Lecture 19 - Sampling

Tuesday, November 15, 2022 3:03 PM

today: sampling - approximate inference

Approximate inference - sampling

example: 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 expectations

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

we want to 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

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

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

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

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

this is still true if $X^{(i)}$ were dependent

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

$\mathrm{tr}(\mathrm{cov}(\hat{\mu}, \hat{\mu}))$ $= \mathbb{E}\left[\frac{1}{n^2}\sum_{i=1, j=1}^{n} \langle f(x_i) - \mu, f(x_j) - \mu \rangle\right]$

by independence ⇒ off-diagonal term are zero

i.e. $\langle \underbrace{\mathbb{E}f(x_i)}_{\mu} - \mu, \mathbb{E}f(x_j) - \mu \rangle = 0$

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

$\mathbb{E}[\|f(X^{(i)}) - \mu\|^2] \triangleq \sigma^2$

$n \quad i=j$

$\mathbb{E}[\|f(X^{(i)}) - \mu\|^2] \triangleq \sigma^2$ $\quad n^2$

$= tr(cov(f(X), f(X)))$ $\quad = \frac{\sigma^2}{n}$

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

note: there is no $\underline{\text{explicit dimension}}$ in rate

(apart σ^2 which could depend implicitly)

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

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

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

How to sample?

1) $X \sim Unif([0,1]) \rightarrow$ pseudo-random generator "rand"

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

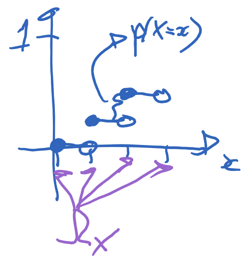
3) <u>inverse transform sampling</u> trick:

let F be target c.d.f. (cum. dist. fct.) of dist. p for X $\qquad F(x) \triangleq P\{X \leq x\}$

(first, suppose F is invertible)

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

claim that X has cdf $F(x)$



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

[if F is <u>not invertible</u>, define $\boxed{X \triangleq \min\{x : F(x) \geq U\}}$] (recall that F is cts. from right)

example: want $X \sim Exp(\lambda)$ $\qquad$ 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 distributions?

can generalize above trick using "chain rule"

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

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

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

use cdf for this conditional to cts.

$$F_{X_i|X_{1:i-1}}(x_i|x_{1:i-1}) \triangleq P\{X_i \leq x_i | X_{1:i-1} = x_{1:i-1}\}$$

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

inverse of this argument

$$\begin{cases} X_1 = F_{X_1}^{-1}(U_1) \\ X_2 = F_{X_2|X_1}^{-1}(U_2|X_1) \\ \vdots \\ X_p = F_{X_p|X_{1:p-1}}^{-1}(U_p|X_{1:p-1}) \end{cases}$$

inverse of $F_{X_2|X_1}(\cdot|X_1)$

is a very complicated fct.

(curse of dimensionality)

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

<u>exception is multivariate Gaussian</u>

$N(\mu, \Sigma)$ $\quad \Sigma = U \Omega U^T$

(where $UU^T = I_p$, Ω is diagonal)

generate $V \sim N(0, I_p)$

($V_p \overset{iid}{\sim} N(0,1)$)

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

(cholesky decomposition)

$\Sigma = LL^T$

$\mathbb{E}X = \mu$

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

Box-Muller transformation to sample (2d) Gaussian

$$\begin{matrix} R^2 \sim \mathrm{Exp}(1) \\ \Theta \sim \mathrm{Unif}([0, 2\pi]) \end{matrix} \Longrightarrow \begin{matrix} X \triangleq R\cos\Theta \\ Y \triangleq R\sin\Theta \end{matrix} \quad \begin{pmatrix} X \\ Y \end{pmatrix} \sim N(0, I)$$

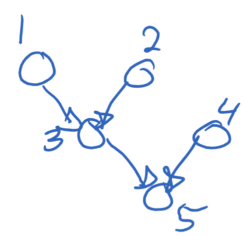
16h14

<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

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

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



<u>ancestral sampling</u>:

for $i = 1, \dots, d$

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

end

those are already observed by top. sort. property end

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

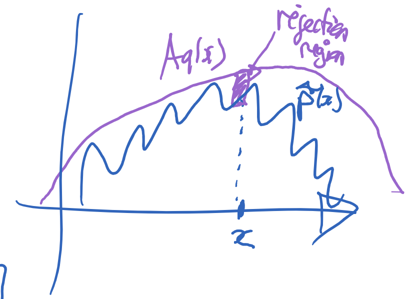
⊛ important sidenote: when you sample from joint you are also sampling from "marginals" by just ignoring joint aspect

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 find a $q(x)$ "proposal" which is easy to sample from

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



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

let's show that accepted samples has correct dist.

(say X is discrete)

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

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

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

application to conditioning in a DGM

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

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

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

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

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

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

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

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

alg.

$$\boxed{\begin{array}{l} \cdot \text{ do ancestral sampling} \\ \cdot \text{ accept if } X_E = \tilde{x}_E \\ \quad \text{o.w. reject} \end{array}}$$

(rejection sampling for DGM sampling)

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

Importance sampling:

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

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

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

s.t. $\operatorname{supp}(q) \supseteq \operatorname{supp}(p)$

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

$$\approx \frac{1}{n}\sum_{i=1}^{n} g(Y^{(i)})$$ 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"

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

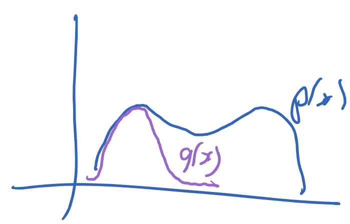
"importance weights"

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

$$\mathbb{E}[\hat{\mu}_{IS}] = \mu$$

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

issues when q is small and p is big



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

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 $\hat{Z}_{p/q} \triangleq \frac{1}{n}\sum_{j=1}^{n} \frac{\tilde{p}(y)}{\tilde{q}(y)}$

$= \frac{1}{n}\sum_{j=1}^{n} w_j$

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

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