- The competition between the two models theoretically leads to both developing better features to outperform each other.
- The unique solution of the adversarial process is the generator G capturing train data distribution and discriminator D predicting 0.5 for all images.
- Deep discriminative models are successful due to backpropagation, dropout etc. with ReLU providing proper gradient.
- Deep Generative models faced problem of intractable probability density function that arises during maximum likelihood estimation.
- Restricted Boltzmann machine : Generative stochastic artificial neural network that can learn a probability distribution over its set of inputs.
    - They have a partition function which integrates all its potential functions over all states of random variables, hence it is intractable.
    - This problem can be solved by using Markov chain Monte Carlo methods, but mixing approaches makes the algorithm problematic.
- Training Discriminator D in an inner loop will cause overfitting on training data.
- Minimizing log(1-D(G(z))) gives low gradient during the initial stage when G is learning. So maximising log(D(G(z))) gives good gradient at an early stage.
- While training D, it tries to discriminate between training data and generated data distributions. When G is training, it changes mapping from z to x distribution so that there is more chance of generated data being classified as training data.
- V (G, D) = ∫ pdata(x) log(D(x))dx + ∫ pz(z) log(1 − D(g(z)))dz = ∫ pdata(x) log(D(x)) + pg(x) log(1 − D(x))dx. D wants maximizing V(G,D).
- Maximum value of a.log(y) + b.log(1-y) is achieved at a/(a+b) for domain [0,1].
    - So D*(x) = pdata(x)/(pdata(x) + pg(x)). It is the optimal discriminator for a fixed G.
- Max V(G,D*) = C(G) = Ex~pdata [log D∗ G(x)] + Ez~pz [log(1 − D∗ G(G(z)))] = Ex~pdata [log D∗ G(x)] + Ex~pg [log(1 − D∗ G(x))].
- For G's objective, minimum of C(G) can be achieved only when pg = pdata with C*(G) = -[log4].
    - C(G) can be rewritten as C(G) = − log(4) + KL (pdata || pdata + pg /2) + KL (pg || pdata + pg /2) with two KL divergence terms. C(G) = − log(4) + 2 · JSD (pdata || pg) with Jenson-Shannon divergence which is non-negative. So, global minimum is obtained when pdata = pg with JSD = 0.
- V(G,D) can be expressed as U(pg,D) as func of pg, it is convex. supD U(pg, D) is convex in pg with a unique global optima which can be reached by performing gradient descent update of pg at every optimum D i.e supD U(pg,D).
- Probability of the test set was estimated as log-likelihood by fitting a Gaussian parzen window to samples generated by G. σ parameter was cross validated on validation set. This method of estimating log-likelihood has high variance but was the best method available then.
- Nearest neighbour of randomly picked generated images taken from training data is shown to show that training data images have no been memorized but original data is generated.

- Disadvantages :
  - No explicit representation of pg.
  - D must be synchronized with G while training as if many training steps of G without training D may lead to G producing the same images from different z to fool discriminator.
- Advantages :
  - No markov chains needed and hence no inference at learning.
  - Any differentiable func can be incorporated in the model.
  - Learning is not directly from training images, but from backprop which reduces chances of copied training images.
- Features from the discriminator or inference net could improve performance of classifiers when limited labeled data is available.
- Learned approximate inference can be performed by training an auxiliary network to predict z given x with fixed generator net.