Penn
UNIVERSITY of PENNSYLVANIA

Typicality and OOD Detection

Eric Lei

Introduction

- Typicality is a tool from information theory that provides properties about the structure of the n-fold product distribution $P^{\otimes n}$
- Allows straightforward existence proofs in information theory, for channel coding and source coding
- Recently, typicality has received attention in the OOD detection literature
- Idea is to test whether a batch of samples is typical w.r.t $P^{\otimes n}$, rather than seeing if they have a high likelihood

Warm Up

Warm Up

- Let $X_i \sim \mathrm{Ber}(3/4)$ be i.i.d., and consider the following sequences of $n = 10$ realizations $(X_1, \ldots, X_n) = X^n$:
  - x^n= 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
  - y^n=0, 1, 0, 1, 1, 0, 1, 1, 1, 1
  - z^n=1, 1, 1, 1, 1, 1, 1, 1, 1, 1

Warm Up

- Let $X_i \sim \text{Ber}(3/4)$ be i.i.d., and consider the following sequences of $n = 10$ realizations $(X_1, \ldots, X_n) = X^n$:
  - x^n= 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
  - y^n=0, 1, 0, 1, 1, 0, 1, 1, 1, 1
  - z^n=1, 1, 1, 1, 1, 1, 1, 1, 1, 1
- Which sequence is more likely?
  - $\Pr(X^n = x^n) = (1/4)^{10}$
  - $\Pr(X^n = y^n) = (1/4)^3(3/4)^7$
  - $\Pr(X^n = z^n) = (3/4)^{10}$

Explanation

- The individual sequence y^n=0, 1, 0, 1, 1, 0, 1, 1, 1, 1 has smaller probability than z^n=1, 1, 1, 1, 1, 1, 1, 1, 1, 1
- There is only one sequence of all ones, but there are $\binom{10}{7}$ sequences with 3 zeros and 7 ones $\{X^n : X^n \text{ has 7 ones and 3 zeros}\}$
- This set has probability $\Pr(\{X^n : X^n \text{ has 7 ones and 3 zeros}\}) = \binom{10}{7}(3/4)^7(1/4)^3 \approx 0.25$
- Is this set that much larger than a single sequence?
  - No. It takes up $\dfrac{\binom{10}{7}}{2^{10}} \approx 11.7\,\%$ of the space
  - This effect becomes greater as n increases

Typical Sets

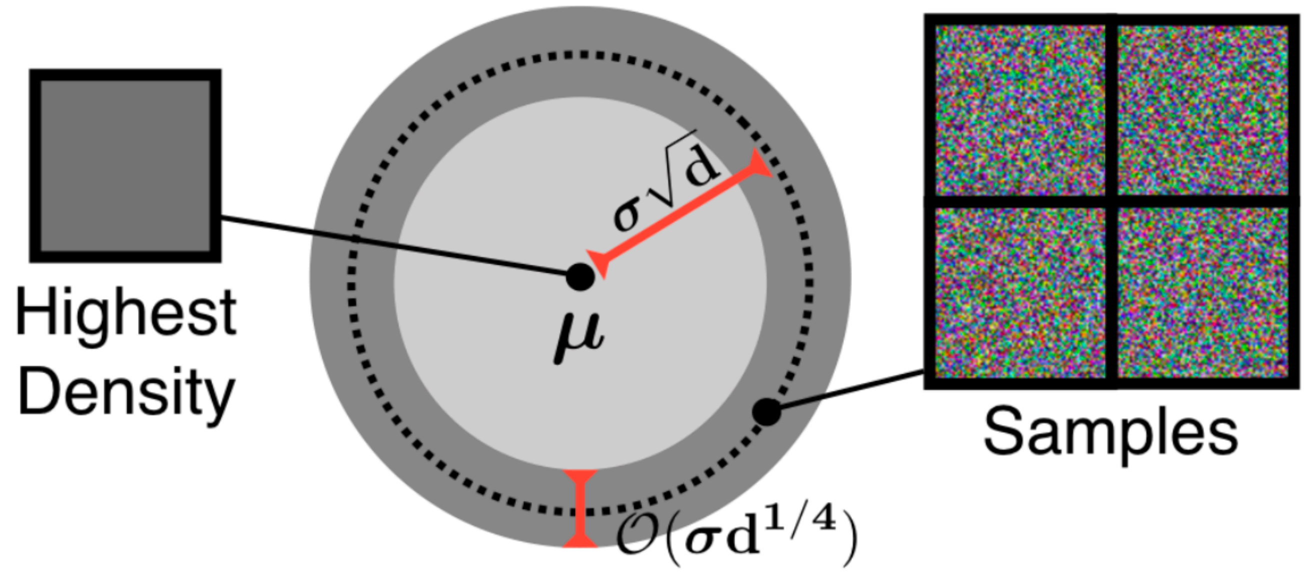
- Asymptotic Equipartiton Property (AEP): For $\{X_i\}_{i=1}^n \sim P \in \mathscr{P}(\mathscr{X})$ i.i.d.,
$-\frac{1}{n}\sum_{i=1}^{n}\log P(X_i) \longrightarrow H(X) := -\mathbb{E}_{X\sim P}[\log P(X)]$ in probability.

- The typical set of P considers sequences of $\mathscr{X}^n$ that approximately satisfy the AEP.

- ϵ-typical set of P: $A_\epsilon^{(n)}(P) := \left\{ x^n \in \mathscr{X}^n : \left| -\frac{1}{n}\log P^{\otimes n}(x^n) - H(P) \right| < \epsilon \right\}$

Properties of $A_\epsilon^{(n)}$

- $A_\epsilon^{(n)}(P) := \left\{ x^n \in \mathcal{X}^n : \left| -\frac{1}{n} \log P^{\otimes n}(x^n) - H(P) \right| < \epsilon \right\}$



1. $x^n \in A_\epsilon^{(n)} \implies 2^{-n(H(P)+\epsilon)} \le P^{\otimes n}(x_1, \ldots, x_n) \le 2^{-n(H(P)-\epsilon)}$

2. $P^{\otimes n}(A_\epsilon^{(n)}) \ge 1 - \epsilon$ for n sufficiently large.

3. $(1-\epsilon)2^{n(H(P)-\epsilon)} \le |A_\epsilon^{(n)}(P)| \le 2^{n(H(P)+\epsilon)}$ for n sufficiently large.

Interpretation: As n grows, the product distribution $P^{\otimes n}$ concentrates on $A_\epsilon^{(n)}$, and is approximately uniform (assigns probability $\approx 2^{-nH(P)}$ to sequences on a small set of size $\approx 2^{nH(P)}$).

Proofs

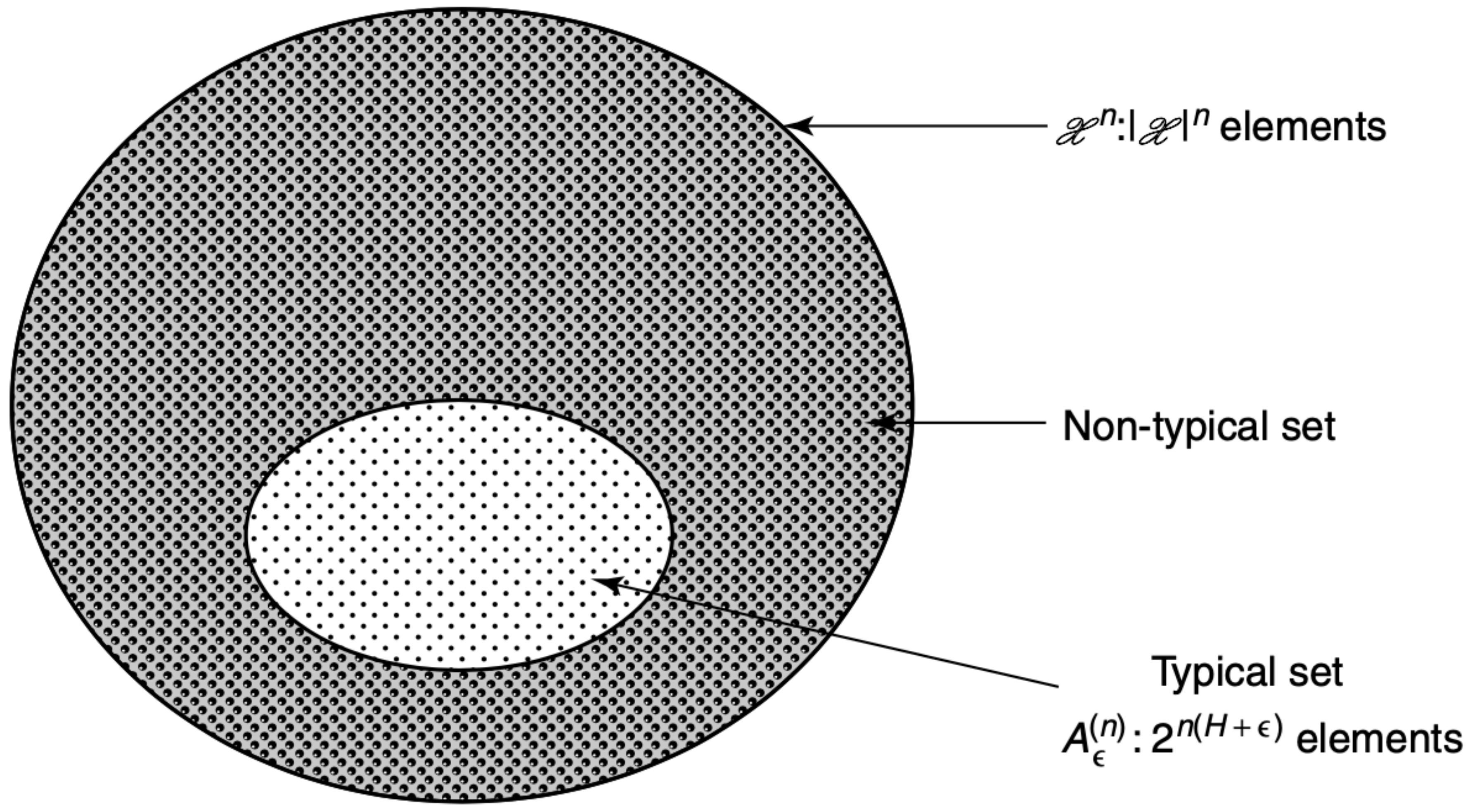
- First two properties are due to WLLN and definition of the typical set
- For the size:

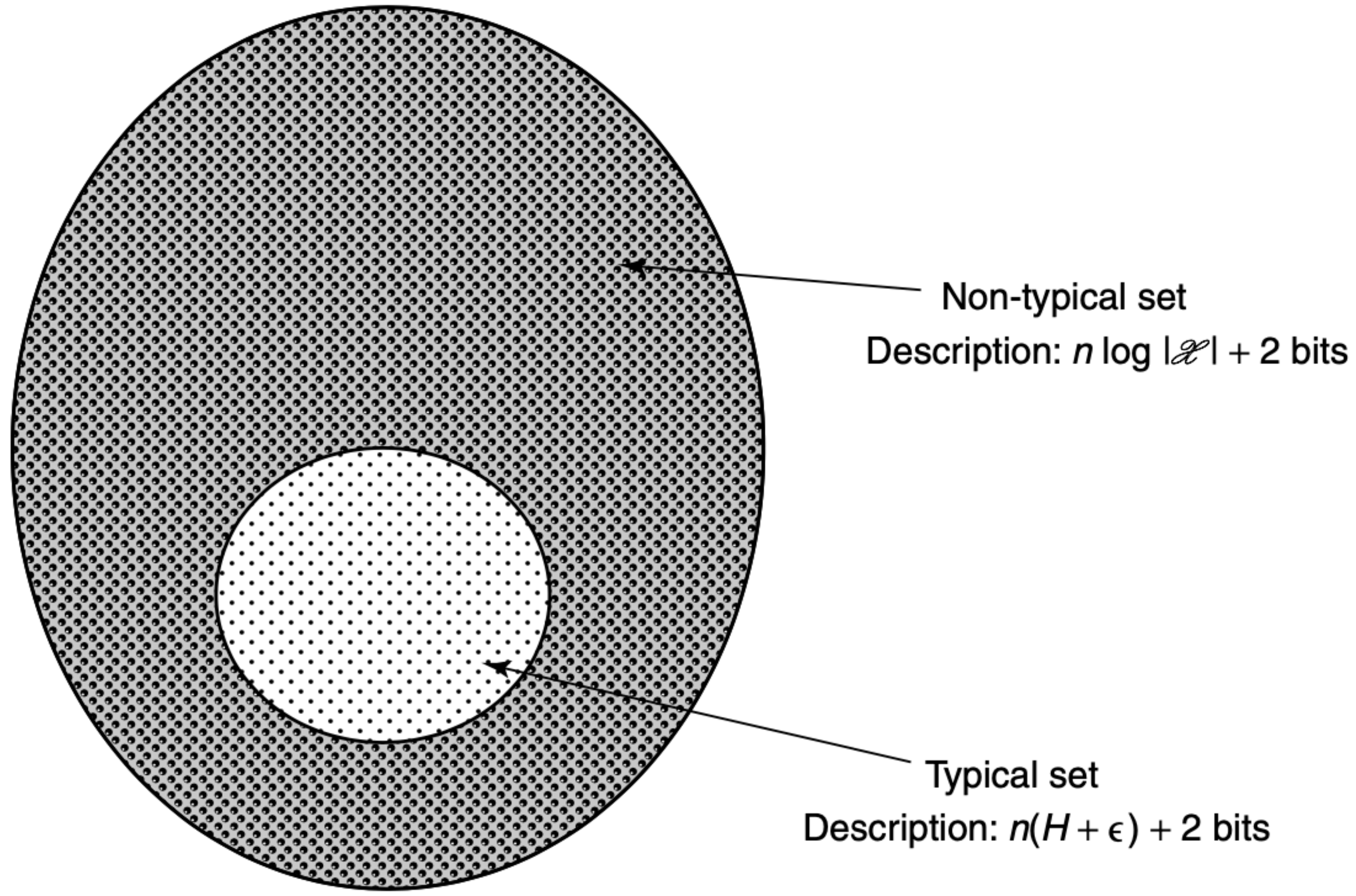
$$\begin{aligned} 1 &= \sum_{x^n \in \mathcal{X}^n} P^{\otimes n}(x^n) \\ &\geq \sum_{x^n \in A_\epsilon^{(n)}} P^{\otimes n}(x^n) \\ &\geq \sum_{x^n \in A_\epsilon^{(n)}} 2^{-n(H(P)+\epsilon)} \\ &= 2^{-n(H(P)+\epsilon)} |A_\epsilon^{(n)}| \end{aligned}$$

$$\begin{aligned} 1 - \epsilon &< P^{\otimes n}(A_\epsilon^{(n)}) \\ &\leq \sum_{x^n} 2^{-n(H(P)-\epsilon)} \\ &= 2^{-n(H(P)-\epsilon)} |A_\epsilon^{(n)}| \end{aligned}$$

Usage in Information Theory

- Q: Minimum number of bits to represent $X_1, \ldots, X_n \sim P$, iid?
- Technique: break up $\mathscr{X}^n$ into typical and non-typical sequences, and order them
- If X^n is typical, its index requires no more than $n(H(P) + \epsilon) + 1$ bits. Prepend with 0.
- Otherwise, index requires no more than $n \log |\mathscr{X}| + 1$ bits. Prepend with 1.
- $\mathbb{E}[\ell(X^n)] \leq \Pr(A_\epsilon^{(n)})(n(H(P) + \epsilon) + 2) + \Pr(A_\epsilon^{(n)^\complement})(n \log |\mathscr{X}| + 2)$
- $\mathbb{E}[\ell(X^n)] \leq n(H(P) + \epsilon) + \epsilon n \log |\mathscr{X}| + 2 = n(H(P) + \epsilon + \epsilon \log |\mathscr{X}| + \frac{2}{n})$
- $\forall \epsilon' > 0, \quad \mathbb{E}[\frac{1}{n}\ell(X^n)] \leq H(P) + \epsilon'$ for n sufficiently large.





Non-product distributions

- Non-product distributions on $\mathcal{X}^n$ may not satisfy AEP:
  - Let $\mathcal{X}^n = \{0,1,2\}^n$. Define $Q_n(x^n) = \begin{cases} \frac{1}{2}2^{-n} & \text{if } x^n \text{ contains only 0s or 1s} \\ \frac{1}{2} \cdot \frac{1}{3^n - 2^n} & \text{o.w.} \end{cases}$
  - When n is large, $Q_n(x^n) \approx \begin{cases} \frac{1}{2}2^{-n} & \text{if } x^n \text{ contains only 0s or 1s} \\ \frac{1}{2} \cdot \frac{1}{3^n} & \text{o.w.} \end{cases}$
  - But $Q_n(\{x^n : Q_n(x^n) = 2^{-n+o(n)}\}) = Q_n(\{x^n : Q_n(x^n) = 3^{-n+o(n)}\}) = 1/2$
- In general, ergodic distributions on $\mathcal{X}^n$ satisfy the AEP

Other High Probability Sets on $\mathcal{X}^n$

- Consider $C_\epsilon(P^{\otimes n}) := \min\{\,|B| : B \subseteq \mathcal{X}^n, P^{\otimes n}(B) > 1-\epsilon\}$, the size of the smallest $1-\epsilon$ probability set under $P^{\otimes n}$.

- Fact: $C_\epsilon(P^{\otimes n}) = 2^{nH(P)+o(n)}$ for n sufficiently large.

- Remark: The smallest high probability set has the same size as the typical set, up to first order in the exponent.

- Proof: (UB) Choose sequence ϵ_n such that $P^{\otimes n}(A_{\epsilon_n}^{(n)}) \to 1$ as $n \to \infty$. For n large enough, $C_\epsilon(P^{\otimes n}) \leq |A_{\epsilon_n}^{(n)}(P)|$. But
$1 \geq P^{\otimes n}(A_{\epsilon_n}^{(n)}(P)) = \sum_{x^n \in A_{\epsilon_n}^{(n)}(P)} P^{\otimes n}(x^n) \geq |A_{\epsilon_n}^{(n)}(P)|\, 2^{-nH(P)-\epsilon_n n}$.

(LB): Similar argument.

DGMs

- Consider flow-based generative models
- Can sample from the data distribution *and* estimate densities
- Experiment: when trained on one distribution, can they detect OOD samples?
- Surprisingly, the likelihoods of OOD samples are higher than on in-distribution samples
    - But: the DGM never generates OOD samples

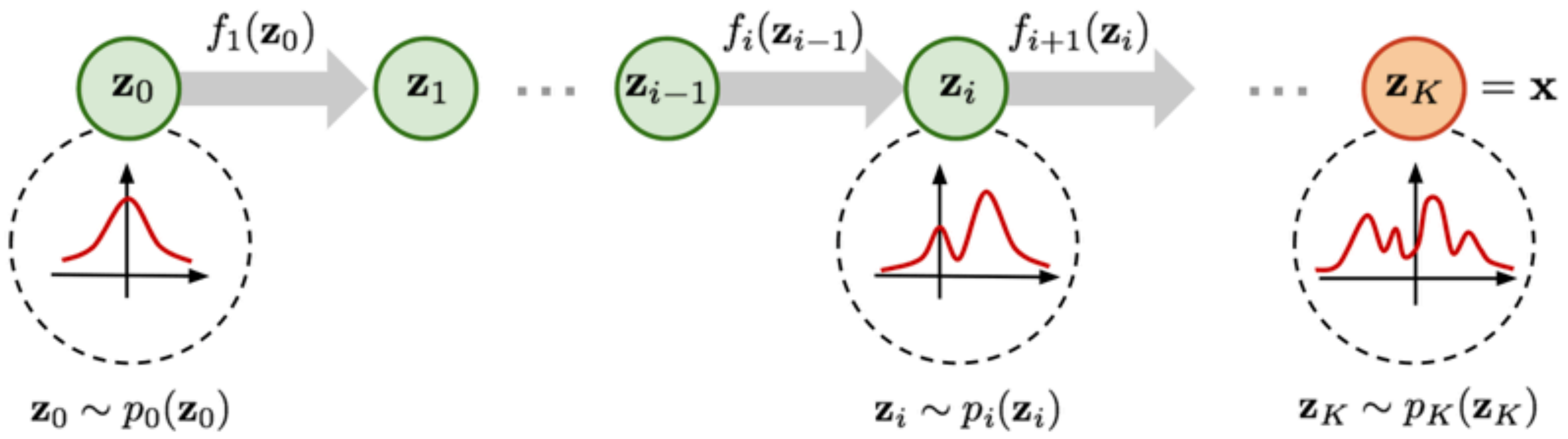
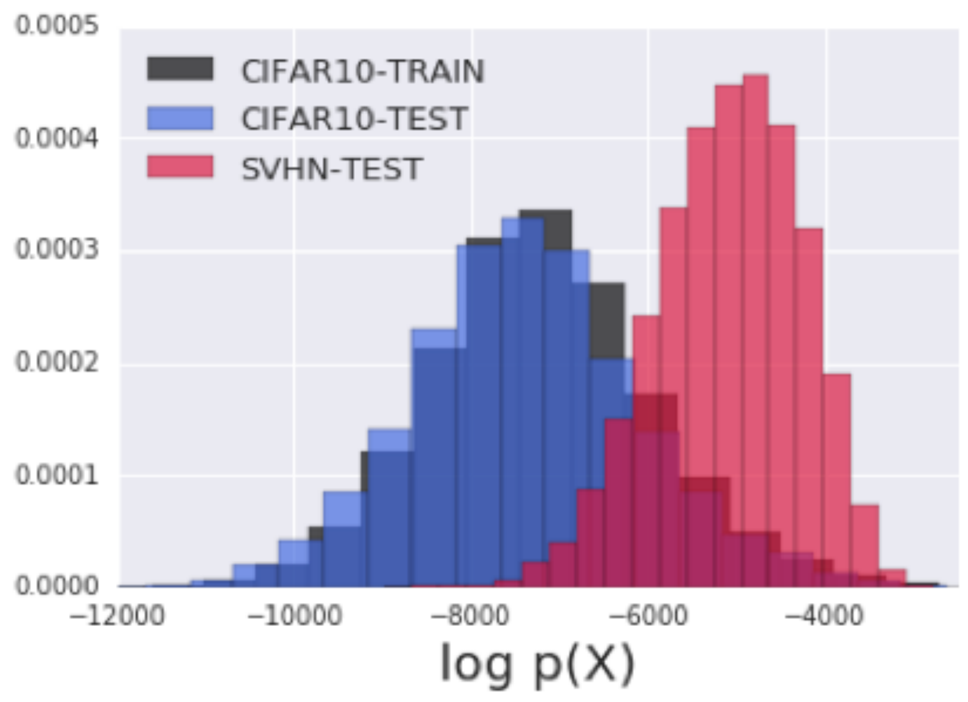


Fig. 2. Illustration of a normalizing flow model, transforming a simple distribution $p\_0(\mathbf{z}\_0)$ to a complex one $p\_K(\mathbf{z}\_K)$ step by step.

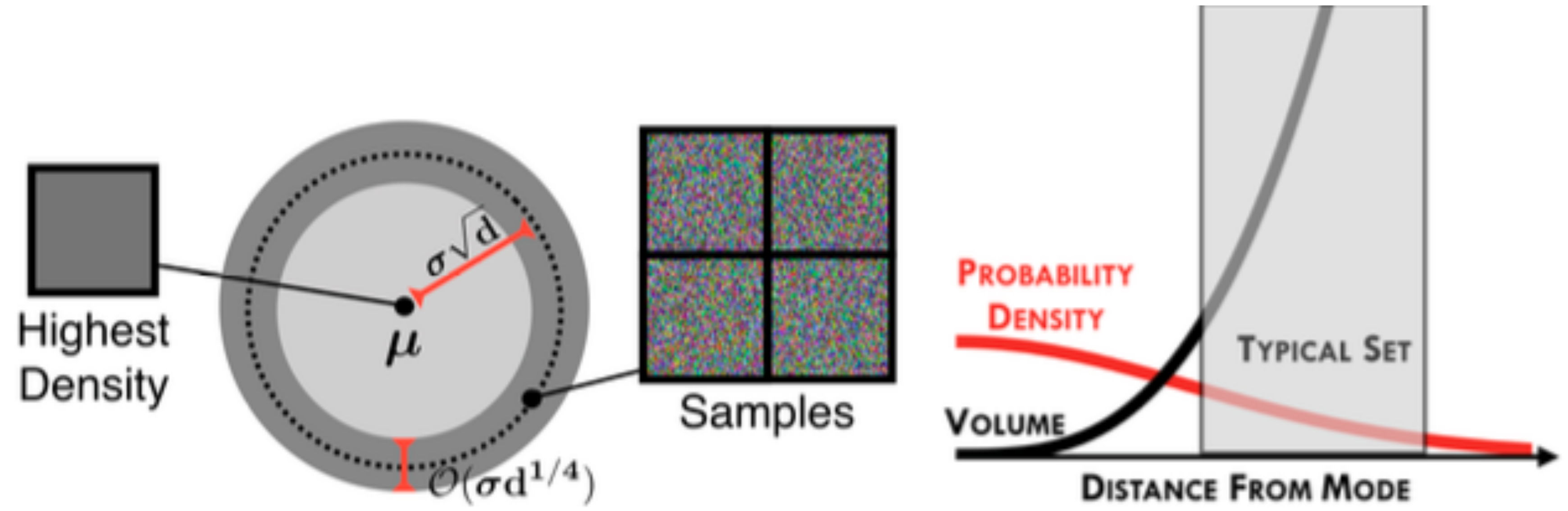
Data Set	Avg. Bits Per Dimension
Glow Trained on CIFAR-10	
CIFAR10-Train	3.386
CIFAR10-Test	3.464
SVHN-Test	**2.389**
Glow Trained on SVHN	
SVHN-Test	2.057



Do Deep Generative Models Know What They Don't Know? *Nalisnick et. al.*

The Typical Set Hypothesis

- Conjecture: The failure of DGMs to detect OOD samples via likelihood estimates is due to typicality.
  - Recall: $A_\epsilon^{(n)}$ does not necessarily intersect with regions of the highest likelihood
  - Ex: Under $\mathcal{N}(0,\sigma^2)$, typical set concentrates on the sphere of radius $\sigma\sqrt{n}$.
  - But doesn't include regions around the mean.
- Idea: instead of comparing likelihoods, test OOD samples for typicality



Nalisnick et. al.., Detecting Out-of-Distribution Inputs to Deep Generative Models Using Typicality.

Testing for Typicality

- We have a trained DGM P_θ, trained from data $\{X_s\}_{s=1}^{S}$
- Want to determine if $\tilde{X}^M = \{\tilde{X}_1, \ldots, \tilde{X}_M\}$ is OOD from P_θ
- Hypothesis test: $H_0 : \tilde{X}^M \in A_\epsilon^{(M)}(P_\theta) \quad H_1 : \tilde{X}^M \notin A_\epsilon^{(M)}(P_\theta)$
- to determine if $\tilde{X}^M$ came from P i.i.d., check if $\left| \frac{1}{M} \sum_{m=1}^{M} -\log P_\theta(\tilde{X}_i) - H(P_\theta) \right| < \epsilon$, where $H(P_\theta)$ is estimated from samples, i.e. $H(P_\theta) \approx -\frac{1}{S} \sum_{s=1}^{S} \log P_\theta(X_s)$

Bootstrap Test

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 6)*

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`

Results

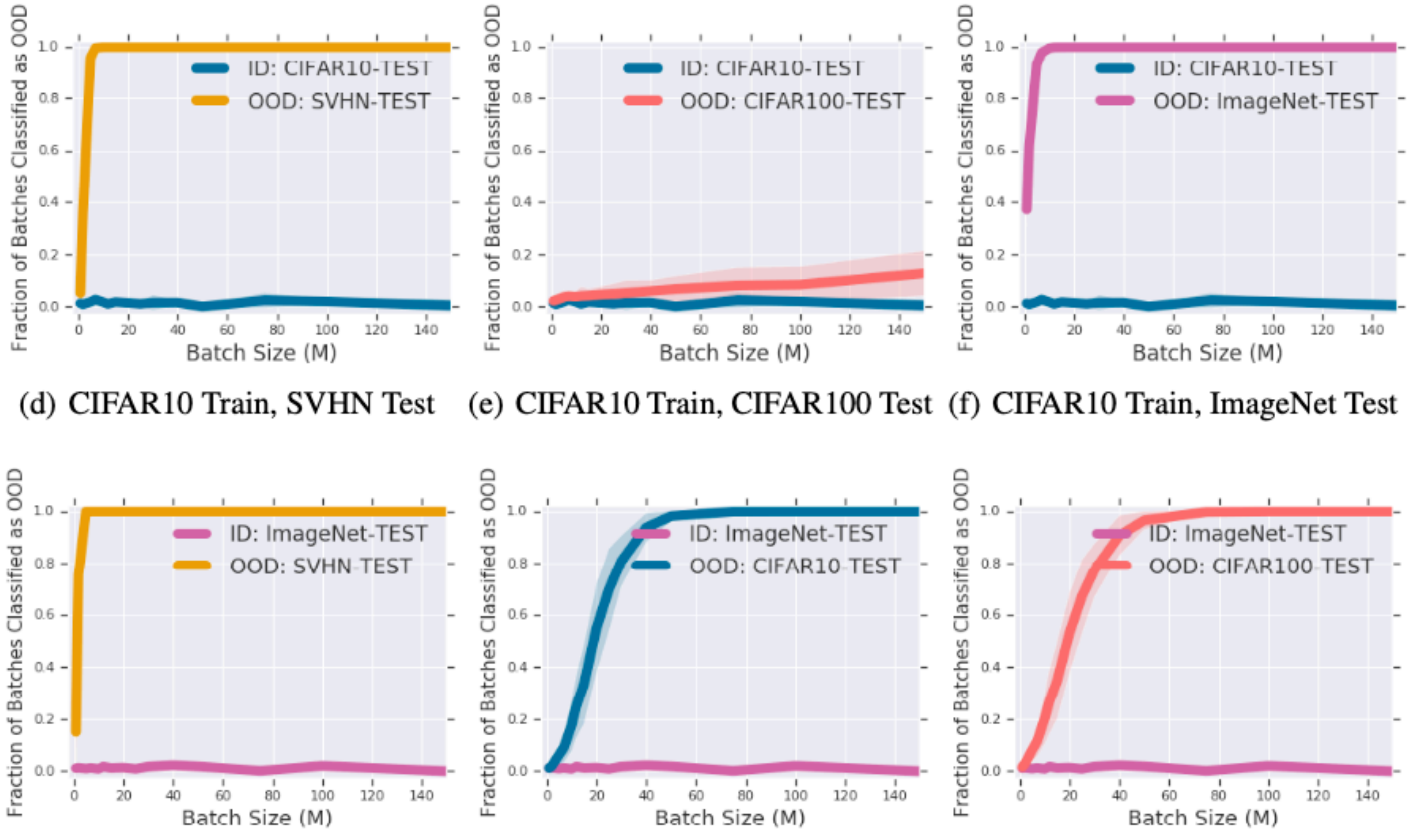
- SVHN, CIFAR-10, ImageNet
- Train DGM on one dataset, evaluate the other two as OOD
- ImageNet is toughest

Table 2: *Natural Images: Fraction of M-Sized Batches Classified as OOD.*

	M = 2			M = 10			M = 25		
METHOD	SVHN	CIFAR-10	IMAGENET	SVHN	CIFAR-10	IMAGENET	SVHN	CIFAR-10	IMAGENET
				*Glow Trained on **SVHN***					
Typicality Test	0.01±.00	**0.98**±.00	**1.00**±.00	**0.00**±.00	**1.00**±.00	**1.00**±.00	0.02±.00	**1.00**±.00	**1.00**±.00
t-Test	**0.00**±.00	0.95±.00	**1.00**±.00	0.04±.00	**1.00**±.00	**1.00**±.00	0.03±.00	**1.00**±.00	**1.00**±.00
KS-Test	**0.00**±.00	0.00±.00	0.00±.00	0.08±.00	**1.00**±.00	**1.00**±.00	0.03±.00	**1.00**±.00	**1.00**±.00
Annulus Method	0.02±.01	0.70±.05	**1.00**±.00	0.02±.01	**1.00**±.00	**1.00**±.00	**0.00**±.00	**1.00**±.00	**1.00**±.00
				*Glow Trained on **CIFAR-10***					
Typicality Test	0.42±.09	0.01±.01	0.64±.04	**1.00**±.00	**0.01**±.01	**1.00**±.00	**1.00**±.00	**0.01**±.01	**1.00**±.00
t-Test	**0.44**±.01	0.01±.00	0.65±.00	**1.00**±.00	0.02±.00	**1.00**±.00	**1.00**±.00	0.02±.00	**1.00**±.00
KS-Test	0.00±.00	**0.00**±.00	0.00±.00	**1.00**±.00	**0.01**±.00	0.98±.00	**1.00**±.00	**0.01**±.00	**1.00**±.00
Annulus Method	0.09±.03	0.02±.00	**0.87**±.05	0.19±.01	0.03±.00	**1.00**±.00	0.35±.02	0.04±.00	**1.00**±.00
				*Glow Trained on **ImageNet***					
Typicality Test	**0.78**±.08	0.02±.01	0.01±.00	**1.00**±.00	0.20±.06	**0.01**±.01	**1.00**±.00	0.74±.05	**0.01**±.01
t-Test	0.76±.00	0.02±.00	0.01±.00	**1.00**±.00	0.18±.01	**0.01**±.00	**1.00**±.00	0.72±.01	**0.01**±.00
KS-Test	0.00±.00	0.00±.00	**0.00**±.00	**1.00**±.00	**0.29**±.01	**0.01**±.00	**1.00**±.00	**0.89**±.01	0.02±.00
Annulus Method	0.00±.00	**0.03**±.00	0.02±.01	0.02±.02	0.15±.04	0.02±.00	0.16±.04	0.57±.12	0.02±.00

Varying M

- Recall that typicality properties hold "for M sufficiently large"



(d) CIFAR10 Train, SVHN Test (e) CIFAR10 Train, CIFAR100 Test (f) CIFAR10 Train, ImageNet Test

(g) ImageNet Train, SVHN Test (h) ImageNet Train, CIFAR10 Test (i) ImageNet Train, CIFAR100 Test

Figure 4: *Natural Image OOD Detection for Glow*. The above plots show the fraction of M-sized batches rejected for three Glow models trained on SVHN, CIFAR-10, and ImageNet. The OOD distribution data sets are these three training sets as well as CIFAR-100.

Detractors

- Zhang, Lily, Mark Goldstein, and Rajesh Ranganath. "Understanding failures in out-of-distribution detection with deep generative models." ICML 2021.
    - Previous work is testing the model's typical set, not the data distribution's typical set —> model estimation errors
    - Mismatch typicality: What probability does $P^{\otimes n}$ assign to another distribution's typical set?
    - $P^{\otimes n}(A_\epsilon^{(n)}(Q)) \approx 2^{-nD_{KL}(P||Q)}$

Detractors (cont)

- Ambiguity of high probability sets
  - Recall: $C_\epsilon(P^{\otimes n}) := \min\{\,|B|\,:\, B \subseteq \mathcal{X}^n, P^{\otimes n}(B) > 1-\epsilon\} = 2^{nH(P)+o(n)}$
  - The B that achieved the min should include the highest likelihood sequence
  - There are other small sets containing most of the probability
  - No reason to prefer the typical set
  - (The "equipartition" property is also not really used)
- Instead: high likelihood samples that are never generated are due to model misestimation, not typicality.
  - Good DGMs are not sufficient for good OOD detection

Conclusion

- Typical sets absorb most of the probability, but are small relative to the size of the entire space
- Typical sets don't always align with regions of high likelihood
- DGMs fail to detect OOD using likelihood, but somewhat work with typicality
- But there are still potential flaws: model misestimation and arbitrariness of typical sets

Alternative Idea Using Compressors

- E. Sabeti, A. Host-Madsen, "Data Discovery and Anomaly Detection Using Atypicality: Theory".

- Suppose you have access to a text compressor, e.g. ZIP

- You have a "training" string $x^n \sim P^{\otimes n}$. Compress it using ZIP, achieving an length $\bar{\ell} = \frac{1}{n}\ell(x^n)$.

- To see if a new string y^m is OOD, append it to x^n and compress using ZIP. If $y^m \sim P^{\otimes m}$, then $\frac{1}{n+m}\ell(x^n y^m) \approx \bar{\ell}$.

Other Limitations

- Recall: $x^n \in A_\epsilon^{(n)} \implies 2^{-n(H(P)+\epsilon)} \leq P^{\otimes n}(x_1, \ldots, x_n) \leq 2^{-n(H(P)-\epsilon)}$
- $P^{\otimes n}$ is approximately uniform
- Says that x^n, y^n cannot have probability under $P^{\otimes n}$ that differ exponentially in *n,* but nothing more.
- Fact: The r.v. $P^{\otimes n}(X^n)$ exhibits fluctuations exponential in $\sqrt{n}$.