Variational Autoencoder

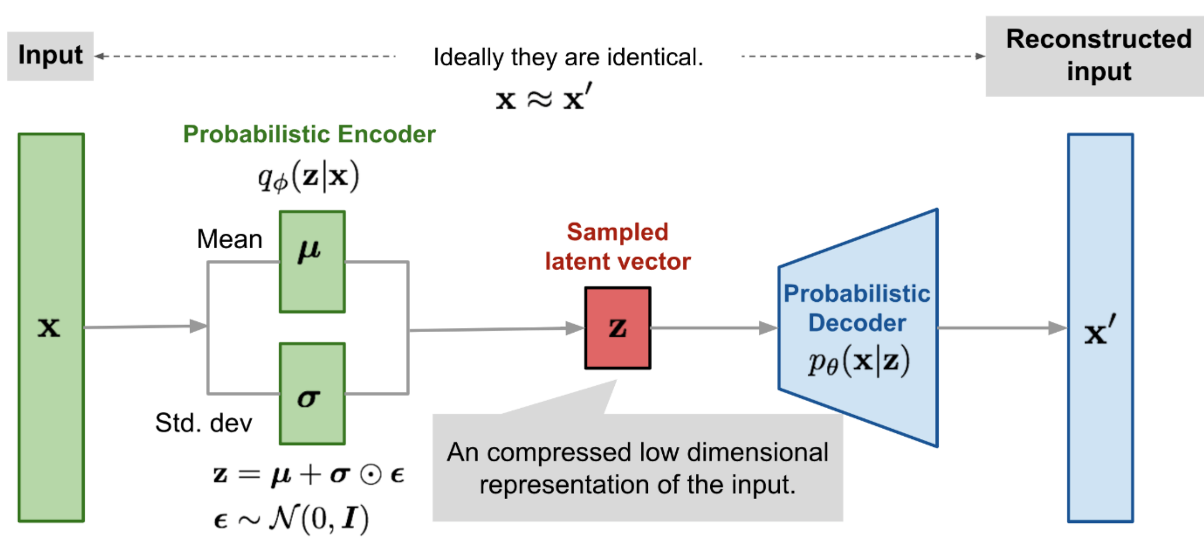
1- $(Motivation)\ Model\ P_D(x)\ Distribution\ with\ NN(\theta)\ to\ take\ new\ samples\ from\ it.$
$assume\ there\ is\ latent\ representation\ (we\ don't\ know\ it)\ which\ data\ points\ generated\ from\ it.$

2- $(objective)\ ML:\ \max_{\theta}\ \mathbb{E}_{P_D(x)}[\log p_\theta(x)]\ \ (marginalized\ on\ z:\ p_\theta(x) = \int p_{\theta(x,z)}\,dz\ )$

$$\mathbb{E}_{P_D(x)}[\log p_\theta(x)] = \mathbb{E}_{P_D(x)}\left[\log \int_z p_\theta(x|z)\, p(z)dz\right]\quad (intractable)$$

$$\log \int_z p_\theta(x|z)\, p(z)dz = \log \int_z \frac{q_{\phi(z|x)}}{q_{\phi(z|x)}}\ p_\theta(x|z)\, p(z)dz = \log \int_z \frac{p_\theta(x|z)\, p(z)}{q_{\phi(z|x)}}\ q_{\phi(z|x)}dz$$
$$= \log \mathbb{E}_{q_{\phi(z|x)}}\left[\frac{p_\theta(x|z)\, p(z)}{q_{\phi(z|x)}}\right] \geq \mathbb{E}_{q_{\phi(z|x)}}\left[\log \frac{p_\theta(x|z)\, p(z)}{q_{\phi(z|x)}}\right]\quad (jensen's\ inequality)$$
$$= \mathbb{E}_{q_{\phi(z|x)}}[\log\ p_\theta(x|z)] + \mathbb{E}_{q_{\phi(z|x)}}\left[\log \frac{p(z)}{q_{\phi(z|x)}}\right] = \mathbb{E}_{q_{\phi(z|x)}}[\log\ p_\theta(x|z)] - KL[q_{\phi(z|x)}||p(z)]$$

$$\boldsymbol{\mathcal{L}_{ELBO}(x) = \mathbb{E}_{P_{D(x)}}[\mathbb{E}_{q_{\phi(z|x)}}[\log\ p_\theta(x|z)] - KL[q_{\phi(z|x)}||p(z)]}$$



Generative Adversarial Network

3- $\underset{G}{\operatorname{argmin}}\,\underset{D}{\operatorname{argmax}}\,\mathbb{E}_{x\sim P_{Data}}[\log D(x)] + \mathbb{E}_{x\sim P_G}[\log{(1-D(x))}]$

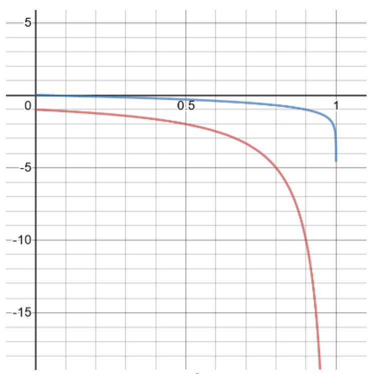
$\boldsymbol{gradient\ ascent\ on\ discriminator}:\ \underset{D}{\operatorname{argmax}}\,\mathbb{E}_{x\sim P_{Data}}[\log D(x)] + \mathbb{E}_{x\sim P_G}[\log{(1-D(x))}]$

$\boldsymbol{gradient\ descent\ on\ generator}:\ \underset{G}{\operatorname{argmin}}\,\mathbb{E}_{x\sim P_{Data}}[\log D(x)] + \mathbb{E}_{x\sim P_G}[\log{(1-D(x))}]$

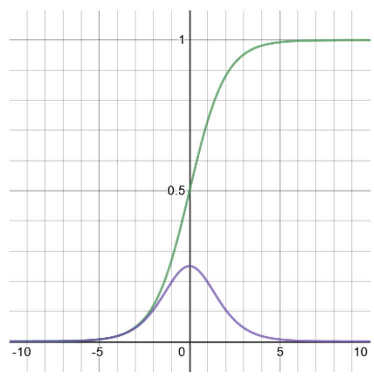
$= \underset{G}{\operatorname{argmin}}\,\mathbb{E}_{x\sim P_G}[\log{(1-D(x))}] = \underset{G}{\operatorname{argmin}}\,\mathbb{E}_{z\sim P_z}[\log{(1-D(G(z)))}]$

4- Problems:

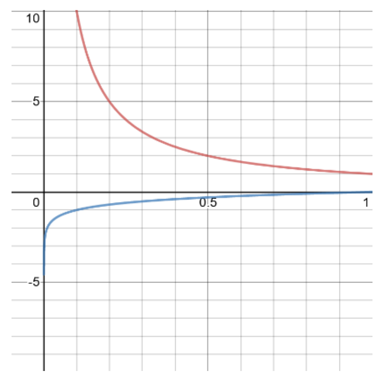
1- Vanishing Gradient:



$log(1-x)$ $\frac{\partial}{\partial x}(\log{(1-x)})$



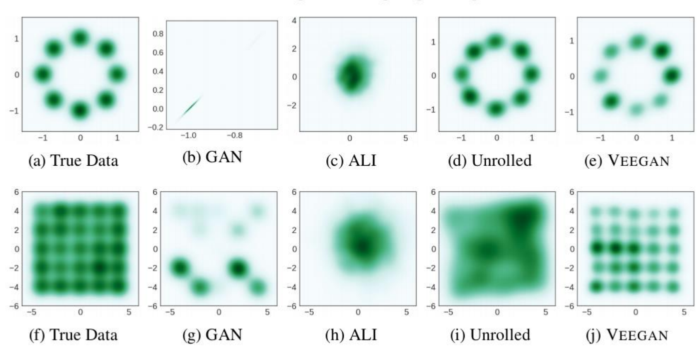
$\sigma(x)$ $\frac{\partial}{\partial x}(\sigma(x))$



$log(x)$ $\frac{\partial}{\partial x} log(x)$

2- Mode Collapse:

Figure 2: Density plots of the true data and generator distributions from different GAN methods trained on mixtures of Gaussians arranged in a ring (top) or a grid (bottom).



3- Convergence / Stability

4- Evaluation Metric (Quality + Diversity)

Problems

1- $\boldsymbol{\mathcal{V}(G,D) = \mathbb{E}_{x\sim P_{Data}}[\log D(x)] + \mathbb{E}_{z\sim P_z}\left[\log\left(1 - D(G(z))\right)\right]}$

If discriminator has infinite capacity, what is the optimal Discriminator.

$$\mathcal{V}(G,D) = \int P_{data}(x)\log\big(D(x)\big)\,dx + \int \log\big(1 - D(x)\big)\; p_z(z)\,dz$$

$$= \int (p_{data}(x)\log\big(D(x)\big) + \; p_G(x)\log\,(1 - D(x)))\,dx$$

$$\frac{\partial}{\partial D}\,\big(p_{data}(x)\log\big(D(x)\big) + \; p_G(x)\log\big(1 - D(x)\big)\big) = \frac{p_{data}(x)}{D(x)} - \frac{p_G(x)}{1 - D(x)} = 0$$

$$D^*(x) = \frac{p_{data}(x)}{p_G(x) + p_{data}(x)}$$

2- If discriminator update optimally ($\boldsymbol{D^*}$). Show that $\boldsymbol{\underset{G}{\operatorname{argmin}}\, \mathcal{V}(G,D^*) = \underset{G}{\operatorname{argmin}}\, JSD(p_{data}\,||\,p_G)}$.

$$\mathcal{V}(G,D^*) = \mathbb{E}_{x\sim P_{Data}}[\log D^*(x)] + \mathbb{E}_{x\sim P_G}[\log\left(1 - D^*(x)\right)]$$

$$\mathbb{E}_{x\sim P_{Data}}\left[\log\frac{p_{data}(x)}{p_G(x) + p_{data}(x)}\right] + \mathbb{E}_{x\sim P_G}[\log\left(1 - \frac{p_{data}(x)}{p_G(x) + p_{data}(x)}\right)]$$

$$= \int (p_{data}(x)\log\left(\frac{p_{data}(x)}{p_G(x) + p_{data}(x)}\right) + \; p_G(x)\log\,(1 - \frac{p_{data}(x)}{p_G(x) + p_{data}(x)})\,dx$$

$$= \int (p_{data}(x)\log\left(\frac{p_{data}(x)}{p_G(x) + p_{data}(x)}\right) + \; p_G(x)\log\,(\frac{p_G(x)}{p_G(x) + p_{data}(x)})\,dx$$

$$= \; -2log2 + \int (p_{data}(x)\log\left(\frac{p_{data}(x)}{\frac{p_G(x) + p_{data}(x)}{2}}\right) + \; p_G(x)\log\,(\frac{p_G(x)}{\frac{p_G(x) + p_{data}(x)}{2}})\,dx$$

$$= \; -2log2 + KL(p_{Data}(x)||\frac{p_G(x) + p_{data}(x)}{2}) + \; KL(p_G(x)||\frac{p_G(x) + p_{data}(x)}{2})$$

$$= \; -2log2 + 2\,JS(p_{data}(x)||p_G(x))$$

3- If input data in VAE are binary, Multivariate Bernoulli have used on decoder output distribution. If the output of last layer of decoder is “a” (after apply sigmoid). show that binary cross entropy appears in loss function.

$$p_\theta(x|z') = Bern(x, a) = \prod_{i=1}^{d} a_i^{x_i}(1 - a_i)^{1-x_i}$$

$$\rightarrow log\ p_\theta(x|z') = \sum_{i=1}^{d} x_i \log(a_i) + (1 - x_i)\log(1 - a_i)$$

$$\rightarrow \mathbb{E}_{q_{\phi(z|x^{(k)})}}\left[\log p_\theta\left(x^{(k)}|z\right)\right] \approx \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{d} x_i^{(k)}\log(a_i^{(j)}) + \left(1 - x_i^{(k)}\right)\log(1 - a_i^{(j)})$$

$$= \frac{1}{n}\sum_{j=1}^{n} -BCE(x^{(i)}, a^{(j)})$$