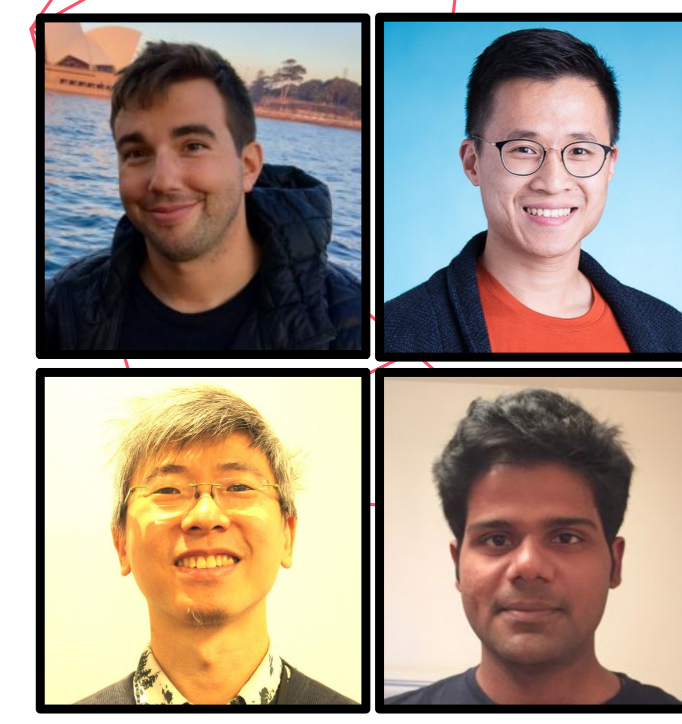
DeepMind

Detecting Out-of-Distribution Inputs to Deep Generative Models Using Typicality

Eric Nalisnick, Aki Matsukawa, Yee Whye Teh, Balaji Lakshminarayanan

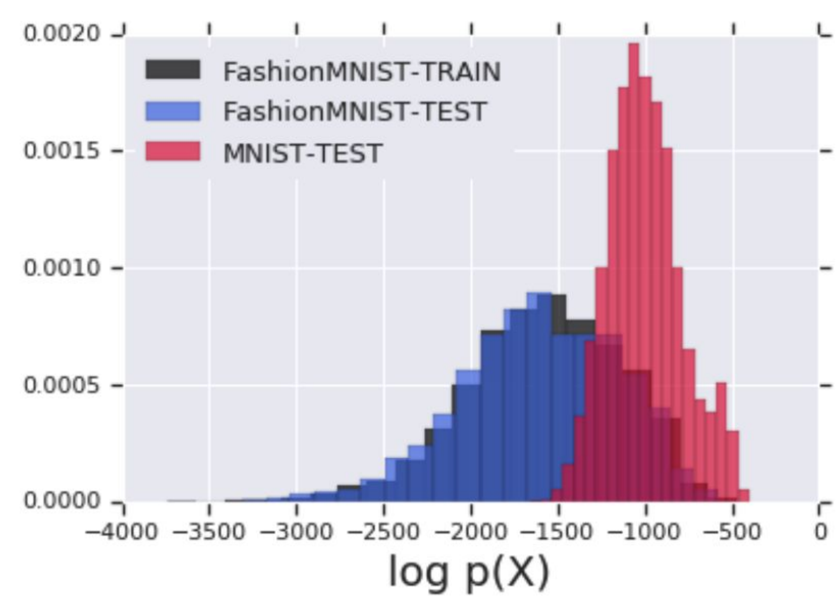


Motivation — Deep Generative Models and Out-of-Distribution Inputs

Nalisnick et al. [ICLR 2019] showed that the likelihood of deep generative models cannot distinguish the training data from out-of-distribution (OOD) inputs.

Data Set	Avg. Bits Per Dimension
Glow Trained on FashionMNIST	
FashionMNIST-Train	2.902
FashionMNIST-Test	2.958
MNIST-Test	**1.833**
Glow Trained on MNIST	
MNIST-Test	1.262

(Lower BPD is Better)

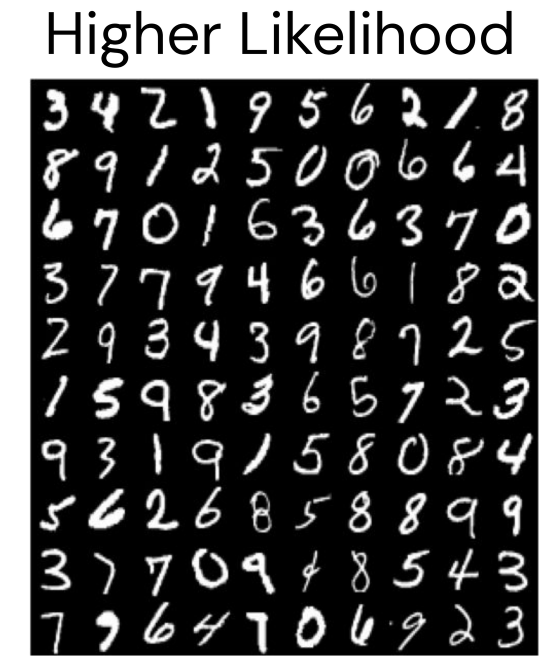


Histogram of Glow Log-Likelihoods

Yet when we sample from the generative model, the outputs conspicuously resemble the training data, not the OOD inputs.

FashionMNIST: Training Set

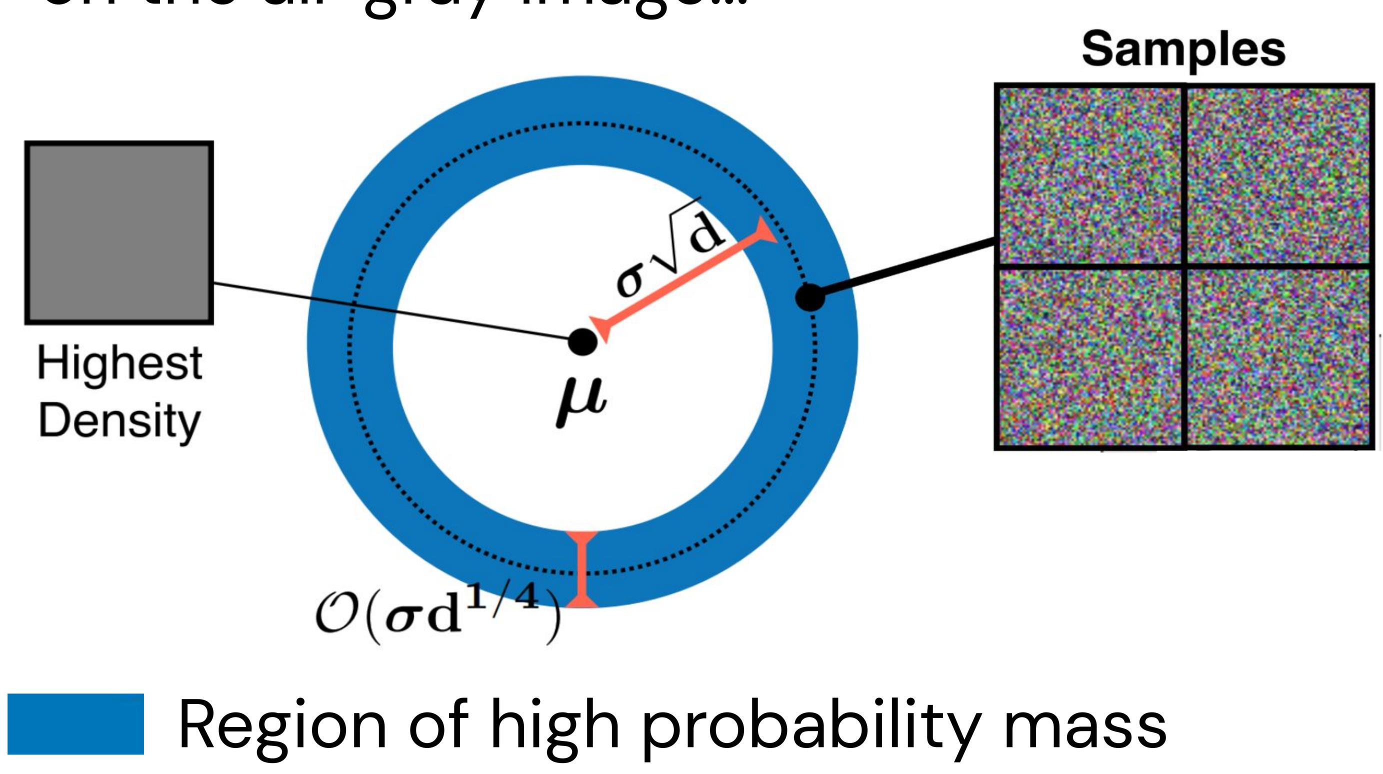
MNIST: Higher Likelihood

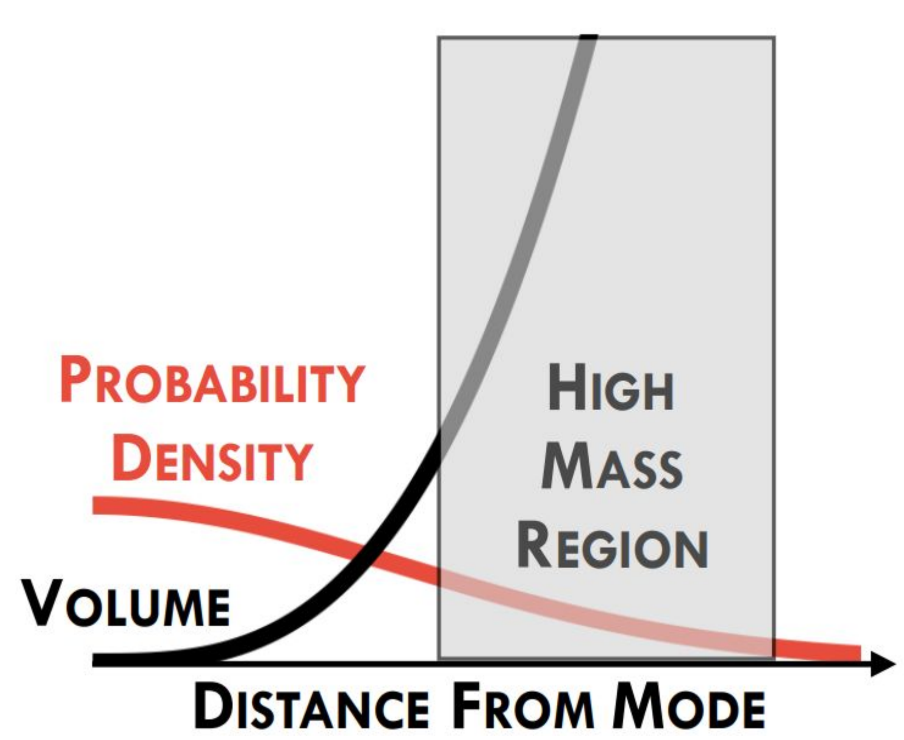


Samples from Generative Model

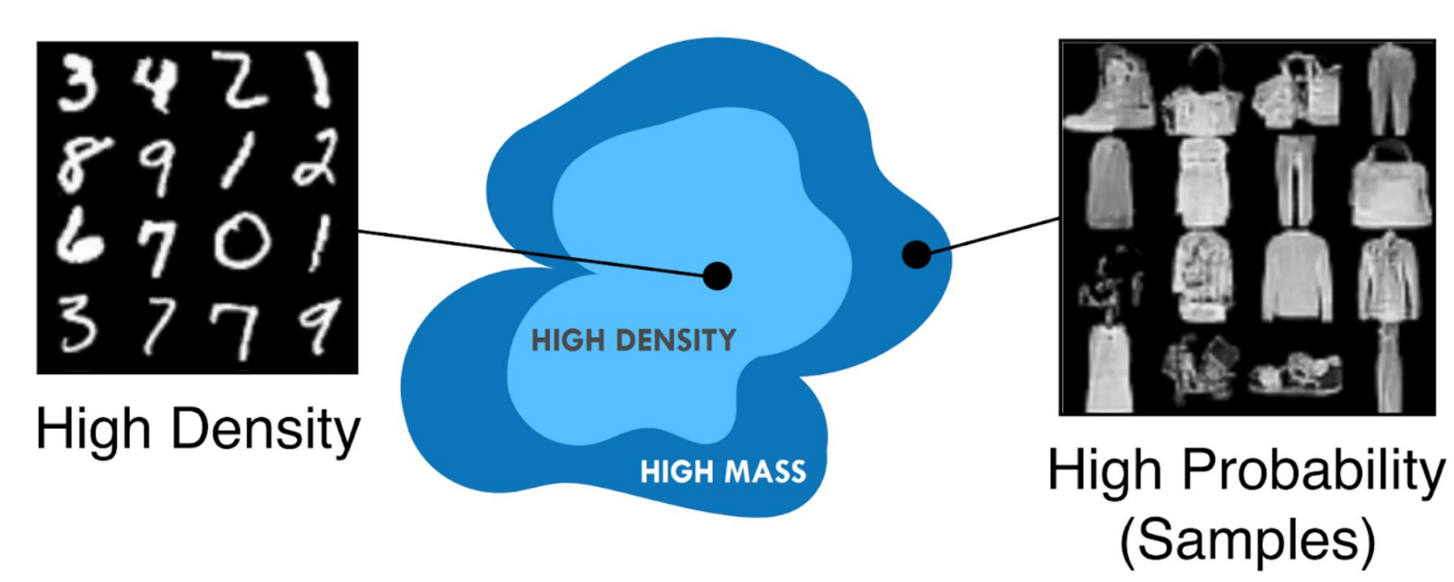
Background — Concentration and Typicality

Consider a high-dimensional Gaussian centered on the all-gray image...





We conjecture that a similar phenomenon is happening with high-dimensional deep generative models...



Methodology — A Goodness-of-Fit Test for OOD Detection

Due to deep generative models having an intractable cumulative distribution function (CDF), we propose detecting OOD inputs via a hypothesis based off of Shannon's [1948] definition of typical sets.

Definition 2.1. ϵ-Typical Set [11] *For a distribution $p(\mathbf{x})$ with support $\mathbf{x} \in \mathcal{X}$, the ϵ-typical set $\mathcal{A}_\epsilon^N[p(\mathbf{x})] \in \mathcal{X}^N$ is comprised of all N-length sequences that satisfy*

$$\mathbb{H}[p(\mathbf{x})] - \epsilon \leq \frac{-1}{N}\sum_{n=1}^{N}\log p(\boldsymbol{x}_n) \leq \mathbb{H}[p(\mathbf{x})] + \epsilon$$

where $\mathbb{H}[p(\mathbf{x})] = \int_{\mathcal{X}} p(\mathbf{x})[-\log p(\mathbf{x})]d\mathbf{x}$ and $\epsilon \in \mathbb{R}^+$ is a small constant.

For an *M*-sized test batch $\widetilde{\boldsymbol{X}} = \{\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_M\}$

`if` $\widetilde{\boldsymbol{X}} \in \mathcal{A}_\epsilon^M[p(\mathbf{x};\boldsymbol{\theta})]$ `then` $\widetilde{\boldsymbol{X}} \sim p(\mathbf{x};\boldsymbol{\theta})$, `otherwise` $\widetilde{\boldsymbol{X}} \not\sim p(\mathbf{x};\boldsymbol{\theta})$

Algorithm 1 A Bootstrap Test for Typicality

Input: Training data $\boldsymbol{X}$, validation data $\boldsymbol{X}'$, trained model $p(\mathbf{x};\boldsymbol{\theta})$, number of bootstrap samples K, significance level α, M-sized batch of possibly OOD inputs $\widetilde{\boldsymbol{X}}$.

Offline prior to deployment

1. Compute $\hat{\mathbb{H}}^N[p(\mathbf{x};\boldsymbol{\theta})] = \frac{-1}{N}\sum_{n=1}^{N}\log p(\boldsymbol{x}_n;\boldsymbol{\theta})$.

2. Sample K M-sized data sets from $\mathbf{X}'$ using bootstrap resampling.

3. For all $k \in [1, K]$:

Compute $\hat{\epsilon}_k = \left|\frac{-1}{M}\sum_{m=1}^{M}\log p(\boldsymbol{x}'_{k,m};\boldsymbol{\theta}) - \hat{\mathbb{H}}^N[p(\mathbf{x};\boldsymbol{\theta})]\right|$ (*Equation 7*)

4. Set $\epsilon_\alpha^M = \texttt{quantile}(F(\epsilon), \alpha)$ (*e.g.* $\alpha = .99$)

Online during deployment

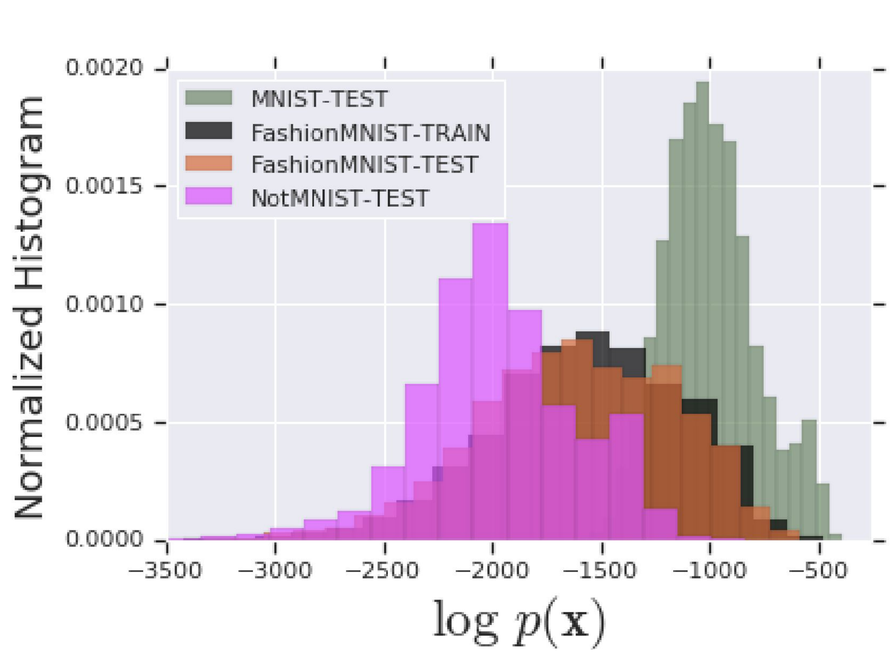
If $\left|\frac{-1}{M}\sum_{m=1}^{M}\log p(\tilde{\boldsymbol{x}}_m) - \hat{\mathbb{H}}^N[p(\mathbf{x};\boldsymbol{\theta})]\right| > \epsilon_\alpha^M$:

Return $\widetilde{\mathbf{X}}$ `is out-of-distribution`

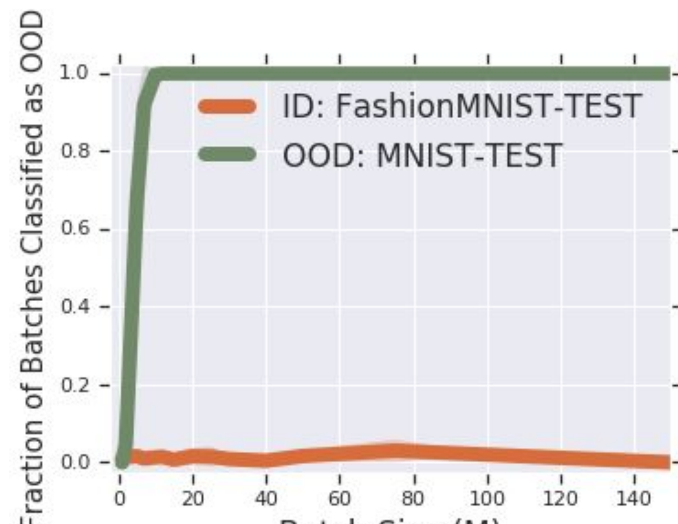
Else:

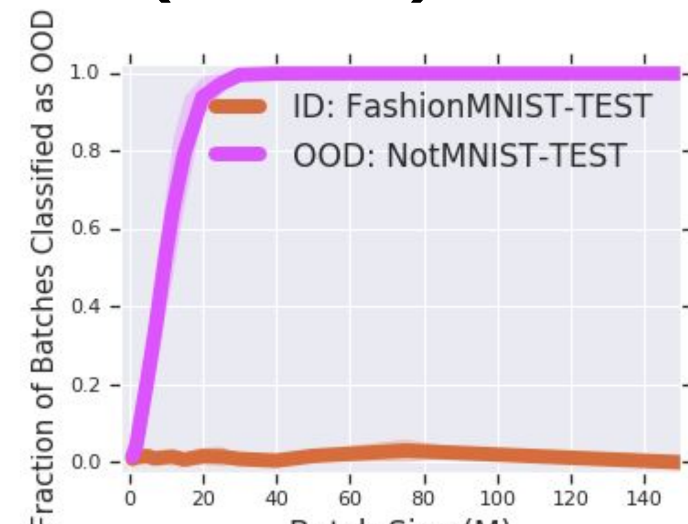
Return $\widetilde{\mathbf{X}}$ `is in-distribution`

Experiment



Fraction of M-Sized Batches Classified as OOD (α=99%)





- Our test for typicality **detects many of the OOD sets found problematic in Nalisnick et al.** [ICLR 2019].
- Relies on **the distribution of model likelihoods being well separated.**
- **ArXiv Link:** https://arxiv.org/abs/1906.02994