UNDERSTANDING OUT-OF-DISTRIBUTION DETECTION WITH DEEP GENERATIVE MODELS

Lily H. Zhang
New York University
lily.h.zhang@nyu.edu

Mark Goldstein
New York University
goldstein@nyu.edu

Rajesh Ranganath
New York University
rajeshr@cims.nyu.edu

ABSTRACT

Deep generative models (DGMs) seem a natural fit for out-of-distribution (OOD) detection. However, they have been shown assign higher probabilities to OOD images than in-distribution ones. Recent works try to explain this phenomenon as a problem with considering only low probability or density points as OOD, contending that even samples in high probability or density regions should be OOD in certain situations. In this work, we present the consequences of broadening the definition of OOD: the task becomes impossible, and even a perfect model can perform worse than a misestimated one. We consider instead that the observed phenomenon is evidence that current DGMs are not yet good enough for OOD detection of certain out-distributions. We also suggest that improving DGMs for OOD detection may require considerations other than those used to enable high likelihoods or good sample quality.

1 INTRODUCTION

OOD detection is an important step towards safe and reliable machine learning (Amodei et al., 2016). Since neural network classifiers have been shown to behave unpredictably on samples that differ significantly from the training data (Nguyen et al., 2015), it is critical to identify examples which should not be fed into a model in the first place.

One method for OOD detection is to train a probability model on observations from the data distribution and to classify examples with low probability or density as OOD (Bishop, 1994). However, DGMs have been shown to assign higher probability or density to known OOD inputs than to training ones (e.g. OOD MNIST vs. in-distribution Fashion-MNIST, OOD SVHN vs. in-distribution CIFAR-10) (Nalisnick et al., 2019b; Hendrycks et al., 2019). Recent works have subsequently argued that it is insufficient to consider only low density or probability points as OOD (Nalisnick et al., 2019a; Choi et al., 2018; Wang et al., 2020). In this work we highlight the consequences implied by this argument, including the impossibility of OOD detection, and consider the alternative perspective that observed OOD failures are due to model estimation error.

2 BACKGROUND AND RELATED WORK

A deep generative model (DGM), which we denote p_θ, estimates a data distribution p. This means that any phenomenon observed in p_θ is either a property of p (if the model is an accurate representation) or is due to estimation error. Several works attempt to explain the observations in (Nalisnick et al., 2019b; Hendrycks et al., 2019) via the former, positing that some out-distributions can overlap in support with the data distribution, including in high probability or density regions (Choi et al., 2018; Nalisnick et al., 2019a; Wang et al., 2020). These works assert that OOD detection should also detect samples from such out-distributions.

We can formalize this task definition as a single-sample hypothesis test (Nalisnick et al., 2019a; Serrà et al., 2020; Wang et al., 2020). OOD detection methods decide whether to reject the hypothesis that a single sample came from the data distribution p in favor of an alternative hypothesis $H_A : \mathbf{x} \sim$

$q, q \neq_d p$. This is done by defining a test statistic $\phi : \mathcal{X} \to \mathbb{R}$ and rejecting any sample whose test statistic value falls outside an accepted set of values. The test statistic is a function of a single sample $\mathbf{x}$ since we wish to make decisions on an individual sample basis.

Most existing methods in OOD detection can be defined by the choice of the test statistic. Some methods utilize a property of a learned classifier as the test statistic (e.g. maximum softmax probability) (Hendrycks & Gimpel, 2017; Liang et al., 2018; Hendrycks et al., 2019), but 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 for very different inputs, and even the true conditional $p(y \mid \mathbf{x})$ is not defined for inputs where $p(\mathbf{x}) = 0$. A separate line of work models the input distribution via a DGM and utilizes this model to derive a test statistic (Nalisnick et al., 2019b; Ren et al., 2019; Serrà et al., 2020; Kirichenko et al., 2020; Nalisnick et al., 2019a; Wang et al., 2020; Schirrmeister et al., 2020). We focus on these methods in our analysis below.

3 The Impossibility of OOD Detection as a Single-Sample Distributional Test

OOD detection as defined in Section 2 is a challenging classification task given that the out-distributions are unknown and can range over infinite possibilities. In this section we establish limits on how well any OOD detection method can perform under this broad definition. Our limits are established using the true distribution p and do not rely on any estimation error.

3.1 All Test Statistics Perform No Better than Chance on Some Out-Distributions

We first show an impossibility result: no test can do well against all alternatives. This result holds absent any considerations of statistical estimation: we allow our test statistic ϕ_p to make use of knowledge of the true in-distribution p.

Proposition 1 (*Informal*)**.** *For any choice in test statistic ϕ_p, there exists a set of alternative distributions $q \in \mathcal{Q}$ where the test does no better than random guessing.*

The formal proposition and proof can be found in Appendix A. The proof sketch is as follows: for a given p, we can construct distributions q where $p(\phi_p(\mathbf{x})) = q(\phi_p(\mathbf{x}))$ but $q \neq_d p$. Intuitively, such constructions can be achieved by moving around mass within a given conditional $p(\mathbf{x} \mid \phi_p(\mathbf{x}))$ for a sufficient set of such conditionals. Because the distribution of the test statistic is the same, any rejection rule will incorrectly reject samples from p and correctly reject samples from q at the same rate, analogous to random guessing.

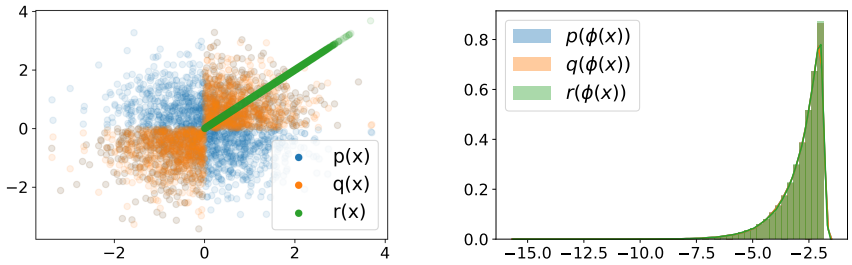
This proposition states that in the context of single-sample distributional testing, all proposed test statistics are trading off power against different out-distributions.

An Example. To build intuition, consider the test statistic $\phi_p = \log p$. For this statistic, 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 density. The analogue for continuous distributions $\mathbb{R}^d$ is collapsing level sets of dimension $\mathbb{R}^{d-1}$. We illustrate the phenomenon in the continuous case in Figure 1.

3.2 Single-Sample OOD Detection Incurs Error When Supports Overlap

Even if we do not consider all possible out-distributions, the probability of classification error is non-zero when the support of a given q overlaps with that of p. Therefore, even with exact knowledge of the in-distribution, no method can achieve perfect OOD detection against such out-distributions.

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



(a) Samples from three different distributions (b) Log densities assigned to samples

Figure 1: Distributions p, q, r have the same distribution of log densities, making them indistinguishable based on the test statistic $\phi(\mathbf{x}) = \log p(\mathbf{x})$. Refer to Appendix B to see how these distributions are constructed.

Proof. We illustrate this via a proof by contradiction. Assume there exists a classification rule $\phi_p(\mathbf{x}) \in \Phi$ that perfectly separates samples from p and q. That is,

$$\text{Pr}_q(\phi_p(\mathbf{x}) \in \Phi) = 0, \text{ and } \text{Pr}_p(\phi_p(\mathbf{x}) \in \Phi) = 1$$

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 $\text{Pr}_q(\mathbf{x} \in \text{supp}(p(\mathbf{x}))) > 0$, $\text{Pr}_q(\phi_p(\mathbf{x}) \in \text{supp}(p(\phi_p(\mathbf{x})))) > 0$. This means that there exists no subset Φ that perfectly separates p and q. □

Proposition 2 demonstrates that if the supports of two distributions (e.g. SVHN and CIFAR-10) overlap, then there exists no solution which can guarantee perfect discrimination between single samples from these two distributions. This result is analogous to the upper 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.

In summary, Propositions 1 and 2 emphasize the fundamental limitations of performing a distributional test given a single sample. If OOD detection is defined as such, then it is impossible.

3.3 A Perfect Model Can Yield Worse Detection Than a Misestimated One

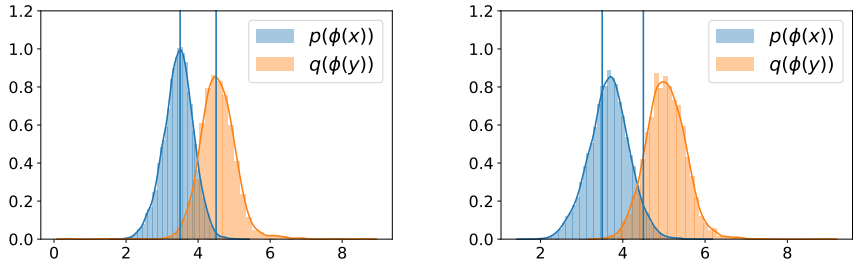
An additional consequence of including support overlap cases in OOD detection is that a perfect model can perform worse than a misestimated one.

Define ϕ_p using on the true model p such that the rejection rule is of the form $\phi_p(\mathbf{x}) > k$ (see Appendix C for details). An example of a test statistic of this form is $\phi_p = -\log p$. Perfect discrimination is achieved when $\text{Pr}(\phi_p(\mathbf{x}) < \phi_p(\mathbf{y})) = 1$ for $\mathbf{x} \sim p, \mathbf{y} \sim q$. We can relate the quality of OOD detection under p_θ, i.e. $\text{Pr}(\phi_{p_\theta}(\mathbf{x}) < \phi_{p_\theta}(\mathbf{y}))$, to the performance under the true null distribution p, i.e. $\text{Pr}(\phi_p(\mathbf{x}) < \phi_p(\mathbf{y}))$ (full derivation is in Appendix D):

$$\begin{aligned} \mathbb{P}\Big(\phi_{p_\theta}(\mathbf{x}) < \phi_{p_\theta}(\mathbf{y})\Big) &= \mathbb{P}\Big(\phi_p(\mathbf{x}) < \phi_p(\mathbf{y}) + D_p - D_q\Big), \\ \text{where} \quad D_p &= \big[\phi_p(\mathbf{x}) - \phi_{p_\theta}(\mathbf{x})\big], D_q = \big[\phi_p(\mathbf{y}) - \phi_{p_\theta}(\mathbf{y})\big]. \end{aligned} \tag{1}$$

A model estimate p_θ can yield better performance than the true distribution p when $D_p - D_q$ is positive, making the right-hand side of inequality under p_θ, i.e. $\phi_p(\mathbf{y}) + D_p - D_q$, larger than its counterpart under p, i.e. $\phi_p(\mathbf{y})$. This occurs when the estimated model creates greater separation in the values of the test statistic between samples from p and samples from q. Note that this result—p_θ improving detection over p—is limited to the scenario where the supports of p and q overlap; otherwise, if all $\mathbf{y} \sim q$ have zero probability or density under p, then $\phi_p = -\log p$ yields the optimal solution since $P(\phi_p(\mathbf{x}) < \phi_p(\mathbf{y})) = 1$.

An Example. We compare a partially trained GLOW model p_θ (Kingma & Dhariwal, 2018) with a pretrained GLOW model which we use as our data distribution p. We train p_θ on samples generated



(a) True model, area under curve (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 $\phi =$ bits per dimension.

from p, intentionally limiting the training (50 epochs with 10 epochs of warmup) to make the model misestimation clear (see Appendix E for further details). Our estimate p_θ achieves an average bits per dimension of 3.67 on the test samples versus 3.45 for the true model p (lower is better). However, when the test statistic is the bits per dim, p_θ performs better OOD detection for an out-distribution of CelebA images (see Figure 2).

4 Understanding Existing OOD Detection Failures in DGMs

Section 3 reveals issues associated with a task definition for OOD detection that considers out-distributions which can lie anywhere in the support of the in-distribution. Rather than expand the definition of what should be considered OOD, we stick to the original formulation of Bishop (1994) and consider instead that the results in Nalisnick et al. (2019b); Hendrycks et al. (2019) are due to model misestimation, namely the failure of existing DGMs to push down probability in certain regions. In fact, while it is untestable whether the supports of dataset pairs such as CIFAR-10 and SVHN overlap, it is reasonable to assume they are disjoint: for instance, we would not expect to draw a house number from the true CIFAR-10 distribution even given infinite samples.

This persepctive suggests that existing DGMs are mistakenly assigning high probability or density in places where they should be assigning zero probability or density, and improving DGMs for OOD detection requires being able to 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 primarily with other applications 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; for instance, 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.

5 Conclusion

The failures of existing DGMs to detect certain out-distributions has prompted some to wonder whether OOD detection based on probability models requires additional considerations in high dimensions. Our analysis suggests that it is the model that is at fault, not the method for OOD detection. To reach this conclusion, we first examine the argument that $\log p$ is a bad test statistic. This argument assumes that certain out-distributions must overlap in support with the in-distribution—otherwise, $\log p$ would be able to correctly distinguish between them—and we reveal several issues with this assumption, including the impossibility of OOD detection. We also describe why it is natural for existing dataset distributions to be out-of-support with respect to one another, meaning models should not place high probability of density on these out-distribution samples. Based on this perspective, we encourage further modeling efforts aimed at improving DGMs for the application of OOD detection.

REFERENCES

Dario Amodei, Chris Olah, J. Steinhardt, Paul F. Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *ArXiv*, 2016.

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

Hyunsun Choi, Eric Jang, and Alexander Amir Alemi. Waic, but why? generative ensembles for robust anomaly detection. *ArXiv*, 2018.

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

Dan Hendrycks, Mantas Mazeika, and Thomas G. Dietterich. Deep anomaly detection with outlier exposure. In *ICLR*, 2019.

Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *NeurIPS*, 2018.

Polina Kirichenko, Pavel Izmailov, and Andrew Wilson. Why normalizing flows fail to detect out-of-distribution data. In *NeurIPS*, 2020.

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

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, and Balaji Lakshminarayanan. Detecting out-of-distribution inputs to deep generative models using typicality. In *NeurIPS Workshop on Bayesian Deep Learning*, 2019a.

Eric T. Nalisnick, A. Matsukawa, Y. Teh, D. Görür, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *ICLR*, 2019b.

Anh M Nguyen, J. Yosinski, and J. Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. 2015.

J. Ren, Peter J. Liu, E. Fertig, Jasper Snoek, Ryan Poplin, Mark A. DePristo, Joshua V. Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. *NeurIPS*, 2019.

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

J. Serrà, David Álvarez, V. Gómez, Olga Slizovskaia, J. F. Núñez, and J. Luque. Input complexity and out-of-distribution detection with likelihood-based generative models. In *ICLR*, 2020.

Ziyu Wang, Bin Dai, D. Wipf, and Jun Zhu. Further analysis of outlier detection with deep generative models. In *NeurIPS*, 2020.

A PROOF OF PROPOSITION 1

Proposition (Informal) For any choice in test statistic ϕ_p, there exists a set of alternative distributions $q \in \mathcal{Q}$ where the test does no better than random guessing.

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 $\mathbf{x} \mid \phi_p(\mathbf{x})$ 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

$$\mathbb{E}_{p(\mathbf{x})}(f_p) - \mathbb{E}_{q(\mathbf{x})}(f_p) = \mathbb{E}_{p(\phi_p(\mathbf{x}))}\Big[\mathbb{E}_{p(\mathbf{x}|\phi_p(\mathbf{x}))}(f_p) - \mathbb{E}_{q(\mathbf{x}|\phi_p(\mathbf{x}))}(f_p)\Big] \tag{2}$$

We then identify $q(\mathbf{x}|\phi_p(\mathbf{x}))$ 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 $p(\phi_p(\mathbf{x}))$. We finally define $q(\mathbf{x}) = p(\phi_p(\mathbf{x}))q(\mathbf{x}|\phi_p(\mathbf{x}))$.

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 $\phi_p(\mathbf{x})$, with probability measure $\mu = \mathbb{P}_{p(\phi_p(\mathbf{x}))}$. By assumption, $p(\mathbf{x} \mid \phi_p(\mathbf{x}))$ is non-degenerate on some μ non-measure zero set. This means there exists a set $\Phi \in \mathcal{F}_{\phi_p(\mathbf{x})}$ with $\mu(\Phi) > 0$ such that $\forall \phi_p(\mathbf{x}) \in \Phi$, $\exists A_{\phi_p(\mathbf{x})} \subset \text{supp}(p(\mathbf{x} \mid \phi_p(\mathbf{x})))$ such that $0 < \mathbb{P}_{p(\mathbf{x} \mid \phi_p(\mathbf{x}))}(A_{\phi_p(\mathbf{x})}) < 1$.

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

$$f_p(\mathbf{x}) \triangleq \mathbb{1}\left[\phi_p(\mathbf{x}) \in \Phi\right] \mathbb{1}\left[\mathbf{x} \in A_{\phi_p(\mathbf{x})}\right] + \mathbb{1}\left[\phi_p(\mathbf{x}) \notin \Phi\right] g(\mathbf{x}) \tag{3}$$

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

$$\begin{aligned} q(\mathbf{x}|\phi_p(\mathbf{x})) \triangleq \mathbb{1}\left[\phi_p(\mathbf{x}) \in \Phi\right] \Big[\frac{1}{C_{\phi_p(\mathbf{x})}}\Big(\lambda p(\mathbf{x}|\phi_p(\mathbf{x}))\mathbb{1}\left[\mathbf{x} \in A_{\phi_p(\mathbf{x})}\right] + p(\mathbf{x}|\phi_p(\mathbf{x}))\mathbb{1}\left[\mathbf{x} \notin A_{\phi_p(\mathbf{x})}\right]\Big)\Big] \\ + \mathbb{1}\left[\phi_p(\mathbf{x}) \notin \Phi\right] p(\mathbf{x}|\phi_p(\mathbf{x})) \end{aligned} \tag{4}$$

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

For simplicity, we construct q such that $\text{supp}(q(\mathbf{x} \mid \phi_p(\mathbf{x}))) = \text{supp}(p(\mathbf{x} \mid \phi_p(\mathbf{x})))$. This is to avoid any issues with an invalid joint distribution $q(\mathbf{x}, \phi_p(\mathbf{x})) \neq q(\mathbf{x})$ if $q(\mathbf{x}|\phi_p(\mathbf{x})) = 0$ (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.

$$\begin{aligned}
&\mathbb{E}_{p(\mathbf{x})}(f_p) - \mathbb{E}_{q(\mathbf{x})}(f_p) && (5)\\
=&\mathbb{E}_{p(\phi_p(\mathbf{x}))}\Big[\mathbb{E}_{p(\mathbf{x}|\phi_p(\mathbf{x}))}(f_p) - \mathbb{E}_{q(\mathbf{x}|\phi_p(\mathbf{x}))}(f_p)\Big] && (6)\\
=&\mathbb{E}_{p(\phi_p(\mathbf{x}))}\mathbb{1}\left[\phi_p(\mathbf{x}) \in \Phi\right]\Big[\mathbb{E}_{p(\mathbf{x}|\phi_p(\mathbf{x}))}(f_p) - \mathbb{E}_{q(\mathbf{x}|\phi_p(\mathbf{x}))}(f_p)\Big] && (7)\\
=&\mathbb{E}_{p(\phi_p(\mathbf{x}))}\mathbb{1}\left[\phi_p(\mathbf{x}) \in \Phi\right]\Big[\mathbb{E}_{p(\mathbf{x}|\phi_p(\mathbf{x}))}(\mathbb{1}\left[\mathbf{x} \in A_{\phi_p(\mathbf{x})}\right]) - \mathbb{E}_{q(\mathbf{x}|\phi_p(\mathbf{x}))}(\mathbb{1}\left[\mathbf{x} \in A_{\phi_p(\mathbf{x})}\right])\Big] && (8)\\
=&\mathbb{E}_{p(\phi_p(\mathbf{x}))}\mathbb{1}\left[\phi_p(\mathbf{x}) \in \Phi\right]\Big[\int_{A_{\phi_p(\mathbf{x})}} p(\mathbf{x}|\phi_p(\mathbf{x}))d\mathbf{x} - \int_{A_{\phi_p(\mathbf{x})}} q(\mathbf{x}|\phi_p(\mathbf{x}))d\mathbf{x})\Big] && (9)\\
=&\mathbb{E}_{p(\phi_p(\mathbf{x}))}\mathbb{1}\left[\phi_p(\mathbf{x}) \in \Phi\right]\Big[\int_{A_{\phi_p(\mathbf{x})}} p(\mathbf{x}|\phi_p(\mathbf{x}))d\mathbf{x} - \int_{A_{\phi_p(\mathbf{x})}} \frac{1}{C_{\phi_p(\mathbf{x})}}\lambda p(\mathbf{x}|\phi_p(\mathbf{x}))d\mathbf{x})\Big] && (10)\\
>&0 && (11)
\end{aligned}$$

Line 10 follows from the substitution of $q(\mathbf{x}|\phi_p(\mathbf{x}))$ defined in Equation (4). Line 11 follows from the fact that $\frac{\lambda}{C_{\phi_p(\mathbf{x})}} < 1$, shown below:

$$C_{\phi_p(\mathbf{x})} = \int_{\mathcal{X}} \lambda p(\mathbf{x}|\phi_p(\mathbf{x}))\mathbb{1}\left[\mathbf{x} \in A_{\phi_p(\mathbf{x})}\right] + p(\mathbf{x}|\phi_p(\mathbf{x}))\mathbb{1}\left[\mathbf{x} \notin A_{\phi_p(\mathbf{x})}\right] d\mathbf{x} \tag{12}$$

$$= \lambda \mathbb{P}_{p(\mathbf{x}|\phi_p(\mathbf{x}))}(A_{\phi_p(\mathbf{x})}) + \mathbb{P}_{p(\mathbf{x}|\phi_p(\mathbf{x}))}(A^C_{\phi_p(\mathbf{x})}) \tag{13}$$

$$= \lambda \mathbb{P}_{p(\mathbf{x}|\phi_p(\mathbf{x}))}(A_{\phi_p(\mathbf{x})}) + 1 - \mathbb{P}_{p(\mathbf{x}|\phi_p(\mathbf{x}))}(A_{\phi_p(\mathbf{x})}) \tag{14}$$

$$\frac{\lambda}{C_{\phi_p(\mathbf{x})}} = \frac{\lambda}{\lambda \mathbb{P}_{p(\mathbf{x}|\phi_p(\mathbf{x}))}(A_{\phi_p(\mathbf{x})}) + 1 - \mathbb{P}_{p(\mathbf{x}|\phi_p(\mathbf{x}))}(A_{\phi_p(\mathbf{x})})} \tag{15}$$

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

$$< 1 \tag{17}$$

Line 17 holds since the denominator in the previous line is greater than 1: Since $0 < \lambda < 1$, $\frac{1}{\lambda} > 1$. Then, $\mathbb{P}_{p(\mathbf{x}|\phi_p(\mathbf{x}))}(A_{\phi_p(\mathbf{x})}) + \frac{1}{\lambda}\left[\mathbb{P}_{p(\mathbf{x}|\phi_p(\mathbf{x}))}(A^c_{\phi_p(\mathbf{x})})\right] > \mathbb{P}_{p(\mathbf{x}|\phi_p(\mathbf{x}))}(A_{\phi_p(\mathbf{x})}) + \mathbb{P}_{p(\mathbf{x}|\phi_p(\mathbf{x}))}(A^c_{\phi_p(\mathbf{x})}) = 1$.

Having constructed the density 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:

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

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 FIGURE 1 DETAILS

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

Consider an in-distribution p that is bivariate Gaussian with an identity covariance matrix. There are a variety of distributions q 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 $z^2 = (x_1^2 + x_2^2)/2$ (i.e. preserving distance from origin).

The out-distributions q and r maintain the same distribution of log densities as p since they distribute mass similarly across the upper level sets $\{\mathbf{x} : p(\mathbf{x}) > t\}$.

C ALL REJECTION RULES CAN BE WRITTEN AS $\phi_p(\mathbf{x}) > k$

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

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 rejects when the negative log likelihood is above a

certain threshold k, whereas the typicality test rejects when the distance to the training set entropy is above k.

D DECOMPOSITION

Here we derive the decomposition in Section 3.3, Equation (1).

$$\begin{aligned}
&\mathbb{P}\Big(\phi_{p_\theta}(\mathbf{x}) < \phi_{p_\theta}(\mathbf{y})\Big) \\
=&\mathbb{P}\Big(\phi_{p_\theta}(\mathbf{x}) + \phi_p(\mathbf{x}) - \phi_p(\mathbf{x}) < \phi_{p_\theta}(\mathbf{y}) + \phi_p(\mathbf{y}) - \phi_p(\mathbf{y})\Big) \\
=&\mathbb{P}\Big(\phi_p(\mathbf{x}) < \phi_p(\mathbf{y}) + D_p - D_q\Big), \text{where} \\
&D_p = \big[\phi_p(\mathbf{x}) - \phi_{p_\theta}(\mathbf{x})\big], D_q = \big[\phi_p(\mathbf{y}) - \phi_{p_\theta}(\mathbf{y})\big]
\end{aligned}$$

E EXPERIMENT DETAILS: OOD DETECTION UNDER MODEL MISESTIMATION

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 pretrained GLOW model * with temperature 1. We generate 40,000 samples for training and 10,000 samples for evaluation. These are the analogous sizes to the train and test sets of the CIFAR-10 dataset, which is the dataset the model p was pretrained on.

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.

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

*https://openaipublic.azureedge.net/glow-demo/logs/abl-1x1-aff.tar