Topology-Aware Robust Representation Balancing for Estimating Causal Effects

Amirhossein Farzam A.FARZAM@DUKE.EDU
Duke University

Ahmed Aloui AHMED.ALOUI@DUKE.EDU
Duke University

Vahid Tarokh VAHID.TAROKH@DUKE.EDU
Duke University

Guillermo Sapiro GUILLERMOS@PRINCETON.EDU
Princeton University and Apple

Abstract

Representation learning in high-dimensional spaces faces significant robustness challenges with noisy inputs, particularly with heavy-tailed noise. Arguing that topological data analysis (TDA) offers a solution, we leverage TDA to enhance representation stability in neural networks. Our theoretical analysis establishes conditions under which incorporating topological summaries improves robustness to input noise, especially for heavy-tailed distributions. Extending these results to representation-balancing methods used in causal inference, we propose the *Topology-Aware Treatment Effect Estimation* (TATEE) framework, through which we demonstrate how topological awareness can lead to learning more robust representations. A key advantage of this approach is that it requires no ground-truth or validation data, making it suitable for observational settings common in causal inference. The method remains computationally efficient with overhead scaling linearly with data size while staying constant in input dimension. Through extensive experiments with α-stable noise distributions, we validate our theoretical results, demonstrating that TATEE consistently outperforms existing methods across noise regimes. This work extends stability properties of topological summaries to representation learning via a tractable framework scalable for high-dimensional inputs, providing insights into how it can enhance robustness, with applications extending to domains facing challenges with noisy data, such as causal inference.

1. Introduction

Robust representation learning is critical across domains, including those involving high-dimensional data, yet developing methods that handle noisy inputs remains challenging, especially for heavy-tailed noise—prevalent in finance and signal processing [2, 23, 29, 50, 52, 57]. Techniques for improving robustness without requiring ground-truth are particularly valuable for observational settings where validation data is unavailable, typical in causal inference applications. Topological Data Analysis (TDA) offers a principled solution through the stability properties of topological summaries, enabling a purely structural approach to robustness. We leverage these properties to enhance the robustness of representation learning, demonstrating their effectiveness through representation-balancing neural networks for treatment effect estimation [26, 28, 46, 55]. Our approach addresses the challenge of representations' robustness to noise, and building on our theoretical results, we introduce a topology-aware framework for treatment effect estimation that demonstrates these stability benefits while maintaining computational scalability with high-dimensional inputs.

Machine learning methods have shown promise for causal inference [13, 20, 35, 39, 47, 48, 59], with representation-balancing approaches gaining popularity for treatment effect estimation [26,

© A. Farzam, A. Aloui, V. Tarokh & G. Sapiro.

46]. However, noisy observations present a key challenge [32, 56], and despite progress [27, 49], techniques for handling noise remain limited, with significant gaps in addressing non-Gaussian noise. Our work addresses this gap by integrating topological summaries in treatment effect estimation, enhancing robustness in a scalable fashion. Topological Data Analysis (TDA) offers a compelling approach to robust representation learning by characterizing the shape of data through its global topology, which remains stable under local perturbations [6]. A standard tool for this characterization is persistent homology, which effectively summarizes topological invariants [15, 16], with well-established stability theorems demonstrating how the resulting topological signatures can enhance robustness in data analysis pipelines [10–12]. Recent work suggests that incorporating these topological summaries into deep learning frameworks can improve their resilience to noise [19, 51], motivating our work and making it the first to bridge TDA and treatment effect estimation.

We leverage the stability properties of topological summaries to enhance the robustness of representations against noise through an approach that is both purely structural and computationally scalable—critical advantages for observational and high-dimensional settings. In Section 3, we derive conditions under which topological summaries enhance representations' robustness by improving metric stability, especially with heavy-tailed noise. This underpins our *Topology-Aware Treatment Effect Estimation (TATEE)* framework proposed in Section 4, which improves the robustness of counterfactual regression (CFR) [46] by imposing both topological and distributional similarities between the representations of treatment and control groups. TATEE's implementation is scalable, with complexity linear in data size and constant in input dimension, once the dimensionality of the representations is fixed—making it suitable for high-dimensional applications. Our experiments confirm that TATEE consistently outperforms existing methods across a range of α-stable noise distributions, including Gaussian and heavy-tailed cases. In conclusion, we establish how topological properties enhance representations' robustness under noise, provide theoretical conditions for this improvement, and demonstrate the benefits in practice. Our main contributions are:

A) Conceptual. We argue that incorporating topological awareness into representation learning offers a principled path to robust deep learning. Our approach is computationally scalable, making it suitable for high-dimensional data, and requires no ground-truth or validation data—a critical advantage for observational settings. To our knowledge, this is the first work to integrate TDA with representation-balancing neural networks for causal inference, improving robustness to noise.

B) Theoretical. We establish new stability-type results for representations learned from noisy data, extending foundational stability theorems from TDA to deep learning. We identify regimes where persistent homology enhances metric stability in representation learning, particularly under heavy-tailed noise, providing a rigorous foundation for the robustness gains achieved by our method.

C) Methodological. We introduce *Topology-Aware Treatment Effect Estimation* (TATEE), a scalable framework that integrates persistence diagrams into representation balancing to improve counterfactual regression's robustness to noise. Extensive experiments across α-stable noise distributions, including Gaussian and heavy-tailed, validate TATEE's ability to meet the conditions for robustness in practical settings, outperforming existing methods.

2. Preliminaries and Related Work

Topological Data Analysis. Topological Data Analysis (TDA) utilizes algebraic topology to extract shape-based features from data across scales [6]. Persistent homology, a central tool in TDA, captures topological features, such as connected components and holes [16]. These features are characterized by their lifespan through varying scales, represented as points in a persistence dia-

gram, where each point corresponds to a feature's birth and death in a filtration of the input—a series of nested simplicial complexes determined by the image of a filtration function [21]. Figure 5 visualizes this concept, with further details in Appendix A. Resulting from *the stability theorem* and fundamental to our work, stability of persistence diagrams is a key property ensuring robustness of these topological summaries to perturbations in data [11]. Consider a triangulable compact metric space $(\mathcal{Z}, d)$ for some metric d, and let $f, g : \mathcal{Z} \to \mathbb{R}$ be Lipschitz filtration functions. The following generalization of the stability theorem holds under conditions in Appendix A.3.

Theorem 1 (Cohen-Steiner et al. [12]) *For some constants $k \geq 1$ and C, we have for all l,*

$$W_p\big(\mu_f^l, \mu_g^l\big) \leq C^{\frac{1}{p}} \, \|f - g\|_\infty^{1-\frac{k}{p}} \,, \tag{1}$$

where μ_f^l and μ_g^l are measures on the space of persistence diagrams corresponding to the l-dimensional homology classes [38], W_p denotes the Wasserstein-p distance, and constants C and k are described in Appendix A.3. Since this inequality holds for all l, we shall drop the superscript l from here on.

Causal Inference. Causal inference aims to determine the effect of a treatment on an outcome Y given covariates X. This effect can be quantified by the conditional average treatment effect (CATE), when conditioned on features. For individuals with covariates x, CATE is given by

$$\tau(x) = \mathbb{E}\left[Y(1) - Y(0) | X = x\right], \tag{2}$$

where $Y(\cdot)$ denotes the potential outcome. The challenge in estimating $\tau(x)$ arises from the unobservability of *counterfactual* outcomes, leading to *the fundamental problem of causal inference*.

Related Work. Causal inference and deep learning have been integrated in various contexts [13, 36, 44], with representation-balancing frameworks showing effectiveness in treatment effect estimation [35, 46] and counterfactual reasoning [26, 41]. Prior work on robust causal effect estimation [27, 49] provides important advances but assumes finite variance noise and often requires large observation counts or validation data. Despite recent progress [33, 43], methods for robust treatment effect estimation with heavy-tailed noise or limited data remain underdeveloped. TDA, which captures the intrinsic shape of data [6, 8], has been integrated into machine learning to improve robustness [4, 18, 40]. We extend these applications to representation balancing for causal inference, introducing a framework that enhances robustness without ground-truth, clean, or large datasets, and accommodates noise beyond finite-variance distributions.

3. Learning Robust Representations via Persistence Diagrams

How can we improve the robustness of representations learned by neural networks under noise—particularly heavy-tailed noise? Here, we introduce a new stability result characterizing when incorporating persistent homology into learning improves metric stability. Our theorem establishes a condition on neural networks' Lipschitz constants that leads to persistence diagrams of representations being more robust to noise than the raw representations themselves.

Problem Setup. Let X and E be random variables representing features and noise, respectively, and define the noise-corrupted features as $\tilde{X} := X + E$. Denote by $\mathbf{X}$, $\mathbf{E}$, $\tilde{\mathbf{X}}$ the corresponding finite-sample matrices. We are interested in the stability of the representations $\varphi(X)$ and $\varphi(\tilde{X})$ where $\varphi : \mathcal{X} \to \mathcal{Z}$, is a neural network mapping to a representation space $\mathcal{Z}$. Observe that $\varphi(X)$ and $\varphi(\tilde{X})$ induce measures on $\mathcal{Z}$, which we denote by μ and $\tilde{\mu}$. Furthermore, suppose that there exist filtration functions f and $\tilde{f}$, which satisfy $\tilde{f}(\varphi(X)) = f(\varphi(\tilde{X}))$, yielding persistence diagrams for both clean and noisy representations and allowing us to invoke the stability theorem. The explicit construction of such filtration functions is provided in Appendix B. While the filtration is typically

fixed once the data is given, the network φ remains trainable. We therefore seek conditions on φ that improve robustness of the persistence diagrams. Importantly, we exploit the Lipschitz continuity of standard neural networks [24, 54] to derive the intended condition on φ.

Finite-Sample Stability and Noise Distribution. Let φ be a K_φ-Lipschitz network. Then Lipschitz continuity arguments yield a finite-sample upper bound $\hat{\mathsf{M}}$ on the Wassertein distance $W_p(\mu, \tilde{\mu})$ between the representations of clean and noisy inputs. Similarly, under standard assumptions, Theorem 1 provides a bound $\hat{\mathsf{K}}_{\text{topo}}$ on the Wasserstein distance between their persistence diagrams, $W_p(\mu_f, \mu_{\tilde{f}})$. These bounds, detailed in Appendix B.1, lead to the following.

Theorem 2 *If $K_\varphi < \Lambda$, then $\hat{K}_{topo} < \hat{M}$, where Λ depends on the noise distribution: Λ is increasing in $\|\mathbf{E}\|_\infty^{p/k-1} \|\bar{\mathbf{E}}\|^{-p/k}$, where $\|\bar{\mathbf{E}}\|$ and $\|\mathbf{E}\|_\infty$ are the average and ∞-norm of the error.*

In the statement above, p is the degree of the Wasserstein distance and k is the constant in Theorem 1. The proof and details about Λ are provided in Appendix B.1. Intuitively, the condition asserts that if the Lipschitz constant of φ is smaller than Λ, the upper bound $\hat{\mathsf{K}}_{topo}$ is smaller than that on the representations, $\hat{\mathsf{M}}$. This suggests that with the appropriate neural network, the Wasserstein space over the persistence diagrams of the representations is more robust than that over the representations themselves–a metric stability which enhances robustness to input noise. Notably, the condition that determines if a neural network is an 'appropriate' one depends on the distribution of the error, particularly, its tail. Theorem 1 leads to the promised stability properties for $p > k$. In this case, the ratio $\|\mathbf{E}\|_\infty^{p/k-1} \|\bar{\mathbf{E}}\|^{-p/k}$ is larger for heavy-tailed distributions. Since Λ is linear in this ratio, a slow-decaying tail of the empirical distribution of the noise corresponds to a more easily achievable neural network that satisfies the condition in Theorem 2.

Implications. Theorem 2 establishes that topological summaries can enhance representations' robustness, particularly with heavy-tailed noise. This insight directly informs our approach to robust causal inference through TATEE. Our experiments in Section 5 confirm that neural networks in TATEE can be trained to satisfy the theoretical conditions and achieve the predicted robustness.

4. Topology-Aware Treatment Effect Estimation

Building on the stability results from Section 3, we introduce ***T**opology-**A**ware **T**reatment **E**ffect **E**stimation* (TATEE), which incorporates topological awareness into representation-balancing neural networks for estimating causal effects. TATEE enhances robustness to input noise by leveraging the stability properties of persistence diagrams in a scalable fashion, showcasing the benefits of topological summaries for robust representation learning. Here, we discuss the main aspects of TATEE's design, analysis, and implications. More details are in Appendix C.

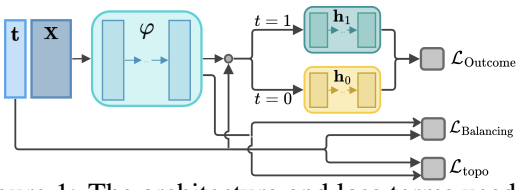

Figure 1: The architecture and loss terms used in TATEE, adopting the two-headed network, outcome loss, and balancing loss from CFR [46]. The topological signature is incorporated as the regularization term $\mathcal{L}_{\text{topo}}$, based on the output of φ.

Architecture and Training. TATEE incorporates topological awareness in representation-balancing neural networks for counterfactual regression (CFR) [46] through a regularization term in the training objective, using the Wasserstein distance between persistence diagrams of treatment and control representations. The architecture follows CFR's two-headed design with a shared encoder φ followed by two separate heads h_1 and h_0 estimating the potential outcomes for the treatment and

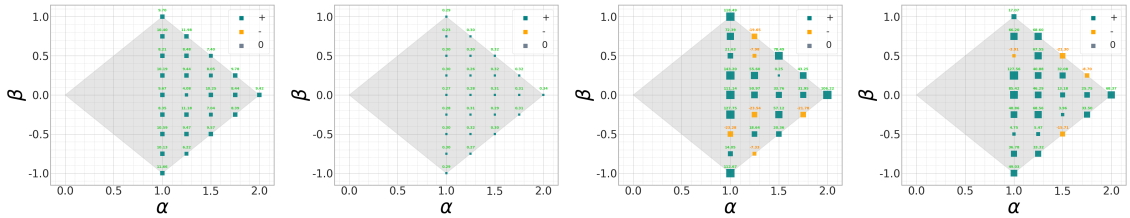

Figure 3: Evaluation of TATEE's robustness when estimating CATE on noise-corrupted features from the IHDP, Twins, Jobs, and ACIC datasets (from left to right) for a range of (α, β) parameters of α-stable distributed noise. The Feller-Takayasu diamond (shaded) marks valid (α, β) values. The relative gain in robustness from using TATEE, quantified by $\rho_{\text{TATEE}} - \rho_{\text{CFR}}$. The green and orange squares mark positive/negative gain; the size of each square is proportional to the magnitude.

control groups. In addition to prediction accuracy, the training objective $\mathcal{L}_{\text{TATEE}} = \mathcal{L}_{\text{Outcome}} + \lambda\mathcal{L}_{\text{Balance}} + \lambda_{\text{topo}}\mathcal{L}_{\text{topo}}$ encourages distributional and topological similarities between the inputs of h_1 and h_0. Complete implementation details are provided in Appendix C.2. This implementation ensures computational scalability by computing persistence diagrams on mini-batches and applying topological regularization to representations rather than raw inputs, resulting in linear overhead with respect to data volume and constant overhead with input dimensionality, as long as the dimensionality of the representations is fixed. Figure 6 illustrates this via a simulation, where overhead elapsed time remains nearly constant as input dimensionality increases from 16 to 256 (with a fixed representation dimensionality), and scales linearly as data volume varies from 800 to 12800 samples.

Robustness of TATEE. Using our stability results, we show that TATEE can improve the robustness of counterfactual regression under the conditions of Theorem 2. In particular, this theorem implies that TATEE's training objective is more stable under input noise than the original CFR's. This is shown by Proposition 9 in Appendix C, informally stated below.

Proposition 3 (Informal) *If φ satisfies the constraint from Theorem 2, then the upper bound of the noise-induced change in TATEE's training objective is smaller than CFR's.*

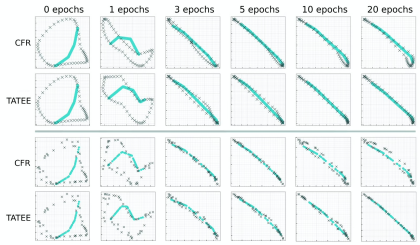

Figure 2: The representations learned by φ in CFR and TATEE throughout 20 epochs of training for the control (gray) and treatment (turquoise) groups, starting without input noise (top rows) and with Gaussian noise (bottom rows).

As in Theorem 2, the constraint on the neural network for achieving this robustness becomes more permissive when the noise distribution has a heavier tail. Figure 2 demonstrates TATEE's enhanced robustness where treatment and control groups have distinct topologies (line vs. circle). While CFR enforces distributional similarity but allows topological divergence, TATEE enforces distributional and topological similarities. Notably, noise considerably impacts CFR's ability to enforce distributional similarity, while TATEE achieves its objective in noisy and clean environments.

5. Experimental Results

We evaluate TATEE's capability to enhance robustness in neural networks for causal effect estimation, as indicated by our theoretical analysis. Our experiments across standard causal inference benchmarks confirm that incorporating topological awareness consistently improves robustness to

input noise compared to CFR and other deep learning methods for treatment effect estimation. Experimental setup details, results discussion, and additional experiments are in Appendix D.

Experimental Setup. We evaluate TATEE's robustness to various input noises sampled from the family of α-stable distributions, characterized with tail and skewness parameters—α and β. We use standard causal inference benchmarks (IHDP, Twins, Jobs, and ACIC) and measure both performance and robustness. The performance is evaluated via CATE estimation error, which is quantified by the *Precision in Estimation of Heterogeneous Effect* (PEHE), denoted ϵ_{PEHE}, following the conventions in causal inference (e.g., Hill [25], Louizos et al. [35], Shi et al. [47]). We evaluate robustness using $\rho = 1 - (\epsilon_{\text{PEHE}} \text{ with noise}/\epsilon_{\text{PEHE}} \text{ without noise})$, quantifying resistance to performance degradation under noise. Complete details, an elaborate discussion of these results, and additional results are provided in Appendix D.

Main Results. Our experiments confirm that TATEE achieves superior robustness to input noise while maintaining comparable performance without noise. Figure 3 shows this for our main comparison against TATEE's counterpart without topological awareness—the original CFR—with α-stable noise distributions across 25 valid (α, β) values on four datasets. The gains are most pronounced with heavier-tailed noise (α closer to 1), reaching near 12, 0.3, 51.0, and 68.6 percents for the IHDP, Twins, Jobs, and ACIC datasets (respectively). Moreover, pairwise comparisons against seven other causal inference methods further confirm TATEE's superior robustness across various noise conditions and datasets. This is demonstrated via the matrix visualized in Figure 4, whose (i, j) entry shows the proportion of dataset-parameter pairs in which the model corresponding to row i is more robust than the one for column j, evaluated by having a weakly larger ρ. Complete performance metrics and detailed comparisons are provided in Appendix D.

6. Discussion

We proposed incorporating topological awareness into representation learning to enhance robustness against input noise, particularly with heavy-tailed distributions. By leveraging the stability properties of persistence diagrams, we showed that using topological summaries can improve representation stability in a scalable fashion without requiring ground-truth or validation data—critical advantages for observational settings. This concept is rooted in foundational stability theorems from TDA and extended through our work to neural networks and the TATEE framework we introduced for treatment effect estimation. Our theoretical analysis establishes conditions for topological awareness to enhance metric stability, especially with heavy-tailed noise, and our experiments validate that TATEE meets these conditions in practice, consistently outperforming existing methods across noise regimes. While we demonstrated our arguments through a causal inference framework, the theoretical results underpinning TATEE have broader implications for robust deep learning. Future work can explore additional topological features and investigate topology-aware methods across a wider range of representation learning problems.

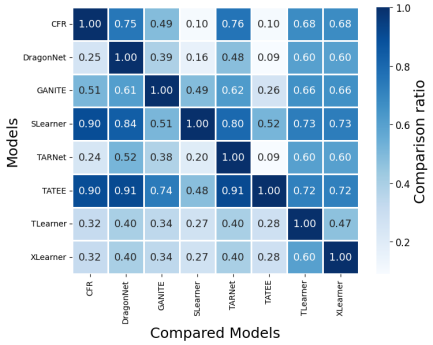

Figure 4: Comparing the robustness of TATEE with 7 other methods. Each entry shows the proportion of cases (over 25 (α, β) values for 4 datasets) where the row model is at least as robust as the column model. TATEE has the largest row average.

Acknowledgments

Work partially supported by NSF, ONR, and the Simons Foundation.

References

[1] Susan Athey and Stefan Wager. Estimating treatment effects with causal forests: An application. Observational Studies, 5(2):37–51, 2019.

[2] Braham Barkat and LJubiša Stanković. Analysis of polynomial fm signals corrupted by heavy-tailed noise. Signal Processing, 84(1):69–75, 2004.

[3] Ulrich Bauer. Ripser: efficient computation of Vietoris-Rips persistence barcodes. Journal of Applied and Computational Topology, 5(3):391–423, 2021.

[4] Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. IEEE Signal Processing Magazine, 34(4):18–42, 2017.

[5] Peter Bubenik et al. Statistical topological data analysis using persistence landscapes. Journal of Machine Learning Research, 16(1):77–102, 2015.

[6] Gunnar Carlsson. Topology and data. Bulletin of the American Mathematical Society, 46(2): 255–308, 2009.

[7] Mathieu Carrière, Marco Cuturi, and Steve Oudot. Sliced wasserstein kernel for persistence diagrams. In International Conference on Machine Learning, pages 664–673. JMLR.org, 2017.

[8] Frédéric Chazal and Bertrand Michel. An introduction to topological data analysis: fundamental and practical aspects for data scientists. Frontiers in Artificial Intelligence, 4:108, 2021.

[9] Frédéric Chazal, Vin de Silva, Marc Glisse, and Steve Oudot. Persistence stability for geometric complexes. Geometriae Dedicata, 173(1):193–214, 2014.

[10] David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. In Proceedings of the Twenty-first Annual Symposium on Computational Geometry, pages 263–271, 2005.

[11] David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. Discrete & Computational Geometry, 37(1):103–120, 2007.

[12] David Cohen-Steiner, Herbert Edelsbrunner, John Harer, and Yuriy Mileyko. Lipschitz functions have l p-stable persistence. Foundations of Computational Mathematics, 10(2):127–139, 2010.

[13] Peng Cui, Zheyan Shen, Sheng Li, Liuyi Yao, Yaliang Li, Zhixuan Chu, and Jing Gao. Causal inference meets machine learning. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 3527–3528, 2020.

[14] Rajeev H Dehejia and Sadek Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. Journal of the American Statistical Association, 94(448): 1053–1062, 1999.

[15] Edelsbrunner, Letscher, and Zomorodian. Topological persistence and simplification. Discrete & Computational Geometry, 28:511–533, 2002.

[16] Herbert Edelsbrunner and John Harer. Persistent homology—a survey. Contemporary Mathematics, 453:257–282, 2008.

[17] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouve, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In International Conference on Artificial Intelligence and Statistics, pages 2681–2690. PMLR, 2019.

[18] Rickard B Gabrielsson and Gunnar Carlsson. Topology of deep neural networks. Journal of Machine Learning Research, 20(196):1–40, 2019.

[19] Rickard Brüel Gabrielsson, Bradley J Nelson, Anjan Dwaraknath, and Primoz Skraba. A topology layer for machine learning. In International Conference on Artificial Intelligence and Statistics, pages 1553–1563. PMLR, 2020.

[20] Shantanu Ghosh, Zheng Feng, Jiang Bian, Kevin Butler, and Mattia Prosperi. Dr-vidal-doubly robust variational information-theoretic deep adversarial learning for counterfactual prediction and treatment effect estimation on real world data. In AMIA Annual Symposium Proceedings, volume 2022, page 485, 2023.

[21] Robert Ghrist. Barcodes: The persistent topology of data. Bulletin of the American Mathematical Society, 45(1):61–75, 2008.

[22] B.V. Gnedenko and A.N.A.N. Kolmogorov. Limit Distributions for Sums of Independent Random Variables. Creative Media Partners, LLC, 2021. English Translation from the Russian edition (1954).

[23] Eduard Gorbunov, Marina Danilova, and Alexander Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. In Advances in Neural Information Processing Systems, volume 33, pages 15042–15053, 2020.

[24] Henry Gouk, Eibe Frank, Bernhard Pfahringer, and Michael J Cree. Regularisation of neural networks by enforcing lipschitz continuity. Machine Learning, 110:393–416, 2021.

[25] Jennifer L Hill. Bayesian nonparametric modeling for causal inference. Journal of Computational and Graphical Statistics, 20(1):217–240, 2011.

[26] Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In International Conference on Machine Learning, pages 3020–3029. PMLR, 2016.

[27] Nathan Kallus, Xiaojie Mao, and Madeleine Udell. Causal inference with noisy and missing covariates via matrix factorization. In Advances in Neural Information Processing Systems, volume 31, 2018.

[28] Amirreza Kazemi and Martin Ester. Adversarially balanced representation for continuous treatment effect estimation. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pages 13085–13093, 2024.

[29] Young Shin Kim, Svetlozar T Rachev, Michele Leonardo Bianchi, and Frank J Fabozzi. Financial market models with lévy processes and time-varying volatility. Journal of Banking & Finance, 32(7):1363–1378, 2008.

[30] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In International Conference on Learning Representations, 2015.

[31] Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. Proceedings of the National Academy of Sciences, 116(10):4156–4165, 2019.

[32] Manabu Kuroki and Judea Pearl. Measurement bias and effect restoration in causal inference. Biometrika, 101(2):423–437, 2014.

[33] Kai Lagemann, Christian Lagemann, Bernd Taschler, and Sach Mukherjee. Deep learning of causal structures in high dimensions under data limitations. Nature Machine Intelligence, 5 (11):1306–1316, 2023.

[34] Robert J LaLonde. Evaluating the econometric evaluations of training programs with experimental data. The American Economic Review, pages 604–620, 1986.

[35] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In Advances in Neural Information Processing Systems, volume 30, 2017.

[36] Yunan Luo, Jian Peng, and Jianzhu Ma. When causal inference meets deep learning. Nature Machine Intelligence, 2(8):426–427, 2020.

[37] Francesco Mainardi. Lévy stable distributions in the theory of probability. Lecture Notes on Mathematical Physics, 2007.

[38] Yuriy Mileyko, Sayan Mukherjee, and John Harer. Probability measures on the space of persistence diagrams. Inverse Problems, 27(12):124007, 2011.

[39] Stephen L Morgan and Christopher Winship. Counterfactuals and Causal Inference. Cambridge University Press, 2015.

[40] Theodore Papamarkou, Tolga Birdal, Michael M Bronstein, Gunnar E Carlsson, Justin Curry, Yue Gao, Mustafa Hajij, Roland Kwitt, Pietro Lio, Paolo Di Lorenzo, et al. Position: Topological deep learning is the new frontier for relational learning. In International Conference on Machine Learning. PMLR, 2024.

[41] Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. In Advances in Neural Information Processing Systems, volume 33, pages 857–869, 2020.

[42] Marcin Pitera, Aleksei Chechkin, and Agnieszka Wyłomańska. Goodness-of-fit test for α-stable distribution based on the quantile conditional variance statistics. Statistical Methods & Applications, 31(2):387–424, 2022.

[43] Antti Pöllänen and Pekka Marttinen. Identifiable causal inference with noisy treatment and no side information. arXiv preprint arXiv:2306.10614, 2023.

[44] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. Proceedings of the IEEE, 109(5):612–634, 2021.

[45] Thibault Séjourné, Jean Feydy, François-Xavier Vialard, Alain Trouvé, and Gabriel Peyré. Sinkhorn divergences for unbalanced optimal transport. arXiv preprint arXiv:1910.12958, 2019.

[46] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In International Conference on Machine Learning, pages 3076–3085. PMLR, 2017.

[47] Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. In Advances in Neural Information Processing Systems, volume 32, 2019.

[48] Claudia Shi, Victor Veitch, and David M Blei. Invariant representation learning for treatment effect estimation. In Uncertainty in Artificial Intelligence, pages 1546–1555. PMLR, 2021.

[49] Di Shu and Grace Y Yi. Causal inference with noisy data: Bias analysis and estimation approaches to simultaneously addressing missingness and misclassification in binary outcomes. Statistics in Medicine, 39(4):456–468, 2020.

[50] Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In International Conference on Machine Learning, pages 5827–5837. PMLR, 2019.

[51] Joshua Southern, Jeremy Wayland, Michael M. Bronstein, and Bastian Rieck. Curvature filtrations for graph generative model evaluation. In Advances on Neural Information Processing Systems, 2023.

[52] Stoyan V Stoyanov, Svetlozar T Rachev, Boryana Racheva-Yotova, and Frank J Fabozzi. Fat-tailed models for risk estimation. The Journal of Portfolio Management, 37(2):107–117, 2011.

[53] Christopher Tralie, Nathaniel Saul, and Rann Bar-On. Ripser.py: A lean persistent homology library for python. The Journal of Open Source Software, 3(29):925, Sep 2018.

[54] Aladin Virmaux and Kevin Scaman. Lipschitz regularity of deep neural networks: analysis and efficient estimation. In Advances in Neural Information Processing Systems, volume 31, 2018.

[55] Hao Wang, Jiajun Fan, Zhichao Chen, Haoxuan Li, Weiming Liu, Tianqiao Liu, Quanyu Dai, Yichao Wang, Zhenhua Dong, and Ruiming Tang. Optimal transport for treatment effect estimation. In Advances in Neural Information Processing Systems, volume 36, 2024.

[56] Michael R Wickens. A note on the use of proxy variables. Econometrica: Journal of the Econometric Society, pages 759–761, 1972.

[57] Haibo Yang, Peiwen Qiu, and Jia Liu. Taming fat-tailed ("heavier-tailed" with potentially infinite variance) noise in federated learning. In Advances in Neural Information Processing Systems, volume 35, pages 17017–17029, 2022.

[58] David S Yeager, Carissa Romero, Dave Paunesku, Christopher S Hulleman, Barbara Schneider, Cintia Hinojosa, Hae Yeon Lee, Joseph O'Brien, Kate Flint, Alice Roberts, et al. Using design thinking to improve psychological interventions: The case of the growth mindset during the transition to high school. Journal of Educational Psychology, 108(3):374, 2016.

[59] Jinsung Yoon, James Jordon, and Mihaela Van Der Schaar. Ganite: Estimation of individualized treatment effects using generative adversarial nets. In International Conference on Learning Representations, 2018.

[60] Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. Discrete & Computational Geometry, 33(2):249–274, 2005.

Appendix A. Topological Data Analysis, Persistent Homology, and Stability Theorems

In this appendix we provide further details on topological data analysis (TDA) and in particular the key concept of *persistent homology*, which crucially helps with understanding complex data structures [16]. Before doing so, we provide a brief review of the basics of TDA. While a more concise review of the preliminaries on TDA was provided in Section 2, here we repeat that review with additional elaborations in order to better familiarize interested readers.

A.1. Review of the Preliminaries

Topological Data Analysis (TDA) applies the principles of algebraic topology to extract informative features from data. It is particularly adept at uncovering invariants such as the shape and connectivity of data across multiple scales [6, 40]. Persistent homology, a central tool in TDA, provides a multiscale representation of topological features [16]. Applications of persistent homology span from feature extraction in computer vision to the analysis of complex datasets in machine learning, which benefit from its intrinsic metric and coordinate-free approach [7].

Persistent homology captures the persistence of topological features such as connected components and holes as a scale parameter varies. These features are represented in a persistence diagram, a collection of points in the plane, each point corresponding to a feature's birth and death in a filtration of the input [21]. The construction of a filtration, a series of nested simplicial complexes determined by the image of a filtration function, is the first step in applying persistent homology to a dataset. The persistent homology of this filtration is then computed, yielding a persistence diagram, which, again, it represents the lifespan of topological features as points marking their birth and death [60]. Figure 5 visualizes this in a simple example. We delve into further details on persistent homology in the following subsection.

Resulting from *the stability theorem*, stability of the persistence diagram is a key property that ensures the robustness of these topological summaries to perturbations in the data [11]. This theorem states a bound on the bottleneck distance between two persistence diagrams obtained via filtration functions f and g.[1] Perhaps regarded as the most central theorem underlying the applicability of persistent homology, the stability theorem provides a guarantee that the persistence diagram is stable under small perturbations, making it particularly conducive to analyzing noisy data [9]. A generalization of the stability theorem used in our analysis is stated in Section 2, and further details are discussed in the Appendix section A.3.

A.2. Persistent Homology

Persistent homology, a fundamental tool in TDA, helps in quantifying the topological features of data. This section aims to define and elucidate key concepts related to filtration functions and persistence diagrams, providing a background for discussing the stability of these constructs and the assumptions behind our theoretical results. We begin with filtration, which yields a multi-scale representation of data, essential for understanding the evolution of topological features.

1. The bottleneck distance between two diagrams is the cost of the optimal matching between their points.

A.2.1. FILTRATION

A filtration is a nested collection of subspaces $\{\mathcal{Z}_a\}_{a\in\mathbb{R}}$ of a topological space $\mathcal{Z}$, such that $\mathcal{Z}_a \subseteq \mathcal{Z}_b$ whenever $a \leq b$. This process can be driven by a real-valued function $f : \mathcal{Z} \to \mathbb{R}$, referred to as the *filtration function*, which assigns a real number to each point in $\mathcal{Z}$. Considering a bounded continuous function $f : \mathcal{Z} \to \mathbb{R}$, the sublevel sets $f^{-1}(-\infty, a_i]$ at various thresholds $a_0 \leq a_1 \leq \ldots \leq a_n$ give rise to a filtration. The filtration captures the evolution of the topological structure of $\mathcal{Z}$ as the threshold varies, revealing critical values where topological features appear or disappear.

Given a filtration function, for every pair of threshold values $a \leq b$, the inclusion relationship between their corresponding subspaces, $\mathcal{Z}_a \subset \mathcal{Z}_b$, induces homomorphisms of the l dimensional homology groups $\mathsf{H}_l(\mathcal{Z}_a)$ and $\mathsf{H}_l(\mathcal{Z}_b)$. If there exists a dimension l, a threshold value $c \in \mathbb{R}$, and a value $\delta > 0$, such that for all $\epsilon \in (0, \delta)$ the homomorphism induced by $\mathsf{H}_l(\mathcal{Z}_{c-\epsilon}) \subset \mathsf{H}_l(\mathcal{Z}_{c+\epsilon})$ is not an isomorphism, c is called a *homological critical value*. These critical values mark the levels where the homology of the sublevel sets changes [10].

Definition 4 (Cohen-Steiner et al. [10]) *A filtration function, $f : \mathcal{Z} \to \mathbb{R}$, is tame, if it only has a finite number of homological critical values, and if for all threshold values $a \in \mathbb{R}$ and dimensions l, the homology groups $\mathsf{H}_l\left(f^{-1}(-\infty, a]\right)$ are finite dimensional.*

The filtration function could be any mapping to a meaningful real-valued representation of the data. Studying how topology changes through the filtration gives insight into the structure of data at different scales. The stability theorem assumes the filtration function is tame.

A.2.2. PERSISTENCE DIAGRAM

Given a filtration, the persistence diagram compactly represents the lifespan of homological features through their birth and death thresholds. Homology classes are born at critical threshold values where new features appear in the filtration. Subsequently, some classes die at larger thresholds. The lifespan of a class that is born at threshold a and dies entering threshold b is characterized by the *persistence* value of the corresponding point, x, in the persistence diagram, defined as $\text{pers}(x) := b - a$ [15]. Classes with larger persistence values are considered to be more prominent features. The birth and death of homology groups can be represented in a 2-dimensional *persistence diagram* as follows [11]: Points (a, b) denoting classes born at threshold value a and dying at threshold value b, and points of the form (a, ∞) representing essential homology classes that never die. This low-dimensional representation allows us to easily interpret and analyze the topological features of the data over different scales. Figure 5 visualizes an example of a Vietoris-Rips filtration [6] and the corresponding persistence diagram of 0- and 1-dimensional homology classes. As the figure demonstrates, the persistence diagram shows the birth and death of topological features of the data with the points farther from the diagonal marking more persistent features.

Before discussing key stability properties of persistence diagrams, we state two definitions related to properties of the filtration function. These definitions relate to *Degree-k total persistence* of a persistence diagram corresponding to a filtration function f. Degree-k total persistence, which sets one of the assumptions for the stability theorem, is defined as the sum of the kth powers of the persistence values of all points in the persistence diagram of f.

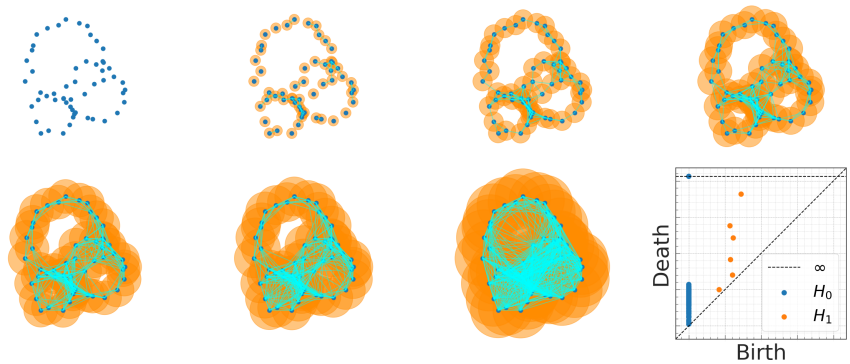

Figure 5: A visualization of the Vietoris-Rips filtration of an example pointcloud and the corresponding persistence diagram.

Definition 5 (Cohen-Steiner et al. [12]) *Given a filtration function $f : \mathcal{Z} \to \mathbb{R}$, let Dgm_f denote the corresponding persistence diagram. The degree-k total persistence is given by*

$$Pers_k(f) = \sum_{x \in Dgm_f} pers(x)^k.$$

Definition 6 (Cohen-Steiner et al. [12]) *We say that a space $\mathcal{Z}$ implies bounded degree-k total persistence, if there exists a constant $C_{\mathcal{Z}}$ that depends only on $\mathcal{Z}$, such that for every tame Lipschitz function $f : \mathcal{Z} \to \mathbb{R}$ with Lipschitz constant K_f, we have*

$$Pers_k(f) \leq C_{\mathcal{Z}} K_f^k. \tag{3}$$

A.3. Stability Theorems

The stability properties of persistence diagrams are crucial, as they imply that the topological signature captured by these diagrams is robust to small perturbations and noise in the data. As such, they serve as the main justification for using persistence diagrams in order to improve the robustness of deep learning frameworks to input noise.

The *stability theorem* —the main theoretical result regarding the stability of persistence diagrams— bounds the bottleneck distance between two diagrams as follows.

Theorem 7 (Cohen-Steiner et al. [10]) *Let $\mathcal{Z}$ be a triangulable space with continuous tame filtration functions $f, g : \mathcal{Z} \to \mathbb{R}$. Then the corresponding persistence diagrams satisfy*

$$d_B(\mu_f, \mu_g) \leq \|f - g\|_\infty,$$

where d_B is the bottleneck distance, and μ_f and μ_g are the probability measures induced by Dgm_f and Dgm_g.

Stating that the distance between the persistence diagrams is controlled by the L_∞ distance between the corresponding filtration functions, this theorem ultimately suggests that the persistence diagram is more stable than the geometry of the data it represents.

In a generalization of this statement, *the Wasserstein Stability Theorem*, extends this result to the Wasserstein-p distance between the diagrams for $p \geq k \geq 1$, when $\mathcal{Z}$ implies bounded degree-k total persistence.

Theorem 8 (Cohen-Steiner et al. [12]) *Let $\mathcal{Z}$ be a triangulable, compact metric space that implies bounded degree-k total persistence, for some $k \geq 1$. Let $f, g : \mathcal{Z} \to \mathbb{R}$ be two tame Lipschitz filtration functions with Lipschitz constants K_f and K_g. Then for all l-dimensional homology classes and all $p \geq k$ we have*

$$W_p\left(\mu_f^l, \mu_g^l\right) \leq C^{\frac{1}{p}} \|f - g\|_\infty^{1-\frac{k}{p}}.$$

The constant C is given by $C = C_{\mathcal{Z}} \max\{K_f^k, K_g^k\}$, where $C_{\mathcal{Z}}$ is the constant in Equation 3.

Note that as $p \to \infty$, this generalized formulation gives the statement in Theorem 7. To conclude, these stability results show persistence diagrams are robust topological summaries for analyzing complex data.

Appendix B. Learning Robust Representations via Persistence Diagrams

Here we provide additional details regarding the theoretical results discussed in Section 3. In particular, while providing the proofs and the background leading to the theorems, we detail the conditions and elaborate on the formulations of the bounds stated in the theorems. Theorem 2 offer insight into the conditions under which using persistence diagrams corresponds to enhanced robustness of the representations, depending on the properties of the neural network. In particular, this theorem constrains the Lipschitz constant of the neural network by a term that contains information on the empirical distribution of the noise, including the sample average and the infinity norm of the sample noise. Importantly, the combination of sample average and infinity norm establishes a connection with the tail of the noise distribution, enabling our analysis of TATEE's robustness through Proposition 9 in Section C.5, which in turn provides theoretical grounds for TATEE's advantage, and validated in empirical settings by the experimental results observed in Section 5. Below, we provide the proof and additional details about the bounds in Theorem 2.

Notation. Following the notation in Section 3, X, E, and $\tilde{X} := X + E$ denote the random vectors of features, noise, and noise-corrupted features on a sample space Ω; and we use $\mathbf{X}$, $\tilde{\mathbf{X}}$, and $\mathbf{E}$ to refer to their corresponding sample matrices. $\varphi : \Omega \to \mathcal{Z}$ is the neural network mapping X and $\tilde{X}$ to $\varphi(X)$ and $\varphi(\tilde{X})$ in the representation space $\mathcal{Z}$, μ and $\tilde{\mu}$ are the measures induced by these representations, and $f, \tilde{f} : \mathcal{Z} \to \mathbb{R}$ are tame and Lipschitz filtration functions used to compute the persistence diagrams inducing the measures μ_f and $\mu_{\tilde{f}}$ on the space of persistence diagrams. For ease of notation, we denote the distances $W_p(\mu, \tilde{\mu})$ and $W_p(\mu_f, \mu_{\tilde{f}})$ by Δ and Δ_{topo}, and their corresponding sample equivalents by $\hat{\Delta}$ and $\hat{\Delta}_{topo}$.

Main Assumptions. The main assumptions for Theorem 2 are related to Lipschitz continuity of the functions involved and the bounds of their co-domains. In particular, following the standard assumptions for stability theorems, we assume the filtration functions are Lipschitz. We also assume that $\mathcal{Z}$ implies bounded degree-k total persistence, as defined in Appendix A; another assumption that is made for the stability theorems. Additionally, we assume the neural network φ is Lipschitz, which, as we discussed in Section 3, is not a restrictive assumption and is satisfied in many standard scenarios. The restrictive assumption for the proofs of our theorem, which is only technical and for simplicity, is due to the formulation of $\tilde{f}$ and its Lipschitz continuity. Recall that we consider f and $\tilde{f}$ which satisfy $\tilde{f}(\varphi(X)) = f(\varphi(\tilde{X}))$. For simplicity and to avoid auxiliary constructions, we require $\tilde{f} = f \circ \varphi \circ S \circ \varphi^{-1}$, where $S(X) := X + E$, so that $\tilde{f}(\varphi(X)) = f(\varphi(\tilde{X}))$ holds exactly.

We remind the reader that these assumptions are technical and serve to enable the derivation of the proofs through simple expressions of the quantities of interest. With that in mind, for clarity, we discuss the assumption that allows $\tilde{f} = f \circ \varphi \circ S \circ \varphi^{-1}$ to hold exactly. Strictly speaking, this is restrictive since it requires $\tilde{f}$ to be Lipschitz continuous and φ to have a Lipschitz inverse, which is typically violated by commonly-used neural networks. However, the specific formulation of $\tilde{f}$ that requires this assumption is only to simplify the derivation and expression of the bounds on $\|f - \tilde{f}\|_\infty$. That is, the assumption allows us to avoid tedious constructions and keep the formulation of the variables used and the steps of the proof simple, concise, clear, and focused, and otherwise could be replaced with a less restrictive assumptions, to derive effectively similar results. While this relaxation is not trivial to the authors' knowledge and hence beyond the scope of this discussion, one such relaxation could be done by assuming that $\phi(.)$ is only locally invertible on the input sample and adding an approximation error term with the $\tilde{f}$ that is defined for the local Lipschitz pseudo-inverse. Moreover, our experimental results indicate that while the assumptions facilitate the steps for our theoretical analysis in a simplified setting, the overall intuition and implications hold more broadly in practical scenarios. With this in mind, we state the theorem, which provide significant insights, validated and confirmed by our experiments.

B.1. Finite Sample Stability and Error Distribution

Let us denote the empirical estimators of the Wasserstein distance of $W_p(\mu, \tilde{\mu})$ by $\hat{\Delta}$, and its upper bound by $\hat{\mathsf{M}}$, as defined in Section 3. We denote the upper bound on $C^{\frac{1}{p}} \left\| f - \tilde{f} \right\|_\infty^{1-\frac{k}{p}}$ by K_{topo}, which bounds $W_p(\mu_f, \mu_{\tilde{f}})$, as stated in Section 3. Furthermore, let us denote its finite sample equivalent from Theorem 1 by $\hat{\mathsf{K}}_{topo}$. Under the assumptions discussed above, Theorem 2 states the following: Given a bounded degree-k total persistence and a K_f-Lipschitz filtration function, if the Lipschitz constant of the neural network is smaller than Λ for a value Λ, then $\hat{\mathsf{K}}_{topo} < \hat{\mathsf{M}}$. Importantly, the value Λ, described below, is linear in $\|\mathbf{E}\|_\infty^{p/k-1} \|\bar{\mathbf{E}}\|^{-p/k}$, which speaks to the impact of the empirical distribution of noise on the gain in robustness through use of persistence diagrams. The proof of this theorem is presented next.

Proof. The upper bound $\hat{\mathsf{M}}$ can be derived using the Lipschitz continuity of φ. In the finite sample regime, this bound becomes

$$\hat{\Delta} \leq \hat{\mathsf{M}} = K_\varphi \hat{\epsilon}, \tag{4}$$

where $\hat{\epsilon}$ is the sample average. For $\mathsf{B}_{topo} = C^{\frac{1}{p}} \left\| f \circ \varphi - \tilde{f} \circ \varphi \right\|_\infty^{1-\frac{k}{p}}$, using Lipschitz continuity of f and φ, we can derive the following finite sample bound,

$$\hat{\Delta}_{topo} \leq \hat{\mathsf{B}}_{topo} \leq \hat{\mathsf{K}}_{topo} = C^{\frac{1}{p}} \left(K_f K_\varphi \bar{\epsilon}\right)^{1-\frac{k}{p}}, \tag{5}$$

where, for ease of notation, we use $\bar{\epsilon} = \|\mathbf{E}\|_\infty$ to denote the finite sample infinity norm of the noise matrix. Additionally, note that $C = C_{\mathcal{Z}} L$, where $L := \max\left\{K_f^k, K_{\tilde{f}}^k\right\}$, as explained in Appendix Section A. Comparing the right hand sides of the inequalities 4 and 5, it follows that if

$$K_\varphi < C^{1/k} K_f{}^{\frac{p}{k}-1} \frac{\bar{\epsilon}^{\frac{p}{k}-1}}{\hat{\epsilon}^{\frac{p}{k}}},$$

then, $\hat{\mathsf{K}}_{topo} < \hat{\mathsf{M}}$, which completes the proof of Theorem 2. □

Appendix C. Topology-Aware Treatment Effect Estimation

Building on the stability results established in Section 3, in Section 4 we introduced ***T**opology-**A**ware **T**reatment **E**ffect **E**stimation* (TATEE)—a framework designed to improve the robustness of deep learning approaches to counterfactual regression. TATEE incorporates topological regularization by embedding the persistence diagram of the learned representations into the training objective. We briefly described the main aspects of TATEE's design, analysis, and implications in Section 4. In this aspect, we elaborate and expand on the discussion in Section 4, delving into details of the structure of the CATE estimation network used in TATEE, along with its training procedure, scalability, and conceptual motivation and implications. We further provide a detailed theoretical analysis of TATEE's robustness properties formally stating results mentioned in the informal Proposition 3, which connects the stability benefits of topological summaries to improved treatment effect estimation. Note that, as in prior works [35, 46, 47], we focus on the binary treatment setting for clarity, though TATEE is in principle extensible to multi-valued treatments.

C.1. Representation-Balancing Neural Network for CFR

TATEE builds on the two-headed architecture introduced in the CFR framework Shalit et al. [46], consisting of a shared encoder followed by separate branches for the treatment and control groups. The shared encoder φ maps inputs to a representation space where the distributions of treated and control units are approximately aligned—enabling the learning of *treatment-agnostic* representations. These shared representations are then passed to two distinct heads, h_0 and h_1, which learn the potential outcomes under control and treatment, respectively. This structure corresponds to a T-Learner [31] —the potential outcome heads for each group— to learn the factual and counterfactual outcomes separately for estimating CATE. Figure 1 visualizes this architecture, showing the shared neural network, φ, and two separate heads h_0 and h_1 which learn the potential outcomes from φ's output. As we detail in Section C.3, the weights are trained with respect to the outcome, the distribution of the representations, as well as the topological summaries of the representations as captured by the persistence diagram of the $1-$dimensional homology class.

C.2. Implementation

The architecture of the neural network in TATEE is described in Section 4, where we explain the role of each component of the model in the CFR-type architecture shown in Figure 1. We also specified the training objective in Section C.3, which, critically, incorporates topological awareness into the CFR framework and is the core distinguishing component of TATEE. In this appendix, we include the details of the implementation of the neural network and state the full algorithm.

The skeleton of the neural network in TATEE follows that of CFR, described in Shalit et al. [46]. As we mention in Appendix G, adopting the hyper-parameters used by Shalit et al. [46], we use three fully connected layers with ELU (for *exponential linear unit*) activation functions for all three components —φ, h_0, and h_1. Each shared representation layer has 200 neurons, while the layers in h_0 and h_1 have 100 neurons each. Since the outcome in the Twins dataset takes binary values, we use a sigmoid activation function on the final layer, this affecting the loss function used as $\mathcal{L}_{Outcome}$ in Equation 6, which is binary cross entropy for Twins dataset and mean squared error for the others. The central term in the loss function, $\mathcal{L}_{topo}$, uses the Wasserstein-2 distance between the persistence landscapes of the representations of a mini-batch of size 256 for the IHDP dataset, and size 128 for the others. We use the Vietoris-Rips complex [6] for computing the persistence diagrams, which

are obtained using the Ripser package [3, 53]. The distance between the persistence landscapes for the 1-dimensional homology class are then approximated using the Sinkhorn divergence [45] —an efficient and differentiable approximation of the Wasserstein distance which is amenable to gradient descent for training the neural network. The GeomLoss package [17] is used for computing the Sinkhorn divergence. The full algorithm is stated in Algorithm 1, which clarifies how the topological signature, as described in Section 4, is incorporated in the CFR framework to obtain TATEE. Note that we used a fixed number of epochs of training, hence, in our implementation, the convergence criterion in Algorithm 1 is simply completing the specified number of epochs.

Algorithm 1 TATEE Training

Input: Neural network composed of the components φ and $h(\cdot, t_i) := (1 - t_i)h_0(\cdot) + t_i h_1(\cdot)$ with initial weights θ_φ and θ_h, sample data $(x_1, t_1, y_1), \ldots, (x_N, t_N, y_N)$, regularization parameters $\lambda, \lambda_{topo} > 0$, and loss function $\mathcal{L}$.
Compute $N_1 = \sum_{i=1}^{N} t_i$ and $N_0 = N - N_1$.
Compute sample weights $w_i = \frac{N t_i}{2N_1} + \frac{N(1-t_i)}{2N_0}$ for $i = 1 \ldots n$.
while not converged **do**
  Take mini-batch $I_M := \{i_1, i_2, \ldots, i_M\} \subseteq \{1, 2, \ldots, N\}$
  Compute representations $\Phi_0 := \{\varphi(x_j) : t_j = 0, j \in I_M\}$ and $\Phi_1 := \{\varphi(x_j) : t_j = 1, j \in I_M\}$
  Compute the predicted outcomes $\{\hat{y}_j = h(\varphi(x_j), t_j) : j \in I_M\}$
  Compute the persistence diagrams $\mathrm{Dgm}_1(\Phi_0)$ and $\mathrm{Dgm}_1(\Phi_1)$
  Compute the gradient of the empirical $\mathcal{L}_{\text{Balancing}}$ as $\delta_{\text{Balancing}} = \nabla_{\theta_\varphi} W_p(\Phi_0, \Phi_1)$
  Compute the gradient of the empirical $\mathcal{L}_{\text{topo}}$ as $\delta_{\text{topo}} = \nabla_{\theta_\varphi} W_p(\mathrm{Dgm}_1(\Phi_0), \mathrm{Dgm}_1(\Phi_1))$
  Compute the gradients of the empirical $\mathcal{L}_{\text{Outcome}}$ as
  $\delta_{\varphi,\text{Outcome}} = \nabla_{\theta_\varphi} \frac{1}{M} \sum_{j \in I_M} w_j \mathcal{L}(y_j, \hat{y}_j)$
  $\delta_{h,\text{Outcome}} = \nabla_{\theta_h} \frac{1}{M} \sum_{j \in I_M} w_j \mathcal{L}(y_j, \hat{y}_j)$
  Determine the step size η using Adam
  Update weights
  $\theta_\varphi \leftarrow \theta_\varphi - \eta \left[\delta_{\varphi,\text{Outcome}} + \lambda \delta_{\text{Balancing}} + \lambda_{\text{topo}} \delta_{\text{topo}}\right]$
  $\theta_h \leftarrow \theta_h - \eta(\delta_{h,\text{Outcome}})$
  Check for convergence
end while

C.3. Training TATEE

The effectiveness of TATEE in enhancing the robustness of learning treatment-agnostic representations is primarily owed to incorporating the topology of the shared representations in the training process. This is achieved by adding a topological regularization term based on the Wasserstein distance between the persistence diagrams of the treatment and control groups. The CFR network architecture in Shalit et al. [46] is predicated on the minimization of a loss function that encapsulates both the prediction accuracy and the distributional balance between treated and control groups. Incorporating the topological regularization, for a sample of size N, TATEE's training objective is as follows:

$$\mathcal{L}_{\text{TATEE}} = \mathcal{L}_{\text{Outcome}} + \lambda \mathcal{L}_{\text{Balance}} + \lambda_{\text{topo}} \mathcal{L}_{\text{topo}}, \tag{6}$$

The three components of $\mathcal{L}_{\text{TATEE}}$ are given by:

$$\mathcal{L}_{\text{Outcome}} = \frac{1}{N}\sum_{i=1}^{N} w_i \ell\big(\hat{y}_i, y_i\big),\ \mathcal{L}_{\text{Balance}} = W_p\big(\mathbf{\Phi}_0, \mathbf{\Phi}_1\big),\ \mathcal{L}_{\text{topo}} = W_p\big(\text{Dgm}_1\big(\mathbf{\Phi}_0\big), \text{Dgm}_1\big(\mathbf{\Phi}_1\big)\big),$$

where $\mathbf{\Phi}_0$ and $\mathbf{\Phi}_1$ are the representations of the control and treated groups, respectively; y_i the true outcome of unit i, $\hat{y}_i \coloneqq h\big(\varphi(\mathbf{x}_i), t_i\big)$ the predicted outcome with the feature vector $\mathbf{x}_i$ and treatment $\mathbf{t}_i$, $\ell(\cdot)$ the outcome prediction loss function, $\text{Dgm}_1(\cdot)$ the 1-homology class persistence diagram, and $W_p(\cdot,\cdot)$ the Wasserstein-p distance. The weight w_i aims to deal with the imbalance in the size of the treatment and control groups, and is given by $\frac{Nt_i}{2N_1} + \frac{N(1-t_i)}{2N_0}$ where N_1 and N_0 are the sample sizes of the two groups. The function $h(\cdot, t_i) \coloneqq (1-t_i)h_0(\cdot) + t_i h_1(\cdot)$ combines the two potential outcome functions. While our theoretical analysis in Equation (1) holds for any homology class, we use the first homology group in practice. This choice is motivated by simplicity and the goal of capturing holes, thereby characterizing topological features beyond connected components. We include the full algorithm in Appendix C.2, describing the implementation of TATEE in details.

C.4. Scalability of TATEE

TATEE incorporates topological summaries in a computationally scalable fashion, due to two key factors in our implementation: First, topological regularization operates on representations rather than raw inputs, making computational cost constant with respect to input dimensionality, once the dimensionality of the representations are fixed. Second, we compute persistence diagrams on mini-batches, hence, with a batch size of b and for N total data points, the cost of computing $\mathcal{L}_{\text{topo}}$ for each of the N/b batches remains fixed, allowing the overhead to scale linearly with data volume. As a result, despite topological methods typically being expensive for high-dimensional or large inputs, TATEE's implementation ensures the topological component does not become a computational bottleneck. Figure 6 illustrates this scalability via a simulation, confirming that overhead elapsed time remains nearly constant as input dimensionality increases from 16 to 256 (with a fixed representation dimensionality), and scales linearly as data volume varies from 800 to 12800 samples. These results further support the fact that the topological component introduces a scalable computational overhead while providing significant robustness benefits, making TATEE practical for real-world applications.

C.5. Robustness of TATEE's Training to Noise

Using the stability results presented before, we now show that under the conditions of Theorem 2, TATEE can improve the robustness of counterfactual regression to input noise. In particular, we show that Theorem 2 implies that TATEE's training objective in Equation 6 is more stable than the original CFR's objective. Consider the problem setup and notation in Section 3, and let us label the variables corresponding to the treatment and control groups by superscripts $\cdot^0$ and $\cdot^1$, respectively. By the triangle inequality, the terms $\lambda\big(\hat{\mathsf{M}}^1 + \hat{\mathsf{M}}^0\big) + \lambda_{\text{topo}}\big(\hat{\mathsf{K}}^1_{topo} + \hat{\mathsf{K}}^0_{topo}\big)$ and $\lambda_{\text{CFR}}\big(\hat{\mathsf{M}}^1 + \hat{\mathsf{M}}^0\big)$ upper-bound the noise-induced change in $\lambda\mathcal{L}_{\text{Balance}} + \lambda_{\text{topo}}\mathcal{L}_{\text{topo}}$ and $\lambda_{\text{CFR}}\mathcal{L}_{\text{Balance}}$, for loss coefficients λ, λ_{topo}, and λ_{CFR}. Using Theorem 2, we derive the following result (proof in Appendix E).

Proposition 9 *If the Lipschitz constant of φ satisfies the constraint in Theorem 2, for any given $\lambda_{CFR} > 0$, with sufficiently small choices of λ and λ_{topo}, we have*

$$\lambda\big(\hat{\mathsf{M}}^1 + \hat{\mathsf{M}}^0\big) + \lambda_{topo}\big(\hat{\mathsf{K}}^1_{topo} + \hat{\mathsf{K}}^0_{topo}\big) \leq \lambda_{CFR}\big(\hat{\mathsf{M}}^1 + \hat{\mathsf{M}}^0\big).$$

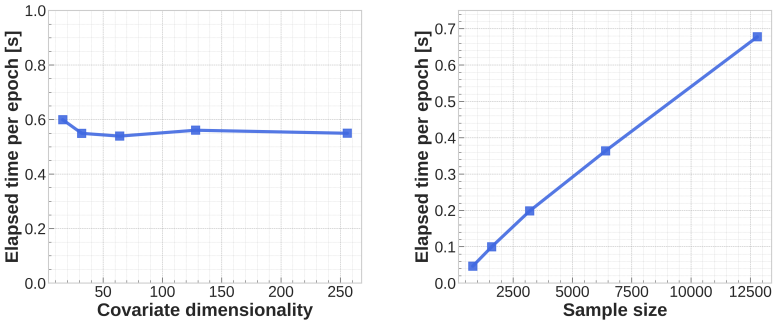

Figure 6: The average overhead elapsed time for computing $\mathcal{L}_{\text{topo}}$ per epoch of training TATEE. As the dimensionality of the input (left) and input volume (right) increases, the overhead cost remains constant and scales linearly, respectively. The elapsed times are averaged over 20 runs.

The inequality in Proposition 9 indicates that the upper bound on the sum of the balancing and topological terms in TATEE's loss undergoes a smaller change due to input noise, compared to that on the balancing term in CFR's loss. In other words, this proposition implies that, when φ in TATEE's architecture (Figure 1) satisfies the conditions of Theorem 2, TATEE is trained using a loss that can be more stable under additive noise than its counterpart CFR. As in Theorem 2, this robustness is easier to attain when the noise is heavy-tailed—i.e., the constraint on φ's Lipschitz constant becomes more permissive. This highlights TATEE's particular suitability for robustness under heavy-tailed perturbations. Our experiments in Section 5 provide empirical evidence that TATEE can meet the theory-indicated conditions and gain the anticipated robustness in practical settings, outperforming the original CFR from Shalit et al. [46] and other standard causal inference baselines.

C.6. TATEE in Action

According to our theoretical results, the topological regularization term in Equation (6) could help train φ such that the total loss becomes more robust to noise. To complement the theoretical analysis, we examine the intuition behind this improvement through a simple example of how the network learns the representations. As we explained earlier, the CFR framework is based on the principle of balancing the representations of treated and control groups through enforcing a distributional similarity between the two. TDA on the other hand, characterizes data using topological features which are invariant to smooth deformations. This allows TDA-based methods to capture qualitative properties of the shape of the data at a global level, while limiting their sensitivity to local geometric perturbations, hence leading to the resulting robustness. This understanding provides an explanation as to why informing a representation balancing framework with the topology of the representations could enhance robustness.

To see how this works in a simple example, we simulate features such that the 2-dimensional representations corresponding to the treated and control groups have distinct topologies: one forming a line and the other a circle—a simple difference in their 1-dimensional homology class. Figure 2 compares how φ learns to balance these representations in TATEE versus standard CFR, with and without noise. Visually confirming our understanding, CFR forces distributional similarity between groups while topological differences persist, even widening after 10 epochs of training without

topological awareness. TATEE, however, enforces both distributional and topological similarities through the distance between the persistence diagrams, resulting in representations that converge to similar shapes with matching topologies. In other words, while the Wasserstein loss of $\mathcal{L}_{\text{Balance}}$ may allow the representations to qualitatively diverge, Figure 2 illustrates that the topological signature captured by $\mathcal{L}_{\text{topo}}$ effectively prevents that. Notably, noise considerably impacts CFR's ability to enforce distributional similarity, while TATEE achieves its objective equally well in both noisy and clean settings.

Appendix D. Experiments

Section 5 discusses our main experimental results, showing TATEE's superior robustness. In this appendix, we provide comprehensive details on our experimental setup, elaborate on our main results, and present additional results that complement the findings described in the main paper. We begin by describing the experimental setup, including datasets, evaluation metrics, noise distributions, and implementation. We then review our main experimental results on robustness of TATEE in more detail, followed by a report on its performance. We also elaborate on the distinction between evaluating robustness and performance on noisy inputs and comment on an expected tradeoff therein. We then discuss our experiments benchmarking TATEE against other treatment effect estimation methods, showing its superior robustness. Additional details on the implementation of these experiments including hyperparameter values, and a more detailed description of each benchmark dataset are provided in appendices G and H.

D.1. Experimental Setup

Models and Evaluation Metrics. We implement TATEE with the network architecture and training algorithm described in Section 4. In this implementation, we use the Wasserstein-2 distance in the balancing and topological terms in the training objective. We also use persistence landscapes [5], which maintain a one-to-one correspondence with persistence diagrams while offering differentiability and better statistical tractability. Other implementation details and hyperparameters are provided in Appendices C.2 and G. Since TATEE incorporates a topological signature in CFR [46], to evaluate the resulting improvement in robustness, we compare their performances with and without input noise. We complement our experiments by several other causal inference models included in Figure 4 and listed in Appendix G. Following the conventions in causal inference (e.g., Hill [25], Louizos et al. [35], Shalit et al. [46], Shi et al. [47]), we use the Precision in Estimation of Heterogeneous Effect (PEHE) to quantify the CATE estimation error. The empirical PEHE is given by $\hat{\epsilon}_{\text{PEHE}} = \frac{1}{N}\sum_{i=1}^{N}\big(\tau(x_i) - \hat{\tau}(x_i)\big)^2$, where $\tau(x_i)$ is the true CATE from Equation (2) and $\hat{\tau}$ is the estimated CATE. While PEHE can be used on noisy input to assess the performance in noisy regimes, we are primarily interested in judging TATEE's robustness. To evaluate the robustness of the models to noise, we use the following metric, denoted by ρ,

$$\rho = 1 - \frac{\tilde{\epsilon}_{\text{PEHE}}}{\hat{\epsilon}_{\text{PEHE}}}. \tag{7}$$

ρ compares PEHE on noisy and noise-free data, which we denote here by $\tilde{\epsilon}_{\text{PEHE}}$ and $\hat{\epsilon}_{\text{PEHE}}$. Unless the noise itself helps the training, ρ should take negative values; more negative when less robust. In order to quantify the gain in robustness from TATEE, we compare the increase in ρ by computing $\rho_{\text{TATEE}} - \rho_{\text{CFR}}$, which takes positive values if TATEE is more robust than CFR. Given the

main objective of our work, our evaluation of the experiments is primarily focused on robustness, while we also inspect the performance of TATEE on both clean and noisy datasets to make sure the enhanced robustness does not cost a considerable decline in performance. Note the distinction between the robustness and performance metrics: a larger (worse) $\tilde{\epsilon}_{\text{PEHE}}$ can still yield a higher (better) ρ, when the noise-induced deterioration of performance, captured by $\frac{\tilde{\epsilon}_{\text{PEHE}}}{\hat{\epsilon}_{\text{PEHE}}}$, is lower. In other words, $\rho_{\text{TATEE}} - \rho_{\text{CFR}}$ can be positive, reflecting better robustness, even if both $\tilde{\epsilon}_{\text{PEHE}}$ and ϵ_{PEHE} are larger for TATEE, as long as the increase from $\hat{\epsilon}_{\text{PEHE}}$ to $\tilde{\epsilon}_{\text{PEHE}}$ remains smaller.

Data. We use four standard benchmark datasets in causal inference: IHDP, Twins, Jobs, and ACIC [1, 25, 35, 46]. The Twins dataset has the rare quality of having real-world values for both potential outcomes. The other datasets are semi-synthetic, with empirical values for features and simulated potential outcomes. All datasets are described in details in Appendix H. Note that using semi-synthetic data in experiments on estimating CATE is standard, and inevitable due to the *fundamental problem of causal inference*. This also dictates the use of synthetic noise as the only means to introduce noise to the treatment effect estimates. The Twins dataset stands out in this regard, allowing us to evaluate TATEE on intrinsic noise in empirical measurements.

Noise Distribution. We use the family of stable distributions [37] to simulate input noise for the feature matrix. This allows us to evaluate the robustness of TATEE in learning from features corrupted with various noise distributions of interest in empirical contexts [23, 29, 52, 57]. In particular, $\alpha-$stable distributions are characterized by a tail parameter α and an asymmetry/skewness parameter β. Larger values of α correspond to slower decay of the tail of the distribution, and positive/negative values of β correspond to positive/negative asymmetry. The valid parameter values fall within the *Feller-Takayasu diamond*, where $\alpha \in (0, 2]$ and $|\beta| \leq \min\{2 - \alpha, \alpha\}$ [37], giving the Gaussian distributions at $\alpha = 2$ and $\beta = 0$. Appendix F contains a more detailed review of stable distributions.

Noise Configuration. For $\alpha < 2$, the noise follows a heavy-tailed distribution with infinite moments of integer order higher than 1, and for $\alpha \leq 1$, the mean is infinite. We therefore focus on noise distributions with finite mean, as well as $\alpha = 1$, which corresponds to the Cauchy distribution for $\beta = 0$. Varying the two parameters of this family of distributions allows us to empirically test the theoretical implications of Theorem 2. Here we report the average results over 30 trials. Since the Central Limit Theorem in its classical formulation does not hold for $\alpha < 2$, the standard sample size criteria for statistical confidence does not apply, as discussed in Appendix F.

D.2. Main Results on Enhanced Robustness

Our experiments showcase the utility of TATEE and speak to its superior robustness, validating the applicability of our theoretical analysis in empirical settings. Reviewing our main empirical results from Section 5 with more details, we begin by confirming that TATEE has a performance comparable with CFR in the absence of synthetic noise, and then turn our attention to the main focus of this paper, showing the superior robustness of TATEE.

Performance without Noise. TATEE is consistently on-par with the original CFR in the noise-free regime, with 1.977×10^{-2}, 4.976×10^{-4}, and 2.631×10^{-3} increase in ϵ_{PEHE} for IHDP, Jobs, and ACIC datasets (respectively), and 1.007×10^{-2} decrease for Twins. Note that the theory-informed advantage of TATEE is its robustness. In fact, adding $\mathcal{L}_{topo}$ in the loss is expected to lead to a more challenging path for optimizing the remaining terms, and hence, an increase in ϵ_{PEHE} is

not unexpected. Nevertheless, the evaluation mentioned above indicates that TATEE's superior robustness does not come at the cost of a noticeable decline in performance. Additional experiments in Appendix D.3 indicate that TATEE also maintains a comparable or better performance on noisy data in most cases, with up to 6.9% improvement on IHDP, and 1.37%, 45.1%, and 42.3% on the Twins, Jobs, and ACIC datasets, as measured by reduction in PEHE. As clarified in our discussion on the evaluation metrics following Equation (7), a better performance in the noisy regime could still correspond to a lower robustness, if the noise-driven degradation is larger. While due to the semi-synthetic nature of all but one dataset we can only experiment with synthetic noise, the Twins dataset contains empirical treatment effects, which are computed under the assumption that a pair of twins have the same covariates. This assumption, as well as empirical measurement noise in the features, are likely to lead to real intrinsic noise in the Twins dataset, even before injecting synthetic noise. Notably, on this dataset TATEE achieves the lowest (best) $\hat{\epsilon}_{\text{PEHE}}$ than all other benchmarks without simulated noise (see Appendix D.4) and outperforms CFR in all cases, including performance without synthetic noise, as well as both performance and robustness with simulated noise for all noise distribution parameters.

Robustness to Noise. Sampling additive feature noise from $\alpha-$stable distributions for 25 (α, β) pairs in the right half of the Feller-Takayasu diamond, we compute the relative gain in ρ on the test set after training on noisy features. The results (Figure 3) strongly confirm that TATEE is more robust than CFR. Aligned with the theoretical discussion following Theorem 2, the experiments also show that the largest values of gain in robustness (larger $\rho_{\text{TATEE}} - \rho_{\text{CFR}}$) are observed when α is closer to 1, corresponding to a heavier tail of the noise distribution, reaching near 12, 0.3, 51.0, and 68.6 percents of gain for the IHDP, Twins, Jobs, and ACIC datasets (respectively), when $\alpha > 1$. [2]

D.3. Performance on Clean and Noisy Data

While TATEE aims to improve the robustness of treatment effect estimation, our experiments confirm that TATEE also has a performance comparable with CFR both in the absence of noise and on noisy inputs. Here we report the experimental results on the performance of TATEE compared to CFR, complementing the discussion on robustness. The main performance metric measuring CATE estimation error is PEHE, which we denote by ϵ_{PEHE} for noise-free input and by $\tilde{\epsilon}_{\text{PEHE}}$ for input with synthetic additive noise, as described in Appendix D.1. To compare the performances of TATEE and CFR on noisy features, we assess the relative gain in reduction of PEHE due to TATEE, given by $1 - \frac{\tilde{\epsilon}_{\text{PEHE,TATEE}}}{\tilde{\epsilon}_{\text{PEHE,CFR}}}$ and report the percentage point gains in PEHE. Figure 7 visualizes the performance of TATEE across all four datasets under various noise distributions. These results underscore the efficacy of incorporating topological awareness into representation learning, enhancing robustness against diverse noise distributions while maintaining performance on both clean and noisy data. Altogether, the performance and robustness improvements observed across all four datasets align with our theory-informed expectations and demonstrate the practical utility of the TATEE framework in robust treatment effect estimation.

Empirical Results on Performance. On clean data, TATEE is consistently on par with CFR, with minimal deviations in $\hat{\epsilon}_{\text{PEHE}}$ across all four datasets, showing 1.977×10^{-2}, 4.976×10^{-4},

2. Recall that the first absolute moment of the noise distribution is infinite for $\alpha = 1$, hence, we need to be cautious about interpreting the average of trials at $\alpha = 1$.

and 2.631×10^{-3} increase for IHDP, Jobs, and ACIC (respectively), and 1.007×10^{-2} decrease for Twins. This indicates that the integration of topological summaries does not compromise performance in noise-free scenarios. When subjected to additive noise from α-stable distributions in the inputs, TATEE demonstrates superior performance by reducing $\hat{\epsilon}_{\text{PEHE}}$ compared to CFR in most cases. On the IHDP and Twins datasets, TATEE consistently outperforms CFR across the entire range of noise parameters. For the Jobs and ACIC datasets, while the improvements are less consistent, TATEE still offers better performance in almost as many cases as CFR, while consistently exhibiting more robustness, as illustrated in Figure 3 and discussed in Section 5 and Appendix D.2.

Distinguishing Performance and Robustness. Note the distinction between the performance and robustness metrics: a worse (higher) $\tilde{\epsilon}_{\text{PEHE}}$ on noisy inputs can still yield a better (higher) ρ if the noise-induced deterioration, $\tilde{\epsilon}_{\text{PEHE}}/\hat{\epsilon}_{\text{PEHE}}$, is comparatively smaller. In other words, $\rho_{\text{TATEE}} - \rho_{\text{CFR}}$ can be positive even if both $\tilde{\epsilon}_{\text{PEHE}}$ and $\hat{\epsilon}_{\text{PEHE}}$ for TATEE exceed those of CFR, as long as TATEE's increase in error due to noise remains more modest. This subtlety underscores why TATEE might show higher $\hat{\epsilon}_{\text{PEHE}}$ overall, yet still achieve superior robustness as reflected by a larger ρ. Consequently, we observe scenarios where TATEE exhibits a smaller gap between its clean and noisy performances, reinforcing the principle that our method's primary goal is to mitigate noise-driven degradation.

Expected Performance Trade-offs. The theory-informed advantage of TATEE lies in improving robustness to noise by including $\mathcal{L}_{\text{topo}}$ in the loss. Since this regularization term can complicate the path toward minimizing other objectives, such as $\hat{\epsilon}_{\text{PEHE}}$ on clean inputs, making a slight performance decline is not unexpected. Nonetheless, the evaluation above affirms that TATEE's enhanced robustness does not come at the cost of a significant decline in performance: our experiments show only minimal increases in $\hat{\epsilon}_{\text{PEHE}}$ without noise, and in many settings, TATEE even outperforms CFR when noise is injected in terms of $\tilde{\epsilon}_{\text{PEHE}}$. Additionally, the Twins dataset—featuring empirical treatment effects and inherent real noise—exemplifies TATEE's robustness, as it achieves the lowest (best) $\hat{\epsilon}_{\text{PEHE}}$ among all methods tested, with or without synthetic noise injection. This finding implies that in contexts where empirical noise exists, TATEE can achieve lower treatment effect estimation error, aligning with our theoretical claims.

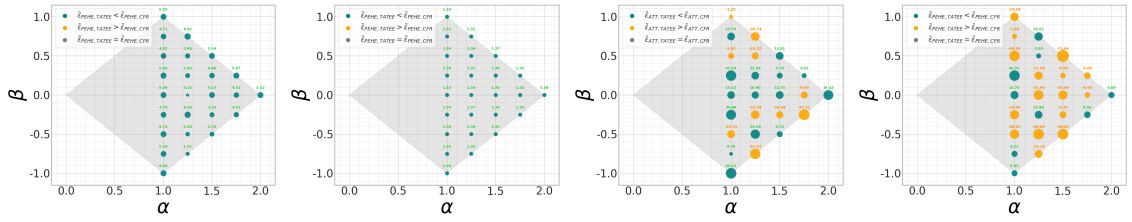

Figure 7: Evaluation of TATEE's performance when estimating CATE on noise-corrupted features from the IHDP, Twins, Jobs, and ACIC datasets (from left to right) for a range of (α, β) parameters of α-stable distributed noise. The Feller-Takayasu diamond (shaded) marks valid (α, β) values. The relative gain in $\hat{\epsilon}_{\text{PEHE}}$ from using TATEE is visualized by the circles. The green and orange circles mark improvement/deterioration; the size of the circles is proportional to the magnitude of relative change.

D.4. Benchmark Comparisons

Benchmark Models. To conduct a more thorough evaluation of the performance and robustness of TATEE, we compare it against several causal inference models. These benchmark models encompass a range of methodologies, ensuring a thorough assessment across different approaches. In the main text of the paper we discussed CFR [46], the most closely related model to TATEE, on which we base our main empirical evaluations. Here we name the other benchmark models used in our experiments described below and in Section 5. Dragonnet [47] is a neural network architecture designed to jointly model treatment assignments and potential outcomes by estimating propensity scores from the features, and GANITE [59] leverages generative adversarial networks to generate counterfactual outcomes. SLearner, TLearner, and XLearner [31] utilize meta-learning approaches that adapt base learners to estimate treatment effects. TARNet [46] is a CFR-type framework, learning separate representations for treated and control units, without the balancing loss term. Each of these models introduces unique mechanisms for addressing causal inference challenges. By benchmarking TATEE against these diverse models, we demonstrate its effectiveness and superior robustness across a broad spectrum of causal inference methods.

Results. Our experiments confirm TATEE's enhanced robustness compared to the benchmark models in pairwise comparisons. As mentioned in Section 5, this superior robustness is shown in Figure 4, whose (i, j) entry shows the proportion of dataset-parameter pairs in which the model corresponding to row i is more robust than the one for column j, TATEE consistently demonstrates better robustness across most datasets and noise parameters. Figure 8 provides additional evidence by showing the average robustness rank of each model (among the 8 models) for each pair of noise distribution parameters, as measured by ρ, averaged over the four datasets. A smaller number in row i and column j means, on average, the model corresponding to column j is more robust than others for the noise distribution parameters in row i. These results further confirm TATEE's superior robustness. Meanwhile, the $\hat{\epsilon}_{\text{PEHE}}$ values reported in Table 1 show that this gain in robustness does not come at the cost of a noticeable decline in performance. While the main purpose of TATEE is enhancing robustness, and in general, TATEE is not expected to perform better than CFR on noise-free inputs, the results indicate that TATEE has competitive performance in noise-free regimes as well, achieving the best or second best $\hat{\epsilon}_{\text{PEHE}}$ on three out of four datasets. Notably, on the Twins dataset–the only dataset where we suspect intrinsic noise exists without injecting synthetic noise–TATEE achieves the best $\hat{\epsilon}_{\text{PEHE}}$ without simulated noise.

Appendix E. Analysis of TATEE's Robustness

In Appendix C.5 we present a detailed analysis of the stability of TATEE's learning objective, relying on Theorem 2 to derive conditions which correspond to TATEE's enhanced robustness in counterfactual regression, stated in Proposition 9. This proposition suggests that under the assumptions of Theorem 2 and when the constraint stated in this theorem is satisfied, for a suitable choice of training parameters, TATEE's loss function (Equation 6) is more robust than the CFR's loss, which does not account for the topology of the representations. Here we provide a proof for this proposition. Unless stated otherwise, the derivation here follows the problem setup and the notation defined in Section 3 and Appendix B.

Notation. Following a similar notation as in Theorem 2, let μ^i be the measure over the representations of the control ($i = 0$) and treatment ($i = 1$) groups and ${\mu_f}^i$ the measure over the persistence

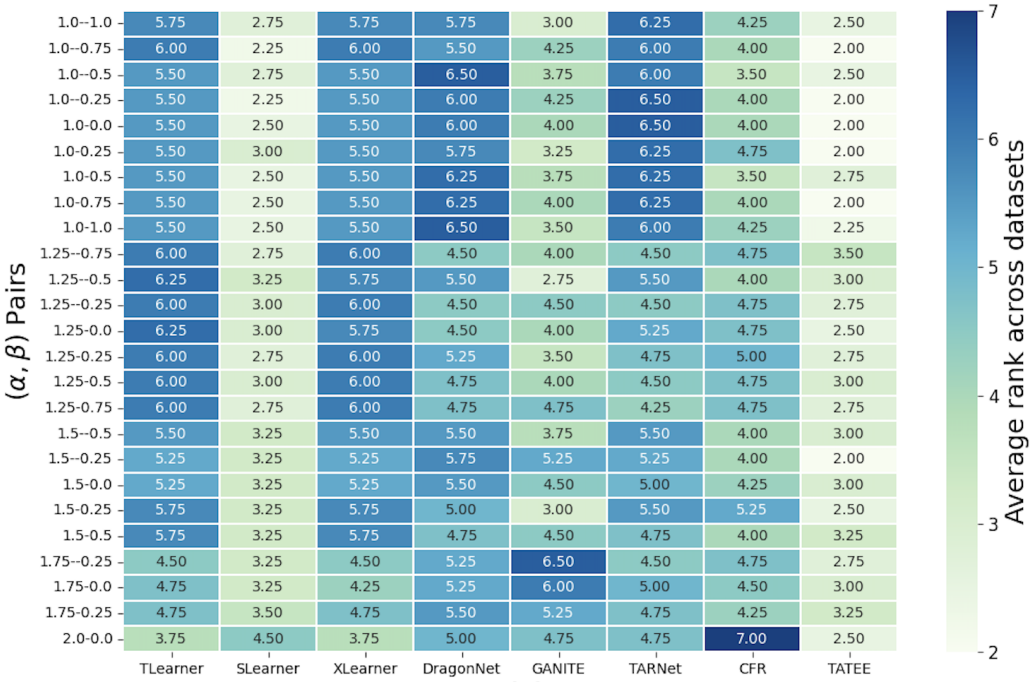

Figure 8: Comparing robustness of TATEE and the benchmark models: The quantity in column j and row i shows the average rank of the model corresponding to column j (out of 8), averaged over the model's rank across all 4 datasets, for the noise distribution's (α, β) values in row i.

	IHDP	Twins	Jobs	ACIC
TLearner	**0.596 ± 0.000**	0.454 ± 0.000	4.546 ± 0.000	0.264 ± 0.000
SLearner	0.863 ± 0.000	0.418 ± 0.000	0.004 ± 0.000	0.255 ± 0.000
XLearner	<u>0.596 ± 0.000</u>	0.453 ± 0.000	4.546 ± 0.000	0.264 ± 0.000
DragonNet	1.234 ± 0.435	0.493 ± 0.105	0.457 ± 0.409	0.334 ± 0.259
GANITE	4.184 ± 0.216	0.727 ± 0.271	0.013 ± 0.013	0.457 ± 0.229
TARNet	1.268 ± 0.456	0.504 ± 0.147	0.456 ± 0.462	0.220 ± 0.152
CFR	0.870 ± 0.089	<u>0.414 ± 0.017</u>	**0.002 ± 0.003**	**0.006 ± 0.005**
TATEE	0.888 ± 0.088	**0.404 ± 0.029**	<u>0.003 ± 0.004</u>	<u>0.009 ± 0.010</u>

Table 1: $\hat{\epsilon}_{\text{PEHE}}$ on input without synthetic noise. Lower means better performance, with best in bold and second best underlined. TATEE is best or second best (second to CFR) across all but one dataset. Notably, TATEE performs best on the Twins dataset—the only dataset which contains empirical values for ground-truth treatment effects, and hence can have real-world noise.

diagrams with a filtration function f. Also, similar to the notation used in Section 2, we use $\tilde{\cdot}$ to denote the variables corresponding to the noisy input. Let $W^{0,1} := W_p(\mu^0, \mu^1)$, $\tilde{W}^{0,1} := W_p(\tilde{\mu}^0, \tilde{\mu}^1)$, $\tilde{W}^0 := W_p(\tilde{\mu}^0, \mu^0)$, $\tilde{W}^1 := W_p(\tilde{\mu}^1, \mu^1)$, ${W_{\text{topo}}}^{0,1} := W_p({\mu_f}^0, {\mu_f}^1)$, $\tilde{W}^{0,1}_{\text{topo}} := W_p({\mu_{\tilde{f}}}^0, {\mu_{\tilde{f}}}^1)$, $\tilde{W}^0_{\text{topo}} := W_p({\mu_{\tilde{f}}}^0, {\mu_f}^0)$, $\tilde{W}^1_{\text{topo}} := W_p({\mu_{\tilde{f}}}^1, {\mu_f}^1)$ denote the Wasserstein-p distances. Following the notation used for the upper bounds in Section 3, for $i \in \{0, 1\}$, we use M^i and $\mathsf{K}^i_{\text{topo}}$ to denote the upper bounds on $\tilde{W}^i$ and $\tilde{W}^i_{\text{topo}}$, and $\hat{\mathsf{M}}^i$ and $\hat{\mathsf{K}}^i_{\text{topo}}$ for their final sample equivalents.

Proof. Given any positive balancing loss coefficient for the CFR loss, $\lambda_{\text{CFR}} > 0$, a suitable choice of λ and λ_{topo} loss term coefficients of TATEE can always lead to satisfying the inequality below.

$$1 \leq \frac{\lambda_{\text{CFR}} - \lambda}{\lambda_{\text{topo}}}. \tag{8}$$

Meanwhile, by Theorem 2, when φ satisfies the constraint stated in the theorem, we have $\hat{\mathsf{K}}^0_{\text{topo}} < \hat{\mathsf{M}}^0$ and $\hat{\mathsf{K}}^1_{\text{topo}} < \hat{\mathsf{M}}^1$, hence, using Inequality 8, we can write

$$1 \leq \frac{\lambda_{\text{CFR}} - \lambda}{\lambda_{\text{topo}}} \frac{\hat{\mathsf{M}}^0 + \hat{\mathsf{M}}^1}{\hat{\mathsf{K}}^0_{\text{topo}} + \hat{\mathsf{K}}^1_{\text{topo}}}.$$

Rearranging the terms in this inequality gives

$$\lambda \left(\hat{\mathsf{M}}^1 + \hat{\mathsf{M}}^0\right) + \lambda_{\text{topo}} \left(\hat{\mathsf{K}}^1_{topo} + \hat{\mathsf{K}}^0_{topo}\right) \leq \lambda_{\text{CFR}} \left(\hat{\mathsf{M}}^1 + \hat{\mathsf{M}}^0\right), \tag{9}$$

completing the proof for Proposition 9. □

Notice that the result in Inequality 9 shows that the upper bound on the noise-induced change in the sum of the balancing and topological loss terms of TATEE is smaller than the upper bound on the change in the balancing term of CFR due to noise. These upper bounds are due to the triangle and quadrilateral inequalities (or applying the reverse triangle inequality and then the triangle inequality), which yield the following.

$$\begin{aligned}
\left\|\tilde{W}^{0,1} - W^{0,1}\right\| &\leq \tilde{W}^0 + \tilde{W}^1 \leq \mathsf{M}^0 + \mathsf{M}^1, \\
\left\|\tilde{W}^{0,1}_{\text{topo}} - {W_{\text{topo}}}^{0,1}\right\| &\leq \tilde{W}^0_{\text{topo}} + \tilde{W}^1_{\text{topo}} \leq \mathsf{K}^0_{\text{topo}} + \mathsf{K}^1_{\text{topo}}, \\
\left|\left(\lambda\tilde{W}^{0,1} + \lambda_{\text{topo}}\tilde{W}^{0,1}_{\text{topo}}\right) - \left(\lambda W^{0,1} + \lambda_{\text{topo}}{W_{\text{topo}}}^{0,1}\right)\right| &\leq \lambda \left|\lambda\tilde{W}^{0,1} - W^{0,1}\right| \\
&\quad + \lambda_{\text{topo}} \left|\lambda_{\text{topo}}\tilde{W}^{0,1}_{\text{topo}} - {W_{\text{topo}}}^{0,1}\right|, \\
&\leq \lambda \left(\mathsf{M}^0 + \mathsf{M}^1\right) + \lambda_{\text{topo}} \left(\mathsf{K}^0_{\text{topo}} + \mathsf{K}^1_{\text{topo}}\right).
\end{aligned}$$

Appendix F. α-Stable Distributions

α-stable distributions are a rich class of probability distributions that allow modeling data with heavy tails and asymmetry. Their flexible parametric form is particularly suitable for data that exhibit extreme values. Random variables with α-stable distributions do not necessarily have finite mean or variance. This family of distributions facilitate a generalization of the central limit theorem (CLT): While the normal distribution arises as the limit distribution when summing independent, thin-tailed random variables with finite second moment, α-stable distributions with the tail parameter $\alpha < 2$ arise as the limit when the random variables have infinite variance [22, 37].

The characteristic function of a random variable with an α-stable distribution is as follows,

$$\varphi(t) = \begin{cases} \exp\left\{itm - |ct|^\alpha \left(1 - i\beta \text{sgn}(t) \tan\frac{\pi\alpha}{2}\right)\right\} & \alpha \neq 1 \\ \exp\left\{itm - |ct|^\alpha \left(1 + i\beta\frac{2}{\pi}\text{sgn}(t) \log|t|\right)\right\} & \alpha = 1. \end{cases} \tag{10}$$

The parameters α, β, $c \geq 0$, and $m \in \mathbb{R}$, which we refer to as the *characteristic exponent* or the *tail* parameter, *skewness* or *asymmetry* parameter, *scale* parameter, and *location* parameter, determine the distribution of the random variable.

Note that the main parameters of interest, determining the shape of the distribution, are the tail and skewness parameters. These parameters take values in a region of the plane where $\alpha \in (0, 2]$ and $|\beta| \leq \min\{\alpha, 2-\alpha\}$, which is referred to as the *Feller-Takayasu diamond* [37]. The skewness parameter $\beta \in [-1, 1]$ introduces asymmetry; a symmetric α-stable distribution has $\beta = 0$. The parameter α determines how fast the tail decays, with smaller values meaning heavier tails. The normal distribution is the special case when $\alpha = 2$, and $\alpha = 1$ corresponds to the Cauchy distribution. Except the Gaussian case, other $\alpha-$stable distributions have infinite moments of order greater than or equal to α, i.e., infinite variance for all $\alpha < 2$, and infinite absolute mean for $\alpha \leq 1$. Most statistical models rely on the CLT to justify using Gaussian distributions as the asymptotic distribution of the sum/mean of an independent and identically distributed (i.i.d.) sequence of random variables. However, the CLT does not hold for α-stable distributions with $\alpha < 2$. This fact is consequential for the sample size criterion and significance level of statistical tests [42].

Appendix G. Implementation Details

Recall that the main purpose of our experiments is to evaluate the change in robustness to input noise due to our proposed topological regularization in TATEE. To this end, in order to assess the impact of incorporating the regularization term based on persistence diagrams into the training of the model, and not the CFR-type model architecture, we adopted the same parameters and overall configuration utilized by Shalit et al. [46] for CFR, for all implementation and training purposes of TATEE, as well as the CFR model we compared against TATEE. Each of φ, h_0, and h_1 have 3 fully connected layers of size 200 for φ and 100 for h_0 and h_1. Following Shalit et al. [46], we used the value $\sqrt{10}$ for the λ regularization coefficient for the balancing term of the training objective in Equation 6, which Shalit et al. [46] found to yield the lowest ϵ_{PEHE} on the IHDP dataset. After a standard grid search [3] fine-tuning of λ_{topo}, we used $\lambda_{\text{topo}} = \frac{\sqrt{10}}{4}$ for the topological regularization coefficient in Equation 6, in the case of IHDP dataset. We also performed a grid search fine tuning over batch size, λ, and λ_{topo} for the Twins and Jobs datasets, with a parameters grid of size 80, from which we obtained the values 128, 10, and $2\sqrt{10}$ for the Twins dataset, 128, 0.1, and 1 for the Jobs dataset, and 256, 10, and 0.1 for the ACIC dataset. The learning rate is 10^{-2} for mini-batch training using an Adam optimizer [30] with weight decay parameter value of 10^{-5}. The system specification for the computations is provided in Table 2.

CPU	Intel(R) Xeon(R) CPU @ 2.20GHz
GPU	Nvidia A100 SXM4 40GB
OS	Ubuntu 22.04.3 LTS
Architecture	x86_64

Table 2: System specifications for the computations.

3. In this case the search is in fact over a line, as we fine tuned only a single parameter.

Appendix H. Datasets

Infant Health and Development Program (IHDP). The IHDP dataset is a semi-synthetic benchmark widely used in causal inference studies [25]. It combines real-world covariates from a randomized experiment with simulated counterfactual outcomes, providing a ground truth for evaluating treatment effect estimation methods.

Twins. The Twins dataset uniquely contains both factual and counterfactual outcomes, as it includes data on twin pairs [35]. By treating one twin as treated and the other as control, the dataset provides real-world values for both potential outcomes, eliminating the need for synthetic counterfactuals. This characteristic allows for evaluation of treatment effects estimation against empirical ground-truth, making it the only dataset for assessing the robustness of methods in the presence of intrinsic empirical noise, without injecting synthetic noise. Notably, TATEE outperforms all benchmarks on this dataset (see Appendix D.4) in experiments without simulated noise, achieving the lowest (best) ϵ_{PEHE}.

Jobs. The Jobs dataset, originally collected by LaLonde [34] and later curated for causal inference benchmarking by Shalit et al. [46], includes both randomized controlled trial (RCT) and observational data. The dataset combines treated units from the RCT subset with a control sample from Dehejia and Wahba [14], ensuring that treatment assignment depends on covariates.

Atlantic Causal Inference Conference (ACIC) 2018. The ACIC-18 dataset [1] is based on data from the National Study of Learning Mindsets [58]. Assuming an RCT design, the dataset allows for the computation of the true average treatment effect on the treated (ATT). To introduce covariate shift, a mask variable sampled from a Bernoulli distribution based on a feature is applied to control units, making treatment assignment dependent on the covariates.