Minimax optimal differentially private synthetic data for smooth queries

Rundong Ding RUNDONGD@USC.EDU
Department of Mathematics, University of Southern California

Yiyun He YIH130@UCSD.EDU
Department of Mathematics, University of California San Diego

Yizhe Zhu YIZHEZHU@USC.EDU
Department of Mathematics, University of Southern California

Editors: Steve Hanneke and Tor Lattimore

Abstract

Differentially private synthetic data enables the sharing and analysis of sensitive datasets while providing rigorous privacy guarantees for individual contributors. A central challenge is to obtain strong utility guarantees for meaningful downstream analysis. Many existing methods provide uniform accuracy over broad query classes, such as all Lipschitz functions, but this level of generality can lead to suboptimal rates for statistics of practical interest. Since many data-analysis queries possess smoothness beyond what is captured by worst-case Lipschitz bounds, it is natural to ask whether this additional structure can be exploited to improve the utility of private synthetic data.

We study the problem of generating (ε, δ)-differentially private synthetic data from a dataset of size n supported on the hypercube $[-1, 1]^d$. Utility is measured uniformly over all k-smooth queries, namely functions with bounded derivatives up to order k. We propose a polynomial-time algorithm that achieves a minimax error rate of $O_{k,d}(n^{-\min\{1, \frac{k}{d}\}})$, up to a $\log(n)$ factor. This characterization reveals a phase transition at $k = d$: increasing the smoothness order improves the sample-size rate until k reaches the ambient dimension, after which the optimal parametric rate is attained.

Our algorithm is based on private Chebyshev moment matching. It generalizes the Chebyshev moment-matching framework of (Musco et al., 2025; Wang et al., 2016) and strictly improves the error rates for k-smooth queries established in (Wang et al., 2016). The analysis relies on two approximation-theoretic ingredients: a higher-order global coefficient-decay bound for multivariate Chebyshev expansions of k-smooth functions, and a multivariate Jackson-type theorem adapted to the query class. Together, these results reduce uniform control over smooth queries to the control of finitely many weighted Chebyshev moments, which can then be released privately using the Gaussian mechanism.

For a broad range of privacy parameters (ε, δ), we prove a matching minimax lower bound that any (ε, δ)-differentially private synthetic data algorithm must incur error at least $(n\varepsilon)^{-\min\{1,k/d\}}$, up to constants depending only on k. The proof uses a privacy-aware Assouad construction with localized smooth bump functions supported on disjoint grid cells. Consequently, our upper bound is minimax optimal in its dependence on n, and the phase transition at $k = d$ is intrinsic.

Keywords: Differential privacy, synthetic data, smooth queries, minimax lower bound

Acknowledgments

The authors thank Roman Vershynin for helpful comments. Y.Z. was partially supported by the Simons Grant MPS-TSM-00013944.

1. Extended abstract. Full version appears as https://arxiv.org/abs/2602.01607v3.

References

Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Differentially private assouad, fano, and le cam. In *Algorithmic Learning Theory*, pages 48–78. PMLR, 2021.

March Boedihardjo, Thomas Strohmer, and Roman Vershynin. Covariance's loss is privacy's gain: Computationally efficient, private and accurate synthetic data. *Foundations of Computational Mathematics*, pages 1–48, 2022.

March Boedihardjo, Thomas Strohmer, and Roman Vershynin. Private measures, random walks, and synthetic data. *Probability theory and related fields*, 189(1):569–611, 2024.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Yiyun He, Roman Vershynin, and Yizhe Zhu. Algorithmically effective differentially private synthetic data. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 3941–3968. PMLR, 2023.

Yiyun He, Thomas Strohmer, Roman Vershynin, and Yizhe Zhu. Differentially private low-dimensional synthetic data from high-dimensional datasets. *Information and Inference: A Journal of the IMA*, 14(1):iaae034, 2025.

Cameron Musco, Christopher Musco, Lucas Rosenblatt, and Apoorv Vikram Singh. Sharper bounds for chebyshev moment matching, with applications. In *The Thirty Eighth Annual Conference on Learning Theory*, pages 4309–4358. PMLR, 2025.

Natalia Ponomareva, Zheng Xu, H Brendan McMahan, Peter Kairouz, Lucas Rosenblatt, Vincent Cohen-Addad, Cristóbal Guzmán, Ryan McKenna, Galen Andrew, Alex Bie, et al. How to dp-fy your data: A practical guide to generating synthetic data with differential privacy. *arXiv preprint arXiv:2512.03238*, 2025.

Jonathan Ullman and Salil Vadhan. Pcps and the hardness of generating private synthetic data. In *Theory of Cryptography Conference*, pages 400–416. Springer, 2011.

Ziteng Wang, Chi Jin, Kai Fan, Jiaqi Zhang, Junliang Huang, Yiqiao Zhong, and Liwei Wang. Differentially private data releasing for smooth queries. *The Journal of Machine Learning Research*, 17(1):1779–1820, 2016.