# Lecture 21 - sampling

Tuesday, November 17, 2020 14:29

today: sampling - approximate inference

## Approximate inference

**Sampling** example: NP hard to do exact inference in Ising model → need approximate

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

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

b) approximate $p(x_i)$

→ special case of expectations

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

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

e.g. if $f(x) \triangleq \mathbb{1}\{X_A = x_A\}$ $\quad \mathbb{E}_p[f(X)] = p(X_A = x_A)$

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

$\hat{p}$: empirical dist $\quad \hat{p}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{x = X^{(i)}\}$

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

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

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

2) expected error ($\ell_2$-error) $\mathbb{E}[\|\hat{\mu} - \mu\|_2^2] = \mathbb{E}\left[\left\langle \frac{1}{n}\sum_i f(X^{(i)}) - \mu, \frac{1}{n}\sum_j f(X^{(j)}) - \mu \right\rangle\right]$

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

by independence ⇒ off diagonal term are zero

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

$= \frac{1}{n^2} n\sigma^2$ $\quad \mathbb{E}[\|f(X^{(i)}) - \mu\|^2] \triangleq \sigma^2$

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

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

note : that no dimension in rate (apart $\sigma^2$ which could depend implicitly)

## How to sample?

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

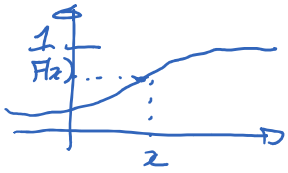
2) $X \sim \mathrm{Bernoulli}(p)$ $\quad X = \mathbb{1}\{U \leq p\}$ where $U \sim \mathrm{Unif}([0,1])$

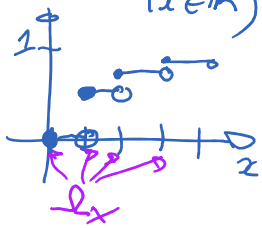
3) inverse transform sampling — cum. dist. fct.

let $F$ be target c.d.f. of distribution $p$ for $X$ $\quad F(x) \triangleq P\{X \leq x\}$ $(x \in \mathbb{R})$

(first, suppose $F$ is invertible)

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




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

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

[ if $F$ is not invertible, define $\boxed{X \triangleq \min\{x : F(x) \geq U\}}$ ] (recall $F$ 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 $p$)  cdf $F(x_{1:p}) \triangleq P\{X_1 \le x_1, \dots, X_p \le x_p\}$

change from class scribbles

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

use cdf for this conditional:

$F_{X_p | X_{1:p-1}}(x_p | x_{1:p-1}) \triangleq P\{X_p \le x_p \mid X_{1:p-1} = x_{1:p-1}\}$

(not $P\{X_p \le x_p \mid X_{1:p-1} \le x_{1:p-1}\}$ as I had written in class)

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

inverse for this argument slice

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

$X_2 = F_{X_2|X_1}^{-1}(U_2 | X_1)$ — 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]

14h31

exception is multivariate Gaussian:

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

(Cholesky decomposition) $L = U \Lambda^{1/2}$, $\Sigma = LL^T$

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

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

$\mathbb{E}X = \mu$

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

Box-Muller transformation to sample (2d) Gaussian

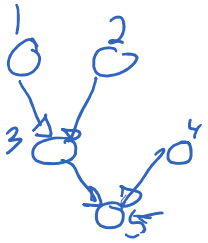
$R^2 \sim \text{Exp}(1)$, $\theta \sim \text{Unif}([0, 2\pi])$ $\Rightarrow$ $X \triangleq R\cos\theta$, $Y \triangleq R\sin\theta$  $\begin{pmatrix} X \\ Y \end{pmatrix} \sim N(0, I)$

sampling from a DGM is (relatively) easy: ancestral sampling:

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

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

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




ancestral sampling

for $i = 1, \dots, p$

sample $X_i \sim p(X_i = \cdot \mid X_{\pi_i})$ — these are already observed by top sort.

end

Can show by induction that $(X_1, \dots, X_p)$ has dist $p$

(*) important remark: when you sample from joint,

you are also sampling from "marginals" by just ignoring the joint aspect

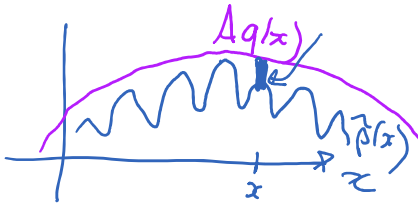
i.e. $(X, Y) \sim p(x, y)$

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

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

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




rule:

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

let's show that accepted sample has correct dist,

(say $X$ is discrete)

$$P\{X = x, X \text{ is accepted}\} = P\{X \text{ is accepted} \mid X = x\} \, P\{X = x\}$$

($P$: sampling mechanism)

$P\{X=x, X \text{ is accepted}\} = P\{X \text{ is accepted} \mid X=x\}\, P\{X=x\}$

(sampling mechanism)

$P\{X \text{ is accepted} \mid X=x\} = \frac{\tilde{p}(x)}{Aq(x)}$ $\qquad P\{X=x\}$ proposal $q(x)$

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

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

**application to conditioning in a DGM:**

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

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

$Z_p = p(\bar{x}_E)$ $\qquad p(x) = p(x_{E^c} \mid \bar{x}_E)\,\delta(x_E, \bar{x}_E)$

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

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

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

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

alg: · do ancestral sampling
· accept if $x_E = \bar{x}_E$ (rejection sampling for DGM conditioning)

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

**Importance sampling:**

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

↝ can "weight" sample $x^{(i)}$

$\mathbb{E}_p[f(x)] = \sum_x f(x)\,p(x) = \sum_x f(x)\frac{p(x)}{q(x)}\,q(x)$ for some dist $q$ s.t. $\text{supp}(q) \supseteq \text{supp}(p)$

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

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

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

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

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

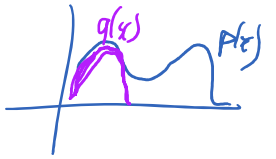
"importance weights"

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

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

issues here when $q$ is small & $p$ is big

indeed $\mathrm{Var}[\cdot]$ can sometimes be $\infty$




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

extension to un-normalized dist.:

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

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

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

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

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

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

(normalization "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