Position: **What makes an image realistic?**

Lucas Theis [1]

Abstract

The last decade has seen tremendous progress in our ability to *generate* realistic-looking data, be it images, text, audio, or video. Here, we discuss the closely related problem of *quantifying* realism, that is, designing functions that can reliably tell realistic data from unrealistic data. This problem turns out to be significantly harder to solve and remains poorly understood, despite its prevalence in machine learning and recent breakthroughs in generative AI. Drawing on insights from algorithmic information theory, we discuss why this problem is challenging, why a good generative model alone is insufficient to solve it, and what a good solution would look like. In particular, we introduce the notion of a *universal critic*, which unlike adversarial critics does not require adversarial training. While universal critics are not immediately practical, they can serve both as a North Star for guiding practical implementations and as a tool for analyzing existing attempts to capture realism.

1. Introduction

What distinguishes realistic images from unrealistic images? Humans are able to detect a wide variety of flaws in images and other sensory data, yet there are no robust losses which could be used to penalize unrealistic images across a broad set of tasks in machine learning, and no widely accepted formal notion of realism exists today. In particular, we are interested in real-valued functions U producing a low value $U(\mathbf{x})$ when some data $\mathbf{x}$ is *realistic* and a large value when $\mathbf{x}$ is *unrealistic*. Here, $\mathbf{x}$ could be a single image, a small set of images, or a video. But our discussion will also be relevant for other types of data such as text of arbitrary length or more generally any data drawn from some distribution which we will denote P.

[1]Google DeepMind, London, UK. Correspondence to: Lucas Theis <theis@google.com>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

Potential applications of such functions are plentiful and include anomaly detection (Ruff et al., 2021), deepfake detection (Sha et al., 2023; Pondoc et al., 2023), generative model evaluation (Theis et al., 2016; Heusel et al., 2017; Borji, 2019), model distillation (van den Oord et al., 2018; Yin et al., 2023), neural compression (Ballé et al., 2021; Yang et al., 2023), computational photography (Fang et al., 2020), and computer graphics (Herzog et al., 2012; Reinhard et al., 2013; Poole et al., 2023). Unfortunately, their implementation is extremely challenging. Our ability to *generate* realistic data is rapidly improving (e.g., Dhariwal and Nichol, 2021) yet no reliable candidates or recipes for constructing U exist in machine learning today. This is not for a lack of trying. While some progress has been made in the *detection* of unrealistic examples, the design of functions that are robust to *optimization* (for tasks involving generation) has been less successful. The latter problem is significantly harder because our function now not only has to detect a limited set of artefacts but has to anticipate any unrealistic examples an optimization might run into. Weaknesses in a function's design often only make themselves known once we start optimizing (Ding et al., 2021). Complicating the matter is the fact that the optimization depends on U itself.

To give a more concrete example of the kind of tasks we are interested in, consider the following loss which naturally comes up in lossy compression. If $\mathbf{x} = g(\mathbf{z})$ is the output of a neural network, we may want to find a representation $\mathbf{z}$ such that

$$R(\mathbf{z}) + \alpha d(\mathbf{x}, \mathbf{x}^*) + \beta U(\mathbf{x}) \tag{1}$$

is minimal, where d measures the distance to some target image $\mathbf{x}^*$ and R is the number of bits required to encode $\mathbf{z}$.

In this paper we will take the view that $\mathbf{x}$ is realistic if it appears to have come about in a particular way, which is another way of saying that $\mathbf{x}$ is a plausible sample of a distribution P capturing the data generating process. What is considered realistic therefore depends on P. If P is a distribution over natural images then most photos would qualify as realistic. While an MNIST image (LeCun and Cortes, 2010) would not be considered a realistic example of a natural image, we would still consider it to be realistic if P is the distibution of MNIST digits.

In Section 2 we will first review why common approaches to formalizing realism in terms of probability and typicality

fail. This will highlight the challenges involved in defining realism and provide motivation for later sections. In Section 3 we will review much more successful notions of realism based on divergences, adversarial losses, and feature statistics, and discuss how they still fall short of our goal. In Section 4 we will make the case that ***randomness deficiency* (Li and Vitányi, 1997) captures realism** and introduce the concept of a *universal critic*. Finally, in Section 5 we will apply our newly gained understanding of realism to examples from the machine learning literature.

What has been referred to as *realism* (e.g., Fan et al., 2018; Theis and Wagner, 2021; Careil et al., 2023) is also often referred to as *perceptual quality* (e.g., Blau and Michaeli, 2018; Fang et al., 2020; Salehkalaibar et al., 2023). It is therefore natural to wonder to what extent human perception should factor into its formalization. Our approach to defining realism is normative, that is, we consider how an idealized observer would judge realism. Similar to how Bayesian inference does not take inspiration from neuroscience but Bayesian decisions resemble human decisions (e.g., Knill and Pouget, 2004), we too can hope that human perception agrees with our definition of realism because it addresses a similar task as that faced by humans. In Section 4.4, we will further make the case that *batched universal critics* not only generalize no-reference metrics and divergences—which represent the prevalent ways of formalizing realism—but are also a better model of a human observer.

2. Probability and Typicality

In this section we review the two most common approaches to capture realism found in machine learning, namely those based on probability and typicality, and their failures. Similar failures of probability and typicality have been documented in the anomaly detection literature (e.g., Choi et al., 2019; Le Lan and Dinh, 2021; Osada et al., 2023) but are worth repeating as they continue to be a source of confusion.

2.1. Probability

If $\mathbf{x}$ is discrete, it is natural to consider its probability under P to determine whether it is a realistic example of P. After all, if $\mathbf{x}$ has low probability then it seems unlikely to have come from P. This intuition is widespread in machine learning. Unsupervised anomaly detection, for instance, generally defines anomalies as those data points having low probability or density under a distribution of *normal* examples (Ruff et al., 2021), where the probability is often measured in some feature space (e.g., Zong et al., 2018). Probability density is also frequently maximized in an attempt to guide synthetic images towards more realistic examples (e.g., Sønderby et al., 2017; Graikos et al., 2022). To see how this approach might fail, consider the following simple example.

Example 1 (Probability). Consider a computer program simulating a sequence of independent and nearly unbiased coin tosses, $\mathbf{x}^N = (x_1, \dots, x_N)$ with $P(x_n = 1) = 0.5+\varepsilon$ for some very small $\varepsilon > 0$. For reasonably large N, we would expect the program to output a number of 1s which is close to N/2 and we would suspect a bug if the program outputs a sequence of only 1s, yet this is the most probable sequence.

Example 1 shows that maximizing $P(\mathbf{x})$ can lead to unrealistic examples. It also shows that $P(\mathbf{x})$ would not detect a bug which causes a program to only output 1s. If instead we count the number of 1s, $k = \sum_{n=1}^{N} x_n$, and measure the probability of k, this bug could be detected. Does this mean we only need to find the right set of features? By ignoring some aspects of the data, we risk not detecting unrealistic examples. We might therefore conclude that we simply need to test sufficiently many features. Unfortunately this approach also runs into trouble. Consider testing whether $\mathbf{x}$ has 1s in even places and 0s in odd places, $\mathbf{x} = \texttt{0101..01}$. The probability of this sequence is approximately 2^{-N} so that we would reject it with high confidence if we happen to observe it. However, since all sequences have roughly the same probability, we would reject *every* sequence as unrealistic if we tested *all* features identifying a specific sequence.

Using densities instead of probabilities introduces an additional challenge, namely that our answer now depends on the parametrization of the data. If P is an exponential distribution with rate 1, say, then values of $\mathbf{x}$ close to zero seem preferable over larger values if judged by their density. But if we consider $\mathbf{y} = e^{-\mathbf{x}}$ instead, then all values of $\mathbf{y}$ would now be considered equally preferable.

2.2. Weak Typicality

Many readers will not have been surprised by the inability of probabilities to capture realism thanks to the widely known *asymptotic equipartition property* (AEP) of random sequences (Cover and Thomas, 2006). This property is such that if $\mathbf{x}^N = (\mathbf{x}_1, \dots, \mathbf{x}_N)$ is a sequence of i.i.d. random variables drawn from P, then with probability 1 we have

$$\lim_{N\to\infty} -\tfrac{1}{N} \log P(\mathbf{x}_1, \dots, \mathbf{x}_N) = H[\mathbf{x}_n] \qquad (2)$$

almost surely, where $H[\mathbf{x}_n]$ is the entropy of P. The *typical set* is defined as (Cover and Thomas, 2006)

$$A_\delta^N = \{\mathbf{x} : |-\tfrac{1}{N} \log P(\mathbf{x}^N) - H[\mathbf{x}_n]| < \delta\} \qquad (3)$$

and elements from this set are considered *weakly typical*. While other notions of typicality exist, weak typicality is the one most commonly encountered in the machine learning literature (Nalisnick et al., 2019b; Choi et al., 2019; Dieleman, 2020). The AEP implies that as N increases, the probability

that a randomly drawn sequence is contained in the typical set A_δ^N approaches 1 for any $\delta > 0$. That is, a realistic sequence is likely to be typical. (While we have stated the AEP for sequences of independent and discrete random variables, generalizations to dependent and continuous sources exist and are well known; e.g., Algoet and Cover, 1988).

The above suggests that instead of expecting the probability to be large, we should expect realistic $\mathbf{x}$ to have negative log-probability close to the entropy—or the probability to be roughly $2^{-H[\mathbf{x}]}$, especially if $\mathbf{x}$ is high-dimensional. It therefore appears that $|-\log P(\mathbf{x}) - H[\mathbf{x}]|$ would be a good candidate for a measure of realism (Choi et al., 2019; Nalisnick et al., 2019b). Unfortunately, also this definition fails to quantify realism as the following examples demonstrate.

Example 2 (Typicality). Consider again a sequence of independent coin tosses. If the coin is unbiased, then the log-probability of any sequence is exactly the entropy, $-\log_2 P(\mathbf{x}^N) = N$. In other words, in this case the probability of $\mathbf{x}^N$ under P is completely uninformative and the typical set contains *every* sequence. Does this mean that every sequence of coin flips is realistic? Clearly, there is a sense in which the sequence `0000000000` is less realistic than `1100010100` which is not captured by weak typicality.

Example 3 (Typicality). As another example, consider a multivariate Gaussian distribution with density $p(\mathbf{x}) \propto \exp(-\|\mathbf{x}\|^2)$. With high probability, the negative log-density of a random sample will be close to the differential entropy, which amounts to the norm $\|\mathbf{x}\|$ being roughly constant. While we would expect realistic examples from our distribution to look like uncorrelated noise, optimizing for typicality will only constrain the norm. If $\mathbf{x}$ represents an image, our optimization will merely adjust its contrast but will not decorrelate pixels as one might hope.

Weak typicality may be a necessary criterion for realism but it is clearly not sufficient. Put differently, the typical set contains the realistic sequences we care about but also many sequences which are unrealistic, such as long sequences of fair coin flips which all come up heads.

Probability and typicality both fail as a measure of realism because they address the wrong question. They tell us something about $\mathbf{x}$ *assuming that* $\mathbf{x}$ *has distribution* P. However, we cannot make this assumption, since whether or not $\mathbf{x}$ follows P is precisely the question we are trying to answer. That is, **we are not interested in the probability (or typicality) of $\mathbf{x}$ given P, but in the probability of P given $\mathbf{x}$.**

An extended discussion of typicality can be found in Appendix A.

3. Divergences

More successful notions of realism are based on divergences between a ground-truth data distribution P and a distribution Q which we are trying to evaluate. In line with our intuitive notion of realism, if a divergence is zero, then instances of Q are indistinguishable from instances of P, that is, we have perfect realism.

In coding theory, formalizing realism in terms of divergences (Matsumoto, 2018; Blau and Michaeli, 2019; Chen et al., 2022) has resulted in an improved understanding of the lossy compression problem and novel methods to solve them (e.g., Theis and Agustsson, 2021). In practical applications, generative adversarial networks (GANs; Goodfellow et al., 2014) trained with adversarial losses (which approximate divergences) significantly advanced the state of the art in the perceptual quality of generated images (e.g., Denton et al., 2015; Ledig et al., 2017). For the evaluation of generated images, the Fréchet inception distance (Heusel et al., 2017) has established itself as the method of choice and is based on a divergence between distributions over feature activations.

In the following, we review two approaches to approximating divergences based on samples.

3.1. Adversarial Losses

Adversarial losses provide lower bounds on divergences. For the broad class of f-divergences (Rényi, 1961) between two distributions with densities p and q, we can write

$$D_f[q \,\|\, p] = \int p(\mathbf{x}) f\left(\frac{q(\mathbf{x})}{p(\mathbf{x})}\right) d\mathbf{x}, \tag{4}$$

where f is a convex function with $f(1) = 0$. This class of divergences includes the Jensen-Shannon divergence, the total variation distance, and the Kullback-Leibler divergences. For a real-valued function T (with an appropriately limited output range), we obtain the lower bound (Nguyen et al., 2010; Nowozin et al., 2016)

$$D_f[q \,\|\, p] \geq \mathbb{E}_q[T(\mathbf{x})] - \mathbb{E}_p[f^*(T(\mathbf{x}))], \tag{5}$$

where f^* is the convex conjugate of f. T acts as a *critic* whose purpose is to produce values which are large for samples drawn from q and small for samples drawn from p. In practice, the critic may be a neural network $T_{\boldsymbol{\theta}}$ and adversarial training amounts to alternating between maximizing the lower bound with respect to its parameters, $\boldsymbol{\theta}$, and minimizing the bound with respect to the parameters of q (although practical implementations often deviate from this basic recipe).

For the Kullback-Leibler divergence, for instance, we have

$$f(u) = u \log u, \qquad f^*(t) = \exp(t-1), \tag{6}$$

and the bound is tight for

$$T_q(\mathbf{x}) = \log q(\mathbf{x}) - \log p(\mathbf{x}) + 1. \tag{7}$$

Note that this optimal critic depends on the distribution q that we are trying to evaluate. In contrast, in our setting we may only have access to a single instance or a few instances drawn from q. Furthermore, the dependence of the critic on q is responsible for optimization instabilities that are known to plague adversarial training and which we would like to avoid. In Section 4 we will discuss critics which are *universal* in the sense that they do not depend on q and therefore do not require adversarial training.

3.2. Maximum Mean Discrepancy

Maximum mean discrepancy (MMD; Gretton et al., 2012) refers to a class of divergences which have been used for hypothesis testing as well as for generating realistic images (Li et al., 2015; Dziugaite et al., 2015). Given two sets of i.i.d. examples—$\mathbf{x}_1, \dots, \mathbf{x}_M$ and $\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_N$—estimates of MMD can be used to decide whether the two sets were drawn from the same distribution. Formally, we compute

$$\mathrm{MMD}^2(\mathbf{x}^M, \tilde{\mathbf{x}}^N) = \left\| \tfrac{1}{M} \textstyle\sum_m \Phi(\mathbf{x}_m) - \tfrac{1}{N} \sum_n \Phi(\tilde{\mathbf{x}}_n) \right\|^2$$

in some potentially very high (even infinite) dimensional feature space Φ to estimate a squared MMD. Notably, the estimator depends on the two distributions only through examples and unlike adversarial losses does not require optimization of any critic. This makes it worthwhile to consider as a candidate for our function U, especially in regimes where we have access to at least a small number of unrealistic examples. The basic idea is that we would fix a relatively large number of realistic examples and compare it to a small batch of examples we wish to test for realism. Support for this idea also comes from Amir and Weiss (2021) who have shown that MMD can be used to construct an effective full-reference perceptual metric[1] which agrees with human judgments in determining the similarity of pairs of images. To construct the metric, each image was treated as a distribution over small patches.

It remains unclear how to use MMD to quantify the realism of a single data point without a reference. For an image, one might compare features averaged over image patches to the features obtained from patches of a larger dataset of images, and similar ideas have shown promise in image quality assessment (e.g., Mittal et al., 2013; Zhang et al., 2015). But the limitations of this approach are also clear as not all realistic images have statistics representative of the entire data distribution.

A bigger concern perhaps is that the statistical power of MMD can drop quickly as the dimensionality of the problem increases[2] (Ramdas et al., 2015), suggesting that we might need a very large number of examples if we want to identify defects in reasonably sized images or videos.

The MMD estimator makes fewer assumptions than is necessary for us. In particular, it seems reasonable to assume access to P (or a good approximation) both from a conceptual and a practical point of view, given the power of today's generative models. By incorporating P into our definition of realism, we can hope to quantify realism more efficiently. MMD leaves it to us to choose Φ and does not provide a clear mechanism for incorporating P.

4. Universal Critics

In this section, we introduce an alternative notion of realism based on concepts from algorithmic information theory (AIT) (Martin-Löf, 1966; Chaitin, 1987; Li and Vitányi, 1997). AIT is concerned with whether a given sequence of bits is a *random* sequence of independent coin flips. If we can answer this question, then the answer to the more general question of whether $\mathbf{x}$ is an instance of P directly follows, since if we use P to (losslessly) compress $\mathbf{x}$ then the resulting bits should appear random. Several notions of randomness have been proposed and studied in AIT. Some have been rejected on the basis of flaws, such as von Mises randomness (Mises, 1919). Other notions survived scrutiny and turned out to be equivalent (Chaitin, 2001, Chapter 3), namely Martin-Löf randomness (Martin-Löf, 1966), Solovay randomness (Solovay, 1975), incompressibility (Li and Vitányi, 1997), and Chaitin randomness (Chaitin, 2001). The fact that multiple authors converged to essentially the same answer should give us hope that there is something fundamental about the concepts they discovered. Instead of reviewing the different (equivalent) definitions of randomness, we start with the conclusion relevant for us and then develop a justification for it below. In particular, AIT suggests the following measure of randomness to decide whether $\mathbf{x}$ was drawn from a distribution P:

$$U(\mathbf{x}) = -\log P(\mathbf{x}) - K(\mathbf{x}) \tag{8}$$

Here, $K(\mathbf{x})$ is the *Kolmogorov complexity* of $\mathbf{x}$ which is defined as the length of a shortest program (in some Turing complete programming language) which outputs $\mathbf{x}$. The quantity $U(\mathbf{x})$ is also known as *randomness deficiency*[3] (Li and Vitányi, 1997) but for reasons that will become clear soon, we will refer to U as a *universal critic*.

The following characterization of Kolmogorov complexity

[1] A full-reference metric takes two images as arguments where a no-reference metric only has a single input.

[2] This assumes that the difficulty of the estimation problem remains constant, as measured by the KL divergence between the two distributions being tested.

[3] A more accurate definition of randomness deficiency would be over sequences of arbitrary length but for simplicity we will work with Eq. 8.

will be more convenient for us,

$$K(\mathbf{x}) = -\log S(\mathbf{x}), \quad S(\mathbf{x}) = \textstyle\sum_n \pi_n Q_n(\mathbf{x}), \quad (9)$$

where $S(\mathbf{x})$ is *Solomonoff's probability* (Solomonoff, 1960) and requires some explanation. Consider the set of all discrete probability distributions implementable in a programming language of your choice. Each program corresponds to a sequence of bits and we are free to interpret those bits as a natural number. In other words, the set of computable probability distributions is countable and we can assign each such distribution Q_n a number n. S is a mixture of all of these. The choice of weights π_n is not critical for now and we can choose $\pi_n \propto 1/n^2$ or $\pi_n = 2^{-C(n)}$ where $C(n)$ is the number of bits assigned to n by some universal code.

A similar argument holds for continuous sample spaces (Li and Vitányi, 1997, Chapter 4.5). That is, there is a corresponding S for continuous sample spaces which sums over measures, or lower semicomputable semimeasures[4] to be precise. A measure is *semicomputable* if it can be approximated from below to arbitrary precision, that is, it is enough to be able to compute approximations of a measure for it to be included in the mixture S. For simplicity, we will focus on discrete spaces even though continuous spaces are relevant in practice if we want to optimize for realism.

For a more thorough treatment of these concepts, see the excellent introduction to Kolmogorov complexity by Li and Vitányi (1997). Here we will try to not get hung up on technical details since we are ultimately interested in practical applications and—as some readers may already rightfully object—Kolmogorov complexity and S are uncomputable. Nevertheless, we will argue that universal critics as defined in Eq. 8 correctly formalize realism, and that it is useful to understand practical approaches as (good or bad) approximations of it—similar to how deriving Bayesian posteriors is useful even when they are intractable since they can guide us towards better approximations.

As a first step, note that if P is computable (or just lower semi-computable), then there exists an m with $Q_m = P$. If π_n is our prior belief that $\mathbf{x}$ was generated by Q_n, then

$$\begin{aligned} -U(\mathbf{x}) &= \log P(\mathbf{x}) - \log S(\mathbf{x}) && (10)\\ &= \log \frac{\pi_m Q_m(\mathbf{x})}{\sum_n \pi_n Q_n(\mathbf{x})} - \log \pi_m && (11)\\ &= \log \Pr(m \mid \mathbf{x}) - \log \pi_m && (12) \end{aligned}$$

can be seen as the log-posterior probability of P given $\mathbf{x}$ up to a constant, consistent with our earlier notion of realism.

[4] A semimeasure integrates to a value less or equal 1. S itself is an example of a semimeasure with $\sum_{\mathbf{x}} S(\mathbf{x}) < 1$. This is due to the *halting problem* causing some unknowable set of indices n to correspond to programs which never stop running. For these n, we set $Q_n(\mathbf{x}) = 0$.

4.1. Batched Universal Critics

How does our new notion of realism compare to existing notions of realism? U is a particular instance of a no-reference metric since it can be applied to a single instance $\mathbf{x}$. But it turns out that we can also use it to approximate divergences by taking averages, as we will demonstrate. Consider evaluating the distribution Q based on its average realism score as assigned by U. We have

$$\begin{aligned} \mathbb{E}_Q[U(\mathbf{x})] &= \mathbb{E}_Q[\log S(\mathbf{x}) - \log P(\mathbf{x})] && (13)\\ &\le \mathbb{E}_Q[\log Q(\mathbf{x}) - \log P(\mathbf{x})] && (14)\\ &= D_{\mathrm{KL}}[Q\|P], && (15) \end{aligned}$$

where Eq. 14 is due to Q minimizing cross-entropy when the data is distributed according to Q. On the other hand, if Q is computable (or just lower semicomputable), we have

$$\begin{aligned} \mathbb{E}_Q[U(\mathbf{x})] &= \mathbb{E}_Q[\log \textstyle\sum_n \pi_n Q_n(\mathbf{x}) - \log P(\mathbf{x})] && (16)\\ &\ge \mathbb{E}_Q[\log(\pi_m Q_m(\mathbf{x})) - \log P(\mathbf{x})] && (17)\\ &= D_{\mathrm{KL}}[Q\|P] - \log \tfrac{1}{\pi_Q} && (18) \end{aligned}$$

since we must have $Q_m = Q$ for some m. For ease of notation, we also write π_Q to refer to π_m. The inequality follows because the terms we dropped from the sum are all non-negative. What this sandwich bound implies is that our universal critic works well as a replacement for the optimal critic T_Q (Eq. 7) if the *complexity of* Q, $\log(1/\pi_Q)$, is low relative to the KL divergence between Q and P. This agrees with our intuition for realism. In particular, we are more likely to accept an alternative explanation of the data if the explanation is simple, that is, if it can be described in a few words (or bits). A sequence of zeros (Examples 1 and 2) is easy to detect because it is cheap to describe ("always output 0"). While the critic U depends on P, it is *universal* in the sense that it does not depend on Q.

Example 4 (Low Complexity). Consider a distribution over natural images P and a distribution Q_0 which assigns all its mass to a single flat image, $Q_0(\mathbf{x} = 0) = 1$. Based on our bounds above, we should expect U to detect Q_0 as unrealistic since it is cheap to describe, that is, $\log(1/\pi_{Q_0})$ is small for any reasonable coding scheme. In contrast, using $-\log P(\mathbf{x})$ instead of $U(\mathbf{x})$ would fail to detect Q_0 since natural image distributions generally assign high probability to flat images. Similarly, images of Gaussian white noise would be detected since their distribution is cheap to describe as independent copies of a simple distribution.

Note from Example 4 that low-complexity distributions can have both low or high entropy, that is, the complexity (or coding cost) $\log(1/\pi_Q)$ of a distribution Q is different from its entropy.

Example 5 (High Complexity). As another example, consider a distribution which has memorized a training set of natural images, $Q_{\mathcal{D}}(\mathbf{x}) \propto \sum_{\mathbf{x}' \in \mathcal{D}} \delta_{\mathbf{x}'}(\mathbf{x})$. This distribution will remain undetected since its complexity is high. To describe $Q_{\mathcal{D}}$, we would have to encode every image in the training set $\mathcal{D}$. On the one hand, this means that U may perform poorly as an approximation of the KL divergence between $Q_{\mathcal{D}}$ and P (due to the loose lower bound, Eq. 17). On the other hand, this behavior is in line with our intuitive notion of realism since we would also fail to tell a single example generated by P from a single example selected from the training set. Like the universal critic, we consider training set images to be realistic[5].

As a side note, a tighter bound can be obtained by choosing m which maximizes $\pi_m Q_m(\mathbf{x})$ instead of choosing m with $Q_m = Q$ as in Eq. 17. This would correspond to the *minimum description length* (MDL) principle of selecting models based on the total cost of describing the data and the model (Rissanen, 1978). That is, where adversarial training uses objectives such as maximum likelihood to select a critic, the universal critic can be viewed as selecting a critic based on MDL.

We can further improve the critic's odds of detecting Q by feeding it multiple independent examples. We define a *batched universal critic* as a critic of the form

$$U^B(\mathbf{x}^B) = \log \sum_n \pi_n \prod_b Q_n(\mathbf{x}_b) - \log \prod_b P(\mathbf{x}_b), \quad (19)$$

where $\mathbf{x}^B = (\mathbf{x}_1, \ldots, \mathbf{x}_B)$. In the following, let Q^B indicate the product measure, that is, a distribution over B independent samples from Q. Then

$$\begin{aligned}
&\tfrac{1}{B}\mathbb{E}_{Q^B}[U^B(\mathbf{x}^B)] && (20)\\
&\geq \tfrac{1}{B}\mathbb{E}_{Q^B}\left[\log\left(\pi_m Q_m^B(\mathbf{x}^B)\right) - \log P^B(\mathbf{x}^B)\right] && (21)\\
&= \tfrac{1}{B}\textstyle\sum_b \mathbb{E}_Q\left[\log Q_m(\mathbf{x}_b) - \log P(\mathbf{x}_b)\right] + \tfrac{1}{B}\log \pi_m\\
&= \mathbb{E}_Q\left[\log Q(\mathbf{x}_b) - \log P(\mathbf{x}_b)\right] + \tfrac{1}{B}\log \pi_Q && (22)\\
&= D_{\mathrm{KL}}[Q\|P] - \tfrac{1}{B}\log \tfrac{1}{\pi_Q} && (23)
\end{aligned}$$

for some m where $Q_m = Q$. Compared to Eq. 17, we now obtain a tighter bound, which agrees with our intuition that upon observing multiple examples we should be able to do a better job of discriminating Q from P. In the limit of large B we recover the KL divergence. In this sense our notion of realism generalizes prior notions of realism based on no-reference metrics or divergences, and allows us to interpolate between the two.

[5] More concretely, we can say that an average training set image would be considered realistic in the sense that $\mathbb{E}_{\mathcal{D}}[\mathbb{E}_{Q_{\mathcal{D}}}[U(\mathbf{x})]] = \mathbb{E}_P[U(\mathbf{x})] \leq 0$ (Eq. 14).

4.2. Universal Tests

Deciding whether $\mathbf{x}$ is realistic or not means deciding between two hypotheses. The null hypothesis is that $\mathbf{x}$ is realistic, by which we mean that $\mathbf{x}$ came about in a particular way, modelled by $\mathbf{x}$ being drawn from the distribution P. Our alternative hypothesis is that $\mathbf{x}$ is unrealistic, or that it came about by some other process Q. For example, P may be a distribution over photos but an alternative explanation could involve heavy compression with JPEG, corresponding to a distribution over images with blocking artefacts. If there are multiple ways in which $\mathbf{x}$ can fail to be realistic, Q_n, then it is natural to assign probabilities π_n to these events and to consider a mixture distribution as our alternative hypothesis. We end up with S as our alternative hypothesis if the only assumption we are willing to make is that $\mathbf{x}$ was generated by some computable process. By the well-known *Neyman-Pearson lemma* (Neyman et al., 1933), the most powerful test is then a likelihood ratio test of the form

$$\log S(\mathbf{x}) - \log P(\mathbf{x}) > \eta, \quad (24)$$

where η is a parameter which controls the trade-off between false positives and false negatives. Note that the left-hand side is our universal critic. If we accept the Neyman-Pearson lemma then it is easy to accept that our measure of realism should take the form of a likelihood ratio instead of just $P(\mathbf{x})$. However, this does not yet explain why our choice of alternative hypothesis should be S.

We can provide the following additional justification for the universal critic. Assume that instead of S we decide to use another alternative hypothesis corresponding to a computable (or just lower semicomputable) measure Q. Then it is not difficult to see that

$$U^B(\mathbf{x}^B) \geq \log Q^B(\mathbf{x}^B) - \log P^B(\mathbf{x}^B) - \log \tfrac{1}{\pi_Q} \quad (25)$$

for all $\mathbf{x}^B$ and all B (following the same reasoning as in Eqs. 20-23). That is, U^B additively dominates any computable likelihood ratio test and the constant $\log(1/\pi_Q)$ becomes negligible for sufficiently large B. Asymptotically, the universal critic is as sensitive to unrealistic examples as any other test based on an alternative hypothesis Q.[6]

4.3. MCMC

When optimizing data for realism it is natural to look to Markov chain Monte Carlo (MCMC) methods for solutions. In MCMC, the data is stochastically perturbed until it converges to a sample from our target distribution P (at which point it would appear realistic). For example, for a continuous distribution with differentiable density p, a simple

[6] Li and Vitányi (1997, Chapter 4.3) proved the stronger result that randomness deficiency additively dominates any so-called sum-P test.

MCMC strategy based on *Langevin diffusion* uses updates of the form

$$\mathbf{x}_{t+\varepsilon} = \mathbf{x}_t + \varepsilon \left(\nabla \log p(\mathbf{x}_t) + \sqrt{2}\boldsymbol{\eta}_t \right), \tag{26}$$

where $\boldsymbol{\eta}_t \sim \mathcal{N}(0, \mathbf{I})$ is independent Gaussian noise. For infinitesimal ε, the sequence of $\mathbf{x}_t$ converges to the distribution P. For a fixed $\varepsilon > 0$ the stationary distribution will only approximate P, but this can be addressed by performing additional Metropolis-Hastings accept/reject steps (Besag, 1994; Welling and Teh, 2011).

While MCMC produces realistic examples, it is not directly applicable to problems of the form of Eq. 1, since it is unclear how to translate an MCMC algorithm into a loss function U. If we naively interpreted Eq. 26 as a noisy gradient update, then this would correspond to using p as a measure of realism and is bound to fail (Section 2.1).

In a second attempt to make MCMC work for us, consider the sequence of distributions generated by Eq. 26. Let q^0 be the density used to initialize $\mathbf{x}_0$. Then each update produces a new density q^t which approaches p as t goes to infinity. Maoutsa et al. (2020) and Song et al. (2021) showed that the deterministic updates

$$\mathbf{x}_{t+\varepsilon} = \mathbf{x}_t + \varepsilon \left(\nabla \log p(\mathbf{x}_t) - \nabla \log q^t(\mathbf{x}_t) \right) \tag{27}$$

follow the same sequence of distributions q^t (for infinitesimal ε, or approximately for $\varepsilon > 0$). Eq. 27 suggests moving $\mathbf{x}_t$ towards high-density regions of p but away from high-density regions of its *current distribution* q^t. When optimizing for realism, we do not know q^t. But assuming an underlying q^t exists, a Bayesian approach would be to estimate the missing gradient in Eq. 27 by assigning prior probabilities π_n to candidate densities q_n and then to form the posterior expectation

$$\textstyle\sum_n P(n \mid \mathbf{x}_t) \nabla \log q_n(\mathbf{x}_t) = \nabla \log \sum_n \pi_n q_n(\mathbf{x}_t) \tag{28}$$

where $P(n \mid \mathbf{x}_t) \propto \pi_n q_n(\mathbf{x}_t)$ (Appendix B). Note the resemblance of the right-hand side to Solomonoff's probability. If we restrict the universal critic to distributions with differentiable densities, then gradient descent on its density can be viewed as a Bayesian's attempt to simulate Eq. 27.

4.4. Limited-Memory Observer

We demonstrated useful statistical properties of universal critics and discussed connections to adversarial critics, significance testing, and MCMC. However, did we capture anything about *how humans perceive inputs*? In this section we will argue that batched universal critics not only generalize no-reference metrics and divergences, but also represent a more realistic model of human observers.

No-reference metrics are motivated by the idea that humans can look at a single image and decide whether it is realistic or not. It should therefore be possible to design a function which performs this task similarly well. However, in practice, even human observers often have access to not just a single image but a number of images. When evaluating the quality of image codecs or generative models, for example, human raters typically receive a stream of images and are asked to rate them. Mean opinion score tests ask raters to assign a score between 1 and 5 to each image while an alternative approach asks raters to classify between real and generated images (Denton et al., 2015). A generative model which always produces the same output would easily be identified by humans in such a task, even when the image appears realistic when viewed in isolation. While humans would be able to better detect a faulty generative model over time, no-reference metrics continue to produce the same output no matter how many examples they receive. That is, a no-reference metric is memoryless. While it may have been obtained through training on a set of realistic and unrealistic examples, it is unable to adapt to the method(s) currently under evaluation once it has been fixed.

Divergences represent the other extreme as they have access to the entire distribution. This corresponds to a human observer who has received an infinite stream of examples of either real or generated data. The total variation distance, for example, measures the probability of an *optimal observer* correctly classifying real from generated data (Nguyen et al., 2009; Blau and Michaeli, 2018),

$$p_{\text{success}} = \frac{1}{2} D_{\text{TV}}[Q, P] + \frac{1}{2}, \tag{29}$$

that is, an observer who has had access to infinitely many training examples. Other divergences can be similarly interpreted as classifiers which are optimal but with respect to different losses (Nguyen et al., 2009).

Like other no-reference metrics and human observers, universal critics provide a score for individual examples. Like divergences they can also be viewed as the score of a classifier deciding between two hypotheses, but unlike divergences they only have access to a finite set of training examples. This limitation means that prior assumptions become more important. Alternatively, universal critics can be viewed as measuring the performance of an ideal observer with limited memory (Appendix C). In this sense, batched universal critics are a better model of human observers than either no-reference metrics (memoryless) or divergences (infinite memory).

Universal critics as defined in Eq. 8 depend on an uncomputable Kolmogorov complexity and therefore could be implemented neither by humans nor computers. Given sufficient evidence, it will detect any failures a human observer might detect (Section 4.2) but will also detect any unrealistic properties that would be missed by us. In this sense, universal critics provide a sufficient but not necessary criterion for

high perceptual quality (unlike typicality, which is necessary but not sufficient). The limitations of human observers can be incorporated naturally into universal critics by limiting S to a mixture over fewer components. However, characterizing the limitations and abilities of human observers is beyond the scope of this paper. We refer to Griffiths and Tenenbaum (2003), who studied the ability of humans to detect randomness in binary sequences, and compared it to algorithmic notions of randomness.

5. Related Work

Given the wide range of related fields and the vast amount of work in them (Section 1), it is impossible to review any meaningful fraction of related work here. Instead, we will focus on two successful examples with interesting connections to universal critics.

5.1. Input Complexity

Several papers on outlier detection made the puzzling observation that generative models trained on one dataset of images can assign higher probability to other datasets (Choi et al., 2019; Nalisnick et al., 2019a; Hendrycks et al., 2019). Serrà et al. (2020) found that the issue virtually disappears if instead of measuring log-probabilities, the negative log-probability under the model is compared with the coding cost of a lossless image compression method such as PNG,

$$-\log P(\mathbf{x}) - C(\mathbf{x}), \tag{30}$$

where $C(\mathbf{x})$ is the coding cost obtained via compression. The authors found that this signal performed significantly better for outlier detection, providing support for our definition of realism (Eq. 8) by viewing $C(\mathbf{x})$ as an approximation to Kolmogorov complexity. It is further enouraging that a simple but flexible compression scheme can provide a useful signal. An interesting question for future research is what a differentiable analogue of C would look like, and whether it can be made robust enough for optimization.

We note that input complexity has also been considered in statistics for its applications in hypothesis testing, including as an approximation to universal tests (Ryabko et al., 2006).

5.2. Score Distillation Sampling

Score distillation sampling (SDS; Poole et al., 2023) is a technique which has gained a lot of popularity for training 3D generative models. Training 3D generative models is challenging due to the high cost associated with collecting 3D data. SDS tries to overcome these limitations by leveraging diffusion models (Sohl-Dickstein et al., 2015) trained on large amounts of 2D images to guide text-to-3D models towards realistic outputs. Briefly, diffusion models define latent variables $\mathbf{z}_t = \alpha_t \mathbf{x} + \sigma_t \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ and a function $\hat{\epsilon}_t(\mathbf{z}_t)$ is trained to predict $\boldsymbol{\epsilon}$. For a conditional diffusion model whose outputs depend on text y, we have the important relationship (Robbins, 1956)

$$\hat{\epsilon}_t(\mathbf{z}_t; y) \approx \mathbb{E}[\boldsymbol{\epsilon} \mid \mathbf{z}_t, y] = -\sigma_t \nabla_{\mathbf{z}_t} \log p_t(\mathbf{z}_t \mid y), \tag{31}$$

where p_t is the distribution of $\mathbf{z}_t$ so that $\hat{\epsilon}_t$ can also be used to estimate the gradient of these log-densities.

Simplifying a bit, Poole et al. (SDS; 2023) propose the following gradient,

$$\nabla_{\mathbf{x}} \mathcal{L}_{\text{SDS}}(\mathbf{x}; y) = \mathbb{E}_{t,\boldsymbol{\epsilon}}[w(t)(\hat{\epsilon}_t(\mathbf{z}_t; y) - \boldsymbol{\epsilon})] \tag{32}$$

where $w(t)$ are hyperparameters assigning weights to the different noise levels. Is $\mathcal{L}_{\text{SDS}}$ a good candidate for U? We can see that SDS tries to find $\mathbf{x}$ such that $\mathbf{z}_t$ is near modes of p_t. Note that p_t is essentially the density of $\mathbf{x}$ smoothed via convolution with a Gaussian kernel, and so SDS appears fundamentally similar to using p as a measure of realism (Section 2.1) and susceptible to similar failures. Indeed, if the data distribution was Gaussian, then p_t would also be Gaussian and the optimal $\mathbf{x}$ would be the mean, which tends to be unrealistic. This raises the question of why SDS performs well in practice. The key to its success lies in classifier-free guidance (CFG; Ho and Salimans, 2021). Instead of using $\hat{\epsilon}_t$ directly, this common trick is to use

$$\hat{\epsilon}_t^v(\mathbf{z}_t; y) = (1 + v)\hat{\epsilon}_t(\mathbf{z}_t; y) - v\hat{\epsilon}_t(\mathbf{z}_t), \tag{33}$$

where $\hat{\epsilon}_t(\mathbf{z}_t)$ is an unconditional prediction of $\boldsymbol{\epsilon}$ and the guidance weight $v \geq 0$ is a hyperparemeter. This corresponds to a gradient signal proportional to

$$v\nabla_{\mathbf{z}_t} \log p_t(\mathbf{z}_t) - (1 + v)\nabla_{\mathbf{z}_t} \log p_t(\mathbf{z}_t \mid y). \tag{34}$$

Implicit in the marginal density $p_t(\mathbf{z}_t)$ is a large mixture over all possible texts y,

$$p_t(\mathbf{z}_t) = \textstyle\sum_y p(y) p_t(\mathbf{z}_t \mid y). \tag{35}$$

Note the resemblance of Eq. 34 to our universal critic. For large v, the constant 1 becomes negligible and we are left with a density ratio between the target distribution and a large mixture distribution over alternative explanations. Indeed, Poole et al. (2023) found that SDS without CFG produced blurry 3D scenes and very large guidance weights worked best.

We therefore submit that the reason SDS works well is not explained by its ability to find modes in densities or its connections to model distillation techniques, but by its ability to approximate universal critics. Reinterpreting SDS in this way suggests new ways of overcoming its weaknesses (e.g., its tendency to produce oversaturated images), such as a more intentional design of the mixture of alternatives, or batched losses analogous to Eq. 19.

6. Discussion

In this position paper we have argued that the question of realism is equivalent to the question of randomness, that is, whether observations originated from a particular distribution. This allowed us to draw on insights from algorithmic information theory and to propose universal critics, or randomness deficiency (Li and Vitányi, 1997), as a rational answer. Perceptual quality can be viewed as the result of a (necessarily) imperfect approximation of universal critics. However, despite the relevance of these concepts to problems in machine learning, discussions of randomness deficiency are notably absent from its literature. Instead, dominant notions of realism continue to be based on probability (e.g., Ruff et al., 2021; Poole et al., 2023), typicality (e.g., Nalisnick et al., 2019b) or divergences (e.g., Blau and Michaeli, 2018; Theis and Wagner, 2021).

A divergence of zero is a sufficient condition for perfect realism but corresponds to an ideal observer with access to an infinite stream of examples. As such, it is stronger than required for most practical applications where observers only have access to one or a few examples. At the other end of the spectrum, weak typicality is an example of a criterion which only considers a necessary criterion, while most no-reference metrics correspond to neither a necessary nor a sufficient criterion (e.g., high probability in some feature space). Universal critics enable principled relaxations of divergence-based constraints. While weaker than divergences (in the desired way), they are still strong in the sense that they are as strong as other likelihood-ratio tests for realism, up to a constant which depends on the complexity of the competing test (Section 4.2).

Many interesting practical and theoretical questions remain. For example, what is the impact of different choices of π_n on the sample efficiency of universal critics? What are the implications of using universal critics in rate-distortion-realism trade-offs? Most importantly, what do practical approximations to universal critics (Eqs. 8 or 19) look like that can serve as optimization targets?

Acknowledgements

I would like to thank Aaron B. Wagner and Johannes Ballé for many helpful discussions shaping the arguments presented in this paper, Andriy Mnih, Matthias Bauer, Jörg Bornschein, Iryna Korshunova, Emilien Dupont, Eirikur Agustsson, and Alexandre Galashov for valuable feedback on the manuscript, and Daniel Severo for exploring various ideas to make universal critics practical.

Impact statement

This position paper presents ideas and arguments to broaden the understanding of realism in the machine learning community. The potential consequences of a better understanding of realism are difficult to predict, none of which we feel must be specifically highlighted here.

References

Paul H. Algoet and Thomas M. Cover. A Sandwich Proof of the Shannon-McMillan-Breiman Theorem. *The Annals of Probability*, 16:899—-909, 1988.

Dan Amir and Yair Weiss. Understanding and simplifying perceptual distances. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12226–12235, June 2021.

Johannes Ballé, Philip A. Chou, David Minnen, Saurabh Singh, Nick Johnston, Eirikur Agustsson, Sung Jin Hwang, and George Toderici. Nonlinear transform coding. *IEEE Journal of Selected Topics in Signal Processing*, 15(2):339–353, 2021. doi: 10.1109/JSTSP.2020.3034501.

J. Besag. Comments on 'Representations of knowledge in complex systems' by U. Grenander and M. I. Miller. *Journal of the Royal Statistical Society, Series B*, 56:591–592, 1994.

Y. Blau and T. Michaeli. The perception-distortion tradeoff. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6228–6237, 2018.

Y. Blau and T. Michaeli. Rethinking lossy compression: The rate-distortion-perception tradeoff. In *International Conference on Machine Learning*, 2019.

A. Borji. Pros and Cons of GAN Evaluation Measures. *Computer Vision and Image Understanding*, 179:41—65, 2019.

Marlène Careil, Matthew J Muckley, Jakob Verbeek, and Stéphane Lathuilière. Towards image compression with perfect realism at ultra-low bitrates. *arXiv preprint arXiv:2310.10325*, 2023.

Gregory. J. Chaitin. *Algorithmic Information Theory*. Cambridge Tracts in Theoretical Computer Science. Cambridge University Press, 1987.

Gregory J. Chaitin. *Exploring RANDOMNESS*. Discrete Mathematics and Theoretical Computer Science. Springer London, 2001. ISBN 9781852334178.

Jun Chen, Lei Yu, Jia Wang, Wuxian Shi, Yiqun Ge, and Wen Tong. On the rate-distortion-perception function.

IEEE Journal on Selected Areas in Information Theory, 3(4):664–673, 2022. doi: 10.1109/JSAIT.2022.3231820.

Hyunsun Choi, Eric Jang, and Alexander A. Alemi. WAIC, but Why? Generative Ensembles for Robust Anomaly Detection, 2019.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*, volume 2. John Wiley & Sons, 2006.

Emily L. Denton, Soumith Chintala, Arthur Szlam, and Rob Fergus. Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

Sander Dieleman. Musings on typicality, 2020.

K. Ding, K. Ma, S. Wang, and E. P. Simoncelli. Comparison of Full-Reference Image Quality Models for Optimization of Image Processing Systems. *International Journal of Computer Vision*, (129):1258–1281, 2021.

Gintare Karolina Dziugaite, Daniel M. Roy, and Zoubin Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, UAI'15, page 258–267, Arlington, Virginia, USA, 2015. AUAI Press. ISBN 9780996643108.

Shaojing Fan, Tian-Tsong Ng, Bryan Lee Koenig, Jonathan Samuel Herberg, Ming Jiang, Zhiqi Shen, and Qi Zhao. Image visual realism: From human perception to machine computation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(9):2180–2193, 2018. doi: 10.1109/TPAMI.2017.2747150.

Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3677–3686, 2020.

Pierre Glaser, Michael Arbel, and Arthur Gretton. KALE flow: A relaxed KL gradient flow for probabilities with disjoint support. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27, 2014.

Alexandros Graikos, Nikolay Malkin, Nebojsa Jojic, and Dimitris Samaras. Diffusion models as plug-and-play priors. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

Thomas Griffiths and Joshua Tenenbaum. From algorithmic to subjective randomness. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems*, volume 16. MIT Press, 2003.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2019.

Robert Herzog, Martin Čadík, Tunç O. Aydın, Kwawng In Kim, Karol Myszkowski, and Hans-Peter Seidel. NoRM: no-reference image quality metric for realistic image synthesis. *Computer Graphics Forum*, 31(2):545–554, 2012. doi: 10.1111/j.1467-8659.2012.03055.x.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

David C. Knill and Alexandre Pouget. The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, 27(12):712–719, 2004. ISSN 0166-2236. doi: https://doi.org/10.1016/j.tins.2004.10.007.

Charline Le Lan and Laurent Dinh. Perfect density models cannot guarantee anomaly detection. *Entropy*, 23(12), 2021. ISSN 1099-4300. doi: 10.3390/e23121690.

Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010.

C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and

W. Shi. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In *CVPR*, 2017.

Ming Li and Paul Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer, 1997.

Yujia Li, Kevin Swersky, and Rich Zemel. Generative moment matching networks. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1718–1727, Lille, France, 07–09 Jul 2015. PMLR.

Dimitra Maoutsa, Sebastian Reich, and Manfred Opper. Interacting Particle Solutions of Fokker–Planck Equations Through Gradient–Log–Density Estimation. *Entropy*, 22 (8), 2020. ISSN 1099-4300. doi: 10.3390/e22080802.

Per Martin-Löf. The definition of random sequences. *Information and Control*, 9(6):602–619, 1966. ISSN 0019-9958. doi: https://doi.org/10.1016/S0019-9958(66)80018-9.

Ryutaroh Matsumoto. Introducing the perception-distortion tradeoff into the rate-distortion theory of general information sources. *IEICE Communications Express*, advpub:2018XBL0109, 2018. doi: 10.1587/comex.2018XBL0109.

R. v. Mises. Grundlagen der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 5(1):52–99, 1919. doi: 10.1007/BF01203155.

A. Mittal, R. Soundararajan, and A. C. Bovik. Making a "Completely Blind" Image Quality Analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, March 2013. doi: 10.1109/LSP.2012.2227726.

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*, 2019a.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, and Balaji Lakshminarayanan. Detecting out-of-distribution inputs to deep generative models using typicality, 2019b.

Jerzy Neyman, Egon Sharpe Pearson, and Karl Pearson. IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933. doi: 10.1098/rsta.1933.0009.

X. L. Nguyen, M. J. Wainwright, and M. I. Jordan. On surrogate loss functions and f-divergences. *The Annals of Statistics*, 37(2):876 – 904, 2009. doi: 10.1214/08-AOS595.

XuanLong Nguyen, Martin J. Wainwright, and Michael I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010. doi: 10.1109/TIT.2010.2068870.

S. Nowozin, B. Cseke, and R. Tomioka. f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

G. Osada, T. Takahashi, B. Ahsan, and T. Nishide. Out-of-distribution detection with reconstruction error and typicality-based penalty. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5540–5552, Los Alamitos, CA, USA, jan 2023. IEEE Computer Society. doi: 10.1109/WACV56688.2023.00551.

Christopher Pondoc, Joseph C O'Brien, and Joseph Guman. Seeing through the facade: Understanding the realism, expressivity, and limitations of diffusion models. 2023.

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D Diffusion. In *The Eleventh International Conference on Learning Representations*, 2023.

Aaditya Ramdas, Sashank J. Reddi, Barnabás Póczos, Aarti Singh, and Larry Wasserman. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 3571–3577. AAAI Press, 2015. ISBN 0262511290.

Erik Reinhard, Alexei A. Efros, Jan Kautz, and Hans-Peter Seidel. On visual realism of synthesized imagery. *Proceedings of the IEEE*, 101(9):1998–2007, 2013. doi: 10.1109/JPROC.2013.2260711.

Alfréd Rényi. On Measures of Entropy and Information. In *The 4th Berkeley Symposium on Mathematics, Statistics and Probability*, page 547–561. University of California Press, 1961.

J. Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978. ISSN 0005-1098. doi: https://doi.org/10.1016/0005-1098(78)90005-5.

H. Robbins. An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 157–163, 1956.

Lukas Ruff, Jacob R. Kauffmann, Robert A. Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G. Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection.

Proceedings of the IEEE, 109(5):756–795, 2021. doi: 10.1109/JPROC.2021.3052449.

Boris Ryabko, Jaakko Astola, and Alex Gammerman. Application of kolmogorov complexity and universal codes to identity testing and nonparametric testing of serial independence for time series. *Theoretical Computer Science*, 359(1):440–448, 2006. ISSN 0304-3975. doi: https://doi.org/10.1016/j.tcs.2006.06.004. URL https://www.sciencedirect.com/science/article/pii/S0304397506003537.

Sadaf Salehkalaibar, Truong Buu Phan, Jun Chen, Wei Yu, and Ashish J Khisti. On the choice of perception loss function for learned video compression. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

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.

Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. DE-FAKE: Detection and Attribution of Fake Images Generated by Text-to-Image Generation Models, 2023.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, Lille, France, 07–09 Jul 2015. PMLR.

Ray J. Solomonoff. A preliminary report on a general theory of inductive inference. Technical report, Cambridge, MA, 1960.

R. M. Solovay. Lecture notes. 1975.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

C. Sønderby, J. Caballero, L. Theis, W. Shi, and F. Huszár. Amortised map inference for image super-resolution. In *International Conference on Learning Representations*, 2017.

L. Theis and E. Agustsson. On the advantages of stochastic encoders. In *Neural Compression Workshop at ICLR*, 2021.

L. Theis and A. B. Wagner. A coding theorem for the rate-distortion-perception function. In *Neural Compression Workshop at ICLR*, 2021.

L. Theis, A. van den Oord, and M. Bethge. A note on the evaluation of generative models. In *International Conference on Learning Representations*, Apr 2016.

Aaron van den Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George van den Driessche, Edward Lockhart, Luis Cobo, Florian Stimberg, Norman Casagrande, Dominik Grewe, Seb Noury, Sander Dieleman, Erich Elsen, Nal Kalchbrenner, Heiga Zen, Alex Graves, Helen King, Tom Walters, Dan Belov, and Demis Hassabis. Parallel WaveNet: Fast high-fidelity speech synthesis. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3918–3926. PMLR, 10–15 Jul 2018.

Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 681–688, Madison, WI, USA, 2011. Omnipress. ISBN 9781450306195.

Y. Yang, S. Mandt, and L. Theis. An introduction to neural data compression. *Foundations and Trends in Computer Graphics and Vision*, 15(2):113–200, 2023.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. *arXiv preprint arXiv:2311.18828*, 2023.

Lin Zhang, Lei Zhang, and Alan C. Bovik. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24(8):2579–2591, 2015. doi: 10.1109/TIP.2015.2426416.

Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. Deep Autoencoding Gaussian Mixture Model for Unsupervised Anomaly Detection. In *International Conference on Learning Representations*, 2018.

A. Typicality

In Appendix A.1 we extend our discussion of weak typicality. In particular, we elaborate on how a majority of examples in the typical set can be unrealistic. In Appendix A.2 we additionally consider *strong typicality*.

A.1. Bounded size of weakly typical sets

As discussed in the main text, the typical set contains sequences $\mathbf{x}^N \sim P^N$ with high probability, that is, A_δ^N is large enough that $P^N(A_\delta^N)$ approaches 1 as N goes to infinity. On the other hand, the typical set is small in the sense that the number of elements is bounded by (Cover and Thomas, 2006)

$$|A_\delta^N| \leq 2^{H[\mathbf{x}^N]+N\delta}. \tag{36}$$

This fact is exploited in information theory to build simple but efficient codes for data compression. Using

$$\log_2 |A_\delta^N| + 1 \leq H[\mathbf{x}^N] + N\delta + 1 \tag{37}$$

bits, we can address each element in the typical set. Normalized by the number of elements, this becomes

$$\frac{1}{N} H[\mathbf{x}^N] + \delta + \frac{1}{N} = H[\mathbf{x}_n] + \delta + \frac{1}{N}, \tag{38}$$

approaching the entropy of P as N increases and δ decreases. Counter-intuitively, this suggests that the typical set cannot contain too many unrealistic sequences, or else our compression scheme would be inefficient. However, note that while the coding rate overhead is only $(\delta + 1/N)$ above $H[\mathbf{x}_n]$ (Eq. 38), the number of elements in the typical set already exceeds $2^{H[\mathbf{x}]}$ by a factor of up to $2^{N\delta}$ (Eq. 36). If we relax the threshold δ so that $N\delta$ increases by 1, then this would increase the total coding cost of a sequence by only 1 bit, yet the number of elements in the typical set increases by a factor of up to 2.

A.2. Strong typicality

Here we consider strong typicality. A closely related notion (P-typicality) was considered by Chen et al. (2022) to quantify the realism of a batch of examples. Strong typicality is also similar in spirit to maximum mean discrepancy (MMD; Gretton et al., 2012), which was discussed in Section 3.2.

Let $\mathcal{X}$ be a finite set and let $\#(x, \mathbf{x}^N)$ be the number of occurrences of x in a sequence $\mathbf{x}^N = (x_1, \ldots, x_N)$, that is, a histogram. The set of *strongly typical* sequences is defined as (Cover and Thomas, 2006)

$$T_\delta^N = \left\{ \mathbf{x}^N \in \mathcal{X}^N : \sum_{x \in \mathcal{X}} \left| \frac{1}{N} \#(x, \mathbf{x}^N) - P(x) \right| < \delta \right\}. \tag{39}$$

As for weakly typical sets A_δ^N, the probability that a randomly drawn sequence $\mathbf{x}^N \sim P^N$ is strongly typical approaches 1 for any $\delta > 0$ as N increases. Strong typicality requires the empirical distribution of elements in a sequence to be close to the distribution of interest, P. For large N, a randomly selected element of a strongly typical sequence will appear like a sample from P, that is, it will appear realistic. However, the main challenge we are trying to overcome is to define realism for short sequences and individual x. If we naively apply strong typicality to a single element ($N = 1$), we obtain

$$\begin{aligned}
\textstyle\sum_{x \in \mathcal{X}} |\#(x, (x_1)) - P(x)| &= |1 - P(x_1)| + \textstyle\sum_{x \neq x_1} |0 - P(x)| && (40)\\
&= 1 - P(x_1) + \textstyle\sum_{x \neq x_1} P(x) && (41)\\
&= 1 - P(x_1) + 1 - P(x_1) && (42)\\
&= 2 - 2P(x_1), && (43)
\end{aligned}$$

that is, we are effectively back to measuring the probability of x_1. It is therefore unclear how strong typicality could be used to evaluate objects as high-dimensional as images. One might consider dividing an image into patches of lower dimensionality and treating the image as a sequence of these. However, this would ignore dependencies between patches and we would further have to assume that the statistics of each realistic image is representative of the entire distribution (i.e., ergodicity), which may not be the case.

B. Expected gradient of log-density

Let $P(n \mid \mathbf{x}) \propto \pi_n q_n(\mathbf{x})$ be the posterior probability that $\mathbf{x}$ was drawn from q_n. Then the expected gradient of the log-density is:

$$\begin{aligned}
\textstyle\sum_n P(n \mid \mathbf{x}) \nabla \log q_n(\mathbf{x}) &= \textstyle\sum_n P(n \mid \mathbf{x}) \frac{1}{q_n(\mathbf{x})} \nabla q_n(\mathbf{x}) && (44)\\
&= \textstyle\sum_n \frac{\pi_n q_n(\mathbf{x})}{\sum_m \pi_m q_m(\mathbf{x})} \frac{1}{q_n(\mathbf{x})} \nabla q_n(\mathbf{x}) && (45)\\
&= \textstyle\frac{1}{\sum_n \pi_n q(\mathbf{x})} \nabla \sum_n \pi_n q_n(\mathbf{x}) && (46)\\
&= \nabla \log \sum_n \pi_n q_n(\mathbf{x}) && (47)
\end{aligned}$$

C. Limited-memory observer

Here we elaborate on the relationship between the batched universal critic and an ideal observer in a sequential prediction task. Assume an observer assigns a value $T(\mathbf{x})$ to an image $\mathbf{x}$. Further assume we ask the observer to

$$\underset{T}{\text{maximize}} \quad \mathbb{E}_Q[T(\mathbf{x})] - \mathbb{E}_P[\exp(T(\mathbf{x}))], \tag{48}$$

that is, the observer receives a reward of $T(\mathbf{x})$ if $\mathbf{x} \sim Q$ and a penalty of $\exp(T(\mathbf{x}))$ if $\mathbf{x} \sim P$. The optimal output is then given by (Glaser et al., 2021)

$$T_Q(\mathbf{x}) = \log Q(\mathbf{x}) - \log P(\mathbf{x}). \tag{49}$$

Note that

$$\mathbb{E}_P[\exp(T_Q(\mathbf{x}))] = \mathbb{E}_P[Q(\mathbf{x})/P(\mathbf{x})] = \sum_{\mathbf{x}} Q(\mathbf{x}) = 1. \tag{50}$$

T_Q remains the optimal solution if we solve the following closely related constrained optimization problem,

$$\underset{T}{\text{maximize}} \quad \mathbb{E}_Q[T(\mathbf{x})] \quad \text{subject to} \quad \mathbb{E}_P[\exp(T(\mathbf{x}))] \leq 1, \tag{51}$$

and so we can use $\mathbb{E}_Q[T(\mathbf{x})]$ to evaluate T if we fix P and restrict the class of allowed T in this way. In other words, an equivalent task presents raters only with examples from Q, but applies restrictions to the scores that can be assigned.

Unlike typical classification problems where both P and Q are unknown and must be learned, we can assume P to be known to the observer through prior experience while Q still needs to be learned. A rational observer who expects $\mathbf{x}$ to be distributed according Q_n with probability π_n would maximize the expected reward by using

$$U(\mathbf{x}) = \log P(\mathbf{x}) - \log S(\mathbf{x}), \quad \text{where} \quad S(\mathbf{x}) = \textstyle\sum_n \pi_n Q_n(\mathbf{x}). \tag{52}$$

After receiving B examples from Q, $\mathbf{x}^B = (\mathbf{x}_1, \ldots, \mathbf{x}_B)$, a rational observer would update those beliefs to

$$\pi(n \mid \mathbf{x}^B) \propto \pi_n Q_n^B(\mathbf{x}^B) = \pi_n \textstyle\prod_{b=1}^B Q_n(\mathbf{x}_b) \tag{53}$$

and receive a reward of

$$U(\mathbf{x} \mid \mathbf{x}^B) = \log P(\mathbf{x}) - \log S(\mathbf{x} \mid \mathbf{x}^B), \quad \text{where} \quad S(\mathbf{x} \mid \mathbf{x}^B) = \textstyle\sum_n \pi(n \mid \mathbf{x}^B) Q_n(\mathbf{x}) \tag{54}$$

for a subsequent example $\mathbf{x}$ from the unknown Q. This score is slightly different from the batched universal critic, which is more readily interpreted as the combined value assigned to an entire batch of examples. However, the following relationship holds:

$$\begin{aligned}
U(\mathbf{x}^B) &= \log P^B(\mathbf{x}^B) - \log \textstyle\sum_n \pi_n Q_n^B(\mathbf{x}^B) && (55)\\
&= \log P^B(\mathbf{x}^B) - \log \textstyle\sum_n \pi_n Q_n^{B-1}(\mathbf{x}^{B-1}) Q_n(\mathbf{x}_B) && (56)\\
&= \log P^B(\mathbf{x}^B) - \log \textstyle\sum_n \frac{\pi_n Q_n^{B-1}(\mathbf{x}^{B-1})}{\sum_m \pi_m Q_m^{B-1}(\mathbf{x}^{B-1})} Q_n(\mathbf{x}_B) - \log \textstyle\sum_m \pi_m Q_m^{B-1}(\mathbf{x}^{B-1}) && (57)\\
&= \log P(\mathbf{x}_B) - \log \textstyle\sum_n \pi(n \mid \mathbf{x}^{B-1}) Q_n(\mathbf{x}_B) + \log P^{B-1}(\mathbf{x}^{B-1}) - \log \textstyle\sum_m \pi_m Q_m^{B-1}(\mathbf{x}^{B-1}) && (58)\\
&= U(\mathbf{x}^B \mid \mathbf{x}^{B-1}) + U(\mathbf{x}^{B-1}) && (59)\\
&= \textstyle\sum_{b=1}^B U(\mathbf{x}^b \mid \mathbf{x}^{b-1}) && (60)
\end{aligned}$$

That is, the output of the batched universal critic can be viewed as the sum of scores achieved in B sequential prediction tasks.