Model-Agnostic Searches for New Physics with Normalizing Flows

Justin Tan [1]

Abstract

Experimental searches in high-energy physics for signatures of phenomena predicted by new physics models are typically structured as a simple hypothesis test for which the likelihood ratio between the background-only and signal-plus-background hypothesis forms the optimal test statistic. Current discovery searches at particle colliders are dominated by model-specific searches, which are bound to specific new physics scenarios and tend to poorly generalize outside a highly specialized scope. In this note we review model-agnostic approaches to searches for new physics signatures using normalizing flow models to approximate the likelihood ratio between two hypotheses. We also propose a bootstrap method that allows us to estimate the continuous densities that form the likelihood ratio in an unsupervised manner with minimal a priori knowledge of the signal structure.

1. Introduction

There are many outstanding questions in high-energy physics related to experimental phenomena that cannot be explained by the canonical theoretical framework of high-energy physics, the 'Standard Model', thought to be a low-energy approximation of a more fundamental theory. Experimental searches in this field are mainly concerned with searches for signatures of 'new physics' at particle colliders to test theoretical extensions to the Standard Model. These collider searches are typically model-specific, by assuming a given new physics 'signal' model with a particular experimental signature that is additive on top of well-understood 'background' processes. Discovery is then formulated in terms of a hypothesis test between two competing explanations of the observed data. Such approaches bind each analysis to a particular model of new physics, which there may be an arbitrary number of. Furthermore, model-specific approaches typically rely on supervision to distinguish between background and signal processes observed at collider experiments. As a consequence, they rely on expensive and potentially unreliable simulations of collider data. As a result, there has been growing interest in model-agnostic searches that decouple experimental analyses from individual models, instead searching for anomalous patterns in recorded data that may be incompatible with predictions made by the Standard Model. In this note we briefly review how anomaly detection may be phrased as a hypothesis test, and explore how normalizing flow models may be used to approximate the test statistic leading to the asymptotically most powerful test between two competing hypotheses.

2. Background

We will refer to physical processes detected at particle colliders as 'events'. In what follows events will be synonymous with their feature representations $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^n$. These features are derived from measurements of raw sensor data in particle colliders and typically describe kinematic and geometric properties of reconstructed particles in some recorded physical process. We may be interested in specific 'signal' classes of events, or wish to test if an observed data sample of events is compatible with the background-only hypothesis. Let $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ be a dataset consisting of i.i.d. observations, θ be a parameterization of a hypothesis on the process responsible for generating $\mathcal{D}$ and $p(\mathbf{x}|\theta)$ be a conditional density over phase space of observations. For discovery searches of hypothetical signal processes, we are interested in discriminating between the null hypothesis $\theta = \theta_0$ that the data consists of background processes only and some alternative hypothesis $\theta = \theta_1$. A claim of discovery of new physics requires establishing that the observed data is incompatible with the background-only hypothesis θ_0. The Neyman-Pearson lemma states that the likelihood ratio statistic λ forms the optimal test statistic between a simple null hypothesis θ_0 and simple alternate hypothesis θ_1:[1]:

[1]School of Physics, University of Melbourne, Australia. Correspondence to: Justin Tan <justin.tan@unimelb.edu.au>.

Second workshop on *Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models* (ICML 2020), Virtual Conference

[1]That is, maximizes the probability of rejecting the null hypothesis when the alternate is true for some fixed probabiilty of rejecting the null hypothesis when it is true.

$$\lambda\left(\mathcal{D};\theta_0,\theta_1\right)=\prod_{\mathbf{x}\in\mathcal{D}}\frac{p(\mathbf{x}|\theta_0)}{p(\mathbf{x}|\theta_1)}. \tag{1}$$

Unfortunately, it is not possible to directly evaluate $p(\mathbf{x}|\theta)$ due to the complexity of modelling the physical processes generating $\mathcal{D}$. Instead it is common in experimental high-energy physics to learn a proxy for λ by training a binary discriminator to distinguish between samples $\mathbf{x}\sim p(\mathbf{x}|\theta_0)$ and $\mathbf{x}\sim p(\mathbf{x}|\theta_1)$. It is well known that the optimal discriminator D^* minimizing the cross-entropy functional $\mathcal{L}=\mathbf{E}_{\mathbf{x}\sim p_{\theta_0}}\left[\log D(\mathbf{x})\right]+\mathbf{E}_{\mathbf{x}\sim p_{\theta_1}}\left[\log(1-D(\mathbf{x}))\right]$ can be used to approximate the likelihood ratio:

$$D^*(\mathbf{x})=\frac{p(\mathbf{x}|\theta_1)}{p(\mathbf{x}|\theta_0)+p(\mathbf{x}|\theta_1)},\quad \lambda(\mathbf{x})=\frac{1-D^*(\mathbf{x})}{D^*(\mathbf{x})}. \tag{2}$$

The drawback behind this approach is that it binds the analysis to the model explaining the hypothetical signal process, represented by the alternate hypothesis θ_1. Model-specific analyses may generalize poorly in the sense of an inability to detect experimental signatures that arise from new physics scenarios not predicted by the proposed model. These analyses typically rely on supervision to distinguish signal from background events - requiring the specialized simulation of events generated by each model. Such approaches strongly rely on the fidelity of the simulations for each model class to real-world data, which can be difficult to assess. Lastly, there can be an arbitrary number of explanatory models, as in, for example, proposed Supersymmetric extensions to the Standard Model, making coverage of the complete volume of phase space considered by collider experiments impractical.

2.1. Estimating the Likelihood Ratio

In typical experimental analyses, the null hypothesis θ_0 is taken to be that the data originates from known Standard-Model background processes. In other words, the distribution of data $\mathbf{x}$ in phase space is given by some background-only generative process $p(\mathbf{x}|\text{bkg})$. The alternate hypothesis θ_1 is then taken to be that the data $\mathbf{x}$ are distributed according to the observed data distribution $p(\mathbf{x}|\text{data})$. These are simple hypotheses in the sense that there are no unknown parameters associated with either distribution. Then the likelihood ratio statistic maximizes the probability of rejecting the background-only hypothesis, when the data hypothesis is true (which it is by definition), for some fixed probability of rejecting the background-only hypothesis when it is true. So we would like to estimate the quantity

$$\lambda(\mathbf{x})=\frac{p(\mathbf{x}|\text{bkg})}{p(\mathbf{x}|\text{data})}. \tag{3}$$

The natural question is how to reliably estimate the densities which occur in the likelihood ratio, without the need for explicit supervision. The usefulness of the likelihood ratio to identify anomalous samples in data depends on how well the phase space densities appearing in the ratio can be approximated. If these densities are estimated from data, then it is important that the model for $p(\mathbf{x}|\text{bkg})$ is estimated using the appropriate subpopulation from the total data sample $\mathcal{D}$. We review one such approach and propose an alternative below.

2.1.1. Localization

Let m be some feature in which signal events are known to exhibit different characteristics from background events. For simplicity we will assume $m\in\mathbb{R}$. In most experimental searches in high-energy physics, m is typically taken to be the invariant mass of some resonance predicted by theoretical models. In this case the resonance is the true progenitor of the particles in the signal event, As signal events originate from the physical decay of some resonance, we expect the distribution of the invariant mass of the constituents of a signal event to be strongly peaked in a localized region around the true resonance mass m_0. This is in contrast to background events, which should exhibit a featureless, 'smoothly falling' shape in this region - these characteristic shapes may be observed in Figure 1. Nachman & Shih (2020) make the physically motivated assumption that signal events are localized in some *signal region* (SR) of the resonant mass m with $m\in[m^*-\Delta_m, m^*+\Delta_m]$. The density of the observables given the background-only hypothesis, $p(\mathbf{x}|\text{bkg})$, is then approximated through some density estimation method in the regions outside the SR with $m\notin[m^*-\Delta_m, m^*+\Delta_m]$, sometimes referred to as the *sideband region* (SB). This approximate density may be interpolated into the signal region and compared against an approximation of $p(\mathbf{x}|\text{data})$ obtained through density estimation in the signal region, to form the likelihood ratio $\lambda(\mathbf{x})$. In the absence of signal events in the given signal region, $\lambda(\mathbf{x})\approx 1$. In the presence of signal, the magnitude of the density $p(\mathbf{x}|\text{data})$ should increase relative to $p(\mathbf{x}|\text{bkg})$, causing $\lambda(\mathbf{x})$ to fall below 1. In principle one can scan across a range of different signal regions with this technique to account for different localizations in m^*.

2.1.2. Bootstrap

While physically well-motivated, defining the signal region requires implicit supervision when fixing the central point m^* as well as the width Δm. This may be challenging without some a priori model-specific knowledge of the characteristics of the hypothesized signal process, such as the resonance width. Here we propose a simple method that bootstraps the current estimate of the likelihood ratio to achieve the desired separation. First we note that any test statistic

$t(\mathbf{x}) \in \mathbb{R}$ defines a rejection region $\mathcal{R} \subseteq \mathcal{X}$ where the null background-only hypothesis θ_0 is rejected if $\lambda(\mathbf{x}) \leq \epsilon$ for some threshold ϵ:

$$\mathcal{R} = \{\mathbf{x} \in \mathcal{X} \mid t(\mathbf{x}) \leq \epsilon\}\,. \tag{4}$$

The threshold ϵ can be determined by fixing an upper bound on the probability α of rejecting the null hypothesis when it is true and using the following approximation:

$$\begin{aligned}
\alpha &= \mathbb{P}\left(\mathbf{x} \in \mathcal{R} | \theta_0\right) \\
&= \int_{\mathbf{x}} d\mathbf{x}\, p(\mathbf{x}|\theta_0) \cdot \mathbb{I}\left[t(\mathbf{x}_n) \leq \epsilon\right] \\
&= \int_{\mathbf{x}} d\mathbf{x}\, p(\mathbf{x}|\theta_1)\lambda(\mathbf{x}) \cdot \mathbb{I}\left[t(\mathbf{x}_n) \leq \epsilon\right] \\
&= \int_{\mathbb{R}} dt\, p(t(\mathbf{x})|\theta_1)\lambda(\mathbf{x}) \cdot \mathbb{I}\left[t(\mathbf{x}_n) \leq \epsilon\right] \\
&\approx \frac{1}{N}\sum_{n=1}^{N} t(\mathbf{x}_n) \cdot \mathbb{I}\left[t(\mathbf{x}_n) \leq \epsilon\right],\ \mathbf{x}_n \sim p(\mathbf{x}|\theta_1).
\end{aligned} \tag{5}$$

Where $\mathbb{I}\left[\cdot\right]$ is the indicator function and the samples in the Monte Carlo estimate in the last line are drawn from the full data distribution. $p(\mathbf{x}|\theta_1)$ is estimated by some density model trained on the entirety of the dataset, without restriction. Events passing the selection requirement $t(\mathbf{x}) \leq \epsilon$ are classified as background events for the purposes of estimation of the likelihood ratio. During training, at each iteration, the background-only density $p(\mathbf{x}|\theta_0)$ is estimated by some density model trained only on samples $\mathbf{x}$ which pass this criterion. Then the current estimate of the likelihood ratio $\lambda(\mathbf{x})$ may be used to find the threshold ϵ demarcating the background events for the next iteration via Equation 5, given some fixed Type I error rate α, which is treated as a hyperparameter.

2.2. Normalizing Flows

Normalizing flows enable a complex density over continuous random variables $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^D$ to be expressed as a transformation of a simple base density $p_{\mathbf{z}}(\mathbf{z})$, where $\mathbf{z} \in \mathcal{Z} \subseteq \mathbb{R}^D$. This transformation is represented by an injective, differentiable function $f : \mathcal{Z} \to \mathcal{X}$ parameterized by θ (overloading previously defined notation). The resulting density $p_{\mathbf{x}}(\mathbf{x};\theta)$ may be related to the base density through a change of variables:

$$\begin{aligned}
\log p_{\mathbf{x}}(\mathbf{x};\theta) &= \log p_{\mathbf{z}}\left(f^{-1}(\mathbf{x};\theta)\right) + \log\left|\det \frac{\partial f^{-1}(\mathbf{x};\theta)}{\partial \mathbf{x}}\right| \\
&= \log p_{\mathbf{z}}(\mathbf{z}) - \log\left|\det \frac{\partial f(\mathbf{z};\theta)}{\partial \mathbf{z}}\right|.
\end{aligned} \tag{6}$$

Intuitively, the first term yields the log-likelihood of the transformed sample $f^{-1}(\mathbf{x})$ under the base density, while the second term accounts for the change in volume for regions in $\mathbb{R}^D$ under the transformation f to ensure normalization of the transformed $p_{\mathbf{x}}(\mathbf{x};\theta)$. This model density may be used to approximate the true data-generating distribution $p^*_{\mathbf{x}}(\mathbf{x})$ over $\mathbf{x}$ through minimization of the KL-divergence $D_{\mathrm{KL}}\left(p^*_{\mathbf{x}}(\mathbf{x}) \| p_{\mathbf{x}}(\mathbf{x};\theta)\right)$ with respect to θ. This is equivalent to optimizing θ to maximize the log-likelihood of samples $\mathbf{x} \sim p^*_{\mathbf{x}}(\mathbf{x})$.

In practice several tradeoffs exist when designing the transformation f (Papamakarios et al., 2019). This mapping is typically parameterized by some neural network architecture restricted to enable efficient evaluation of the determminant of the Jacobian of the transformation. This limitation can be partially mitigated by nothing that invertibility and differentiability are preserved under function composition. This allows us to build a more expressive transformation by composing appropriate transformations: $f = f_K \circ f_{K-1} \circ \cdots \circ f_1$. By the chain rule, the resultant log-density is given by:

$$\log p_{\mathbf{x}}(\mathbf{x};\theta) = \log p_{\mathbf{z}}(\mathbf{z}) - \sum_{k=1}^{K} \log\left|\det \frac{\partial f(\mathbf{z}_{k-1};\theta)}{\partial \mathbf{z}_{k-1}}\right|. \tag{7}$$

Further discussion on normalizing flows for density estimation is given in Appendix C. Nachman & Shih (2020) propose to use normalizing flow models to estimate the densities that appear in the likelihood ratio $\lambda(\mathbf{x})$. The authors use the Masked Autoregressive Flow model (Papamakarios et al., 2017) to compute the necessary densities over $\mathbf{x}$. Here the individual elements x_i are built as an autoregressive sequence of location-scale transformations:

$$\begin{aligned}
x_i &= f(z_i, \mathbf{x}_{<i};\theta) \\
&= z_i \exp\left(f_\sigma(\mathbf{x}_{<i})\right) + f_\mu\left(\mathbf{x}_{<i}\right),\ z_i \sim \mathcal{N}(0,1).
\end{aligned}$$

Where $\mathbf{x}_{<i}$ denotes all elements of $\mathbf{x}$ before element x_i according to some predefined ordering and f_μ and f_σ are neural networks with collective parameters θ. f is invertible by construction and hence the model density may be efficiently calculated using change of variables (Eq. 6), as the Jacobian of f is lower triangular.

2.3. Latent Variable Models

Density estimation is also possible using latent variable models. Here unobserved or 'latent' variables $\mathbf{z} \in \mathcal{Z}$ are introduced, together with a factorized joint density over the data and these auxiliary variables, parameterized by θ: $p_{\mathbf{x},\mathbf{z}}(\mathbf{x},\mathbf{z};\theta) = p_{\mathbf{x}}(\mathbf{x}|\mathbf{z};\theta)p_{\mathbf{z}}(\mathbf{z})$. In what follows we suppress this θ-dependence and the random variable subscript on densities for readability. The marginal log-likelihood, or log-model evidence, can be obtained through marginalization of the latent variables.

$$\log p(\mathbf{x}) = \log \int_{\mathbf{z}} d\mathbf{z}\, p(\mathbf{x}|\mathbf{z}) p(\mathbf{z}). \tag{8}$$

Direct evaluation of the model evidence is intractable in general as the conditional $p(\mathbf{x}|\mathbf{z})$ is typically parameterized through some nonlinear mapping. Instead a proposal, or variational, distribution $q(\mathbf{z};\mathbf{x})$ which admits tractable sampling and evaluation, is introduced. This leads to the following lower bound on the log-model evidence:

$$\begin{aligned}
\log p(\mathbf{x}) &= \log \int_{\mathbf{z}} d\mathbf{z}\, p(\mathbf{x}|\mathbf{z}) p(\mathbf{z}) \frac{q(\mathbf{z};\mathbf{x})}{q(\mathbf{z};\mathbf{x})} \\
&= \log \mathbf{E}_{q(\mathbf{z};\mathbf{x})} \left[\frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{q(\mathbf{z};\mathbf{x})} \right] \\
&\geq \mathbf{E}_{q(\mathbf{z};\mathbf{x})} \left[\log \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{q(\mathbf{z};\mathbf{x})} \right] \triangleq \mathcal{L}(q)
\end{aligned}$$

Where the last inequality comes from Jensen's inequality for concave functions. $\mathcal{L}(q)$ may be efficiently optimized under maximum likelihood with respect to the distribution q to obtain a biased estimator of $\log p(\mathbf{x})$. The distributions $p(\mathbf{x}|\mathbf{z})$ and $q(\mathbf{z};\mathbf{x})$ are typically parameterized by neural networks. Burda et al. (2016) demonstrate that the tightness of this bound can be improved by using the following multisample importance-weighted lower bound:

$$\log p(\mathbf{x}) \geq \mathbf{E}_{q(\mathbf{z};\mathbf{x})} \left[\log \frac{1}{K} \sum_{k=1}^{K} \frac{p(\mathbf{x}|\mathbf{z}_k)p(\mathbf{z}_k)}{q(\mathbf{z}_k|\mathbf{x})} \right] \triangleq \mathcal{L}_K(q) \tag{9}$$

Where $K = 1$ corresponds to the single-sample bound $\mathcal{L}(q)$. The estimator $\mathcal{L}_K(q)$ is consistent in the sense that $\log p(\mathbf{x}) = \lim_{K\to\infty} \mathcal{L}_K(q)$ and is monotonically non-decreasing: $\mathcal{L}_{K+1}(q) \geq \mathcal{L}_K(q)$. This allows us to tradeoff computation time for tightness in the lower bound. Building on this framework, Luo et al. (2020) develop an unbiased estimator of $\log p(\mathbf{x})$ - the Stochastically Unbiased Marginalization Objective, abbreviated SUMO. Concretely, they observe that the marginal log-likelihood can be expressed as a telescoping series:

$$\log p(\mathbf{x}) = \mathcal{L}_m(q) + \lim_{K\to\infty} \sum_{k=m}^{K} \left(\mathcal{L}_{k+1}(q) - \mathcal{L}_k(q) \right). \tag{10}$$

An unbiased Monte Carlo estimate of the series, and therefore $\log p(\mathbf{x})$, can be constructed via the 'Russian Roulette' procedure - randomly truncating a modified version of this series with appropriately weighted terms to account for the truncation (Forsythe & Leibler, 1950).

The mappings used to parameterize the conditional $p(\mathbf{x}|\mathbf{z})$ and variational distribution $q(\mathbf{z};\mathbf{x})$ do not have to be designed to maintain a tractable Jacobian as in the case of discrete normalizing flows as the lower bound $\mathcal{L}_K(q)$ is tractable by construction, allowing the use of very expressive architectures to represent these mappings. However, the quantity being optimized is a lower bound on $\log p_{\mathbf{x}}(\mathbf{x};\theta)$ instead of the true log-model evidence (excepting the SUMO estimator). As the variational distribution is restricted to a family of distributions that supports efficient sampling and evaluation, an overly simple choice of family may result in a highly biased estimator of the marginal likelihood (van den Berg et al., 2018).

3. Experiments

Here we evaluate empirically the performance of different density estimation models used to estimate $p(\mathbf{x}|\text{bkg}) = p(\mathbf{x}|\theta_0)$ and $p(\mathbf{x}|\text{data}) = p(\mathbf{x}|\theta_1)$ in their respective regions of the mass spectrum.

3.1. Dataset

We consider the same simulated dataset from the Large Hadron Collider Olympics 2020 Challenge (Kasieczka et al., 2019) considered in Nachman & Shih (2020). This consists of one million large-radius quantum chronomdynamic QCD dijet events as well-understood background processes. The signal process is defined by the decay of a hypothetical electroweak boson W' with mass $M_{W'} = 3.5$ TeV into a (hypothetical) X boson ($M_X = 500$ GeV) and Y boson ($M_Y = 100$ GeV), $W' \to XY$. The X and Y bosons are unstable and rapidly decay into a single large-radius 'jet' of stable hadronic particles. The observable quantities $\mathbf{x}$ in this dataset originate from simulated collider sensor measurements of properties of these hadronic decay products. We consider four high-level observables derived from the jet substructure of each event:

- M_{J_1}: The invariant mass of the most massive reconstructed jet, referred to as the leading jet.
- $\Delta M_{J_1,J_2}$: The difference in invariant mass between the leading and subleading jets.
- $\tau_{21}^{J_1}, \tau_{21}^{J_2}$: The n-subjettiness ratio (Thaler & Van Tilburg, 2012) for the leading and subleading jets, respectively.

The localizing feature m is taken to be the invariant mass M_{JJ} of the two leading jets in the event (Figure 1). Note the peak in signal events near the mass of the progenitor W'. The signal region is defined as $M_{JJ} \in [3.3, 3.7]$ TeV, and the sideband region as the complement. To simulate

the relative sparsity of signal events, 1000 randomly selected signal events are injected into the full background sample. The full dataset is divided into a train and test sample in an equal ratio. Within the train sample approximately 50000 events are held out for validation. As noted in Dinh et al. (2017), we scale each feature in $\mathbf{x}$ to the unit interval $[0, 1]$ and apply the smoothing transformation $g : [0, 1]^D \to \mathbb{R}^D, g(\mathbf{x}) = \text{Logit}\,(\epsilon + (1 - \epsilon)\mathbf{x})$ to circumvent boundary effects during density estimation. The Jacobian of this map is accounted for when reporting results and computing the test statistic $\lambda(\mathbf{x})$.

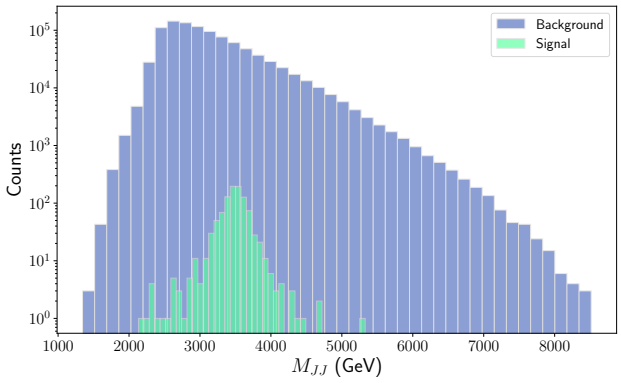

Figure 1. Distribution of invariant mass of the two leading jets for each event.

3.2. Results

To assess a fairly representative subset of flow models applied to density estimation, we evaluate one model based on partitioned affine transformations (Real-NVP, Dinh et al. (2017)), one model based on autoregressive flows (Masked Autoregressive Flows, Papamakarios et al. (2017)) and another based on continuous flows (FFJORD, Grathwohl et al. (2019)). Appendix C provides further discussion of each of these models. For the latter we use exact evaluation of the Jacobian trace at test time, as opposed to the use of the Monte Carlo Hutchinson trace estimator during training. In addition to the aforementioned flow methods, we consider both the class of importance-weighted estimators $\mathcal{L}_K(q)$ and SUMO for the task of density modelling. Further details of our experimental setup and reporting of results may be found in Appendix A.

To begin, we consider a simple baseline where the model log-likelihood $\log p_{\mathbf{x}}(\mathbf{x}; \theta)$ is used to directly score samples $\mathbf{x}$. The rationale here is that out of distribution data should be poorly modelled and assigned a low likelihood by the density model. As hypothetical signal events are likely to be absent or very scarcely represented in the empirical training distribution, these should be scored lower on average under the model. A decision function can be defined through a threshold on $\log p(\mathbf{x}; \theta)$, chosen to achieve an acceptable false positive rate. For this baseline, each model is trained on the entirety of the training set. Results are reported in Table 1. Despite using a continuous dataset, we report results in Bits per Dimension: BPD $= -\log_2 p_{\mathbf{x}}(\mathbf{x}; \theta)/D$, where D is the dimension of $\mathbf{x}$. This can be interpreted as a relative measure of the description length - a high BPD indicates that the model poorly explains the given data. We observe that the marginal likelihood provides a strong baseline, with all density models achieving similar performance for all metrics reported. In general, this method underperforms methods based on using the likelihood ratio as a test statistic, because it does not receive the advantage of implicit supervision afforded the latter. However, for this same reason, it is more widely applicable.

Returning to the separation through the likelihood ratio (Eq. 3), Figures 3 and 4 show qualitative results in the signal region in the test dataset for the signal/sideband region method. Here we consider the inverse of the test statistic for easier visualization. In Figure 3 the distribution of $\log p(\mathbf{x}|\text{bkg})$ against $\lambda^{-1}(\mathbf{x})$ is visualized. We observe that $\lambda(\mathbf{x}) \approx 1$ for the majority of background samples under each model. We observe that signal and background events are largely separated in this plane - background events are mostly concentrated around $\lambda(\mathbf{x}) = 1$, as expected, while signal events have a longer 'tail' at lower values of $\lambda(\mathbf{x})$ where background is less probable. While the MAF and Real-NVP models learn to keep background events localized well in this plane, background events are more diffusely distributed in the case of FFJORD and the importance-weighted autoencoder models. This can be attributed to the stochasticity present during optimization for these models. The training procedure for FFJORD involves Monte Carlo estimation of the Jacobian trace through the Hutchinson trace estimator, while stochasticity is inherent when sampling importance particles to construct the importance-weighted lower bound. This leads to predictions that are higher in variance relative to the considered discrete normalizing flow models, although we speculate this may have a regularizing effect on generalization.

Figure 4 makes the separation of classes more explicit by visualizing the $(\log p(x|\theta_0), \log p(x|\theta_1))$ plane. Again, most events are concentrated along the diagonal of this plane, corresponding to $\lambda(\mathbf{x}) = 1$, with signal events being concentrated at low values of and background events concentrated at higher values of $\lambda(\mathbf{x})$.

More quantitatively, in Figure 2 we examine the outcome of using $\lambda(\mathbf{x})$ as a threshold to distinguish between signal and background events. The curves are obtained by scanning over $\lambda(\mathbf{x})$ and computing the signal efficiency and background rejection rates (the inverse of background efficiency)

Table 1. Model performance using $\log p(\mathbf{x})$ as a scoring function, reported in the average Bits per Dimension for all events (BPD), for signal and background events (Signal BPD and Bkg. BPD, respectively) over the test dataset, and the area under the curve (AUC) obtained by thresholding $\log p(\mathbf{x})$. For the purposes of density modelling, lower BPD is better.

MODEL	BPD	SIGNAL BPD	BKG. BPD	$\lvert\Delta_{\text{BPD}}\rvert$	AUC
$\mathcal{L}_{16}(q)$	-0.739 ± 0.001	-0.076 ± 0.007	-0.743 ± 0.001	0.667 ± 0.007	0.814 ± 0.002
SUMO	-0.748 ± 0.001	-0.149 ± 0.006	-0.751 ± 0.001	0.602 ± 0.006	0.818 ± 0.004
$\mathcal{L}_{1024}(q)$	-0.742 ± 0.001	-0.067 ± 0.007	-0.746 ± 0.001	$\mathbf{0.679 \pm 0.007}$	0.815 ± 0.002
FFJORD	-0.744 ± 0.001	-0.060 ± 0.018	-0.747 ± 0.001	$\mathbf{0.687 \pm 0.017}$	0.817 ± 0.003
REAL-NVP	-0.747 ± 0.000	-0.136 ± 0.001	-0.750 ± 0.000	0.614 ± 0.001	$\mathbf{0.826 \pm 0.001}$
MAF	$\mathbf{-0.758 \pm 0.001}$	-0.215 ± 0.001	-0.761 ± 0.000	0.547 ± 0.000	0.807 ± 0.001

at each step. Again, FFJORD and the importance-weighted estimator outperform the MAF and Real-NVP models significantly at high signal efficiencies, but is less performant at lower signal efficiencies due to the high variance of predictions leading to a higher level of background contamination in the $\lambda(\mathbf{x}) < 1$ region. The bootstrap method is competitive with the signal/sideband region method at high signal efficiencies, but suffers poor performance in the low signal efficiency region, possibly as the supervision signal from self-training is too weak to distinguish background samples which mimic the signal in the considered phase space.

In Table 2 we summarize the results of methods which use the likelihood ratio as a test statistic, observing that the bootstrap method is able to significantly outperform the baseline and achieve comparable performance to the signal/sideband region method without receiving any implicit supervision in the form of hand-selected regions of high background purity.

4. Conclusion

We have shown in this preliminary study that normalizing flows and latent variable models are viable candidates for unsupervised model-independent search methods for signatures of new physics at collider experiments. Unlike standard model-dependent search methods based on supervised learning, unsupervised methods decouple the search procedure from the exact model choice, mitigating a potentially problematic reliance on expensive Monte Carlo simulations of hypothesized signal processes, which are tied to a particular model and may not reflect the true data-generating distribution. Instead of a replacement for standard search techniques, the models studied herein may be used in the early stages of the data processing stream at collider experiments to isolate potentially interesting events for further analysis. Directions for future work may consider alternative weak supervision methods that can be used to find regions of the phase space with high background purity, and investigate if superior density estimation directly translates to an increase in signal significance.

Acknowledgements

This research was undertaken using the LIEF HPC-GPGPU Facility hosted at the University of Melbourne. This Facility was established with the assistance of LIEF Grant LE170100200.

References

Burda, Y., Grosse, R., and Salakhutdinov, R. Importance weighted autoencoders. In *International Conference on Learning Representations (ICLR)*, 2016.

Chen, R. T. Q., Li, X., Grosse, R., and Duvenaud, D. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, 2018a.

Chen, R. T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. Neural ordinary differential equations. *Advances in Neural Information Processing Systems*, 2018b.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using real NVP. In *International Conference on Learning Representations (ICLR)*, 2017.

Forsythe, G. E. and Leibler, R. A. Matrix inversion by a monte carlo method. *Mathematical Tables and Other Aids to Computation*, 4(31):127–129, 1950. ISSN 08916837.

Grathwohl, W., Chen, R. T. Q., Bettencourt, J., Sutskever, I., and Duvenaud, D. FFJORD: Free-form continuous dynamics for scalable reversible generative models. In *International Conference on Learning Representations (ICLR)*, 2019.

Kasieczka, G., Nachman, B., and Shih, D. R&D Dataset for LHC Olympics 2020 Anomaly Detection Challenge, April 2019. URL https://doi.org/10.5281/zenodo.3832254.

Luo, Y., Beatson, A., Norouzi, M., Zhu, J., Duvenaud, D., Adams, R. P., and Chen, R. T. Q. Sumo: Unbiased estimation of log marginal probability for latent variable models.

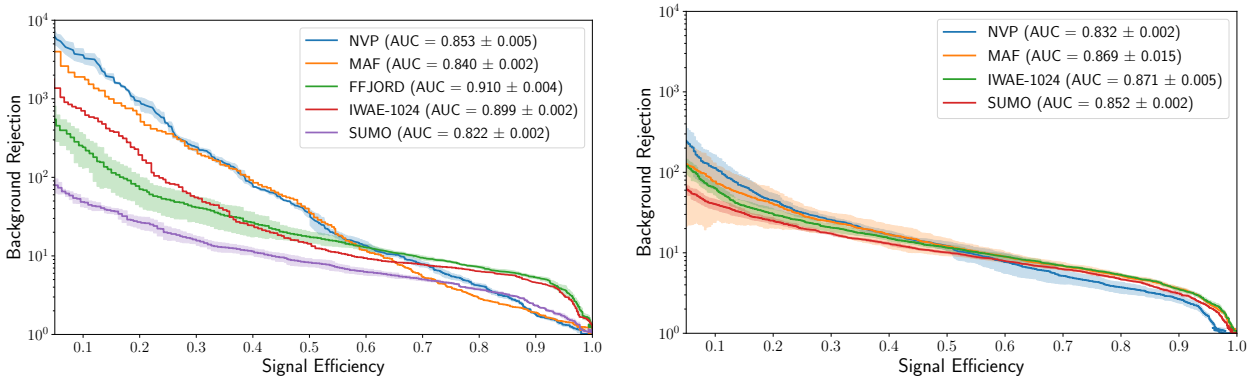

Figure 2. Background rejection versus signal efficiency curves over the signal region in the test set using $\lambda(\mathbf{x})$ obtained using the signal/sideband region method (left) and the bootstrap method (right) as a decision function. Results are given as the mean and standard deviation over 3 trials with different random seeds. The uncertainty bands show the 1σ deviation from the mean.

In *International Conference on Learning Representations*, 2020.

Nachman, B. and Shih, D. Anomaly detection with density estimation. *Phys. Rev. D*, 101, 2020.

Papamakarios, G., Pavlakou, T., and Murray, I. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems 30*, 2017.

Papamakarios, G., Nalisnick, E., Jimenez Rezende, D., Mohamed, S., and Lakshminarayanan, B. Normalizing Flows for Probabilistic Modeling and Inference. *arXiv*, art. arXiv:1912.02762, 2019.

Thaler, J. and Van Tilburg, K. Maximizing boosted top identification by minimizing n-subjettiness. *Journal of High Energy Physics*, (2), 2012.

van den Berg, R., Hasenclever, L., Tomczak, J., and Welling, M. Sylvester normalizing flows for variational inference. In *proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2018.

A. Experimental Details

We used Adam with a learning rate of 1×10^{-4}, $\beta_1 = 0.9$ and $\beta_2 = 0.999$ for all experiments, trained with early stopping if no improvements have been observed for 8 epochs on the validation set, with a maximum of 42 epochs permitted (except for FFJORD, which is trained for a maximum of 16 epochs). Hyperparameters were not significantly tuned. as we observed that the 'default' settings for each model yielded reasonable results. The specific architectures used for each model are the following:

- MAF: 16 discrete flows. The autoregressive location and scale transformations are parameterized by neural networks with 3 hidden layers with hidden dimension 64 with ReLU activation. The autoregressive ordering to generate the conditional $p(x_i|\mathbf{x}_{<i})$ is the same as the original data ordering. Total parameter count: 586,240.

- Real-NVP: 16 discrete flows. The partitioned location and scale transformations are parameterized by neural networks with 3 hidden layers with hidden dimension 64 with LeakyReLU activation. Total parameter count: 276,544.

- FFJORD: 2 integrating ODE blocks, with 3 hidden layers per block with hidden dimension 256 and Tanh activation. Batch normalization is applied between blocks. Total parameter count: 272,482.

- Importance-weighted estimator: Both the inference and generative networks have the same architecture: 2 hidden layers with hidden dimension 128 and ELU activation. The dimension of the latent space is set to 4. The prior is an isotropic Gaussian with identity covariance, the proposal distribution $q_\phi(z|x)$ is a Gaussian with mean and diagonal covariance parameterized by ϕ, and similarly for the conditional distribution $p_\theta(x|z)$. We scan over the number of importance samples K in the range $\{2^i\}_{i=5}^{10}$ and report the results in Appendix B. Provided the data, together with the associated importance samples, can be held in memory, the memory requirement will scale as $O(K)$ while computation time scales sublinearly. Total parameter count: 74,024.

- SUMO: Identical architecture to the importance-weighted estimator. The stopping time, denoted by K, is a random variable sampled for each batch. We use the CDF proposed by the authors:

$$\mathbb{P}(\mathcal{K} \geq k) = \begin{cases} \frac{1}{k}, & \text{for } k < 80 \\ \frac{1}{\alpha}\,(0.9)^{k-\alpha}, & k \geq 80 \end{cases} \tag{11}$$

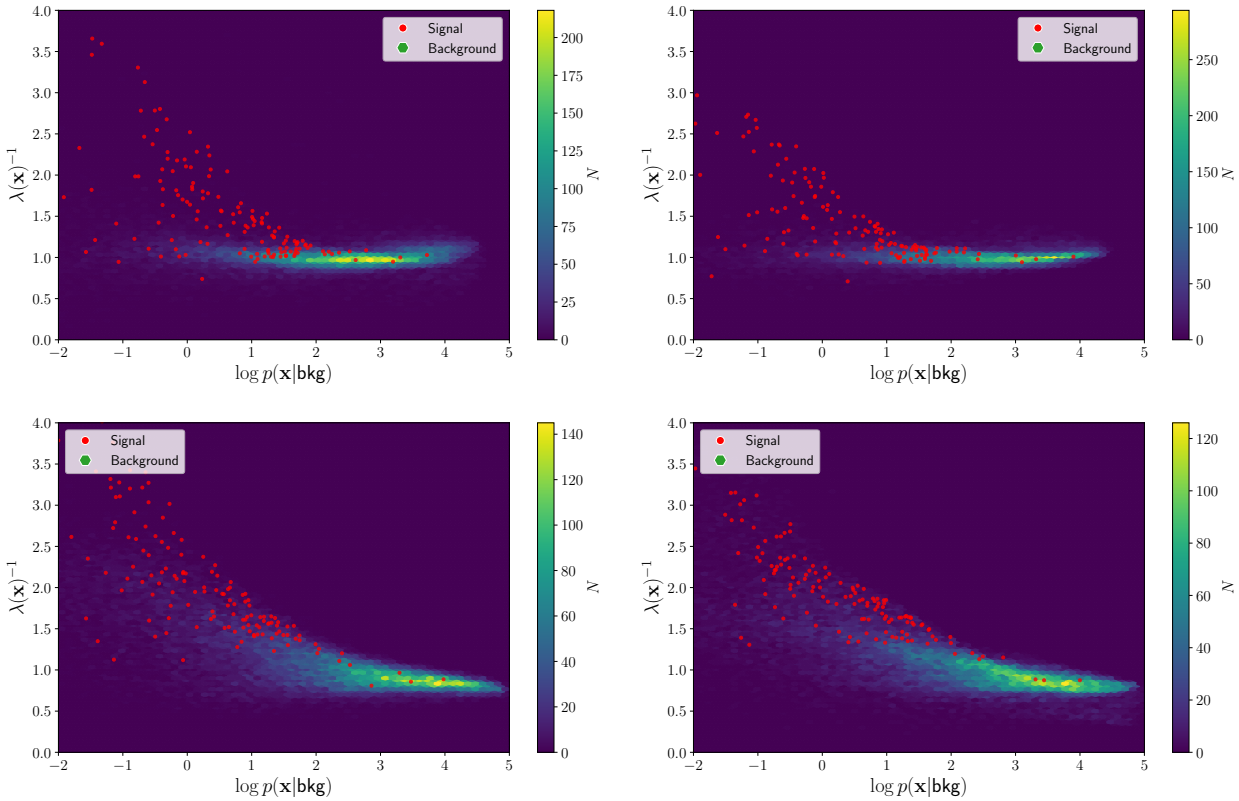

Figure 3. Distribution of $\log p(\mathbf{x}|\theta_0)$ versus $\lambda(\mathbf{x})^{-1}$ in the signal region in the test dataset. Background events are shown as a 2D histogram while signal events are overlaid in red. Clockwise from top left quadrant: MAF, Real-NVP, FFJORD, and the Importance-Weighted Estimator with 1024 samples.

Note the memory and computation requirements will depend on the sampled value of K. We use a minimum stopping time of 16, giving an expected stopping time $K \approx 16$.

When reporting results models are ordered, ascending, by number of parameters, with ties broken by expected computation time. Results are given as the mean and standard deviations over 3 trials with different random seeds. Bolded entries indicate results that are statistically insignificant from the the best performing model within each column, according to an unpaired t-test with significance level 0.05.

Unlike Nachman & Shih (2020), we do not provide the resonant mass M_{JJ} as contextual information to each density model in our reported results using the signal region/sideband approach, finding that this hurt training stability. We also find that training is sufficiently stable for all models to omit the multi-epoch averaging they used to calculate the density estimate returned by the background model.

When training the bootstrap estimator we set the Type I error rate $\alpha = 0.1$, although we find that values of α from 0.05 to 0.2 do not significantly change $\lambda(\mathbf{x})$ estimated on the validation set on each of the models. We use bisection search to find the value of the threshold ϵ corresponding to α using Equation 5 where the current batch is used as the empirical sample.

During optimization of the bootstrap procedure for discrete normalizing flow models (Real-NVP, MAF), we find that the log-likelihood assigned to samples in the signal region could assume very large negative values due to near-zero probability estimates of these datapoints under the base standard Gaussian distribution. This meant that the background density model was subject to large gradients during training. As an alternative to gradient norm clipping, we resolved the issue by using the heavy-tailed Student's t-distribution with number of degrees of freedom $\nu = 1$ as the base distribution $p_{\mathbf{z}}(\mathbf{z})$ for discrete normalizing flows. Density estimation based on latent variable models suffered this problem to a lesser extent, and we apply gradient clipping with a maximum 2-norm of 0.1 to the background model. We find that continuous normalizing flows do not suffer this issue, likely due to the regularizing effect of the Monte Carlo trace estimator preventing any datapoint from being assigned an unstably low probability during training.

Table 2. Model performance using the likelihood ratio $\lambda(\mathbf{x})$ as a scoring function for the signal/sideband region method exploiting localization of signal in the resonant mass (Nachman & Shih, 2020) (left subtable), and the bootstrap method (right subtable) proposed in Section 2.1.2. Results are reported in the average $\lambda(\mathbf{x})$ of signal and background events over the test dataset (Sig. $\lambda(\mathbf{x})$ and Bkg. $\lambda(\mathbf{x})$, respectively), and the area under the curve (AUC) obtained by thresholding $\log \lambda(\mathbf{x})$.

	Mass Localization				Bootstrap			
Model	Sig. $\lambda(\mathbf{x})$	Bkg. $\lambda(\mathbf{x})$	$\lvert\Delta_{\lambda(\mathbf{x})}\rvert$	AUC	Sig. $\lambda(\mathbf{x})$	Bkg. $\lambda(\mathbf{x})$	$\lvert\Delta_{\lambda(\mathbf{x})}\rvert$	AUC
SUMO	0.681 ± 0.022	1.084 ± 0.010	0.403 ± 0.018	0.822 ± 0.002	0.266 ± 0.012	1.007 ± 0.010	0.741 ± 0.005	0.852 ± 0.002
$\mathcal{L}_{1024}(q)$	0.549 ± 0.006	1.040 ± 0.045	$\mathbf{0.491 \pm 0.040}$	0.899 ± 0.002	0.147 ± 0.028	0.916 ± 0.008	0.769 ± 0.021	0.871 ± 0.005
FFJORD	0.560 ± 0.026	1.031 ± 0.009	$\mathbf{0.471 \pm 0.027}$	$\mathbf{0.910 \pm 0.004}$				
Real-NVP	0.761 ± 0.011	1.009 ± 0.001	0.248 ± 0.010	0.853 ± 0.005	0.326 ± 0.022	1.017 ± 0.005	0.691 ± 0.026	0.832 ± 0.002
MAF	0.740 ± 0.006	1.063 ± 0.007	0.323 ± 0.007	0.840 ± 0.002	0.062 ± 0.002	1.030 ± 0.004	0.968 ± 0.005	0.869 ± 0.015

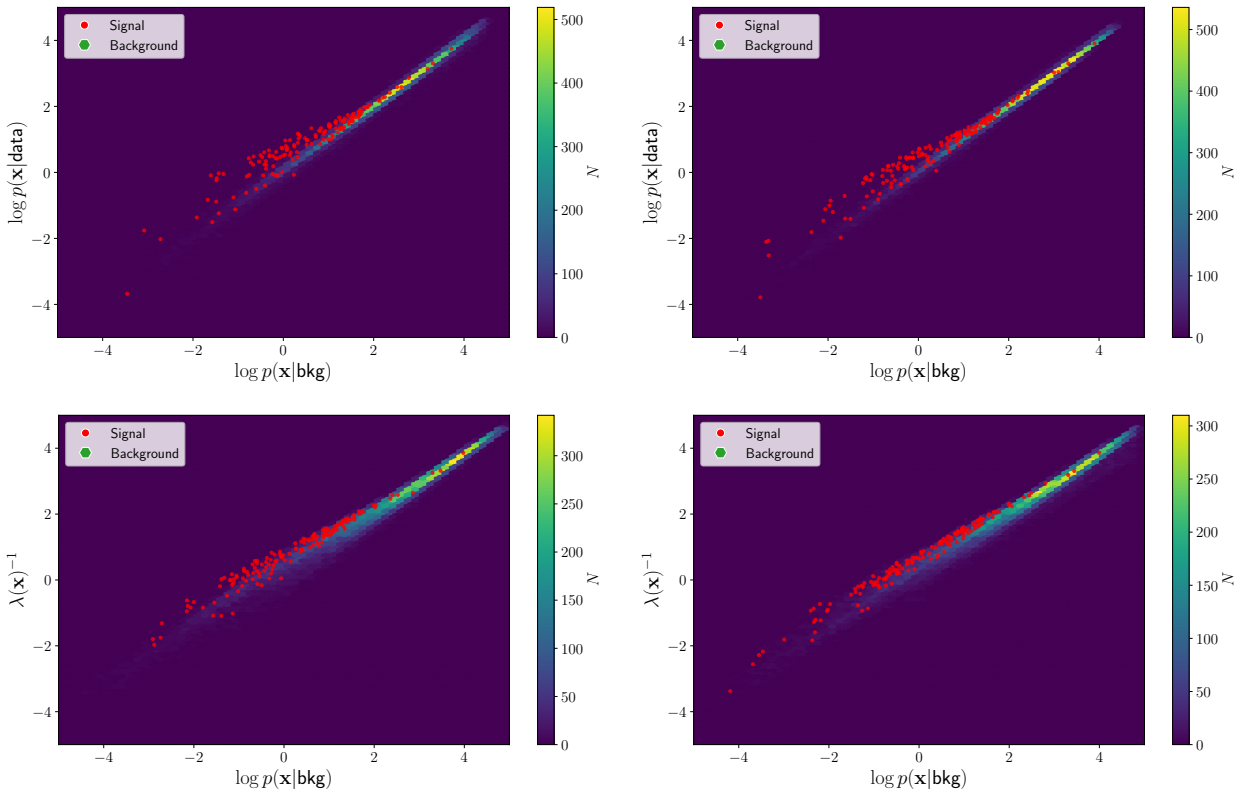

Figure 4. Distribution of $\log p(\mathbf{x}|\theta_0)$ (background-only hypothesis) versus $\log p(\mathbf{x}|\theta_1)$ (signal-plus-background hypothesis) in the signal region in the test dataset. Background events are shown as a 2D histogram while signal events are scattered in red. Clockwise from top left quadrant: MAF, Real-NVP, FFJORD, and the Importance-Weighted Estimator with 1024 samples.

B. Effect of K on $\mathcal{L}_K(q)$ estimator

In Figure 5 we observe the effect of increasing the number of importance samples K when estimating the required densities using the importance-weighted estimator $\mathcal{L}_K(q)$, using the signal region/sideband method. Increasing K appears to result in modest improvements in separation power, albeit at significantly diminishing returns, suggesting that the variance of the importance weighted estimator becomes sufficiently low at high K such that the gap between $\mathcal{L}_K(q)$ and $\log p(\mathbf{x})$ becomes negligible.

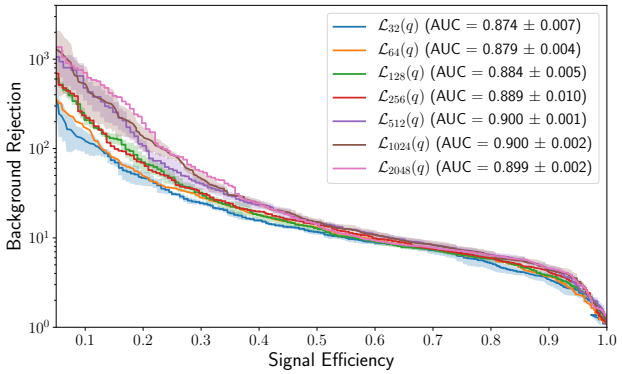

Figure 5. Background rejection versus signal efficiency curves over the signal region in the test set for the $\mathcal{L}_K(q)$ estimator with different number of importance samples K. Results are given as the mean and standard deviation over 3 trials with different random seeds. The uncertainty bands show the 1σ deviation from the mean.

C. Normalizing Flows for Density Estimation

Normalizing flows can be divided into three main categories (Papamakarios et al., 2019):

- *Autoregressive flows:* Here each element of the output of the transformation, z_i' is represented as a mapping τ from the corresponding input index z_i, parameterized by a family of functions $\mathbf{g}$:

$$z_i' = \tau\left(z_i; \mathbf{g}_i(\mathbf{z}_{<i})\right)$$

  Here g_i is constrained to depend only on elements of the input $\mathbf{z}$ before element z_i according to some predefined ordering, denoted by $\mathbf{z}_{<i}$ (hence the name *autoregressive*). The mapping τ is constrained to be a strictly monotonic function of z_i' to ensure invertibility. Because of the autoregressive dependence of z_i' on only $\mathbf{z}_{<i}$, the Jacobian is upper triangular by construction and can be computed in $O(D)$. A popular choice of mapping τ is a location-scale transformation, due to their ease of inversion (Dinh et al., 2017):

$$\tau(z_i; \mathbf{g}_i) = \alpha_i(\mathbf{z}_{<i})z_i + \beta(\mathbf{z}_{<i})$$

  Where invertibility of the flow requires that α_i is nonzero. Because of the existence of an analytical inverse and tractable Jacobian by construction, autoregressive flow models have been widely applied to density estimation.

- *Residual flows:* Here the transformation f is defined as the identity mapping plus a correction term with parameters ϕ.

$$\mathbf{z}' = \mathbf{z} + g_\phi(\mathbf{z})$$

  Residual flows are not used for direct density estimation due to the absence of an analytical inverse for this general class of flows. However, residual flows can be indirectly applied to density estimation through latent variable models, where they can be used to derive a more expressive variational posterior to reduce the bias in the optimized bound on the marginal log likelihood (van den Berg et al., 2018).

- *Continuous-time flows:* Here we assume that the flow state $\mathbf{z}$ is indexed by a continuous scalar parameter t, referred to as 'time'. Continuous time flows parameterize the state time derivative with a neural network with parameters ϕ, $g_\phi : \mathbb{R}^D \times \mathbb{R} \to \mathbb{R}^D$, through the following ordinary differential equation, with corresponding boundary conditions:

$$\frac{d\mathbf{z}(t)}{dt} = g_\phi(\mathbf{z}(t), t), \;\; \mathbf{z}(t_0) = \mathbf{u}, \;\; \mathbf{z}(t_1) = \mathbf{x}$$

  The 'forward' transformation can be computed via numerical integration:

$$\mathbf{z}(t_1) = \mathbf{x} = \mathbf{z}(t_0) + \int_{t_0}^{t_1} dt\, g_\phi(\mathbf{z}(t), t).$$

  This may be easily inverted by exchanging the limits of integration. The change in log-density obeys a second differential equation (Chen et al., 2018a):

$$\frac{d \log p(\mathbf{z}(t))}{dt} = -\nabla \cdot g(\mathbf{z}(t), t)$$

  Where the divergence of the vector field g_ϕ, appears on the right hand side, which results in the following continuous version of the change of variables:

$$\log p_{\mathbf{x}}(\mathbf{x}) = \log p_{\mathbf{u}}(\mathbf{u}) - \int_{t_0}^{t_1} dt\, \nabla \cdot g(\mathbf{z}(t), t)$$

  This may be computed using a numerical integrator and optimized through the adjoint sensitivity method (Chen et al. (2018b), Grathwohl et al. (2019)).