Deep Variational Inference for q with Gamma Distribution

Author: Natan Katz

Bayesian Inference

In the world of ML, Bayesian inference is often treated as the peculiar enigmatic uncle. On one hand Bayesian inference offers a massive exposure to theoretical scientific tools from mathematics statistics and physics. Furthermore it carries an impressive historical legacy of scientific breakthroughs. On the other hand, will tell you nearly every data scientist: it hardly ever works.

This post is comprised of two parts:

1. A brief survey on the appearance of variational inference (VI)
2. A description of my trials to solve a VI problem using DL techniques for data with Beta likelihood

The Bayesian problem

Bayesian problem can be described as follow: We have an observed data **X**

$x_1, x_2, x_3, \ldots \quad x_N$, where data can be numbers, categories or any other type of data that one can imagine. We assume that this data has been generated using a latent variable **Z.** We have four distributions:

1. P(**Z**) The <u>prior distribution</u> of the latent variable
2. P(**X**| **Z**) <u>The likelihood</u> – The distribution type of this function is determined by the data, (for example: if we have integers we may think of Poisson ,if we have positive numbers we may use Gamma).
3. P(**X**) – The distribution of the observed data.
4. P(**Z**|**X**) <u>The posterior distribution -</u> The probability to have a value of **Z** given a value of **X**

These distributions are bound together by <u>Bayes formula</u> :

$$\mathrm{P(Z|X)} = \frac{P(\mathrm{X|Z})P(\mathrm{Z})}{P(X)}$$

The Bayesian problem is therefore about finding the posterior distribution and the values of **Z**. The obstacle is that in real life models, P(X) is seldom tractable.

Until 1999 the modus operandi to solve Bayesian problems was sampling. Algorithms such as Metropolis Hasings or Gibbs used successfully for resolving posterior functions in wide scientific domains. Although they worked well they suffered from two significant problems:

- High variance
- Slow convergence

In 1999 Michael Jordan published a paper "An introduction to variational graphic model" (This is the year that the more famous MJ quitted, which raises the question whether universe is governed by "MJ preservation law")

In this paper he described an analytical solution to Bayes problem. In the core of this solution he suggested that rather chasing the posterior function with sampling, we can introduce a distribution function **q** of the latent variable **Z .** By setting apriorically **q**'s family functions and a metric we can approximate the posterior function using this **q**. For this solution he used Euler Lagrange equation which is a fundamental equation in **Calculus of Variations** ,that brought the notion**: Variational Inference (VI)**.

What is variational inference?

In this section we describe VI in more details. Recall that we aim to approximate a posterior function P(**Z**|**X**) using a function **q** which is the distribution of **Z**. The metric that Jordan suggested was KL divergence (https://projecteuclid.org/download/pdf_1/euclid.aoms/1177729694)

This idea of analytical solution reduces the obstacles of high variance and the slow convergence. On the other hand, since we set a family of functions we introduce some bias. The following formula illustrates the offered solution:

$$\begin{aligned}\log p(\mathbf{x}) = {} & \mathrm{KL}(q(\mathbf{z}\,;\,\lambda)\;\|\;p(\mathbf{z}\mid\mathbf{x})) \\ & + \;\mathbb{E}_{q(\mathbf{z}\,;\,\lambda)}\big[\log p(\mathbf{x},\mathbf{z}) - \log q(\mathbf{z}\,;\,\lambda)\big]\end{aligned}$$

The LHS does not depend on **Z** therefore it can be considered as a constant. Minimizing KL divergence is equivalent to maximizing the second term: "Evidence Lowe Bound" (ELBO)

$$\mathrm{ELBO}(\lambda) = \;\mathbb{E}_{q(\mathbf{z}\,;\,\lambda)}\big[\log p(\mathbf{x},\mathbf{z}) - \log q(\mathbf{z}\,;\,\lambda)\big]$$

As a function with pre-defined shape, **q** has constants that are denoted by λ.

We can consider VI as an ELBO maximization problem:

$$\lambda^* = \arg\max_{\lambda} \mathrm{ELBO}(\lambda).$$

Physics!

The motivation of Jordan to use this solution came from works of Helmohltz and Boltzmann in thermo dynamics. The ELBO function is extremely similar to Helmholtz free energy where the first term is the energy and second is the entropy of **q.** Maximizing the ELBO is about increasing the energy and reducing the entropy. Jordan used mean field theorem (MFT) from Ising model, there magnetic spins are assumed to be uncorrelated which simplified the way to handle **q**. Ising problem has an exponential solution (Boltzmann distribution), which became the common choice for the shape of **q**

VI – Example

I will now give an example of how one formulate a VI problem. We will assume that we have a data of real numbers with a Gaussian likelihood. The latent variable **Z ,** is therefore the pair

Z ={μ, τ }

Where the setting of the problem is as follow:

$$\begin{aligned} \tau &\sim \mathrm{Gamma}(a_0, b_0) \\ \mu &\sim \mathcal{N}(\mu_0, (\lambda_0 \tau)^{-1}) \\ \{x_1, \ldots, x_N\} &\sim \mathcal{N}(\mu, \tau^{-1}) \end{aligned}$$

$$p(\mathbf{X}, \mu, \tau) = p(\mathbf{X} \mid \mu, \tau) p(\mu \mid \tau) p(\tau)$$

Using mean field theorem, **q** can be written as a product of **q** (μ, τ) =$q_{\mu}\ q_{\mathrm{T}}$

Where

$$\ln q_{\mu}^*(\mu) = \mathrm{E}_{\tau}[\ln p(\mathbf{X} \mid \mu, \tau) + \ln p(\mu \mid \tau) + \ln p(\tau)] + C$$

$$\ln q_{\tau}^*(\tau) = \mathrm{E}_{\mu}[\ln p(\mathbf{X} \mid \mu, \tau) + \ln p(\mu \mid \tau)] + \ln p(\tau) + \text{constant}$$

A detailed formulation can be found here https://towardsdatascience.com/variational-inference-in-bayesian-multivariate-gaussian-mixture-model-41c8cc4d82d7)

One can see that such formulation requires massive analytical calculations for every model. We will discuss this in one of the sections below when we will talk about BBVI.

VI has been strongly boosted in 2002 when Blei published his "Latent Dirichlet allocation", where he used VI for topics extraction

There are several available tools to be used for VI such as Edward, Stan, PyMC3 and Pyro

Using VI for distribution tail's Inference

The problem I had to handle was identifying a tail of an observed data. The data itself is a set of probability values thus we assume that the likelihood has a **Beta** distribution. Before I discuss the modeling steps I will give an example for tail's typical problem.

Let's assume that we have a model that predicts a disease with accuracy of 99%. We can assume that 0.1% of the population is sick. Each day 100,000 people are tested. 990 healthy people will hear that they are sick. This mistake is costly, since they are going to do medical procedures. Such problems occur mainly because models are often trained to optimize accuracy under symmetric assumptions. If we rely on our models, tiny changes in the threshold may cause huge damages. We need therefore to understand the 1% tail as accurately as possible.

VI as analytical solution seems to be an interesting tool for achieving this purpose. Surprisingly I found out that the world of DL didn't massively embraced VI, which indeed encouraged me to walk on this way.

Deep learning for Variational Inference

In this part I will describe my trials to use DL for solving VI problems.

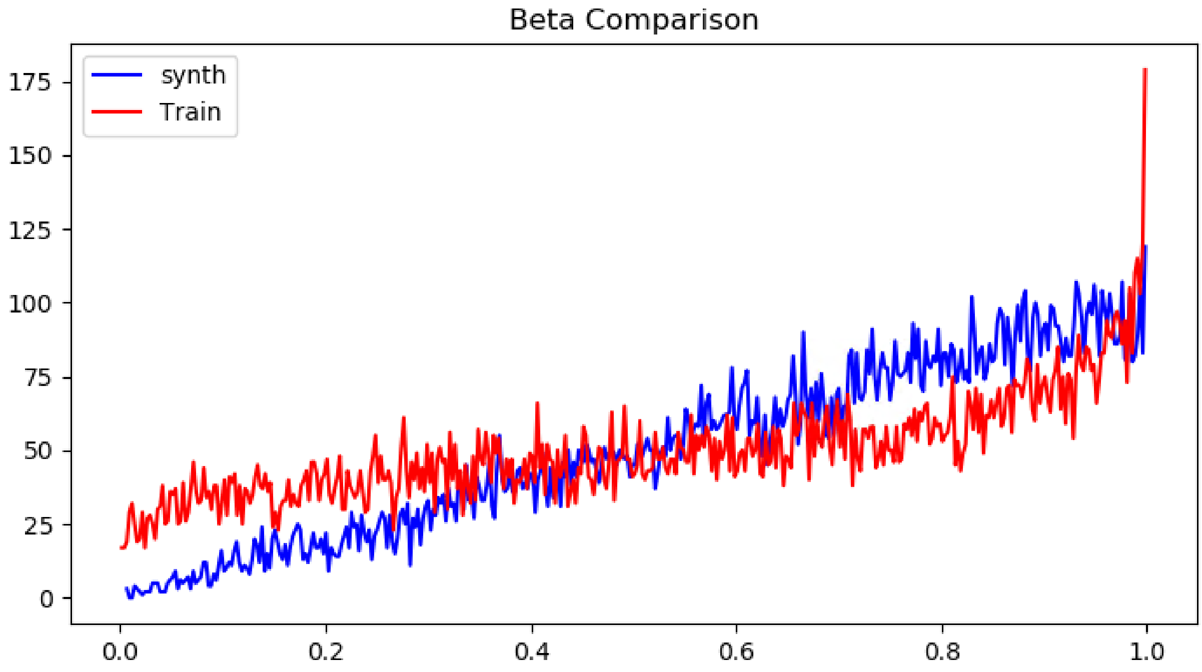
Recall that our data's likelihood has a Beta distribution. This distribution has an extremely difficult prior. We therefore take Gamma as a prior since it has a support on the positive real line. In VI terminology Gamma is the prior and **q** distribution. Hence we train Gamma to sample **Z** which itself is the pair {α, β} of Beta.

Mixture Density Networks -MDN

The first framework to be considered was MDN (Mixture Density Networks) . This class of networks that proposed by Bishop (http://publications.aston.ac.uk/id/eprint/373/) aims to learn parameters of a given distribution. We modify the loss: rather using likelihood we use ELBO.

In all the plots, red represents our VI based sample and blue is a synthetic sample that has been created by a random engine.

Here is atypical outcome of our MDN trials:



The VI based sample is relatively close to the synthetic sample. However, it is definitely improvable. Moreover while in the inner part and the left side (which is the interested tail from my perspective) are fairly similar, in the upper tail we suffer from huge differences.

We wish to try additional tools.

Reinforcement Learning – Actor Critic

Variational inference is often interpreted as a reinforcement learning problem (see http://www0.cs.ucl.ac.uk/staff/d.silver/web/Publications_files/viral.pdf) . We consider the **q** function as the policy function, **Z** is the option and the reward is the ELBO. When we use RL tools for such problem, we must notice that two modifications are needed:

- We don't have an episode -We have IID samples
- The options are not discrete but Gamma distributed

In order to resolve the first clause we have to modify the Q learning update formula:

$$Q_{k+1}(s,a) \;= r\,(s,a) + \gamma \max_{a'} Q_k(s',a')$$

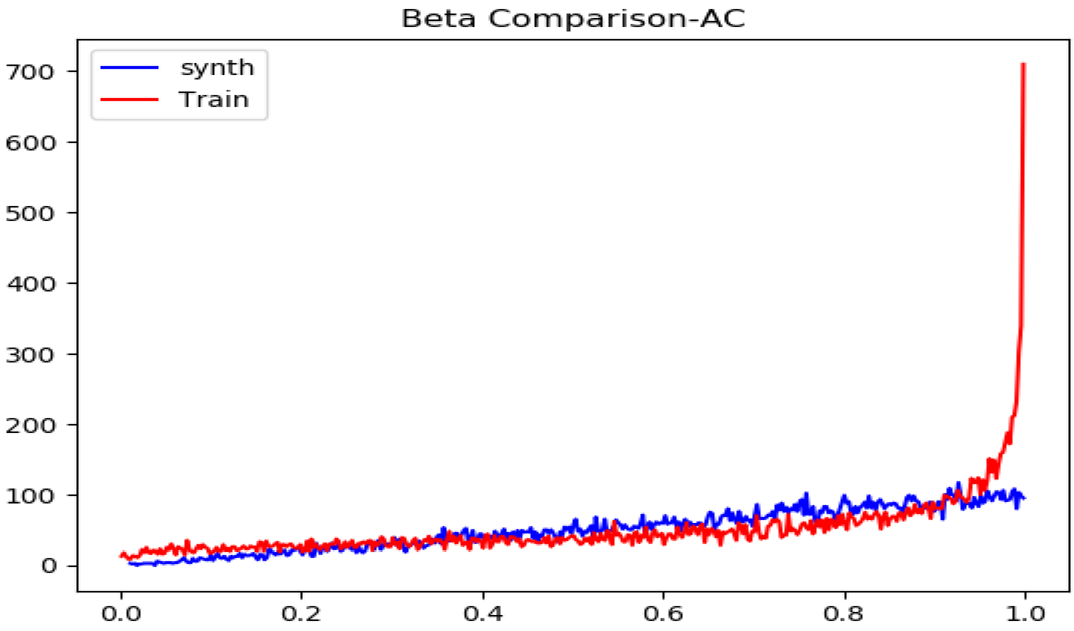
As there is no real episode one has to remove the second term, since it indicates the future expected rewards.

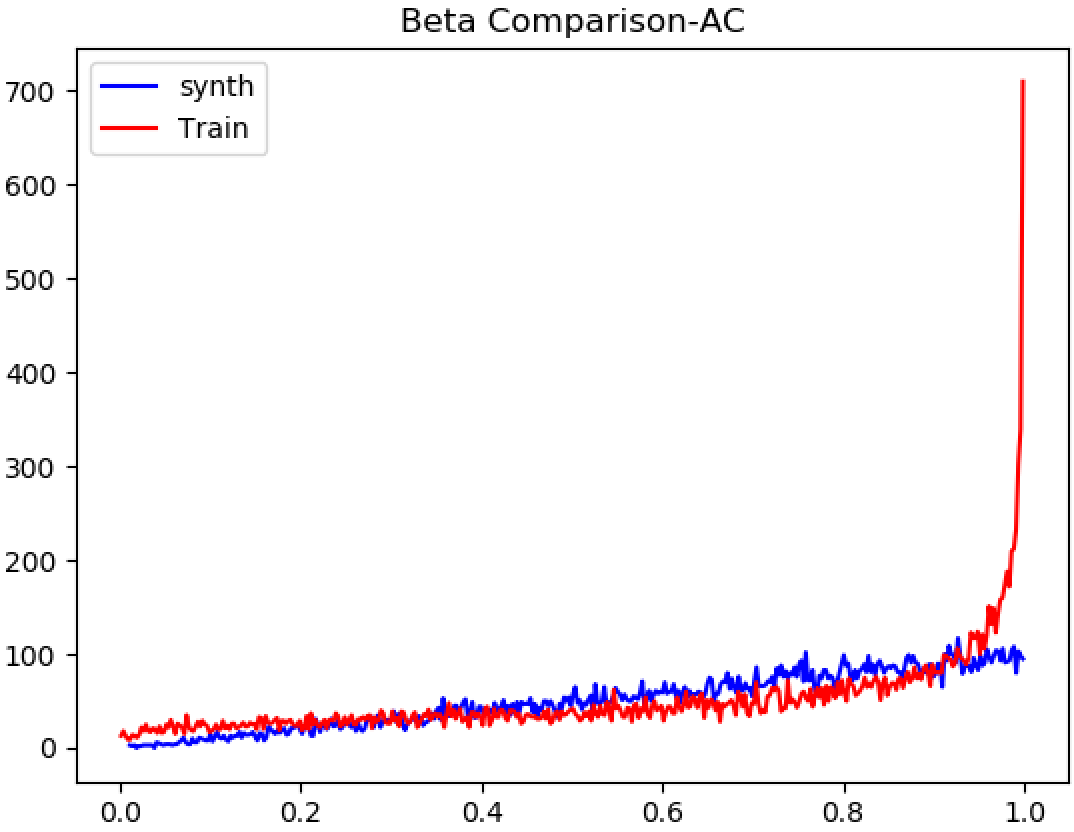
Regarding the second clause, we have to train a Gamma form policy function (a nice motivation exists here http://proceedings.mlr.press/v70/chou17a/chou17a.pdf)

In order to train policy function we will use actor-critic with experience replay https://towardsdatascience.com/understanding-actor-critic-methods-931b97b6df3f

and add our modifications.

These are the plots that we got:

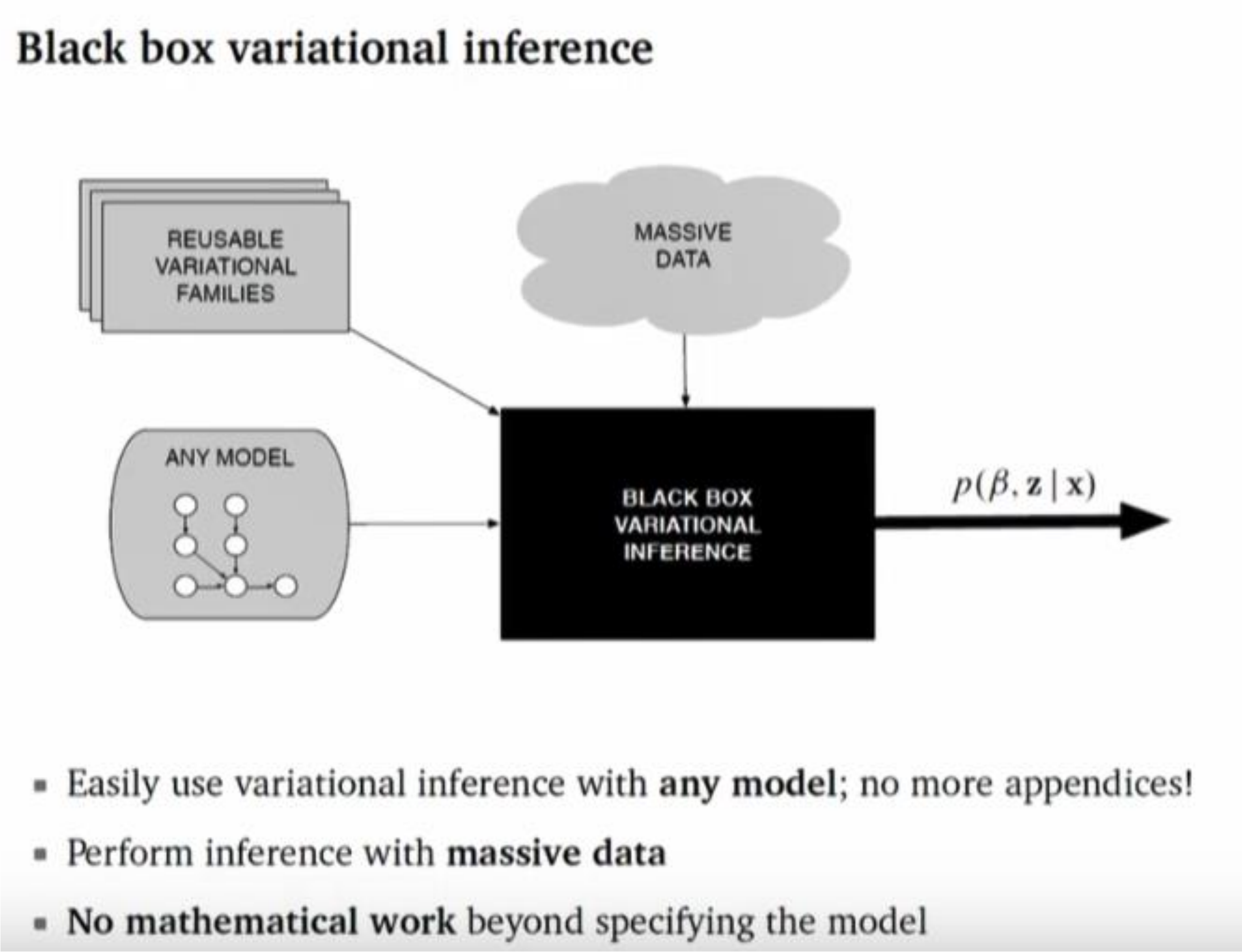

Beta Comparison-AC
synth
Train
700
600
500
400
300
200
100
0
0.0
0.2
0.4
0.6
0.8
1.0


Beta Comparison-AC
synth
Train
700
600
500
400
300
200
100
0
0.0
0.2
0.4
0.6
0.8
1.0

The outcomes are fairly good in an interval[0,1-a) for a small positive. We totally fail to model the upper tail.

Black-BOX Variational Inference

The appearance of VI provided new directions for studying posterior distributions. Nevertheless it imposed a massive analytical computations of the expectations (see the examples of Gaussian dist. Above). There was a clear objective to find a more generic scheme that reduce the amount of analysis and better handles huge amount of data. The slide below is Blei's description for this need:



Black box variational inference

REUSABLE VARIATIONAL FAMILIES

MASSIVE DATA

ANY MODEL

BLACK BOX VARIATIONAL INFERENCE

$p(\beta, \mathbf{z} \mid \mathbf{x})$

- Easily use variational inference with **any model**; no more appendices!
- Perform inference with **massive data**
- **No mathematical work** beyond specifying the model

The offered solution used *stochastic optimization*. A class of algorithms that has been presented by Robbins and Monro (A stochastic Approximation method 1951)

Robbins-Monro Algorithm

The Robbins Monro algorithm aims to solve a root problem: Let F a function and α a constant. We assume a unique solution of the eq:

$$F(x) = \alpha$$

We wish to find the x^* such that solves this equation.

Assume that F is not observable but there exists a random variable T such that

$$E[T(x)] = F(x)$$

Robbins and Monro have shown that for certain regulations the following algorithm converges with L^2 :

$$x_{t+1} = x_t - a_n \ (T(x_t) - \alpha)$$

Where

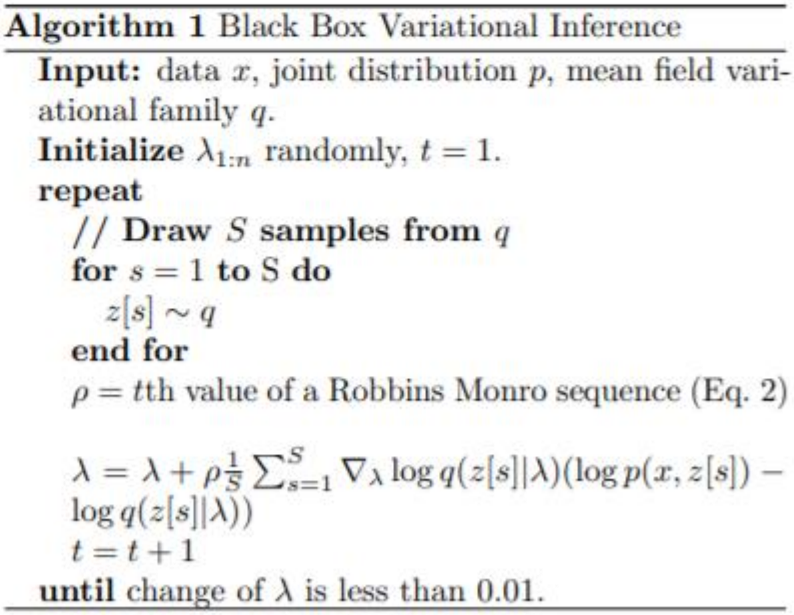
a_1, a_2 ………..a_n are positive numbers They satisfy:

$$\sum_{i=0}^{\infty} a_i = \infty \ , \ \sum_{i=0}^{\infty} a_i^2 < \infty$$

This sequence is called Robbins Monro sequence.

Back to BBVI

In a paper from 2013 https://arxiv.org/pdf/1401.0118.pdf ,Blei suggested two steps improvement for VI by using Robbins-Monro sequence. This algorithm has been denoted "Black Box Variational Inference" (BBVI). The first step is presented below:



```
Algorithm 1 Black Box Variational Inference
Input: data $x$, joint distribution $p$, mean field variational family $q$.
Initialize $\lambda_{1:n}$ randomly, $t = 1$.
repeat
    // Draw $S$ samples from $q$
    for $s = 1$ to S do
        $z[s] \sim q$
    end for
    $\rho$ = $t$th value of a Robbins Monro sequence (Eq. 2)
    $\lambda = \lambda + \rho \frac{1}{S} \sum_{s=1}^{S} \nabla_\lambda \log q(z[s] || \lambda)(\log p(x, z[s]) - \log q(z[s] || \lambda))$
    $t = t + 1$
until change of $\lambda$ is less than 0.01.
```

The main idea is to construct a Monte Carlo estimator to the ELBO gradient and use Robbins Monro sequence as described in the section beyond. This algorithm (as many other Monte Carlo algorithms) suffers from high variance. Blei suggested two methods to reduce this variance:

- <u>Control Variates</u> – for the estimator $f^{\wedge}$ he searches a function **g** such that:

$$E_q\,[f^{\wedge}] = E_q\,[\mathbf{g}] \quad \text{and} \quad Var_q\,[f^{\wedge}] < Var_q\,[\mathbf{g}]$$

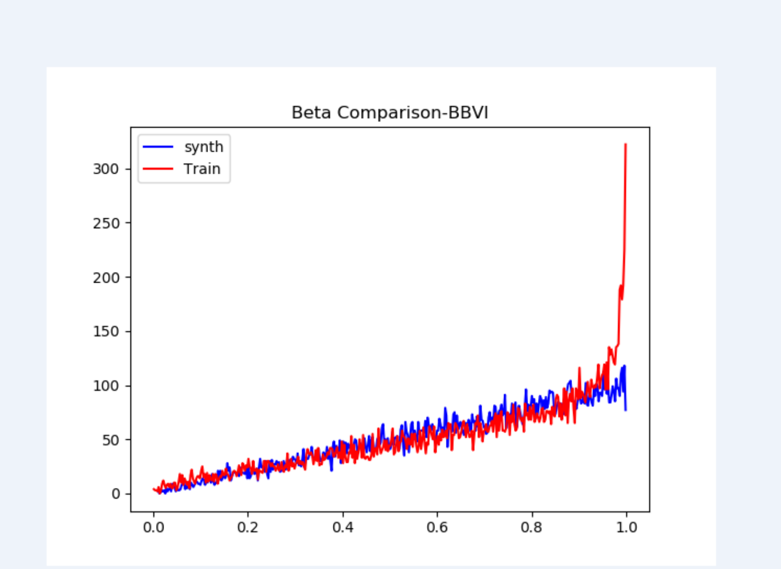
2 <u>Rao- Blackwell</u> – Rao Blackwell theorem claims that if X is an estimator of Z and T is a sufficient statistics then for Y = E(X|T)

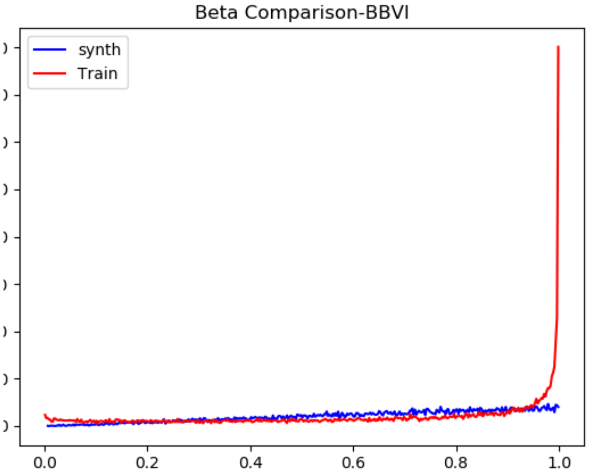
$$\mathrm{E}[(Y-Z)^2\,] \le \mathrm{E}[(X-Z)^2\,]$$

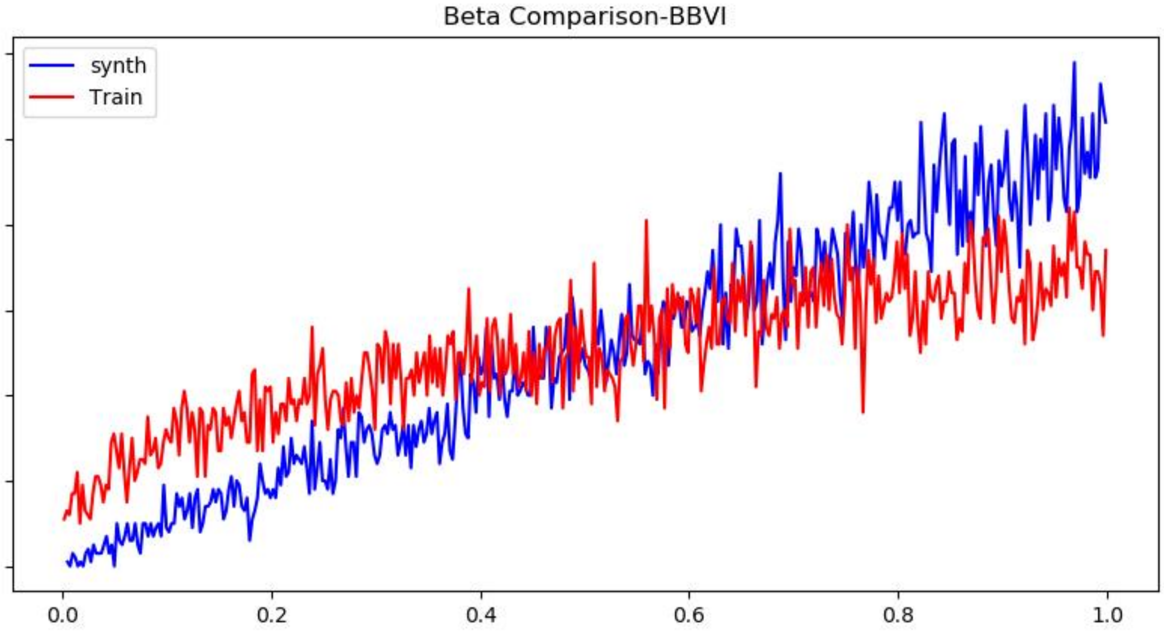
In the paper Blei uses Rao Blackwell to create an estimator for each value of λ_i by using the Rao Blackwell property for conditional distribution.

<u>Now Plots</u>

I used the BBVI without the variance reduction techniques .


Beta Comparison-BBVI
synth
Train
300
250
200
150
100
50
0
0.0
0.2
0.4
0.6
0.8
1.0


Beta Comparison-BBVI
synth
Train
0.0
0.2
0.4
0.6
0.8
1.0



We can see that we reduced the tail issues in comparison to AC, but still suffer from this problem. In the last plot we did not have tail issues but the approximation became weaker

Results Summary

We saw that using Deep learning tools for VI works with partial success. We performed fairly good approximation within the interval but fail to approximate the upper tail. Moreover in all the frameworks I suffered a certain level of instability.

There are several plausible reasons:

DL- Issues

- Architecture is not sophisticated enough
- More Epochs

Gamma- Beta issues

- Traditionally, DL problems are studied on common distributions such as Gaussian or Uniform. These distributions don't have parameters dependent Skewness and Kurtosis. This is not the case in Gamma, where the parameters that are trained by the engine fully determine the Skewness and Kurtosis. It may require different techniques
- We use ELBO as loss function. This function relies strongly on the prior shapes. Gamma is not the real prior of Beta. Which may suggest that loss modifications are required

- Beta has a compact support. The fact that we have issues only near the tail (sort of a "singular point" of the distribution) may raise obstacles. This "singularity" issues are to be studied

We can summarize that these experiments have shown that VI can be resolved using DL mechanism. However, it requires further work. Nevertheless, we got some clarifications for the enigmatic uncle.

A torch code that mimics my trials (the real code has been used for some commercial work I did, sorry) exists here https://github.com/natank1/DL-_VI

Acknowledgements

I wish to acknowledge Uri Itai for fruitful discussions before and during the work and for his comprehensive reviewing, to Shlomo Kashani and Yuval Shachaf for their ideas and comments and to Leonard Newnham, for clarifying issues in the reinforcement learning world

References

- http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf
- https://people.eecs.berkeley.edu/~jordan/papers/variational-intro.pdf
- https://towardsdatascience.com/a-hitchhikers-guide-to-mixture-density-networks-76b435826cca
- https://publications.aston.ac.uk/id/eprint/373/1/NCRG_94_004.pdf
- https://www.ri.cmu.edu/wp-content/uploads/2017/06/thesis-Chou.pdf
- http://www.michaeltsmith.org.uk/#
- https://github.com/trevorcampbell/ubvi/tree/master/examples
- https://arxiv.org/pdf/1811.01132.pdf
- http://proceedings.mlr.press/v70/chou17a/chou17a.pdf
- https://towardsdatascience.com/understanding-actor-critic-methods-931b97b6df3f
- https://www.freecodecamp.org/news/an-intro-to-advantage-actor-critic-methods-lets-play-sonic-the-hedgehog-86d6240171d/
- http://incompleteideas.net/book/the-book-2nd.html
- https://arxiv.org/pdf/1401.0118.pdf
- https://towardsdatascience.com/variational-inference-in-bayesian-multivariate-gaussian-mixture-model-41c8cc4d82d7
- https://projecteuclid.org/download/pdf_1/euclid.aoms/1177729586
- http://edwardlib.org/tutorials/klqp
- https://mc-stan.org/users/interfaces/
- https://github.com/stan-dev/pystan/tree/develop/pystan
- https://docs.pymc.io/notebooks/getting_started.html
- https://projecteuclid.org/download/pdf_1/euclid.aoms/1177729694
- https://www.cs.princeton.edu/courses/archive/fall11/cos597C/lectures/variational-inference-i.pdf
- http://publications.aston.ac.uk/id/eprint/373/