Week 5: 3 March, 2019

Scribed by: Sasha Naidoo

5.1 Momemtum Based Optimisers

An interesting improvement of WGAN-GP is its ability to use `Adam` and other momentum based optimisers as opposed to WGAN, which cannot (according to the paper). This is due to the critic loss being non-stationary i.e. the fundamental data distribution you are learning is changing over time. However, it was discussed that while WGAN-GP has been shown to work well with momentum based optimisers, WGAN is likely capable of working with momentum just generally not as well.

5.2 Weight Clipping vs Exploding/Vanishing Gradients (WGAN)

Weight clipping is introduced to maintain the 1-Lipschitz constraint as this has been seen to affect the network gradients in two ways.

- **Vanishing Gradients:** If you have a deep network and you clip all the weights (making them small or close to zero), the error signal propagated back through the network decays exponentially (as they are being multiplied by this small-valued weights matrix), which naturally results in vanishing gradients.
- **Exploding Gradients:** When using weight clipping, the loss function typically results in the weights being very close to the upper/lower bounds of the clipping interval (Figure 1b WGAN-GP), which exhausts the clipping constraint. If this constraint is below the ideal weight initialisation then you get vanishing gradients (as explained above) but conversely, if they are larger than the ideal point this results in exploding gradients.

In order to avoid issues with the gradients, the clipping constraint needs to be balanced between the vanishing case and the exploding case, which is typically very difficult.

5.3 Definition of $\hat{x}$

For WGAN-GP, we need to have the penalty on the gradients for both the real data and the generated data. $\hat{x}$ defines an interpolation between these two things where $\hat{x} \leftarrow \epsilon x + (1-\epsilon)\tilde{x}$ with x being the sampled data, $\tilde{x}$ being the generated data and ϵ, the mixing proportion. Based on Proposition 1 (in WGAN-GP paper), we note that if there is an optimal coupling between $\mathbb{P}_r$ (real data distribution) and $\mathbb{P}_g$ (generated data distribution), our optimal solution f^* is differentiable and if $\hat{x} \leftarrow \epsilon x + (1-\epsilon)\tilde{x}$ then we can conclude that f^* has gradient norm 1 everywhere under $\mathbb{P}_r$ and $\mathbb{P}_g$ and we have a 1-Lipschitz function f^*. In short, Proposition 1 tells you that the function that you are looking to recover f^*, exhausts the Lipschitz constraint at least on the straight lines between coupled points between the sampled data and the real data (which $\hat{x} \leftarrow \epsilon x + (1-\epsilon)\tilde{x}$ defines).

5.4 Batch Normalisation with WGAN-GP

Based on experiments done, it was highlighted that while batch-norm does seem to work with WGAN-GP (contrary to what is indicated in the paper), it does not seem to work reliably (i.e. $sim50\%$ of the time). An intuition into why it sometimes fails is that by using batch-norm it results in missing terms needed for the gradient penalty. This can cause certain terms to become arbitrarily large and as such the network could "cheat".

5.5 Overfitting

With WGAN, it is possible to train the critic to optimality and therefore if the optimal critic is struggling to tell the difference between real images and generated images, this tells us that the generated images are becoming more similar to the real images. This substantiates the claim that the WGAN (using weigh clipping) loss correlates with the sample quality. Therefore, the loss value is used a metric to measure overfitting for the WGAN as when the sample quality starts to increase, the loss goes up.

5.6 General Comments

- The aim of the GP method is to approximate the Kantorovich potential function that compares the distributions. Imagine two functions that are arbitrarily close in terms of mean-squared-error but give you completely different gradients and in essence, weight clipping forces the worst case scenario to happen a large percentage of the time, which leads poorer generator gradients. This is due to the network being incentivised to maximise the difference in expectation between value of the critic at the fake point and the value of the critic at the real points.
- During implementation, it was found that WGAN-GP was typically easier to get working for various architectures and produced fairly good results, in comparison to a standard WGAN. Overall, WGAN was found to need more iterations to produce similar results to standard GANs.