The continuous Bernoulli: fixing a pervasive error in variational autoencoders

Gabriel Loaiza-Ganem and John P. Cunningham
Department of Statistics, Columbia University

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Introduction

- Variational autoencoders (VAE [KW14]) are a central tool in probabilistic modeling.
- MNIST is the most widely used testbed for VAE improvements.
- While MNIST is not binary, Bernoulli likelihoods are commonly used.
- We introduce a new distribution: the continuous Bernoulli.
- We interpret the practice as ignoring a normalizing constant.

Contributions of the paper

- Introducing and characterizing the continuous Bernoulli.
- Showing it outperforms the Bernoulli, even when the data is close to binary.
- Showcasing the importance of correctly modeling data type.

Variational autoencoders

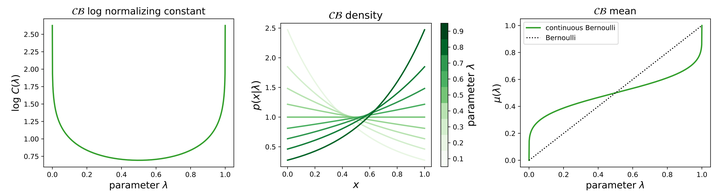
- Inference is performed on the model:
$$Z_n \sim p_0(z) \text{ and } X_n|Z_n \sim p_\theta(x|z_n)$$
by maximizing the ELBO with respect to the generative θ and approximate posterior parameters, θ and ϕ:
$$\mathcal{E}(\theta,\phi) = \sum_{n=1}^{N} E_{q_\phi}[\log p_\theta(x_n|z)] - KL(q_\phi||p_0)$$

The pervasive error in Bernoulli VAE

- In the Bernoulli case where $p_\theta(x|z) = \mathcal{B}(\lambda_\theta(z))$ the reconstruction term is given by:
$$E_{q_\phi}\Big[\sum_{d=1}^{D} x_{n,d}\log\lambda_{\theta,d}(z_n) + (1-x_{n,d})\log(1-\lambda_{\theta,d}(z_n))\Big]$$
- This loss is often used with real-valued data in $[0,1]$.

$\mathcal{CB}$: the continuous Bernoulli

- To study the effect of this practice, we introduce the $[0,1]$-supported continuous Bernoulli distribution, where $\tilde{p}(\cdot|\lambda)$ denotes the Bernoulli distribution:
$$X \sim \mathcal{CB}(\lambda) \iff p(x|\lambda) \propto \tilde{p}(x|\lambda) = \lambda^x(1-\lambda)^{1-x}$$
- The continuous Bernoulli is well defined and its pdf and expected value are given by:
$$p(x|\lambda) = C(\lambda)\lambda^x(1-\lambda)^{1-x}, \text{where } C(\lambda) = \begin{cases} \dfrac{2\tanh^{-1}(1-2\lambda)}{1-2\lambda} & \text{if } \lambda \neq 0.5 \\ 2 & \text{otherwise} \end{cases}$$
$$\mu(\lambda) := E[X] = \begin{cases} \dfrac{\lambda}{2\lambda-1} + \dfrac{1}{2\tanh^{-1}(1-2\lambda)} & \text{if } \lambda \neq 0.5 \\ 0.5 & \text{otherwise} \end{cases}$$



- We fully characterize the continuous Bernoulli distribution.
- The shape of $C(\lambda)$ reveals that ignoring it hurts performance the most near the extrema.

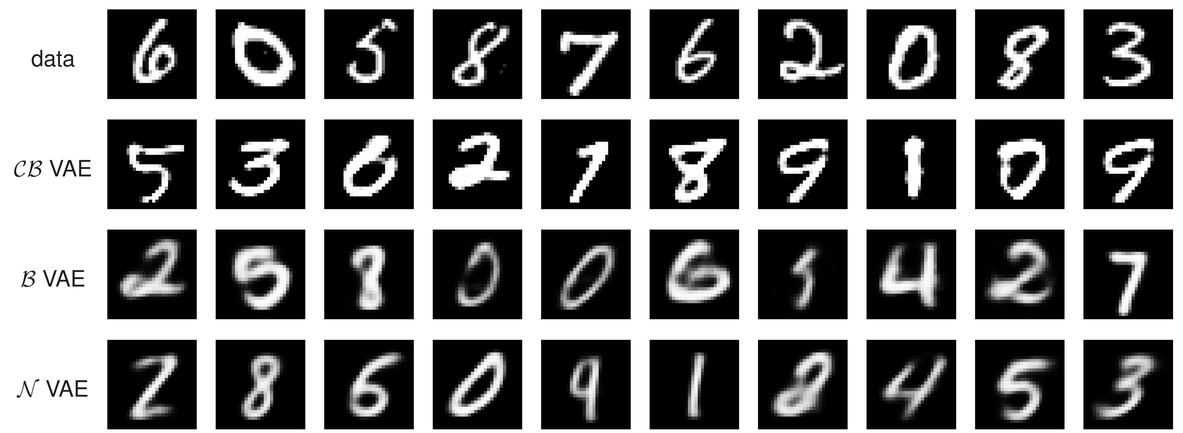
The continuous Bernoulli VAE

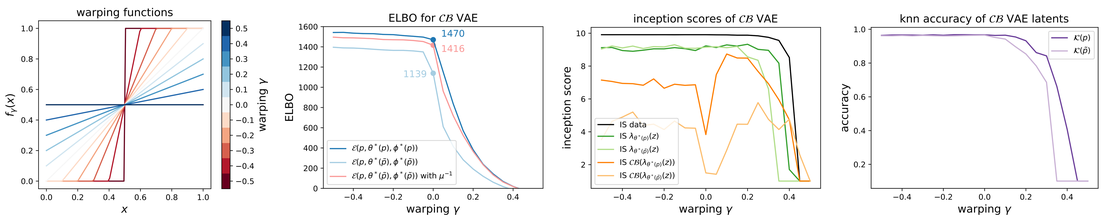
- When using a continuous Bernoulli the ELBO becomes:
$$\mathcal{E}(p,\theta,\phi) = \sum_{n=1}^{N} -KL(q_\phi||p_0) + E_{q_\phi}\Big[\sum_{d=1}^{D} x_{n,d}\log\lambda_{\theta,d}(z_n) + (1-x_{n,d})\log(1-\lambda_{\theta,d}(z_n)) + \log C(\lambda_{\theta,d}(z_n))\Big]$$
$$\mathcal{E}(\tilde{p},\theta,\phi) = \sum_{n=1}^{N} -KL(q_\phi||p_0) + E_{q_\phi}\Big[\sum_{d=1}^{D} x_{n,d}\log\lambda_{\theta,d}(z_n) + (1-x_{n,d})\log(1-\lambda_{\theta,d}(z_n))\Big]$$
- We find most explanations of using a Bernoulli likelihood lacking.
- We show that $\mathcal{E}(\tilde{p},\theta,\phi) < \mathcal{E}(p,\theta,\phi)$.

Conclusions

- Besides allowing the study of using Bernoulli likelihoods with real-valued data, the continuous Bernoulli outperforms the Bernoulli in a variety of metrics.
- Careful consideration of data type matters.

Experiments





distribution	objective	map	$\mathcal{E}(p,\theta^*,\phi^*)$	IS w/ samples	IS w/ parameters	$\mathcal{K}(\phi^*)$
$\mathcal{CB}/\mathcal{B}$	$\mathcal{E}(p,\theta,\phi)$	·	**1007**	1.15	2.31	**0.43**
	$\mathcal{E}(\tilde{p},\theta,\phi)$	μ^{-1}	916	**1.49**	**4.55**	0.42
	$\mathcal{E}(\tilde{p},\theta,\phi)$	·	475	1.41	1.39	0.42
Gaussian	$\mathcal{E}(p,\theta,\phi)$	·	**1891**	**1.86**	**3.04**	**0.42**
	$\mathcal{E}(\tilde{p},\theta,\phi)$	·	-42411	1.24	1.00	0.1
beta	$\mathcal{E}(p,\theta,\phi)$	·	**3121**	**2.98**	**4.07**	**0.47**
	$\mathcal{E}(\tilde{p},\theta,\phi)$	·	-38913	1.39	1.00	0.1

Acknowledgments and References

We thank Yixin Wang, Aaron Schein, Andy Miller, and Keyon Vafa for helpful conversations, and the Simons Foundation, Sloan Foundation, McKnight Endowment Fund, NIH NINDS 5R01NS100066, NSF 1707398, and the Gatsby Charitable Foundation for support.

[KW14] Diederik P Kingma and Max Welling, *Auto-encoding variational bayes*, International Conference on Learning Representations, 2014.

Email: gl2480@columbia.edu