Lecture 19 - scribbles

Tuesday, November 8, 2016
14:28

today: approximate inference - sampling

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

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

b) approximate $p(x_j)$; more generally,

$f: \mathbb{R}^p \to \mathbb{R}^d$, approximate $\mu = \mathbb{E}_p[f(X)]$

e.g. $f(X) = \mathbb{1}\{X_A = x_A\}$ $\mathbb{E}_p[f(X)] = P\{X_A = x_A\}$

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

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

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

- estimate $\hat{\mu} \triangleq \frac{1}{n} \sum_{i=1}^{n} f(X^{(i)})$

properties: 1) unbiased $\mathbb{E}[\hat{\mu}] = \frac{1}{n} \sum_i \mathbb{E} f(X^{(i)}) = \mu$ (this even true if $X^{(i)}$ are dependent)

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

(variance in 1d) $\operatorname{tr}(\operatorname{cov}(\hat{\mu}, \hat{\mu}))$

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

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

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

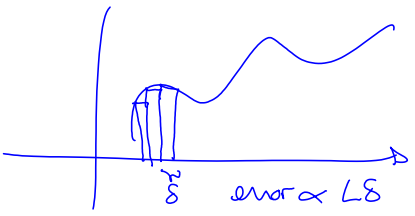
$$\mathbb{E}[\|\hat{\mu} - \mu\|^2] = \frac{\sigma^2}{n}$$

Aside: asymptotic analysis LLN $\hat{\mu} \xrightarrow{a.s.} \mu$

CLT ($\mu \in \mathbb{R}$, $\sigma^2 < \infty$) $\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{d} N(0, \sigma^2)$

aside on numerical computing: 1d is "easy"

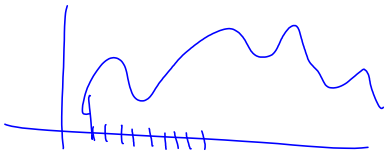
- numerical integration in 1d L-Lipschitz functions



error $\leq \varepsilon \Rightarrow$ complexity $O(\frac{1}{\varepsilon})$

but grid in high d, $O(\frac{1}{\varepsilon^d})$ → curse of dimensionality

- global optimization in 1d



error $\propto L\delta \Rightarrow$ complexity $O(\frac{1}{\varepsilon})$ (# of evaluations)

How to sample?

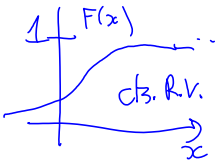
1) $X \sim \text{Unif}([0,1])$ → pseudo-random generator "rand"

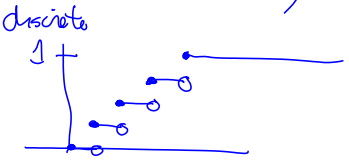
2) $X \sim \text{Bernoulli}(p)$ $\quad X = \mathbb{1}\{U \leq p\}$ where $U \sim \text{Unif}([0,1])$

3) ~~inverse transform sampling~~:

Let F be target cdf of distribution $F(x) \triangleq P\{X \leq x\}$ $(x \in \mathbb{R})$

(say F is invertible)





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

$\Rightarrow$ X has cdf $F(x)$

Proof: $P\{X \leq y\} = P\{F^{-1}(U) \leq y\} = P\{U \leq F(y)\} = F(y)$

(aside if F not invertible, define $X = \min\{x : F(x) \geq U\}$)

example: $X \sim \text{Exp}(\lambda)$ $\quad$ density $p(x) = \lambda e^{-\lambda x} \mathbb{1}_{\mathbb{R}^+}(x)$

cdf $F(x) = 1 - e^{-\lambda x}$

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

Multivariate distribution?

use "chain rule"

$$F_{X_{1:p}}(x_{1:p}) = F_{X_1}(x_1)\, F_{X_2|X_1}(x_2|x_1) \cdots F_{X_p|X_{1:p-1}}(x_p|x_1 \cdots x_{p-1})$$

could use $U_1, \ldots, U_p \overset{\text{iid}}{\sim} \text{Unif}$

could use $u_1, \dots, u_p \overset{iid}{\sim} \text{Unif}$

$x_1 = F_{x_1}^{-1}(u_1)$

$\dots$

$x_p = F^{-1}_{x_p | x_{1:p-1}}(u_p)$ ← is very complicated fct. (curse of dim.)

[aside: "copulas" → model to multivariate dependence with uniform marginals]

exception: multivariate Gaussian

$N(\mu, \Sigma)$ $\quad \Sigma = U \Lambda U^T$ $\quad$ generate $V \sim N(0, I)$

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

$X \sim N(\mu, \Sigma)$

$\Sigma = LL^T$

$L = U \Lambda^{1/2}$

Box-Muller transform 2d Gaussian $x = r\cos\theta$, $y = r\sin\theta$ $\quad \begin{pmatrix} x \\ y \end{pmatrix} \sim N(0, I)$

$\rightarrow r^2$ is $\text{Exp}(1)$

θ is $\text{Unif}([0, 2\pi])$

sampling for DGM is easy: ancestral sampling

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

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

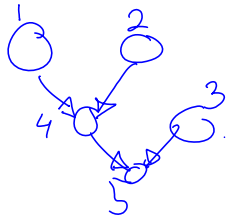
suppose WLOG, $1, \dots, p$ is a top sort of G

ancestral sampling:

for $i = 1, \dots, p$ do

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

end



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

2 node example:

(X, Y) $\quad X \sim p(x)$

$Y|X \sim \tilde{p}(y|x)$

can show that two R.V. are equal in distribution $U \overset{d}{=} V$

$\iff \mathbb{E}_U[f(U)] = \mathbb{E}_V[f(V)]$ for all functions in big enough class (e.g. cts & bounded functions)

here

$$\mathbb{E}[f(\tilde{X},\tilde{Y})] = \mathbb{E}\big[\mathbb{E}[f(\tilde{X},\tilde{Y})|\tilde{X}]\big]$$

$$= \int_{\tilde{x}} \Big[\int_{\tilde{y}} f(\tilde{x},\tilde{y}) p_{Y|X}(\tilde{y}|\tilde{x})\, d\tilde{y}\Big] p_X(\tilde{x})\, d\tilde{x}$$

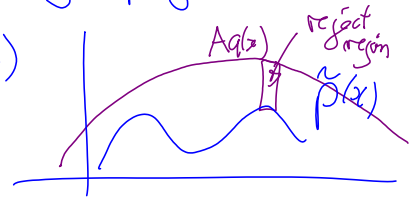
$$= \int_{\tilde{x}}\int_{\tilde{y}} f(\tilde{x},\tilde{y}) \underbrace{\big[p_{Y|X}(\tilde{y}|\tilde{x}) p_X(\tilde{x})\big]}_{p_{X,Y}(\tilde{x},\tilde{y})} d\tilde{y}\, d\tilde{x}$$

thus $(\tilde{X},\tilde{Y})$ has correct joint //

rejection sampling:

say $p(x) = \frac{\tilde{p}(x)}{Z_p}$; say we can find $q(x)$ a distribution we can easily sample from

s.t. $Aq(x) \geq \tilde{p}(x)$



rule: • sample $X \sim q(x)$
• Accept with probability $\frac{\tilde{p}(x)}{Aq(x)} \in [0,1]$
(reject o.w.)

Let's show that accepted draws from rejection sampling have distribution p:
(say X is discrete)

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

$$= \frac{\tilde{p}(x)}{Aq(x)}\, q(x)$$

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

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

Importance sampling:

suppose $X \sim p$, we want to compute $\mathbb{E}_p[f(X)]$

$$\mathbb{E}_p[f(X)] = \int f(x)p(x)dx = \int \frac{f(x)p(x)}{q(x)} q(x)dx \quad \text{for some distribution } q$$

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

$$\approx \frac{1}{n}\sum_{i=1}^{n} g(Y_i) \quad \text{where } Y_i \overset{iid}{\sim} q \text{ and } g(y) \triangleq f(y)\frac{p(y)}{q(y)}$$

$$= \frac{1}{n}\sum_{i=1}^{n} f(X_i) w_i \quad \text{where } w_i \triangleq \frac{p(y_i)}{q(y_i)}$$

"importance weights"

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} f(Y_i) w_i \qquad \mathbb{E}[\hat{\mu}] = \mu = \mathbb{E}_p[f(X)]$$

$$\mathrm{Var}(\hat{\mu}) = \frac{1}{n}\mathrm{Var}_{q(x)}\left[\frac{f(x)p(x)}{q(x)}\right]$$

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

Extension to unnormalized distributions:

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

weights $w(x)$

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

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

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} f(y_i) w_i \quad \text{where } y_i \sim q$$

$\left[\frac{1}{n}\sum_{i=1}^{n} w_i\right] \hat{Z}_{p/q}$

idea behind Markov chain Monte-Carlo (MCMC)

idea is to relax independence assumptions;

i.e. we'll run a chain $X_t \mid X_{t-1}$ s.t. $X_t \xrightarrow[t\to\infty]{} p$ (correct distribution)

"stationary dist." of the chain

then, we can approximate τ

$E_p[f(x)]$ as $\frac{1}{T-T_0} \sum_{t=T_0+1} f(x_t)$

T_0 is "burn-in" period → depends on "mixing time"

⊛ no need to thin [use Δt between samples to get more independence]

as this yields higher variance

→ better to use all samples after T_0 (unless too expensive)