|_\rho$ on the space $(L^2_{\rho_X})^n$ with norm $\|\vec{f}\|_\rho = (\sum_{l=1}^n \|f^l\|_\rho^2)^{1/2}$. Before proceeding, we provide some necessary assumptions which have been used extensively in machine learning literature, e.g., [24,25].

Assumption 1. *Supposing that $\nabla f_\rho \in \mathcal{H}_K^n$ and the kernel K is C^3, there exists a constant $c_v > 0$ such that*

$$|f_\rho(x) - f_\rho(\tilde{x}) - \nabla f_\rho(\tilde{x})^\top (x - \tilde{x})| \le c_v |x - \tilde{x}|^2, \quad \forall x, \tilde{x} \in X. \tag{6}$$

Assumption 2. *Assume $|y| \le M$, $|x| \le M_X$ almost surely. Suppose that, for some $\varsigma \in (0, \frac{2}{3})$, $c_l, c_h > 0$, the marginal distribution ρ_X satisfies*

$$\rho_X(\{x \in X : \inf_{\tilde{x} \in \mathbb{R}^n \setminus X} |x - \tilde{x}| \le s\}) \le c_h^2 s^{4\varsigma}, \quad \forall s > 0, \tag{7}$$

and the density $p(z)$ of $d\rho(z)$ exists and satisfies

$$c_l \le p(z) \le c_h, \quad |p(z) - p(\tilde{z})| \le c_h |z - \tilde{z}|^\varsigma, \quad \forall z, \tilde{z} \in Z. \tag{8}$$

Taking the functional derivatives of (5), we know that $\vec{f}_{\lambda,t}$ can be expressed in terms of the following integral operator on the space $(L^2_{\rho_X})^n$.

Definition 1. *Let integral operator $L_{K,s} : (L^2_{\rho_X})^n \to (L^2_{\rho_X})^n$ be defined by*

$$L_{K,s}\vec{f} = \int_Z \int_Z \phi(z, \tilde{z}) \omega(x, \tilde{x}) \big(\vec{f}(\tilde{x})^\top (x - \tilde{x})\big) K_{\tilde{x}}(x - \tilde{x}) d\rho(\tilde{z}) d\rho(z), \tag{9}$$

where

$$\phi(z, \tilde{z}) = \Big(\int_Z \int_Z e^{tV(\vec{f}_{\lambda,t}, u, v)} d\rho(u) d\rho(v) \Big)^{-1} e^{tV(\vec{f}_{\lambda,t}, z, \tilde{z})}.$$

The operator $L_{K,s}$ has its range in $\mathcal{H}_K^n$. It can also be regarded as a positive operator on $\mathcal{H}_K^n$. We shall use the same notion for the operators on these two different domains. Given the definition of integral operator $L_{K,s}$, we can write $\vec{f}_{\lambda,t}$ in the following equation.

Theorem 1. *Given the integral operator $L_{K,s}$, we have the following relationship:*

$$\vec{f}_{\lambda,t} = (L_{K,s} + \lambda I)^{-1} \vec{f}_{\rho,s}, \tag{10}$$

where $\vec{f}_{\rho,s} = \int_Z \int_Z \phi(z, \tilde{z}) \omega(x, \tilde{x}) \big(f_\rho(x) - f_\rho(\tilde{x})\big) K_{\tilde{x}}(x - \tilde{x}) d\rho(\tilde{z}) d\rho(z)$, and I is the identity operator.

Proof of Theorem 1. To solve the scheme (5), we take the functional derivative with respect to $\vec{f}$, apply it to an element $\delta\vec{f}$ of $\mathcal{H}_K^n$ and set it equal to 0. We obtain

$$\int_Z \int_Z \phi(z, \tilde{z}) \omega(x, \tilde{x}) \big(\tilde{y} - y + \vec{f}_{\lambda,t}(\tilde{x})^\top (x - \tilde{x})\big) \delta\vec{f}(\tilde{x})^\top (x - \tilde{x}) d\rho(\tilde{z}) d\rho(z) + \lambda \langle \vec{f}_{\lambda,t}, \delta\vec{f} \rangle_K = 0.$$

Since it holds for any $\delta\vec{f} \in \mathcal{H}_K^n$, it is trivial to obtain

$$\int_Z \int_Z \phi(z, \tilde{z}) \omega(x, \tilde{x}) \big(\tilde{y} - y + \vec{f}_{\lambda,t}(\tilde{x})^\top (x - \tilde{x})\big) K_{\tilde{x}}(x - \tilde{x}) d\rho(\tilde{z}) d\rho(z) + \lambda \vec{f}_{\lambda,t} = 0$$

and

$$\lambda \vec{f}_{\lambda,t} + L_{K,s} \vec{f}_{\lambda,t} = \vec{f}_{\rho,s}.$$

The desired result follows by shifting items. □

On this basis, we propose to bound the error $\|\vec{f}_{\lambda,t} - \nabla f_\rho\|_\rho$ by a functional analysis approach and present the error decomposition as following proposition. The proof is straightforward and omitted for brevity.

Proposition 1. *For the $\vec{f}_{\lambda,t}$ defined in (5), it holds that*

$$\|\vec{f}_{\lambda,t} - \nabla f_\rho\|_\rho \leq \|\vec{f}_{\lambda,t} - \nabla f_\rho + \lambda(L_{K,s} + \lambda I)^{-1}\nabla f_\rho\|_\rho + \|\lambda(L_{K,s} + \lambda I)^{-1}\nabla f_\rho\|_\rho. \tag{11}$$

In the sequel, we focus on bounding $\|\vec{f}_{\lambda,t} - \nabla f_\rho + \lambda(L_{K,s} + \lambda I)^{-1}\nabla f_\rho\|_\rho$ and $\|\lambda(L_{K,s} + \lambda I)^{-1}\nabla f_\rho\|_\rho$, respectively. Before we embark on the proof, we single out a important property regarding $\phi(z, \tilde{z})$ that will be useful in later proofs.

Lemma 1. *Under the Assumptions 1 and 2, there exists B_t and A_t dependent on t satisfying*

$$B_t = e^{-8|t|(M^2 + C_K M_X)} \leq \phi(z, \tilde{z}) \leq A_t = e^{8|t|(M^2 + C_K M_X)}. \tag{12}$$

Proof of Lemma 1. Since the kernel K is C^3 and $\vec{f}_{\lambda,t} \in \mathcal{H}_K^n$, we know from Zhou [26] that $f_{\lambda,t}^l$ is C^1 for each l. There exists a constant C_K satisfying $|\vec{f}_{\lambda,t}(x)|^2 \leq C_K, \forall x \in X$. Hence, using Cauchy inequality, we have

$$\begin{aligned} V(\vec{f}_{\lambda,t}, z, \tilde{z}) &= \omega(\tilde{x}, x)\big(\tilde{y} - y + \vec{f}_{\lambda,t}(\tilde{x})^\mathsf{T}(x - \tilde{x})\big)^2 \\ &\leq 2(4M^2 + |\vec{f}_{\lambda,t}(\tilde{x})|^2|x - \tilde{x}|^2) \\ &\leq 8(M^2 + C_K M_X). \end{aligned}$$

By a direct computation, we obtain

$$e^{-8|t|(M^2 + C_K M_X)} \leq \Big(\int_Z\int_Z e^{tV(\vec{f}_{\lambda,t}, u, v)} d\rho(u) d\rho(v)\Big)^{-1} e^{tV(\vec{f}_{\lambda,t}, z, \tilde{z})} \leq e^{8|t|(M^2 + C_K M_X)}.$$

The desired result follows. □

Denote $\kappa = \sup_{x \in X} K(x, x)$ and the moments of the Gaussian as $J_p = \int_{\mathbb{R}^n} e^{-\frac{|x|^2}{2}}|x|^p dx$, $p = 1, 2, 3, \cdots$, we establish the following Lemma.

Lemma 2. *Under Assumptions 1 and 2, we have*

$$\|\vec{f}_{\lambda,t} - \nabla f_\rho + \lambda(L_{K,s} + \lambda I)^{-1}\nabla f_\rho\|_K \leq 2M\frac{s}{\lambda}\kappa c_v c_h J_3 A_t. \tag{13}$$

Proof of Lemma 2. Taking notice of (10), it follows that

$$\vec{f}_{\lambda,t} - \nabla f_\rho + \lambda(L_{K,s} + \lambda I)^{-1}\nabla f_\rho = (L_{K,s} + \lambda I)^{-1}(\vec{f}_{\rho,s} - L_{K,s}\nabla f_\rho).$$

Then, we have

$$\begin{aligned} \|\vec{f}_{\lambda,t} - \nabla f_\rho + \lambda(L_{K,s} + \lambda I)^{-1}\nabla f_\rho\|_K \leq& \|(L_{K,s} + \lambda I)^{-1}\|_K \|\vec{f}_{\rho,s} - L_{K,s}\nabla f_\rho\|_K \\ \leq& \frac{1}{\lambda}\|\vec{f}_{\rho,s} - L_{K,s}\nabla f_\rho\|_K. \end{aligned}$$

We note that

$$J_p s^{p-2} = \int_{\mathbb{R}^n} \omega(x, \tilde{x})|x - \tilde{x}|^p d\tilde{x} = \int_{\mathbb{R}^n} \frac{1}{s^{n+2}} e^{\frac{-|x - \tilde{x}|^2}{2s^2}}|x - \tilde{x}|^p d\tilde{x}, \quad p = 2, 3, \cdots.$$

From Assumptions 1 and 2, we have

$$\|\vec{f}_{\rho,s} - L_{K,s}\nabla f_\rho\|_K \leq \int_Z\int_Z \omega(x, \tilde{x})|x - \tilde{x}|^3\phi(z, \tilde{z})\|K_{\tilde{x}}\|_K c_v d\rho(z) d\rho(\tilde{z}) \leq 2Ms\kappa c_v c_h J_3 A_t.$$

The desired result follows. □

As for $\|\lambda(L_{K,s}+\lambda I)^{-1}\nabla f_\rho\|_\rho$, the multivariate mean value theorem ensures that there exists $R_t(\tilde{z})=\phi(\tilde{z},\eta_z),\eta_z\in\mathbb{R}^n\times Y$, such that

$$\begin{aligned}&\int_Z\int_{\mathbb{R}^n\times Y}e^{-\frac{|x-\tilde{x}|^2}{2s^2}}\frac{|x-\tilde{x}|^2}{s^{2+n}}\phi(z,\tilde{z})K_{\tilde{x}}\vec{f}(\tilde{x})p(\tilde{z})dz d\rho(\tilde{z})\\ =&\int_Z\int_{\mathbb{R}^n\times Y}e^{-\frac{|x-\tilde{x}|^2}{2s^2}}\frac{|x-\tilde{x}|^2}{s^{2+n}}R_t(\tilde{z})K_{\tilde{x}}\vec{f}(\tilde{x})p(\tilde{z})dz d\rho(\tilde{z}).\end{aligned}\tag{14}$$

From (14), we can define the integral operator associated with the Mercer kernel K which is related to $L_{K,s}$. Using Lemma 16 and Lemma 18 in [10], we establish the following Lemma.

Lemma 3. *Under the Assumption 2, denote* $c_\rho = \left(2MA_t\kappa^2 c_h(2J_{2+\varsigma}+J_4+c_hJ_2)\right)^{\frac{1}{\varsigma}}$ *and* $V_p=\int_Z(p(z))^2R_t(z)dz$. *For any* $0<s\le\min\{c_\rho\lambda^{\frac{1}{\varsigma}},1\}$, *we have*

$$\|\lambda(L_{K,s}+\lambda I)^{-1}\nabla f_\rho\|_\rho\le 2\sqrt{\lambda}(V_pn(2\pi)^{\frac{n}{2}}M)^{-\frac{1}{2}}\|L_K^{-\frac{1}{2}}\nabla f_\rho\|_\rho,\tag{15}$$

where L_K *is a positive operator on* $(L^2_{\rho_X})^n$ *defined by*

$$L_K\vec{f}=\int_Z K_x\vec{f}(x)\frac{p(z)R_t(z)}{V_p}d\rho(z),\vec{f}\in(L^2_\rho)^n.$$

Proof of Lemma 3. To estimate (15), we need to consider the convergence of $L_{K,s}$ as $s\to 0$. Denote the stepping stone

$$\vec{g}=\int_Z\int_Z\omega(x,\tilde{x})(x-\tilde{x})R_t(\tilde{z})K_{\tilde{x}}(x-\tilde{x})^{\mathsf{T}}\vec{f}(\tilde{x})p(\tilde{z})dz d\rho(\tilde{z}),$$

we deduce that

$$\begin{aligned}\|L_{K,s}\vec{f}-2MV_pn(2\pi)^{\frac{n}{2}}L_K\vec{f}\|_K&\le\|L_{K,s}\vec{f}-\vec{g}+\vec{g}-2MV_pn(2\pi)^{\frac{n}{2}}L_K\vec{f}\|_K\\&\le\|L_{K,s}\vec{f}-\vec{g}\|_K+\|\vec{g}-2MV_pn(2\pi)^{\frac{n}{2}}L_K\vec{f}\|_K.\end{aligned}$$

Using the multivariate mean value theorem, there exists $z_\zeta,z_\sigma\in\mathbb{R}^n\times Y$, such that

$$\begin{aligned}\|L_{K,s}\vec{f}-\vec{g}\|_K=&\Bigg\|p(z_\zeta)\int_Z\int_{\mathbb{R}^n\times Y}R_t(\tilde{z})\omega(x,\tilde{x})\left(\vec{f}(\tilde{x})^{\mathsf{T}}(x-\tilde{x})\right)K_{\tilde{x}}(x-\tilde{x})dz d\rho(\tilde{z})\\&-\int_Z\int_Z\omega(x,\tilde{x})(x-\tilde{x})R_t(\tilde{z})K_{\tilde{x}}(x-\tilde{x})^{\mathsf{T}}\vec{f}(\tilde{x})p(\tilde{z})dz d\rho(\tilde{z})\Bigg\|_K\\ \le&\Bigg\|p(z_\zeta)\int_Z\int_{\mathbb{R}^n\times Y}R_t(\tilde{z})\omega(x,\tilde{x})\left(\vec{f}(\tilde{x})^{\mathsf{T}}(x-\tilde{x})\right)K_{\tilde{x}}(x-\tilde{x})dz d\rho(\tilde{z})\\&-\int_Z\int_{\mathbb{R}^n\times Y}R_t(\tilde{z})\omega(x,\tilde{x})\left(\vec{f}(\tilde{x})^{\mathsf{T}}(x-\tilde{x})\right)K_{\tilde{x}}(x-\tilde{x})p(z)dz d\rho(\tilde{z})\Bigg\|_K\\&+\Bigg\|\int_Z\int_Z R_t(\tilde{z})\omega(x,\tilde{x})\left(\vec{f}(\tilde{x})^{\mathsf{T}}(x-\tilde{x})\right)K_{\tilde{x}}(x-\tilde{x})(p(z)-p(\tilde{z}))dz d\rho(\tilde{z})\Bigg\|_K\\ \le&\Bigg\|p(z_\zeta)-p(z_\sigma)\int_Z\int_{\mathbb{R}^n\times Y}R_t(\tilde{z})\omega(x,\tilde{x})\left(\vec{f}(\tilde{x})^{\mathsf{T}}(x-\tilde{x})\right)K_{\tilde{x}}(x-\tilde{x})dz d\rho(\tilde{z})\Bigg\|_K\\&+\Bigg\|\int_Z\int_Z R_t(\tilde{z})\omega(x,\tilde{x})\left(\vec{f}(\tilde{x})^{\mathsf{T}}(x-\tilde{x})\right)K_{\tilde{x}}(x-\tilde{x})(p(z)-p(\tilde{z}))dz d\rho(\tilde{z})\Bigg\|_K\\ \le&4Ms^\varsigma\kappa c_hJ_{2+\varsigma}\|\vec{f}\|_\rho A_t.\end{aligned}$$

Noticing $n(2\pi)^{\frac{n}{2}} = J_2$, we have

$$2V_p n(2\pi)^{\frac{n}{2}} M L_K \vec{f} = \int_Z \int_{\mathbb{R}^n \times Y} \omega(x,\tilde{x}) R_t(\tilde{z}) K_{\tilde{x}} \vec{f}(\tilde{x})(x-\tilde{x})^{\mathsf{T}}(x-\tilde{x}) p(\tilde{z}) d\!z d\rho(\tilde{z}).$$

Then, by (7), we can obtain the following conclusion from Lemma 16 in [10] when $0 \le s \le 1$,

$$\begin{aligned} \|\vec{g} - 2MV_p n(2\pi)^{\frac{n}{2}} L_K \vec{f}\|_K \quad & \le \left\| \int_Z \int_{(\mathbb{R}^n \times Y)\backslash Z} \omega(x,\tilde{x}) R_t(\tilde{z}) K_{\tilde{x}} \vec{f}(\tilde{x}) p(\tilde{z}) |x-\tilde{x}|^2 d\!z d\rho(\tilde{z}) \right\|_K \\ & \le 2Mc_\rho A_t \int_X \int_{\mathbb{R}^n \backslash X} \omega(x,\tilde{x}) K_{\tilde{x}} |\vec{f}(\tilde{x})| |(x-\tilde{x})|^2 d\!x d\rho_X(\tilde{x}) \\ & \le 2Ms^{\varsigma} \kappa c_h (J_4 + c_h J_2) \|\vec{f}\|_\rho A_t. \end{aligned}$$

Combining the above two estimates, there holds for any $0 \le s \le 1$,

$$\|L_{K,s} - 2MV_p n(2\pi)^{\frac{n}{2}} L_K\|_K \le 2MA_t \kappa^2 c_h s^{\varsigma} (2J_{2+\varsigma} + J_4 + c_h J_2). \tag{16}$$

Using Lemma 18 in [10] and (16), the desired result follows. □

Since the measure $d\tilde{\rho} = \int_Y \frac{p(z)R_t(z)}{V_p} d\rho$ is probability one on X, we know that the operator L_K can be used to define the reproducing kernel Hilbert space [22]. Let $L_K^{1/2}$ be the $\frac{1}{2}$-th power of the positive operator L_K on $(L^2_{\tilde{\rho}})^n$ with norm $\|\vec{f}\|_{\tilde{\rho}} = (\sum_{l=1}^n \|f^l\|^2_{\tilde{\rho}})^{1/2}$ having a range in $\mathcal{H}_K^n$, where $\|f^l\|_{\tilde{\rho}} = (\int_X |f^l(x)|^2 d\tilde{\rho})^{1/2}$. Then, $\mathcal{H}_K^n$ is the range of $L_K^{1/2}$:

$$\|\vec{f}\|_{\tilde{\rho}} = \|L_K^{1/2} \vec{f}\|_K, \quad \vec{f} \in (L^2_{\tilde{\rho}})^n. \tag{17}$$

The assumption we shall use is $\|L_K^{-1/2} \nabla f_\rho\|_{\tilde{\rho}} < \infty$. It means that ∇f_ρ lies in the range of $L_K^{1/2}$. Finally, we can give the upper bound of the error $\|\vec{f}_{\lambda,t} - \nabla f_\rho\|_\rho$.

Theorem 2. *Under the Assumptions 1 and 2, choose* $\lambda = m^{-\frac{\tau}{n+2+3\tau}}$ *and* $s = (\kappa c_h)^{\frac{2}{\varsigma}} m^{-\frac{1}{n+2+3\tau}}$. *For any* $m \ge (\kappa c_h)^{2(n+2+3\tau)/\tau}$, *there exists a constant* $C_{\rho,K}$ *such that we have*

$$\|\vec{f}_{\lambda,t} - \nabla f_\rho\|_\rho \le C_{\rho,K} \frac{A_t}{\sqrt{B_t}} \left(\frac{1}{m}\right)^{\frac{\zeta}{2n+4+6\zeta}}. \tag{18}$$

Proof of Theorem 2. Using Cauchy inequality, for $\vec{f} = (f^1, f^2, \ldots, f^n)^{\mathsf{T}} \in (L^2_{\rho_X})^n$, we have

$$\begin{aligned} \int_X \Big(f^l(x)\Big)^2 d\rho_X(x) & \le \Big(\int_Z \Big(f^l(x)\Big)^2 \frac{p(z)R_t(z)}{V_p} d\rho(z)\Big)^{\frac{1}{2}} \Big(\int_Z \Big(f^l(x)\Big)^2 \frac{V_p}{p(z)R_t(z)} d\rho(z)\Big)^{\frac{1}{2}} \\ & \le \sqrt{\frac{V_p}{c_l B_t}} \Big(\int_Z \Big(f^l(x)\Big)^2 \frac{p(z)R_t(z)}{V_p} d\rho(z)\Big)^{\frac{1}{2}} \Big(\int_X \Big(f^l(x)\Big)^2 d\rho_X(x)\Big)^{\frac{1}{2}}. \end{aligned}$$

It means that

$$\Big(\int_X \Big(f^l(x)\Big)^2 d\rho_X(x)\Big)^{\frac{1}{2}} \le \sqrt{\frac{V_p}{c_l B_t}} \Big(\int_Z \Big(f^l(x)\Big)^2 \frac{p(z)R_t(z)}{V_p} d\rho(z)\Big)^{\frac{1}{2}}.$$

According to the definitions of $\|f^l\|_\rho$ and $\|f^l\|_{\tilde{\rho}}$, it is trivial to obtain

$$\|\vec{f}\|_\rho \le \sqrt{\frac{V_p}{c_l B_t}} \|\vec{f}\|_{\tilde{\rho}}. \tag{19}$$

Since $s = (\kappa c_h)^{\frac{2}{\zeta}} \lambda^{\frac{1}{\zeta}}, \lambda = (\frac{1}{m})^{\frac{\zeta}{n+2+3\zeta}}$, we see from the fact $J_2 > 1$ that the restriction $0 < s \leq \min\{c_\rho \lambda^{\frac{1}{\zeta}}, 1\}$ in Lemma 3 is satisfied for $m \geq (\kappa c_h)^{2(n+2+3\tau)/\tau}$. Then, combining Lemma 2, Lemma 3, Equation (17) and inequality (19), we have

$$\begin{aligned}\|\vec{f}_{\lambda,t} - \nabla f_\rho\|_\rho \leq & \|\vec{f}_{\lambda,t} - \nabla f_\rho + \lambda(L_{K,s} + \lambda I)^{-1} \nabla f_\rho\|_\rho + \|\lambda(L_{K,s} + \lambda I)^{-1} \nabla f_\rho\|_\rho \\ \leq & \kappa \|\vec{f}_{\lambda,t} - \nabla f_\rho + \lambda(L_{K,s} + \lambda I)^{-1} \nabla f_\rho\|_K + \|\lambda(L_{K,s} + \lambda I)^{-1} \nabla f_\rho\|_\rho \\ \leq & 2M \frac{s}{\lambda} \kappa^2 c_v c_h J_3 A_t + 2\sqrt{\frac{V_p}{c_l B_t}} \sqrt{\lambda} (M V_p n (2\pi)^{\frac{n}{2}})^{-\frac{1}{2}} \|\nabla f_\rho\|_K \\ \leq & C_{\rho,K} \frac{A_t}{\sqrt{B_t}} \left(\frac{1}{m}\right)^{\frac{\zeta}{2n+4+6\zeta}},\end{aligned}$$

where $C_{\rho,K} = ((2\kappa c_h)^{\frac{2}{\zeta}} + 2) \max\{M\kappa^2 c_v c_h J_3, \sqrt{\frac{V_p}{c_l}} (M V_p n (2\pi)^{\frac{n}{2}})^{-\frac{1}{2}} C_K\}$. □

Remark 2. *Theorem 2 shows when $m \to +\infty$, $\|\vec{f}_{\lambda,t} - \nabla f_\rho\|_\rho \to 0$. This means that the scheme (5) is consistent. In addition, A_t and B_t tend to 1 as t tends 0, we can see that the convergence rate of Scheme (5) is $-\frac{\zeta}{2n+4+6\zeta}$, which is consistent with previous result in [10]. It means that the proposed method can be regarded as an extension of traditional GL.*

3. Computing Algorithm

In this section, we present the GL model under TERM and propose to use the gradient descent algorithm to find the minimizer. Finally, the convergence of the proposed algorithm is also guaranteed.

Given a set of observations $\mathbf{z} = \{z_i = (x_i, y_i)\}_{i=1}^m \in Z^m$ independently drawn according to ρ and assume that the RKHS are rich that the kernel matrix $\mathbf{K} = (K(x_i, x_j))_{i,j=1}^m$ is strictly positive definite [27]. According to the Representer Theorem of kernel methods [28], we assert the approximation of $\vec{f}_{\lambda,t}$ has the following form: $\sum_{i=1}^m c_i K_{x_i}, c_i = (c_i^1, \dots, c_i^n)^\top \in \mathbb{R}^n$. Let $c = (c_1^\top, \dots, c_m^\top)^\top \in \mathbb{R}^{mn}$, the empirical version of (4) is formulated as follows:

$$c_{\mathbf{z},\lambda} := \arg\min_{c \in \mathbb{R}^{mn}} \left\{ \mathcal{E}_{\mathbf{z}}(c,t) + \lambda \| \sum_{i=1}^m c_i K_{x_i} \|_K^2 \right\}, \tag{20}$$

where

$$\mathcal{E}_{\mathbf{z}}(c,t) = \frac{1}{t} \log \left(\frac{1}{m^2} \sum_{i,j=1}^m \exp \left\{ t\omega(x_i, x_j) (y_i - y_j + \sum_{p=1}^m K(x_p, x_i) \hat{x}_{ij} c_p)^2 \right\} \right),$$

with $\hat{x}_{ij} = (x_j - x_i)^\top$. For simplicity, we denote

$$V_{\mathbf{z}}(c, z_i, z_j) = \omega(x_i, x_j) (y_i - y_j + \sum_{p=1}^m K(x_p, x_i) \hat{x}_{ij} c_p)^2$$

and

$$\phi_{\mathbf{z}}(c, z_i, z_j) = \exp \left\{ t \big(V_{\mathbf{z}}(c, z_i, z_j) - \mathcal{E}_{\mathbf{z}}(c,t) \big) \right\}.$$

The gradients of $\mathcal{E}_{\mathbf{z}}(c,t)$ and $\| \sum_{i=1}^m c_i K_{x_i} \|_K^2$ at c are given by

$$\begin{aligned}\nabla_c \mathcal{E}_{\mathbf{z}}(c,t) = & \frac{1}{m^2} \sum_{i,j=1}^m \phi_{\mathbf{z}}(c, z_i, z_j) 2\omega(x_i, x_j) (y_i - y_j + \sum_{p=1}^m K(x_p, x_i) \hat{x}_{ij} c_p) \times \\ & \big(K(x_1, x_i) \hat{x}_{ij}, \dots, K(x_m, x_i) \hat{x}_{ij} \big)^\top,\end{aligned}$$

and

$$\nabla_c \Big\| \sum_{i=1}^{m} c_i K_{x_i} \Big\|_K^2 = 2\sum_{i=1}^{m} \big(K(x_i,x_1)c_i^{\top},\ldots,K(x_i,x_m)c_i^{\top}\big)^{\top}.$$

Correspondingly, scheme (20) can be solved via the following gradient method:

$$c^k = c^{k-1} - \alpha\left(\nabla_c \mathcal{E}_{\mathbf{z}}(c^{k-1},t) + \lambda \nabla_c \Big\| \sum_{i=1}^{m} c_{i,k-1} K_{x_i} \Big\|_K^2\right), \tag{21}$$

where $c^k = (c_{1,k}^{\top},\ldots,c_{m,k}^{\top})^{\top} \in \mathbb{R}^{mn}$ is the calculated solution at iteration k, and α is the step-size. The detailed gradient descent scheme is stated in Algorithm 1. To prove the convergence, we introduce the following lemma derived from Theorem 1 in [29].

Lemma 4. *When $h(c)$ has an γ-Lipschitz continuous gradient (γ-smoothness) and is μ-strongly convex, for the basic unconstrained optimization problem $c^* = \arg\min h(c)$, the gradient descent algorithm $c^k = c^{k-1} - \frac{1}{\gamma}\nabla h(c^{k-1})$ with a step-size of $1/\gamma$ has a global linear convergence rate*

$$h(c^k) - h(c^*) \le \big(1 - \frac{\mu}{\gamma}\big)^k \big(h(c^0) - h(c^*)\big).$$

Algorithm 1 Gradient descent for the Gradient Learning under TERM

```
Input  : data ${(x_i, y_i)}_{i=1}^m$, regularization parameter $\lambda > 0$, initial guess
         $c^0 = (0, 0, \cdots, 0)^{\top}$, $\epsilon > 0$, step-size $\alpha$, $t$, weight matrix $(\omega(x_i, x_j))_{i,j=1}^m$,
         kernel matrix $(K(x_i, x_j))_{i,j=1}^m$.
Output : the learned gradient coefficients $c^k$.
while the stopping criterion $|c^k - c^{k-1}| \le \epsilon$ is not satisfied do
    • Compute the loss for $i, j = 1, \ldots, m$
          $V_{\mathbf{z}}(c^k, z_i, z_j) = \omega(x_i, x_j)\big(y_i - y_j + \sum_{p=1}^m K(x_p, x_i)\hat{x}_{ij}c_p^k\big)^2.$
    • Compute the gradient of the loss for $i, j = 1, \ldots, m$
          $\nabla_c V_{\mathbf{z}}(c^k, z_i, z_j) = 2\omega(x_i, x_j)\big(y_i - y_j + \sum_{p=1}^m K(x_p, x_i)\hat{x}_{ij}c_p^k\big) \times \big(K(x_1, x_i)\hat{x}_{ij}, \ldots, K(x_m, x_i)\hat{x}_{ij}\big)^{\top}.$
    • Compute the gradient of the $\| \sum_{i=1}^m c_i^k K_{x_i} \|_K^2$
          $\nabla_c \| \sum_{i=1}^m c_i^k K_{x_i} \|_K^2 = 2\sum_{i=1}^m \big(K(x_i, x_1){c_i^k}^{\top}, \ldots, K(x_i, x_m){c_i^k}^{\top}\big)^{\top}.$
    • Compute the descent step:
          $c^{k+1} \leftarrow c^k - \alpha\big(\frac{1}{m^2}\sum_{i,j=1}^m \phi_{\mathbf{z}}\big(c^k, z_i, z_j)\nabla_c V_{\mathbf{z}}(c^k, z_i, z_j) + \lambda\nabla_c \| \sum_{i=1}^m c_i^k K_{x_i} \|_K^2\big),$
      and set $k = k + 1$.
end
```

From Lemma 4, we obtain the following conclusion which states that the proposed algorithm converges to (20) by choosing a suitable step size α.

Theorem 3. *Denote $L(c,t) = \mathcal{E}_{\mathbf{z}}(c,t) + \lambda \| \sum_{i=1}^m c_i K_{x_i} \|_K^2$, β_{max}, β_{min} are the maximum and minimum eigenvalues of kernel matrix $\mathbf{K}$, respectively. There exist $\mu \in \mathbb{R}^+$ and $\gamma \in \mathbb{R}^+$ dependent* *on t such that $L(c^k,t)$ is γ-smoothness and μ-strongly convex for any $t > (-n\lambda\beta_{min}/64(M^2 + C_K M_X)M_X^2 m\kappa^4)$. In addition, let the minimizer $c_{\mathbf{z},\lambda}$ defined in scheme (20) and $\{c^k\}$ be the sequence generated by Algorithm 1 with $\alpha = 1/\gamma$, we have*

$$L(c^k,t) - L(c_{\mathbf{z},\lambda},t) \leq \big(1 - \frac{\mu}{\gamma}\big)^k \big(L(c^0,t) - L(c_{\mathbf{z},\lambda},t)\big). \tag{22}$$

Proof of Theorem 3. Note that the strong convexity and the smoothness are related to the Hessian Matrix, and we provide the proof by dividing the Hessian Matrix into three parts:

$$\begin{aligned}\nabla^2_{cc^\top} L(c,t) = & \underbrace{\frac{t}{m^2}\sum_{i,j=1}^{m} \phi_{\mathbf{z}}(c,z_i,z_j)\big(\nabla_c V_{\mathbf{z}}(c,z_i,z_j) - \nabla_c \mathcal{E}_{\mathbf{z}}(c,t)\big)\nabla_c V_{\mathbf{z}}(c,z_i,z_j)^\top}_{E_1} \\ & + \underbrace{\frac{1}{m^2}\sum_{i,j=1}^{m} \phi_{\mathbf{z}}(c,z_i,z_j)\nabla^2_{cc^\top} V_{\mathbf{z}}(c,z_i,z_j)}_{E_2} + \underbrace{\lambda \nabla^2_{cc^\top} \| \sum_{i=1}^{m} c_i K_{x_i} \|_K^2}_{E_3}.\end{aligned} \tag{23}$$

(1) Estimation on E_1: Note that $m^2\nabla_c \mathcal{E}_{\mathbf{z}}(c,t) = \sum_{i,j=1}^m \phi_{\mathbf{z}}(c,z_i,z_j)\nabla_c V_{\mathbf{z}}(c,z_i,z_j)$ and $\sum_{i,j=1}^m \phi_{\mathbf{z}}(c,z_i,z_j) = m^2$. It follows that

$$\sum_{i,j=1}^{m} \phi_{\mathbf{z}}(c,z_i,z_j)(\nabla_c V_{\mathbf{z}}(c,z_i,z_j) - \nabla_c \mathcal{E}_{\mathbf{z}}(c,t))\nabla_c^\top \mathcal{E}_{\mathbf{z}}(c,t) = 0.$$

Hence, we can get the following equation:

$$E_1 = \frac{t}{m^2}\sum_{i,j=1}^{m} \phi_{\mathbf{z}}(c,z_i,z_j)(\nabla_c V_{\mathbf{z}}(c,z_i,z_j) - \nabla_c \mathcal{E}_{\mathbf{z}}(c,t))(\nabla_c V_{\mathbf{z}}(c,z_i,z_j) - \nabla_c \mathcal{E}_{\mathbf{z}}(c,t))^\top. \tag{24}$$

Similar to the proof of Lemma 1, for $i,j = 1,\ldots,m$, it directly follows that

$$\omega(x_i,x_j)\big(y_i - y_j + \sum_{p=1}^{m} K(x_p,x_i)\hat{x}_{ij}c_p\big) \leq 2\sqrt{2(M^2 + C_K M_X)}.$$

Note that, for $i,j = 1,\ldots,m$, $\nabla_c V_{\mathbf{z}}(c,z_i,z_j)\nabla_c V_{\mathbf{z}}(c,z_i,z_j)^\top$ has a sole eigenvalue, it means

$$\nabla_c V_{\mathbf{z}}(c,z_i,z_j)\nabla_c V_{\mathbf{z}}(c,z_i,z_j)^\top \preceq 32(M^2 + C_K M_X)M_X^2 m\kappa^4 I_{mn}, \tag{25}$$

and we have

$$(\nabla_c V_{\mathbf{z}}(c,z_i,z_j) - \nabla_c \mathcal{E}_{\mathbf{z}}(c,t))^\top(\nabla_c V_{\mathbf{z}}(c,z_i,z_j) - \nabla_c \mathcal{E}_{\mathbf{z}}(c,t)) \leq 128(M^2 + C_K M_X)M_X^2 m\kappa^4.$$

It means that the maximum eigenvalue of E_1 is $128t(M^2 + C_K M_X)M_X^2 m\kappa^4$. Then, the following inequations are satisfied

$$\begin{cases} 0_{mn} \preceq E_1 \preceq 128t(M^2 + C_K M_X)M_X^2 m\kappa^4 I_{mn}, & t > 0; \\ 128t(M^2 + C_K M_X)M_X^2 m\kappa^4 I_{mn} \preceq E_1 \preceq 0_{mn}, & t < 0, \end{cases} \tag{26}$$

where 0_{mn} is the $mn \times mn$ matrix with all elements zero.

(2) Estimation on E_2: Note that $\nabla^2_{cc^\top} V_{\mathbf{z}}(c,z_i,z_j)$ can be rewritten as

$$2\omega(x_i,x_j)\big(K(x_1,x_i)\hat{x}_{ij},\ldots,K(x_m,x_i)\hat{x}_{ij}\big)\big(K(x_1,x_i)\hat{x}_{ij},\ldots,K(x_m,x_i)\hat{x}_{ij}\big)^\top.$$

Similar to (25), we have $\nabla^2_{cc^\top} V_{\mathbf{z}}(c, z_i, z_j) \preceq 2\kappa^4 M_x^2 I_{mn}$. It follows

$$0_{mn} \preceq E_2 \preceq 2\kappa^4 M_x^2 I_{mn}. \tag{27}$$

(3) Estimation on E_3: By a direct computation, we have

$$E_3 = 2\lambda \begin{bmatrix} I_n K(x_1, x_1) & I_n K(x_1, x_2) & \dots & I_n K(x_1, x_m) \\ I_n K(x_2, x_1) & I_n K(x_2, x_2) & \dots & I_n K(x_2, x_m) \\ \vdots & \vdots & \ddots & \vdots \\ I_n K(x_m, x_1) & I_n K(x_m, x_2) & \dots & I_n K(x_m, x_m) \end{bmatrix}.$$

Setting $Q = (q_{11}, q_{21}, \dots, q_{n1}, \dots, q_{1m}, q_{2m}, \dots, q_{nm})^\top \in \mathbb{R}^{mn}$, we deduce that

$$Q^\top E_3 Q = 2\lambda \sum_{l=1}^{n} \sum_{i=1}^{m} \sum_{j=1}^{m} K(x_i, x_j) q_{li} q_{lj}.$$

Note that the matrix of quadratic form $\sum_{i=1}^m \sum_{j=1}^m K(x_i, x_j) q_{li} q_{lj}$ is $\mathbf{K}$, then we can obtain

$$2\lambda n \beta_{min} I_{mn} \preceq E_3 \preceq 2\lambda n \beta_{max} I_{mn}. \tag{28}$$

Combining (26), (27) and (28), there exist two constants

$$\mu = \min\{2n\lambda\beta_{min} + 128t(M^2 + C_K M_X) M_X^2 m\kappa^4, 2n\lambda\beta_{min}\}$$

and

$$\gamma = \max\{128t(M^2 + C_K M_X) M_X^2 m\kappa^4 + 2n\lambda\beta_{max}, 2\kappa^4 M_x^2 + 2n\lambda\beta_{max}\}$$

satisfying that

$$\mu I_{mn} \preceq \nabla^2_{cc^\top} L(c, t) \preceq \gamma I_{mn}.$$

Note $\mu > 0$ as $t > -n\lambda\beta_{min}/64(M^2 + C_K M_X) M_X^2 m\kappa^4$, and it means that $L(c, t)$ is γ-smoothness and μ-strongly convex. The desired result follows by Lemma 4. □

4. Simulation Experiments

In this section, we carry out simulation studies with the TGL model ($t < 0$ for robust) on a synthetic data set in the robust variable selection problem. Let the observation data set $\mathbf{z} = \{z_i = (x_i, y_i)\}_{i=1}^m$ with $x_i = (x_i^1, \cdots, x_i^n)$ be generated by the following linear equations:

$$y_i = x_i \cdot w + \epsilon,$$

where ϵ represents the outliers or noises. To be specific, three different noises are used: Cauchy noise with the location parameter $a = 2$ and scale parameter $b = 4$, Chi-square noise with 5 DOF scaled by 0.01 and Gaussian noise $\mathcal{N}(0, 0.3)$. Three different proportions of outliers including 0%, 20%, or 40% are drawn from the Gaussian noise $\mathcal{N}(0, 100)$. Meanwhile, we consider two different cases with $(m, n) = (50, 50), (30, 80)$ corresponding to $m = n$ and $m < n$, respectively. The weighted vector $w = (w^1, \cdots, w^n)$ over different dimensions is constructed as follows:

$$w^l = 2 + 0.5\sin(\tfrac{2\pi l}{10}), \text{ for } l = 1, \dots, N_n \text{ and } 0, \text{ otherwise}.$$

Here, $N_n = 30$ means the number of effective variables. Two situations including uncorrelated variables $x \sim \mathcal{N}(0_n, I_n)$ and correlated variables $x \sim \mathcal{N}(0_n, \Sigma_n)$ are implemented for x, where the covariance matrix Σ_n is given with the (l, p)th entry $0.5^{|l-p|}$.

For the variable selection algorithms, we perform the TGL with $t = 6 \times 10^{-6}, -1, -10$ and compare the traditional GL model [10] and RGL model [20]. For the GL and TGL models, N_n variables are selected by ranking

$$r_l = \frac{\|f^l_{\mathbf{z},\lambda}\|^2_K}{\sum_{p=1}^n \|f^p_{\mathbf{z},\lambda}\|^2_K}, \quad l = 1, \cdots, n.$$

For the RGL model, N_n variables are selected by ranking

$$r_l = \frac{\sum_{i=1}^m (c^l_i)^2}{\sum_{q=1}^n \sum_{i=1}^m (c^q_i)^2}, \quad l = 1, \cdots, n.$$

A model selecting more effective variables $(\leq N_n)$ means a better algorithm.

We repeat experiments for 30 times with the observation set $\mathbf{z}$ generated in each circumstance. The average selected effective variables for different circumstances are reported in Table 1, and the optimal results are marked in bold. Several useful conclusions can be drawn from Table 1.

Table 1. Variable selection results for different circumstances.

	Methods	**Uncorrelated Variables**			**Correlated Variables**		
		0%	**20%**	**40%**	**0%**	**20%**	**40%**
Cauchy noise	GL	28.70	24.27	19.03	20.27	17.53	16.53
$(m, n) = (50, 50)$	RGL	29.00	**26.57**	**27.7**	20.80	15.40	14.16
	$TGL_{t=6\times10^{-6}}$	**29.63**	24.06	18.04	20.67	17.00	16.23
	$TGL_{t=-1}$	29.53	26.07	26.00	**21.07**	**17.6**	**17.13**
	$TGL_{t=-10}$	29.53	24.23	24.03	16.93	15.78	15.67
Chi-square noise	GL	29.40	24.73	20.37	18.40	17.93	16.03
$(m, n) = (50, 50)$	RGL	29.63	**26.90**	**27.60**	19.90	16.10	14.67
	$TGL_{t=6\times10^{-6}}$	**29.84**	24.4	20.90	18.20	17.30	17.20
	$TGL_{t=-1}$	29.14	24.56	25.18	**21.10**	**18.77**	**17.93**
	$TGL_{t=-10}$	25.13	24.10	24.93	20.83	17.10	16.60
Gaussian noise	GL	28.83	25.16	20.13	18.04	16.70	15.93
$(m, n) = (50, 50)$	RGL	**29.40**	**26.70**	**27.20**	19.87	16.40	14.36
	$TGL_{t=6\times10^{-6}}$	29.23	25.23	20.20	18.37	17.76	16.3
	$TGL_{t=-1}$	27.63	26.20	25.90	21.06	**18.40**	**17.90**
	$TGL_{t=-10}$	22.9	25.23	25.06	**21.43**	17.13	16.23
Cauchy noise	GL	29.60	11.33	12.30	11.93	11.57	10.97
$(m, n) = (30, 80)$	RGL	**29.87**	**29.97**	**29.93**	16.50	**16.97**	**15.20**
	$TGL_{t=6\times10^{-6}}$	28.47	10.67	10.49	11.13	11.03	10.93
	$TGL_{t=-1}$	27.06	20.67	11.3	**17.08**	14.4	11.56
	$TGL_{t=-10}$	16.66	16.23	15.12	13.97	13.92	13.54
Chi-square noise	GL	29.83	11.47	12.57	12.57	11.67	11.33
$(m, n) = (30, 80)$	RGL	**29.93**	**29.93**	**29.71**	**19.87**	**18.80**	**17.50**
	$TGL_{t=6\times10^{-6}}$	29.03	11.10	12.90	12.50	10.87	11.43
	$TGL_{t=-1}$	29.37	23.60	23.53	16.08	14.4	11.40
	$TGL_{t=-10}$	28.17	23.33	23.23	13.97	13.92	13.54
Gaussian noise	GL	**29.77**	11.83	12.27	12.92	12.44	11.54
$(m, n) = (30, 80)$	RGL	29.70	**29.93**	**29.93**	**19.73**	13.67	9.83
	$TGL_{t=6\times10^{-6}}$	28.47	10.67	10.49	13.06	9.79	8.73
	$TGL_{t=-1}$	27.06	20.67	11.3	16.08	**14.4**	11.90
	$TGL_{t=-10}$	16.66	16.23	15.12	13.97	13.92	**13.54**

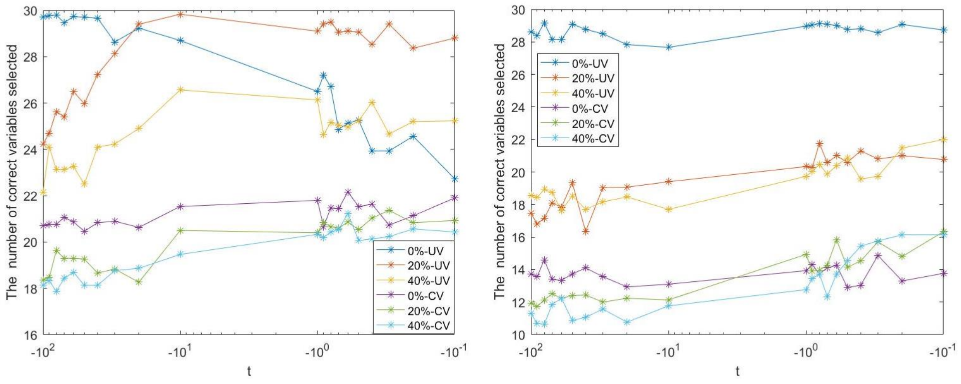
(1) When the input variables are uncorrelated, the three models have similar performance under different noise conditions and can provide satisfactory variable selection results (approaching N_n) without outliers. However, the performance degrades severely for

GL and a little for TGL ($t < 0$ for robust) with the increasing proportions of outliers, especially in case $(m, n) = (30, 80)$. In contrast, RGL can always provide satisfying performance. This is consistent with the previous phenomenon [20].

(2) When the input variables are correlated, the three models also have similar performance under different noise conditions but only can select partial effective variables ranging from $N_n/3$ to $2N_n/3$. In general, they degrade slowly with the increasing proportions of outliers and perform better in case $(m, n) = (50, 50)$ than in $(m, n) = (30, 80)$. Specifically, the TGL model with $t = -1$ gives slightly better selection results than GL and RGL in case $(m, n) = (50, 50)$. It supports the superiority of TGL to some extent.

(3) It is worth noting that the TGL model with $t = 6 \times 10^{-6}$ has similar performance to GL. This phenomenon supports the theoretical conclusion that TGL recovers the GL as $t \to 0$ and the algorithmic effectiveness that the proposed gradient descent method can converge to the minimizer.

(4) Noting that the TGL model with different parameters t has great differences in the variable selection results, we further conduct some simulation studies to investigate the influence. Figure 1 shows the variable selection results of different parameters t ranging from -100 to -0.1. We can see that the satisfying performance can be achieved when the parameter t is near -1. It does not turn out well when $|t|$ is too large. This coincides with our previous discussion that $L(c, t)$ is strongly convex with limited t.



(**a**) $(m, n) = (50, 50)$ (**b**) $(m, n) = (30, 80)$

Figure 1. The influence of different t on the variable selection results.

5. Conclusions

In this paper, we have proposed a new learning objective TGL by embedding the t-tilted loss into the GL model. On the theoretical side, we have established its consistency and provided the convergence rate with the help of error decomposition and operator approximation technique. On the practical side, we have proposed a gradient descent method to solve the learning objective and provided the convergence analysis. Simulated experiments have verified the theoretical conclusion that TGL recovers the GL as $t \to 0$ and the algorithmic effectiveness that the proposed gradient descent method can converge to the minimizer. In addition, they also demonstrated the superiority of TGL when the input variables are correlated. Along the line of the present work, several open problems deserve further research—for example, using the random feature approximation to scale up the kernel methods [30] and learning with data-dependent hypothesis space to achieve a tighter error bound [31]. These problems are under our research.

Author Contributions: All authors have made a great contribution to the work. Methodology, L.L., C.Y., B.S. and C.X.; formal analysis, L.L. and C.X.; investigation, C.Y., Z.P. and C.X.; writing—original draft preparation, L.L., B.S. and W.L.; writing—review and editing, W.L. and C.X.; visualization, C.Y. and Z.P.; supervision, C.X.; project administration, B.S.; funding acquisition, B.S. and W.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Fundamental Research Funds for the Central Universities of China (2662020LXQD002), the Natural Science Foundation of China (12001217), the Key Laboratory of Biomedical Engineering of Hainan Province (Opening Foundation 2022003), and the Hubei Key Laboratory of Applied Mathematics (HBAM 202004).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Schwarz, G. Estimating the Dimension of a Model. *Ann. Stat.* **1978**, *6*, 461–464. [CrossRef]
2. Fan, J.; Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **2001**, *96*, 1348–1360. [CrossRef]
3. Efron, B.; Hastie, T.; Johnstone, I.; Tibshirani, R. Least angle regression. *Ann. Stat.* **2004**, *32*, 407–499. [CrossRef]
4. Chen, H.; Guo, C.; Xiong, H.; Wang, Y. Sparse additive machine with ramp loss. *Anal. Appl.* **2021**, *19*, 509–528. [CrossRef]
5. Chen, H.; Wang, Y.; Zheng, F.; Deng, C.; Huang, H. Sparse Modal Additive Model. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 2373–2387. [CrossRef]
6. Deng, H.; Chen, J.; Song, B.; Pan, Z. Error bound of mode-based additive models. *Entropy* **2021**, *23*, 651. [CrossRef] [PubMed]
7. Engle, R.F.; Granger, C.W.J.; Rice, J.; Weiss, A. Semiparametric Estimates of the Relation Between Weather and Electricity Sales. *J. Am. Stat. Assoc.* **1986**, *81*, 310–320. [CrossRef]
8. Zhang, H.; Cheng, G.; Liu, Y. Linear or Nonlinear? Automatic Structure Discovery for Partially Linear Models. *J. Am. Stat. Assoc.* **2011**, *106*, 1099–1112. [CrossRef]
9. Huang, J.; Wei, F.; Ma, S. Semiparametric Regression Pursuit. *Stat. Sin.* **2012**, *22*, 1403–1426. [CrossRef] [PubMed]
10. Mukherjee, S.; Zhou, D. Learning Coordinate Covariances via Gradients. *J. Mach. Learn. Res.* **2006**, *7*, 519–549.
11. Mukherjee, S.; Wu, Q. Estimation of Gradients and Coordinate Covariation in Classification. *J. Mach. Learn. Res.* **2006**, *7*, 2481–2514.
12. Jia, C.; Wang, H.; Zhou, D. Gradient learning in a classification setting by gradient descent. *J. Approx. Theory* **2009**, *161*, 674–692.
13. He, X.; Lv, S.; Wang, J. Variable selection for classification with derivative-induced regularization. *Stat. Sin.* **2020**, *30*, 2075–2103. [CrossRef]
14. Dong, X.; Zhou, D.X. Learning gradients by a gradient descent algorithm. *J. Math. Anal. Appl.* **2008**, *341*, 1018–1027. [CrossRef]
15. Mukherjee, S.; Wu, Q.; Zhou, D. Learning gradients on manifolds. *Bernoulli* **2010**, *16*, 181–207. [CrossRef]
16. Borkar, V.S.; Dwaracherla, V.R.; Sahasrabudhe, N. Gradient Estimation with Simultaneous Perturbation and Compressive Sensing. *J. Mach. Learn. Res.* **2017**, *18*, 161:1–161:27.
17. Ye, G.B.; Xie, X. Learning sparse gradients for variable selection and dimension reduction. *Mach. Learn.* **2012**, *87*, 303–355. [CrossRef]
18. He, X.; Wang, J.; Lv, S. Efficient kernel-based variable selection with sparsistency. *arXiv* **2018**, arXiv:1802.09246.
19. Guinney, J.; Wu, Q.; Mukherjee, S. Estimating variable structure and dependence in multitask learning via gradients. *Mach. Learn.* **2011**, *83*, 265–287. [CrossRef]
20. Feng, Y.; Yang, Y.; Suykens, J.A.K. Robust Gradient Learning with Applications. *IEEE Trans. Neural Netw. Learn. Syst.* **2016**, *27*, 822–835. [CrossRef]
21. Li, T.; Beirami, A.; Sanjabi, M.; Smith, V. On tilted losses in machine learning: Theory and applications. *arXiv* **2021**, arXiv:2109.06141.
22. Cucker, F.; Smale, S. On the mathematical foundations of learning. *Bull. Am. Math. Soc.* **2002**, *39*, 1–49. [CrossRef]
23. Chen, H.; Wang, Y. Kernel-based sparse regression with the correntropy-induced loss. *Appl. Comput. Harmon. Anal.* **2018**, *44*, 144–164. [CrossRef]
24. Feng, Y.; Fan, J.; Suykens, J.A. A Statistical Learning Approach to Modal Regression. *J. Mach. Learn. Res.* **2020**, *21*, 1–35.
25. Yang, L.; Lv, S.; Wang, J. Model-free variable selection in reproducing kernel hilbert space. *J. Mach. Learn. Res.* **2016**, *17*, 2885–2908.
26. Zhou, D.X. Capacity of reproducing kernel spaces in learning theory. *IEEE Trans. Inf. Theory* **2003**, *49*, 1743–1752. [CrossRef]
27. Belkin, M.; Niyogi, P.; Sindhwani, V. Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples. *J. Mach. Learn. Res.* **2006**, *7*, 2399–2434.
28. Schölkopf, B.; Smola, A.J.; Bach, F. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*; MIT Press: Cambridge, MA, USA, 2002.
29. Karimi, H.; Nutini, J.; Schmidt, M. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Riva del Garda, Italy, 19–23 September 2016; Springer: Cham, Switzerland, 2016; pp. 795–811.

30. Dai, B.; Xie, B.; He, N.; Liang, Y.; Raj, A.; Balcan, M.F.F.; Song, L. Scalable Kernel Methods via Doubly Stochastic Gradients. In *Advances in Neural Information Processing Systems;* Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2014; Volume 27.
31. Wang, Y.; Chen, H.; Song, B.; Li, H. Regularized modal regression with data-dependent hypothesis spaces. *Int. J. Wavelets Multiresolution Inf. Process.* **2019**, *17*, 1950047. [CrossRef]