statistical inference.

Model (with unknown parameters) $\xrightarrow{\text{inference}}$ Model with inferred parameters

① MLE, EM for ~~HN~~ Gaussian Mixture Model ⟶ point estimation

② But in many learning problems. you are actually interested in the posterior $\boxed{\text{distribution}}$ $p(\theta|data) \propto p(data|\theta)\,p(\theta)$

conjugate posterior : 如果 $p(\theta|data)$ 和 $p(\theta)$ 有 ~~一样~~ 的分布

同一种类型

(例如 Gaussian)

~~if $p(\theta|x)$ is very too complex to estimate~~

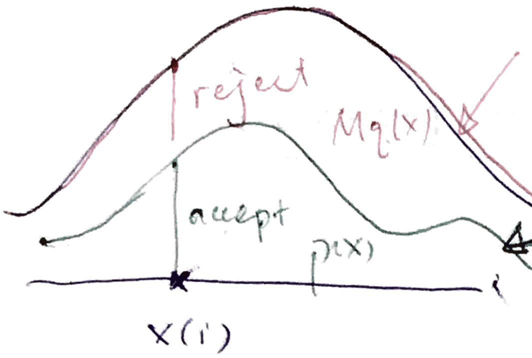
if $p(\theta|x)$ is very complex → hard to obtain the analytical solution of ~~$\mathbb{E}(p(\theta|x))$~~

we can use sampling to approximate → $\mathbb{E}_{\sim p(\theta|x)}(\theta)$

$\theta^{(i)} \overset{iid}{\sim} p(\theta|x)$ $\quad \hat{\mathbb{E}}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\theta^{(i)}$

③

1°. ~~adapted~~ ~~res~~ rejection sampling



relatively easy to sample

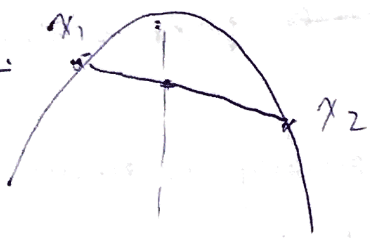
difficult to sample

i.e. q(·) is more computible

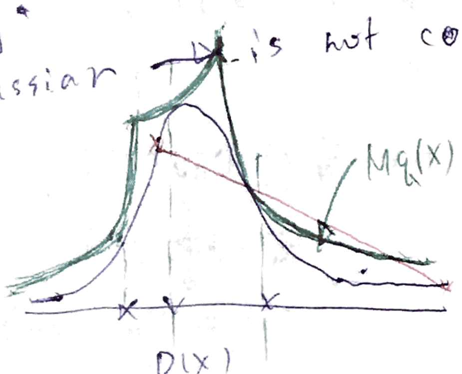
more efficient than vanilla RS

(ARS) adaptive rejection sampling

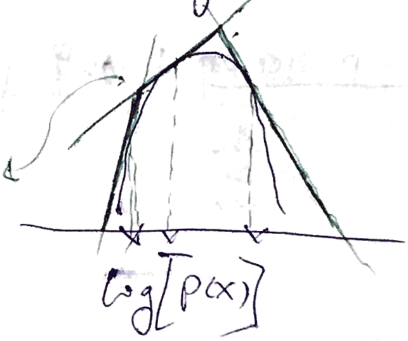
concave



log · Gaussian is not concave



but log-Gaussian is.



适用条件: only if Log[p(x)] is concave.

Requirement

2°. importance sampling

using sampling to approximate

$$E_{p(x)}[f(x)] = \int_x f(x)\, p(x)\, dx$$

$$\mathbb{E}_{p(x)}[f(x)] \quad \hat{\mathbb{E}}_{p(x)}[f(x)] = \frac{1}{N}\sum_{i=1}^{N} f(x^{(i)}), \quad x^{(i)} \overset{iid}{\sim} p(x)$$

(Sampling)

但是 p(x) 很复杂 太难 sample

$$\mathbb{E}[f(x)] = \int_x \underbrace{\left[f(x)\cdot\frac{p(x)}{q(x)}\right]}_{g(x)} q(x)\, dx = \int_x g(x)\, q(x)\, dx, \quad x^{(i)} \overset{iid}{\sim} q(x)$$

选择一个易sample的q(x)

$$\Rightarrow \hat{E}[f(x)] = \frac{1}{N}\sum_{i=1}^{N}\left[f(x^{(i)})\,\frac{p(x^{(i)})}{q(x^{(i)})}\right]$$

"importance weights" $w(x^{(i)})$

comparison

① in rejection sampling, the ~~sam~~ weights of each ~~sampling~~ sample are identical. sampling is achieved by rejecting some sample points

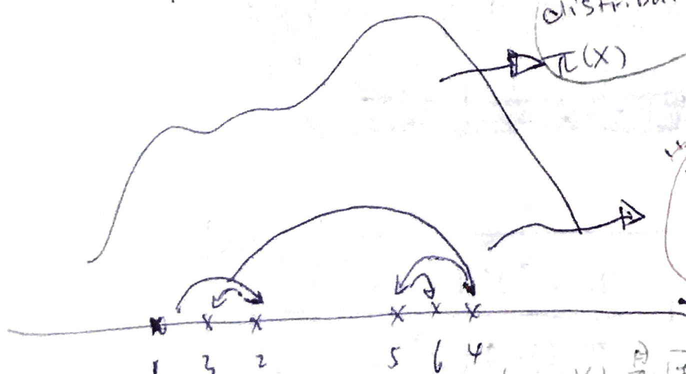
② in importance sampling, ~~samp~~ there is no rejection of samples. sampling is achieved by giving samples different weights.

3° MCMC. (Markov Chain Monte Carlo)

与1°.2°不同，不能并行地决定 sample ~~data~~ points in [illegible].

是一种 sequential 的算法

distribution $\pi(x)$ ← 最后形成的 sample points 服从 $\pi(x)$ 分布



"transition kernel" $K(x^{t-1}\rightarrow x^t)$ ← 如何得到？

因为 $\pi(x)$ 是用 $K(x^{t-1}\rightarrow x^t)$ 这种方式形成的

因此 $\pi(x)$ 和 $K(x^{t-1}\rightarrow x^t)$ 有某种关系

那么他们的关系是什么呢？(如何从 $\pi(x)$ 得到 $K(\cdot)$?)

p(x,y)? ← x-y 应是对称的 Transition Matrix

$$p(y) = \int_x p(y|x) \cdot p(x)\, dx$$

$$\boxed{\pi_t(x^*) = \int_x \pi_{t-1}(x) * K(x \mapsto x^*)\, dx}$$ → 或者写成 $$\boxed{\pi_t(x^*) = \sum_x \pi_{t-1}(x) K[x \mapsto x^*]}$$

Obviously, at equilibrium ~~st.~~, that stationary distribution why? (到了稳定状态后)

$\pi_t(x^*)$ 和 $\pi_{t-1}(x)$ 一定是同一种分布. 故下标 t 可以去掉

$$\Rightarrow \boxed{\pi(x^*) = \int_x \pi(x)\, K(x \mapsto x^*)\, dx}$$ ← 即，选择的 $K(x \mapsto x^*)$ 矩阵必须满足这个式子

所选择的 $K[x \mapsto x^*]$ 须满足的条件

The detailed balance condition holds when

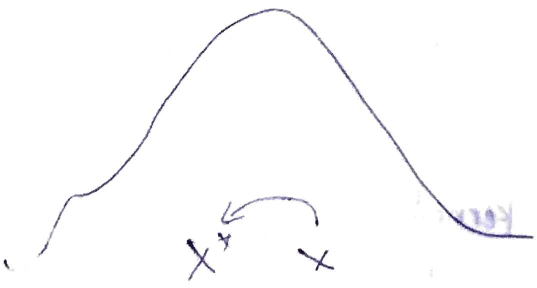
$$\boxed{\pi(x) K(x^*|x) = \pi(x^*) K(x|x^*)}$$

$$\Rightarrow \int \pi(x) K(x^*|x)\, dx = \int \pi(x^*) K(x|x^*)\, dx = \pi(x^*) \int K(x|x^*)\, dx = \pi(x^*)$$

所以我们一般去证明所选的 $K(x \mapsto x^*)$ 满足 detailed balance 就

Metropolis ~~Metropolitan~~ - Hastings Algo. → (a variation ~~variant~~ of MCMC)

MH Algo.



$$\boxed{K(x \mapsto x^*) = q(x^*|x)\, \min\left(1, \frac{\pi(x^*) q(x|x^*)}{\pi(x) \cdot q(x^*|x)}\right)} = \alpha(x^*)$$

propose x^* from $q(x^*|x)$

then accept x^* with ratio $\alpha(x^*)$

Metropolis-Hastings Algorithm

1. Initialize $x^{(0)}$

2. For $i = 0$ to $N-1$

   - Sample $u \sim \mathcal{U}_{[0,1]}$
   - Sample $x^* \sim q(x^* | x^{(i)})$ ← For 1st-order markov assumption
   - If $u < A(x^{(i)}, x^*) = \min\left\{1, \frac{\pi(x^*)\, q(x^{(i)} | x^*)}{\pi(x^{(i)})\, q(x^* | x^{(i)})}\right\}$

     $x^{(i+1)} = x^*$

   else

     $x^{(i+1)} = x^{(i)}$

$(x^* \to x^{(i)})$ → the prob of staying in x^* and then moving to $x^{(i)}$

the prob of staying in $x^{(i)}$ and then moving to x^* $(x^{(i)} \to x^*)$

Question: How to chose $q(x^* | x^{(i)})$?

Why does ~~为什么~~ MH Algo. work? To prove it, we simply need to verify the "detailed balance" condition.

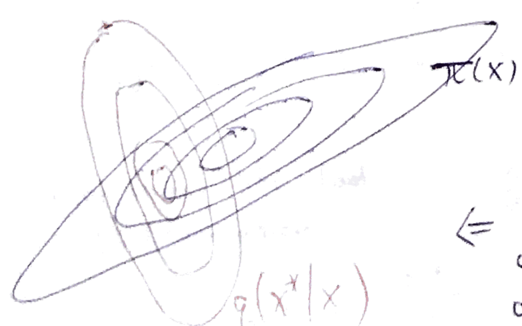
Proof $\quad \pi(x) \cdot K(x^* | x) \stackrel{?}{=} \pi(x^*)\, K(x | x^*)$

$$LHS = \pi(x) \cdot \left[q(x^* | x) \cdot \min\left(1, \frac{\pi(x^*) \cdot \cancel{K(x^* \to x)}}{\pi(x) \cdot \cancel{K(x \to x^*)}} \frac{q(x | x^*)}{q(x^* | x)}\right)\right]$$

$$= \min\left(\pi(x)\, q(x^* | x),\ \pi(x^*)\, q(x | x^*)\right)$$

$$= \pi(x^*)\, q(x | x^*) \min\left(1, \frac{\pi(x)\, q(x^* | x)}{\pi(x^*)\, q(x | x^*)}\right) = RHS.$$

We need to choose $q(x^* | x)$ carefully!



⇐ an example of bad choice of $q(x^* | x)$