HHS Public Access

Author manuscript

Proc Mach Learn Res. Author manuscript; available in PMC 2022 July 19.

Published in final edited form as:

Proc Mach Learn Res. 2021 July ; 139: 12427–12436.

Understanding Failures in Out-of-Distribution Detection with Deep Generative Models

Lily H. Zhang[1], **Mark Goldstein**[1], **Rajesh Ranganath**[1]

[1]New York University.

Abstract

Deep generative models (DGMs) seem a natural fit for detecting out-of-distribution (OOD) inputs, but such models have been shown to assign higher probabilities or densities to OOD images than images from the training distribution. In this work, we explain why this behavior should be attributed to model misestimation. We first prove that no method can guarantee performance beyond random chance without assumptions on which out-distributions are relevant. We then interrogate the *typical set hypothesis*, the claim that relevant out-distributions can lie in high likelihood regions of the data distribution, and that OOD detection should be defined based on the data distribution's typical set. We highlight the consequences implied by assuming support overlap between in- and out-distributions, as well as the arbitrariness of the typical set for OOD detection. Our results suggest that estimation error is a more plausible explanation than the misalignment between likelihood-based OOD detection and out-distributions of interest, and we illustrate how even minimal estimation error can lead to OOD detection failures, yielding implications for future work in deep generative modeling and OOD detection.

1. Introduction

Predictive models have little guarantee in performance on inputs that differ from the training distribution. Thus, detecting such out-of-distribution (OOD) inputs is an important step towards safe and reliable machine learning (Amodei et al., 2016). OOD detection has been formalized as the task of identifying points with low likelihood[1] under the training distribution, estimated via a model (Bishop, 1994).

Deep generative models (DGMs) estimate complex distributions from often high-dimensional inputs and produce high-quality simulations (Salimans et al., 2017; Chen et al., 2018; Kingma & Dhariwal, 2018). However, explicit likelihood DGMs (e.g. autoregressive models, normalizing flows) have been shown to assign higher likelihoods to unrelated inputs than even those from the training distribution. For instance, a model trained on Fashion-MNIST, an image dataset of clothing items, assigns higher likelihoods to MNIST images. The same is true for the training distribution (or in-distribution) CIFAR-10, a dataset of animals and vehicles, and the OOD distribution (or out-distribution) SVHN, a dataset of house numbers. For such

Correspondence to: Lily H. Zhang < lily.h.zhang@nyu.edu>.

[1]As is common in the literature, we will use "likelihood" to refer to the probability or density of a sample under a distribution, even though the statistical definition of the term refers to a function of the parameters given fixed data.

dataset pairs, OOD detection based on explicit likelihood DGMs performs worse than random chance (Nalisnick et al., 2019a; Hendrycks et al., 2019).

This observation has motivated many alternative methods for OOD detection which employ the same DGMs but modify how they are used (Ren et al., 2019; Serrà et al., 2020; Schirrmeister et al., 2020; Choi et al., 2018; Nalisnick et al., 2019b; Wang et al., 2020; Morningstar et al., 2021). While these methods have been successful in empirical benchmarks, we prove that all methods are powerless against some set of out-distributions (Section 2.2). This result applies to any detection method, regardless of whether DGMs are involved, and highlights the need to specify the out-distributions of interest for the task.

Some works have suggested that the failure of deep generative models to assign low likelihoods to OOD points is not a model failure; rather, out-distributions of interest can lie in high likelihood regions of the data distribution. To explain why points with high likelihood are never observed among samples from the data distribution, these works mention that points assigned high density or probability under the training distribution may lie within regions of small overall probability. A method that identifies low likelihood points as OOD will fail to detect such out-distributions, so existing works suggest instead to flag as OOD any point which falls outside a distribution's typical set, a set that contains the majority of the probability mass of a distribution but not necessarily the highest density or probability points (Choi et al., 2018; Nalisnick et al., 2019b; Wang et al., 2020; Morningstar et al., 2021). This *typical set hypothesis*—the idea that relevant out-distributions are determined based on the typical set of a distribution—assumes that OOD regions can lie in the support of the data distribution.

In this work, we highlight problems with the typical set hypothesis (Section 3.1). First, the hypothesis assumes that relevant out-distributions (e.g. SVHN) can overlap in support with the data distribution (e.g. CIFAR-10). However, when the in- and out-distribution overlap, there is an irresolvable upper bound on OOD detection performance (Section 3.2), and even a perfect model of the in-distribution can yield worse OOD detection than a misestimated one (Section 3.3). A preference for the typical set over other similar sets is also arbitrary for OOD detection (Section 3.4). These results highlight the implausibility of the typical set hypothesis and its support overlap assumption.

In our experiments, we offer empirical demonstrations of the analyses presented. First, given an OOD detection method and a specific in-distribution, we provide examples of out-distributions that the method fails to distinguish from the in-distribution (Section 4.1). Then, we showcase an instance where a partially-trained DGM yields better OOD detection than the true distribution of the data when supports overlap between the in- and out-distribution (Section 4.2).

Based on the implausible implications of the typical set hypothesis, we conclude that the high likelihoods assigned to certain OOD images are instead due to model estimation error. First, it is reasonable to believe that existing dataset pairs have disjoint (rather than overlapping) support, as one would not expect to draw a house number from the CIFAR-10 distribution, or a digit from the Fashion-MNIST distribution, even given infinite samples.

This implies that existing models mistakenly assign high probability or density where they should be assigning zero. We demonstrate how even a model with good generation quality and heldout likelihood can still exhibit OOD failures (Section 5.1). We then discuss what this perspective of estimation error implies for DGMs and OOD detection (Section 5.2). We illustrate how recent methods that were motivated by the typical set hypothesis may instead correct for model estimation error, and we suggest future modeling directions to improve DGMs for OOD detection.

2. Defining OOD detection

OOD detection has been defined as the task of identifying "whether a test example is from a different distribution from the training data" (Hendrycks & Gimpel, 2017). Here, we illustrate why it is critical to specify the out-distributions to consider. In fact, without any constraints on out-distributions, the task of OOD detection is impossible.

In this section, we first formalize the broadest form of OOD detection as a single-sample goodness-of-fit test (Section 2.1). We then prove that no method can guarantee better than random chance performance under this task definition (Section 2.2). We conclude that any formal analysis of an OOD detection method must take into account the out-distributions which define the task.

2.1. OOD Detection as Goodness-of-fit Testing

In its unconstrained form, OOD detection can be formalized as a single-sample hypothesis test (Nalisnick et al., 2019b; Serrà et al., 2020; Wang et al., 2020); given a sample $\mathbf{x}$, the test decides whether to reject the null hypothesis that a sample was drawn from the data distribution P, in favor of an alternative hypothesis that the sample came from a distribution other than P:

$$H_0 : \mathbf{x} \sim P$$

$$H_A : \mathbf{x} \sim Q \in \mathbb{Q},\ P \notin \mathbb{Q}.$$

The decision to reject or not reject the null hypothesis (i.e. mark a sample OOD) is based on the value of a predetermined test statistic ϕ, which can be any arbitrary function of a single sample $\mathbf{x}$. In our analysis, we focus on test statistics which directly utilize knowledge of the input distribution P or an estimate of it via a deep generative model P_θ. P_θ can either be a continuous distribution, as is the case for normalizing flows (Dinh et al., 2015; 2017; Kingma & Dhariwal, 2018), or a discrete distribution, as is the case for existing autoregressive models (Salimans et al., 2017; Oord et al., 2016a;b; Chen et al., 2018). An example of a test statistic is $\phi = \log p_\theta$, where $p_\theta(\mathbf{x})$ denotes either the probability of an observation for a discrete distribution or the density of an observation with respect to the Lebesgue measure for a continuous distribution. This test statistic is often accompanied by the rejection rule $\phi(\mathbf{x}) < k$, i.e. reject as OOD points where $\log p(\mathbf{x})$ is low. We discuss various

choices of test statistics, including those which do and do not use an estimate of P, in the related work in Section 6.

In order to determine whether an input should be processed through a given classifier, OOD tests must make decisions on a single sample at a time. This stands in contrast with most goodness-of-fit testing setups that make a decision based on a collection of samples. We discuss the challenges of this single-sample formulation in Section 2.2.

The quality of a test is measured by its ability to correctly detect OOD samples without flagging in-distribution samples as OOD. The proportion of OOD samples detected, known as the true positive rate or power of a test, can be plotted as a function of the proportion of in-distribution samples incorrectly rejected, known as the false positive rate or size of a test. This is equivalent to a receiver operating characteristic (ROC) curve, and the area under the curve (AUC) is the area under the power vs. size curve.

Using this general formulation of OOD detection, we can now interrogate OOD detection methods more broadly by abstracting away the choice of test statistic ϕ.

2.2. OOD Detection as a Single-Sample Distributional Test is Impossible

OOD detection defined as a single-sample goodness-of-fit test is a challenging classification task given that the out-distributions are unknown. To remove the effect of misestimation, we consider test statistics which can use knowledge of the true in-distribution P via its density or probability function, denoted $\phi_p : \mathcal{X} \to \mathbb{R}$. We now present an impossibility result: no test can do well against all alternatives.

Proposition 1. *Let P be the distribution under the null hypothesis H_0. Let μ be the measure associated with the distribution of test statistic $\phi_p(\mathbf{x})$ under the null. Then, assuming the conditional $\mathbf{x} \mid \phi_p(\mathbf{x})$ is not degenerate on a μ-non-measure zero set, there exists a set of alternative distributions $Q \in \mathcal{Q}$ where $Q \neq_d P$ and the test has power equal to the false positive rate. In other words, the test does no better than random guessing.*

Proof. See Appendix A. The proof sketch is as follows: First we construct distributions $Q \in \mathcal{Q}$ for which the distribution of $\phi_p(\mathbf{x})$ is the same but the distribution of $\mathbf{x} \mid \phi_p(\mathbf{x})$ differs when $\mathbf{x} \sim P$ and $\mathbf{x} \sim Q$ for all $\phi_p(\mathbf{x})$ in a non-measure-zero set Φ. This implies $q(\mathbf{x}) \neq_d p(\mathbf{x})$. We show that the power of the test for any rejection rule for such a pair P, Q is equal to the false positive rate for all false positive rates, which is equivalent to random guessing. □

Proposition 1 demonstrates that no test statistic can be useful for all possible out-distributions. In the context of single-sample distributional testing, all proposed test statistics trade off power against different out-distributions. This means that, without additional assumptions on the family of alternative hypotheses for OOD detection, no test statistic can be uniformly better across out-distributions than another. To build intuition behind the proposition, imagine that $\mathcal{X}$ is the space of d-dimensional reals $\mathbb{R}^d$ and the in-distribution has a density with respect to the d-dimensional Lebesgue measure. The test statistic is a function that maps from $\mathbb{R}^d \to \mathbb{R}$; thus, the statistic is a one-dimensional projection of the distribution. In the same way that not all differences in two multivariate distributions can be assessed by

looking at a single marginal, not all the differences between P and Q can be assessed by looking at their projections on the test statistic. This result is focused on the single-sample formulation of OOD detection and holds even for test statistics which are consistent in power asymptotically.

An Example: Using log p as a test statistic.—When the test statistic is the log probability or density, the set of alternative distributions $Q \in \mathcal{Q}$ that cannot be distinguished from P are those which yield the same distribution of log probabilities or densities under P. These are distributions which collapse any of the level sets of P. As an example in the discrete case, imagine a countable sample space and a distribution P where c of the elements are given the same probability. Any distribution Q which moves the total probability of the c elements in P to any subset of these elements will share the same distribution of probability under P. The analogue for continuous distributions $\mathbb{R}^d$ is collapsing level sets of dimension $\mathbb{R}^{d-1}$. We illustrate the phenomenon with an example in the continuous case in Section 4.1.

Proposition 1 emphasizes the need to specify the family of relevant out-distributions for OOD detection. For instance, likelihood ratio test statistics (Ren et al., 2019; Serrà et al., 2020; Schirrmeister et al., 2020) are optimal when the alternative hypothesis is correctly specified, but like all test statistics, they trade off power in some other alternative; therefore, comparing different likelihood ratios (and test statistics in general) is only useful when the family of out-distributions is formalized and standardized.

Like any test statistic, $\phi = \log p$ works well for some out-distributions (those whose samples have zero or low likelihood under the data distribution) but poorly for others (those whose samples have high likelihood under the data distribution). Whether the latter such out-distributions are relevant for OOD detection is a central question underlying our analysis of the typical set hypothesis.

3. The Implausibility of the Typical Set Hypothesis

Nalisnick et al. (2019a); Hendrycks et al. (2019) observed that DGMs trained on CIFAR-10 samples assign higher likelihoods to SVHN images, and DGMs trained on FashionMNIST samples assign higher likelihoods to MNIST images. The explanation for these observations can either be A. such OOD samples do have high likelihoods under the *data* distribution, or **B**. these OOD samples only have high likelihoods under the *model* distribution due to estimation error. The typical set hypothesis argues for the former, that out-distributions can lie in high probability or density regions of the data distribution. A test based on the log p test statistic and $\phi(\mathbf{x}) < k$ rejection rule lacks power, even under the perfect model, to detect out-distributions whose samples have high likelihood under the data distribution, and the typical set hypothesis assumes that SVHN is such an out-distribution relative to CIFAR-10, as is MNIST to Fashion-MNIST. In this section, we detail the typical set hypothesis (Section 3.1), reveal consequences which fall from its assumptions (Section 3.2, Section 3.3), and discuss its relevance for OOD detection (Section 3.4).

3.1. The Typical Set Hypothesis

The typical set hypothesis posits that 1. out-distributions of interest can lie in regions of high likelihood but small overall probability under the data distribution, and 2. to detect such distributions, OOD detection should take into account the data distribution's typical set[2] (Choi et al., 2018; Nalisnick et al., 2019a; Wang et al., 2020; Morningstar et al., 2021). Tests that reject low likelihood points will perform worse than random chance on out-distributions whose samples have high in-distribution likelihoods, but tests that consider the data distribution's typical set can have power over such out-distributions. Quoting Wang et al. (2020):

> Samples from a high-dimensional distribution will often fall on a typical set with high probability, but the typical set itself does not necessarily have the highest probability density at any given point. Per this line of reasoning, to determine if a test sample is an outlier, we should check if it falls on the typical set of the inlier distribution rather than merely examining its likelihood under a given deep generative model.

Given a distribution P, the typical set $A_\epsilon^{(n)}$ is the set of n-length sequences $(x_{i1}, \ldots, x_{in})$, $x_{ij} \overset{\text{i.i.d.}}{\sim} P$ whose empirical entropy is close to the entropy of P, i.e. $H(P) = -\mathbb{E}_{x_{ij}\sim P}\left[\log p\left(x_{ij}\right)\right]$, within a neighborhood determined via the constant ϵ (Cover & Thomas, 1991):

$$H(P) - \epsilon \leq -\frac{1}{n}\sum_{j=1}^{n} \log p\left(x_{ij}\right) \leq H(P) + \epsilon\,. \tag{1}$$

The typical set can be viewed as a set of elements from the sample space of the product measure $P^n = P \times P \times \ldots \times P$ (n copies). For a sufficiently large n, i.e. a sufficiently high-dimensional distribution P^n, the typical set is small relative to the total number of possible elements in P^n, yet the probability of the set under P^n is close to one.

The idea underlying the typical set hypothesis is the following: If nearly all of the total probability mass of a distribution is concentrated in a small set, then it is unlikely that a sample generated from the distribution will fall outside of this set. For instance, samples from an out-distribution such as SVHN could have high likelihood in the CIFAR-10 distribution but fall outside its typical set, which would explain why SVHN samples are not seen in the finite CIFAR-10 dataset.

Tests based on the typical set have power to detect out-distributions concentrated in high likelihood regions of the data distribution which have small overall probability, but the assumption that an out-distribution like SVHN lies within the support of a data distribution like CIFAR-10 yields questionable consequences, illustrated in the next two sections.

[2]Nalisnick et al. (2019b) discuss the *model's* typical set rather than the data distribution's but do not mention model estimation error. Subsequent works have interpreted their message in the context of the data distribution's typical set (Morningstar et al., 2021).

3.2. No Method Can Guarantee Perfect Detection When Supports Overlap

In order to explain why DGMs trained on CIFAR-10 or Fashion-MNIST place high likelihood on SVHN or MNIST samples respectively, the typical set hypothesis must assume that the out-distributions overlap in support with the in-distributions. However, the probability of classification error is non-zero when the support of a given out-distribution Q overlaps with that of the in-distribution P. Therefore, even with exact knowledge of the in-distribution, no method can achieve perfect detection against out-distributions which overlap in support with the in-distribution.

Proposition 2. *Let P and Q have overlapping support: $Pr_Q(\mathbf{x} \in supp(p(\mathbf{x}))) > 0$. Then, any test has non-zero probability of error.*

Proof. Assume there exists a rejection rule $\phi_p(\mathbf{x}) \notin \Phi$ that perfectly separates samples from P and Q i.e.

$$\Pr_Q\left(\phi_p(\mathbf{x}) \in \Phi\right) = 0, \text{ and } \Pr_P\left(\phi_p(\mathbf{x}) \in \Phi\right) = 1.$$

The above condition requires Φ to encompass all values in $\{\phi(\mathbf{x})|\mathbf{x} \in \text{supp}(p(\mathbf{x}))\}$ and none in $\{\phi(\mathbf{x})|\mathbf{x} \in \text{supp}(q(\mathbf{x}))\}$. However, since $\mathbf{Pr}_Q(\mathbf{x} \in \text{supp}(p(\mathbf{x}))) > 0$, $\mathbf{Pr}_Q(\phi_p(\mathbf{x}) \in \text{supp}(p(\phi_p(\mathbf{x})))) > 0$. By contradiction, there exists no subset Φ that perfectly separates P and Q. □

Proposition 2 states that if the supports of two distributions (e.g. SVHN and CIFAR-10) overlap, then no solution can guarantee perfect discrimination between single samples from these two distributions. This relates to the bound on performance given by the Bayes optimal classifier: even the optimal classifier has non-zero error when the covariate distributions from two classes overlap.

3.3. A Wrong Model Can Perform Better Than a Perfect One When Supports Overlap

An additional consequence of including support overlap cases in OOD detection is that for a given out-distribution Q, a perfect model can perform worse than a misestimated one. Define ϕ_p using the data distribution P (e.g. $\phi_p = \log p$) such that the rejection rule is of the form $\phi_p(\mathbf{x}) < k$. (We can recast existing test statistic and rejection rule pairings to follow this form, even if the original pairing does not use rejection rule $\phi_p(\mathbf{x}) < k$. See Appendix B for details.) We can write the AUC of an OOD detection procedure using ϕ_p as $\Pr(\phi_p(\mathbf{x}) > \phi_p(\mathbf{y}))$ for $\mathbf{x} \sim P$, $\mathbf{y} \sim Q$. Perfect discrimination is achieved when $\Pr(\phi_p(\mathbf{x}) > \phi_p(\mathbf{y})) = 1$.

We now show that it is possible for OOD detection based on a misestimated model to perform better than detection using the true in-distribution when the supports of P and Q overlap. Let $\phi_p = p$, and let Q have support over the entire sample space $\mathcal{X}$. We can construct a P_θ proportional to the likelihood ratio of P and Q:

$$p_\theta(\mathbf{x}) = \frac{1}{C}\frac{p(\mathbf{x})}{q(\mathbf{x})}, \; C = \int_{\mathcal{X}} \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x},$$

assuming integrability. Then, ϕ_{p_θ} is proportional to the likelihood ratio, and by the Neyman-Pearson Lemma (Neyman & Pearson, 1933), a likelihood ratio test statistic is uniformly most powerful for a simple hypothesis (i.e. yields the highest power against any single alternative hypothesis Q for a specified test size). Since the ratio is most powerful at every false positive rate specified, OOD detection via ϕ_{p_θ} achieves the maximal AUC possible for a given pair P, Q. Since a uniformly most powerful decision rule is unique up to sets of measure zero, given $P_\theta \neq P$, OOD detection using P_θ is strictly better than detection using P. In summary,

$$\Pr\left(\phi_{p_\theta}(\mathbf{x}) > \phi_{p_\theta}(\mathbf{y})\right) > \Pr\left(\phi_p(\mathbf{x}) > \phi_p(\mathbf{y})\right).$$

The same idea applies even when Q does not place strictly positive density or probability across the sample space $\mathcal{X}$ or when the test statistic is a function other than $\boldsymbol{\phi}_p = \log p$: A model P_θ which makes values of $\phi_{p_\theta}(\mathbf{x})$, $\mathbf{x} \sim P$ higher relative to $\phi_{p_\theta}(\mathbf{y})$, $\mathbf{y} \sim Q$ will improve OOD detection. We illustrate this phenomenon with an empirical example in Section 4.2, comparing the OOD performance for a specific P, Q pair when the test statistic utilizes the true distribution P versus a (poor) estimate of it.

Note that this result does not apply when distributions P and Q have disjoint support and the test statistic used is $\boldsymbol{\phi}_p = p$. Concretely, $\boldsymbol{\phi}_p(\mathbf{y}) = 0$ for all $\mathbf{y} \sim Q$ and $\boldsymbol{\phi}_p(\mathbf{x}) > 0$ for all $\mathbf{x} \sim P$, which implies that $\Pr(\boldsymbol{\phi}_p(\mathbf{x}) > \boldsymbol{\phi}_p(\mathbf{y})) = 1$; in this setting, likelihood-based OOD detection using a perfect model yields optimal performance.

3.4. OOD Detection based on the Typical Set is Arbitrary

The previous two sections (Section 3.2 and Section 3.3) highlight two consequences that result from the assumption in the typical set hypothesis that out-distributions of interest (e.g. SVHN, MNIST) overlap in support with the in-distribution (e.g. CIFAR-10, FashionMNIST). Beyond the issues resulting from the support overlap assumption, the typical set hypothesis relies on the idea that the typical set is the preferred subset to demarcate what is in- and out-of-distribution. Here, we question this idea, explaining why the properties of the typical set do not suffice to explain its relevance to OOD detection.

As discussed in Section 3.1, the typical set hypothesis uses the fact that the typical set can be small yet high probability to justify why points outside the set should be considered OOD. However, there can exist other similarly small sets that also contain nearly all of the probability mass, meaning it is arbitrary to prefer the typical set based on its small size and high probability properties alone.

As an example, consider a high-dimensional distribution of i.i.d. Bernoullis, each with 75% probability of success. A vector of 100% ones, denoted $\mathbf{z}_{100}$, has the highest probability but is not part of the typical set $\mathcal{A}_\epsilon$ which consists of sequences with close to 75% ones and 25% zeros. The property $\Pr(\mathbf{x} \in \mathcal{A}_\epsilon) \approx 1$ is used to justify why it is okay to consider any points outside this set as OOD, including $\mathbf{z}_{100}$. However, we also can define a new set $\mathcal{A}'_\epsilon$

which substitutes in $\mathbf{z}_{100}$ in place of one of the 72/25 sequences—for instance, the sequence whose first 75% of elements are ones and last 25% are zero, which we will denote $\mathbf{z}_{75}$. This same-sized set also satisfies $\Pr(\mathbf{x} \in \mathcal{A}'_\epsilon) \approx 1$ and in fact has strictly greater probability than the typical set, since a sequence of ones is more likely than the particular 75/25 sequence that was removed. Under this newly constructed set, the sequence $\mathbf{z}_{75}$ would be considered OOD since it is not contained in $\mathcal{A}'_\epsilon$. Since we randomly selected one of the 75/25 sequences to replace, the decision to mark this sequence OOD based on set membership is arbitrary. Yet, it is just as arbitrary to exclude $\mathbf{z}_{100}$ from the set of in-distribution points; after all, $\Pr(\mathbf{z}_{100}) > \Pr(\mathbf{z}_{75})$.

The unique property of the typical set relative to other small-volume, high-probability sets is that its elements are close to equiprobable. However, the need for this property does not follow from the motivation for OOD detection.

In summary, there are several issues with the typical set hypothesis. First, the hypothesis relies on the assumption that out-distributions can overlap in support with the in-distribution; this support overlap assumption implies that perfect discrimination between image datasets distibutions such as CIFAR-10/SVHN and Fashion-MNIST/MNIST is impossible (Section 3.2), and that one could perform better OOD detection with a wrong model of the data distribution than the right one (Section 3.3). Additionally, there is no clear motivation for preferring the typical set over other small volume, high probability sets. These issues suggest the implausibility of the typical set hypothesis as the explanation and solution for existing OOD failures of DGMs.

Consequently, the phenomenon observed in Nalisnick et al. (2019a); Hendrycks et al. (2019) is likely a result of model estimation error, rather than a property of existing image data distributions requiring an alternative task definition. We detail what this perspective implies about existing models and provide guidance for future work in Section 5.

4. Experiments

The following experiments demonstrate the theoretical properties shown in the prequel. To build intuition around the impossibility result of Proposition 1, we give an example of different distributions with the same distribution of densities under P, meaning a test based on a likelihood statistic cannot distinguish them even with access to a perfect model of the in-distribution. We then demonstrate an instance of a partially-trained DGM performing better OOD detection than the actual model of the data, providing an empirical example for the analysis in Section 3.3 that an erroneous model can perform better OOD detection than a perfect one.

4.1. Any Test Statistic has Failure Modes When Possible Out-distributions are Unrestricted

Proposition 1 states that any test statistic gives up power over certain alternative hypotheses. We show that the test statistic $\phi(\mathbf{x}) = \log p(\mathbf{x})$ cannot detect as OOD single samples drawn from out-distributions Q, R which are contained within P but collapse any of its level sets.

Consider an in-distribution P that is bivariate Gaussian with an identity covariance matrix. There are a variety of out-distributions whose samples yield the same distribution of the test statistic $\phi(\mathbf{x}) = \log p(\mathbf{x})$. We consider two in Figure 1: Q, the distribution obtained by sampling (x_1, x_2) from a standard bivariate normal and then flipping the sign of x_2 if it is in the second or fourth quadrant, and R, the distribution obtained by sampling (x_1, x_2) from a standard bivariate normal and mapping it to the point (z, z) where $z^2 = \left(x_1^2 + x_2^2\right)/2$ (i.e. preserving distance from origin). The out-distributions Q and R maintain the same distribution of log-densities under P as the distribution P since they distribute mass similarly across the upper level sets $\{\mathbf{x} : p(\mathbf{x}) > t\} \forall t$. We can use similar logic to construct problematic out-distributions for other test statistics which are a function of $p(\mathbf{x})$, including the test statistic associated with the typicality test in Nalisnick et al. (2019b). In fact, this test statistic, $\phi(\mathbf{x}) = \left|-\log p(\mathbf{x}) - \widehat{H}_p\right|$ where $\widehat{H}_p = -\frac{1}{|\mathcal{D}_{tr}|}\sum_{\mathbf{x} \in \mathcal{D}_{tr}} \log p(\mathbf{x})$, is no better than random chance whenever a log-probability test statistic is no better than chance. To see this, note that $\widehat{H}_p$ is a constant, so the resulting distributions over $\phi(\mathbf{x})$ are simply shifted and scaled relative to the distributions over the log-likelihoods.

4.2. A Bad Model Can Beat a Perfect One When the Out- and In-Distributions Overlap in Support

We now provide an empirical example where misestimation can result in better OOD detection of a particular out-distribution when supports of the in- and out-distribution overlap. Because we require access to the true model density, we designate a pretrained DGM as our in-distribution P—specifically the GLOW model of Kingma & Dhariwal (2018) trained on CIFAR-10. Next, we train a separate GLOW model P_θ on 40, 000 samples from P. See Appendix C for model and training details. We then compare performance of an OOD detection method using p versus p_θ. We choose the CelebA dataset of celebrity faces (Liu et al., 2015) as our out-distribution Q. The in-distribution P represented by the flow overlaps in support with the out-distribution Q, as evidenced by the fact that the pretrained model assigns positive densities to all CelebA images. Our partially-trained model P_θ (only 50 epochs) achieves an average bits-per-dimension (BPD) of 3.67 on the test samples from P, versus 3.45 for the true model (lower is better). However, P_θ improves OOD performance relative to the true model for this choice of Q. The misestimation has increased BPD for both the in-distribution and out-distribution samples relative to the true model; however, since the extent of the increase is higher for the out-distribution samples, the result is better separation, meaning better OOD detection under the misestimated model. See Figure 2.

5. An Alternative Perspective

Rather than conclude that existing image data distributions assign high likelihoods to certain OOD images, we now consider an alternative explanation, that the phenomenon observed in Nalisnick et al. (2019a); Hendrycks et al. (2019) is a result of model estimation error.

First, it is reasonable to assume that the supports of dataset pairs such as CIFAR-10 and SVHN are disjoint: for instance, we would not expect to draw a house number from the true CIFAR-10 distribution even given infinite samples. This assumption is untestable, but if it does

hold, then existing DGMs are mistakenly assigning high probability or density in places where they should be assigning none.

Such misestimation may seem surprising given that a DGM trained on CIFAR-10 never seems to generate SVHN images, as previous works have noted. We can understand this as poor estimation in small-volume regions of the sample space with negligible total probability mass. In fact, even good generators can be very wrong in this regard. From this perspective, training a DGM for OOD detection can require accurate estimation in regions which are unimportant for good generation. We demonstrate this with an example.

5.1. A Good Generator Can Still Exhibit OOD Detection Failures

A model P_θ can output samples from the true data distribution P with probability close to 1 (i.e. good generation) yet still assign higher probability or density to certain out-of-support samples (i.e. bad OOD detection). As an illustration, consider a finite, discrete sample space where distributions P and Q have disjoint support. If the size of the support of P is much greater than that of the support of Q, i.e. $|\text{supp}(P)| \ggg |\text{supp}(Q)|$, then a model P_θ could place higher probability mass on each element of Q than any element in P yet assign negligible probability denoted by ϵ in total to the elements in $\text{supp}(Q)$, i.e.

$$\begin{aligned} &\Pr_{p_\theta}(\text{supp}(P)) = 1 - \epsilon, \quad \Pr_{p_\theta}(\text{supp}(Q)) = \epsilon, \\ &\Pr_{p_\theta}(\mathbf{x}) < \Pr_{p_\theta}(\mathbf{y}),\ \forall \mathbf{x} \in \text{supp}(P),\ \mathbf{y} \in \text{supp}(Q) \end{aligned} \tag{2}$$

As an example, let P be a distribution which assigns the same probability to each element in its support, i.e. $\Pr_p(\mathbf{x}) = 1/|\text{supp}(P)|$. Consider P_θ which moves $\epsilon/|\text{supp}(P)|$ probability away from each element in $\text{supp}(P)$ and splits it equally across all $\mathbf{y} \in \text{supp}(Q)$, i.e. $\Pr_{p_\theta}(\mathbf{y}) = \epsilon / \left|\text{supp}(Q)\right|$. To meet the criteria of Equation (2),

$$\epsilon > \frac{|\text{supp}(Q)|}{|\text{supp}(P)| + |\text{supp}(Q)|}.$$

When $|\text{supp}(P)|$ is much larger than $|\text{supp}(Q)|$, the ϵ needed for P_θ to assign higher probabilities to points in the support of Q is small. Such a model would also have very good in-distribution probabilities, smaller than those of P only by an $\epsilon/|\text{supp}(P)|$ amount. In other words, even small differences in held-out log probabilities can matter for OOD detection of small-volume out-distributions. For instance, using $\epsilon = {|\text{supp}(Q)|}/{|\text{supp}(P)|}$ if $|\text{supp}(P)| = 10^6$ and $|\text{supp}(Q)| = 10^4$, then $\epsilon = 10^{-2}$ and $\Pr_p(\mathbf{x}) = 10^{-6}$ while $\Pr_{p_\theta}(\mathbf{x}) = 10^{-6} - 10^{-8}$. The resulting negative log-likelihoods are −13.8155 and −13.8255, respectively. When $|\text{supp}(Q)|$ is even smaller relative to $|\text{supp}(P)|$, the differences can become even smaller. See Table 1 for examples.

While the above example considers discrete distributions, the same idea holds for continuous probability measures with bounded support.[3] We can construct examples similar to the one above by transferring a small amount of mass, originally spread over a large volume within $\text{supp}(P)$, to a region of small volume outside of $\text{supp}(P)$, e.g. $\text{supp}(Q)$.

Choi et al. (2018); Nalisnick et al. (2019b); Wang et al. (2020) have made similar points to explain how there can exist elements with high density or probability that are almost never generated. While these works have sought to characterize the true in-distribution P, we describe this phenomenon strictly in terms of model misestimation $P_\theta \neq P$.

This scenario can apply to existing OOD failures if the volume of the support of the out-distribution is small in comparison to that of the in-distribution. While it is difficult to determine whether this difference in volumes exists in these image distributions, there is reason to believe it might. For one, there are many more pixel combinations which generate a textured pattern than a smooth one. As an extreme example, consider a distribution over solid-color images versus one of random noise images. The former consists of elements where the image is perfectly predictable from the first pixel, and the size of the support is bounded by the number of possible first-value pixels. The latter has much larger support. Likewise, it is plausible that image distributions containing varying textures (e.g. CIFAR-10, Fashion-MNIST) have larger support than distributions with smooth textures (e.g. SVHN, MNIST). This suggests that volume differences can exist in real image distributions.

The fact that a good generator can still experience OOD detection failures suggests that good generation (and a high held-out likelihood) is not sufficient for good OOD detection. Poor model likelihoods over relatively small-volume regions may be a concern for OOD detection.

5.2. Improving OOD Detection with DGMs

Given this perspective that model estimation error is the problem, we list several ways to improve OOD detection. We first discuss how alternative test statistics, including existing ones, can correct for known errors. We then turn to potential future directions to address model bias.

First, given knowledge of a particular model bias, we can construct alternative test statistics to ameliorate the bias for the application of OOD detection. For instance, consider the issue of DGMs placing high probability or density where they should place zero. Assuming the distribution of $\log p_\theta(\mathbf{x})$ is the same for a test set drawn from the same distribution as the training set, OOD images which are assigned higher likelihoods than training images will be further away from the average training likelihood than the in-distribution test images will be. Consequently, a test statistic which considers distance to the average training likelihood —the statistic proposed by Nalisnick et al. (2019b)—will perform better OOD detection than $\phi = \log p_\theta$. We can improve this further by fitting a non-parametric density estimator over the probabilities assigned to the training images and rejecting when a particular likelihood value has not yet been seen or is rare—this OOD procedure is a simplified version of the density of states estimator proposed in Morningstar et al. (2021), who consider a several statistics jointly, not just the likelihood under the model. While test statistics such as those in Nalisnick et al. (2019b); Morningstar et al. (2021) were introduced to approximate an alternative definition of OOD based on the typical set, a more plausible viewpoint is that these test statistics correct for estimation error off the support of the in-distribution.

[3]Let densities p and q be defined with respect to probability measures P and Q and base Lebesgue measure μ. Consider $\mu(\text{supp}(P)) \gg \mu(\text{supp}(Q))$.

Alternative test statistics can help in certain cases, but they are not guaranteed to improve detection results over log p across out-distributions, as shown empirically (Morningstar et al., 2021). An alternative fix involves improving the models themselves. To avoid the issues observed in Nalisnick et al. (2019a); Hendrycks et al. (2019), it is important that models can sufficiently push down probability or density outside of the support of the in-distribution. To do so may require different modeling preferences than the ones developed with other applications, like generation, of DGMs in mind. As an example, certain inductive biases such as convolutional layers which benefit image modeling in general may make OOD detection between image datasets more difficult; Schirrmeister et al. (2020) found that replacing the convolutional layers in a GLOW model with fully connected layers improved OOD detection of problematic image dataset pairs, even though it resulted in worse likelihoods.

The choice of objective may matter as well. Maximum likelihood estimation (MLE) minimizes $\mathrm{KL}(p \| p_\theta)$ which does not allow p_θ to be zero anywhere p is non-zero (Jerfel et al., 2021). This means maximum likelihood favors overdispersed solutions in the finite-data regime, thus posing a challenge to learning good supports. Kirichenko et al. (2020) show that one can improve a MLE-trained normalizing flow architecture that assigns high likelihoods to OOD images by directly minimizing the density of OOD images. After showing the same for PIXELCNN++ in Appendix D, we show that in-distribution likelihoods can remain high even with the additional constraint of forcing down probabilities. This result suggests that existing model classes contain similarly good solutions which do not suffer from the problem of high likelihoods for particular OOD inputs; however, the combination of the model architecture coupled with existing gradient descent-based maximum likelihood optimization does not seem to find such solutions.

6. Related Work

Likelihood Ratio Test Statistics.

Given the poor results of DGMs seen in Nalisnick et al. (2019a), several works propose likelihood ratio test statistics (Ren et al., 2019; Serrà et al., 2020; Schirrmeister et al., 2020). Ren et al. (2019) propose a likelihood ratio where the alternative distribution is the same DGM model class trained on perturbed samples; Serrà et al. (2020) use a distribution induced by a general image compressor; and Schirrmeister et al. (2020) train a DGM on a general image distribution such as 80 Million Tiny Images. Per the discussion in Section 2, we can conclude that the optimal alternative depends on the set of out-distributions of interest.

Test Statistics Motivated by Typicality.

Nalisnick et al. (2019b) devise a goodness-of-fit test based on how close the empirical entropy of a sample deviates from the empirical entropy of the training set. Morningstar et al. (2021) learn a kernel density estimator or one-class SVM over multiple statistics jointly, similarly accounting for whether a test example's statistics deviate from values seen in the training set. Choi et al. (2018) suggest a typicality test in the latent space of a normalizing flow but see better performance using a likelihood-based test statistic which takes into account variance across an ensemble of DGMs. Wang et al. (2020) learn a function to map training samples to a white noise sequence and classify as OOD any input that is not white

noise after being transformed via this function. The empirical success of some of these test statistics has been described as evidence in favor the need to take typicality into account, but we offer an alternative conclusion in Section 5.2: these test statistics may compensate for a particular model misestimation.

Modifying the DGM.

Kirichenko et al. (2020) and Schirrmeister et al. (2020) both modify the invertible layers of a normalizing flow, improving OOD performance at the expense of worse likelihoods. Maaløe et al. (2019) replace the posterior distributions of lower-level latents in their variational autoencoder with their priors, showing improved OOD performance but worse approximate likelihoods.

Investigating Density for OOD Detection.

Le Lan & Dinh (2020) suggest that density may be limited in its use for anomaly detection because a transformation applied to a continuous random variable with strictly positive density can arbitrarily re-rank the density. However, points outside of the support will still be outside of the support under transformations, meaning that such out-distribution samples cannot be re-ranked to be higher than points in the in-distribution.

Test Statistics Based on Discriminative Models.

Methods based on discriminative models use some property of a learned classifier for discriminating in-distribution classes, such as the maximum softmax probability (Hendrycks & Gimpel, 2017). Extensions incorporate OOD loss terms, such as encouraging the softmax probabilities of OOD samples to be uniform or utilizing temperature scaling to increase the sharpness of the in-sample probabilities (Hendrycks et al., 2019; Liang et al., 2018). These methods can be seen as learning a direct mapping from inputs to some OOD score based on minimizing risk with respect to a given distribution of in- and out-samples. However, these test statistics suffer from the same issue that motivates OOD detection in the first place: the learned conditional $\hat{p}(y \mid \mathbf{x})$ must perform extrapolation when the input is unlike what was seen during training, and even the true conditional $p(y \mid \mathbf{x})$ is not defined for inputs where $p(\mathbf{x}) = 0$.

Directly Learning a Decision Boundary.

One-class Support Vector Machines (Schölkopf et al., 1999), Support Vector Data Description (Tax & Duin, 2004), and their deep variants (Ruff et al., 2018) learn to separate a subset of the input space with its complement. These methods estimate the distribution's support rather than the density or probability over that set. Note that the support boundary is all that matters for detection if the family of relevant out-distributions only include those disjoint in support with the in-distribution.

7. Discussion

The failures of existing DGMs to detect certain out-distributions based on log-likelihood has prompted some to wonder whether OOD detection based on probability models requires additional considerations in high dimensions. The results of our analysis suggest that it is

the model that is at fault, not the method for OOD detection. We additionally highlight the importance of formalizing the out-distributions of interest for OOD detection in general, as well as the arbitrary choice of the typical set for OOD detection.

Understanding the OOD detection failures of DGMs as estimation error introduces avenues for future work. We suggest that existing models are incorrectly assigning higher probability or density to certain natural images even when such images should have zero probability or density, and we hypothesize that the issue arises due to not just a single modeling choice, but the combination of the model architecture and maximum likelihood objective.

The extent of misestimation in existing DGMs could be a relatively small amount of total probability mass, if the total volume of the out-distribution support (e.g. those of SVHN, MNIST) is relatively small in comparison to that of the in-distribution support (e.g. those of CIFAR-10, FashionMNIST). Under this scenario, a model could be near-perfect yet still assign higher probability or density to samples from an out-distribution with disjoint support. This possibility illustrates the additional considerations required for OOD detection beyond for instance what is necessary for good generation or held-out likelihood. That said, the bias exhibited by existing models may affect more than a negligible probability set (including if the problem exists for multiple out-distributions), meaning that future work directed towards correcting this bias across existing model classes could benefit not only OOD detection, but other applications for generative models as well. Further work comparing generative modeling and support detection may also provide insight. Finally, the interplay between OOD detection and epistemic uncertainty is worth further study, based on their shared relevance to predictive modeling.

Acknowledgements.

This work was supported by NIH/NHLBI Award R01HL148248, NSF Award 1922658 NRT-HDR: FUTURE Foundations, Translation, and Responsibility for Data Science, NSF Award 1514422 TWC, and a DeepMind Fellowship. We thank the reviewers for their very helpful feedback and Kingma & Dhariwal (2018) and Ren et al. (2019) for open-sourcing their code.

A.: Proposition 1

Proposition Let P be the distribution under the null hypothesis H_0. Let μ be the measure associated with the distribution of test statistic $\phi_p(\mathbf{x})$ under the null. Then, assuming conditional $\mathbf{x}|\phi_p(\mathbf{x})$ is not degenerate on μ-non-measure zero set, there exists a set of alternative distributions $Q \in \mathcal{Q}$ where $Q \neq_d P$ and the test has power equal to the false positive rate. In other words, the test does no better than random guessing.

Proof. We first construct a distribution $q(\mathbf{x}) \neq_d p(\mathbf{x})$ but where $q(\phi_p(\mathbf{x})) = p(\phi_p(\mathbf{x}))$.

The roadmap for this part of the proof is as follows: for some function f, we write

$$\mathbb{E}_{p(\mathbf{x})}(f_p) - \mathbb{E}_{q(\mathbf{x})}(f_p) = \mathbb{E}_{p(\phi_p(\mathbf{x}))}\left[\mathbb{E}_{p(\mathbf{x} \mid \phi_p(\mathbf{x}))}(f_p) - \mathbb{E}_{q(\mathbf{x} \mid \phi_p(\mathbf{x}))}(f_p)\right] \quad (3)$$

We then identify $q(\mathbf{x}|\boldsymbol{\phi}_p(\mathbf{x}))$ and f_p such that the inner difference of expectations is non-zero, which implies inequality in distribution via $\mathbb{E}_{q(\mathbf{x})}(f_p) \neq \mathbb{E}_{p(\mathbf{x})}(f_p)$. We do not change the distribution in the outer expectation $p(\boldsymbol{\phi}_p(\mathbf{x}))$. We finally define $q(\mathbf{x}) = p(\boldsymbol{\phi}_p(\mathbf{x}))q(\mathbf{x} \mid \boldsymbol{\phi}_p(\mathbf{x}))$.

We now show how to construct f_p, q. Let $(\Omega_{\phi_p(\mathbf{x})}, \mathcal{F}_{\phi_p(\mathbf{x})})$ be the probability space associated with $\boldsymbol{\phi}_p(\mathbf{x})$, with probability measure $\mu = \mathbb{P}_{p(\phi_p(\mathbf{x}))}$. By assumption, $p(\mathbf{x} \mid \boldsymbol{\phi}_p(\mathbf{x}))$ is non-degenerate on some μ non-measure zero set. This means there exists a set $\boldsymbol{\Phi} \in \mathcal{F}_{\phi_p(\mathbf{x})}$ with $\mu(\Phi) > 0$ such that $\forall \boldsymbol{\phi}_p(\mathbf{x}) \in \Phi$, $\exists A_{\phi_p(\mathbf{x})} \subset \text{supp}\left(p\left(\mathbf{x} \mid \phi_p(\mathbf{x})\right)\right)$ such that $0 < \mathbb{P}_{p(\mathbf{x} \mid \phi_p(\mathbf{x}))}\left(A_{\phi_p(\mathbf{x})}\right) < 1$.

Let g be any function for which $\mathbb{E}_{p(\mathbf{x} \mid \phi_p(\mathbf{x}))}(g) < \infty \forall \phi_p(\mathbf{x}) \notin \boldsymbol{\Phi}$. Then define

$$f_p(\mathbf{x}) \triangleq \mathbb{1}\left[\phi_p(\mathbf{x}) \in \boldsymbol{\Phi}\right]\mathbb{1}\left[\mathbf{x} \in A_{\phi_p(\mathbf{x})}\right] + \mathbb{1}\left[\phi_p(\mathbf{x}) \notin \boldsymbol{\Phi}\right]g(\mathbf{x}) \tag{4}$$

Define the conditional $q(\mathbf{x} \mid \boldsymbol{\phi}_p(\mathbf{x}))$ with normalization constant $C_{\phi_p(\mathbf{x})}$ and $0 < \lambda < 1$:

$$\begin{aligned} q\left(\mathbf{x} \mid \phi_p(\mathbf{x})\right) \triangleq \mathbb{1}\left[\phi_p(\mathbf{x}) \in \boldsymbol{\Phi}\right]\Bigg[\frac{1}{C_{\phi_n(\mathbf{x})}}\Big(\lambda p\left(\mathbf{x} \mid \phi_p(\mathbf{x})\right)\mathbb{1}\left[\mathbf{x} \in A_{\phi_p(\mathbf{x})}\right] \\ + p\left(\mathbf{x} \mid \phi_p(\mathbf{x})\right)\mathbb{1}\left[\mathbf{x} \notin A_{\phi_p(\mathbf{x})}\right]\Big)\Bigg] + \mathbb{1}\left[\phi_p(\mathbf{x}) \notin \boldsymbol{\Phi}\right]p\left(\mathbf{x} \mid \phi_p(\mathbf{x})\right) \end{aligned} \tag{5}$$

For $\boldsymbol{\phi}_p(\mathbf{x}) \notin \Phi$, $q(\mathbf{x}|\boldsymbol{\phi}_p(\mathbf{x})) = p(\mathbf{x} \mid \boldsymbol{\phi}_p(\mathbf{x}))$. Therefore, $\mathbb{1}\left[\phi_p(\mathbf{x}) \notin \boldsymbol{\Phi}\right]\left[\mathbb{E}_{p(\mathbf{x} \mid \phi_p(\mathbf{x}))}(f_p) - \mathbb{E}_{q(\mathbf{x} \mid \phi_p(\mathbf{x}))}(f)\right] = 0$. For $\boldsymbol{\phi}_p(\mathbf{x}) \in \Phi$, $q(\mathbf{x} \mid \boldsymbol{\phi}_p(\mathbf{x}))$ moves density away from points in $A_{\phi_p(\mathbf{x})}$ relative to $p(\mathbf{x}|\boldsymbol{\phi}_p(\mathbf{x}))$, given that $0 < \lambda < 1$.

For simplicity, we construct q such that $\text{supp}(q(\mathbf{x}|\boldsymbol{\phi}_p(\mathbf{x}))) = \text{supp}(p(\mathbf{x}|\boldsymbol{\phi}_p(\mathbf{x})))$. This is to avoid any issues with an invalid joint distribution $q(\mathbf{x}, \boldsymbol{\phi}_p(\mathbf{x})) \neq q(\mathbf{x})$ if $q(\mathbf{x} \mid \boldsymbol{\phi}_p(\mathbf{x})) = 0$ (the left-hand side would be 0 while the right-hand side would be greater than 0 $\forall \mathbf{x} \in \text{supp}(p(\mathbf{x}))$.

We now show that this construction leads to $\mathbb{E}_{p(\mathbf{x})}(f_p) - \mathbb{E}_{q(\mathbf{x})}(f_p) > 0$, implying inequality in distribution.

$$\mathbb{E}_{p(\mathbf{x})}(f_p) - \mathbb{E}_{q(\mathbf{x})}(f_p) \tag{6}$$

$$= \mathbb{E}_{p(\phi_p(\mathbf{x}))}\left[\mathbb{E}_{p(\mathbf{x} \mid \phi_p(\mathbf{x}))}(f_p) - \mathbb{E}_{q(\mathbf{x} \mid \phi_p(\mathbf{x}))}(f_p)\right] \tag{7}$$

$$= \mathbb{E}_{p(\phi_p(\mathbf{x}))}\mathbb{1}\left[\phi_p(\mathbf{x}) \in \boldsymbol{\Phi}\right]\left[\mathbb{E}_{p(\mathbf{x} \mid \phi_p(\mathbf{x}))}(f_p) - \mathbb{E}_{q(\mathbf{x} \mid \phi_p(\mathbf{x}))}(f_p)\right] \tag{8}$$

$$= \mathbb{E}_{p(\phi_p(\mathbf{x}))} \mathbb{1}\left[\phi_p(\mathbf{x}) \in \boldsymbol{\Phi}\right] \left[\mathbb{E}_{p(\mathbf{x} \mid \phi_p(\mathbf{x}))}\left(\mathbb{1}\left[\mathbf{x} \in A_{\phi_p(\mathbf{x})}\right]\right) - \mathbb{E}_{q(\mathbf{x} \mid \phi_p(\mathbf{x}))}\left(\mathbb{1}\left[\mathbf{x} \in A_{\phi_p(\mathbf{x})}\right]\right)\right] \tag{9}$$

$$= \mathbb{E}_{p(\phi_p(\mathbf{x}))} \mathbb{1}\left[\phi_p(\mathbf{x}) \in \boldsymbol{\Phi}\right] \left[\int_{A_{\phi_p(\mathbf{x})}} p(\mathbf{x} \mid \phi_p(\mathbf{x})) d\mathbf{x} - \int_{A_{\phi_p(\mathbf{x})}} q(\mathbf{x} \mid \phi_p(\mathbf{x})) d\mathbf{x}\right] \tag{10}$$

$$= \mathbb{E}_{p(\phi_p(\mathbf{x}))} \mathbb{1}\left[\phi_p(\mathbf{x}) \in \boldsymbol{\Phi}\right] \left[\int_{A_{\phi_p(\mathbf{x})}} p(\mathbf{x} \mid \phi_p(\mathbf{x})) d\mathbf{x} - \int_{A_{\phi_p(\mathbf{x})}} \frac{1}{C_{\phi_p(\mathbf{x})}} \lambda p(\mathbf{x} \mid \phi_p(\mathbf{x})) d\mathbf{x}\right] \tag{11}$$

$$> 0 \tag{12}$$

Line 11 follows from the substitution of $q(\mathbf{x}|\phi_p(\mathbf{x}))$ defined in Equation (5). Line 12 follows from the fact that $\frac{\lambda}{C_{\phi_p(\mathbf{x})}} < 1$, shown below:

$$C_{\phi_p(\mathbf{x})} = \int_{\mathcal{X}} \lambda p(\mathbf{x} \mid \phi_p(\mathbf{x})) \mathbb{1}\left[\mathbf{x} \in A_{\phi_p(\mathbf{x})}\right] + p(\mathbf{x} \mid \phi_p(\mathbf{x})) \mathbb{1}\left[\mathbf{x} \notin A_{\phi_p(\mathbf{x})}\right] d\mathbf{x} \tag{13}$$

$$= \lambda \mathbb{P}_{p(\mathbf{x} \mid \phi_p(\mathbf{x}))}\left(A_{\phi_p(\mathbf{x})}\right) + \mathbb{P}_{p(\mathbf{x} \mid \phi_p(\mathbf{x}))}\left(A^C_{\phi_p(\mathbf{x})}\right) \tag{14}$$

$$\frac{\lambda}{C_{\phi_p(\mathbf{x})}} = \frac{\lambda}{\lambda \mathbb{P}_{p(\mathbf{x} \mid \phi_p(\mathbf{x}))}\left(A_{\phi_p(\mathbf{x})}\right) + \mathbb{P}_{p(\mathbf{x} \mid \phi_p(\mathbf{x}))}\left(A^C_{\phi_p(\mathbf{x})}\right)} \tag{15}$$

$$= \frac{1}{\mathbb{P}_{p(\mathbf{x} \mid \phi_p(\mathbf{x}))}\left(A_{\phi_p(\mathbf{x})}\right) + \frac{1}{\lambda}\left[\mathbb{P}_{p(\mathbf{x} \mid \phi_p(\mathbf{x}))}\left(A^c_{\phi_p(\mathbf{x})}\right)\right]} \tag{16}$$

$$< 1 \tag{17}$$

Line 17 holds since the denominator in the previous line is greater than 1: Since $0 < \lambda < 1$, $\frac{1}{\lambda} > 1$. Then,

$$\mathbb{P}_{p(\mathbf{x} \mid \phi_p(\mathbf{x}))}\left(A_{\phi_p(\mathbf{x})}\right) + \frac{1}{\lambda}\left[\mathbb{P}_{p(\mathbf{x} \mid \phi_p(\mathbf{x}))}\left(A^c_{\phi_p(\mathbf{x})}\right)\right] > \mathbb{P}_{p(\mathbf{x} \mid \phi_p(\mathbf{x}))}\left(A_{\phi_p(\mathbf{x})}\right) + \mathbb{P}_{p(\mathbf{x} \mid \phi_p(\mathbf{x}))}\left(A^c_{\phi_p(\mathbf{x})}\right) = 1.$$

Having constructed the distribution Q, we now proceed with the second part of the proposition: for any specified false positive rate, any test based on ϕ_p has power equal to the false positive rate when the OOD samples come from Q.

Recall that $q(\phi_p(\mathbf{x})) = p(\phi_p(\mathbf{x}))$. Then, for any rejection rule $\phi_p(\mathbf{x}) \notin \Phi_{\text{Accept}}$, the probability of rejection is the same regardless of whether the sample $\mathbf{x}$ is drawn from P or q.

$$\forall \boldsymbol{\Phi}_{\text{Accept}}, \quad \mathbb{P}_{\mathbf{x}\sim q}\left(\phi_p(\mathbf{x}) \notin \boldsymbol{\Phi}_{\text{Accept}}\right) = \mathbb{P}_{\mathbf{x}\sim p}\left(\phi_p(\mathbf{x}) \notin \boldsymbol{\Phi}_{\text{Accept}}\right). \tag{18}$$

Therefore, the power of the test (i.e. rejecting under the $H_A : \mathbf{x} \sim q$) is equal to the false positive rate (i.e. rejecting under $H_0 : \mathbf{x} \sim p$). When power and false positive rate are equal for all possible values of the false positive rate, then the result is an ROC curve $y = x$ with AUC 0.5. This is equivalent to random guessing with rejection rate based on the false positive rate chosen for the test. □

B.: Rejection Rules Can Be Written in the Form $\phi_p(\mathbf{x}) < k$

Lemma 1 *Any rejection rule involving intervals, i.e.* $\phi(\mathbf{x}) \notin \Phi$ *can be recast as a rule of the form* $\phi'(\mathbf{x}) < k$.

Proof If we have a one-sided rule, i.e. an interval Φ where one of the endpoints is $-\infty$ or ∞, we simply reverse the sign if necessary, and for two-sided rules, i.e. a bounded interval, we can find the midpoint of the interval m, where $\Phi = [m - k, m + k]$, and recast the rule to $|\phi(\mathbf{x}) - m| < k$.

Rejection rules of this form match the same "rejection" rules used for binary classification more broadly. For added clarity, we define some OOD detection methods based on their rejection rules in this form. For instance, the likelihood-based test (Bishop, 1994) rejects when the negative log likelihood is above a certain threshold k, whereas the typicality test (Nalisnick et al., 2019b) rejects when the distance to the training set entropy is above k.

C.: Details for Experiment 5.2

In this experiment, we compare a partially trained GLOW model p_θ with a pretrained GLOW model (Kingma & Dhariwal, 2018) which we use as our data distribution P. First, we generate samples from P by sampling from the GLOW model pretrained on CIFAR-10[4]. We use temperature 1 for sampling to ensure our samples come from the distribution specified by the model. We generate 40,000 samples for training and 10,000 samples for evaluation, matching the train and test set sizes of the CIFAR-10 dataset.

The glow (GLOW) model p_θ is made of 3 blocks, each with 8 affine coupling layers with 400 hidden units per layer. The network is trained with Adamax at learning rate 0.001, which stays constant after 10 epochs of warmup. We use batch size 64 during training. We intentionally limit the training (50 epochs with 10 epochs of warmup) to make the model mis-estimation clear. Our model achieves an average bits per dimension of 3.67 on the test samples, versus 3.45 for the true model (lower is better).

The true model is a larger model than p_θ, consisting of 3-blocks each with 32 affine coupling layers with 400 units each.

[4]The pretrained model is available here: https://openaipublic.azureedge.net/glow-demo/logs/abl-1x1-aff.tar

We evaluate OOD performance on the test set of the model samples and the test set of CelebA.

D.: Existing Model Architectures Can Yield Good OOD Detectors

We directly optimize a PIXELCNN++ to distinguish between FashionMNIST and MNIST by replacing the maximum likelihood training objective with one which simultaneously maximizes likelihood on FashionMNIST images while minimizing likelihood of MNIST images. Our objective is similar to that of Kirichenko et al. (2020), who show that flows can distinguish problematic OOD dataset pairs when optimized directly to do so.

$$\frac{1}{N_{in}} \sum_{x \in \mathcal{D}_{in}} \log p_\theta(x) - \frac{1}{N_{ood}} \sum_{x' \in \mathcal{D}_{ood}} \min(\log p_\theta(x'), c) \tag{19}$$

Replicating the architecture of Ren et al. (2019), we train our model across five random seeds using the MLE objective and five seeds with the above objective. We use the same training hyperparameters as Ren et al. (2019): 50,000 steps at a learning rate of 0.0001 with exponential decay rate of 0.999995 per step, batch size of 32, and Adam optimizer with momentum parameters 0.95 and 0.9995. Our results, shown in Table 2, demonstrate that the PIXELCNN++ architecture has the capacity to push down probabilities on problematic OOD samples while maintaining high in-distribution likelihoods.

Table 2.

Existing models can be optimized to distinguish datasets. PIXELCNN++ trained via the negative training (NT) objective in Equation (19) can achieve near-perfect OOD detection while maintaining comparable held-out log likelihoods (LL) with models trained via maximum likelihood estimation (MLE). We report mean and standard deviation of the results over 5 random seeds.

	Fashion LL	OOD AUC
MLE	−1550± 6	0.097 ± 0.004
NT	−1562± 7	1.000 ± 0.000

References

Amodei D, Olah C, Steinhardt J, Christiano PF, Schulman J, and Mané D Concrete problems in ai safety. ArXiv, 2016.

Bishop CM Novelty detection and neural network validation. IEE Proceedings-Vision, Image and Signal processing, 141(4):217–222, 1994.

Chen X, Mishra N, Rohaninejad M, and Abbeel P Pixelsnail: An improved autoregressive generative model. In ICML, volume abs/1712.09763, 2018.

Choi H, Jang E, and Alemi AA Waic, but why? generative ensembles for robust anomaly detection. ArXiv, 2018.

Cover T and Thomas J Elements of Information Theory. 1991.

Dinh L, Krueger D, and Bengio Y Nice: Non-linear independent components estimation. ArXiv, 2015.

Dinh L, Sohl-Dickstein J, and Bengio S Density estimation using real nvp. ArXiv, abs/1605.08803, 2017.

Hendrycks D and Gimpel K A baseline for detecting misclassified and out-of-distribution examples in neural networks. In ICLR, 2017.

Hendrycks D, Mazeika M, and Dietterich TG Deep anomaly detection with outlier exposure. In ICLR, 2019.

Jerfel G, Wang SL, Fannjiang C, Heller KA, Ma Y, and Jordan M Variational refinement for importance sampling using the forward kullback-leibler divergence. In Third Symposium on Advances in Approximate Bayesian Inference, 2021.

Kingma DP and Dhariwal P Glow: Generative flow with invertible 1×1 convolutions. In NeurIPS, 2018.

Kirichenko P, Izmailov P, and Wilson A Why normalizing flows fail to detect out-of-distribution data. In NeurIPS, 2020.

Le Lan C and Dinh L Perfect density models cannot guarantee anomaly detection. In NeurIPS I Can't Believe It's Not Better Workshop, 2020.

Liang S, Li Y, and Srikant R Enhancing the reliability of out-of-distribution image detection in neural networks. In ICLR, 2018.

Liu Z, Luo P, Wang X, and Tang X Deep learning face attributes in the wild. In Proceedings of International Conference on Computer Vision (ICCV), December 2015.

Maaløe L, Fraccaro M, Liévin V, and Winther O Biva: A very deep hierarchy of latent variables for generative modeling. In NeurIPS, 2019.

Morningstar W, Ham C, Gallagher AG, Lakshminarayanan B, Alemi AA, and Dillon JV Density of states estimation for out-of-distribution detection. 2021.

Nalisnick E, Matsukawa A, Teh Y, Görür D, and Lakshminarayanan B Do deep generative models know what they don't know? In ICLR, 2019a.

Nalisnick E, Matsukawa A, Teh YW, and Lakshminarayanan B Detecting out-of-distribution inputs to deep generative models using typicality. In NeurIPS Workshop on Bayesian Deep Learning, 2019b.

Neyman J and Pearson E On the problem of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society A, 231:289–337, 1933.

Oord A, Kalchbrenner N, Espeholt L, Kavukcuoglu K, Vinyals O, and Graves A Conditional image generation with pixelcnn decoders. In NeurIPS, 2016a.

Oord A, Kalchbrenner N, and Kavukcuoglu K Pixel recurrent neural networks. In ICML, 2016b.

Ren J, Liu PJ, Fertig E, Snoek J, Poplin R, DePristo MA, Dillon JV, and Lakshminarayanan B Likelihood ratios for out-of-distribution detection. In NeurIPS, 2019.

Ruff L, Görnitz N, Deecke L, Siddiqui S, Vander-meulen RA, Binder A, Müller E, and Kloft M Deep one-class classification. In ICML, 2018.

Salimans T, Karpathy A, Chen X, and Kingma DP Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. 2017.

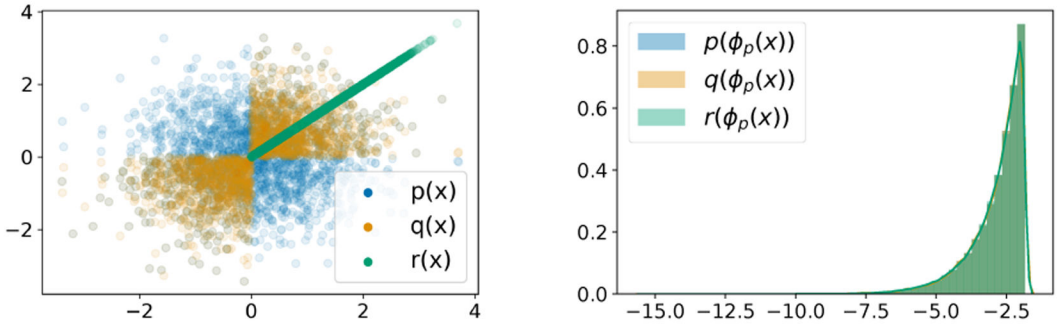
Schirrmeister RT, Zhou Y, Ball T, and Zhang D Understanding anomaly detection with deep invertible networks through hierarchies of distributions and features. In NeurIPS, 2020.

Schölkopf B, Williamson RC, Smola A, Shawe-Taylor J, and Platt JC Support vector method for novelty detection. In NeurIPS, 1999.

Serrà J, Álvarez D, Gómez V, Slizovskaia O, Núñez JF, and Luque J Input complexity and out-of-distribution detection with likelihood-based generative models. In ICLR, 2020.

Tax D and Duin R Support vector data description. Machine Learning, 54:45–66, 2004.

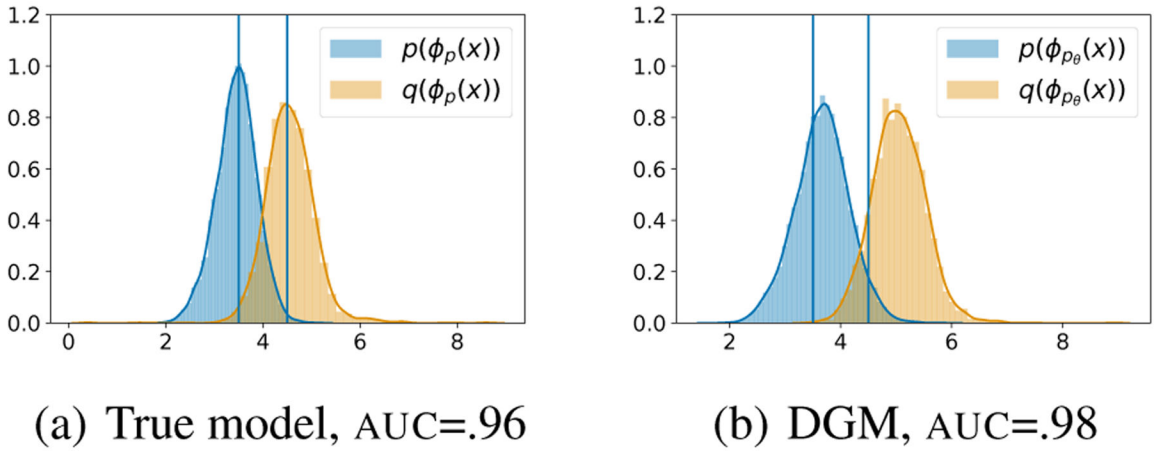
Wang Z, Dai B, Wipf D, and Zhu J Further analysis of outlier detection with deep generative models. In NeurIPS, 2020.



(a) Samples from three different distributions (b) Distributions of the log-density under P

Figure 1.

For a given in-distribution P and choice of test statistic, we can construct out-distributions where OOD detection will perform no better than random chance. In the above example, given test statistic $\phi_p(\mathbf{x}) = \log p(\mathbf{x})$, we can construct out-distributions Q, R (left) such that the distribution of log-density under P, i.e. $\phi_p(\mathbf{x})$, is the same regardless of whether $\mathbf{x}$ comes from P, Q or R (right).



(a) True model, AUC=.96 (b) DGM, AUC=.98

Figure 2.

A perfect model can perform worse than a misestimated one when supports of the in- and out-distributions overlap. The DGM yields better OOD detection performance than the true model because the misestimation has decreased the amount of overlap in the distributions of the test statistic bits per dimension.

Table 1.

A model P_θ can place higher probabilities on an out-distribution Q even with in-distribution negative log-likelihoods (NLL) close to the true model P (Oracle). All calculations are based on a uniform P with $|\text{supp}(P)| = 10^6$ and a uniform transfer of $\epsilon = {}^{|\text{supp}(Q) \setminus \text{supp}(P)|}/_{|\text{supp}(P)|}$ from P to be divided evenly across $|\text{supp}(Q)|$ of different size (10^4, 10^3, 10^2).

	$\lvert\text{supp}(Q)\rvert$		
Oracle	10^4	10^3	10^2
−13.8155	−13.8255	−13.8165	−13.8156