Paper 7283 | The continuous Bernoulli: fixing a pervasive error in variational autoencoders

We sincerely thank the reviewers for helping us improve this work, as well as for finding it "inspiring" and "very well written". We will make all minor changes requested, and here we focus on the major points of feedback.

- **More attention/detail to continuous Bernoulli vs beta (raised by all reviewers)**: Thank you, this is important. There are two main points; we will add a thorough discussion of them, and new results, to the paper.
  1. As a device for study: the continuous Bernoulli is the natural and necessary distribution to understand the impact of erroneously using a Bernoulli likelihood on $[0, 1]$-valued data, thus providing critical insight into one main and important question raised in this work.
  2. Concentration of mass at the extrema: the notably different behavior of beta and continuous Bernoulli near the extrema warrant observation. Denoting $p(x|\lambda)$ and $p(x|\alpha, \beta)$ as continuous Bernoulli and beta densities, respectively, we can show that $\lim_{x\to 0} p(x|\lambda)/p(x|\alpha, \beta) \to 0$ if $\alpha < 1$, and $\lim_{x\to 0} p(x|\lambda)/p(x|\alpha, \beta) \to \infty$ if $\alpha > 1$; a finite positive limit is only achieved if $\alpha = 1$. (An analogous result holds as $x \to 1$ depending on the sign of β, but we focus on a neighborhood of 0 here.) Next, for a given $p(x|\lambda)$, denote $p(x|\alpha_\lambda, \beta_\lambda)$ as the corresponding beta distribution with matching mean and variance. Then, $\lambda < 0.5$ implies $\alpha_\lambda < 1$, which in other words shows that, for comparable moments, the beta distribution places much more mass in a neighborhood of 0 (see Figure 1 below). Note that this does not imply that the continuous Bernoulli is not placing most of its mass around 0, just not as much as the beta. This key insight highlights that even "similar looking" continuous Bernoullis and betas behave considerably differently at the extrema, which is precisely the most important part of the densities when modeling almost binary data (such as MNIST, which while almost binary, does have grayscale pixels). Empirically, we find that the beta distributed VAE produces means that are *less* extremal (i.e. grayer images, as seen in the appendix materials), which implies that each beta was shifted away from 0 and 1 to reduce adding too much mass to the extrema (precisely the effect of Figure 1).
- **State of the art architectures (reviewer 2)**: To further show the meaningful performance gains of the continuous Bernoulli, we have run additional experiments with a normalizing flow architecture in the approximate posterior (Rezende and Mohammed, 2015). Figure 2 (below) shows that not only does the continuous Bernoulli still outperform the Bernoulli, but the gap is even larger than in simpler architectures. For CIFAR-10 the beta distribution outperforms the continuous Bernoulli, which we suspect to be because beta can concentrate mass anywhere in RGB space. However, even in this case, the continuous Bernoulli outperforms the Bernoulli in a non-marginal way: again, ignoring normalizing constants significantly hurts performance. This gives an exciting first answer to this comment, and we will include more experiments with more complicated architectures on CIFAR-10 by publication time. Of course, the advantages of the continuous Bernoulli for data with values close to the extrema might suggest future distributions more suitable for CIFAR-10, which we will also explore.
- **More analysis of the continuous Bernoulli distribution (reviewer 3)**: As you pointed out, the continuous Bernoulli distribution has a closed form moment generating function; thank you. We have computed its characteristic function, entropy, and the KL divergence between two continuous Bernoullis, all of which also have closed form expressions (omitted here due to space constraints, but we will include them in the paper). These additions will make for a very thorough characterization of our distribution.
- **Why μ^{-1} does not equate the Bernoulli to the continuous Bernoulli (reviewer 2)**: In a maximum likelihood setting with no latent variables, μ^{-1} does offer an equivalence. While one might expect this to fully carry over in a setting with latent variables, that is incorrect; the full argument for which is presented in section 4.5 of the paper. Even still, using μ^{-1} still recovers some amount of the lost performance. Is this finding a fundamental suboptimality of the Bernoulli model, or simply an artifact of approximate inference (training with the ELBO)? We have run the following additional experiment: we use the EM algorithm to estimate the parameters of a mixture of continuous Bernoullis (loosely, the VAE can be thought of as an infinite-component extension) on simulated data. Figure 3 (below) shows that correcting with μ^{-1} still does not fully correct for using a Bernoulli likelihood (as measured by distance between the estimated distribution and the ground truth). This is thus a fundamental modeling issue.

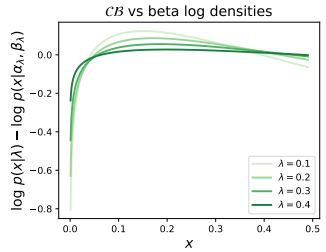

Figure 1

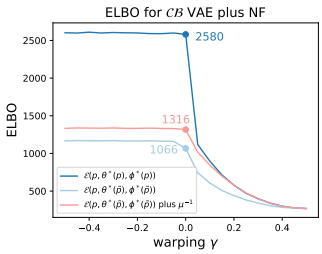

Figure 2

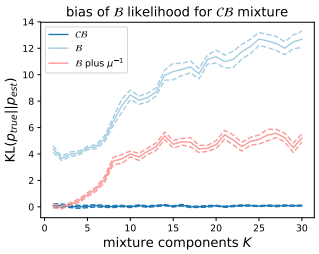

Figure 3