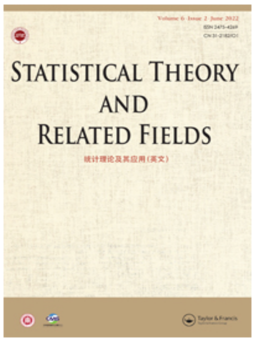
Statistical Theory and Related Fields

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tstf20



Discussion of: a review of distributed statistical inference

Zheng-Chu Guo

To cite this article: Zheng-Chu Guo (2022) Discussion of: a review of distributed statistical inference, Statistical Theory and Related Fields, 6:2, 104-104, DOI: 10.1080/24754269.2021.2022998

To link to this article: https://doi.org/10.1080/24754269.2021.2022998

STATISTICAL THEORY AND RELATED FIELDS
2022, VOL. 6, NO. 2, 104
https://doi.org/10.1080/24754269.2021.2022998

Taylor & Francis
Taylor & Francis Group

SHORT COMMUNICATION

OPEN ACCESS Check for updates

Discussion of: a review of distributed statistical inference

Zheng-Chu Guo

Zhejiang University, Hangzhou, People's Republic of China

ARTICLE HISTORY Received 11 November 2021; Accepted 20 November 2021

Analysing and processing massive data is becoming ubiquitous in the era of big data. Distributed learning based on divide-and-conquer approach has attracted increasing interest in recent years, since it not only reduces computational complexity and storage requirements, but also protects the data privacy when data subsets are distributively stored on different local machines. This paper provides a comprehensive review for distributed learning with parametric models, nonparametric models and other popular models.

As mentioned in this paper, nonparametric regression in reproducing kernel Hilbert spaces is popular in machine learning; however, theoretical analysis for distributed learning algorithms in reproducing kernel Hilbert spaces mainly focuses on the least-square loss functions, and results for some other loss functions are limited; it would be interesting to conduct error analysis for distributed regression with general loss functions and distributed classification in reproducing kernel Hilbert spaces.

In distributed learning, a standard assumption is that the data are identically and independently drawn from some unknown probability distribution; however, this assumption may not hold in practice since data are usually collected asynchronously throughout time. It is of great interest to study distributed learning algorithms with non-i.i.d. data. Recently, Sun and Lin (2020) considered distributed kernel ridge regression for strong mixing sequences. The mixing conditions are very common assumptions in the stochastic processes and the mixing coefficients can be estimated in some cases such as Gaussian and Markov processes. In the community of machine learning, the strong mixing conditions are used to quantify the dependence of samples. It is assumed in Sun and Lin (2020) that D_k $(1 \leq k \leq m)$ is a strong mixing sequence with α-mixing coefficient α_j, and there exists a suitable arrangement of $D_1, D_2, \ldots, D_m$ such that $D = \bigcup_{k=1}^{m} D_k$ is also a strong mixing sequence with α-mixing coefficient α_j; in addition, under some mild conditions on the regression function and the hypothesis spaces, it is shown in Sun and Lin (2020) that as long as the number of the local machines is not too large, an almost optimal convergence rate can be derived, which is comparable to the result under i.i.d. assumptions.

Disclosure statement

No potential conflict of interest was reported by the author.

Reference

Sun, Z., & Lin, S. B. (2020). *Distributed learning with dependent samples*. Preprint. arXiv:2002.03757

CONTACT Zheng-Chu Guo guozhengchu@zju.edu.cn