High-dimensional and Permutation Invariant Anomaly Detection with Diffusion Generative Models

Vinicius Mikuni

National Energy Research Scientific Computing Center,
Lawrence Berkeley National Laboratory, Berkeley, CA 94720 vmikuni@lbl.gov

Benjamin Nachman
Physics Division,
Lawrence Berkeley National Laboratory, Berkeley, CA 94720
Berkeley Institute for Data Science,
University of California, Berkeley, CA 94720, USA bpnachman@lbl.gov

Abstract

Methods for anomaly detection of new physics processes are often limited to low-dimensional spaces due to the difficulty of learning high-dimensional probability densities. Particularly at the constituent level, incorporating desirable properties such as permutation invariance and variable-length inputs becomes difficult within popular density estimation methods. In this work, we introduce a permutation-invariant density estimator for particle physics data based on diffusion models, specifically designed to handle variable-length inputs. We demonstrate the efficacy of our methodology by utilizing the learned density as a permutation-invariant anomaly detection score, effectively identifying jets with low likelihood under the background-only hypothesis. To validate our density estimation method, we investigate the ratio of learned densities and compare to those obtained by a supervised classification algorithm.

1 Introduction

Anomaly detection (AD) has emerged as a complementary strategy to classical model-dependent searches for new particles at the Large Hadron Collider and elsewhere. These tools are motivated by the current lack of excesses and the vast parameter space of possibilities [1, 2]. Machine learning (ML) techniques are addressing these motivations and also allowing for complex particle physics data to be probed holistically in their natural high dimensionality [3]. Nearly all searches for new particles begin by positing a particular signal model, simulating the signal and relevant Standard Model (SM) backgrounds, and then training (with or without ML) a classifier to distinguish the signal and background simulations. Machine learning–based AD tries to assume as little as possible about the signal while also maintaining the ability to estimate the SM background. Two main classes of ML approaches are unsupervised and weakly/semi-supervised. Unsupervised methods use 'no' information about the signal in training

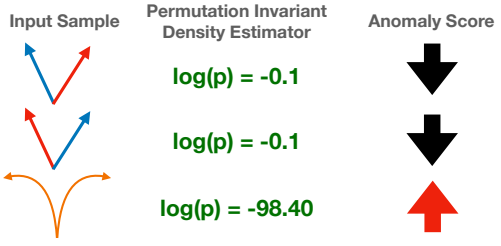

Figure 1: Diagram of the proposed anomaly detection model.

Machine Learning and the Physical Sciences Workshop, NeurIPS 2023.

while weakly/semi-supervised methods use limited or noisy labels. The 'no' is in quotes because there is often implicit signal information used through event and feature selection. At their core, unsupervised methods select events that are rare, while weakly/semi supervised methods focus on events that have a high likelihood ratio with respect to some reference(s). Even though a number of weakly supervised methods have statistical guarantees of optimality that unsupervised method lack [4, 5], there has been significant interest in unsupervised AD because of its flexibility. The flexibility of unsupervised learning leads to a number of challenges. There is no unique way to estimate the probability density of a given dataset, with some methods offering only an implicit approximation through proxy quantities like the reconstruction fidelity of compression algorithms. Even though particle physics data are often described by high- (and variable-)dimensional, permutation-invariant sets ('point clouds'), there has not yet been a proposal to use explicit density estimation techniques for AD that account for all of these properties.

We propose to use point cloud diffusion models combined with explicit density estimation for AD, summarized in Fig. 1. Our approach is based on Ref. [6], and inherits the ability to process variable-length and permutation-invariant sets. From the learned score function, we estimate the data density and provide results for two different diffusion models; one trained with standard score-matching objective and one trained using maximum likelihood estimation. Since the true density is not known, we quantify the performance of the density estimation with likelihood ratios. Finally, we demonstrate the performance of the density as an anomaly score for top quark jets as well as jets produced from dark showers in a hidden valley model. Other tasks that require access to the data density could also benefit from our method.

2 Methodology

Score-based generative models are a class of generative algorithms that aim to generate data by learning the score function, or gradients of the logarithm of the probability density of the data. The training strategy presented in Ref. [7] introduces the idea of denoising score-matching, where data can be perturbed by a smearing function and matching the score of the smeared data is equivalent to matching the score of the smearing function Ref. [8]. Given some high-dimensional distribution $\mathbf{x} \in \mathbb{R}^{\mathrm{D}}$, the score function we want to approximate, $\nabla_{\mathbf{x}} \log p_{\text{data}}$, with $\mathbf{x} \sim p_{\text{data}}$, is obtained by minimizing the following quantity

$$\frac{1}{2}\mathbb{E}_t\mathbb{E}_{p_t(\mathbf{x})}\left[\lambda(t)\left\|\mathbf{s}_\theta(\mathbf{x_t}, t) - \nabla_{\mathbf{x_t}} \log p_t(\mathbf{x}_t|x_0)\right\|_2^2\right]. \tag{1}$$

The goal of a neural network $\mathbf{s}_\theta(\mathbf{x_t}, t)$ with trainable parameters θ and evaluated with data $\mathbf{x}_t$ that have been perturbed at time t is to give a time-dependent approximation of the score function. The time dependence of the score function is introduced to address the different levels of perturbation used in each time step. At times near 0, at the beginning of the diffusion process ($\mathbf{x}(0) := \mathbf{x}_0 := \mathbf{x}$), the smearing applied to data is small, gradually increasing as time increases and ensures that at longer time scales the distribution is completely overridden by noise. Similarly, the positive weighing function $\lambda(t)$ can be chosen independently and determines the relative importance of the score-matching loss at different time scales. The score function of the perturbed data is calculated by using a Gaussian perturbation kernel $p_\sigma(\tilde{x}|x) := \mathcal{N}(\mathbf{x}, \sigma^2)$ and $p_\sigma(\tilde{\mathbf{x}}) := \int p_{\mathrm{data}}(\mathbf{x}) p_\sigma(\tilde{\mathbf{x}}|\mathbf{x}) \mathrm{d}\mathbf{x}$, simplifying the last term of Eq. 1

$$\nabla_{\tilde{\mathbf{x}}} \log p_\sigma(\tilde{\mathbf{x}}|\mathbf{x}) = \frac{\mathbf{x} - \tilde{\mathbf{x}}}{\sigma^2} \sim \frac{\mathcal{N}(0,1)}{\sigma}. \tag{2}$$

The learned approximation to the score function can then be used to recover the data probability density by solving the following equation:

$$\log p_0(\mathbf{x}_0) = \log p_T(\mathbf{x}_T) + \int_0^T \nabla \cdot \tilde{\mathbf{f}}_\theta(\mathbf{x}_t, t) \mathrm{d}t, \tag{3}$$

with

$$\tilde{\mathbf{f}}_\theta(\mathbf{x}_t, t) = [f(t)\mathbf{x}_t - \frac{1}{2} g(t)^2 s_\theta(\mathbf{x}_t, t)]. \tag{4}$$

The drift (f) and diffusion (g) coefficients are associated with the parameters of the Gaussian perturbation kernel. In our studies, we use the VPSDE [9] framework with velocity parameterization

as used in [6]. In this parameterization, the score function of the perturbed data reads:

$$s_\theta(\mathbf{x}_t, t) = \mathbf{x}_t - \frac{\alpha_t}{\sigma_t}\mathbf{v}_\theta(\mathbf{x}_t, t), \tag{5}$$

where the outputs of the network prediction, $\mathbf{v}_\theta(\mathbf{x}_t, t)$, are combined with the perturbed data, $\mathbf{x}_t$, and the mean and standard deviation of the induced perturbation kernel $\mathcal{N}(\mathbf{x}(0)\alpha, \sigma^2)$. A cosine schedule is used with $\alpha_t = \cos(0.5\pi t)$ and $\sigma_t = \sin(0.5\pi t)$. The resulting drift and diffusion coefficients are also identified based on the perturbation parameter as

$$f(\mathbf{x}, t) = \frac{\mathrm{d}\log\alpha_t}{\mathrm{dt}}\mathbf{x}_t, \qquad g^2(t) = \frac{\mathrm{d}\sigma_t^2}{\mathrm{dt}} - 2\frac{\mathrm{d}\log\alpha_t}{\mathrm{dt}}\sigma_t^2. \tag{6}$$

While the estimation of the data probability density is independent from the choice of the weighing function $\lambda(t)$ described in Eq. 1, different choices can enforce different properties to the learned score function. For example, the velocity parameterization in Eq. 5 implicitly sets $\lambda(t) = \sigma(t)^2$, which avoids the last ratio in Eq. 2 that diverges as $\sigma(t) \to 0$ at times near 0. On the other hand, Ref. [10] shows that choosing $\lambda(t) = g(t)^2$ turns the training objective in Eq. 1 into an upper bound to the negative log-likelihood of the data, effectively allowing the maximum likelihood training of diffusion models and possibly leading to more precise estimates of the data probability density. The negative aspect of this choice is that the lack of the multiplicative σ^2 term can lead to unstable training. This issue can be mitigated by using an importance sampling scheme that reduces the variance of the loss function. During the training of the likelihood weighted objective we implement the same importance sampling scheme based on the log-SNR implementation defined in [11] where the time parameter is sampled uniformly in $-\log(\alpha^2/\sigma^2)$while in the standard implementation the time component itself is sampled from an uniform distribution.

3 Results

The top quark tagging dataset is the widely-used community standard benchmark from Ref. [12, 13]. The background consists of dijets produced via Quantum Chromodynamics (QCD) and the signal is top quark pair production with all-hadronic decays. All jets in the range $550\ \text{GeV} < p_T < 650\ \text{GeV}$ and $|\eta| < 2$ are saved for processing. Each jet is represented with up to 100 constituents (zero-padded if fewer; truncated if more). To illustrate the anomaly detection abilities of our approach, we also simulate jets produced from a dark shower within a hidden valley model [14–17]. Our dark showers are motivated by Ref. [18], and consist of a Z' with a mass of 1.4 TeV that decays to two dark fermions charged under a strongly coupled U(1)'. These fermions have a mass of 75 GeV and hadronize into dark pion and ρ mesons, each of which can decay back to the Standard Model.

The network implementation and training scheme used to train the diffusion model are the same ones introduced in Ref. [6], based on the DEEPSETS [19] architecture with Transformer layers [20]. This model is trained to learn the score function of the jet constituents in $(\Delta\eta, \Delta\phi, \log(1 - \mathrm{p_{T}}_{rel}))$ coordinates, with the relative particle coordinates $\Delta\eta = \eta_{\text{part}} - \eta_{\text{jet}}$, $\Delta\phi = \phi_{\text{part}} - \phi_{\text{jet}}$, and $\mathrm{p_{T}}_{\text{rel}} = \mathrm{p_{T}}_{\text{part}}/\mathrm{p_{T}}_{\text{jet}}$ calculated based on the jet kinematic information. The particle generation model is conditioned on the overall jet kinematics described by $(\mathrm{p_{T}}_{\text{jet}}, \eta_{\text{jet}}\text{mass}, N_{\text{part}})$ The overall jet kinematic information is learned (simultaneously) by a second diffusion model as done in Ref. [6] using a model based on the RESNET [21] architecture.

All features are first normalized to have mean zero and unit standard deviation before training. The probability density is calculated with Eq. 3. The integral is solved using SCIPY [22] with explicit Runge-Kutta method of order 5(4) [23, 24] with absolute and relative tolerances of 5×10^{-5} and 10^{-4}, respectively. Lower and higher values of the absolute and relative tolerances were tested with overall results remaining unchanged. We define the anomaly score in this work as

$$\text{anomaly score} = -\log(p(\text{jet})p(\text{part}|\text{jet})^{1/N}), \tag{7}$$

with the model learning the likelihood in the particle space conditioned on the jet kinematic information ($p(\text{part}|\text{jet})$) normalized by the particle multiplicity. We show the distribution of the anomaly score in Fig. 2 for diffusion models trained exclusively on QCD or top quark jets.

The diffusion model training using maximum likelihood ($\lambda(t) = g(t)^2$) also presents, on average, lower anomaly score compared to the standard diffusion approach ($\lambda(t) = \sigma(t)^2$). With this choice

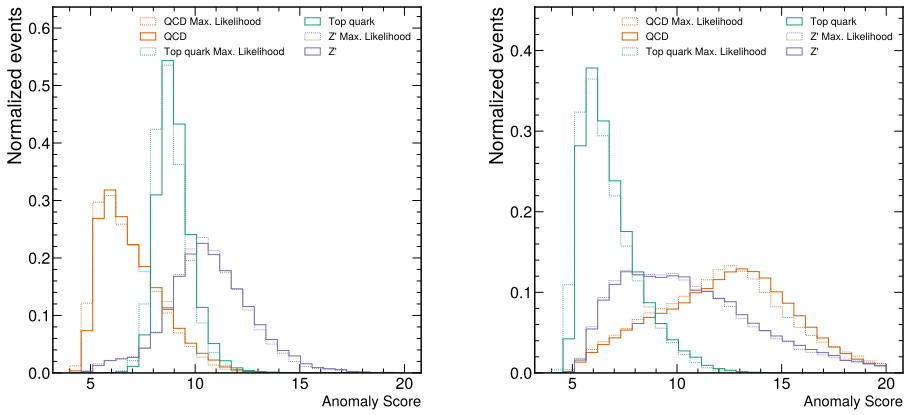

Figure 2: Anomaly score for QCD, top quark, and Z' jets evaluated on the model trained exclusively on QCD (left) and top quark (right) jet events.

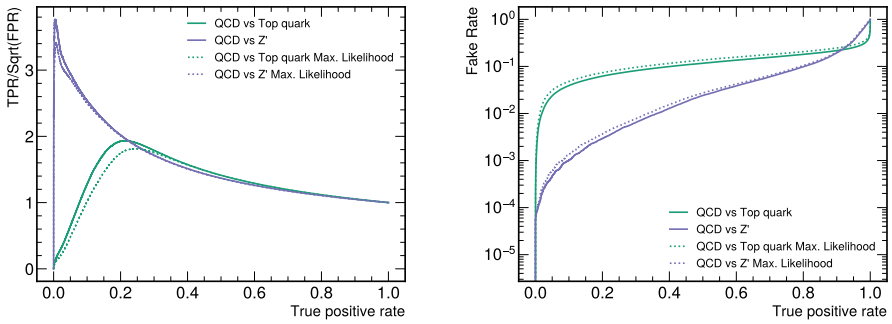

Figure 3: Significance improvement characteristic curve (left) and receiver operating characteristic curve (right) for different classes of anomalies investigated in this work.

of anomaly score, we investigate the the significance improvement characteristic curve (SIC), and Receiver operating characteristic (ROC) curve shown in Fig. 3.

For both classes of anomalies we observe maximum values for the SIC curve above 1, supporting the choice of metric for anomaly detection. Conversely, the maximum-likelihood training results in slightly lower SIC curve for anomalous jets containing the decay products of top quarks.

4 Conclusions and Outlook

In this work we presented an unsupervised anomaly detection methodology based on diffusion models to perform density estimation. Our method approximates the score function to estimate the probability density of the data. The diffusion model is trained directly on low-level objects, represented by particles clustered inside jets. The model for the score function is equivariant with respect to permutations between particles, leading to a permutation invariant density estimation. We test different strategies to train the diffusion model, including a standard implementation and a maximum-likelihood training of the score model. The maximum-likelihood training presents on average a lower negative-log-likelihood, indicating improved probability density estimation. However, when applied for anomaly detection, we do not observe notable improvements.

References

[1] N. Craig, P. Draper, K. Kong, Y. Ng, and D. Whiteson, Acta Phys. Polon. B **50**, 837 (2019), arXiv:1610.09392 [hep-ph] .

[2] J. H. Kim, K. Kong, B. Nachman, and D. Whiteson, JHEP **04**, 030 (2020), arXiv:1907.06659 [hep-ph] .

[3] G. Karagiorgi, G. Kasieczka, S. Kravitz, B. Nachman, and D. Shih, (2021), arXiv:2112.03769 [hep-ph] .

[4] E. M. Metodiev, B. Nachman, and J. Thaler, JHEP **10**, 174 (2017), arXiv:1708.02949 [hep-ph] .

[5] J. H. Collins, P. Martín-Ramiro, B. Nachman, and D. Shih, (2021), arXiv:2104.02092 [hep-ph] .

[6] V. Mikuni, B. Nachman, and M. Pettee, (2023), arXiv:2304.01266 [hep-ph] .

[7] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, ArXiv **abs/2011.13456** (2021).

[8] P. Vincent, Neural Computation **23**, 1661 (2011).

[9] J. Ho, A. Jain, and P. Abbeel, Advances in Neural Information Processing Systems **33**, 6840 (2020).

[10] Y. Song, C. Durkan, I. Murray, and S. Ermon, Advances in Neural Information Processing Systems **34**, 1415 (2021).

[11] D. Kingma, T. Salimans, B. Poole, and J. Ho, in *Advances in Neural Information Processing Systems*, Vol. 34, edited by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan (Curran Associates, Inc., 2021) pp. 21696–21707.

[12] A. Butter *et al.*, SciPost Phys. **7**, 014 (2019), arXiv:1902.09914 [hep-ph] .

[13] G. Kasieczka, T. Plehn, J. Thompson, and M. Russel, "Top quark tagging reference dataset," (2019).

[14] M. J. Strassler and K. M. Zurek, Phys. Lett. B **651**, 374 (2007), arXiv:hep-ph/0604261 .

[15] L. Carloni and T. Sjostrand, JHEP **09**, 105 (2010), arXiv:1006.2911 [hep-ph] .

[16] L. Carloni, J. Rathsman, and T. Sjostrand, JHEP **04**, 091 (2011), arXiv:1102.3795 [hep-ph] .

[17] S. Knapen, J. Shelton, and D. Xu, Phys. Rev. D **103**, 115013 (2021), arXiv:2103.01238 [hep-ph] .

[18] T. Buss, B. M. Dillon, T. Finke, M. Krämer, A. Morandini, A. Mück, I. Oleksiyuk, and T. Plehn, (2022), arXiv:2202.00686 [hep-ph] .

[19] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Póczos, R. Salakhutdinov, and A. J. Smola, CoRR **abs/1703.06114** (2017), 1703.06114 .

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, CoRR **abs/1706.03762** (2017), 1706.03762 .

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," (2015), arXiv:1512.03385 [cs.CV] .

[22] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, Nature Methods **17**, 261 (2020).

[23] J. R. Dormand and P. J. Prince, Journal of computational and applied mathematics **6**, 19 (1980).

[24] L. F. Shampine, Mathematics of computation **46**, 135 (1986).