Report

Q1. To solve w, b for the global minimum of C (w, b, θ), we differentiate C w.r.t. to w and b respectively and let the derivatives be zero.

$$\frac{\partial \mathrm{C}}{\partial \mathrm{w}} = -\frac{(-1)(\exp(-wx-b))(-x)}{(1+\exp(-wx-b))^2} + \frac{(-1)(\exp(-w\theta z-b))(-\theta z)}{(1+\exp(-w\theta z-b))^2} = 0$$

$$\frac{\partial \mathrm{C}}{\partial \mathrm{w}} = \frac{-x(\exp(-wx-b))}{(1+\exp(-wx-b))^2} + \frac{\theta z(\exp(-w\theta z-b))}{(1+\exp(-w\theta z-b))^2} = 0 \qquad \dots\dots (1)$$

$$\frac{\partial \mathrm{C}}{\partial \mathrm{b}} = \frac{-\exp(-wx-b)}{(1+\exp(-wx-b))^2} + \frac{\exp(-w\theta z-b)}{(1+\exp(-w\theta z-b))^2} = 0 \qquad \dots\dots (2)$$

From (2), we have

$$\frac{\exp(-wx-b)}{(1+\exp(-wx-b))^2} = \frac{\exp(-w\theta z-b)}{(1+\exp(-w\theta z-b))^2}$$

Let $\mathrm{p} = \frac{\exp(-wx-b)}{(1+\exp(-wx-b))^2} = \frac{\exp(-w\theta z-b)}{(1+\exp(-w\theta z-b))^2} > 0$. Substituting p & q into (1), we have

$$-xp + \theta zp = 0$$
$$x = \theta z$$

We observe that w and b are missing in our final result after making derivative zero. Hence, we are unable to solve w and b for the global minimum of C. This shows that w and b do not tend to be a fixed point in Step 2.

In Step 3, to solve θ for the global maximum of C (w, b, θ), we differentiate C w.r.t. to θ and let the derivative be zero.

$$\frac{\partial \mathrm{C}}{\partial \theta} = 0 + \frac{(-1)(\exp(-w\theta z-b))(-wz)}{(1+\exp(-w\theta z-b))^2} = 0$$

$$\frac{\partial \mathrm{C}}{\partial \theta} = \frac{wz(\exp(-w\theta z-b))}{(1+\exp(-w\theta z-b))^2} = 0 \qquad \dots\dots (3)$$

As exp > 0, from (3), we have

$$wz = 0$$

This result doesn't involve θ and we are unable to solve θ for the global maximum of C. This suggests that the value of θ doesn't tend to a fix point.

Based on the calculations above, we see iterative steps 2 &3 do not converge.

Q2. WGAN is proposed to solve GAN's training instability problem as GAN does not converge in general. WGAN makes use of Wasserstein distance in its loss function and requires the discriminator (critic) to lie within the space of 1-Lipschitz functions, which is enforced by weight clipping. Gulrajani et al. [1] proved that weight clipping can lead to optimization difficulties and undesired behaviors. The first problem caused by weight clipping is that weight clipping tends to bias the critic towards much simpler functions. As proved by [1], optimal critic under the WGAN loss function has unit gradient norm almost everywhere. Under weight clipping, in order to achieve their maximum gradient norm of k, most neural network architectures will have to learn very simple functions. This makes the critic to ignore higher moments of the data distribution and be biased towards simple approximations. The second problem is exploding and vanishing gradients which are hardly avoidable given the optimization difficulty introduced by interactions between the weight constraint and the cost function.

References

[1] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. arXiv preprint arXiv:1704.00028, 2017.