Lecture 19 - Sampling

Tuesday, November 14, 2023 2:40 PM

today: sampling - approximate inference

Approximate inference - sampling

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

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

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

b) approximate marginal $p(x_i)$

$\to$ special case of expectations

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

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

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

Monte Carlo integration/estimation

$\to$ appears in physics, applied math., ML, statistics

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

$$\boxed{\begin{array}{l} \text{MC estimation} \quad \text{alg.:} \cdot\ n \text{ samples } X^{(i)} \overset{\text{iid}}{\sim} p \\ \qquad\qquad\qquad\quad \cdot \text{ estimate } \hat{\mu} = \frac{1}{n}\sum_{i=1}^n f(X^{(i)}) = \mathbb{E}_{\hat{p}_n}[f(X)] \end{array}}$$

$\hookrightarrow$ emp. dist.

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

Properties:

1) unbiased estimator

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

expectation over $(X^{(i)})_{i=1}^n$

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

2) expected error (l_2-error)

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

$$\| \quad \text{tr}(\text{cov}(\hat{\mu}, \hat{\mu})) \qquad = \mathbb{E}\left[\frac{1}{n^2}\sum_{i=1, j=1}^n \langle f(X^{(i)}) - \mu, f(X^{(j)}) - \mu \rangle\right]$$

$$\ldots \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \ldots$$

by independence $\Rightarrow$ off-diagonal terms are zero

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

$$\langle \underbrace{\mathbb{E}f(X^{(i)})}_{\mu} - \mu, \; \mathbb{E}(\quad) \rangle = 0$$

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

$\operatorname{tr}\left(\operatorname{cov}(f(X), f(X))\right)$

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

note: there is no explicit dimension in rate

(apart σ^2 which could depend implicitly)

eg $f(X) = X$

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

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

How to sample?

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

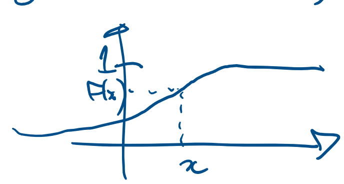
2) $X \sim \mathrm{Bernoulli}(p)$ $\quad X = \mathbb{1}\{U \leq p\}$ where $U \sim \mathrm{unif}([0,1])$

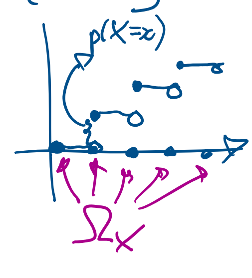
3) inverse transform sampling trick

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)

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





claim that X has cdf $F(x)$

F is invertible (and monotone)

proof: $\mathbb{P}\{X \leq y\} = \mathbb{P}\{F^{-1}(U) \leq y\} \overset{\downarrow}{=} \mathbb{P}\{U \leq F(y)\} = F(y)$ //

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

example:

F is cts from right)

example: want $X \sim \mathrm{Exp}(\lambda)$ 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)

from $p(x_{1:p}) = \prod_{i=1}^{p} p(x_i \mid x_{1:i-1})$

$\hookrightarrow$ use cdf for this conditional

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

$$F_{X_i \mid X_{1:i-1}}(x_i \mid x_{1:i-1}) \triangleq \mathbb{P}\{X_i \leq x_i \mid X_{1:i-1} = x_{1:i-1}\} \triangleq \int_{-\infty}^{x_i} p(x \mid x_{1:i-1})\, dx$$

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

$X_1 = F_{X_1}^{-1}(U_1)$

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

$\vdots$

$X_p = F_{X_p|X_{1:p-1}}^{-1}(U_p \mid X_{1:p-1})$ is a very complicated function (curse of dimensionality)

[aside: "copulas" $\longrightarrow$ model for multivariate data with uniform marginals]

exception is multivariate Gaussian

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

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

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

$$\boxed{X = \underbrace{U \Lambda^{1/2}}_{L} V + \mu}$$

$\mathbb{E}X = \mu$

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

Box-Muller transformation to sample (2d) Gaussian

$R^2 \sim \mathrm{Exp}(1)$ $\quad \Rightarrow \quad X \triangleq R\cos\theta \qquad \begin{pmatrix} X \\ Y \end{pmatrix} \sim N(0, I)$

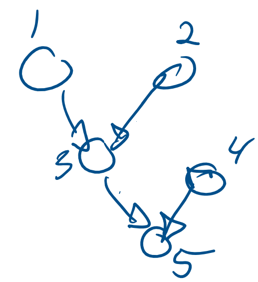
$\theta \sim \mathrm{Unif}([0,2\pi])$ $\qquad\quad Y \triangleq R\sin\theta$

15h32

<u>sampling from a DGM is (relatively) easy: ancestral sampling</u>

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

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



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

<u>ancestral sampling:</u>

for $i = 1, \dots, d$

  sample $X_i \sim p(X_i = \cdot \mid X_{\pi_i})$ these are already sampled by top-sort property

end

can show by induction $(X_1, \dots, X_d)$ has dist. p

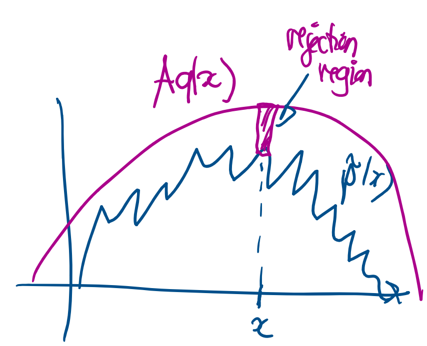
(*) important side note! when you sample from joint you are also sampling from "marginals" by just <u>ignoring</u> joint aspect

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

<u>rejection sampling:</u>

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

$Z_p \triangleq \int_x \tilde{p}(x)\,dx$ $\qquad$ s.t. $A q(x) \geq \tilde{p}(x)\ \forall x$



rule: $\boxed{\begin{array}{l} \cdot \text{ sample } X \sim q(x) \\ \cdot \text{ Accept with prob } \frac{\tilde{p}(x)}{A\,q(x)} \in [0,1] \\ \cdot \text{ reject} \to \text{start again} \end{array}}$

let's show that accepted samples have correct dist.

(say X is discrete)

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

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

(marginal prob of acceptance → want this to be high)

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

<u>application to conditioning in a DGM</u>

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

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

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

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

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

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

<u>alg.</u>: $\boxed{\begin{array}{l} \cdot \text{ do ancestral sampling} \\ \cdot \text{ accept if } x_E = \bar{x}_E \\ \cdot \text{ o.w. reject} \end{array}}$ (rejection sampling for DGM sampling)

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

Importance sampling:

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

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

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

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

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

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

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

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

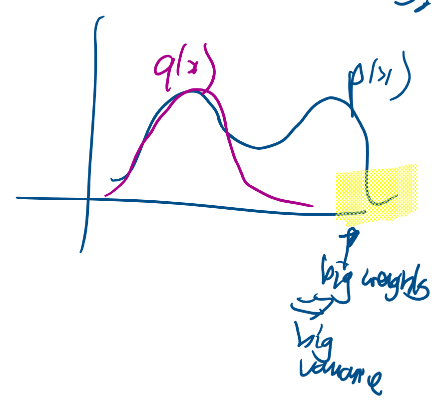
"importance weight"

indeed $\mathrm{var}(\hat{\mu})$ can be ∞ sometimes!

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

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

issues when q is small and p " big



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

extension to un-normalized dist.

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

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

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

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

$$= \frac{1}{n}\sum^{n} w_i$$

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

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

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

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

(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