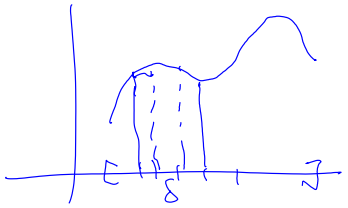
Lecture 18 - scribbles

Friday, November 10, 2017
13:18

today : • sampling
• MCMC

aside on numerical computing : 1d "is easy"

- numerical integration in 1d
for function L-Lipschitz



error $\propto L\delta$

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

use $\delta \approx \varepsilon$

but grid in dimension d $O(\frac{1}{\varepsilon^d})$

- global optimization in 1d

error $\propto L\delta$

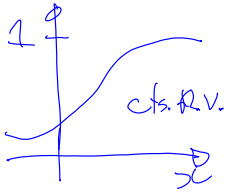
complexity (# of evaluations) $O(\frac{1}{\varepsilon})$

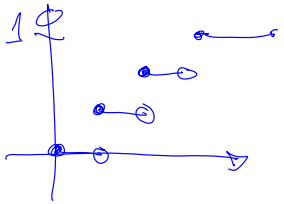
How to sample ?

1) $X \sim \text{Unif}([0,1])$ ⟶ pseudo-random generator "rand"
2) $X \sim \text{Bernoulli}(p)$ $X = \mathbb{1}\{U \leq p\}$ where $U \sim \text{Unif}([0,1])$
3) universe transform sampling :

let F be target cdf of distribution for X $F(x) \triangleq P\{X \le x\}$ $(x \in \mathbb{R})$

(first, suppose F is invertible)





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

claim that X has cdf $F(x)$

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

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

example:

want $X \sim \text{Exp}(1)$ 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 distributions?

generalize above trick using "chain rule"

$X_{1:p}$ $(\dim p)$ cdf $F(x_{1:p}) \triangleq P\{X_1 \le x_1, \dots, X_p \le x_p\}$

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

$F_{X_2|X_1}(x_2|x_1) \triangleq P\{X_2 \le x_2 \mid X_1 \le x_1\}$

recall use $U_{1:p} \overset{\text{iid}}{\sim} \text{Unif}$.

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

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

$\vdots$

$X_p = F^{-1}_{X_p | X_{1:p-1}}(U_p | X_{1:p-1})$ ← is very complicated function (curse of dimensionality?)

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

exception is multivariate Gaussian:

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

(Cholesky decomposition)

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

$\Sigma = L L^T$

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

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

$\mathbb{E}X = \mu$

$\text{Cov}(X) = \Sigma$

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

Box-Muller transform to sample Gaussian (2d):

$x = r \cos\theta$

$y = r \sin\theta$

$\left(\begin{smallmatrix} x \\ y \end{smallmatrix}\right) \sim N(0, I)$

$\rightsquigarrow r^2 \sim \text{Exp}(1)$

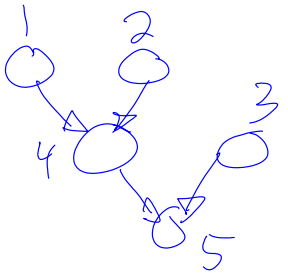
θ is $\text{Unif}([0, 2\pi])$

sampling for DGM is easy: ancestral sampling

$(X_1, \dots, X_p) \sim p \in \mathcal{G}(G)$ 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 $i = 1, \dots, p$ do

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

end

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

aside: formal def to show that (X_1, X_2) has right distribution

2 node example:

$$X_1 \sim p(x_1)$$

$$X_2 \mid X_1 \sim p(x_2 \mid x_1)$$

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

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

here, I want to show that

$$\mathbb{E}[f(X_1, X_2)] = \mathbb{E}_{X_1}\left[\mathbb{E}_{X_2 \mid X_1}[f(X_1, X_2) \mid X_1]\right]$$

$$= \int_{\mathbb{R}} p_{X_1}(x_1)\, dx_1 \left[\int f(x_1, x_2)\, p(x_2 \mid x_1)\, dx_2\right]$$

$$= \int_{x_1} p_{X_1}(x_1)\,dx_1 \left[\int_{x_2} f(x_1, x_2)\, p(x_2|x_1)\,dx_2 \right]$$

$$= \int_{x_1}\int_{x_2} f(x_1, x_2) \underbrace{\left[p_{X_1}(x_1)\, p(x_2|x_1) \right]}_{p(x_1, x_2)} dx_1 dx_2$$

⊛ important sidenote: when sample from joint, you are also sampling from marginal

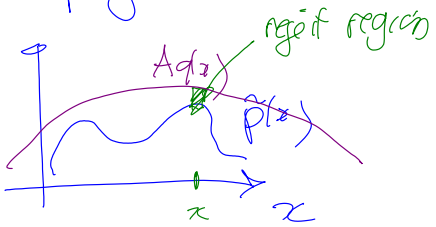
i.e. $(X, Y) \sim p(x, y)$

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

rejection sampling:

say $p(x) = \frac{\tilde{p}(x)}{Z_p}$; let's say we can find $q(x)$, a dist. we can easily sample from

s.t. $A\,q(x) \geq \tilde{p}(x)$ (proposal, target)



rule:

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

let's show accepted samples have correct dist.

(say X is discrete)

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

$$= \frac{\tilde{p}(x)}{A}$$

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

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

application to DGM:

say we want to sample from $p(x \mid \bar{x}_E)$ and $p \in \mathcal{G}(G)$

here $\tilde{p}(x) = p(x_E, \bar{x}_E)\,\delta(x_E, \bar{x}_E)$

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

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

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

alg:
- do ancestral sampling
- accept if $x_E = \bar{x}_E$

$$\text{acceptance prob} = \frac{\tilde{p}(x)}{q(x)} = \begin{cases} 1 & \text{if } x_E = \bar{x}_E \\ 0 & \text{o.w.} \end{cases}$$

[i.e. reject when $x_E \neq \bar{x}_E$]

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

Importance sampling:

in context of computing $E_p[f(X)] = \mu \quad X \sim p$

$\sim$ can "weight" sample $x^{(i)}$

$$E_p[f(X)] = \int f(x)p(x)\,dx = \int f(x)\frac{p(x)}{q(x)}q(x)\,dx \quad \text{for some distribution } q \text{ s.t. } \text{supp}(q) \supseteq \text{supp}(p)$$

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

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

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

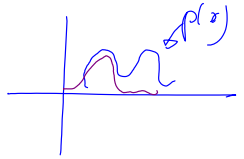
where $w(y) \triangleq \frac{p(y)}{q(y)}$ "weight"

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

"importance weights" $w_i = \frac{p(y_i)}{q(y_i)}$

$$E[\hat{\mu}] = \mu$$

$$\text{Var}(\hat{\mu}) = \frac{1}{n}\text{Var}_{q(x)}\left[f(x)\frac{p(x)}{q(x)}\right] = \frac{1}{n}\left[E_p\left[f(x)^2\frac{p(x)}{q(x)}\right] - \mu^2\right]$$



issues here when q small and p big

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

extension to un-normalized distributions:

extension to un-normalized distributions:

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

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

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

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

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

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

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

(normalization "stabilizes" $\hat{\mu}_{IS}$) new weights $\hat{w}_i = \frac{w_i}{\frac{1}{n}\sum_j w_j} \in [0, \cdot]$

Variance reduction:

control variate trick

want $\hat{\mu}$ for $\mu = \mathbb{E}[X]$; say have Y s.t. $\mathbb{E}Y$ is cheap to compute and Y correlated to X

consider estimator $\hat{\mu}_\alpha \triangleq \alpha(X - Y) + \mathbb{E}Y$ (1-sample version)

Consider estimator $\delta_\alpha \triangleq \alpha(X-Y) + \mathbb{E}Y$ (1-sample version)

$\alpha \in [0,1]$

$\mathbb{E}\delta_\alpha = \alpha \mathbb{E}X + (1-\alpha)\mathbb{E}Y$ (convex combo)

if $\alpha = 1$ then δ_α is unbiased

$\delta_\alpha = X + (\mathbb{E}Y - Y)$ correction

$\mathrm{Var}(\delta_\alpha) = \alpha^2 [\mathrm{Var}X + \mathrm{Var}Y - 2\,\mathrm{cov}(X,Y)]$

smaller α, reduce the variance

if big enough, then $\mathrm{var}(\delta_\alpha) \leq \mathrm{var}(X)$

recall $\mathbb{E}((\delta_\alpha - \mu)^2) = \mathrm{var}(\delta_\alpha) + (\mathbb{E}[\delta_\alpha] - \mu)^2$ bias

$\alpha = 1 \rightsquigarrow$ SAGA

$\alpha = \frac{1}{n} \rightsquigarrow$ SAG

(see lecture 8)

Rao-Blackwellization

motivation: Rao-Blackwell thm: $\mathrm{var}(\mathbb{E}[\delta(X) \mid Z]) \leq \mathrm{var}(\delta(X))$

so if can compute $\mathbb{E}[\delta(X) \mid Z]$; get better estimator

e.g. say $X = (Y, Z)$ $Y|Z$ is simple (e.g. Gaussian)

but Z is complicated

moral of story: if can integrate out analytically some parts, do it!