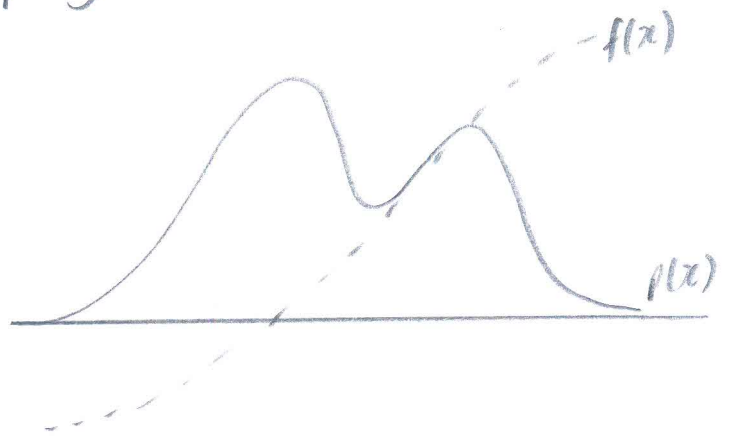
CUED Jordan (U3 material on sampling methods). 05/09/2020

Sampling Methods

$$\langle f \rangle = \int f(x) p(x) dx \quad (1.1)$$

$$\hat{f} = \frac{1}{M} \sum_{m=1}^{M} f(x^{(m)}) \quad (1.2)$$



- $x^{(m)} \sim p(x)$

$\Rightarrow \langle \hat{f} \rangle = \langle f \rangle$

$var[\hat{f}] = \frac{\sigma^2}{M}$ $\quad \sigma^2 = \langle (f - \langle f \rangle)^2 \rangle$ $\quad var[\hat{f}] \downarrow$ as $M \to \infty$.

- Issues:- independent samples $\{x^{(m)}\}_{m=1}^{M}$?
  - $f(x_1)$ large, $p(x_1)$ small
  - $f(x_0)$ small, $p(x_0)$ large.
  - $\}$ → interaction of magnitude of f with probability weighting.

Ancestral Sampling in GM: GM joint:- $p(x) = \prod_{i=1}^{d} p(x_i | x_{\pi(i)})$

- sample following GM flow
- no observed variables.

Logic sampling in GM

- some nodes instant. at observed values
- sample x_i and observed agree → continue sampling.
- else discard.

1.0.1 - Sampling and EM.

- direct impl. of Bayesian network.
- Frequentist → E-step ML est. with intract. posterior.

- x_H - hidden
- x_V - visible
- θ - param.

M-step optimises :- ECLL

$$Q(\theta, \theta_{old}) = \mathbb{E}_{p(x_H | x_V, \theta_{old})} [\ln p(x_H, x_V | \theta)]$$

$$= \int p(x_H | x_V, \theta_{old}) \ln p(x_H, x_V | \theta) \, dx_H \quad (1.5)$$

- Approx I.S via finite sum over samples $\{x_H^{(l)}\}$ drawn from current est. for posterior $p(x_H|x_V, \theta_{old})$:-

$$Q(\theta, \theta_{old}) \simeq \frac{1}{M} \sum_{m=1}^{M} \ln p(x_H^{(m)}, x_V|\theta) \qquad (1.6)$$

- Then optimise Q in usual way in M-step

(*) That was Monte Carlo EM

(*) Augment to find mode of posterior over θ (MAP estimate) by placing prior $p(\theta)$ and adding $\ln p(\theta)$ to $Q(\theta, \theta_{old})$ in M-step.

(*) stochastic EM.

(*) Full Bayesian treatment (IP/data aug algo) :-

- Sample from post over θ
- Draw from joint post $p(\theta, x_H|x_V)$ ⟶ computationally difficult.
- Imputation-Post

I-step : Sample from $p(x_H|x_V)$ → can't directly.

- Note: $p(x_H|x_V) = \int p(x_H|\theta, x_V) p(\theta|x_V) \, d\theta$

- For $m = 1, \ldots, M$:-
  i) sample $\theta^{(m)}$ from current est. for $p(\theta|x_V)$
  ii) Use this to sample $x_H^{(m)}$ from $p(x_H|\theta^{(m)}, x_V)$.

P-step: Note $p(\theta|x_V) = \int p(\theta|x_H, x_V) p(x_H|x_V) \, dx_H$

- Use $\{x_H^{(m)}\}_{m=1}^{M}$ from I-step to compute revised est. of post. over θ given by

- $p(\theta|x_V) \simeq \frac{1}{M} \sum_{m=1}^{M} p(\theta|x_H^{(m)}, x_V)$

1.1. Basic Sampling Algos

- sample from distn defined over univariate $x \in \mathbb{R}$
- Pseudo-random $\longrightarrow$ deterministic, passes tests for randomness
- Pseudo-random over Unif(0,1)
- this is basis for other algos.

1.1.1. standard distns

- PRNG - successive app. of trans $D(\cdot)$
- sequence :- $(x^{(1)}, x^{(2)}, \ldots, x^{(N)})$ via recurrence $x^{(n)} = D(x^{(n-1)})$
- seed/initial value :- $x^{(1)}$
- choice of $D(\cdot)$ - finite rep of digit. computers.
- integer representation :-
  $x \in \{0, \ldots, N\}$. N - largest poss. integer in comp.
- interval [0,1) generated by :- $\frac{x}{N+1}$

- Period? :- smallest integer T:
  $x^{(n+T)} = x^{(n)} \quad \forall n.$ or : $D(\cdot)^T$ is identity op.
  $x^{(n)} = D(x^{(n-1)}) \not\exists (M+1)$ (?)

(?) - largest value which simplegen :- of period

linear congruential: $D(x) = (ax+b) \bmod (M+1)$ - (a, b) - select with care.

- non-uniform PRNG: - assume source of uniformly distn random nos.
  $x \sim U(0,1)$.
- transform x via $f(\cdot)$: $y = f(x)$.
  $p(y) = p(x)\left|\frac{dx}{dy}\right| \qquad p(x) = 1$ (?) $\int p(x)dx = 1$?

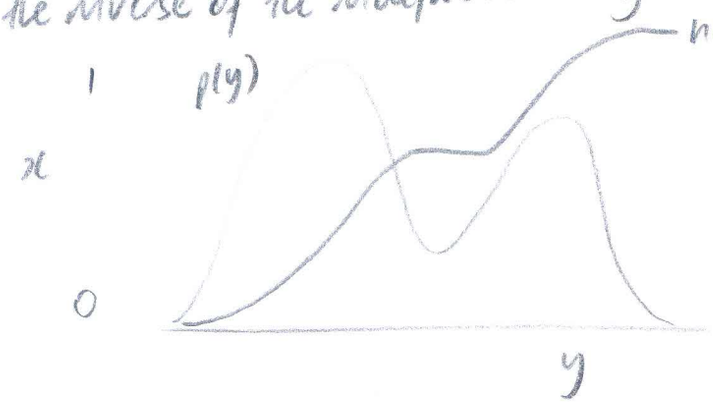
(w):
- is this similar to transforms of r.v.s?

- select $f(\cdot)$ such that resulting values of y have desired distn p(y). (0/5/1)
  inverse transform. tighten this.

(?) - missing a step.
- $x = h(y) \equiv \int_{-\infty}^{y} p(y')\,dy'$ (indefinite integral of p(y))

$\Rightarrow y = h^{-1}(x)$

(*) transform uniformly distri random nos using a function $f(\cdot)$ that is the inverse of the indefinite integral of desired distri.



e.g: Exponential.

$$p(y) = \lambda \exp(-\lambda y) \qquad (1.12)$$

$$h(y) = 1 - \exp(-\lambda y)$$

- use $y = h^{-1}(x) = -\lambda^{-1} \ln(1-x)$ then $y \sim \exp$

e.g. Cauchy

$$p(y) = \frac{1}{\pi} \frac{1}{1+y^2} \qquad (1.13)$$

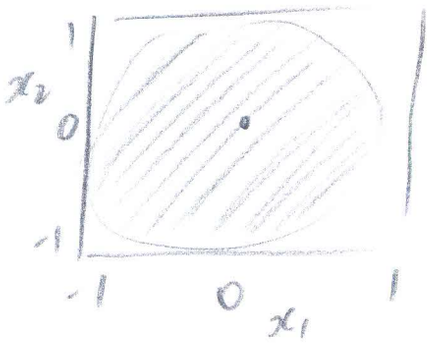
$$y = h^{-1}(x) = \tan(x)$$

Multiple variables

$$p(y_1, \dots, y_d) = p(x_1, \dots, x_d) \left| \frac{\partial(y_1, \dots, y_d)}{\partial(x_1, \dots, x_d)} \right|$$? ✗

Box Muller:

- Generate uniformly distri $(x_1, x_2) \in (-1, 1)$
- Discard pair unless $x_1^2 + x_2^2 \leq 1$
- uniform distri



$$p(x_1, x_2) = \frac{1}{\pi}$$?

- For each pair (x_1, x_2), evaluate:-

$$y_1 = x_1\left(\frac{-2\ln x_1}{r^2}\right)^{1/2} \qquad y_2 = x_2\left(\frac{-2\ln x_2}{r^2}\right)^{1/2} \qquad (1.15)(1.16)$$

OIS 2

$r^2 = x_1^2 + x_2^2$

- Joint distn: $p(y_1, y_2) = p(x_1, x_2)\left|\frac{\partial(y_1, y_2)}{\partial(x_1, x_2)}\right|$ (1.17) (?)

$$= \left[\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-y_1^2}{2}\right)\right]\left[\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-y_2^2}{2}\right)\right] \qquad (1.18)$$

- y_1, y_2 indep; zero-mean, unit variance.
- Finally use transformation: $y \to \sigma y + \mu$ to get Normal with μ, σ^2

- vector-valued multi-variate Normal, mean $\underline{\mu}$, cov $\underline{\Sigma}$.
- eigenvalue/vector decomp of $\underline{\Sigma}$:-

$$\underline{\Sigma}\,\underline{u}_i = \lambda_i \underline{u}_i \qquad (1.19)$$

OIS 3

- If x_i are univariate, normally distri, then

$$y = \underline{\mu} + \sum_i \lambda_i^{1/2} x_i \underline{u}_i \text{ has required distri}$$

- use Cholesky decomp of $\underline{\Sigma} = LL^T$ in practice.

- Rejection + importance sampl. $\longrightarrow$ univariate distns.

1.12. Rejection sampling

- sample from $p(x) = \frac{1}{z}\hat{p}(x)$ is difficult (analytic. intract.)

$\hat{p}(x)$ - readily evaluated

z - unknown normalising constant.

- introduce a proposal distri $q(x)$. and constant k. :

$$kq(x) \geq \hat{p}(x) \quad \forall x.$$

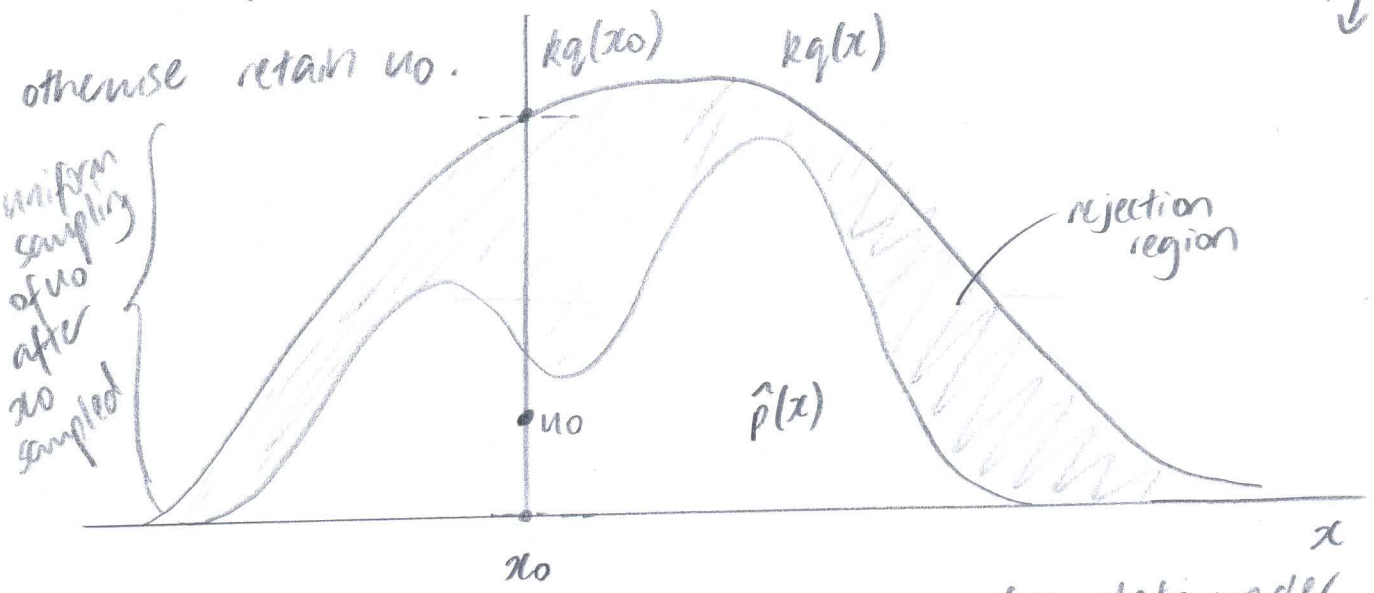
- $Kq(x)$ - <u>comparison function</u>
- Generate 2 random nos.

  $x_0 \sim q(x)$

  $u_0 \sim unif(0, Kq(x_0))$ $(u_0|x_0)$

- (x_0, u_0) have uniform distri under curve of $Kq(x)$
- If $u_0 > \hat{p}(x_0)$ then reject sample

otherwise retain u_0.

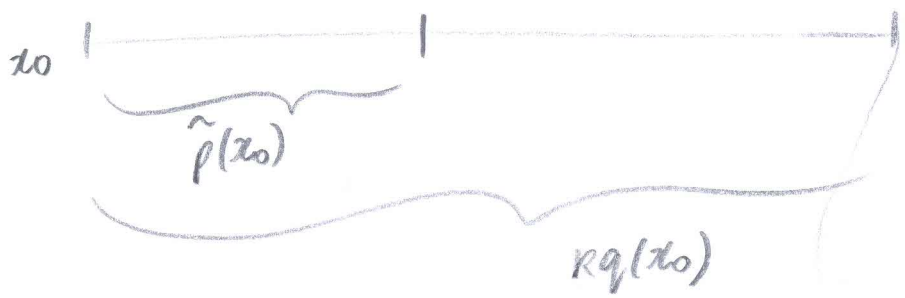


- Remaining pairs in unshaded region ⟶ have uniform distri under the curve of $\hat{p}(x)$
- corresponding x values distri according to $p(x)$ (?)

(*) correctness of rejection sampling (from $p(x)$)

- Original values of x generated from $q(x)$ distri.
- These samples are accepted with probability $\frac{\tilde{p}(x)}{Kq(x)}$

↳ 0 accept $\tilde{p}(x_0)$ reject $Kq(x_0)$



(*) Resulting distn of x after norm:-

$$\frac{[\hat{p}(x)/kq(x)]\,q(x)}{\int[\hat{p}(x)/kq(x)]\,q(x)\,dx} = \frac{k\hat{p}(x)}{k\int\hat{p}(x)\,dx} = \frac{\hat{p}(x)}{\int\hat{p}(x)dx} = p(x).$$

(*) Prob. sample accepted:- typo?

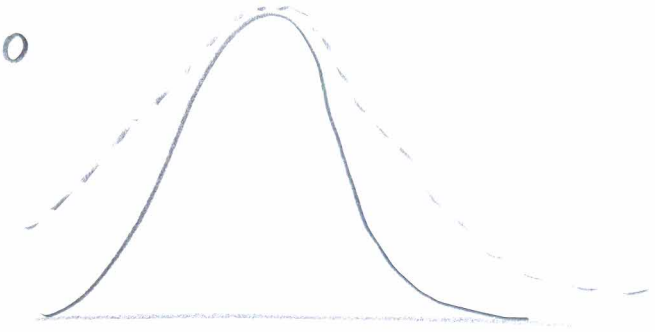
$$p(\text{accept}) = \int[\hat{p}(x)/kq(x)]\,q(x)\,dx = \frac{1}{k}\int\hat{p}(x)\,dx$$ (?) (0/5 3)

- fraction of points rejected depends on ratio of area under unnormalised $\hat{p}(x)$ compared to area under curve $kq(x)$.
- difficulty seeing this intuitively
- (*): constant k must be small as possible subject to constraint that $kq(x)$ is nowhere less than $\hat{p}(x)$

Example:

- sample from Gamma:-

$$\text{Gam}(x|a) = \frac{x^{a-1}\exp(-ax)}{\Gamma(a)} \qquad a>0$$



- use cauchy as $q(x)$.
- Generalise cauchy to ensure that nowhere is smaller than Gamma distri.
- Transform r.v. z:-

$$x = b\tan z + c$$

- Yielding:-

$$q(x) = \frac{k}{1+\frac{(x-c)^2}{b^2}}$$

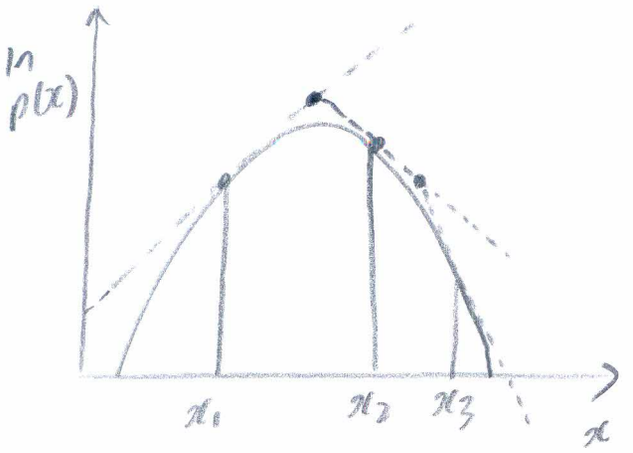
(0/5 4)

(?)

- min rejection rate:-
- set $c = a-1$ $b^2 = 2a-1$
- set k as small as possible while ensuring $kq(x) \geq \hat{p}(x)\ \forall x$

11.3. Adaptive Rejection Sampling

- univariate
- log concave $p(x)$
- construction of envelope fn $q(x)$ on the fly using grid points, tangent lines.



- function ln p(x) evaluated at set of grid points
- use tangent line intersections to construct envelope functions.
- draw samples from envelope distri.

- envelope distri $q(x)$ - piecewise exponent.:-

$$q(x) = k_i \lambda_i \exp\{-\lambda_i (x - x_{i-1})\} \qquad x \in (x_{i-1}, x_i]$$

→ (155) show this.

- usual rejection criterion.
  1. Acceptance → draw from desired distri.
  2. Rejection → incorporate into grid, compute tagent, refine envelope
- more grid points, envelope is better approx of p(x), p(reject)↓.
- adaptive rejection Metropolis sampling.

<u>other points on rejection sampling</u>

(X) comparison function be close to required distri → minimise rejection rate (interests of an efficient algorithm)

- <u>dimensionality issues</u> (illustration)
- sample from zero-mean MVG.

$$\underline{x} \sim N(\underline{0}, \sigma_p^2 \underline{I}) \qquad \underline{x} \in \mathbb{R}^d \quad \text{(isotropic)}$$

- do this by using an MVG proposal :-

$$Q(\underline{x}) = N(\underline{0}, \sigma_Q^2 \underline{I})$$

- for $kq(\underline{x}) \geq p(\underline{x})$ $\forall \underline{x}$; require $\sigma_Q^2 > \sigma_p^2$ (see diagram)
- d-dimensions, optimal value of k is given by (156) - show this

$$k = \left(\frac{\sigma_q}{\sigma_p}\right)^d$$

06/09/2020

- Acceptance rate:
- Ratio of volumes under $p(x)$ and $Kq(x)$ ② $\tilde{p}(x)$?
- Since both distns normalised, this will be $\frac{1}{K}$.
- Acceptance rate $= \frac{1}{K} = \frac{1}{\left(\sigma_Q/\sigma_P\right)^d}$

(0/5 6) - Assess why you couldn't reproduce below argument - Arithmetic error!

- As $\sigma_Q^2 > \sigma_P^2 \Rightarrow \sigma_Q > \sigma_P \Rightarrow \frac{\sigma_Q}{\sigma_P} > 1$
- As $d \to \infty$, acceptance rate $\to 0$ (exponential)
- Numerically: $\sigma_Q = (1.1\sigma_P) \times$ required to ind. acg. $\sigma_Q = 1.01\sigma_P$!

(w). Can't get these numbers/back of envelope calcs to work.

MacKay: - what value of K is required if dimensionality $d = 1000$?

- density of $q(\underline{x})$ at origin is ($\underline{x} = \underline{0}$) $\frac{1}{(2\pi\sigma_Q^2)^{d/2}}$

- for $Kq(\underline{x}) \geq p(\underline{x})$ requires $\Rightarrow$

$$\frac{K}{(2\pi\sigma_Q^2)^{d/2}} \geq \frac{1}{(2\pi\sigma_P^2)^{d/2}} \Rightarrow K \geq \frac{(2\pi\sigma_Q^2)^{d/2}}{(2\pi\sigma_P^2)^{d/2}} = \frac{(2\pi)^{d/2}(\sigma_Q^2)^{d/2}}{(2\pi)^{d/2}(\sigma_P^2)^{d/2}}$$

$$\Rightarrow K \geq \left(\frac{\sigma_Q}{\sigma_P}\right)^d \quad * \Rightarrow K \geq \exp\left(d \ln \frac{\sigma_Q}{\sigma_P}\right)$$ (exp. lower bound on K).

- Set $d = 1000$, $\left(\frac{\sigma_Q}{\sigma_P}\right) = 1.01$. $\Rightarrow K \geq \exp(10) \Rightarrow K \simeq 22000$

- Acceptance rate for this value of K?
- Ratio of volume under $p(\underline{x})$ to volume under $Kq(\underline{x})$.
- As p and q both appropriately normalised i.e. $\int p(\underline{x})\, d\underline{x} = 1$ and $\int q(\underline{x})\, d\underline{x} = 1$

Then acceptance rate is $\frac{1}{K}$.

∴ Not suitable in high-dimensional cases.

- Other issues → multi-modal, sharply peaked, difficult to find good proposal
- Exponential decrease in acceptance is a feature
- But subroutine for more sophis. algos

+ comparison.

11.1.1 Importance Sampling

- Recall there are two problems in sampling:-
  i) Evaluating expectations under a distri $p(x)$.
  ii) Sampling from $p(x)$
- Importance sampling solves (i) but not (ii). - Assume diffi. to draw samples from $p(x)$ directly - supplement this with Markov
- Note: we discuss:-

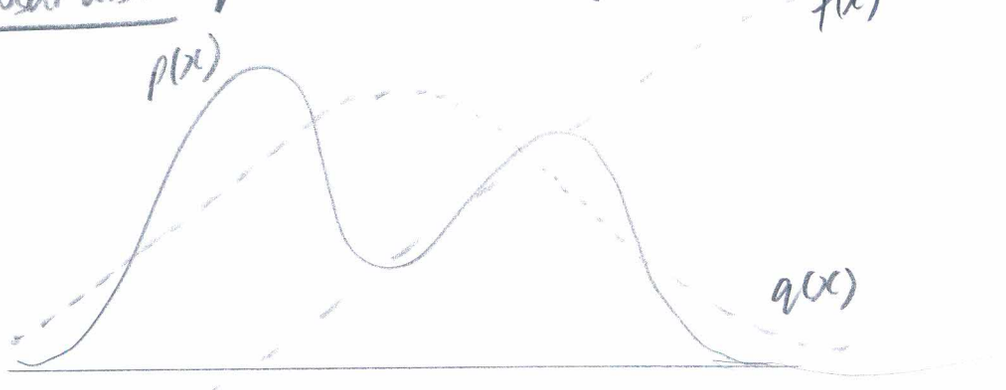
$$\langle f \rangle = \int f(x)\, p(x)\, dx \xrightarrow{\text{approx by}} \langle f \rangle \simeq \sum_{m=1}^{M} p(x^{(m)})\, f(x^{(m)})$$

- using a discretisation of $X \rightarrow \{x^{(1)}, x^{(2)}, \ldots, x^{(M)}\}$ uniformly.
- Issue: No. of terms grows exponentially with dimensionality of x. (?)
  - contradicts claim about accuracy of estimator ind dim?
- This is an example of uniform sampling

- Jordan - to save typos under clarity.

(x) Inefficiency of uniform sampling in high-dim space:
probability dist's of interest have much mass confined in small regions of ~~support~~ x-space.
- In high-dims, only a very small propor. of samples contrib. to sum.

(x) Ideally, select sample points to fall in regions where $p(x)f(x)$ is large.
- use a proposal distri $q(x)$ to draw samples (similar to rejection-accept.)



(*) Draw samples $\{x^{(m)}\}$ from simple $q(x)$ and weight corres. terms in summation with $p(x^{(m)})/q(x^{(m)})$.

$$\langle f \rangle = \int f(x) p(x)\, dx$$

$$= \int f(x) \frac{p(x)}{q(x)} q(x)\, dx \qquad \text{- IS 'trick' similar to V.I.}$$

$$\simeq \frac{1}{M} \sum_{m=1}^{M} \frac{p(x^{(m)})}{q(x^{(m)})} f(x^{(m)}) \qquad x^{(m)} \sim q(x) \qquad (1.28)$$

(w): computing $\left\langle \frac{p(x)}{q(x)} f(x) \right\rangle_q$?

- importance weights:- $r_m = \frac{p(x^{(m)})}{q(x^{(m)})}$ - correct bias introduced by sampling from wrong distribution.

- All samples retained.

Normalised importance sampling:-

- $p(x)$ can only be evaluated up to a normalisation constant.

$$p(x) = \frac{\hat{p}(x)}{z_p}$$

$\hat{p}(x)$ - easily eval.

z_p - unknown norm constant.

- use similar proposal distri $q(x) = \frac{\hat{q}(x)}{z_q}$

$$\langle f \rangle = \int f(x) p(x)\, dx$$

$$= \int f(x) \frac{p(x)}{q(x)} q(x)\, dx = \int f(x) \frac{\hat{p}(x)/z_p}{\hat{q}(x)/z_q} q(x)\, dx$$

$$= \frac{z_q}{z_p} \int f(x) \frac{\hat{p}(x)}{\hat{q}(x)} q(x)\, dx$$

- Hence approx:-

$$\langle f \rangle \simeq \frac{Z_q}{Z_p} \frac{1}{M} \sum_{m=1}^{M} r_m f(x^{(m)}) \qquad r_m = \frac{\hat{p}(x^{(m)})}{\hat{q}(x^{(m)})}$$

should be $\hat{p}(x)$ (typo).

- evaluation of Z_p/Z_q :-

$$\frac{Z_p}{Z_q} = \frac{1}{Z_q} \int \hat{p}(x)\, dx = \int \frac{1}{Z_q} \hat{p}(x)\, dx = \int p(x) \frac{q(x)}{\hat{q}(x)}\, dx$$

(?) (*)

$$\simeq \frac{1}{L} \sum_{m=1}^{L} r_m$$

(*)

(?) typo?

- Hence: $\langle f \rangle \simeq \frac{\sum_m r_m f(x^{(m)})}{\sum_m r_m}$

$L = M$ as same sample set used

(*): think there is a typo and $p(x)$ replaced with $\hat{p}(x)$.

$$\frac{Z_p}{Z_q} = \frac{1}{Z_q} \cdot Z_p = \frac{1}{Z_q} \int \hat{p}(x)\, dx = \frac{q(x)}{\hat{q}(x)} \int \hat{p}(x)\, dx = \int \frac{\hat{p}(x)}{\hat{q}(x)} q(x)\, dx$$

$$\Rightarrow \frac{Z_p}{Z_q} = \left\langle \frac{\hat{p}(x)}{\hat{q}(x)} \right\rangle_{q(x)} = \langle r_m \rangle_q$$

- use finite/empirical approx of $\frac{Z_p}{Z_q}$:-

$$\frac{Z_p}{Z_q} \simeq \frac{1}{L} \sum_{m=1}^{L} r_m \Rightarrow \frac{Z_q}{Z_p} = \frac{1}{\frac{1}{L} \sum_{m=1}^{L} r_m}$$

- As $\langle f \rangle \simeq \frac{Z_q}{Z_p} \frac{1}{M} \sum_{m=1}^{M} r_m f(x^{(m)})$

$$= \frac{\frac{1}{M} \sum_{m=1}^{M} r_m f(x^{(m)})}{\frac{1}{L} \sum_{m=1}^{M} r_m} = \frac{\sum_{m=1}^{M} r_m f(x^{(m)})}{\sum_{m=1}^{M} r_m}$$

0158

- A little confused (need to review again when fresher)
- reconciling with 10-708 L13.
- may just be difference mass. ie. proposal/norm constant as required.

caveats

- dependent on how well sampling distri $q(x)$ matches desired distri $p(x)$.
- $p(x)f(x)$ strongly varying, pmass concentrated on small regions of x
- set of importance weights $\{r_m\}$ dom by few weights having large values
- Effective sample size $<<$ apparent samp size M.
- Even worse $\longrightarrow$ none of samples fall where $p(x)f(x)$ large.
- Apparent variances of r_m and $r_m f(x^{(m)})$ may be small even if est. of exp very wrong.

(*): sampling distri $q(x)$ should not be zero or small where $p(x)$ sig.

- use heavy-tailed $q(\cdot)$ continuous distributions

(*) provides a mechanism for sampling from BN.

(*) and also likelihood-weighted sampling

I.I.S. Sampling - Importance Resampling (weighted resampling)

- Rejection sampling $\rightarrow$ determine k in comparison.
- Issue: - impractical for many $p(x)$, $q(x)$ to determine k which is sufficiently large to guarantee bound or impractically small accept.
- SIR/weighted resampling $\longrightarrow$ use sampling distri, avoid determining k.

2 stage scheme:

- draw N samples $x^{(1)}, \ldots, x^{(N)}$ from $q(x)$. — 1st stage
- construct weights $w^{(1)}, \ldots, w^{(N)}$: — 2nd stage

$$w^{(n)} = \frac{\hat{p}(x^{(n)}) / q(x^{(n)})}{\sum_{n=1}^{N} \tilde{p}(x^{(n)}) / q(x^{(n)})}$$

- 2nd set of M samples drawn from discrete distri $(x^{(1)}, \ldots, x^{(N)})$ with probabilities given by $(w^{(1)}, \ldots, w^{(N)})$. ($M >> N$)

characterising these M samples

- the M samples are only approximately distributed according to $p(x)$.
- that is distri becomes correct as $N \to \infty$ (converged in distri?)

CDF of resampled values:-

$$Pr(x \leq a) = \sum_{n: x_n \leq a} w^{(n)}$$

$$= \frac{\sum_{n=1}^{N} \mathbb{I}(x_n \leq a)\, \hat{p}(x_n)/q(x_n)}{\sum_{n=1}^{N} \hat{p}(x_n)/q(x_n)}.$$

- take limit as $N \to \infty$, assuming regularity conditions:-
- Replace summations with integrals weighted according to orig sampling distri $q(x)$. ↙ (?)

$$\cdot\ Pr(x \leq a) = \frac{\int [\mathbb{I}(x \leq a)\, \hat{p}(x)/q(x)]\, q(x)\, dx}{\int [\hat{p}(x)/q(x)]\, q(x)\, dx}$$

(p154)

$$= \frac{\int \mathbb{I}(x \leq a)\, \tilde{p}(x)\, dx}{\int \hat{p}(x)\, dx}$$ ↙ (?)

$$= \int \mathbb{I}(x \leq a)\, p(x)\, dx$$

→ CDF of $p(x)$.
→ Normalisation of $p(x)$ not required.

- finite N, given initial sample set, resampled values only approx. drawn from desired distn $p(x)$.
- approx. improves (à la rejection sampling) as

$$q(x) \longrightarrow p(x)$$

- $q(x) = p(x)$, initial samples $(x^{(1)}, \ldots, x^{(N)}) \sim p(x)$

and weights $w^{(n)} = \frac{1}{N}$.

- Resampled values then have desired distr.

- computation of moments:

- use original samples and weights:-

$$\langle F(x) \rangle = \int F(x) p(x)\, dx$$

$$= \frac{\int F(x) [\hat{p}(x)/q(x)]\, q(x)\, dx}{\int [\hat{p}(x)/q(x)]\, q(x)\, dx}$$

$$\simeq \sum_{n=1}^{N} f(x_n) w_n$$

(?) review. intuitively not 'clicking'

1.1. Particle filtering

- use weighted resampling ideas to obtain SMC algorithm - particle filter
- particularly for online applications in which we make posterior inferences / updates in light of new observations.
- HMM + KF → use discrete or linear Gaussian distri.
- Particle filtering allows for more complex choices of observation/emission $p(y_t|x_t)$, and where posteriors are analytically intractable.
- recall time and measurement updates from Jordan (2003).

Time update: $p(x_t|y_0, \ldots, y_t) \longrightarrow p(x_{t+1}|y_0, \ldots, y_t)$

Measurement update: $p(x_{t+1}|y_0, \ldots, y_t) \longrightarrow p(x_{t+1}|y_0, \ldots, y_{t+1})$

- Given observed values $y_{(t)} = (y_1, \ldots, y_t)$
- Draw M samples from posterior $p(x_t|y_{(t)})$ to eval. exp. of $f(x)$ wrt posterior.

- Bayes: $\langle f(x_t) \rangle = \int f(x_t)\, p(x_t|y_{(t)})\, dx_t$

$$= \int f(x_t)\, p(x_t|y_t, y_{(t-1)})\, dx_t$$

$$= \int f(x_t) \cdot \frac{p(y_t, x_t|y_{(t-1)})}{p(y_t|y_{(t-1)})}\, dx_t$$

$$= \int f(x_t) \cdot \frac{p(y_t|x_t)\,p(x_t|y_{(t-1)})}{\int p(y_t|x_t)\,p(x_t|y_{(t-1)})\,dx_t}\,dx_t$$

- constant + x_t integrated out.

$$= \frac{\int f(x_t)\,p(y_t|x_t)\,p(x_t|y_{(t-1)})\,dx_t}{\int p(y_t|x_t)\,p(x_t|y_{(t-1)})\,dx_t}$$

$$\simeq \sum_{m=1}^{M} \left(\frac{p(y_t|x_t^{(m)})}{\sum_{m=1}^{M} p(y_t|x_t^{(m)})} \right) f(x_t^{(m)}) \quad \text{where } \{x_t^{(m)}\}_{m=1}^{M} \sim p(x_t|y_{(t-1)})$$

- use of C.I. $p(y_t|x_t, y_{(t-1)}) = p(y_t|x_t)$ ⑤ - insure how this features (check prob rules)

- defining: $w_t^{(m)} = \dfrac{p(y_t|x_t^{(m)})}{\sum_{m=1}^{M} p(y_t|x_t^{(m)})}$ yields

$$\Rightarrow \langle f(x_t) \rangle \simeq \sum_{m=1}^{M} w_t^{(m)} f(x_t^{(m)})$$

set of samples

⑧ (*) posterior distri $p(x_t|y_t)$ (*) is represented by $\{x_t^{(m)}\}_{m=1}^{M}$ and weights $\{w_t^{(m)}\}_{m=1}^{M}$

- weights: $0 \leq w_t^{(m)} \leq 1$ $\sum_m w_t^{(m)} = 1$

(*) Typo:- $p(x_t|y_{(t)})$!

- Suppose:

- we have a set of:-

1) samples $\{x_t^{(m)}\}_{m=1}^{M} \sim p(x_t|y_{(t-1)})$

2) weights $\{w_t^{(m)}\}_{m=1}^{M} = \dfrac{p(y_t|x_t^{(m)})}{\sum_{m=1}^{M} p(y_t|x_t^{(m)})}$

} obtained at time step t, represents post. distri $p(x_t|y_{(t)})$

⑧ use weighted resampling rep. of $p(x_t|y_{(t)})$ to iteratively infer $p(x_{t+1}|y_{(t+1)})$.

- Then we observe y_{t+1} and wish to find weights and samples at time $(t+1)$ i.e. $p(x_{t+1}|y_{(t+1)})$

PTO. ⟶

- Note have used C.I properties:-

$p(x_{t+1}|x_t, y_{(t)}) = p(x_{t+1}|x_t)$ via d-sep.

$p(y_t|x_t, y_{(t-1)}) = p(y_t|x_t)$

- similar to mixture distribution

- samples drawn by selecting component m with prob given by mixing coefficients $w^{(m)}$ and drawing sample from corresp. component.

(W): Jordan (2003) is very ref here! + measurement update unclear

- Also using lecture slides 10-708 L13. to supplement.

- Summary: - At time step t, have sample-rep of post $p(x_t|y_{(t)})$ contained in $\{w_t^{(m)}\}_{m=1}^M$ and $\{x_t^{(m)}\}_{m=1}^M$. (mixture rep.)

- For next step, draw M samples from mixture distri. and for each sample use y_{t+1} to evaluate corres weights $w_{t+1}^{(m)} \propto p(y_{t+1}|x_{t+1}^{(m)})$?

(*) develop measurement update as unclear!

(W): made a decision to re-present particle filtering, with a view to pseudo will clarity (effectively merging Jordan (2003) and 10-708 L13).

- sequential weighted resampling for particle filtering consists of:-

1) starting point - sample based representation of the posterior $p(x_t|y_{(t)})$ using weights $\{w_t^{(m)}\}_{m=1}^M$ and samples $\{x_t^{(m)}\}_{m=1}^M$

2) time-update - computation of $p(x_{t+1}|y_{(t)})$ using above sampling representation ← Jordan covers this out this is very vague

3) measurement update - computation of $p(x_{t+1}|y_{(t+1)})$ using time-update and also newly evaluated weights.

GOAL: sample from an intractable posterior.

First sample from $p(x_{t+1}|y_{(t)})$ via Bayes:

Time update

$$p(x_{t+1}|y_{(t)}) = \int p(x_{t+1}|x_t, y_{(t)})\, p(x_t|y_{(t)})\, dx_t$$

$$= \int p(x_{t+1}|x_t)\, p(x_t|y_{(t)})\, dx_t$$

$$= \int p(x_{t+1}|x_t)\, p(x_t|y_t, y_{(t-1)})\, dx_t$$

$$= \int p(x_{t+1}|x_t) \cdot \frac{p(y_t, x_t|y_{(t-1)})}{p(y_t|y_{(t-1)})}\, dx_t$$

$$= \int p(x_{t+1}|x_t) \cdot \frac{p(y_t|x_t)\, p(x_t|y_{(t-1)})}{\int p(y_t|x_t)\, p(x_t|y_{(t-1)})\, dx_t}\, dx_t$$

(no x_t terms)

$$= \frac{\int p(x_{t+1}|x_t)\, p(y_t|x_t)\, p(x_t|y_{(t-1)})\, dx_t}{\int p(y_t|x_t)\, p(x_t|y_{(t-1)})\, dx_t}$$

- using similar principles to earlier.

$$\simeq \sum_{m=1}^{M} \left(\frac{p(y_t|x_t^{(m)})}{\sum_{n=1}^{M} p(y_t|x_t^{(n)})} \right) p(x_{t+1}|x_t^{(m)})$$

- mean

$$= \sum_{m=1}^{M} w_t^{(m)}\, p(x_{t+1}|x_t^{(m)})$$

(x) like mixture model!

That is, we use representation of posterior $p(x_t|y_{(t)})$ given by weights and samples:-

1) samples $\{x_t^{(m)}\}_{m=1}^{M} \sim p(x_t|y_{(t-1)})$

2) weights $\{w_t^{(m)}\}_{m=1}^{M}$

1. Starting point

- At time step t, we assume a sampling-based representation of the posterior $p(x_t|y_{(t)})$, and assume this is available to us.
- Recall this is because $p(x_t|y_{(t)})$ is analytically intractable, due to it being likely that there is no conjugacy we can exploit (e.g. in K.F.).
- The posterior $p(x_t|y_{(t)})$ has following expression:-

$$p(x_t|y_{(t)}) = p(x_t|y_t, y_{(t-1)}) = \frac{p(x_t|y_{(t-1)})p(y_t|x_t)}{\int p(x_t|y_{(t-1)})p(y_t|x_t)dx_t} \quad \left(= \frac{p(y_t, x_t|y_{(t-1)})}{p(y_t|y_{(t-1)})}\right)$$

- We represent $p(x_t|y_{(t)})$ using:-

$$\left\{ x_t^{(m)} \sim p(x_t|y_{(t-1)}),\ w_t^{(m)} = \frac{p(y_t|x_t^{(m)})}{\sum_{m=1}^{M} p(y_t|x_t^{(m)})} \right\}_{m=1}^{M}$$

$x_t^{(m)}$ drawn from ?

- weights computed using observation distn $p(y_t|x_t^{(m)})$

2. Time update

- Note that:-

$$p(x_{t+1}|y_{(t)}) = \int p(x_{t+1}|x_t)p(x_t|y_{(t)})dx_t$$

→ see earlier for details on how C.I. assumptions are used. + derivations

- Which is represented by:-

$$p(x_{t+1}|y_{(t)}) \approx \sum_{m=1}^{M} w_t^{(m)} p(x_{t+1}|x_t^{(m)})$$

(akin to mixture model)

- We can draw samples of x_{t+1} from this mixtre distn using $x_t^{(m)}$ and $w_t^{(m)}$!

3. Measurement update

- Note that measurement update is the posterior $p(x_t|y_{(t)})$ (starting point) propagated forward by one time-step and using new evidence y_{t+1} i.e:-

$$p(x_{t+1}|y_{(t+1)}) = p(x_{t+1}|y_{t+1}, y_{(t)}) = \frac{p(x_{t+1}|y_{(t)})p(y_{t+1}|x_{t+1})}{\int p(x_{t+1}|y_{(t)})p(y_{t+1}|x_{t+1})dx_{t+1}}$$

· to get the weighted, resampled vesion of $p(x_{t+1}|y_{(t+1)})$, we can represent it using:-

$$\left\{ x_{t+1}^{(m)} \sim p(x_{t+1}|y_{(t)}) \;,\; w_{t+1}^{(m)} = \frac{p(y_{t+1}|x_{t+1}^{(m)})}{\sum_{m=1}^{M} p(y_{t+1}|x_{t+1}^{(m)})} \right\}_{m=1}^{M}$$

· we draw samples $x_{t+1}^{(m)}$ using the mixture model sample representation of the time update $p(x_{t+1}|y_{(t)})$

· Then compute new weights $w_{t+1}^{(m)}$ using $x_{t+1}^{(m)}$, new evidence y_{t+1}, and our observation distri $p(y|x)$. (plug values in)

(*) This concludes particle filtering → now see diagram. (doesn't make sense in Jordan - diffs from 10-708)

- corrected diagram: