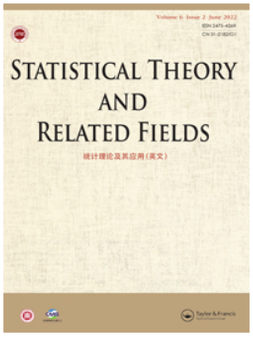

Statistical Theory and Related Fields

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

Rejoinder on 'A review of distributed statistical inference'

Yuan Gao, Weidong Liu, Hansheng Wang, Xiaozhou Wang, Yibo Yan & Riquan Zhang

To cite this article: Yuan Gao, Weidong Liu, Hansheng Wang, Xiaozhou Wang, Yibo Yan & Riquan Zhang (2022) Rejoinder on 'A review of distributed statistical inference', Statistical Theory and Related Fields, 6:2, 111-113, DOI: 10.1080/24754269.2022.2035304

To link to this article: https://doi.org/10.1080/24754269.2022.2035304

Published online: 09 Feb 2022.

STATISTICAL THEORY AND RELATED FIELDS
2022, VOL. 6, NO. 2, 111–113
https://doi.org/10.1080/24754269.2022.2035304

Taylor & Francis
Taylor & Francis Group

SHORT COMMUNICATION

Rejoinder on 'A review of distributed statistical inference'

Yuan Gao[a], Weidong Liu[b], Hansheng Wang[c], Xiaozhou Wang[a], Yibo Yan[a] and Riquan Zhang[a]

[a]School of Statistics and Key Laboratory of Advanced Theory and Application in Statistics and Data Science – MOE, East China Normal University, Shanghai, People's Republic of China; [b]School of Mathematical Sciences – School of Life Sciences and Biotechnology – MOE Key Lab of Artifcial Intelligence, Shanghai Jiao Tong University, Shanghai, People's Republic of China; [c]Guanghua School of Management, Peking University, Beijing, People's Republic of China

ARTICLE HISTORY Received 26 November 2021; Accepted 10 January 2022

We thank the editor, Professor Jun Shao, for organizing this stimulating discussion. We are grateful to all discussants for their insightful comments on our review article on the distributed statistical inference. Due to the urgent need to process the datasets with massive sizes, various distributed computing methods have been proposed for the large-scale statistical problems. Meanwhile, some important theoretical results were established. While we want to give a relatively comprehensive overview on this hot topic, there are still some important works that have been missed in our review written over a year ago. However, we are glad to see the discussants provide reviews of some new works and references. We hope that these discussions and our review would serve as a stimulus for further studies in this rapidly developing area.

In the following discussions, we focus on some main issues raised by the discussants. We refer to Yu and Cheng as YC, Lv and Zhou as LZ, Guo as G, Lian as L, and Zhao as Z.

1. Bias-variance trade-off

As commented by Lv and Zhou (LZ), our review follows the principle of bias-variance trade-off. The bias-variance trade-off plays an important role in the modern statistical learning problems. Different from classical statistical estimation problems, an strictly unbiased estimator is usually hard to derive in many learning tasks, such as nonparametric regression and shrinkage estimation. In these scenarios, an asymptotic unbiased estimator is often satisfactory. However, the order of asymptotic bias usually depends on the sample size. Consequently, when we use the one-shot averaging strategy, the order of the bias of the distributed estimator would depend on the local sample size n, since simple average does not help in reducing bias. The situation will be even worse if we divide the whole sample into too many parts. In this case, the bias term of the averaging estimator could dominate the estimation error (e.g., mean squared error). This can lead to significant efficiency loss when compared with the whole sample estimator. Therefore, we often need a condition to restrict the number of local machines (e.g., $K \ll \sqrt{N}$), which is also pointed out in LZ's discussion. To eliminate the efficiency loss due to non-negligible bias caused by simple average, several debiasing techniques were developed, as discussed in our review. Another line of works adopted the iterative approach, which allows more information to be communicated between different machines. This gives us more freedom to better balance the bias and variance of the resulting estimator. Once an efficient estimator is obtained, some direct inference procedures can be subsequently conducted (Jordan et al., 2018).

2. Inference

As pointed out by Yu and Cheng (YC), we acknowledge that our review focuses mainly on the distributed estimation problems, and does not mention much about the details of the statistical inference. In this regard, we thank YC and Zhao (Z) for introducing some interesting works on the distributed inference. Bootstrap is a useful and flexible tool that applies to various inference problems. However, the heavy computational cost hampers its wide application in large-scale tasks. Bag of Little Bootstraps (BLB), which is briefly introduced in our review, provides a useful distributed approach to ease the computational burden. The `k-grad` and `n+k-1-grad` bootstrap methods (Yu et al., 2020, 2021), mentioned by YC, provide a novel distributed framework for simultaneous inference and remove the constraint on the number of local machines. It is worthwhile to note that the two methods can collaborate well with the CSL method (Jordan et al., 2018), which can be useful to save the communication cost in the high-dimensional problems.

CONTACT Riquan Zhang rqzhang@stat.ecnu.edu.cn

Hypothesis testing of the one sample mean vector is a classical and basic inference problem. We appreciate that Z provides some interesting results on this problem in the distributed framework (Du & Zhao, 2021). In this regard, they find that data splitting does decrease the power of the test. As emphasized by Z, the testing problem could be very different from the estimation problem in the distributed setting. Another inference tool mentioned by Z is the linear discriminant analysis (LDA). In the distributed setting, a one-shot and a two-round methods are proposed in the work that Z refers to. One of the most interesting findings is that, the prior probability of the two classes affects the relative efficiency of the two-round estimator. Beyond LDA, we think it would be also interesting to investigate the quadratic discriminant analysis (Cai & Zhang, 2021) in a distributed framework.

3. Nonparametric learning

Nonparametric distributed learning is one of the most mentioned issues by the discussants. We agree with the difficulties of multivariate-nonparametric distributed learning pointed out by LZ. It is known that the classical local smoothing and spline methods are not very convenient to be used for high-dimensional regression models. In this regard, reproducing kernel Hilbert space (RKHS)-based methods seem to be more popular in the distributed setting. However, as carefully discussed by LZ, making predictions for newly observed data is not a trivial task for RKHS-based algorithm, even in the one-shot averaging manner. Guo (G) also mentions the problems of distributed regression with general loss functions and distributed classification in RKHS. For the former problem, we refer to Xu et al. (2016) as a valuable work. As mentioned by LZ, the existing works on nonparametric distributed learning are almost one-shot methods. Although the pioneering work by Lin et al. (2020) can be seen as an exception, there still are some issues (e.g., communication-efficiency) to be addressed in future works.

Another problem raised by LZ is related to the artificial neural networks, which are particularly popular and have achieved many impressive results in recent years. It is known that the objective function of a neural network is often highly non-convex (Choromanska et al., 2015). As LZ's discussion, the one-shot averaging strategy cannot guarantee the validity of the final distributed estimator in this scenario. Hence, an iterative approach should be adopted. In fact, the so-called *federated learning* technique uses the gradients returned by client machines to update the model parameters (McMahan et al., 2017). However, modern deep neural networks often involve millions of trainable parameters. In these cases, even transferring gradients could lead to a significant communication cost. The issue of communication cost will be discussed in the next section.

4. Communication cost

Communication cost is another common issue concerning the discussants. As pointed out in our review, one major difference between the distributed computing and the traditional parallel computing is that the communication cost cannot be ignored in the former case. To tackle this problem, many communication-efficient methods have been proposed. One of the most successful strategy is the approximate Newton-type method (Shamir et al., 2014), which uses the local high-order information (e.g., Hessian matrices) to approximate the global one. This inspires a lot of works, including Wang et al. (2017), Jordan et al. (2018), Battey et al. (2021), Shi et al. (2021), Zhou et al. (2021), and many others. In the literature, communication-efficiency usually means the communication cost is of the order linear in the parameter dimension p. However, for the high-dimensional models, e.g., deep neural networks, the parameter dimension could be extremely high. In this case, the communication cost should be carefully budgeted. A more realistic and practical approach is to let the communication cost come into play, as discussed by Lian (L) and Z. In this regard, we fully agree with the comments by L and Z, and hope there will be more works taking the statistical, computational, and communicational efficiency into account as a whole.

In addition, as commented by L, the existing statistical studies focus mainly on the centralized distributed framework (i.e., one central machine connected to several local machines). In such a framework, the central machine bears much more communication cost than the local machines. This is because each local machine needs to communicate with the central machine. Consequently, the local machines may queue for the communication with the central machine. Furthermore, the performance of a centralized system depends crucially on the state of the central machine: the breakdown of the central machine could lead to collapse of the whole system. Hence, as suggested by L, it is of great interest to investigate the decentralized distributed system from a statistical perspective. Some related works are Ormándi et al. (2013), Lalitha et al. (2018), Tang et al. (2018), and those in L's discussion.

5. Other issues

As pointed out by L, the distributed framework can also cooperate with some efficient algorithms to further accelerate the local computation process. This is particularly useful when the local sample size is still considerable, due to limited computing resources or other concerns. In addition to the sketching methods (Lian

et al., 2021) mentioned by L, stochastic optimization methods (Chen et al., 2021) and subsampling methods (Yu et al., 2021) can be combined with the distributed framework as well.

The robustness is another important issue that needs to be considered in the distributed learning problems. In practice, it is not unusual that some local machines send totally wrong data to the central machine due to, for example, external attacks or unstable networks. How to cope with these unpredictable consequences and preserve the efficiency of the resulting estimator deserves an in-depth study. We sincerely appreciate that LZ provide a careful review on the Byzantine-robust distributed learning problems.

Last but not the least, as pointed out by G, many works on distributed learning assume that the data are independently and identically distributed (i.i.d.) among different machines. This is often an unrealistic assumption. In this regard, Sun and Lin (2020) consider the dependent samples and Pan et al. (2021) consider the non-randomly distributed samples. Moreover, we think that analysing network data in a distributed system is also a challenging but meaningful work. A recent attempt at community detect can be found in Wu et al. (2020).

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Battey, H., Tan, K. M., & Zhou, W.-X. (2021). *Communication-efficient distributed quantile regression with optimal statistical guarantees*. Preprint. arXiv:2110.13113

Cai, T. T., & Zhang, L. (2021). A convex optimization approach to high-dimensional sparse quadratic discriminant analysis. *The Annals of Statistics*, *49*(3), 1537–1568. https://doi.org/10.1214/20-AOS2012

Chen, X., Liu, W., & Zhang, Y. (2021). First-order Newton-type estimator for distributed estimation and inference. *Journal of the American Statistical Association*, 1–17. https://doi.org/10.1080/01621459.2021.1891925

Choromanska, A., Henaff, M., Mathieu, M., Arous, G. B., & LeCun, Y. (2015). The loss surfaces of multilayer networks. In *Artificial Intelligence and Statistics* (pp. 192–204). PMLR. http://proceedings.mlr.press/v38/choromanska15.pdf

Du, B., & Zhao, J. (2021). *Hypothesis testing of one-sample mean vector in distributed frameworks*. Preprint. arXiv:2110.02588

Jordan, M. I., Lee, J. D., & Yang, Y. (2018). Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 114(526), 668–681. https://doi.org/10.1080/01621459.2018.1429274

Lalitha, A., Shekhar, S., Javidi, T., & Koushanfar, F. (2018). Fully decentralized federated learning. In *Third Workshop on Bayesian Deep Learning (NeurIPS)*. http://bayesiandeeplearning.org/2018/papers/140.pdf

Lian, H., Liu, J., & Fan, Z. (2021). Distributed learning for sketched kernel regression. *Neural Networks*, *143*, 368–376. https://doi.org/10.1016/j.neunet.2021.06.020

Lin, S.-B., Wang, D., & Zhou, D.-X. (2020). Distributed kernel ridge regression with communications. *Journal of Machine Learning Research*, *21*(93), 1–38. https://jmlr.org/papers/volume21/19-592/19-592.pdf

McMahan, B., Moore, E., Ramage, D., & Hampson, S. (2017). Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics* (pp. 1273–1282). PMLR.

Ormándi, R., Hegedös, I., & Jelasity, M. (2013). Gossip learning with linear models on fully distributed data. *Concurrency and Computation: Practice and Experience*, *25*(4), 556–571. https://doi.org/10.1002/cpe.v25.4

Pan, R., Ren, T., Guo, B., Li, F., Li, G., & Wang, H. (2021). A note on distributed quantile regression by pilot sampling and one-step updating. *Journal of Business & Economic Statistics*, 1–10. https://doi.org/10.1080/07350015.2021.1961789

Shamir, O., Srebro, N., & Zhang, T. (2014). Communication-efficient distributed optimization using an approximate newton-type method. In *International Conference on Machine Learning* (pp. 1000–1008). PMLR.

Shi, J., Qin, G., Zhu, H., & Zhu, Z. (2021). Communication-efficient distributed m-estimation with missing data. *Computational Statistics & Data Analysis*, *161*, Article 107251. https://doi.org/10.1016/j.csda.2021.107251

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

Tang, H., Lian, X., Yan, M., Zhang, C., & Liu, J. (2018). D^2: Decentralized training over decentralized data. In *International Conference on Machine Learning* (pp. 4848–4856). PMLR.

Wang, J., Kolar, M., Srebro, N., & Zhang, T. (2017). Efficient distributed learning with sparsity. In *International conference on machine learning* (pp. 3636–3645). PMLR.

Wu, S., Li, Z., & Zhu, X. (2020). *Distributed community detection for large scale networks using stochastic block model*. Preprint. arXiv:2009.11747

Xu, C., Zhang, Y., Li, R., & Wu, X. (2016). On the feasibility of distributed kernel regression for big data. *IEEE Transactions on Knowledge and Data Engineering*, *28*(11), 3041–3052. https://doi.org/10.1109/TKDE.2016.2594060

Yu, Y., Chao, S.-K., & Cheng, G. (2020). Simultaneous inference for massive data: Distributed bootstrap. In *International Conference on Machine Learning* (pp. 10892–10901). PMLR.

Yu, Y., Chao, S.-K., & Cheng, G. (2021). *Distributed bootstrap for simultaneous inference under high dimensionality*. Preprint. arXiv:2102.10080

Zhou, X., Chang, L., Xu, P., & Lv, S. (2021). *Communication-efficient byzantine-robust distributed learning with statistical guarantee*. Preprint. arXiv:2103.00373