Automatic Differentiation Variational Inference with Mixtures

Warren R. Morningstar
Google Research

Sharad M. Vikram
Google Research

Cusuh Ham
Georgia Institute of Technology

Andrew G. Gallagher
Google Research

Joshua V. Dillon
Google Research

Abstract

Automatic Differentiation Variational Inference (ADVI) is a useful tool for efficiently learning probabilistic models in machine learning. Traditionally, approximate posteriors learned by ADVI are forced to be unimodal in order to facilitate use of the reparameterization trick. In this paper, we show how stratified sampling may be used to enable mixture distributions as the approximate posterior, and derive a new lower bound on the evidence analogous to the importance weighted autoencoder (IWAE). We show that this "SIWAE" is a tighter bound than both IWAE and the traditional ELBO, both of which are special instances of this bound. We verify empirically that the traditional ELBO objective disfavors the presence of multimodal posterior distributions and may therefore not be able to fully capture structure in the latent space. Our experiments show that using the SIWAE objective allows the encoder to learn more complex distributions which contain multimodality, resulting in higher accuracy, better calibration, and improved generative model performance in the presence of incomplete, limited, or corrupted data.

1 Introduction

Variational inference has become a powerful tool for Bayesian modeling using deep neural networks, with successes including image generation [Kingma et al., 2014], classification [Alemi et al., 2017], uncertainty quantification [Snoek et al., 2019] and outlier detection [Bishop, 1993, Nalisnick et al., 2018]. Much of the recent success in variational inference have been driven by the relative ease of fitting models using ADVI, where small numbers of samples can be used for individual forward passes through a model, and noisy but unbiased gradients can be determined using the reparameterization trick, allowing the use of backpropagation in training and enabling traditional stochastic gradient methods [Rezende et al., 2014, Kingma et al., 2014]. Currently, one major limitation of ADVI is that it is only possible if the posterior distribution is reparameterizable. This has to date forced ADVI methods to utilize a limited set of possible distributions. While there have been developments in extending reparameterization to broader classes of distributions [e.g., gamma and beta distributions; Ruiz et al., 2016], multimodal distributions have remained elusive.

This paper explores using ADVI with mixture posterior distributions. Mixture distributions present an advantage over unimodal distributions due to their flexibility [Bishop et al., 1998, West, 1993]. The contributions of this paper are as follows:

1. We propose the SIWAE, a new lower bound on the evidence for the specific case of a mixture variational posterior. When applicable, the SIWAE is tighter than the evidence or importance-weighted evidence lower bounds.

2. We demonstrate on toy problems that SIWAE is better suited to approximate a known multimodal posterior distribution than the traditional ELBO or the score function estimator.

3. We empirically show that models trained using the traditional ELBO objective often fail to discover multimodality in the latent space even if mixtures are used for the posterior. We also show

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s).

that SIWAE allows models to more easily infer multimodality when it exists.

4. We demonstrate that models trained with SIWAE achieve higher classification accuracy and better model calibration than ELBO using incomplete feature information.

2 Approach

Consider a simple latent variable model with a single observed data point x and corresponding latent variable z along with a prior distribution $r(z)$ and likelihood $p(x|z)$. In probabilistic modeling, we are interested in the posterior distribution $p(z|x)$, but generally, computing the posterior analytically is intractable. Variational inference is a strategy that reframes Bayesian inference as an optimization problem by first introducing a surrogate variational posterior $q_\phi(z|x)$, where ϕ are free parameters, and then maximizing the evidence lower bound (ELBO) with respect to ϕ. The ELBO is defined as,

$$\mathcal{L}_{\text{ELBO}}(\phi) \triangleq \mathbb{E}_{q_\phi(z|x)}\left[\log p(x|z)\right] - D_{\text{KL}}(q_\phi(z|x), r(z)) \tag{1}$$

and is a lower bound on the marginal probability of the data $\log p(x)$ [Jordan et al., 1999]. In ADVI, we aim to compute $\nabla_\phi \mathcal{L}(\phi)$, but computing the exact ELBO is generally analytically intractable. Both terms in $\mathcal{L}(\phi)$ are expectations over $q_\phi(z|x)$, so we approximate the gradient by first drawing samples from $q_\phi(z|x)$ and computing the gradient of a Monte-Carlo approximation of the ELBO, i.e., for a single sample $z' \sim q_\phi(z|x)$, we see that $\mathcal{L}_{\text{ELBO}}(\phi) \approx \log p(x|z') - \log q_\phi(z'|x) + \log r(z')$.

When computing the gradient, ADVI differentiates through the sampling procedure itself, utilizing the *reparameterization trick* [Kingma et al., 2014, Rezende et al., 2014]. The reparameterization trick expresses sampling a random variable z from its distribution as a transformation of noise drawn from a base distribution $\epsilon \sim p(\epsilon)$, where the transformation is a deterministic function of the parameters of the sampling distribution ϕ. In ADVI, we are restricted to "reparameterizable" posterior distributions – distributions whose sampling procedure can be expressed in this way. Although there has been notable work in growing this class of distributions, such as in Figurnov et al. [2018] and Jankowiak and Obermeyer [2018], the choice of posterior in ADVI remains limited.

In this paper, we consider mixture posteriors for ADVI, specifically mixtures whose component distributions are reparameterizable. Mixture distributions are a powerful class of posteriors, as growing the number of components can make them arbitrarily expressive, but are challenging to use as posteriors in ADVI as sampling from a mixture is not naively reparameterizable due to the discrete categorical variable that is sampled to assign a data point to a mixture component. As seen in [Roeder et al., 2017], *stratified sampling* can address this issue. In stratified sampling, we compute expectations by sampling evenly over component distribuions ("strata") and averaging using the weights of each stratum. For a mixture distribution, the natural stratification is each of the mixture component distributions. Rather than initially drawing an assignment and then drawing a sample from the corresponding component distribution, we draw one sample from each component individually and compute a weighted average over the samples. Formally, for any continuous and differentiable function $f(z)$ and mixture distribution $q(z) \triangleq \sum_{k=1}^K \alpha_k q_k(z)$, where α_k are the mixture weights and $q_k(z)$ are the components, we can compute the expectation $\mathbb{E}_{q(z)} f(z)$ as follows:

$$\begin{aligned}\mathbb{E}_{q(z)} f(z) &= \int f(z) \left(\sum_{k=1}^K \alpha_k q_k(z)\right) dz \\ &= \sum_{k=1}^K \alpha_k \int f(z) q_k(z) dz \\ &= \sum_{k=1}^K \alpha_k \mathbb{E}_{q_k(z)}\left[f(z)\right]\end{aligned} \tag{2}$$

By pulling the sum over the mixture components outside of the integral over z and sampling from each of the K mixture components, we are able to compute the expectation using the reparameterization trick, so long as the component distributions from the mixture are themselves reparameterizable. Returning to ADVI, when the posterior $q_\phi(z|x)$ is a mixture distribution with weights $\{\alpha_{k,\phi}(x)\}_{k=1}^K$ and components $\{q_{k,\phi}(z|x)\}_{k=1}^K$, we can compute the "stratified ELBO," or SELBO:

$$\mathcal{L}_{\text{SELBO}}(\phi) \triangleq \sum_{k=1}^K \alpha_{k,\phi}(x) \mathbb{E}_{q_{k,\phi}(z|x)}\left[\log \frac{p(x|z_k) r(z_k)}{q_\phi(z_k|x)}\right]$$

While SELBO is technically the same objective as the ELBO but specialized to mixtures, we draw this distinction to imply that we are drawing K reparameterizable samples to compute a differentiable, Monte-Carlo estimate of the SELBO whereas the traditional ELBO formulation implies we take a single sample to compute a non-differentiable estimate. While this is an increase in compute budget (needing to draw K samples v.s. just one), we are now able to compute gradients w.r.t. more expressive variational posterior.

2.1 A tighter bound for mixture posteriors

While the SELBO objective allows us to fit a mixture posterior using ADVI, it falls prey to the same issues that make fitting multimodal distributions with the ELBO difficult, namely the ELBO's mode-seeking behavior. Furthermore this mode-seeking behavior actively works against the goal of learning a multimodal posterior. Consider fitting a multimodal variational distribution $q(z)$ to a multimodal distribution $p(z)$ using the ELBO. Since maximizing the ELBO corresponds to minimizing $D_{\mathrm{KL}}(q(z), p(z))$, the ELBO only meaningfully depends on regions where $q(z)$ has significant mass. This results in the well-known phenomenon where a $q(z)$ fit via variational inference only captures one of the modes of the target distribution $p(z)$.

Now consider optimizing the SELBO, which is a weighted average of the ELBO for each mixture component, with ADVI. For components that produce low ELBO values, gradients of SELBO will downweight those components, potentially all the way to 0. Since the ELBO is content with $q(z)$ fitting just a single mode of $p(z)$, there is no reason for the components to ever be upweighted, resulting in mode collapse. Unless the variational distribution is initialized perfectly (i.e. it has significant density at each of the true posterior's modes), we argue that ADVI on SELBO will collapse mixture components and learn an overly conservative approximate posterior.

To combat this harmful exploration penalty, we can use importance sampling. An importance-weighted estimate of the ELBO first draws T i.i.d. samples from the posterior $\{z_t\}_{t=1}^T \sim q_\phi(z|x)$, computing a lower bound using the ratio of the densities of a sample under the joint distribution and posterior (i.e., importance weights) for each sample (called "IWAE" in Burda et al. [2015]):

$$\mathcal{L}_{\mathrm{IWAE}}^T(\phi) \triangleq \mathbb{E}_{\{z_t \sim q_\phi(z|x)\}_{t=1}^T} \left[\log \frac{1}{T} \sum_{t=1}^{T} \frac{p(x|z_t) r(z_t)}{q_\phi(z_t|x)} \right]$$

Burda et al. [2015] shows that if the importance weights are bounded, then as T increases the IWAE grows tighter and approaches $\log p(x)$ as $T \to \infty$. Unlike the regular ELBO, the posterior in the IWAE is less penalized for generating samples that are unlikely.

Our main contribution is a novel importance-weighted estimator for the ELBO when using mixture posteriors. To incorporate importance sampling into the SELBO, we first draw T samples from each of the mixture components, $\{z_{kt}\}_{k=1,t=1}^{K,T}$. We then compute importance weights that are themselves weighted by the mixture weights, arriving at the "stratified IWAE," or SIWAE:

$$\mathcal{L}_{\mathrm{SIWAE}}^T(\phi) \triangleq \mathbb{E}_{\{z_{kt} \sim q_{k,\phi}(z|x)\}_{k=1,t=1}^{K,T}} \Bigg[\log \frac{1}{T} \sum_{t=1}^{T} \sum_{k=1}^{K} \alpha_{k,\phi}(x) \frac{p(x|z_{kt}) r(z_{kt})}{q_\phi(z_{kt}|x)} \Bigg]$$

```
def siwae(prior, likelihood, posterior, x, T):
  q = posterior(x)
  z = q.components_dist.sample(T)
  z = tf.transpose(z, perm=[2, 0, 1, 3])
  loss_n = tf.math.reduce_logsumexp(
    (- tf.math.log(T) + tf.math.log_softmax(
        mixture_dist.logits)[:, None, :]
     + prior.log_prior(z) + likelihood(z).
         log_prob(x) - q.log_prob(z)),
    axis=[0, 1])
  return tf.math.reduce_mean(loss_n, axis=0)
```

Figure 1: TF Probability implementation of SIWAE loss for local latent variable models (e.g., VAE).

Intuitively, in SIWAE, "bad" samples contribute less to the objective thanks to importance weighting. Therefore, the "bad" components responsible for those samples will contribute less to the objective, and thus while the gradients w.r.t. their parameters might be smaller, their mixture weights will not collapse to zero. We thus conjecture that SIWAE encourages mixture components to increase their variance and spread their mass in more regions of the true posterior. This is a desirable property to avoid component collapse when fitting mixture distributions using ADVI and it better allows components to explore distinct modes.

By repeated application of Jensen's equality, we demonstrate that $\mathcal{L}_{\mathrm{SIWAE}}^T$ is a valid lower bound that is tighter than $\mathcal{L}_{\mathrm{IWAE}}^T$ when $K > 1$ (see theorems and proofs in Appendix A). $\mathcal{L}_{\mathrm{SIWAE}}$ is also equivalent to $\mathcal{L}_{\mathrm{IWAE}}^T$ and $\mathcal{L}_{\mathrm{SELBO}}$ under certain circumstances ($K = 1$ and $K = T = 1$, respectively). Because $\mathcal{L}_{\mathrm{IWAE}}$ is tighter than $\mathcal{L}_{\mathrm{SELBO}}$ even when $T = 1$, $\mathcal{L}_{\mathrm{SIWAE}}$ is also tighter than $\mathcal{L}_{\mathrm{SELBO}}$. Furthermore the importance sampling step enables higher-variance posteriors, as it mitigates the penalty for low-likelihood samples. Consequently, the implicit posterior [Cremer et al., 2017] (defined by importance sampling the learned posterior) can better capture different modes. Furthermore, SELBO and SIWAE are both easy to implement and are simple augmentations of existing variational inference code. See Figure 1 for a code snippet in TensorFlow [Abadi et al., 2016] which evaluates the SIWAE for a latent variable model.

3 Related Work

Salimans and Knowles [2013] and Kingma et al. [2014]

show that sampling from a distribution can be reparameterized as a deterministic function of the parameters of the distribution and some auxiliary variable, thereby facilitating the propagation of gradients through the distribution. They also introduce the Variational Auto Encoder (VAE), which uses an amortized variational posterior for a deep generative model. Burda et al. [2015] showed that the bound on the evidence could be tightened using importance sampling, and that the tightness of the bound was improved by the number of importance samples. Cremer et al. [2017] suggest that the IWAE can be viewed as fitting the traditional ELBO, but using a richer latent space distribution defined by the importance-reweighting of samples from the posterior, and further explore the functional forms of these implicit posteriors.

While our work explores mixtures for the variational posterior, we'd like to draw distinctions from other works that have studied the use of (trainable) mixtures for the *prior*. Dilokthanakul et al. [2016], Johnson et al. [2016], Jiang et al. [2017] introduce a VAE which uses a learnable mixture of Gaussians to represent the prior distribution of a latent variable. Learning a mixture prior does not require differentiating through the prior's sampling procedure as draws ADVI draws samples from the posterior, not prior. Dilokthanakul et al. [2016] and Jiang et al. [2017] find that their models achieve competitive performance on unsupervised clustering, with the mixture components learning clusters that approximate the different classes present in the data. Similarly, Tomczak and Welling [2017] use a mixture of Gaussians trained on learnable pseudo-inputs as the prior, which allows them to introduce greater flexibility in the latent space distribution. They find that their generative performance improves on a number of benchmarks using this procedure. While using a mixture distribution as a prior enables modeling global structure in the latent space, it does not explicitly model ambiguity or competing explanations for a single observation. The uses of mixture distributions for either the prior or posterior are orthogonal and complementary, and a mixture distribution in either part of the model is a valid option.

Domke and Sheldon [2019] propose to use alternative sampling schemes (including stratified) from a uniform distribution defined over a state space, along with a coupling transformation to the latent space in order to design a sampling scheme which results in better coverage of the approximating posterior distribution. They also show that the divergence of this approximation from the true posterior is bounded by the looseness of the evidence bound.

When using mixture distributions as the posterior, the typical strategy is to fix component weights [Oh et al., 2019], or by using a continuous relaxation (e.g., the concrete relaxation of the categorical distribution [Poduval et al., 2020]). Graves [2016] proposes an algorithm that allows for gradients to be backpropagated through the mixture distribution when the component distribution have diagonal covariances by composing the sampling procedure as a recursion over the dimensions. Our method only requires that the component distributions is subject to reparameterization, and therefore can be used with a wider class of distributions. Furthermore it does not require explicit specification of the gradient updates to be hard-coded, making it easy to integrate mixtures into existing models. Roeder et al. [2017] derives a pathwise gradient extension to the SELBO that lowers the variance of gradient estimates, but still suffers from the mode-seeking properties of the SELBO.

4 Experimental Results

In this section, we aim to demonstrate that SIWAE not only works as conjectured, i.e. it enables capturing distinct modes where SELBO does not, but also that SIWAE can improve the quality and calibration of posterior distributions on large scale tasks. First we evaluate SIWAE against a suite of baselines on two synthetic examples that have explicitly multimodal posteriors. We then evaluate SIWAE on an image generative modeling task and an image classification task, comparing to both non-probabilistic models, and unimodal posteriors, showing that multimodality enables more expressive and better calibrated posteriors. We also provide additional experiments exploring extensions to the those in the main body of the paper in Appendix C.6 and Appendix C.7 .

4.1 Toy Problems

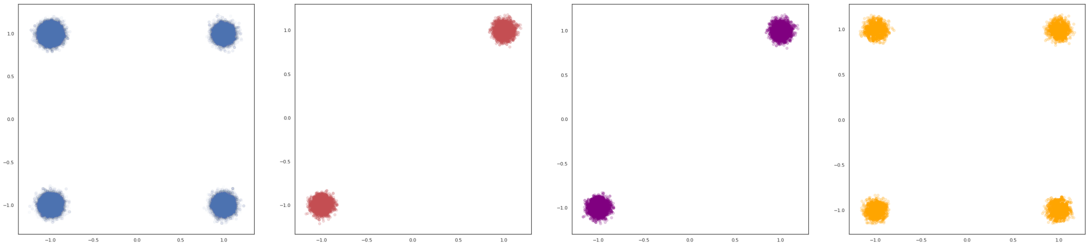
We construct two illustrative examples, each designed to have a multimodal ground-truth posterior distribution. Our goal is to demonstrate situations in which not only are multimodal variational posteriors necessary, but that good performance is dependent on being trained with SIWAE as opposed to SELBO.

4.1.1 Generative Model

We define a latent variable model where the true posterior is multimodal by construction, with the hope of recovering the distinct modes. Specifically, we sample 1000 datapoints from the following two-dimensional generative model:

$$z \sim \mathcal{N}(0, I) \qquad x \sim \mathcal{N}\left(|z|, \sigma^2 I\right)$$

where $\sigma^2 = 0.005$, i.e., we first sample a latent z from an isotropic normal, but observe $|z|$ with some Gaussian



(a) True posterior

(b) SELBO posterior

(c) Score function posterior

(d) SIWAE posterior

Figure 2: We sample the true posterior along with each of the learned posteriors for the observed data point $(1, 1)$. We see that the SELBO- and score function-trained posteriors are unable to capture all 4 modes of the true posterior.

noise. For an observed x, there are 4 distinct modes in z-space that could have generated it, since z is two-dimensional. We initialize the variational posterior $q_\phi(z|x)$ as a multilayer perceptron (MLP) with 2 layers of 100 hidden units that outputs a 4-component mixture of Gaussians distribution. We evaluate three different estimators of the ELBO: (1) SELBO, (2) SIWAE, and (3) a score function estimator as a baseline. We fit the posterior for 1000 epochs, with a batch size of 32 and using the Adam [Kingma and Ba, 2014] optimizer with a learning rate of 0.001, using 10 importance samples for SIWAE and 100 for both SELBO and score function. Each baseline was initialized and trained identically (same initial weights and order of batches).

We measure performance using a 10^6-sample SIWAE estimate, and observe that the SIWAE-trained estimator achieves the highest value of -1.505, compared to -2.024 and -2.038 from the SELBO and score function estimators, respectively. Investigating further, we plot samples from each of the posteriors in the latent space. We find that in many cases, the SELBO and score function posteriors are unable to capture the four distinct modes (see Figure 2), whereas the higher-variance SIWAE posterior is able to cover the modes successfully. We also observe similar results to those found in Rainforth et al. [2018], where tighter variational bounds result in lower signal-to-noise ratios in the gradients to the posterior. This is reflected by on-average higher-variance gradients while training a SIWAE posterior vs. a SELBO posterior (1.16 vs. 0.48 average elementwise variance, respectively). However, the score function estimator has significantly higher empirical variance (261.4) than that of both SIWAE and SELBO, indicating that the variance reduction coming from the use of the reparameterization trick offsets the additional variance from a tighter variational bound. We also found that using the "sticking-the-landing" (stl) estimator [Roeder et al., 2017] (Figure B.2, Figure B.3) does not significantly improve the SELBO or SIWAE in the toy experiment.

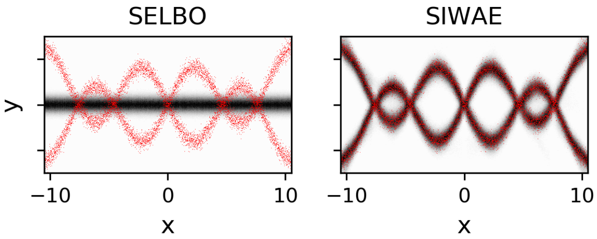


Figure 3: Red points show the observed data. The black histogram shows the posterior predictive distribution. The left panel shows a model trained with SELBO, and the right panel shows a model trained with SIWAE.

4.1.2 Predictive Model

We also consider a predictive task where the latent variable must contain multimodality in order to produce accurate predictions. Specifically, we sample data from the following one dimensional model:

$$\begin{aligned} x &\sim \mathsf{Uniform}(-10, 10) \\ z &\sim \mathsf{Rademacher}() \\ m &\sim \mathsf{Normal}(\mu(x), \sigma_1^2) \\ y &\sim \mathsf{Normal}\left(zm, \sigma_2^2\right) \end{aligned}$$

where $\mu(x) = 7\sin(3x/4) + x/2$ is a sinusoidal function. We set $\sigma_1^2 = 0.1$, and $\sigma_2 = 0.9$ so that the total variance of data generated from a single mode is 1.

We set up our model in a Variational Information Bottleneck (VIB) architecture [Alemi et al., 2017], a variant of the VAE in which the decoder predicts a distribution over an output y which is assumed to be different from the input x. We use a 2 dimensional encoding, and a mixture of $K = 10$ multivariate normal distributions for our posterior. We use a single Normal Distribution for the prior. For the decoder, we use a single affine layer to predict the means of the likelihood, and fix the variance to 1, which corresponds to the observed empirical variance for a given mode. Additional details of the training procedure can be found in the supplement.

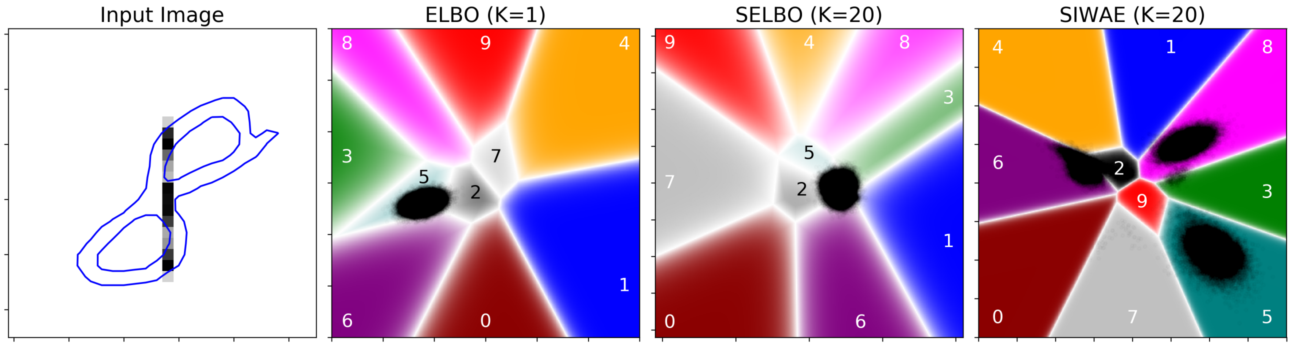


Figure 4: An illustration of the difference between the latent spaces learned by mixtures versus those found using unimodal variational posterior distributions. The left panel shows the input to the network, for which all but the center-most column has been discarded. The contour shows the true image of the data. The second column shows a model trained using a unimodal latent space distribution and optimizing the ELBO. The third column shows the latent space learned using the SELBO objective with 20 mixture components. On the right, we show the latent space for the same example found using the SIWAE objective. The latent space is colored by the predicted class from that position in the latent space, and the transparency of that color indicates the confidence of the predicted class relative to the second most probable class.

In Figure 3, we show the data generated from 10k samples from the model alongside the posterior predictive distribution trained using either SIWAE or SELBO. The difference is striking: SELBO appears to make meaningless predictions, while SIWAE makes nearly perfect predictions. We attribute this to a harmful property of SELBO, namely the fact that SELBO requires that all of the samples from the posterior predict the output data equally well (by weighting them equally in the computation of the loss). This specifically predisposes models against learning multimodal posteriors which offer multiple competing (and distinct) explanations for the observed data, for which all but one will always be a “bad” explanation. In contrast, SIWAE offers penalties relative to the best samples from the predictive distribution, softening the blow against modes which do not explain a single instance of data and thereby facilitating exploration in the model. This exploration is aided by different training examples which in this case have similar input features but different observed outcomes.

4.2 Single Column MNIST Classification

To evaluate SIWAE's efficacy on a more challenging problem, we trained a classifier on the benchmark dataset MNIST [LeCun et al., 1998]. We again use a VIB model for our classifier. To induce multimodality in this problem, we give the model incomplete information about the input. In particular, Doersch [2016] shows that training a VAE using only the centermost column of the image introduces multimodality into the dataset that is difficult to capture using a unimodal encoder. We replicate this multimodality in the classification setting by taking the centermost column of each training image. An example of a corrupted input can be seen in Figure 4. In general, it can be difficult even for a human to correctly classify the image given this type of corruption. In this scenario, we look for not only accurate predictions but also well-calibrated uncertainty for those predictions.

In the middle two columns of Figure 4, we visualize samples from the posterior of a single validation set example learned by optimizing the ELBO/SELBO objective. We find that, while SELBO enables the use of multiple mixture components in the variational posterior distribution, the model only learns a unimodal representation of the latent variable. This is a direct consequence of the ELBO objective, which disincentivizes exploration and encourages mode-seeking in the variational posterior. In this case, we observe the posterior “hedging its bets,” where the single mode sits across several decision boundaries. These decision boundaries are also quite wide, suggesting that the model is using variance in the decoder as a source of uncertainty. We find this behavior undesirable, and show later that it negatively affects how well calibrated the model is.

The rightmost column of Figure 4 shows the latent space learned by optimizing the SIWAE objective. In stark contrast to models trained with SELBO, we find that SIWAE learns posteriors that have many active and distinct modes. This implies that rather than “‘hedging its bets” as in the SELBO, a SIWAE-trained posterior offers multiple competing explanations, moving the uncertainty in the final prediction into the latent space rather than the output of the decoder. This can be directly seen by looking at the lightness of the background colors in Figure 4, which indicate the

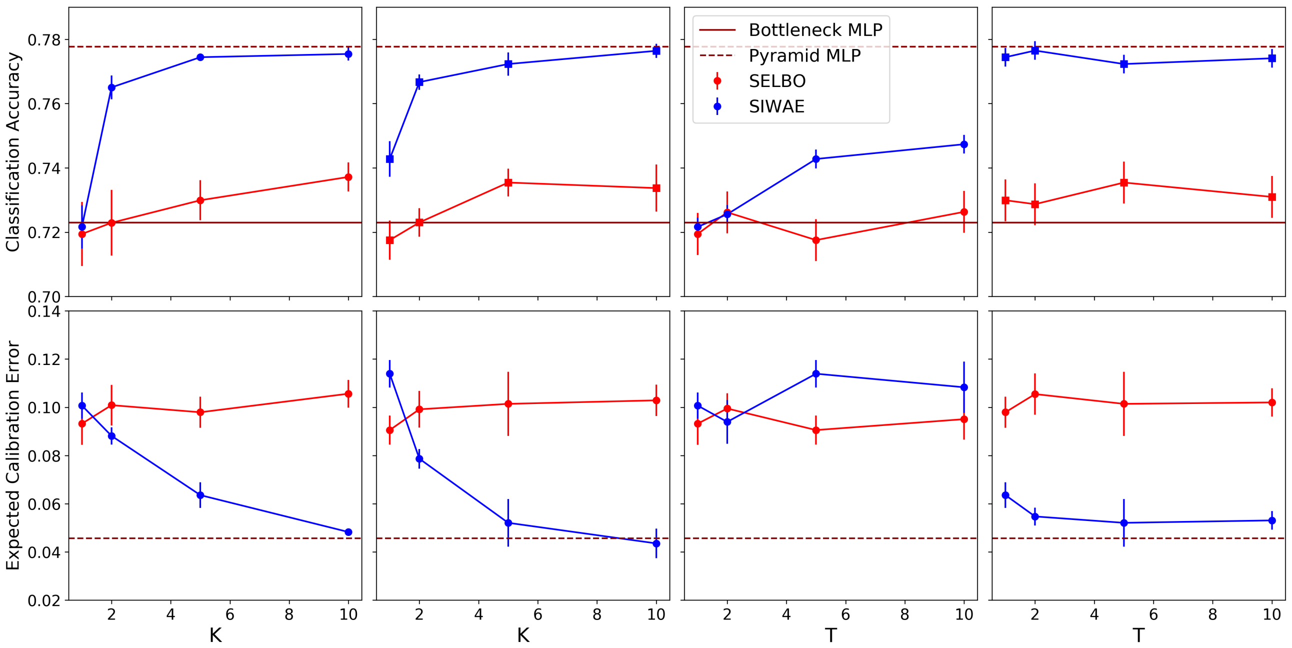


Figure 5: **Top row:** Classification accuracy of a model trained using the SELBO and SIWAE objectives as we vary the number of mixture components (K, left two panels) and the number of samples per component (T, right two panels). **Bottom row:** Expected Calibration Error as a function of the number of mixture components K, or samples T.

confidence in the decoder prediction (less transparent, more saturated colors indicate more confidence in a prediction and vice-versa). Where the SELBO-trained decoder tends to have fuzzier, more transparent decision regions, the SIWAE-trained decoder has sharper, more confident decision boundaries. We later see how this property is critical for well-calibrated predictions. Furthermore, while it is difficult to evaluate the interpretability of the latent space quantitatively, the SIWAE models are qualitatively easier to interpret using the latent space, with the model very clearly predicting the example shown as either a 5, an 8, or a 6 (with some additional limited probability that it is a 3). This appears to reflect our own intuition of the output class of this example.

To quantitatively compare SIWAE and SELBO, we consider how classification performance varies as the number of mixture components are varied. For this, we train models using $K = [1, 2, 5, 10]$ for the number of mixture components. We also use $T = [1, 2, 5, 10]$ for the number of samples drawn *per component*. For a single component model, we optimize both the traditional evidence lower bound (ELBO), as well as the importance weighted estimate of the evidence (IWAE). For the mixture models, we use stratified sampling to compute the ELBO (SELBO), as well as SIWAE. To evaluate the accuracy of the model, we first compute the predictive distribution by decoding 10^4 samples from $q_\phi(z|x)$ and averaging the class probabilities returned by each sample. This marginalizes over the uncertainty in the latent variables and if our prior beliefs are correct, nominally produces calibrated probabilities. The predicted class is the one with the largest probability under the predictive distribution, and accuracy of these predictions is measured on the test set. Because one-column MNIST does not have an established benchmark, we also train two deterministic models to use as baselines: (1) a "pyramid" MLP with 5 layers of 256 hidden units to approximate the peak deterministic accuracy, and (2) a "bottleneck" MLP with the same architecture as our VIB models, therefore containing a two dimensional "latent space." Additional experimental details can be found in the supplement.

Figure 5 shows the classification accuracy and expected calibration error of our VIB model over a range of K and T. For this figure, we use 10^4 samples from $q_\phi(z|x)$ to compute the posterior predictive distribution. From these results we make several observations: 1) We find that SELBO disfavors multimodality (as seen in Figure 4), and therefore offers no improvement (or only marginal improvement) with additional mixture components or samples. 2) We find that SIWAE overcomes these deficiencies and therefore offer increased accuracy with additional mixture components (and samples for $K = 1$). For large K, T, the performance approaches the deterministic baseline, but does

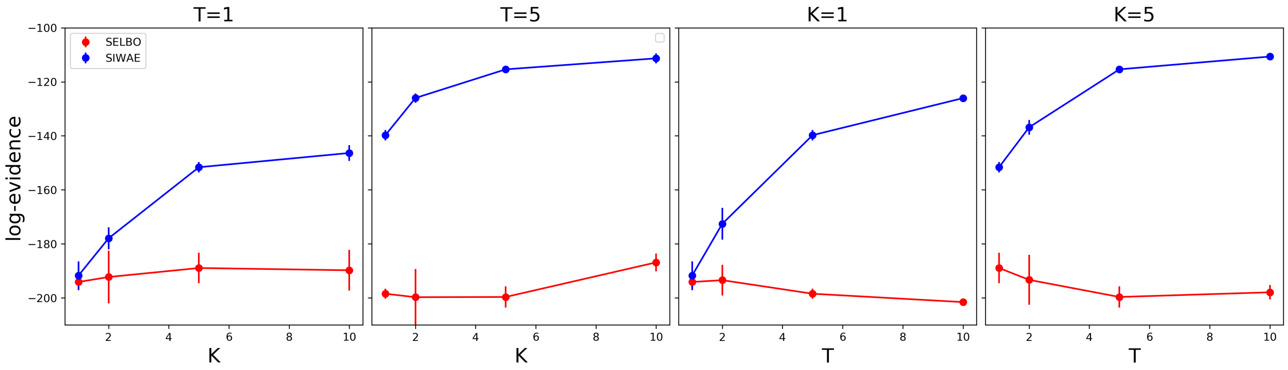


Figure 6: Model evidence as a function of the number of mixture components K or the number of samples per component T. The evidence was measured using a single SIWAE estimate with 100 samples. We find that models trained with SELBO appear to offer little to no noticeable improvement with either K or T, while SIWAE offers substantial improvements with both.

so using far fewer parameters. 3) Optimizing the SIWAE loss with a larger number of components leads to an improvement in calibration, as measured by a reduction in the expected calibration error [Guo et al., 2017], which measures the difference between the probability of an outcome and the observed frequency at which that outcome occurs. This is important, since real-world decision-making systems not only require accurate models, but also ones which quantify their uncertainty correctly. It is also important to note that arbitrarily growing the number of importance samples may also be harmful, a phenomenon observed by Rainforth et al. [2018]. We do not see any evidence for this over the range of $T = 1 \rightarrow 10$ importance samples, suggesting that positive effect of importance sampling enabling fitting better mixture models outweighs the negative effect of worse gradients. However we also speculate that since the gradient variance scales as $T^{0.5}$, the performance may turn over for sufficiently large T.

4.3 Single Column MNIST VAE

The SIWAE objective appears to successfully infer latent structure indicative of class boundaries using only a single column of the image. However, a different and equally intriguing question is if this representation is also sufficient to reconstruct the image itself. This question was explored by Doersch [2016], who showed that a class-conditional VAE was necessary to break the class degeneracy that can exist when the images are a single column. Our hypothesis was that the use of a mixture posterior distribution can replicate this conditionality, without using the class labels.

Our test setup is the same: train a model with either the SIWAE or the SELBO loss, and observe performance as a function of K and T. This time, we use the log-evidence to measure performance, computed with a SIWAE estimate using 100 samples from the surrogate posterior. We thought this was the most fair comparison, as it holds the total sample number fixed, and therefore highlights the difference based solely on the posterior expressiveness.

Figure 6 shows the model evidence as a function of K and T. We find that for SIWAE trained models, the log evidence increases substantially with increasing K, indicative of the model successfully leveraging representational multimodality. For SELBO-like losses, we observe no improvement with K or T, indicating unimodality and unsuccessful posterior approximation. Consequently SELBO shifts uncertainty into the decoder, resulting in fuzzy, low confidence outputs (see Appendix C.3). We replicated this result using the FashionMNIST and CIFAR-10 datasets (see Appendix C.4 and Appendix C.5). For comparison, in Appendix C.6 we run the same experiment using full-image MNIST; results indicate that SIWAE provides benefits over SELBO in lower dimensional latent spaces and these benefits diminish as the dimensionality increases.

5 Conclusion

We demonstrate that although stratified sampling enables ADVI with mixture posterior distributions, the ELBO impedes surrogate posterior multimodality. SIWAE, a tighter evidence lower bound analogous to the IWAE, utilizes stratification over posterior mixture components to make the bound tighter. We experimentally verify that SIWAE facilitates discovery of multimodality in the latent space, stratified ELBO does not, and that multimodality improves generative model performance, particularly for incomplete input data or low dimensionality representations. We also show that SIWAE enables better classifier accuracy and calibration error and that both improve as as the number of components is increased.

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283, 2016.

Alex Alemi, Ian Fischer, Josh Dillon, and Kevin Murphy. Deep variational information bottleneck. In *ICLR*, 2017. URL `https://arxiv.org/abs/1612.00410`.

Christopher M Bishop. Novelty detection and neural network validation. *ICANN '93*, 1993.

Christopher M Bishop, Neil D Lawrence, Tommi Jaakkola, and Michael I Jordan. Approximating posterior distributions in belief networks using mixtures. In *Advances in neural information processing systems*, pages 416–422, 1998.

Yuri Burda, Roger Grosse Roger, and Ruslan Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

Chris Cremer, Quaid Morris, and David Duvenaud. Reinterpreting importance-weighted autoencoders. *arXiv preprint arXiv:1704.02916*, 2017.

Nat Dilokthanakul, Pedro AM Mediano, Marta Garnelo, Matthew CH Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. Deep unsupervised clustering with Gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*, 2016.

Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.

Justin Domke and Daniel R Sheldon. Divide and couple: Using Monte Carlo variational objectives for posterior approximation. In *Advances in Neural Information Processing Systems*, pages 338–347, 2019.

Mikhail Figurnov, Shakir Mohamed, and Andriy Mnih. Implicit reparameterization gradients. In *Advances in Neural Information Processing Systems*, pages 441–452, 2018.

Alex Graves. Stochastic backpropagation through mixture density distributions. *arXiv preprint arXiv:1607.05690*, 2016.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1321–1330. JMLR. org, 2017.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. Beta-VAE: Learning basic visual concepts with a constrained variational framework. *ICLR*, 2(5):6, 2017.

Martin Jankowiak and Fritz Obermeyer. Pathwise derivatives beyond the reparameterization trick. *arXiv preprint arXiv:1806.01851*, 2018.

Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou. Variational deep embedding: An unsupervised and generative approach to clustering. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 1965–1972. AAAI Press, 2017. ISBN 9780999241103.

Matthew J Johnson, David K Duvenaud, Alex Wiltschko, Ryan P Adams, and Sandeep R Datta. Composing graphical models with neural networks for structured representations and fast inference. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2946–2954. Curran Associates, Inc., 2016. URL `http://papers.nips.cc/paper/6379-composing-graphical-models-with-neural-networks-for-structured-representations-and-fast-inference.pdf`.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Diederik P. Kingma, Max Welling, et al. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, volume 1, 2014.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324, 1998.

Warren R. Morningstar, Alexander A. Alemi, and Joshua V. Dillon. Pacm-bayes: Narrowing the empirical risk gap in the misspecified bayesian regime. *arxiv preprint arXiv:2010.09629*, 2020.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? *arXiv preprint arXiv:1810.09136*, 2018.

Seong Joon Oh, Andrew C. Gallagher, Kevin P. Murphy, Florian Schroff, Jiyan Pan, and Joseph Roth. Modeling uncertainty with hedged instance embeddings. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=r1xQQhAqKX`.

Pranav Poduval, Hrushikesh Loya, Rajat Patel, and Sumit Jain. Mixture distributions for scalable Bayesian inference, 2020. URL `https://openreview.net/forum?id=S1x6TlBtwB`.

Tom Rainforth, Adam R Kosiorek, Tuan Anh Le, Chris J Maddison, Maximilian Igl, Frank Wood, and Yee Whye Teh. Tighter variational bounds are not necessarily better. *arXiv preprint arXiv:1802.04537*, 2018.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, pages 1278–1286, 2014.

Geoffrey Roeder, Yuhuai Wu, and David K Duvenaud. Sticking the landing: Simple, lower-variance gradient estimators for variational inference. In *Advances in Neural Information Processing Systems*, pages 6925–6934, 2017.

Francisco J. R. Ruiz, Michalis K. Titsias, and David M. Blei. The generalized reparameterization gradient. In *Advances in neural information processing systems*, pages 460–468, 2016.

Tim Salimans and David A Knowles. Fixed-form variational posterior approximation through stochastic linear regression. *Bayesian Analysis*, 8(4):837–882, 2013.

Jasper Snoek, Yaniv Ovadia, Emily Fertig, Balaji Lakshminarayanan, Sebastian Nowozin, D Sculley, Joshua Dillon, Jie Ren, and Zachary Nado. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, pages 13969–13980, 2019.

Jakub M Tomczak and Max Welling. VAE with a VampPrior. *arXiv preprint arXiv:1705.07120*, 2017.

George Tucker, Dieterich Lawson, Shixiang Gu, and Christopher Maddison. Doubly reparameterized gradient estimators for monte carlo objectives. 2019. URL `https://openreview.net/pdf?id=HkG3e205K7`.

Mike West. Approximating posterior distributions by mixtures. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(2):409–422, 1993.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.