Stat 521A
Lecture 6

Outline

- Exponential family: what?(8.2)
- Why? (Extra)
- Connection with GMs (8.3)
- Entropy (8.4)
- Projections (8.5)
- Querying a distribution ("inference") – 2.1.5
- Worst case complexity of exact inference (9.1)

Exponential family

- Def 8.2.2. The exponential family is a set of distributions of the form

$$p(\mathbf{x}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} h(\mathbf{x}) \exp\left(\mathbf{t}(\boldsymbol{\theta})^T \mathbf{T}(\mathbf{x})\right)$$

$$Z(\boldsymbol{\theta}) = \sum_{\mathbf{x} \in S} h(\mathbf{x}) \exp\left(\mathbf{t}(\boldsymbol{\theta})^T \mathbf{T}(\mathbf{x})\right)$$

Where $x \in X$ are the variables, h(x) defines the support (must not depend on θ), $T(x) \in R^K$ are the sufficient statistics, $\theta \in \Theta \subseteq R^M$ are the parameters, $t(\theta)$ in R^K are the natural parameters, and $Z(\theta) \in R^+$ is the partition function.

We would like Θ to be a convex open subset of R^M, and to be non-redundant (iff $t(\theta)$ is invertible).

Examples

- $X \sim \mathrm{Ber}(\theta)$.

$$\begin{aligned} \mathbf{T}(x) &= [I(x=0), I(x=1)] \\ \mathbf{t}(\boldsymbol{\theta}) &= [\log\theta, \log(1-\theta)] \\ Z(\theta) &= 1 \\ p(x) &= \exp\left(\mathbf{T}(x)^T \mathbf{t}(\boldsymbol{\theta})\right) \end{aligned}$$

$$\Theta = [0,1], \mathcal{X} = \{0,1\}$$

- $X \sim N(\mu,\sigma^2)$.

$$\begin{aligned} p(x) &= \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}\mu^2) \\ \mathbf{T}(x) &= [x, x^2] \\ \mathbf{t}(\mu,\sigma^2) &= [\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}] \\ Z(\mu,\sigma^2) &= \sqrt{2\pi}\sigma\exp(\frac{\mu^2}{2\sigma^2}) \end{aligned}$$

$$\Theta = \mathbb{R} \times \mathbb{R}^+, \mathcal{X} = \mathbb{R}$$

Non-examples

- Let X ~ Unif(a,b). Then

$$p(x|\boldsymbol{\theta}) \quad = \quad \frac{1}{b-a} I(a \leq x \leq b) = \exp(\log \frac{1}{b-a}))I(a \leq x \leq b)$$

- Support depends on \theta.
- Let X ~ $\sum_k \pi_k$ f(x,ϕ_k) – mixture model. Cannot be written in required form.

Linear exponential family

- Consider the set

$$\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^K : \int \exp(\boldsymbol{\theta}^T \mathbf{T}(\mathbf{x}))d\mathbf{x} < \infty\}$$

- If Θ is open and convex, and $t(\theta)=\theta$, we say it is a linear exponential family.
- We write

$$\begin{aligned} p(\mathbf{x}|\boldsymbol{\eta}) &= \frac{1}{Z(\boldsymbol{\eta})} h(\mathbf{x}) \exp[\boldsymbol{\eta}^T \mathbf{T}(\mathbf{x})] \\ Z(\boldsymbol{\eta}) &= \int h(\mathbf{x}) \exp[\boldsymbol{\eta}^T \mathbf{T}(\mathbf{x})] d\mathbf{x} \end{aligned}$$

- Or

$$\begin{aligned} p(\mathbf{x}|\boldsymbol{\eta}) &= h(\mathbf{x}) \exp[\boldsymbol{\eta}^T \mathbf{T}(\mathbf{x}) - A(\boldsymbol{\eta})] \\ A(\boldsymbol{\eta}) &= \log Z(\boldsymbol{\eta}) \end{aligned}$$

Bernoulli try 1

- .

$$\mathbf{T}(x) = [I(x=0), I(x=1)]$$
$$\boldsymbol{\eta} = [\log\theta, \log(1-\theta)]$$
$$p(x) = \exp\left(\boldsymbol{\eta}^T \mathbf{T}(x)\right)$$

- However, (log \theta, log (1-\theta)) is a curve, not a convex subset. Also, it is redundant.

Bernoulli try 2

- Define

$$
\begin{aligned}
T(x) &= [I(x=1)] \\
\eta &= \log \frac{\theta}{1-\theta} \qquad\qquad \Theta = \mathbb{R} \\
Z(\eta) &= 1 + \frac{\theta}{1-\theta} = \frac{1}{1-\theta} \\
p(x) &= \frac{1}{Z(\eta)} \exp(\eta T(x)) = (1-\theta) \exp(x \log \frac{\theta}{1-\theta}) \\
p(x=0) &= (1-\theta) \\
p(x=1) &= (1-\theta) \frac{\theta}{1-\theta} = \theta
\end{aligned}
$$

Gaussian – natural params

$$\boldsymbol{\eta} = [\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}]$$

$$\mathbf{T}(x) = [x, x^2]$$

The natural parameter space is $\mathbb{R} \times \mathbb{R}^-$

Finite sufficient statistics

- Defn. A statistic is a function of the data, T(D), where D=(x1,…,xn). A sufficient statistic is one that contains all the information in the data. More formally, T is sufficient for θ if θ -> T(D) -> D.
- Let Xi ~ ExpFam. The likelihood is given by

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{x}_i|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})^n}[\prod_i h(\mathbf{x}_i)] \exp(\mathbf{t}(\boldsymbol{\theta})^T \sum_{i=1}^{n} \mathbf{T}(\mathbf{x}_i))$$

- Hence the distribution has sufficient statistics of size K, independent of n $\mathbf{T}(D) = \sum_{i=1}^{n} \mathbf{T}(\mathbf{x}_i))$

- Thm (**Pitman-Koopman-Darmois).** The expfam is the only family (amongst those where support is indep of theta) with fixed sized suff stat.

Non-parametric models

- Parametric = fixed sized theta
- Exp fam = fixed size suff stat

	T fixed	T growing
θ fixed	exp fam	eg finite mixtures
θ growing	X	eg DP mixtures

LogZ is MGF

- Consider a linear expfam

$$p(\mathbf{x}|\boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})} h(\mathbf{x}) \exp[\boldsymbol{\eta}^T \mathbf{T}(\mathbf{x})]$$

- Define

$$\frac{1}{g(\boldsymbol{\eta})} \stackrel{\text{def}}{=} Z(\boldsymbol{\eta}) = \int h(\mathbf{x}) \exp[\boldsymbol{\eta}^T \mathbf{T}(\mathbf{x})] d\mathbf{x}$$

- Then

$$\begin{aligned}
1 &= g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp[\boldsymbol{\eta}^T \mathbf{T}(x)] d\mathbf{x} \\
0 &= \nabla g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp[\boldsymbol{\eta}^T \mathbf{T}(x)] d\mathbf{x} \\
&\quad + g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp[\boldsymbol{\eta}^T \mathbf{T}(\mathbf{x})] \mathbf{T}(\mathbf{x}) d\mathbf{x} \\
\int p(\mathbf{x}|\boldsymbol{\eta}) \mathbf{T}(\mathbf{x}) d\mathbf{x} &= -\nabla g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp[\boldsymbol{\eta}^T \mathbf{T}(x)] d\mathbf{x}
\end{aligned}$$

LogZ is MGF

$$
\begin{aligned}
\int p(\mathbf{x}|\boldsymbol{\eta})\mathbf{T}(\mathbf{x})d\mathbf{x} &= -\nabla g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp[\boldsymbol{\eta}^T \mathbf{T}(x)] d\mathbf{x} \\
-\nabla \log g(\boldsymbol{\eta}) &= -\frac{\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} = -(\nabla g(\boldsymbol{\eta}))(\int h(\mathbf{x}) \exp[\boldsymbol{\eta}^T \mathbf{T}(\mathbf{x})] d\mathbf{x}) \\
E[\mathbf{T}(\mathbf{X})] &= -\nabla \log g(\boldsymbol{\eta}) = \nabla \log Z(\boldsymbol{\eta})
\end{aligned}
$$

MLE is moment matching

- Proof

$$
\begin{aligned}
\log p(\mathcal{D}|\boldsymbol{\theta}) &= -n \log Z(\boldsymbol{\theta}) + \boldsymbol{\theta}^T \mathbf{T}(\mathcal{D}) \\
\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) &= -n \nabla_{\boldsymbol{\theta}} \log Z(\boldsymbol{\theta}) + \mathbf{T}(\mathcal{D}) = \mathbf{0} \\
E\mathbf{T}(\mathbf{X}) &= \frac{1}{n} \mathbf{T}(\mathcal{D})
\end{aligned}
$$

- Example. Gaussian, T(X) = (X, X^2).

$$
\begin{aligned}
E[X] &= \mu = \frac{1}{n} \sum_i x_i \\
\text{Var}\,[X] &= (EX^2) - (EX)^2 \\
E[X^2] &= \sigma^2 + \mu^2 = \frac{1}{n} \sum_i x_i^2 \\
\sigma^2 &= \frac{1}{n} \sum_i x_i^2 - \mu^2
\end{aligned}
$$

Conjugate priors

- Defn. A prior $p(\theta) \in F$ is conjugate to a likelihood $p(D|\theta)$ if the posterior satistifes $p(\theta|D) \in F$, i.e., has the same functional form as the prior.
- Thm. All dist in expfam have conj prior.
- Most distrib with conj prior are in exp fam.

Maximum entropy principle

- Defn. The entropy of a pmf is

$$H(p) \stackrel{\text{def}}{=} -\sum_x p(x) \log p(x), H(p) \geq 0$$

- The differential entropy of a pdf can be –ve

$$h(p) \stackrel{\text{def}}{=} -\int_S p(x) \log p(x) dx$$

- The relative entropy, or KL divergence, from p to q is given by

$$KL(p,q) \stackrel{\text{def}}{=} \sum_x p(x) \log \frac{p(x)}{q(x)}$$

- KL is always >= 0, even for pdf's.

Maxent principle

- Suppose we want to pick the most uncertain distribution (principle of least commitment) subject to the constraints that

$$\sum_x f_k(x)p(x) = F_k$$

- Optimize the Lagrangian

$$\begin{aligned}
J(p) &= -\sum_x p(x)\log p(x) + \lambda_0(1-\sum_x p(x)) + \sum_k \lambda_k(F_k - \sum_x p(x) f_k(x)) \\
\frac{\partial J}{\partial p(x)} &= -1 - \log p(x) - \lambda_0 - \sum_k \lambda_k f_k(x) = 0 \\
p(x) &= \frac{1}{Z}\exp(-\sum_k \lambda_k f_k(x)) \\
Z &= e^{1+\lambda_0} \\
1 &= \sum_x p(x) = \frac{1}{Z}\sum_x \exp(-\sum_k \lambda_k f_k(x)) \\
Z &= Z(\boldsymbol{\lambda}) = \sum_x \exp(-\sum_k \lambda_k f_k(x))
\end{aligned}$$

Gaussian maximizes entropy

- MVN is in expfam.

$$
\begin{aligned}
p(\mathbf{x}) &= \frac{1}{Z}\exp(-\tfrac{1}{2}\mathbf{x}^T\mathbf{K}\mathbf{x}) = \frac{1}{Z}\exp(\sum_k \lambda_k f_k(\mathbf{x})) \\
f_{ij}(\mathbf{x}) &= x_i x_j,\ \lambda_{ij} = \tfrac{1}{2}K_{ij}
\end{aligned}
$$

Theorem 0.1. *Let $g(\mathbf{x})$ be any density satisfying $\int g(\mathbf{x})x_i x_j = \Sigma_{ij}$. Let $\phi = \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. Then $h(g) \leq h(\phi)$.*

Proof. (From (**?**, p234).) We have

$$
\begin{aligned}
0 &\leq KL(g||\phi) && (1)\\
&= \int g(\mathbf{x}) \log \frac{g(\mathbf{x})}{\phi(\mathbf{x})} d\mathbf{x} && (2)\\
&= -h(g) - \int g(\mathbf{x}) \log \phi(\mathbf{x}) d\mathbf{x} && (3)\\
&= -h(g) - \int \phi(\mathbf{x}) \log \phi(\mathbf{x}) d\mathbf{x} \text{ (**)} && (4)\\
&= -h(g) + h(\phi) && (5)
\end{aligned}
$$

where the line marked (**) follows since g and ϕ yield the same moments for the quadratic form $\log \phi(\mathbf{x})$. ∎

Some GMs are expfam models

- We showed earlier that many +ve UGM can be represented as an expfam

$$p(\mathbf{x}) = \frac{1}{Z}\exp(\sum_i \boldsymbol{\theta}_i^T f_i(\mathbf{x}))$$

- Most CPDs can be represented as expfam
- Eg table p(X|U). T(X,U)=[I(X=x), I(U=u)], t(\theta) = [\log p(x|u)].
- Eg lingauss.

$$p(x|\mathbf{u}) = \frac{1}{\sqrt{2\pi\sigma}}\exp\left(-\frac{1}{2\sigma^2}(x-(w_0+w_1u_1+\cdots+w_ku_k))^2\right)$$

$$\mathbf{T}(x,\mathbf{u}) = [1, x, u_1, \ldots, u_k, xu_1, \ldots, xu_k, u_1^2, u_1u_2, \ldots, u_k^2]$$

- Product of expfam is expfam.

DGMs are curved expfam

- In general, the fact that CPDs sum to 1 locally means that they are not linear expfam
- See p248 of K&F
- Geiger'01 shows that DGMs are curved expfam models (curved means the params are not linearly indep, so \theta is smaller than t(\theta)).
- Geiger'01 also shows that GMs with hidden variables are stratified exponential families (SEFs) - a finite union of CEFs of various dimensions satisfying some regularity conditions.

Stratified exponential families: Graphical models and model selection
Dan Geiger, David Heckerman, Henry King, and Christopher Meek
Source: Ann. Statist. Volume 29, Number 2 (2001), 505-529.

Entropy of an expfam model

- Thm 8.4.1. If X ~ ExpFam(theta), then

$$H(P_{\boldsymbol{\theta}}(\mathbf{x})) = \log Z(\boldsymbol{\theta}) - E[\mathbf{T}(\mathbf{x})^T \mathbf{t}(\boldsymbol{\theta})]$$

- Ex 8.4.2. Gaussian.

$$\begin{aligned}
p(x) &= \frac{1}{\sqrt{2\pi\sigma}} \exp(-\frac{1}{2\sigma^2} x^2 + \frac{\mu}{\sigma^2} x - \frac{1}{2\sigma^2} \mu^2) \\
\mathbf{T}(x) &= [x, x^2] \\
\mathbf{t}(\mu, \sigma^2) &= [\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}] \\
Z(\mu, \sigma^2) &= \sqrt{2\pi}\sigma \exp(\frac{\mu^2}{2\sigma^2}) \\
H &= \tfrac{1}{2} \ln(2\pi\sigma^2) + \frac{\mu^2}{2\sigma^2} - \frac{\mu}{\sigma^2} E[x] + \frac{1}{2\sigma^2} E[x^2] \\
&= \tfrac{1}{2} \ln(2\pi\sigma^2) + \frac{\mu^2}{2\sigma^2} - \frac{2\mu^2}{2\sigma^2} + \frac{1}{2\sigma^2} (\mu^2 + \sigma^2) \\
&= \tfrac{1}{2} \ln(2\pi\sigma^2) + \tfrac{1}{2} = \tfrac{1}{2} \ln(2\pi\sigma^2) + \tfrac{1}{2} \ln e = \tfrac{1}{2} \ln(2\pi\sigma^2 e)
\end{aligned}$$

Entropy of a GM

- Thm 8.4.3. If P(X) = 1/Z $\prod_c \phi_c(X)$ is a UGM, then

$$H(P_{\boldsymbol{\theta}}(\mathbf{x})) = \log Z(\boldsymbol{\theta}) + \sum_c E[-\ln \phi_c(\mathbf{x}_c)]$$

- Thm 8.4.5. If P(X) is a DGM, then

$$H(P(\mathbf{X})) = \sum_i H(P(X_i|X_{\pi_i}))$$

- Pf.

$$\begin{aligned} H(P(\mathbf{X})) &= E[-\log p(\mathbf{X})] = E[-\sum_i \log p(X_i|\mathbf{X}_{\pi_i})] \\ &= \sum_i E[-\log p(X_i|\mathbf{X}_{\pi_i})] = \sum_i H(P(X_i|\mathbf{X}_{\pi_i})) \\ &= \sum_i \sum_{\mathbf{x}_{\pi_i}} p(\mathbf{x}_{\pi_i}) H(P(X_i|\mathbf{x}_{\pi_i})) \end{aligned}$$

- Thm 8.4.6. If P(X) is a DGM, then

$$\sum_i \min_{\mathbf{x}_{\pi_i}} H(P(X_i|\mathbf{x}_{\pi_i})) \le H(P(\mathbf{X})) \le \sum_i \max_{\mathbf{x}_{\pi_i}} H(P(X_i|\mathbf{x}_{\pi_i}))$$

Projections

- Def 8.5.1. Let P b a distribution and Q a convex set of distributions.
- The I-projection (information) is

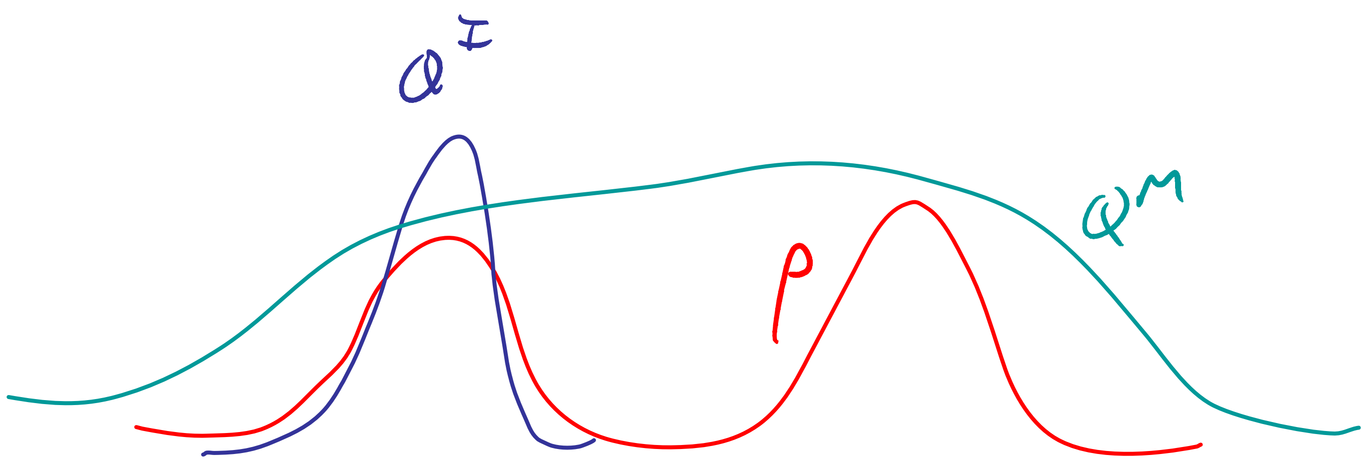
$$Q^I = \arg\min_{Q \in \mathcal{Q}} D(Q||P)$$

Zero forcing: P=0 => Q=0 Mode seeking

- The M-projection (moment) is

$$Q^M = \arg\min_{Q \in \mathcal{Q}} D(P||Q)$$

Q=0 => P=0 High variance



M-projection is moment matching

- Thm 8.5.5. Let P be any distrib over X, and let Q be expfam. If there is a set of params θ st $E_Q(\theta)[\tau(X)] = E_P[\tau(X)]$, then the M-projection of P onto Q is Q_θ.
- Ex. Let Q = fully factorized distribution. Then Q^M is given by product of marginals.

$$Q^M(\mathbf{x}) = p(X_1) \ldots p(X_d)$$

- Ex. Let P = mix Gaussians, Q = single Gaussian.

$$\begin{aligned}
p(\mathbf{x}) &= \sum_k \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\
Q^M(\mathbf{x}) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_Q, \boldsymbol{\Sigma}_Q) \\
\boldsymbol{\mu}_Q &= \sum_k \pi_k \boldsymbol{\mu}_k \\
\boldsymbol{\Sigma}_Q &= \sum_k \pi_k (\boldsymbol{\Sigma}_k + (\boldsymbol{\mu}_k - \boldsymbol{\mu}_Q)(\boldsymbol{\mu}_k - \boldsymbol{\mu}_Q)^T)
\end{aligned}$$

I-projection

- I-projection requires computing expectations of log(P) – which often factorizes - wrt Q, and the entropy of Q.

$$Q^I = \arg\min_{Q \in \mathcal{Q}} D(Q||P) = \arg\min \sum_x Q(x) \log \frac{Q(x)}{P(x)}$$

- We can choose Q to be “simple”, so that it is easy to compute these expectations and entropy terms.
- This is the basis of variational inference.
- By contrast, M-projections require expectations wrt P. Usually this can only be done locally, as in expectation propagation.

Querying a distribution ("inference")

- Suppose we have a joint $p(X_1,\ldots,X_d)$. Partition the variables into E (evidence), Q (query), and H (hidden/ nuisance). We might pose the following queries
- Conditional probability (posterior):

$$p(\mathbf{X}_Q|\mathbf{x}_E) \propto \sum_{\mathbf{x}_H} p(\mathbf{X}_Q, \mathbf{x}_E, \mathbf{x}_H)$$

- MAP estimate (H=∅) (posterior mode)

$$\mathbf{x}_Q^* = \arg\max_{\mathbf{x}_Q} p(\mathbf{x}_Q|\mathbf{x}_E) = \arg\max_{\mathbf{x}_Q} p(\mathbf{x}_Q, \mathbf{x}_E)$$

- Marginal MAP estimate (mode of marginal post):

$$\mathbf{x}_Q^* = \arg\max_{\mathbf{x}_Q} p(\mathbf{x}_Q|\mathbf{x}_E) = \arg\max_{\mathbf{x}_Q} \sum_{\mathbf{x}_H} p(\mathbf{x}_Q, \mathbf{x}_E, \mathbf{x}_H)$$

MAP vs marginal MAP

- Max max ≠ max sum
- Ex 2.1.12. Joint is

$$a^* = \arg\max_a \sum_b p(a,b) = 1$$

$$b^* = \arg\max_b \sum_a p(a,b) = 1$$

$$(a,b)^* = \arg\max_{a,b} p(a,b) = (0,1)$$

	A=0	A=1	
B=0	0.04	0.3	0.34
B=1	0.36	0.3	0.66
	0.4	0.6	

- One can show that max sum is strictly computationally harder than sum, which is in turn harder than max

Speech recognition

- Eg speech recognition. Let Q=words, H = pronunciation (phonemes sequence), E = signal.
- We often make the following approximation, which lets us use the Viterbi algorithm

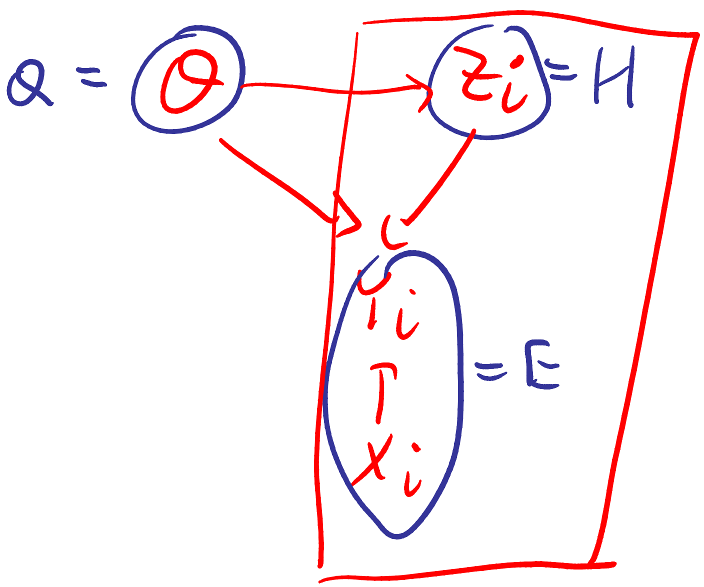
$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \sum_{\mathbf{h}} p(\mathbf{w}, \mathbf{h}|\mathbf{e}) \approx \arg\max_{\mathbf{w}} \max_{\mathbf{h}} p(\mathbf{w}, \mathbf{h}|\mathbf{e})$$

- Eg. Consider W1="a back", vs W2="aback". There might be 10 alternative state sequences for W1, each with prob 0.03, but just one sequence for W2, with prob 0.2. Viterbi would choose W2, but W1 is actually more likely.

Bayesian statistics

- Bayesian statistics amounts to defining a single joint distribution for both “variables” – latent and observed - and “parameters” (often fixed in number), and then querying the parameters.

$$p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Y}) \quad \propto \quad p(\boldsymbol{\theta}) \prod_i \int p(\mathbf{z}_i|\boldsymbol{\theta}) p(\mathbf{y}_i|\mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\theta}) d\mathbf{z}_i$$



Probability of evidence

- To compute conditional queries, we need to evaluate $p(x_E)$

$$p(\mathbf{X}_Q|\mathbf{x}_E) = \frac{\sum_{\mathbf{x}_H} p(\mathbf{X}_Q, \mathbf{x}_E, \mathbf{x}_H)}{p(\mathbf{x}_E)}$$

$$p(\mathbf{x}_E) = \sum_{\mathbf{x}_Q} \sum_{\mathbf{x}_H} p(\mathbf{x}_Q, \mathbf{x}_E, \mathbf{x}_H)$$

- This may be a high dimensional integral

$$p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Y}) = \frac{p(\boldsymbol{\theta}) \prod_i \int p(\mathbf{z}_i|\boldsymbol{\theta}) p(\mathbf{y}_i|\mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\theta}) d\mathbf{z}_i}{p(\mathbf{X}, \mathbf{Y})}$$

$$p(\mathbf{X}, \mathbf{Y}) = \int p(\boldsymbol{\theta}) \left[\prod_i \int p(\mathbf{z}_i|\boldsymbol{\theta}) p(\mathbf{y}_i|\mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\theta}) d\mathbf{z}_i \right] d\boldsymbol{\theta}$$

- $p(x_E)$ can be used to decide how likely x_E is to have come from this model (classification and model selection)

Sampling

- Often the posterior is too big to even store explicitly.
- Marginals and MAP estimates are one summary, but may be unrepresentative.
- Samples may provide a better summary.
- eg Attractive Ising model has 2 modes, all 0 and all 1. The marginals are [0.5, 0.5].
- We want to be able to sample from p(xQ|xE)
- Sometimes we can do this even if we cannot evaluate p(xE) – this is the key idea behind MCMC

Monte Carlo integration

- Sometimes we want to E[f(xQ)|xE], where f() depends on global properties of Q, so we cannot use marginal distributions.
- However, if we sample from p(XQ|xE), we can use

$$E[f(\mathbf{X}_Q)|\mathbf{x}_E] = \int f(\mathbf{x}_Q)p(\mathbf{x}_Q|\mathbf{x}_E)d\mathbf{x}_Q \approx \frac{1}{N}\sum_{i=1}^{n} f(\mathbf{x}_Q^i)$$

Inference in discrete state spaces

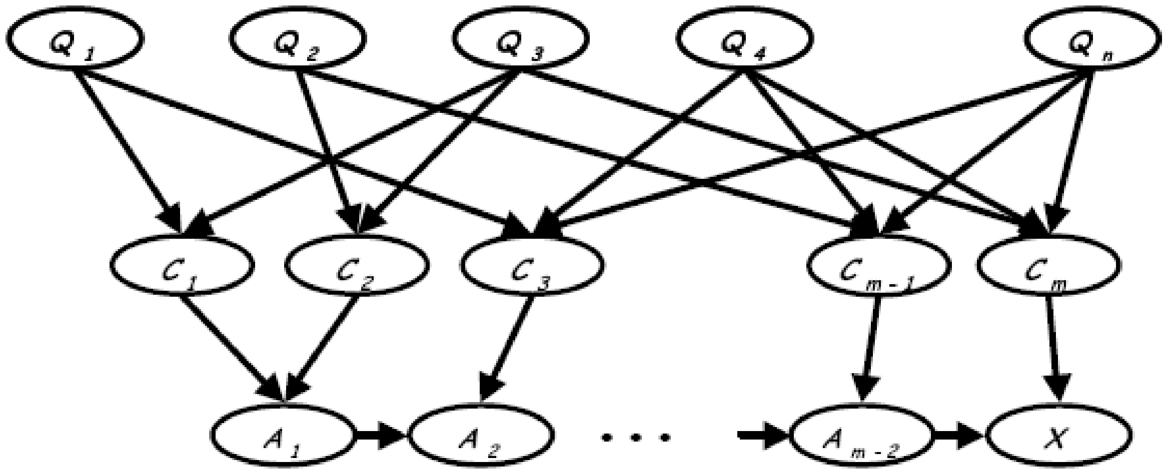
- We will mostly focus on the case where Q and H are discrete rv's (E can be cts or discrete).
- Thus everything amounts to computing a large number of sums as quickly as possible.
- We will also consider the case where Q, H and E are all jointly Gaussian, where exact answers can also be obtained.
- For general distributions (eg for applications in Bayesian statistics), exact inference is usually not possible (except 1 layer of parameters with conjugate priors and no latent variables).

Complexity of inference

- Consider computing p(X_Q), p(X_Q|x_E), or p(x_E) for a discrete state space.
- Later we will show that if P is representable by a GM, then we can compute these quantities efficiently, if the graph has special properties.
- However, in general, the problem is computationally expensive.

Complexity of exact inference

- Thm 9.1.1. Given a DGM, deciding if p(X=x)>0 is NP-complete.
- Pf. Easy to see is in NP (linear time to check if p(x)>0.) Can show is NP-hard by showing how to reduce 3-SAT to a poly-sized DGM.



$$X = (Q_1 \vee \neg Q_2 \vee Q_3) \wedge (Q_2 \vee Q_5 \vee Q_3) \cdots$$

P(X=1) = #satisfying assignments/ 2^n

Complexity of exact inference

- Defn. NP is the class of problems of the form “are there any solutions x such that f(x) is true”. #P is the class of problems “Count the number of solutions x st f(x) is true”.
- Thm 9.1.2. Given a DGM, computing p(X=x) is #P-complete.

Complexity of approximate inference

- Def 9.1.3. A estimate ρ has absolute error ε if

$$|p(\mathbf{x}_Q|\mathbf{x}_e) - \rho| \leq \epsilon$$

- Def 9.1.4. An estimate ρ has relative error ε if

$$\frac{\rho}{1+\epsilon} \leq p(\mathbf{x}_Q|\mathbf{x}_e) \leq \rho(1+\epsilon)$$

- Thm 9.1.5. Given a DGM, finding a number ρ which as relative error ε for p(X=x) is NP-hard.
- Thm 9.1.6. Given a DGM, finding a number ρ that has absolute error ε for p(X|e) is NP-hard for any $0 \leq \varepsilon \leq 0.5$.