概率图模型理论及应用

Theory and Applications of Probabilistic Graphical Models

(Lesson 13)

欧智坚

清华大学电子工程系

Addr: 罗姆楼 6-104

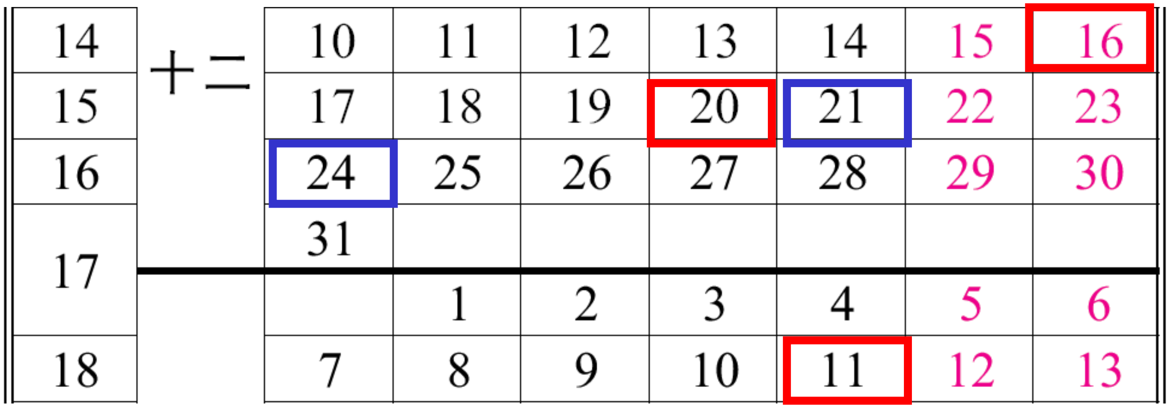
Tel: 62796193

Email: ozj@tsinghua.edu.cn

电子系海外学者短期讲学课程

- 语音处理中的机器学习（Machine Learning for Speech Processing）
  - 2012年12月17日—21日；
  - 周一：15:00-17:00；
  - 周二：9:30-11:30, 15:00-17:00；
  - 周三：15:00-17:00；
  - 周四：9:30-11:30, 15:00-17:00；
  - 周五：9:30-11:30, 15:00-17:00；
  - 电子系罗姆楼8-208
- Dr. Shinji Watanabe，研究员，Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA
- 发邮件给wb.th08@gmail.com，或于第一次课现场报名

通知 @14th week（严格时间点）



14	十二	10	11	12	13	14	15	16
15		17	18	19	20	21	22	23
16		24	25	26	27	28	29	30
17		31						
			1	2	3	4	5	6
18		7	8	9	10	11	12	13

- 第14周周末（12月16日）23:59前：每位同学递交评估版报告
  - 按要求的（中文）模板书写
- 第15周周四（12月20日）23:59前：每位同学返回互评表
  - 书写清晰, 新意及深入程度, 工作量及完善程度
- 第15周周五（12月21日）23:59前：网络学堂上公布选做口头报告的同学名单
- 第16周周一（12月24日）the last lesson：口头报告
- 课程大作业提交截止
  - 现场检查时间：1月11日，每人10分钟ppt汇报（含演示）

课程章节

- 第一章 引言（**1**）
- 第二章 图模型的表示理论（**3**）
  - **DGM-UGM**
  - **Semantics**
  - **HMM-CRF**
- 第三章 图模型的推理理论（**6**）
  - 精确推理：**variable-elimination**，**cluster-tree**，**triangulate**
  - 连续变量：**Kalman**
  - 采样近似：**sampling**
  - 变分近似：**variational**
- 第四章 图模型的学习理论（**3**）
  - 参数学习：**maxlikelihoodEstimate**，**BayesEstimate**
  - 结构学习：**StructureLearning**
- 第五章 一个综合例子（1）

引言

- 假设我们所关心变量 x 的联合分布由概率图来表达:

$$p\left(x \mid g, \theta^g\right)$$

  - g 表示结构假设（structure hypothesis），假设 x 的联合分布可以依图 g 分解
  - θ^g 表示图中的参数

- 基于独立同分布样本集 $D = (x[1], \ldots, x[M])$
  - 估计出 g：结构学习
  - 固定 g 估计出 θ^g：参数学习

Structure learning: how ?

Two approaches

① Constraint-based approach

- 从数据出发做CI检验（$X_1 \perp X_2 | X_3$ ？）
- 构造出与CI检验结果相符的结构

1、如何做CI检验（ CI test ）？

$$I(x_1, x_2 | x_3) = \sum_{x_1, x_2, x_3} p(x_1, x_2, x_3) \log \frac{p(x_1, x_2 | x_3)}{p(x_1 | x_3) p(x_2 | x_3)} \mathbf{= 0 \, ?}$$

2、检验的顺序如何选择？一旦某个CI检验有问题，引发连锁错误。

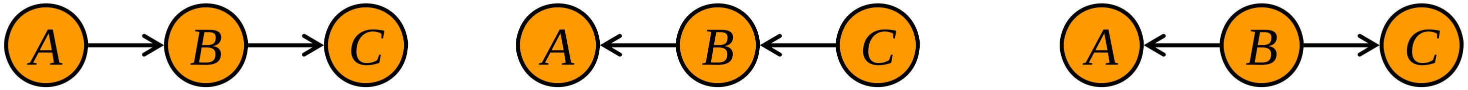
局部调整/选择, e.g. “Switching auxiliary chains for speech recognition”, IEEE SPL 2007.

② Score-and-search approach

- 定义一个得分函数(一种准则函数)，评估一个结构与数据的匹配程度（BIC得分，贝叶斯得分）
- 然后搜索有最大得分的结构

结构的可辨识性(Identifiability)

- 我们称两个BN结构是独立等价的（*independence equivalent*，I-equivalent），如果它们代表完全相同的条件独立性。

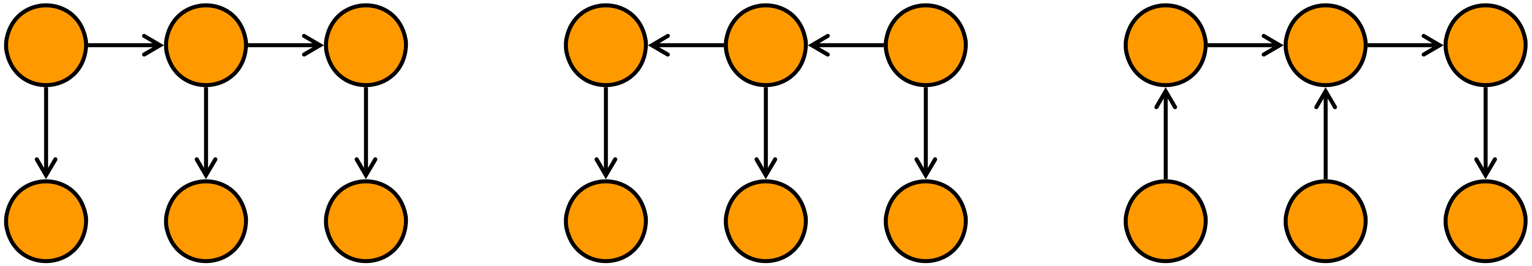


I-equivalent Bayes Nets：represent $A \perp C \mid B$

  - 另一方面，这些独立等价的bn确实代表了不同的因果模型，是 A 影响 B，还是 B 影响 A ？
  - 运用知识，我们才能分辩那种因果关系是对的。

结构的可辨识性(Identifiability)

- 定理：两个bn结构是独立等价的，当且仅当它们具有相同的无向图版本和相同的v结构。



  - 在结构学习中，通常对一个结构假设 g 关联一个结构等价类，而不是单一的一个结构；
  - 学习得到一个结构等价类。

Structure learning

— from complete data

Focus on score-and-search approach

结构与数据的匹配程度

Likelihood score

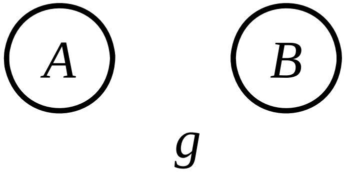
- IID样本集 $D = (x[1],\ldots, x[M])$ 下似然函数

$$\max_{\theta^g} p\left(D \mid g, \theta^g\right) = \prod_{m=1}^{M} p\left(x[m] \mid g, \theta^g\right)$$

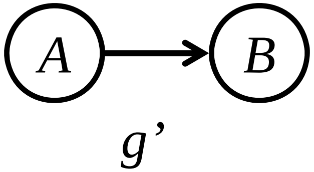
- 结构*g*的似然得分

$$\underbrace{\text{最大似然参数估计}: \theta^g_{ML} = \arg\max_{\theta^g} p\left(D \mid g, \theta^g\right)}$$

$$likelihood\,(g : D) \triangleq \max_{\theta^g} \log p\left(D \mid g, \theta^g\right) = \log p\left(D \mid g, \theta^g_{ML}\right)$$



g



g'

  - 引理：*g* 添加边得到更复杂结构 *g'*，$likelihood(g' : D) \geq likelihood(g : D)$

$$\max_{\theta^{g'}} p\left(D \mid g', \theta^{g'}\right) \geq \max_{\theta^g} p\left(D \mid g, \theta^g\right)$$

  - 全连接结构具有最大的似然得分
  - 似然得分不适合于做模型选择

Bayesian score

- Bayesian approach: 将未知量视为随机变量
  - 将 g 视为一个随机变量
  - 给定数据 D 下结构 g 的后验分布

$p(g)$: 结构先验分布, e.g. $p(g) \propto c^{|g|}$, $c<1$

$$\max_{g} p(g \mid D) = \frac{p(D \mid g)\, p(g)}{p(D)}$$

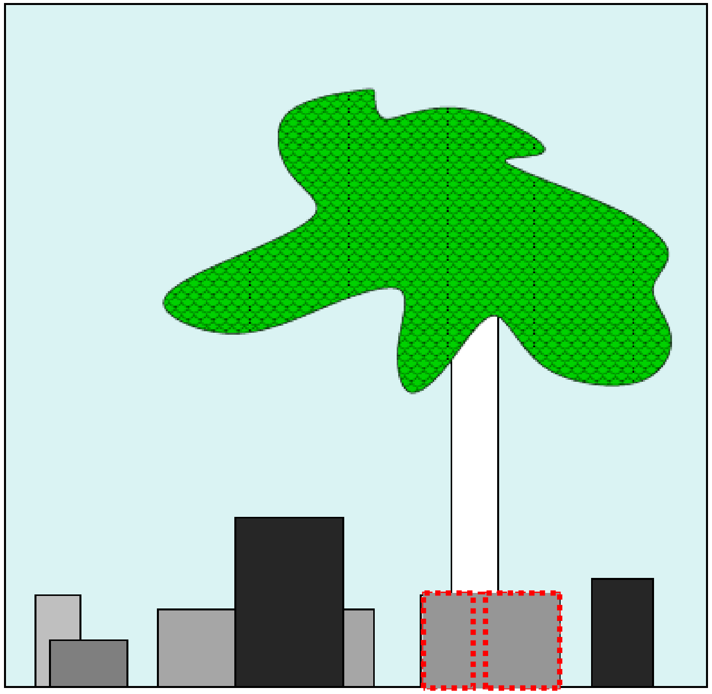
$p(D \mid g)$: marginal likelihood

$$p(D \mid g) = \int p(D \mid g, \theta^g)\, p(\theta^g \mid g)\, d\theta^g$$

- Bayesian score: $Bayes(g : D) \triangleq \log p(D \mid g) + \log p(g)$

贝叶斯模型选择

Occam’s Razor : 接受能拟合数据的最简单模型



树后是一个盒子，

还是两个盒子?

- 贝叶斯模型选择 体现 奥克姆剃须刀原理

贝叶斯模型选择 体现 奥克姆剃须刀原理

- 接受能拟合数据的最简单解释

Data: $D = \{ x[1], x[2], \ldots x[M] \}$

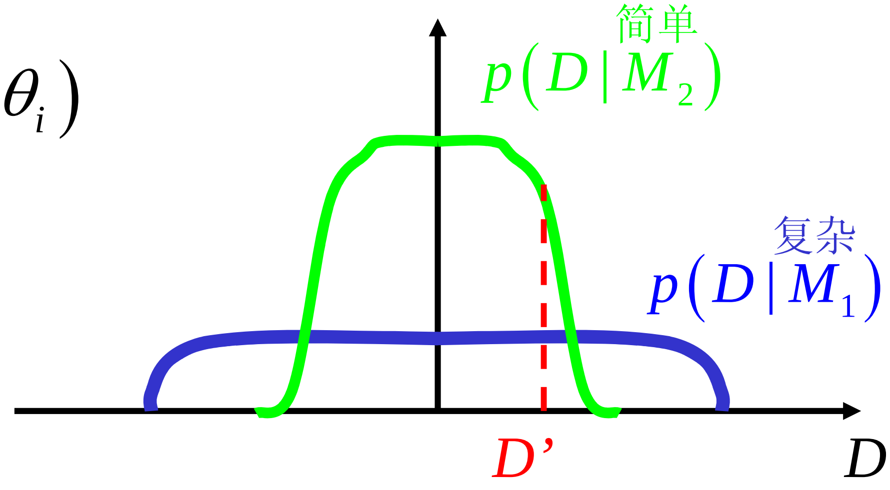
Models $M_i : p(\theta_i | M_i), p(x | M_i, \theta_i)$

贝叶斯模型选择

$$\max_{M_i} p(M_i | D) = \frac{p(M_i)\, p(D | M_i)}{p(D)}$$

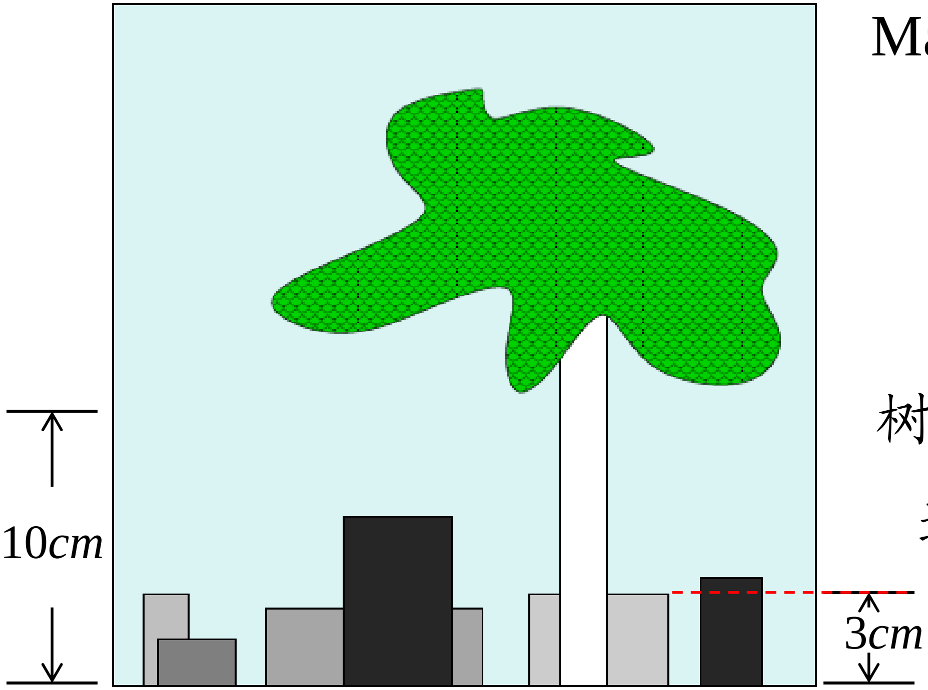
$$\max_{M_i} \; p(D | M_i)$$

Marginal likelihood



贝叶斯模型选择 体现 奥克姆剃须刀原理

Occam’s Razor：接受能拟合数据的最简单模型



Mackay书, Section 28

树后是一个盒子，M_1, α

还是两个盒子？M_2, β_1, β_2

$$p\left(D \mid M_1\right)=\int p\left(D \mid M_1, \alpha\right) \overset{=3cm}{p\left(\alpha \mid M_1\right)} d\alpha \quad =0.1$$

$$p\left(D \mid M_2\right)=\int p\left(D \mid M_2, \beta_1, \beta_2\right) \overset{=3cm}{p\left(\beta_1 \mid M_2\right)} \overset{=3cm}{p\left(\beta_2 \mid M_2\right)} d\beta_1 d\beta_2 \quad =0.01$$

Computation of the marginal likelihood

- Marginal likelihood for Multinomial BN Has closed form
  - 在完备数据、全局参数独立、局部参数独立条件下，
  - 对结点 n 的父结点集 pa_n 的每个可能取值组合 i，有一个多元分布 $p(x_n \mid pa_n=i,\theta_{n,i})$

$$p(D|g)=\frac{p\left(\theta^g|g\right)p\left(D|g,\theta^g\right)}{p\left(\theta^g|g,D\right)}=\frac{\prod_n\prod_i p\left(\theta_{n,i}\right)\cdot\prod_n\prod_i p\left(D_{n,i}|g,\theta_{n,i}\right)}{\prod_n\prod_i p\left(\theta_{n,i}|g,D_{n,i}\right)}$$

$$=\prod_n\prod_i p\left(D_{n,i}|g\right)$$

Marginal likelihood for multinomial $p(x_n \mid pa_n=i)$

possible value of x_n $1 \le k \le K_n$

$$p(D|g)=\prod_{n=1}^{N}\prod_{i}\frac{\Gamma\left(\alpha_{n,\,i}\right)}{\Gamma\left(\alpha_{n,\,i}+N_{n,\,i}\right)}\prod_{k=1}^{K_n}\frac{\Gamma\left(\alpha_{n,\,i,\,k}+N_{n,\,i,\,k}\right)}{\Gamma\left(\alpha_{n,\,i,\,k}\right)}$$

Node n

possible value of pa_n

Computation of the marginal likelihood

- Marginal likelihood for Multinomial Has closed form

- Likelihood for data D
  with sufficient statistics $(N_1,\ldots, N_K)$

$$p(D|\theta)=\prod_k \theta_k^{N_k}$$

- Dirichlet prior over parameters

$$p(\theta)=\frac{\Gamma\left(\sum \alpha_k\right)}{\prod \Gamma(\alpha_k)}\prod_{k=1}^{K}\theta_k^{\alpha_k-1}$$

- Parameter posterior

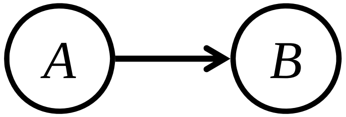
$$p(\theta|D)=\frac{\Gamma\left(\sum(\alpha_k+N_k)\right)}{\prod \Gamma(\alpha_k+N_k)}\prod_{k=1}^{K}\theta_k^{\alpha_k+N_k-1}$$

- Note that

$$p(D)=\frac{p(\theta)\,p(D|\theta)}{p(\theta|D)}=\frac{\Gamma\left(\sum \alpha_k\right)}{\prod \Gamma(\alpha_k)}\frac{\prod \Gamma(\alpha_k+N_k)}{\Gamma\left(\sum(\alpha_k+N_k)\right)}$$

$D = (x[1],\ldots, x[M]),\ x[m]\in\{1, 2, \ldots, K\},\ m=1,\ldots, M$

Example: Multinomial BN



$$p(D \mid g) = \text{Marginal likelihood for multinomial } p(A)$$

$$\times \text{Marginal likelihood for multinomial } p(B \mid A=0)$$

$$\times \text{Marginal likelihood for multinomial } p(B \mid A=1)$$

$$p(D \mid g) = \text{Marginal likelihood for multinomial } p(A)$$

$$\times \text{Marginal likelihood for multinomial } p(B)$$

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

$$p(D) = \frac{\Gamma\left(\sum \alpha_k\right)}{\prod \Gamma(\alpha_k)} \frac{\prod \Gamma(\alpha_k + N_k)}{\Gamma\left(\sum(\alpha_k + N_k)\right)}$$

Bayesian score : Large-sample Behavior

- Bayesian score: $Bayes(g:D) \triangleq \log p(D \mid g) + \log p(g)$

- For large amounts of data, i.e., large M

$$\log p(D \mid g) \approx \underbrace{\log p\left(D \mid g, \theta^g_{ML}\right)}_{\text{Likelihood score}} - \underbrace{\frac{\log M}{2} \dim\left(\theta^g\right)}_{\text{Complexity penalty}}$$

- BIC (Bayesian Information Criterion) score
  - BIC得分定义为：大样本下贝叶斯得分的近似

$$BIC(g:D) \triangleq \log p\left(D \mid g, \theta^g_{ML}\right) - \frac{\log M}{2} \dim\left(\theta^g\right)$$

Scoring function - summary

- Key property: Decomposability
  - Score of a Bayesian network g is a sum of scores of families

$$score(g:D)=\sum_{n=1}^{N} score\left(\left(x_n, pa_n\right):D_n\right) \qquad D_n=\begin{pmatrix} x_n[1] \\ pa_n[1] \end{pmatrix},\cdots,\begin{pmatrix} x_n[M] \\ pa_n[M] \end{pmatrix}$$

一、假设各个条件分布 $p(x_1\,|\,pa_1), \ldots, p(x_N\,|\,pa_N)$ 有各自表征参数 $\{\theta_1, \ldots, \theta_N\}$

$$BIC(g:D)\triangleq \log p\left(D\,|\,g,\theta_{ML}^{g}\right)-\frac{\log M}{2}\dim\left(\theta^{g}\right)$$

$$=\sum_{n}\log p\left(D_n\,|\,g,\theta_{n,ML}^{g}\right)-\frac{\log M}{2}\sum_{n}\dim\left(\theta_{n}^{g}\right)$$

二、假设全局参数独立

$$Bayes(g:D)\triangleq \log p\left(D\,|\,g\right)+\log p\left(g\right)$$

$$=\sum_{n}\log p\left(D_n\,|\,g\right)+\log p\left(g\right)$$

Structure search

- Goal: search for the network structure that maximizes the score.

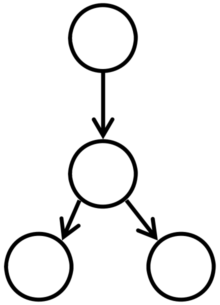
- Theorem:

  Finding maximal scoring structure with at most k parents per node is NP-hard for $k > 1$.

  考虑每个结点的父结点数 k>1 的结构，在这样的结构中搜索最大得分结构是NP难。

- In general, we need to use heuristic search.

Search for tree-structure



- Tree-strcuture: at most one parent per node
  - we can solve the search problem in polynomial time.
- We can write the score as:

$$score(g) = \sum_{n} score(x_n, pa_n)$$

$$= \boxed{\begin{array}{l} \displaystyle\sum_{n:\,|pa_n|>0} score(x_n, pa_n) \\ \displaystyle - \sum_{n:\,|pa_n|>0} score(x_n) \end{array}} \begin{array}{c} + \\ \\ + \end{array} \boxed{\begin{array}{l} \displaystyle\sum_{n:\,|pa_n|=0} score(x_n) \\ \displaystyle\sum_{n:\,|pa_n|>0} score(x_n) \end{array}}$$

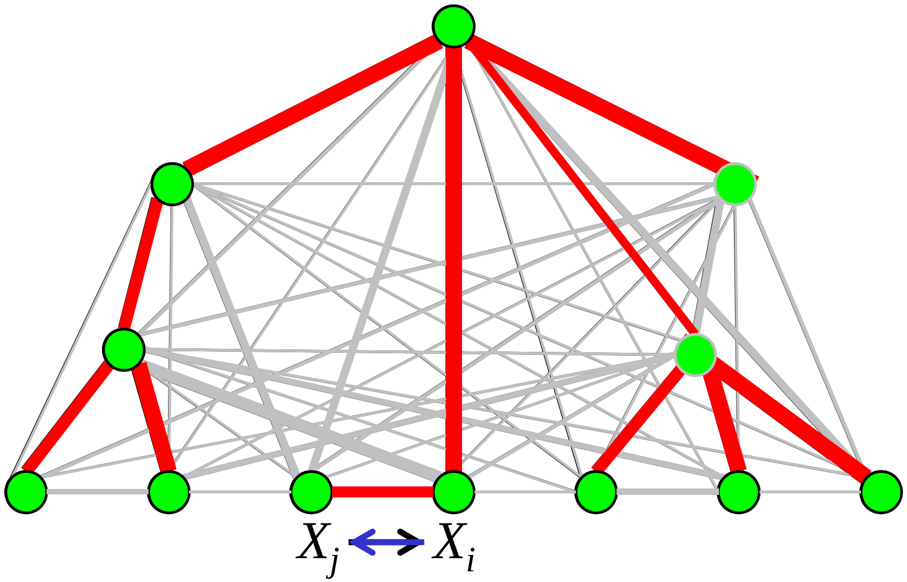
$$= \sum_{n:\,|pa_n|>0} \left\{ score(x_n, pa_n) - score(x_n) \right\} + \sum_{n} score(x_n)$$

change over “empty” network

Score of “empty” network

Score = sum of edge scores + constant

Search for tree-structure



- Algorithm
  - Construct graph with nodes $1, \ldots, N$
  - Set $w(j \sim i) = Score(\ X_j \rightarrow X_i\ ) - Score(X_i)$

    $= Score(\ X_j \leftarrow X_i\ ) - Score(X_j)$
  - Find tree with maximal weight.

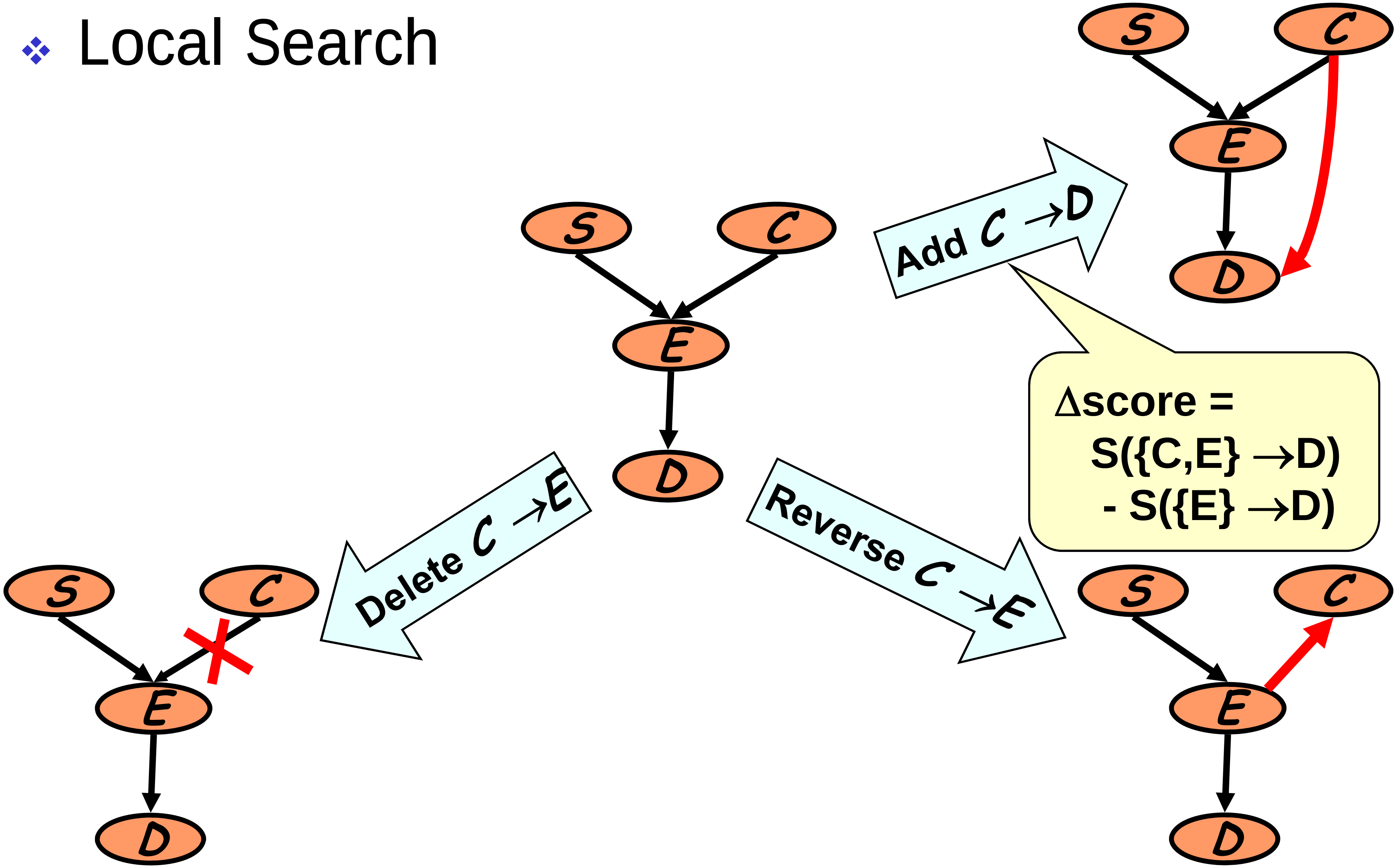
    Standard max spanning tree algorithm — $O(N^2 \log N)$

Beyond trees

- When we consider general DAG, the problem is not easy.
- Need to resort to heuristic search
  - Start with a given network (e.g., the best tree , a random network)
  - **Successive local search (逐步局部搜索)**:
  - Stop when no modification improves score.

在某个局部(某个结点处)修改当前的网络结构(添加、删除边，或改变边的方向)，(利用得分的分解性去)计算网络结构改变所带来的得分差异，以此得到局部看来最好的结构，然后再进行下一处局部搜索

- Local Search



Δscore =
S({C,E} →D)
- S({E} →D)

Structure learning

— from incomplete data

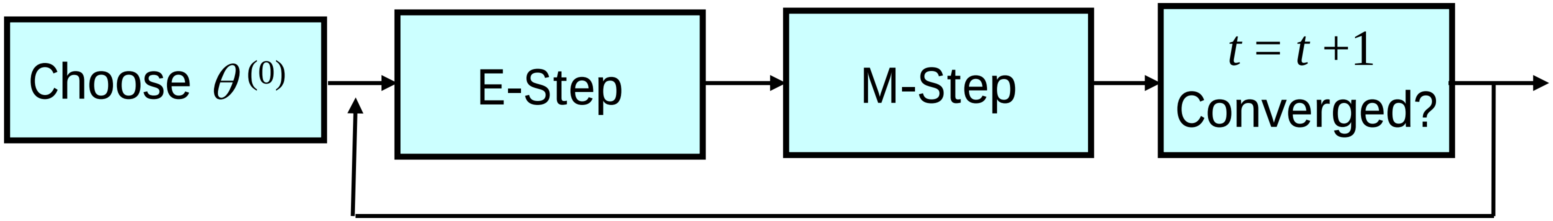
Focus on score-and-search approach

EM algorithm (review)

- 总体分布 $p(x, z \mid \theta)$
  - 总体分布的IID 样本集 $D = (x[1],\ldots, x[M])$： 观测数据
    $H = (z[1],\ldots, z[M])$
  - 目标: $\theta_{ML} = \arg\max_{\theta} \log p\left(D \mid \theta\right)$
- 定义一个辅助函数 $Q\left(\theta \mid \theta^{(old)}\right) = E_{p\left(H \mid \theta^{(old)}, D\right)}\left[\log p\left(D, H \mid \theta\right)\right]$

$$\theta^{(new)} = \arg\max_{\theta} Q\left(\theta \mid \theta^{(old)}\right)$$

$$p\left(D \mid \theta^{(new)}\right) \geq p\left(D \mid \theta^{(old)}\right)$$



Choose $\theta^{(0)}$

E-Step

M-Step

$t = t + 1$

Converged?

Why processing complete data is ‘easy’ ?

- With complete data, BICScore of a network decomposes.

$$BIC(g:D) \triangleq \log p\left(D \mid g, \theta_{ML}^{g}\right) - \frac{\log M}{2}\dim\left(\theta_{ML}^{g}\right)$$

$$= \sum_{n=1}^{N}\left\{\log p\left(D_n \mid g, \theta_{ML,n}^{g}\right) - \frac{\log M}{2}\dim\left(\theta_{ML,n}^{g}\right)\right\}$$

- With incomplete data, we loose decomposability of score.

$$\theta_{ML}^{g} = \arg\max_{\theta^g}\left[\log p\left(D \mid g, \theta^g\right)\right] = \arg\max_{\theta^g}\left[\log \sum_{H} p\left(D, H \mid g, \theta^g\right)\right]$$

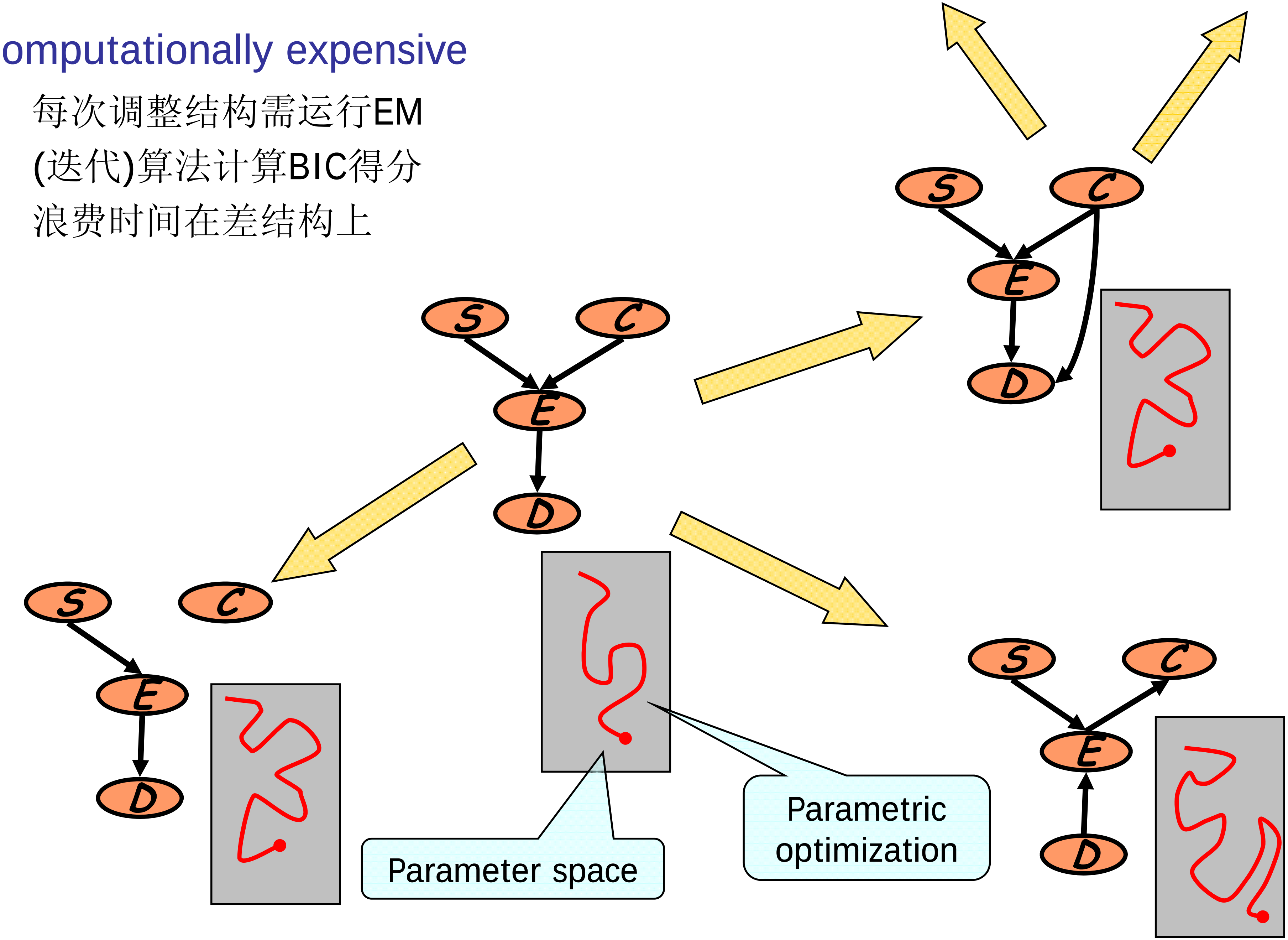
$$BIC(g:D) \triangleq \log p\left(D \mid g, \theta_{ML}^{g}\right) - \frac{\log M}{2}\dim\left(\theta_{ML}^{g}\right)$$

需要使用EM迭代算法去计算一个结构 g 的BIC得分

Naive approach

Computationally expensive

- 每次调整结构需运行EM (迭代)算法计算BIC得分
- 浪费时间在差结构上



固定结构，EM迭代

$$BIC(g:D) \triangleq \log p\left(D \mid g, \theta_{ML}^{g}\right) - \frac{\log M}{2}\dim\left(\theta_{ML}^{g}\right)$$

$$\max_{\theta^g} BIC\left(g, \theta^{g} : D\right) \triangleq \log p\left(D \mid g, \theta^{g}\right) - \frac{\log M}{2}\dim\left(\theta^{g}\right)$$

不完备数据的广义BIC得分

$$\max_{\theta^g} BIC\left(g, \theta^{g} : D\right) \qquad \left(g, \theta^{(0)}\right) \to \cdots \to \left(g, \theta^{(t)}\right) \to \left(g, \theta^{(t+1)}\right) \to$$

EM迭代

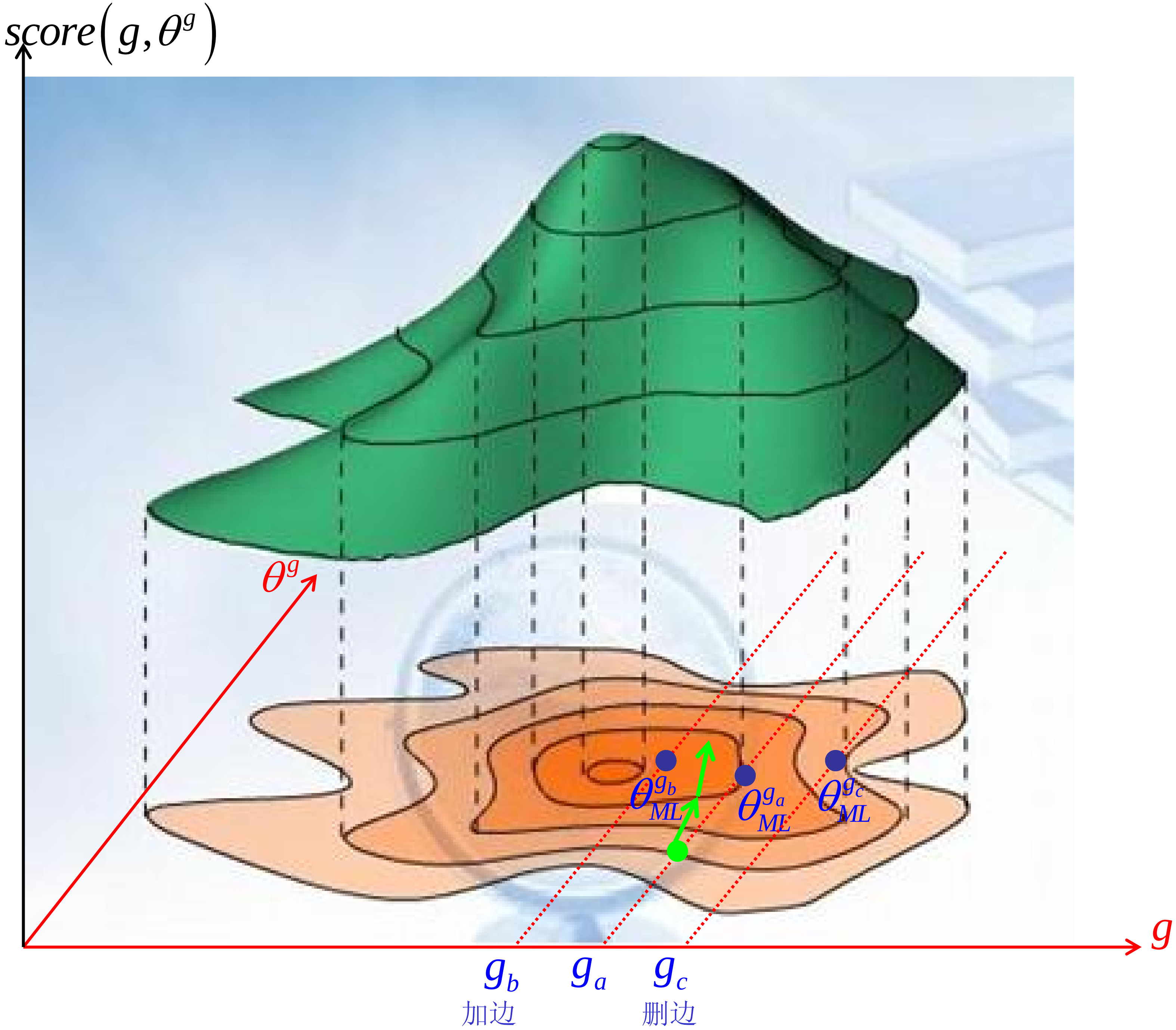
- 定义一个辅助函数

完备数据的广义BIC得分

$$Q\left(g, \theta \mid g, \theta^{(old)}\right) \triangleq E_{p\left(H \mid g, \theta^{(old)}, D\right)}\left[BIC\left(g, \theta : D, H\right)\right]$$

$$\theta^{(new)} = \arg\max_{\theta} Q\left(g, \theta \mid g, \theta^{(old)}\right)$$ **一步EM**

$$BIC\left(g, \theta^{(new)} : D\right) \geq BIC\left(g, \theta^{(old)} : D\right)$$


$score\left(g,\theta^g\right)$
θ^g
g
$\theta_{ML}^{g_b}$
$\theta_{ML}^{g_a}$
$\theta_{ML}^{g_c}$
g_b
加边
g_a
g_c
删边

Structural EM (Friedman, ICML97,98)

$$\max_{g} BIC(g:D) \triangleq \log p\left(D\,|\,g,\theta^{g}_{ML}\right) - \frac{\log M}{2}\dim\left(\theta^{g}_{ML}\right)$$

$$\max_{g}\left\{\max_{\theta^{g}} BIC\left(g,\theta^{g}:D\right)\right\} \triangleq \log p\left(D\ \,|\,g,\theta^{g}\right) - \frac{\log M}{2}\dim\left(\theta^{g}\right)$$

不完备数据的广义BIC得分

$$\max_{g,\theta^{g}} BIC\left(g,\theta^{g}:D\right) \qquad \left(g^{(0)},\theta^{(0)}\right)\to\cdots\to\left(g^{(t)},\theta^{(t)}\right)\to\left(g^{(t+1)},\theta^{(t+1)}\right)\to$$

把调整结构与参数EM算法以一种更紧密的方式结合

结构EM迭代

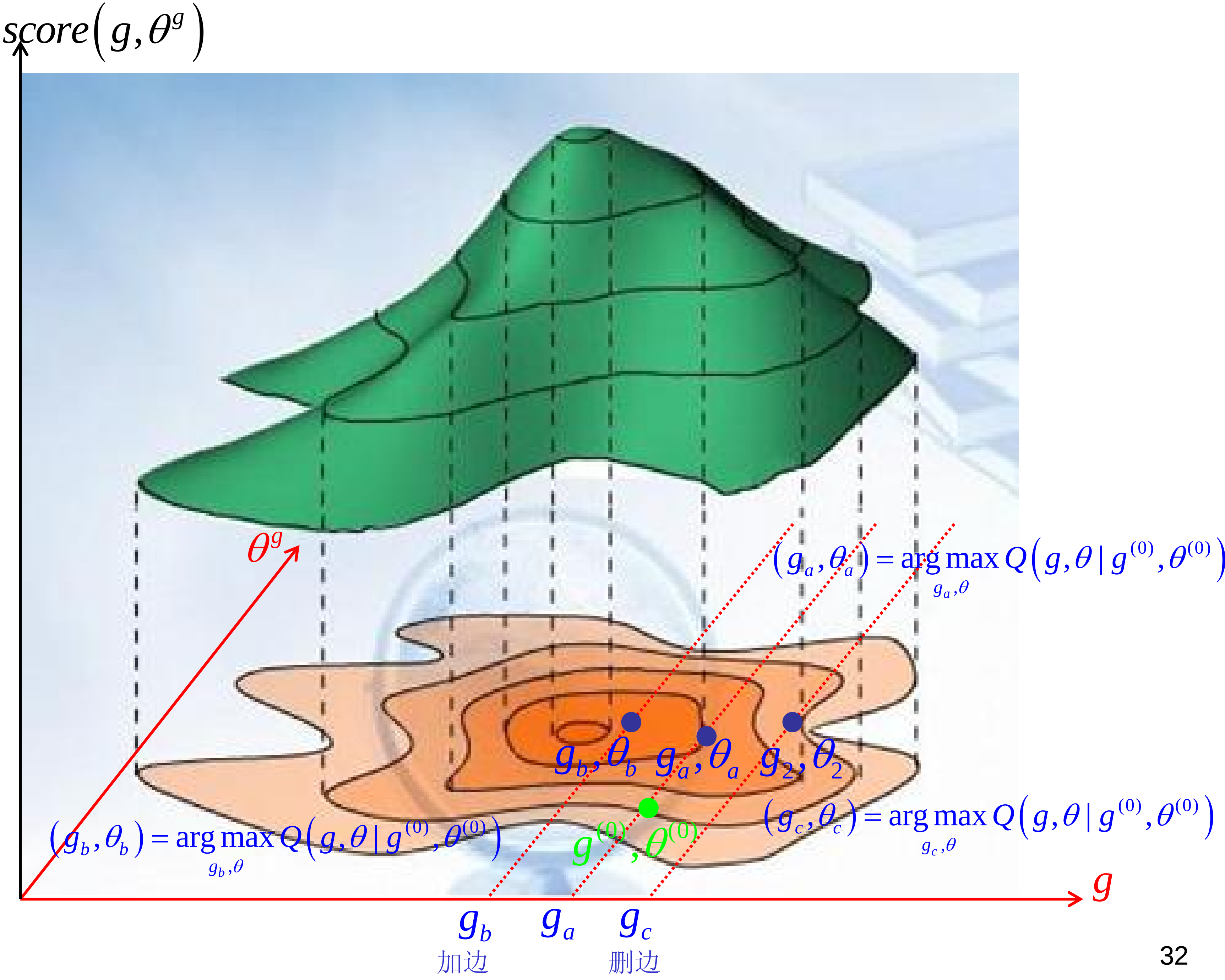
- 定义一个辅助函数

完备数据的广义BIC得分

$$Q\left(g,\theta\,|\,g^{(old)},\theta^{(old)}\right) \triangleq E_{p\left(H|g^{(old)},\theta^{(old)},D\right)}\left[BIC\left(g,\theta:D,H\right)\right]$$

$$\left(g^{(new)},\theta^{(new)}\right) = \arg\max_{g,\theta} Q\left(g,\theta\,|\,g^{(old)},\theta^{(old)}\right)$$ **一步结构EM**

$$BIC\left(g^{(new)},\theta^{(new)}:D\right) \geq BIC\left(g^{(old)},\theta^{(old)}:D\right)$$



$$(g_a,\theta_a)=\underset{g_a,\theta}{\arg\max}\,Q\left(g,\theta\,|\,g^{(0)},\theta^{(0)}\right)$$

$$(g_b,\theta_b)=\underset{g_b,\theta}{\arg\max}\,Q\left(g,\theta\,|\,g^{(0)},\theta^{(0)}\right)$$

$$(g_c,\theta_c)=\underset{g_c,\theta}{\arg\max}\,Q\left(g,\theta\,|\,g^{(0)},\theta^{(0)}\right)$$

Structural EM for BIC score

- 定理

$$BIC\left(g^{(new)},\theta^{(new)}:D\right)-BIC\left(g^{(old)},\theta^{(old)}:D\right)$$
$$\geq Q\left(g^{(new)},\theta^{(new)}\mid g^{(old)},\theta^{(old)}\right)-Q\left(g^{(old)},\theta^{(old)}\mid g^{(old)},\theta^{(old)}\right)$$

- SEM Algorithm
  - Choose $g^{(0)}$, $\theta^{(0)}$ as initial structure and parameters.
  - Loop for $t = 0, 1, \ldots$ until convergence **结构EM迭代**

一步结构EM

Find model $g^{(t+1)}$ with $\theta^{(t+1)}$: $\left(g^{(t+1)},\theta^{(t+1)}\right)=\arg\max_{g,\theta} Q\left(g,\theta\mid g^{(t)},\theta^{(t)}\right)$

结构不变 g_a $\quad \theta_a=\arg\max_{\theta} Q\left(g_a,\theta\mid g^{(t)},\theta^{(t)}\right)$

添加边 g_b $\quad \theta_b=\arg\max_{\theta} Q\left(g_b,\theta\mid g^{(t)},\theta^{(t)}\right)$

删除边 g_c $\quad \theta_c=\arg\max_{\theta} Q\left(g_c,\theta\mid g^{(t)},\theta^{(t)}\right)$

改变边的方向 g_d $\quad \theta_d=\arg\max_{\theta} Q\left(g_d,\theta\mid g^{(t)},\theta^{(t)}\right)$

The learning problem

	Known structure		Unknown structure
Complete data	ML	Bayesian	Learning tree Score-and-search
Incomplete data	ML	Bayesian	Structural EM