A NOTE ON THE EVALUATION OF GENERATIVE MODELS

Lucas Theis[*]
University of Tübingen
72072 Tübingen, Germany
lucas@bethgelab.org

Aäron van den Oord[*†]
Ghent University
9000 Ghent, Belgium
aaron.vandenoord@ugent.be

Matthias Bethge
University of Tübingen
72072 Tübingen, Germany
matthias@bethgelab.org

ABSTRACT

Probabilistic generative models can be used for compression, denoising, inpainting, texture synthesis, semi-supervised learning, unsupervised feature learning, and other tasks. Given this wide range of applications, it is not surprising that a lot of heterogeneity exists in the way these models are formulated, trained, and evaluated. As a consequence, direct comparison between models is often difficult. This article reviews mostly known but often underappreciated properties relating to the evaluation and interpretation of generative models with a focus on image models. In particular, we show that three of the currently most commonly used criteria—average log-likelihood, Parzen window estimates, and visual fidelity of samples—are largely independent of each other when the data is high-dimensional. Good performance with respect to one criterion therefore need not imply good performance with respect to the other criteria. Our results show that extrapolation from one criterion to another is not warranted and generative models need to be evaluated directly with respect to the application(s) they were intended for. In addition, we provide examples demonstrating that Parzen window estimates should generally be avoided.

1 INTRODUCTION

Generative models have many applications and can be evaluated in many ways. For density estimation and related tasks, log-likelihood (or equivalently Kullback-Leibler divergence) has been the de-facto standard for training and evaluating generative models. However, the likelihood of many interesting models is computationally intractable. For example, the normalization constant of unnormalized energy-based models is generally difficult to compute, and latent-variable models often require us to solve complex integrals to compute the likelihood. These models may still be trained with respect to a different objective that is more or less related to log-likelihood, such as contrastive divergence (Hinton, 2002), score matching (Hyvärinen, 2005), lower bounds on the log-likelihood (Bishop, 2006), noise-contrastive estimation (Gutmann & Hyvärinen, 2010), probability flow (Sohl-Dickstein et al., 2011), maximum mean discrepancy (MMD) (Gretton et al., 2007; Li et al., 2015), or approximations to the Jensen-Shannon divergence (JSD) (Goodfellow et al., 2014).

For computational reasons, generative models are also often compared in terms of properties more readily accessible than likelihood, even when the task is density estimation. Examples include visualizations of model samples, interpretations of model parameters (Hyvärinen et al., 2009), Parzen window estimates of the model's log-likelihood (Breuleux et al., 2009), and evaluations of model performance in surrogate tasks such as denoising or missing value imputation.

In this paper, we look at some of the implications of choosing certain training and evaluation criteria. We first show that training objectives such as JSD and MMD can result in very different optima than

---

[*]These authors contributed equally to this work.
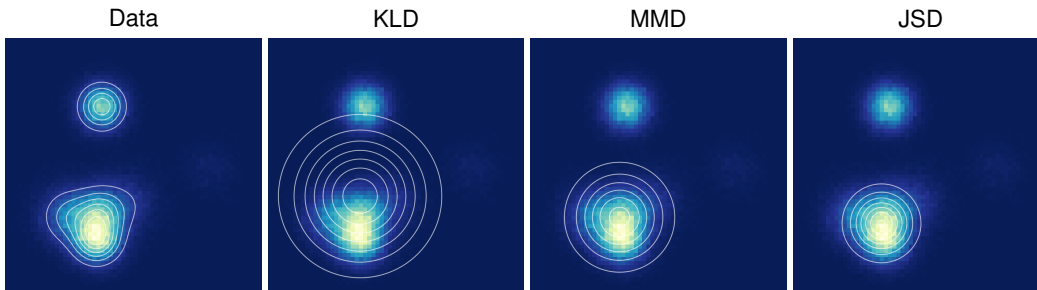[†]Now at Google DeepMind.

Figure 1: An isotropic Gaussian distribution was fit to data drawn from a mixture of Gaussians by either minimizing Kullback-Leibler divergence (KLD), maximum mean discrepancy (MMD), or Jensen-Shannon divergence (JSD). The different fits demonstrate different tradeoffs made by the three measures of distance between distributions.

log-likelihood. We then discuss the relationship between log-likelihood, classification performance, visual fidelity of samples and Parzen window estimates. We show that good or bad performance with respect to one metric is no guarantee of good or bad performance with respect to the other metrics. In particular, we show that the quality of samples is generally uninformative about the likelihood and vice versa, and that Parzen window estimates seem to favor models with neither good likelihood nor samples of highest possible quality. Using Parzen window estimates as a criterion, a simple model based on k-means outperforms the true distribution of the data.

2 TRAINING OF GENERATIVE MODELS

Many objective functions and training procedures have been proposed for optimizing generative models. The motivation for introducing new training methods is typically the wish to fit probabilistic models with computationally intractable likelihoods, rendering direct maximum likelihood learning impractical. Most of the available training procedures are consistent in the sense that if the data is drawn from a model distribution, then this model distribution will be optimal under the training objective in the limit of an infinite number of training examples. That is, if the model is correct, and for extremely large amounts of data, all of these methods will produce the same result. However, when there is a mismatch between the data distribution and the model, different objective functions can lead to very different results.

Figure 1 illustrates this on a simple toy example where an isotropic Gaussian distribution has been fit to a mixture of Gaussians by minimizing various measures of distance. Maximum mean discrepancy (MMD) has been used with generative moment matching networks (Li et al., 2015; Dziugaite et al., 2015) and Jensen-Shannon divergence (JSD) has connections to the objective function optimized by generative adversarial networks (Goodfellow et al., 2014) (see box for a definition). Minimizing MMD or JSD yields a Gaussian which fits one mode well, but which ignores other parts of the data. On the other hand, maximizing average log-likelihood or equivalently minimizing Kullback-Leibler divergence (KLD) avoids assigning extremely small probability to any data point but assigns a lot of probability mass to non-data regions.

Understanding the trade-offs between different measures is important for several reasons. First, different applications require different trade-offs, and we want to choose the right metric for a given application. Assigning sufficient probability to all plausible images is important for compression, but it may be enough to generate a single plausible example in certain image reconstruction applications (e.g., Hays & Efros, 2007). Second, a better understanding of the trade-offs allows us to better interpret and relate empirical findings. Generative image models are often assessed based on the visual fidelity of generated samples (e.g., Goodfellow et al., 2014; Gregor et al., 2015; Denton et al., 2015; Li et al., 2015). Figure 1 suggests that a model optimized with respect to KLD is more likely to produce atypical samples than the same model optimized with respect to one of the other two measures. That is, plausible samples—in the sense of having large density under the target

MMD (Gretton et al., 2007) is defined as,

$$\text{MMD}[p, q] = (\text{E}_{p,q}[k(\mathbf{x}, \mathbf{x}') - 2k(\mathbf{x}, \mathbf{y}) + k(\mathbf{y}, \mathbf{y}')])^{\frac{1}{2}}, \tag{1}$$

where $\mathbf{x}, \mathbf{x}'$ are indepent and distributed according to the data distribution p, and $\mathbf{y}, \mathbf{y}'$ are independently distributed according to the model distribution q. We followed the approach of Li et al. (2015), optimizing an empirical estimate of MMD and using a mixture of Gaussian kernels with various bandwidths for k.

JSD is defined as

$$\text{JSD}[p, q] = \frac{1}{2}\text{KLD}[p \,||\, m] + \frac{1}{2}\text{KLD}[q \,||\, m], \tag{2}$$

where $m = (p+q)/2$ is an equal mixture of distributions p and q. We optimized JSD directly using the data density, which is generally not possible in practice where we only have access to samples from the data distribution. In this case, generative adversarial networks (GANs) may be used to approximately optimize JSD, although in practical applications the objective function optimized by GANs can be very different from JSD. Parameters were initialized at the maximum likelihood solution in all cases, but the same optimum was consistently found using random initializations.

distribution—are not necessarily an indication of a good density model as measured by KLD, but may be expected when optimizing JSD.

3 EVALUATION OF GENERATIVE MODELS

Just as choosing the right training method is important for achieving good performance in a given application, so is choosing the right evaluation metric for drawing the right conclusions. In the following, we first continue to discuss the relationship between average log-likelihood and the visual appearance of model samples.

Model samples can be a useful diagnostic tool, often allowing us to build an intuition for why a model might fail and how it could be improved. However, qualitative as well as quantitative analyses based on model samples can be misleading about a model's density estimation performance, as well as the probabilistic model's performance in applications other than image synthesis. Below we summarize a few examples demonstrating this.

3.1 LOG-LIKELIHOOD

Average log-likelihood is widely considered as the default measure for quantifying generative image modeling performance. However, care needs to be taken to ensure that the numbers measured are meaningful. While natural images are typically stored using 8-bit integers, they are often modeled using densities, i.e., an image is treated as an instance of a continuous random variable. Since the discrete data distribution has differential entropy of negative infinity, this can lead to arbitrary high likelihoods even on test data. To avoid this case, it is becoming best practice to add real-valued noise to the integer pixel values to dequantize the data (e.g., Uria et al., 2013; van den Oord & Schrauwen, 2014; Theis & Bethge, 2015).

If we add the right amount of *uniform* noise, the log-likelihood of the continuous model on the dequantized data is closely related to the log-likelihood of a discrete model on the discrete data. Maximizing the log-likelihood on the continuous data also optimizes the log-likelihood of the discrete model on the original data. This can be seen as follows.

Consider images $\mathbf{x} \in \{0, ..., 255\}^D$ with a discrete probability distribution $P(\mathbf{x})$, uniform noise $\mathbf{u} \in [0, 1[^D$, and noisy data $\mathbf{y} = \mathbf{x} + \mathbf{u}$. If p refers to the noisy data density and q refers to the model density, then we have for the average log-likelihood:

$$\int p(\mathbf{y}) \log q(\mathbf{y}) \, d\mathbf{y} = \sum_{\mathbf{x}} P(\mathbf{x}) \int_{[0,1[^D} \log q(\mathbf{x} + \mathbf{u}) \, d\mathbf{u} \tag{3}$$

$$\leq \sum_{\mathbf{x}} P(\mathbf{x}) \log \int_{[0,1[^D} q(\mathbf{x} + \mathbf{u}) \, d\mathbf{u} \tag{4}$$

$$= \sum_{\mathbf{x}} P(\mathbf{x}) \log Q(\mathbf{x}), \tag{5}$$

where the second step follows from Jensen's inequality and we have defined

$$Q(\mathbf{x}) = \int_{[0,1[^D} q(\mathbf{x} + \mathbf{u}) \, d\mathbf{u} \tag{6}$$

for $\mathbf{x} \in \mathbb{Z}^D$. The left-hand side in Equation 3 is the expected log-likelihood which would be estimated in a typical benchmark. The right-hand side is the log-likelihood of the probability mass function Q on the original discrete-valued image data. The negative of this log-likelihood is equivalent to the average number of bits (assuming base-2 logarithm) required to losslessly compress the discrete data with an entropy coding scheme optimized for Q (Shannon, 2001).

SEMI-SUPERVISED LEARNING

A second motivation for using log-likelihood comes from semi-supervised learning. Consider a dataset consisting of images $\mathcal{X}$ and corresponding labels $\mathcal{Y}$ for some but not necessarily all of the images. In classification, we are interested in the prediction of a class label y for a previously unseen query image $\mathbf{x}$. For a given model relating $\mathbf{x}$, y, and parameters $\boldsymbol{\theta}$, the only correct way to infer the distribution over y—from a Bayesian point of view —is to integrate out the parameters (e.g., Lasserre et al., 2006),

$$p(y \mid \mathbf{x}, \mathcal{X}, \mathcal{Y}) = \int p(\boldsymbol{\theta} \mid \mathcal{X}, \mathcal{Y}) p(y \mid \mathbf{x}, \boldsymbol{\theta}) \, d\boldsymbol{\theta}. \tag{7}$$

With sufficient data and under certain assumptions, the above integral will be close to $p(y \mid \mathbf{x}, \hat{\boldsymbol{\theta}}_{\text{MAP}})$, where

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \text{argmax}_{\boldsymbol{\theta}} \, p(\boldsymbol{\theta} \mid \mathcal{X}, \mathcal{Y}) \tag{8}$$

$$= \text{argmax}_{\boldsymbol{\theta}} \, [\log p(\boldsymbol{\theta}) + \log p(\mathcal{X} \mid \boldsymbol{\theta}) + \log p(\mathcal{Y} \mid \mathcal{X}, \boldsymbol{\theta})]. \tag{9}$$

When no training labels are given, i.e., in the unsupervised setting, and for a uniform prior over parameters, it is therefore natural to try to optimize the log-likelihood, $\log p(\mathcal{X} \mid \boldsymbol{\theta})$.

In practice, this approach might fail because of a mismatch between the model and the data, because of an inability to solve Equation 9, or because of overfitting induced by the MAP approximation. These issues can be addressed by better image models (e.g., Kingma et al., 2014), better optimization and inference procedures, or a more Bayesian treatment of the parameters (e.g., Lacoste-Julien et al., 2011; Welling & Teh, 2011).

3.2 SAMPLES AND LOG-LIKELIHOOD

For many interesting models, average log-likelihood is difficult to evaluate or even approximate. For some of these models at least, generating samples is a lot easier. It would therefore be useful if we could use generated samples to infer something about a model's log-likelihood. This approach is also intuitive given that a model with zero KL divergence will produce perfect samples, and visual inspection can work well in low dimensions for assessing a model's fit to data. Unfortunately these intuitions can be misleading when the image dimensionality is high. A model can have poor log-likelihood and produce great samples, or have great log-likelihood and produce poor samples.

POOR LOG-LIKELIHOOD AND GREAT SAMPLES

A simple lookup table storing enough training images will generate convincing looking images but will have poor average log-likelihood on unseen test data. Somewhat more generally we might

consider a mixture of Gaussian distributions,

$$q(\mathbf{x}) = \frac{1}{N} \sum_n \mathcal{N}(\mathbf{x}; \mathbf{x}_n, \varepsilon^2 \mathbf{I}), \tag{10}$$

where the means $\mathbf{x}_n$ are either training images or a number of plausible images derived from the training set (e.g., using a set of image transformations). If ε is small enough such that the Gaussian noise becomes imperceptible, this model will generate great samples but will still have very poor log-likelihood. This shows that plausible samples are clearly *not sufficient* for a good log-likelihood.

Gerhard et al. (2013) empirically found a correlation between some models' log-likelihoods and their samples' ability to fool human observers into thinking they were extracted from real images. However, the image patches were small and all models used in the study were optimized to minimize KLD. The correlation between log-likelihood and sample quality may disappear, for example, when considering models optimized for different objective functions or already when considering a different set of models.

GREAT LOG-LIKELIHOOD AND POOR SAMPLES

Perhaps surprisingly, the ability to produce plausible samples is not only not sufficient, but also *not necessary* for high likelihood as a simple argument by van den Oord & Dambre (2015) shows: Assume p is the density of a model for d dimensional data $\mathbf{x}$ which performs arbitrarily well with respect to average log-likelihood and q corresponds to some bad model (e.g., white noise). Then samples generated by the mixture model

$$0.01 p(\mathbf{x}) + 0.99 q(\mathbf{x}) \tag{11}$$

will come from the poor model 99% of the time. Yet the log-likelihood per pixel will hardly change if d is large:

$$\log \left[0.01 p(\mathbf{x}) + 0.99 q(\mathbf{x}) \right] \geq \log \left[0.01 p(\mathbf{x}) \right] = \log p(\mathbf{x}) - \log 100 \tag{12}$$

For high-dimensional data, $\log p(\mathbf{x})$ will be proportional to d while $\log 100$ stays constant. For instance, already for the 32 by 32 images found in the CIFAR-10 dataset the difference between log-likelihoods of different models can be in the thousands, while $\log(100)$ is only about 4.61 nats (van den Oord & Dambre, 2015). This shows that a model can have large average log-likelihood but generate very poor samples.

GOOD LOG-LIKELIHOOD AND GREAT SAMPLES

Note that we could have also chosen q (Equation 11) such that it reproduces training examples, e.g., by choosing q as in Equation 10. In this case, the mixture model would generate samples indistinguishable from real images 99% of the time while the log-likelihood would again only change by at most 4.61 nats. This shows that any model can be turned into a model which produces realistic samples at little expense to its log-likelihood. Log-likelihood and visual appearance of samples are therefore largely independent.

3.3 SAMPLES AND APPLICATIONS

One might conclude that something must be wrong with log-likelihood if it does not care about a model's ability to generate plausible samples. However, note that the mixture model in Equation 11 might also still work very well in applications. While q is much more likely a priori, p is going to be much more likely a posteriori in tasks like inpainting, denoising, or classification. Consider prediction of a quantity y representing, for example, a class label or missing pixels. A model with joint distribution

$$0.01 p(\mathbf{x}) p(y \mid \mathbf{x}) + 0.99 q(\mathbf{x}) q(y \mid \mathbf{x}) \tag{13}$$

may again generate poor samples 99% of the time. For a given fixed $\mathbf{x}$, the posterior over y will be a mixture

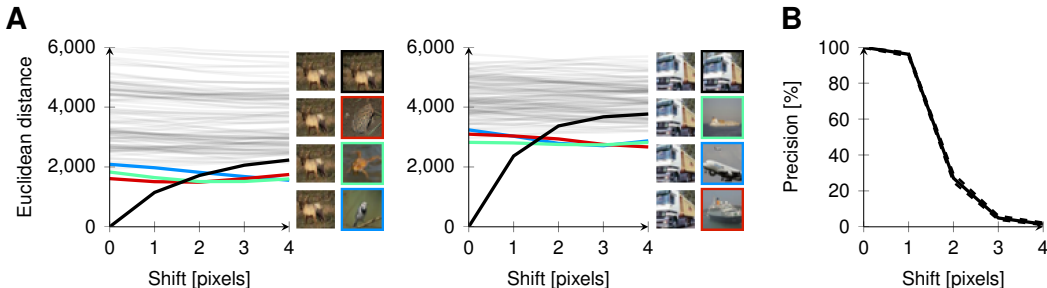$$\alpha p(y \mid \mathbf{x}) + (1 - \alpha) q(y \mid \mathbf{x}), \tag{14}$$

Figure 2: **A:** Two examples demonstrating that small changes of an image can lead to large changes in Euclidean distance affecting the choice of nearest neighbor. The images shown represent the query image shifted by between 1 and 4 pixels (left column, top to bottom), and the corresponding nearest neighbor from the training set (right column). The gray lines indicate Euclidean distance of the query image to 100 randomly picked images from the training set. **B:** Fraction of query images assigned to the correct training image. The average was estimated from 1,000 images. Dashed lines indicate a 90% confidence interval.

where a few simple calculations show that

$$\alpha = \sigma\left(\ln p(\mathbf{x}) - \ln q(\mathbf{x}) - \ln 99\right) \tag{15}$$

and σ is the sigmoidal logistic function. Since we assume that p is a good model, q is a poor model, and $\mathbf{x}$ is high-dimensional, we have

$$\ln p(\mathbf{x}) \gg \ln q(\mathbf{x}) + \ln 99 \tag{16}$$

and therefore $\alpha \approx 1$. That is, mixing with q has hardly changed the posterior over y. While the samples are dominated by q, the classification performance is dominated by p. This shows that high visual fidelity of samples is generally not necessary for achieving good performance in applications.

3.4 EVALUATION BASED ON SAMPLES AND NEAREST NEIGHBORS

A qualitative assessment based on samples can be biased towards models which overfit (Breuleux et al., 2009). To detect overfitting to the training data, it is common to show samples next to nearest neighbors from the training set. In the following, we highlight two limitations of this approach and argue that it is unfit to detect any but the starkest forms of overfitting.

Nearest neighbors are typically determined based on Euclidean distance. But already perceptually small changes can lead to large changes in Euclidean distance, as is well known in the psychophysics literature (e.g., Wang & Bovik, 2009). To illustrate this property, we used the top-left 28 by 28 pixels of each image from the 50,000 training images of the CIFAR-10 dataset. We then shifted this 28 by 28 window one pixel down and one pixel to the right and extracted another set of images. We repeated this 4 times, giving us 4 sets of images which are increasingly different from the training set. Figure 2A shows nearest neighbors of corresponding images from the query set. Although the images have hardly changed visually, a shift by only two pixels already caused a different nearest neighbor. The plot also shows Euclidean distances to 100 randomly picked images from the training set. Note that with a bigger dataset, a switch to a different nearest neighbor becomes more likely. Figure 2B shows the fraction of query images assigned to the correct training image in our example. A model which stores transformed training images can trivially pass the nearest-neighbor overfitting test. This problem can be alleviated by choosing nearest neighbors based on perceptual metrics, and by showing more than one nearest neighbor.

A second problem concerns the entropy of the model distribution and is harder to address. There are different ways a model can overfit. Even when overfitting, most models will not reproduce perfect or trivially transformed copies of the training data. In this case, no distance metric will find a close match in the training set. A model which overfits might still never generate a plausible image or might only be able to generate a small fraction of all plausible images (e.g., a model as in Equation 10 where instead of training images we store several transformed versions of the
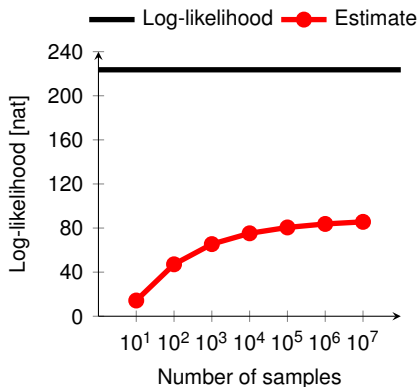
Figure 3: Parzen window estimates for a Gaussian evaluated on 6 by 6 pixel image patches from the CIFAR-10 dataset. Even for small patches and a very large number of samples, the Parzen window estimate is far from the true log-likelihood.

Table 1: Using Parzen window estimates to evaluate various models trained on MNIST, samples from the true distribution perform worse than samples from a simple model trained with k-means.

Model	Parzen est. [nat]
Stacked CAE	121
DBN	138
GMMN	147
Deep GSN	214
Diffusion	220
GAN	225
True distribution	**243**
GMMN + AE	282
k-means	313

training images, or a model which only describes data in a lower-dimensional subspace). Because the number of images we can process is vanishingly small compared to the vast number of possible images, we would not be able to detect this by looking at samples from the model.

3.5 EVALUATION BASED ON PARZEN WINDOW ESTIMATES

When log-likelihoods are unavailable, a common alternative is to use Parzen window estimates. Here, samples are generated from the model and used to construct a tractable model, typically a kernel density estimator with Gaussian kernel. A test log-likelihood is then evaluated under this model and used as a proxy for the true model's log-likelihood (Breuleux et al., 2009). Breuleux et al. (2009) suggested to fit the Parzen windows on both samples and training data, and to use at least as many samples as there are images in the training set. Following Bengio et al. (2013a), Parzen windows are in practice commonly fit to only 10,000 samples (e.g., Bengio et al., 2013b; Goodfellow et al., 2014; Li et al., 2015; Sohl-Dickstein et al., 2015). But even for a large number of samples Parzen window estimates generally do not come close to a model's true log-likelihood when the data dimensionality is high. In Figure 3 we plot Parzen window estimates for a multivariate Gaussian distribution fit to small CIFAR-10 image patches (of size 6 by 6). We added uniform noise to the data (as explained in Section 3.1) and rescaled between 0 and 1. As we can see, a completely infeasible number of samples would be needed to get close to the actual log-likelihood even for this small scale example. For higher dimensional data this effect would only be more pronounced.

While the Parzen window estimate may be far removed from a model's true log-likelihood, one could still hope that it produces a similar or otherwise useful ranking when applied to different models. Counter to this idea, Parzen window estimates of the likelihood have been observed to produce rankings different from other estimates (Bachman & Precup, 2015). More worryingly, a GMMN+AE (Li et al., 2015) is assigned a higher score than images from the training set (which are samples from the true distribution) when evaluated on MNIST (Table 1). Furthermore it is relatively easy to exploit the Parzen window loss function to achieve even better results. To illustrate this, we fitted 10,000 centroids to the training data using k-means. We then generated 10,000 independent samples by sampling centroids with replacement. Note that this corresponds to the model in Equation 10, where the standard deviation of the Gaussian noise is zero and instead of training examples we use the centroids. We find that samples from this k-means based model are assigned a higher score than any other model, while its actual log-likelihood would be $-\infty$.

7

4 CONCLUSION

We have discussed the optimization and evaluation of generative image models. Different metrics can lead to different trade-offs, and different evaluations favor different models. It is therefore important that training and evaluation match the target application. Furthermore, we should be cautious not to take good performance in one application as evidence of good performance in another application.

An evaluation based on samples is biased towards models which overfit and therefore a poor indicator of a good density model in a log-likelihood sense, which favors models with large entropy. Conversely, a high likelihood does not guarantee visually pleasing samples. Samples can take on arbitrary form only a few bits from the optimum. It is therefore unsurprising that other approaches than density estimation are much more effective for image synthesis (Portilla & Simoncelli, 2000; Dosovitskiy et al., 2015; Gatys et al., 2015). Samples are in general also an unreliable proxy for a model's performance in applications such as classification or inpainting, as discussed in Section 3.3.

A subjective evaluation based on visual fidelity of samples is still clearly appropriate when the goal is image synthesis. Such an analysis at least has the property that the data distribution will perform very well in this task. This cannot be said about Parzen window estimates, where the data distribution performs worse than much less desirable models[1]. We therefore argue Parzen window estimates should be avoided for evaluating generative models, unless the application specifically requires such a loss function. In this case, we have shown that a k-means based model can perform better than the true density. To summarize, our results demonstrate that for generative models there is no one-fits-all loss function but a proper assessment of model performance is only possible in the the context of an application.

REFERENCES

Bachman, P. and Precup, D. Variational Generative Stochastic Networks with Collaborative Shaping. *Proceedings of the 32nd International Conference on Machine Learning*, pp. 1964–1972, 2015.

Bengio, Y., Mesnil, G., Dauphin, Y., and Rifai, S. Better mixing via deep representations. In *Proceedings of the 30th International Conference on Machine Learning*, 2013a.

Bengio, Y., Thibodeau-Laufer, E., Alain, G., and Yosinski, J. Deep generative stochastic networks trainable by backprop, 2013b. arXiv:1306.1091.

Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer, 2006.

Breuleux, O., Bengio, Y., and Vincent, P. Unlearning for better mixing. Technical report, Universite de Montreal, 2009.

Denton, E., Chintala, S., Szlam, A., and Fergus, R. Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks. *arXiv.org*, 2015.

Dosovitskiy, A., Springenberg, J. T., and Brox, T. Learning to Generate Chairs with Convolutional Neural Networks. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2015.

Dziugaite, G. K., Roy, D. M., and Ghahramani, Z. Training generative neural networks via maximum mean discrepancy optimization, 2015. arXiv:1505.0390.

Gatys, L. A., Ecker, A. S., and Bethge, M. Texture synthesis and the controlled generation of natural stimuli using convolutional neural networks, 2015. arXiv:1505.07376.

---

[1]In decision theory, such a metric is called an *improper scoring function.*

Gerhard, H. E., Wichmann, F. A., and Bethge, M. How sensitive is the human visual system to the local statistics of natural images? *PLoS Computational Biology*, 9(1), 2013.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, 2014.

Gregor, K., Danihelka, I., Graves, A., and Wierstra, D. DRAW: A recurrent neural network for image generation. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

Gretton, A., Borgwardt, K. M., Rasch, M., Schölkopf, B., and Smola, A. J. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems 20*, 2007.

Gutmann, M. and Hyvärinen, A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, 2010.

Hays, J. and Efros, A. A. Scene completion using millions of photographs. *ACM Transactions on Graphics (SIGGRAPH)*, 26, 2007.

Hinton, G. E. Training Products of Experts by Minimizing Contrastive Divergence. *Neural Computation*, 14(8):1771–1800, 2002.

Hyvärinen, A., Hurri, J., and Hoyer, P. O. *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision*. Springer, 2009.

Hyvärinen, A. Estimation of non-normalized statistical models using score matching. *Journal of Machine Learning Research*, pp. 695–709, 2005.

Kingma, D. P., Rezende, D. J., Mohamed, S., and Welling, M. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems 27*, 2014.

Lacoste-Julien, S., Huszar, F., and Ghahramani, Z. Approximate inference for the loss-calibrated Bayesian. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, 2011.

Lasserre, J. A., Bishop, C. M., and Minka, T. P. Principled hybrids of generative and discriminative models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2006.

Li, Y., Swersky, K., and Zemel, R. Generative moment matching networks. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

Portilla, J. and Simoncelli, E. P. A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, 40:49–70, 2000.

Shannon, C. E. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

Sohl-Dickstein, J., Battaglino, P., and DeWeese, M. R. Minimum Probability Flow Learning. In *Proceedings of the 28th International Conference on Machine Learning*, 2011.

Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

Theis, L. and Bethge, M. Generative Image Modeling Using Spatial LSTMs. In *Advances in Neural Information Processing Systems 28*, 2015.

Uria, B., Murray, I., and Larochelle, H. RNADE: The real-valued neural autoregressive density-estimator. In *Advances in Neural Information Processing Systems 26*, 2013.

van den Oord, A. and Dambre, J. Locally-connected transformations for deep GMMs, 2015. Deep Learning Workshop, ICML.

van den Oord, A. and Schrauwen, B. Factoring Variations in Natural Images with Deep Gaussian Mixture Models. In *Advances in Neural Information Processing Systems 27*, 2014.

Wang, Z. and Bovik, A. C. Mean squared error: Love it or leave it? *IEEE Signal Processing Magazine*, 2009.

Welling, M. and Teh, Y. W. Bayesian Learning via Stochastic Gradient Langevin Dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, 2011.