06/07/2020

Lec - Approx. Inf: MCMC

- Recap of Monte Carlo
- generate samples from p.d. $x^{(i)}$
- estimate expectations of functions under a p.d.
$\mathbb{E}[f(x)]$ under $p(x)$.

Ex. How does this relate to inference covered in course?

e.g. $p(x_a|e)$, $\log(Z)$, Z

- marginal query given evidence
- log-likelihood
- partition

- $p(x_a|e) = \mathbb{E}[\mathbb{I}(x_a)|e]$
- Fn. $\log(Z)$, Z, recall that the normalisation constants correspond to weights of samples (A1) - using sampling can address inf. tasks

(*) Limitations of Monte Carlo

- Related to quality of proposal distri Q.
- Tried approach to boost acceptance rates of samples gen by Q.
- Cause they (samples) are mostly accepted, they ensure different regions in distributions which are multi-modal etc.
- Focus on augmenting Q via a more aggressive Q, which updates
- adaptive.

(*) Markov Chain Monte Carlo

- An example of importance sampling with a bad proposal $Q(x)$
- However the samples are manipulated, it will be hard to 'cover' the distri $P(x)$ using $Q(x)$.
- A long tail
- Use adaptive Q
- Instead of $Q(x')$, use $Q(x'|x)$ where x' - new state being sampled
x - previous sample.

(?) As x changes,
$Q(x'|x)$ can also change (as a fn of x')

- see diagram
- Draw sample $x^{(1)}$; generate $Q(x^{(2)}|x^{(1)})$ i.e. proposal distn. on $x^{(1)}$
- (centred on $x^{(1)}$)
- Draw sample $x^{(2)}$; generate $Q(x^{(3)}|x^{(2)})$
- Probability mass shifts to the right progressively; allowing us to better capture the target distr.

EX: will study how to operationalise this high-level idea; ensure correctness to make sure it is 'correct'. i.e. after many samples, they are coming from true target distr P.

(*) Metropolis-Hastings (MH) (transitional kernel)

- Draw a sample x' from $Q(x'|x)$, where x is the previous sample.
- The new sample x' is accepted or rejected with some prob. $A(x'|x)$.
- Acceptance prob. $A(x'|x) = \min\left(1, \frac{P(x')\,Q(x|x')}{P(x)\,Q(x'|x)}\right)$
- $A(x'|x)$ ensures that after sufficiently many draws, samples will come from true distr $P(x)$.
- $A(x'|x)$ - ratio of importance weights
- $\frac{P(x')}{Q(x'|x)}$ - importance weight for x'
- $\frac{P(x)}{Q(x|x')}$ - importance weight for x
- This ratio can be smaller or greater than 1

(*) If ratio > 1, accept the sample with prob 1

(*) MH Algorithm

(12). review pseudocode

- simple
- Ex: Need to remember acceptance rate for new samples
- Illustration of how a proposal distri covers both modes of multimodal distr.

(Q11): In practice, we don't know P

(*) When we reject a sample, take older example and put in bag, re-use proposal

(13) Review this example

(14) not very likely under proposal, but we are 'lucky' to sample it.

- However, it gets 'boost' due to high prob under target distri P.

(*) Therefore it has a better importance weight and is accepted.

(*) Using this dynamic, we amplify the tend of moving to 'pope modes' under target distri P; even though the proposal doesn't give a good chance of accepting that point.

(*) (Q7): Adaptive proposal allows us to capitalise and get a good sample.

(*) Previous examples: fixed proposal + some kind of augmentation

(*) Can still sample from bimodal target with unimodal proposal, as the (or bino) latter moves around!

(*) Theoretical aspects of MCMC

(*) Burn-in: first 1,000/10,000 samples have to be thrown away.

- Nature of 'burn-in' is a little mysterious (art)

(Q11): Why are MH samples guaranteed to be from P(x) - prove

(*) MC and MCMC:-

- Adaptively change Q based on previous example.
- In principle depends on all previous examples, but we assume a 1st order Markov property.

(*) Markov chain:-

- sequence of r.v.s. $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$ with M.P.:-

$$p(x^{(n)} = x \mid x^{(1)}, \ldots, x^{(n-1)}) = p(x^{(n)} = x \mid x^{(n-1)})$$

(07) $p(x^{(n)} = x \mid x^{(n-1)})$ - transition kernel

- x can be a scalar or vector

(*) Assume that next state only depends on previous state (?)

(*) Homogeneous-time invariant transition matrix.

Hence $T(x'|x)$ is not indexed by time.

(*) MC concepts

Ex: What is effect of conditioning on previous sample (A3)-review.

(07): $\pi^{(t+1)}(x') = \sum \pi^{(t)}(x) T(x'|x)$

- MCMC: Hope for a protocol generating samples from a target distri.

⇒ Stationary distri: $\pi(x') = \sum_x \pi(x) T(x'|x) \quad \forall x'$

- distribution after transitioning (using kernel) leads to a distri that is identical to 'starting point'.
- this generates constraints on T.
- not every T allows for this stationarity

Ex: When do we have a stationary distri?

- characterise properties (should be familiar)

(A4) review intuition on these properties.

(*) Irreducibility

(*) Aperiodicity

(*) Ergodicity (irreducible + aperiodic) (07)

↳ Allows stationary distri.

Ex: What transition kernel T satisfies these conditions?

(*) Reversible MC

- An MC is reversible if there exists a distn $\pi(x)$ such that the detailed balance equation is satisfied

$$\pi(x')\,T(x|x') = \pi(x)\,T(x'|x)$$

- Probability $x' \to x$ is same as $x \to x'$

(TH): reversible MCs always have a stationary distn

- simple proof: start with detailed balance, derive stationary.

(*) Why does Metropolis-Hastings algorithm work?

- Sh: MH satisfies detailed balance
- (PS) review the proof (from setting $Q = TA$
- (OH): - MH algorithm defines a transition kernel that satisfies detailed bal.
  - Because of the way we defined our proposal Q and acceptance A, MH allows to get distn to be the stationary distn
- It has theoretical guarantees

Caveats

- No guarantees on when convergence occurs, only that it does occur.
- Knowing when to halt burn-in → artform
- rely on heuristics

(*) Gibbs Sampling

- special case of MCMC.
- simplification of MCMC.
- $Q(\underline{x}^{(t=2)} | \underline{x}^{(t=1)})$

$$\underline{x}^{(t=1)} = \begin{pmatrix} x_1^{(t=1)} \\ x_2^{(t=1)} \\ \\ x_n^{(t=1)} \end{pmatrix} \quad \underline{x}^{(t=2)} = \begin{pmatrix} x_1^{(t=2)} \\ x_2^{(t=1)} \\ \vdots \\ x_n^{(t=1)} \end{pmatrix}$$

- Note for $\underline{x}^{(t=2)}$, we use $Q(\cdot)$ in such a way that $x_1^{(t=2)}$ is a new value; $x_{i \neq 1}^{(t=2)} = x_{i \neq 1}^{(t=1)} \quad \forall i$

- $X_1^{(t=2)} \sim P(X_1 | X_2^{(t=1)}, X_3^{(t=1)}, \ldots, X_n^{(t=1)})$

- This $P(\cdot)$ is the proposal distri $Q(\underline{x}^{(2)} | \underline{x}^{(1)})$

(*) Sample next example such that all but one dimension/component is same as previous example.

(*) That different dimension will be sampled conditional on all the dimensions that have already been sampled.

(67): Note we can further reduce $P(X_1 | X_2^{(t=1)}, X_3^{(t=1)}, \ldots, X_n^{(t=1)})$ in a graphical model.
By using $P(X_1 | MB(X_1))$ where MB is Markov blanket

(68): A specialised way to draw the next sample where you only change one dimension and keep the others; the condition on those.

Ex: Interesting property; this proposal $Q(\cdot)$ leads to an acceptance of 1

(*) Gibbs sampling → (16) - review

(*) Distinction between epoch → is it one dim or the whole vector that is refreshed to count as one 'epoch'?

- varies in literature

(*) Gibbs sampling - example

- Previous alarm network
  - Not many effective samples

- T=0 - Random guess/intentional start point

(47) - Review this (unclear)

(*) Exploit C.I. prop of MB; rather than high-dim proposal of all states

(?): As acceptance rate is 1, we use all the examples?

(*) Topic Models:- Collapsed Gibbs

(*) Gibbs sampling became well-known in ML due to mentioned paper (Griffiths + Steyvers)

- LDA: used MFA in V.I.
- ML-paper: integrate out variables; only sample word-topic assignments z

(*) (PB) - review plate, algo

- Algo: For all variables $\underline{z} = z_1, z_2, \ldots, z_n$
  - Draw $z_i^{(t+1)}$ from $P(z_i | \underline{z}_{-i}, \underline{w})$
  - with $\underline{z}_{-i} = z_1^{(t+1)}, z_2^{(t+1)}, \ldots, z_{i-1}^{(t+1)}, z_{i+1}^{(t)}, \ldots z_n^{(t)}$

$P(z_i | \underline{z}_{-i}, \underline{w})$:-

$$P(z_i = j | \underline{z}_{-i}, \underline{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i}^{(d_i)} + T\alpha}$$

'word-topic' 'doc-topic'

- Two types of weighted counts (product).
- word-topic → no. of words appearing under a particular topic over all words
  topic-specific word frequency
- doc-topic → no. of topics appearing under a particular document over all topics
  document specific topic freq.

(PB) - detailed in slides, how nos. can be evaluated from zs.

(*) collapsed Gibbs Illustration

- shows indexed words of documents
- z_i refers to z_i of document j (z_{ij})

(*) Iteration 1: Given from initialisation

- z_1? - 1st word of 1st document → topic?
- use proposal dist? of one single word given all others

$$P(z_i = j | \underline{z}_{-i}, \underline{w}) \propto$$

(*) Draw $z_{1,1}^{(2)}$ (or draw conditioning on $z_{2,1}^{(1)}, z_{3,1}^{(1)}, z_{4,1}^{(1)} \cdots z_{50,5}^{(1)}$) — (*) Incorrect indexing here

(*) Draw $z_{2,1}^{(2)}$ conditioning on $\{z_{3,1}^{(1)}, z_{4,1}^{(1)} \ldots, z_{50,5}^{(1)}\}$ - but works idea

⋮

(*) Large-scale topic models implement sophisticated indexing programs (to store and retrieve efficiently)

Ex: Extremely simple, liked by practitioners

- Allows for additional variables in TM design
- can index word in different ways e.g. introduce sentiment index s

$z_{i,j,s}$ where s indexes sentiment.

(*) Gibbs sampling as special case of MH.

GS proposal:- $Q(x_i, x_{-i} | x_i, x_{-i}) = P(x_i' | x_{-i})$

x_{-i} - all variables except x_i

(A11) - Review maths here

Ex: Now have 2 algorithms with proof of correctness

(*) Practical aspects of MCMC

- Proposal distn quality $Q(x'|x)$ ⟶ acceptance rate, autocorr
- when to stop burn-in? ⟶ sample values vs time
  log-like —"—

(*) Acceptance Rate

(*) Issue is you don't know P; you have to rely on samples blindly to stop. from Q(·)

(*) Slides illustrate some shortcomings & tradeoff

(*) There are tradeoffs between exploration (variance) and acceptance.

(*) some guidelines are proposed

- see papers for provenance Neal (1993), Roberts, Gelman (1994)

(*) Autocorrelation function

(*) MCMC examples are always correlated (you are using previous examples to generate transition for new).

- autocorrelation function:- $R_x(k) = \dfrac{\sum_{t=1}^{n-k} (x_t - \bar{x})(x_{t+k} - \bar{x})}{\sum_{t=1}^{n-k} (x_t - \bar{x})^2}$

(*) when samples are correlated, there exist formulae to compute effective sample size

(*) e.g. for 2 samples where 2nd is twice the magnitude of the 2nd (does this count as 2 'samples'?) - really 1.

- no info other than linear scaling (n).

EX: Compute autocorrelation function to encourage good proposal.

- want low autocorrelation

(*) Monitoring convergence

- interesting statistics

- e.g. if you want Gaussian parameter estimates

  - monitor.

(*) note well-mixed and poorly-mixed chains

(*): in practice, use multiple chains due to poor initial random guesses (rather like multiple initialisations of EM)

(*) issue: - high dim GMs (difficult to visualise all r.v. chains at once)

(*) log-likelihood vs time.

- see profile of log-likelihood over time

- these are main MCMC diagnostic tools

(*) that covers:-
i) Metropolis-Hastings algo
ii) Gibbs sampling

Q: How to choose the random variable of interest to be sampled conditioned on rest (ordering)
- Ex: No formal ordering / theory behind it.
  - Guided by implementation

(#) Supplementary

Q) Key struggle of MCMC → random walk behavior (the spirit of wiener!)

- Random walk in MCMC
- symmetric, limited bandwidth proposal $Q(x_{new}|x_{old})$
- Indep? - large bandwidth Gaussian proposal with wide variance
  - Low acceptance rate

Ex: Give access to the ground truth target distri $P(\cdot)$ which is not known in practice; do we have a preference for the direction?
- we want to explore modalities of target distri (goes against R.W behavior)
- All standard vanilla MCMC are r.w. behaviours based on proposal distri
- can we steer the behavior towards mode?

(*) Gradients/derivatives is only feature that gives some info on direct.

(*) Hamiltonian Monte Carlo

- Hamiltonian → sum of potential, kinetic energy

Ex: Interesting connections (not going to review physics).

→ see how this yields new sampling method

- introduce more r.v.s into auxiliary distribution
- introduce kinetic energy K

(*) (57): $\frac{dx_i}{dt} = \frac{\partial H}{\partial p_i}$ $\frac{dp_i}{dt} = -\frac{\partial H}{\partial x_i}$ (alt. notation)

- Radford Neal, MacKay

Ex: Point is to define change of state as gradient of loss function

(*) leverage gradient info into transitions.

- position & momentum updates $p(\cdot), q(\cdot)$.

(*) Multiple mechanisms for updates.

- leapfrog method preferred to avoid over/undershooting.

(*) MCMC from Hamiltonian dynamics (57)

- let q be variable of interest

Define: $P(q,p) = \frac{1}{Z} \exp(-U(q)/T) \exp(-K(p)/T)$

where $U(q) = -\log[\pi(q) L(q|D)]$ (Bayesian setup)

$K(p) = \sum_{i=1}^{d} \frac{p_i^2}{2m_i}$ (kinetic term auxiliary r.v.)

(57): use Hamiltonian dynamics to propose next step

- Given q_0 (starting state)
- Draw $p \sim N(0, I)$
- use L steps to propose next state
- accept/reject based on change in Hamiltonian

(57): $\min[1, \exp(-U(q^*) + U(q) - K(p^*) + K(p))]$

(A13) - review accept-reject part based on Hand

(*) broader generalisation:-

combine optimisation + MCMC

gradients \ sampling

(*) 2D Gaussian

- Hamiltonian MCMC vs R.W. Metropolis

(*) Better convergence of to target distri

(*) Better mixing of samples

R.W. - high correlation does not allow us to move away from initial pop.

(*) see empirical results on 20/100D Gaussian comparing Hamil MCMC, RW.

(*) Langevin dynamics

- Analogous to Gibbs being special case of M-H
- Langevin MCMC is special case of Hamiltonian MCMC
- use more sophisticated updates than leapfrog using 2nd order updates
- Eg constraints on regions you don't want to enter
  - can further augment optim. technique to make proposal more efficient.

Ex: to improve acceptance rate, better mixing; incorporate optimisation techniques to improve proposals

---

(*) variational inference

- can be comp with MCMC.
- optimisation based method yielding approx of distri based on evidence.
- Render proposal distri a variational distri; subject it to optimisation
- use M-H for acceptance/rejection of samples.

(*) This is all classical material in PGMs

- representat. in DGM UGM
- Exact inf & learning (with, without latent)
- Approx Inference

(*) Next: use PGMs as a foundation for thinking about modern forms of deep learning