Adaptive Constrained Optimization for Tabular Synthetic Data Generation

Saheed Obitayo
J.P. Morgan AI Research
New York
saheed.o.obitayo@jpmorgan.com

Niccolò Dalmasso
J.P. Morgan AI Research
New York
niccolo.dalmasso@jpmchase.com

Vamsi Potluru
J.P. Morgan AI Research
New York
vamsi.k.potluru@jpmchase.com

Manuela Veloso
J.P. Morgan AI Research
New York
manuela.veloso@jpmchase.com

Abstract

Synthetic data generation has recently emerged as a solution in data-scarce regulated industries, such as finance and healthcare. While synthetic data requires navigating various tradeoffs including fidelity, utility, fairness or privacy properties, business objectives are usually focused on a single dimension. Although recent optimization approaches such as SCGOAT [18] enable Bayesian optimization to explore tradeoffs in synthetic data generation, determining the boundaries for constraints remains challenging, as it relies on both the original dataset and the trained generator(s). To tackle this issue, we propose a novel Adaptive Constrained Threshold (ACT) strategy within the SCGOAT framework. Our method starts by relaxing the constraints to identify feasible regions and progressively tightens them, thereby minimizing wasted evaluations in infeasible spaces. Experiments on tabular datasets demonstrate that our approach achieves a competitive tradeoffs between various synthetic data dimensions such as downstream performance, fidelity and privacy, improving over fixed-constraints baseline approaches.

1 Introduction

Synthetic data, as in data artificially generated rather than recorded from real world events, has recently emerged as a solution across regulated and data-scarce sectors such as finance [4, 29] and healthcare [16, 17], as well as an approach to avoid the high cost of acquiring real data, especially for large language models training and fine-tuning [24]. In financial fraud detection, synthetic datasets have shown to be useful to augment fraud-detection pipelines and improve trading research [5, 9]. In healthcare, synthetic data can be used to improve public health models for prediction of infectious diseases [14], investigate healthcare policy implications [13, 3, 12] and to improve the downstream performance of automatic imaging detection models [21, 11]. These technical advances have unfolded alongside sectoral guidance, such as FERPA [27], GDPR [19] and HIPAA [10], highlighting how synthetic data can become a key component in machine learning development infrastructures [22].

While synthetic data generation can be driven by many factors, such as safeguarding privacy of the training set records, fidelity of the generated data with respect to the training data distribution and utility when using synthetic data for downstream tasks, in many applications the business objective typically prioritizes a single aspect. Modelers and developers frequently focus on a single dimension, such a privacy guarantee strong enough to unlock a use case, a fairness constraint required by

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Generative AI in Finance.

law, or a downstream revenue indicator, while expecting remaining axes like fidelity, diversity and generalization properties to be "as good as possible". While recent evaluation frameworks provide a way of evaluating synthetic data across these multiple axes [36, 1] and effectively find tradeoffs across these different dimensions, they do not provide guidance for practitioners on what to prioritize during model development. Additionally, deep generative models are trained to approximate the empirical data distribution and are not amenable to incorporating constraints during training and generation without significant structural changes; see for instance, [33, 23, 34].

Recent meta-approaches such as SCGOAT [18] focus on the optimization of a single downstream metric, but extending such approaches to multiple dimensions becomes cumbersome, as approaches such as augmented-Lagrangian formulation might suffer from feasibility issues and non-trivial convergence behavior for non-convex objective and constraints [8, 25]. In this work, we improve over SCGOAT by proposing an adaptive-threshold optimization framework to generate synthetic data that (i) optimizes for the business-critical objective, while (ii) iteratively tightening the constraints on all other secondary criteria. Inspired by continuation methods [2] and adaptive constraint handling in evolutionary optimization [31], we utilize an adaptive constraints threshold (ACT) strategy within the SCGOAT framework [18] for generating tabular data. In practice, we use Bayesian optimization for optimizing the primary criterion, while initializing secondary criteria (fidelity, coverage) with loose constraints that are adaptively reduced across iterations, shrinking the feasible set until we land at a solution that minimizes the primary objective while achieving the best tradeoff elsewhere. Our approach complements existing work on constraint layers [30], which enforces sample level feasibility by providing an outer-loop mechanism that adaptively guides optimization, as well as strategies that incorporate constraints during either training or hyper-parameter tuning [33, 23, 28].

2 ACT-SCGOAT: Tabular Synthetic Data Generation with Adaptive Constraints

In synthetic data generation, properties such as fidelity and privacy are often formalized as hard constraints [28, 18], but whether the constraints (i) are feasible and, if so, (ii) how big is the feasible region, is often data-dependent and hard to know in advance. This leads to a potential large number of evaluations to be wasted on infeasible regions, which is a known challenge in the Bayesian Optimization literature [15]. To address this, we propose to integrate ACT within the SCGOAT framework [18], which optimizes synthetic tabular data generators on a downstream metric performance by composing multiple models. Given M pre-trained generators, at each iteration k, a synthetic dataset D^k_{syn} is created using a mixture of synthetic data points from the generators with proportions $\boldsymbol{\alpha}^{(k)} = \{\alpha_1, \alpha_2, ..., \alpha_M\}^{(k)}$, with $\sum_m \alpha_m = 1$. Once assembled, we evaluate the synthetic dataset D^k_{syn} on a downstream model (e.g., a binary classifier) and evaluate the main objective $l(\alpha)$ and constraints. Crucially, ACT does not require setting fixed constraints boundaries that would require domain knowledge or data-dependent analysis (as in [15]), but rather evolves the constraints over the iterations as detailed in Section 2.1 and Algorithm 1 below.

2.1 Adaptive Constraints Threshold

Let $\boldsymbol{\alpha}$ denote the mixture weights over a set of M tabular generative models and let g_j represents constraint metrics $g_j(\boldsymbol{\alpha})$, such as privacy and fidelity. Let C^*_j indicate a constraint target bound, and define the feasible region:

$$\mathcal{F} = \{\boldsymbol{\alpha} : s_j \cdot g_j(\boldsymbol{\alpha}) \leq s_j \cdot C^*_j, \forall j\},$$

with $s_j = +1$ for upper bound constraints and $s_j = -1$ for lower bound constraints. Ideally, we want the feasible region to either include the global maximum of the objective, i.e., $\boldsymbol{\alpha}^\star = \max_\alpha l(\alpha)$, or values which are close to the global maximum, i.e., $\boldsymbol{\alpha}_\epsilon = \{\alpha : \text{ s.t. } |l(\alpha) - l(\alpha^\star)| \leq \epsilon\}$.

However, for each dataset and set of generative models, the target bounds are not known in advance: setting them too loose might sacrifice other aspects outside of the objective too aggressively, and setting them too tight might resolve in infeasibility or computational inefficiency in identifying a feasible region. Instead of enforcing C^*_j directly, our approach ACT-SCGOAT begins with a relaxed bound $C^{(0)}_j$ and updates thresholds at iteration t via a tightening schedule:

$$C_j^{t+1} = \begin{cases} \max(\overline{C_j}, C_j^{(t)} - \eta_j e^{-t/T}(C_j^{(t)} - \overline{C_j})) & \text{if upper bound constraint,} \\ \min(\underline{C_j}, C_j^{(t)} + \eta_j e^{-t/T}(\underline{C_j} - C_j^{(t)})) & \text{if lower bound constraint.} \end{cases} \tag{1}$$

with an auto shrink factor $\eta_j e^{-t/T} \in (0,1)$, over T iterations. $\overline{C_j}$ and $\underline{C_j}$ represent the upper and lower bound of the constraints, which can be by taking into consideration the definition of the constraint (or partial domain knowledge). For instance, one might want to set $\underline{C_j} = 0.5$ for a downstream classifier AUC, to make sure the downstream classifier performs better then random. In the early iterations, ACT tightens the constraints quickly, allowing for greater feasibility and encouraging exploration. As the iterations progress, the tightening schedule gradually shrinks. This design is inspired by homotopy methods in optimization [2] and constraint adaptation in evolutionary computation [31]. To ensure feasibility is maintained throughout the process, we use a Guassian process surrogate to evaluate the feasibility over the exploration space, as common in the Bayesian optimization literature [15]. For a candidate $\boldsymbol{\alpha}$, the probability of feasibility under thresholds $C_j^{(t)}$ can be computed by using a separate Gaussian process $\mathcal{G}g_j$ for each constraint:

$$\prod(\boldsymbol{\alpha}; C^{(t)}) = \prod_{j=1}^{J} \Pr_{\mathcal{G}g_j(\boldsymbol{\alpha})}(s_j \cdot g_j(\boldsymbol{\alpha}) \leq s_j \cdot C_j^{(t)}).$$

To evaluate infeasibility, we check if $\max_{\boldsymbol{\alpha}} \prod(\boldsymbol{\alpha}; C^{(t)}) < \rho$, where ρ is a feasibility floor. If the maximum probability is below the feasibility floor, the algorithm reverts to the last feasible threshold. This ensures robustness against infeasibility failures while maintaining diversity in the search. Finally, the acquisition function for each $\boldsymbol{\alpha}$ is then modeled by $\text{CEI}(\boldsymbol{\alpha}) = \text{EI}(\boldsymbol{\alpha}) \cdot \prod(\boldsymbol{\alpha}; C^{(t)})$, where EI $(\boldsymbol{\alpha})$ is the expected improvement in the objective.

3 Experiment

We evaluate our proposed framework on two benchmark datasets: the FICO dataset[1] and Adult dataset [6][2]. Following the SCGOAT setup [18], we optimize a mixtures of 4 synthetic data generators – Gaussian Copula[20], CopulaGAN, CTGAN anb TVAE [35] – to maximize downstream AUC. We enforce dataset-level constraints on privacy, using a nearest-neighbor-based re-identification score (the number of synthetic samples for which the nearest neighbor is a real sample, akin to distance-based privacy metrics in [26]) and summary statistics privacy (SSP intersection, [32]), and on fidelity, using the Kolmogorov-Sminorv test (KS, [7]). We compare ACT-SCGOAT against unconstrained SCGOAT optimization, as well as two baseline which simulate different level of domain knowledge: (a) SCGOAT-Medium and (b) SCGOAT Tight. We run both baselines by incorporating the constrained Bayesian optimization approach by [15] within SCGOAT, with the difference being the upper and lower bounds on constraints are set in advance based on a separate run of unconstrained optimization. Constraints are set tighter for the SCGOAT-Tight variant, see Table 1 for details. All variants sample an initial $n = 20$ points and are given a budget of $n = 200$ samples, which ACT-SCGOAT divides equally over $K = 5$ rounds. Table 2 reports the results for the two datasets. As expected, we observe a utility-privacy tradeoff when moving from unconstrained to constraint optimization. However, using ACT-SCGOAT provides a better downstream performance (higher AUC) than the constrained versions that required setting the threshold in advance (SCGOAT-Medium and SCGOAT-Tight), while landing on final constraints not too dissimilar from the SCGOAT-Tight ones (see Table 1).

4 Conclusions and Future Work

This paper demonstrates the effectiveness of ACT as a strategy for constrained Bayesian optimization that relaxes constraints early and tightens them over time within the SCGOAT framework for synthetic tabular generation. On FICO and Adult datasets, ACT-SCGOAT achieves a superior balance between downstream performance and compliance with constraint metrics compared to fixed-constraint baselines. Future work will: (i) extend these experiments to additional data modalities and generators, such as time series, (ii) evaluate algorithmic fairness constraints and (iii) compare ACT's scheduling strategies systematically for optimization of budget and number of iterations.

[1] Dataset available at `https://community.fico.com/s/explainable-machine-learning-challenge`

[2] Data available on the UCI platform at `https://archive.ics.uci.edu/dataset/2/adult`

Algorithm 1: ACT-SCGOAT

Input: Real data D_{real}, Generators $\theta^m\ \forall m \in M$ initial thresholds $C(0)$, upper and lower bounds $\overline{C}_j$ and $\underline{C}_i$, iteration numbers K, feasibility floor ρ, time constant T, shrink setting η_j.

Create partition $\{D_{train}, D_{val}, D_{test}\} = \{(X_{train}, Y_{train}), (X_{val}, Y_{val}), (X_{test}, Y_{test})\}$

Initialize $\boldsymbol{\alpha} = \{\alpha_m\}_{m\in M}$

for $k = 1, \ldots, K$ **do**

Sample $D_m^k = S_m([\alpha_m^k N]; \theta^m) \quad \forall m \in M$

Create $D_{syn}^k = \begin{pmatrix} D_{GC}^k \\ \hline \vdots \\ \hline D_{TVAE}^k \end{pmatrix} = \left(\begin{array}{c|c} X_{GC}^k & Y_{GC}^k \\ \hline \vdots & \vdots \\ \hline X_{TVAE}^k & Y_{TVAE}^k \end{array}\right)$

Train $\hat{\mu}^k = f(Y_{syn}^k \sim X_m^k)$

Compute $\hat{Y}_{val}^k = \hat{\mu}^k(X_{val})$ and $l^k = \mathcal{L}(\hat{Y}_{val}^k, Y_{val})$

Apply constraints:

- Evaluate constraint metrics $g_j^{(k)}$
- Compute feasibility probability $\prod(\boldsymbol{\alpha}) = \prod_j \Pr[(s_j g_j(\alpha) \le s_j C_j^{(t)}]$
- **If** $\max_{\boldsymbol{\alpha}} \prod(\boldsymbol{\alpha}) < \rho$ **then** Revert to last feasible $C(t) \leftarrow C$
- Mark $\boldsymbol{\alpha}^k$ feasible if $g_j^{(k)}$ within $C_j(t)$

Suggest $\boldsymbol{\alpha}^{k+1}$ using Bayesian optimization approach based on $\{\boldsymbol{\alpha}^0, ..., \boldsymbol{\alpha}^k\}$ and $\{l^0, ..., l^k\}$, with acquisition weighted by feasibility under $C(t)$

Threshold update:

- Let U be indices of upper-bound constraints; L of lower-bound constraints
- $step \leftarrow \eta_j \cdot e^{-k/T}$
- $C_U(t+1) \leftarrow \max(\overline{C_U}, C_U(t) - step \cdot (C_U(t) - \overline{C_U}))$
- $C_L(t+1) \leftarrow \min(\underline{C_L}, C_L(t) + step \cdot (\underline{C_L} - C_L(t)))$

return $D_{syn}^{k^*}$ where $k^* = \arg\min_k l^k$

Table 1: Constraints for SCGOAT-Medium and SCGOAT-Tight, with final ACT-SCGOAT constraints.

Metric	Type of Bound	Medium	Tight	ACT (Adult)	ACT (FICO)
Re-ID Score	Upper bound (↓)	≤ 0.20	≤ 0.10	≤ 0.094	≤ 0.087
KS Test	Lower Bound (↑)	≥ 0.60	≥ 0.70	≤ 0.712	≥ 0.747
SSP	Lower Bound (↑)	≥ 0.05	≥ 0.10	≥ 0.093	≥ 0.117

Table 2: Results across the Adult and FICO datasets, bolding the best constrained optimization model.

Dataset	Method	AUC - Objective (↑)	re-ID Score (↓)	KS Test (↑)	SSP (↑)
Adult	SCGOAT	0.911	0.106	0.935	0.096
	SCGOAT Medium	0.896	0.079	0.746	0.108
	SCGOAT Tight	0.887	0.075	0.722	0.114
	ACT-SCGOAT	**0.902**	0.087	0.776	0.103
FICO	SCGOAT	0.791	0.118	0.935	0.092
	SCGOAT Medium	0.771	0.035	0.922	0.393
	SCGOAT Tight	0.769	0.033	0.912	0.423
	ACT-SCGOAT	**0.781**	0.092	0.931	0.213

Disclaimer

This paper was prepared for informational purposes by the Artificial Intelligence Research group of JPMorgan Chase & Co. and its affiliates ("JP Morgan") and is not a product of the Research Department of JP Morgan. JP Morgan makes no representation and warranty whatsoever and disclaims all liability, for the completeness, accuracy or reliability of the information contained herein. This document is not intended as investment research or investment advice, or a recommendation, offer or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction, and shall not constitute a solicitation under any jurisdiction or to any person, if such solicitation under such jurisdiction or to such person would be unlawful.

References

[1] A. Alaa, B. Van Breugel, E. S. Saveliev, and M. Van Der Schaar. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In *International conference on machine learning*, pages 290–306. PMLR, 2022.

[2] E. L. Allgower and K. Georg. *Numerical continuation methods: an introduction*, volume 13. Springer Science & Business Media, 2012.

[3] A. T. Amoon, O. A. Arah, and L. Kheifets. The sensitivity of reported effects of emf on childhood leukemia to uncontrolled confounding by residential mobility: a hybrid simulation study and an empirical analysis using caps data. *Cancer Causes & Control*, 30(8):901–908, 2019.

[4] S. A. Assefa, D. Dervovic, M. Mahfouz, R. E. Tillman, P. Reddy, and M. Veloso. Generating synthetic data in finance: opportunities, challenges and pitfalls. In *Proceedings of the First ACM International Conference on AI in Finance*, pages 1–8, 2020.

[5] F. L. Becerra-Suarez, H. Alvarez-Vasquez, and M. G. Forero. Improvement of bank fraud detection through synthetic data generation with gaussian noise. *Technologies*, 13(4):141, 2025.

[6] B. Becker and R. Kohavi. Adult. *UCI Machine Learning Repository*, 10:C5XW20, 1996.

[7] V. W. Berger and Y. Zhou. Kolmogorov–smirnov test: Overview. *Wiley statsref: Statistics reference online*, 2014.

[8] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

[9] D. Carvajal-Patiño and R. Ramos-Pollán. Synthetic data generation with deep generative models to enhance predictive tasks in trading strategies. *Research in International Business and Finance*, 62:101747, 2022.

[10] I. G. Cohen and M. M. Mello. Hipaa and protecting health information in the 21st century. *Jama*, 320(3):231–232, 2018.

[11] H. P. Das, R. Tran, J. Singh, X. Yue, G. Tison, A. Sangiovanni-Vincentelli, and C. J. Spanos. Conditional synthetic data generation for robust machine learning applications with limited pandemic data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11792–11800, 2022.

[12] P. Davis, R. Lay-Yee, and J. Pearson. Using micro-simulation to create a synthesised data set and test policy options: The case of health service effects under demographic ageing. *Health Policy*, 97(2-3):267–274, 2010.

[13] W. T. Enanoria, F. Liu, J. Zipprich, K. Harriman, S. Ackley, S. Blumberg, L. Worden, and T. C. Porco. The effect of contact investigations and public health interventions in the control and prevention of measles transmission: a simulation study. *PloS one*, 11(12):e0167160, 2016.

[14] S. Gallagher, L. F. Richardson, S. L. Ventura, and W. F. Eddy. Spew: synthetic populations and ecosystems of the world. *Journal of Computational and Graphical Statistics*, 27(4):773–784, 2018.

[15] J. R. Gardner, M. J. Kusner, Z. E. Xu, K. Q. Weinberger, and J. P. Cunningham. Bayesian optimization with inequality constraints. In *ICML*, volume 2014, pages 937–945, 2014.

[16] M. Giuffrè and D. L. Shung. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *NPJ digital medicine*, 6(1):186, 2023.

[17] A. Gonzales, G. Guruswamy, and S. R. Smith. Synthetic data in health care: A narrative review. *PLOS Digital Health*, 2(1):e0000082, 2023.

[18] F. Hamad, S. Nakamura-Sakai, S. Obitayo, and V. Potluru. A supervised generative optimization approach for tabular data. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pages 10–18, 2023.

[19] C. J. Hoofnagle, B. Van Der Sloot, and F. Z. Borgesius. The european union general data protection regulation: what it is and what it means. *Information & Communications Technology Law*, 28(1):65–98, 2019.

[20] T. Janke, M. Ghanmi, and F. Steinke. Implicit generative copulas. *Advances in Neural Information Processing Systems*, 34:26028–26039, 2021.

[21] Y. Jiang, H. Chen, M. Loew, and H. Ko. Covid-19 ct image synthesis with a conditional generative adversarial network. *IEEE Journal of Biomedical and Health Informatics*, 25(2):441–452, 2020.

[22] J. Jordon, L. Szpruch, F. Houssiau, M. Bottarelli, G. Cherubin, C. Maple, S. N. Cohen, and A. Weller. Synthetic data–what, why and how? *arXiv preprint arXiv:2205.03257*, 2022.

[23] J. Jordon, J. Yoon, and M. Van Der Schaar. Pate-gan: Generating synthetic data with differential privacy guarantees. In *International conference on learning representations*, 2018.

[24] Z. Li, H. Zhu, Z. Lu, and M. Yin. Synthetic data generation with large language models for text classification: Potential and limitations. *arXiv preprint arXiv:2310.07849*, 2023.

[25] R. Nishihara, L. Lessard, B. Recht, A. Packard, and M. Jordan. A general analysis of the convergence of admm. In *International conference on machine learning*, pages 343–352. PMLR, 2015.

[26] W. Niu, A. H. Celdran, K. Siarsky, and B. Stiller. Fest: A unified framework for evaluating synthetic tabular data. In *Proceedings of the International Conference on Information Systems Security and Privacy ICISSP*, number 1, pages 434–444. SCITEPRESS-Science and Technology Publications, 2025.

[27] C. Parks. Beyond compliance: Students and ferpa in the age of big data. *Journal of Intellectual Freedom & Privacy*, 2(2):23–33, 2017.

[28] V. Perrone, M. Donini, M. B. Zafar, R. Schmucker, K. Kenthapadi, and C. Archambeau. Fair bayesian optimization. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 854–863, 2021.

[29] V. K. Potluru, D. Borrajo, A. Coletta, N. Dalmasso, Y. El-Laham, E. Fons, M. Ghassemi, S. Gopalakrishnan, V. Gosai, E. Kreačić, et al. Synthetic data applications in finance. *arXiv preprint arXiv:2401.00081*, 2023.

[30] M. C. Stoian, S. Dyrmishi, M. Cordy, T. Lukasiewicz, and E. Giunchiglia. How realistic is your synthetic data? constraining deep generative models for tabular data. *arXiv preprint arXiv:2402.04823*, 2024.

[31] H. Takagi. Interactive evolutionary computation: Fusion of the capabilities of ec optimization and human evaluation. *Proceedings of the IEEE*, 89(9):1275–1296, 2002.

[32] S. Wang, R. Wei, M. Ghassemi, E. Kreacic, and V. K. Potluru. Guarding multiple secrets: Enhanced summary statistic privacy for data sharing. *arXiv preprint arXiv:2405.13804*, 2024.

[33] L. Xie, K. Lin, S. Wang, F. Wang, and J. Zhou. Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739*, 2018.

[34] B. Xin, W. Yang, Y. Geng, S. Chen, S. Wang, and L. Huang. Private fl-gan: Differential privacy synthetic data generation based on federated learning. In *Icassp 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 2927–2931. IEEE, 2020.

[35] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni. Modeling tabular data using conditional gan. *Advances in neural information processing systems*, 32, 2019.

[36] Y. Yuan, Y. Liu, and L. Cheng. A multi-faceted evaluation framework for assessing synthetic data generated by large language models. *arXiv preprint arXiv:2404.14445*, 2024.

A Experimental Details

The experiments were conducted using CUDA version 11.8 and Python version 3.11.5. Bayesian optimization was performed with the `Bayes OPT` library, using 20 initial points, and 200 optimization iterations. The constraint hyperparameters were set to $\eta_j = 0.6$, $T = 25$, $\rho = 0.05$. We split the data into training, validation, and test sets with a ratio of 70%/15%/15% respectively. All experiments were run on ml.r5.xlarge hardware, with the total runtime per dataset being approximately 4 hours for the Adult dataset and 3 hours for the FICO dataset.