Lecture 19 - scribbles

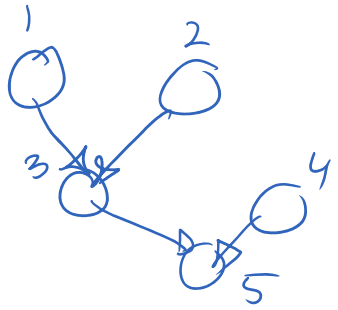
Tuesday, November 13, 2018 14:30

today: more sampling
MCMC

sampling for DGM is easy: ancestral sampling

$(X_1, \dots, X_p) \sim P \in \mathcal{L}(G)$ where G is a DAG

$$p(x_1, \dots, x_p) = \prod_{i=1}^{p} p(x_i \mid x_{\pi_i})$$



suppose WLOG, $1, \dots, p$ be a top. sort for G

ancestral sampling:

for $i = 1, \dots, p$

these are already sampled by top. sort.

sample $X_i \sim p(X_i = \cdot \mid X_{\pi_i})$

end

(can show (by induction) that $(X_1, \dots, X_p)$ has dist. P)

aside: formal test to show that (X_1, X_2) has right joint dist

2 nodes example: $U_1 \sim p(x_1)$

$U_2 | U_1 \sim p(x_2 | X_1 = U_1)$

to show that two R.Vectors are equal in dist. ie. $U \overset{d}{=} V$

$\Leftrightarrow \mathbb{E}_U[f(U)] = \mathbb{E}_V[f(V)]$ for all functions in a big enough class

(eg. cts, & bounded functions)

here, I want to show that:

want to show this

$$\mathbb{E}[f((X_1, X_2))] \overset{?}{=} \mathbb{E}[f(U_1, U_2)]$$

$$\overset{\text{tower property of expect.}}{=} \mathbb{E}_{U_1}\left[\mathbb{E}_{U_2|U_1}[f(U_1,U_2) \mid U_1]\right]$$

$$\overset{\text{by sampling mechanism}}{=} \int_{U_1} p(u_1)\, du_1 \left[\int f(u_1,u_2)\, p(u_2|u_1)\, du_2\right]$$

$$= \int_{U_1}\int_{U_2} f(u_1,u_2) \underbrace{\left[p(u_2|u_1)\, p(u_1)\right]}_{p(u_1,u_2)} du_1 du_2 \quad //$$

(*) important sidenote: when sample from joint, you are also sampling from "marginal" bcs just ignoring the joint aspect...

ie. $(X, Y) \sim p(x, y)$

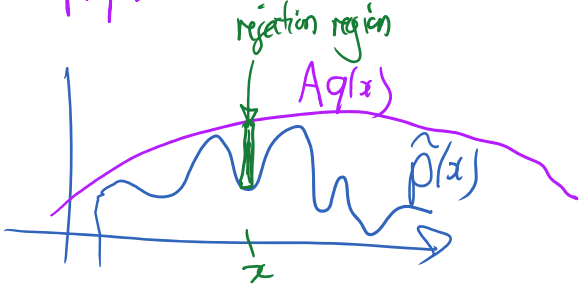
then looking at X itself, you have $X \sim p(x)$

rejection sampling:

say $p(x) = \frac{\tilde{p}(x)}{Z_p}$; let's say we can find a $q(x)$ ("proposal" dist.) which easy to sample from

s.t. $A\,q(x) \geq \tilde{p}(x)\ \forall x$

($q(x)$: proposal; $\tilde{p}(x)$: target (unnorm.))



rule:

- sample $X \sim q(x)$
- Accept with prob $\frac{\tilde{p}(x)}{Aq(x)} \in [0,1]$
- (reject Otw. → start again)

Let's show that accept samples have correct dist.

(say X is discrete)

$$P\{X = x,\ X \text{ is accepted}\} = \underbrace{P\{X \text{ is accepted} \mid X = x\}}_{\propto \tilde{p}(x)} \underbrace{P(X = x)}_{\text{sampling mechanism } q(x)}$$

(P: sampling mechanism)

$Aq(x)$

$$= \frac{\tilde{p}(x)}{A}$$

$$P\{X \text{ is accepted}\} = \sum_x \frac{\tilde{p}(x)}{A} = \frac{Z_p}{A}$$

(marginal prob. of acceptance → want this high)

$$P\{X = x \mid X \text{ is accepted}\} = \frac{\tilde{p}(x)}{A} \Big/ \frac{Z_p}{A} = p(x)$$

//

application to DGM:

say we want to sample from $p(x \mid \bar{x}_E)$

here, could use $\tilde{p}(x) = p(x_E^c, \bar{x}_E)\, \delta(x_E, \bar{x}_E) \propto p(x \mid \bar{x}_E)$ (x_E^c: complement of E)

$$Z_p = p(\bar{x}_E)$$

let $q(x)$ be original joint in DGM (sample from using ancestral sampling)

$$q(x) = p(x_E^c, \bar{x}_E)$$

$$\Rightarrow q(x) \geq \tilde{p}(x) \quad \forall x \qquad [\text{take } A = 1]$$

$$\text{acceptance prob.} = \frac{\tilde{p}(x)}{A\, q(x)} = \delta(x_E, \bar{x}_E)$$

alg.: • do ancestral sampling (rejection sampling

alg. : • do ancestral sampling
• accept if $x_E = \tilde{x}_E$

(rejection sampling for DGM)

$$P\{\text{accept}\} = \frac{Z_p}{A} = p(\tilde{x}_E)$$

15h33

Importance sampling :

in context of computing $\mathbb{E}_p[f(X)] = \mu \qquad X \sim p$

$\rightsquigarrow$ can "weight" sample $x^{(i)}$

$$\mathbb{E}_p[f(X)] = \int f(x)\, p(x)\, dx = \int f(x) \frac{p(x)}{q(x)}\, q(x)\, dx \qquad \text{for some dist. } q \text{ st. } \text{supp}(q) \supseteq \text{supp}(p)$$

$$= \mathbb{E}_q\left[f(Y) \frac{p(Y)}{q(Y)}\right] \quad \text{where } Y \sim q$$

$$\approx \frac{1}{n} \sum_{i=1}^{n} g(Y_i) \quad \text{where } Y_i \overset{\text{iid.}}{\sim} q$$

and $g(y) \triangleq f(y)\, w(y)$

where $w(y) \triangleq \frac{p(y)}{q(y)}$ "weight"

$$\hat{\mu}_{IS} = \frac{1}{n} \sum_{i=1}^{n} f(Y_i)\, w_i \qquad Y_i \sim q$$

"importance weights"

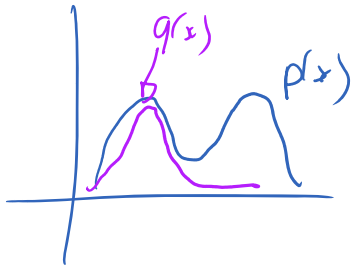
$$w_i = \frac{p(Y_i)}{q(Y_i)}$$

"importance weights" $q(y_i)$

$\mathbb{E}[\hat{\mu}] = \mu$ (to be checked)

$\mathrm{Var}(\hat{\mu}) = \frac{1}{n}\mathrm{Var}_q\left[f(x)\frac{p(x)}{q(x)}\right] = \frac{1}{n}\left[\mathbb{E}_p\left[f(x)^2\frac{p(x)}{q(x)}\right] - \mu^2\right]$

issues here when q is small and p is big



intuitively, you want $q(x) \propto f(x)p(x)$

extension to un-normalized dist'n.

$p(x) = \frac{\tilde{p}(x)}{Z_p}$ $\quad q(x) = \frac{\tilde{q}(x)}{Z_q}$

$\mathbb{E}_q\left[f(x)\frac{p(x)}{q(x)}\right]$

$= \mathbb{E}_q\left[f(x)\frac{\tilde{p}(x)}{\tilde{q}(x)}\right]\frac{Z_q}{Z_p}$

$= \mu \cdot \frac{Z_q}{Z_p}$

estimate $\frac{Z_q}{Z_p}$ with $\hat{Z}_{q/p} \triangleq \frac{1}{n}\sum_{i=1}^{n}\frac{\tilde{p}(y_i)}{\tilde{q}(y_i)} = \frac{1}{n}\sum_{i=1}^{n} w_i$

$$\hat{\mu}_{ISU} = \frac{1}{n}\sum_{i=1}^{n} f(y_i)\, w_i \qquad y_i \sim q$$

$$\hat{\mu}_{ISU} = \frac{\frac{1}{n}\sum_{i=1} f(y_i)\, w_i}{\frac{1}{n}\sum_{j=1} w_j} \qquad y_i \sim q \qquad w_i \triangleq \frac{\tilde{p}(y_i)}{\tilde{q}(y_i)}$$

note: • $\hat{\mu}_{ISU}$ is (slightly) biased, but asymptotically unbiased (as $n \to \infty$)

• this estimator often has lower variance than $\hat{\mu}_{IS}$ even when $Z_p = Z_q = 1$

(normalization "stabilizes" $\hat{\mu}_{IS}$) new weights $\tilde{w}_i = \frac{w_i}{\frac{1}{n}\sum_j w_j} \in [0, n]$

See 2017 notes for
- variance reduction (link with STAT)
- Rao-Blackwellization

MCMC

Markov Chain Monte-Carlo :

Idea: is to relax indep. assumption between samples

to allow adaptive proposal dist.

i.e. we'll run a chain $X_t \mid X_{t-1}$ s.t. $X_t \xrightarrow{t \to \infty}$ target dist. p

"stationary dist. of chain"

then, we can approximate

$$\mathbb{E}_p[f(x)] \text{ as } \frac{1}{T-T_0} \sum_{t=T_0+1}^{T} f(x_t)$$

T_0 is called "burn in" period → depends on "mixing time" of Markov chain

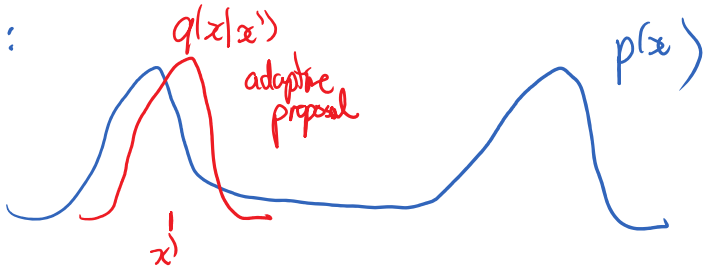
⊛ no need to thin the samples [i.e. use Δt between samples to get more independance]

as this yield higher variance

→ better to use all samples after T_0

(unless it is too expansive)

Motivation:



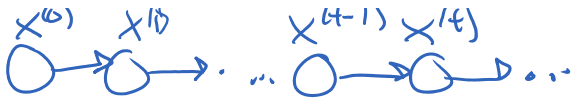
before: samples were $x^{(t)} \sim q$

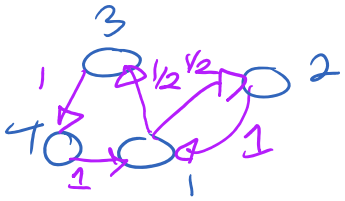
MCMC: $x^{(t)} | x^{(t-1)} \sim q(\cdot | x^{(t-1)})$

↑ Markov transition prob

Review of (finite state space) Markov chains [$|\mathcal{X}| = k$]

- as a DGM, $x^{(0)} \to x^{(1)} \to \cdots \to x^{(t-1)} \to x^{(t)} \to \cdots$

- as a DGM, 
- there is also a transition prob. point of view: use one node per state

(probabilistic FSA) 4 states example 

[homogeneous M.C.]

↳ ie. $P\{X_t = i \mid X_{t-1} = j\} = A_{ij}$ (no time dep.)

A is a $K \times K$ matrix s.t. $\mathbb{1}_K^T A = \mathbb{1}_K$ ← vector of K ones

"left-stochastic matrix"

⊛ (as in HMM), suppose $P\{X_{t-1} = j\} = \pi(j)$

$$P\{X_t = i\} = \sum_j \underbrace{P\{X_t = i \mid X_{t-1} = j\}}_{A_{ij}} \underbrace{P(X_{t-1} = j)}_{\pi_j}$$

$$\pi_{t+1} = A\pi_t$$

$$\Rightarrow \boxed{\pi_t = A^t \pi_0}$$

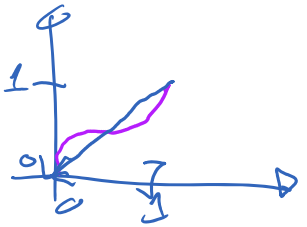
$$\Rightarrow \boxed{\pi_t = A^t \pi_0}$$

Stationary dist π of A is a dist. π s.t. $A\pi = \pi$

[note that π is a right e-vector of A with e-value of 1]

fact: every stochastic matrix has at least 1 stat. dist.

(by Brouwer's fixed pt. thm.)



16h30