CSE6243: Advanced Machine Learning **Fall 2024**

Lecture 8: Sampling: MCMC (MH, Gibbs & Hamiltonian)

Lecturer: Bo Dai *Scribes: Abinav Chari, Siddharth Pamidi*

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

8.1 Recap

- Frequentist Learning vs. Bayesian Learning
- Motivation for Sampling
- Types of Sampling: Inverse Probability Trasnformation, Acceptance-Rejection Sampling
- Recapped how to find best $q(x)$ (Arbitrary distribution to sample from)

Proof:

$$\text{Best } q(x) \rightarrow \underset{q}{\text{argmin}}\ Var[\mathbb{E}_q[\frac{p(x)}{q(x)}f(x)]]$$

$$= \underset{q}{\text{argmin}}\ Var[\frac{1}{n}\sum_{i=1}^{n} f(x_i)\frac{p(x_i)}{q(x_i)}]$$

$$= \underset{q}{\text{argmin}}\ \frac{1}{n}(\mathbb{E}_q[\sum_{i=1}^{n} f^2(x_i)\frac{p^2(x_i)}{q^2(x_i)}] - (\mathbb{E}_q[f(x)\frac{p(x)}{q(x)}])^2)$$

The first expectation term can be rewritten as: $\underset{argmin}{q} \int f^2(x)\frac{p^2(x)}{q^2(x)}$ and the second expectation term can be rewritten as $\mathbb{E}_p[f(x)]$, which is constant in q and thus removed from the optimization. Then by applying the Cauchy–Schwarz inequality we can rewrite it as

$$\int f^2(x)\frac{p^2(x)}{q^2(x)}dx * \int q(x)dx \geq (\int f(x)p(x)dx)^2$$

($\int q(x)dx = 1$ because it is a probability distribution)

$$q(x) = \frac{f(x)p(x)}{\int f(x)p(x)dx}$$

∎

8.2 New Content

8.2.1 Motivation

The problem with Acceptance and Rejection Sampling is that $q(x)$ needs to meet certain conditions which is restrictive and sometimes difficult.

8.2.2 MCMC Sampling

Psuedocode for MCMC sampling:

```
Sample x_0 ~ q(x) \\
for t = 1....
    x_t ~ T(x|x_{t - 1})
```

(This is known as a Markovian process because the current sample only depends on the previous sample).

As t goes to infinity the sample converges to the target distribution $x_\infty \sim P(x)$.

We can write this condition mathematically as

$$P(x) = \lim_{t \to \infty} \int \prod_{t=1}^{t} T(x_i | x_{i-1} q(x_0)) d\{x_i\}_{i=1}^{t} \tag{8.1}$$

Designing T so that it converges to the target distribution $p(x)$

Thm1. $\begin{cases} P(x') = \int T(x'|x)p(x)dx \\ T(x'|x) \text{ has only one unique stationary distribution} \end{cases}$

Theorem 1 is a sufficient condition to check equation 8.1

Thm2. $\begin{cases} P(x')T(x'|x) = P(x)T(x'|x) \\ \text{Irreducable } (\forall x, y T(x|y), T(y|x) \geq 0) \text{ and a-periodic } (pt(x|x) \neq 0 \forall t) \end{cases}$

Thm2. is a sufficient condition to check for thm1.
Proof to show Detailed balance (Thm2.) is sufficient to prove stationary distirbution (thm1.)

Proof:

$$\begin{gathered} \int T(x'|x)p(x)dx \\ = \int T(x|x')p(x')dx \\ = p(x') \int T(x|x')dx \\ = p(x') * 1 \\ p(x') = p(x') \end{gathered}$$

■

The pros of MCMC sampling is that our choice of q(x) is less restricted but the cons are that the sampled data generates is dependent.

8.2.3 MH - Metropolis-Hasting Algorithm

The MH algorithm is one such MCMC method. In the case of the MH algorithm, $T(\cdot|x_{t-1})$ is as follows:
$\begin{cases} i) & y \sim \tilde{P}(\cdot|x_{t-1}) \\ ii) & \mu \sim U[0,1] \end{cases}$
Accept sample if $\mu \leq A(x,y) := min(1, \frac{P(y)\tilde{P}(x|y)}{P(x)\tilde{P}(y|x)})$

Proof to check if MH algorithms satisfies Thm2.

Proof: First Condition:

$$p(x)T(y|x) = p(x)A(x,y)\tilde{p}(y|x) = p(x)\tilde{p}(y|x) * \min(1, \frac{p(y)\tilde{p}(x|y)}{p(x)\tilde{p}(y|x)})$$

$$= \min(p(x)\tilde{p}(y|x), p(y)\tilde{p}(x|y)) = p(y)\tilde{p}(x|y) * \min(\frac{p(x)\tilde{p}(y|x)}{p(y)\tilde{p}(x|y)}, 1) = p(y)T(x|y)$$

Proving the second condition is tedious, so it was not covered in class. ■

8.2.4 Hit and Run Algorithm

Hit and Run algorithm is a modification on the MH operator, it samples y from a normal distribution centered around the previous sample.

$$y \sim \tilde{P}(\cdot|x_{t-1}) \propto exp(\frac{||x - x_{t-1}||^2}{2\sigma^2})$$

Example 1. Given $P(x) \propto \exp(||x||^2)$. How do we calculate A(x, y)?

$$= min(1, \frac{P(y)\tilde{P}(x|y)}{P(x)\tilde{P}(y|x)})$$

$$= exp(-||y||^2 + ||x||^2)$$

(note: $y = x + \epsilon$)

$$= exp(||x||^2 - ||x + \epsilon||^2)$$

$$= exp(\epsilon^2 - 2x^T\epsilon)$$

As we can see, when ϵ is small there is a high chance to accept the sample whereas if it is large there is a high chance to reject sample.

8.2.5 Gibbs Sampling

Gibbs sampling is another MCMC algorithm.
Given $x \in \mathbb{R}^d$

$$x_t \sim T(\cdot|x_{t-1})$$

(Find permutation of d)
Repeat d times to obtain x_t:

$$y_i \sim P(x_i|x_{-i})$$

Unlike the previous algorithms, we accept every sample. $A((x_i, x_{-i}), x_{t-1}) = 1$

Proof:

$$A(x, y) = \min(1, \frac{p(y)\tilde{p}(x|y)}{p(x)\tilde{p}(y|x)})$$

Let $z = \frac{p(y)\tilde{p}(x|y)}{p(x)\tilde{p}(y|x)}$. Also, define $x = \{x_1, x_{-1}\}$ and $y = \{y_1, x_{-1}\}$. Then,

$$p(y) = p(x_{-1})p(y_1|x_{-1})$$
$$p(x) = p(x_{-1})p(x_1|x_{-1})$$
$$\tilde{p}(x|y) = \tilde{p}(x_1|x_{-1})$$
$$\tilde{p}(y|x) = \tilde{p}(y_1|x_{-1})$$
$$z = \frac{p(x_{-1}) * p(y_1|x_{-1}) * \tilde{p}(x_1|x_{-1})}{p(x_{-1}) * p(x_1|x_{-1}) * \tilde{p}(y_1|x_{-1})} = 1$$

Thus, $A(x, y) = \min(1, 1) = 1$ ∎