Developing a Framework for Assessing Synthetic Tabular Data *

Clayton McLamb and Scott Spurlock
[1]Department of Computer Science
Elon University
Elon, NC 27244
{cmclamb, sspurlock}@elon.edu

Abstract

While text and images are at the forefront of generative artificial intelligence, one of the most common forms of data is less represented. Organized into rows and columns, tabular data is used in a multitude of applications such as medical trials and credit scoring. However, the heterogeneity of data types and other challenges have made it difficult to model and generate realistic tabular data. This work aims to provide a comprehensive framework for analyzing generative tabular models. This work evaluates several recent generative models in terms of the statistical quality and machine learning efficacy of synthetic data. Statistical quality, or the "realness" of the data, falls into two groups: intra-feature and inter-feature quality. Intra-feature quality refers to how well each individual feature is modeled and can be quantified using statistical tests, namely the Kolmogorov-Smirnov (KS) test. Inter-feature quality measures how well the synthetic data captures dependencies between features. To measure the inter-feature quality of a synthetic dataset, this work proposes a novel metric: the normalized Pairwise Correlation Difference (nPCD). This work also evaluates the utility of synthetic data in terms of its machine learning efficacy. Following a train-on-synthetic-data, test-on-real-data protocol, this work analyzes the downstream capabilities of synthetic tabular data.

1 Introduction

Machine learning (ML) is increasingly utilized in real-world applications such as healthcare and finance. To build and deploy accurate ML models, these applications need large datasets to train predictive models. However, access to high-quality data is often restricted due to privacy concerns, data scarcity, or the cost of annotating unlabeled data.

Generative artificial intelligence (GenAI) refers to the use of machine learning models to create new and synthetic observations, with applications in data augmentation and differential privacy. Although GenAI has achieved remarkable success in generating text and images, modeling tabular data remains a challenge [13]. Tabular data refers to structured datasets organized in rows and columns, where each row represents an individual record and each column corresponds to a distinct feature or variable. Unlike data types such as images and text, which may be represented as an array of pixels or tokens, tabular data typically contains a mix of continuous and categorical features. In addition, while the quality of synthetic text or images can be assessed through human evaluation or standardized benchmarks, evaluating the quality of synthetic tabular data is not easily accomplished by human evaluation and is less researched. This paper proposes a comprehensive framework for assessing the quality of synthetic tabular data created by various GenAI models.

This framework assesses the quality of synthetic tabular data across multiple dimensions, including utility, variability, and statistical quality. Evaluation with eight benchmark datasets target three research questions:

- **RQ1** How does the statistical quality of synthetic tabular data vary among different generative models? In particular, this paper examines the success of models across multiple types of statistical quality.

- **RQ2** To what extent does the use of synthetic tabular data influence machine learning efficacy?

- **RQ3** What characteristics of generative model architecture contribute to the production of high-quality synthetic tabular data?

In the next section, this paper summarizes recent GenAI models to create, and metrics to assess, synthetic tabular data. In Section 3, this paper outlines the framework and methodology, including a description of the benchmark datasets chosen, followed by a review of the experiments in Section 4. This paper concludes in Section 5 and offers some insight into the utility of synthetic tabular data.

2 Related Work

Unlike images or text, tabular data presents unique challenges for generative modeling due to its heterogeneous structure [13]. Each row in a tabular dataset represents a distinct observation, while columns may include continuous or categorical features. These complexities hinder the application of techniques developed for image or text generation, where data types are typically homogeneous and spatially structured.

To address these challenges, researchers have developed specialized GenAI models for tabular data. This work will investigate models that incorporate the generative adversarial network (GAN), variational autoencoder (VAE), and diffusion model frameworks. While some recent work is beginning to explore large language models (LLMs) for tabular data generation [5], these approaches are beyond the current scope.

2.1 GAN-based Models

The deep-generative GAN-based models reviewed in this work include CTAB-GAN, CTGAN, and TableGAN. TableGAN, one of the first deep learning models attempting to generate synthetic tabular data for differential privacy, introduces a classifier neural network and a convolutional neural network to generate synthetic tabular data [11]. The classifier neural network preserves the semantics of the dataset; e.g., a generated sample with age 5 and occupation "doctor" would be corrected [11]. A conditional tabular GAN (CTGAN) addresses several challenges of tabular data, including multi-modal distributions in continuous columns and imbalanced categorical columns [15]. To generate more realistic data, CTGAN utilizes a variational Gaussian model for multimodal continuous distributions and a “training-by-sampling” methodology for categorical imbalance columns [15]. Another conditional tabular GAN (CTAB-GAN) has similar features to both TableGAN and CTGAN. Like TableGAN, CTAB-GAN utilizes a classifier to preserve the semantics of the dataset [17]. In addition to employing a condition-based architecture like CTGAN, CTAB-GAN also utilizes the “training-by-sampling” methodology [17]. However, CTAB-GAN emphasizes the statistical quality of the generated data while training, focusing on maintaining similar statistical measures for the generated compared to the real data [17].

2.2 VAE-based Models

The variational autoencoder (VAE) counterpart of CTGAN, the tabular variational autoencoder (TVAE), uses the same approach with a VAE architecture [15]. However, the authors report that the TVAE model outperforms the CT-

GAN model across several datasets [15]. A recent method, TabSyn, combines the VAE and diffusion model architectures [16]. TabSyn utilizes a transformer-based encoder and decoder to deal with the challenges of tabular data, while using a diffusion model to learn the latent embeddings and generate realistic tabular data [16].

2.3 Synthetic Data Evaluation

Existing approaches to generative models for tabular data differ significantly in their choice of datasets and evaluation metrics. TableGAN focuses on the privacy of synthetic data, using distance-based metrics and graphical comparisons of cumulative distributions across four datasets [11]. In contrast, CTGAN and TVAE are evaluated using likelihood-based fitness and machine learning efficacy (MLE), reporting Area Under the Curve (AUC) scores across eight real datasets [15]. However, the evaluation of CTGAN and TVAE emphasizes predictive performance over statistical similarity.

The more recent models, CTAB-GAN and TabSyn, introduce more rigorous and comprehensive evaluation metrics. The evaluation of CTAB-GAN assesses not only ML utility (how useful synthetic data is for downstream machine learning tasks) across five datasets, but also compares synthetic to real data based on statistical similarity using Jensen-Shannon divergence, Wasserstein distance, and correlation differences [17]. To evaluate the TabSyn models, the authors evaluate synthetic data using the Kolmogorov-Smirnov test for numeric columns, total variation distance for categorical columns, and pairwise correlation errors [16]. The paper also incorporates MLE through AUC-based comparisons. These variations highlight the absence of a unified evaluation standard, motivating the need for a comprehensive framework that systematically measures both statistical fidelity and machine learning utility.

One recent paper introduces a new metric to describe the statistical quality of synthetic tabular data for medical data [6]. The pairwise correlation difference (PCD) measures how well the synthetic data incorporates similar dependencies between features as the original data. A synthetic dataset that captures the correlation between features well will have a lower PCD. However, PCD is unlike other statistical metrics and is difficult to interpret because it is not bounded, making it challenging to compare across datasets with different dimensionalities or to determine whether a given value represents good or poor performance. In Section 3 this work will describe a normalized PCD to address this challenge.

Machine learning efficacy is the primary utility metric for evaluating the utility of synthetic tabular data [4]. By implementing a train-on-synthetic-data, test-on-real-data protocol, the performance of a classifier should reflect the generative model's ability to preserve the validity of the synthetic data [4].

In the following sections, this project describes the approach to evaluating synthetic data and discusses how these techniques capture various aspects of data quality and usefulness.

3 Methodology

This work aims to provide a framework for a comprehensive analysis of generative tabular models. The framework is designed to assess two critical dimensions of data quality: statistical quality and utility. Statistical quality captures how well the synthetic data preserves the underlying distributional properties and relationships found in the original dataset. This includes intra-feature quality, which evaluates the similarity of individual feature distributions, and inter-feature quality, which measures how well dependencies between features are maintained. In contrast, machine learning efficacy focuses on the practical utility of the synthetic data by measuring the performance of classifiers trained on synthetic data and tested on real data. Together, these two perspectives provide a robust and holistic evaluation of synthetic data quality. The hierarchy of the framework for evaluating tabular synthetic data can be seen in figure 1.

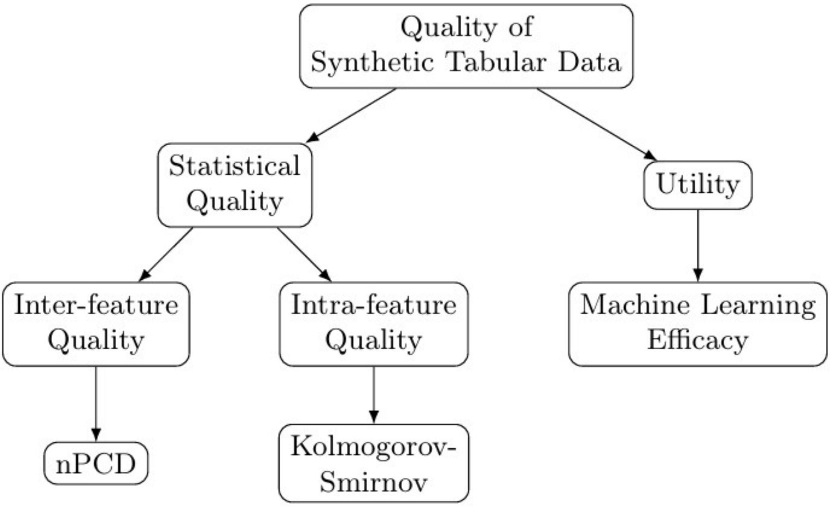
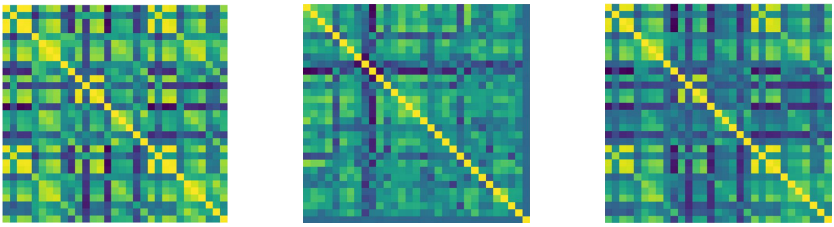


Figure 1: Hierarchy for assessing the quality of synthetic tabular data.



(a) Real (b) Synthetic 1 (c) Synthetic 2

Figure 2: The normalized pairwise correlation difference (nPCD) measures how well a synthetic dataset matches a real dataset. The correlation matrix for the real dataset (a) is less similar to the first synthetic dataset (b), corresponding to an nPCD of 0.199, while it is more similar to the second synthetic dataset (c), with an nPCD of 0.066 (lower is more similar).

3.1 Inter-feature Quality

Inter-feature quality refers to how well the synthetic data captures dependencies between features. While the pairwise correlation difference (PCD) provides a single metric to evaluate the inter-feature quality of the dataset, the interpretation of PCD can be difficult. Adapting the PCD [6], this work proposes the normalized pairwise correlation difference (nPCD), more in line with standard statistical metrics. An nPCD of zero implies that the partial dependencies between features are perfectly captured, while a higher nPCD suggests that a generative model is less capable of capturing dependencies between features. This project defines the nPCD between a real data matrix, X_{real}, and a synthetic data matrix, X_{syn}, as:

$$\text{nPCD}(X_{\text{real}}, X_{\text{syn}}) = \frac{\|C_{\text{real}} - C_{\text{syn}}\|_F}{2\sqrt{n^2 - n}} \tag{1}$$

Here C_{real} and C_{syn} refer to the matrices of Pearson correlation coefficients for each dataset, $\|\cdot\|_F$ refers to the Frobenius Norm, and n is the number of features in a dataset. Note that the upper bound of the Frobenius Norm between two correlation matrices is $2\sqrt{n^2 - n}$, which implies that the nPCD is in the range $[0, 1]$. Figure 2 shows three example correlation matrices; the first is less similar to the second (nPCD of 0.199) than to the third (nPCD of 0.066).

3.2 Intra-feature Quality

Intra-feature quality refers to how well a generative model preserves the distribution of individual features in the dataset. This metric evaluates each feature independently, comparing the synthetic and real data to determine whether they exhibit similar distributions. For each feature, the Kolmogorov-Smirnov (KS) test is used. The KS test is a non-parametric test that compares distributions, determining if two samples are the same. A high p-value ($\alpha > 0.05$) indicates that the synthetic distribution is indistinguishable from its real counterpart. In this work, intra-feature quality is reported as the proportion of features (both continuous and discrete) that are not significantly different from the real data. While future work could explore additional metrics, such as total variation distance for discrete features, this project focuses on developing a holistic and unified framework for evaluating the overall quality of synthetic tabular data.

3.3 Data Utility

As in prior work, machine learning efficacy (MLE) will be used to evaluate the utility of synthetic tabular data. MLE measures how well synthetic data supports downstream predictive tasks by training a classifier on synthetic data and testing it on real data. Performance is measured using two widely used metrics: F1-score, which balances precision and recall, and accuracy, which reflects the proportion of correct predictions. Together, these metrics provide a practical indication of the usability of synthetic data for machine learning tasks. Higher values for F1-score and accuracy indicate that the synthetic data retains predictive structure similar to the real dataset. By comparing these metrics across datasets and generative models, this evaluation highlights which models produce synthetic data that is not only statistically similar but also effective for real-world classification tasks.

3.4 Data

To evaluate the performance of generative models across a variety of conditions, this study uses eight publicly available benchmark datasets spanning different domains, data types, and complexities. These datasets, shown in Table 1, vary widely in size (ranging from 208 to 1,728 observations), number of features (4 to 61), and composition of discrete versus continuous variables. This diversity allows for a robust assessment of how well each generative model handles different types of tabular structures and feature distributions. Importantly, all datasets require minimal preprocessing, which ensures consistency across experiments and avoids introducing variation due to imputation or encoding.

Dataset	N	Features	Discrete	Continuous
Wisconsin Breast Cancer [14]	569	31	1	30
Pima Diabetes [9]	768	9	2	7
Heart Disease [10]	1,025	15	8	7
Sonar [7]	208	61	1	60
Ionosphere [12]	351	35	1	34
Haberman's Survival [8]	306	4	1	3
Vertebral Column [1]	310	7	1	6
Car Evaluation [2]	1,728	7	7	0

Table 1: Overview of datasets and their characteristics.

Model	Breast Cancer	Diabetes	Heart Disease	Sonar	Ionosphere	Haberman's	Vertebral	Car
CTGAN	0.170	0.102	0.059	0.159	0.120	0.147	0.138	0.048
TVAE	0.076	0.048	0.054	**0.085**	**0.072**	0.122	**0.039**	0.034
TabSyn	**0.066**	**0.047**	**0.038**	0.101	0.094	**0.041**	0.052	**0.032**
CTAB-GAN	0.124	0.070	0.044	0.107	0.085	0.100	0.054	0.057
TableGAN	0.199	0.148	0.150	0.201	-	0.212	0.158	0.079

Table 2: Inter-feature quality of each model across datasets in terms of nPCD (lower is better). Note that the TableGAN model failed to converge for the Ionosphere dataset so no result is available.

4 Results and Discussion

This section presents results from applying the proposed evaluation framework to eight benchmark datasets and five generative models. For each dataset, each of the generative model types is trained to generate synthetic data, which is then evaluated with the proposed framework. Findings are organized around the three research questions (RQ1–RQ3) proposed in Section 1.

4.1 RQ1: Statistical Quality of Synthetic Data

Inter-feature quality is measured using the normalized pairwise correlation difference (nPCD) (Section 3.1), where lower values indicate stronger preservation of feature relationships. As shown in Table 2, TabSyn and TVAE consistently achieve the lowest nPCD scores across most datasets, indicating superior performance in capturing inter-feature dependencies. Along with the CTAB-GAN model, the TabSyn and TVAE models perform the most consistently, having the lowest variation in scores. CTGAN and TableGAN perform poorly in this regard, especially on datasets with more continuous features.

This project assesses Intra-feature quality using the Kolmogorov-Smirnov

Model	Breast Cancer	Diabetes	Heart Disease	Sonar	Ionosphere	Haberman's	Vertebral	Car
CTGAN	0.064	0.444	0.428	0.344	0.085	0.500	0.428	0.428
TVAE	**0.741**	0.444	0.571	0.639	**0.771**	0.500	0.714	**0.857**
TabSYN	0.677	0.444	0.500	**0.770**	0.114	**1.000**	**1.000**	0.714
CTAB-GAN	0.354	**0.555**	**0.714**	**0.770**	0.542	0.750	0.571	0.285
TableGAN	0.064	0.111	0.357	0.147	0.200	0.500	0.571	0.428

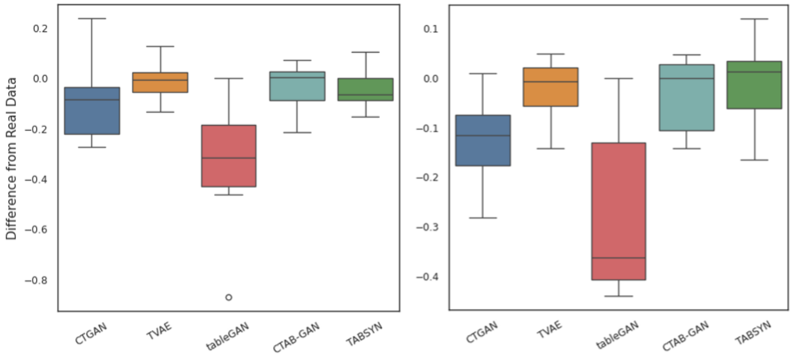
Table 3: Intra-feature quality of synthetic data generated by each model for each dataset in terms of the proportion of the features that are not significantly different from the original data (higher is better).

test across features, reporting the proportion of features with distributions not significantly different from the original data. The results, shown in Table 3, indicate that TabSyn, TVAE, and CTAB-GAN consistently achieve the highest proportions of statistically similar characteristics across the datasets. TabSyn, in particular, performs strongly on datasets such as Haberman's and Vertebral Column, where the dimensionality is lowest. In contrast, CTGAN and TableGAN frequently demonstrate a lower alignment with real data, particularly on datasets with predominantly continuous variables.

4.2 RQ2: Machine Learning Efficacy

Machine learning efficacy (MLE) evaluates the practical utility of synthetic data by measuring how well models trained on synthetic data generalize to real-world tasks. For each combination of dataset and generative model, this work generates a synthetic dataset and uses it to train a separate classifier. Each classifier is evaluated in terms of F1-score and accuracy on a held-out testing set of real data that was not previously used. These metrics can then be compared to the performance of a classifier that was trained on real data. In the experiments, this project tried several classifiers with little difference among them; the results below are based on the XGBoost classifier [3].

As shown in Figure 3, the F1-scores and accuracy metrics vary considerably across both datasets and generative models. TabSyn, CTAB-GAN, and TVAE consistently deliver stronger downstream performance. Notably, TabSyn was able to demonstrate neutral (no difference in accuracy) or improved (positive difference in accuracy) performance across five datasets. Furthermore, TabSyn was able to achieve over an 11% increase in accuracy when tested on the Sonar dataset. Among the top-performing models, TabSyn resulted in an average accuracy drop of only 1.2%, while CTAB-GAN and TVAE saw modest decreases of 3.3% and 2.2%, respectively. This indicates that these models are better able to preserve predictive structure when generating synthetic tabular data. In contrast, TableGAN and CTGAN frequently underperform.



(a) F1-score

(b) Accuracy

Figure 3: Each model (x-axis) generates synthetic data that is used to train a classifier. Classifier performance is compared to a classifier trained with real data (y-axis) in terms of F1-score (a) and accuracy (b). (Higher is better.)

4.3 RQ3: Characteristics of Architecture

Based on the results, which includes both statistical quality and machine learning efficacy, the top-performing models are CTAB-GAN, TVAE, and TabSyn. Compared to the two underperforming GAN-based models, TableGAN and CTGAN, the top-performing models incorporate mechanisms that explicitly condition on, reconstruct, or directly model the real data distribution during generation, allowing them to retain meaningful structure and dependencies from the original dataset. Both the variational autoencoder models directly attempt to reconstruct real data, while CTAB-GAN is aware of the statistics of the real data while training.

5 Conclusion

This study presents a comprehensive framework for evaluating the quality of synthetic tabular data, addressing both statistical quality and downstream utility. Through experiments across eight diverse datasets, the results demonstrate that generative models such as CTAB-GAN, TVAE, and TabSyn consistently outperform other approaches by better preserving feature distributions and inter-feature relationships, while maintaining strong machine learning efficacy. As synthetic data becomes increasingly important for machine learning, this framework offers a standardized approach for future evaluations and comparisons of generative models.

References

[1] G. Barreto and A. Neto. *Vertebral Column Data Set.* https://archive.ics.uci.edu/ml/datasets/Vertebral+Column. UCI Machine Learning Repository, Accessed: 2025-06-18. 2010.

[2] CL Blake and CJ Mertz. *UCI Repository of machine learning database, Irvine, CA: University of California.* 1998.

[3] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* KDD '16. San Francisco, California, USA: Association for Computing Machinery, 2016, pp. 785–794. DOI: 10.1145/2939672.2939785. URL: https://doi.org/10.1145/2939672.2939785.

[4] Yuntao Du and Ninghui Li. *Systematic Assessment of Tabular Data Synthesis Algorithms.* 2024. arXiv: 2402.06806 [cs.CR]. URL: https://arxiv.org/abs/2402.06806.

[5] Xi Fang et al. "Large Language Models (LLMs) on Tabular Data: Prediction, Generation, and Understanding - A Survey". In: *Transactions on Machine Learning Research* (2024). ISSN: 2835-8856. URL: https://openreview.net/forum?id=IZnrCGF9WI.

[6] Andre Goncalves et al. "Generation and evaluation of synthetic patient data". In: *BMC Medical Research Methodology* (2020). DOI: https://doi.org/10.1186/s12874-020-00977-1.

[7] R. Paul Gorman and Terrence J. Sejnowski. *Sonar, Mines vs. Rocks Data Set.* https://archive.ics.uci.edu/ml/datasets/sonar. UCI Machine Learning Repository, Accessed: 2025-06-18. 1988.

[8] S. Haberman. *Haberman's Survival Data Set.* https://archive.ics.uci.edu/dataset/43/haberman+s+survival. UCI Machine Learning Repository, Accessed: 2025-06-18. 1999.

[9] M. Kahn. *Pima Indians Diabetes Database.* https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database. Accessed: 2025-06-17. 2017. URL: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database.

[10] David Lapp. *Heart Disease Dataset.* Kaggle dataset. Available at https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset (based on the UCI Machine Learning Repository). 2019. URL: https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset.

[11] Noseong Park et al. “Data synthesis based on generative adversarial networks”. In: *Proc. VLDB Endow.* 11.10 (June 2018), pp. 1071–1083. ISSN: 2150-8097. DOI: 10.14778/3231751.3231757. URL: https://doi.org/10.14778/3231751.3231757.

[12] V. Sigillito and S. Wing. *Ionosphere Data Set.* https://archive.ics.uci.edu/ml/datasets/ionosphere. UCI Machine Learning Repository, Accessed: 2025-06-18. 1989.

[13] Alex X Wang et al. “Challenges and opportunities of generative models on tabular data”. In: *Applied Soft Computing* (2024), p. 112223.

[14] W. Wolberg and O. Mangasarian. *Breast Cancer Wisconsin (Diagnostic) Data Set.* https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic). UCI Machine Learning Repository, Accessed: 2025-06-18. 1995.

[15] Lei Xu et al. “Modeling tabular data using conditional GAN”. In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems.* Red Hook, NY, USA: Curran Associates Inc., 2019.

[16] Hengrui Zhang et al. *Mixed-Type Tabular Data Synthesis with Score-based Diffusion in Latent Space.* 2024. arXiv: 2310.09656 [cs.LG]. URL: https://arxiv.org/abs/2310.09656.

[17] Zilong Zhao et al. “CTAB-GAN: Effective Table Data Synthesizing”. In: *Proceedings of The 13th Asian Conference on Machine Learning.* Ed. by Vineeth N. Balasubramanian and Ivor Tsang. Vol. 157. Proceedings of Machine Learning Research. PMLR, 17–19 Nov 2021, pp. 97–112. URL: https://proceedings.mlr.press/v157/zhao21a.html.