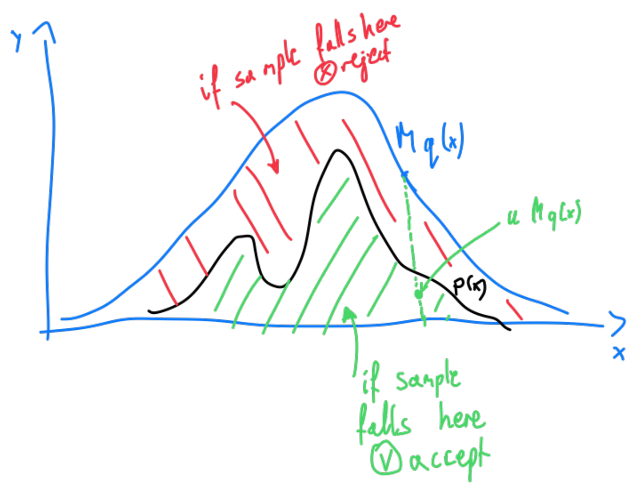
Tutorial 7

1. Show/explain why it is sufficient to know $p(x)$ up to a normalizing constant.
   it is enough to construct $q(x) \cdot M$, since probability of accepting sample does not depend on Z
2. Rejecting too many samples will slow the sampling, Derive the acceptance probability $p(u < \frac{\tilde{p}(x)}{Mq(x)})$ for the unnormalized case, where $p(x) = \frac{\tilde{p}(x)}{Z}$. Prob $Q(x) \cdot P(u < p^*) = Q(x) \cdot p(x) / (M Q(x))$ $u \simeq Z$?

1.



Since $Mq(x)$ is the approximation of more complex distribution $p(x)$. Since we want the distribution $q(x)$ to be close to $p(x)$.

Probability of accepting sample is conditioned by

$$\underbrace{p\left(u_i < \frac{p(x_i)}{M \cdot q(x_i)}\right)}, \text{ where } \underline{u} \text{ is}$$

sampled from uniform distribution $U_{(0,1)}$. Since $\frac{p(x)}{q(x)} < M$, we could see by analyzing equation and a figure that just knowing $\underline{M}$ is enough to and we do not need to know what is a exact $p(x)$.

2.

$$\int p\left(u < \frac{\tilde{p}(x)}{M q(x)}\right) q(x)\, dx = \int \frac{\tilde{p}(x)}{M q(x)} q(x)\, dx$$

$$= \int p\left(u < \frac{\tilde{p}(x)}{M q(x)} \,\middle|\, X = x\right) dx = \int \underbrace{p(x)}_{\substack{\shortparallel \\ 1}} \frac{Z}{M}\, dx$$

$$= \underline{\underline{\frac{Z}{M}}}$$

→ probability of acceptance can be denounced as

$$p(\text{acceptance} \mid X = x) = \frac{\tilde{p}(x)}{M q(x)}$$

→ Since $x \sim q(x)$, $p(X = x) = q(x)$, thus $\frac{p(x)}{q(x)} \approx 1$

$$\to p(x) = \frac{\tilde{p}(x)}{Z} \longrightarrow \tilde{p}(x) = Z \cdot p(x)$$

$$\to \int p(x)\, dx = 1$$

1. Show that $E_{p(x)}[f(x)] \approx \frac{1}{L}\sum_{l=1}^{L} f(x^{(l)})w(x^{(l)})$, where $x^{(l)} \sim q(x)$ (equation $\boxed{4}$).

2. Why do we need importance weights? *

1.

$\boxed{\text{proof}}$

$p(x)$ - probability of generating samples
$f(x)$ - distribution we want to approximate.

$$E_{p(x)}[f(x)] \approx \frac{1}{L}\sum_{l=1}^{L} f(x^{(l)})\,\omega(x^{(l)})$$

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

$$E_{p(x)}[f(x)] = \int f(x)p(x)\,dx = \int f(x)\frac{p(x)}{q(x)}q(x)\,dx \approx \frac{1}{L}\sum_{l=1}^{L} f(x)\,\omega(x)$$

2. weights $\omega(x) := \frac{p(x)}{q(x)}$. We need the weights to see how close the sample is to the actual distribution $p(x)$. Increasing/decreasing weight can also allow us on sampling from a particular area of interest.

1. Show, that the Markov Chain transition kernel (equation $\boxed{7}$) in Metropolis-Hastings Algorithm satisfy the detailed balance condition (equation $\boxed{1}$).

2. Why would you want to use an asymmetrical proposal distribution, i.e MH and not just Metropolis?
MH rejection rate does not grow exponentially with dims, the distribution we are sampling from my simply be not sym, i.e. (f
3. For a model with PGM $p(y,x) = p(x)p(y|x)$, what could be a reasonable proposal distribution for the *Independent Sampler* when we want to sample from posterior $p(x|y)$? *(Hint y is known but we want a proposal that is independent of y.)

1.

ow that:

$$p(x^{(i)})T(x^{(i-1)}|x^{(i)}) = p(x^{(i-1)})T(x^{(i)}|x^{(i-1)})$$

balance equation

sibility that this invariant distribution is not unique. A s

transition kernel

q is cancelled in acceptance

The transition kernel for MH is

$$T(x^{(i+1)}|x^{(i)}) = q(x^{(i+1)}|x^{(i)})A(x^{(i)}, x^{(i+1)}) + \delta_{x^{(i)}}(x^{(i+1)})r(x^{(i)}),$$

$\rightarrow A(x^i, x^{i+1}) \rightarrow$ acceptance prob of moving to the new state

$$\begin{cases} A(x^i, x^{i+1}) = \left\{ 1; \frac{p(x^{(i+1)})\, q(x^i \mid x^{i+1})}{p(x^i) \cdot q(x^{i+1} \mid x^i)} \right\} \\ r(x) = \int q(x^{i+1} \mid x)\, (1 - A(x^i, x^{i-1})\, dx \end{cases}$$

$\rightarrow$ since A have 2 possible outcomes, multiplying it by $q(x^{i+1} \mid x^i)$

- $T(x^{i+1} \mid x) = \min\left\{ q(x^{i+1} \mid x^i), \frac{p(x^{i+1})\, q(x^i \mid x^{i+1})}{p(x^i)} \right\} + \underbrace{\delta_x^{(i)}(x^i)\, r(x^i)}_{*}$

$\rightarrow$ Then $*$ can be extended to

- $\delta_x^i(x^i) r(x^i) = \delta_x^i(x) \int q(x^{i+1} \mid x)\, (1 - A(x^i, x^{i-1})\, dx =$ | since $\int q(x^{i+1} \mid x)\, dx = 1$

$$= \delta_x^i(x^i) \cdot \left(1 - \int q(x^{i+1} \mid x)\, A(x^i, x^{i+1})\, dx\right)$$

$\rightarrow$ Thus we end up with $T(x^{i+1} \mid x^i)$

- $\min\left\{ q(x^{i+1} \mid x^i), \frac{p(x^{i+1})\, q(x^i \mid x^{i+1})}{p(x^i)} \right\} + \delta_x^i(x^i) \cdot \left(1 - \int q(x^{i+1} \mid x) A(x^i, x^{i+1})\, dx\right)$

$\rightarrow$ Multiplying it by $p(x_i) \Longrightarrow p(x_i) T(x_{i+1} \mid x_i)$

- $\min\left\{ q(x^{i+1} \mid x^i)\, p(x^i), p(x^{i+1})\, q(x^i \mid x^{i+1}) \right\} + p(x_i) \delta_x^i(x^i) \cdot \left(1 - \int q(x^{i+1} \mid x) A(x^i, x^{i+1})\, dx\right)$

since (as we can read in Tutorial description) dystribution is symmetric $q(x^{i+1}|x^i) = q(x^i|x^{i+1})$, we could see that two sides will be equal i.e.

$$p(x^{i+1})\,T(x^i|x^{i+1}) = p(x^i)\,T(x^{i+1}|x^i)$$

2) The distribution we are going to sampling from may not be symmetric in which case, approximating it with a symmetric distribution could cause too many samples to get rejected.

3.) If two distributions are independent, then $p(y,x) = p(x)\,p(y|x)$ and current state $(i+1)$ is independent from the previous one (i) then $q(x^{i+1}|x^i) = q(x^i|x^{i+1}) = q(x^{i+1})$

The acceptance rate is given by $A = \min\left\{1, \frac{p(x^{i+1})}{p(x^i)}\right\}$

If we want to sample from $p(y|x)$,

$$\underbrace{\min\left\{1, \frac{p(y^{i+1}, x^{i+1})\; p(x^i)}{p(x^{i+1})\; p(y^i, x^i)}\right\}}$$,

That would be a good proposal distribution

Exercises (pen-and-paper) 1pts

Show that $\tilde{w}_t^i \propto \tilde{w}_{t-1}^i \frac{p(y_t|x_t^{(i)})p(x_t^{(i)}|x_{t-1}^{(i)})}{\pi(x_t^{(i)}|x_{0:t-1}^{(i)}, y_{1:t})}$ (Equation 10 from Section 4.2).

In the equations from tutorial, we can see that:

- unnormalize weights $\omega_t = \dfrac{p(x_{0:t} \mid y_{1:t})}{\pi(x_{0:t} \mid y_{1:t})}$, where π is a proposal distribution
- normalize importance weights $\tilde{\omega}_t = \dfrac{\omega_t^i}{\sum_j \omega_t^j}$ → sum of all the weights at timestep $\underline{t}$

$$\frac{\tilde{\omega}_t}{\tilde{\omega}_{t-1}} = \frac{p(x_{0:t} \mid y_{1:t})\, \pi(x_{0:t-1} \mid y_{1:t-1})}{\pi(x_{0:t} \mid y_{1:t})\, p(x_{0:t-1} \mid y_{1:t-1})}$$

We can see in the tutorial, that:

$$\pi(x_{0:t} \mid y_{1:t}) = \pi(x_0) \prod_{k=1}^{t} \pi(x_k \mid x_{0:k-1}, y_{1:k})$$

$$\frac{\pi(x_{0:t} \mid y_{1:t})}{\pi(x_{0:t-1} \mid y_{1:t-1})} = \pi(x_t \mid x_{0:t-1}, y_{1:t})$$

$$\frac{\tilde{\omega}_t}{\tilde{\omega}_{t-1}} = \frac{p(x_{0:t} \mid y_{1:t})}{\pi(x_t \mid x_{0:t-1}, y_{1:t})\, p(x_{0:t-1} \mid y_{1:t-1})} \quad *$$

→ Now, we can note that according to recursive definition of joint distribution presented in tutorial:

$$p(x_{0:t} \mid y_{1:t}) = p(x_{0:t-1} \mid y_{1:t-1}) \frac{p(y_t \mid x_t)\, p(x_t \mid x_{t-1})}{p(y_t \mid y_{1:t})}, \text{ thus}$$

$$* \quad \frac{p(y_t \mid x_t)\, p(x_t \mid x_{t-1})}{p(y_t \mid y_{1:t})}$$

and

$$\frac{\tilde{\omega}_t}{\tilde{\omega}_{t-1}} = \frac{p(y_t|x_t)\,p(x_t|x_{t-1})}{\pi(x_t|x_{0:t-1}, y_{1:t})\,p(y_t|y_{1:t})}$$

What proves that

$$\frac{\tilde{\omega}_t}{\omega_{t-1}} \propto \frac{p(y_t|x_t)\,p(x_t|x_{t-1})}{\pi(x_t|x_{0:t-1}, y_{1:t})} \Rightarrow \tilde{\omega}_t \propto \tilde{\omega}_{t-1} \frac{p(y_t|x_t)\,p(x_t|x_{t-1})}{\pi(x_t|x_{0:t-1}, y_{1:t})}$$

$\blacksquare$