

Robust Classification and Purification with a Single Tensor Network Born Machine

Martin L. Nissen Gonzalez¹ José R. Pareja Monturiol²

¹*GREYC, University of Caen* ²*Universidad Politécnica de Madrid*

Motivation and contribution. Classifiers trained to be accurate on natural data can be fooled by imperceptibly small perturbations [5], and adversarial robustness is linked to how well a classifier captures the data distribution [6], renewing interest in *generative classifiers*. Diffusion- and flow-based classifiers improve robustness [2] but keep the generative model and classifier separate, so neither exposes the input density directly; energy-based models such as JEMs improve robustness [4], but their intractable partition function leaves the input density only approximate for detection or purification.

We argue that matrix product state (MPS) Born machines [3] are a uniquely clean instrument: via the Born rule a single MPS defines a joint density with *tractable* partition function. With no architectural change, one model supports (i) a continuous discriminative→generative interpolation that regularizes the classifier, (ii) density-based detection of adversarial inputs, and (iii) purification from the model’s own density, letting us study the generative↔robust-classification relationship *within one tractable model* whose only moving part is the loss, a regime unexplored for energy-based classifiers. Even a small generative component improves the density and strengthens these defenses over discriminative training.

Tractable density. Each input $x \in [-1, 1]^n$ is embedded feature-wise by a local map $\phi : [-1, 1] \rightarrow \mathbb{C}^d$ [7] and the label by a one-hot vector $e_c \in \mathbb{C}^K$; a weight tensor T , decomposed as an MPS, defines the amplitude $\psi_T(x, c) = \langle \bigotimes_i \phi(x_i) \otimes e_c, T \rangle$, contracted without forming T . The Born rule gives a joint density $p_T(x, c) = |\psi_T(x, c)|^2 / Z_T$ with tractable normalizer: for an isometric (Legendre) embedding, $Z_T = \|T\|_F^2$. Conditioning on x gives the classifier $p_T(c|x)$; marginalizing over labels gives the input density $p_T(x)$.

One model, only the loss changes. Following Bishop et al. [1], we interpolate the discriminative loss $L_D = -\mathbb{E}_{p_{\text{data}}} \log p_T(c|x)$ and the generative loss $L_G = -\mathbb{E}_{p_{\text{data}}} \log p_T(x, c)$ with $\alpha \in [0, 1]$,

$$L_\alpha = (1 - \alpha)L_D + \alpha L_G = -\mathbb{E}_{p_{\text{data}}} \log |\psi_T(x, c)|^2 + (1 - \alpha) \mathbb{E}_{p_{\text{data}}} \log \sum_{c'} |\psi_T(x, c')|^2 + \alpha \log \|T\|_F^2. \quad (1)$$

The discriminative term normalizes over classes, the generative term over input space; raising α adds generative regularization. Since $p_T(x)$ is available, we flag low-likelihood inputs as adversarial (a training-free detector) and purify perturbed inputs within an ℓ_∞ ball, either by *likelihood* ascent on $\log p_T$ via PGD or by *resampling* each coordinate from its truncated model conditional, reusing the same model.

Results. On a two-dimensional spirals dataset, the purely discriminative model learns a sharp, accurate boundary, but its input density barely resembles the data; a small generative component concentrates $p_T(x)$ along the spiral geometry, while the adversarially trained model is intermediate. On MNIST (resized to 12×12 pixels), generative regularization improves robustness at high clean accuracy over purely discriminative training, most clearly under likelihood-based purification. The regularization pushes the model toward a more faithful density, which benefits both purification and detection; but when model capacity is too small, robust-accuracy gains can vanish and clean accuracy can suffer. Adversarial training, by contrast, raises robust accuracy substantially even at small capacity, yet gains nothing from purification, suggesting that generative regularization and adversarial training play complementary roles.

A single tractable MPS Born machine thus unifies classification, detection, and purification, with a mildly generative loss improving both density and defenses.

References

- [1] C. M. Bishop et al. Discr. or gen.? getting the best of both. In *Bayesian Statistics 8*. Oxford University Press, 2007.
- [2] H. Chen et al. Robust classification via a single diffusion model. In *ICML*, 2024.
- [3] Z.-Y. Han et al. Unsupervised generative modeling using matrix product states. *Phys. Rev. X*, 8(3), 2018.
- [4] K. Jiang et al. Your classifier can do more: Towards balancing the gaps in classification, robustness, and generation, 2026.
- [5] A. Madry et al. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.
- [6] S. Santurkar et al. Image synthesis with a single (robust) classifier. In *NeurIPS*, volume 32, 2019.
- [7] E. Stoudenmire et al. Supervised learning with tensor networks. In *NeurIPS*, volume 29, pages 4799–4807, 2016.