Adversarial Estimation of Topological Dimension with Harmonic Score Maps

Eric Yeats (Duke), Cameron Darwin, Frank Liu (ODU), Hai Li (Duke)

Contact: eric.yeats@duke.edu

NEURAL INFORMATION PROCESSING SYSTEMS

Duke

OAK RIDGE National Laboratory

OLD DOMINION UNIVERSITY

Dirichlet energy regularization of score maps boosts their adversarial robustness while revealing learned topological dimension.

Dirichlet Energy (DE): a functional measuring how *variable* a function is. A harmonic map is a critical point of the DE subject to a boundary constraint.

Definition 2.2. (Dirichlet Energy) The Dirichlet energy (DE) of a map $\phi : X \to Y$ is:

$$\mathcal{D}[\phi] = \frac{1}{2}\int_X \|d\phi_x\|^2 d\mathrm{Vol}(x),$$

where $\|d\phi_x\|$ is the spectral norm of the differential of the map at x.

Adversarial Robustness: Deep neural nets are notoriously vulnerable to malicious input perturbations [2]. Small changes to the input can result in very large changes in output.

Topological Dimension (TD): Number of dimensions data occupies near a point. It is important for:

- Data compression and dimension reduction
- Generalization ability of classifiers

Existing statistical estimators of TD require ample data and well-picked hyperparameters, esp. for high-dimensional, noisy data.

Method Contributions

We identify the *local averaging property* as key to instilling robustness of learned score maps, and we prove that DE regularization corresponds exactly to added variance in the normal subspace of the learned density.

- **Locally Averaging** score maps are more robust to adversaries. Reducing Dirichlet energy of score maps makes them closer to locally averaging.
- **Dirichlet Energy (DE)** Regularization of score maps corresponds to learning additional variance in the normal subspace of the learned density.
- **Topological dimension** is revealed by exactly how much additional variance is learned, and we can measure variance with adversarial attacks.

Training Implementation We augment weighted denoising score matching with DE regularization:

$$\theta^* = \arg\min_\theta \mathbb{E}_t \lambda(t)\mathbb{E}_{x_0}\mathbb{E}_{x_t|x_0}\underbrace{\|s_\theta(x_t,t) - \nabla_{x_t}\log p_{0t}(x_t|x_0)\|^2}_{\text{Denoising Score Matching}} + \underbrace{n\gamma\|ds_\theta(x_t,t)\|^2}_{\text{DE Reg.}}$$

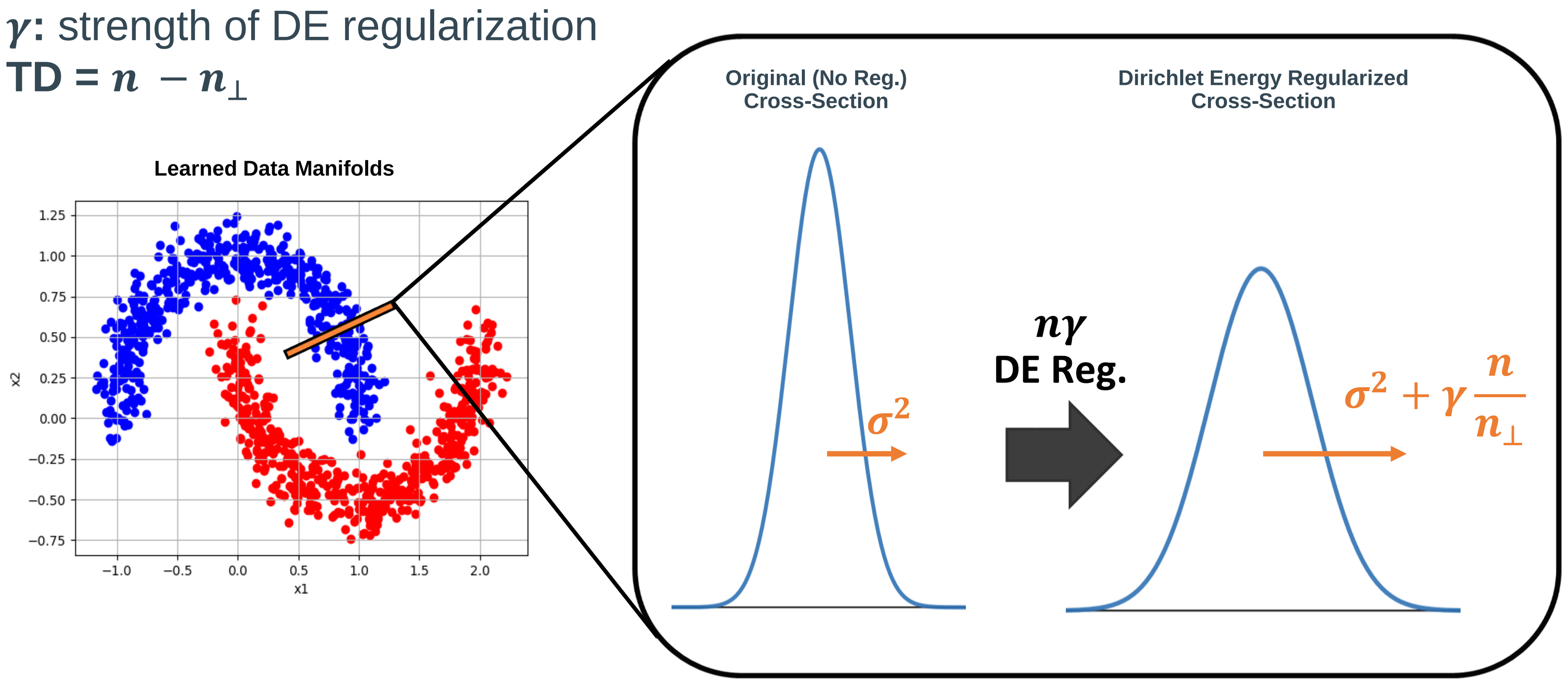
Additive Variance Property of DE Regularization Reveals Topological Dimension (TD)

$\boldsymbol{n}$: ambient dimension

$\boldsymbol{n_\perp}$: normal (off-manifold) dimension

$\boldsymbol{\gamma}$: strength of DE regularization

TD = $\boldsymbol{n - n_\perp}$



Algorithm 1 Topological Dimension Estimate

Require: $x \in \mathbb{R}^n$, $s_\theta : \mathbb{R}^n \to \mathbb{R}^n$, $\tilde{g} : \mathbb{R}^n \times \mathbb{R}^+ \to \mathbb{R}^n$, $\gamma \in \mathbb{R}^+$, $\sigma \in \mathbb{R}^+$
 ▷ $s_\theta(x)$: score map with γ DE reg. and σ noise scale, $\tilde{g}(x,\epsilon)$: attack func. with $L2$ budget ϵ

Ensure: $\hat{n}_{\mathcal{M}} \approx n_{\mathcal{M}}$

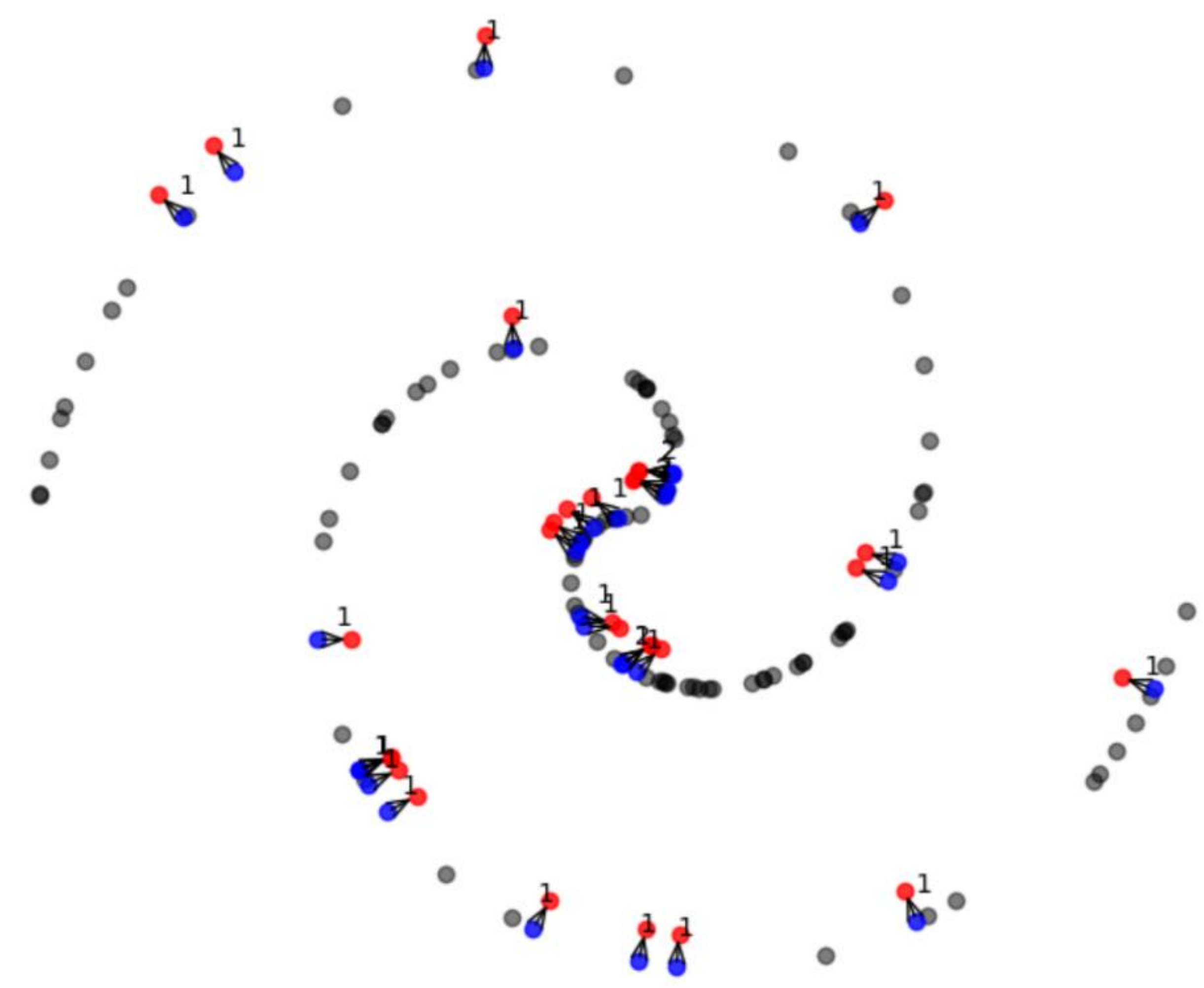
 $\tilde{x} \leftarrow \tilde{g}(x,\sigma)$ ▷ $\tilde{g}$ returns an adversarial example near x (e.g., by following $-s_\theta(x)$)

 $\delta \leftarrow \|s_\theta(\tilde{x}) - s_\theta(x)\| \,/\, \|\tilde{x} - x\|$ ▷ δ stores learned score "slope"

 $\hat{n}_{\mathcal{M}} \leftarrow n - n\gamma/(\delta^{-1} - \sigma^2)$ ▷ δ should be approximately $1/(\sigma^2 + n\gamma/n_\perp)$

 return $\hat{n}_{\mathcal{M}}$

Given the ambient dimension n, denoising scale σ, and level of DE regularization γ, one can recover the topological dimension (TD) using adversarial attacks.



Depiction of our topological dimension estimation method applied to the "Swirl" manifold. Some randomly selected original data and adversarial attacks are plotted with associated estimates of topological dimension.

Results: Adversarial Robustness

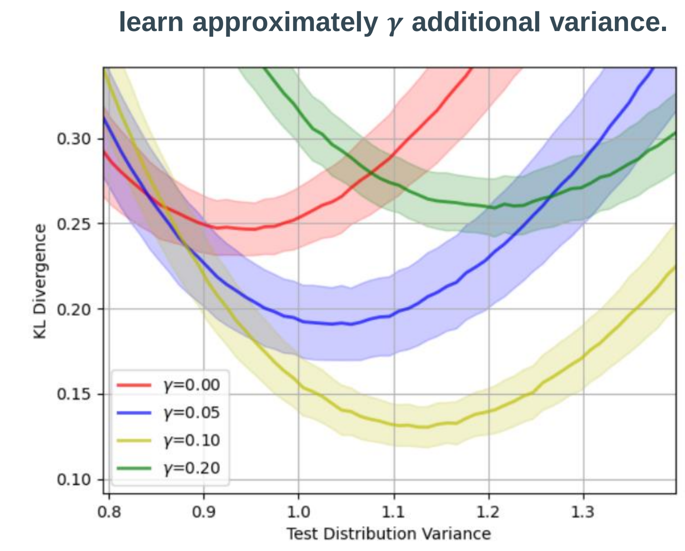
Table 1: BPD(↓) for L_2 attacks on CIFAR-10 test set. **Key:** Method(iters, ϵ)

DE Reg. (γ)	Clean	Random(1, 0.2)	PGD(1, 0.2)	Random(1, 0.8)	PGD(20, 0.8)
0	**3.288**	3.81	4.182	4.834	5.282
1e-4	3.387	**3.776**	4.127	4.803	5.211
2e-4	3.464	3.793	**4.118**	**4.787**	**5.187**

The likelihood of DE regularized DDPMs is more robust to adversarial attacks.

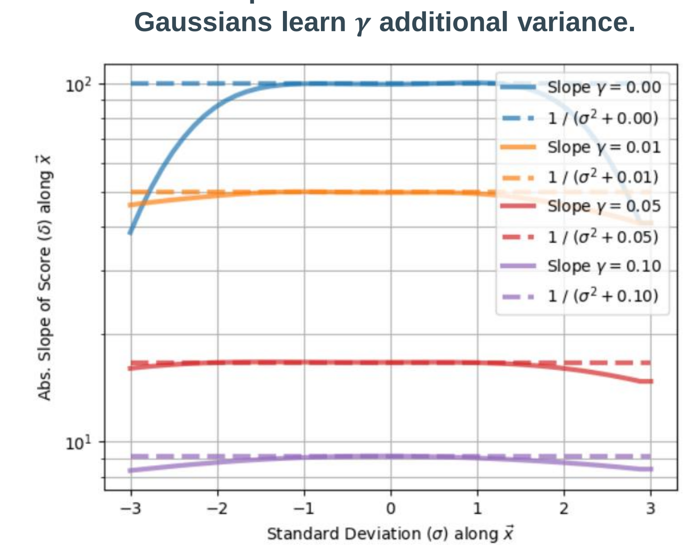
Results: Additive Variance Property

DDPMs trained on 8-dim Standard Gaussians learn approximately γ additional variance.



(a) KL divergence of isotropic Gaussian distributions of different variance with the learned distributions of DDPMs. Solid line: average, shaded: one stdev.

Score maps trained on 16-dim Standard Gaussians learn γ additional variance.



(b) Comparison of learned score "slope" along the vector $\vec{x} = \vec{1}$ of score maps trained on an isolated point in $\mathbb{R}^{16}$ with $\sigma = 0.1$ and various levels of DE reg. (γ).

Results: Topological Dimension Estimation

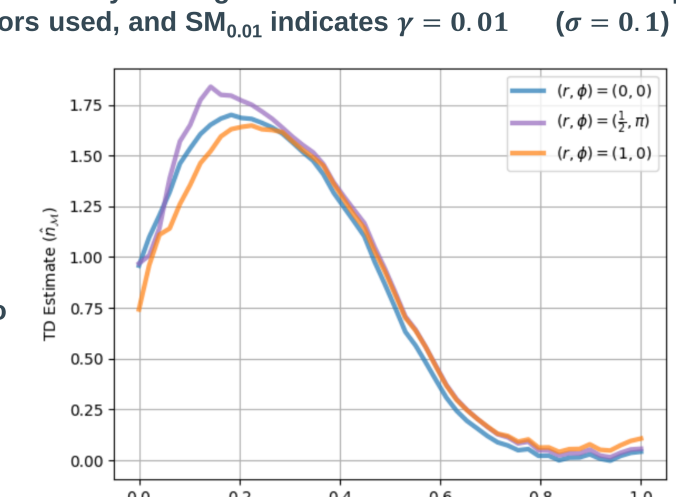
Table 2: MSE (↓) of topological dimension prediction averaged over 5 independent trials

Benchmark	TD(s)	MLE_{10}	MLE_{20}	MiND_{10}	MiND_{20}	$\mathrm{SM}_{0.01}$
Swirl	1	0.152	0.063	0.171	**0.001**	0.046
Swirl $\sigma_{0.01}$	1	0.495	0.272	0.974	0.486	**0.080**
LineDiskBall	1-3	0.308	0.226	**0.118**	0.479	0.312
HyperTwinPeaks	10	5.931	5.006	13.755	25.735	**0.084**
HyperTwinPeaks	30	90.008	89.830	402.091	437.310	**0.162**

Our method (SM) is competitive with statistical estimators (MLE and MiND [3,4]) on simple manifolds [5] (Swirl, LineDiskBall) but is much more accurate on noisy and high-dimensional manifolds. Subscripts of statistical methods are number of neighbors used, and $\mathrm{SM}_{0.01}$ indicates $\gamma = 0.01$ ($\sigma = 0.1$)

Topological dimension is estimated for various points on the "Swirl" manifold as they are decayed through time (forward VP diffusion process without noise).

The TD estimates clearly depict how the locally 1D swirl structure is first collapsed to a disk then eventually to a point (the Gaussian prior).



Conclusion

This work connects adversarial vulnerability of score models with the geometry of the underlying manifold they capture. We show that minimizing the Dirichlet energy of learned score maps simultaneously boosts their robustness while revealing topological dimension. Leveraging this, we introduce a novel method to measure the topological dimension of manifolds captured by score models using adversarial attacks.

References

[1] Yang Song et al. (2020). "Score-based generative modeling through stochastic differential equations." *arXiv preprint arXiv:2011.13456*

[2] Christian Szegedy et al. (2013). "Intriguing properties of neural networks." *arXiv preprint arXiv:1312.6199*

[3] Alessandro Rozza et al. (2012). "Novel high intrinsic dimensionality estimators." *Machine Learning*, 89:37-65

[4] Elizaveta Levina et al. (2004). "Maximum likelihood estimation of intrinsic dimension." *Advances in neural information processing systems,* 17 2004

[5] Jonathan Bac et al. (2021). "Scikit-dimension: a python package for intrinsic dimension estimation. *Entropy*, 23

Acknowledgements

Eric Yeats is supported by a U.S. Department of Energy, Office of Science Graduate Student Research (SCGSR) award. This research is further supported by U.S. Army Research W911NF2220025, U.S. Air Force Research Lab FAA8750-21-1-1015, and through the U.S. Department of Energy Office of Advanced Scientific Computing Research's "Data-Driven Decision Control for Complex Systems (DnC2S)" project.