1

Day 11 in 2021s

SPR Day 20

Sampling Methods

Reading Bishop PRML

11.1.1 Standard distrib.

Skim for broader knowledge — 11.1.2-5 rejection importance

11.2.1 Markov chains

Topics

Monte Carlo estimation

easy way to approx. hard expectations

Graphical Models + Ancestral Sampling

Method: Uniform Sample + Inverse CDF

Transformations of Sampled Variables

Markov Chain Monte Carlo

Why samples?

- Represent distributions that have **no** simple analytical form
- Easily **estimate** expectations

Task: You have model for r.v. z (continuous or discrete) given by $p(z)$

You want to know expected value of function $f(z)$

Ideal: $$\bar{f} = \mathbb{E}_{z \sim p(z)}[f(z)]$$
$$= \int p(z) f(z) dz$$

integral might be hard!

Monte Carlo Estimate: $$\hat{f} = \frac{1}{S} \sum_{s=1}^{S} f(z^s)$$

where $z^s \overset{iid}{\sim} p(z)$

Uses S samples, $z^1, z^2, \ldots z^S$ each iid from $p(z)$

What is <u>expected value</u> of MC estimate?

$$\mathbb{E}[\hat{f}] = \mathbb{E}_{z^1,\ldots z^S \sim p(z)}\left[\frac{1}{S}\sum_{s=1}^{S} f(z^s)\right]$$

$$= \frac{1}{S}\sum_{s=1}^{S} \mathbb{E}_{z^s \sim p(z)}\left[f(z^s)\right]$$ by linearity of expectations and iid assumption

$$= \frac{1}{S}\sum_{s=1}^{S} \int p(z) f(z)\, dz$$

$$= \frac{1}{S}\sum_{s=1}^{S} \bar{f}$$ by definition of $\bar{f}$

$$= \bar{f}$$

MC estimate is <u>UNBIASED</u>.

What is variance of MC estimate?

$$\mathrm{Var}[\hat{f}] = \frac{1}{S}\underbrace{\mathrm{Var}[f(z)]}_{\text{variance of function } f \text{ under original } p(z)}$$

So, for large S, MC will be "close" to ideal $\bar{f}$!

Variance of MC estimate: Derivation

$$\mathrm{Var}[\hat{f}] = \mathbb{E}_{z^1,\dots z^S}\left[(\hat{f}-\bar{f})^2\right]$$

$$= \mathbb{E}_{z^1\dots z^S}\left[\hat{f}^2 - 2\hat{f}\bar{f} + \bar{f}^2\right]$$

$$= \mathbb{E}_{z^1\dots z^S}\left[\hat{f}^2\right] - \bar{f}^2$$

$$= \mathbb{E}\left[\frac{1}{S}\sum_{s=1}^{S} f(z^s)\right]^2 - \bar{f}^2$$

self product

cross product

$$= \frac{1}{S^2}\,\mathbb{E}\left[\left(f(z^1)+\dots f(z^S)\right)\left(f(z^1)+\dots+f(z^S)\right)\right]$$

Simplify cross terms

bc z^1 is indep of z^S

$$\mathbb{E}\left[f(z^1)f(z^S)\right] = \mathbb{E}[f(z^1)]\,\mathbb{E}[f(z^S)] = \bar{f}^2$$

expand + group into self + cross terms

$$= \frac{1}{S^2}\Big[\underbrace{S\,\mathbb{E}[f(z)^2]}_{\substack{\text{self terms}\\ f(z^i)f(z^i)}} + \underbrace{(S^2-S)\,\bar{f}^2}_{\substack{\text{cross terms } i\neq j\\ f(z^i)f(z^j)}}\Big] - \bar{f}^2$$

$$= \frac{1}{S}\,\mathbb{E}[f(z)^2] + \cancel{\frac{S^2-S}{S^2}}\,\bar{f}^2 - \cancel{\frac{S^2}{S^2}}\,\bar{f}^2$$

cancel these

$$= \frac{1}{S}\,\mathbb{E}_{p(z)}[f(z)^2] - \frac{1}{S}\bar{f}^2 = \frac{1}{S}\,\mathrm{Var}_{p(z)}[f(z)]$$

How to Sample from Models with Multiple Variables

Suppose we have model of many r.v., we wish to sample from joint distribution.

$p(z_1, z_2, \cdots z_T)$

e.g. draw from Markov model.

Suppose further we know conditional independence assumptions, and can use these to define a graph (must be directed, acyclic)

where vertices/nodes are random vars
directed edges represent conditional dependence assumptions

$$z_1 \to z_2 \to z_3 \to z_4$$

Markov model with $T=4$

Can use graph to re-write joint as:

$$p(z_{1:4}) = \prod_{i \in V} p(z_i \mid z_{pa(i,E)})$$

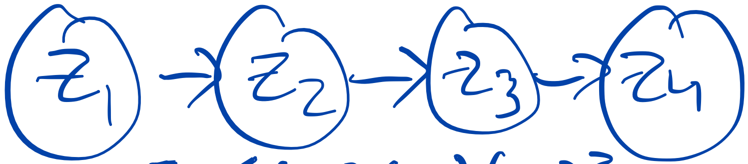
V is vertex list

E is edge list

$pa(i,E)$: "parents" of node i

$\{j : (j,i) \in E\}$

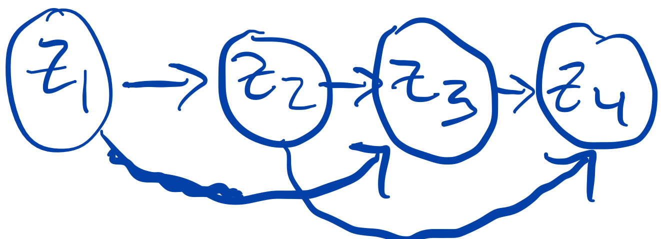
So for our 1^{st} order Markov model



$$p(z_{1:4}) = p(z_1)\,p(z_2|z_1)\,p(z_3|z_2)\,p(z_4|z_3)$$

$E = \{(1,2),(2,3),(3,4)\}$

2^{nd} order Markov



$$p(z_{1:4}) = p(z_1)\,p(z_2|z_1) \cdot p(z_3|z_2,z_1) \cdot p(z_4|z_3,z_2)$$

$E = \{(1,2),(2,3),(3,4),(1,3),(2,4)\}$

We call this representation a <u>directed graphical model</u>

Useful for <u>sampling</u> and many other things
(computing marginals $p(z_t)$
joints $p(z_1, z_2)$
conditionals $p(z_2|z_1)$)

How to sample?

Assume we have way to sample from the simple conditional

$$p(z_i \mid z_{pa(i,E)})$$

for all nodes i

Then, we can sample $z_{1:T} \sim p(z_{1:T})$ using ancestral sampling.

"TOPOLOGICAL SORTED"

Arrange node indices in order $1, 2, \dots T$ s.t. for any index $j \in \{1, 2, \dots T\}$ if i is a parent of j, then $i < j$.

This order is always achievable if graph has no cycles (remember, our directed graph is acyclic)

Ancestral Sampling: ← assume TOPO SORT order!

for i in $1, 2, \dots T$:

$$z_i \sim p(z_i \mid z_{pa(i,E)})$$

return $[z_1, z_2 \dots z_T]$

Guaranteed to sample from joint $p(z_{1:T})$!

What can we do with samples from a joint distribution?

(0) Compute a Monte Carlo (MC) expectation

(1) <u>Sample from a marginal</u>

Given S samples $z_{1:T}^{(1)}, \cdots z_{1:T}^{(S)}$

we can get samples of z_t

by just keeping $z_t^{(1)}, z_t^{(2)}, \cdots z_t^{(S)}$

(2) <u>Sample from a conditional</u>

Given S samples $z_{1:T}^{(1)}, z_{1:T}^{(2)}, \cdots z_{1:T}^{(S)}$

we can get samples from $p(z_t \mid z_u = k)$

by keeping $\{ z_t^{(s)} : z_u^{(s)} = k \}$

and discarding others

Sampling via inverse CDF

Consider real valued random var x
with pdf: $p(x)$
and cumulative distribution function

$$cdf(a) = p(x \leq a) = \int_{-\infty}^{a} p(x)\,dx$$

Properties of cdf: $0 \leq cdf(a) \leq 1$
for all $a \in \mathbb{R}$

If cdf function F is invertible analytically, we can sample using the simple transformation

① $u \sim \mathrm{Unif}([0,1])$ ② $x \leftarrow F^{-1}(u)$

Why? $p(x \leq a) = \int_0^{F(a)} 1\,du = F(a)$

thus, x by construction must have cdf F

Sampling via transformations
"change of variables"

Suppose we have a "target" rand var X
and a "source" rand. var. U

U has known pdf $f : \mathbb{R} \to [0, +\infty]$
that is easy to evaluate

U has known sampling procedure DRAW_U
so we can easily generate
$u^1, u^2, \ldots u^S$

We have an invertible transformation function $T : \mathbb{R} \to \mathbb{R}$

We generate samples x via: (1) $u \sim$ DRAW_U
(2) $x \leftarrow T(u)$

What is pdf of rand. var. X?

Remember CHANGE of VARIABLES from calculus

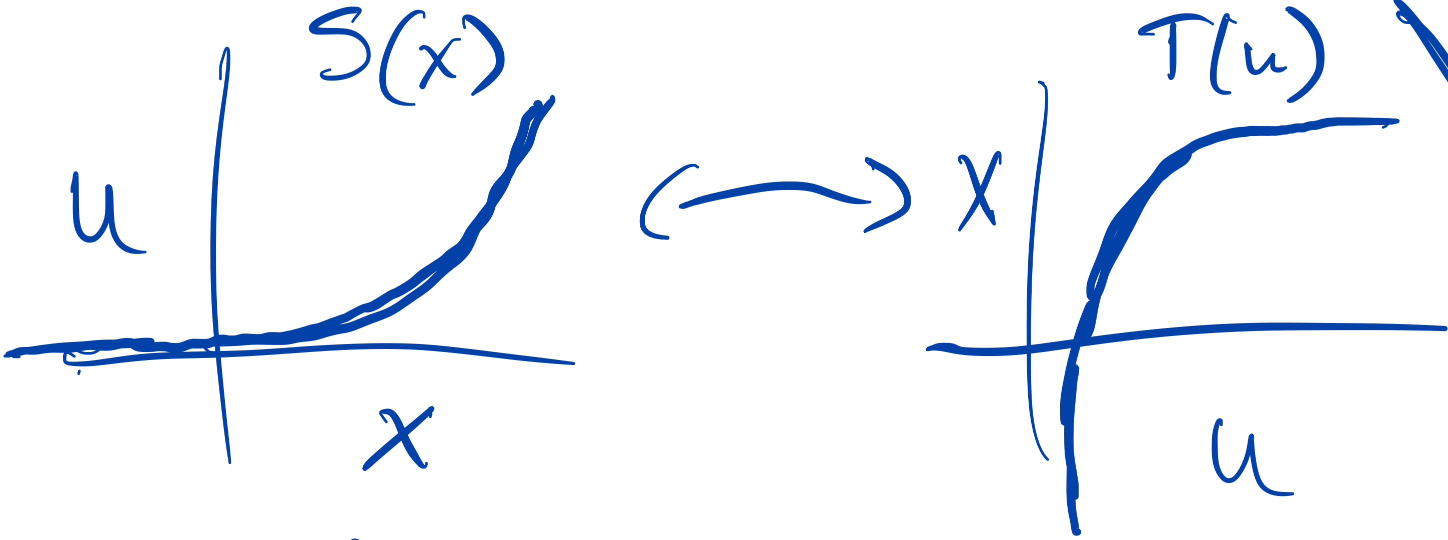
For any smooth function g,

$$\int_a^b g(u)\,du = \int_{T(a)}^{T(b)} g(S(x))\, S'(x)\,dx$$

where $u = S(x)$
$\frac{\partial u}{\partial x} = S'(x) = \frac{\partial}{\partial x} S(x)$

$$\begin{tikzcd} U \arrow[r, bend right, "T"'] & X \arrow[l, bend right, "S"'] \end{tikzcd}$$

S & T are inverses



Invertible functions are <u>Monotonic</u> bc they need to be one-to-one.

So interval (a,b) in u domain is either from $\cancel{\min} T(a), \max T(b)$ if $S'(x) \geq 0$ (pos slope)

or $T(b), T(a)$ or $S'(x) \leq 0$ (neg slope)

So, using change of vars, we have:

$$p(a \leq u \leq b) = \int_a^b f(u)\,du$$

f known pdf for u

$$= \int_{T(a)}^{T(b)} f(S(x))\, S'(x)\, dx$$

Use abs value to fix <u>sign issue</u>

<u>Thus</u>

$$p(\min \leq x \leq \max) = \int_{\min}^{\max} f(S(x))\,|S'(x)|\,dx$$

$$\text{pdf}(x) = f(S(x))\,|S'(x)|$$

uses known simpler pdf f & transform S

Markov Chain Monte Carlo

Goal: Want sample z from a target distribution $p^*(z)$

but we may only know the pdf up to a constant c (does not depend on z)

$$p^*(z) = \underset{\text{unknown}}{c} \; \underset{\text{known function, easy to evaluate}}{\tilde{p}(z)}$$

$$\log p(z) = \log c + \log \tilde{p}(z)$$

Insight: We can <u>sample</u> a sequence

$$z_1 \xrightarrow{T} z_2 \xrightarrow{T} \cdots \xrightarrow{T} z_S$$

by $z_1 \leftarrow$ reasonable guess

$z_t \mid z_{t-1} \sim T$ Markov proposal distribution

if we are careful about choosing proposal distrib T, then marginal $p(z_S) = p^*(z)$

$\underbrace{\text{Markov Chain}}$ $\underbrace{\text{Monte Carlo}}$

uses Markov distrib. T to propose $z_t | z_{t-1}$

purpose is to draw samples from a target distrib. $p^*(z)$

We want $p^*(z)$ to be the stationary distribution of the Markov chain.

Stationary distribution

$p(z)$ is stationary distrib. of Markov operator T if

Discrete case $z \in \{1, 2, \dots K\}$

$$p(z_{t+1}=k) = \sum_{j=1}^{K} \underbrace{p(z_t=j)\, T(z_{t+1}=k \mid z_t=j)}_{\text{joint prob of being in state } j \text{ \& transition to state } k}$$

Continuous case

$$p(z_{t+1}) = \int_{z_t \in \Omega} p(z_t)\, T(z_{t+1} | z_t)\, dz_t$$

When does a unique stationary distribution exist? When the Markov chain is **ergodic**, which means

if we start in state i at $t=0$, then for some time $T_0 > 0$, we have for every state k

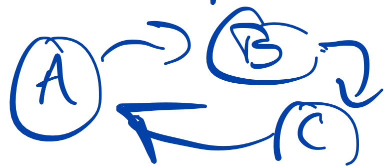
$$p(z_\tau = k \mid z_0 = 1) > 0$$

for all $\tau > T_0$

Intuition: Need to be able to get from any state to any other state

Key Technical Conditions:

- T must be irreducible: path exists between any two states
- T must be aperiodic: no cycles without shortcuts or "long" cuts



need to transition $A \to B$ in diff. number of steps between any states