EVALUATION OF CATEGORICAL GENERATIVE MODELS - BRIDGING THE GAP BETWEEN REAL AND SYNTHETIC DATA

Florence Regol, Anja Kroon, Mark Coates

Dept. of Electrical and Computer Engineering, McGill University, Montréal, Québec, Canada

ABSTRACT

The machine learning community has mainly relied on real data to benchmark algorithms as it provides compelling evidence of model applicability. Evaluation on synthetic datasets can be a powerful tool to provide a better understanding of a model's strengths, weaknesses and overall capabilities. Gaining these insights can be particularly important for generative modeling as the target quantity is completely unknown. Multiple issues related to the evaluation of generative models have been reported in the literature. We argue those problems can be avoided by an evaluation based on ground truth. General criticisms of synthetic experiments are that they are too simplified and not representative of practical scenarios. As such, our experimental setting is tailored to a realistic generative task. We focus on categorical data and introduce an appropriately scalable evaluation method. Our method involves tasking a generative model to learn a distribution in a high-dimensional setting. We then successively bin the large space to obtain smaller probability spaces where meaningful statistical tests can be applied. We consider increasingly large probability spaces, which correspond to increasingly difficult modeling tasks, and compare the generative models based on the highest task difficulty they can reach before being detected as being too far from the ground truth. We validate our evaluation procedure with synthetic experiments on both synthetic generative models and current state-of-the-art categorical generative models.

Index Terms— Evaluation of generative models, categorical generative models

1. INTRODUCTION

In the machine learning community, evaluation of generative models is an ongoing topic of research [1, 2, 3, 4, 5, 6, 7, 8]. The most common method is to evaluate the (usually upper bounded) log-likelihood on held out test data; the model with the highest log-likelihood is declared the better model. Although principled, this method of evaluation has some known drawbacks as highlighted in [6]. A simple example from [9] shows how good likelihood models can generate poor samples. Simply put, the learned model is far from the true distribution.

Another observation, possibly symptomatic of this issue, is the out-of-distribution (OOD) problem. A notable finding by [10] is that high likelihood and good sample generation do not guarantee good OOD detection. The task of OOD detection is an important application of generative modeling. If a
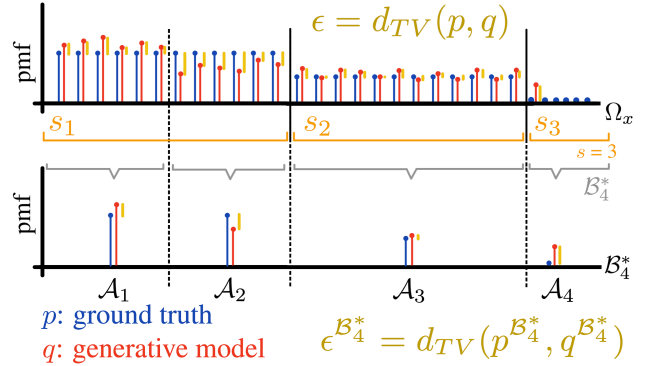

Fig. 1: Overview of our proposed binning procedure. The initial distribution p has $s = 3$ flat regions on the Ω space contained in $\mathbf{S} = \{\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3\}$. For a granularity level $k = 4$, our binning procedure finds that cutting $\mathcal{S}_1 = \{\mathcal{A}_1, \mathcal{A}_2\}$ to form $\mathcal{B}_4^* = \{\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, \mathcal{A}_4\}$ maximises $d_{TV}(p^{\mathcal{B}_4}, q^{\mathcal{B}_4})$. Hypothesis testing can then be conducted over the induced distributions $p^{\mathcal{B}_4^*}, q^{\mathcal{B}_4^*}$ on the 4 event probability space $\mathcal{B}_4^*$.

model is an appropriate approximation of the true distribution, then it should be able to detect low-probability samples.

One approach taken in the literature to mitigate this problem is to adopt a more comprehensive and task-oriented methods of evaluation. By assessing over multiple metrics and designing task-oriented metrics, one can better assess the true capabilities of a generative model [6, 3, 11]. Nonetheless, this remains a heuristic approach that is unavoidably tied to the application at hand and therefore cannot be applied to the general problem of generative model evaluation.

In this work, we propose to return to a synthetic setting where the evaluation of a generative model can be framed as a statistical identity testing problem. This allows us to draw from statistical testing literature. This approach alleviates the problems associated with log-likelihood evaluation. Rather than testing whether high log-likelihood is assigned to unseen samples, we can directly evaluate whether the model assigns the correct probability mass to each element in the space. We can then directly answer questions about OOD capability, sample generation quality, and overall generative modeling performance. This approach also provides better insight into which types of distributions a particular generative model can fit.

The synthetic task must be representative of the actual task a generative model would be called to solve. In almost all cases, generative models in machine learning target a (very) high dimensional setting. One challenge induced by such set-

tings is that statistical tests based on samples become meaningless. We simply cannot collect enough samples for a test to reject a null hypothesis. In this work, to address this issue, we propose an algorithm to bin the probability space at various granularity levels. We then consider the induced distributions on the smaller probability spaces as a testing proxy for our generative model quality, observing that the error for the induced distributions is a lower bound on the error in the original space. Figure 1 provides an overview of our procedure. Random binning is not successful and we explain how we construct bins to maximize the lower bound on the error.

We validate our binning procedure through synthetic experiments, ensuring it can preserve the correct ranking of synthetic generative models. We then show that our testing procedure correctly ranks state-of-the-art generative models trained on our synthetic task by comparing with the empirical total variation on the initial space Ω and with visual inspection of the learned distributions.

2. METHODOLOGY

We aim to design an evaluation method or metric that can rank generative models reliably for a *very large scale* discrete distribution space $|\Omega|$, *given the ground truth distribution p*. Given samples from unknown distributions q_1, q_2, we wish to distinguish which distribution is closer to a reference distribution p (thus achieving a ranking of generative models).

2.1. Distribution Testing

This setting is analogous to a fundamental problem in statistical distribution testing: identity testing. In this setting, given sample access to an unknown distribution q, an identity test declares if q is ϵ-close to a reference distribution p with probability at least $1 - \delta$. See [12, 13] for a more formal definition and [14] for a review on this topic. In this field, closeness of discrete distributions p, q on a discrete sample space Ω is formalized as the total variation distance:

$$\epsilon = d_{TV}(p, q) \triangleq \frac{1}{2}||p - q||_1 = \frac{1}{2}\sum_{x \in \Omega}|p_x - q_x| \quad (1)$$

(with p_x used as a shorthand for $p(x)$). The ℓ^2 or Hellinger distance $d^2(p, q) \triangleq ||p - q||_2 = \sum_{x \in \Omega}(p_x - q_x)^2$ can also be used. For relatively small sized Ω, results from the identity testing literature can be applied directly. However, we target large Ω where the practical number of samples is $m \ll |\Omega|$. At this scale, algorithms for identity testing are unusable; the provably most powerful test has a $\sqrt{|\Omega|}$ term in its required number of samples [13].

2.2. Binning the space

We propose to move from the probability space Ω to a smaller space $\mathcal{B}$ by binning the elements x in partitioning sets of Ω: $\mathcal{B} = \{\mathcal{A}_1, \mathcal{A}_2, \dots\}$ with $\cup_i \mathcal{A}_i = \Omega$ and $\cap \mathcal{A}_i = \emptyset$. We then assess the distributions induced by this binning to evaluate the models.

We denote all partitions of Ω by $\rho(\Omega)$. The 'binned' distribution of the ground truth p is $p^{\mathcal{B}}$, and the 'binned' distribution

of the generative model q becomes $q^{\mathcal{B}}$, where $p_{\mathcal{A}}^{\mathcal{B}} = \sum_{x \in \mathcal{A}} p_x$. This binning operation naturally introduces a new total variation error on the $\mathcal{B}$ space :

$$\epsilon^{\mathcal{B}} = d_{TV}(p^{\mathcal{B}}, q^{\mathcal{B}}) = \frac{1}{2}\sum_{\mathcal{A} \in \mathcal{B}}|p_{\mathcal{A}}^{\mathcal{B}} - q_{\mathcal{A}}^{\mathcal{B}}| = \frac{1}{2}\sum_{\mathcal{A} \in \mathcal{B}}|\sum_{x \in \mathcal{A}} p_x - q_x|. \quad (2)$$

Binning also introduces an error between the initial distribution p and its binned version $p^{\mathcal{B}}$ that can be defined as:

$$\epsilon^{p \to p^{\mathcal{B}}} = \frac{1}{2}\sum_{\mathcal{A} \in \mathcal{B}}\sum_{x \in \mathcal{A}}|p_x - \frac{p_{\mathcal{A}}^{\mathcal{B}}}{|\mathcal{A}|}|. \quad (3)$$

At the cost of introducing error, binning allows us to reduce the space to a workable size, so that we can use tools from the distribution testing literature. In particular, we can design a hypothesis test that aims to reject $\mathcal{H}_0 : d(p^{\mathcal{B}}, q^{\mathcal{B}}) < \epsilon$ at a probability significance level δ.

We can see that $d_{TV}(p^{\mathcal{B}}, q^{\mathcal{B}})$ is a lower bound for $d_{TV}(p, q)$, but the result of such test in the binned space *does not* provide an direct indication of the outcome in the original space. The intuition is that, if the test fails for some binning $\mathcal{B}$, this suggests that a generative model is poor, in the sense that it cannot achieve the easier task of representing the binned distribution to within a required error. In the following section, we describe how we select the binning.

2.2.1. Choosing the bins

When moving from p, q to $p^{\mathcal{B}}, q^{\mathcal{B}}$, information is inevitably lost. Additionally, a smaller numbers of bins $k = |\mathcal{B}|$ will result in greater information loss. Therefore, for a given granularity level k, bins are chosen such that the least possible error on $\epsilon^{\mathcal{B}_k}$ is introduced for the reference distribution p.

Minimizing the binning error of $\epsilon^{p \to p^{\mathcal{B}}}$: We start by considering bins that minimize the error from going to p to $p^{\mathcal{B}}$. Denote by $\mathcal{B}_k^p(\lambda)$ the set of all k partitions of Ω s.t. $\epsilon^{p \to p^{\mathcal{B}}}$ is at most λ:

$$\mathcal{B}_k^p(\lambda) = \{\mathcal{B}; \epsilon^{p \to p^{\mathcal{B}}} \leq \lambda, \mathcal{B} \in \rho^k(\Omega)\}, \quad (4)$$

where $\rho^k(\Omega)$ denotes the set of all partitions of Ω of size k.

It is easy to test whether a particular $\mathcal{B}$ lies in $\mathcal{B}_k^p(\lambda)$ for any p, but $\mathcal{B}_k^p(\lambda)$ may be very large solution space. For the particular case where no error is tolerated $\mathcal{B}_k^p(\lambda = 0)$, however, the space is much smaller (and may be empty).

We take advantage of our ability to specify the synthetic distribution p, and construct p so that the pmf has s unique values and places all elements with a common value in bins $\mathcal{S}_i$ to form the binning $\mathbf{S} = \{\mathcal{S}_1, \dots, \mathcal{S}_s\}$. With a judicious choice of s, such a reference distribution is sufficient to provide a stern test of generative models. Given a p with $s = |\mathbf{S}|$ flat regions, we have that :

$$\mathcal{B}_k^p(0) = \begin{cases} \emptyset & \text{if } s > k \\ \mathbf{S} & \text{if } s = k \\ \{\bigcup_{\mathcal{S}_i \in \mathbf{S}} \mathcal{P}_i; \mathcal{P}_i \in \rho^{c_i}(\mathcal{S}_i), \sum_{i=1}^s c_i = k\} & \text{if } s < k \end{cases} \quad (5)$$

where $c_i \in \mathbb{N}$ is the cardinal of the partitioning that can be made on $\mathcal{S}_i$. The set $\{\bigcup_{\mathcal{S}_i \in \mathbf{S}} \mathcal{P}_i; \mathcal{P}_i \in \rho^{c_j}(\mathcal{S}_i), \sum_j c_j = k\}$ is all possible

Algorithm 1 Highest granularity level k of q
___
Input: p with flat regions $s = |\mathcal{S}|$, sampling access from q, error threshold ϵ_{test}, significance level δ.
Sample the unknown distribution $\{x_i\}_{i=1}^m; x \sim q$.
for $k \in [s, \dots, 2s]$ **do**
  Solve for $\mathcal{B}_k^*$.
  Compute $p^{\mathcal{B}_k^*}$.
  Construct $\hat{q}_{emp}^{\mathcal{B}_k^*}$ from the m samples.
  Test for $d_{TV}(p^{\mathcal{B}_k^*}, \hat{q}_{emp}^{\mathcal{B}_k^*}) < \epsilon$ at significance δ.
  if the test FAILS **then**
    Return the granularity k at which the test failed.
  end if
end for
___

union of partitions of each flat region $\mathcal{S}_i$, s.t. we end up with the correct number of bins at the end (enforced by $\sum_{i=1}^s c_i = k$).

Maximizing the binning error of $p^{\mathcal{B}}, q^{\mathcal{B}}$: By the triangle inequality, we can see that true error ϵ will always be greater than $\epsilon^{\mathcal{B}}$. Hence, given the choice of binning from $(\mathcal{B}_k^p(0))$, we can select the binning that **maximizes** the error between the induced binned distribution $p^{\mathcal{B}}$ and the induced binned distribution $q^{\mathcal{B}}$:

$$\mathcal{B}_k^* = \underset{\mathcal{B} \in \mathcal{B}_k^p(0)}{\arg\max} \, d(p^{\mathcal{B}}, q^{\mathcal{B}}). \quad (6)$$

This concludes how we select bins at a specific granularity level $\mathcal{B}_k^*$. For a given flat region $\mathcal{A}_i$, if we seek a $k = 2$ bin-granularity: $\rho^2(\mathcal{S}_i)$, the unique solution that maximises the error is to split the positive error region $\{x \in \mathcal{S}_i; q_x - p_x > 0\}$ from the negative error region $\{x \in \mathcal{S}_i; q_x - p_x < 0\}$. If there is no positive or negative error, then every bin is equivalent. Thus, different and tractable solutions for the granularity $k = \{s, \dots, 2s\}$ can be obtained:

$$\mathcal{B}_k^* = \underset{\mathcal{B} \in \mathcal{B}_k^p(0)}{\arg\max} \, d(p^{\mathcal{B}}, q^{\mathcal{B}}). \quad (7)$$

We have presented the discussion in terms of q and $q^{\mathcal{B}}$, but we assume that we only have sample access to q (and hence to $q^{\mathcal{B}}$). Hence, only the empirical pmf of q is accessible. We can approximate with: $\mathcal{B}_k^* \approx \arg\max_{\mathcal{B} \in \mathcal{B}_k^p(0)} d(p^{\mathcal{B}}, \hat{q}_{emp}^{\mathcal{B}})$.

A limitation of this procedure is that it is tied to the number of flat regions s of the distribution p. However, since we have control over p, we can choose p with a number of flat regions that achieves a trade off between (i) posing a suitably demanding challenge and (ii) generating a test that does not require too many samples. In general, in very high dimensions, such as those of protein sequences, for example, structured distributions can be very challenging to learn even if there are only three unique values.

2.3. Hypothesis testing

We now have a sequence of binnings of increasingly difficult granularity levels, ranging from $k = s$ to $k = 2s$: $\mathcal{B}_s^*, \mathcal{B}_{s+1}^*, \dots, \mathcal{B}_{2s}^*$. At each granularity level k, we can use a hypothesis test to detect poor approximation of the ground truth. Since the sampling complexity of the test scales with the probability space and we control the granularity

level k, we can easily generate $O(k)$ samples. Our null hypothesis is that the binned generative model $q^{\mathcal{B}_k^*}$ is within some error ϵ_{test} of the ground truth distribution $p^{\mathcal{B}_k^*}$; i.e. $\mathcal{H}_0 : d^2(p^{\mathcal{B}_k^*}, q^{\mathcal{B}_k^*}) < \epsilon_{test}$. We use the ℓ^2 distance as it is more practical to test for. If the test is able to reject at some granularity k, we stop and return this k as the highest granularity level reached by the generative model q. For the hypothesis testing, we use the closeness test from [15], which tests the distance between two distributions, both of which only have sample access. To make the test applicable to our context, we simply replace the statistical estimates for unknown p by our known statistics for p. The complete testing procedure is outlined in Algorithm 1.

3. EXPERIMENTS

3.1. Datasets

Our synthetic experiments are designed to emulate real world datasets. The most important characteristic is that the sampling space is several orders larger than the number of generated samples m. Second, the support of the targeted distribution (that we denote $\Omega^+ = \{x \in \Omega; p_x > 0\}$) is believed to be much smaller than the whole space, i.e. $\Omega^+/\Omega \approx 0$. The intuition behind this feature is that uniform random sampling is very unlikely to give a valid sample. This is a characteristic of most applications of high-dimensional generative models. Nonetheless, the positive space is still generally believed to be very large and scales with the size of the space. We choose the ratio of approximately $|\Omega^+|/|\Omega| \approx c!/c^c$ as it loosely approaches an estimate given in [16] of the size of a family's sequence space. Since we want to keep the number of flat regions small (as explained in the previous section), we choose to model p by a stair distribution with s stairs, where the last stair is a assigned zero probability mass. Hence for a given s, we have

$$p(x) = \begin{cases} p_{\mathcal{A}_i} & \text{if } x \in \mathcal{A}_i \quad (i \in \{1, \dots, s-1\}) \\ 0 & \text{if } x \in \mathcal{A}_s \end{cases}. \quad (8)$$

3.2. Baselines

Synthetic q: Even though we could scale to high Ω, for visualisation purposes, we set $|\Omega| = 6^6$ and generate a smaller number of samples $m = 1000$. We define a stair distribution p with $s = 4$ flat regions. We produce 4 synthetic generative models q_1, q_2, q_3, q_4 by increasingly perturbing the ground truth distribution: $d_{TV}(p, q_1) = 0$, $d_{TV}(p, q_2) = 0.1$, $d_{TV}(p, q_3) = 0.15$, $d_{TV}(p, q_4) = 0.2$.

Generative models q: We evaluate our evaluation method on state-of-the-art generative models for categorical data. We set $|\Omega| = 6^6$ and define a stair distribution p with $s = 3$ flat regions. **CNF** [17] is a normalizing flow method that learns a mapping from the categorical space to a continuous representation. **CDM** [18] is a diffusion-based model that operates in the discrete space. **argmaxAR** [19] is a normalizing flow method that uses an argmax operation to map a continuous representation to the discrete space. For each trial, we generate $m = 10,000$ samples.
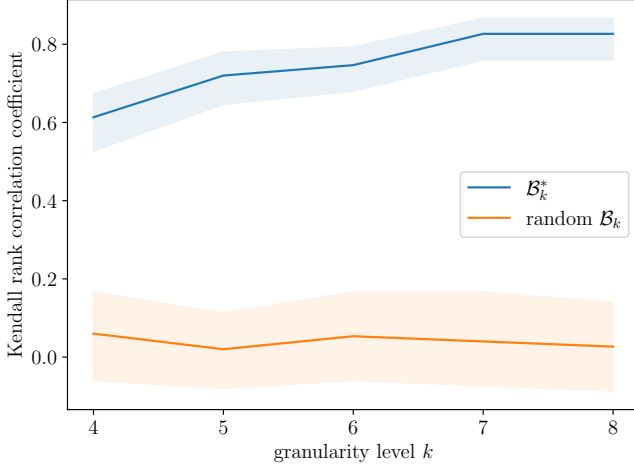
Fig. 2: Accurate ranking measure (Kendall Tau) of our proposed binning algorithm vs a random binning baseline. The proposed algorithm is able to correctly rank the baselines consistently and outperforms a random binning. We report the mean over 50 trials with a 90% bootstrapped CI.
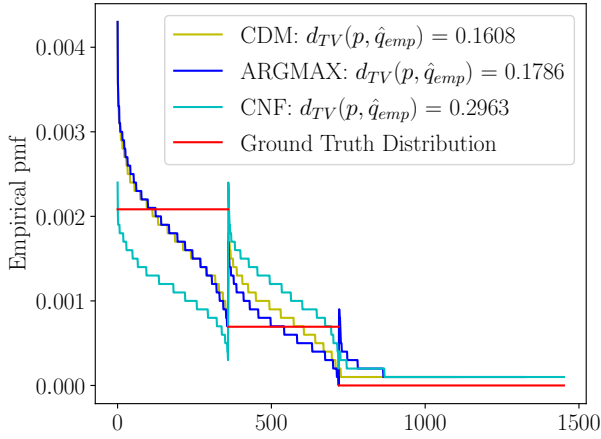

Fig. 3: Empirical pmf of the CDM, argmaxAR, and CNF models over a subset of Ω. The ground truth p is shown in red. The x-axis is sorted differently for each models. For a given model, within a flat region, we sort the $x \in \mathcal{A}_i$ on the x-axis from the most overestimating $\hat{q}_x^{emp}$ to the most underestimating $\hat{q}_x^{emp}$ w.r.t. ground truth. For each model, we compute the empirical total variation error $d_{TV}(p, \hat{q}^{emp})$.

3.3. Experiment Details

We begin by validating our algorithm for choosing $\mathcal{B}_k^*$. We verify the ordering of the empirical total variation given by $\epsilon_{emp}^{\mathcal{B}} \triangleq d_{TV}(p^{\mathcal{B}}, \hat{q}_{emp}^{\mathcal{B}})$ of the synthetically created q_1, q_2, q_3, q_4 aligns with the correct ranking: $q_1 < q_2 < q_3 < q_4$. We compare the ordering provided by the empirical total variation derived form our binning algorithm $\mathcal{B}_k^*$ $d_{TV}(p^{\mathcal{B}_k^*}, \hat{q}_{emp}^{\mathcal{B}_k^*})$ with the empirical total variation derived from some random binning $\mathcal{B}_k$ $d_{TV}(p^{\mathcal{B}_k}, \hat{q}_{emp}^{\mathcal{B}_k})$. To compare raking accuracy, we compute the Kendall Tau correlation coefficient which measures the piece-wise displacements between two rankings.
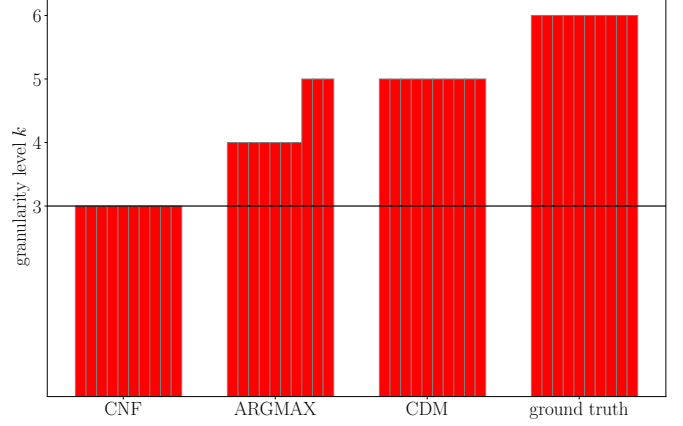

Fig. 4: Highest granularity level k obtained by the generative models for 10 trials.

A higher value corresponds to a more closely matched rank. Results can be viewed in Figure 2. Next, we test our evaluation procedure outlined in Algorithm 1 on state-of-the-art categorical generative models. We validate that the generative model reaching the highest granularity level k with our test is the generative model with the lowest empirical total variation on Ω $d_{TV}(p, \hat{q}^{emp})$. We set the significance level of the hypothesis testing to $\delta = 0.05$ and test for $\epsilon_{test} = 0.1$. We train each model on a dataset of $10,000$ training samples generated from p for 1000 epochs. For each baseline, a hyperparameter search is conducted over the hidden size = $[8, \ldots 64]$, depth = $[1, \ldots, 4]$ and learning rate = $[1e-1, \ldots, 1e-5]$ and we set the remaining parameters to the default values provided in each respective paper.

4. DISCUSSION

Figure 2 shows the binning procedure we propose is sensible. In Figure 3, we show the empirical distribution generated from one trial of the generative models that were trained on the synthetic dataset as described in Section 3.3. Visually, we see that the CNF baseline distribution is a very poor approximation of the ground truth distribution. This is also reflected in its empirical total variation error $d_{TV}(p, \hat{q}^{emp})$. The argmaxAR and CDM baselines are closer, with the CDM slightly outperforming argmaxAR based on $d_{TV}(p, \hat{q}^{emp})$. This is reflected in our hypothesis testing ranking in Figure 4, where we see that all trials of the CDM pass the test at $k = 4$, whereas only a fraction of the argmaxAR trials pass the same test. Every CNF trial is rejected.

5. CONCLUSION

In conclusion, we have introduced an alternative way to evaluate generative models for categorical data. Even though our approach is based on statistical tests, it remains applicable for very large distribution space which is the setting of interest for those models. In future work, we plan on providing theoretical guarantee of our procedure and to further exploit our control over p to design an even more powerful test. pagebreak

6. COPYRIGHT FORMS

7. REFERENCES

[1] Cristina Garbacea, Samuel Carton, Shiyan Yan, and Qiaozhu Mei, "Judge the judges: A large-scale evaluation study of neural language models for online review generation," in *Proc. Conf. on Empirical Methods in Natural Language Process. and Int. Joint Conf. on Natural Language Process EMNLP-IJCNLP*, 2019.

[2] Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao, "Evaluation of text generation: A survey," arXiv preprint: arXiv 2006.14799, 2020.

[3] Sharon Zhou, Mitchell L. Gordon, Ranjay Krishna, Austin Narcomey, Li Fei-Fei, and Michael S. Bernstein, "HYPE: A benchmark for human eye perceptual evaluation of generative models," in *Proc. Adv. Neural Info. Process. Syst. NeurIPS*, 2019.

[4] Ali Borji, "Pros and cons of gan evaluation measures," *Computer Vision and Image Understanding*, vol. 179, pp. 41–65, 2019.

[5] Rylee Thompson, Boris Knyazev, Elahe Ghalebi, Jungtaek Kim, and Graham W. Taylor, "On evaluation metrics for graph generative models," in *Proc. Int. Conf. Learning Representations ICLR*, 2022.

[6] L. Theis, A. van den Oord, and M. Bethge, "A note on the evaluation of generative models," in *Proc. Int. Conf. Learning Representations ICLR*, 2016.

[7] Yuhuai Wu, Yuri Burda, Ruslan Salakhutdinov, and Roger B. Grosse, "On the quantitative analysis of decoder-based generative models," in *Proc. Int. Conf. Learning Representations ICLR*, 2017.

[8] Vaishnavh Nagarajan, Anders Andreassen, and Behnam Neyshabur, "Understanding the failure modes of out-of-distribution generalization," in *Proc. Int. Conf. Learning Representations ICLR*, 2021.

[9] Aaron van den Oord and Benjamin Schrauwen, "Factoring variations in natural images with deep gaussian mixture models," in *Proc. Adv. Neural Info. Process. Syst. NeurIPS*, 2014.

[10] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan, "Do deep generative models know what they don't know?," in *Proc. Int. Conf. Learning Representations ICLR*, 2019.

[11] Massimo Caccia, Lucas Caccia, William Fedus, Hugo Larochelle, Joelle Pineau, and Laurent Charlin, "Language gans falling short," in *Proc. Int. Conf. Learning Representations ICLR*, 2020.

[12] T. Batu, E. Fischer, L. Fortnow, R. Kumar, R. Rubinfeld, and P. White, "Testing random variables for independence and identity," in *Proc. IEEE Symp. on Foundations of Computer Sci.*, 2001.

[13] Ilias Diakonikolas, Themis Gouleakis, Daniel M. Kane, John Peebles, and Eric Price, "Optimal testing of discrete distributions with high probability," in *Proc. ACM SIGACT Symp. on Theory of Comput.*, 2021.

[14] Clément L. Canonne, *A Survey on Distribution Testing: Your Data is Big. But is it Blue?*, Number 9 in Graduate Surveys. 2020.

[15] Tuğkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White, "Testing closeness of discrete distributions," *J. ACM*, vol. 60, no. 1, 2013.

[16] Jeanne Trinquier, Guido Uguzzoni, Andrea Pagnani, Francesco Zamponi, and Martin Weigt, "Efficient generative modeling of protein sequences using simple autoregressive models," *Nature Communications*, vol. 12, 2021.

[17] Phillip Lippe and Efstratios Gavves, "Categorical normalizing flows via continuous transformations," in *Proc. Int. Conf. Learning Representations ICLR*, 2021.

[18] Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling, "Argmax flows and multinomial diffusion: Learning categorical distributions," in *Proc. Adv. Neural Info. Process. Syst. NeurIPS*, 2021.

[19] Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling, "Argmax flows: Learning categorical distributions with normalizing flows," in *Proc. Symposium on Adv. in Appr. Bayesian Inference*, 2021.