Neurocomputing 000 (2017) 1–10

Contents lists available at ScienceDirect

Neurocomputing

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

Refined bounds for online pairwise learning algorithms☆

Xiaming Chen[a], Yunwen Lei[b,*]

[a] Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong, China
[b] Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China

ARTICLE INFO

Article history:
Received 2 April 2017
Revised 7 September 2017
Accepted 17 November 2017
Available online xxx

Communicated by Dr Yiming Ying

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

ABSTRACT

Motivated by the recent growing interest in pairwise learning problems, we study the generalization performance of Online Pairwise lEaRning Algorithm (OPERA) in a reproducing kernel Hilbert space (RKHS) without an explicit regularization. The convergence rates established in this paper can be arbitrarily closed to $O(T^{-\frac{1}{2}})$ within T iterations and largely improve the existing convergence rates for OPERA. Our novel analysis is conducted by showing an almost boundedness of the iterates encountered in the learning process with high probability after establishing an induction lemma on refining the RKHS norm estimate of the iterates.

1. Introduction

Machine learning often refers to a process of inferring the relationship underlying some examples $\{z_t = (x_t, y_t)\}_{t=1}^T$ drawn from a probability measure ρ defined over $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$ with a compact input space $\mathcal{X} \subset \mathbb{R}^d$ and an output space $\mathcal{Y} \subset \mathbb{R}$. For many machine learning problems, the relationship can be expressed by a function from $\mathcal{X}$ to $\mathcal{Y}$ and the quality of a model $f : \mathcal{X} \to \mathbb{R}$ can be quantified by a local error $V(y, f(x))$ induced by a function $V : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$. For example, a binary classification problem aims to build a classifier f from $\mathcal{X}$ to $\mathcal{Y} = \{\pm 1\}$ and typical choices of V include the zero-one loss $V(y, a) = \mathbf{1}_{\{ya<0\}}$ and its surrogates $V(y, a) = \phi(ya)$ with a convex nonnegative function ϕ. Here $\mathbf{1}_{\{\cdot\}}$ is the indicator function. Regression problems aim to estimate the output value y by a function $f : \mathcal{X} \to \mathcal{Y}$ and the quality of f at (x, y) can be measured by an increasing function of the discrepancy between $f(x)$ and y. We refer to these learning problems as "pointwise learning" since the local error $V(y, f(x))$ only involves a single example $z = (x, y) \in \mathcal{Z}$.

Recently, there is growing interest in another important class of learning problems which we refer to as "pairwise learning" problems. For pairwise learning problems, the associated local error depends on a pair of examples $z = (x, y), \tilde{z} = (\tilde{x}, \tilde{y})$ and we wish to build an estimator $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. To be precise, the local error of f at $(z, \tilde{z})$ can be typically quantified by $V(r(y, \tilde{y}), f(x, \tilde{x}))$, where $r : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is a reducing function whose specific realization depends on the application domain. Many machine learning problems can be incorporated into the framework of pairwise learning by choosing appropriate reducing functions r and loss functions V, including ranking [6,23], similarity and metric learning [2,4,11], AUC maximization [32], gradient learning [22] and learning under minimum error entropy criterion [8,13,14]. For example, the problem of ranking aims to learn a good order $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ between z and $\tilde{z}$ such that we predict $y \le \tilde{y}$ if $f(x, \tilde{x}) \le 0$. The local error of f at $(z, \tilde{z})$ can be naturally quantified by $\mathbf{1}_{\{(y-\tilde{y})f(x,\tilde{x})<0\}}$, which is of the form $V(r(y, \tilde{y}), f(x, \tilde{x}))$ by taking $V(r, a) = \mathbf{1}_{\{ra<0\}}$ and $r(y, \tilde{y}) = y - \tilde{y}$. A convex surrogate of this $0-1$ loss is the so-called *least squares ranking loss* $V^{\mathrm{sq}}(r(y, \tilde{y}), f(x, \tilde{x})) := \big(|y - \tilde{y}| - \mathrm{sgn}\big(y - \tilde{y}\big)\big(f(x) - f(\tilde{x})\big)\big)^2$ studied in [1,3,34,36], where $\mathrm{sgn}(a)$ denotes the sign of $a \in \mathbb{R}$.

Training examples in some machine learning problems become available in a sequential manner. Online learning provides an efficient method to handle these learning tasks by iteratively updating the model f_t upon the arrival of an example $z_t = (x_t, y_t), t \in \mathbb{N}$. As the counterpart of batch learning which handles training samples at the same time, online learning enjoys an additional advantage in computational efficiency. This computational advantage is especially appealing in the pairwise learning context since the objective function for batch pairwise learning over T examples involves $O(T^2)$ terms. Motivated by these observations, the generalization analysis of online pairwise learning has recently received consid-

☆ The work described in this paper is partially supported by the Research Grants Council of Hong Kong [Project No. CityU 11303915] and by National Natural Science Foundation of China under Grants 11461161006, 11471292 and 11771012.

* Corresponding author.
E-mail addresses: xiamichen2-c@my.cityu.edu.hk (X. Chen), leiyw@sustc.edu.cn (Y. Lei).

https://doi.org/10.1016/j.neucom.2017.11.049
0925-2312/

erable attention [3,12,15,18,29,34]. In particular, an error bound of the order $O(T^{-\frac{1}{3}}\log T)$ was established in [34] for an Online Pairwise lEaRning Algorithm (OPERA) in a reproducing kernel Hilbert space (RKHS) after T iterations. Unlike existing work requiring the iterates to be restricted to a bounded domain or the loss function to be strongly convex [15,29], OPERA is implemented in an RKHS, without constraints on the iterates, to minimize a non-strongly convex objective function. This paper aims to refine these theoretical results. To be precise, we give an error bound of the order arbitrarily closed to $O(T^{-\frac{1}{2}})$ for OPERA in [34]. This improvement is achieved by establishing the "boundedness" of iterates encountered in the learning process, which was shown to grow polynomially with respect to the iteration number in [34]. Our novel analysis is based on an induction lemma showing that $\{f_t\}_{t=1}^T$ would belong to a ball of radius $O(t^{\alpha-\nu})$ with high probability if one can show that it belongs to a ball of radius $O(t^\alpha)$, where ν is a positive constant depending only on the step size sequence. The "boundedness" of iterates can then be derived by applying repeatedly this induction lemma.

2. Main results

Throughout this paper, we assume that the training examples $\{z_t = (x_t, y_t)\}_{t\in\mathbb{N}}$ are independently drawn from ρ in an online manner. We consider online pairwise learning in an RKHS defined on the product space $\mathcal{X}^2 = \mathcal{X}\times\mathcal{X}$. Let $K : \mathcal{X}^2\times\mathcal{X}^2 \to \mathbb{R}$ be a Mercer kernel, i.e., a continuous, symmetric and positive semidefinite kernel. The associated RKHS $\mathcal{H}_K$ is defined as the completion of the linear combination of functions $\{K_{(x,\tilde{x})}(\cdot) := K((x,\tilde{x}),(\cdot,\cdot)) : (x,\tilde{x}) \in \mathcal{X}^2\}$ under an inner product satisfying the following reproducing property

$$\left\langle K_{(x,\tilde{x})}, g\right\rangle = g(x,\tilde{x}), \quad \forall x,\tilde{x}\in\mathcal{X} \text{ and } g\in\mathcal{H}_K.$$

Denote $\kappa := \sup_{x,\tilde{x}\in\mathcal{X}}\sqrt{K((x,\tilde{x}),(x,\tilde{x}))}$, and throughout the paper we assume that $|y|\le M$ almost surely for some $M>0$.

We study a specific pairwise learning problem with the local error taking the least squares form $V(r(y,\tilde{y}), f(x,\tilde{x})) = (f(x,\tilde{x}) - y + \tilde{y})^2$, which coincides with the *least squares ranking loss* V^{sq} with applications to ranking problems [1] if $y\neq\tilde{y}$. These two loss functions would be identical almost surely if the set $\{(z,\tilde{z})\in\mathcal{Z}\times\mathcal{Z} : y=\tilde{y}\}$ is of measure zero under the probability measure $\rho\times\rho$. For this specific pairwise learning problem studied in [3,34,36], an efficient OPERA starting with $f_1 = f_2 = 0$ was introduced in [34] as follows

$$f_{t+1} = f_t - \frac{\gamma_t}{t-1}\sum_{j=1}^{t-1}\big(f_t(x_t,x_j) - y_t + y_j\big)K_{(x_t,x_j)}, \quad t = 2,3,\ldots. \tag{2.1}$$

Here $\{\gamma_t > 0 : t\in\mathbb{N}\}$ is usually referred to as the sequence of step sizes. This paper only considers polynomially decaying step sizes of the form $\gamma_t = \frac{t^{-\theta}}{\mu}$ with $\theta\in\left(\frac{1}{2},1\right)$ and $\mu\ge\kappa^2$, which implies $\gamma_t\kappa^2\le 1$ for all $t\in\mathbb{N}$.

The generalization error of a function $f : \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ is defined by

$$\mathcal{E}(f) = \iint_{\mathcal{Z}\times\mathcal{Z}}(f(x,\tilde{x}) - y + \tilde{y})^2 d\rho(x,y)d\rho(\tilde{x},\tilde{y}).$$

Define the *pairwise regression function* $\widetilde{f}_\rho$ as the difference between two standard regression functions

$$\widetilde{f}_\rho(x,\tilde{x}) = f_\rho(x) - f_\rho(\tilde{x}), \quad f_\rho(x) = \int_{\mathcal{X}} y d\rho(y|x).$$

We denote $L^2_\rho(\mathcal{X}^2)$ the space of square integrable functions on the domain $\mathcal{X}\times\mathcal{X}$, i.e.,

$$L^2_\rho(\mathcal{X}^2) = \Big\{f : \mathcal{X}\times\mathcal{X}\to\mathbb{R} : \|f\|_\rho = \Big(\iint_{\mathcal{X}\times\mathcal{X}}|f(x,\tilde{x})|^2 d\rho_{\mathcal{X}}(x)d\rho_{\mathcal{X}}(\tilde{x})\Big)^{\frac{1}{2}} < \infty\Big\},$$

where $\rho_{\mathcal{X}}$ is the marginal distribution of ρ over $\mathcal{X}$. Analogous to the standard least square regression problem [7], the following identity holds for any $f\in L^2_\rho(\mathcal{X}^2)$ [13,34]

$$\mathcal{E}(f) - \mathcal{E}\big(\widetilde{f}_\rho\big) = \|f - \widetilde{f}_\rho\|^2_\rho, \tag{2.2}$$

from which we also see clearly that $\widetilde{f}_\rho$ minimizes the functional $\mathcal{E}(\cdot)$ among all measurable functions.

Our generalization analysis requires a standard regularity assumption on the pairwise regression function in terms of the integral operator $L_K : L^2_\rho(\mathcal{X}^2)\to L^2_\rho(\mathcal{X}^2)$ defined by

$$L_K f = \iint_{\mathcal{X}\times\mathcal{X}} f(x,\tilde{x})K_{(x,\tilde{x})}d\rho_{\mathcal{X}}(x)d\rho_{\mathcal{X}}(\tilde{x}).$$

The integral operator L_K is compact and positive since K is a Mercer kernel, from which the fractional power L_K^β $(\beta>0)$ can be well defined by $L_K^\beta f = \sum_{j=1}^\infty \lambda_j^\beta\alpha_j\psi_j$ for $f = \sum_{j=1}^\infty\alpha_j\psi_j\in L^2_\rho(\mathcal{X}^2)$, where $\{\lambda_j\}_{j\in\mathbb{N}}$ are the positive eigenvalues and $\{\psi_j\}_{j\in\mathbb{N}}$ are the associated orthonormal eigenfunctions. Furthermore, we know $L_K^{\frac{1}{2}}(L^2_\rho(\mathcal{X}^2)) = \mathcal{H}_K$ with the norm satisfying

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

Here $L_K^\beta(L^2_\rho(\mathcal{X}^2)) = \{L_K^\beta f : f\in L^2_\rho(\mathcal{X}^2)\}$ for any $\beta>0$. Throughout this paper, we always assume the existence of an $f^*\in\mathcal{H}_K$ such that $f^* = \arg\min_{f\in\mathcal{H}_K}\mathcal{E}(f)$, which implies that

$$\frac{1}{2}\nabla\mathcal{E}(f^*) = \iint_{\mathcal{Z}\times\mathcal{Z}}\big(f^*(x,\tilde{x}) - y + \tilde{y}\big)K_{(x,\tilde{x})}d\rho(x,y)d\rho(\tilde{x},\tilde{y}) = L_K\big(f^* - \widetilde{f}_\rho\big) = 0, \tag{2.3}$$

where ∇ is the gradient operator. This assumption, weaker than assuming $\widetilde{f}_\rho\in\mathcal{H}_K$, was also imposed in the literature, see, e.g., [5,35]. The following theorem to be proved in Section 4.3 establishes learning rates for the last iterate f_{T+1} generated by OPERA. For any $a\in\mathbb{R}$, let $\lfloor a\rfloor$ and $\lceil a\rceil$ denote the largest integer not larger than a and the smallest integer not smaller than a, respectively.

Theorem 1. *Let $\{f_t : t\in\mathbb{N}\}$ be given by OPERA. Suppose $\widetilde{f}_\rho\in L_K^\beta(L^2_\rho(\mathcal{X}^2))$ with some $\beta>0$. For any $0<\delta<1$, with probability $1-\delta$, there holds*

$$\mathcal{E}(f_{T+1}) - \mathcal{E}\big(\widetilde{f}_\rho\big) \le \bar{C}T^{-\min(2\beta,1)(1-\theta)}\left(\log\frac{T^{\left\lceil\frac{1-\theta}{2\theta-1}\right\rceil}}{\delta}\right)^{\left\lceil\frac{2-2\theta}{2\theta-1}\right\rceil}\log^2\frac{T}{\delta}\log^3(eT),$$

where $\bar{C}$ is a constant independent of T and δ.

Our proof of Theorem 1 is based on the following key observation to be proved in Section 4.2 on the boundedness of $\{f_t\}_{t=1}^T$ (up to factors $\log T$) with high probability.

Proposition 2. *Let $\{f_t : t\in\mathbb{N}\}$ be given by OPERA and assume $\gamma_t\kappa^2\le 1$ for any $t\in\mathbb{N}$. Then for any $\delta\in(0,1)$, with probability at least $1-\delta$, the following inequality holds for all $t = 2,\ldots,T$*

$$\|f_t\|_K \le \widetilde{C}\left[\log\frac{8T^{\left\lceil\frac{1-\theta}{2\theta-1}\right\rceil}}{\delta} + 1\right]^{\left\lceil\frac{1-\theta}{2\theta-1}\right\rceil}\log(eT), \tag{2.4}$$

where $\widetilde{C}$ is a constant independent of T and δ.

Please cite this article as: X. Chen, Y. Lei, Refined bounds for online pairwise learning algorithms, Neurocomputing (2017), https://doi.org/10.1016/j.neucom.2017.11.049