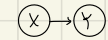
· Maximum Likelihood Estimation for Bayesian Network

eg. 

X	θ_X
x^0	θ_{x^0}
x^1	θ_{x^1}

X	Y	
x^0	y^0	$\theta_{y^0\|x^0}$
x^0	y^1	$\theta_{y^1\|x^0}$
x^1	y^0	$\theta_{y^0\|x^1}$
x^1	y^1	$\theta_{y^1\|x^1}$

given a sample $D=\{\langle x[m], y[m]\rangle\}$

$$L(\theta;D)=\prod_{m=1}^{M} P(x[m], y[m];\theta)$$

$$=\prod_{m=1}^{M} P(x[m];\theta)\cdot P(y[m]\mid x[m];\theta)$$

each of these terms is a "local likelihood" that measures how well the variable is predicted given its parents

$$=\left(\prod_{m=1}^{M} P(x[m];\theta)\right)\cdot\left(\prod_{m=1}^{M} P(y[m]\mid x[m];\theta)\right)$$

$$\ell(\theta;D)=\sum_{m=1}^{M}\log P(x[m];\theta)+\sum_{m=1}^{M}\log P(y[m]\mid x[m];\theta)$$

$$=M[x^0]\cdot\log\theta_{x^0}+M[x^1]\cdot\log\theta_{x^1}+\sum_{x,y\in Val(X)} M[x,y]\cdot\log\theta_{y|x}$$

$$\min_{\theta} -c^T\log\theta \qquad 1^T\theta=1$$

$$L(\theta,v)=-c^T\log\theta+v(1^T\theta-1)$$

$$\nabla_\theta L=-c\cdot\frac{1}{\theta}+v\cdot 1=0 \qquad v=\frac{c}{\theta}$$

$$\theta=\frac{c}{v1}$$

$$g(v)=-c^T\left(\log c-\log v1\right)+v\left(1^T\frac{c}{v1}-1\right)$$

$$=-c^T\log c+c^T\log v1+v\cdot 1^T\frac{c}{v1}-v$$

$$=-c^T\log c+c^T\log v1+1^Tc-v$$

$$\nabla_v g=\frac{1^Tc}{v}-1=0$$

$$v=1^Tc$$

$$\theta=c/1^Tc$$

$$\theta_{x^0}=\frac{M[x^0]}{M[x^0]+M[x^1]} \qquad \theta_{x^1}=\frac{M[x^1]}{M[x^0]+M[x^1]}$$

$$\theta_{y|x}=\frac{M[y,x]}{\sum_{y\in Val(Y)} M[y,x]}=\frac{M[y,x]}{M[x]}$$

$$\hat{\theta}_{x|u}=\frac{M[x,u]}{\sum_{x\in Val(X)} M[x,u]}=\frac{M[x,u]}{M[u]} \qquad \forall\, x\in Val(X),\ u\in Val(Parents(X))$$

· Gaussian Bayesian Networks

$$P(X\mid U)=\mathcal{N}(\beta_0+\beta_1U_1+\cdots+\beta_kU_k;\sigma^2)$$

$$\ell(\theta_{X|U};D)=\sum_{i=1}^{M}\log\mathcal{N}(x[i];\beta_0+\cdots+\beta_kU_k,\sigma^2)$$

$$=\sum_{i=1}^{M}-\frac{1}{2}\log(2\pi\sigma)-\frac{1}{2}\frac{1}{\sigma^2}\left(\beta_0+\cdots\beta_kU_k[i]-x[i]\right)^2$$

$$\nabla_{\beta_i}\ell=\sum_{i=1}^{M}-\frac{1}{\sigma^2}\left(\beta_0+\cdots\beta_kU_k[i]-x[i]\right)=0$$

$$\frac{1}{M}\sum_{i=1}^{M}x[i]=\frac{1}{M}\sum_{i=1}^{M}\beta_0+\cdots+\beta_kU_k[i]=\beta_0+\beta_1\frac{1}{M}\sum_i U_1[i]+\cdots\beta_k\frac{1}{M}\sum_i U_k[i]$$

$$E_D[X]=\beta_0+\beta_1\cdot E[U_1]+\cdots\beta_k\cdot E[U_k]$$

$$\nabla_{\beta_i} \ell = \sum_{m=1}^{M} -\frac{1}{\sigma^2}\left(\beta_0 + \cdots \beta_k U_k[m] - X[m]\right)\cdot U_i[m] := 0$$

$$1/M \sum_{m=1}^{M} X[m]\cdot U_i[m] = 1/M \sum_{m=1}^{M} \beta_0 \cdot U_i[m] + \cdots + 1/M \sum_{m=1}^{M} \beta_k U_k[m]\cdot U_i[m]$$

$$E_{\hat{P}_D}[X\cdot U_i] = \beta_0 E_{\hat{P}_D}[U_i] + \beta_1 E_{\hat{P}_D}[U_1\cdot U_i] + \cdots + \beta_k E_{\hat{P}_D}[U_k U_i]$$

$$\begin{bmatrix} 1 & E[U_1] & \cdots & E[U_k] \\ E[U_1] & E[U_1 U_1] & \cdots & E[U_k U_1] \\ E[U_2] & E[U_1 U_2] & \cdots & E[U_k U_2] \\ \vdots & \vdots & & \vdots \\ E[U_k] & E[U_1 U_k] & \cdots & E[U_k U_k] \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} = \begin{bmatrix} E[X] \\ E[XU_1] \\ E[XU_2] \\ \vdots \\ E[XU_k] \end{bmatrix}$$

Solve for β by solving linear equation

$$\begin{aligned} Cov_D[X, U_i] &= E_D[X\cdot U_i] - E_D[X]\cdot E_D[U_i] \\ &= \beta_0 E_D[U_i] + \beta_1 E_D[U_1\cdot U_i] \cdots + \beta_k E_D[U_k U_i] \\ &\quad - \beta_0 E_D[U_i] - \beta_1 E_D[U_1]\cdot E_D[U_i] \cdots - \beta_k E_D[U_k] E_D[U_i] \\ &= \beta_1 Cov_D[U_1, U_i] + \cdots \beta_k Cov_D[U_k, U_i] \end{aligned}$$

$$\ell(\theta_{MLE}; D) = \sum_{i=1}^{M} -\frac{1}{2}\log(2\pi\theta) - \frac{1}{2}\frac{1}{\sigma^2}\left(\beta_0 + \cdots \beta_k U_k[i] - X[i]\right)^2$$

$$\nabla_\sigma \ell = \sum_{m=1}^{M} -\frac{1}{2\pi\sigma^2}\cdot 2\pi\sigma + \frac{1}{2}\frac{1}{(\sigma^2)^2}\left(\beta_0 + \cdots \beta_k U_k[m] - X[m]\right)^2 := 0$$

$$\sum_{m=1}^{M} \frac{1}{(\sigma^2)^2}\left(\beta_0 + \cdots \beta_k U_k[m] - X[m]\right)^2 - \frac{1}{\sigma^2} = 0$$

$$\sigma^2 = Cov_D[X, X] - \sum_i \sum_j \beta_i \beta_j Cov_D[U_i, U_j]$$

· MLE as M-Projection

Let D be a dataset, $\ell(\theta; D) = \log L(\theta; D) = M\cdot E_{x\sim\hat{P}_D}[\log P(x;\theta)]$

$$\begin{aligned} \ell(\theta; D) &= \sum_{m=1}^{M} \log P(X[m]; \theta) \\ &= \sum_{x\in val(X)} \left[\sum_m 1\{X[m] = x\}\right]\cdot \log P(x;\theta) \\ &= \sum_{x\in val(X)} M\cdot \hat{P}_D(x) \log P(x;\theta) \\ &= M\cdot E_{x\sim\hat{P}_D}[\log P(x;\theta)] \end{aligned}$$

$$D(P\|P') = \int_x P(x)\log\frac{P(x)}{P'(x)}\,dx = E_{x\sim P}\left[\log\frac{P(x)}{P'(x)}\right] = -H_P(X) - E_{x\sim P}[\log P'(x)]$$

$$\ell(\theta; D) = M\cdot\left(-H_{\hat{P}_D}(X) - D(\hat{P}_D\|P)\right)$$

the maximum likelihood estimation θ in a parametric family is the param of M-projection of $\hat{P}_D$ on P_θ

Joint Probabilistic Model

encode the prior knowledge about θ with a probability distribution,
which represents how likely different choices of θ are
then create a joint distribution of parameter θ and observation $\{X[m]\}$

eg. coin toss

$P(X[m]=1|\theta) = \theta$ $\quad \theta \sim \text{Uniform}(0,1)$

$P(X[m]=0|\theta) = 1-\theta$

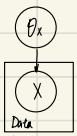
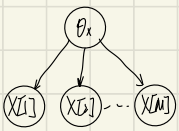


plate model



Ground Bayesian Network

$X[i] \perp X[j] \mid \theta_X$

$$P(\theta, X[1]\cdots X[M]) = P(\theta)\cdot P(X[1]\cdots X[M]|\theta)$$
$$= P(\theta)\cdot \prod_{m=1}^{M} P(X[m]|\theta)$$
$$= P(\theta)\cdot \theta^{M[1]}\cdot(1-\theta)^{M[0]}$$

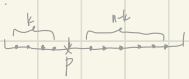
posterior: $P(\theta|X[1]\cdots X[M]) = \dfrac{P(\theta, X[1]\cdots X[M])}{P(X[1]\cdots X[M])}$

In prediction, instead of selecting a single value for θ, we use it for predicting the probability over the next toss

$$P(X[M+1]|X[1]\cdots X[M]) = \int P(X[M+1], \theta|X[1]\cdots X[M])\,d\theta$$
$$= \int P(X[M+1]|\theta, X[1]\cdots X[M])\cdot P(\theta|X[1]\cdots X[M])\,d\theta$$
$$= \int P(X[M+1]|\theta)\cdot P(\theta|X[1]\cdots X[M])\,d\theta$$
$$= \int P(X[M+1]|\theta)\cdot \frac{P(\theta)\cdot P(X[1]\cdots X[M]|\theta)}{P(X[1]\cdots X[M])}\,d\theta$$
$$= \frac{1}{P(X[1]\cdots X[M])}\int P(X[M+1]|\theta)\cdot\theta^{M[1]}(1-\theta)^{M[0]}\,d\theta$$

$$\int_0^1 p^k(1-p)^{n-k}\,dp = \frac{k!\,(n-k)!}{(n+1)!}$$

Strong proof:



throw k white balls, n-k black balls and 1 red.

$P($ white $|$ black $)$

$$P(X[M+1]=1|X[1]\cdots X[M]) = \frac{1}{P(X[1]\cdots X[M])}\int_0^1 \theta^{M[1]+1}(1-\theta)^{M[0]}\,d\theta = \frac{1}{Z}\frac{(M[1]+1)!\,M[0]!}{(M[0]+M[1]+2)!}$$

$$P(X[M+1]=0|X[1]\cdots X[M]) = \frac{1}{Z}\int_0^1 \theta^{M[1]}(1-\theta)^{M[0]+1}\,d\theta = \frac{M[1]!\,(M[0]+1)!}{(M[0]+M[1]+2)!}$$

$$\begin{cases} a+b=1 \\ \frac{a}{b} = \frac{(M[1]+1)!\cdot M[0]!}{M[1]!\cdot(M[0]+1)!} = \frac{M[1]+1}{M[0]+1} \end{cases} \Rightarrow P(X[M+1]=1|X[1]\cdots X[M]) = \frac{M[1]+1}{M[0]+M[1]+2}$$

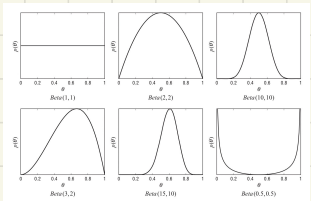
Laplace Smooth

Priors

appropriate distribution is beta distribution

$\theta \sim Beta(\alpha_1, \alpha_0) = \gamma\, \theta^{\alpha_1 - 1} (1-\theta)^{\alpha_0 - 1} \quad \theta \in [0,1]$

where $\gamma = \frac{\Gamma(\alpha_1 + \alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_0)}$ is a normalizing constant, $\Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t}\, dt$ is the Gamma function



Gamma function $\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha} e^{-x} \frac{1}{x}\, dx \quad (\alpha > 0)$

$\Gamma(n) = (n-1)! \quad n$ integer $\alpha > 0$

$\Gamma(\frac{1}{2}) = \sqrt{\pi}$

$\Gamma(\frac{1}{2}) = \int_0^{+\infty} x^{-\frac{1}{2}} e^{-x}\, dx \quad u = x^{1/2}$

$= \int_0^{+\infty} u^{-1} e^{-u^2}\, du^2$

$= 2\int_0^{+\infty} e^{-u^2}\, du$

$= \int_{-\infty}^{+\infty} e^{-u^2}\, du$

$= \sqrt{\pi}$

$\Gamma(\alpha+1) = \int_0^{+\infty} x^{\alpha} e^{-x}\, dx = \int_0^{+\infty} x^{\alpha}\, de^{-x}$

$= -\left[x^{\alpha} e^{-x} \Big|_0^{+\infty} - \int_0^{+\infty} e^{-x}\, dx^{\alpha} \right]$

$= \alpha \int_0^{+\infty} e^{-x} x^{\alpha - 1}\, dx$

$\begin{cases} \Gamma(\alpha+1) = \alpha \cdot \Gamma(\alpha) \\ \Gamma(1) = 1 \end{cases}$

$\int_0^1 \theta^{\alpha_1 - 1} (1-\theta)^{\alpha_0 - 1}\, d\theta = \frac{(\alpha_1 - 1)!\,(\alpha_0 - 1)!}{(\alpha_1 - 1 + \alpha_0 - 1 + 1)!} \qquad \gamma = \frac{1}{\int d\theta} = \frac{(\alpha_1 + \alpha_0 - 1)!}{(\alpha_1 - 1)!\,(\alpha_0 - 1)!} = \frac{\Gamma(\alpha_1 + \alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_0)}$ (for discrete α_1, α_0)

eg. $(X|\theta) \sim Binomial(M, \theta) \quad \sim Beta(a, b) \qquad$ find $P(\theta|X)$

$P(\theta | X=k) = \frac{P(\theta) \cdot P(X=k)}{P(X)} = \frac{\binom{M}{k} \theta^k (1-\theta)^{M-k} \cdot \gamma \cdot \theta^{a-1} (1-\theta)^{b-1}}{\int_0^1 P(\bar{\theta})\, P(X=k|\bar{\theta})\, d\bar{\theta}} \propto \theta^{a+k} (1-\theta)^{b-1+M-k}$

$(\theta|X) \sim Beta(a+X,\ b+n-X)$

prior: $p \sim$ Beta, $X|p \sim$ Binomial posterior: $p|x \sim$ Beta

Beta is "conjugate prior" to Binomial

$P(X[M+1] | X[1] \cdots X[M]) = \int_{\theta=0}^{1} P(X[M+1], \theta | X[1] \cdots X[M])\, d\theta$

$= \int_0^1 P(X[M+1] | \theta, X[1] \cdots X[M])\ P(\theta | X[1] \cdots X[M])\, d\theta$

$P(X[M+1]=1 | X[1] \cdots X[M]) = \int_0^1 \theta \cdot \gamma\, \theta^{a+X-1} (1-\theta)^{b+M-X-1}\, d\theta$

$= \gamma \int_0^1 \theta^{a+X} (1-\theta)^{b+M-X-1}\, d\theta = \gamma \frac{(a+X)!\,(b+M-X-1)!}{(a+b+M)!} \qquad X = \sum_1^M 1\{X[i]=1\}$

$P(X[M+1]=0 | X[1] \cdots X[M]) = \int_0^1 (1-\theta)\, \gamma\, \theta^{a+X-1} (1-\theta)^{b+M-X-1}\, d\theta$

$= \gamma \int_0^1 \theta^{a+X-1} (1-\theta)^{b+M-X}\, d\theta = \gamma \frac{(a+X-1)!\,(b+M-X)!}{(a+b+M)!}$

$P(X[M+1]=1 | X[1] \cdots X[M]) = \frac{a+X}{a+b+M}$

intuitively, a, b corresponds to the num of imaginary heads and tails we "have seen" before the experiment $\{X[1] \cdots X[M]\}$

· Prior and Posteriors

$D = \{x[1] \cdots x[m]\}$ iid $x[i] \sim P_\theta$

$$P(\theta|D) = \frac{P(D|\theta)\cdot P(\theta)}{P(D)} = \frac{P(D|\theta)\cdot P(\theta)}{\int_\theta P(D|\theta)P(\theta)\,d\theta}$$

(likelihood · prior; denominator: marginal likelihood — a priori probability of seeing the particular dataset D given prior beliefs $P(\theta)$)

A Dirichlet distribution is specified by a set of hyperparameters $\alpha_1 \cdots \alpha_k$

$\theta \sim \text{Dirichlet}(\alpha_1 \cdots \alpha_k)$ if $p(\theta) \propto \prod_k \theta_k^{\alpha_k - 1}$

If $P(\theta)$ is Dirichlet $(\alpha_1 \cdots \alpha_k)$, then $P(\theta|D)$ is Dirichlet$(\alpha_1 + M[1], \cdots \alpha_k + M[k])$

A family of priors $P(\theta;\alpha)$ is conjugate to a model $P(x|\theta)$ if for any $D = \{x[1] \cdots x[m]\}$ of iid samples $x[i] \sim P(x|\theta)$, $\forall \alpha$, there are hyperparameter α' that describe the posterior

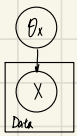
$P(\theta;\alpha') \propto P(D|\theta)\,P(\theta;\alpha)$ (prior and posterior are in the same family)

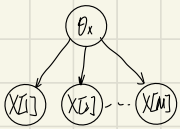
eg. multinomial likelihood and Dirichlet prior

$\theta \in R^k$ $\sum \theta = 1$ $P(\theta) \propto \prod \theta_k^{\alpha_k - 1}$

$P(X=k|\theta) = \theta_k$

$L(\theta;D) = P(D|\theta) = \prod_k \theta_k^{M[k]}$





$$P(\theta|D) = \frac{P(D|\theta)\cdot P(\theta)}{\int_\theta P(D|\theta)P(\theta)\,d\theta} \propto \prod_k \theta_k^{M[k]} \cdot \theta_k^{\alpha_k - 1}$$

$\theta|D \sim \text{Dirichlet}(\alpha_1 + M[1], \cdots \alpha_k + M[k])$

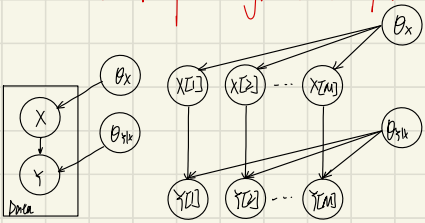
Dirichlet prior are conjugate of multinomial model

Special case: Beta prior are conjugate of Binomial model

· Parameter Independence and Global Decomposition

the instances are independent given unknown parameters

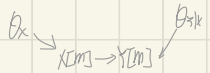
eg.



$X[m], Y[m] \perp X[m'], Y[m'] \mid \theta_x, \theta_{y|x}$

$\theta_x \perp \theta_{y|x} \mid (X[1], Y[1]), \cdots (X[M], Y[M])$

knowing the value of one parameter tells nothing about another



$(X[m], Y[m])$ block the v-structure

let G be a bayesian network with parameter $\theta = (\theta_{X_i|Pa(X_i)} \cdots \theta_{X_n|Pa(X_n)})$

$P(\theta)$ satisfies global parameter independence if it has form $P(\theta) = \prod_i P(\theta_{X_i|Pa(X_i)})$

complete data d-separates the parameter for different CPDs

$P(\theta_X, \theta_{Y|x} | D) = P(\theta_X|D) \cdot P(\theta_{Y|x}|D)$

Once we can solve each parameter separately, we can combine the results (like in MLE)

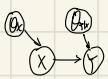
$$P(\theta|D) = \frac{P(D|\theta)\cdot P(\theta)}{P(D)}$$

$$= \frac{1}{P(D)} \underbrace{\prod_i L_i(\theta_{X_i|Pa(X_i)}; D)}_{\text{local likelihoods}} \cdot \underbrace{\prod_i P(\theta_{X_i|Pa(X_i)})}_{\text{global parameter independence}}$$

$$= \frac{1}{P(D)} \prod_i \left[L_i(\theta_{X_i|Pa(X_i)}; D) \cdot P(\theta_{X_i|Pa(X_i)}) \right]$$

let G be a bayesian network over $\mathcal{X}$. $D = \{X[1] \cdots X[M]\}$ if $P(\theta)$ satisfies global parameter independence

eg.



$$P(x[M+1], y[M+1] | D) = \int_\theta P(x[M+1], y[M+1] \theta_i | D)\, d\theta = \int_\theta P(x[M+1], y[M+1] | \theta) \cdot P(\theta|D)\, d\theta$$

$$= \int_{\theta_X} \int_{\theta_{Y|x}} P(x[M+1]|\theta_X) \cdot P(y[M+1] | \theta_{Y|x}, x[M+1]) \, P(\theta_X|D) \, P(\theta_{Y|x}|D)\, d\theta_X \, d\theta_{Y|x}$$

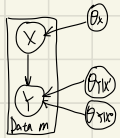
$$= \int_{\theta_X} P(x[M+1]|\theta_X) \, P(\theta_X|D)\, d\theta_X \cdot \int_{\theta_{Y|x}} P(y[M+1] | \theta_{Y|x}, x[M+1]) \, P(\theta_{Y|x}|D)\, d\theta_{Y|x}$$

if the prior for parms of different CPDs are independent

$$P(X[M+1] \cdots X_n[M+1] | D) = \prod_i \int P(X_i[M+1] \mid Pa(X_i)[M+1], \theta_{X_i|Pa(X_i)}) \cdot P(\theta_{X_i|Pa(X_i)}|D)\, d\theta_{X_i|Pa(X_i)}$$

· Local Decomposition

eg.



$$P(y[m] = y \mid x[m], \theta_{Y|x}, \theta_{Y|x'}) = \begin{cases} \theta_{y|x} & x[m] = x \\ \theta_{y|x'} & x[m] = x' \end{cases}$$

$\theta_{Y|x}$ and $\theta_{Y|x'}$ are dependent given the y (not given x)

$\theta_{Y|x}$ and $\theta_{Y|x'}$ are independent given the data

$P(\theta_{Y|x}|D) = P(\theta_{Y|x'}|D) \cdot P(\theta_{Y|x}|D)$

$$P(\theta, D) = P(\theta) \cdot P(D|\theta)$$
$$= P(\theta_X) \cdot P(\theta_{Y|x}) \cdot \prod_{m=1}^{M} P(x[m]|\theta_X) \cdot P(y[m]|x[m], \theta_{Y|x})$$
$$= P(\theta_X) \cdot L_X(\theta_X; D) \cdot P(\theta_{Y|x^1}) \cdot \prod_{m: x[m]=x^1} P(y[m]|x[m], \theta_{Y|x^1}) \cdots P(\theta_{Y|x^k}) \cdot \prod_{m: x[m]=x^k} P(y[m]|x[m], \theta_{Y|x^k})$$

if $\theta \sim \text{Dirichlet}(\alpha_1, \dots \alpha_k)$ $\quad P(\theta) = \prod_k \theta_k^{\alpha_k - 1}$ $\quad P(\theta_{Y|x^k}) = \prod_i \theta_{Y|x^k(i)}^{\alpha_{Y|x^k(i)} - 1}$

$$P(\theta_{Y|x}|D) = \int_{\theta_X} \int_{\theta_{Y|x}} P(\theta_{Y|x}|D) \cdot P(\theta_X|D) \cdot P(\theta_{Y|x^k}|D)\, d\theta_X\, d\theta_{Y|x^k}$$
$$= \int_{\theta_X} \int_{\theta_{Y|x^k}} P(\theta|D)\, d\theta_X\, d\theta_{Y|x^k}$$
$$\propto P(\theta_{Y|x^k}) \prod_{m: x[m]=x^k} P(y[m]|x[m], \theta_{Y|x^k})$$
$$= \left(\prod_i \theta_{Y|x^k(i)}^{\alpha_{Y|x^k(i)} - 1} \right) \cdot \theta_{y^1|x^k}^{M[x^k, y^1]} \cdots \theta_{y^l|x^k}^{M[x^k, y^l]}$$
$$= \text{Dirichlet}\left(\alpha_{y^1|x^k} + M[x^k, y^1], \dots \alpha_{y^l|x^k} + M[x^k, y^l] \right)$$

let X be a variable with parents U, $P(\theta_{X|u})$ satisfies local parameter independence if
$P(\theta_{X|u}) = \prod_{u' \in \text{val}(U)} P(\theta_{X|u'})$

If the prior $P(\theta)$ satisfies global and local parameter independence, then
$P(\theta|D) = \prod_i \prod_{u \in \text{val}(pa_{X_i})} P(\theta_{X_i|u}|D)$

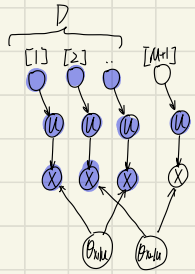
$P(\theta|D) = \prod_i P(\theta_{X_i|pa_{X_i}}|D)$ // global parameter independence
$= \prod_i \prod_{u \in \text{val}(pa_{X_i})} P(\theta_{X_i|u}|D)$ // local parameter independence

If $P(\theta_{X|u})$ is a Dirichlet distribution $P(\theta_{X|u}) \propto \prod_{k \in \text{val}(X)} \theta_{X|u(k)}^{\alpha_{X|u(k)} - 1}$
then $P(\theta_{X|u}|D)$ is a Dirichlet distribution $P(\theta_{X|u}) \propto \prod_k \theta_{X|u(k)}^{\alpha_{X|u(k)} + M[u, x_k]}$

If U is a parent of X $\quad \theta_{X|u} \sim \text{Dirichlet}(\alpha_{x_1|u}, \dots \alpha_{x_k|u})$, D is complete data set over $\mathcal{X}$

$$P(X[M+1] = x_i \mid U[M+1] = u, D) = \frac{\alpha_{x_i|u} + M[x_i, u]}{\sum_j (\alpha_{x_j|u} + M[x_j, u])}$$

$$P(X[M+1] = x_i \mid U[M+1] = u, D) = \int_{\theta_{X|u} \in [0,1]} P(X[M+1] = x_i, \theta_{X|u} \mid U[M+1] = u, D)\, d\theta_{X|u}$$
$$= \int P(X[M+1] = x_i \mid \theta_{X|u}, U[M+1] = u, D) \cdot P(\theta_{X|u} \mid U[M+1] = u, D)\, d\theta_{X|u}$$
$$= \int \underbrace{P(X[M+1] = x_i \mid \theta_{X|u}, U[M+1] = u)}_{(X[M+1] \perp D \mid \theta_{X|u}) \text{ } \theta \text{ blocks the trail}} \cdot \underbrace{P(\theta_{X|u} \mid D)}_{\text{unknown } X[m] \text{ blocks the V-structure}} d\theta_{X|u}$$
$$= \int \theta_{X|u} \frac{P(\theta_{X|u}) P(D|\theta_{X|u})}{P(D)}\, d\theta_{X|u}$$



$$= \frac{\gamma}{P(D)} \int \theta_{x|u} \prod_j \theta_{y|u}^{\alpha_{y|u}-1} \prod_j \theta_{y|u}^{M[y,u]} \, d\theta_{x|u}$$

$$P(D|\theta_{x|u}) \begin{cases} \nearrow P(\{D|U=u\}) \to \text{chain rule} \\ \searrow P(\{D|U\neq u\}) \to \text{chain rule} \end{cases}$$

$$= \frac{\gamma}{P(D)} \int \theta_{x|u}^{\alpha_{x|u}+M[x,u]} \cdot \prod_{j\neq i} \theta_{y|u}^{\alpha_{y|u}+M[y,u]} \, d\theta_{x|u}$$

// integrating a polynomial over a simplex

$$= \frac{\gamma}{P(D)} \cdot \frac{\Gamma(\alpha_{x|u}+M[x,u]+1) \cdot \prod_{j\neq i} \Gamma(\alpha_{y|u}+M[y,u])}{\Gamma(\alpha_{x|u}+M[x,u]+1+\sum_{j\neq i}(\alpha_{y|u}+M[y,u]))}$$

$$= \frac{1}{Z} \alpha_{x|u} + M[x,u]$$

$$P(X[M+1]=x_i \mid U[M+1]=u, D) = \frac{\alpha_{x_i|u} + M[x_i,u]}{\sum_j \alpha_{y|u} + M[y,u]}$$

// Integrating over a Simplex

let $S_n = \{(\theta_1 \cdots \theta_n) \mid \theta_i \geq 0, \ \sum \theta = 1\}$

$$\int_{S_n} \theta_1^{k_1} \cdots \theta_n^{k_n} \, d\theta$$

$$I(t) = \int_{[0,\infty)^n} \theta_1^{k_1} \cdots \theta_n^{k_n} \, \delta(t-\theta_1-\cdots-\theta_n) \, d\theta$$

Laplace transform: $\mathcal{L}(s) = \int_0^{+\infty} I(t) e^{-st} \, dt$

$$= \int_0^{+\infty} \int_{[0,\infty)^n} \theta_1^{k_1} \cdots \theta_n^{k_n} \, \delta(t-\theta_1-\cdots-\theta_n) \, d\theta \; e^{-st} \, dt$$

$$= \int_{[0,\infty)^n} \theta_1^{k_1} \cdots \theta_n^{k_n} \int_t^{+\infty} \delta(t-\theta_1-\cdots-\theta_n) \, e^{-st} \, dt \, d\theta$$

$$= \int_{[0,\infty)^n} \theta_1^{k_1} \cdots \theta_n^{k_n} \exp(-s(\theta_1+\cdots+\theta_n)) \, d\theta$$

$$= \prod_{i=1}^n \int_0^{+\infty} \theta^{k_i} \exp(-s\theta) \, d\theta$$

$$= \prod_{i=1}^n \frac{k_i!}{s^{k_i+1}}$$

$$= \frac{\prod_{i=1}^n k_i!}{s^{\sum k_i + n}}$$

$$= \frac{(\sum k_i + n)!}{s^{\sum k_i + n}} \cdot \frac{\prod_{i=1}^n k_i!}{(\sum k_i + n)!}$$

$$= \frac{\prod_{i=1}^n k_i!}{(\sum_i k_i + n)!} \cdot \mathcal{L}\left(t^{\sum k_i + n}\right)(s)$$

$$\int_\theta \theta_1^{k_1} \cdots \theta_n^{k_n} \, d\theta = I(1) = \frac{\prod_{i=1}^n k_i!}{(\sum_i k_i + n)!} = \frac{\prod_{i=1}^n \Gamma(k_i+1)}{\Gamma(\sum_i (k_i+1))}$$

$$\int_\theta \theta_1^{k_1} \cdots \theta_n^{k_n} \, d\theta = \frac{\prod_{i=1}^n \Gamma(k_i)}{\Gamma(\sum_i k_i)}$$

$$\int_0^{+\infty} \theta^k \exp(-s\theta) \, d\theta$$
$$= \frac{-1}{s} \int_0^{+\infty} \theta^k \, d\exp(-s\theta)$$
$$= -\frac{1}{s} \left[\theta^k \exp(-s\theta) \Big|_0^{+\infty} - k \int_0^{+\infty} \exp(-s\theta) \cdot \theta^{k-1} \, d\theta \right]$$
$$= \frac{k}{s} \int_0^{+\infty} \theta^{k-1} \exp(-s\theta) \, d\theta$$
$$= \frac{k!}{s^{k+1}}$$
$$\mathcal{L}(\theta^k)(s) = \frac{k!}{s^{k+1}}$$

- **Priors for Bayesian Networks**

  In a Bayesian network $\mathcal{B}$ over $\mathcal{X}$, for each var $U \to X \in \mathcal{B}$, X has a set of multinomial distributions $\theta_{X|u}$ $\forall u \in \text{val}(U)$

  Can set the prior of $\theta_{X|u}$ as a Dirichlet prior

  $\theta_{X|u} \sim \text{Dirichlet}(\#[x_1,u], \cdots \#[x_n,u])$

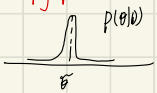
· MAP Estimation

the parameters that maximize the posterior probability

$\tilde{\theta} = \arg\max_\theta \log P(\theta|D)$

when we have a large amount of data, the posterior is often sharply peaked around $\tilde{\theta}$

$P(x_{[N+1]}|D) = \int P(x_{[N+1]}|\theta)\, P(\theta|D)\, d\theta \approx P(x_{[N+1]}|\tilde{\theta})$



$\arg\max_\theta \log p(\theta|D) = \arg\max_\theta \log \frac{p(\theta)\,p(D|\theta)}{p(D)}$

$= \arg\max_\theta \log p(\theta) + \log p(D|\theta)$ use the prior to provide regularization over log-likelihood.

$= \arg\max_\theta \log p(\theta) + \sum_{x\in D} \log p(x|\theta)$ when D is a big dataset, $\log p(\theta)$ is negligible

eg. Beta prior and Bernoulli model

$p(\theta; \alpha_0, \alpha_1) = c\cdot \theta^{\alpha_1 - 1}(1-\theta)^{\alpha_0 - 1}$

$p_B(x) = (1-\theta)\exp\langle \mathbb{I}\{x=1\}, \ln\frac{\theta}{1-\theta}\rangle$

let $\eta = \ln\frac{\theta}{1-\theta}$ $\quad \theta = \frac{1}{1+e^{-\eta}}$ $\quad 1-\theta = \frac{1}{1+e^{\eta}}$

$p(\eta) = p(\theta(\eta))\left|\frac{d\theta}{d\eta}\right|$ $= \frac{d\theta}{d\eta}\, p(\theta(\eta)) = 1\cdot\frac{1}{(1+e^{-\eta})^2}\cdot e^{-\eta}\cdot\frac{1}{c}\cdot\left(\frac{1}{1+e^{-\eta}}\right)^{\alpha_1-1}\left(\frac{1}{1+e^{\eta}}\right)^{\alpha_0-1}$

$= \frac{1}{1+e^{-\eta}}\cdot\frac{e^{-\eta}}{1+e^{-\eta}}\cdot c\cdot\left(\frac{1}{1+e^{-\eta}}\right)^{\alpha_1-1}\left(\frac{1}{1+e^{\eta}}\right)^{\alpha_0-1}$

$= c\cdot\left(\frac{1}{1+e^{-\eta}}\right)^{\alpha_1}\left(\frac{1}{1+e^{\eta}}\right)^{\alpha_0}$

when $\alpha_0 = \alpha_1 = 1$, $p(\theta) = 1$ for $\theta\in[0,1]$, uniform

but $p(\eta) = \frac{1}{1+e^{\eta}}\cdot\frac{1}{1+e^{-\eta}} = \frac{1}{1+1+e^{\eta}+e^{-\eta}} = \frac{1}{2+e^{\eta}+e^{-\eta}}$ is far from uniform

both MLE and Bayesian estimation are "representation-independent". (by change of variable in the integral)

but MAP estimation is not

$\tilde{\theta} = \arg\max_\theta \log p(\theta) = \arg\max_\theta (\alpha_1-1)\log\theta + (\alpha_0-1)\log 1-\theta = \frac{\alpha_1-1}{\alpha_0+\alpha_1-2}$

$\tilde{\eta} = \arg\max_\eta p(\eta) = \log\frac{\alpha_1}{\alpha_0}$ $\quad \theta(\tilde{\eta}) = \frac{\alpha_1}{\alpha_0+\alpha_1}$ $\quad \tilde{\theta} \neq \theta(\tilde{\eta})$

MAP estimation is more sensitive to choices in formalizing the likelihood and the prior than MLE or Bayesian inference