Making Generative Models Know What They Don't Know via Hypothesis Testing

Leander Kurscheidt [1] **Antonio Vergari** [1]

Abstract

Out-of-distribution (OOD) detection is a fundamental problem in machine learning, and normalizing flows (Dinh et al., 2015) are natural candidates due to their explicit density. However, likelihood alone is known to be unreliable for OOD detection, as flows may assign higher likelihood to OOD samples than to in-distribution data (Nalisnick et al., 2019a). In this work, we propose a goodness-of-fit (GoF) test for normalizing flows based on latent-space statistics. Our method enables both direct assessment of model fit on a held-out dataset via uniformity testing of the p-values, and single-sample OOD detection using a recalibrated test statistics. In contrast to existing likelihood- and score-based approaches, this allows principled evaluation of whether the learned flow itself fits the data distribution. We further show that recalibrated p-values provide an interpretable and flexible scoring mechanism for OOD detection when the underlying flow model achieves a sufficiently good fit.

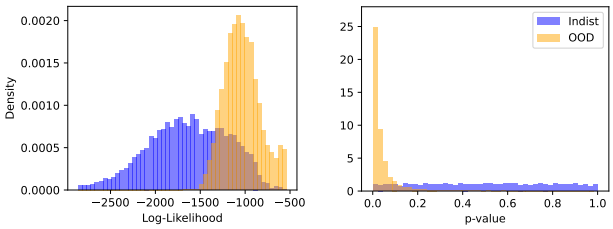


Figure 1. **The radial hypothesis test correctly separates in- from out-of-distribution samples, while density scores are higher on the out-of-distribution** as shown for a Glow model trained on FashionMNIST and evaluated on MNIST (OoD), where MNIST samples obtain higher likelihood than in-distribution samples (left, the density paradox (Nalisnick et al., 2019a)), while the distribution of p-values correctly separates them (right).

1. Introduction

In safety-critical systems, knowing whether to rely on predictions from machine learning (ML) models is crucial (Hendrycks & Gimpel, 2016; Marconato et al., 2024). While the performance on a hold-out dataset is a reasonable estimate of the performance of a machine learning system on unseen data, this assumes that the distribution of both datasets stays stationary. As neural networks can provide arbitrary predictions with high-confidence on unseen data (Hein et al., 2019), it is crucial to detect when a datapoint is unlikely to belong to a dataset. This task is known as Out-of-Distribution (OoD) detection.

A common approach in ML is to use score-based methods such as Deep SVDD (Ruff et al., 2018) or autoencoders (Zhou & Paffenroth, 2017). However, these approaches often suffer from poorly interpretable scores, training hyperparameters that are difficult to tune, and black-box models whose fit to the data distribution is difficult to assess, especially if there is no access to outlier-examples beforehand.

In comparison, density based-models like normalizing flows (Dinh et al., 2015) combine a number of advantages. First off, especially for models with an explicit density, training them via maximizing the log-likelihood is a well understood and straightforward process. Additionally, the generative nature of these models makes it possible to assess the fit to the dataset purely using expert knowledge, which is not possible with a purely classification-based approach. And since these models are trained to capture all aspects of the in-distribution, they are promising candidates for the outlier-detection task. Despite these promising properties, reliably leveraging density-based models for outlier detection remains challenging, as high likelihood alone does not necessarily correspond to semantic in-distribution membership (Nalisnick et al., 2019a). This lead to a recent interest in goodness-of-fit approaches to tackle the OoD-problem (Nalisnick et al., 2019b; Bergamin et al., 2022). These approaches, however, often rely on estimators with limited power in the single-sample OoD regime, making reliable detection of individual outliers challenging. At the same time, comparatively little attention has been given to directly

[1]School of Informatics, University of Edinburgh, UK. Correspondence to: Leander Kurscheidt <l.kurscheidt@sms.ed.ac.uk>.

ICML 2026 Workshop on Hypothesis Testing, Seoul, South Korea, 2026. Copyright 2026 by the author(s).

assessing the statistical fit of the generative model itself.

In this paper, we propose goodness-of-fit tests for normalizing flows and investigate the following research-questions:

R1 Can the tests scale to high-dimensional settings?
R2 Are the tests applicable in the single-sample regime?
R3 Can the tests directly assess the flow model fit?
R4 Are the tests effective in out-of-distribution detection?

2. Background

Notation. We denote random variables by uppercase letters (e.g., X, Z), and their assignments with lowercase ones (e.g., x, z). Bold symbols denote sets of variables (e.g., $\mathbf{X}, \mathbf{Z}$), and their joint assignments (e.g., $\mathbf{x}, \mathbf{z}$). We denote the domain of $\mathbf{X}$ with $\mathcal{X}$ and of $\mathbf{Z}$ with $\mathcal{Z}$.

Normalizing Flows. Normalizing flows (Dinh et al., 2015) are a flexible and powerful class of generative models that combine expressive transformations with exact and tractable density evaluation. They are based on the multivariate change-of-variables formula and assume that the neural network f from data-space $\mathcal{X}$ into a latent-space $\mathcal{Z}$ is a diffeomorphism with a tractable Jacobian determinant (or log-determinant). This is usually achieved by restricting the layers to have an upper or lower-triangular Jacobian, in which case the determinant reduces to the product of the diagonal (Dinh et al., 2016). By choosing a simple latent-distribution with a tractable density p_z, which we assume to be $\mathcal{N}(0, I)$, a very common choice (Dinh et al., 2016; Kingma & Dhariwal, 2018), we have a tractable, exact log-likelihood objective we can use to train the model:

$$\log p_{model}(\mathbf{x}) = \log p_z(f(\mathbf{x})) + \sum_{i \in \text{Layers}} \log \det J_{f_i}(\mathbf{x})$$

OoD-Detection. Out-of-Distribution detection is the task of differentiating between unusual and usual samples from your data-distribution p_{data} (Aggarwal, 2016). It is usually approached in an unsupervised setting, with only access to samples from your in-distribution p_{data} and no or only a very limited notion about the OoD-distribution p_{out}. Common application-domains include fraud-detection (Phua et al., 2010), intrusion detection in cybersecurity (Garcia-Teodoro et al., 2009) or visual anomaly detection in industrial inspection scenario (Bergmann et al., 2019). In this paper, we assume to have only access to p_{data}, no access to samples of p_{out} and want to classify a single sample for In- or OoD.

Goodness-of-Fit Tests. We assume a model distribution p_{model} fitted to samples from p_{data}. The setting in this paper is to decide between $\mathcal{H}_0 : \mathbf{X} \sim p_{model}$ and $\mathcal{H}_1 : \mathbf{X} \not\sim p_{model}$ given a previously unseen sample $\mathbf{x}$. In hypothesis-testing, this is typically approached via a test statistic T, whose distribution under $\mathcal{H}_0$ is known exactly or asymptotically (D'Agostino, 2017). This allows for the the construction of a rejection rule based on $T(\mathbf{x}) > \tau_\alpha$, where τ_α is chosen to control the Type-I error (i.e., the probability of incorrectly rejecting $\mathcal{H}_0$) at level α. In some settings, the test statistic T is only implicit, as the decision rule operates directly on derived quantities such as recalibrated p-values (Sellke et al., 2001).

3. Related Work

A number of statistical GoF test for normalizing flows have been developed. Nalisnick et al. (2019b) introduces a hypothesis test for outlier detection based on typical sets (Cover, 1999), which is based on thresholding the empirical entropy for a number of samples based on observed upper and lower-bounds. While it can be used in the single-sample setting, as it used the empirical entropy $-\frac{1}{N}\sum_i \log p(\mathbf{x}_n)$ it is theoretically more suited for a multiple-sample setting. Bergamin et al. (2022) build upon this work and combines the typicality test with a parametric score-based tests (Rao, 1948) using Fisher's-method (Fisher, 1970). An important drawback of these approaches is that they require estimating the distribution under H_0 using a hold-out validation set as an analytical distribution is not given, and therefore is constrained to the OoD-detection task.

A concurrent line of research tries to establish outlier-scores based on normalizing flows. Choi et al. (2018) proposes a way to combine ensembles using the Watanabe Akaike Information Criterion (WAIC) (Watanabe & Opper, 2010). Ren et al. (2019) proposes a test based on the ratio between the likelihood for the images and the likelihood for the background estimated by a separate background model.

4. Going Into The Latent-Space

A main advantage of normalizing flows (Dinh et al., 2015) is the availability of the an exact likelihood. It is therefore natural to focus on the likelihood for OoD-Detection (Bishop, 1993). Unfortunately, as show by (Nalisnick et al., 2019a), normalizing flows exhibit a paradox: When trained on complex image-distributions, they routinely assign higher likelihood to OoD-datasets that have a simpler structure. For example, when comparing likelihoods for the test-set of MNIST (lec, 2002) to the test-set of Fashion-MNIST (Xiao et al., 2017) for a model trained on Fashion-MNIST, normalizing flows tend to assign higher-likelihood to the out-of-distribution samples, as shown in figure 1. This phenomenon was further analyzed by Kirichenko et al. (2020) and explained by the strong, local pixel-correlations in these simpler datasets leading to high likelihoods even if the global structure is violated.

In this paper, we want to switch the perspective. Let us

first assume a perfect flow f with no approximation error. Due to the learning-objective, the flow f transforms the data-distribution p_{data} to the latent-distribution p_z, which we assume to be independent gaussians, so $p_z = \mathcal{N}(0, I)$. We can now switch our perspective, as the flow f is a diffeomorphism:

$$H_0 : \mathbf{X} \sim p_{data} \tag{1}$$
$$\iff f(\mathbf{X}) \sim p_z = \mathcal{N}(0, I) \tag{2}$$

As GoF-test for the normal-distribution with fixed parameters is a well-studied concept, we have a range of test-statistics to approach this hypothesis test. In this paper, we focus on a selection of powerful tests (Anderson, 2003) for this scenario:

Radial. This test uses the test-statistic $T(x) = ||x||^2$ with $||x||^2 \sim \chi_2$.

Max-Order. Here, we use the test-statistic $T(x) = \max_i |x_i|$ with $P(\max_i |x_i| \leq t) = (2\Phi(t) - 1)^d$.

Kurtosis. In the kurtiosis test, we have $T(x) = \frac{1}{d}\sum_{i=1}^{d} x_i^4$ and under H_0 $T(x) \approx \mathcal{N}\left(3, \frac{96}{d}\right)$ using the CLT.

Hermite (Neyman, 1937; Rayner & Best, 1990). This test is categorized as a smooth GoF-test, which is suitable to detect smooth derivations in the distribution from $\mathcal{N}(0, I)$. It is based on the (orthogonal) Hermite polynomials H_k. Now, set

$$\tilde{X}_k = \frac{1}{\sqrt{d}} \sum_{i=1}^{d} H_k(x_i), \qquad T(x) = \sum_{k=1}^{4} \frac{\tilde{X}_k^2}{\sigma_k^2}. \tag{3}$$

By the CLT, $\tilde{X}_k \Rightarrow \mathcal{N}(0, \sigma_k^2)$, and orthogonality of Hermite polynomials implies asymptotic independence across k. Hence $T(x) \approx \chi_4^2$ $(d \to \infty)$.

As these goodness-of-fit tests are quick to compute even for very high-dimensional gaussian distributions, we can positively answer **R1** (scalability). They are also easily applicable to the single-sample regime, especially for higher-dimensional latent-distributions, and hence fulfill **R2**.

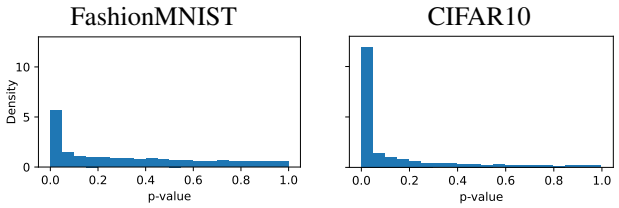


Figure 2. **By using the GoF-test on a hold-out dataset, one can assess the fit of the model to the data distribution.** We show a histogram over the observed p-values for two Glow-models (Kingma & Dhariwal, 2018) for the hermite-GoF test on the unseen validation data for FashionMNIST (Left) and CIFAR10 (Right).

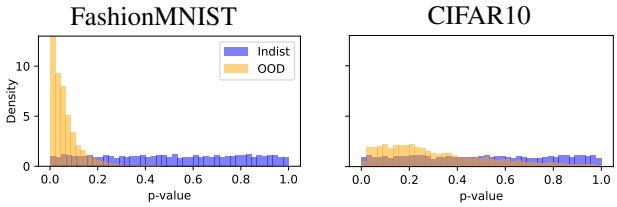


Figure 3. **On a properly trained model, we are able to successfully distinguish between in- and outliers.** We show the same setting as in figure 2 for the recalibrated p-values on an unseen test-set and compare them to the recalibrated p-values on MNIST (Left) and SVHN (Right).

By assessing the distribution of the p-values on a holdout-dataset, one can judge the fit of the model. A perfect fit would result in uniformly-distributed p-values. In comparison, by only comparing likelihoods on the train- vs validation dataset, only overfitting but not underfitting can be shown. This is visualized in figure 2, where we show the p-value distribution for the hermite GoF-test on two Glow-models trained on FashionMNIST and CIFAR-10 (Hinton et al., 2012). One can see that the model on FashionMNIST, while having a spike on the left, exhibits significantly more uniformly distributed p-values compared to the model trained on CIFAR10, a hard to fit dataset. In addition to using expert-knowledge on generative samples, a principled GoF-test on a holdout-dataset is another important measure to judge the fit of the distributions and not possible using the typicality or score-based methods for normalizing flows.

In case the fit is judged as sufficient, this hold-out dataset can be used to recalibrate the p-values to achieve a uniform distribution. This is done by using the empirical CDF of the observed p-values. Using these recalibrated p-values, we can define a significance-level to flag samples for out-of-distribution. This is very effective for properly trained models, for example the Glow-model trained on FashionMNIST, but less effective for models not close to a uniform distribution in the validation-dataset, as shown in figure 3.

The recalibration step transforms the goodness-of-fit test from relying purely on the model-implied null distribution to empirically calibrated p-values based on a held-out in-distribution dataset, and therefore integrates well with the approximate nature of machine-learning. The availability of a held-out dataset is a common assumption and a requirement of the typicality- (Nalisnick et al., 2019b) and score-based (Bergamin et al., 2022) tests to yield p-values.

5. Results

In this-section, we want to investigate research-question **R3** and **R4**, so assess whether the GoF-tests can be used to directly asses the fit of the model and whether they are effective for out-of distribution detection.

Table 1. **Goodness-of-Fit tests achieve strong outlier-detection performance for well-fitted generative models, but their effectiveness deteriorates when the underlying model fit is poor.** We show AUROC↑ for single-sample OOD detection for a Glow-model trained on FashionMNIST or CIFAR-10 vs MNIST or either SVHN or CIFAR100.

		Proposed GoF Tests				**Prior Work**		
In-Distribution	OoD	Radial	Max-Coordinate	Kurtosis	Hermite	Typicality	Score-Stat.	Fisher
FashionMNIST	MNIST	**0.961747**	0.804272	0.941513	0.933804	0.745519	0.830463	0.856139
CIFAR-10	SVHN	0.607916	0.412003	0.476175	0.676845	0.887873	0.832897	**0.897248**
CIFAR-10	CIFAR100	0.544030	0.532759	0.548439	0.545382	0.548296	0.559352	**0.576427**

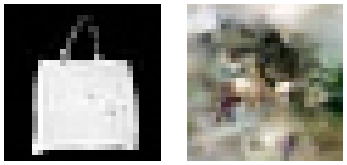
In this section, we compare the performance of two Glow-models trained on FashionMNIST and CIFAR10, we provide details about the architecture and training setup in the appendix in A.

Assessing The Fit Of The Model. We will first turn our attention to **R3** and investigate whether the GoF-tests can be used to judge the fit of the normalizing flows to the held-out validation datasets, in our setting unseen by the model. We provide KS statistic for the p-values of the proposed GoF-tests in table 2. One can see a significantly better p-value distribution according to the KS-statistics for the FashionMNIST-model, which corresponds to a better image quality as shown by two random samples in figure 4. We can therefore positively answer **R3**, providing a statistically principled approach to assessing model fit and thereby contributing a useful tool to the toolbox of machine learning practitioners in density estimation.

Table 2. **Goodness-of-Fit tests enable direct assessment of model fit for normalizing flow-based models.** Calibration results (KS statistic ↓) for in-distribution p-values for the Glow-model.

	Radial	**Max-Coord.**	**Kurtosis**	**Hermite**
FashionMNIST	0.21	0.25	0.26	0.26
CIFAR-10	0.47	0.28	0.47	0.57

FashionMNIST **CIFAR10**



We show two random samples from Glow models trained on FashionMNIST and CIFAR10. The observed KS statistics on in-distribution p-values (Table 2) correlates to image-fidelity.

Figure 4. **Goodness-of-fit reflects image fidelity.**

Out-Of-Distribution Detection. In order to investigate **R4**, we show the out-of-distribution performance of our hypothesis tests on the glow-models against MNIST for the in-distribution FashionMNIST and against SVHN (Netzer et al., 2011) and CIFAR100 (Krizhevsky et al., 2009) for CIFAR-10. We use the negative p-values induced by the hypothesis tests as outlier scores and we evaluate the performance by computing the area-under-the-curve for the ROC-characteristic (Hastie et al., 2009). Here, we compare against typicality (Nalisnick et al., 2019b), the score-based GoF-test (Rao, 1948) and the combination of both using Fisher's-methods proposed by Bergamin et al. (2022) (denoted by Fisher). Here, a mixed picture emerges. One can see a very strong outlier-detection performance on the well-fitted FashionMNIST-model. With a AUROC of 0.96 for the Radial-test, our method improves significantly on the Bergamin et al. (2022) with a AUROC of 0.85. But for the model trained CIFAR-10, a dataset that is very hard to fit for generative models, and evaluated against SVHN, the performance of all our GoF-tests degrades compared to the baselines. This might indicate that the model does not manage to achieve a latent-space distribution close to the normal-distribution, as also indicated by Table 2. In general, one can see that Radial, Kurtiosis and Hermite have significantly better AUROC-scores than Max-Coordinate, indicating that it is less suitable for the goodness of fit of normalizing flows.

6. Conclusion

In this paper, we propose a range of goodness-of-fit tests to assess the fit of normalizing flows by going into the latent-space. These tests are principled, scalable and applicable to the single-sample-regime. We empirically show that the distribution of the p-values on a hold-out set corresponds to the visual fidelity of the generative samples, which enables practitioners to assess underfitting of density models even on large-scale datasets where manual inspections does not scale. For the outlier-detection performance, a mixed picture emerges. The proposed test achieve superior outlier-detection performance for the well-fitted FashionMNIST model, but for the CIFAR10-model, which has a latent-distribution that is less normal, we degrade significantly while the baselines still achieve strong performance against SVHN. A promising avenue to improve the performance of the goodness of fit-tests in the latent-space is to built upon the insights by Kirichenko et al. (2020). Currently, the influence of all latent-variables is weighted equally, but in images most latents encode local correlations. A principled way to weigh them differently might unlock significantly better OoD-performance by focusing more on latents modeling semantic information.

Contribution

LK conceived the initial idea of the project and is responsible for all theoretical contributions, illustrations, algorithms and the implementation. LK wrote the paper with help from AV.

Acknowledgements

AV was supported by the "UNREAL: Unified Reasoning Layer for Trustworthy ML" project (EP/Y023838/1) selected by the ERC and funded by UKRI EPSRC.

References

Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 2002.

Aggarwal, C. C. An introduction to outlier analysis. In *Outlier analysis*, pp. 1–34. Springer, 2016.

Anderson, T. W. Introduction to multivariate statistical analysis,. 11, 2003. doi: 10.1017/S0013091500021842.

Bergamin, F., Mattei, P., Havtorn, J. D., Sénétaire, H., Schmutz, H., Maaløe, L., Hauberg, S., and Frellsen, J. Model-agnostic out-of-distribution detection using combined statistical tests. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I. (eds.), *International Conference on Artificial Intelligence and Statistics, AISTATS 2022, 28-30 March 2022, Virtual Event*, Proceedings of Machine Learning Research, pp. 10753–10776. PMLR, 2022. URL https://proceedings.mlr.press/v151/bergamin22a.html.

Bergmann, P., Fauser, M., Sattlegger, D., and Steger, C. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9592–9600, 2019.

Bishop, C. M. Novelty detection and neural network validation. In *International Conference on Artificial Neural Networks*, pp. 789–794. Springer, 1993.

Choi, H., Jang, E., and Alemi, A. A. Waic, but why? generative ensembles for robust anomaly detection. *arXiv preprint arXiv:1810.01392*, 2018.

Cover, T. M. *Elements of information theory*. John Wiley & Sons, 1999.

D'Agostino, R. B. *Goodness-of-fit-techniques*. Routledge, 2017.

Dinh, L., Krueger, D., and Bengio, Y. NICE: non-linear independent components estimation. In Bengio, Y. and LeCun, Y. (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015. URL http://arxiv.org/abs/1410.8516.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using real NVP. *CoRR*, abs/1605.08803, 2016. URL http://arxiv.org/abs/1605.08803.

Fisher, R. A. Statistical methods for research workers. In *Breakthroughs in statistics: Methodology and distribution*, pp. 66–70. Springer, 1970.

Garcia-Teodoro, P., Diaz-Verdejo, J., Maciá-Fernández, G., and Vázquez, E. Anomaly-based network intrusion detection: Techniques, systems and challenges. *computers & security*, 28(1-2):18–28, 2009.

Hastie, T., Tibshirani, R., Friedman, J., et al. The elements of statistical learning, 2009.

Hein, M., Andriushchenko, M., and Bitterwolf, J. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 41–50. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00013. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Hein_Why_ReLU_Networks_Yield_High-Confidence_Predictions_Far_Away_From_the_CVPR_2019_paper.html.

Hendrycks, D. and Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *CoRR*, abs/1610.02136, 2016. URL http://arxiv.org/abs/1610.02136.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

Kingma, D. P. and Dhariwal, P. Glow: Generative flow with invertible 1x1 convolutions. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 10236–10245, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/d139db6a236200b21cc7f752979132d0-Abstract.html.

Kirichenko, P., Izmailov, P., and Wilson, A. G. Why normalizing flows fail to detect out-of-distribution data. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/ecb9fe2fbb99c31f567e9823e884dbec-Abstract.html.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Marconato, E., Bortolotti, S., van Krieken, E., Vergari, A., Passerini, A., and Teso, S. Bears make neuro-symbolic models aware of their reasoning shortcuts. In Kiyavash, N. and Mooij, J. M. (eds.), *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pp. 2399–2433. PMLR, 15–19 Jul 2024. URL https://proceedings.mlr.press/v244/marconato24a.html.

Nalisnick, E. T., Matsukawa, A., Teh, Y. W., Görür, D., and Lakshminarayanan, B. Do deep generative models know what they don't know? In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019a. URL https://openreview.net/forum?id=H1xwNhCcYm.

Nalisnick, E. T., Matsukawa, A., Teh, Y. W., and Lakshminarayanan, B. Detecting out-of-distribution inputs to deep generative models using a test for typicality. *CoRR*, abs/1906.02994, 2019b. URL http://arxiv.org/abs/1906.02994.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A. Y., et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, pp. 4. Granada, 2011.

Neyman, J. Smooth test for goodness of fit. *Scandinavian Actuarial Journal*, 1937(3-4):149–199, 1937.

Phua, C., Lee, V., Smith, K., and Gayler, R. A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119*, 2010.

Rao, C. R. Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. In *Mathematical proceedings of the cambridge philosophical society*, volume 44, pp. 50–57. Cambridge University Press, 1948.

Rayner, J. and Best, D. Smooth tests of goodness of fit: an overview. *International Statistical Review/Revue Internationale de Statistique*, pp. 9–17, 1990.

Ren, J., Liu, P. J., Fertig, E., Snoek, J., Poplin, R., Depristo, M., Dillon, J., and Lakshminarayanan, B. Likelihood ratios for out-of-distribution detection. *Advances in neural information processing systems*, 32, 2019.

Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S. A., Binder, A., Müller, E., and Kloft, M. Deep one-class classification. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4393–4402. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/ruff18a.html.

Sellke, T., Bayarri, M. J., and Berger, J. O. Calibration of ρ values for testing precise null hypotheses. *The American Statistician*, 55(1):62–71, 2001.

Watanabe, S. and Opper, M. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of machine learning research*, 11(12), 2010.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017. URL http://arxiv.org/abs/1708.07747.

Zhou, C. and Paffenroth, R. C. Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 665–674, 2017.

A. Architecture and Training Details

Both models are glow-normalizing flows (Kingma & Dhariwal, 2018).

We train using the following hyperparameters, following the convention of Kirichenko et al. (2020).

On FashionMNIST.

- epochs: 151
- lr: $1e{-}5$
- batch-size: 32
- optimizer: Adam
- num-scales: 2
- num-coupling-layers-per-scale: 16
- st-type: highway
- num-blocks: 3
- num-mid-channels: 200

On CIFAR-10.

- epochs: 161
- lr: $1e{-}5$
- batch-size: 128
- optimizer: RMSprop
- num-scales: 3
- num-coupling-layers-per-scale: 8
- st-type: highway
- num-blocks: 3
- num-mid-channels: 400