Project ML2 - 97209
Taming VAE

David Parnas 337977045
Tamir Shazman 316250877

Quick Overview

The paper's main goals are:

1. Offer a deeper understanding of common problems with VAEs
2. Discuss the relationship between the prior and marginal posterior of the latent space
3. Relate VAEs to other fields-Spectral Clustering and Statistical Mechanics
4. GECO - a principled approach to managing the importance given to reconstruction error versus the KL term during training

1. Common problems with VAEs

There are two primary problems that affect the generation of random samples with VAEs. The first is "holes"-provided with a sample point in the latent space the decoder fails to construct a meaningful data point in the observed space. The most intuitive example of this is a VAE trained on pictures. A hole occurs when a sampled point is decoded as an almost totally black picture. There is a "hole" in the latent space resulting in an empty picture being reconstructed.
The second common problem is blurred reconstruction. Unlike a "hole," the VAE succeeds in constructing a meaningful data point in the observed space. Returning to the earlier example, the picture is not black. It is, however, blurry. The literature has often attributed this phenomena to the use of Gaussian posteriers. However, the authors claim that blurred reconstruction occurs as a result of the latent sample being found in the cross section of the supports of several decoders.

2. The relationship between the prior and marginal posterior of the latent space

The authors continue building a deeper understanding of VAEs' behavior. They show that after making certain assumptions about the decoders, it can be shown that the marginal posterior equals the prior.

3. Relating VAEs to other fields

The authors develop a better understanding of the decoders' fixed points by framing the problem in terminology borrowed from Statistical Mechanics.

4. GECO

Having expanded the theory on current VAE models, the authors offer the GECO algorithm. GECO offers a more general optimization problem and a method for controlling the trade-off between the reconstruction error and the KL term during training.
Instead of ELBO:

$$ELBO = \mathbb{E}_{\rho(x)}\left[[\mathbb{E}_{q(z|x)}[\ln p(x|z)] - KL[q(z|x)\|\pi(z)]\right]$$

GECO uses:

$$L_\lambda = \mathbb{E}_{\rho(x)}\left[KL[q(z|x)\|\pi(z)] + \lambda^T\,\mathbb{E}_{q(z|x)}[\mathcal{C}(x, g(z))\right]$$

Although both contain a term controlling the reconstruction error and the KL divergence, GECO changes two things. 1. It allows for a variety of reconstruction error constraints, generalized as $\mathcal{C}(x, g(z))$. 2. Similar to β-VAEs it formulates the reconstruction error as a Lagrangian constraint with a Lagrangian multiplier.
The paper "Beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework" shows improved results on ELBO by formulating the reconstruction error as a Lagrangian constraint with a Lagrangian multiplier. This allows for the trade-off between the reconstruction error and KL divergence to be controlled by β, the Lagrangian multiplier. This gives the ML practitioner more customizability in the β-VAE model relative to the simple VAE model. However, it adds a hyperparameter that must be found.
The authors offer the GECO algorithm using the above optimization problem and, critically, a method for adjusting the Lagrangian multiplier during training. They claim that this prevents "over-optimizing" of the reconstruction error at the expense of the KL term, resulting in a suboptimal latent distribution space.

We summarize these ideas in more detail below with proofs.

Understanding Holes and Blurred Reconstructions

The authors show that the optimal decoder, g(z), is a convex linear combination of the training data weighted by the encoders, $q(z|x)$. As a result, when $q(z|x) \approx 0$ there is a "hole" in the decoder and that blurring occurs when there is overlap between the supports of multiple x's decoders, $q(z|x)$. In order to do so, they use the stationary points of ELBO.

Stationary Points Using ELBO

$$ELBO = \mathbb{E}_{\rho(x)}\left[[\mathbb{E}_{q(z|x)}[\ln p(x|z)] - KL[q(z|x)\|\pi(z)]\right]$$

We will further develop the expression using the assumptions that the authors make in their derivations.
They make the following assumptions:

$$\rho(x) = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \delta^*(x - x_i)$$
$$\pi(z) = \mathcal{N}(\mathbf{0}, \mathcal{I})$$
$$p(x|z) = \mathcal{N}(g(z), \sigma^2)$$

*δ represents the Dirac measure.

$$ELBO = \mathbb{E}_{\rho(x)}\left[[\mathbb{E}_{q(z|x)}[\ln p(x|z)] - KL[q(z|x)\|\pi(z)]\right] \approx \sum_x \rho(x)\left(\sum_z q(z|x)\ln p(x|z) - \sum_z ln\frac{q(z|x)}{\pi(z)}\right)$$

By the definition of Dirac measure, $\forall x$, $\rho(x) = \frac{\mathcal{I}_{(x=x_i)}}{n}$ Therefore, we receive the following:

$$\frac{1}{n}\sum_x\sum_z q(z|x)\ln p(x|z) - \sum_z ln\frac{q(z|x)}{\pi(z)}$$

We will now find the partial derivatives of the ELBO function by decoder, $g(z)$, and encoder, $q(z|x)$.

Decoder

$$\frac{\partial}{\partial g(z)}\frac{1}{n}\sum_x\sum_z q(z|x)\ln p(x|z) - ln\frac{q(z|x)}{\pi(z)} \tag{1}$$

We will plug in the distribution for $p(x|z) = \mathcal{N}(g(z), \sigma^2)$

$$\frac{\partial}{\partial g(z)}\frac{1}{n}\sum_x\sum_z q(z|x)\ln\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{\|x-g(z)\|^2}{2\sigma^2}} - \ln\frac{q(z|x)}{\pi(z)} \tag{2}$$

$$\frac{\partial}{\partial g(z)}\frac{1}{n}\sum_x\sum_z q(z|x)\ln\frac{1}{\sqrt{2\pi\sigma^2}} - \frac{\|x-g(z)\|^2}{2\sigma^2} - \ln\frac{q(z|x)}{\pi(z)} \tag{3}$$

$$\frac{1}{n}\sum_x -\frac{q(z|x)(x-g(z))}{\sigma^2} \tag{4}$$

Therefore, setting the partial derivative to zero to find the stationary point, we find the following relationship between decoder and encoders:

$$g(z) = \frac{\sum_x q(z|x)x}{\sum_x q(z|x)}$$

Thus the decoder is a convex linear combination of the training data weighted by the encoders, $q(z|x)$.

Encoder

$$\frac{\partial}{\partial q(z|x)}\frac{1}{n}\sum_x\sum_z q(z|x)\ln p(x|z) - ln\frac{q(z|x)}{\pi(z)} \tag{1}$$

$$\frac{1}{n}\left[\ln\frac{1}{\sqrt{2\pi\sigma^2}} - \frac{\|x-g(z)\|^2}{2\sigma^2} + \ln\frac{q(z|x)}{\pi(z)} + 1\right] \tag{2}$$

The paper only shows a proportional relationship, ignoring the constants. The simplified equation is:

$$-\frac{\|x-g(z)\|^2}{2\sigma^2} + \ln\frac{q(z|x)}{\pi(z)} \tag{3}$$

Therefore, setting the partial derivative to zero to find the stationary point, we find the following relationship between encoder and decoder:

$$q(z|x) \propto \pi(z)e^{\frac{-\|x-g(z)\|^2}{2\sigma^2}}$$

Holes

Having seen that $g(z) = \frac{\sum_x q(z|x)x}{\sum_x q(z|x)}$, it easy to see that if $\forall x, q(z|x) \approx 0$ then also $g(z) \approx 0$ resulting in holes.

Blurred Reconstructions

The authors claim that blurred reconstruction is not primarily caused by using Gaussian models in the likelihood. Rather, it is from overlap in the support of multiple $q(z|x_i)$.
They show there results by replacing $q(z|x_i)$ with $\frac{\pi(z)\mathcal{I}(z\in\Omega_i)}{\pi_i}$ s.t. $\pi_i = \mathbb{E}_{\pi(z)}[\mathcal{I}(z \in \Omega_i)]$ Therefore,

$$g(z) = \sum_i x_i\frac{q(z|x_i)}{\sum_j q(z|x_j)} \tag{1}$$

Because $\pi(z)$ appears in the numerator and denominator it is canceled out and we get

$$\sum_i x_i\frac{\frac{\mathcal{I}(z\in\Omega_i)}{\pi_i}}{\sum_j\frac{\mathcal{I}(z\in\Omega_j)}{\pi_j}} \tag{2}$$

As we can see, when z is uniquely found within the support of a specific x_i then we get:

$$\sum_i x_i\frac{\frac{\mathcal{I}(z\in\Omega_i)}{\pi_i}}{\sum_j\frac{\mathcal{I}(z\in\Omega_j)}{\pi_j}} = x_i\frac{\frac{1}{\pi_i}}{\frac{1}{\pi_i}} = x_i$$

However, if z is found in the support of multiple x then g(z) is a weighted average of these x's, resulting in blurred reconstructions.

Conclusion

In conclusion, the authors use the stationary point $g(z) = \frac{\sum_x q(z|x)x}{\sum_x q(z|x)}$ in order to show that "holes" occur when $q(z|x) \approx 0$. They also show that the decoder $g(z)$ is a linear combination of $q(z|x)$. Therefore, when a z is part of the support of multiple $q(z|x)$, it is reconstructed as a blurry, weighted average of the respective x.

Relationship of Optimal Encoders to the Prior

The authors claim that when the encoders $q(z|x_i) = \frac{\pi(z)\mathcal{I}(z\in\Omega_i)}{\pi_i}$ s.t. $\pi_i = \mathbb{E}_{\pi(z)}[\mathcal{I}(z \in \Omega_i)]$, they are fixed points and, when optimized, form an equiprobable partition of the latent probability space. This in turn results in the marginal posterior being equal to the prior.

Proving Fixed Points

Before we can show that $q(z|x)$ are fixed points, we must show that σ^2 approaches zero as the encoders and decoders are optimized.

1. σ^2 approaches zero

$$ELBO = \mathbb{E}_{\rho(x)} \left[[\mathbb{E}_{q(z|x)}[\ln p(x|z)] - KL[q(z|x) \| \pi(z)] \right] = \sum_x \rho(x) \left(\sum_z q(z|x) \ln p(x|z) - \sum_z ln \frac{q(z|x)}{\pi(z)} \right) \tag{1}$$

We will note that the KL term is not dependent on σ^2. Therefore, we will ignore it as it will become zero in the partial derivative.

$$\frac{\partial}{\partial \sigma^2} \mathbb{E}_{\rho(x)q(z|x)} - \frac{1}{2} \left[\ln 2\pi - \ln \sigma^2 - \frac{\|x - g(z)\|^2}{\sigma^2} \right] \tag{2}$$

$$\mathbb{E}_{\rho(x)q(z|x)} \left[-\frac{1}{2} \left(-\frac{1}{\sigma^2} + \frac{\|x - g(z)\|^2}{(\sigma^2)^2} \right) \right] \tag{3}$$

Setting the partial derivative to zero and noticing that σ^2 is not a random variable related to the expectation, we find:

$$\sigma^2 = \mathbb{E}_{\rho(x)q(z|x)} \left[\|x - g(z)\|^2 \right] \tag{4}$$

Therefore, as optimization brings g(z) closer to x, σ^2 approaches zero.

2. Simplifying g(z)

As we saw earlier, $g(z) = \frac{\sum_x q(z|x)x}{\sum_x q(z|x)}$

We also previously mentioned that in their proofs the authors replace $q(z|x_i)$ with $\frac{\pi(z)\mathcal{I}(z\in\Omega_i)}{\pi_i}$ s.t. $\pi_i = \mathbb{E}_{\pi(z)}[\mathcal{I}(z \in \Omega_i)]$

Let us note that assuming that $\{\Omega_i\}_{i=1}^n$ forms a partition of the latent probability space then:

$$\forall x_i, \frac{q(z|x_i)}{\sum_j q(z|x_j)} = \frac{\frac{\pi(z)\mathcal{I}(z\in\Omega_i)}{\pi_i}}{\sum_j \frac{\pi(z)\mathcal{I}(z\in\Omega_j)}{\pi_j}} = \frac{\frac{\pi(z)\mathcal{I}(z\in\Omega_i)}{\pi_i}}{\frac{\pi(z)\mathcal{I}(z\in\Omega_i)}{\pi_i}} = \mathcal{I}(z \in \Omega_i)$$

Therefore,

$$g(z) = \frac{\sum_i q(z|x_i)x_i}{\sum_i q(z|x_i)} = \sum_i \mathcal{I}(z \in \Omega_i)x_i$$

3. $q(z|x)$ are fixed points

Now we can continue to show that $q(z|x)$ are fixed points.

$$q(z|x) \propto \pi(z)e^{\frac{-\|x-g(z)\|^2}{2\sigma^2}}$$

Combining these equations and dividing by the necessary constant, we get:

$$q(z|x_i) = \lim_{\sigma^2 \to 0} \frac{\pi(z)e^{\frac{-\|x_i-g(z)\|^2}{2\sigma^2}}}{\sum_z \pi(z)e^{\frac{-\|x_i-g(z)\|^2}{2\sigma^2}}} \tag{1}$$

$$q(z|x_i) = \lim_{\sigma^2 \to 0} \frac{\pi(z)\sum_k e^{\frac{-\|x_i-x_k\|^2}{2\sigma^2}} \mathcal{I}(z \in \Omega_k)}{\sum_z \pi(z)\sum_j e^{\frac{-\|x_i-x_j\|^2}{2\sigma^2}} \mathcal{I}(z \in \Omega_j)} \tag{2}$$

$$q(z|x_i) = \frac{\pi(z)\mathcal{I}(z \in \Omega_i)}{\pi_i} \tag{3}$$

And this is a fixed-point for any prior $\pi(z)$ constrained by a partition such as $\{\Omega_i\}_{i=1}^n$

Proving Equiprobable Partition of the latent probability space

Plugging these fixed points into the KL term we get:

$$KL = \mathbb{E}_{\rho(x)q(z|x)}\left[\ln \frac{q(z|x)}{\pi(z)}\right] = \frac{1}{n}\sum_i \sum_z \frac{\pi(z)\mathcal{I}(z \in \Omega_i)}{\pi_i} \ln \frac{\pi(z)\mathcal{I}(z \in \Omega_i)}{\pi(z) * \pi_i} \tag{1}$$

$$= \frac{1}{n}\sum_i \sum_z \frac{\pi(z)\mathcal{I}(z \in \Omega_i)}{\sum_z \pi(z)\mathcal{I}(z \in \Omega_i)} \ln \frac{\pi(z)\mathcal{I}(z \in \Omega_i)}{\pi(z) * \pi_i} \tag{2}$$

$$= \frac{1}{n}\sum_i \sum_z \frac{\mathcal{I}(z \in \Omega_i)}{\sum_z \mathcal{I}(z \in \Omega_i)} \ln \frac{\mathcal{I}(z \in \Omega_i)}{\pi_i} \tag{3}$$

$$= \frac{1}{n}\sum_i \ln \frac{1}{\pi_i} = -\frac{1}{n}\sum_i \ln \pi_i \tag{4}$$

Solving for the maximal values of π_i using the Lagrangian with the usual constraints that assume that π_i a probability, we find:

$$-\frac{1}{n}\sum_i \ln \pi_i + \eta(\sum_i \pi_i - 1) - \sum_i \lambda_i^1 \pi_i + \sum_i \lambda_i^2(\pi_i - 1) \tag{5}$$

For some i

$$\frac{\partial}{\partial \pi_i}\left(-\frac{1}{n}\sum_i \ln \pi_i + \eta(\sum_i \pi_i - 1) - \sum_i \lambda_i^1 \pi_i + \sum_i \lambda_i^2(\pi_i - 1)\right) = \frac{1}{n * \pi_i} + \eta - \lambda_i^1 + \lambda_i^2 \tag{6}$$

When setting the partial derivative to zero and assuming that $\pi_i \neq 0$ or 1 we get

$$\frac{1}{n * \eta} = \pi_i \tag{7}$$

Using the condition that $\sum_i \pi_i = 1$,

$$\sum_i \frac{1}{n * \eta} = \sum_i \pi_i \tag{8}$$

$$\frac{1}{\eta} = 1 \tag{9}$$

Therefore, $\eta = 1$ and $\forall i, \pi_i = \frac{1}{n}$
We remind the reader that $\pi_i = \mathbb{E}_{\pi(z)}[\mathcal{I}(z \in \Omega_i)]$, which is to say the proportion of z in Ω_i.
Therefore, if $\forall i, \pi_i = \frac{1}{n}$ then the z are equally distributed between the partition $\{\Omega_i\}_{i=1}^n$.

Marginal Posterior Equal to Prior

In light of the previous mathematical developments, we look at the marginal posterior $q(z)$ given a partition $\{\Omega_i\}_{i=1}^n$.

$$g(z) = \frac{1}{n}\sum_i q(z|x_i) = \frac{1}{n}\sum_i \frac{\pi(z)\mathcal{I}(z \in \Omega_i)}{\pi_i} = \frac{n * \pi(z)}{n}\sum_i \mathcal{I}(z \in \Omega_i) = \pi(z)$$

Therefore, optimal solutions for VAE's encoders are inference models that cover the latent space in such a way that their marginal is equal to the prior.

Conclusion

The authors show that $q(z|x_i) = \frac{\pi(z)\mathcal{I}(z\in\Omega_i)}{\pi_i}$ s.t. $\pi_i = \mathbb{E}_{\pi(z)}[\mathcal{I}(z \in \Omega_i)]$ are fixed points and upon convergence to these fixed points, ELBO is maximized when the marginal posterior equals the prior.

Relationship of β-VAE to Other Fields

β-VAEs use a weighted optimization function allowing for customizable trade-off between the reconstruction error and the KL term.
In this part of the paper, the authors present β-VAEs in the terminology of Statistical Mechanics in order to discuss the convergence of decoders.
They model the statistical mechanics problem in the following way:

- **order parameter** $u(\beta)\ = \mathbb{E}\left[\|x - g(z)\|^2\right]$
- **critical temperature points** β_c
- **phase transitions** are detected by areas of high-curvature, i.e. a large second derivative $(\frac{\partial^2 u(\beta)}{\partial^2 \beta})$

This re-framing of the problem allows β-VAEs to be compared to kernel-PCA used with normalized Gaussian kernels. These kernels are used for dimension reduction and their reconstructions are equivalent to the fixed-point equations that maximize the ELBO for the β-VAE decoder.
We will show this result:
The decoder $g(z)$ can be presented as $\sum_i \frac{q(z|x_i)x_i}{\sum_j q(z|x_j)}$.
In this part of the paper, we represent the latent space using an orthogonal basis ϕ_a and denote $\phi_a : \mathcal{R}^{d_z} \to \{0,1\}$.
We denote the set of weights that maintain the equivalence to the original basis as $m_{i,a}$ s.t. i relates to an index relative to the sum and a refers to the new basis.
If we label $g(z) = \psi^T\phi(z)$ to be an equivalent representation of the decoder under the new basis ϕ, then we can reformulate the fixed-points in the following way:

$$q_i^{t+1} = \pi(z)\sum_b \frac{e^{\frac{-\|x_i-\psi_b^t\|^2}{2\beta}}}{\sum_b \pi_b e^{\frac{-\left\|x_i-\psi_b^t\right\|^2}{2\beta}}}$$

In order to simplify the following equation, let us define $m_{i,b}$ relative to basis $\phi(b)$:

$$m_{i,b}^t = \sum_b \frac{e^{\frac{-\|x_i-\psi_b^t\|^2}{2\beta}}}{\sum_b \pi_b e^{\frac{-\left\|x_i-\psi_b^t\right\|^2}{2\beta}}}$$

If so, then by the fixed-point equation mentioned above for $g(z)$, it must be that:

$$\psi_b^{t+1} = \sum_i \frac{m_{i,b}x_i}{\sum_j m_{j,b}}$$

Let us note that $m_{i,b}$ is the same term used in a normalized Gaussian kernel. Thereby, showing the equivalency claimed by the author between the decoder's fixed point and the reconstructions of a kernel-PCA model using a normalized Gaussian Kernel with standard deviation$\sqrt{\beta}$.

Equipartition of energy for β-VAEs

For β-VAEs, we proved in the previous section that the reconstruction vectors ψ_b converge to fixed-points.
Therefore, presenting the problem in terms of Statistical Mechanics, we define the reconstruction error, $\mathcal{C}(x, g(z))$, as the Hamiltonian function $H(x, z)$. The Hamiltonian function is used to

measure the total energy of a thermodynamic system. Because the reconstruction vectors converge to fixed-points, there are areas within the latent space where the Hamiltonian is approximately constant.

Let us define the following: $\Omega(x, z_0)$ is the set of points in the latent space where the Hamiltonian is approximately constant. That is to say $\Omega(x, z_0) = \{z' || H(x, z') - H(x, z_0)| \leq \epsilon\}$. We define Ω_a as the set $\{\Omega(x, z_0), \forall x \text{ and } z_0\}$.

Based on these observations, the authors suggest viewing the tiling of the latent space as the creation of different levels of the Hamiltonian function. We proved earlier that when using ELBO as the optimization function, the latent space is divided into an equiprobable partition under the prior. Therefore, these levels are equiprobable and this is similar to the *equipartition of energy theorem*. They believe that this view will lead to the development of more meaningful constraints for VAE models.

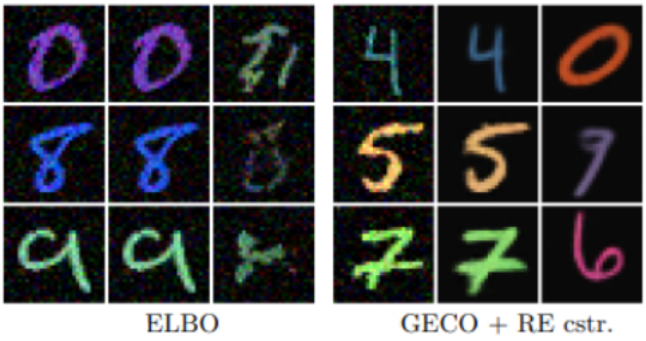
The GECO Algorithm

The main contribution of the authors in this paper is the GECO algorithm.
After explaining the "holes" and blurred reconstruction phenomena and enumerating the qualities of the β-VAE, the authors offer a principled approach to controlling the balance between the reconstruction error and the KL divergence.
Unlike β-VAE in which the machine learning practitioner must search for an elusive, optimal β to control the role of the unintuitive KL term, GECO optimizes the trade-off during training.

GECO Algorithm

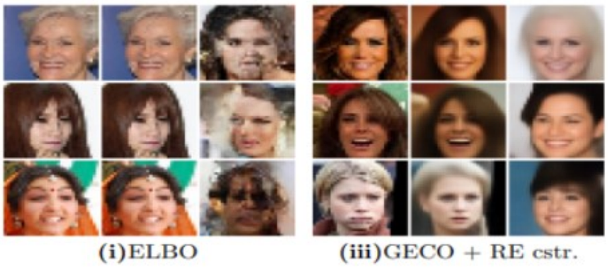
```
Initialize $t = 0$;
Initialize $\lambda = 1$;
**while** *training* **do**:
    Process current batch, x;
    Sample from variational posterior $z \sim q(z|x)$
    Compute empirical expectation of the reconstruction error $C^t = \mathcal{C}(x^t, g(z^t))$;
    **if t is 0** then initialize the moving average of the reconstruction error, $C^0_{ma} = C^0$
    **else** $C^t_{ma} = \alpha C^{t-1}_{ma} + (1 - \alpha) C^t$;
    $C^t + StopGradient(C^t_{ma} - C^t)$
    Update the parameters of the encoders and decoders using gradient descent
    Update $\lambda$ using the update rule $\lambda^t = \lambda^{t-1} e^{\propto C^t}$
    $t = t + 1$
```

As stated earlier, the GECO algorithm optimizes the trade-off parameter that controls the relative importance of the KL term and the reconstruction error.
Additionally, by using e, λ stays a positive number for all updates. This is a necessary requirement of optimization when using the Lagrangian, the mathematical basis of GECO.
The GECO algorithm allows the trade-off to change during training. For example, if the reconstruction error is high then more weight may be given to the reconstruction error; thereby, minimizing it.
This is important because often VAEs achieve the necessary reconstruction error, i.e. no noticeable blurriness, early on in training. After that point, when they continue to reduce the total error, these VAEs minimize the reconstruction error at the expense of the KL term.
The authors show the significance of this in their empirical trials using ConvDraw trained on the CelebA and Color-MNIST data sets.
Below, we can see that using ELBO creates a latent space with blurriness and holes-results of the unnecessary reduction of the KL term in favor of the reconstruction error. In contrast, GECO achieves reconstruction that is visually equivalent with no holes or blurriness.



ELBO GECO + RE cstr.

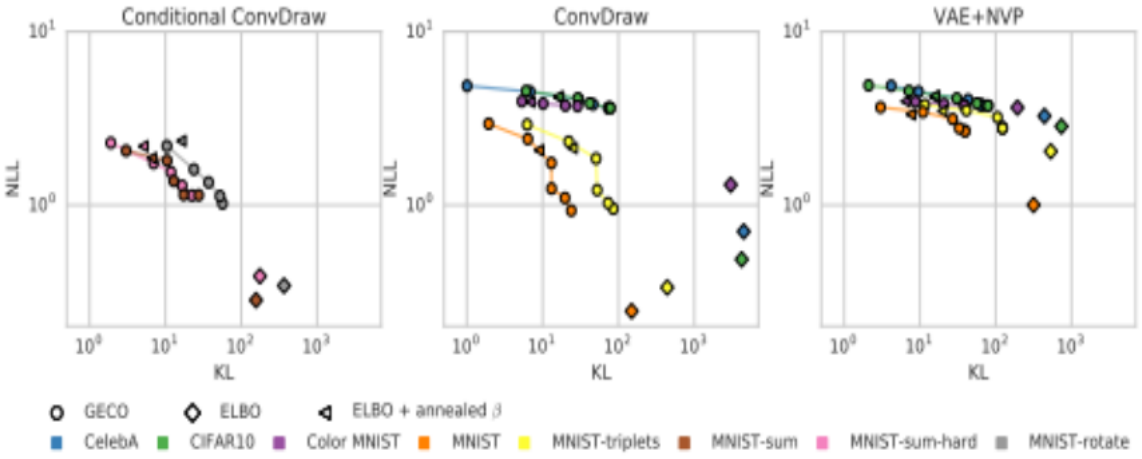
Figure 1: Note the differences in the last column of the ELBO dataset



(i)ELBO (iii)GECO + RE cstr.

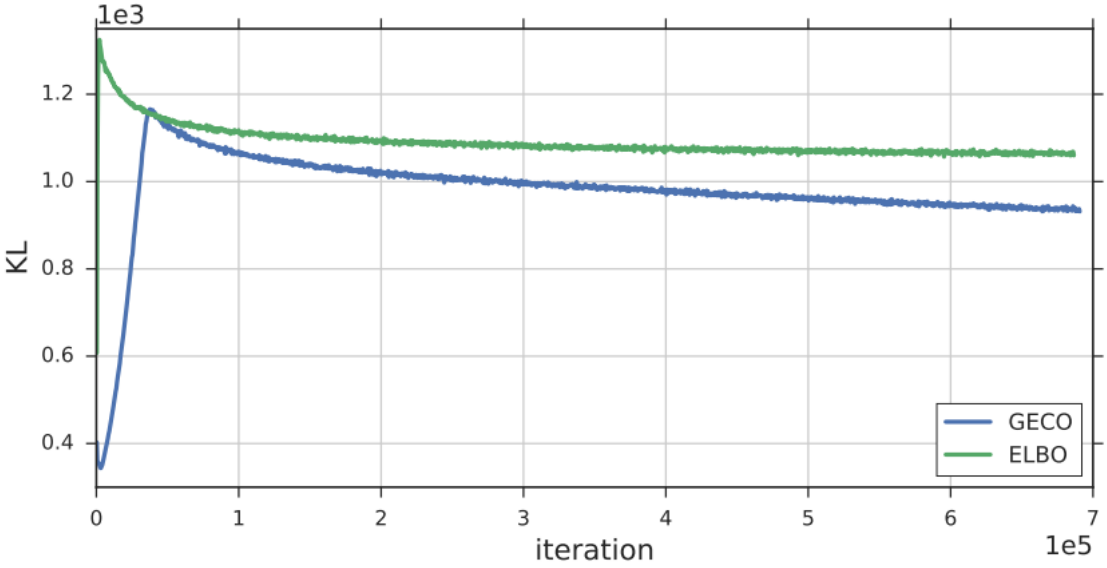
Figure 2: Note the differences in the last column of the ELBO dataset

The authors show with the following graphs how GECO succeeds in maintaining low KL during training.



We notice that the circles, representing GECO, are always farthest to the left. GECO has the lowest KL term.
In conjunction with the pictures presented earlier, we conclude that GECO achieves well reconstructed images **and** low KL divergence.

Additionally, the authors present the following graph, showing the steady decrease of KL for GECO relative to ELBO.



Our Experiments

In our experiment, we use the Fashion-MNIST data set consisting of a training set of 60,000 examples and a test set of 10,000 examples. Each example is a 28x28 gray scale image associated with a label from one of ten classes.
This data set requires relatively little computation, making it relevant to train without access to additional resources.
The authors present experiments trained on Color-MNIST and CelebA using a simple VAE model and conditional ConvDraw and ConvDraw. We trained the simple VAE model on Fashion-MNIST. However, due to computational limitations it was unfeasible to train the ConvDraw model. [We received permission for this in a meeting with the course's staff.]

We trained the simple VAE model in three different way:

1. GECO with the following reconstruction error - $||x - g(x)|| - \kappa^2$
2. Beta with Negative Log-Likelihood as a reconstruction error with a proper adjustment, i.e $NLL - \kappa^2$
3. Beta with the following reconstruction error - $||x - g(x)|| - \kappa^2$

For each Beta training (from the two above) we choose the following β : 1 (a.k.a ELBO), 0.5 and 2 and the following κ: 3 , 4 and 5.

Information plane analysis

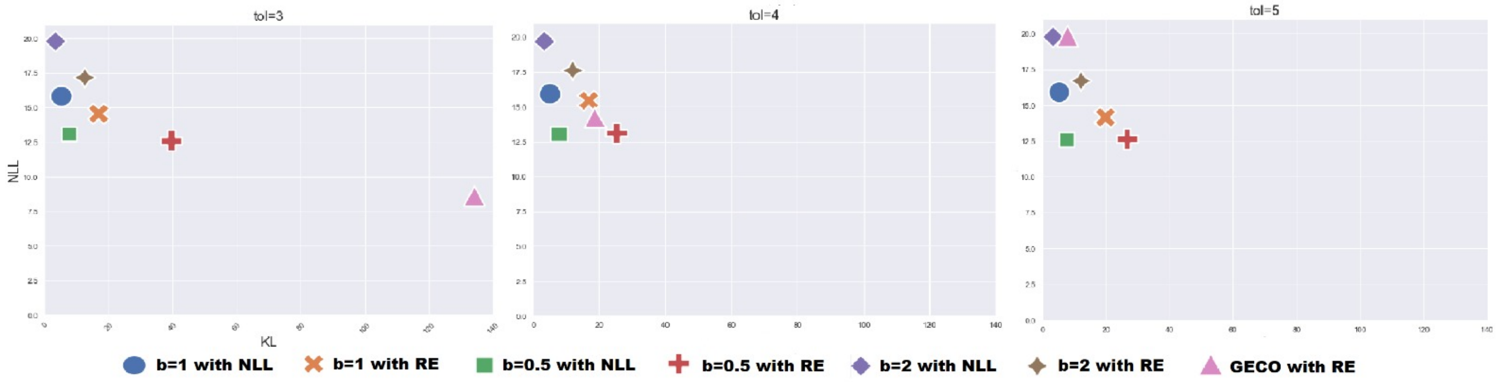


Figure 3: Each graph shows a different κ (tol=κ). Each model is trained using 200 epochs. We use $\sigma^2 = \sigma^2_{opt}\mathcal{I}_n$ for all the models to calculate the NLL s.t. $\sigma^2_{opt} = \frac{1}{M}\sum_i^M \|x_i - g(z_i)\|^2$

In our findings, GECO does not achieve the lowest KL term. We will address this divergence from the paper's findings in the following section.
Here we note the significance of the κ hyper-parameter. As we increase κ, the GECO model minimizes the KL term and increases the NLL. The larger κ causes the reconstruction error, $||x - g(x)|| - \kappa^2$, to become negative early on in training. This in turn decreases λ due to GECO's update method, giving more importance to the KL term.
This is clear from the objective function:

$$\mathbb{E}_{\rho(x)}\left[KL[q(z|x)\|\pi(z)] + \lambda^T \mathbb{E}_{q(z|x)}[\mathcal{C}(x, g(z))]\right]$$

This phenomena can be observed by following the pink triangle in the above graphs. As κ increases from 3 to 5, the pink triangle moves from the lower right side of the graph, i.e. low NLL & high KL, to the upper left side of the graph, i.e. high NLL & low KL.
The authors do not address the importance of picking the appropriate κ. However, we find it important to note. Because although GECO removes the need to tune the β hyper-parameter, κ remains an important hyper-parameter that must be handpicked.

Image Reconstruction and Generation

In this section, we present the visual output of the models, test the models' ability to reconstruct an image and observe the models' ability to generate new images based points sampled from the latent space, z.
From the β model's we chose to present $\beta = 0.5$ based on the information plane analysis. $\beta = 0.5$ performed the best both in terms of KL and NLL.

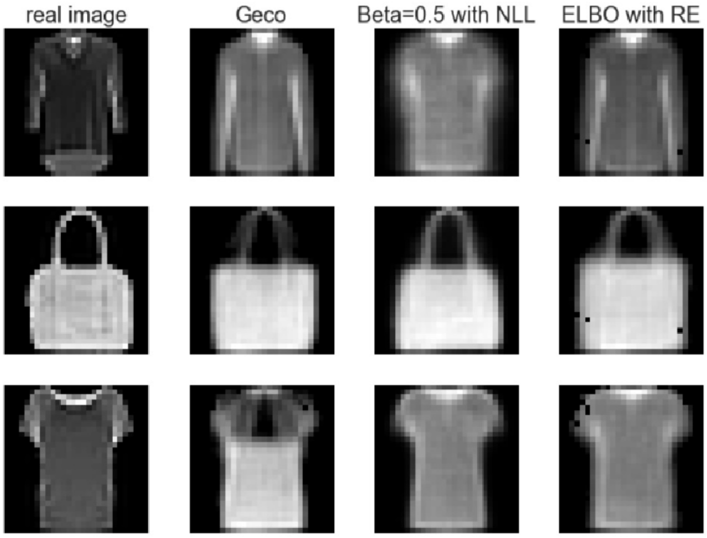


Figure 4: Reconstructed images of each model

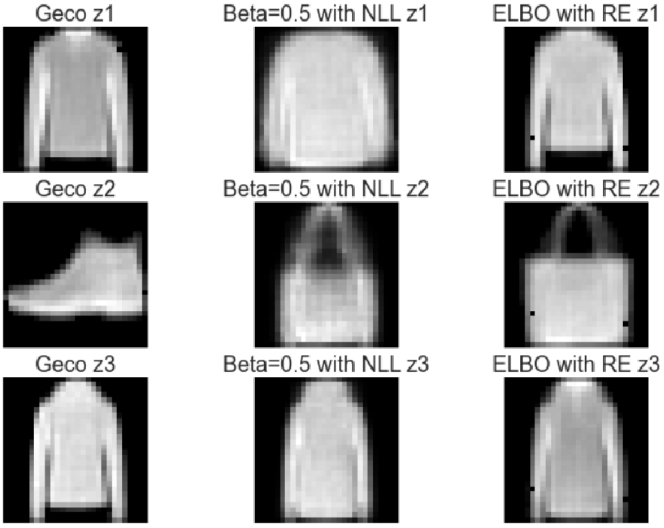


Figure 5: Three images generated by each model decoding the random Guassian variables $z_1, z_2, z_3 \in \mathcal{R}^{200}$

An interesting phenomenon that we can see from the above images is the similarity between the β and ELBO models. In the second line, both interpret z_2 as bags, while GECO interprets the point as a shoe. We witnessed this phenomena in other cases not presented here also.
Additionally, the GECO model generates clear images relative to the other models, exemplifying the advantage of controlling the trade-off between the KL term and NLL.

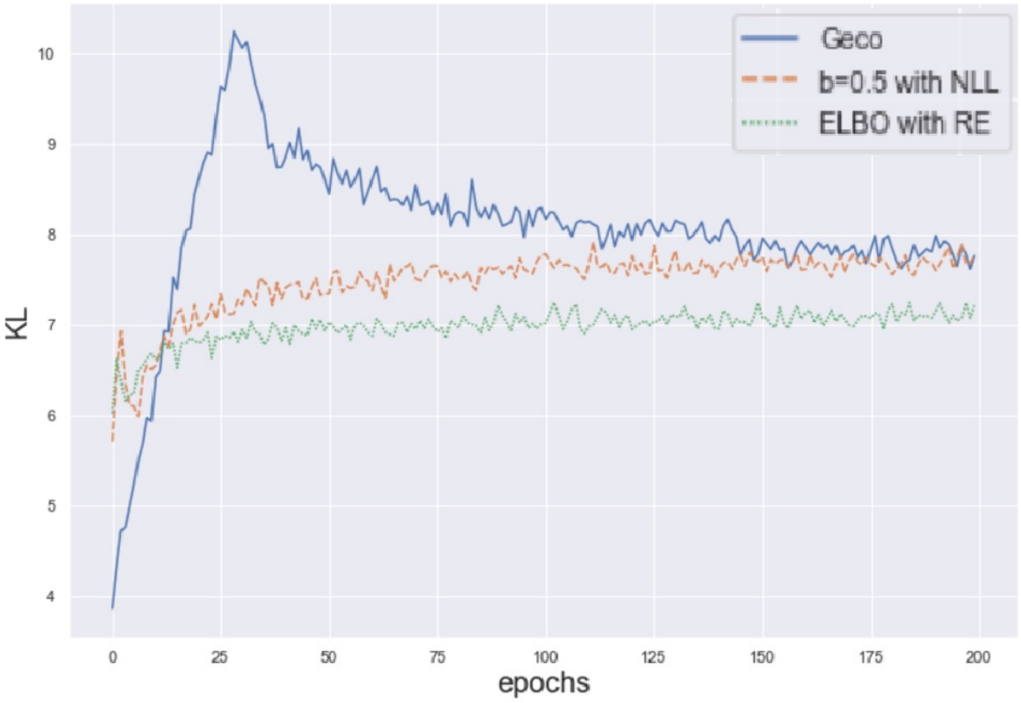


Figure 6: Here we can see the behavior of the KL term while training in each model. GECO's KL term rises quickly due to the changes in λ.

Comparing Results

Both the paper and our experiments find that GECO succeeds in striking a good balance between the KL term and the reconstruction error.

Relative to ELBO and β-VAE, GECO maintains aesthetically pleasing reconstructions and generated images, while surpassing them in the creation of robust latent spaces. In the GECO section, we show distorted numbers and faces that ELBO and β-VAE fail to generate while GECO succeeds. Our findings also support this. In the previous section, GECO produces coherent images of shirts and footwear, while ELBO and β-VAE create fuzzy approximations.

The information plane that we found after training GECO, ELBO and β-VAE differs from the findings in the paper. Unlike in the paper, the KL term for the GECO models under all κ is not smaller than the other models in our findings. We suggest that this does not indicate a failure of GECO. This is a result of training the model for less epochs. Because of limited computational resources, we trained our model for only 200 epochs. Whereas, the authors trained their models for several thousand epochs. This is significant because of the mechanics of the GECO algorithm. After reaching a reconstruction error of less than κ, λ shifts the objective function's priority to the KL term. The more epochs that occur after reaching this critical point, the lower the KL term. Therefore, given more epochs, GECO would also achieve a low KL term on this data set.

For similar reasons, the graphs showing the change in KL during training are different. Both in our findings and in the paper, the KL term explodes in the beginning because λ grows. However, in our findings even at the end of training the KL term is high. Certainly, higher than the KL term of ELBO. We believe that GECO would achieve the lowest KL term relative to the other models given more training epochs.

In conclusion, our results from the image reconstruction and generation on the Fashion-MNIST data set support the paper and show the advantages of GECO VAE. We claim that the differences found in the graphs showing the KL term in our findings are insignificant.