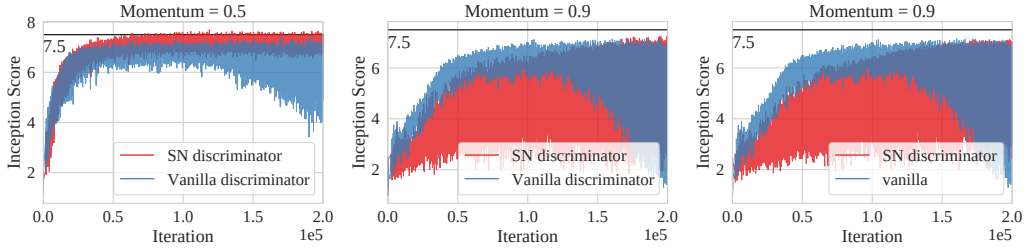
A Appendix

A.1 Additional experimental results and experimental methodology

Architectures. In all the two moons experiments, the DeepMLP has 4 layers and 100, 100, 100 and 1 output units respectively. The shallow MLP has 2 layers of 100 and 1 unit. All methods were trained for 100 iterations on 50 datapoints. The MNIST plots are obtained from MLP classifiers having 4 layers of 1000, 1000, 1000 and 10 units each and are trained for 500 iterations at batch size 100, reaching an accuracy of 95% on the entire test set. For the GAN CIFAR-10 experiments, we use the architectures specified in the Spectral normalization paper [5]. Unless otherwise specified, we use the default Adam optimizer [72] β_1 and β_2 parameters.

Computing the local Lipschitz constant in Figure 4. To compute the local Lipschitz function of the decision surface learned on two moons, we split the space into small neighborhoods (2500 equally sized grids). For each grid, we sample 2500 random pairs of points in the grid and report $\max \|f(\mathbf{x}) - f(\mathbf{y})\| \,/\, \|\mathbf{x} - \mathbf{y}\|$.

Spectral normalization. In Figure 6 we show that the effect of momentum on spectral normalization is independent of whether caching of the initialization vector for power iteration is performed or not.



(a) Low momentum: 0.5. (b) High momentum: 0.9. (c) High momentum with 10 power iteration caching and no caching.

Figure 6: The effect of momentum on spectral normalization on GAN performance. This shows that the iteration between momentum and spectral normalization is not due to the caching between iterations done for computational reasons.