Depth First Learning **Wasserstein GAN**

Week 1: 3 Feb, 2019

Scribed by: Timothy Reeder, Julia Rozanova & Sasha Naidoo

1.1 Overview

The desired outcome for this week is that there is a good understanding of *Kullback-Leibler Divergence*, maximising Likelihood, minimising Mean-Square error, and how they all relate to one another.

1.2 Brief Feedback

The course work for week 1 was very theoretical which confused the more practically inclined participants. In future it might be of benefit to have a broad discussion on the "bigger picture", maybe even a short practical approach. This would allow for a *fill the gaps* learning approach rather than a *build from the bottom up* approach, the latter being currently adopted.
These concerns would be valuable to have on the collaborative Google document. This will alow for constructive conversations before the short time we have when meeting.

1.3 KL Divergence

The weeks reading materials and exercises included the formula for *KL Divergence* and *R-KL Divergence*. A point of interest to someone first seeing these formulas is that: for the forward KL Divergence formula, the term that depends on the true distribution can be omitted - which is convenient since the true distribution is unknown. However, the Reverse KL Divergence formula depends on the true distribution. How then is this calculated?

This was left here to be taken up later.

An informal definition of KL Divergence is that it gives a distance/difference metric between two probability distributions.

The formula (1.1) was discussed and the simillarity it has with log-likelihood. P is your data and Q is your model. The formula can be re-written as (1.2) which shows you have the differential entropy of your data distribution and the expected value the model prediction. Which allows you to ignore the first term with a forward KL divergence. So when minimsing KL divergence, the goal is to maximise the remainf term (1.3). Which leaves us with the log likelihood of the model which is then maximum likelihood estimation.

$$D_{KL}(P||Q) = \int_{\infty}^{-\infty} p(x)log(\frac{p(x)}{q(x)})dx \tag{1.1}$$

$$D_{KL}(P||Q) = \int_{\infty}^{-\infty} p(x)log(p(x))dx - \int_{\infty}^{-\infty} p(x)log(q(x))dx \tag{1.2}$$

$$min(D_{KL}(P||Q)) = -max(\int_{\infty}^{-\infty} p(x)log(q(x))dx) \tag{1.3}$$

1.4 Side Notes

- When you have a probability density, you can have densities greater than 1. This is unintuitive since when considering a probability mass, it is between 0 and 1 always. However, this is not the case with a probability density function. This is when scaling occurs. Since if you scale the x-axis, the random variable will need to scale to fit. This could result in a probability density function being greater than 1, even though the mass will always remain between 0 and 1.
- The KL divergence of a distribution to itself is 0.

1.5 Equivalence between minimizing KL and maximizing likelihood

- It was deliberated upon for some time that we could say that

$$\frac{1}{N}\sum_{i=0}^{N}\log(q(x_i)) = \mathbb{E}_{x\sim p}[\log(q(x_i))]. \tag{1.4}$$

- Martin pointed out that the above equality is a consequence of the "Law of large numbers" (for a large number of data points, the average value of $f(x)$ (in this case $(q(x_i))$) is a good approximation of its expectation).
- He also advised us that problems arise when you try and take the KL divergence between a discrete and a continuous distribution.
- It is actually a bit more complicated: basically, one probability distribution should be absolutely continuous with respect to another if we wish to meaningfully take the KL divergence between them. There are certain requirements that need to be met, such as simultaneously assigning a 0 probability to sets of measure 0 (the details should be looked into more formally, if the reader is interested).

1.6 Reverse KL divergence

- The formula for reverse KL divergence was inspected closely, and it was remarked that the

$$\mathbb{E}_{x\sim q_\theta}[\log(p(x))] \tag{1.5}$$

  term was a spanner in the works, as we do not have access to the "true" distribution $p(x)$ of the underlying data.
- Martin asked a probing question: "Can you think of any real-life examples where reverse KL is used to fit a model?"
- This isn't really done, as this isn't practically possible without an explicit p(x), which we wouldn't have in such a context. However, reverse KL can be used in some situations: distillation problems, autoregressive models, parallel wavenet (where we have a good density model which is hard to sample from).

1.7 Equivalence between minimising MSE and maximising likelihood

- It was suggested that a Gaussian distribution was a useful thing to assume, especially so as to avoid assigning zero probability to any values that aren't exactly some observed data point.
- Writing out the expression for the likelihood explicitly, $p(x|\theta)$ was shown to be proportional to an exponential function with a sum of square errors on the exponent. By monotonicity, we could maximise the likelihood by minimising the sum of square errors in this exponent, which is equivalent up to a constant factor to the MSE technique.
- The key observation was how, in this particular problem with the assumption of a Gaussian distribution, the following are equivalent:
  - Maximising Likelihood
  - Minimising the KL Divergence
  - Minimising the Mean Square Error

1.8 Probability Distance Metrics

There are two classes of *probability distance metrics*: F-divergences and Integral Probability Metrics (IPMs). Loosely, the F-divergences determine distance using division of two probability distributions, $\frac{P(x)}{Q(x)}$ and the IPMs use the difference, $P(x) - Q(x)$. There are various metrics that fall within each class. For F-divergences, we have the Kullback-Liebler (KL) divergence, Reverse-KL divergence and Jenson-Shannon Divergence (JSD). For the IPMS, we have the Wassterstein distance (which is used in the WGAN) and the Maximum Mean Discrepancy (MMD). The Total Variation (TV) metric overlaps between these two classes. These metrics are depicted in Figure 1.1.

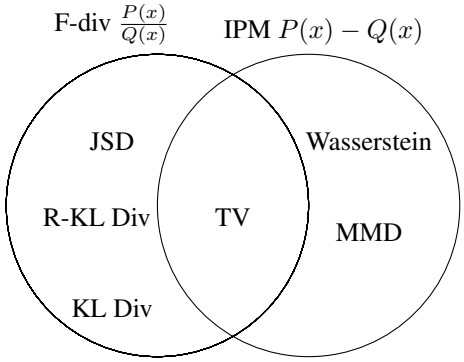

Figure 1.1: Classes of Probability Metrics.

The different metrics of measuring distance between two probability distributions forms the main diversity for many GAN algorithms. The various GAN techniques take advantage of the different properties of each distance to suit the application.

1.9 Maximum Likelihood

When trying to model a dataset, we are trying to find the parameters, θ, of a model that best represents the data. To do this, we take advantage of the relationship defined in Equation (1.1). For Maximum Likelihood Estimation, we are maximising the $p(x|\theta)$ in order to find the parameters θ. By focusing only on maximising this probability distribution, this often leads to over-fitting. In order to address this issue, we typically include a form of regularisation which takes into account our prior belief of the parameters, $p(\theta)$. For example, L2-regularisation imposes a restriction that the parameters of the model are Guassian distributed, such that the model is penalised for having large valued parameters.

$$p(\theta|x) \propto p(x|\theta) \tag{1.6}$$

$$= \frac{p(x|\theta)p(\theta)}{p(x)} \text{ (Bayes Rule)} \tag{1.7}$$

Additionally, determining $p(x)$ can be very computationally challenging. In order to deal with this, it is possible to compute a point estimate of this distribution, which is called a *MAP estimate*.