Key points of this chapter are that:

① Knowing the motivation and concepts.

② Clarification of terms (Sampling, monte carlo, Gibbs Sampling, MCMC)

③ Sampling methods analysis.

Sampling and Monte Carlo Methods

- why sampling? { Calculating sum/integral efficiently (considering them as expectation) / We want to generate samples from training distribution.

- Basic idea of Monte Carlo Methods

Aim: we want to calculate sum/integral:

$$S = \sum_x p(x) f(x) = E_p[f(x)] \quad \text{or} \quad S = \int_x p(x) f(x) = E_p[f(x)]$$

we draw samples from $p(x)$

$$\hat{S}_n = \frac{1}{n}\sum_i f(x_i) \rightarrow E[\hat{S}_n] = \frac{1}{n}\sum_i E[f(x_i)] = S$$

if $n \to +\infty$, $\hat{S}_n = S$. unbiased estimation

About the variance: $Var(\hat{S}_n) = \frac{Var[f(x)]}{n}$ can be seen as an indicator of confidence

Importance Sampling

Motivation: Always, we cannot directly sampling from $P(x)$. we re-write:

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

sampling from $q(x)$, and finally adding with importance weights $\frac{p(x)}{q(x)}$.

$$E_p(f(x)) = E_q\left(\frac{p}{q} f(x)\right)$$

About the variance: $Var(\hat{S}_q) = Var[\frac{pf}{q}]/n$, when $q^* = \frac{pf}{Z}$, we got minimum variance, which means that q is sufficiently close to p.

Bias importance sampling:

$$\hat{S}_{BIS} = \frac{\sum_{i=1}^{n} \frac{\tilde{p}(x_i)}{q(x_i)} f(x_i)}{\sum_{i=1}^{n} \frac{\tilde{p}(x_i)}{\tilde{q}(x_i)}}$$

Avoid to calculate partition function Z.

– Deep Learning use sampling: To compute some intractable calculations, making them a representation of expectation first and then do sampling.

Markov Chain Monte Carlo Methods

– Unlike directed graph, undirected graph cannot find a explicit dependcy relation, we cannot do ancestral sampling.

– MCMC methods build a markov chain to decide q, according to x^{t-1}:

$$q_t(x^t) = \sum_x T(x'|x) \cdot q_t(x^{t-1})$$

we sampling $q_t(x^t)$ at timestep t.

Transition / Stochastic matrix

$t-1$ step probability of $q_t(x)$

– Stationary Distribution of Markov Chain.

$v^t = A v^{(t-1)}$ consists of $T(x'|x)$

$v^t = A^t v^0$, then if $\underline{v = Av}$, the distribution is stationary

A has eigenvalue 1, and v is eigenvector

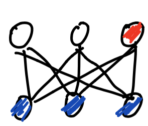
– MCMC operation: First burn into a stationary distribution, then sampling according to $q_t^t = A q_t^{t-1}$, sometimes fetching samples with n interval to ensure non-correlation between samples.

Gibbs Sampling

$q(x)$ is quite important in MCMC and importance sampling, how we decide $q(x)$ is related to how we choose $T(x'|x)$ / A, means what kind of transition matrix is appropriate. two ways proposed

- Gibbs sampling
- Directly parametrize T.

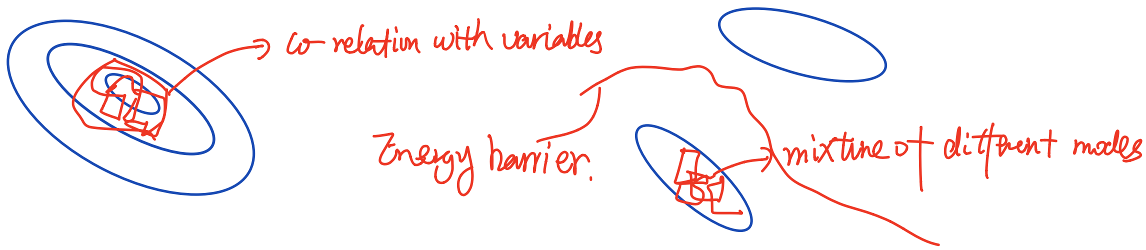
– Gibbs Sampling for $T(x'|x)$ means that $T(x'|x)$ with only one variable x_i is different in x' from x. such as RBM:



$T(x'|x)$ can be seen as shadowed part sampling from , and we get $T(x'|x)$, if = 1,2,3, then can be ....

The challenge of Mixing between Separated Modes (Models and Variables)

- MCMC: A significant pattern, we sample according to last time step, it will yield we are trapped within one mode.



- How we solve this issue:
  - Tempering: $p(x) \propto \exp(-\beta \mathcal{E}(x))$ → Control to not be "so" peak!
  - Depth may help:

O O O → x observed → $p(x|h)$

→ Deeper structure

O O O → h latent → $p(h|x)$

means sample within huge "gap".