Sampling Methods :

Different cases :

1. Given $p(x)$, draw samples $x \sim p(x)$

   e.g. $\underset{\in \mathbb{R}^d}{x} \sim \mathcal{N}(\underset{d\times 1}{\mu}, \underset{d\times d}{\Sigma})$, $x = \mu + Lz$, $z \sim \mathcal{N}(0,$

   $\Sigma = LL^T$ Cholesky decomposition

   $\hookrightarrow$ Cholesky factor

   e.g. $p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)\, p(\theta)$

   we want to generate $\theta \sim p(\theta|\mathcal{D})$.

2. Given $x_1, x_2, \dots, x_n$, $x_i \in \mathbb{R}^d$, learn $p(x)$

3. Estimate statistics, e.g. given a r.v $x \sim p(x)$:

$$\mathbb{E}_{x\sim p(x)}[f(x)] = \int f(x)\, p(x)\, dx \quad , \quad x \in \mathbb{R}^d$$

4. Perform Bayesian inference.

(*) Many sums or integrals can be written as expectations :

e.g. • marginal likelihood : $p(\mathcal{D}) = \int p(\mathcal{D}|\theta)\, p(\theta)\, d\theta =$

$$= \mathbb{E}_{\theta\sim p(\theta)}[\, p(\mathcal{D}|\theta)\,]$$

• predictive posterior distribution :

$$p(y^*|x^*, \mathcal{D}) = \int p(y^*|x^*, \mathcal{D}, \theta)\, p(\theta|\mathcal{D})\, d\theta$$

$$= \mathbb{E}_{\theta\sim p(\theta|\mathcal{D})}[\, p(y^*|x^*, \mathcal{D}, \theta)\,]$$

Example :

Q1 : What is the avg height of students in ENM 360 ?

ans: $\mathbb{E}_{h\sim G}[h] := \frac{1}{|G|} \sum_{p \in G} h(p)$

Q2: What is the avg height of people in Center city?

ans: $\mathbb{E}_{h\sim G}[h] \approx \frac{1}{S} \sum_{i=1}^{S} h(p_i)$

Monte Carlo approximation:

Goal: Approximate an expectation/integral using samples.

$$\mathbb{E}_{x\sim p(x)}[f(x)] = \underbrace{\int f(x)\, p(x)\, dx}_{\text{intractable}}$$

$x \in \mathbb{R}^d$, $d >> 1$

Definition: If $x_1, x_2, \ldots, x_n \overset{i.i.d.}{\sim} p(x)$ then:

$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} f(x_i)$ is a basic Monte Carlo estimator

(this is just the sample average)

Remarks:

1.) $\mathbb{E}[\hat{\mu}_n] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[f(x_i)] \xrightarrow{n\to\infty} \mathbb{E}[f(x)]$,

hence $\hat{\mu}_n$ is an unbiased estimator.

2.) $\hat{\mu}_n \xrightarrow{P} \mathbb{E}[f(x)]$ as $n \to \infty$, convergence in probability

i.e.: $\forall \varepsilon > 0$, $P(|\hat{\mu}_n - \mathbb{E}[f(x)]| < \varepsilon) \xrightarrow{n\to\infty} 1$

3.) because x_i are i.i.d.

$\mathrm{Var}[\hat{\mu}_n] = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}[f(x_i)] \xrightarrow{n\to\infty} \frac{1}{n} \mathrm{Var}[f(x)]$

$|\hat{\mu}_n - \mathbb{E}[f(x)]|^2 = \cancel{\text{bias}^2}^{\,0} + \text{var} \xrightarrow{n\to\infty} \frac{1}{n} \mathrm{Var}[f(x)]$

$\xrightarrow{n\to\infty} \frac{1}{\sqrt{n}} \mathrm{std}[f(x)]$

Therefore, $\hat{\mu}_n$ converges to $\mathbb{E}[f(x)]$ at a rate $\Theta\left(\frac{1}{\sqrt{n}}\right)$

Q: How many samples do we need in practice?

Note: Despite the fact that we know this rate of

1.) convergence, it may be very difficult what the actual error is in practice, because we don't know the variance $\mathbb{V}ar[f(x)] = \int (f(x) - \mathbb{E}[f(x)])^2 p(x) d$

2.) Practical limitation: One needs to be able to efficiently generate i.i.d. samples x_i from $p(x)$.