2 Sep:

The big picture: - probabilistic model

[RECAP]

1. Specify a likelihood func.: $p(x|\vec{\theta})$ → $x$: observed data ($N$ samples, $D$ features): $x \in \mathbb{R}^{N \times D}$; → $\vec{\theta}$: vector of model params. $\theta \in \mathbb{R}^p$ ($p$: # params)
2. Max. likelihood estimation (frequentist approach) → if likel. $\sim$ Gauss., ML estim. = least squares

a) $\hat{\theta}_{ML} := \underset{\vec{\theta}}{\operatorname{argmax}}\ p(x|\vec{\theta})$ ($\hat{\theta}_{ML}$: estimator of ML) where, $p(x|\vec{\theta}) = \prod_i p(x_i|\theta)$ [i.i.d. assump.: indep. & ident. distrib.]

b) Add regularizer for smoother estimates: MAP (= penalized ML)

c) Specify a prior on model params: $p(\vec{\theta})$

$$\theta_{MAP} = \underset{\vec{\theta}}{\operatorname{argmax}}\ p(x, \vec{\theta}) = \underset{\vec{\theta}}{\operatorname{argmax}}\ p(x|\theta)p(\theta) = \underset{\vec{\theta}}{\operatorname{argmax}}\ \frac{p(x|\theta)p(\theta)}{p(x)} = \underset{\theta}{\operatorname{argmax}}\ p(\theta|x)$$

($p(x, \vec{\theta})$: maximize joint probab.; $p(x|\theta)$: likel.; $p(\theta)$: prior; $p(\theta|x)$: posterior; dividing by this does not affect the 'argmax')

★ Bayesian approach: - specify prior $p(\theta)$
- Calcul. posterior $p(\theta|x)$

Supervised learning:

ip : $X \in \mathbb{R}^{N \times D}$

o/p : $Y \in \mathbb{R}^N$ (regression), $Y \in \{C_1, \dots, C_K\}^N$ (classif.)

Conditional likel.: $p(Y|X, \vec{\theta}) = \prod_{i=1}^{N} p(y_i|x_i, \vec{\theta})$

↳ Learning same as un-conditional case:
1) ML or,
2) introduce prior and MAP.

3. After learning → make preds:

frequentist: $\hat{\vec{\theta}}_{ML} = \underset{\vec{\theta}}{\operatorname{argmax}}\ p(Y|X, \theta)$

predictive distrib. is: $p(y^*|x^*, \theta_{ML})$ ($y^*$: new $Y$; $x^*$: new $X$)

→ Best/chosen prediction = $\hat{y} = \underset{y^*}{\operatorname{argmax}}\ p(\vec{y}^*|\vec{x}^*, \hat{\vec{\theta}}_{ML})$ (i.e. has max. probab.)

→ But sometimes we don't want the one that has max. probab. BUT minimizes expected loss.

in medicine loss func. is very asymm. eg: TP, FP, etc

(Bishop 1.5) → we have a loss func. $L(y, y_{pred})$ (this quantifies loss)
→ Min. expected loss func.:

$$\hat{y} = \underset{y^*}{\operatorname{argmin}} \sum_y p(y|x^*, \vec{\theta})\, L(y, y^*)$$ → scalar?

↳ prob. of true $Y$ given new $X$ & $\hat{\vec{\theta}}_{ML}$?

**Bayesian** predic. use complete posterior: (over params $\theta$)
$p(\vec{\theta} | X, Y)$

predictive distr. $p(y^* | x^*, X, Y) = \int p(y^* | x^*, \vec{\theta}) \; p(\vec{\theta} | X, Y) \, d\theta$

Then proceed as before ... –

★ Sometimes we have latent var. in our prob. model:
$\vec{z} = (z_1, \dots z_n)$

So in case (1):
Sum/integrate latent var. out:

$k^N$ $2^2$ $2^2$

$p(x | \vec{\theta}) = \sum_z p(x, z | \vec{\theta}) = \sum_{z_1} \cdots \sum_{z_n} p(x, z | \theta)$ [discrete]

$p(x | \theta) = \int p(x, z | \theta) \, dz$ [contin.]

In practice, how do we learn in this case??
1) EM algo.
2) variational / sampling methods for approx. Bayesian learning & predic.
(Bayes)

## Exponential Family Distrib.: (Bishop 2.4)

Def: $p(\vec{x} | \vec{\eta}) = h(\vec{x}) \, g(\vec{\eta}) \exp\left(\vec{\eta}^T \cdot \vec{u}(\vec{x})\right)$

choose 3 func.
$\vec{\eta}$ = natural params | $h(x)$ = backgr. measure
$g(\cdot)$ = normal. term | $\vec{u}(\vec{x})$ = sufficient/natural stats.
$z(\vec{\eta}) = \frac{1}{g(\vec{\eta})} = \int \exp\left(\vec{\eta}^T \cdot \vec{u}(\vec{x})\right) h(\vec{x}) \, d\vec{x}$

$\frac{\partial}{\partial \vec{\eta}} \log(z(\vec{\eta})) = -\frac{\partial}{\partial \vec{\eta}} (\log(g(\vec{\eta}))) = \frac{1}{z(\vec{\eta})} \int \exp(\;) \cdot \vec{u}(\vec{x}) \, h(\vec{x}) \, d\vec{x} = \mathbb{E}\left(\vec{u}(\vec{x}) | \vec{\eta}\right)$

Eg: Gaussian:

(multivariate gauss.) $\mathcal{N}(\vec{x} | \vec{\mu}, \Sigma) = \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} (\vec{x} - \vec{\mu})^T \Sigma^{-1} (\vec{x} - \vec{\mu})\right)$

using $\mathrm{Tr}(ABC) = \mathrm{Tr}(BCA) = \mathrm{Tr}(CAB)$

$= \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} \vec{\mu}^T \Sigma^{-1} \vec{\mu}\right) \exp\left(\frac{-1}{2} \mathrm{Tr}(\Sigma^{-1} \vec{x} \vec{x}^T) + (\vec{\mu}^T \Sigma^{-1} \vec{x})\right)$

$(\Sigma^{-1}\mu)^T \Sigma (\Sigma^{-1}\mu)$; $\vec{\eta}_2$; $u_2(\vec{x})$; actually $\mathrm{Tr}(\vec{x}^T \Sigma^{-1} \vec{x})$; $\vec{\eta}_1$; $u_1(\vec{x})$

$g(\vec{\eta})$; func. of $\vec{\eta}_1$ & $\vec{\eta}_2$

$\Rightarrow \begin{pmatrix} \vec{\eta}_1 \\ \vec{\eta}_2 \end{pmatrix} = \begin{pmatrix} \Sigma^{-1} \vec{\mu} \\ -\frac{1}{2} \Sigma^{-1} \end{pmatrix} \quad \| \quad \begin{pmatrix} \vec{\mu}_1(x) \\ \vec{\mu}_2(x) \end{pmatrix} = \begin{pmatrix} \vec{x} \\ \vec{x} \vec{x}^T \end{pmatrix} \quad \| \quad h(x) = 1$

$(D + D^2$ elements$)$

$\mathrm{Tr}(AB) = \sum_{ij} (AB)_{ii}$
$= \sum_{ij} A_{ij} B_{ij}$ [bcz, $\Sigma$ is symm., so ...]
$= \mathrm{vec}(A^T) \, \mathrm{vec}(B)$

**Q.** if $\vec{x} \sim \mathcal{N}(\vec{\mu}, \Sigma)$ then $\mathbb{E}(\vec{x}) = \mu$

$$\Rightarrow \mathbb{E}(\vec{x}|\vec{\eta}) = \mathbb{E}(u_1(\vec{x})|\vec{\eta}) = \frac{\partial}{\partial \vec{\eta}_1} \log z(\vec{\eta}) = \frac{\partial}{\partial \vec{\eta}_1}\left(\frac{1}{2}\vec{\eta}_1^T \Sigma \vec{\eta}_1\right) = \Sigma \vec{\eta}_1 = \Sigma\Sigma^{-1}\vec{\mu} = \mu$$

Proved

$\Rightarrow$ Derive second moment on your own (var.).

**2** ML of m.v. Gaussian (2.3 Bishop)

given Dataset $\mathcal{D} = \{\vec{x}_1, \dots, \vec{x}_n\}$, log-likel.

$$L(\vec{\eta}, \mathcal{D}) = \sum_{n=1}^{N} \ln p(\vec{x}_n|\vec{\eta})$$

$$\frac{\partial L}{\partial \vec{\eta}} = \sum_n \ln\left[h(\vec{x})\, g(\vec{\eta}) \exp(\vec{\eta}^T \vec{u}(\vec{x}_n))\right] = N \frac{\partial}{\partial \vec{\eta}} \ln(g(\vec{\eta})) + \sum_n \vec{u}(\vec{x}_n)$$

$$\Leftrightarrow \boxed{-\frac{\partial}{\partial \vec{\eta}} \ln g(\vec{\eta}) = \frac{1}{N}\sum_{n=1}^{N} u(\vec{x}) = \overline{u(\vec{x})}} \quad \text{(empirical mean)}$$

$$\boxed{\vec{\eta}_{ML} = F(\vec{x})}$$

★ For Gaussian:

(first moment) $\vec{\mu}_{ML} = \mathbb{E}(\vec{x}) = \overline{u(x)} = \bar{\vec{x}} = \frac{1}{N}\sum_n \vec{x}_n$

(2nd moment) $\vec{\Sigma}_{ML} = \mathbb{E}(\vec{x}\vec{x}^T) + \vec{\mu}_{ML}\vec{\mu}_{ML}^T$

$$= \frac{1}{N}\sum_n \vec{x}_n \vec{x}_n^T + \vec{\mu}_{ML}\vec{\mu}_{ML}^T$$

**3** Marginal Distrib. & condit. distrib.:

$\vec{x} \sim \mathcal{N}(\vec{\mu}, \Sigma)$

Suppose $\vec{x}$ split into $(\vec{x}_a, \vec{x}_b)$, Def: $\vec{\mu} = (\vec{\mu}_a, \vec{\mu}_b)$, $\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}$

$\Sigma$ is symm. $\rightarrow$ so, $\Sigma_{ab} = \Sigma_{ba}^T$, $\Sigma_{aa} = \Sigma_{aa}^T$, ⋮

Marg. Distr. $p(\vec{x}_a) = \mathcal{N}(\vec{x}_a | \vec{\mu}_a, \Sigma_{aa})$

conditional. $p(\vec{x}_a|\vec{x}_b) = \mathcal{N}(\vec{x}_a | \vec{\mu}_{a|b}, \Sigma_{a|b})$ where, $\Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}$

$$\mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(\vec{x}_b - \vec{\mu}_b)$$

**4** Another useful identity

product of 2 Gaussians: $\mathcal{N}(\vec{x}|\vec{a}, A)\,\mathcal{N}(\vec{x}|\vec{b}, B) = \underbrace{\mathcal{N}(\vec{a}|\vec{b}, A+B)}_{\text{normalizing}}\,\mathcal{N}(\vec{x}|\vec{c}, C)$ with $C = (A^{-1}+B^{-1})^{-1}$, $\vec{c} = C(A^{-1}a + B^{-1}b)$

"splitting of square"

## Student t-distrib. :

→ with Gauss., density falls rapidly ($\propto x^2$) for $x \to \pm\infty \Rightarrow \mathcal{N}(\,) \propto e^{-\frac{1}{2\sigma^2}x^2}$
→ sometimes we need heavier tails., $St \propto |x|^{-\alpha}$ (power law)

Mixt. of Gauss. : $St(x|\mu, \lambda=\frac{a}{b}, \nu=2a)$

1. Draw precision $\tau \sim \text{gamma}(a,b)$
2. Draw $x \sim \mathcal{N}(\mu, \tau^{-1})$

gamma distrib. : $\text{gamma}(\tau|a,b) = \frac{b^a}{\Gamma(a)} \tau^{a-1} e^{-b\tau}$

gamma func. : $\Gamma(x) = \int_0^\infty u^{x-1} e^{-u}\, du$ , $\Gamma(x+1) = \Gamma(x)\cdot x$ (factorial)

★ $p(x|\mu,a,b) = St(x|\mu,\lambda,\nu) = \frac{b^a}{\Gamma(a)\,2\pi} \int_0^\infty \tau^{a-\frac{1}{2}} \exp\left(-\left(b + \frac{1}{2}(x-\mu)^2\right)\tau\right) d\tau$

Lab 1:

(1) $\tanh a = \frac{e^a - e^{-a}}{e^a + e^{-a}} = \frac{1-e^{-2a}}{1+e^{-2a}} \Rightarrow \int \phi(a)\,da = -\int \frac{1-e^{-2a}}{1+e^{-2a}}$

Let, $y = e^{-2a}$ ⇒

$-\frac{1}{2}\ln y = a$

$\frac{da}{dy} = \frac{-1}{2y}$

$= +\frac{1}{2}\int \frac{1-y}{(1+y)y}\,dy = \frac{1}{2}\int \frac{1+y-2y}{(1+y)y}$

$= \frac{1}{2}\int \frac{1+y}{(1+y)y} - \frac{1}{2}\int \frac{2y}{(1+y)y}$

$= \frac{1}{2}\int \frac{1}{y}\,dy - \int \frac{1}{1+y}\,dy$

$= \frac{1}{2}\ln y - \ln(1+y) + C$

$= \frac{1}{2}\cdot -2a - \ln(1+e^{-2a}) + C$

$p(a) = e^{-a - \ln(1+e^{-2a}) + C}$

(2) $\int -a^2\,da = \ln p(a))$

$p(a) = e^{-\frac{1}{4}a^4 + C}$

$\int \frac{e^a - e^{-a}}{e^a + e^{-a}}\,da$

$x = e^a - e^{-a} \Rightarrow \ln x = 2a$, $a = \frac{\ln x}{2}$, $\frac{dx}{2x} = da$

$\int \frac{1}{u}\,du$

$= \ln(u) + C$

9th Sept

2.4.2 Conjugate priors (read)

2.3.6 Bayesian inference for Gaussian:

1-dim data $X = \{x_1, \dots, x_n\}$

① Var. known, mean estimated
$\sigma^2$ given, $\mu = ?$

conj. prior $\Rightarrow p(\mu) = \mathcal{N}(\mu \mid \mu_0, \sigma_0^2)$
posterior $\Rightarrow p(\mu|D) = \mathcal{N}(\mu \mid \mu_N, \sigma_N^2)$

$\mu_N = (\quad)$, $\frac{1}{\sigma_N^2} = (\quad)$

② $\sigma^2 = ?$, $\mu$ = known ($\lambda$ = precision)

likelihood $p(X|\lambda) = \prod_{i=1}^{N} \mathcal{N}(x_i \mid \mu, \lambda^{-1})$

$\propto \lambda^{N/2} \exp\left(\frac{-\lambda}{2} \sum_i (x_i - \mu)^2\right)$

$\propto \text{gamma}\left(\lambda \mid \frac{N}{2} + 1, \frac{1}{2}\sum_n (x_n - \mu)^2\right)$

prior: $p(\lambda) = \text{gamma}(\lambda \mid a_0, b_0)$
poster.: $p(\lambda|X) = \text{gamma}(\lambda \mid a_N, b_N)$

$a_N = a_0 + \frac{N}{2}$, $b_N = b_0 + \frac{1}{2}\sum_n (x_n - \mu)^2$

③ $\sigma^2 = ?$, $\mu = ?$

"Normal gamma distr."

$p(\mu, \lambda \mid \mu_0, \beta, a, b) = \mathcal{N}(\mu \mid \mu_0, (\beta\lambda)^{-1}) \cdot \text{gamma}(\lambda \mid a, b)$

- $\mathcal{N}(\mu \mid \mu_0, (\beta\lambda)^{-1})$: prior on $\mu$
- $\text{gamma}(\lambda \mid a, b)$: prior on $\lambda$
- not indep.

Sample $\lambda$ then use it to sample $\mu$.

# Inform. Theory :

inform (in bits) = $-\log_2(\text{prob.})$

∴ inform. of A $\quad h(A) = -\log_2 p(A)$
or,
$-\ln p(A)$

## Shannon entropy :

$$H(X) = -\sum_{x \in p(x)} p(x) \cdot \ln_2 p(x) \qquad = -E\left(\ln_2 p(x)\right)$$

(X is a r.v. → in discrete domain)

Eg: coin flip : $X = \{0, 1\} = D_x$

$$H_x = -\frac{1}{2}\log_2 \frac{1}{2} - \frac{1}{2}\log_2 \frac{1}{2}$$

$$= 1 \text{ bit}$$

if K indep. fair coins : K bits

## Differential Entropy :

$$H(X) = -\int p(x) \log_2 p(x)\, dx \qquad (X \text{ is r.v. in cont. domain, i.e., } x \in \mathbb{R}^d)$$

## KL divergence : (relative entropy)

$$KL(p(x) \| q(x)) = -\int p(x) \log \frac{q(x)}{p(x)}\, dx$$

Imp. properties -

- ↳ KL( ) ≥ 0
- ↳ KL(p||q) = 0 ⟹ p = q (if p, q sufficiently regular)
- ↳ KL(p||q) ≠ KL(q||p)
- ↳ KL(p||q) + KL(q||r) ≠ KL(p||r)

Eg: 2 coin tosses (indep.)

mutual info = 0
cond. info = same as info. of 1 coin.
H(X) = H(X|Y)

## Conditional entropy :

$$H(y|x) = -\underbrace{\int p(x)}_{\substack{\text{average} \\ \text{over all} \\ \text{values of } x}} \cdot \int p(y|x) \ln p(y|x)\, dx$$

(joint) $\quad H(x, y) = H(x) + H(y|x)$
$= H(y) + H(x|y)$

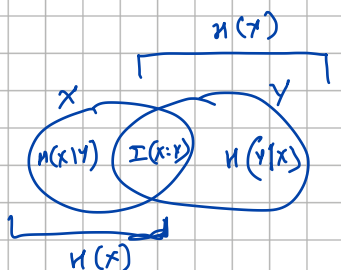
## Mutual infor. : (shared info. betw. 2 distr.)

$p(x,y) = p(x)p(y)$ ← if they are indep.

⇓

If not indep., then we want to know how much info. they share.

$$I(x:y) = KL\left(p(x,y) \| p(x) \cdot p(y)\right)$$
$$= H(x) - H(x|y)$$
$$= H(y) - H(y|x)$$



Info. theory interpretation of MLE :

$$\min_{\theta} KL(p(x) \,\|\, q(x|\theta))$$

$p(x)$ ↑ empirical distrib. ; $q(x|\theta)$ ↑ model distrib.

# Ch-8 Probabilistic Graphical Models

- graphical way of expressing prob. dist.
- useful when many vars / bid data.
- useful for causal reasoning / modelling
- design / communicate stat models.
- cond. indep. relations are encoded in graph.
- directed / undirected [DAG / MRF]
- inference efficiently, calcul. cond. dist. (Eg. Bayes theo.)

* more useful if have less edges.

Eg: directed graph. model

$$p(A,B,C) = p(C) \cdot p(B|C) \cdot p(A|B,C) \quad \text{or,} \quad p(A) \cdot p(B|A) \cdot p(C|A,B)$$



(nodes are r.v.)

- A depends on B & C
- B " " C

* more useful if have less edges:

$$p(A,B,C) = p(C) \cdot p(B|C) \cdot p(A|B)$$


Generalizing for N-variables :

1. Determine ordering of var. (topological ordering)
2. In this ordering, call parents of $x_i$ $P_{a_i}$ : $(P_a(x_i))$

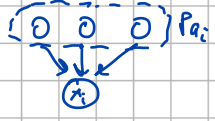
$$P(x_1, \dots x_m) = \prod_{i=1}^{m} p(x_i | P_{a_i})$$

$P_{a_i}$ does not have to be all, can be a subset.

→ [if no parent, take marginal: $p(x_i|\phi) = p(x_i)$]

$$\Sigma / \int p(x_i | P_{a_i}) = 1$$

$P_{a_i} \subseteq \{1, \dots i-1\}$

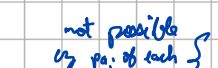



Directed graph. model or Bayesian netw.

$\forall j \in P_{a_i}$ $j \to i$ in graph (DAG)

acyclic coz can't have a cycle.

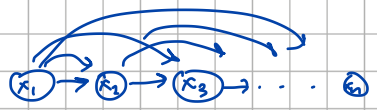
not possible coz $P_{a_i}$ of each var. is not earlier.



**Eg:** Markov chain $x_1 \to x_2 \to \cdots \to x_m$

$$p(x_1 \cdots x_m) = p(x_1) \prod_{i=2}^{m} p(x_i | x_{i-1}) \Rightarrow$$


**Eg:** Fully connected:




$$p(x_1 \cdots x_m) = \prod_{i=1}^{m} p(x_i | x_1 \cdots x_{i-1})$$

Assume $D_{x_i} = \{1, \ldots, K\}$ (but $\xi_k = 1$)

# params:

$$p(x_1) \prod_{i=2}^{m} p(x_i | x_{i-1})$$

$K-1$ + $(m-1)(K-1)K = O(K^2 \cdot m)$

# params

$$p(x_1 \cdots x_m) = \prod_{i=1}^{m} p(x_i | x_1 \cdots x_{i-1})$$

$O(K^m)$ ⇐ can be redundancy here

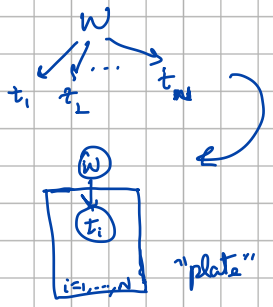
$p(x_1) = K - 1$
$p(x_2, x_2) = K^2 - 1$
$p(x_1 \cdots x_m) = K^m - 1$

} # params

**Eg:** 1) Draw param $w$
2) Draw $N$ pts $t_1 \cdots t_N$

$$p(w) \cdot \prod_{i=1}^{N} p(T_i | w)$$

prior likelih.

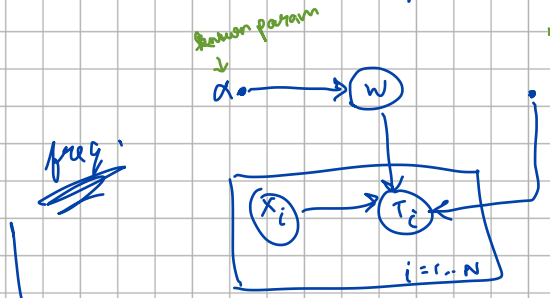



"plate" repetition of part inside plate indexed by $i$.

⇓

Add params, say i/p, $X = (x_1, \ldots, x_N)$, $\alpha$, $\sigma^2$

Learn $w$ from $T$, conditioning on $X$, for known $\alpha, \sigma^2$
(Regress. / classif.)



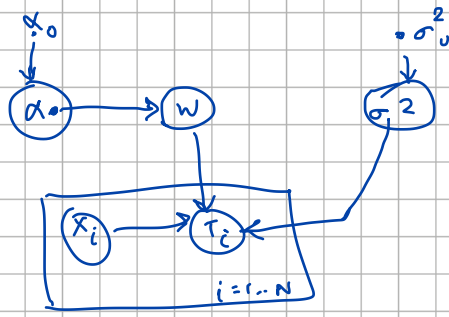

$$p(w | \alpha) \cdot \prod_{i=1}^{N} \left[ p(T_i | x_i, w, \sigma^2) \cdot p(x_i) \right]$$

↑ nodes in graph

Predictive distr.:

$$p(T^* | x^*, X, \alpha, \sigma^2)$$

* Add nodes for $T^*$ & $x^*$




Bayesian

$$p(\sigma^2 | \sigma_0^2) \cdot p(\alpha | \alpha_0) \cdot [\,,\,]$$

For ass./lab:

$X, Y$ indep. if $p(X,Y) = p(X) \cdot p(Y)$ $\quad \boxed{X \perp\!\!\!\perp Y}$

$X$ conditionally indep. of $Y$ given $Z$ if $p(X,Y|Z) = p(X|Z) \cdot p(Y|Z)$ $\quad \boxed{X \perp\!\!\!\perp Y | Z}$

Als valid for sets. ie, $\vec{X}, \vec{Y}, \vec{Z} \ldots$

## D-separation:

if $X_A$ is d-separated from $X_B$ & $X_C$ then,

$$p(X_A, X_B | X_C) = p(X_A|X_C)\, p(X_B|X_C)$$

ie, $X_A \perp\!\!\!\perp X_B | X_C$

How to check d-separ. holds (from graph):

1) Consider all paths (seq. of nodes, connected by edges) betw. node A & a node in B
   - ↳ - does not have to be all
   - ↳ such that no node repeats

2) A node is <u>blocked</u> by $X_C$ if:

   a) it contains a <u>collidor</u>: $\ldots \rightarrow u \leftarrow \ldots$
   such that $u$ is not <u>ancestor</u> of a node in $C$.

   ↳ transitive closure of parents

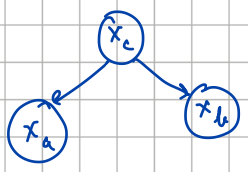
   $X \rightarrow \ldots \rightarrow V$
   H is ances. of V
   U is also ances. of H

   b) it contains a <u>non-collidor</u>

   $\ldots \rightarrow u$
   $\ldots \rightarrow u \rightarrow \ldots$
   $\ldots \leftarrow u \leftarrow \ldots$
   $\ldots \leftarrow u \rightarrow \ldots$
   $u \leftarrow \ldots$

   such that $U$ is in $C$ (when C is a set)
   else $u = C$

**Eg:**




A is not d-separ. from {B}
$A \leftarrow C \rightarrow B$
↑ non-collidor

{A} is d-sep. {B} given {C}

$X_A \not\perp\!\!\!\perp X_B$

$X_A \perp\!\!\!\perp X_B | X_C$

12th Sep

**Bayesian Netw.:** $\langle G, \{p(x_i \mid x_{Pa_G(i)})\} \rangle$

→ Prob. distr.: $p(x_1, \dots x_N) = \prod_{i=1}^{N} p(x_i \mid x_{Pa_G(i)})$

↓ d-separ.

$A \perp B \mid C$

↓ markov prop. ⇒ $x_A \perp\!\!\!\perp x_B \mid x_C$

↓ conditional / marginal independencies

$x_A \perp\!\!\!\perp$ (or $\not\perp\!\!\!\perp$) $x_B \mid x_C$

* dsep ⇒ CI
* no dsep ⇏ no CI
= dsep ⇐ CI (faithfulness)

Eg: deterministic relations ⇒ $x \to y \to z$

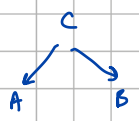
① $Y \perp z \mid x$

② $x \to y \to z$ where, $p(y \mid x) = \delta_{y,x}$ (Y is a copy of x)

$Y \not\perp z \mid x$ $\quad$ $Y \perp\!\!\!\perp z \mid x$

↓ faithfulness (holds in most cases)

Eg:




$A \perp B \mid C$

(A is d-sep. from B → cz C is a non-collider → cz no 2-way arrows & no other path)

∴ $x_A \perp\!\!\!\perp x_B \mid x_C$

$$P(x_A, x_B \mid x_C) = \frac{P(x_A, x_B, x_C)}{P(x_C)}$$

$$= \frac{P(C) \cdot P(A \mid C) \cdot P(B \mid C)}{P(C)}$$

(P(A|C) · P(B|C) = A & B indep. | C)

(22.) ★ $A \not\perp B \mid \phi$ in general $x_A \not\perp\!\!\!\perp x_B \mid x_C$

Eg: $A \to B \to C$

Here, $A \perp C \mid B$

∴ $A \perp\!\!\!\perp C \mid B$

$A \not\perp C$ so, not $P(A,C) = P(A)P(C)$

↑ if we don't condition on B

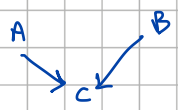
$$P(A, C \mid B) = \frac{P(A,B,C)}{P(B)}$$

$$= \frac{P(A)\,P(B \mid A)\,P(C \mid B)}{P(B)}$$

$$= P(A \mid B)\,P(C \mid B)$$

proved

Eg:




$A \perp B \mid \phi$ → C is not

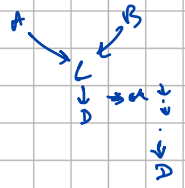
$A \not\perp B \mid C$

Typically $P(A,B|C) \neq P(A|C)\,P(B|C)$

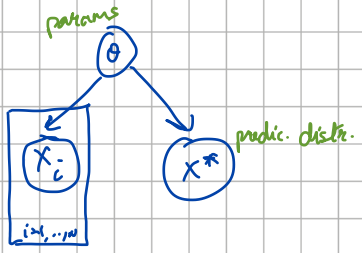
$$P(A,B) = \sum_C P(A,B,C)$$
$$= \sum_C P(A)\,P(B)\,P(C|A,B)$$
$$= P(A)P(B) \underbrace{\sum_C P(C|A,B)}_{=1}$$
$$= P(A)\,P(B)$$

∴ A indep. of B.

Eg.




$A \not\perp B \mid D$




$$p(\theta)\cdot p(x^*|\theta)\prod_{i=1}^{N} p(x_i|\theta)$$

$x^* \not\perp x_i$

typically, $x^* \perp\!\!\!\perp x_i$

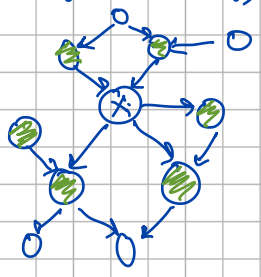
But $x^* \perp x_i \mid \theta$

$\forall x_1 \ldots x_N \quad x^* \perp\!\!\!\perp x_i \mid \theta$

(Once we learn θ & fix it ...

## Markov Blanket :

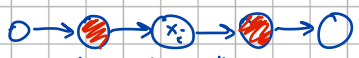
In Bayesian Netw., MB of node $x_i$ : $MB(x_i) = Pa_i \cup ch_i \cup (Pa_{ch_i} \setminus i)$




$$p(x_i | x_{MB_i}, x_{rest}) = p(x_i | x_{MB_i})$$

↳ useful during inference.
Need to know only $x_{MB_i}$ of $x_i$.
Not the rest.

Eg:




"Markov chains"

## Undirected Graph. Models (MRFs) :

$$MRF = \langle G, \{\psi_A\}_{A \in \text{cliques}(G)} \rangle \longrightarrow p(x_1 \ldots x_N) = \frac{1}{Z}\prod_A \psi_A(x_A)$$

$G$: undirected

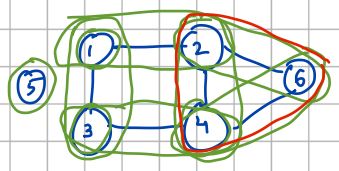
$\psi_A$: clique potential

$Z$: partition sum (normaliz. const.)

$x_A$: nodes appearing in clique.

where, a "clique" in undirn. 'G' (with nodes $\{1,..,N\}$) is a fully connected subset of nodes.

Eg:




— cliques

A clique is maximal if there is no clique that strictly contains it.

Here, max. cliques: $\{5\}$, $\{1,2\}$, $\{1,3\}$, $\{3,4\}$, $\{2,4,6\}$

$$P(x_1,..x_6) = \frac{1}{Z}\,\psi_5(x_5)\,\psi_{1,2}(x_1,x_2)\,\psi_{1,3}(x_1,x_3)\,\psi\cdots\psi_{2,4,6}(x_2,x_4,x_6)$$

where,

$$Z = \underbrace{\sum_{x_1\in D_{x_1}}\sum_{x_2\in D_{x_2}}\cdots\sum_{x_6}}_{\text{sum over domains}}\left(\psi\cdots\psi_{2,4,6}\right)$$

} thus, $\sum_{x_1}\cdots\sum_{x_6} P(x_1..x_6)=1$

★ calcul. of Z a bottleneck in MRF. → use approx.

Separation:

In undir. graph 2 subset of nodes A & B are "separated" given a third subset C, if each path betw. a node in A to a node in B passes through (at least) one node in C.

$A \perp B \mid C$

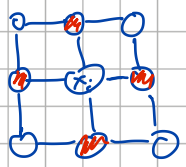
markov prop.: MRF $\langle G, \{\psi_v\}\rangle$ if $A \perp B \mid C$ in G, then $x_A \perp\!\!\!\perp x_B \mid x_C$ in

$$P(x_1\ldots x_n) = \frac{1}{Z}\prod_u \psi_u$$

Markov blanket: $MB(i) = nb_i$

↓

neighb. = nodes adj. to i / edge connec. with i (i – j).




Hammersley-Cliff. theor:

Give $p(\vec{x})>0$, then this can be represented as MRF.

Eg:


same cond. indep. as the directed one:

Earlier, $P(x_1)\cdot P(x_2|x_1)P(x_3|x_2)\cdots$

Here, $\frac{1}{Z}\,\psi_{1,2}(x_1,x_2)\,\psi_{2,3}(x_2,x_3)\cdots$

no big. prob. are not uniquely

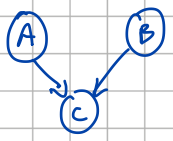
more flexible

$$\tilde{\psi}_{1,2}(x_1,x_2) = \psi_{1,2}(x_1,x_2)\cdot\psi(x_2)$$

given we write,

$$\tilde{\psi}_{2,3}(\ ) = \frac{\psi_{2,3}(x_2,x_3)}{\psi(x_2)}$$

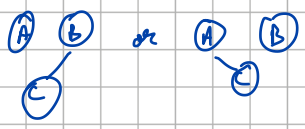
Eg:




$A \perp\!\!\!\perp B$

but typically, $A \not\perp\!\!\!\perp B \mid C$

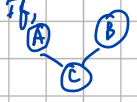
if we model this as MRF:




$A \perp\!\!\!\perp B \mid C$ | $A \perp\!\!\!\perp B \mid C$

no paths from A to B that pass through C (no such paths)

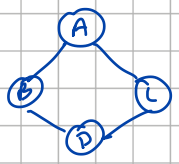
If,




$A \not\perp\!\!\!\perp B$

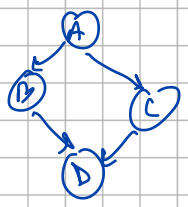
↳ depends on choice of $\psi$ to make it $A \perp\!\!\!\perp B$

Eg: MRF but not BN:



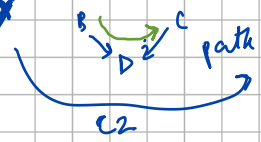

$A \perp D \mid \{B, C\}$
$B \perp C \mid \{A, D\}$

} can't find a BN satisfying this




$A \perp D \mid \{B, C\}$ ✓ (eg B & C are non-colliders)

$B \perp C \mid \{A, D\}$ ✗




path is open. Other is blocked

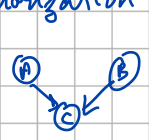
C2


No conditional indep.

## Factor Graphs :

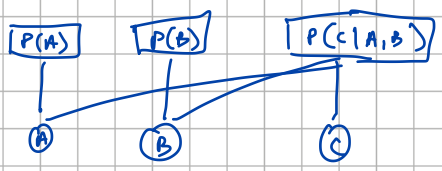
represent factorization struct. of a prob. dist.

BN:




$p(A, B, C) = p(A) \cdot p(B) \cdot p(C \mid A, B)$ → can be different factorizs. of this cond. prob. dists. Eg: $P(B \mid C, A)$

Factor graph:



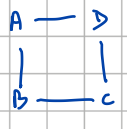

□ = factors
○ = variables

} bi-partite graph (i.e. 2 types of nodes)

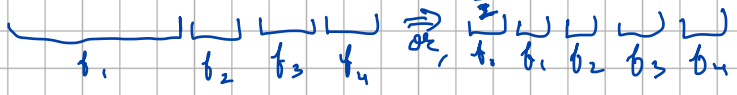
⇓

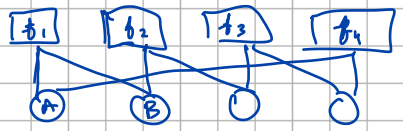
undirected edges betw. 2 node types only.

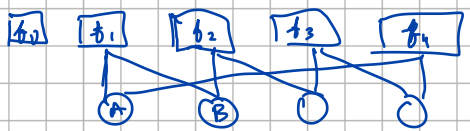
MRF:



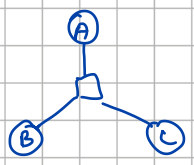

$p(A, B, C, D) = \frac{1}{Z} \cdot \psi_{A,B}(A, B) \cdot \psi_{BC} \cdot \psi_{CD} \cdot \psi_{DA}$



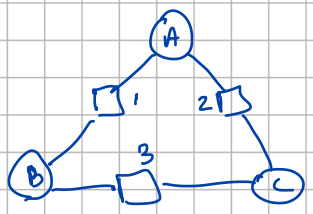



Eg:




$P(A,B,C) = f(A,B,C)$

[uninformative]




$P(A,B,C) = f_1(A,B) \cdot f_2(A,C) \cdot f_3(B,C)$

↳ Factor fyn.

same thing in MRF will be in terms of clique potential.

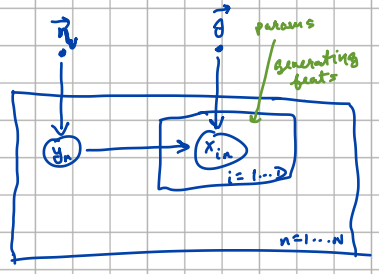
$$P(A,B,C) = \frac{1}{Z}\,\psi_{A,B,C}(A,B,C)$$

uniform. compared to factor graph. Less expressive.

## Naive Bayes Classif. :

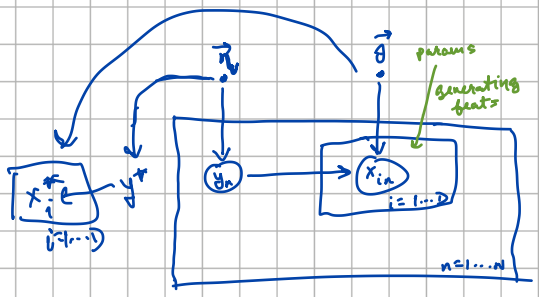
labels : $Y_N$

feats : $X_{1n} \cdots X_{Dn}$ $\quad n = 1 \cdots N$



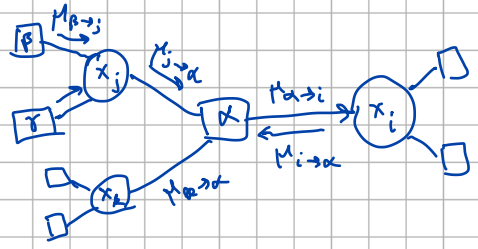

$$\Rightarrow \underbrace{\prod_{n=1}^{N}}_{\text{outer plate}} p(y_n \mid \vec{\pi}) \underbrace{\prod_{i=1}^{D}}_{\text{inner plate}} p(x_{in} \mid Y_n, \vec{\theta})$$

In N.B. $\quad X_{in} \perp\!\!\!\perp X_{in} \mid Y_n$



# 19th Sep

Bishop 8.4.4 **Sum product Algo / Belief propag.**




**Factor → variable messages :**

$$\mu_{\alpha\to i}(x_i) = \sum_{x_\alpha} f_\alpha(x_\alpha) \prod_{j\in \alpha\setminus i} \mu_{j\to\alpha}(x_j)$$

$x_\alpha = (x_i, x_j, x_k)$

$x_\alpha := (x_i)_{i \in \text{neigh}(\alpha)}$

**variable → factor messages :**

$$\mu_{j\to\alpha}(x_j) = \prod_{\beta\in ne(j)\setminus\alpha} \mu_{\beta\to j}(x_j)$$

#messages = 2 × no. of factors in graph.

**variable beliefs :**

$$p(x_i) = \sum_{\vec{x}\setminus i} p(\vec{x}) = \frac{1}{Z} \prod_{\alpha\in ne(i)} \mu_{\alpha\to i}(x_i) = \prod_{\alpha\in ne(i)} \mu_{\alpha\to i}(x_i)$$

$$Z = \sum_{x_i} \prod_{\alpha\in ne(i)} \mu_{\alpha\to i}(x_i)$$

either calcul. global normaliz. → or

$$p(x) = \frac{1}{Z} \prod_\alpha f_\alpha(x_\alpha)$$

**factor beliefs :**

$$p(x_\alpha) = \frac{1}{Z} f_\alpha(x_\alpha) \prod_{i\in ne(\alpha)} \mu_{i\to\alpha}(x_i)$$

**PRO :**

So from exp. in #nodes to → #edges

**complexity :** $O(2EK^M)$ where, $E$ = edges

$|D_{x_i}| \le K$ (domains upper bounded by K)

$M$ = max. variables in the factors = $|ne(i)| \le M$

(for eg, $\alpha$ here has 3. if no other has more than 3 then $M=3$)

**CONS :**

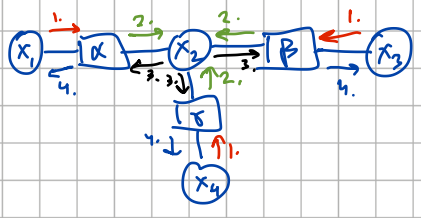
Loops of factors OK → can be absorbed

Loops of vars. not OK in this algo. (if loops then its a brute approxim.) → loopy belief propag.

if trees / forests (ie, no loops) = exact result

* alternate for exact inference : var. elimination.

**Eg :**




$$p(x_1, x_2, x_3, x_4) = \frac{1}{Z} f_\alpha(x_1, x_2)\, f_\beta(x_2, x_3)\, f_\gamma(x_2, x_4)$$

**Step 1 :** start at leaf node

$\mu_{1\to\alpha}(x_1) = 1$ (eg no neigh.)

$\mu_{3\to\beta}(x_3) = 1$

$\mu_{4\to\delta}(x_4) = 1$

Step 2 :

$$\mu_{\alpha\to 2}(x_2) = \sum_{x_1} f(x_1, x_2)\, \mu_{1\to\alpha}(x_1)$$

$$\mu_{\beta\to 2}(x_2) = \sum_{x_3} f_\beta(x_2, x_3)\, \mu_{3\to\beta}(x_3)$$

$$\mu_{\delta\to 2}(x_2) = \sum_{x_4} f_\delta(x_2, x_4)\, \mu_{4\to\delta}(x_4)$$

Incoming done $(x_2)$ ... Now start with outgoing :

Step 3 :

$$\mu_{2\to\alpha}(x_2) = \mu_{\beta\to 2}(x_2) \cdot \mu_{\delta\to 2}(x_2) \quad \text{(product of messages)}$$

$$\mu_{2\to\beta}(x_2) = \mu_{\alpha\to 2}(x_2) \cdot \mu_{\delta\to 2}(x_2)$$

$$\mu_{2\to\delta}(x_2) = \mu_{\alpha\to 2}(x_2) \cdot \mu_{\beta\to 2}(x_2)$$

---

$$\mu_{\alpha\to 1}(x_1) = \sum_{z} f_\alpha(x_1, x_2)\, \mu_{2\to\alpha}(x_2)$$

$$\mu_{\beta\to 3}(x_1) = \sum_{z} f_\beta(x_1, x_2)\, \mu_{2\to\beta}(x_2)$$

$$\mu_{\delta\to 4}(x_1) = \sum_{z} f_\delta(x_1, x_2)\, \mu_{2\to\delta}(x_2)$$

(checking)

$$p(x_2) = \frac{1}{Z}\, \mu_{\alpha\to 2}(x_2)\, \mu_{\beta\to 2}(x_2)\, \mu_{\delta\to 2}(x_2)$$

(substitute)

$$= \frac{1}{Z} \sum_{x_1} \sum_{x_3} \sum_{x_4} f_\alpha \cdot f_\beta \cdot f_\delta$$

$p(x_2 \mid x_3 = 5) = ?$

Introduce evidence factors :

Give F.G. $p(x_1 \dots x_d) = \frac{1}{Z} \prod_\alpha f_\alpha(x_\alpha)$

then, $p(x_i \mid x_j = \xi_j) = ?$

Evidence factor $f_{\xi_j}(x_j) = \delta_{x_j, \xi_j} = \begin{cases} 1, & x_j = \xi_j \\ 0, & x_j \neq \xi_j \end{cases}$

$$p(x_i \mid x_j = \xi_j) = \frac{p(x_i, x_j = \xi_j)}{\sum_{x_i} p(\dots)} \propto p(x_i, x_j = \xi_j) = \sum_{x_j} p(x_i, x_j)\, \delta_{x_j, \xi_j}$$

$$= \sum_{x_j} p(x_i, x_j) \cdot f_{\xi_j}$$

Soln : extend F.G. $\tilde{p}(x_1 \dots x_d) = \frac{1}{\tilde{Z}} \left[ \prod_\alpha f_\alpha(x_\alpha) \cdot f_{\xi_j}(x_j) \right]$

In our eg : extend the graph as : $\boxed{\beta}$ — $(x_3)$ — $\boxed{f_{x_3 = 5}}$

Run the same algo. again on this F.G.

(Ex P2)

(intuitively same as not summing over all $\sum_{x_3}$)

# Max-Sum algo :

want to calcul. $x^* = \arg\max_x \frac{1}{Z} \prod_a f_a(x_a)$ → state that gives max. prob.
drop cz const. not affects argmax.

[joint assignment of $x_i$, $x_j$ & $x_k$ (3 values) that maximize ]

$$p(x^*) = \frac{1}{Z} \max_x \prod_a f_a(x_a)$$

→ value of max. prob.

↓ in log domain

$$\log p(x^*) = -\log Z + \max_x \log \left[\prod_a f_a(x_a)\right]$$

$$= -\log Z + \max_x \sum_a \log f_a(x_a)$$

$a + \max(b,c) = \max\{a+b, a+c\}$

ie, replacing { prods. by sums & sums by max & factors → log factors

sum-prod. algo → max-sum algo. (helps us to)

## Factor → variable messages :

$$\mu_{a\to i}(x_i) = \sum_{x_a | x_i \in D_{a|i}} f_a(x_a) \prod \mu_{j\to a}(x_j)$$

↓

$$\nu_{a\to i}(x_i) = \max_{x_{a|i}} \log f_a + \sum_{j\in a|i} \nu_{j\to a}(x_j)$$

Illy,

$$\nu_{j\to a}(x_j) = \sum_{p\in ne(j)|a} \nu_{p\to j}(x_j)$$

max beliefs / max marginals :

$$q(x_i) = \sum_{a\in ne(i)} \nu_{a\to i}(x_i) - \log Z$$

not a prob. its max-belief. (a func. of $x_i$)

↓ used for

Given max-marginals $\{q_i(x_i)\}_{i=(1,\dots d)}$, run Viterbi algo. (Bishop 8.4.5) to get global optimum $x^*$

$$x^* = \arg\max_x \prod_a f_a(x_a)$$

if $q_i(x_i)$ has unique max. then can use : $x_i^* = \arg\max_{x_i} q_i(x_i)$ (don't have to run Viterbi)

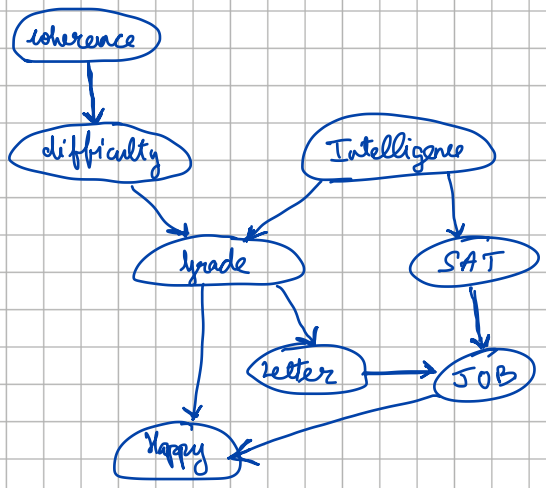
For eg : $|D_{x_i}| = 5$ : $q_i(x_i) = (6, 8, -1, 2, \pi)$ ($x_i = 1, 2, 3, 4, 5$)

$\Rightarrow x^* = 2$

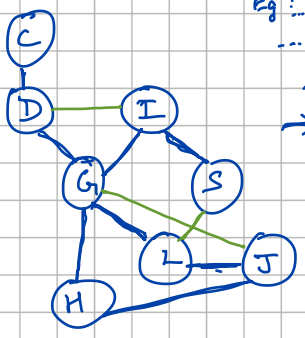
# Variable elimin. algo : (9.8.4.6 exact inference on general graphs)

Ch-20, Murphy for var. elimin. & tree junc. algo.

B.N. : (not F.G. but M.R.F)




Step 1: convert B.N to M.R.F → make all edges undirected
→ marry parents (moralig.)
↓ done eg




Eg : $P(G|D,I)$ .....
..... $\psi_{DG}(D,G)\,\psi_{GI}(G,I)$ .....
slight: less expressive
→ if marry D & I
.... $\psi_{DGI}(D,G,I)$ ....
max slight

Step 2 : $P(J) = \sum_L \sum_S \sum_G \sum_H \sum_I \sum_D \sum_C p(L,S,G,H,I,D,C)$

$= \sum_L \sum_S \sum_G \sum_H \sum_I \sum_D \sum_C \underbrace{\psi_C(C)\,\psi_D(C,D)}_{\text{absorb}\ \tilde{\psi}_D(C,D)} \underbrace{\psi_I(I)\,\psi_G(G,I,D)}_{\tilde{\psi}_G(G,I,D)} \psi_S(S,I)\,\psi_L(L,G)\,\psi_J(J,L,S)\,\psi_H(H,G,J)$

using distributive law:

Choose an order for elimin. : C, D, I, H, G, S, L } put them to right as far as possible

$= \sum_L \sum_S \psi_J(J,L,S) \sum_G \psi_L(L,G) \sum_H \psi_H(H,G,J) \sum_I \psi_S(S,I) \sum_D \tilde{\psi}_G(G,I,D) \sum_C \tilde{\psi}_C(C,D)$

→ only depends on D as summed over C.

$\tau_1(D)$

$\tau_2(G,I)$

$\tau_3(G,S)$

$\tau_4(G,S,J)$

$\tau_5(L,S,J)$
$\tau_6(L,J)$
$\tau_7(J)$
↳ what we wanted

| Step | Elim. | Factors Used | Var. induced | New factor |
|---|---|---|---|---|
| 1 | C | $\tilde{\psi}_C = \psi_C \cdot \psi_D$ | C, D | $\tau_1(D)$ |
| 2 | D | $\tau_1, \tilde{\psi}_G = \psi_G \cdot \psi_I$ | G, I, D | $\tau_2(G,I)$ |
| 3 | I | $\tau_2, \psi_S$ | | |

$P(J) \propto \tau_7(J)$
$= \frac{\tau_7(J)}{Z=1}$

# 26th Sep

## VAE's :

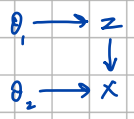
**Model** : $p(x, z|\theta) = p(x|z,\theta_2) \cdot p(z|\theta_1)$ (params)

$x$ : observed, $z$ : latent

$p(x|z,\theta_2)$ : decoder (NN)

$p(z|\theta_1)$ : prior (gauss.) } less indep. ⇝ non-linear PCA.

[if student's t ⇝ non-linear ICA (more indep. enforced)]




**Aim** : find lower dim representation. (encode)

Use variational inference. Hence have approx. posterior $\underbrace{q(z|x,\lambda)}_{\text{encoder}} \approx p(z|x,\theta_1,\theta_2)$ (NN)

Thus, we get an exact repres. of decoder but only an approx. of encoder.

**Log-likel.** : $L(\theta) = \ln p(x|\theta) = \ln \int p(x|z,\theta)\, p(z|\theta)\, dz$

hard to approx. ∴ var. infer.

**ELBO** : $L(\theta,\lambda) = \mathbb{E}_{q(z|x,\lambda)}\left[\ln p(x|z,\theta_2) + \ln p(z|\theta_1)\right] + H(q(z|x,\lambda)) \leq L(\theta)$

diffic. integral so approx. using sampling.

so optimize this ⇝ Hard to optimize

optimize $\max_{\theta,\lambda} L(\theta,\lambda)$ using SGD

what we actually optimize is $\max_{\theta,\lambda} \hat{L}(\theta,\lambda) = \frac{1}{K}\sum_{k=1}^{K}\left[\ln p(x|z^k,\theta_2) + \ln p(z^k|\theta_1)\right]$

where, $z^k \sim q(z|x,\lambda)$ ($k$ times)

**Use** : $\mathbb{E}_{q(z|x,\lambda)} f(z) \approx \frac{1}{K}\sum_{k=1}^{K} f(z^k)$ $\quad z^k \sim q(z|x,\lambda)$ (fix $x$, sample)

like change of variable: reparam: $z = g_\lambda(x,\varepsilon)$, $\varepsilon \sim p(\varepsilon)$

$\approx \frac{1}{K}\sum_{k=1}^{K} f\left(g_\lambda(x,\varepsilon^k)\right)$, $\varepsilon^k \sim p(\varepsilon)$

**decoder** : $p(x|z,\theta_2) = \mathcal{N}\left(x \mid \mu_{\theta_2}(z), \Sigma_{\theta_2}(z)\right)$

where, $\mu$ & $\Sigma$ are o/p of MLP with i/p = $z$ & wts. $\theta_2$.

**encoder** :

The approx. also assumed gaussian :

$q(z|x,\lambda) = \mathcal{N}\left(z \mid \mu_\lambda(x), \Sigma_\lambda(x)\right)$

where, $\mu$ & $\Sigma$ are o/p of a MLP with i/p $x$ & wts $\lambda$.

$g_\lambda(x,\varepsilon) = \mu_\lambda(x) + \sigma_\lambda(x) \odot \varepsilon$ ($\varepsilon \sim \mathcal{N}(0,I)$, fixed (scalar))

$\sigma_\lambda(x)$ moves to sample (multi-variate)

also gives less variance in grads. so faster convergence

$\begin{bmatrix} \sigma^2_{\lambda,1}(x) & \cdots & 0 \\ \vdots & & \\ 0 & \cdots & \sigma^2_{\lambda,d}(x) \end{bmatrix}$

## Training procedure :

- sample minibatch $(x_1 \cdots x_m)$ from data
- for $m = 1, \dots, M$ sample $z_m^k \sim q(z | x^m, \lambda^t)$ (use reparam. Trick)

  actually, $z_m^k = g_{\lambda^t}(x^m, \epsilon^k)$, $\epsilon^k \sim p(\epsilon)$ ($K$ samples)
- compute $\hat{L}(\theta^t, \lambda^t)$
- compute $\nabla_{\theta,\lambda} \hat{L}(\theta^t, \lambda^t)$ on all $x_m, z_m^k$ $\quad m = 1 \dots M$, $k = 1 \dots K$
- SGD update:

$$\theta^{t+1} \leftarrow \theta^t + \alpha \nabla_\theta \hat{L}(\theta^t, \lambda^t)$$

$$\lambda^{t+1} \leftarrow \lambda^t + \alpha \cdot \nabla_\lambda \hat{L}(\theta^t, \lambda^t)$$

# 30th Sept

## Sampling Methods :

### Monte-Carlo :

goal : compute $\mathbb{E}_{p(z)}[f(z)] = \int f(z) p(z) dz \Big/ \int f(z) p(z) dz$

applic. :

- prediction : $p(y^* | x^*) = \int p(y^* | x^*, \theta) p(\theta | x, y) d\theta$
- Estim. evidence : $p(y | x) =$

### ① Regular sampling :

( $\hat{}$ or $< >$ means estim. this value )

1) draw $N$ samples from $p(z)$ $\quad z_i \sim p(z) \quad i = 1, \dots, N$

2) Calcul. $\mathbb{E}[f] \approx \hat{f} = \frac{1}{N} \sum_{i=1}^{N} f(z_i)$ $\quad$ as $N \to \infty$ $\hat{f} \to \mathbb{E}[f]$

properties :

→ Unbiased estimator : $\mathbb{E}[\hat{f}] = \mathbb{E}[f]$

$$\mathbb{E}[\hat{f}] = \mathbb{E}\left[\frac{1}{N} \sum_i f(z_i)\right] = \frac{1}{N} \sum_i \mathbb{E}[f(z_i)] = \mathbb{E}[f]$$

→ Variance :

$$\mathrm{var}[\hat{f}] = \mathbb{E}\left[(\hat{f} - \mathbb{E}[\hat{f}])^2\right]$$

$$= \mathrm{Var}\left[\frac{1}{N} \sum_i f(z_i)\right]$$

$$= \frac{1}{N^2} \sum_i \mathrm{var}\left[f(z_i)\right]$$

$$= \frac{1}{N^2} N \sum_i \mathrm{var}\left[f(z)\right]$$

$$= \frac{1}{N} \mathrm{var}\left[f(z)\right]$$

more samples, less var.

( $\mathrm{var}(a+b) = \mathrm{var}(a) + \mathrm{var}(b)$ if $a, b$ indep. Here we have indep. samples. )

( $\mathrm{var}(\lambda x) = \lambda^2 \mathrm{var}(x)$ )

$\Longrightarrow$ RMS error $= \sqrt{\mathrm{var}[\hat{f}]} = \frac{1}{\sqrt{N}} \sqrt{\mathrm{var}[f(z)]} \Big\} \downarrow 0$ as $N \to \infty$

## Discrete Random Variables Sampling :

Say $z \in \{1, 2, \dots K\}$ . Given $p(z)$, how to sample from $p(z)$ ?

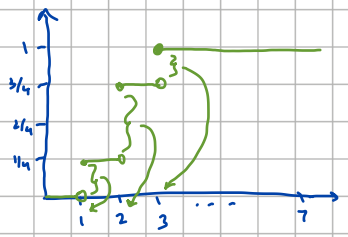
**Steps :**

1) Calcul. CDF/CMF

$K = 3$

$p(z=1) = 1/4$
$p(z=2) = 2/4$
$p(z=3) = 1/4$

$p(z \le 0) = 0$
$p(z \le 1) = 1/4$
$p(z \le 2) = 3/4$
$p(z \le 3) = 1$
$p(z \le 4) = 1$




2) Sample from uniform $u_i \sim U(0,1)$

Eg : $u = 0.35 \rightarrow z_i = 2$

$0 \le u_i \le 1/4 \quad z_i = 1$
$1/4 \le u_i \le 3/4 \quad z_i = 2$
$3/4 \le u_i \le 1 \quad z_i = 3$

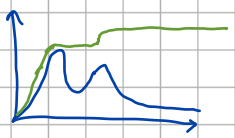
3) Look up corresp. value of $z_i$

## Contin. random sampling :

variables in $\mathbb{R}$ with density $p(z)$

1) Calcul. CDF $p(z \le \varepsilon) = \int_{-\infty}^{\varepsilon} p(z)\, dz = F(\varepsilon)$ $\quad F : \mathbb{R} \to [0,1]$

2) Calcul. $F^{-1}$ $\quad F^{-1} : [0,1] \to \mathbb{R}$
3) sample $u_i \sim U(0,1)$
4) $z_i \sim F^{-1}(u_i)$



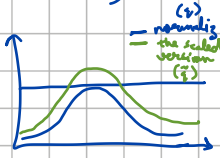
## ② Rejection Sampling

Above was for low dimen. If high dimens. than $\int$ can't calcul.

Given : unnormalized density $\tilde{p}(z) \propto p(z)$

1) Find proposal distrib. $\tilde{q}$ (have unnormalized density $\tilde{q}(z) \propto q(z)$ that we can sample from such that :)
   - $\rightarrow \int \tilde{q}(z)\, dz < \infty$
   - $\rightarrow \tilde{p}(z) \le \tilde{q}(z) \quad \forall z$

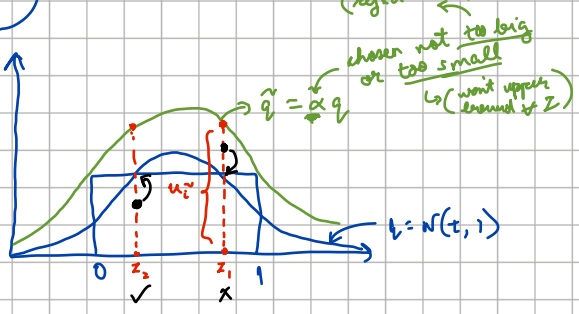
Eg. : Gaussian




2) For $i = 1, \dots, N$ :
   - $\rightarrow$ sample $z_i \sim q$ (or $\tilde{q}$)
   - $\rightarrow$ sample $u_i \sim U(0, \tilde{q}(z_i))$ (has to be $\tilde{q}$)
   - $\rightarrow$ if $\tilde{p}(z) > u_i$ keep sample $z_i$
     else throw away $z_i$ & re-sample

3) $q(z_i) \cdot \frac{\tilde{p}(z_i)}{\tilde{q}(z_i)} \propto p(z_i)$

★ in high dimens., prob. to reject samples is high.



## (3) Importance Sampling

In above we have to find the upper bound. Here, not.

Need: – unnorm. density $\tilde{p}(z)$ to sample from
– unnorm. proposal density $\tilde{q}(z)$ that we can sample from.
↳ (does not need to upper bound $\tilde{p}$)

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

for $i = 1, \dots, N$:

1) Sample $z_i \sim q$
2) Calcul. weight $\frac{\tilde{p}(z_i)}{\tilde{q}(z_i)} = w_i$ } if $\tilde{p} \gg \tilde{q}$ (high var.) Ex: $\tilde{q} \approx 0$ but $\tilde{p} \gg 0 \Rightarrow$ dividing by small no.
3) Calcul. $\mathbb{E}[f] = \frac{\sum_i w_i f(z_i)}{\sum_i w_i}$ (in ①, $w_i = 1$)

$$\mathbb{E}[f] = \int p(z) f(z) dz = \frac{\int q(z) \frac{p(z)}{q(z)} f(z) dz}{\int q(z) \frac{p(z)}{q(z)} dz} = \frac{\int q(z) \frac{\tilde{p}(z)}{\tilde{q}(z)} \cdot \frac{Z_q}{Z_p} f(z) dz}{\int q(z) \frac{\tilde{p}(z)}{\tilde{q}(z)} \frac{Z_q}{Z_p} dz} \simeq \frac{\frac{1}{N} \sum_i f(z_i) w_i}{\frac{1}{N} \sum_i w_i}$$

(denominator $= 1$; $\because$ unnormalized densities can cancel these off)

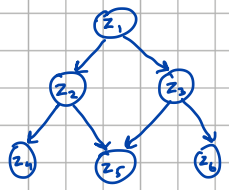
∎

- ★ Imp. sampling faster than rej. (no throwing away of samples, use all)
- ★ But imp. samp. has higher var. than rej. samp.
- ★ Imp. sampling also not so good in high dimens. (diffic. to find dens. $q$ s.t both $p$ & $q$ are large together)

## (4) Ancestral Sampling

Given Bayesian NN, if want to sample from it $p(z) = \prod_{i=1}^{N} p(z_i | pa(z_i))$, use topological ordering $z_1 \dots z_d$ (ie, $z_j < z_i$ if $j \in pa(i)$)

For $k = 1, \dots, d$ (# variables)
1) Sample $z_k \sim p(z_k | z_{pa(k)})$




$\tilde{z}_1 \sim p(z_1)$
$\tilde{z}_2 \sim p(z_2 | z_1 = \tilde{z}_1)$
$\tilde{z}_3 \sim p(z_3 | \tilde{z}_1)$
$\tilde{z}_4 \sim p(z_4 | \tilde{z}_2)$
⋮

## (5) MCMC Sampling


(typically 'T' same at every time step)

Given target distr. $p(x)$ that we want to sample from, setup a Markov chain s.t. $p(x_N) \to p(x)$ as $N \to \infty$ ie, s.t. $p(x)$ is its equilibrium distrib.

1) $x_1 \sim q(x_1)$ ($q$ = proposal distrib.)
2) $x_2 \sim T(x_2 | x_1)$ ← Transition kernel T or trans. prob. $T(x_2 | x_1)$
3) $q(x_1, x_2) = q(x_1) T(x_2 | x_1)$
   $q(x_2) = \int q(x_1) T(x_2 | x_1) dx_1$
   $x_i \sim T(x_i | x_{i-1})$

4) $q(x_1, \dots x_N) = q(x_1) \prod_{i=2}^{N} q(x_i | x_{i-1})$

$= q(x_1) \prod_{i=2}^{N} T(x_i | x_{i-1})$

Suppose after convergence: if $x_N \sim p(x)$ then like to have $x_{N+1} \sim p(x)$ ("invariance")

$p(x_{N+1}) = \int T(x_{N+1} | x_N) p(x_N) \, dx_N$

this gives a constraint on $T$.

eigen value eq. with $\lambda = 1$

<u>Detailed balance implies invariance</u>:

$T(x_{t+1} | x_t) p(x_t) = T(x_t | x_{t+1}) p(x_{t+1}) \quad \forall x_{t+1}, x_t$

[$T(y|x) p(x) = T(x|y) p(y) \quad \forall x, y$]

↳ detailed balance property (in equilib. state)

$\int T(x_{N+1} | x_N) p(x_N) = \int T(x_N | x_{N+1}) p(x_{N+1}) \, dx_N = 1 \cdot p(x_{N+1})$

proved

One can show: given Markov chain with initial distrib. $q$ & trans. kernel $T$ { $p$ is invariant; ergodicity holds ($T(y|x) > 0 \; \forall x, y$) }

then $p$ is unique equilib. distrib. of Markov chain.

3rd Oct

<u>MCMC sampling</u>: either in high dim. as can string along many $Q(x_{t+1} | x_t)$ to estim. target distr. without relying too much on its unknow. to come up with 1 single good proposal distr.

<u>Detailed balance</u>: $p^*(x_{t+1}) = \int T(x_{t+1} | x_t) p^*(x_t) \, dx_t \quad \forall x_{t+1}$

$\Downarrow$ or $\Sigma$

or,

$p^*(x_t) T(x_{t+1} | x_t) = p^*(x_{t+1}) T(x_t | x_{t+1})$

<u>Valid kernels</u>:

If $T_1$ & $T_2$ valid:

1) $\alpha T_1 + (1-\alpha) T_2 \quad (\alpha \in [0,1])$ ⇒ with prob. $\alpha$ choose $T_1$ at 't' & $1-\alpha$ choose $T_2$.
2) $T_2 \circ T_1$ (composition)

ie, $T_3 = T_2 \circ T_1$ : $T_3(x_{t+1} | x_t) = \int T_2(x_{t+1} | x_t) T_1(x' | x_t) \, dx'$ ($x'$ → $x_{t+1/2}$)

The <u>MCMC sampler</u> uses those transition prob.:

choose state $x_0$ from some $p(x_0)$

for $t = 0 \dots N$:

sample $x_{t+1} \sim T(x_{t+1} | x_t)$

output $x_N$

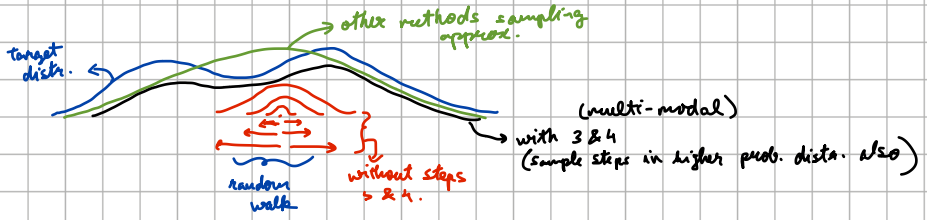
→ How do we sample from an arbitrary kernel / distr. $T$ ?

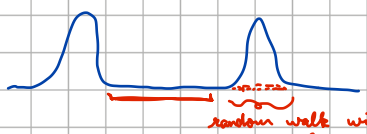
## Metropolis Hastings algo : (inner loop of MCMC sampler)

given a proposal trans. kernel $Q(x_{t+1}|x_t)$ (something that we can sample from)

To sample $x_{t+1} \sim T(\ )$ :

1) sample $\tilde{x}_{t+1} \sim Q(\ )$
2) compute acceptance prob. $\alpha(\tilde{x}_{t+1}|x_t) = \min\left(1, \frac{p(\tilde{x}_{t+1})\, Q(x_t|\tilde{x}_{t+1})}{p(x_t)\, Q(\tilde{x}_{t+1}|x_t)}\right)$
3) sample $u_t \sim U(0,1)$
4) If $u \le \alpha$ : accept the sample $\rightarrow x_{t+1} = \tilde{x}_{t+1}$
   else : reject it but don't throw it away, use it as new state : $x_{t+1} = x_t$




} combine random + walk with another algo that can take longer steps.

### Sketch that is satisfies detailed balance :

★ if a new sample is accepted - $(x_{t+1} = \tilde{x}_{t+1})$

$p^*(x_t)\, Q(x_{t+1}|x_t) \min(\ )$
- $p^*(x_t)$ : start in $x_t$
- $Q(x_{t+1}|x_t)$ : transition to $x_{t+1}$ with this prob.
- $\min(\ )$ : happens with prob. $\alpha$

$= \min\left(p^*(x_t)\, Q(x_{t+1}|x_t),\ p^*(x_{t+1})\, Q(x_t|x_{t+1})\right)$

$= \min\left(p^*(x_t)\, Q(x_t|x_{t+1}),\ p^*(x_t)\, Q(x_{t+1}|x_t)\right)$ ↔ $x_t \leftrightarrow x_{t+1}$ to prove detail bal.

$= p^*(x_{t+1})\, Q(x_t|x_{t+1}) \min\left(1, \frac{p^*(x_t)\ldots}{\ldots}\right)$

proved

## (6) Gibbs Sampler (good in high dim.)

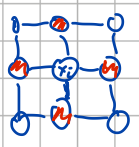
D-dimen. : $\vec{x}^t = (x_1^t \ldots x_D^t)$ ($p^*$ = target distr.)

1 step in MC → repeat =

$$\begin{cases} x_1^{t+1} \sim p^*(x_1 | x_2^t \ldots x_D^t) \\ x_2^{t+1} \sim p^*(x_2 | x_1^{t+1}, x_3^t \ldots x_D^t) \\ x_3^{t+1} \sim p^*(x_3 | x_1^{t+1}, x_2^{t+1}, \ldots x_D^t) \\ \vdots \\ x_D^{t+1} \sim p^*(x_D | x_1^{t+1} \ldots x_{D-1}^{t+1}) \end{cases}$$

★ can change ordering. If ordering is fixed detailed balance does not hold.

★ Here we are sampling from conditionals (not joint target distr.) & this sometimes possible. Eg :



$p(x_i | x_1 \ldots x_{i-1}, x_{i+1} \ldots x_D)$

⇓ still tough but can reduce this to sampling from just markov blanket

$p(x_i | \mathcal{E}_m)$

NOTE : $x \sim \mathcal{N}(\mu, \sigma^2)$ (distrib.), $p(x) = \mathcal{N}(x|\mu, \sigma^2)$ (density)

↳ not $p(x) = \mathcal{N}(\mu, \sigma^2)$ (func. of $x$ but no $x$ in the RHS)

$X \sim p(x)$, $Y \sim q(y)$

$p(x+y) \neq p(x) + p(y)$

↳ $= p(x) * q(y)$

$= \int p(t)\, q(y + x - t)\, dt$

# 10th Oct

## EM for HMM

1. Calcul. initial values $\theta^{old}$ with $\theta = (\pi, A, \phi)$
2. Iterate until convergence:

**E step**: (calcul. $Q(\theta, \theta^{old})$)

Baum-Welch algo (message passing step):
1) Run forw. $\alpha$ recursion to calcul. $\alpha(z_1), \dots \alpha(z_N)$
2) Run backw. $\beta$ recursion ... $\beta(z_N) \dots \beta(z_1)$
3) Calcul. suff. stats $\gamma(z_n)$, $\xi(z_{n-1}, z_n)$ $n = 1, \dots, N$ $p(X|\theta^{old})$

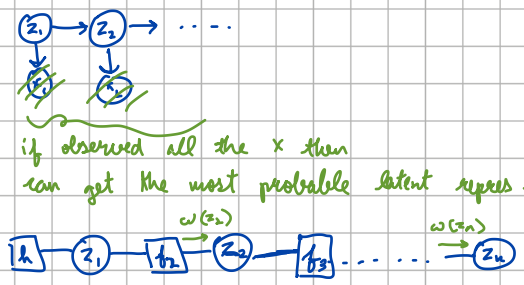
**M-step**: Calcul. $\theta^{new} = \arg\max_\theta Q(\theta, \theta^{old})$

$\pi^{new}$ : (13.18)
$A^{new}$ : (13.19)

categ. emissions : (13.20, 13.21)
Gauss. emissions : (13.23)
other emissions : $Q(\theta, \theta^{old}) = \dots$

## Viterbi algo (max-product for HMMs)

Calcul. most prob. seq. of latent states given X.




$h(z_1) = p(z_1|\pi)\, p(x_1|z_1, \phi)$
$f_n(z_{n-1}, z_n) = p(z_n|z_{n-1}, A)\, p(x_n|z_n, \phi)$

def $\omega(z_n) = \nu_{f_n \to z_n}(z_n)$ $2 \le n \le N$ } do forw. pass ($\omega$ messages) but no backw. pass. Instead do dynamic programming.

$\omega(z_1) = \ln p(z_1) + \ln p(x_1|z_1)$
for $n = 1 \dots N-1$
$\omega(z_{n+1}) = \ln p(x_{n+1}|z_{n+1}) + \max_{z_n} [\ln p(z_{n+1}|z_n) + \omega(z_n)]$

$f_n$ ; the $z_n$ that gave this max ; incoming message

max-prod:
$\dots\dots\to \psi_n(z_{n+1}) = \arg\max_{z_n} [\ln p(z_{n+1}|z_n) + \omega(z_n)]$ } keep track of this.

end
$z_n^{max} = \arg\max_{z_n} \omega(z_n)$

(reason backw.) dynam. prog.
for $n = N-1 \dots 1$
$z_n^{max} = \psi_n(z_{n+1}^{max})$

## Linear Dynamical Sys. (13.3)

Same as HMM's but var. continuous & prob. = Gaussian.
(same graph)

Now: $\{x_n, z_n\}$ are continuous $x_n \in \mathbb{R}^{d_1}$, $z_n \in \mathbb{R}^{d_2}$
Linear Gauss. model: all condi. distr. in Bay. Netw. are Gaussian with means that depend linearly on parents.

---

Viterbi: for Gauss. mean = median = MAP (mean is the most probable)
// Am.

→ no viterbi necessary.

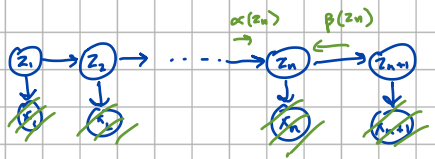
---

Forw. passing message eqs: Kalman filter eqs.
Backw. " " " : Kalman smoother eqs.

Transition: $p(z_n|z_{n-1}) = \mathcal{N}(z_n | Az_{n-1}, \Gamma)$ ← trans. uncertainty
Emission: $p(x_n|z_n) = \mathcal{N}(x_n | Cz_n, \Sigma)$ ← observ. noise
$\Gamma, \Sigma$ } const. in time } ⇒ mean shifts & covar. gets larger & larger.
$C$ = I if good recording instrument

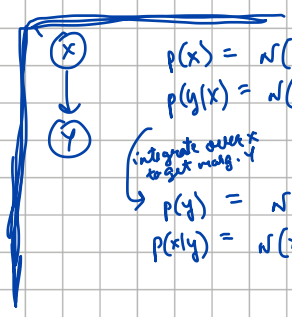
Initial State: $p(z_1) = \mathcal{N}(z_1 | \mu_0, V_0)$
Params: $\theta = (A, \Gamma, C, \Sigma, \mu_0, V_0)$

## Inference in LDS (for E-step) (13.3.1) (prob. of latent given observed)



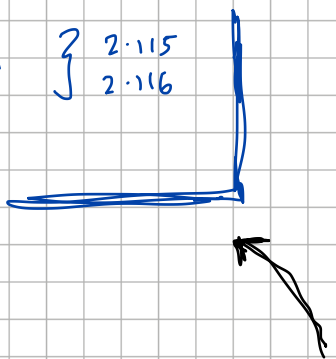

Forw. eqs.: $\hat{\alpha}(z_n) = \mathcal{N}(z_n | \mu_n, V_n)$
not normalized (messages)
update these in message update step.

USE TRICKS




$p(x) = \mathcal{N}(x | \mu, \Lambda^{-1})$
$p(y|x) = \mathcal{N}(y | Ax+b, L^{-1})$
integrate over x to get marg. ↓
$p(y) = \mathcal{N}(y | A\mu+b, L^{-1} + A\Lambda^{-1}A^T)$
$p(x|y) = \mathcal{N}(x | \Sigma(A^T L(y-b) + \Lambda\mu), \Sigma)$, $\Sigma = (\Lambda + A^T L A)^{-1}$ } 2.115 2.116

Basic idea: $p(x)p(y|x) = p(y)p(x|y)$



## Forw. eqs:

$$\hat{\alpha}(z_n) = \frac{1}{c_n} p(x_n|z_n) \int \hat{\alpha}(z_{n-1}) p(z_n|z_{n-1}) \, dz_{n-1}$$

$c_n$: normal.

$$c_n \cdot \underbrace{\mathcal{N}(z_n|\mu_n, V_n)}_{p(y)} = \underbrace{\mathcal{N}(x_n|Cz_n, \Sigma)}_{p(y|x)} \int \underbrace{\mathcal{N}(z_{n-1}|\mu_{n-1}, V_{n-1})}_{p(x)} \mathcal{N}(z_n|Az_{n-1}, \Gamma) \, dz_{n-1}$$

$p(x|y)$ with $y = x_n$ → not normalized

$$\int \mathcal{N}(z_{n-1}|\mu_{n-1}, V_{n-1}) \mathcal{N}(z_n|Az_{n-1}, \Gamma) \, dz_{n-1} = \mathcal{N}(z_n|A\mu_{n-1}, \underbrace{\Gamma + AV_{n-1}A^T}_{P_{n-1}})$$

now just product of 2 Gauss.

Translation Table:

| | |
|---|---|
| $x$ | $z_n$ |
| $y$ | $x_n$ |
| $\mu$ | $A\mu_{n-1}$ |
| $\Lambda^{-1}$ | $P_{n-1}$ |
| $A$ | $C$ |
| $b$ | $0$ |
| $L^{-1}$ | $\Sigma$ |
| $\Sigma$ | $(P_{n-1}^{-1} + C^T\Sigma^{-1}C)^{-1}$ |

$$c_n = \mathcal{N}(x_n | CA\mu_{n-1}, \Sigma + CP_{n-1}C^T)$$
$$\mu_n = (P_{n-1}^{-1} + C^T\Sigma^{-1}C)^{-1}(C^T\Sigma^{-1}x_n + P_{n-1}^{-1}A\mu_{n-1})$$
$$V_n = (P_{n-1}^{-1} + C^T\Sigma^{-1}C)^{-1}$$

Apply Bishop C.7 (Woodbury identity)

$$(A + BD^{-1}C)^{-1} = A^{-1} - A^{-1}B(D + CA^{-1}B)^{-1}CA^{-1}$$

less inverses

$$V_n = P_{n-1} - \underbrace{P_{n-1}C^T(\Sigma + CP_{n-1}C^T)^{-1}}_{K_n}CP_{n-1} = (I - K_nC)P_{n-1}$$

($K_n$: Kalman gain matrix)

Apply Bishop C.5:

$$(P^{-1} + B^TR^{-1}B)^{-1}B^TR^{-1} = PB^T(BPB^T + R)^{-1}$$

$$\therefore \mu_n = K_nx_n + (I - K_nC)A\mu_{n-1}$$
$$= A\mu_{n-1} + K_n(x_n - CA\mu_{n-1})$$

new observ. − predicted observ.

## Backw. eqs:

not in $\hat{\beta}$.

formulated in terms of variable beliefs (prod. of incoming msgs)

$$\gamma(z_n) = \hat{\alpha}(z_n)\hat{\beta}(z_n) = \mathcal{N}(z_n|\hat{\mu}_n, \hat{V}_n)$$

factor beliefs

$$\xi(z_{n-1}, z_n) = \mathcal{N}\left(\begin{pmatrix} z_{n-1} \\ z_n \end{pmatrix} \middle| \begin{pmatrix} \hat{\mu}_{n-1} \\ \hat{\mu}_n \end{pmatrix} \begin{pmatrix} \hat{V}_{n-1} & J_{n-1}\hat{V}_n \\ \hat{V}_nJ_{n-1}^T & \hat{V}_n \end{pmatrix}\right)$$

symm. w.r.t. $^T$

} need for EM

## Learning in LDS using EM: (13.3.2)

complete data log-likelihood:

$$\ln p(x, z|\theta) = \ln(z_1|\mu_0, V_0) + \sum_{n=2}^{N} \ln p(z_n|z_{n-1}, A, \Gamma) + \sum_{n=1}^{N} p(x_n|z_n, C, \Sigma)$$

$$Q(\theta, \theta^{old}) = \mathbb{E}_{z|\theta^{old}}\left[\ln p(x, z|\theta)\right]$$

The inference results: $\mathbb{E}[z_n|\theta^{old}] = \hat{\mu}_n$ (mean of $\gamma(z_n)$)

(13.105 – 13.107 skipped)

$$\mathbb{E}[z_n z_{n-1}^T|\theta^{old}] = J_{n-1}\hat{V}_n + \hat{\mu}_n\hat{\mu}_{n-1}^T$$

$$\mathbb{E}[z_n z_n^T|\theta^{old}] = \hat{V}_n + \hat{\mu}_n\hat{\mu}_n^T$$

(variance of $\mathcal{E}(z_n)$)

$$\begin{bmatrix} \mathcal{N}\left(\begin{pmatrix} x \\ y \end{pmatrix}, \begin{pmatrix} \mathcal{E}_{xx} & \mathcal{E}_{xy} \\ \mathcal{E}_{yx} & \mathcal{E}_{yy} \end{pmatrix}\right) \\ \mathbb{E}[xx^T] = \mathbb{E}\left[(x-\mu_x)(x-\mu_x)^T + \cdots\right] \\ = \mathcal{E}_{xx} + \mu_x\mu_x^T \end{bmatrix}$$

**M-step:**

$$\theta^{new} = \underset{\theta}{\operatorname{argmax}}\, Q(\theta, \theta^{old}) \quad x, y^T$$

Eg: $\mu_0, V_0$ : $Q(\theta, \theta^{old}) = -\frac{1}{2}\ln|V_0| - \frac{1}{2}\mathbb{E}_{z_1|\theta^{old}}\left[(z_1-\mu_0^T)V_0^{-1}(z_1-\mu_0)\right] + \text{const}$ (quadratic in $\mu_0$, optimum in $z_1$)

$$\mu_0^{new} = \mathbb{E}[z_1|\theta^{old}]$$

$$V_0^{new} = \mathbb{E}[z_1 z_1^T|\theta^{old}] - \mathbb{E}[z_1|\theta^{old}]\,\mathbb{E}[z_1^T|\theta^{old}]$$

⋮ updates for $A, \Gamma, C, \Sigma$

14th Oct

**Causal Relations:** A causes B if change in A → change in B.


$x_1$ causes $x_2$