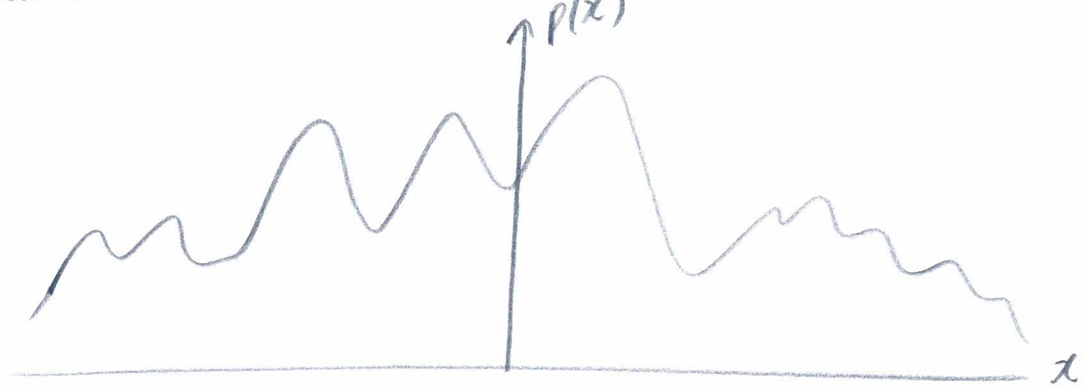
10-708 24/06/2020

L13 - Approx Inference: Monte Carlo and sequential Monte Carlo

(*) How to represent a joint m.d.g.

- $\mathbb{E}_p[f(x)] = \int f(x)p(x)\,dx$
- sample based representation:
- conside a multimodal distribution (and density):-



(?) How to integrate?
- It may not be feasible (analytically)? to integrate this.
- sol: draw samples from distn
- computationally: will use a uniform PRNG to draw a sample $U(0,1)$ and transform to get Gaussian.

(*) Draw N samples $x^{(i)} \sim P$:-

$i = 1, \dots, N$

compute $\mathbb{E}_p[f(X)] = \dfrac{\sum_{i=1}^{N} f(x^{(i)})}{N}$

(*) original distn P expressed as a set of samples

Monte Carlo methods

(*) Monte Carlo methods

(*) replace integration with summation (sample-based averages)

(*) Asymptotically exact → more samples drawn $(N \to \infty)$ arbitrarily close to expectation we wish to approx.

Challenges:

1) How to draw samples from distr?
  - How to do this efficiently? from an arbitrary
2) How to make better use of samples?
  - asymptote - how many samples can we draw ...

(*) Naive sampling

$p(x_A, \ldots, x_m) = p(x_B)p(x_A|x_B) \ldots$

$$\underline{x}^{(k)} = \begin{bmatrix} x_A \\ \vdots \\ x_m \end{bmatrix} \sim p$$

- How to draw samples?
  - traverse graph
- (*) ex walks through realisation of sample in BN
  - sample r.v.s. from B.N.

(*) Rows of table correspond to a 'sample' from B.N.

(*) Computation of queries: check samples where condition is met.

(*) Rare events ⟶ conditioning on rare events?

(*) Shows the limitations of sampling

(*) Monte Carlo methods (Rejection sampling)

- Attempt to get around above deficiencies
- $\pi(x) = \dfrac{\pi'(x)}{Z}$

  $\pi(x)$ - difficult to sample from

  $\pi'(x)$ - easy to evaluate (eg. unnormalised clique pot. product)

- Sample from simpler distr $Q(x)$

(5) : $x^* \sim Q(x)$ (draw samples from Q)

- Accept x^* with probability $\frac{\pi'(x^*)}{kQ(x^*)}$ k-arbitrary constant

(*) Ratio of unnormalised part of given (evaluatable distn) and a constant multiplied by proposal distribution.

(*) with these two steps:-

$x^* \sim \pi(x)$

Ex: How to prove this?

$$P(x^*) = \frac{Q(x^*)\frac{\pi'(x^*)}{kQ(x^*)}}{\int Q(x^*)\frac{\pi'(x^*)}{kQ(x^*)}dx^*} = \frac{k\pi'(x^*)}{k\int \pi'(x^*)dx^*} = \pi(x)$$

(*) A solution to sampling from a distn $\pi(x)$ which is difficult to normalise (using a proposal distrib)

Pitfall: (Q1): Geometric intuition / explanation of too many rejections ?

(*) Rejection sampling

- conceptual experiment to illustrate rejection sampling
- Small differences between proposal distn and distn of interest caused by scaling constant can cause a large difference in rejection rate
- $k = \left(\frac{\sigma_q}{\sigma_p}\right)^d \approx \frac{1}{20{,}000}$

(*) The actual vol. of 2 distns are vastly different in high-dim space → reject samples most of time even though distns are close

(*) solution: Don't use one Q-distn (proposal distn) to cover the target distn, but use a piece-wise envelope

(*) Adaptive rejection sampling is an example of this
- So rejection sampling doesn't work well in a high-dim space

(*) unnormalised importance sampling
- Draw, as before, samples $x^{(i)} \sim Q(x)$ from a proposal distr. N times
- instead of accepting/rejecting as earlier, compute weight based on new sample

$$w^{(i)} = \frac{P(x^{(i)})}{Q(x^{(i)})}$$

- store $\{x^{(i)}, w^{(i)}\}_{i=1}^{N}$ as representation of true distr.

samples
(*) Why is P suddenly available/tractable?
- Assume in simplest case that P can be eval (up to norm. constant)

(A2): Assess proof, picture.
- Proposal distr introduced as a dummy in proof
- from likelihood ratio
(*) unnormalised as weights directly comp. from likelihood ratio (do not sum to one).

Ex. what if I can only evaluate true distr up to a constant? (?)
e.g. no exact value of norm. constant (define to be arbitrary α)

(*) Normalised importance sampling
- only can evaluate $p'(x) = \alpha P(x)$ (e.g. MRF)
- Get around normalisation constant:-
(*) take expectation of the ratio

$$r(x) = \frac{p'(x)}{Q(x)} \Rightarrow \langle r(x) \rangle_Q = \int \frac{p'(x)}{Q(x)} Q(x)\, dx = \int p'(x)\, dx = \alpha$$

- expectation of ratio is normalisation constant $\alpha = \frac{r^{(i)}}{\sum_{i=1}^{N} r^{(i)}}$

- Now draw $x^{(i)} \sim Q(x)$ $i=1,\dots,N$ (N samples)

- compute $r^{(i)}$ for each i^{th} sample.

- compute α via $\alpha = \sum_{i=1}^{N} r^{(i)}$

A3 - confused here
- review derivations

Proof:

(*) compute expectations of function using samples drawn as above
(*) instructor presentation lost me. (*) don't need to know norm constant or part function of target

---

(*) normalised vs unnormalised importance sampling
A4 - review at home (instructor does not emphasise).

(*) efficiency of likelihood weighting
- so far we have used likelihood weighting → may not get 'true' answer in peculiar/patho scenarios
- consider practical implementation of MCMC. when do we stop sampling?
typically stop when stability/convergence observed as empir. strategy for controlling iteration/algo duration.
- in a technically equivalent way; can view this as terminating when variance ↓ of current solution.
- makes sense. (i.e. convergent behavior through variance)
(*) So variance in a sense can be viewed as measure of closeness to true answer (of current algo output)

(*) In importance sampling; we can have a pathological case:-
A5: Review; watching + record intuition
(*) Example of poor proposal distn (note we do not know its poor);
does not envelop true distn.
(*) Hoping weights $w^{(i)}$ will correct this.
Solution: use heavy tailed Q → (but many samples wasted)
· weighted resampling from $\{x^{(i)}, w^{(i)}\}_{i=1}^{N}$

- we resample from $\{w^{(i)}\}_{i=1}^{N}$ N' times where $N' >> N$.

(*) 'Amplify' distr using ~~resa~~ weighted resampling (Ab): Review weighted resampling

(*) weighted resampling

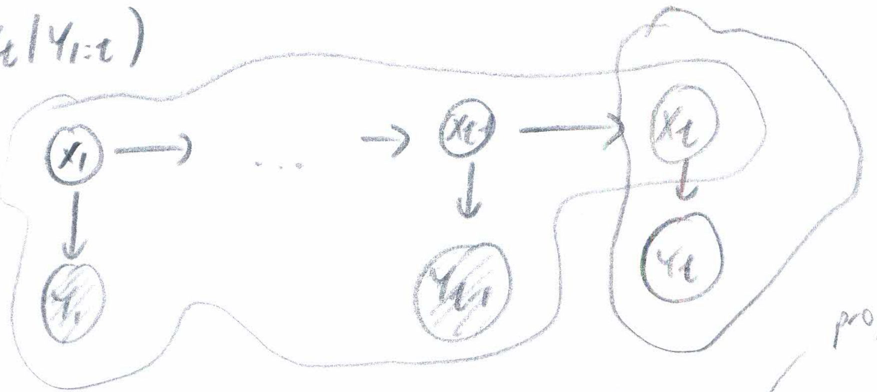
(*) Particle filtering

- use resampling idea to elegant, efficient inference.
- make a fast sampling based algorithm for SSMs.
- KF demanding to implement in high-dim space, or non-Gaussian (Taylor approx)

(*) Sketch of particle filtering

- use KF to establish recursive procedure.

$p(X_t | Y_{1:t})$



- Break down $p(X_t|Y_{1:t}) \rightarrow$ 2 parts.

$$p(X_t|Y_{1:t}) = p(X_t|Y_t, Y_{1:t-1}) = \frac{\overbrace{p(X_t|Y_{1:t-1})}^{\text{prop.}}\,\overbrace{p(Y_t|X_t)}^{\text{'weights'}}}{\int p(X_t|Y_{1:t-1})\,p(Y_t|X_t)\,dX_t}$$

- Inspect $p(X_t|Y_{1:t-1})$ and use this as a proposal distrib. from which
- we sample
- Represent $p(X_t|Y_{1:t})$ using above decomp:-

$$\left\{ X_t^{(i)} \sim p(X_t|Y_{1:t-1}),\; w_t^{(i)} = \frac{p(Y_t|X_t^{(i)})}{\sum_{i=1}^{N} p(Y_t|X_t^{(i)})} \right\}$$

† Hence $p(X_t|Y_{1:t})$ is expressed as weighted resamples from proposal distr.

(*) Goal:- use weighted resample representation of $p(X_t|Y_{1:t})$ to iteratively infer $p(X_{t+1}|Y_{1:t+1})$

(*) sequential weighted resampler

Time update: keep evidence; propagate forward with no new measurement

$$p(X_{t+1} \mid Y_{1:t}) = \int \underbrace{p(X_{t+1} \mid X_t)}_{(i)} \underbrace{p(X_t \mid Y_{1:t})}_{} \, dX_t$$

↳ given by weighted res. rep.

· i) Given via model

(*) Replace integration with sampling :-

$$p(X_{t+1} \mid Y_{1:t}) = \sum_{i=1}^{N} w_t^{(i)} p(X_{t+1} \mid x_t^{(i)})$$

- weighted sum of transition prob. condit. on samples of previous states.

(*) Similar to mixture model (sampling from).

(*) Given new evidence :-

new prop.

reweights

measurement update :- $p(X_{t+1} \mid Y_{1:t+1}) = \dfrac{\overset{(*)}{p(X_{t+1} \mid Y_{1:t})} \, p(Y_{t+1} \mid X_{t+1})}{\int p(X_{t+1} \mid Y_{1:t}) \, p(Y_{t+1} \mid X_{t+1}) \, dX_{t+1}}$

(*) note (*) is computed from time update

- Replace by weighted resampling version :-

⇒ $p(X_{t+1} \mid Y_{1:t+1})$ has representation :-

$$\left\{ X_{t+1}^{(i)} \sim p(X_{t+1} \mid Y_{1:t}), \; w_{t+1} = \frac{p(Y_{t+1} \mid X_{t+1}^{(i)})}{\sum_{i=1}^{N} p(Y_{t+1} \mid X_{t+1}^{(i)})} \right\}$$ reweighted

(*) Particle filtering graph (illustrating resampling)

- Size of balls correspond to weights

(?) - review

(*) Particle filtering
gives online results for predictive distri

(*) PF for SSMs (switching)

- Allows drawing from more complex distri e.g switching SSM
- multiple hidden distns.

(*) Rao-Blackwellised sampling

(RS): Review slides (not emphasized)

(*) Statistical properties of unnormalised/normalised importance sampling.

- unnormalised importance sampling is unbiased

i.e. $\mathbb{E}_Q[f(x)w(x)] = \mathbb{E}_P[f(x)]$

- Normalised importance sampling is biased for finite samples e.g. $M=1$

$$\mathbb{E}_Q\left[\frac{f(x')\,r(x')}{\sum r(x')}\right] \neq \mathbb{E}_P[f(x)]$$