CSE6243: Advanced Machine Learning **Fall 2023**

Lecture 11: EBM and Diffusion

Lecturer: Bo Dai *Scribes: Nikhil Shanbhogue, Takahiro Furuya*

11.1 Recap

In the previous class, we learned about energy-based models and applying score matching and contrastive divergence (Fischer) in an attempt to reduce the learning and evaluation complexity.

$$p(x) = \frac{\exp f_\theta(x)}{z(\theta)} \xrightarrow[duality]{optimization} \text{Max Entropy Model}$$

$$\text{Learning} \rightarrow \text{Minimize Divergence} \begin{cases} \text{K.L} \rightarrow \text{M.L.E.} \rightarrow \text{C.D.} \\ \text{Fischer} \rightarrow \text{Score Matching} \end{cases}$$

We further explored different sampling techniques such as Gibbs, Langevin, and Metropolis-Hastings (x_{t+1} sampled from $p(\cdot|x_t)$), and understood how sampling is analogous to generation.

11.2 New Content

11.2.1 Generation/ Sampling

We can start sampling using a Langevin Dynamic Sampler. The first step is to sample from an initial data distribution.Next, we start an iterative process for different time steps to generate a sample and accept it using Metropolis-Hastings.

$x_0 \sim p_0(x)$

for t = 1, ...

$y \sim p(\cdot|x_t)$
$y = x_t + \eta \nabla_{x_t} \log p_\theta(x_t) + \sqrt{\eta}\epsilon$
$u \in \mathcal{U}[0, 1]$
$x_{t+1} = y$ if $u \leq A(x, y) = min(1, \frac{p(y)p(x|y)}{p(x)p(y|x)})$

Let us consider the score of of $p_\theta(x)$ as follows:

$$\nabla_x \log p_\theta(x) = \nabla_x f_\theta(x) \tag{11.1}$$

11.2.2 Score Matching

Recall the score-matching expression and substitute the score function

$$\int \hat{p}(x)||\nabla_x f_\theta(x) - \nabla_x \log \hat{p}(x)||^2 dx \tag{11.2}$$

Here, $s_w(x) = \nabla_x f_\theta(x) : \mathbb{R}^d \to \mathbb{R}^d$. We can expand out the expression for score by writing:

$$\int \hat{p}(x)(s_w(x))^2 dx + \int \hat{p}(x)(\nabla_x \log \hat{p}(x))^2 dx - 2\int \hat{p}(x) s_w(x) \nabla_x \log \hat{p}(x) dx$$

The second integral term is constant with respect to the optimization and therefore the expression reduces to:

$$\mathbb{E}_x[s_w(x)^2 + 2\nabla_x s_w(x)]$$

11.2.3 Why score matching fails

1. Difficult to learn the gradient in the $s_w(x) = \nabla_x f_\theta(x) : \mathbb{R}^d \to \mathbb{R}^d$ dimension
2. The Langevin dynamics need to go to infinite steps to accurately sample

11.2.4 Diffusion Model Design

In the above intractable expression, it gets challenging to estimate the gradient of a second order equation. This in turn leads to the generation process being expensive. Hence, let us try conditioning the energy-based model with noise and utilise a perturbed distribution.

$$p_\theta(x) = \frac{\exp f_\theta(x)}{Z(\theta)} \tag{11.3}$$

$$p(x'|x) = N(x\sqrt{1-\beta}, \beta I) \tag{11.4}$$

$$x' = \sqrt{1-\beta}x + \sqrt{\beta}\epsilon, \quad \epsilon \sim N(0,1) \tag{11.5}$$

$$p_\beta(x') = \int p_\theta(x) p_\beta(x'|x) dx, \quad \beta \to 0 \tag{11.6}$$

For sampling each data point and obtaining a sequence of data points $\{x_0, x_1, \ldots, x_N\}$, relation (11.4) (discrete Markov chain model) is used. Equation 11.5 is the relation between consecutive data points.

$$\nabla_{x'} \log p_\beta(x') = \frac{\nabla_{x'} \int p_\theta(x) p_\beta(x'|x) dx}{p_\beta(x')} \tag{11.7}$$

$$= \frac{\int p_\theta(x) \nabla_{x'} p_\beta(x'|x) dx}{p_\beta(x')} \tag{11.8}$$

$$= \int p(x|x') \nabla_{x'} \log p_\beta(x'|x) dx \tag{11.9}$$

$$= \mathbb{E}_{x|x'}\left[\nabla_{x'} \log p_\beta(x'|x)\right] \tag{11.10}$$

$$= \mathbb{E}_{x|x'}\left[\nabla_{x'}\left(-\frac{||x' - \sqrt{1-\beta}x||^2}{2\beta}\right)\right] \tag{11.11}$$

Equation (11.9) is obtained by considering the joint probability of x' and (11.11) is derived after substituting (11.4).

$$x' + \beta\nabla_{x'} \log p_\beta(x') = \mathbb{E}_{x|x'}[\sqrt{1-\beta}x] \tag{11.12}$$

$$x' + \beta \underbrace{\nabla_{x'} \log p_\beta(x')}_{\text{parametrized as } S_w(x',\beta)} = \sqrt{1-\beta}\mathbb{E}_{x|x'}[x] \tag{11.13}$$

$$\tag{11.14}$$

11.2.5 Parametrization

Now, the objective function to optimize the score function (minimize w) is defined as follows:

$$\min_w \mathbb{E}_\beta \mathbb{E}_{x|x'}[||x' + \beta S_w(x,\beta) - \sqrt{1-\beta}\,\mathbb{E}_{x|x'}[x]\,||^2] \tag{11.15}$$

$$\min_w \mathbb{E}_\beta \mathbb{E}_{x|x'}\left[||x' + \beta S_w(x',\beta) - \sqrt{1-\beta}x||^2\right] \tag{11.16}$$

The optimal model $s_w(x,\beta)$ (w is parameters) is the one to minimize expression (11.15), so we will find it by considering expression (11.16) by showing the equivalence below:

Let $b = \sqrt{1-\beta}\mathbb{E}_{x|x'}[x]$.

$$\mathbb{E}_\beta \mathbb{E}_{x|x'}\left[||x' + \beta S_w(x',\beta) - \sqrt{1-\beta}x||^2\right] \tag{11.17}$$

$$= \mathbb{E}_\beta \mathbb{E}_{x|x'}\left[||x' + \beta S_w(x',\beta) - b + b - \sqrt{1-\beta}x||^2\right] \tag{11.18}$$

$$= \mathbb{E}_\beta \mathbb{E}_{x|x'}\left[||x' + \beta S_w(x',\beta) - b||^2\right] + \mathbb{E}_\beta \mathbb{E}_{x|x'}\left[||b - \sqrt{1-\beta}x||^2\right] \tag{11.19}$$

$$+ 2\mathbb{E}_\beta \mathbb{E}_{x|x'}\left[(x' + \beta S_w(x',\beta) - b)\right] \mathbb{E}_\beta \mathbb{E}_{x|x'}\left[(b - \sqrt{1-\beta}x)\right] \tag{11.20}$$

We can show that the cross term goes to 0 by the Tower property:

$$\begin{aligned}
&\mathbb{E}_\beta \mathbb{E}_{x|x'}[(x' + \beta S_w(x',\beta)^\top (b - \sqrt{1-\beta}x)] \\
&= \mathbb{E}_\beta \mathbb{E}_{x|x'}[(x' + \beta S_w(x',\beta)^\top (b - \sqrt{1-\beta}\mathbb{E}_{x|x'}[x])] \\
&= \mathbb{E}_\beta \mathbb{E}_{x|x'}[(x' + \beta S_w(x',\beta)^\top \mathbf{0}] \\
&= 0.
\end{aligned} \tag{11.21}$$

The above expression gives an insight into how the diffusion process is related to the energy-based model by introducing a noise perturbed distribution.