Lecture 19 - Sampling

Monday, November 18, 2024 12:33 PM

today : sampling - approximate inference

Approximate inference - sampling

motivation : NP hard to do exact inference in Ising model → need approximation

why sampling? $X = (X_1, \dots, X_p)$

a) simulation $X^{(i)} \sim p$

b) approximate marginal $p(x_i)$

→ special case of expectation's

consider $f : \mathbb{R}^p \to \mathbb{R}$

we want approximate $\mu = \mathbb{E}_p[f(X)]$

special case : if $f(X) \triangleq \mathbb{1}\{X_A = x_A\} \Rightarrow \mathbb{E}_p[f(X)] = p(X_A = x_A)$

Monte Carlo integration / estimation → appears in physics, applied math, ML, statistics

to approximate $\mu = \mathbb{E}_p[f(X)]$

MC estimation: alg : • n samples $X^{(i)} \overset{iid}{\sim} p$

• estimate $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} f(X^{(i)}) = \mathbb{E}_{\hat{p}_n}[f(X)]$

$\hat{p}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X^{(i)} = x\}$

properties :

1) unbiased estimator $\mathbb{E}[\hat{\mu}] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[f(X^{(i)})] = \frac{1}{n} n\mu = \mu$

($\mathbb{E}$: expectation over $(X^{(i)})_{i=1}^n$; $\mathbb{E}[f(X^{(i)})] = \mu$)

this is still true even if $X^{(i)}$'s are dependent

2) expected error (ℓ_2-error) $\mathbb{E}[\|\mu - \hat{\mu}\|_2^2] = \mathbb{E}\left[\left\langle \frac{1}{n}\sum_{i=1}^{n} f(X^{(i)}) - \mu, \frac{1}{n}\sum_{j=1}^{n} f(X^{(j)}) - \mu \right\rangle\right]$

$= \operatorname{tr}(\operatorname{cov}(\hat{\mu}, \hat{\mu}))$

$\to \mathbb{E}\left[\frac{1}{n^2}\sum_{i=1,j=1} \langle f(X^{(i)}) - \mu, f(X^{(j)}) - \mu \rangle\right]$

by independence ⇒ off diagonal terms are zero

i.e. $\mathbb{E}(\langle f(X^{(i)}) - \mu, f(X^{(j)}) - \mu \rangle)$

$(i \neq j)$ terms $= \langle \mathbb{E} f(X^{(i)}) - \mu, \mathbb{E} f(X^{(j)}) - \mu \rangle$

$(i=j)$ terms $\to$ $\frac{1}{n^2}\sum_{i=j}\mathbb{E}[\langle f(x^{(i)})-\mu, f(x^{(i)})-\mu\rangle] = \frac{n\sigma^2}{n^2}$

$\mathbb{E}[\langle f(x^{(i)})-\mu, f(x^{(i)})-\mu\rangle] = \mathbb{E}[\|f(x^{(i)})-\mu\|^2] \triangleq \sigma^2$

$= \mathrm{tr}(\mathrm{cov}(f(x), f(x)))$

$(i\neq j)$ terms $\to$ $\langle \mathbb{E}f(x^{(i)})-\mu, \mathbb{E}f(x^{(j)})-\mu\rangle$, $\mathbb{E}f(x^{(i)}) = \mu$, $\mathbb{E}f(x^{(j)}) = \mu$, Δ_0

$= \frac{\sigma^2}{n}$

$$\mathbb{E}[\|\hat{\mu}-\mu\|_2^2] = \frac{\sigma^2}{n}$$

(apart σ^2 which could depend implicitly)

note: there is no explicit dimension in rate

e.g. $f(x) = x$

$x_0 \sim N(0, \tilde{\sigma}^2)$

$\mathbb{E}[\|f(x)-\mu\|^2] = p\tilde{\sigma}^2$

How to sample?

1) $X \sim \mathrm{Unif}([0,1]) \to$ pseudo-random generator "rand"

2) $X \sim \mathrm{Bernoulli}(p)$ $\quad X = \mathbb{1}\{U \leq p\}$ where $U \sim \mathrm{Unif}([0,1])$

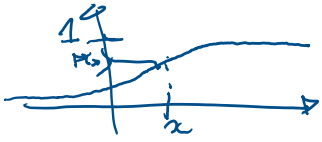
3) inverse transform sampling trick

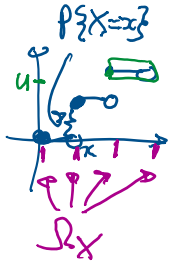
Let F be target c.d.f. of dist p for X $\quad F(x) \triangleq P\{X \leq x\}$

(first, suppose F is invertible)

Let $X = F^{-1}(U)$ with $U \sim \mathrm{Unif}([0,1])$

claim is that X has cdf $F(x)$ (F is monotone)





proof: $P\{X \leq y\} = P\{F^{-1}(U) \leq y\} \overset{F \text{ is invertible}}{=} P\{F(F^{-1}(U)) \leq F(y)\}$

$= P\{U \leq F(y)\} = F(y)$

[if F is not invertible, define $X \triangleq \min\{x : F(x) \geq U\}$ (recall that F is cts from right)

example:

want $X \sim \mathrm{Exp}(\lambda)$ $\quad$ density $p(x) = \lambda \exp(-\lambda x)\,\mathbb{1}_{\mathbb{R}_+}(x)$

$F(x) = 1 - \exp(-\lambda x)$

inverse $F^{-1}(u) = -\frac{1}{\lambda}\log(1-u)$

multivariate distribution?

can generalize above trick using "chain rule"

$X_1 : p$ (dim p)

$X_{1:p}$ (dim p)

from $p(x_{1:p}) = \prod_{i=1}^{p} p(x_i | x_{1:i-1})$ ← use cdf for this conditional

"conditional density sense" for $X_{1:i-1} = x_{1:i-1}$

$$F_{X_i|X_{1:i-1}}(x_i | x_{1:i-1}) \triangleq \mathbb{P}\{X_i \le x_i \mid X_{1:i-1} = x_{1:i-1}\} \triangleq \int_{-\infty}^{x_i} p(x_i' | x_{1:i-1})\, dx_i'$$

could use $U_1, \dots, U_p \overset{iid.}{\sim} \text{Unif}([0,1])$

$X_1 = F_{X_1}^{-1}(U_1)$ ← inverse of this argument

$X_2 = F_{X_2|X_1}^{-1}(U_2 | X_1)$ ← inverse of $F_{X_2|X_1}(\cdot | X_1)$

$\vdots$

$X_p = F_{X_p|X_{1:p-1}}^{-1}(U_p | X_{1:p-1})$ ← is a very complicated function

(curse of dimensionality)

[aside: "copulas" → model for multivariate data with uniform marginals]

<u>exception is multivariate Gaussian</u>

$N(\mu, \Sigma)$ $\quad \Sigma = U \Lambda U^T$ (where $UU^T = I_p$, Λ is diagonal)

(Cholesky decomposition $\Sigma = LL^T$)

generate $V \sim N(0, I_p)$ $\quad (V_p \overset{iid.}{\sim} N(0,1))$

$$X = U \Lambda^{1/2} V + \mu$$

$\mathbb{E}X = \mu$

$\text{cov}(X) = U \Lambda^{1/2} \underbrace{\text{cov}(V)}_{I_p} \Lambda^{1/2} U^T = \Sigma$

Box-Muller transformation to sample (2d) Gaussian

$R^2 \sim \text{Exp}(1)$ $\quad \Theta \sim \text{Unif}([0, 2\pi])$

$\Rightarrow X = R\cos\Theta$, $Y = R\sin\Theta$ $\quad \begin{pmatrix} X \\ Y \end{pmatrix} \sim N(0, I)$

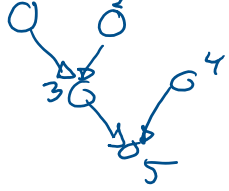
13h34

<u>sampling from a DGM is (relatively) easy c.q. ancestral sampling</u>

$(X_1, \dots, X_d) \sim p \in \mathcal{G}(G)$ where G is a DAG

$$p(x_1, \dots, x_d) = \prod_{i=1}^{d} p(x_i | x_{\pi_i})$$

suppose WLOG $1, \dots, d$ is a top-sort of G



<u>ancestral sampling:</u>

for $1, \dots, d$ sample $x_i \sim p(x_i = \cdot \mid x_{\pi_i})$ ← these are already sampled by top-sort property

[for $1,\dots,d$ sample $x_i \sim p(x_i = \cdot \mid x_{\pi_i})$ end] ← these are already sampled by top-sort property

can show by induction $(x_1, \dots, x_d)$ has dist. p

⊛ important side note: when you sample from a joint you are also sampling from "marginals" by just ignoring joint aspect

i.e. $(x, y) \sim p(x, y)$ then look at x by itself $x \sim p(x)$

rejection sampling:

say $p(x) = \frac{\tilde{p}(x)}{Z_p}$; let's say can form a $q(x)$ "proposal" that is easy to sample from

$Z_{\tilde{p}} \triangleq \int_x \tilde{p}(x)\,dx$

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

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

let's show that accepted samples have correct dist.

(say x discrete)

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

$P\{X = x\}$ ← sampling mechanism; $P\{X \text{ is accepted} \mid X = x\} = \frac{\tilde{p}(x)}{Aq(x)}$; $P\{X = x\} = q(x)$

$P\{X \text{ is accepted}\} = \sum_x \frac{\tilde{p}(x)}{A} = \frac{Z_{\tilde{p}}}{A}$ (marginal prob. of acceptance → want this to be high)

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

application to conditioning in a DGM

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

here could use $\tilde{p}(x) = p_1(x_E, x_{E^c})\,\delta(x_E, \bar{x}_E) \Rightarrow p(x) = p_1(x_{E^c} \mid \bar{x}_E)\,\delta(x_E, \bar{x}_E)$

let $q(x)$ be original joint in DGM (i.e. $p_1(x) = q(x)$)

↳ can sample using ancestral sampling

$q(x) = p_1(x_E, x_{E^c})$

$q(x) \geq \tilde{p}(x) \;\forall x$ [take $A=1$]

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

alg:

- do ancestral sampling
- accept if $X_E = \bar{x}_E$
- o.w. reject

(rejection sampling for DGM conditional)

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

Importance sampling

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

↝ can "weighted" sample $x^{(i)}$ [add q]

$$\mathbb{E}_p[f(x)] = \sum_x f(x)p(x) = \sum_x f(x)\frac{p(x)}{q(x)}q(x) \quad \text{for some dist } q$$

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

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

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

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

$$\hat{\mu}_{I.S.} = \frac{1}{n}\sum_{i=1}^{n} f(y_i)\,w_i \qquad Y_i \overset{iid}{\sim} q$$

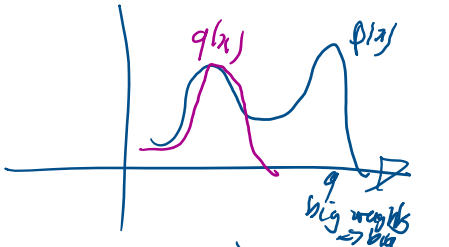
"importance weights"

$w_i \triangleq \frac{p(y_i)}{q(y_i)}$

$\mathbb{E}[\hat{\mu}_{I.S.}] = \mu$

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

issues when q is small and p is big



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

indeed $\mathrm{var}(\hat{\mu})$ can be so something?

extension to un-normalized dist:

extension to un-normalized dist:

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

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

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

$= \frac{1}{n}\sum_{i=1}^{n} w_i$

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

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

• this estimator has often lower variance that $\hat{\mu}_{IS}$ even when $Z_{\tilde{p}} = Z_{\tilde{q}} = 1$

(normalize "stabilizes" estimator 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 SAGA)
- Rao-Blackwellization

Good reference on sampling:
Monte Carlo Statistical Methods, Robert & Casella, 2004