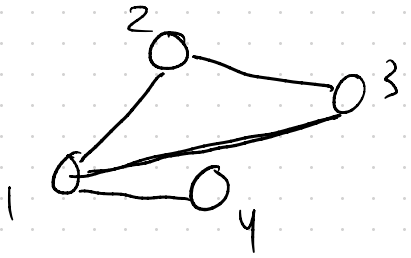
Some Probability Review

Network Analysis $\{V = \text{webpages}, E = \text{links in webpages}\}$



$f : V \longrightarrow \mathbb{R}$

View f as a probability function. We can construct f as a centrality measure.

Most classical example page-Rank:

$$C_R(u) = \sum_{\substack{v \in V \\ \{u,v\} \in E}} \frac{C_R(v)}{\deg(v)}$$

How do we determine C_R? Start with uniform

$f(u) = \frac{1}{|V|}$ for all $u \in V$, then

look at **Markov Process**:

↳ X_0 = first webpage

X_1 = second webpage chosen by clicking link.

~ Probability of clicking link independent of previous path

~ can only move to a webpage if $\exists$ link

~ Probability does not depend on how long we've been clicking.

Example

(3) Metropolis–Hastings Algorithm and C.D.:

Define a matrix $P = (p_{uv}) \in \mathbb{R}^{n \times n}$

n = # vertices on connected graph

with

$$p_{uv} = \begin{cases} \frac{1}{d} \min\left\{1, \frac{\pi(u)}{\pi(v)}\right\} & \{u,v\} \in E, u \neq v \\ 0 & \{u,v\} \notin E, u \neq v \\ 1 - \sum_{i \neq v} p_{iv} & u = v \end{cases}$$

$d = \max_{v \in V} \deg(v)$

$\pi : V \to \mathbb{R}$ probability with $\pi(u) > 0$

$\frac{\pi(u)}{\pi(v)}$ only need ratio → i.e. know more likely link to click.

Thm 2.43 P is transition matrix of

- aperiodic → $\gcd(\{k \in \mathbb{N} \mid (p^k)_{uu} > 0\}) = 1$
- irreducible → $\forall u, v \in V$, $\exists k$ with $(p^k)_{u,v} > 0$
- Markov process

with unique stationary distribution π. $P\pi = \pi$

Why care about stationary distributions?

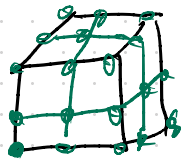
Theorem 7.40 For any probability distribution $f: V \to \mathbb{R}$,

$$\lim_{k \to \infty} P^k f = \pi,$$

For example f = uniform, $\pi = c_R$ page Rank

Toy example Suppose we want to sample points in the n-dimensional cube $[-1,1]^n$, and integrate a density/probability function ρ on this cube C

3-d cube



integrate density ρ = mass.

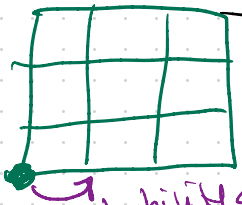
Idea $\int_C \rho = \sum_i \rho(x_i)\,\mathrm{Vol}(\mathrm{cube}_i)$

Split C into smaller cubes C_i

① Problem a uniform sampling x_i doesn't see unit ball → what if density is concentrated inside the unit ball.

② Problem # sample points for uniform sampling grows exponentially

Solution given a probability density f Sample x_i according to $f(i)$. with f_i concentrated on unit cube.



probability of getting here chosen by f.

Long term $P^k f \approx \pi$ = "correct" distribution of data.

Art of Data Science

- Need to choose initial f to match our problem
- we don't know how many k to sample.
  - ↳ Let run for $k=100$ (burn-in period)

  Start sampling at $k=101 \rightarrow$

Why use Markov Chains?

- ↳ Limit curse of dimensionality
- ↳ Don't need to know exact distribution of data
- ↳ only depend on current state, not all previous states [need less data to make future predictions]

For exam

Types of problems include

- Given a sample graph, describe transition matrix.
- Given transition matrix + distribution, determine if aperiodic, strongly, irreducible.

Machine Learning + Probabistic functions

Recall normal distribution, $x \in \mathbb{R}^n$

$$\Phi(x \mid \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} \exp\left(\frac{-1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

When $n=1$, $\Sigma = \sigma^2 =$ variance

$\mu =$ mean

$$\Phi(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-1}{2\sigma^2}(x-\mu)^2\right)$$

$f_\theta : \mathbb{R}^D \to \mathbb{R}$, $\theta = \langle a, b \rangle \in \mathbb{R} \times \mathbb{R}^D$

$f_\theta(x) = a + \langle b, x \rangle$. Linear model

Thm 3.12

$$P_\theta(y \mid x) := \Phi(y \mid f_\theta(x), \sigma^2)$$

normal distribution with variance σ^2 and mean $f_\theta(x)$.

$y = f_\theta(x) + \varepsilon$ for $\varepsilon \sim N(0, \sigma^2)$

Thm 3.13

Nonlinear Compose with $\phi : \mathbb{R}^D \to \mathbb{R}^P$.

Now $\theta \in \mathbb{R}^P$

$$p_\theta(y|x) = \Phi(y | \theta^T \phi(x), \sigma^2)$$

Thm 3.14 We can make σ^2 variance random too!

$$\theta_{ML} = \Omega^+ y, \quad \sigma^2_{ML} = \frac{1}{n} \|y - \Omega\theta_{ML}\|^2$$

The Curse of Dimensionality: $n \ll D$ (less samples ↓, many features ↙)

All theorems require rank $\approx D$ or P (↑ nonlinear.)

Add MAP Maximum A posteriori $\theta \in \mathbb{R}^P$

prior $\theta \sim N(\mu, \Sigma)$ $\mu \in \mathbb{R}^P$, $\Sigma \in \mathbb{R}^{P \times P}$

$\Omega \in \mathbb{R}^{n \times P}$

MAP

Theorem 3.15

$$\theta_{MAP} = \left(\Omega^T\Omega + \sigma^2\Sigma^{-1}\right)^{-1}\left(\Omega^T y + \sigma^2\Sigma^{-1}\mu\right)$$

from idea of pseudo inverse

adds randomness make full rank matrix, i.e. invertible

adds randomness to account for random θ.

Recall Proof uses Bayes' Theorem + Def'n of Φ.

$$P(\theta \mid X, Y) = P(\theta)\frac{P(Y \mid X, \theta)}{P(Y \mid X)}$$

$$P(\theta \mid Y) = P_\theta(\theta)\,P(Y \mid \theta)\,P_Y(Y)$$

$P_\theta(\theta)$: independent of X.

$P(Y \mid \theta)$, $P_Y(Y)$: both depend on X.

Moreal of Story Probabilistic approach is a key to obtaining unique solutions to regression problems

→ as in verifies computer output is correct/ only one possible output.