CSE6243: Advanced Machine Learning **Fall 2023**

Lecture 16: EBMs, GANs, and Divergences

Lecturer: Bo Dai *Scribes: Aditya Akula, Jiahong Zhang*

Note: *LaTeX template courtesy of UC Berkeley EECS Department.*

16.1 Recap

EBM: Energy-based Model:

$$\frac{\exp(g_0|x)}{Z_0}, Z_0 = \int \exp(g_0|x)$$

Autoregressive model (combine Transformer in ChatGPT):

$$P\left(\{x_i\}_{i=1}^d\right) = \prod_{i=1}^d P\left(x_i|x_{<i}\right)$$

This model generates probabilities at a location as a function of all the values before it.

VAE (latent model):

$$P_\theta(x) = \int P_\theta(x|z)P(z)dz,$$

Here, we explicitly find the latent space and classify each data point by generating a probability through integration over the whole latent space.

Diffusion model (image/video modeling):

$$P_\theta(x) = \int P_\theta\left(x|z_0\right) \prod_{i=1}^k p\left(z_{i-1}|z_i\right) dz_{i=0}^k$$

We combine aspects of VAEs and autoregressive models by generating probabilities through integrating across all paths of diffusion timesteps

16.2 New Content

16.2.1 Generative Adversarial Net(GAN)

Generative Adversarial Net(GAN) learns samplers instead of explicitly learning distribution. This could be turned to

$$\varepsilon \sim P(\varepsilon), \quad N\left(0, \theta^i I\right)|x = g_\theta\left(\varepsilon_i\right)$$

which generates samples as a function of random noise ϵ.

Objective function: Represents a generator network trying to minimize loss and discriminator network trying to maximize loss.

$$\min_{\theta}\max_{\phi}\mathcal{L}(\theta,\phi)=\mathbb{E}_{P_d}\left[\log D_{\phi}(x)\right]+\mathbb{E}_{x=g_\theta(x)}[\log\left(1-D_{\phi}(x)\right)] \quad \leftarrow \text{MLE for logistic regression}$$

Step 1. Fix a ϕ and max $D_\phi(x)\mathcal{L}(\phi,\theta)$ when $\nabla_{D_\phi(x)}\mathcal{L}(\phi,\theta)=0$

$\Rightarrow \forall x$ calculate

$$\begin{aligned}&\nabla_{D_\phi(x)}\left(P_d(x)\cdot\log D_\phi(x)+D_{g_\theta}(x)\cdot\log\left(1-D_\phi(x)\right)\right)\\ =&P_d(x)\cdot\frac{1}{D_\phi(x)}-P_{g_\theta}(x)\cdot\frac{1}{1-D_\phi(x)}=0\end{aligned}$$

Then,

$$D_\phi(x)=\frac{P_d(x)}{P_d(x)+P_{g_\theta}(x)}$$

Step 2. Find the optimal θ under fixed ϕ.

We put the ϕ into the objective function.

$$\begin{aligned}L\left(\theta,\phi^*\right)&=\mathbb{E}_{P_d}\left[\log\cdot D^*(\phi)\right]+\mathbb{E}_{P_{g_\theta}}\left[\log\left(1-D^*_\phi(x)\right)\right]\\ &=\int P_d(x)\cdot\log\frac{P_d(x)}{P_d(x)+P_{g_\theta}(x)}dx+\int P_{g_\theta}(x)\cdot\log\frac{P_{g_\theta}(x)}{P_d(x)+P_{d_\theta}(x)}dx\\ &\propto 2\mathrm{JS}\left(P_d\cdot P_{g_\theta}\right)+\mathrm{const}\end{aligned}$$

Recall Jensen -Shannon divergence

$$\mathrm{JS}(p,q)=\frac{1}{2}\mathrm{KL}\left(p\|\frac{p+q}{2}\right)+\frac{1}{2}\mathrm{KL}\left(q\|\frac{p+q}{2}\right)$$

Generator's goal is to minimize this divergence by recovering the original data distribution.

Algorithm.

Init $x=g_{\theta_0}\left(\varepsilon_i\right)$

For $i=1,\ldots,T$

1. Sample $x\sim P_d$ – take instance of real data
2. Sample $x'=g_\theta(\varepsilon),\quad \varepsilon_i\sim P_0(\varepsilon)$
3. For $k=1,\ldots,K,\quad \phi_{k+1}:\phi_k+\eta_k\hat{\nabla}_\phi\mathcal{L}(\theta,\phi)$ – gradient descent on ϕ
4. $\theta_{t+1}=\theta_t-\lambda_t\hat{\nabla}_\phi L\left(\theta,\phi_k\right)$ – gradient descent on θ. Note that we don't update θ every time we update ϕ, as the two networks are adversaries, and so we need to make sure we don't simply make one better by weakening the other.

Based on JS-Divergence, we also call above algorithm as JS -GAN (minimizing JS divergence)

16.2.2 Introduction to f-GAN (min f divergence)

Recall f divergence:

$$D_f(p,q) = \mathbb{E}_q\left[f\left(\frac{p}{q}\right)\right] = \int q(x) f\left(\frac{p}{q}\right) dx$$

where $f(\cdot)$ is a convex function, $f(1) = 0$.

Choice for $f(\cdot)$.

- KL-divergence

$$\begin{aligned} f(a) = a \cdot \log a \to D_f(p,q) &= \int q(x) \cdot \frac{p(x)}{q(x)} \cdot \log \frac{p(x)}{q(x)} \cdot dx \\ &= \int p(x) \cdot \log \frac{p(x)}{q(x)} dx \end{aligned}$$

- χ^2-divergence

$$\begin{aligned} f(a) = (a-1)^2, \quad D_f(p,q) &= \int q(x) \left(\frac{p(x)}{q(x)} - 1\right)^2 dx \\ &= \int q(x) \left[\left(\frac{p(x)}{q(x)}\right)^2 - 2\frac{p(x)}{q(x)} + 1\right] dx \\ &= \int q(x) \left(\frac{p(x)}{q(x)}\right)^2 dx - \underbrace{\int 2 \cdot p(x) dx + \int q(x) dx}_{=-2+1=-1} \\ &= \int q(x) \cdot \left(\frac{p(x)}{q(x)}\right)^2 dx - 1 \end{aligned}$$

- TV

$$f(a) = |a-1|, \quad D_f(p,q) = \int |p(x) - q(x)| dx$$

We have several background knowledge for f-GAN:

(1) $f(\cdot)$ convex

(2) The convex conjugate of a function satisfies the property $f**(x) = f(x)$. The convex conjugae is defined as $f(x)$ where

$$f^*(y) = \sup_{x \in \Omega} \left(x^\top y - f(x)\right)$$

e.g.

$$\begin{aligned} &f(x) = \frac{1}{2}x^2, \quad f^*(y) = \sup_x x^\top y - \frac{1}{2}x^2 = y^\top y - \frac{1}{2}y^\top y = \frac{1}{2}y^\top y \\ &x^* = y \\ &f(x) = \log \sum_{i=1}^{n} \exp(x_i) \quad f^*(y) = \sum_{i=1}^{n} y_i \log y_i \end{aligned}$$

Using the second property:

$$(f^*)^* x = f(x) \Rightarrow f(x) = \max_y y^\top x - f^*(y)$$

using convex conjugate in f-GAN

$$D_f(p,q) = \int q(x) \cdot f\left(\frac{p(x)}{q(x)}\right) dx = \int q(x) \sup_{y_x} \left[y x^\top x - f^*\left(y_x\right)\right] dx$$

denote $y_x = y(x)$, then

$$\begin{aligned}
&= \max_{y(x)} \int q(x) \left[y(x)^\top \frac{p(x)}{q(x)} - f^*(y(x)) \right] dx \\
&= \max_{y(x)} \left(\int y(x) \cdot p(x) dx - \int q(x) \cdot f^*(y(x)) dx \right) \\
&= \max_{D(x)} \mathbb{E}_{p(x)}[D(x)] - \mathbb{E}_{q(x)} \cdot [f^*(D(x))]
\end{aligned}$$

16.2.3 Connection between EBM and GAN

EBM:

$$P_\phi(x) = \frac{\exp\left(D_\phi(x)\right)}{z(\phi)} \Rightarrow \text{MLE} = E_{P_x}\left[D_\phi(x)\right] - \log z(\phi), z(\phi) = \int \exp(D_\phi(x) dx$$

Apply log and rewrite $z(\phi)$,

$$\log z(\phi) = \log \int \exp\left(D_\phi(x)\right) dx = \sup_q \langle q(x) \cdot D(x) \rangle - \underbrace{\int q(x) \cdot \log q(x) dx}_{+H(q)}$$

using convex conjugate property

Then

$$\text{MLE} = \max_\phi \min_q \mathbb{E}_{P_d}[D(x)] - \mathbb{E}_q[D(x)] - H(q)$$

where $H(q)$ is entropy.