Journal of Complexity 47 (2018) 42–61

Contents lists available at ScienceDirect

Journal of Complexity

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

Generalization properties of doubly stochastic learning algorithms☆

Junhong Lin [a,*,1], Lorenzo Rosasco [a,b]

[a] LCSL, Massachusetts Institute of Technology and Istituto Italiano di Tecnologia, Cambridge, MA 02139, USA
[b] DIBRIS, Università degli Studi di Genova, Via Dodecaneso 35, Genova, Italy

ARTICLE INFO

Article history:
Received 4 July 2017
Accepted 30 January 2018
Available online 21 February 2018

Keywords:
Kernel method
Doubly stochastic algorithm
Nonparametric regression

ABSTRACT

Doubly stochastic learning algorithms are scalable kernel methods that perform very well in practice. However, their generalization properties are not well understood and their analysis is challenging since the corresponding learning sequence may not be in the hypothesis space induced by the kernel. In this paper, we provide an in-depth theoretical analysis for different variants of doubly stochastic learning algorithms within the setting of nonparametric regression in a reproducing kernel Hilbert space and considering the square loss. Particularly, we derive convergence results on generalization error for the studied algorithms either with or without an explicit penalty term. To the best of our knowledge, the derived results for the unregularized variants are the first of this kind, while the results for the regularized variants improve those in the literature. The novelties in our proof are a sample error bound that requires controlling the trace norm of a cumulative operator, and a refined analysis of bounding initial error.

1. Introduction

In nonparametric regression, we are given a set of samples of the form $\{(x_i, y_i)\}_{i=1}^T$, where each $x_i \in \mathbb{R}^d$ is an input, y_i is a real-valued output, and the samples are drawn i.i.d. from an unknown

☆ Communicated by S. Pereverzyev.

* Corresponding author.
E-mail addresses: jhlin5@hotmail.com (J. Lin), lrosasco@mit.edu (L. Rosasco).
[1] J. Lin is now with the École Polytechnique Fédérale de Lausanne, Switzerland.

https://doi.org/10.1016/j.jco.2018.02.004
0885-064X/

distribution on $\mathbb{R}^d \times \mathbb{R}$. The goal is to learn a function which can be used to predict future outputs based on the inputs.

Kernel methods [18,5,21] are a popular nonparametric technique based on choosing a hypothesis space to be a reproducing kernel Hilbert space (RKHS). Stochastic/online learning algorithms [9,3] (often called stochastic gradient methods [14,12] in convex optimization) are among the most efficient and fast learning algorithms. At each iteration, they compute a gradient estimate with respect to a new sample point and then updates the current solution by subtracting the scaled gradient estimate. In general, the computational complexities for training are $O(T + Td)$ in space and $O(T^2d)$ in time, due to the nonlinearity of kernel methods. In recent years, different types of online/stochastic learning algorithms, either with or without an explicit penalty term, have been proposed and analyzed, see e.g. [3,23,25,17,22,15,7,11] and references therein.

In classic stochastic learning algorithms, all sampling points need being stored for testing. Thus, the implementation of the algorithm may be difficult in learning problems with high-dimensional inputs and large datasets. To tackle such a challenge, an alternative stochastic method, called doubly stochastic learning algorithm was proposed in [6]. The new algorithm is based on the random feature approach proposed in [13]. The latter result is based on Bochner's theorem and shows that most shift-invariant kernel functions can be expressed as an inner product of some suitable random features. Thus the kernel function at each iteration in the original stochastic learning algorithm can be estimated (or replaced) by a random feature. As a result, the new algorithm allows us to avoid keeping all the sample points since it only requires generating the random features and recovers past random resampling them using specific random seeds [6]. The computational complexities of the algorithm are $O(T)$ (independent of the dimension of the data) in space and $O(T^2d)$ in time. Numerical experiments given in [6], show that the algorithm is fast and comparable with state-of-the-art algorithms. Convergence results with respect to the solution of regularized expected risk minimization were derived in [6] for doubly stochastic learning algorithms with regularization, considering general Lipschitz and smooth losses.

In this paper, we study generalization properties of doubly stochastic learning algorithms in the framework of nonparametric regression with the square loss. Our contributions are theoretical. First, for the first time, we prove generalization error bounds for doubly stochastic learning algorithms without regularization, either using a fixed constant step-size or a decaying step-size. Compared with the regularized version studied in [6], doubly stochastic learning algorithms without regularization do not involve the model selection of regularization parameters, and thus it may have some computational advantages in practice. Secondly, we also prove generalization error bounds for doubly stochastic learning algorithms with regularization. Compared with the results in [6], our convergence rates are faster and do not require the bounded assumptions on the gradient estimates as in [6], see the discussion section for details. The key ingredients to our proof are an error decomposition and an induction argument, which enables us to derive total error bounds provided that the initial (or approximation) and sample errors can be bounded. The initial and sample errors are bounded using properties from integral operators and functional analysis. The difficulty in the analysis is the estimation of the sample error, since the sequence generated by the algorithm may not be in the hypothesis space. The novelty in our proof is the estimation of the sample error involving upper bounding a trace norm of an operator, and a refined analysis of bounding the initial error.

The rest of the paper is organized as follows. In Section 2, we introduce the learning setting we consider and the doubly stochastic learning algorithms. In Section 3, we present the main results on generalization properties for the studied algorithms and give some simple discussions. Section 4–7 are devoted to the proofs of all the main results.

2. Learning setting and doubly stochastic learning algorithms

Learning a function from a given finite number of instances through efficient and practical algorithms is the basic goal of learning theory. Let the input space X be a closed subset of Euclidean space $\mathbb{R}^d$, the output space $Y \subseteq \mathbb{R}$, and $Z = X \times Y$. Let ρ be a fixed Borel probability measure on Z, with its induced marginal measure on X and conditional measure on Y given $x \in X$ denoted by $\rho_X(\cdot)$ and $\rho(\cdot|x)$ respectively. In statistical learning theory, the probability measure ρ is unknown, but only