LECTURE 1
INTRODUCTION TO GM

Types of Graphical Models

Directed models
Bayesian, markov models

Undirected models
markov random fields, factor graphs

If graph is tree, 'message-passing' algos

Graphical models useful when dataset is limited & domain knowledge useful

For high dimensions → Too many parameters
Huge computational time

LECTURE 2:
CONDITIONAL INDEPENCE & DIRECTED MODELS

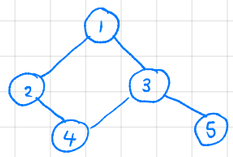
Conditional Inpendence

Definition:

$$X \perp Y \mid Z \iff P(X,Y \mid Z) = P(X \mid Z) \cdot P(Y \mid Z)$$

Directed acyclic graph

Set of nodes, no directed cycles
In a DAG, can ensure parent nodes have lower values (topological ordering)



Chain rule of probability

$$P(x_1, x_2, x_3 \cdots) = P(x_1) \cdot P(x_2 \mid x_1) \cdot P(x_3 \mid x_1, x_2) \cdots$$
$$= \prod P(x_i \mid x_{Pa(i)})$$

Pa → Parent set

Directed model

Assume we have DAG, parents have lower number

$$x_i \perp x_1 \cdots x_{i-1} \mid x_{Pa(i)}$$

Proof:

$$P(x_1 \cdots x_n) = \prod_{i=1}^{N} P(x_i \mid x_1 \ldots x_{i-1})$$
$$= \prod_{i=1}^{N} P(x_i \mid x_{Pa(i)})$$

Upshot: Reduces no. of free parameters

← Factorized distribution of CI

A DAG encodes all assumptions

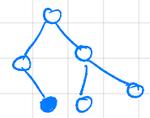
Example: $P(x_1)^1 \; P(x_2 \mid x_1)^2 \; P(x_3 \mid x_1)^2$
Refer Fig (1) $P(x_4 \mid x_2, x_3)^4 \; P(x_5, x_3)^2$

11 free params (original $2^5 - 1$)

Topological ordering allows to drop CI

Assume – A DAG, each var is C.I of its non-descendants given parents

Starting CI's imply others (later)



Conditional Indepedence motivation

The world has true distribution.
& True factorized form maybe

Model may not correspond to True dist.

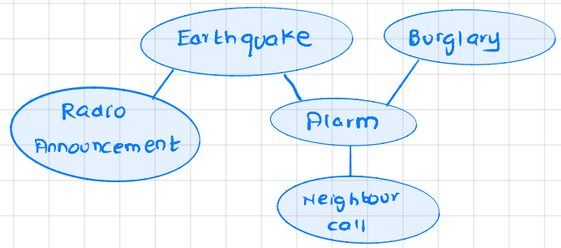
Why assumptions work?

- Domain-Knowledge
- Intuition about causality
  Careful search / Structure learning
- Convenience

Causality

- Causal reasoning implies C.I.
- However, C.I does not imply causality

EG



$R \perp A \mid E$	T	
$E \perp B$	T	
$N \perp E \mid A$	T	
$R \perp B \mid A$	✗	H
$E \perp B \mid A, R$	✗	N
* $E \perp B \mid R, N$	N	

LECTURE 3
BAYES BALL

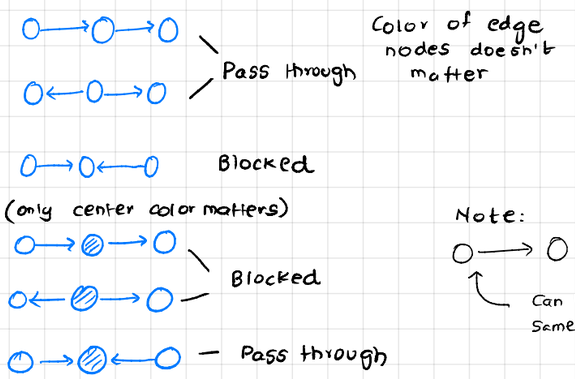
D-separation & Bayes Ball

Given a directed model, is $X_i \perp X_j \mid X_A$? (A: set)

If yes, they are d-separated

① Take graph, color ("observed") nodes in A
Start a ball at X_i

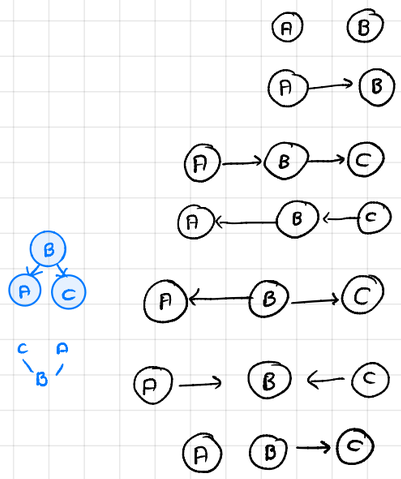
② Bounce ball with these rules:



③ If you hit X_j, not CI

④ If impossible to hit X_j, CI

CI and Graphs



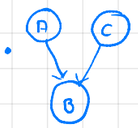
$A \perp B$
None ← Neither dependent or independent

Dist. of 2 vars can be factorized
$p(a,b) = p(a)\,p(b|a)$

$A \perp C \mid B$
$C \perp A \mid B$
$A \perp C \mid B$

$A \perp C$

$A \perp B \quad B \perp C$
$C \perp A \mid B$
$B \perp A$



$A \perp C$ does not imply $A \perp C \mid B$

Example ① Boiler → Noise, Alarm

Noise dependent on Alarm
i.e hearing noise 'influences' estimate of Alarm

② Boiler → Noise, Alarm

Noise ⊥ Alarm
Since we know boiler is on, knowing that noise is present doesn't influence estimate

Examples

$E \perp B \mid R$ yes

$R \perp B \mid A$

SUMMARY

DAGs ⟶ CIs ⟶ Factorized distribution

DAG + Bayes ball ⟶ D separation

To discuss
- Given a dist, how to query it ?
- Estimate dist from data

LECTURE 4

KL Divergence & Maximum Likelihood

Inference vs Learning

Let $\theta \rightarrow$ all params of P_θ

Inference : query $P_\theta(A, B \dots \mid C, D \dots)$

Learning : Given data and P_θ, fixed graph. Find θ

(fixed graph: immutable DAG)

Bayesian doesn't involve "learning"

MLE in Fully observed directed models

(Fully observed: no missing data)

Example

$$P_\theta(S) = \begin{cases} \theta^S_0, & s=0 \\ \theta^S_1, & s=1 \end{cases}$$

$\theta \rightarrow$ vector $\in \mathbb{R}^6$

$$P_\theta(C \mid S) = \begin{cases} \theta^{CIS}_{0,0} \\ \theta^{CIS}_{0,1} \\ \theta^{CIS}_{1,0} \\ \theta^{CIS}_{1,1} \end{cases}$$

$$\theta = [\theta^S_0, \theta^S_1, \theta^{CIS}_{0,0}, \theta^{CIS}_{0,1}, \theta^{CIS}_{1,0}, \theta^{CIS}_{1,1}]$$

Suppose given dataset

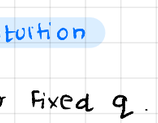
$\theta^S_0 = \frac{1}{2}$ $\theta^S_1 = \frac{1}{2}$

$\theta^{CIS}_{0,0} = \frac{3}{4}$ $\theta^{CIS}_{1,0} = \frac{1}{4}$

$\theta^{CIS}_{0,1} = \frac{1}{2}$ $\theta^{CIS}_{1,1} = \frac{1}{2}$

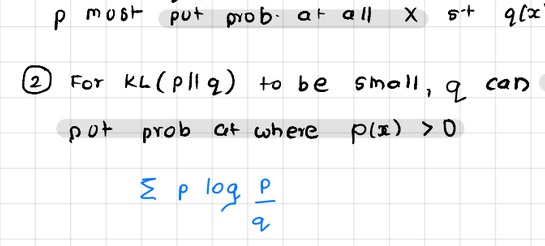
Smoker	Cancer
1	0
0	1
1	1
0	0
0	0
0	0
1	0
1	1

Example ②



Why is counting reasonable?

Parameter selection



Want to find best approximation to true distribution

Kullback - Leibler Divergence

(non-commutative)

$$KL(q \| p) = \sum_x q(x) \log \frac{q(x)}{p(x)}$$

- Lower is better
- Not a distance metric
- Measures the information "lost" when using encoding of one dist to another

Satisfies :

① $KL(q\|p) \geq 0$

② $KL(q\|q) = 0$ iff $p = q$

③ $KL(p\|q) \neq KL(q\|p)$

④ $KL(p\|q) \neq KL(p\|s) + KL(s\|q)$ — Triangle property does not hold

Intuition

For fixed q.

① For $KL(q\|p)$ to be small p must put prob. at all x s.t. $q(x) > 0$

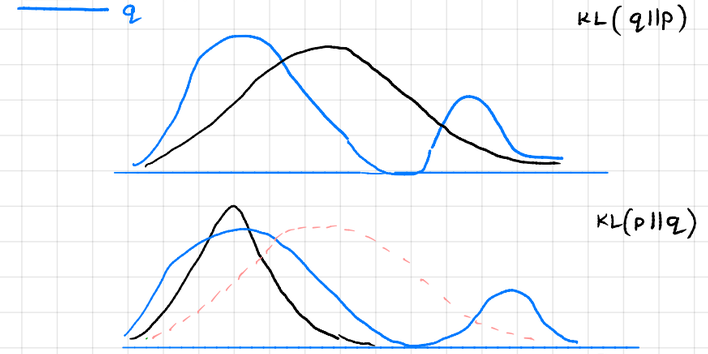
② For $KL(p\|q)$ to be small, q can only put prob at where $p(x) > 0$

$\sum p \log \frac{p}{q}$

Example

best p

q



p doesn't put p over $q(x) = 0$ because of ②

Given data and a directed model, how to minimize divergence from dist to our model ?

Monte Carlo approximation

Eg ⟶ Estimate mean of (height of students)^2

Grab random sample.

If $x_1, x_2 \dots x_n \sim P$ iid

$$\mathop{\mathbb{E}}_P[f(x)] \approx \frac{1}{N} \sum_{i=1}^{N} f(x_i)$$

Maximum Likelihood

Given $x^{(1)} \dots x^{(m)}$, minimize KL divergence wrt P_*

① $$KL(P^* \| P_\theta) = \sum_x P^*(x) \log \frac{P^*(x)}{P_\theta(x)}$$

② $$\approx \frac{1}{m} \sum_{j=1}^{m} \log \frac{P^*(x^{(j)})}{P_\theta(x^{(j)})}$$ (monte carlo)

③ From monte Carlo

$$= \underbrace{\frac{1}{m} \sum_{j=1}^{m} \log P^*(x^{(j)})}_{\text{constant}} - \frac{1}{m} \sum_{j=1}^{m} \log P_\theta(x^{(j)})$$

④ $$\operatorname{argmin}_\theta KL(P^* \| P_\theta) = \operatorname{argmax}_\theta \sum_{j=1}^{m} \log P_\theta(x^{(j)})$$

$$\operatorname{argmax}_\theta \prod_{j=1}^{m} P_\theta(x^{(j)})$$ → Maximum likelihood

Minimizing the monte Carlo approximation of KL-Div of model & true dist is the max likelihood estimate θ

Learning Bernoulli (why counting is enough)

① Data $x^{(1)} \dots x^{(m)}$ IID, $P_\theta(x) = \begin{cases} \theta_0^x & x=0 \\ \theta_1^x & x=1 \end{cases}$

② Log likelihood: $\sum_{j=1}^{m} \log P_\theta(x^{(j)}) = \#_0 \log \theta_0^x + \#_1 \log \theta_1^x$ (count)

Want to maximize, under $\theta_0^x + \theta_1^x = 1 \quad \theta_0^x \geq 0 \quad \theta_1^x \geq 0$

Use Lagrangian multipliers

$$\mathcal{L}(\theta, \lambda) = \#(0) \log \theta_0^x + \#(1) \log \theta_1^x + \lambda(1 - \theta_0^x - \theta_1^x)$$

$$\frac{\delta \mathcal{L}}{\delta \theta_0^x} = \frac{\#(0)}{\theta_0^x} - \lambda \qquad \frac{\delta \mathcal{L}}{\delta \theta_1^x} = \frac{\#(1)}{\theta_1^x} - \lambda$$

LECTURE 5
Extended MLE & Undirected models

$\theta_{ML}^x = \frac{\#(x)}{m}$ → After lagrangian

Learning fully-observed discrete directed model via MLE

① Data $x^{(1)} \dots x^{(m)}$ $x^i \in \mathbb{R}^n$ (discrete)

② $P_\theta(x) = \prod_{i=1}^{n} P_\theta(x_i \mid x_{Pa(i)}) = \prod_{i=1}^{n} \theta^{x \mid x_{Pa(i)}}_{x_i \mid x_{Pa(i)}}$ (R.V: labels; values)

Log Likelihood

$$\sum_{j=1}^{m} \log P_\theta(x^{(j)}) = \sum_{j=1}^{m} \sum_{i=1}^{n} \log P_\theta(x_i^{(j)} \mid x_{Pa(i)}^{(j)})$$

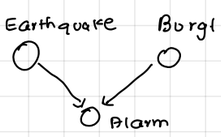
$$= \sum_{j=1}^{m} \sum_{i=1}^{n} \log \theta^{x_i \mid x_{Pa(i)}}_{x_i^{(j)} \mid x_{Pa(i)}^{(j)}}$$

$$\star = \sum_{i=1}^{n} \sum_{x_i} \sum_{x_{Pa(i)}} \#(x_i, x_{Pa(i)}) \log \theta^{x_i \mid x_{Pa(i)}}_{x_i \mid x_{Pa(i)}}$$

($\#(x_i, x_{Pa(i)})$: count of configs)

$x_i, x_{Pa(i)}$: Configurations occur many times in data
↳ Summing over all configs

Eg



A	E	B
0	0	0
1	0	0
0	0	0

$$\sum_{j=1}^{3} \log \theta^{A \mid E,B}_{x_a \mid x_{e,b}}$$

$$\log \theta^{A \mid E,B}_{F \mid F,F} + \log \theta^{A \mid E,B}_{T \mid F,F} + \log \theta^{A \mid E,B}_{F \mid F,F}$$

Constraints: $\forall_i, \forall_{Pa(i)}, \sum_{x_i} \theta^{x_i \mid x_{Pa(i)}}_{x_i \mid x_{Pa(i)}} = 1$

Solving using Lagrangian

$$\mathcal{L} = \sum_i \sum_{x_i} \sum_{x_{Pa(i)}} \#(x_i, x_{Pa(i)}) \log \theta^{x_i \mid x_{Pa(i)}}_{x_i \mid x_{Pa(i)}} + \sum_i \sum_{x_{Pa(i)}} \lambda_i(x_{Pa(i)}) \left(1 - \sum_{x_i} \theta^{x_i \mid x_{Pa(i)}}_{x_i \mid x_{Pa(i)}}\right)$$

$$\frac{\delta \mathcal{L}}{\delta \theta^{x_i \mid x_{Pa(i)}}_{x_i \mid x_{Pa(i)}}} = \frac{\#(x_i, x_{Pa(i)})}{\theta^{x_i \mid x_{Pa(i)}}_{x_i \mid x_{Pa(i)}}} - \lambda_i(x_{Pa(i)}) = 0$$

$$\Rightarrow \theta^{x_i \mid x_{Pa(i)}}_{x_i \mid x_{Pa(i)}} = \frac{\#(x_i, x_{Pa(i)})}{\lambda_i(x_{Pa(i)})}$$

↓ By defn

$$= \frac{\#(x_i, x_{Pa(i)})}{\sum_{x_i} \#(x_i, x_{Pa(i)})} = \frac{\#(x_i, x_{Pa(i)})}{\#(x_{Pa(i)})}$$

Note: $\theta^{x \dots}_{x \dots}$ can be computed by counting

Summary

① Maximizing MLE ≡ minimize Monte Carlo approx. of KL Divergence

② Minimizing KL is reasonable even with modelling error

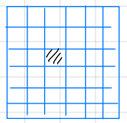
③ To maximize likelihood with fully observed discrete data

$$P_\theta(x_i \mid x_{Pa(i)}) = \theta^{x_i \mid x_{Pa(i)}}_{x_i \mid x_{Pa(i)}} = \frac{\#(x_i, x_{Pa(i)})}{\#(x_{Pa(i)})}$$

Undirected models

Asymmetry in directed models
- $x_i \perp x_{nd(i)} \mid x_{pa(i)}$
- Do not assert $x_i \mid x_{-i} \mid x_{Pa(i)}$ (except i)

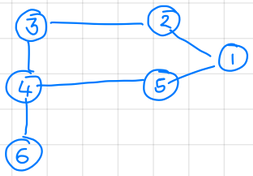
≡ ⊥



x_i → Production of regn i

How to model?

An undirected model asserts $x_A \perp x_B \mid x_C$ if C separates A & B



$x_1 \perp x_3 \mid x_2$ ×

$x_1 \perp x_3 \mid x_2, x_4$ ✓

$x_{6,1} \perp x_{3,5} \mid x_2$ ×

When $P(x) > 0$ an undirected models asserts

$P(x_i \mid x_{-i}) = P(x_i \mid x_{nbr(i)})$

Formula for $P(x)$? ≡ $x_i \perp x_{-i} \mid x_{nb(i)}$

Hammersley - Clifford theorem

A positive dist $P(x) > 0$ satisfies C.Is of an undirected graph iff can be written as

$$P(x_1, \dots x_n) = \frac{1}{Z} \prod_{c \in G} \phi_c(x_c)$$

(set of cliques)

Undirected Graphs assert

① $x_A \perp x_B \mid x_C$ if C separates A & B

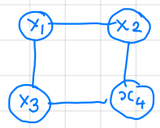
② $P(x_i \mid x_{-i}) = P(x_i \mid x_{nbr(i)})$

HC Theorem

A positive dist. ($P(x) > 0$, domain of x) satisfies CI of UG iff p can be written as ($\phi_c(x_c) > 0$)

$$P(x) = \frac{1}{Z} \prod_{c \in C} \phi_c(x_c)$$

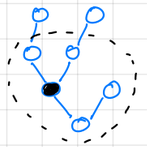
(Z: normalizing factor; C: set of max cliques)



We know $x_1 \perp x_4 \mid x_2, x_3$

$$\therefore P(x) = \frac{1}{Z} \phi_{12}(x_{1,2}) \cdot \phi_{13}(x_{1,3}) \cdot \phi_{24}(x_{2,4}) \cdot \phi_{34}(x_{3,4})$$

Markov blanket - Set of conditions that make x_i C.I from everyone else



MB of node is the only knowledge needed to predict the behaviour of that node & its children

HC Proof

Define $x^* = (0,0\cdots)$ $\quad g(x) = \ln\left(\frac{P(x)}{P(x^*)}\right)$

Step 1 We can uniquely write g as

$$g(x) = \sum_i x_i G_i(x_i) + \sum_{i<j} x_i x_j G_{ij}(x_i, x_j) + \sum_{i<j<k} x_i x_j x_k G_{ijk}(x_i, x_j, x_k) + \cdots + x_1 x_2 \cdots x_n G_{1\cdots n}(x_1, x_2 \cdots x_n)$$

(think tables; 2 for loops; pairs; triples; $x \in \mathbb{R}^n$)

Finding G_i

All other terms disappear

Set $x = (1,0,0\cdots) \longrightarrow G_1(1)$

$x = (2,0,0\cdots) \longrightarrow G_1(2)$

$x = (0,1,0\cdots) \longrightarrow G_2(1)$

Finding $G_{i,j}$

we know G_i, substitute

Set $x = (1,1,0\cdots) \longrightarrow G_{12}(x_1, x_2)$

Upshot → Easy to reason about g, given G_*

Step 2

Define $x^i \rightarrow$ ith component $\leftarrow 0$ $\quad (x_1, x_2 \cdots 0, x_{i+1} \cdots)$

Then $\exp(g(x) - g(x^i)) \equiv \frac{P(x)}{P(x^i)} = \frac{P(x_i \mid x_{-i})}{P(x_i^0 \mid x_{-i})}$ (1) (By conditional Prob)

↓ only depends on nbr(i)

$\frac{P(x_i \mid x_{-i})}{P(x_i^0 \mid x_{-i})} \frac{P(x_{-i})}{P(x_{-i})} = \frac{P(x_i \mid x_{-i})}{P(x_{i=0} \mid x_{-i})} = \frac{P(x_i \mid x_{nb(i)})}{P(x_i^0 \mid x_{nb(i)})}$

Step 3

Pick node 1 without loss of generality (WOLOG)

$g(x) - g(x') = x_1 \left(G_1(x_1) + \sum_{i<j} x_j G_{ij}(x_{ij}) + \sum_{i<j<k} \cdots \right)$

↓ will only have terms involving x_1

Step 4

Suppose j is not a neighbour of 1. (UG assertion)

Then all terms involving x_j must be zero

⇒ Terms containing x_j in decomposition are 0

⇒ The only surviving terms are from cliques. Why?

Because

$g(x) - g(x')$ doesn't depend on x_j

Because $P(x_i \mid x_{-i})$ doesn't depend on j (1) (j is not nbr, & U-G assertion)

Set $x_K = 0 \quad \forall K \notin \{i,j\}$. Set everything except $1, j \rightarrow 0$

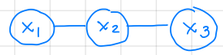
we get, $x_1 (G_1(x_1) + x_j G_{1,j}(x_1, x_j))$

If $G_{ij} \neq 0$ Contradiction { ∵ $g(x) - g(x')$ doesn't depend on x_j

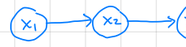
⇒ If j not a neighbour of 1 then $G_{1j} = 0$

For x_1, x_j, x_{1k} we can argue similarly

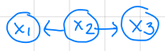
Example



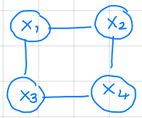
$X_1 \perp X_3 \mid X_2$



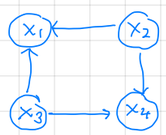
$X_3 \perp X_1 \mid X_2$



$X_1 \perp X_3 \mid X_2$

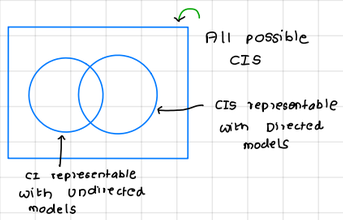


$X_1 \perp X_4 \mid X_{2,3}$
$X_2 \perp X_3 \mid X_{1,4}$



✗ Not possible to get CIs as directed

LECTURE 7
Markov Random Fields & CRFs



Bayesian nets → Represent distributions

Markov Random Field

$$P(x_1, x_2 \cdots x_n) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \phi_c(x_c)$$

$\mathcal{C}$ – set of cliques

ϕ_c – arbitrary positive functions

$$Z = \sum_x \prod_{c \in \mathcal{C}} \phi_c(x_c)$$

(normalization factor)

If we want to emphasize params

$\phi_c(x) = \phi_c(x \mid \theta)$ (can depend on θ in any manner)

Ising model

$\phi_c(x|\theta)$ cliques of size 2

$X_i \in \{-1, 1\}$ binary $\quad P(x) = \frac{1}{Z} \prod_{ij \in pairs} \exp(\theta_{ij} x_i x_j)$

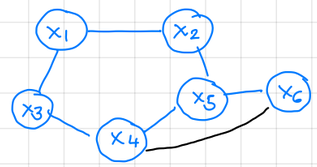
θ_{ij} is how much x_i on x_j "want" to have same value → IF θ_{ij} large & $x_i = x_j = 1$ or $x_i = x_j = -1$ ⇒ High prob.

Ising model is Markov Random Field (MRF)

$$\phi_{\{i,j\}}(x_i, x_j) = \begin{cases} e^{\theta_{ij}} & x_i = x_j \\ e^{-\theta_{ij}} & x_i \neq x_j \end{cases}$$

Conditional Random Field

Consider



$$P(x) = \frac{1}{Z} \phi_{1,2}(x_1, x_2)\, \phi_{1,3}(x_1, x_3)\, \phi_{3,4}(x_3, x_4)\, \phi_{2,5}(x_2, x_5)\, \phi_{4,5}(x_4, x_5)\, \phi_{5,6}(x_5, x_6)$$

what is $P(x_1, x_2, x_3 \mid x_{4,5,6})$ Fix $x_{4,5,6}$ & renormalize

$\phi_{4,5}$ $\phi_{5,6}$ dropped → observed

$$\equiv \frac{1}{Z(x_4, x_5, x_6)} \phi_{12}(x_1, x_2)\, \phi_{13}(x_1, x_3)\, \phi_{3,4}(x_3, x_4)\, \phi_{2,5}(x_2, x_5)$$

$$Z(x_4, x_5, x_6) = \sum_{x_1} \sum_{x_2} \sum_{x_3} \phi_{12}\, \phi_{13}\, \phi_{34}\, \phi_{25}$$

$x_{4,5}$ terms can cel

Now if we add edge 4,6, conditional distribution remains same

⇒ some modelling choice doesn't matter if we only care about conditional distribution ← *

$$P(Y \mid X) = \frac{1}{Z(x)} \prod_{c \in \mathcal{C}} \phi_c(x_c, y_c)$$

$$Z(x) = \sum_y \prod_{c \in \mathcal{C}} \phi_c(x_c, y_c)$$

What you get from MRF & conditioning

No need to define ϕ_c for cliques that only involve x → only y also influences dist

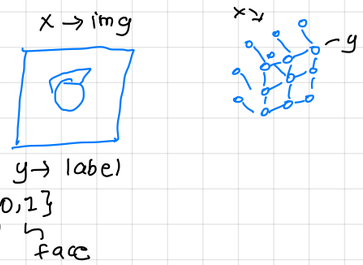
No need to worry about curse of dimensionality (COD) of X, only y?

$$P_\omega(y|x) = \frac{1}{Z(x,\omega)} \prod_{c \in \mathcal{G}} \phi_c(y_c, x; \omega)$$ (not x_c)

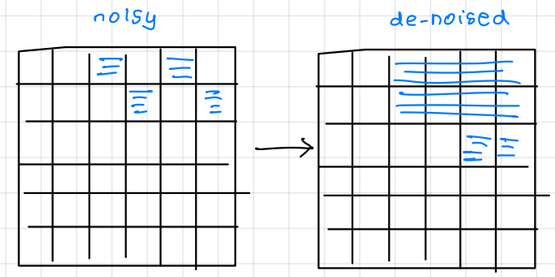
$$Z(x,\omega) = \sum_y \prod_{c \in \mathcal{G}} \phi_c(y_c, x; \omega)$$

ϕ_c [parameterized in application dependent way, can depend on Full input

Eg



Example → Noisy Ising CRF



$x_i \in \{-1, 1\}$ $\quad y_i \in \{-1, 1\}$

$$P_\omega(Y|X) = \frac{1}{Z(x,\omega)} \prod_{(i,j) \text{ pairs}} \exp(\omega_1 y_i y_j) \prod_i \exp(\omega_2 y_i x_i)$$

ω_1: controls how much neighbouring pixels have same value

ω_2: wants output pixel to match input pixel

Why use a CRF?

- we can fit on MRF $P_\theta(x, Y)$ OR
- Fit on CRF $P_\omega(Y|X)$

Suppose $P_*(x)$ is really complex but $P_*(y|x)$ is simple → CRF avoids modelling x ✗

Any reason to use MRF? (more biased, less variance)

Yes → can have superior sample complexity

Summary

1. $X_A \perp X_B | X_C$ if C separates A & B
2. HC Hm → $P(x) = \frac{1}{Z} \prod_{c \in \mathcal{G}} \phi_c(x_c)$ (NO direct probabilistic interpretation)
3. MRFs → $P(x) = \frac{1}{Z} \prod_{c \in \mathcal{G}} \phi_c(x_c)$
4. CRFs → $P(y|x) = \frac{1}{Z(x,\omega)} \prod_c \phi_c(y_c, x; \omega)$

LECTURE 8

Ising model

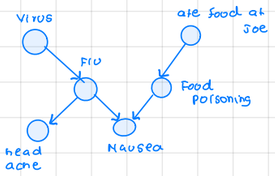
Generative model → $P_\theta(x) = \frac{1}{Z_1} \prod \exp(\theta_{ij} x_i x_j)$

$P_\omega(y|x) = \frac{1}{Z(x,\omega)} \prod_{(i,j) \text{ pairs}} \exp(\omega_1 y_i y_j)$ (noisy ising) $+ \exp(\omega_2 y_i x_i)$

P_ω: clean pixels; $(y|x)$: input img

Inference with model

Why do we need inf.



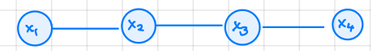
- Patient – nausea, headache

P(ate at Joe | nausea, headache)

$$P(j, n, h) = \sum_v \sum_f \sum_p P(v, f, p, h \cdots)$$

↳ Does not scale

Message passing on a chain



we can use normalizing constant Z! How?

Works with directed models too

$$Z_1 = \sum_{x_1} \sum_{x_2} \sum_{x_3} \sum_{x_4} \phi_1(x_1)\, \phi_2(x_2)\, \phi_3(\,)\, \phi_4 \; \phi_{12}(\,)\, \phi_{23}(\,)\, \phi_{34}(\,)$$

$$\phi_i(x_i) \begin{cases} 1 & x_i = 0 \\ 2 & x_i = 1 \end{cases}$$ → single potential

$$\phi_{ij}(x_i, x_j) = \begin{cases} 1 & x_i \neq x_j \\ 2 & x_i = x_j \end{cases}$$ — pairwise potential

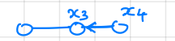
↳ towards symmetry

Process:

Push sums to right

$$Z_1 = \sum_{x_1} \phi_1(x_1) \sum_{x_2} \phi_2(x_2)\, \phi_{12}(x_1, x_2) \sum_{x_3} \phi_3(x_3)\, \phi_{23}(x_2, x_3) \sum_{x_4} \phi_4(x_4)\, \phi_{34}(x_3, x_4)$$

$\sum_{x_4} \phi_4(x_4)\, \phi_{34}(x_3, x_4)$ → $m_{4\to3}(x_3)$



need x_3

Define $m_{4\to3}(x_3)$

$$Z_1 = \sum_{x_1} \phi_1(x_1) \sum_{x_2} \phi_2(x_2)\, \phi_{12}(x_1, x_2) \sum_{x_3} \phi_3(x_3)\, \phi_{23}(x_2, x_3)\, m_{4\to3}(x_3)$$

need to know x_2

Define as $m_{3\to2}(x_2)$

$$Z_1 = \sum_{x_1} \phi_1(x_1) \sum_{x_2} \phi_2(x_2)\, \phi_{12}(x_1, x_2)\, m_{3\to2}(x_2)$$

$\sum_{x_2} \cdots$ → $m_{2\to1}(x_1)$

$$Z_1 = \sum_{x_1} \phi_1(x_1)\, m_{2\to1}(x_1)$$ Sum to 1

$$m_{4\to3}(x_3 = 0) = \phi_4(0)\, \phi_{3,4}(0,0) + \phi_4(1)\, \phi_{3,4}(0,1)$$
$$1 \cdot 2 \qquad\qquad 2 \cdot 1 \qquad = 4$$

$$m_{4\to3}(x_3 = 1) = \phi_4(0)\, \phi_{3,4}(1,0) + \phi_4(1)\, \phi_{3,4}(1,1)$$
$$1 \cdot 1 \qquad\qquad 2 \cdot 2 \qquad = 5$$

$m_{4\to3} = [4 \mid 5]$

m wants $x_3 = 1$

$$m_{3\to2}(x_2 = 0) = \phi_3(0)\, \phi_{2,3}(0,0)\, m_{4\to3}(x_3=0) + \phi_3(1)\, \phi_{2,3}(0,1)\, m_{4\to3}(x_3=1)$$
$$1 \times 2 \times 4 \qquad + \qquad 2 \times 1 \times 5 \qquad = 18$$

$$m_{3\to2}(x_2 = 1) = \phi_3(0)\, \phi_{2,3}(0,1) + \phi_3(1)\, \phi_{2,3}(1,1) \cdot m_{4\to3}(x_3=1)$$
$$\times m_{4\to3}(x_3=0) \qquad\qquad = 24 \; ?$$
$$1 \times 1 \times 4 + 2 \times 2 \times 5 = 24$$

Finally we get $m_{2\to1}$

$$\sum_{x_i} \phi_1(x_1)\, m_{2\to1}(x_1) = \phi_1(x_1=0)\cdot m_{2\to1}(x_1=0) + \phi_1(x_1=1)\, m_{2\to1}(x_1=1)$$

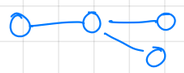
1×84 ; 2×114 ; $= 312$

we can compute forward chain $(L \to R)$

Define

$$\star \to m_{i\to j}(x_j) = \sum_{x_i} \phi(x_i)\,\phi(x_i,x_j) \times \prod_{l \in nb(i) \text{ except } j} m_{l\to i}(x_i)$$

works for acyclic undirected graphs → trees



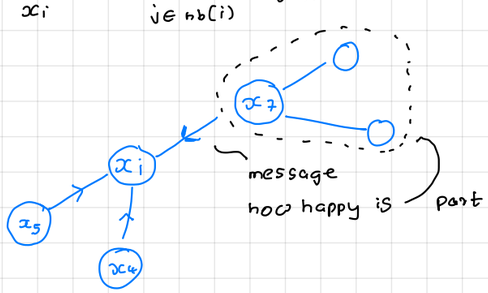
General formula for Z

$$Z = \sum_{x_i} \phi(x_i) \prod_{j \in nb(i)} m_{j\to i}(x_i)$$

$\phi_1\ \phi_{12}\ \phi_2\ \phi_3\ \phi_{23}\ \phi_4\ \phi_{24}$

$m_{3\to2}$; $m_{4\to2}$

$m_{2\to1} = \phi_1\,\phi_{12} \prod_{n \in nb(2)\setminus 1} m_{n\to2}$



we can move sums around

$m_{7\to i}(0)$ = How happy is subgraph starting with x_7 when x_i has value 0

what is $p(x_1)$

$$\frac{\sum_{x_2}\sum_{x_3}\sum_{x_4}\left(\phi_1(x_1)\,\phi_2(x_2)\dots\phi_4(x_4)\,\phi_{12}(x_1,x_2)\dots\right)}{Z}$$

$$= \frac{\phi_1(x_1)\, m_{2\to1}(x_1)}{Z}$$

⇒ From empirical evd, prefers to be 1

Gen formula for univariate marginals

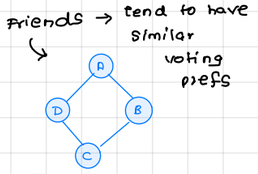
$$P(x_i) = \frac{1}{Z}\,\phi(x_i) \prod_{j \in nb(i)} m_{j\to i}(x_i)$$

NOTES

factor

$$\tilde{p}(A,B,C,D) = \phi(A,B)\,\phi(B,C)\,\phi(C,D)\,\phi(D,A)$$

score (any func)



$$\phi(x,y) \begin{cases} 10 & x=y=1 \\ 5 & x=y=0 \\ 1 & 0.\omega \end{cases}$$

more wt. to consistent votes b/w frnds

pushes x's vote closer to y

$$p(A,B,C,D) = \frac{1}{Z}\tilde{p}(A,B,C,D)$$ — Normalize

probability dist

Indicate a level of coupling b/w variables

MRF

$$p(x_1,\dots x_n) = \frac{1}{Z}\prod_{c\in C}\phi_c(x_c)$$

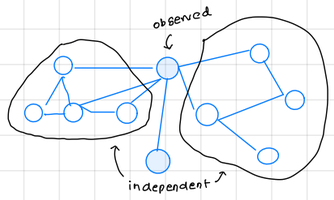
Advantages:

- Applied to many applications → No directionality

Drawbacks

- Z calculation expensive
- Difficult to interpret

Independencies:



Re: Markov blanket

CRF

$$P(y|x) = \frac{1}{Z(x)}\prod_{c\in C}\phi_c(x_c,y_c)$$

partition func [depends on x]

LECTURE 9

EXPONENTIAL FAMILIES

Summary

Take pair-wise MRF $p(x) = \frac{1}{Z}\prod_i \phi_i(x_i) \prod_{(i,j)} \phi_{ij}(x_i,x_j)$

Pass messages

How happy i is for value x_j

$$m_{i\to j}(x_j) = \sum_{x_i} \phi(x_i)\,\phi(x_i,x_j) \prod_{l\in nb(i)\setminus j} m_{l\to i}(x_i)$$

We can find the dist. for x_i & pairs

$$P(x_i) = \frac{1}{Z}\,\phi(x_i) \prod_{l\in nb(i)} m_{l\to i}(x_i)$$

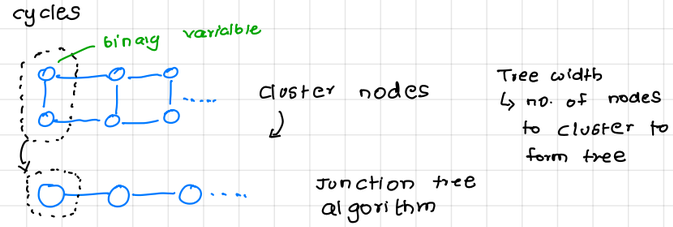
$$P(x_i,x_j) = \frac{1}{Z}\,\phi(x_i)\,\phi(x_j)\,\phi(x_i,x_j) \prod_{l\in nb(i)\setminus j} m_{l\to i}(x_i) \prod_{l\in nb(j)\setminus j} m_{l\to j}(x_j)$$

(x_i, x_j: neighboors)

Discussion:

- Works for any tree → pass messages in right order
- Messages can get very large
  - Use log-space
  - normalize messages so that $\sum_{x_j} m_{i\to j}(x_j) = 1$
- We can use CRFs → no change, fix values
- Can extend to graphs with larger cliques
- Can use same basic alg for directed models

For cycles



You can use message-passing as a heuristic

Loopy Belief Propagation

Iterate graph till graph converges

Exponential Families

- UG are cases of exp fam

$$P_\theta(x) = h(x)\exp\left(\theta^T T(x) - A(\theta)\right)$$

($A(\theta)$: makes dist sum to 1; $T(x)$: feature extractor)

θ – Natural parameters
$T(x)$ – Sufficient statistics
$A(\theta)$ – log partition function
$h(x)$ – Scaling constant (usually 1)
└ does not depend on θ, constant

$$A(\theta) = \log \int h(x)\exp(\theta^T T(x))\,dx$$
$$= \log \sum_x h(x)\exp(\theta^T T(x)) \quad \text{—— } *$$

Why ?

Graphical models are Factorized distributions
└ Product of simple functions

$$\exp(\theta^T T(x)) = \exp(\textstyle\sum \theta_i T_i(x)) = \prod \exp(\theta_i T_i(x))$$

Bernoulli example

$x \in \{0,1\}$ $\quad Ber_\mu(x) = \mu^x(1-\mu)^{1-x}$

$T(x) = \left(\mathbb{I}[x=0] \quad \mathbb{I}[x=1]\right)$

$h(x) = 1$

$\theta = (\theta_1, \theta_2) = (\log\mu, \log(1-\mu))$

$P_\theta(x) = \exp(\theta^T T(x) - A(\theta))$

$A(\theta) = \log \sum_x \exp(\theta^T T(x)) = \log(\theta_0 T(0) + \theta_1 T(1))$

$= \log(e^{\theta_1} + e^{\theta_2})$

$= \exp(\theta_1 T_1(x) + \theta_2 T_2(x) - A(\theta))$

$= \exp(\quad)$

Option B

$T(x) = x \qquad \theta = \log\left(\frac{\mu}{1-\mu}\right)$

$\log \Sigma \exp(\theta^T T(x))$
$= \log(1+e^\theta)$ ← softmax

$P_\theta(x) = \exp(\theta^T T(x) - A(\theta))$

$= \exp\left(\log\frac{\mu}{1-\mu} - A(\theta)\right)$

Ising model

$$P(x) = \frac{1}{Z_1}\prod_{(i,j)} \exp(w x_1 x_2)$$
$$= \frac{1}{Z}\exp(w x_1 x_2)\exp(\cdots)$$

$T(x) = (x_1x_2 + x_2x_3 + \cdots) \qquad \theta = w$

Ising

$$P(x) = \frac{1}{Z}\prod_{(i,j)} \exp(w_{ij} x_i x_j)$$

pairs

$T(x) = [x_i x_j \quad \forall (i,j) \in \text{pairs}]$ → exponential family

$\theta = [w_{ij} \quad \forall (i,j) \in \text{pairs}]$

Properties of exponential families

Critical property #1

Gradients of $A(\theta)$ are expected values

$$\frac{d}{d\theta}A(\theta) = \mathop{\mathbb{E}}_{x\sim P_\theta(x)}[T(x)]$$

$p_\theta(x) = h(x)\exp(\theta^T T(x) - A(\theta))$
$T(x)$: Features of x(sufficient statistics)
$\theta^T T(x)$: Scoring of x
$\exp(\theta^T T(x))$: Makes the scoring/linear combination of x always come out to be positive
$Z(\theta) = \exp(A(\theta))$: Normalizing constant to make this a distribution
$h(x)$: A term we add to make this formulation more general but the above intuition is the real takeaway

Sufficient statistic – A statistic that summarizes all the information in a sample about chosen parameter

Eg → Consider data (1,2,3,4,5). It has sample mean 3.
Imagine we don't know data but know sample mean, finding out data doesn't improve estimate

LECTURE 10
EXPONENTIAL FAMILY CONTD.

Form:

$$P_\theta(x) = h(x)\exp(\theta^T T(x) - A(\theta))$$

(θ: to make +ve; $T(x)$: score; $A(\theta)$: log partition function)

{ Gradient property }

Critical property #2

$$\frac{d^2A(\theta)}{d\theta\, d\theta^T} = \mathop{\mathbb{C}}_{x\sim P_\theta(x)}[T(x)]$$

(Hessian; covariance)

$\frac{d^2A}{d\theta\, d\theta^T}$ is $\text{len}(\theta)\times\text{len}(\theta)$

$s.t.\left(\frac{d^2A}{d\theta\, d\theta^T}\right)_{ij} = \frac{d^2A(\theta)}{d\theta_i\, d\theta_j}$

$$\equiv \frac{d}{d\theta^T}\left(\frac{d}{d\theta}A(\theta)\right) = \frac{d}{d\theta^T}\sum_x P_\theta(x)T(x)$$

$$= \frac{d}{d\theta^T}\sum_x h(x)\exp(\theta^T T(x) - A(\theta))T(x)$$

$$= \sum_x \underbrace{h(x)\exp(\theta^T T(x) - A(\theta))}_{P_\theta(x)} T(x)\,\frac{d}{d\theta^T}\left(\theta^T T(x) - A(\theta)\right)$$

$$= \sum_x P_\theta(x)\,T(x)\left(T(x)^T - \mathbb{E}[T(x)]^T\right)$$

$$= \mathbb{E}[T(x)T(x)^T] - \mathbb{E}[T(x)]\,\mathbb{E}[T(x)]^T$$

Maximum Likelihood

Fix $P_\theta(x) = h(x)\exp(\cdots)$

$$L(\theta) = \frac{1}{n}\sum_{i=1}^n \log P_\theta(x^{(i)})$$

$$= \frac{1}{n}\sum_{i=1}^n \left\{\log\left(h(x^{(i)})\right) + \left(\theta^T T(x^{(i)}) - A(\theta)\right)\right\}$$

(Data lives here)

$$= \theta^T\left(\frac{1}{n}\sum_{i=1}^n T(x^{(i)})\right) - A(\theta) + \frac{1}{n}\sum_{j=1}^n \log\left(h(x^{(i)})\right)$$

(Expected value of data; Independent of θ)

We only need for MLE

Hence $T(x)$ → sufficient statistics

At the minimum $\rightarrow \frac{dL(\theta)}{d\theta} = 0$

$$\frac{dL}{d\theta} = \frac{1}{n}\sum_{i=1}^{n}\left(T(x^{(i)})\right) - \mathbb{E}[T(x)] = 0$$

↑ no analytic soln

Moment matching conditions

$$\Rightarrow \underset{P_\theta(x)}{\mathbb{E}}[T(x)] = \frac{1}{n}\sum_{i=1}^{n}[T(x^{(i)})]$$

MLE finds θ s.t Expectation under data = True expectation

Practical MLE computation

- Compute / Approximate $\underset{P_\theta(x)}{\mathbb{E}}[T(x)]$
- Find $\frac{dL}{d\theta} \rightarrow \theta^T \leftarrow \theta^{T-1} + \alpha \frac{dL}{d\theta}$
- Repeat

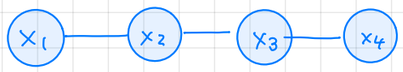
$\frac{d}{d\theta}A(\theta) = [P(x_0=0, x_1=0), \dots]$

So Far

- $P_\theta(x) = h(x)\exp(\theta^T T(x) - A(\theta))$
- Bernoulli, Ising, Gaussian, chi-squared, etc in exp family
- $\frac{d}{d\theta}A(\theta) = \underset{P_\theta(x)}{\mathbb{E}}[T(x)]$ | $\frac{d^2A}{d\theta\, d\theta^T} = \underset{P_\theta(x)}{\mathbb{C}}[T(x)]$
- Likelihood $L(\theta) = \theta^T \bar{T} - A(\theta) + C$

$$\bar{T} = \frac{1}{n}\sum_{i=1}^{n} T(x^{(i)}) \qquad C = \frac{1}{n}\sum_i \log h(x^{(i)})$$

MRF as Exponential Family



$$P(x) = \frac{1}{Z}\phi(x_1, x_2)\ \phi(x_2, x_3)\ \phi(x_3, x_4)$$ binary → 12 free params

Modeling $T(x)$ & θ

$T(x) = [\mathbb{I}(x_1=0, x_2=0), \mathbb{I}(x_1=0, x_2=1), \mathbb{I}(x_1=1, \dots), \dots$ 12 components
$\mathbb{I}(x_2=0, x_3=0) \quad \dots]$ ↖ num 1's = num cliques

$\theta = [\theta(x_1=0, x_2=0), \dots \theta(x_3=1, x_4=1)]$

For a config of X, exactly 3 components in Θ will be 1 corresponding to 3 cliques

$$P_\theta(x) = \exp(\theta^T T(x) - A(\theta))$$

$$= \exp(\theta(x_1,x_2) + \theta(x_2,x_3) + \theta(x_3,x_4) - A(\theta))$$

$$= \frac{\exp(\theta(x_1,x_2))\exp(\theta(x_2,x_3))\cdots}{\exp(A(\theta))} \equiv \frac{1}{Z(\theta)}\ \phi(x_1,x_2)\ \phi(x_2,x_3)\ \phi(x_3,x_4)$$

$$\frac{d}{d\theta}A(\theta) = \mathbb{E}[T(x)] = [P_\theta(x_1=0, x_2=0), P_\theta(\dots)\dots]$$

$\equiv$ marginals

LECTURE 11
MLE in MRF

General MRFs as exponential families

$$P(x) = \frac{1}{Z}\prod_{c\in C}\phi_c(x_c)$$

Note: x_c clique can have many variables $x_c = \{x_1, x_2\}$

Exp family → $T(x)$ has indicator funcs for every config. of every clique

$$T(x) = [\mathbb{I}[x_c = x_{c'}] \quad \forall c \in C,\ \forall x_{c'}] \quad \text{—— } *$$

$$\theta = [\theta(x_{c'}) \quad \forall c \in C, \forall x_{c'}]$$

$\phi_c(x_c)$ has no. of parameters = no. of configuration of x_c

ϕ_c { θ_1 $x_c = \{0,1,3$; θ_2 $x_c = \{0,2\}$; ⋮

(L) $\underset{P_\theta(x)}{\mathbb{E}}[T(x)] = \mathbb{E}[\mathbb{I}(x_c = x_{c'}) \ \forall c \in C, \forall x_c]$

$= [P(x_c = x_c') \ \forall c \in C, \forall x_{c'}]$

↓ model marginals

(R) $\frac{1}{n}\sum_{i=1}^{n} T(x^{(i)}) = \left[\frac{\#(x_{c'})}{n} \ \forall c \in C, \forall x_{c'}\right]$

↓ data marginals

Model marginal match data marginal at MLE

Conditional Exponential family

$$P_\theta(X, Y) = h(x,y)\exp(\theta^T T(x,y) - A(\theta))$$

what is $P_\theta(y|x)$

$$P_\theta(y|x) = \frac{P_\theta(Y,x)}{P_\theta(x)} = \frac{h(x,y)\exp(\theta^T T(x,y) - A(\theta))}{\sum_{y'} h(x,y')\exp(\theta^T T(x,y') - A(\theta))}$$

↳ $A(\theta)$ cancels

$$= h(x,y)\exp(\theta^T T(x,y) - \underbrace{\log\sum_{y'} h(x,y')\exp(\theta^T T(x,y))}_{A'(\theta)})$$

Different exponential family form!

Definition

Conditional Exp. family

$$P_\theta(y|x) = h(x,y)\exp(\theta^T T(x,y) - A(x,\theta))$$ ↳ Depends on X

$$A(x,\theta) = \log\sum_y h(x,y)\exp(\theta^T T(x,y))$$

Note: IF $X = \{\}$, becomes traditional exp family

$$\frac{dA(x;\theta)}{d\theta} = \sum_y P_\theta(y|x)T(x,y) = \underset{P(y|x)}{\mathbb{E}}[T(x,y)]$$

Learning in Conditional Exp family

Given $(x^{(i)}, y^{(i)})$

Likelihood is $L(\theta) = \frac{1}{n}\sum_{i=1}^{n}\log P(y^{(i)}|x^{(i)})$

$$L(\theta) = \frac{1}{n}\left[\sum_{i=1}^{n}\left(\theta^T T(x^{(i)}, y^{(i)}) - A(x^{(i)}, \theta)\right) + \log h(x^{(i)}, y^{(i)})\right]$$

$$\frac{dL(\theta)}{d\theta} = \frac{1}{n}\left(\sum_{i=1}^{n} T(x^{(i)}, y^{(i)}) - \frac{d}{d\theta}A(x^{(i)}, \theta)\right)$$

At optimum $\frac{dL}{d\theta} = 0$

$$\Rightarrow \frac{1}{n}\sum_{i=1}^{n} T(x^{(i)}, y^{(i)}) = \frac{1}{n}\sum_{i=1}^{n}\frac{d}{d\theta}A(x^{(i)}, \theta)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\underset{P(y|x^{(i)})}{\mathbb{E}}[T(x^{(i)}, y)]$$

$$\underset{x,y}{\hat{\mathbb{E}}}[T(x,y)] = \underset{x}{\hat{\mathbb{E}}}\,\underset{P_\theta(y|x)}{\mathbb{E}}[T(x,y)]$$ → No need to model X $P_\theta(x)$

↳ we can use data expectation of x

Latent Variables

$$P_\theta(x,u) = h(x,u)\ \exp\left(\theta^T T(x,u) - A(\theta)\right)$$

u: unobserved

marginalize u

$$\log P_\theta(x) = \log \sum_u P_\theta(x,u) = \log\left(\sum_u h(x,u)\exp\left(\theta^T T(x,u) - A(\theta)\right)\right)$$

$$= \underbrace{\log \sum_u h(x,u)\exp(\theta^T T(x,u))}_{A(x;\theta)} - A(\theta)$$

$$\log P_\theta(x) = A(x;\theta) - A(\theta)$$

$$P_\theta(x) = \exp\left(A(x;\theta) - A(\theta)\right)$$

$$\frac{dA}{d\theta}(x,\theta) = \mathop{\mathbb{E}}_{P_\theta(u|x)}\left[T(u,x)\right]$$

$$\frac{dA(\theta)}{d\theta} = \mathop{\mathbb{E}}_{P_\theta(x,u)}\left[T(x,u)\right]$$

Given $x^{(1)} \dots x^{(n)}$. The marginal log likelihood is

$$L(\theta) = \frac{1}{n}\sum_{i=1}^n \log P_\theta(x^{(i)}) = \frac{1}{n}\sum_{i=1}^n A(x^{(i)};\theta) - A(\theta)$$

If θ is MLE, $\frac{dL}{d\theta} = 0$

$$\frac{1}{n}\sum_{i=1}^n \mathop{\mathbb{E}}_{P_\theta(u|x^{(i)})} T(x^{(i)},u) = \mathop{\mathbb{E}}_{P_\theta(x,u)} T(x,u)$$

$$\mathop{\hat{\mathbb{E}}}_{x} \mathop{\mathbb{E}}_{P_\theta(u|x)} T(x,u) = \mathop{\mathbb{E}}_{P_\theta(x,u)} T(x,u)$$

Using the model to fill in data

Different than prev

LECTURE 12

LEARNING THEORY

$$\mathop{\hat{\mathbb{E}}}_{x}[f(x)] = \frac{1}{n}\sum_{i=1}^n f(x^{(i)})$$

$$\mathop{\mathbb{E}}_{P_\theta(x)}[f(x)] = \sum_x P_\theta(x) f(x)$$

Why not instead of $\max_\theta \frac{1}{n}\sum_{i=1}^n \log P_\theta(x^{(i)})$

Use $\min_\theta \left|\left| \mathop{\hat{\mathbb{E}}}_{x}[T(x)] - \mathop{\mathbb{E}}_{P_\theta(x)}[T(x)] \right|\right|_2^2$

Fill as much model on left & as much data on right?

Max Likelihood vs Conditional Max likelihood

Data $\{x^{(1)},y^{(1)} \dots x^{(n)},y^{(n)}\}$

Create model $P_\theta(x,y)$

Only care about about $P_\theta(y|x)$ accuracy Choosing is subtle!

Could maximize

minimizing the monte carlo approximation of KL divergence

① $\frac{1}{n}\sum_{i=1}^n \log P_\theta(x^{(i)},y^{(i)})$ — full likelihood

same with cond. divergence

② $\frac{1}{n}\sum_{i=1}^n \log P_\theta(y^{(i)}|x^{(i)})$ — cond. likelihood

We know:

cond. likelihood

$$\frac{1}{n}\sum_{i=1}^n \log P_\theta(x^{(i)},y^{(i)}) = \frac{1}{n}\sum_{i=1}^n \log P_\theta(y^{(i)}|x^{(i)}) + \frac{1}{n}\sum_{i=1}^n \log P_\theta(x^{(i)})$$

full

input likelihood

Which should we use?

If $P_\theta(y|x)$ and $P_\theta(x)$ use separate parts of θ, doesn't matter

If we correctly 'model', can be useful.
- higher bias
- lower variance

If model is exact (no error, for some θ), both find true $P^*(y|x)$.

But full likelihood has less variance.

If model has error — full likelihood with infinite data finds

$$\arg\min_\theta \sum_{x,y} P^*(x,y)\log\frac{P^*(x,y)}{P_\theta(x,y)} = \arg\min_\theta KL(P^*,P_\theta)$$

Conditional likelihood finds

$$\arg\min_\theta \sum_{x,y} P^*(x,y)\log\frac{P^*(y|x)}{P_\theta(y|x)} = \arg\min_\theta KL(P^*(y|x)\,||\,P_\theta(y|x))$$

conditional divergence

Gradient ascent

$$\nabla L(\theta) = \mathop{\hat{\mathbb{E}}}_{x,y}[T(x,y)] - \mathop{\mathbb{E}}_{P_\theta(x,y)}[T(x,y)]$$

Joint

Inference over x & y together

For conditional likelihood

Inference over y only with x fixed

$$\nabla L(\theta) = \mathop{\hat{\mathbb{E}}}_{x,y} T(x,y) - \mathop{\hat{\mathbb{E}}}_{x}\mathop{\mathbb{E}}_{P_\theta(Y|x)}[T(x,Y)]$$

Repeated n times — ★

Summary

Next

- True dist P_{θ_0}
- Data $x^{(1)},\dots x^{(n)} \sim P_{\theta_0}$ Don't know θ_0
- maximize likelihood to find $\hat{\theta}_n$
- $\hat{\theta}_n$ will be different from θ_0

How different?

LECTURE 13

FISCHER INFORMATION

We have true dist $P_\theta(x)$. Data $x^{(1)}\dots x^{(n)}$ from $P_\theta(x)$ don't know θ

MLE $\rightarrow \hat{\theta}_n$

How different are θ_0 & $\hat{\theta}_n$

Central limit theorem

If $x_1, x_2 \dots$ are IID variables with mean μ and covariance Σ then $\bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$ $\quad \bar{x}_n \approx \mathcal{N}(\mu, \frac{1}{n}\Sigma)$

$$\sqrt{n}(\bar{x}_n - \mu) \xrightarrow{D} \mathcal{N}(0,\Sigma)$$

D: distribution

"rate" of convergence

converges to

converges in distribution

Fisher Information

$x \sim P_{\theta_0}(x)$

Claim: $\mathbb{E}_{P_{\theta_0}(x)}\left[\nabla_\theta \log P_{\theta_0}(x)\right] = 0$ (1)

Proof:

$\sum_x P_\theta(x) = 1$

$\nabla_\theta \sum_x P_\theta(x) = 0$ ↓ log derivative trick

$\sum_x P_\theta(x) \nabla_\theta \log P_{\theta_0}(x) = 0$

$\mathbb{E}\left[\nabla_\theta \log P_{\theta_0}(x)\right] = 0$

$$I(\theta) = -\mathbb{E}_{P_\theta(x)}\left[\nabla^2 \log P_\theta(x)\right]$$

(2)

Claim: $\mathbb{V}_{P_{\theta_0}(x)}\left[\nabla_\theta \log P_{\theta_0}(x)\right] = -\mathbb{E}_{P_{\theta_0}(x)}\left[\nabla^2 \log P_{\theta_0}(x)\right]$ (the right side is $I(\theta)$, the Fisher Information)

Proof: $\nabla^2 \sum_x P_\theta(x) = 0$ ← no trick

$= \nabla^T \nabla \sum_x P_\theta(x)$

$= \nabla^T \sum_x P_{\theta_0}(x) \nabla \log P_\theta(x)$

$= \sum_x \nabla \log P_\theta(x) (\nabla P_\theta(x))^T + \sum_x P_\theta(x) \nabla^2 \log P_\theta(x)$

$= \sum_x P_\theta(x) \nabla \log P_\theta(x) (\nabla \log P_\theta(x))^T$; $E[\nabla^2 \log P_\theta(x)]$

$E[\log P_\theta(x)^2] \equiv \mathrm{Var}(\dots)$ ∵ $E[x] = 0$

Exponential Family

$P_\theta(x) = h(x) \exp(\theta^T T(x) - A(\theta))$

$\Rightarrow \nabla^2 \log P_\theta(x) = -\nabla^2 A(\theta) \Rightarrow I(\theta) = \nabla^2 A(\theta)$

Asymptotic dist of likelihood grad at θ_0

Get θ_0, Find maximum likelihood (of θ), Find gradient

if grad ≈ 0, we are at mle (of data)

$\nabla L_1(\theta_0) = \nabla \log P_\theta(X)$ is R.V

$\nabla L_n(\theta_0) = \frac{1}{n} \sum_{i=1}^n \nabla \log P_\theta(X^{(i)})$ is a R.V

What is dist of $\nabla L_n(\theta_0)$

$\sqrt{n}\, \nabla L_1(\theta_0) \xrightarrow{D} \mathcal{N}(0, I(\theta))$ ——— *

using prev bounds $\nabla L_1(\theta_0)$ is 0-mean & var $I_\theta(\,)$

Asymptotic dist of max likelihood

Define $\hat{\theta}_n = \arg\max_\theta L_n(\theta)$

We know → $\nabla L_n(\hat{\theta}_n) = 0$

Assume model is consistent. For large n $\hat{\theta}_n \approx \theta_0$

$0 = \nabla L_n(\hat{\theta}_n) \approx \nabla L_n(\theta_0) + \nabla^2 L_n(\theta_0)(\hat{\theta}_n - \theta_0)$ ← Taylor Exp ?!

$\sqrt{n}(\hat{\theta}_n - \theta_0) \approx \underbrace{(-\nabla^2 L_n(\theta_0))^{-1}}_{\to I^{-1}(\theta_0)} \underbrace{\sqrt{n}\, L_n(\theta_0)}_{\xrightarrow{D} \mathcal{N}(0, I(\theta_0))}$

$I^{-1}(\theta_0)$: Fixed value; $\mathcal{N}(0, I(\theta_0))$: distribution

Slutsky's thm

$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{D} I^{-1}(\theta_0)\, \mathcal{N}(0, I(\theta_0))$

$\sqrt{n}(\hat{\theta}_n - \theta_0) \to \mathcal{N}(0, I(\theta_0)^{-1})$ ← $\mathbb{C}[AX] = A\,\mathbb{C}[X]\,A^T$

Very important result in statistics

Bernoulli example

$I(\theta) = -\mathbb{E}\left[\frac{d^2}{d\theta^2} \log P_\theta(x)\right]$ ↓ ?

$= -\theta \frac{d^2}{d\theta^2} \log\theta - (1-\theta) \frac{d^2}{d\theta_2} \log(1-\theta)$

$= \frac{1}{\theta} + \frac{1}{1-\theta} = \frac{1}{\theta(1-\theta)}$

Maximum Likelihood $\hat{\theta}_n = \hat{\mathbb{E}}[x]$

$\theta \to$ high var $\to$ 0 5 ——— ?

$\theta \to$ low var $\to$ ($\to 0 \to 1$)

If we consider tree-structured graph, we can do this quickly

LECTURE 14

CRAMER-RAO & CHOW-LIU

Recall:

$\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} \mathcal{N}(0, I(\theta^*)^{-1})$

$\hat{\theta}_n$: ML params; θ^*: true params

Cramer-Rao bound

1) Take $X^1, X^2, \dots X^n \sim P_{\theta^*}$ i.i.d

2) Let $\hat{\theta}$ be any unbiased estimator of θ^*

An estimator is func of data

$\hat{\theta}(X^1, \dots X^n)$ — Avg of $\hat{\theta} \equiv \theta^* = \mathbb{E}\left[\hat{\theta}(X^1, \dots X^n)\right] = \theta^*$

3) Then, under some regularity assumptions

$\mathbb{V}[\hat{\theta}] \geq \frac{1}{n} I(\theta)^{-1}$ ($A \geq B \iff A - B$ is P.S.D)

Implies ↓

Maximum likelihood is as well as we can do

Caveats

① Need to assume well-specified model ← that the true distribution has some "structure"

② max likelihood is not unbiased, but asymptotically unbiased

If n is large, error is coming from variance, than bias

Chow-Liu structure learning

What if we don't know the graph?

Take dataset $x^{(1)} \dots x^{(n)}$. How to pick graph G?

Naive:

- Loop over all graphs ← computationally expensive
- Find max likelihood params
- Return graph with highest likelihood

It will return fully connected graph → 1 clique

Solution:

① Limit G with multiple edges

Choosing the "best" tree

Max likelihood on trees

Take a tree-structured MRF with "full tables"

$$P_\theta(x) = \prod_i \phi_i(x_i) \prod_{(i,j)\in T} \phi_{ij}(x_i, x_j)$$

fixed

Given T and data $x^{(1)} \dots x^{(n)}$ want ML θ

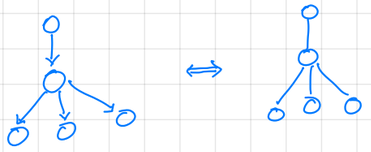
Intuition

Tree structured MRF same structed as Directed models

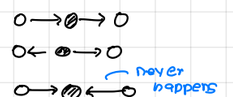
Can learn tree-structured models by counting

Claim

If each variable in a directed model has ≤ 1 parent then it is equivalent to model with same edges



Re: Bayes ball

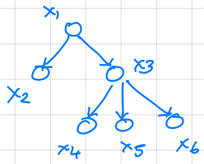


never happens in ≤ 1 parent tree

Empirical marginals

Set $\hat{P}(x_i) = \#(x_i)/n$

$\hat{P}(x_i, x_j) = \#(x_i, x_j)/n$



$P(x_1) = \hat{P}(x_1)$

$P(x_2|x_1) = \hat{P}(x_1, x_2) / \hat{P}(x_1)$

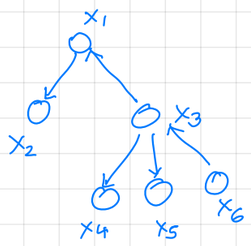
$\vdots$

$P(x_6|x_3) = \hat{P}(x_3, x_6) / \hat{P}(x_3)$

Joint distribution

$$P(x_1, x_2 \dots) = \hat{P}(x_1)^{-1} \hat{P}(x_3)^{-3} \hat{P}(x_1, x_2) \hat{P}(x_1, x_3) \hat{P}(x_3, x_4) \hat{P}(x_3, x_5) \hat{P}(x_3, x_6)$$

What if we had a different tree? Now,



$P(x_6) = \hat{P}(x_6)$

$P(x_3|x_6) = \hat{P}(x_3, x_6) / \hat{P}(x_6)$

$\vdots$

General Formula:

$$P(x) = \prod_i \hat{P}(x_i)^{1-|nb(i)|} \prod_{(i,j)\in T} \hat{P}(x_i, x_j)$$

Equivalently

$$P(x) = \prod_i \hat{P}(x_i) \prod_{(i,j)\in T} \frac{\hat{P}(x_i, x_j)}{\hat{P}(x_i)\hat{P}(x_j)}$$

LECTURE 15

CHOW-LIU + MARKOV CHAIN MONTE CARLO

Previously:

Want tree T with max likelihood for a dataset

Reasonable approach even if true dist. not tree-structured

Learning tree-structured MRF

General Form

$$P(x) = \prod_i \hat{P}(x_i)^{1-|nb(i)|} \prod_{(i,j)\in T} \hat{P}(x_i, x_j)$$

Equivalently

$$P(x) = \prod_i \hat{P}(x_i) \prod_{(i,j)\in T} \frac{\hat{P}(x_i, x_j)}{\hat{P}(x_i)\hat{P}(x_j)}$$

We want to find the tree

To learn a tree-structured MRF, set —

$$\phi_i(x_i) = \hat{P}(x_i)^{1-|nb(i)|} \qquad \phi_{ij}(x_i, x_j) = \hat{P}(x_i, x_j)$$

Or,

$$\phi_i(x_i) = \hat{P}(x_i) \qquad \phi_{ij}(x_i, x_j) = \frac{\hat{P}(x_i, x_j)}{\hat{P}(x_i)\hat{P}(x_j)}$$

Same speed (complexity) as fully-observed directed model

What is final likelihood?

$$L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log P_\theta(x^{(i)}) = \hat{\mathbb{E}} \log P_\theta(x)$$

$$= \hat{\mathbb{E}}\left[\sum_i \log \phi_i(x_i) + \sum_{i,j\in T} \log \phi_{ij}(x_i, x_j) \right]$$

$$= \hat{\mathbb{E}}\left[\sum_i \log \hat{P}(x_i) + \sum_{i,j\in T} \log \frac{\hat{P}(x_i, x_j)}{\hat{P}(x_i)\hat{P}(x_j)} \right]$$

$\log \hat{P}(x_i)$: -ve entropy

$\log \frac{\hat{P}(x_i, x_j)}{\hat{P}(x_i)\hat{P}(x_j)}$: mutual information

$$= -\sum_i H_{\hat{P}}(X_i) + \sum_{(i,j)\in T} I_{\hat{P}}(x_i, x_j)$$

?

$H_{\hat{P}}(X_i)$ does not depend on T

$$H_{\hat{P}}(X_i) = -\sum_{x_i} \hat{P}(x_i) \log \hat{P}(x_i)$$

$$I_{\hat{P}}(X_i, X_j) = \sum_{x_i, x_j} \hat{P}(x_i, x_j) \log \frac{\hat{P}(x_i, x_j)}{\hat{P}(x_i)\hat{P}(x_j)}$$

So Far:

Given dataset D and tree T, we can compute the likelihood for best MRF on tree T in terms of H and I

Define L_T = likelihood of ML params for tree T

We know

$$L_T = -\sum_i H_i + \sum_{(i,j)\in T} I_{i,j}$$

$H_{\hat{P}}(X_i)$ $\quad$ $I_{\hat{P}}(X_i, X_j)$

can be pre-calculated

To get best T? —— max spanning tree

Only true for trees

Chow-Liu algorithm

1. Input $x^{(1)} \dots x^{(n)}$
2. Define $\hat{P}(x_i, x_j) = \frac{\#(x_i, x_j)}{n}$
3. For all pairs (i,j)

$$I_{i,j} = \sum_{x_i} \sum_{x_j} \hat{P}(x_i, x_j) \log \frac{\hat{P}(x_i, x_j)}{\hat{P}(x_i)\hat{P}(x_j)}$$

$$H_i = -\sum_{x_i} \hat{P}(x_i) \log \hat{P}(x_i)$$

4. Find max. spanning tree T with weights $I_{i,j}$
5. Return $P(x) = \prod_i \hat{P}(x_i) \prod_{(i,j)\in T} \frac{\hat{P}(x_i, x_j)}{\hat{P}(x_i)\hat{P}(x_j)}$
6. Optional: Find likelihood $-\sum_i H_i + \sum_{i,j\in T} I_{i,j}$

LECTURE 16
MARKOV chain MONTE CARLO

Monte Carlo

method to approximate expectations

Idea:

Sample $X_1, X_2 \cdots X_n \sim P(x)$ i.i.d, and

$$\mathop{\mathbb{E}}_{P(x)}[f(x)] \approx \frac{1}{N} \sum_{n=1}^{N} f(x_n)$$ — accuracy improves as sample size increases

Properties:

- unbiased
- Variance decreases like $\frac{1}{N}$ — By Central limit thm
- measure arbitrary properties $P(x_3 = 4) \Rightarrow f(x) = \mathbb{I}(x_3 = 4)$

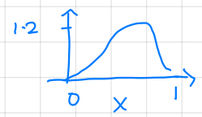
Problem

How do we sample from P? — Hard esp. high dimension

Example:

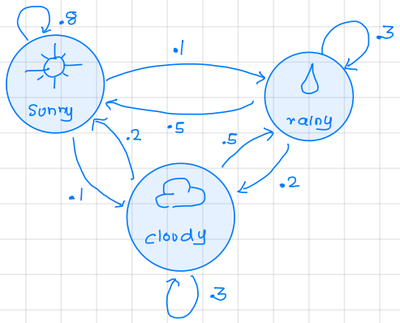
Let $P(x) = 12(x^2 - x^3)$

this is normalized $x \in [0,1]$



Markov Chain

It is a set of states with transition probabilities



← Not a graphical model!

The weather each day only depends on yesterdays weather — markov property

Transition probability

From \ To	☀	☁	🌧
☀	.8	.1	.1
☁	.5	.3	.2
🌧	.2	.5	.3

Discrete markov chain

- set of states $1, 2 \cdots D$
- set of R.V $x_t \in \{1, \cdots D\}$
- Transition Probability $P(x_{t+1} = j \mid x_t = i) = T_{ij}$ — To (From)

  Constraints $T_{ij} > 0$ & $\sum_j T_{ij} = 1$

Simulating a markov chain

$x_0 =$ ☀ $\rightarrow r_1 = .5$ drawn uniformly from $[0,1]$

$x_1 =$ ☀ $\rightarrow r_2 = .92$

$x_2 =$ ☁

Probability of a sequence

Take Markov Chain with Transition Prob T

x_0 is fixed

$$P(x_1, x_2, \cdots x_n \mid x_0) = P(x_1 \mid x_0)\, P(x_2 \mid x_1)\, P(x_3 \mid x_2) \cdots$$ — markov property

$$= T_{x_0 x_1} \cdot T_{x_1 x_2} \cdot T_{x_2 x_3} \cdot T_{x_{n-1}} T_{x_n}$$

Marginal distribution after N steps?

$P(x_t \mid x_0)$?

Now, $P(x_1 \mid x_0) = T_{x_0 x_1}$

$$P(x_2 \mid x_0) = \sum_{x_1} P(x_1, x_2 \mid x_0) = \sum_{x_1} P(x_2 \mid x_1)\, P(x_1 \mid x_0)$$

$$= \sum_{x_1} P(x_1 \mid x_0)\, T_{x_1 x_2}$$

Generalizing,

$$P(x_t \mid x_0) = \sum_{x_{t-1}} P(x_t, x_{t-1} \mid x_0)$$

$$= \sum_{x_{t-1}} P(x_t \mid x_{t-1})\, P(x_{t-1} \mid x_0)$$

$$= \sum_{x_{t-1}} P(x_{t-1} \mid x_0)\, T_{x_{t-1} x_t}$$

we can compute $P(x_t \mid x_0)$ recursively

What if x_0 is distribution?

$$P(x_t) = \sum_{x_{t-1}} P(x_t, x_{t-1}) = \sum_{x_{t-1}} P(x_t \mid x_{t-1})\, P(x_{t-1})$$ — matrix multiplication

P_t is row vector of values $P(x_t)$

$P_t = (P(x_t = 1), P(x_t = 2), \cdots P(x_t = D))$

$P_t = P_{t-1} T$

$= P_{t-2} T \cdot T$

$\vdots$

$= P_0 T^t$

Questions

- What happens when T becomes large
  - Does limit exist? $\lim_{t \to \infty} P_t$ — Not necessary
  - does it converge

Stationary distributions and Linear Algebra

Notation

E = set of states

M = |E|

π = some dist over E

π' = successor dist (row vectors)

$\pi'(y) = \sum_{x \in E} \pi(x)\, P(y \mid x)$

πT

Definition

π is stationary if $\pi = \pi T$

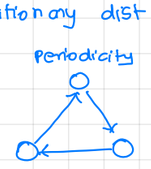
There could be 0, 1, ∞ stationary distribution

0: no convergence; ∞: T = Identity matrix

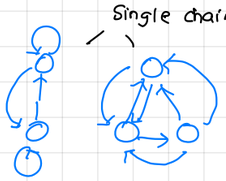
π is stationary on T if π is an eigenvector of T with eigenvalue 1

How can a chain fail to have a unique stationary dist?

Periodicity



Single chain



Reducible — multiple stationary dist

LECTURE 17
Mc MC contd

Idea:

Input P dist. → want to estimate $\mathop{\mathbb{E}}_{p(x)} f(x)$

Create a markov chain with stationary dist p.
Run markov chain for long time — $x_1, \dots x_{t_{max}}$

Use approximation $\mathop{\mathbb{E}}_{p(x)} f(x) \approx \frac{1}{t_{max}} \sum_{t=1}^{t_{max}} f(x_t)$

Note: $x_1 \dots x_{t_{max}}$ not i.i.d → they are correlated
(okay as long as not too correlated decays)

not sampled from P, but from close to P if chain run for long time

Questions:

① How to create markov chain with stationary dist P

② How to be sure that p is the only stationary p

③ How long to run the chain?

Note for a T there can be multiple stationary distribution. T = Identity every p is ↑

Stationary distribution & Linear Algebra

π is stationary on T if $\pi(y) = \sum_{x \in E} \pi(x)\, T(y|x)$

$\pi \equiv \pi T$

Principle,

① T has unique stationary dist iff it has 1 eigenvector with eigenvalue 1

② Speed of convergence is determined by spectral gap ↳ diff b/w 1st & 2nd largest eigen value

Can only do this when state space E is small → in which case we can sample exactly with SVD
→ our applications require large E

Regularity

A markov chain is regular if there exists a finite value t (means — we can strat from any state and can go to any state with non-zero prob in t iterations)

$\forall_{i,j} \quad (T^t)_{i,j} > 0$

Theorem

A regular markov chain has a unique Stationary dist

Detailed Balance

A markov chain T satisfies detailed balance w.r.t distribution π if $\forall x, x'$

$$\pi(x)\, T(x'|x) = \pi(x')\, T(x|x')$$

Intuition

- Put $\pi(x)$ "water" at each node
- For each node x send a fraction $T(x'|x)$ of it's water to x'
- π is stationary if total amount of water at each node is unchanged
- For detailed balance, flow along each pair must be equal
- If T satisfies D.B w.r.t π, then π is a stationary dist of T

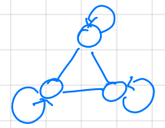
Proof:

Let $\pi' = \pi T$ be result of running chain once

Then $\pi'(x') = \sum_x \pi(x)\, T(x'|x)$ (definition)

$\equiv \sum_{x'} \pi(x')\, T(x|x')$ (Def of DB)

Imagine



D.B does not hold
Stationary dist possible

Gibbs Sampling

Init $X^{(0)}$
For $t = 1, 2 \dots t_{max}$

Randomly pick $i \in \{1, \dots D\}$ $D = \text{len}(x)$
Sample $r \sim p(x_i | x_{-i})$ ← full conditional
(sample one dimen)

Set $x_i \leftarrow r$

Record $X^{(t)} \leftarrow X$

Return $X^{(1)}, \dots X^{(t_{max})}$

T is huge, exp in # vars

Questions

① IS p stationary? → Detailed balance

② Is stationary dist unique? → Regularity

LECTURE 18
GIBBS, METROPOLIS HASTING

Linear Algebra SVD decomp not useful for large state spaces E

Regularity for Gibbs

want to show $\Rightarrow$ t s.t $(T^t)_{ij} > 0 \quad \forall i,j$

Assume $p(X_i = x_i \mid X_{-i} = x_{-i}) > 0$ always

It means its possible to reach any point from any point in finite no. of steps

Thus, regularity holds

Gibbs sampling detailed balance

$T_i(x'|x)$ → Transition prob assuming i chosen

$$\equiv \mathbb{I}[x'_{-i} = x_{-i}]\, P(x'_i | x_{-i})$$

$T(x'|x)$ = Full transition probabilities

$$= \frac{1}{D} \sum_{i=1}^{D} T_i(x'|x)$$

$$P(x) T_i(x'|x) = P(x)\, P(x'_i | x_{-i})\, \mathbb{I}[x'_{-i} = x_{-i}]$$

$$= P(x_{-i})\, P(x_i | x_{-i})\, P(x'_i | x_{-i})\, \mathbb{I}[x'_i = x_{-i}]$$

$$P(x') T_i(x|x') = P(x')\, P(x_i | x'_{-i})\, \mathbb{I}[x_{-i} = x'_{-i}]$$

$$= P(x'_{-i})\, P(x'_i | x'_{-i})\, P(x_i | x'_{-i})\, \mathbb{I}[x_{-i} = x'_{-i}]$$

$$= P(x_{-i}) P(x'_i | x_{-i})\, P(x_i | x_{-i})\, \mathbb{I}[x_{-i} = x'_{-i}]$$

$$= P(x)\, T_i(x'|x)$$

$$P(x) T(x'|x) = P(x) \frac{1}{D} \sum_{i=1}^{D} T_i(x'|x)$$

Defn of T: i chosen uniform

$$= P(x') \underbrace{\frac{1}{D} \sum_{i=1}^{D} T_i(x|x')}_{T(x|x')}$$

$$= P(x')\, T(x|x')$$

Thus detailed balance holds for Gibbs sampling

Summary

Has unique stationary dist because regular

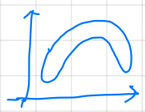
Has p as stationary dist because of detailed balance

Advantages

- Efficient per iteration
- Seems to require little training

Disadvantages

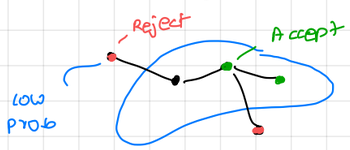
- Need to sample from $P(x_i|x_{-i})$
- mixing slow when strong correlations



To overcome drawbacks

① In each iteration, try some kind of random move

② "Accept" or "Reject" carefully to guarantee regularity & detailed balance



Proposal / Acceptance MCMC

Input $P(x)$

Init x

For $t = 1, 2, \ldots t_{max}$

- Sample $x' \sim g(x'|x)$
- Look at x and x' and calc probability $\alpha(x', x)$ of keeping x' (depend on p)
- Choose $r \in [0, 1]$ uniformly
- If $r < \alpha(x', x)$, $x \leftarrow x'$ else nothing
- $x^{(t)} \leftarrow x$

Return $x^{(1)}, x^{(2)} \ldots x^{(t_{max})}$

How do we choose α?

want to choose α s.t. p is stationary dist

Use detailed balance

What are transition prob?

If $x' \neq x$ (Propose, Accept)

$$T(x'|x) = g(x'|x)\, \alpha(x', x)$$

If $x' = x$ don't care

$$P(x) T(x'|x) = P(x') T(x|x')$$

LECTURE 19
METROPOLIS HASTINGS

Want an acceptance probability α that guarantees detailed balance

If $x' \neq x$ ⇒ $T(x'|x) = g(x'|x) \cdot \alpha(x', x)$

else ⇒ don't care ↳ D.B always satisfied

Proposal 1 — TODO

Proposal 2

π: Prob dist; g: prob. density

$$\alpha(x', x) = \min\left(1, \frac{\pi(x')\, g(x|x')}{\pi(x)\, g(x'|x)}\right)$$

Claim: DB holds

Proof:

If $\pi(x)\, g(x'|x) \geq \pi(x')\, g(x|x')$ then

$$\pi(x) T(x'|x) = \pi(x)\, g(x'|x)\, \alpha(x', x)$$

$$= \pi(x)\, g(x'|x)\, \frac{\pi(x')\, g(x|x')}{\pi(x)\, g(x'|x)}$$

$$= \pi(x')\, g(x|x')$$ (can add because $\alpha(x, x') = 1$)

$$= \pi(x')\, g(x|x')\, \alpha(x, x')$$ (order diff)

$$= \pi(x')\, T(x|x')$$

If $\pi(x)\, g(x'|x) \leq \pi(x')\, g(x|x')$, switch primes ⇒ same results

{ One of $\alpha(x, x')$, $\alpha(x', x)$ will be 1

$\theta_0 \sim \pi(\theta)$

for each iteration: → Need to specify

① $\theta'_t \sim \mathcal{N}(\theta_{t-1}, \sigma)$

θ'_t: proposed; $\mathcal{N}(\theta_{t-1}, \sigma)$: proposal dist (should satisfy) $P(\theta_a \to \theta_b) = P(\theta_b \to \theta_a)$

② $r = \frac{P(x|\theta'_t)\, P(\theta'_t)}{P(x|\theta_{t-1})\, P(\theta_{t-1})}$

③ if $r > U \sim U(0,1)$

$\theta_t = \theta'_t$

else: ignore

Metropolis Hastings Algo

Init x

For $t=1,2,\dots,t_{max}$

$x' \sim g(x'|x)$

Choose $r \in [0,1]$ randomly

IF $r \le \frac{p(x')\,g(x|x')}{p(x)\,g(x'|x)}$, then $x \leftarrow x'$

↳ don't need $\min(1,\dots)$

$x^t \leftarrow x$

Return $x^1, x^2, \dots x^{t_{max}}$

$P(x) = \frac{1}{Z}\tilde{P}(x)$, we don't need Z ↷ $\frac{P(x)}{P(x')} = \frac{\tilde{P}(x)}{\tilde{P}(x')}$

Discussion:

Performance depends critically on choosing proposal distribution

How "wide" should $g(x'|x)$ be? ↙ want to jump far / want to jump often

Paper – Weak convergence & optimal scaling {Gelman et al, 97

Acceptance rate – .234 works well in practice

One iteration of MH touches all variables ⇒ more expensive
↳ Gibbs only 1 dim

Like all MCMC → consider mixing time

Choosing g

Ideal $g(x'|x) \to P(x')$ ↷ Can't sample

this would "mix completely" with single accepted sample

Hamiltonian Monte Carlo (HMC)

Idea: Use gradients to hit a "far away" point that has high $P(x')$

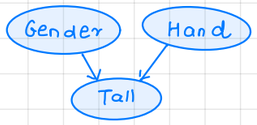
HMC is Metropolis with a particular proposal

Uses augmented space with "momentum"

LECTURE 20
BAYESIAN INFERENCE

Example

Gender	Hand	Tall
F	L	110
M	R	113
F	R	130



$\omega \equiv$ weights $\omega_t = (\mu_{FL}, \mu_{FR}, \mu_{ML}, \mu_{MR})$

$P(g,h,t|\omega) = P(g|\omega_g)\,P(h|\omega_h)\,P(t|g,h,\omega_t)$

① $P(g|\omega_g) = \begin{cases} \omega_g & g=F \\ 1-\omega_g & o.w \end{cases}$ ② $P(h|\omega_h) = \begin{cases} \omega_h \\ 1-\omega_h \end{cases}$

③ $P(t|g,h,\omega) = \mathcal{N}(t|\mu_{gh}, \sigma^2)$

↳ 1 param for diff. configs

Max Likelihood Solution

Frequentist

$$\hat{\omega} = \arg\max_{\omega} \sum_{i=1}^{N} \log P(g^{(i)}, h^{(i)}, t^{(i)} | w)$$

Bayesian Solution

Want probability of ω given data $\to P(\omega|\text{Data})$

$$P(\omega|\text{Data}) = \frac{P(\omega)\,P(\text{Data}|\omega)}{P(\text{Data})}$$

Posterior | Prior | Likelihood | Evidence

Can calculate ↳ $P(\text{Data}|\omega) = \prod_{i=1}^{N} P(x^{(i)}|\omega)$ Data generated independently

$P(\omega) \to$ Prior? Need to specify (Using domain knowledge)

$P(\text{Data}) = \int_{\omega} P(\omega)\,P(\text{Data}|\omega)\,d\omega$ ↷ don't need

Possible Prior

① $P(\omega_g) = \begin{cases} 1 & 0 \le \omega_g \le 1 \\ 0 & o.w \end{cases} = \text{Uniform}(0,1)$

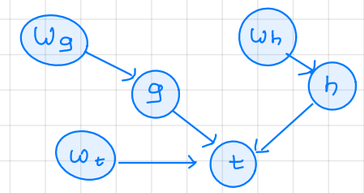
Similarly

② $P(\omega_h) = \text{Uniform}(0,1)$

③ $P(\omega_t) = \prod_g \prod_h \mathcal{N}(\mu_{gh}|\,120, 30^2)$ Empirical ?!

Treat $\mu_{FL}, \mu_{FR}, \mu_{ML}, \mu_{MR}$ is independent

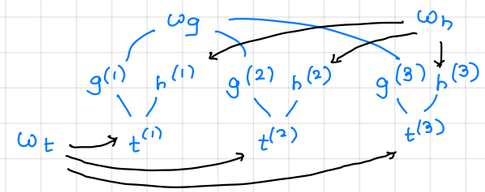
we can specify Graphical model for $P(\omega,x)$



Guideline for checking model

- Specify $P(\omega,x)$
- Sample $\omega,x \sim P$
- Think of x as "synthetic dataset"
- Ask if it looks reasonable
- Repeat

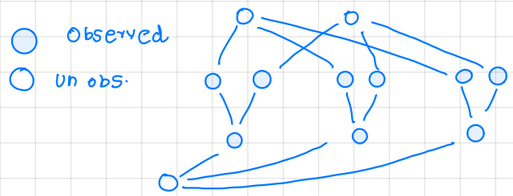
Model for $P(\omega, x^{(1)}, x^{(2)}, x^{(3)})$



For $N=3$

	$g^{(i)}$	$h^{(i)}$	$t^{(i)}$
$i=1$	F	L	110
$i=2$	M	R	113
$i=3$	F	R	130

$P(\omega|\text{Data})$ is just the conditional dist of directed model with certain variables observed



Can draw samples using MCMC

Using direct samples hard because children observed

How Bayesian Works

Design $P(\text{Data},\omega) \approx P(\omega)\prod_i P(x^i|\omega)$
↳ Specifies a Graphical model

Conditional dist $P(\omega|\text{Data})$ is posterior

Use $P(\omega|\text{Data})$ in inference
↳ draw samples using MCMC

No separation of "learning" & "inference"

Game:

80 A coins .6 Head
20 B —"— .4 —"—

Pick random coin, flip 10 times

Utility → $U(\omega_{pred}, \omega_{true})$ ≡ reward

e.g. $\propto (\omega_{pred} - \omega_{true})^2$

How happy to make pred

What strategy maximizes expected reward?

Decision rule $\alpha(x^{(1)} \dots x^{(10)})$

↳ what ω_{pred} to choose for every possible dataset

Expected Utility

$$\mathbb{E}_{P(\omega)} \; \mathbb{E}_{P(x^{(1)} \dots x^{(10)} \mid \omega)} \; U(\alpha(x^{(1)} \dots x^{(10)}), \omega)$$

Prior: $P(\omega)$ — expectation over possible "true" coin bias

Likelihood: $P(x^{(1)} \dots x^{(10)} \mid \omega)$ — all datasets that can be generated using "true" coin

$P(\omega = a) = .8$
$P(\omega = b) = .2$

$P(x \mid \omega)$:

x \ ω	a	b
0	.4	.6
1	.6	.4

Best strategy claim

$$\alpha(x^{(1)}, \dots x^{(10)}) = \underset{\omega'}{\operatorname{argmax}} \; \mathbb{E}_{P(\omega \mid x^{(1)}, \dots x^{(10)})} \; U(\omega', \omega)$$

LECTURE 21
BAYESIAN INF continued

Goal: Choose decision rule (α) to make Expected Utility as high as possible

Expected Utility $\mathbb{E}_{P(\omega)} \; \mathbb{E}_{P(x' \dots x^{10} \mid \omega)} \; U(\alpha(x' \dots x^{10}), \omega)$

↳ how good is decision making

Best Strategy:

$$\alpha(x' \dots x^{10}) = \underset{\omega'}{\arg\max} \; \mathbb{E}_{P(\omega \mid x' \dots x^{10})} \; U(\omega', \omega)$$

Why: Exp Uti equivalently

$$\mathbb{E}_{P(x' \dots x^{10})} \; \mathbb{E}_{P(\omega \mid x' \dots x^{10})} \; U(\alpha(x' \dots x^{10}), \omega)$$

No other strategy can have higher utility on average

Assumptions about Bayesian

- Need to know prior $P(\omega)$
- Need to know model $P(x \mid \omega)$ ⌐ data generation process

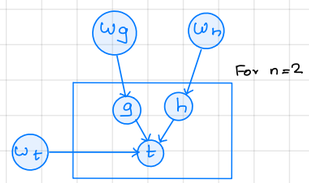
does not influence posterior when data large

$$P(\omega \mid \text{Data}) \propto P(\omega) \prod_{i=1}^{n} P(x^i \mid \omega)$$

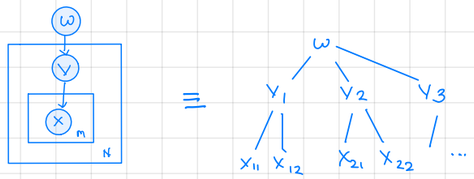
Posterior ∝ Prior × Likelihood

Aside: Plate Notation

Draw a box around repetitions



Plates can be nested.



Emulate MAP with Bayesian.

$$U(\omega_{pred}, \omega) = \mathbb{I}[\omega_{pred} = \omega] \quad \text{if discrete}$$
$$= \delta(\omega_{pred} = \omega) \quad \text{if continuous}$$

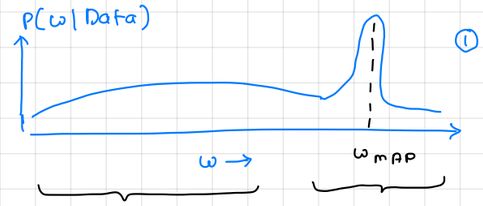
Want $P(g=F, h=R, t=110)$

Given a specific ω, we can compute $P(g=F \dots \mid \omega)$

Bayesians want $\int P(\omega \mid \text{Data}) \; P(g=F, h=R, t=110 \mid \omega)$

Sample $\omega^1 \dots \omega^{t_{max}}$ and approximate

$$\approx \frac{1}{t_{max}} \sum_{t=1}^{t_{max}} P(g=F, h=R, t=110 \mid \omega^t)$$



These ω think many right handed Femoles

These ω think less R, F

P(h=R, g=F)

① Bayes will give high prob. because many samples from left area

② MAP will give low prob

Advantages

- use domain knowledge
- some way "optimal"

Disadvantages

- Hard inference
- Computationally expensive

LECTURE 22
VARIATIONAL INFERENCE

General Strategy

① Input $P(z, x)$ and observed X
② Let $q_\omega(z)$ be a simple dist with params ω
③ Minimize $KL(q_\omega(z) \,\|\, P(z \mid x))$

ELBO decomposition

Evidence lower bound optimization

$$\log P(x) = \sum_z q_\omega(z) \log \frac{P(z,x)}{q_\omega(z)} + \sum_z q_\omega(z) \log \frac{q_\omega(z)}{P(z \mid x)}$$

ELBO — $KL(q_\omega(z) \,\|\, P(z \mid x))$

Proof:

$$= \sum_z q_\omega(z) \log P(z,x) - q_\omega(z) \log q_\omega(z) + \sum_z q_\omega(z) \log q_\omega(z) - \sum_z q_\omega(z) \log P(z \mid x)$$

$q_\omega(z)$ → has to be simple to sample from
eg: Normal

Note:

① Can't compute KL term → don't know $P(z \mid x)$
② Can approx ELBO → $z \sim q_\omega$, compute $\log \frac{P(z,x)}{q_\omega}$
③ $KL \geq 0$
④ ⇒ ELBO ≤ $\log P(x)$ (lower bound)
⑤ maximizing ELBO ≡ minimize KL ∵ $\log P(z,x)$ = constant

Uses of VI

① Want lower bound on $\log P(x)$ → by learning params
② Want $q(z) \approx$ posterior for inference

Why not minimize $KL(P \| q)$

$$KL(q\|P) = \sum_z q(z) \log \frac{q(z)}{P(z)}$$ ← Expectation wrt q

$$KL(P\|q) = \sum_z p(z) \log \frac{p(z)}{q(z)}$$ ← Exp w.r.t p

Posterior is unknown, can't use :)

VI Algorithm

Input $P(x,z)$ and Fixed x

We know $ELBO(q_\omega \| p(x,z))$

Maximize ELBO w.r.t ω somehow

Use $q_\omega(z)$ as proxy for $p(x,z)$

Expectation - Maximization

Given X, want to maximize $P_\theta(x)$ wrt θ

$$P_\theta(x) = \sum_z P_\theta(x,z)$$

Usual derivation

$$\log P_\theta(x) = \log \sum_z P_\theta(x,z)$$

$$= \log \sum_z q(z) \frac{P_\theta(x,z)}{q(z)}$$

$$\geq \sum_z q(z) \log \frac{P_\theta(x,z)}{q(z)}$$ — Jensen's inequality

Algorithm

- Set $q(z) = P_\theta(z|x)$
- Maximize $\sum_z q(z) \log \frac{P_\theta(x,z)}{q(z)}$ w.r.t θ
- Repeat

Produces local maximum of $P_\theta(x)$

Variational EM

learning & inference simultaneously

Goal: maximize $P_\theta(x)$

Strategy

- Define $ELBO(q_\omega(z) \| P_\theta(z,x))$
- we know $\forall \omega, \theta \quad ELBO(q_\omega(z) \| P_\theta(z,x)) \leq \log P_\theta(x)$
- maximize ELBO over both ω, θ simult.

Since ELBO is lowerbound → approach reasonable

IF $q_\omega(z) = P_\theta(z|x) \Rightarrow ELBO = \log P_\theta(x)$

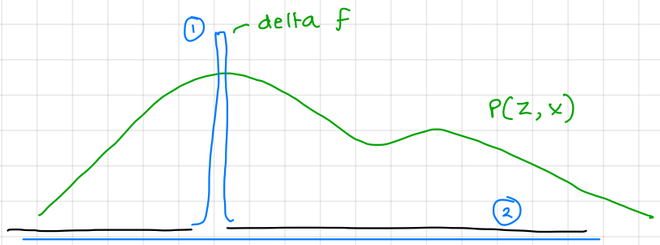
Examples

- Regular EM → where $q_\omega(z)$ can represent $P(z|x)$ exactly
- Reg. VI → where θ is fixed
- VAE $q_\omega(z)$ and P_θ represented with neural nets

Intuition

$$ELBO = \sum_z q_\omega(z) \log P(z,x) - \sum_z q_\omega(z) \log q_\omega(z)$$

① encourages $q_\omega(z)$ to be high where $P(z|x)$ is high
② encourages $q_\omega(z)$ to be random (high entropy)



Stochastic Gradient Variational Inference

Black-box VI

Idea: Need $q_\omega(z)$ to
- sample from
- evaluate
- compute entropy
- maybe compute gradient ← use autodiff

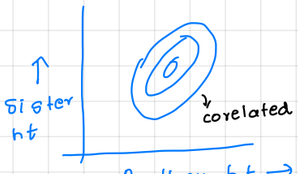
Need $P(z,x)$
- evaluate
- maybe compute gradient

Algorithm

- Start with some ω
- For $t = 1,2 \ldots$ tmax
  - Get unbiased estimate g of ∇_ω ELBO
  - Take gradient step - RMSprop, Adam, etc
- Return ω

How to estimate $\nabla_\omega ELBO$? { should be cheap to compute; low variance

LECTURE 23
SGVI



We have avg. ht of sisters = 64 in
P(Brother, Sister)
we get to sample 1 from P(B,S)
want avg. ht of Brother

Can estimate Brother
Better → Brother + (μ − Sister)

Input $P(z,x)$. Fixed x
Want ω s.t $q_\omega(z) \approx P(z|x)$
↑ posterior estimate

Idea:
For $t = 1,2 \ldots$
- Get unbiased estimate of ∇_ω ELBO
- Take a step

$\approx -KL(q_\omega(z) \| p(z|x))$

$$ELBO = \int q_\omega(z) \log P(z,x)\, dz - \int q_\omega(z) \log q_\omega(z)\, dz$$

find high prob. regions ← q_ω should have high divergence

Often Entropy is known in closed form

For 1-d Gaussian $\omega = (\mu, \sigma)$ → $H = \frac{1}{2} \ln(2\pi e \sigma^2)$ — ind. of μ

n-d Gaussian $\omega = (\mu, \Sigma)$ → $H = \frac{1}{2} \ln |2\pi e \Sigma|$ ← def.

If we have exact entropy → can compute gradient

Mean evidence → $\int_z q_\omega(z) \log P(z,x)\, dz$

$= E_q[\log P(z,x)]$

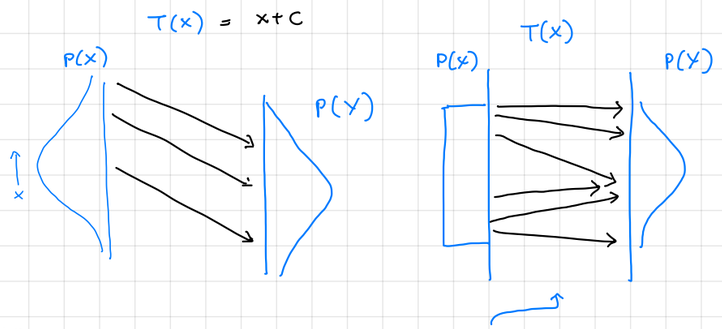
estimate by ① drawing a single sample from $q_\omega(z)$
② Return $\log p(x,z)$

However, want estimate of gradient $\equiv \nabla_\omega \int_z q_\omega(z) \log P(z,x)\, dz$

Aside: Transformation of densities

if $Y = T(X)$ (T: transformation) then

$P(X=x) = P(Y=T(x)) \left|\det(\nabla T(x))\right|$ (Jacobian)



we need det. to model non-linear

If $T(x)$ changes near X quickly

Re-parameterization trick estimator

Find - dist $q(\varepsilon)$ independent of ω
- Transformation $T_\omega(\varepsilon)$

such that $\forall \omega$

$T_\omega(\varepsilon) \overset{d}{=} z, \quad \varepsilon \sim q(\varepsilon), \; z \sim q_\omega(z)$

{ $q(\varepsilon)$ is fixed, transform it non-linearly to get $q(\omega)$

For any f, $\mathbb{E}_{q_\omega(\varepsilon)} f(z) = \mathbb{E}_{q(\varepsilon)} f(T_\omega(\varepsilon))$

Thus: $\nabla_\omega \mathbb{E}_{q_\varepsilon(z)} f(z) = \mathbb{E}_{q(\varepsilon)} \nabla_\omega f(T_\omega(t))$

since, $q(\varepsilon)$ ind. of ω

Example: Gaussian re-parameterization

FACT
$X \sim \mathcal{N}(0, I)$ then $Cx + \mu \sim \mathcal{N}(\mu, CC^T)$

$\omega = (\mu, C)$ with $\Sigma = CC^T$

Then,

$q(\varepsilon) = \mathcal{N}(\varepsilon \mid 0, I)$

$T_\omega(t) = C\varepsilon + \mu$

Compute $\nabla_\omega f(T_\omega(t))$ by auto-diff

Mean-Evidence

$\mathbb{E}\, \nabla_\omega \log P(T_\omega(\varepsilon), x) = \nabla_\omega \int q_\omega(z) \log P(z,x)\, dz$

Given $x = (x^1 \cdots x^n)$

$P(z,x) = P_z(z) \prod P(x^i \mid z)$ (x iid)

Then,

$\nabla_\omega \log P(T_\omega(\varepsilon), x) = \nabla_\omega \log P_z(T_\omega(\varepsilon)) + \sum_{j=1}^{n} \nabla_\omega \log P_{x|z}(x^{(i)} \mid T_\omega(\varepsilon))$

expensive for large n (log calc.)

Reducing expense via sub-sampling

$\log P(z,x) \approx \log P_z(z) + \frac{N}{m} \sum_{i \in \text{mini batch}} \log P_{x|z}(x^i \mid z)$

Doubly-stochastic estimator

$\nabla_\omega \log P_z(T_\omega(\varepsilon)) + \frac{N}{m} \sum \log P_{x|z}(x^i \mid T_\omega(\varepsilon))$

- Stochastic over ε & mini-batch

$q_\omega(z) = \mathcal{N}(z \mid \omega, 0.5 \cdot I)$ (variational distribution)

$\int q_\omega(z) \log p(\omega, \text{Data})\, dz$ (mean Evidence)

LECTURE 24
V·I & REVIEW

Re-parameterization discussion

Pros — Low-ish variance

Cons — Only works for continuous z
Need to find $q(\varepsilon)$ & $T_\omega(\varepsilon)$

Usually use auto. diff

Score function Estimator

We know,

$\nabla_\omega \int q_\omega(z) \log P(z,x)\, dz$ (Expectation)

$= \int q_\omega(z) \log P(z,x)\, \nabla_\omega \log q_\omega(z)\, dz$

$\because \nabla_\omega \log q_\omega(z) = \frac{1}{q_\omega} \nabla_\omega q_\omega(z)$

Estimator

↳ Sample $z \sim q_\omega(z)$

Set $g = \log P(z,x)\, \nabla_\omega \log q_\omega(z)$

Then - $\mathbb{E}\, g = \nabla_\omega \mathbb{E}_{q_\omega(z)} \log P(z,x)$

What if we don't have closed-form entropy?

$-\nabla_\omega \int q_\omega(z) \log q_\omega(z) = -\int (\nabla_\omega q_\omega(z)) \log q_\omega(z)$ (chain Rule)

$= -\int q_\omega(z) (\nabla_\omega \log q_\omega(z))$

Can estimate like mean Evidence

$\nabla \log x = \frac{1}{x} \nabla x$

$= \nabla \frac{q_\omega}{q_\omega} \quad q_\omega = \nabla 1 = 0$

Can estimate

$\nabla_\omega \text{ELBO} = \nabla_\omega \mathbb{E}_{q_\omega(z)} (\log P(z,x) - \log q_v(z))\big|_{v=\omega}$

Discussion of Score function

Pros — simple, Easy, works with discrete z

Cons — Typically high variance

REVIEW

1. Conditional Ind. in directed model

$P(1 \perp 2 \mid 3)$?

$P(x_i \perp X_{nd(i)} \mid X_{Pa(i)})$

$P(x_i \perp X_{-i} \mid X_{nb(i)})$

2. Cond ind in Undirected model

3. HC Theorem — Something involving proof

Change assumption or ... work?

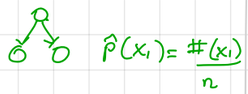
$P(x) > 0$

$G_{ij} > 0$

4. Message Passing efficiency (Time complexity)

5. Msg Pass inference on tiny model

6. Chow Liu — In a tiny dataset what is ML tree



$\hat{P}(x_i) = \frac{\#(x_i)}{n}$

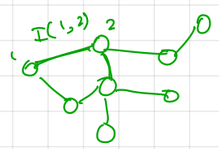
Tree-structured MRFs $\equiv$ DAG $\Rightarrow$ Counting reasonable

$P(x) = \prod \phi_i(x_i) \prod \phi_j(x_i, x_j)$

$\hat{P}(x_i)$ $\hat{P}(x_i, x_j)$

$\log \sum \hat{P}(x_i)$ — Ind tree

$\log \sum x_i x_j$ — mutual Int



7. Exponential Family — cheat sheet

given a simple dist & dataset

is θ an optimal of likelihood

8. Bayesian Inf. — Conceptual

8-10 lsh Q

3-5 parts

Is decision rule optimal?

Compare to ML, MAP

9. MCMC — Conceptual — Regularity, Detailed balance

Given Metropolis / Gibbs change some aspect of algo → check if work

Give argument?

10. VI — Conceptual — ELBO

11. SGVI — Given simple p, q derive gradient estimator

init ω

For $t_1 \dots t_{max}$

$\nabla\omega = \nabla_\omega$ ELBO $\longrightarrow$ $\sum_z q_\omega(z) \log \frac{P(z,x)}{q_\omega(z)}$

$\omega \leftarrow \omega + \alpha \nabla\omega$

$P(z) \prod P(x^i \mid z)$

$\int_\varepsilon q(\varepsilon) \log P(\tau, \omega) \prod P(x_i)$