Entropic Issues in Likelihood-Based OOD Detection

Anonymous Author(s)
Affiliation
Address
email

Abstract

Deep generative models trained by maximum likelihood remain very popular methods for reasoning about data probabilistically. However, it has been observed that they can assign higher likelihoods to out-of-distribution (OOD) data than in-distribution data, thus calling into question the meaning of these likelihood values. In this work we provide a novel perspective on this phenomenon, decomposing the average likelihood into a KL divergence term and an entropy term. We argue that the latter can explain the curious OOD behaviour mentioned above, suppressing likelihood values on datasets with higher entropy. Although our idea is simple, we have not seen it explored yet in the literature. This analysis provides further explanation for the success of OOD detection methods based on likelihood ratios, as the problematic entropy term cancels out in expectation. Finally, we discuss how this observation relates to recent success in OOD detection with manifold-supported models, for which the above decomposition does not hold.

1 Introduction

Deep Generative Models (DGMs) are a staple for probabilistic modelling of high-dimensional data, and are typically trained by maximum likelihood – or approximations thereof – over observations. One possible application of DGMs is identifying whether new data is in- or out-of-distribution (OOD), which naïvely seems quite straightforward given a sufficiently powerful model: average likelihood could always be further increased by raising it for observed points, and thus lowering it for OOD data. However, Nalisnick et al. [19] provided a landmark study showing that DGMs do not behave in this way, as models trained on the CIFAR-10 dataset [13] produce higher likelihoods when tested on Street View House Numbers (SVHN) [21], and similarly for models trained on Fashion-MNIST [32] and tested on MNIST [15]. In either case, we note that the phenomenon only occurs in one direction, with the more complicated dataset of the pair (i.e. CIFAR-10 or Fashion-MNIST) consistently scoring lower likelihoods than the simpler one (i.e. SVHN or MNIST). Various explanations have been posed for this phenomenon, including OOD data residing in high likelihood but low probability regions [19, 20], early layers of the architectures encoding generic information [11, 27], poor model fit [34], or representations of the data being being inadequate [8, 14].

In this paper, we propose a simple, alternative explanation for this phenomenon, centred on a standard decomposition of the average likelihood into (i) a KL divergence term between the true data distribution and the model, and (ii) an entropy term. The latter is constant with respect to the parameters of the DGM, and is thus easily ignored in optimization, but can wildly vary over different datasets. In particular, data distributions with high entropy may be biased towards lower likelihoods, providing a new perspective on how even a perfect DGM could be unable to perform OOD detection. This interpretation further explains the success of OOD detection methods based on likelihood ratios, which we see are invariant to the entropy term in expectation. We also examine how our analysis relates with the typical set approach [20], which itself only considers the entropy of the in-distribution set and not the incoming test data, thus becoming exposed to pathological cases where likelihoods

Submitted to NeurIPS 2021 Workshop on DGMs and Downstream Applications. Do not distribute.

are similar but the data is far different. As an addendum to the main analysis of this work, we note that the likelihood decomposition does not apply to DGMs supported on low-dimensional manifolds embedded in the higher-dimensional ambient space. We conjecture that this may account for some of the recent improvements in OOD detection for models of this type, as they may be less susceptible to the influence of the entropy term.

2 Background

Deep Generative Models DGMs are, loosely speaking, models which combine probabilistic elements with deep neural networks. In this paper, we focus on DGMs which admit densities[1] and are trained by maximum likelihood – or an approximation thereof – such as normalizing flows [23, 12], variational auto-encoders (VAEs) [9, 25], and energy-based models [16, 6]. DGMs are often used in practice as *density estimators*, since, in expectation, maximizing a model's likelihood over an observed dataset also minimizes its KL divergence from the true distribution generating this data (cf. (1)). Likelihood-based DGMs have demonstrated practical success across a wide variety of domains (e.g. Kingma et al. [10], Gao et al. [7], Oord et al. [22], Townsend et al. [29], Vahdat and Kautz [30]), which might suggest that the likelihood values outputted by these models are reliable proxies for the probability that new data is similar to the data on which the model was trained.

Failures on Out-of-Distribution Detection However, Nalisnick et al. [19] discovered that this is *not* the case, as DGMs can often assign higher likelihoods to data which is dissimilar to the training set, or out-of-distribution (OOD), particularly if this new data is considered to be "simpler" than the training set. For example, as mentioned earlier, models trained on CIFAR-10 produce higher likelihoods when evaluated on SVHN data, with similar phenomena observed for other pairs of datasets. This observation has led researchers to question the meaning of likelihood values outputted by DGMs in an attempt to both understand why this phenomenon occurs and what (if anything) can be done about it; we discuss approaches along either of these directions throughout this work.

3 The Influence of Entropy on the Likelihood

In this section, we provide a simple and straightforward explanation for this strange OOD phenomenon based on the standard decomposition of the likelihood into a KL divergence term and an entropy term. We argue that the latter can inflate or deflate likelihoods, thus making these values unreliable on their own. Throughout this section we assume all distributions admit densities, either with respect to the Lebesgue measure on the ambient space for continuous data, or the counting measure for discrete data; our analysis covers both cases. This agrees with standard assumptions made when building full-support DGMs, and we will discussion the implications of relaxing this in section 5.

3.1 Decomposition of the Likelihood

Suppose we have a dataset $\mathcal{D}_P := \{x_i\}_{i=1}^n$, with each x_i sampled i.i.d. from an unknown data-generating distribution P. Suppose also that we have a model for this distribution parametrized by θ, denoted P_θ, which admits a density p_θ. We can write the typical likelihood objective to estimate θ as $\sum_{i=1}^n \log p_\theta(x_i)/n$, which can be rewritten in the limit of infinite data as

$$\mathcal{L}(\theta) \coloneqq \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \log p_\theta(x_i) = \mathbb{E}_{X\sim P}[\log p_\theta(X)] = -D_{\mathrm{KL}}\left(P \;||\; P_\theta\right) - \mathbb{H}[P], \tag{1}$$

where D_{KL} denotes the KL divergence and $\mathbb{H}$ denotes the entropy. This well-known result can motivate maximum likelihood estimation – indeed, Nalisnick et al. [19] provide a similar decomposition in (1) of their work – as we have $\mathrm{argmax}_\theta\, \mathcal{L}(\theta) = \mathrm{argmin}_\theta\, D_{\mathrm{KL}}\left(P \;||\; P_\theta\right)$ since $\mathbb{H}[P]$ is independent of θ.

However, this result also provides a straightforward explanation for why we see strange behaviour when evaluating the likelihood of a model on a *different* dataset $\mathcal{D}_Q := \{y_j\}_{j=1}^m$, with each y_j sampled i.i.d. from an unknown distribution Q with the same support as P. We can again write the

[1]Either with respect to the Lebesgue measure for continuous data, or the counting measure for discrete data

average likelihood of P_θ over $\mathcal{D}_Q$ asymptotically as

$$\lim_{m\to\infty} \frac{1}{m} \sum_{j=1}^{m} \log p_\theta(y_j) = \mathbb{E}_{X\sim Q}[\log p_\theta(X)] = -D_{\mathrm{KL}}\left(Q \mid\mid P_\theta\right) - \mathbb{H}[Q]. \tag{2}$$

Now, suppose $Q \neq P$, so that $\mathcal{D}_Q$ is considered OOD. If $\mathbb{H}[P]$ is much larger than $\mathbb{H}[Q]$, we can easily have $\mathbb{E}_{X\sim Q}[\log p_\theta(X)] > \mathbb{E}_{X\sim P}[\log p_\theta(X)]$ *even if P_θ is a perfect model of P.* Qualitatively, CIFAR-10 (resp. Fashion-MNIST) is far more complex – and thus ostensibly has higher entropy – than SVHN (resp. MNIST), which succinctly explains the observation that models trained on the former score higher likelihoods on the latter. Nalisnick et al. [19] additionally note that constant (and thus low-entropy) inputs score very highly on likelihood, agreeing with the analysis above. At any rate, this discussion provides further insight on the unreliability of using only the likelihood as a test statistic to evaluate whether new data is in- or out-of-distribution.

3.2 Likelihood Ratios for OOD Detection Cancel Out the Entropy

The above analysis suggests that performing likelihood-based OOD detection without somehow accounting for the entropy of incoming data is problematic. On the other hand, OOD detection methods based on likelihood *ratios* have recently demonstrated strong performance [24, 28, 27]. Here, we argue that such approaches intrinsically control for this entropy. Suppose we have a model P_θ with density p_θ trained on data from P, and we also have access to some *reference* model R_ϕ admitting a density r_ϕ; Ren et al. [24] forms this as a model for the background, Schirrmeister et al. [27] trains this on millions of unrelated images, and Serrà et al. [28] uses standard image compression techniques such as PNG as a proxy for this distribution. The above methods assess incoming data x as OOD if they score poorly on the likelihood ratio $\log(p_\theta(x)/r_\phi(x))$. Now, if we assume $x \sim Q$ for some unknown Q, we can take the expectation of the likelihood ratio, following (2):

$$\mathbb{E}_{X\sim Q} \log p_\theta(X) - \mathbb{E}_{X\sim Q} \log r_\phi(X) = -D_{\mathrm{KL}}\left(Q \mid\mid P_\theta\right) - \cancel{\mathbb{H}[Q]} + D_{\mathrm{KL}}\left(Q \mid\mid R_\phi\right) + \cancel{\mathbb{H}[Q]}. \tag{3}$$

In particular, we note that this does not explicitly depend on the entropy of incoming data $\mathbb{H}[Q]$. This might explain the success of likelihood ratio methods in OOD detection, as (3) will be mediated by the distance from Q to P, assuming P_θ is a good model of P and $D_{\mathrm{KL}}\left(Q \mid\mid R_\phi\right) \lessapprox D_{\mathrm{KL}}\left(P \mid\mid R_\phi\right)$.

4 Related Work

As far as we are aware, nobody has considered the effect of the entropy term on OOD detection. We will review some alternative explanations here and provide further discussion.

Perfect Models Le Lan and Dinh [14] also noted that OOD detection methods based on likelihood scoring are unreliable *even when provided with a perfect density model of in-distribution data*, although under the observation that likelihood-based OOD detection techniques are not invariant to different representations of the same data. Our work instead focuses on the impact of the entropy term in OOD detection and thus provides a complimentary analysis; perhaps this suggests that we should look for lower-entropy representations of data before performing likelihood-based OOD detection.

Likelihood Ratios Bishop [2] is the earliest instance of a likelihood ratio approach for novelty detection of which we are aware, with much more recent work heading in this direction as mentioned in subsection 3.2. Of note is that Serrà et al. [28] and Schirrmeister et al. [27] require trained models on data of the same generic type but not the same particular distribution – an image compression algorithm and a density model on 80 million tiny images, respectively – thus requiring additional knowledge besides just the in-distribution training data. However, other works [14, 34] have indeed suggested that some knowledge of the out-distribution is crucial to attacking the OOD detection problem. Ren et al. [24] do not require additional data, but still require training a *background* model on randomly perturbed in-distribution data. Noting that the above do not work well for VAEs, Xiao et al. [33] develop a successful OOD detection method comparing the likelihood of the posterior optimized for a single observation against the likelihood from the entire VAE. All of these approaches control for the entropy of incoming data as previously discussed in subsection 3.2, and further can assess *individual* inputs as in- or out-of-distribution which is an attractive property for practitioners.

Typicality An alternative explanation of the unintuitive OOD behaviour of likelihood-based models is presented in Nalisnick et al. [20], who note that a probability distribution's regions of high likelihood may not be associated with regions of high probability – especially as dimensionality increases. This led to the development of a test characterizing in- and out-of-distribution on the basis of *typicality*, i.e. how similar test (log-)likelihood values are to those from in-distribution training data. This is akin to controlling for the entropy of the in-distribution data. However, as per subsection 3.1, we could imagine a scenario where we observe OOD data with similar likelihoods to the in-distribution data, but which also has far different KL and entropy terms that throw off the comparison. Controlling for the entropy of the *incoming* data would help prevent such a scenario from occurring.

Another approach entirely is to consider test statistics other than the log likelihood, yet still comparing on the basis of typicality. Choi et al. [5] themselves provide an early discussion on typicality, and find that an ensemble-based estimate of the Watanabe-Akaike Information Criterion [31] performs well empirically. Sastry and Oore [26] decide whether or not incoming data is OOD based on how its pairwise feature correlations compare with those on in-distribution data. Morningstar et al. [18] propose an approach which takes any reasonable statistic(s) from a pre-existing DGM and builds a density estimator on this, rejecting points as OOD if they are not likely under this model. These methods are fully unsupervised in that they do not require access to OOD data, but often require additional model training or isolating seemingly random features of the data and thus make their general application less clear.

Model Fit Failure The final perspective on OOD detection that we will cover centres around model fit failure. Kirichenko et al. [11] focus on the inductive biases of normalizing flow methods, remarking that these learn generic image features and thus overpower the maximum likelhood objective as it relates to OOD detection; Schirrmeister et al. [27] find something similar in that the likelihoods on convolutional-based architectures are dominated by low-level features. These methods propose modified architectures or test statistics, respectively, demonstrating improved results. Zhang et al. [34] argue that model fit is the most likely culprit in OOD detection failures, and highlights several issues with the (typical set) idea that the support of OOD data could overlap with that of the training data and thus assumes that this does not occur. This is yet another complementary analysis to ours: in the view of (1) and (2), assuming we have a model P_θ which matches P perfectly, and P and Q have non-overlapping support, then $D_{\mathrm{KL}}\left(Q \,||\, P_\theta\right) = \infty$ and thus $\mathbb{E}_Q \log p_\theta(X) \to -\infty$. Since we do not observe this in practice, it suggests that model fit may indeed be part of the issue.

5 Discussion and Conclusion

In this work, we have presented an alternative explanation for the strange phenomenon wherein a deep generative model trained on one dataset assigns high likelihood to OOD data, particularly when this data is simpler than the training data. By exploiting a well-known decomposition of the average likelihood, we find that these observations can be explained by fluctuations in the entropy term between datasets. This also sheds further light on the success of OOD detection methods based on likelihood *ratios*, as these statistics are not affected by the entropy term (in expectation).

Although we find this observation interesting on its own, one limitation of this work is that we have not devised a new method to overcome issues posed by the entropy term, beyond the suggestion to use pre-existing likelihood ratio approaches. We at least hope that bringing this explanation to light might inspire other researchers to consider the impact of the entropy term when performing OOD detection with deep generative models. We would also like to specifically probe this phenomenon in future work with experiments and compare the relative impacts of the KL and entropy terms on the likelihood, but will need to get around the issues with estimating entropy in high dimensions [17].

This work may also serve as a motivation for further work in OOD detection with likelihood-based methods supported on low-dimensional manifolds embedded in high-dimensional ambient space, as the decomposition (1) does not apply in this case. Recently, Caterini et al. [4] have shown that injective flow methods (e.g. [3]) can accurately assign lower likelihoods to MNIST when trained by maximum likelihood over Fashion-MNIST; perhaps these methods are not as susceptible to fluctuations in the entropy term and can provide further improvements to likelihood-based OOD detection. Indeed, in accordance with the *manifold hypothesis* [1], it seems unlikely that high-dimensional data such as images would even admit densities with respect to the Lebesgue measure on the ambient space, and thus low-dimensional density models appear to be an important direction forward.

References

[1] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828, 2013.

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

[3] Johann Brehmer and Kyle Cranmer. Flows for simultaneous manifold learning and density estimation. *Advances in Neural Information Processing Systems*, 33, 2020.

[4] Anthony L Caterini, Gabriel Loaiza-Ganem, Geoff Pleiss, and John P Cunningham. Rectangular flows for manifold learning. *arXiv preprint arXiv:2106.01413*, 2021.

[5] Hyunsun Choi, Eric Jang, and Alexander A Alemi. Waic, but why? generative ensembles for robust anomaly detection. *arXiv preprint arXiv:1810.01392*, 2018.

[6] Yilun Du and Igor Mordatch. Implicit generation and generalization in energy-based models. *arXiv preprint arXiv:1903.08689*, 2019.

[7] Yuanjun Gao, Evan W Archer, Liam Paninski, and John P Cunningham. Linear dynamical neural population models through nonlinear embeddings. *Advances in neural information processing systems*, 29:163–171, 2016.

[8] John Just and Sambuddha Ghosal. Deep generative models strike back! improving understanding and evaluation in light of unmet expectations for ood data. *arXiv preprint arXiv:1911.04699*, 2019.

[9] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

[10] Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, pages 3581–3589, 2014.

[11] Polina Kirichenko, Pavel Izmailov, and Andrew G Wilson. Why normalizing flows fail to detect out-of-distribution data. *Advances in Neural Information Processing Systems*, 33, 2020.

[12] Ivan Kobyzev, Simon Prince, and Marcus Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[13] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009.

[14] Charline Le Lan and Laurent Dinh. Perfect density models cannot guarantee anomaly detection. In *"I Can't Believe It's Not Better!"NeurIPS 2020 workshop*, 2020.

[15] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[16] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

[17] Damiano Lombardi and Sanjay Pant. Nonparametric k-nearest-neighbor entropy estimator. *Physical Review E*, 93(1):013310, 2016.

[18] Warren Morningstar, Cusuh Ham, Andrew Gallagher, Balaji Lakshminarayanan, Alex Alemi, and Joshua Dillon. Density of states estimation for out of distribution detection. In *International Conference on Artificial Intelligence and Statistics*, pages 3232–3240. PMLR, 2021.

[19] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *International Conference on Learning Representations*, 2019.

[20] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, and Balaji Lakshminarayanan. Detecting out-of-distribution inputs to deep generative models using typicality. *arXiv preprint arXiv:1906.02994*, 2019.

[21] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

[22] Aaron Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George Driessche, Edward Lockhart, Luis Cobo, Florian Stimberg, et al. Parallel wavenet: Fast high-fidelity speech synthesis. In *International conference on machine learning*, pages 3918–3926. PMLR, 2018.

[23] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *arXiv preprint arXiv:1912.02762*, 2019.

[24] Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. *Advances in Neural Information Processing Systems*, 32:14707–14718, 2019.

[25] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and variational inference in deep latent gaussian models. In *International Conference on Machine Learning*, volume 2, page 2. Citeseer, 2014.

[26] Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with gram matrices. In *International Conference on Machine Learning*, pages 8491–8501. PMLR, 2020.

[27] Robin Schirrmeister, Yuxuan Zhou, Tonio Ball, and Dan Zhang. Understanding anomaly detection with deep invertible networks through hierarchies of distributions and features. *Advances in Neural Information Processing Systems*, 33, 2020.

[28] Joan Serrà, David Álvarez, Vicenç Gómez, Olga Slizovskaia, José F Núñez, and Jordi Luque. Input complexity and out-of-distribution detection with likelihood-based generative models. In *International Conference on Learning Representations*, 2020.

[29] James Townsend, Thomas Bird, and David Barber. Practical lossless compression with latent variables using bits back coding. In *International Conference on Learning Representations*, 2019.

[30] Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

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

[32] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[33] Zhisheng Xiao, Qing Yan, and Yali Amit. Likelihood regret: An out-of-distribution detection score for variational auto-encoder. *Advances in Neural Information Processing Systems*, 33, 2020.

[34] Lily Zhang, Mark Goldstein, and Rajesh Ranganath. Understanding failures in out-of-distribution detection with deep generative models. In *International Conference on Machine Learning*, pages 12427–12436. PMLR, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] In the final section
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A] No Experiments
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A] No Experiments
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A] No Experiments
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A] No Experiments
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]