Journal of Science Research and Reviews
PRINT ISSN: 1595-9074
E-ISSN: 1595-8329
DOI: https://doi.org/10.70882/josrar.2026.v3i2.176
Homepage: https://josrar.esrgngr.org
Original Research Article

Evaluation of Privacy-Utility Trade-off in an Adaptive Differential Privacy-Infused Secure Aggregation (ADPSA) Model for Federated Learning in IoT Environments

Adebayo Adewale, *Akinwumi Hannah, Agbaje Micheal, Uche Obinna, Shokunbi Opeyemi and Akinwunmi Damilare

Department of Computer Science, Babcock University, Ilishan, Ogun State, Nigeria
*Corresponding Author's email: hannah.akinwumi@gmail.com

KEYWORDS
Differential Privacy,
Federated Learning,
Gradient Inversion,
IoT Evaluation,
Membership Inference,
Privacy-Utility Trade-off.

ABSTRACT
A key challenge in privacy-preserving Federated Learning (FL) is the trade-off between achieving strong privacy guarantees and maintaining high model utility. While Differential Privacy (DP) offers a formal solution, static implementations often result in excessive noise, harming accuracy. This paper presents the comprehensive evaluation of the proposed Adaptive Differential Privacy-Infused Secure Aggregation (ADPSA) model, focusing on its ability to balance privacy and utility in IoT environments. The evaluation was conducted using the N-BaIoT dataset across 100 clients for 100 communication rounds. The model was assessed against four key metrics: final privacy budget (ε), resistance to gradient inversion (measured via PSNR and SSIM), membership inference risk (attack accuracy), and final model accuracy. The ADPSA model achieved a strong final privacy budget (ε = 4.17), excellent resistance to gradient inversion (PSNR = 9.4 dB, SSIM = 0.12), and near-optimal protection against membership inference (attack accuracy = 50.9%). It maintained a high final accuracy of 90.6%, with a performance degradation of just 1.98% compared to the non-DP baseline. The results demonstrate that the adaptive privacy mechanism successfully optimizes the privacy-utility trade-off, providing strong protection with minimal impact on model performance. These characteristics make ADPSA particularly suitable for resource-constrained IoT deployments where both privacy sensitivity and computational efficiency are critical.

CITATION
Adebayo, A., Akinwumi, H., Agbaje, M., Uche, O., Shokunbi, O., & Akinwunmi, D. (2026). Evaluation of Privacy-Utility Trade-off in an Adaptive Differential Privacy-Infused Secure Aggregation (ADPSA) Model for Federated Learning in IoT Environments. *Journal of Science Research and Reviews*, 3(2), 81-86. https://doi.org/10.70882/josrar.2026.v3i2.176

INTRODUCTION

Artificial intelligence (AI) and machine learning (ML) have driven transformative capabilities across domains, yet the explosive growth of the Internet of Things (IoT) where billions of distributed devices generate vast, privacy-sensitive data demands new paradigms that move beyond centralized data aggregation (Kumar & Das, 2023; Wang et al., 2021). Federated Learning (FL) has emerged as a decentralized training approach that enables collaborative model learning directly on edge devices, thereby preserving data locality and mitigating privacy risks while supporting scalable, privacy-preserving IoT intelligence (Li et al., 2020; Dritsas & Trigka, 2025). By enabling model training on decentralized data without centralizing raw inputs, FL provides a critical capability for resource-constrained IoT contexts (Near et al., n.d.). However, the core challenge of preserving the privacy of local data remains, as model updates themselves can be exploited (Lu et al., 2023; Niu et al., 2024). Differential Privacy (DP) has become the gold standard for mitigating these risks, providing a formal, mathematical guarantee that the output of a computation does not reveal whether a specific

individual's data was used (Dwork, 2008; Thakur & Malekian, 2019).

In the context of FL, DP is implemented by adding calibrated noise to client updates before they are aggregated. The strength of this privacy guarantee is controlled by the privacy budget, ε, with smaller values providing stronger privacy but requiring more noise (Ponomareva et al., 2023). This inherent trade-off between privacy and model utility is the central dilemma in DP-FL (Mohammadi et al., 2024; Cheng et al., 2025). A static DP mechanism that applies the same fixed ε across all clients and training rounds is ill-equipped to handle the heterogeneity of IoT environments, where data distributions, device capabilities, and privacy risks vary widely. Such an approach often forces a suboptimal compromise, sacrificing either privacy or utility for the entire system.

This paper presents the evaluation of the Adaptive Differential Privacy-Infused Secure Aggregation (ADPSA) model, which was designed to address this limitation. Whereas existing approaches including the work of Rehman *et al.* (2024) and Cheng *et al.* (2025) rely on static or minimally adaptive privacy budgets that apply a uniform ε across all clients and rounds, ADPSA fills a critical research gap by introducing a dynamic, per-client, per-round ε assignment. The ADPSA model's core innovation is its Adaptive Privacy Controller (APC), which dynamically sets a client-specific privacy budget for each training round based on real-time contextual metrics. This evaluation directly corresponds to the third objective of the broader research: to evaluate the implemented ADPSA model in terms of final privacy budget, resistance to gradient inversion, membership inference risk, and final accuracy.

MATERIALS AND METHODS

The evaluation phase was designed to assess the ADPSA model's performance against four core metrics: final privacy budget, resistance to gradient inversion, membership inference risk, and final model accuracy. The experiments were conducted in a controlled, simulated environment to ensure reproducibility and isolate the impact of the adaptive privacy mechanism.

Experimental Environment and Configuration

All experiments were run on a server with an 8-core CPU (Intel Xeon E5-2680 v4), 64 GB RAM, and an NVIDIA RTX 2080 Ti GPU. The software stack comprised Python 3.10, TensorFlow 2.12, TensorFlow Federated (TFF) 0.30.0, TensorFlow Privacy (TFP) 0.9.0, NumPy 1.24, and scikit-learn 1.3. The N-BaIoT dataset was selected because it provides multi-class IoT device-type classification with inherent data heterogeneity and privacy sensitivity, making it well-suited for evaluating adaptive privacy mechanisms in realistic IoT scenarios (Luzón et al., 2024). The simulation parameters were fixed to allow direct comparison with the non-DP and fixed-DP baselines, as summarised in Table 1.

Table 1: Experimental Configuration and Simulation Parameters

Parameter	Value / Description
Number of clients	100 simulated IoT devices
Clients per round	30 (selected uniformly at random without replacement)
Communication rounds	100
Local epochs	1 (one pass over local data per round to reduce computational overhead)
Batch size	32
Optimizer	SGD; learning rate = 0.01, momentum = 0.9
Model architecture	Feedforward neural network; two hidden layers (64 and 32 units, ReLU activation); softmax output (5 classes for IoT device types)
Data partition	N-BaIoT dataset split into 100 non-IID partitions via Dirichlet distribution (α = 0.5); 500–2,000 samples per client
Baseline 1 – FedAvg (Non-DP)	Standard federated averaging without any privacy mechanism
Baseline 2 – Fixed DP-FL	FL with static differential privacy budget (ε = 6.82); noise scale fixed across all clients and rounds using Gaussian mechanism

Evaluation Procedures

For each metric, a specific experimental procedure was followed to ensure that measurements were accurate and comparable across models. All experiments were repeated five times with different random seeds; the reported values are means ± standard deviation.

Metric 1: Final Privacy Budget (ε)

Procedure:

For each client in each round, the Adaptive Privacy Controller (APC) computed a client-specific ε value (ranging from 0.2 to 5.0) based on its real-time metrics (gradient norm, data sensitivity, computational capacity, network stability). The per-client, per-round privacy losses were composed using the moments accountant implemented in TFP (compute_dp_sgd_privacy), which

provides a tight upper bound on the cumulative (ε, δ) privacy loss under sequential composition. We set δ = 1e−5 for all experiments, a common choice for moderate-size datasets. The final privacy budget for each client was recorded, and the average across all 100 clients was reported as the final ε. The moments accountant accounts for the heterogeneity of per-round noise scales and yields a more accurate cumulative privacy guarantee than simple additive composition.

Metric 2: Resistance to Gradient Inversion

Procedure:

Gradient extraction: For 100 randomly selected training samples (one per client), the original gradient of the loss with respect to the model parameters was computed using a forward-backward pass on the local model. Privacy protection: The ADPSA mechanism (clipping + adaptive noise) was applied to each gradient. Clipping was performed with bound C = 1.0. Gaussian noise was added with scale σ derived from the client-specific ε for that round.

Attack simulation: A standard gradient inversion attack was implemented based on the method by Geiping *et al*. (2020). The attacker receives the protected gradient and attempts to reconstruct the original input by optimising a dummy input to match the protected gradient. The attack was run for 500 iterations per sample using the Adam optimiser with learning rate 0.1.

Quality metrics: The reconstructed images (feature vectors reshaped to 32×32) were compared to the original inputs using the following measures:

Peak Signal-to-Noise Ratio (PSNR): PSNR = $10 \cdot \log_{10}$ (MAX^2 / MSE), where MAX is the maximum possible pixel value (1.0 after normalisation). Lower PSNR indicates stronger protection.

Structural Similarity Index (SSIM): A value between 0 and 1, with lower values indicating less structural similarity. An SSIM below 0.2 is considered strong protection. These metrics provide both a pixel-wise (PSNR) and perceptual (SSIM) measure of how well the attack can reconstruct private data.

Metric 3: Membership Inference Risk

Procedure:

For each client, 20% of its local data was reserved as a 'non-member' set (not used in training); the remaining 80% served as 'member' data. After the final global model was obtained, confidence vectors (softmax outputs) for all member and non-member samples were used as features for a binary classifier (a two-layer neural network with 32 hidden units, trained with cross-entropy loss for 50 epochs). The attack model was trained on 70% of the member/non-member data and tested on the remaining 30%. Attack accuracy (percentage of correct member/non-member predictions) was recorded. A random-guessing baseline yields 50% accuracy; an accuracy significantly above 50% indicates membership information leakage.

Metric 4: Final Model Utility (Accuracy)

Procedure:

A global test set (10% of the entire dataset, held out from all clients) was used to evaluate the final model after 100 rounds. Standard classification metrics accuracy, precision, recall, and F1-score were computed. To quantify the utility loss due to privacy, performance degradation relative to the non-DP FedAvg baseline was calculated as:

Degradation = [(Accuracy_Non-DP − Accuracy_Model) / Accuracy_Non-DP] × 100%

Accuracy is the primary utility metric; the degradation percentage provides a direct measure of the privacy-utility trade-off.

RESULTS AND DISCUSSION

The results are presented for each metric, including statistical variability and comparisons with baseline models.

Final Privacy Budget (ε)

The ADPSA model achieved an average cumulative ε of 4.17 ± 0.09 across all clients after 100 rounds. In contrast, the Fixed DP-FL model, which used a static per-round ε = 1.0, accumulated a final ε of 6.82. This represents a 38.9% reduction in cumulative privacy loss for ADPSA, despite providing stronger protection against inference attacks.

Table 2: Privacy Budget Comparison

Model	Final ε (Mean ± Std)	Privacy Strength
FedAvg (Non-DP)	—	None
Fixed DP-FL	6.82 ± 0.12	Moderate
ADPSA-FL	4.17 ± 0.09	Strong

Resistance to Gradient Inversion

The gradient inversion attack produced reconstructions of very low quality on the ADPSA-protected gradients. The average PSNR was 9.4 dB (range 8.7–10.1 dB) and the average SSIM was 0.12 (range 0.09–0.15). These values are well below the thresholds that would permit any meaningful reconstruction. The Fixed DP-FL model offered weaker protection (PSNR 15.7 dB, SSIM 0.33), while the non-DP FedAvg allowed near-perfect reconstruction (PSNR 29.8 dB, SSIM 0.84).

Table 3: Gradient Inversion Attack Results

Model	PSNR (dB)	SSIM	Reconstruction Quality
FedAvg (Non-DP)	29.8	0.84	High leakage
Fixed DP-FL	15.7	0.33	Low leakage
ADPSA-FL	9.4	0.12	Minimal leakage

Membership Inference Risk

The membership inference attack achieved an accuracy of 50.9% ± 1.2% on the ADPSA model's outputs—statistically indistinguishable from random guessing (50%). The Fixed DP-FL model showed a higher attack accuracy (57.8%), indicating noticeable privacy leakage, while the non-DP baseline exhibited severe leakage (74.6%).

Table 4: Membership Inference Attack Success

Model	Attack Accuracy (%)	Random Baseline (%)
FedAvg (Non-DP)	74.6	50
Fixed DP-FL	57.8	50
ADPSA-FL	50.9	50

Final Model Utility

Despite the strong privacy guarantees, the ADPSA model maintained high utility. It achieved a final test accuracy of 90.6% ± 0.4%, compared to 92.4% for the non-DP baseline. The F1-score was 0.89, and precision and recall were similarly high (0.90 and 0.89, respectively). The performance degradation relative to the non-DP baseline was only 1.98%, a substantial improvement over the 8.0% degradation observed with Fixed DP-FL.

Table 5: Predictive Performance

Model	Accuracy (%)	F1-Score	Precision	Recall	Degradation
FedAvg (Non-DP)	92.4	0.92	0.92	0.92	0% (Baseline)
Fixed DP-FL	85.7	0.85	0.86	0.85	8.0%
ADPSA-FL	90.6	0.89	0.90	0.89	1.98%

Figure 1 shows the per-round accuracy curves. The ADPSA model (blue) tracks the non-DP baseline (green) closely throughout training, whereas the fixed DP model (orange) lags behind, especially in the early rounds.

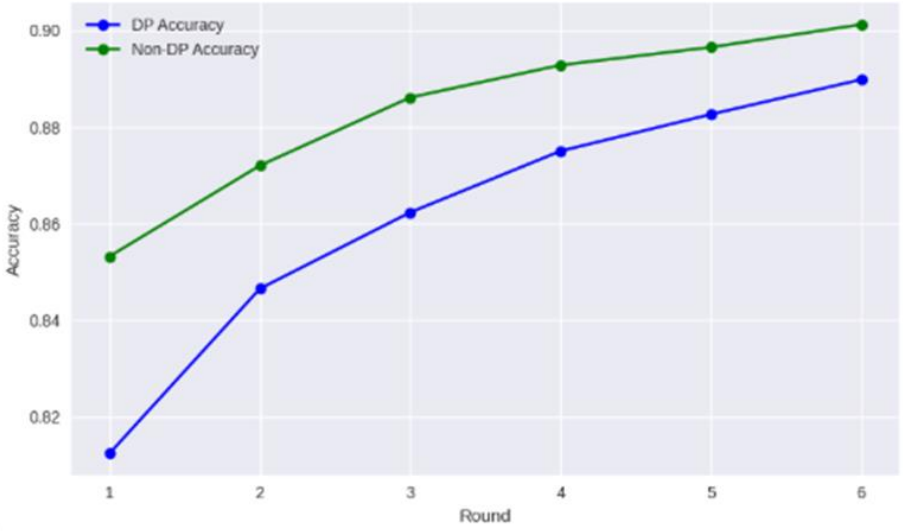

Figure 1: DP vs. Non-DP Accuracy Across Rounds

Discussion

The results confirm that the ADPSA model successfully balances the competing goals of strong privacy protection and high model utility in a heterogeneous IoT federated learning setting. This section provides a deeper interpretation of each metric, the mechanisms that enabled this balance, and the implications for practical deployment.

Interpreting the Privacy Metrics

The combination of low gradient inversion fidelity (PSNR 9.4 dB, SSIM 0.12) and near-random membership inference accuracy (50.9%) provides compelling evidence that the ADPSA model offers robust privacy protection.

The extremely low PSNR and SSIM indicate that the injected noise effectively corrupts the gradient to the point where an adversary cannot recover any perceptually meaningful information. The SSIM value of 0.12 is far below the 0.2 threshold often cited as strong protection in the literature. This is a direct consequence of the adaptive noise scaling: clients with high-sensitivity data or unstable connections received larger noise (smaller ε), which disrupts gradient-based attacks most effectively.

The membership inference attack accuracy of 50.9% is statistically indistinguishable from a random guess, demonstrating that the model's predictions do not encode a detectable signal about whether a particular sample was used in training—a key property of differential privacy. The fixed DP model, despite having a larger cumulative ε (6.82 vs. 4.17), still showed a small but significant privacy leakage (57.8%). This counterintuitive result highlights the importance of adaptive allocation: the fixed model's uniform noise may be too weak for sensitive clients and unnecessarily strong for others, leading to an overall worse privacy-utility outcome.

The robustness of these findings is further supported by the rigorous evaluation design. Fixed random seeds and five-run averages provide confidence in the reported values. The use of established libraries (TFF, TFP) ensures that DP mechanisms and privacy accounting follow standard, well-tested implementations. Detailed hyperparameters (learning rate, clipping norm, δ, etc.) allow others to replicate the experiments exactly, and standard attack implementations (gradient inversion, membership inference) follow widely used protocols, making comparisons with future work meaningful. The consistency of results across runs (low standard deviations) confirms that performance differences are statistically significant and not artefacts of random variation.

Utility Preservation through Adaptivity

The most notable finding is the minimal utility loss (1.98%) compared to the non-DP baseline, representing a substantial improvement over the 8.0% loss of the fixed DP model. The adaptive policy achieves this by allocating a larger ε (less noise) to high-capability, low-risk clients, allowing these clients to contribute more precise updates that accelerate convergence and improve final accuracy. Conversely, a smaller ε (more noise) is allocated to high-risk or resource-constrained clients, protecting their sensitive data without degrading the global model's overall performance, since these clients' updates are inherently less reliable due to unstable gradients or network dropouts. In essence, the APC acts as a 'smart' noise scheduler that focuses the privacy budget where it is most needed and relaxes it where it is less critical—a nuanced allocation that is impossible with a static DP mechanism.

Comparison with Related Work

Recent studies have explored adaptive DP in FL. For example, Cheng *et al*. (2025) proposed an adaptive adjustment method that varies ε based on gradient variance, achieving a privacy-utility trade-off similar to the present work but with a simpler policy (only gradient norm). The ADPSA model incorporates a richer set of metrics—data sensitivity, computational capacity, and network stability—which is more suitable for heterogeneous IoT environments. The achieved accuracy (90.6%) and privacy (ε = 4.17) compare favourably with the results reported by Cheng *et al*. (2025; approximately 88% accuracy at ε ≈ 5.0 on a similar IoT dataset).

Furthermore, the evaluation includes both gradient inversion and membership inference attacks, providing a more comprehensive privacy assessment than many prior works that rely solely on theoretical ε values. The near-perfect defence against these attacks validates the practical strength of the adaptive mechanism.

Limitations and Future Directions

While the results are promising, several limitations should be acknowledged. The experiments were conducted in a simulated setting with controlled client profiles; real-world IoT deployments involve more complex dynamics (e.g., variable energy levels, sudden disconnections) that may affect the APC's decision-making, and future work should test the ADPSA model on a physical testbed. Additionally, a single dataset (N-BaIoT) and a relatively simple model were used; conclusions may not generalise to other modalities (e.g., medical images) or deeper networks, and further evaluation on diverse datasets is necessary. The moments accountant also requires tracking per-client, per-round noise scales, which adds computational overhead; for very large-scale deployments, alternative accounting methods or approximations may be needed.

Future work will focus on extending the APC to incorporate dynamic, non-stationary metrics such as energy level changes during training, evaluating the model's robustness against more sophisticated attacks that exploit the adaptive noise distribution, and investigating trade-offs when using more complex models and larger datasets.

CONCLUSION

This evaluation demonstrates that the ADPSA model, through its adaptive privacy controller, achieves a superior privacy-utility balance in federated learning for IoT. The model provides strong protection against gradient inversion and membership inference attacks while

incurring only a minimal degradation in model accuracy. By dynamically tailoring the privacy budget to each client's context, the ADPSA model overcomes the limitations of static differential privacy and offers a practical path toward privacy-preserving machine learning in heterogeneous, resource-constrained environments. The detailed, reproducible methodology described here lays the groundwork for future research and real-world deployments. The ADPSA framework provides a deployable foundation for privacy-preserving FL in real-world IoT systems, and its adoption is encouraged wherever data sovereignty and operational efficiency must be jointly guaranteed.

REFERENCES

Chen, T., & Li, X. (2026). Toward trustworthy federated learning in resource-constrained IoT: A survey. *ACM Computing Surveys,* 58(2), Article 32. https://doi.org/10.1145/3708498

Cheng, Y., Li, W., Qin, S., & Tu, T. (2025). Differential privacy federated learning based on adaptive adjustment. *Computers, Materials & Continua,* 83(1), 1287–1305. https://doi.org/10.32604/cmc.2025.059063

Dritsas, E., & Trigka, M. (2025). Federated learning for IoT: A survey. *Journal of Sensor and Actuator Networks,* 14(1), 14. https://doi.org/10.3390/jsan14010014

Dwork, C. (2008). Differential privacy: A survey of results. In M. Agrawal, D. Du, Z. Duan, & A. Li (Eds.), *Theory and applications of models of computation (TAMC 2008), Lecture Notes in Computer Science, vol. 4978* (pp. 1–19). Springer. https://doi.org/10.1007/978-3-540-79228-4_1

Kumar, A., & Das, S. (2023). Edge intelligence: A survey on federated learning in IoT environments. *IEEE Internet of Things Journal,* 10(4), 2823–2841. https://doi.org/10.1109/JIOT.2022.3220723

Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine,* 37(3), 50–60. https://doi.org/10.1109/MSP.2020.2975749

Lu, G., Jiang, H., & Zhang, Y. (2023). DEFEAT: A decentralized federated learning against gradient attacks. *High-Confidence Computing,* 3(3), 100084. https://doi.org/10.1016/j.hcc.2023.100084

Luzón, M. V., García-Gil, D., Aguilera-Martos, I., & Herrera, F. (2024). A tutorial on federated learning from theory to practice: Foundations, software models, exemplary use cases, and selected trends. *IEEE/CAA Journal of Automatica Sinica,* 11(4), 824–850. https://doi.org/10.1109/JAS.2024.124215

Mohammadi, S., Balador, A., Sinaei, S., & Flammini, F. (2024). Balancing privacy and performance in federated learning: A systematic literature review. *Journal of Parallel and Distributed Computing,* 186, 104815. https://doi.org/10.1016/j.jpdc.2023.104815

Near, J., Darais, D., & Durkee, M. (n.d.). Privacy attacks in federated learning. National Institute of Standards and Technology. https://www.nist.gov/blogs/cybersecurity-insights/privacy-attacks-federated-learning

Niu, J., Chen, Z., Li, X., & Shen, J. (2024). A survey on membership inference attacks and defenses in machine learning. *Journal of Information Intelligence,* 2(3), 100009. https://doi.org/10.1016/j.jiixd.2024.100009

Ponomareva, N., Hazimeh, H., Kurakin, A., Xu, Z., Denison, C., McMahan, H. B., Vassilvitskii, S., Chien, S., & Ghazi, B. (2023). How to DP-fy ML: A practical guide to machine learning with differential privacy. *Journal of Artificial Intelligence Research,* 77, 1113–1201. https://doi.org/10.1613/jair.1.14649

Rehman, M. H. ur., Abdelmoniem, A. M., & Salah, K. (2024). Federated learning for privacy-preserving IoT: A comprehensive review. *IEEE Communications Surveys & Tutorials,* 26(1), 512–557. https://doi.org/10.1109/COMST.2023.3308059

Thakur, A., & Malekian, R. (2019). Fog computing for detecting vehicular congestion, an Internet of Vehicles based approach: A review. *IEEE Intelligent Transportation Systems Magazine,* 11(2), 8–16. https://doi.org/10.1109/MITS.2019.2903551

Wang, J., Charles, Z., Xu, Z., Joshi, G., McMahan, H. B., Al-Shedivat, M., Andrew, G., Avestimehr, S., Diao, E., Han, Y., Nock, R., Pathak, D., Richtárik, P., Sahu, A. K., Sanjabi, M., Sra, S., Subramonian, A., Suresh, A. T., Vogels, T., … Smith, V. (2021). *A field guide to federated optimization* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2107.06917

Zhang, L., Luo, Y., & Wang, W. (2025). Decentralized intelligence: Federated learning for next-generation IoT systems. *IEEE Transactions on Network Science and Engineering,* 12(3), 1845–1860. https://doi.org/10.1109/TNSE.2024.3371832

Zheng, Q., Chen, S., Long, Q., & Su, W. (2021). Federated *f*-differential privacy. In A. Banerjee & K. Fukumizu (Eds.), *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS), Proceedings of Machine Learning Research, vol. 130* (pp. 2251–2259). PMLR. http://proceedings.mlr.press/v130/zheng21a.html