# Coursework 2

**Tuan Nguyen[1], Gizat Makhanov[2], Jeffrey Man Hong Chu[3]**

[1] *97338*
[2] *97547*
[2] *16841*

December 7, 2018

**As the sequel of the first assignment, this report is focusing on applying both stochastic and deterministic approaches in machine learning with intractable posterior. The pros and cons of respective methods will be illustrated during the evaluation of the processes and the results. This assignment aims to lead the reader to make their own the best decision respective to a specific situation by taking the balance of the performance of both methods.**

## 1 Approximate Inference

### Question 1

The denoised image with the default noise. We need 6 passes through the nodes until we start to get descent results.

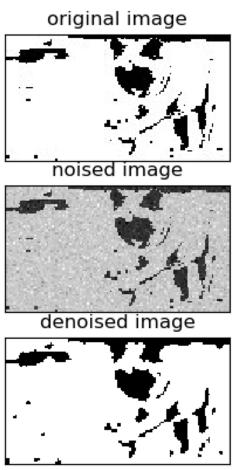



**Figure 1:** *prop = 0.7 and varSigma = 0.1*

The denoised image with more noise. We need 12 passes through the nodes until we reach local minima

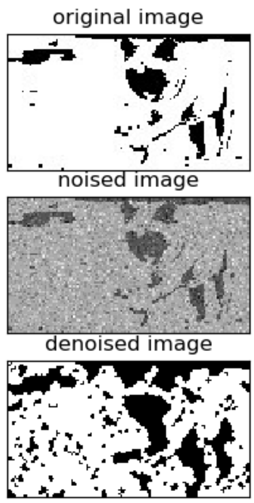



**Figure 2:** *prop = 0.7 and varSigma = 0.3*

We can see that the algorithm does not work well when we increase varSigma from 0.1 to 0.3. This is because the value of noise added to the original image increases, this leads to the the relationship between the noised image and original image is not close. As a result, we do not trust the likelihood much and it causes poor result.

### Question 2

Gibbs sampler for the image denoising with default noise.

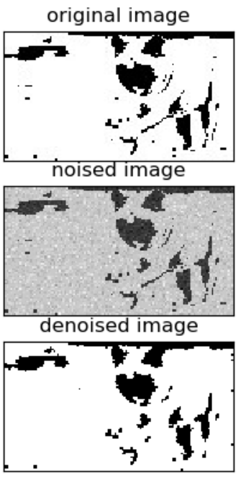



**Figure 3:** *prop=0.7 and varSigma=0.1*

Gibbs sampler for the image denoising with more noise.

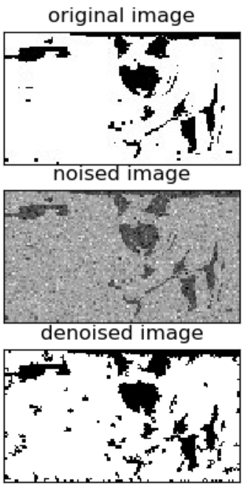



**Figure 4:** *prop=0.7 and varSigma=0.3*

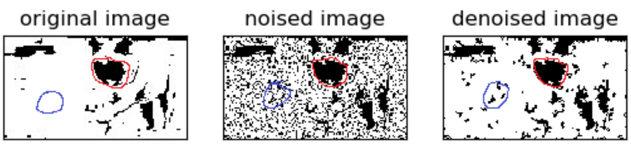



**Figure 5:** *The red circle shows example of where Gibbs sampler works and blue circle shows example of where it doesn't work*

We can see that the Gibbs sampler does not work when the noised image have the small area that most pixels is converted from white to black, it is normally happened when varSigma is large such as 0.3 in the above example . And it works well when most of pixel in medium area is unchanged or with small noise.

## Question 3

When we pick random node in each interation, the final result is approximately same with result by interating whole image. However, the number of iteration is much bigger, 30000 times in our case.

## Question 4

The factors affects number of iterations we run the sampler have on the results:

- Level of noised image, we can see in Question 2, it affects not only the number of iteration but also the quality of denoised image.
- The initialisation of parameter can also make the image very likely or so different.
- The design of probability distribution for the prior and likelihood
- The way we random the threshold (fixed or dynamic) to set pixel value to 1 or -1
- The ways we iterate the image: iterate each pixel in each iteration or choose pixel in image randomly

When we run the sampler for different times, the result does not always get better. This is because when the latent image is close to the original image, the probability of pixel in latent image equal pixel in original image is close to 1, so that it almost hardly change. As a result, the result does not always get better.

## Question 5

We will look at the difference between $KL(q(x)|p(x))$ and $KL(p(x)|q(x))$ in iterm of finding $q(x)$ that fits $p(x)$

- $KL(q(x)|p(x))$: we can see that when $p(x)$ is close to zero, $\frac{q(x)}{p(x)}$ is big number. However, we expect that this fraction is small, so that $p(x)$ must be equal zero. As a result, the probability mass of $q$ is restricted to 0 when the mass probability $p$ is equal 0.
- $KL(p(x)|q(x))$: the $p(x)$ is now in the numerator so there is no restriction for $q(x)$ or $q(x)$ can place probablity mass arbitrarily. As a result, we normally get more general and better approximation of $p(x)$. However, it is more difficult to find $q(x)$ in this case.

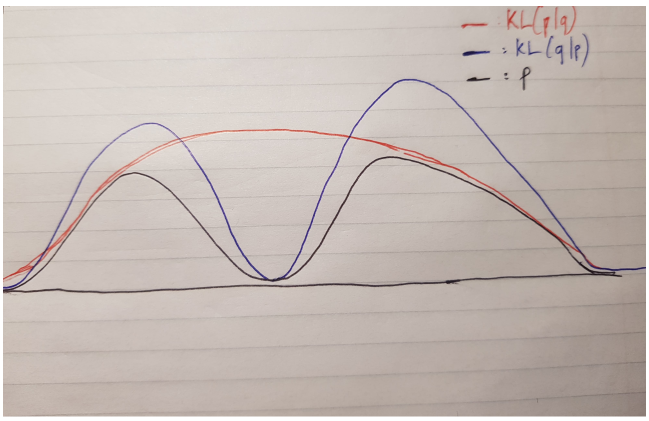



**Figure 6:** *The difference between using $KL(q(x)|p(x))$ and $KL(p(x)|q(x))$*

## Question 6

Variational Bayes for Ising Model in denoising image. The main idea is that we approximate distribution $q(x)$ which is close to the posterior distribution $p(x|y)$ where the distance is measured by KL divergence.

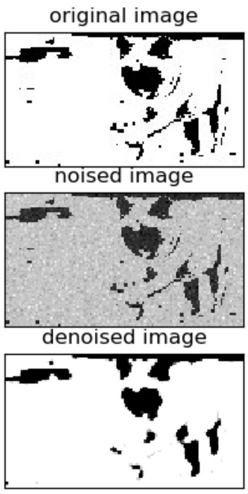



**Figure 7:** *Denoise image*

## Question 7

The result of Variational Bayes method is less noisy than the result of sampling method. However, it fails to recover the pixels in small area (the noise) of original image.

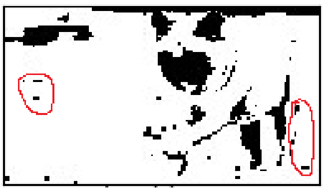


**Figure 8:** *The regions that Variational Bayes method fails to recover in original image*

## Question 8

We use sampling based method to solve image segmentation problem. We use the relationship between the pixel and its neighbours as a prior and the likelihood is does it belong to foreground or background. The probability of being foreground or background is calculated by normalzing the histogram of 3 channels.

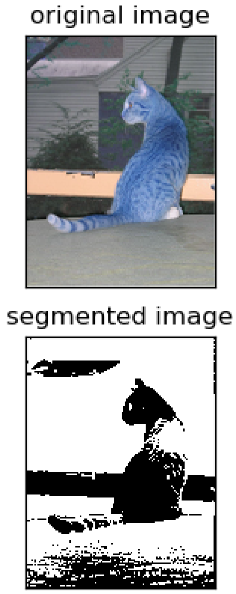



**Figure 9:** *The result of image segmentation*

## Question 9

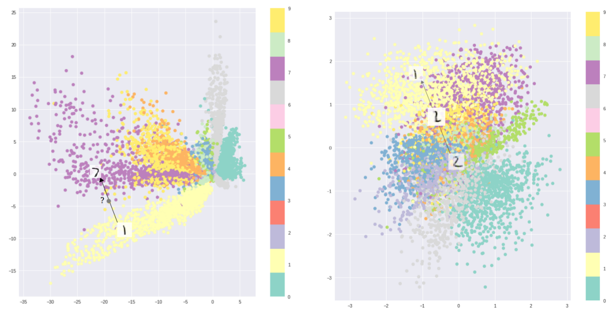
Autoencoders are simply two connected networks — an encoder and a decoder. An encoder takes in an input and converts into a much smaller representation (encoding). The decoder then converts the representation back to the original input.

*Input* $\rightarrow$ **Encoder** $\rightarrow$ Compressed Representation $\rightarrow$ **Decoder** $\rightarrow$ *Output*

In standard autoencoders, the encoders are trained to generate encodings that can only reconstruct the original input. To understand this in detail, we need to look at the space the inputs and encodings lie, i.e. the latent space. The encodings in this latent space are clustered, but the clusters are not continuous. If we pick a sample in this discontinuous region between clusters, the decoder will not know how to model the output. Because of this limitation, the standard autoencoders are mainly used de-noising tasks. The encoder encodes a noisy image and the decoder reconstructs the original image without noise.

Variational Autoencoders (VAEs) have only one main advantage over the standard autoencoders and that advantage makes them useful for generating models. That advantage is the continuity of their latent space. The continuity is achieved by making the encoder output not only an encoding, but also the means and standard deviations of the input, i.e. normal distribution. This allows the decoder to decode not only a single point of a sample, but the variety of points with similar attributes that surround the sample point. As a result,

[1]"Intuitively Understanding Variational Autoencoders", https://towardsdatascience.com/ intuitively-understanding-variational-autoencoders-1bfe67eb5daf



Standard Autoencoder VAE

**Figure 10:** *Latent spaces of the Standard Autoencoder and VAE* [1]

VAE can generate a similar output for a given sample input.

The inference assumption for VAE starts when a decoder has to decide which class the point belongs to before reconstructing a similar output. The decision is based on the mean and variance of an input. These two values are represented in the KL loss function that is used to spread the encodings across the latent space. Another important function that forms the encodings into clusters is the reconstruction loss.

So now we can sample a random vector from the latent space and the decoder will decode it. If there are many clusters of encodings, we could combine several to achieve a totally new output.

## Question 10

For this question we used a Github code [2].

The code generally follows the main idea in Question 9. Specifically, it loads MNIST data, defines the encoder and decoder using TensorFlow, computes losses, trains the network, and generates new data. It took about 40 minutes to train the data on out computer. Following are some example outputs.

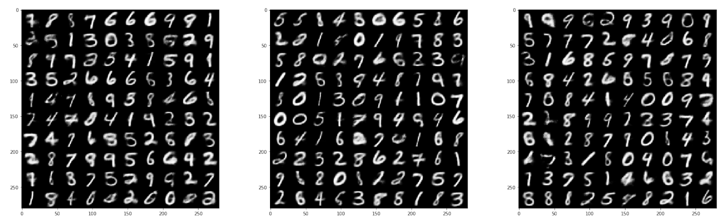


**Figure 11:** *Example numbers generated using VAE in TensorFlow*

This was a simple example, but we can see the huge future for VAE-like models as they could be used to create music, illustrations, books, etc. Both interesting and scary to see the future.

[2]https://github.com/shaohua0116/VAE-Tensorflow

# 2 Final thoughts

To sum up, the stochastic approach is friendly to be formulated but with the difficulties of access. This will be one of the main challenges for machine learning to modify the process of the application in stochastic approach. In other way, deterministic approach achieves the advantage of efficiency but also has the limitation of accuracy. During this assignment, we have learnt how to develop models. This fundamental technique allows us to take a big step to the actual application with real data. We are looking forward to accepting any advanced challenges in machine learning in the future and realising the importance of a strong fundamental background that we learnt today from our incredible lecturer in University of Bristol!