Information Sciences 406–407 (2017) 57–70

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Online pairwise learning algorithms with convex loss functions☆

Junhong Lin, Yunwen Lei*, Bo Zhang, Ding-Xuan Zhou

Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong, China

ARTICLE INFO

Article history:
Received 18 December 2015
Revised 25 March 2017
Accepted 10 April 2017
Available online 13 April 2017

Keywords:
Learning theory
Online learning
Reproducing Kernel Hilbert Space
Pairwise learning

ABSTRACT

Online pairwise learning algorithms with general convex loss functions without regularization in a Reproducing Kernel Hilbert Space (RKHS) are investigated. Under mild conditions on loss functions and the RKHS, upper bounds for the expected excess generalization error are derived in terms of the approximation error when the stepsize sequence decays polynomially. In particular, for Lipschitz loss functions such as the hinge loss, the logistic loss and the absolute-value loss, the bounds can be of order $O(T^{-\frac{1}{3}}\log T)$ after T iterations, while for the least squares loss, the bounds can be of order $O(T^{-\frac{1}{4}}\log T)$. In comparison with previous works for these algorithms, a broader family of convex loss functions is studied here, and refined upper bounds are obtained.

1. Introduction

Many classical learning tasks can be modeled as learning a good estimator or predictor $f: X \to Y$ based on an observed dataset $\{(x_t, y_t)\}_{t=1}^T$ of input-output samples from $X \times Y$, where X is an input space and $Y \subseteq \mathbb{R}$ an output space. Learning algorithms are often implemented by minimizing $\frac{1}{T}\sum_{t=1}^T V(y_t, f(x_t))$ over a hypothesis space of functions in various ways including regularization schemes [26]. Here $V: \mathbb{R}^2 \to \mathbb{R}_+$ is a *loss function* used for measuring the performance of a predictor f. It induces a *local error* $V(y, f(x))$ over an input-output sample $(x, y) \in X \times Y$. For non-parametric regression with $Y = \mathbb{R}$, the least squares loss function $V(y, a) = (y-a)^2$ is often used and, for an input $x \in X$ and an estimator f, the induced local error $V(y, f(x)) = (y - f(x))^2$ measures how well the predicted value $f(x)$ approximates the output value $y \in \mathbb{R}$. For binary classification with $Y = \{1, -1\}$ consisting of the two labels corresponding to the two classes, the misclassification loss function $V(y, a) = \chi_{(-\infty,0)}(ya)$ generated by the characteristic function of the interval $(-\infty, 0)$ is a natural choice, and the induced local error $V(y, f(x)) = \chi_{(-\infty,0)}(yf(x))$ over a sample $(x, y) \in X \times Y$ equals 1 when the sign of $f(x)$ and y correspond to the two different labels in Y (that is, $yf(x) < 0$), while $V(y, f(x)) = 0$ when they correspond to a same label with $yf(x) \geq 0$. But the characteristic function $\chi_{(-\infty,0)}$ is not convex, and the optimization problems involved in the related learning algorithms are not convex. For designing efficient learning algorithms, $\chi_{(-\infty,0)}$ may be replaced by a convex function $\phi: \mathbb{R} \to \mathbb{R}_+$, leading to convex optimization problems involving the local error $V(y, f(x)) = \phi(yf(x))$. One choice of ϕ is the hinge loss $\phi_h(v) = \max\{1 - v, 0\}$ used in the classical support vector machines for solving binary classification problems

☆ The work described in this paper is supported partially by the Research Grants Council of Hong Kong [Project No. CityU 104113]. The corresponding author is Yunwen Lei. Junhong Lin is now within the LCSL, MIT & Istituto Italiano di Tecnologia, Cambridge, MA 02139, USA.

* Corresponding author.
E-mail addresses: jhlin5@hotmail.com (J. Lin), yunwelei@cityu.edu.hk, yunwen.lei@hotmail.com (Y. Lei), bozhang37-c@my.cityu.edu.hk (B. Zhang), mazhou@cityu.edu.hk (D.-X. Zhou).

http://dx.doi.org/10.1016/j.ins.2017.04.022
0020-0255/

[26]. The above learning framework has been well developed within the last two decades [9,26]. It might be categorized as "pointwise learning", as the local error $V(y, f(x))$ takes only one sample point $(x, y) \in X \times Y$ into account.

In this paper, we study another important family of learning problems categorized as "pairwise learning" in which the local error takes a pair $\{(x, y), (x', y')\}$ of two samples from $X \times Y$ into account. Its learning tasks include ranking [1,8], similarity and metric learning [5,28], AUC maximization [34], and gradient learning [19,20,30]. The goal of *pairwise learning* is to learn a good predictor $f : X^2 \to \mathbb{R}$ predicting a value $f(x, x') \in \mathbb{R}$ for each input pair $(x, x') \in X^2$ according to various tasks. To measure the learning performance of a predictor f, we use a loss function $V : \mathbb{R}^2 \to \mathbb{R}_+$ to induce the local error $V(r(y, y'), f(x, x'))$ over two input-output samples $(x, y), (x', y') \in X \times Y$, where $r : Y \times Y \to \mathbb{R}$ is a function, called *reducing function*, chosen according to the learning task. The reducing function r is an essential concept making pairwise learning different from pointwise learning. We demonstrate how to choose the reducing function r by the following examples.

1. For the least squares regression with $Y = \mathbb{R}$ and $V(y, a) = (y - a)^2$, a sample (x, y) is drawn from a probability measure and the expected value of $y \in \mathbb{R}$ given $x \in X$ equals $f^*(x)$, the value of the conditional mean (regression) function f^* at x. So $y - y' = f^*(x) - f^*(x')$ in expectation and we choose the reducing function $r : Y \times Y \to \mathbb{R}$ as the output value difference $r(y, y') = y - y'$. Then the local error $V(r(y, y'), f(x, x')) = \big(y - y' - f(x, x')\big)^2$ measures how well the predicted value $f(x, x')$ for an input pair (x, x') approximates $f^*(x) - f^*(x')$ via the output value difference $y - y'$.
2. For metric learning in binary classification with $Y = \{1, -1\}$, we aim to learn a metric f such that a pair (x, x') of inputs (objects) from the same class $(y = y')$ are close to each other while a pair from different classes $(y \neq y')$ have a large distance $f(x, x')$. A typical choice of the reducing function $r : Y \times Y \to \mathbb{R}$ is given by $r(y, y') = 1$ if $y = y'$ and -1 otherwise [5]. The local error induced by the convex loss function $V(y, a) = \max\{0, 1 + ya\}$ is $V(r(y, y'), f(x, x')) = \max\{0, 1 + r(y, y')f(x, x')\}$. It gives a large local error $1 + f(x, x')$ if the distance $f(x, x')$ between the input pair (x, x') from the same class $(y = y')$ is large.
3. For ranking in a regression framework with $Y = \mathbb{R}$, we aim to learn a good ordering f between objects (inputs) based on their observed features such that $f(x, x') < 0$ if x is preferred over x' meaning that the ranking labels satisfy $y < y'$. A typical choice [21] of the reducing function $r : Y \times Y \to \mathbb{R}$ is given by $r(y, y') = \operatorname{sign}(y - y')$, the sign of $y - y'$. Then the local error induced by the hinge loss ϕ_h is $V(r(y, y'), f(x, x')) = \phi_h(\operatorname{sign}(y - y')f(x, x'))$.

Batch learning and online learning are two kinds of learning algorithms. The former uses an entire dataset to perform learning tasks, while the latter uses the dataset in a stream way. For batch learning algorithms in the pairwise learning framework, theoretical error and robustness analysis have been carried out in [1,5,7,8,21]. One challenge in conducting analysis in pairwise learning is that pairs of training samples are not independent. For example, given the independently and identically distributed (i.i.d.) samples $\{z_t = (x_t, y_t)\}_{t=1}^T$, a batch algorithm for pairwise learning possibly involves a target function

$$\frac{T(T-1)}{2} \sum_{1 \le i < j \le T} V(r(y_i, y_j), f(x_i, x_j)) + \operatorname{pen}(f, \lambda), \tag{1.1}$$

where $\operatorname{pen}(f, \lambda) \ge 0$ is some regularization term used to avoid overfitting. In this case, local errors $V(r(y_i, y_j), f(x_i, x_j))$ and $V(r(y_i, y_{j'}), f(x_i, x_{j'}))$ are indeed dependent. Thus, standard techniques for classification and regression cannot be directly applied, and new tools such as U-statistics [8] or algorithmic stability [1] are necessary for the analysis.

In spite of their good theoretical guarantees, batch algorithms for pairwise learning may be difficult to implement for large-scale learning problems in practice. Indeed, even for the simpler case of pointwise learning, the computational complexity of batch algorithms with many loss functions is of order $O(T^3)$. Moreover, batch algorithms for pairwise learning suffer from extra computational burden of optimizing an objective defined over $O(T^2)$ possible sample pairs.

In practical applications, online learning may be more favorable, due to its scalability to large datasets and applicability to situations where the samples are collected sequentially. Theoretical results for online learning in classification and regression have been well developed (see for example [2,6,18,22,24,31] and references therein), but there is relatively little work for online learning in pairwise learning. Recent research of this direction can be found in [15,27,32]. In particular, online pairwise learning in a linear space was investigated in [15,27], and convergence results were established for the average of the iterates under the assumption of uniform boundedness of the loss function, with a rate $O(1/\sqrt{T})$ in the general convex case, or a rate $O(1/T)$ in the strongly convex case. Online pairwise learning in a RKHS with the least squares loss was studied in [32] where bounds in probability were derived for the excess generalization error.

In this paper, we improve the analysis of online pairwise learning (see Algorithm 1 in the next section) in a RKHS with general convex loss functions. Our main purpose is to develop convergence results for such learning algorithms using polynomially decaying stepsize sequences. Unlike [15,27], we do not assume that the iterates are restricted to a bounded domain or the loss function is strongly convex. In particular, we will provide bounds for the expected excess generalization error, under a mild condition on approximation errors and an increment condition on the loss. For Lipschitz loss functions such as the hinge loss and the logistic loss, our bounds can be of order $O(T^{-\frac{1}{3}} \log T)$, while for the least squares loss, our bounds can be of order $O(T^{-\frac{1}{4}} \log T)$. For general convex loss functions, previous error analysis techniques dealing with the least squares loss in [32], which rely on integral operators, do not apply and are replaced by tools from convex analysis and Rademacher complexity. The key to our proof is an error decomposition, which enables us to study the weighted excess generalization error in terms of the weighted average and the moving weighted average. The novelty lies in an estimate