INFORMATION THEORETIC GENERATIVE MODELING

Anonymous authors - Paper under double-blind review

1 SUMMARY

In this article we use rate-distortion theory, a branch of information theory devoted to the problem of lossy compression introduced by Claude Shannon in 1959 [1], to shed light on an important problem in latent variable modeling of data: is there room to improve the model? One way to address this question is to find an upper bound on the probability (equivalently a lower bound on the negative log likelihood) that the model can assign to some data as one varies the prior and/or the likelihood function in a latent variable model. The core of our contribution is to formally show that the problem of optimizing priors in latent variable models is exactly an instance of the variational optimization problem that information theorists solve when computing rate-distortion functions, and then to use this to derive a lower bound on negative log likelihood. Moreover, we will show that if changing the prior can improve the log likelihood, then there is a way to change the likelihood function instead and attain the same log likelihood, and thus rate-distortion theory is of relevance to both optimizing priors as well as optimizing likelihood functions. The result we present here runs much deeper than the particular modeling problem being solved - in formally connecting the latent variable modeling problem to rate-distortion theory, we have established a bridge where decades of work on either field can now be considered for possible cross-pollination; notably in a subsequent article we intend to ask the question next of whether practical algorithms in data compression can be used to design latent variables. We will experimentally argue for the usefulness of quantities derived from rate-distortion theory in latent variable modeling by applying them to a problem in image modeling.

1.1 ELABORATION

A statistician plans to use a latent variable generative model

$$p(x) = \int p(z)\ell(x|z)dz, \tag{1}$$

where $p(z)$ is known as the prior over the latent variables, and $\ell(x|z)$ is the likelihood of the data given the latent variables. Frequently, both the prior and the likelihood are parametrized and the statistician's job is to find reasonable parametric families for both - an optimization algorithm then chooses the parameter within those families. The task of designing these parametric families can sometimes be time consuming.

In this article we ask the question of how much $p(z)$ can be improved if one fixes $\ell(x|z)$ and viceversa, with the goal of equipping the statistician with tools to make decisions on where to invest her time. One way to answer whether $p(z)$ can be improved for fixed $\ell(x|z)$ is to drop the assumption that $p(z)$ must belong to a particular parametric family and ask how a model could improve this way. Mathematically, given data $\{x_1, \cdots, x_N\}$ the first problem we study is the following optimization problem: for a fixed $\ell(x|z)$,

$$\min_{p(z)} -\frac{1}{N}\sum_{i=1}^{N} \log \int p(z)\ell(x_i|z)dz \tag{2}$$

which as we will show, is also connected to the problem of determining if $\ell(x|z)$ can be improved for a given fixed $p(z)$. The quantity being optimized in (2) is called the average negative log likelihood of the data, and is used whenever one assumes that the data $\{x_1, \cdots, x_N\}$ have been drawn statistically independently at random.

Obviously, for any given $\ell(x|z), p(z)$, from the definition (1) we have the trivial upper bound

$$\min_{p(z)} -\frac{1}{N}\sum_{i=1}^{N} \log \int p(z)\ell(x_i|z)dz \leq -\frac{1}{N}\sum_{i=1}^{N} \log p(x_i) \tag{3}$$

The question we ask here is, can we give a good *lower bound*? A lower bound could tell us how far we can improve the model by changing the prior. The answer turns out to be in the affirmative. By connecting (2) to the computation of rate-distortion functions [2], which is how information theorists study the problem of lossy compression, we will show that

$$\min_{p(z)} -\frac{1}{N}\sum_{i=1}^{N} \log \int p(z)\ell(x_i|z)dz \geq -\frac{1}{N}\sum_{i=1}^{N} \log p(x_i) - \sup_{z} \log\left(\frac{1}{N}\sum_{i=1}^{N} \frac{\ell(x_i|z)}{p(x_i)}\right). \tag{4}$$

This result is very general - it holds for both discrete and continuous latent variable spaces, scalar or vector. It is also *sharp* - if you plug in the right prior, the upper and lower bounds match. It also has the advantage that the lower bound is written as a function of the trivial upper bound (3) - if someone proposes a latent variable model $p(x)$ which uses a likelihood function $\ell(x|z)$, the optimal negative log likelihood value when we optimize the prior is thus known to be within a gap of

$$\sup_z \log \left(\frac{1}{N} \sum_{i=1}^{N} \frac{\ell(x_i|z)}{p(x_i)} \right) \qquad \text{bits.} \tag{5}$$

bits.

More frequently than not, the problem is not to improve a prior, but for a fixed prior, to improve the likelihood function. Interestingly, rate-distortion theory still is relevant to this problem, although the question that we are able to answer with it is smaller in scope. Through a simple change of variable argument, we will argue that if the negative log likelihood can be improved by modifying the prior, exactly the same negative log likelihood can be attained by modifying the likelihood function instead. Thus if rate-distortion theory predicts that there is scope for improvement for a prior, the same holds for the likelihood function but conversely, while rate-distortion theory can precisely determine when it is that a prior can no longer be improved, the same cannot be said for the likelihood function.

In our presentation, we will demonstrate that statistics derived from this lower bound are actually very effective at predicting whether various generative models for priors and likelihood functions of ever increasing complexities [3], [4] can be improved further for a variety of image modeling data sets (MNIST, OMNIGLOT, Caltech 101 Silhouettes, Frey Faces, Histopathology and CIFAR).

We stress that we are not first in noticing that there are relations between rate-distortion theory and the general field of latent variable modeling. The Information Bottleneck method of [5] is a preeminent example of a successful idea that exists in this boundary, having created a subfield of research that remains relevant nowadays, [6] [7]. The autoencoder concept extensively used in the neural network community is arguably directly motivated by the encoder/decoder concepts in lossy/lossless compression. Recently, [8] exploited the β-VAE loss function [9] to explicitly introduce a trade-off between rate and distortion in the latent variable modeling problem, where the notions of rate and distortion have similarities to those used in this article.

In contrast to earlier works, our contribution is more basic - it goes to directly tying formally the most elementary rate-distortion setup, introduced in 1959 by Claude Shannon, to the problem of latent variable modeling which to our knowledge had not been done. The field of rate-distortion theory of course has not been idle since Shannon introduced it, and we think it that there are ideas within it, beyond those exposed in this article, that could be adapted to good use in the problem of latent variable modeling.

REFERENCES

[1] C. E. Shannon. Coding theorems for a discrete source with a fidelity criterion. In *IRE Nat. Conv. Rec., Pt. 4*, pages 142–163. 1959.

[2] T. Berger. *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Prentice-Hall electrical engineering series. Prentice-Hall, 1971.

[3] Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel Recurrent Neural Networks. In *Proceedings of the 33rd International Conference on Machine Learning, New York, NY, USA*. 2016.

[4] Jakub M. Tomczak and Max Welling. VAE with a VampPrior. In *The 21nd International Conference on Artificial Intelligence and Statistics*. 2018.

[5] N. Tishby, F. Pereira, and W. Bialek. The information bottleneck method. In *Proceedings of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377. 1999.

[6] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop*. 2015.

[7] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *CoRR*, abs/1703.00810, 2017.

[8] Alexander A. Alemi, Ben Poole, Ian Fischer, Joshua V. Dillon, Rif A. Saurous, and Kevin Murphy. An information-theoretic analysis of deep latent-variable models. *CoRR*, abs/1711.00464, 2017.

[9] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. Beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *The International Conference on Learning Representations (ICLR), Toulon*. 2017.