Minka 2001 博士论文.

利用贝叶斯公式推断后验:

$$p(x|D)=\frac{p(x,D)}{p(D)}=\frac{p(x,D)}{\int_x p(x,D)dx}$$

计算均值和方差:

$$E[x|D]=\int_x x\,p(x|D)dx=\frac{\int_x x\,p(x,D)dx}{\int_x p(x,D)dx}$$

$$E[x^2|D]=\int_x x^2 p(x|D)dx=\frac{\int_x x^2 p(x,D)dx}{\int_x p(x,D)dx}$$

$$Var(x|D)=E[x^2|D]-E[x|D]^2$$

有多个未知量,计算x的后验:

$$p(x|D)=\frac{\int_{y,z} p(x,y,z,D)dydz}{\int_{x,y,z} p(x,y,z,D)dxdydz}$$

↑ 边缘

注:积分计算难以避免

E[X]的完整形态:

$E_{x\sim p(x)}[X]$ 连续: $\int_x x\,p(x)dx$

离散: $\sum_i x_i\,p(x_i)$

数值积分方法可分为:确定性方法和非确定方法

确定性:ADF, EP, BP

非确定性:MCMC.

ADF: one pass sequential methods.
对观测值顺序敏感,对比batch训练数据采用不一致.

Expection propagation 是ADF算法的扩展,
EP是loopy BP的一种扩展.

Chapter 2. 对积分的逼近计算

方法1:高斯逼近

$$I=\int_A f(x)dx$$

$$\hat{I}=\sum_{i=1}^{n} w_i f(x_i)$$

方法2:变分估界

如:

$$ELBO=\log p(x)$$
$$=\log\int_z p(x,z)dz$$
$$=\log\int_z q(z)\frac{p(x,z)}{q(z)}dz$$
$$=\log E_{z\sim q(z)}\left[\frac{p(x,z)}{q(z)}\right]$$
$$\geq E_{z\sim q(z)}\left[\log\frac{p(x,z)}{q(z)}\right]$$
$$=E_{z\sim q(z)}[\log p(x,z)]-E_{z\sim q(z)}[\log q(z)]$$

Minka 2001 推出

$$P(x|D) = \frac{P(x,D)}{P(D)} = \frac{P(x,D)}{\int_x P(x,D)\,dx}$$

$$E[x|D] = \int_x x\,P(x|D)\,dx = \frac{\int_x x\,P(x,D)\,dx}{\int_x P(x,D)\,dx} \quad (1,2)$$

$$E[x^2|D] = \int_x x^2 P(x|D)\,dx = \frac{\int_x x^2 P(x,D)\,dx}{\int_x P(x,D)\,dx} \quad (1,3)$$

$$P(x|D) = \frac{\int_{y,z} P(x,y,z,D)\,dy\,dz}{\int_{x,y,z} P(x,y,z,D)\,dy\,dz\,dx} \quad (1,4)$$

（当有3个随机变量 x, y, z 时）

边缘 Prob.

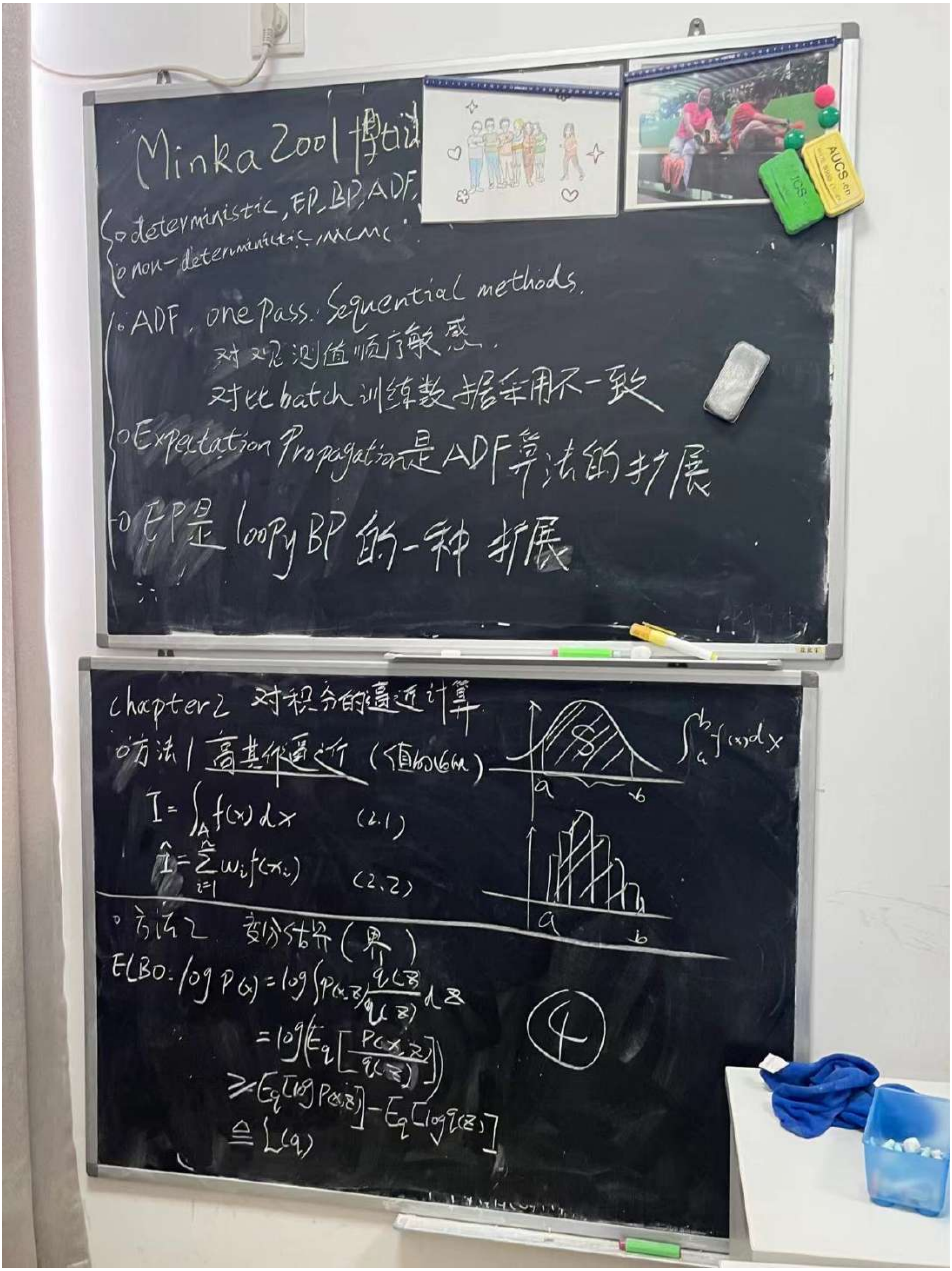
注：积分计算可以避免。

E[X]

完整形态：$E_{X\sim p(x)}[X]$ = $\int_x x p(x)\,dx$（连续）

$\sum_i (x_i \cdot p(x_i))$（离散）

2


Minka 2001 博士论
o deterministic, EP, BP, ADF,
o non-deterministic: MCMC
o ADF. one pass. Sequential methods.
对观测值顺序敏感.
对比batch训练数据采用不一致
o Expectation Propagation是ADF算法的扩展
o EP是loopy BP的一种扩展
chapter2 对积分的逼近计算
o方法1
$I = \int_A f(x)\,dx$ (2.1)
$\hat{I} = \sum_{i=1} w_i f(x_i)$ (2.2)
$\int_a^b f(x)\,dx$
a
b
o 方法2 变分估计(界)
ELBO: $\log P(x) = \log \int P(x,z)\frac{q(z)}{q(z)}\,dz$
$= \log E_q\left[\frac{P(x,z)}{q(z)}\right]$
$\geq E_q[\log P(x,z)] - E_q[\log q(z)]$
$\triangleq L(q)$
AUCS.cn

Chapter 3 Expectation Propagation 期望传播

3.1 Assumed-density filtering ADF方法

ADF也可称为 Online Bayesian learning 在线贝叶斯学习
moment matching 矩匹配
weak marginalization 弱边缘化

随机变量：D（观测变量），θ（隐变量）
目的：计算后验 $p(\theta|D)$ 和证据 $p(D)$

以下例为例来解读ADF：

观测变量的likelihood：$x \in \mathbb{R}^d$

w：已知权重
杂波：$\mathcal{N}(x;0,10I)$
I，10I：协方差矩阵

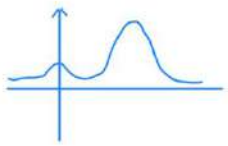
$$p(x|\theta) = (1-w)\mathcal{N}(x;\theta,I) + w\mathcal{N}(x;0,10I) \quad (3.1)$$

$$\mathcal{N}(x;m,V) = \frac{1}{|2\pi V|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(x-m)^T V^{-1}(x-m)\right) \quad (3.2)$$

行列式（不是绝对值）

m：均值向量
V：协方差矩阵

波形图



$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} \quad m = \begin{pmatrix} m_1 \\ m_2 \\ \vdots \\ m_d \end{pmatrix} \quad V = \begin{pmatrix} Cov(x_1,x_1) & \cdots & Cov(x_1,x_d) \\ \vdots & & \vdots \\ Cov(x_d,x_1) & \cdots & Cov(x_d,x_d) \end{pmatrix}$$

二次型：行向量×对称矩阵×列向量

在3.2中即：$(x-m)^T V^{-1}(x-m)$

$$\begin{aligned} Cov(x_i,x_j) &= E[(x_i - E[x_i])(x_j - E[x_j])] \\ &= E[x_i x_j] - 2E[x_i]E[x_j] + E[x_i]E[x_j] \\ &= E[x_i x_j] - E[x_i]E[x_j] \end{aligned}$$

若 $Cov(x_i,x_j)_{(i\neq j)} = 0$，则

$E[x_i x_j] = E[x_i]\cdot E[x_j]$，说明 x_i 与 x_j 不相关

假设θ的先验分布为一个d维高斯分布：

$$p(\theta) \sim N(0, 100 I_d) \qquad (3.3)$$

（0：d维向量）

θ和n个独立观测值 $D=\{x_1, \cdots, x_n\}$（x_n：n个x，每个x是d维）的联合分布为：

$$p(D,\theta) = p(\theta)\prod_i p(x_i|\theta) \qquad (3.4)$$

应用ADF的第一步为因子化 $p(D,\theta)$：

$$p(D,\theta) = \prod_i t_i(\theta) \qquad (3.5)$$

其中：$t_0(\theta) = p(\theta)$，$t_i(\theta) = p(x_i|\theta)$ (3.6), (3.7)

（t_0：先验；t_i：n项似然）

第二步：选择一个关于θ的后验分布逼近（指数族分布：高斯分布）

$$q(\theta) \sim N(m_\theta, v_\theta I) \qquad (3.8)$$

（m_θ：向量；v_θ：数字）

第三步：按照一定顺序吸收 t_i 来逼近后验

$$p(\theta|x_i) = \frac{p(x_i|\theta)p(\theta)}{p(x_i)} = \frac{p(x_i|\theta)p(\theta)}{\int_\theta p(x_i|\theta)p(\theta)d\theta} = \frac{t_i(\theta)p(\theta)}{\int_\theta t_i(\theta)p(\theta)d\theta}$$

在我们所挑选的 $q(\theta)$ 中，必然有一个最佳 $q(\theta)$ 能够替换 $p(\theta)$，在形式上更接近真实后验 $\hat{p}(\theta)$

$$\hat{p}(\theta) = \frac{t_i(\theta)q(\theta)}{\int_\theta t_i(\theta)q(\theta)d\theta} \qquad (3.9)$$

精确后验：

$\hat{p}(\theta) = \frac{t_i(\theta)q(\theta)}{\int_\theta t_i(\theta)q(\theta)d\theta}$

下寻找最好的 $q(\theta)$ 去逼近 $\hat{p}(\theta)$

KL散度：$D(\hat{p}(\theta)\|q^{new}(\theta))$

最好的 $q(\theta) = \underset{q(\theta)}{\operatorname{argmin}}\ KL(p(\theta)\|q(\theta))$

注：KL散度：衡量一个概率分布P和第二个参考的概率分布Q之间的差异。

$$D(P\|Q) = \int_x P(x)\log\left(\frac{P(x)}{Q(x)}\right)dx = E_{x\sim p(x)}\left[\log\frac{p(x)}{q(x)}\right]$$

往往，$p(x)$ 是客观存在（sometimes 仅仅是）

$q(x)$ 是 variational，得挑选好

指数家族分布相关回顾

定义：$p(x|\eta)=h(x)\cdot\exp(T(x)^T\eta-A(\eta))$

- $T(x)$：sufficient statistics
- η：natural parameter
- $A(\eta)$：归一化因子 log-normalizer
- $h(x)$：base measure

很多常见分布都可以被归为指数家族分布

如：高斯分布，伯努利分布，泊松分布，gamma分布，beta分布

狄利克雷分布

命题：一元高斯分布是指数家族分布.

证明：$X\sim N(\mu,\delta)$

$$p(x|\theta)=p(x|\mu,\delta)=\frac{1}{\sqrt{2\pi}\delta}\exp\left(-\frac{(x-\mu)^2}{2\delta^2}\right)$$

$$=\exp\left(-\frac{1}{2}\log(2\pi\delta^2)-\frac{1}{2\delta^2}(x^2-2x\mu+\mu^2)\right)$$

$$=\left(1,x,x^2\right)\begin{pmatrix}\mu/\delta^2\\ -\frac{1}{2\delta^2}\end{pmatrix}-\left(\frac{1}{2}\log(2\pi\delta^2)+\frac{\mu^2}{2\delta^2}\right)$$

$\eta=\begin{pmatrix}\eta_1\\ \eta_2\end{pmatrix}=\begin{pmatrix}\mu/\delta^2\\ -\frac{1}{2\delta^2}\end{pmatrix}$

$$=\exp\left(T(x)^T\cdot\eta-\left(\frac{1}{2}\log(2\pi)+\frac{1}{2}\log(\delta^2)-\frac{\mu^2}{-2\delta^2}\right)\right)$$

$$=\exp\left(T(x)^T\cdot\eta-\left(-\frac{\eta_1^2}{4\eta_2}-\frac{1}{2}\log(-2\eta_2)+\frac{1}{2}\log(2\pi)\right)\right)$$

$A(\eta)$

$T(x)=\begin{pmatrix}x\\ x^2\end{pmatrix}$

3. Chapter 3 - 期望传播(EP)

3.1 Assumed-density filtering

ADF 也可称为：
Online Bayesian learning, moment matching, weak marginalization
(在线贝叶斯学习) (矩匹配) (弱边缘化)

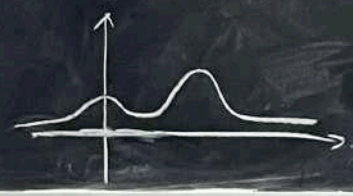
随机变量：D(观测变量), θ(隐变量)

目的：计算后验 $P(\theta|D)$ 和证据 $P(D)$

观测变量的likelihood：(以下仍以为例解读ADF)

$x \in \mathbb{R}^d$, $P(x|\theta) = (1-w)N(x;\theta, I) + wN(x;0, 10I)$, (3.1)

$N(x;m,V) = \frac{1}{|2\pi V|^{\frac{1}{2}}} \exp(-\frac{1}{2}(x-m)^T V^{-1}(x-m))$ (3.2)



w为已知权重，杂波为 $N(x;0,10I)$

V为协方差矩阵

$|2\pi V|$ 为行列式

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix}, \quad m = \begin{pmatrix} m_1 \\ m_2 \\ \vdots \\ m_d \end{pmatrix}, \quad V = \begin{pmatrix} Cov(x_1,x_1) & \cdots & Cov(x_1,x_d) \\ \vdots & & \vdots \\ cov(x_d,x_1) & & Cov(x_d,x_d) \end{pmatrix}$$

$Cov(x_i,x_j) = E[(x_i - E[x_i])(x_j - E[x_j])]$

若 $Cov(x_i,x_j)_{i\neq j} = 0 = E[x_i x_j - x_i E[x_j] - x_j E[x_i] + E[x_i]E[x_j]]$

$\Rightarrow E[x_i x_j] = E[x_i][x_j]$

则 x_i 与 x_j 不相关

3. Chapter 3 一期望传播(EP)

3.1 Assumed-density filtering

ADF 也可称为：online Bayesian learning（在线贝叶斯学习）, moment matching（矩匹配）, weak marginalization（弱边缘化）.

随机变量：D（观测变量），θ（隐变量）

目的：计算后验 $P(\theta|D)$ 和证据 $P(D)$

观测变量的 likelihood：（以下仍为例解读ADF）

$x \in \mathbb{R}^d$，$P(x|\theta) = (1-w)N(x;\theta, I) + wN(x;0, 10I)$，（3.1）

$N(x;m,V) = \frac{1}{|2\pi V|^{\frac{1}{2}}} \exp(-\frac{1}{2}(x-m)^T V^{-1}(x-m))$（3.2）

假设 θ 的先验分布为一个 d 维高斯分布：

$P(\theta) \sim N(0, 100 I_d)$（3.3）

θ 和 n 个独立观测值 $D = \{x_1, \dots, x_n\}$ 的联合分布为：

$P(D,\theta) = P(\theta)\prod_i P(x_i|\theta)$（3.4）

应用 ADF 的第一步为因子化 $P(D,\theta)$：

$P(D,\theta) = \prod_i t_i(\theta)$（3.5）

其中：$t_0(\theta) = \underbrace{P(\theta)}_{\text{先验}}$，$t_i(\theta) = \underbrace{P(x_i|\theta)}_{\text{似然}}$

第二步，选择一个关于 θ 的后验分布逼近

$q(\theta) \sim N(m_\theta, V_\theta I)$（3.8）

第三步，按照一定顺序吸收 t_i 来逼近后验：

$$P(\theta|x_i) = \frac{P(x_i|\theta)P(\theta)}{P(x_i)} = \frac{P(x_i|\theta)P(\theta)}{\int_\theta P(x_i|\theta)P(\theta)d\theta} = \frac{t_i(\theta)P(\theta)}{\int_\theta t_i(\theta)P(\theta)}$$

在我们所挑选的 $q(\theta)$ 中，必然有一个最佳的 $q(\theta)$，能够替换掉 $P(\theta)$，在形式上更接近真实后验 $\hat{P}(\theta)$

$$\hat{P}(\theta) = \frac{t_i(\theta)q(\theta)}{\int_\theta t_i(\theta)q(\theta)d\theta}$$（3.9）

2

精确后验

$$\hat{p}(\theta)=\frac{t_i(\theta)\,q(\theta)}{\int_\theta t_i(\theta)\,q(\theta)\,d\theta}$$

下寻找最好的 $q(\theta)$ 去逼近 $\hat{p}(\theta)$, s.t. $D(\hat{p}(\theta)\|q(\theta))$ 最小

KL散度: $D(\hat{p}(\theta)\|q^{new}(\theta))$

$$=\int \hat{p}(\theta)\log\left(\frac{\hat{p}(\theta)}{q^{new}(\theta)}\right)d\theta=\int\frac{t_i(\theta)q(\theta)}{\int_\theta t_i(\theta)q(\theta)d\theta}\log\left(\frac{\frac{t_i(\theta)q(\theta)}{\int_\theta t_i(\theta)q(\theta)d\theta}}{|2\pi U_\theta|^{\frac{1}{2}}\exp\left(-\frac{1}{2}(\theta-m_\theta^{new})^T(U_\theta^{new})^{-1}(\theta-m_\theta^{new})\right)}\right)d(\theta)$$

$$\nabla_{m_\theta^{new}} D(\hat{p}(\theta)\|q^{new}(\theta))=0$$

则可获得 $q^{new}(\theta)$ 中关于 m_θ^{new} 的迭代公式。

How ?

④

KL散度: 衡量一个概率分布P和第二个参考的概率分布Q之间的差异

$$D(P\|Q)=\int_x p(x)\log\left(\frac{p(x)}{q(x)}\right)dx=\mathop{E}_{x\sim p(x)}\left[\log\frac{p(x)}{q(x)}\right]$$

往往 $p(x)$ 是各状存在, (sometimes 仅仅只是)

- $q(x)$ 是 variational, 待挑选的!

③ 最好的 $q(x)$, $q^*(x)=\arg\min_{q(x)} KL(p\|q)$

指数家族分布：相关回顾

定义：

$$p(x|\eta)=h(x)\cdot\exp(T(x)^T\eta-A(\eta))$$

- $T(x)$: sufficient statistics
- η: natural parameter
- $A(\eta)$ 归一化因子。 log-normalizer
- $h(x)$ base measure

很多常见分布都可以被归为指数家族分布.

如：高斯分布、伯努利分布、泊松分布 gamma、beta分布

狄利克雷分布

5

一元高斯分布是指数家族分布 【结论】

证明：$X\sim N(\mu,\delta)$

$$p(X|\theta)=p(X|\mu,\delta)=\frac{1}{\sqrt{2\pi}\delta}\exp\left(-\frac{(X-\mu)^2}{2\delta^2}\right)$$

$$=\exp\left(-\frac{1}{2}\log(2\pi\delta^2)-\frac{1}{2\delta^2}(X^2-2X\mu+\mu^2)\right)$$

$$=\exp\left((X,X^2)\begin{pmatrix}\mu/\delta^2\\ -\frac{1}{2\delta^2}\end{pmatrix}-\left(\frac{1}{2}\log(2\pi\delta^2)+\frac{\mu^2}{2\delta^2}\right)\right),\quad \eta=\begin{pmatrix}\eta_1\\ \eta_2\end{pmatrix}=\begin{pmatrix}\mu/\delta^2\\ -\frac{1}{2\delta^2}\end{pmatrix}$$

$$=\exp\left(T(x)^T\cdot\eta-\left(\frac{1}{2}\log(2\pi)+\frac{1}{2}\log(\delta^2)-\frac{\mu^2}{2\delta^2}\right)\right)$$

$$=\exp\left(T(x)^T\cdot\eta-\left(-\frac{\eta_1^2}{4\eta_2}-\frac{1}{2}\log(-2\eta_2)+\frac{1}{2}\log(2\pi)\right)\right)$$

$$T(x)=\begin{pmatrix}x\\ x^2\end{pmatrix}$$

$\downarrow A(\eta)$

$N(\mu,\delta^2)$ 是指数族

球形高斯分布也是指数家族分布的一员：

设 $X \sim N(\mu, vI)$，X是一个n维向量，$\mu \in \mathbb{R}^n$，$v \in \mathbb{R}$

$$N(\mu, vI) = \frac{1}{|2\pi vI|^{n/2}} \exp\{-\frac{1}{2}(x-\mu)^T(vI)^{-1}(x-\mu)\}$$

$$= \frac{1}{(2\pi v)^{n/2}} \exp\{-\frac{1}{2v}(x-\mu)^T(x-\mu)\}$$

$$= \exp\{-\frac{n}{2}\log(2\pi v) - \frac{1}{2v}(x^Tx - 2x^T\mu + \mu^T\mu)\}$$

$$= \exp\{\begin{pmatrix} x \\ x^Tx \end{pmatrix}^T \begin{pmatrix} \frac{1}{v}\cdot\mu \\ -\frac{1}{2v} \end{pmatrix} - (\frac{n}{2}\log(2\pi v) - \frac{\mu^T\mu}{2v})\}$$

$$T(x) = \begin{pmatrix} x \\ x^Tx \end{pmatrix}, \quad \eta = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} \frac{1}{v}\cdot\mu \\ -\frac{1}{2v} \end{pmatrix}, \quad A(\eta) = \frac{n}{2}\log(2\pi v) - \frac{\mu^T\mu}{2v}$$

$$v = -\frac{1}{2\eta_2}, \quad A(\eta) = \frac{n}{2}\log(-\frac{1}{2\eta_2}) + \frac{n}{2}\log(2\pi) + \frac{\eta_1^T\eta_1}{4\eta_2}$$
$$\mu = -\frac{\eta_1}{2\eta_2}$$

注：

$$\begin{cases} \eta_1 \in \mathbb{R}^n \\ \eta_2 \in \mathbb{R} \end{cases}, \quad X = \begin{pmatrix} x \\ x^Tx \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \\ x_1^2+\cdots+x_n^2 \end{pmatrix}$$

换言之，$\exp\{\varphi(x_1, x_2, \cdots x_n, \sum_i x_i^2)\}$，weighted sum

当指数里是关于X的二项式时，是一个传统的一元高斯分布，

当是一个加权和时，是一个球型高斯.

下证(3.10) $m_\theta^{new} = \int_\theta \hat{p}(\theta)\theta d\theta$

$$D(\hat{p}(\theta)\|q^{new}(\theta)) = \int_\theta \hat{p}(\theta)\log(\frac{\hat{p}(\theta)}{q^{new}(\theta)})d\theta$$

$$= \int_\theta \hat{p}(\theta)(\log\hat{p}(\theta) - \underbrace{\log[\frac{1}{|2\pi v_\theta^{new} I|^{n/2}}\exp\{-\frac{1}{2}(\theta-m_\theta^{new})^T(v_\theta^{new}I)^{-1}(\theta-m_\theta^{new})\}]}_{\triangle})d\theta$$

$$\nabla_{m_\theta^{new}} D = -\int_\theta \hat{p}(\theta)\frac{1}{\triangle}\triangle(-\frac{1}{2v_\theta^{new}}(-2)(\theta-m_\theta^{new}))d\theta$$

$$= -\frac{1}{v_\theta^{new}}(\int_\theta \hat{p}(\theta)\theta d\theta - m_\theta^{new}\int_\theta \hat{p}(\theta)d\theta) = 0$$

令 $\nabla_{m_\theta^{new}} D = 0 \Rightarrow m_\theta^{new} = \int_\theta \hat{p}(\theta)\theta d\theta$

即 $m_\theta^{new} = E_{\hat{p}(\theta)}[\theta]$

注：

为什么在求 $\nabla_{m_\theta^{new}} D(\hat{p}(\theta)\|q^{new}(\theta))$ 时，无需考虑 $\hat{p}(\theta)$ 在 m_θ^{new} 的梯度？其中：

$\hat{p}(\theta)=\frac{t_i(\theta)q(\theta)}{\int_\theta t_i(\theta)q(\theta)d\theta}$.

因为可将 $D(\hat{p}(\theta)\|q^{new}(\theta)) \xrightarrow{\text{视作}} f_1(\hat{p}(\theta), q^{new}(\theta)) = f_2(\hat{p}(\theta), q(\theta), q^{new}(\theta))$，

当求 $\nabla_{m_\theta^{new}} D(\hat{p}(\theta)\|q^{new}(\theta))$ 时，$q(\theta)$ 的参数是已知的.

下证(3.11) $V_\theta^{new} d + (m_\theta^{new})^T(m_\theta^{new}) = \int_\theta \hat{p}(\theta)\theta^T\theta d\theta$

$$D(\hat{p}(\theta)\|q^{new}(\theta)) = \int_\theta \hat{p}(\theta)\log\frac{\hat{p}(\theta)}{q^{new}(\theta)}d\theta$$

$$= \int_\theta \hat{p}(\theta)\left[\log\hat{p}(\theta) - \log\left[\frac{1}{|2\pi V_I|^{1/2}}\exp\left\{-\frac{1}{2}(\theta-m_\theta^{new})^T(V_\theta^{new}I)^{-1}(\theta-m_\theta^{new})\right\}\right]\right]d\theta$$

$$\frac{\partial D}{\partial V_\theta^{new}} = \left(-\int_\theta \hat{p}(\theta)\left(-\frac{(\theta-m_\theta^{new})^T(\theta-m_\theta^{new})}{2V_\theta^{new}} - \frac{1}{2}\log(2\pi V_\theta^{new})^d\right)d\theta\right)'$$

$$= -\int_\theta \hat{p}(\theta)\left(\frac{(\theta-m_\theta^{new})^T(\theta-m_\theta^{new})}{2(V_\theta^{new})^2} - \left(\frac{d}{2}\log 2\pi V_\theta^{new}\right)'\right)d\theta$$

$$= -\frac{1}{2V_\theta^{new2}}\int_\theta \hat{p}(\theta)(\theta-m_\theta^{new})^T(\theta-m_\theta^{new})d\theta + \frac{d}{2V_\theta^{new}}\int_\theta \hat{p}(\theta)d\theta$$

取其为0：$\int_\theta \hat{p}(\theta)(\theta-m_\theta^{new})^T(\theta-m_\theta^{new})d\theta - V_\theta^{new}d = 0$

$$\Rightarrow \int_\theta \hat{p}(\theta)\theta^T\theta d\theta - 2\int_\theta \hat{p}(\theta)\theta^T m_\theta^{new}d\theta + \int_\theta \hat{p}(\theta)m_\theta^{new^T}m_\theta^{new}d\theta - V_\theta^{new}d = 0$$

$$\Rightarrow \int_\theta \hat{p}(\theta)\theta^T\theta d\theta - 2\int_\theta \hat{p}(\theta)\theta^T d\theta \cdot m_\theta^{new} + (m_\theta^{new})^T m_\theta^{new} - V_\theta^{new}d = 0$$

$$\Rightarrow \int_\theta \hat{p}(\theta)\theta^T\theta d\theta - 2(m_\theta^{new})^T m_\theta^{new} + (m_\theta^{new})^T m_\theta^{new} - V_\theta^{new}d = 0$$

即 $V_\theta^{new}d + (m_\theta^{new})^T(m_\theta^{new}) = \int_\theta \hat{p}(\theta)\theta^T\theta d\theta$ (3.11)

(3.10) $m_\theta^{new} = \int_\theta \hat{p}(\theta)\theta d\theta \Longleftrightarrow$ (3.12) $E_{q^{new}(\theta)}[\theta] = E_{\hat{p}(\theta)}[\theta]$

(3.11)→(3.13) $\underbrace{E_{q^{new}}[\theta^T\theta] = E_{\hat{p}}[\theta^T\theta]}_{?}$

结论：若X为n维向量 $X=\begin{pmatrix}X_1\\X_2\\\vdots\\X_n\end{pmatrix}$，$X\sim N(\mu,vI)$，则：

$$n\cdot v+\mu^T\mu=E_{N(\mu,vI)}[X^TX]$$

下证：

协方差矩阵 $\Sigma=(cov(X_i,X_j))_{n\times n}$

$$=\begin{pmatrix}cov(X_1,X_1)&\cdots\\\vdots&\ddots&\\&\cdots&cov(X_n,X_n)\end{pmatrix}_{n\times n}$$

$$=\left(E[(X_i-E[X_i])(X_j-E[X_j])\right)_{n\times n}$$

$$=E\left[\underset{n\times 1}{(X-E[X])}\underset{1\times n}{(X-E[X])^T}\right]_{n\times n}$$

$$=E[X\cdot X^T-2XE[X]^T+E[X]E[X]^T]$$

$$=E[XX^T]-2E[X]E[X]^T+E[X]E[X]^T$$

$$\Sigma=E[XX^T]-E[X][X]^T$$

当 $\Sigma=vI$，I为单位阵时，

$$\Sigma=vI=\begin{pmatrix}v&&\\&v&\\&&\ddots&v\end{pmatrix}_{n\times n}=\begin{pmatrix}E[X_1X_1]&\cdots&E[X_1X_n]\\\vdots&&\vdots\\E[X_nX_1]&\cdots&E[X_nX_n]\end{pmatrix}_{n\times n}-\begin{pmatrix}E[X_1]E[X_1]&\cdots&E[X_1]E[X_n]\\&\ddots&\\E[X_n]E[X_1]&\cdots&E[X_n]E[X_n]\end{pmatrix}_{n\times n}$$

则 $n\cdot v=Trace(\Sigma)=\sum_{i=1}^{n}E[X_i^2]-\sum_{i=1}^{n}\mu_i^2$

即 $n\cdot v+\mu^T\mu=E[X^TX]$

有关矩估计(moment matching)的一般性结论.

设 $q(x)=h(x)g(\eta)\exp\{\eta^T T(x)\}$ 为指数家族分布.

对 x 积分:

$$1=g(\eta)\int_x h(x)\exp\{\eta^T T(x)\}dx$$

对 η 求梯度:

$$0=\nabla_\eta g(\eta)\int_x h(x)\exp\{\eta^T T(x)\}dx+g(\eta)\int_x h(x)\exp\{\eta^T T(x)\}T(x)dx$$

$$\Rightarrow -\frac{\nabla_\eta g(\eta)}{g(\eta)}=\frac{\int_x g(\eta)h(x)\exp\{\eta^T T(x)\}T(x)dx}{\int_x g(\eta)h(x)\exp\{\eta^T T(x)\}dx}$$

$$\Rightarrow \boxed{-\nabla_\eta \log g(\eta)=\mathop{E}_{x\sim q(x)}[T(x)]}$$ 非常重要的结论.

回顾以前讨论的写法:

$$q(x)=h(x)\exp\{T(x)^T\eta-A(\eta)\}$$

熟知结论: $E_q[T(x)]=\nabla_\eta A(\eta)$

一个例子:

$$x\sim N(u,\sigma^2)$$

$$T(x)=\begin{pmatrix}x\\x^2\end{pmatrix},\ \eta=\begin{pmatrix}\eta_1\\\eta_2\end{pmatrix},\ E\left[\begin{pmatrix}x\\x^2\end{pmatrix}\right]=\nabla_\eta A(\eta)$$

下证 minimize $D(\hat{p}(\theta)\|q^{new}(\theta))$ 的结论.

设 $q^{new}(\theta)\sim N(m_\theta^{new},v_\theta^{new}I)$

$$D(\underset{\text{exact posterior}}{\hat{p}(\theta)}\|q^{new}(\theta))=\int_\theta \hat{p}(\theta)\log\frac{\hat{p}(\theta)}{q^{new}(\theta)}d\theta$$

$$=\int_\theta \hat{p}(\theta)\log\hat{p}(\theta)d\theta-\int_\theta \hat{p}(\theta)\log q^{new}(\theta)d\theta$$

$$=\int_\theta \underbrace{\hat{p}(\theta)\log\hat{p}(\theta)d\theta}_{C}-\int_\theta \hat{p}(\theta)\Big[\underbrace{\log h(\theta)}_{C}+\log g(\eta^{new})+\eta^{new^T}T(\theta)\Big]d\theta$$

$$\nabla_{\eta^{new}} D(\hat{p}(\theta)\|q^{new}(\theta))=\int_\theta -\hat{p}(\theta)\nabla_{\eta^{new}}\log g(\eta^{new})d\theta-\int_\theta \hat{p}(\theta)T(\theta)d\theta$$

令 $\nabla_{\eta^{new}} D(\hat{p}(\theta)\|q^{new}(\theta))=0$, 则

$$-\nabla_{\eta^{new}}\log g(\eta^{new})\int_\theta \hat{p}(\theta)d\theta-\mathop{E}_{\theta\sim\hat{p}(\theta)}[T(\theta)]=0$$

即 $\boxed{\mathop{E}_{\theta\sim q^{new}(\theta)}[T(\theta)]=\mathop{E}_{\theta\sim\hat{p}(\theta)}[T(\theta)]}$ ☆ (3.12) (3.13)

其中 $T(\theta)=\begin{pmatrix}\theta\\\theta^T\theta\end{pmatrix}$.

球形高斯分布也是指数族分布

证：设 $X \sim N(\mu, vI)$

x 是一个 n 维向量 $\mu \in \mathbb{R}^n$ $v \in \mathbb{R}$

$$N(\mu, vI) = \frac{1}{|2\pi vI|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T (vI)^{-1}(x-\mu)\right)$$

$$= \frac{1}{(2\pi v)^{\frac{n}{2}}} \exp\left(-\frac{1}{2v}(x-\mu)^T(x-\mu)\right)$$

$$= \exp\left(-\frac{n}{2}\log(2\pi v) - \frac{1}{2v}(x^Tx - 2x^T\mu + \mu^T\mu)\right)$$

$$= \exp\left(\begin{pmatrix} x \\ x^Tx \end{pmatrix}^T \begin{pmatrix} \frac{1}{v}\mu \\ -\frac{1}{2v} \end{pmatrix} - \left(\frac{n}{2}\log(2\pi v) - \frac{\mu^T\mu}{2v}\right)\right)$$

$$T(x) = \begin{pmatrix} x \\ x^Tx \end{pmatrix}, \quad \eta = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} \frac{1}{v}\mu \\ -\frac{1}{2v} \end{pmatrix}$$

$$A(\eta) = \left(\frac{n}{2}\log(2\pi v) - \frac{\mu^T\mu}{2v}\right)$$

$$\begin{cases} v = -\frac{1}{2\eta_2} \\ \mu = -\frac{\eta_1}{2\eta_2} \end{cases} = \frac{n}{2}\log\left(-\frac{1}{2\eta_2}\right) + \frac{n}{2}\log(2\pi) + \frac{1}{4\eta_1\eta_2}$$

注：

$\eta_1 \in \mathbb{R}^n$

$\eta_2 \in \mathbb{R}$

$$X = \begin{pmatrix} X \\ x^Tx \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \\ x_1^2 + \cdots + x_n^2 \end{pmatrix}$$

换言之，

$\exp(($ weighted sum $(x_1, x_2 \cdots x_n, \sum x_i^2))$

$x_1 x_2$ ✗

下证 3.10. $m_\theta^{new} = \int_\theta \hat{p}(\theta)\cdot\theta\, d\theta$

$D(\hat{p}(\theta)\|q^{new}(\theta)) = \int \hat{p}(\theta)\log\left(\frac{\hat{p}(\theta)}{q^{new}(\theta)}\right)d\theta$

$= \int \hat{p}(\theta)\Big(\log\hat{p}(\theta) - \log\Big[\frac{1}{|2\pi V|^{\frac{1}{2}}}\exp\big(-\frac{1}{2}(\theta-m_\theta^{new})^T(V_\theta^{new}I)^{-1}(\theta-m_\theta^{new})\big)\Big]\Big)d\theta$ → Δ

$\nabla_{m_\theta^{new}} D = -\int_\theta \hat{p}(\theta)\frac{1}{\Delta}\Delta\Big(-\frac{1}{2V_\theta^{new}}(-2)(\theta-m_\theta^{new})\Big)d\theta$

令 $= -\frac{1}{V_\theta^{new}}\Big(\int_\theta \hat{p}(\theta)\cdot\theta\, d\theta - m_\theta^{new}\int_\theta \hat{p}(\theta)d\theta\Big) = 0$

$\Rightarrow m_\theta^{new} = \int_\theta \hat{p}(\theta)\cdot\theta\, d\theta = E_{\hat{p}(\theta)}[\theta]$ (3.10) □

② 2

第2步 $\hat{p}(\theta) = \frac{t_i(\theta)q(\theta)}{\int t_i(\theta)q(\theta)d\theta}$

$X^T(RI)X = kX^TX$

$\frac{\partial(X^TX)}{\partial X} = 2X$

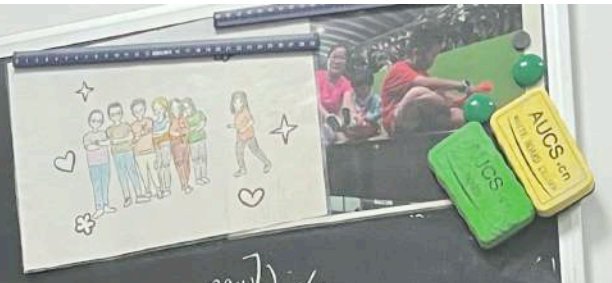
$x\in\mathbb{R}^d$ $f(x)=0$ $\mathbb{R}^3$

$\nabla_x f(x) = 0$ min ⇒ x^* 是极值

x^{new}

$= f(\hat{p}(\theta), q^{new}(\theta)) = f_2(t_i(\theta)\cdot q(\theta), q^{new}(\theta))$

注：当对 $\nabla_{m_\theta^{new}}$ 时，$q(\theta)$ 与 m_θ^{new} 无关 ⇒ m_θ, V_θ



下证 3.11: $V_\theta^{new}\cdot d+(m_\theta^{new})^T(m_\theta^{new})=\int_\theta \hat{p}(\theta)\theta^T\theta\, d\theta$

$$D(\hat{p}(\theta)\|q^{new}(\theta))=\int \hat{p}(\theta)\log\left(\frac{\hat{p}(\theta)}{q^{new}(\theta)}\right)d\theta$$

$$=\int \hat{p}(\theta)\left(\log\hat{p}(\theta)-\log\left[\frac{1}{|2\pi V|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\theta-m_\theta^{new})^T(V_\theta^{new}I)^{-1}(\theta-m_\theta^{new})\right)\right]\right)d\theta$$

$$\frac{dD}{dV_\theta^{new}}=\left(-\int \hat{p}(\theta)\cdot\left(\frac{(\theta-m_\theta^{new})^T(\theta-m_\theta^{new})}{2V_\theta^{new}}-\frac{1}{2}\log\underbrace{|2\pi V_\theta^{new}I|}_{=(2\pi V)^d}\right)d\theta\right)'$$

$$=-\int \hat{p}(\theta)\left(\frac{(\theta-m_\theta^{new})^T(\theta-m_\theta^{new})}{2(V_\theta^{new})^2}-\left(\frac{d}{2}\log 2\pi V_\theta^{new}\right)'\right)d\theta$$

$$=-\frac{1}{2V_\theta^{new\,2}}\int \hat{p}(\theta)(\theta-m_\theta^{new})^T(\theta-m_\theta^{new})d\theta+\frac{d}{2V_\theta^{new}}\int \hat{p}(\theta)d\theta$$

取其为0: $\int \hat{p}(\theta)(\theta-m_\theta^{new})^T(\theta-m_\theta^{new})d\theta-V_\theta^{new}d=0$

$$\int \hat{p}(\theta)\theta^T\theta\,d\theta-2\int \hat{p}(\theta)\theta^T m_\theta^{new}d\theta+\int \hat{p}(\theta)(m_\theta^{new})^T m_\theta^{new}-V_\theta^{new}d=0$$

$$\int \hat{p}(\theta)\theta^T\theta\,d\theta-2\int \hat{p}(\theta)\theta^T d\theta\cdot m_\theta^{new}+(m_\theta^{new})^T(m_\theta^{new})-V_\theta^{new}d=0$$

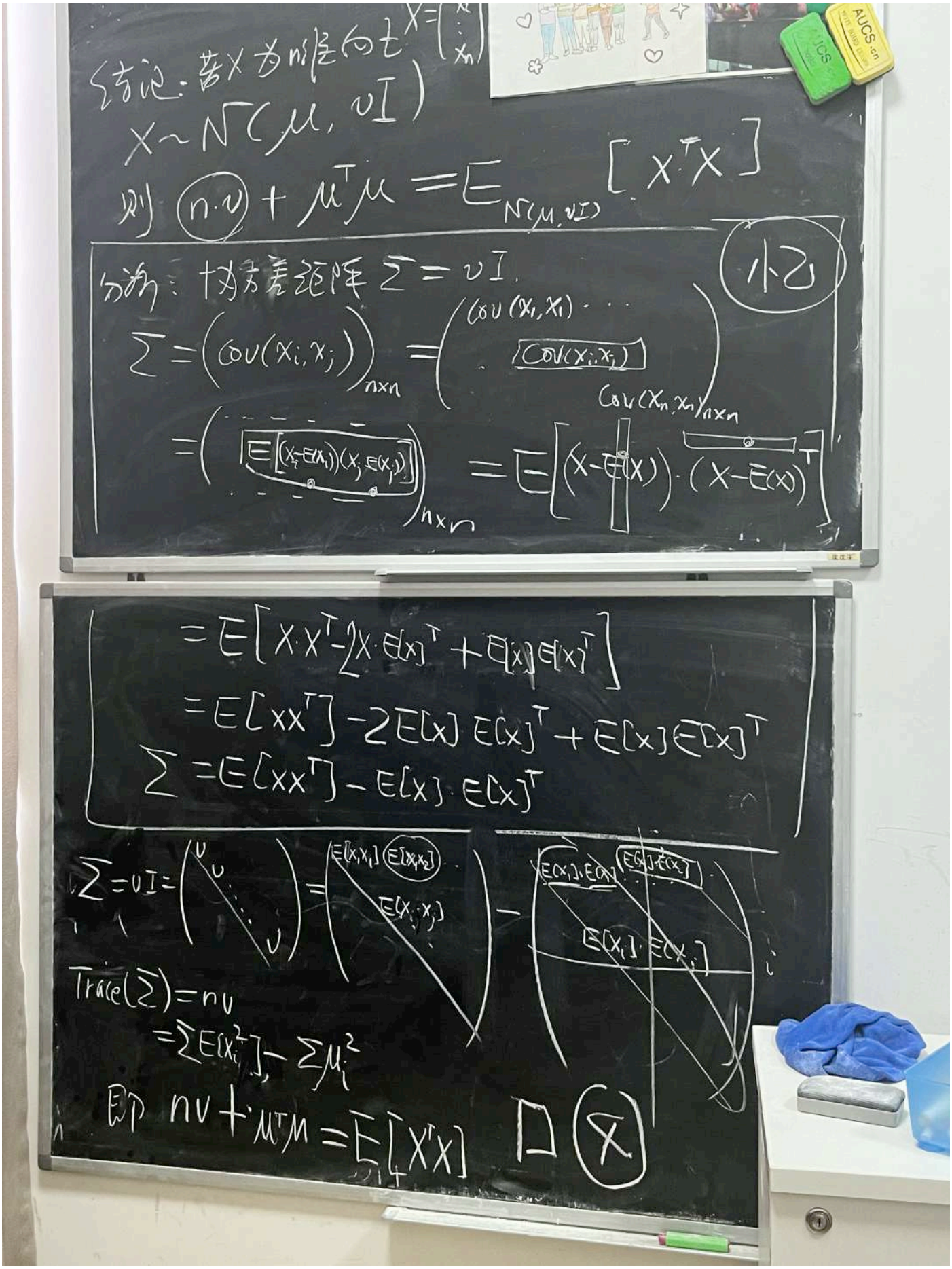
$$\int \hat{p}(\theta)\theta^T\theta\,d\theta-2(m_\theta^{new})^T(m_\theta^{new})+(m_\theta^{new})^T(m_\theta^{new})-V_\theta^{new}d=0$$

$$\Rightarrow \boxed{V_\theta^{new}d+(m_\theta^{new})^T(m_\theta^{new})=\int \hat{p}(\theta)\theta^T\theta\,d\theta \quad (3.11)}$$

(3.10) $m_\theta^{new}=\int_\theta \hat{p}(\theta)\cdot\theta\,d\theta \Longleftrightarrow E_{q^{new}}[\theta]=E_{\hat{p}}[\theta]$ (3.12)

(3.11→3.13) $E_{q^{new}}[\theta^T\theta]=E_{\hat{p}}[\theta^T\theta]$

3


结论：若X为n维向量 $X=\begin{pmatrix}x_1\\ \vdots\\ x_n\end{pmatrix}$
$X\sim N(\mu, vI)$
则 $n\cdot v+\mu^T\mu=E_{N(\mu,vI)}[X^TX]$
协方差矩阵 $\Sigma=vI$.
小记
$\Sigma=(cov(X_i,X_j))_{n\times n}=\begin{pmatrix}cov(X_1,X_1)&\cdots\\ \vdots&cov(X_i,X_j)\\ &cov(X_n,X_n)\end{pmatrix}_{n\times n}$
$=\left(E[(X_i-E(X_i))(X_j-E(X_j))]\right)_{n\times n}=E\left[(X-E(X))\cdot(X-E(X))^T\right]$
$=E[X\cdot X^T-2X\cdot E[X]^T+E[X]E[X]^T]$
$=E[XX^T]-2E[X]E[X]^T+E[X]E[X]^T$
$\Sigma=E[XX^T]-E[X]\cdot E[X]^T$
$\Sigma=vI=\begin{pmatrix}v&&\\&\ddots&\\&&v\end{pmatrix}=\begin{pmatrix}E[X_1X_1]&E[X_1X_2]&\cdots\\ \vdots&E[X_iX_j]&\\ &&\end{pmatrix}-\begin{pmatrix}E[X_1]\cdot E[X_1]&E[X_1]\cdot E[X_2]&\cdots\\ \vdots&E[X_i]\cdot E[X_j]&\\ &&\end{pmatrix}$
$Trace(\Sigma)=nv$
$=\sum E[X_i^2]-\sum\mu_i^2$
即 $nv+\mu^T\mu=E[X^TX]$ □(证)

有关矩估计(moment matching 的一般性结论)

设 $q(x)=h(x)g(\eta)\exp\{\eta^T T(x)\}$ 为指数家族分布.

对 x 积分:

$$1=g(\eta)\int_x h(x)\exp\{\eta^T T(x)\}dx$$

对 η 求梯度:

$$0=\nabla_\eta g(\eta)\int_x h(x)\exp\{\eta^T T(x)\}dx+g(\eta)\int_x h(x)\exp\{\eta^T T(x)\}T(x)dx$$

$$\Rightarrow -\frac{\nabla_\eta g(\eta)}{g(\eta)}=\frac{\int_x g(\eta)h(x)\exp\{\eta^T T(x)\}T(x)dx}{\int_x g(\eta)h(x)\exp\{\eta^T T(x)\}dx}$$

$$\boxed{-\nabla_\eta \log g(\eta)=\mathop{E}_{x\sim q(x)}[T(x)]}$$ IMPORTANT

回顾以前讲过的写法 $q(\eta)=e^{-A(\eta)}$

$$q(x)=h(x)\cdot\exp(T(x)^T\eta-A(\eta))$$

熟知结论:

$$\boxed{E_q[T(x)]=\nabla_\eta A(\eta)}$$

比如 $x\sim N(\mu,\sigma)$

$$T(x)=\begin{pmatrix}x\\x^2\end{pmatrix}\quad \eta=\begin{pmatrix}\eta_1\\\eta_2\end{pmatrix}$$

$$E\begin{bmatrix}x\\x^2\end{bmatrix}=\nabla_\eta A(\eta)$$

5

有关矩估计(moment matching)的一般性结论

设 $q(x)=h(x)g(\eta)\exp\{\eta^T T(x)\}$ 为指数家族分布

对 x 积分:

$1=g(\eta)\int_x h(x)\exp\{\eta^T T(x)\}dx$

对 η 求导两次:

$0=\nabla_\eta g(\eta)\int_x h(x)\exp\{\eta^T T(x)\}dx + g(\eta)\int_x h(x)\exp\{\eta^T T(x)\}T(x)dx$

$\Rightarrow -\frac{\nabla_\eta g(\eta)}{g(\eta)} = \frac{\int_x g(\eta)h(x)\exp\{\eta^T T(x)\}T(x)dx}{\int_x g(\eta)h(x)\exp\{\eta^T T(x)\}dx}$

$$-\nabla_\eta \log g(\eta) = \mathop{E}_{x\sim q(x)}[T(x)]$$

一个关键性质 IMPORTANT

$D(\hat{p}(\theta)\|q^{new}(\theta)) = \int_\theta \hat{p}(\theta)\log\frac{\hat{p}(\theta)}{q^{new}(\theta)}d\theta$

exact posterior

$= \int_\theta \hat{p}(\theta)\log\hat{p}(\theta)d\theta - \int_\theta \hat{p}(\theta)\log q^{new}(\theta)d\theta$

$= \int_\theta \hat{p}(\theta)\log\hat{p}(\theta)d\theta - \int_\theta \hat{p}(\theta)[\log h(\theta) + \log g(\eta^{new}) + \eta^{new\,T}T(\theta)]d\theta$

(c, c, c)

设 $q^{new}(\theta)\sim N(\mu_0^{new}, \Psi_0^{new})$

$\nabla_{\eta^{new}} D(\hat{p}(\theta)\|q^{new}(\theta)) = \int_\theta -\hat{p}(\theta)\nabla_{\eta^{new}}\log g(\eta^{new})d\theta - \int_\theta \hat{p}(\theta)T(\theta)d\theta$

令 $\nabla_{\eta^{new}} D(\hat{p}(\theta)\|q^{new}(\theta)) = 0$

$\Rightarrow -\nabla_{\eta^{new}}\log g(\eta^{new})\int \hat{p}(\theta)d\theta - \mathop{E}_{\theta\sim\hat{p}(\theta)}[T(\theta)] = 0$

6

$$\Rightarrow \mathop{E}_{\theta\sim q^{new}(\theta)}[T(\theta)] = \mathop{E}_{\theta\sim\hat{p}(\theta)}[T(\theta)]$$

☆

$\because T(\theta) = \begin{pmatrix}\theta\\ \theta\theta^T\end{pmatrix}$;

(3.12) + (3.13)

根据矩匹配：

$$E_{q^{new}}[\theta] = E_{\tilde{p}}[\theta] \qquad (3,12)$$

$$E_{q^{new}}[\theta^T\theta] = E_{\tilde{p}}[\theta^T\theta] \qquad (3,13)$$

对于 $\theta \sim q^{new} = \mathcal{N}(m_\theta, V_\theta I)$，只要 (3,12)，(3,13) 满足，则 $KL(\tilde{p}\|q)$ 最小

下述为对一般性的高斯分布结论，对任意 $t(\theta)$ 成立：

$$Z(m_\theta, V_\theta) = \int_\theta t(\theta)\, q(\theta)\, d\theta \qquad q(\theta) \sim \mathcal{N}(m_\theta, V_\theta)$$

Z 事实上是 $\tilde{p}$ 的归一化因子 $\tilde{p}(\theta) = \dfrac{t(\theta)\, q(\theta)}{\int_\theta t(\theta) q(\theta) d\theta}$

注：ADF 的一个 trick

理想后验 $\tilde{p}(\theta) \propto t(\theta) q(\theta)$

$$Z(m_\theta, V_\theta) = \int_\theta \frac{t(\theta)}{(2\pi V_\theta)^{d/2}} \exp\left(-\frac{1}{2V_\theta}(\theta - m_\theta)^T(\theta - m_\theta)\right) d\theta$$

$$\nabla_m \log Z(m_\theta, V_\theta)$$

$$= \frac{1}{Z}\nabla_m Z = \frac{1}{Z}\int_\theta \frac{t(\theta)}{(2\pi V_\theta)^{d/2}} \cdot \frac{1}{V_\theta}[\theta - m_\theta] \cdot \exp(\,)\, d\theta$$

$$= \frac{E_{\tilde{p}}(\theta)}{V_\theta} - \frac{m_\theta}{V_\theta} \qquad (3,17)$$

$$E_{\tilde{p}}[\theta] = m_\theta + V_\theta \nabla_m \log Z(m_\theta, V_\theta) \qquad (3,18)$$

实际上即：$m_\theta^{new} = m_\theta + V_\theta \log Z(m_\theta, V_\theta)$ 为 m_θ^{new} 的递推公式

回到 Minka 2001 论文的特例

由 (3,7)：$t(\theta) = P(x|\theta)$

由 (3,1)：$p(x|\theta) = (1-w)\mathcal{N}(x;\theta, I) + w\mathcal{N}(x;0, 10I)$

令有 $q(\theta) = \mathcal{N}(m_\theta, V_\theta I)$

代入 (3,14)

$$Z(m_\theta, V_\theta) = \int_\theta \left[(1-w)\mathcal{N}(x;\theta, I) + w\mathcal{N}(x;0,10I)\right] \cdot \mathcal{N}(\theta; m_\theta, V_\theta)\, d\theta$$

$$= \int_\theta (1-w)\mathcal{N}(x;\theta, I)\mathcal{N}(\theta; m_\theta, V_\theta) d\theta + \int_\theta w\mathcal{N}(x;0,10I) \cdot \mathcal{N}(\theta, m_\theta, V_\theta) d\theta$$

$$= (1-w)\int_\theta \mathcal{N}(x;\theta, I)\mathcal{N}(\theta; m_\theta, V_\theta) d\theta + w\mathcal{N}(x;0,10I)$$

下证：$\int_\theta N(x;\theta,I)N(\theta;m_\theta,V_\theta)d\theta = N(x;m_\theta,(1+V_\theta)I)$

$$\int_\theta N(x;\theta,I)N(\theta;m_\theta,V_\theta)d\theta$$
$$= \int_\theta \frac{1}{(2\pi)^{d/2}}\exp\left\{-\frac{1}{2}(x-\theta)^T(x-\theta)\right\}\cdot\frac{1}{(2\pi V_\theta)^{d/2}}\exp\left\{-\frac{1}{2V_\theta}(\theta-m_\theta)^T(\theta-m_\theta)\right\}d\theta$$
$$= \frac{1}{2\pi V^{d/2}}\int_\theta \exp\left\{-\left[\frac{(x-\theta)^T(x-\theta)}{2}+\frac{(\theta-m_\theta)^T(\theta-m_\theta)}{2V_\theta}\right]\right\}d\theta \quad (1)$$

下面先关注指数部分：

$$-\left[\frac{(x-\theta)^T(x-\theta)}{2}+\frac{(\theta-m_\theta)^T(\theta-m_\theta)}{2V_\theta}\right]$$
$$= -\frac{(x-\theta)^T(x-\theta)V_\theta+(\theta-m_\theta)^T(\theta-m_\theta)}{2V_\theta}$$
$$= -\frac{\theta^T\theta-2V_\theta x^T\theta+V_\theta\theta^T\theta-2\theta^Tm_\theta}{2V_\theta}-\frac{V_\theta x^Tx+m_\theta^Tm_\theta}{2V_\theta}$$

下面先关注Δ项：

$$\Delta = -\frac{(1+V_\theta)\theta^T\theta-(2V_\theta x^T+2m_\theta^T)\theta}{2V_\theta}$$
$$= -\frac{\theta^T\theta-\frac{2(V_\theta x^T+m_\theta^T)\theta}{1+V_\theta}}{2\frac{V_\theta}{1+V_\theta}}$$
$$= -\frac{(\theta-\frac{V_\theta x+m_\theta}{1+V_\theta})^T(\theta-\frac{V_\theta x+m_\theta}{1+V_\theta})-(\frac{V_\theta x+m_\theta}{1+V_\theta})^T(\frac{V_\theta x+m_\theta}{1+V_\theta})}{2(\frac{V_\theta}{1+V_\theta})}$$

代回(1)式：

$$(1)\text{式} = \frac{1}{(2\pi)^{d/2}V_\theta^{d/2}}\int_\theta \exp\{[\,]\}\cdot\exp\left\{\frac{(\frac{V_\theta x+m_\theta}{1+V_\theta})^T(\frac{V_\theta x+m_\theta}{1+V_\theta})}{2\frac{V_\theta}{1+V_\theta}}-\frac{V_\theta x^Tx+m_\theta^Tm_\theta}{2V_\theta}\right\}d\theta$$

$$= \frac{(2\pi\frac{V_\theta}{1+V_\theta})^{d/2}}{(2\pi)^dV_\theta^{d/2}}\left(\frac{1}{(2\pi\frac{V_\theta}{1+V_\theta})^{d/2}}\int_\theta\exp\{[\,]\}d\theta\right)\cdot\exp\left\{\frac{(V_\theta x+m_\theta)^T(V_\theta x+m_\theta)}{2V_\theta(1+V_\theta)}-\frac{V_\theta x^Tx+m_\theta^Tm_\theta}{2V_\theta}\right\}$$

$$= \frac{1}{[2\pi(1+V_\theta)]^{d/2}}\cdot\exp\left\{\frac{-V_\theta x^Tx+2V_\theta x^Tm_\theta-V_\theta m_\theta^Tm_\theta}{2V_\theta(1+V_\theta)}\right\}$$

$$= \frac{1}{[2\pi(1+V_\theta)]^{d/2}}\cdot\exp\left\{-\frac{(x-m_\theta)^T(x-m_\theta)}{2(1+V_\theta)}\right\}$$

$= N(x;m_\theta,(1+V_\theta)I)$. 得证.

由此可得：

$$Z(m_\theta, V_\theta) = (1-w)\mathcal{N}(x; m_\theta, (V_\theta+1)I) + w\mathcal{N}(x; 0, 10I) \quad (3.20)$$

$$r = \frac{(1-w)\mathcal{N}(x; m_\theta, (V_\theta+1)I)}{(1-w)\mathcal{N}(x; m_\theta, (V_\theta+1)I) + w\mathcal{N}(x; 0, 10I)} \quad (3.21)$$

$$\nabla_m \log Z(m_\theta, V_\theta) = \frac{1}{Z}\nabla_m Z(m_\theta, V_\theta)$$

$$= \frac{1}{Z}\nabla_m \left[(1-w) \frac{1}{|2\pi(V_\theta+1)|^{d/2}} \exp\left(-\frac{1}{2V_\theta+1}(x-m_\theta)^T(x-m_\theta)\right) d\theta + w \right]$$

$$= r \cdot \frac{x-m_\theta}{V_\theta+1} \quad (3.22)$$

代入 (3.18) 得：$m_\theta^{new} = m_\theta + V_\theta r_i \frac{x_i - m_\theta}{V_\theta+1}$ (3.25)

称为：得到了ADF迭代公式

Minka 2001 论文中，ADF有关 m_θ, V_θ 的迭代：

$$\begin{cases} m_\theta^{new} = m_\theta + V_\theta r_i \frac{x_i - m_\theta}{V_\theta+1} \\ V_\theta^{new} = V_\theta - r_i \frac{V_\theta^2}{V_\theta+1} + r_i(1-r_i)\frac{V_\theta^2 (x_i-m_\theta)^T(x_i-m_\theta)}{d(V_\theta+1)^2} \end{cases}$$

简写之：$\begin{cases} m_\theta^{new} \leftarrow m_\theta, V_\theta, x_i \\ V_\theta^{new} \leftarrow m_\theta, V_\theta, x_i \end{cases}$

推出 $q^{new}(\theta)$ 去逼近 θ 的后验 $p(\theta|D)$

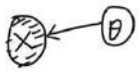
则后面可用 θ 生成 D

$D = \{x_1, \cdots x_n\}$ $D|\theta$

$p(x) = (1-w)\mathcal{N}(x; \theta, I) + w\mathcal{N}(x; 0, 10I)$

(若我写 $D=x$)

图模型 (GM)



$$P(D,\theta) = p(D|\theta)\cdot p(\theta)$$

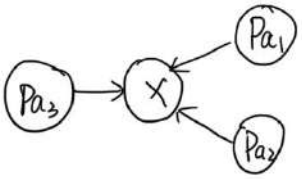
$$= \prod_{i=1}^{n} p(x_i|\theta) p(\theta)$$

$$= \prod_{i=0} t_i(\theta)$$

$\begin{cases} t_0(\theta) = p(\theta) \text{ 先验} \\ t_i(\theta) = p(x_i|\theta) \text{ 似然} \end{cases}$

在一般的有向图模型中：

$$\prod_{nodes} p(x|pa(x))$$



联合概率：$P(X, Parents) = P(x|Pa_1)\cdot P(x|Pa_2)\, P(x|Pa_3) \prod P(a_i)$

形式上等于 $\prod t_i(\cdot)$

3.2 Expectation Propagation 期望传播

$$\tilde{t}_i(\theta) = z\frac{q^{new}(\theta)}{q(\theta)} \quad (3.28)$$

与 $t_i(\theta)$ 进行对比：$t_i(\theta) = p(x_i|\theta) = z\frac{q(\theta)}{q(\theta)}$ 用 $\tilde{t}_i$ 逼近 t_i

回到上帝的 clutter 问题，求出 $\tilde{t}_i$ 的表达式

$$Z_i(m_\theta, V_\theta) = \int_\theta t_i(\theta) q(\theta)\, d\theta \quad (3.29)$$

$$\tilde{t}_i(\theta) = Z_i(m_\theta, V_\theta)\frac{q^{new}(\theta)}{q(\theta)} \quad (3.30)$$

$$= Z_i(m_\theta, V_\theta)\left(\frac{V_\theta}{V^{new}}\right)^{d/2} \exp\left(-\frac{1}{2V_\theta^{new}}(\theta - m_\theta^{new})^T(\theta - m_\theta^{new})\right)\cdot$$

$$\exp\left(\frac{1}{2V_\theta}(\theta - m_\theta)^T(\theta - m_\theta)\right) \quad (3.31)$$

定义 m_i, V_i：

$$V_i^{-1} = (V_\theta^{new})^{-1} - V_\theta^{-1} \quad (3.32)$$

$$m_i = V_i(V_\theta^{new})^{-1} m_\theta^{new} - V_i V_\theta^{-1} m_\theta \quad (3.33)$$

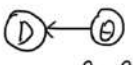
$$= m_\theta + (V_i + V_\theta)V_\theta^{-1}(m_\theta^{new} - m_\theta) \quad (3.34)$$

简化 $\tilde{t}_i(\theta)$：

$$\tilde{t}_i(\theta) = \frac{Z_i(m_\theta, V_\theta)}{N(m_i; m_\theta, (V_i + V_\theta)I)} N(\theta; m_i, V_i I) \quad (3.36)$$

小结：在 EP 中，后验仍是 $\prod_i t_i(\theta)$，但与 ADF 不同的是，每次取一个 $t_i(\theta)$，用 $\tilde{t}_i(\theta)$ 进行替代，得到最好的 $\tilde{t}_i(\theta)$

EP 的算法流程：（一般性流程）



$\theta \sim q^{new}(\theta)$

1. 初始化 t_i 的估计 $\tilde{t}_i$

2. 计算 θ 的后验：

$$q^{new}(\theta) = \frac{\prod_i \tilde{t}_i(\theta)}{\int \prod_i \tilde{t}_i(\theta) d\theta} \quad (3.27)$$

注：作者假设 $q^{new}(\theta) \propto \prod t_i(\theta)$

3. 直到所有的 $\tilde{t}_i$ 收敛

(a). 选择一个 $\tilde{t}_i$

(b) 从后验中去除所选 $\tilde{t}_i$：

$$q^{\setminus i}(\theta) \propto \frac{q^{new}(\theta)}{t_i(\theta)} \quad (3.38)$$

(c). 利用ADF，计算 $q^{new\setminus}(\theta)$，便其逼近 $q^{-i}(\theta)$

说明：$q^{new\setminus}$ = 更新 q^{new} 后，$q^{new\setminus} = \underset{q^{new}}{\arg\min}\, KL(q^{-i} \| q^{new})$

$q^{new\setminus} = h(\theta)\exp(T(\theta)^T \eta^{new\setminus} - A(\eta^{new\setminus}))$

由 $E_{q^{new}}[T(\theta)] = E_{q^{-i}}[T(\theta)$ 导出 $\eta^{new\setminus}$ 的递推公式

(d)：$\tilde{t}_i = Z_i \dfrac{q^{new\setminus}(\theta)}{q^{-i}(\theta)}$

收敛条件：误差小于 10^{-4}

4. $p(D) \approx \int \prod_i \tilde{t}_i(\theta)\, d\theta$ (3.40)

3(b)另一种写法：$q^{-i}(\theta) \propto \prod_{j \neq i} \tilde{t}_i(\theta)$ (3.41)

以上为图模型 $(D) \leftarrow (\theta)$，$\theta \sim q^{new}(\theta)$ 情形下的一般性推导

EP在 Minka 2001 论文中 Clutter 问题求解的示例可直接看(3.42)~(3.62)

3.2.1 The clutter problem

For the clutter problem of the previous section, the EP algorithm is:

1. The term approximations have the form

$$\tilde{t}_i(\theta) = s_i \exp(-\frac{1}{2v_i}(\theta - \mathbf{m}_i)^{\mathrm{T}}(\theta - \mathbf{m}_i)) \tag{3.42}$$

Initialize the prior term to itself:

$$v_0 = 100 \tag{3.43}$$
$$\mathbf{m}_0 = \mathbf{0} \tag{3.44}$$
$$s_0 = (2\pi v_0)^{-d/2} \tag{3.45}$$

Initialize the data terms to 1:

$$v_i = \infty \tag{3.46}$$
$$\mathbf{m}_i = \mathbf{0} \tag{3.47}$$
$$s_i = 1 \tag{3.48}$$

2. $\mathbf{m}_\theta^{new} = \mathbf{m}_0, v_\theta^{new} = v_0$

3. Until all $(\mathbf{m}_i, v_i, s_i)$ converge (changes are less than 10^{-4}):

loop $i = 1, \ldots, n$:

(a) Remove $\tilde{t}_i$ from the posterior to get an 'old' posterior:

$$v_\theta^{-1} = (v_\theta^{new})^{-1} - v_i^{-1} \tag{3.49}$$
$$\mathbf{m}_\theta = v_\theta(v_\theta^{new})^{-1}\mathbf{m}_\theta^{new} - v_\theta v_i^{-1}\mathbf{m}_i = \mathbf{m}_\theta^{new} + v_\theta v_i^{-1}(\mathbf{m}_\theta^{new} - \mathbf{m}_i) \tag{3.50}$$

(b) Recompute $(\mathbf{m}_\theta^{new}, v_\theta^{new}, Z_i)$ from $(\mathbf{m}_\theta, v_\theta)$ (this is the same as ADF):

$$r_i = \frac{(1-w)\mathcal{N}(\mathbf{x}_i; \mathbf{m}_\theta, (v_\theta+1)\mathbf{I})}{(1-w)\mathcal{N}(\mathbf{x}_i; \mathbf{m}_\theta, (v_\theta+1)\mathbf{I}) + w\mathcal{N}(\mathbf{x}_i; \mathbf{0}, 10\mathbf{I})} \tag{3.51}$$

$$\mathbf{m}_\theta^{new} = \mathbf{m}_\theta + v_\theta r_i \frac{\mathbf{x}_i - \mathbf{m}_\theta}{v_\theta + 1} \tag{3.52}$$

$$v_\theta^{new} = v_\theta - r_i \frac{v_\theta^2}{v_\theta+1} + r_i(1-r_i)\frac{v_\theta^2(\mathbf{x}_i - \mathbf{m}_\theta)^{\mathrm{T}}(\mathbf{x}_i - \mathbf{m}_\theta)}{d(v_\theta+1)^2} \tag{3.53}$$

$$Z_i = (1-w)\mathcal{N}(\mathbf{x}_i; \mathbf{m}_\theta, (v_\theta+1)\mathbf{I}) + w\mathcal{N}(\mathbf{x}_i; \mathbf{0}, 10\mathbf{I}) \tag{3.54}$$

(c) Update $\tilde{t}_i$:

$$v_i^{-1} = (v_\theta^{new})^{-1} - v_\theta^{-1} \tag{3.55}$$

$$= \left(\frac{r_i}{v_\theta+1} - r_i(1-r_i)\frac{(\mathbf{x}_i - \mathbf{m}_\theta)^{\mathrm{T}}(\mathbf{x}_i - \mathbf{m}_\theta)}{d(v_\theta+1)^2}\right)^{-1} - v_\theta \tag{3.56}$$

$$\mathbf{m}_i = \mathbf{m}_\theta + (v_i + v_\theta)v_\theta^{-1}(\mathbf{m}_\theta^{new} - \mathbf{m}_\theta) \tag{3.57}$$

$$= \mathbf{m}_\theta + (v_i + v_\theta) r_i \frac{\mathbf{x}_i - \mathbf{m}_\theta}{v_\theta + 1} \tag{3.58}$$

$$s_i = \frac{Z_i}{(2\pi v_i)^{d/2}\mathcal{N}(\mathbf{m}_i; \mathbf{m}_\theta, (v_i + v_\theta)\mathbf{I})} \tag{3.59}$$

4. Compute the normalizing constant:

$$B = \frac{(\mathbf{m}_\theta^{new})^{\mathrm{T}}\mathbf{m}_\theta^{new}}{v_\theta^{new}} - \sum_i \frac{\mathbf{m}_i^{\mathrm{T}}\mathbf{m}_i}{v_i} \tag{3.60}$$

$$p(D) \approx \int \prod_i \tilde{t}_i(\theta) d\theta \tag{3.61}$$

$$= (2\pi v_\theta^{new})^{d/2} \exp(B/2) \prod_{i=0}^{n} s_i \tag{3.62}$$

根据矩匹配

$$\begin{cases} E_{q^{new}}[\theta] = E_{\hat{q}}[\theta] & (3.12) \\ E_{q^{new}}[\theta\theta^T] = E_{\hat{q}}[\theta\theta^T] & (3.13) \end{cases}$$

对于 $\theta \sim q^{new} = N(m_\theta, V_\theta I)$，只要 (3.12), (3.13) 满足，则 $KL(\hat{q} \| q)$ 最小

下述为对一般性 Gaussian 结论，对任意 $t(\theta)$ 成立。

$$Z(m_\theta, V_\theta) = \int_\theta t(\theta) q(\theta) d\theta \quad (3.14)$$

$q(\theta) \sim N(m_\theta, V_\theta)$

$$= \int_\theta t(\theta) \cdot \frac{1}{(2\pi V_\theta)^{d/2}} \cdot \exp\left(-\frac{1}{2V_\theta}(\theta - m_\theta)^T(\theta - m_\theta)\right) d\theta \quad (3.15)$$

Z 事实上是个的归一化因子

$$\hat{p} = \frac{t_i(\theta) q(\theta)}{\int_\theta t_i(\theta) q(\theta) d\theta}$$

$$\nabla_{m_\theta} \log Z(m_\theta, V_\theta)$$

$$= \frac{1}{Z} \nabla_{m_\theta} Z$$

$$= \frac{1}{Z} \int_\theta \frac{t(\theta)}{(2\pi V_\theta)^{d/2}} \cdot \frac{1}{V_\theta}[\theta - m_\theta] \cdot \exp[\dots] d\theta$$

$$= \frac{E_{\hat{q}}[\theta]}{V_\theta} - \frac{m_\theta}{V_\theta} \quad (3.17)$$

注：ADF的一个 trick

理想形态 $\hat{p}(\theta) \propto t(\theta) q(\theta)$

$E_{\hat{q}}[\theta] = m_\theta + V_\theta \nabla \log Z(m_\theta, V_\theta)$ (3.18) 实际上就为 m_θ^{new} 的迭推公式

$$m_\theta^{new} = m_\theta + V_\theta \nabla \log Z(m_\theta, V_\theta)$$

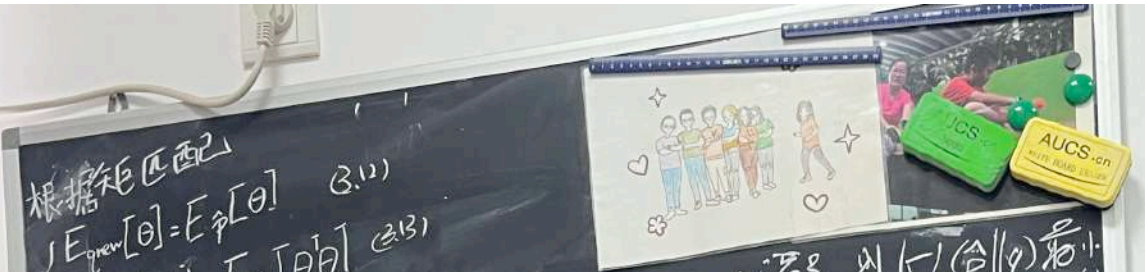
回到 Minka 2001 论文的特例

由(3.7) $t(\theta) = p(x|\theta)$

由(3.1) $p(x|\theta) = (1-w) N(x|\theta, I) + w N(x|0, I)$

另有 $q(\theta) = N(m_\theta, V_\theta I)$

下代入(3.14)



$$Z(m_\theta, v_\theta) = \int_\theta [(1-w)N(x;\theta,I) + wN(x;0,\alpha I)] \cdot N(\theta; m_\theta, v_\theta)\, d\theta$$

$$= \int_\theta (1-w)N(x;\theta,I)N(\theta;m_\theta,v_\theta)\,d\theta + \int_\theta wN(x;0,\alpha I)N(\theta;m_\theta,v_\theta)\,d\theta$$

$$= (1-w)\int_\theta N(x;\theta,I)N(\theta;m_\theta,v_\theta)\,d\theta + wN(x;0,\alpha I)$$

下证：$\int_\theta N(x;\theta,I)N(\theta;m_\theta,v_\theta)\,d\theta = N(x;m_\theta,(1+v_\theta)I)$

$$\int_\theta N(x;\theta,I)N(\theta;m_\theta,v_\theta)\,d\theta = \int_\theta \frac{1}{(2\pi)^{d/2}}\exp\left\{-\frac{1}{2}(x-\theta)^T(x-\theta)\right\} \cdot \frac{1}{(2\pi v_\theta)^{d/2}}\exp\left\{-\frac{1}{2v_\theta}(\theta-m_\theta)^T(\theta-m_\theta)\right\} d\theta$$

$$= \frac{1}{2\pi v_\theta^{d/2}}\int_\theta \exp\left\{-\left[\frac{(x-\theta)^T(x-\theta)}{2} + \frac{(\theta-m_\theta)^T(\theta-m_\theta)}{2v_\theta}\right]\right\} d\theta \quad (1)$$

下面先只关注指数部分：

$$-\left[\frac{(x-\theta)^T(x-\theta)}{2} + \frac{(\theta-m_\theta)^T(\theta-m_\theta)}{2v_\theta}\right] = -\frac{(x-\theta)^T(x-\theta)v_\theta + (\theta-m_\theta)^T(\theta-m_\theta)}{2v_\theta}$$

$$= -\underbrace{\frac{\theta^T\theta - 2v_\theta x^T\theta + v_\theta\theta^T\theta - 2\theta^T m_\theta}{2v_\theta}}_{*} - \frac{v_\theta x^T x + m_\theta^T m_\theta}{2v_\theta}$$

(2)

下面只关注 $*$ 项：

$$* = -\frac{(1+v_\theta)\theta^T\theta - (2v_\theta x^T + 2m_\theta^T)\theta}{2v_\theta} = -\frac{\theta^T\theta - \frac{2(v_\theta x^T + m_\theta^T)\theta}{1+v_\theta}}{2\cdot\frac{v_\theta}{1+v_\theta}}$$

$$= -\left[\frac{(\theta - \frac{v_\theta x + m_\theta}{1+v_\theta})^T(\theta - \frac{v_\theta x + m_\theta}{1+v_\theta})}{2\left(\frac{v_\theta}{1+v_\theta}\right)}\right] - \frac{\left(\frac{v_\theta x + m_\theta}{1+v_\theta}\right)^T\left(\frac{v_\theta x + m_\theta}{1+v_\theta}\right)}{2\left(\frac{v_\theta}{1+v_\theta}\right)}$$

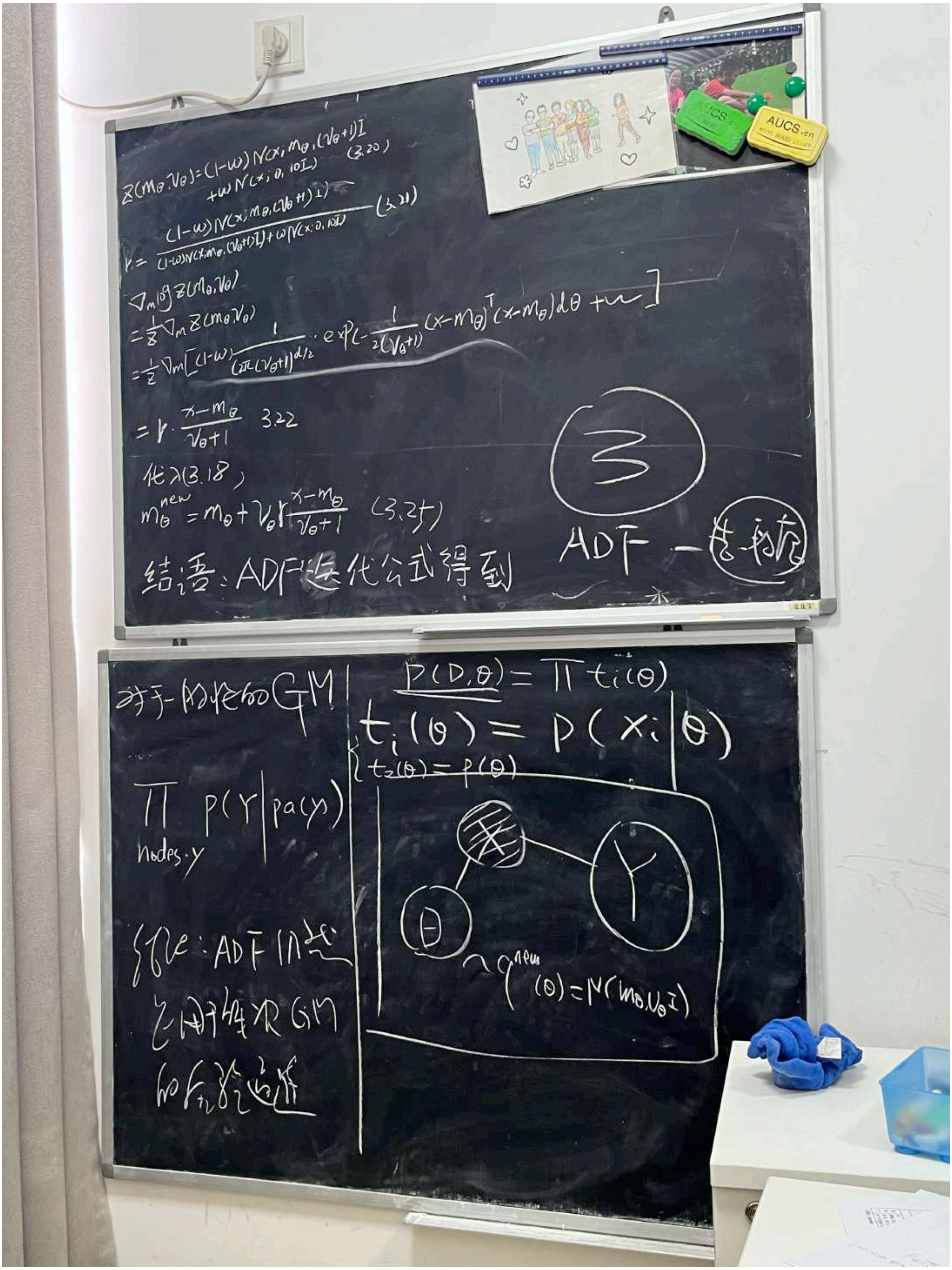
代回(1)式 $= \frac{1}{2\pi v_\theta^{d/2}}\int_\theta \exp\{[\quad]\}\exp\left\{\frac{\left(\frac{v_\theta x + m_\theta}{1+v_\theta}\right)^T\left(\frac{v_\theta x + m_\theta}{1+v_\theta}\right)}{2\frac{v_\theta}{1+v_\theta}} - \frac{v_\theta x^T x + m_\theta^T m_\theta}{2v_\theta}\right\}$

$$= \frac{(2\pi\frac{v_\theta}{1+v_\theta})^{d/2}}{2\pi v_\theta^{d/2}}\left(\frac{1}{(2\pi\frac{v_\theta}{1+v_\theta})^{d/2}}\int_\theta \exp\{[\quad]\}d\theta\right) \cdot \exp\left\{\frac{(v_\theta x + m_\theta)^T(v_\theta x + m_\theta)}{2v_\theta(1+v_\theta)} - \frac{v_\theta x^T x + m_\theta^T m_\theta}{2v_\theta}\right\}$$

$$= \frac{1}{[2\pi(1+v_\theta)]^{d/2}} \cdot \exp\left\{\frac{-v_\theta x^T x + 2v_\theta x^T m_\theta - v_\theta m_\theta^T m_\theta}{2v_\theta(1+v_\theta)}\right\}$$

$$= \frac{1}{[2\pi(1+v_\theta)]^{d/2}} \exp\left\{-\frac{(x-m_\theta)^T(x-m_\theta)}{2(1+v_\theta)}\right\}$$

$$= N(x; m_\theta, (1+v_\theta)I)$$


$Z(m_\theta, v_\theta) = (1-\omega) N(x; m_\theta, (v_\theta+1)I) + \omega N(x; 0, 10I)$ (3.20)
$r = \frac{(1-\omega) N(x; m_\theta, (v_\theta+1)I)}{(1-\omega) N(x; m_\theta, (v_\theta+1)I) + \omega N(x; 0, 10I)}$ (3.21)
$\nabla_m \log Z(m_\theta, v_\theta)$
$= \frac{1}{Z} \nabla_m Z(m_\theta v_\theta)$
$= \frac{1}{Z} \nabla_m [(1-\omega) \frac{1}{(2\pi(v_\theta+1))^{d/2}} \cdot \exp(-\frac{1}{2(v_\theta+1)} (x-m_\theta)^T (x-m_\theta)) d\theta + \ldots]$
$= r \cdot \frac{x-m_\theta}{v_\theta+1}$ 3.22
代入(3.18)
$m_\theta^{new} = m_\theta + v_\theta r \frac{x-m_\theta}{v_\theta+1}$ (3.24)
结语：ADF迭代公式得到
3
ADF
对于一般的GM
$\prod_{nodes\ y} p(Y|pa(y))$
$p(D,\theta) = \prod t_i(\theta)$
$t_i(\theta) = p(x_i|\theta)$
$t_0(\theta) = p(\theta)$
Θ
Y
$q^{new}(\theta) = N(m_\theta, v_\theta I)$
AUCS.cn

Minka2001年的中，ADF有关 m_θ, v_θ

的迭代

$$\begin{cases} m_\theta^{new} = m_\theta + v_\theta(\cdot)\dfrac{x_i - m_\theta}{v_\theta + 1} \\ v_\theta^{new} = \dots \end{cases}$$

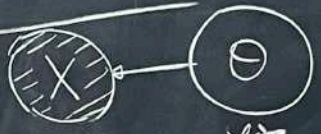
简写之：$\begin{cases} m_\theta^{new} \leftarrow m_\theta, v_\theta, x_i \\ v_\theta^{new} \leftarrow m_\theta, v_\theta, x_i \end{cases}$

$D = \{x_1, \dots, x_n\}$

$D|\theta$

$p(x) = (1-w)N(x; \theta, I) + wN(x; 0, 10I)$

GM （若改写 $D = X$）



$p(D, \theta) = p(D|\theta) \cdot p(\theta)$

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

$= \prod_{i=0} t_i(\theta)$

$t_0(\theta) = p(\theta)$ 先验

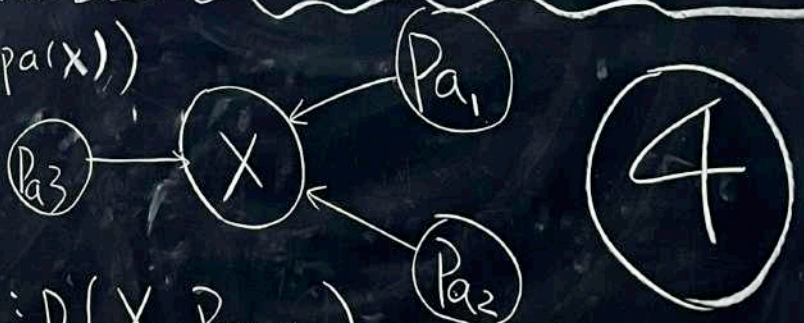
$t_i(\theta) = p(x_i|\theta)$ 后验

推出 $q^{new}(\theta)$ 再次 → θ 的后验 $p(\theta|D)$

则后面可用 θ 生成 D。

在一般的有向图模型

$\prod_{nodes} p(X|pa(X))$



joint Prob: $p(X, Parents)$

$= p(x|pa_1)p(x|pa_2)p(x|pa_3)\prod p(pa_i)$

形式上 $= \prod t_i(\cdot)$

3.2 Expectation Propagation

$$\tilde{t}_i(\theta) = Z \frac{q_{new}(\theta)}{q(\theta)} \quad (3.28)$$

与 $t_i(\theta)$ 对比

$$t_i(\theta) = p(x_i|\theta) = Z \frac{\hat{p}(\theta)}{q(\theta)}$$

用 $\tilde{t}_i$ 逼近 t_i

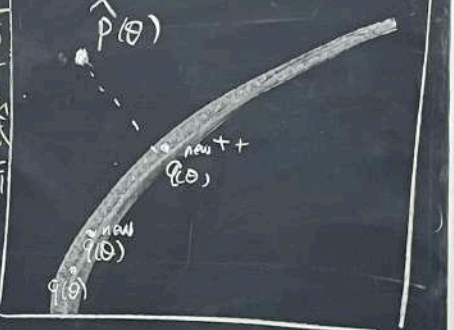
$$Z_i(m_\theta, V_\theta) = \int_\theta t_i q(\theta)\, d\theta \quad (3.29)$$

$$\tilde{Z}_i(\theta) = Z_i(m_\theta, V_\theta) \frac{q_{new}(\theta)}{q_i(\theta)} \quad (3.30)$$

$$= Z_i(m_\theta, V_\theta) \left(\frac{V_\theta}{V_\theta^{new}}\right)^{d/2} \exp\left(-\frac{1}{2V_\theta^{new}}(\theta - m_\theta^{new})^T\right.$$

$$\left.(\theta - m_\theta^{new})\right) \exp\left(\frac{1}{2V_\theta}(\theta - m_\theta)^T(\theta - m_\theta)\right) \quad (3.31)$$

因剔去节点 clutter 而逼近 $\tilde{t}_i$ 的方式



定义 (m_i, V_i)

$$V_i^{-1} = (V_\theta^{new})^{-1} - V_\theta^{-1} \quad (3.32)$$

$$m_i = V_i(V_\theta^{new})^{-1} m_\theta^{new} - V_i V_\theta^{-1} m_\theta \quad (3.33)$$

$$= m_\theta + (V_i + V_\theta) V_\theta^{-1}(m_\theta^{new} - m_\theta) \quad (3.34)$$

简化 $\tilde{t}_i(\theta)$

$$\tilde{t}_i(\theta) = \frac{Z_i(m_\theta, V_\theta)}{N(m_i; m_\theta, (V_i + V_\theta)I)} N(\theta; m_i, V_i I) \quad (3.36)$$

在EP中，后验是 $\prod_i \tilde{t}_i(\theta)$，但与ADF不同的是，每次取一个 $t_i(\theta)$，用 $\tilde{t}_i(\theta)$ 进行替代，得到最好的 $\tilde{t}_i(\theta)$，

5

3.2 Expectation Propagation

EP 算法流程

D ← θ

$\theta \sim q^{new}(\theta)$

1. 初始化 $\tilde{t}_i$ 的估计 $\tilde{t}_i$

2. 计算 θ 的后验

$$q^{new}(\theta) = \frac{\prod_i \tilde{t}_i(\theta)}{\int \prod_i \tilde{t}_i(\theta) d\theta} \quad (3.27)$$

（注：仍需附论 $q^{new}(\theta) \propto \prod t_i(\theta)$ ）

3. Until all $\tilde{t}_i$ converge:

(a) 选择一个 $\tilde{t}_i$

(b) 从后验中去除所选 $\tilde{t}_i$

$$q^{\setminus i}(\theta) \propto \frac{q^{new}(\theta)}{\tilde{t}_i(\theta)} \quad (3.38)$$

(c) 利用 ADF 计算 $q^{new}(\theta)$，使其逼近 $q^{\setminus i}(\theta)$

{ "嗯哼一句"： $q^{new} = \arg\min_{q^{new}} KL(q^{\setminus i} \| q^{new})$.

{ "通俗版"： $q^{new} = h(\theta)\exp(T(\theta)^T \eta^{new} - A(\eta^{new}))$, $E_{q^{new}}[T(\theta)] = E_{q^{\setminus i}}[T(\theta)]$ 求出 η^{new} 后再反求

(d) $\tilde{t}_i^{new} = z_i \frac{q^{new}(\theta)}{q^{\setminus i}(\theta)}$

4. $P(D) \approx \int \prod_i \tilde{t}_i(\theta) d\theta$ (3.40)

3(b) 另一种写法

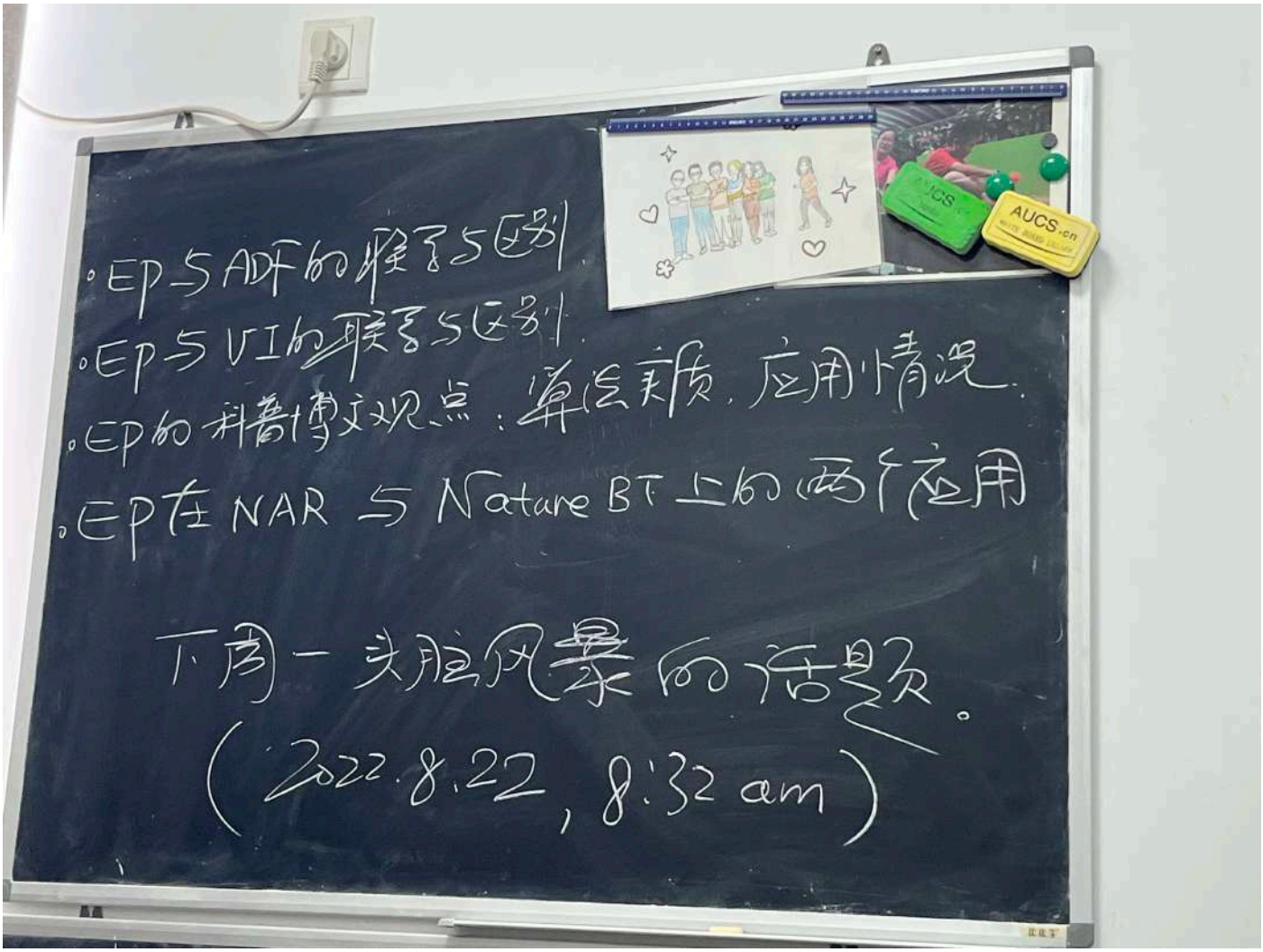
$$q^{\setminus i}(\theta) \propto \prod_{\eta_i} \tilde{t}_i(\theta) \quad (3.41)$$

以上为 D ← θ 情形下的一般性推导.

$\theta \sim q^{new}(\theta)$

EP 在 Minka2001 的论文中 clutter 问题的求解为示例，可直接查看 (3.42)～

6


EP与ADT的联系与区别
EP与VI的联系与区别
EP的科普博文观点：算法新质，应用情况
EP在NAR与Nature BT上的两个应用
下周一头脑风暴的话题。
(2022.8.22, 8:32 am)
AUCS-cn

2022/8/22 ADF/EP算法"头脑风暴"

- EP与ADF的联系与区别
- EP与VI的联系与区别
- EP的科普博文观点
- EP在NAR与Nature BT上的两篇应用.

子轮：ADF对x_t的顺序敏感
EP是ADF的扩展，适应性更高

庄逊：ADF每个样本只更新一次.
EP每个样本更新多次，它的选择是随机的

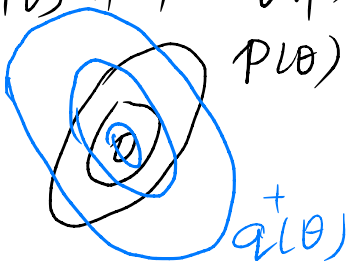
钱钱：ADF近似后验逼近精确后验。
EP近似 $\tilde{t}_i$，计算精确后验 $\hat{p}^{old}(\theta)$，再逼近 $\hat{p}^{old}(\theta)$
ADF是极小化 $KL(\hat{p}||q^{new})$，VI是极小化 $KL(q(\theta)||p(\theta|x))$

思卓：ADF受数据顺序和质量的影响比EP大，EP在考虑全部数据的情况下，进行单个数据的更新，具有更好的鲁棒性。

七北：ADF one-Pass算法，EP存在迭代收敛，(但不一定能成功)
VI的每次迭代是为了提高证据下界(ELBO)
VI使用mean-field方法，假设所有数据独立
EP在估计单个数据点时考虑全部数据点.
ADF是EP的online版本
"ADF和EP"都是增量学习

根据 Bishop 教材，思考 $KL(p\|q)$ 和 $KL(q\|p)$ 的区别

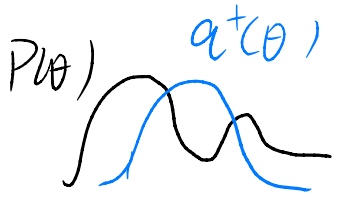
$$q^*(\theta) = \min_{q(\theta)} KL(q\|p)$$



$$q^+(\theta) = \min_{q(\theta)} KL(p\|q)$$

考虑 $p(\theta)$ 为双峰分布情况：

$$q^*(\theta) = \min_{q(\theta)} KL(q\|p)$$



$$q^+(\theta) = \min_{q(\theta)} KL(p\|q)$$

考虑离散的情况

- $KL(q\|p) = \sum q(\theta_i) \log \frac{q(\theta_i)}{p(\theta_i)}$

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

- $q(\theta_i)$ 小 → KL 不大
- $q(\theta_i)$ 大 → $p(\theta_i)$ 必须大

- $KL(p\|q) = \sum p(\theta_i) \log \frac{p(\theta_i)}{q(\theta_i)}$

$\min KL(p\|q) \Rightarrow$

- $p(\theta_i)$ 小 → KL 不大
- $p(\theta_i)$ 大 → $q(\theta_i)$ 必须大

字面理解

思考 VI 和 ADF 中的 KL

* VI 中的 KL

obs.
Data: x
Hidden / Latent Variable: z

$$\log P(x) = E_{z\sim q(z)}[\log P(x)]$$

Jensen不等式 $\geq E_q[\log P(x,z)] - E_q[\log q(z)]$

ELBO

$$ELBO = E_q[\log P(x) + \log P(z|x)] - E_q[\log q(z)]$$

$$= E_q[\log P(x)] + E_q\left[\log \frac{P(z|x)}{q(z)}\right]$$

$$= E_q[\log P(x)] - KL(q\|P)$$

拉高 ELBO，相当于（被动）减少 $KL(q\|P)$

P: unknown 不需要知道

---

* ADF 中的 KL

先选 exact posterior: $\hat{P}(\theta) = \frac{\text{新}\ t_i(\theta)\ \text{旧}\ q(\theta)}{Z}$

主动 极小化 $KL(\hat{P}(\theta)\|q^{new}(\theta))$

↳ 获取 矩匹配 q^{new} → $\hat{P}(\theta)$：部分知道，由 $q(\theta)$ 表征

$q^{new}(\theta)$ 通过矩匹配 由 $t_i, q^{old}(\theta)$ 组合

"ADF和EP都是增量学习"观点和记忆理解如下：

增量贝叶斯公式

$$p(\theta|D_{old}\cup D_{new}) \propto \underbrace{p(\theta|D_{old})}_{q(\theta)}\underbrace{p(D_{new}|\theta)}_{\prod t_i(\theta)}$$

传统贝叶斯公式

$$p(\theta|D)=\frac{p(\theta)p(D|\theta)}{p(\theta)} \propto p(\theta)\,p(D|\theta)$$

$p(\theta)$ 记为 $t_0(\theta)$

$p(D|\theta)$ 记为 $\prod t_i(\theta)$

$D=D^{old}\cup D^{new}$

由增量贝叶斯可直接得到

多种 $\hat{p}(\theta)=\prod t_i(\theta)\cdot q(\theta)$

〔ADF记忆与上述记忆思想一致性〕

主动设计 exact posterior $\hat{p}(\theta) \propto \underbrace{t_i(\theta)}_{新}\underbrace{q(\theta)}_{旧}$

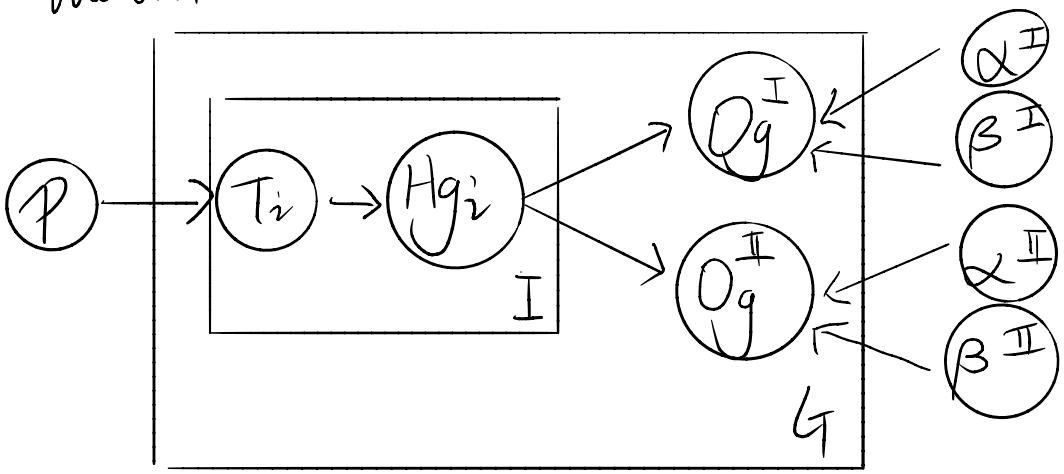
$\propto p(x_i|\theta)\cdot q(\theta)/Z$

$(Z=\int t_i(\theta)\,q(\theta)\,d\theta)$

Z是恒定的

EP算法在生信中的应用

[A modular framework for gene set analysis integrating multilevel omics data] NAR, 2013.



"I" for Items (Go term)

"G" for Genes

Latent Variable: T (Go条目，取 {0 inactive; 1 active)

H (Hg_i=1，当条目i与基因g关联，否之，Hg_i=0)

model parameters: α, β

{①数据融合，考虑数据不同系值性.
②给固定切一些.

NAR 2013
做基因集预测

输出一些条目i
和基因集{g}

- EP与ADF的联系与区别
- EP与VI的联系与区别
- EP的科普性文观点：算法本质，应用情况
- EP在NAR与Nature BT上的两个应用

对不起，迟到了一分钟

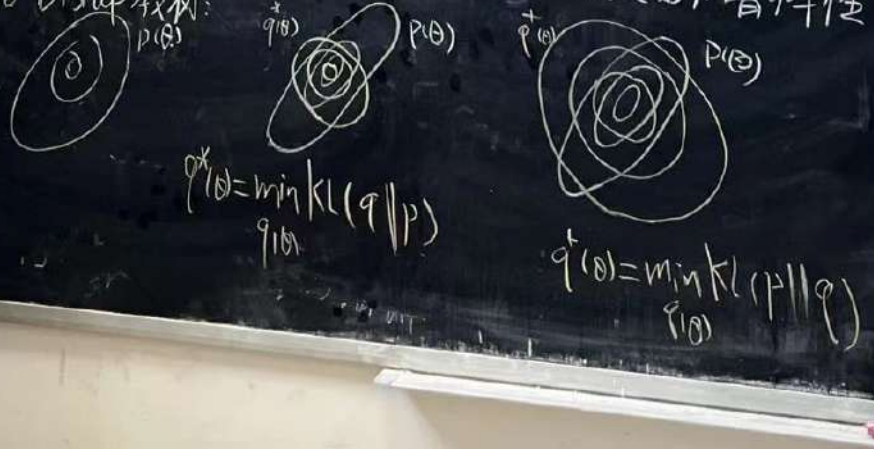
子铭：ADF对x_i的顺序敏感
EP是ADF扩展，适应性更高 (1)

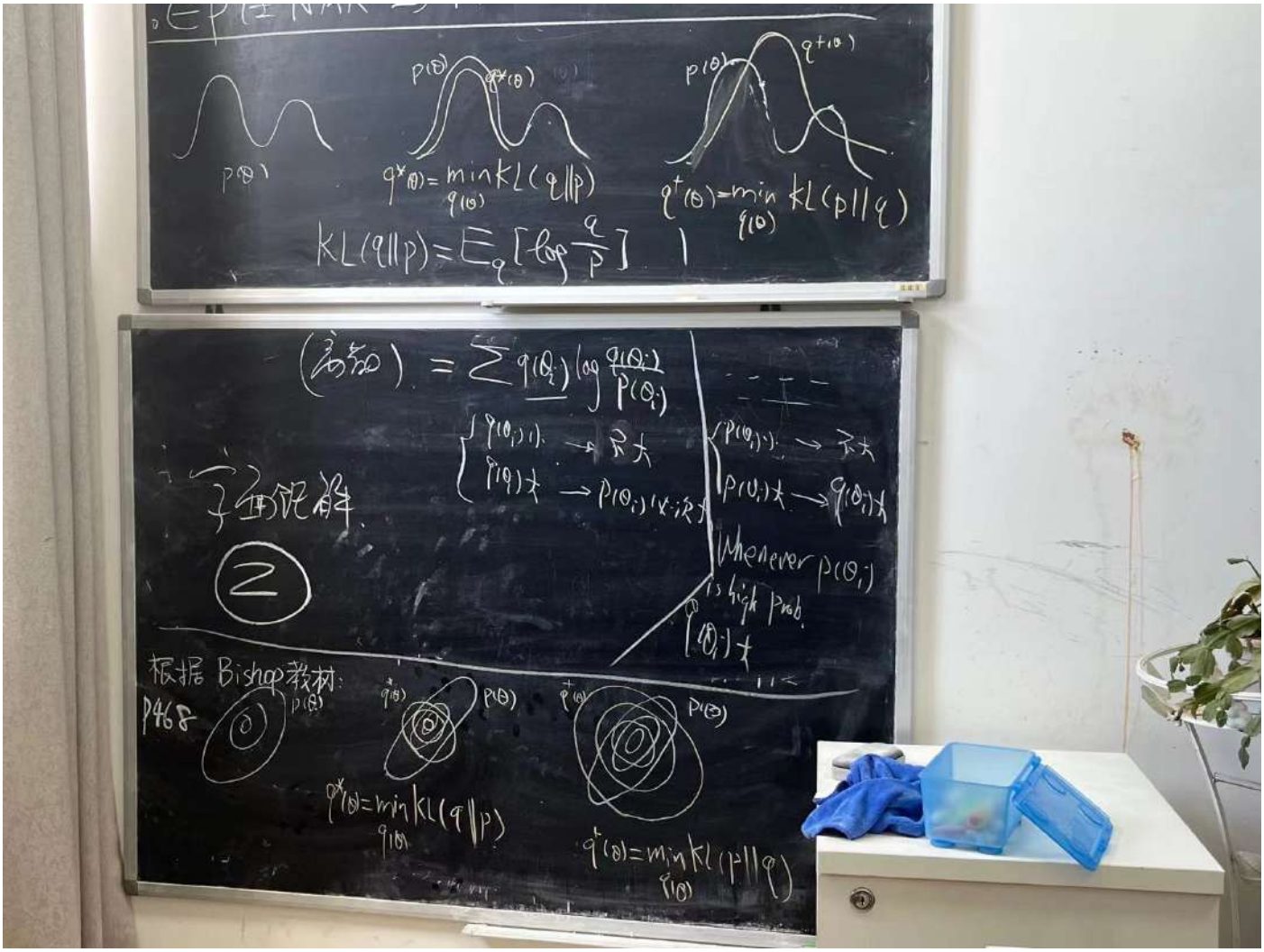
芷涵：ADF每个样本只更新一次
EP每个样本更新多次，t_i的选择是随机的

钱钱：ADF：近似后验逼近精确后验 $\hat{p}(\theta)$
EP：近似 t_i 计算精确后验 $\hat{p}(\theta)$，再逼近 $\hat{p}(\theta)$
ADF是极小化 $KL(\hat{p}\|q^{new})$，VI是极小化 $KL(q(\theta)\|p(\theta|D))$

思卓：ADF受顺序和质量的影响比EP大，EP在考虑全部数据的情况下考虑单个数据的更新，具有更好的鲁棒性

根据Bishop教材：




p(θ)
q*(θ)
q*(θ)=min KL(q||p)
q(θ)
KL(q||p)=E_q[log q/p]
q+(θ)
q+(θ)=min KL(p||q)
q(θ)
(高斯) = Σ q(θ_i) log q(θ_i)/p(θ_i)
子问题解
2
Whenever p(θ_i) is high prob.
板据 Bishop教材:
P468
q*(θ)=min KL(q||p)
q(θ)
q+(θ)=min KL(p||q)
q(θ)

- EP与ADF的联系与区别
- EP与VI的联系与区别
- EP的科普博文观点：算法本质，应用情况
- EP在NAR与Nature BT上的两个应用

姚：ADF：one-pass算法 EP存在迭代收敛(但不一定能成功)
VI的每次迭代是为了提高证据下界(ELBO)
VI使用mean-field方法，假设D都是独立的
EP在估计时考虑所有数据点
ADF只期望使搭online版本
"ADF和EP都是增量学习" —— 观点的论证如下！

增量贝叶斯公式

$$p(\theta|D_{old}\cup D_{new}) \propto p(\theta|D_{old})\,p(D_{new}|\theta)$$

$$p(\theta|D)=\frac{p(\theta)p(D|\theta)}{p(D)} \propto p(\theta)p(D|\theta)$$

$D = D^{old} \cup D^{new}$

由于各自独立 直接相乘

$p(\theta) = \prod_{i=1}^{n} t_i(\theta)$

$\prod_{i=1}^{n} t_i(\theta)$

[ADF的论文与上述公式的思想一致]：

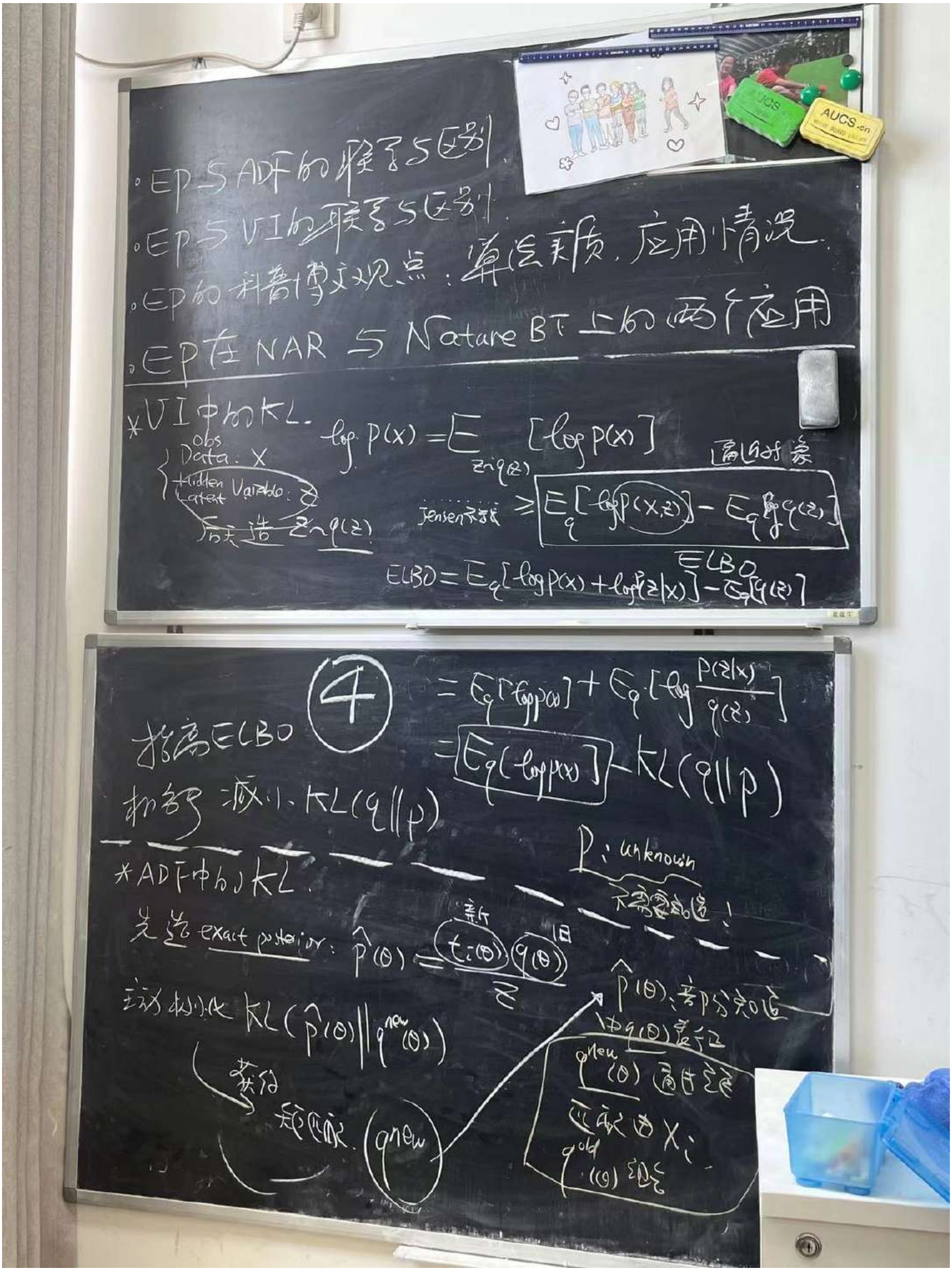
主动设计 exact posterior

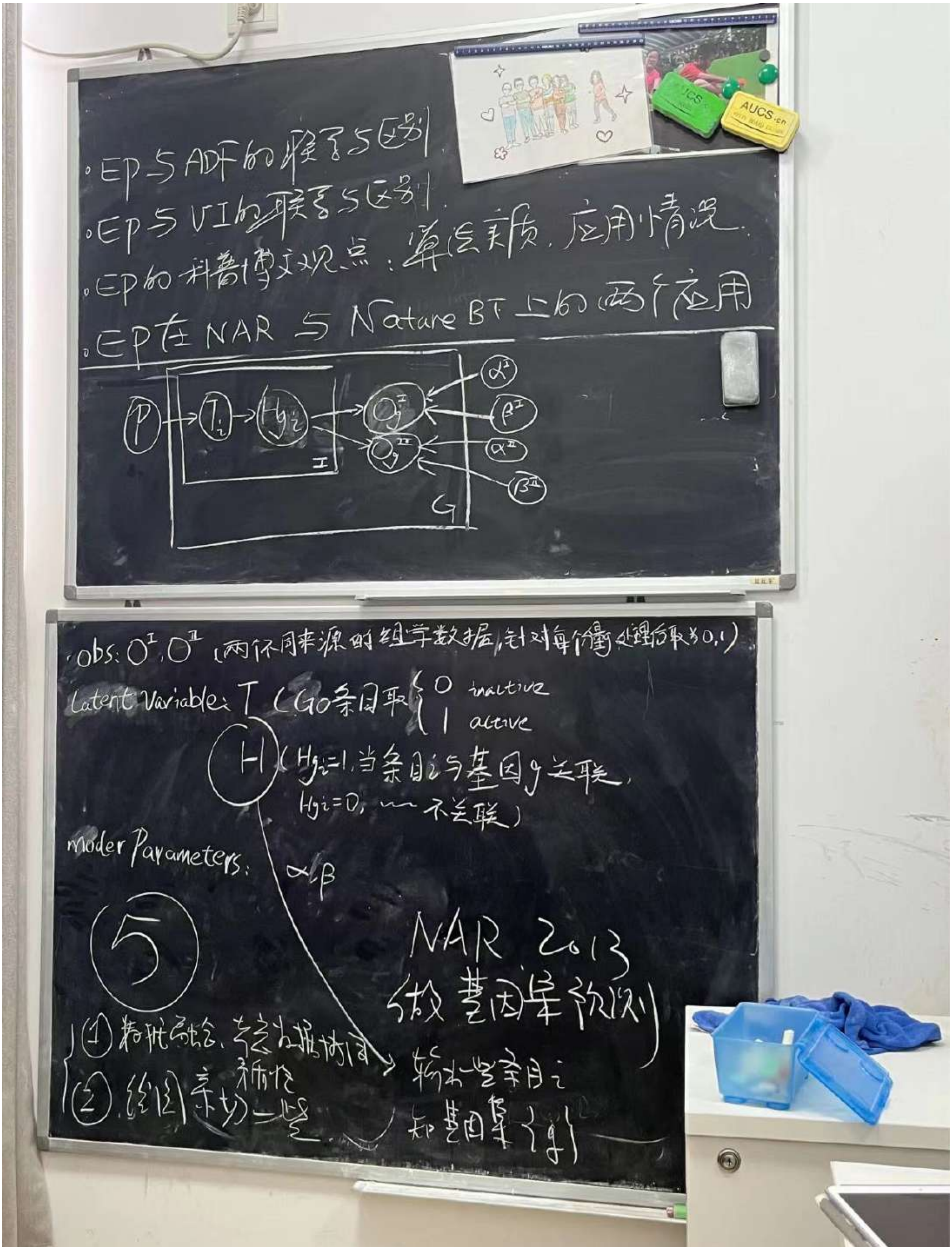
$$\hat{p}(\theta) \propto \frac{t_i(\theta)}{\text{新}}\,\underbrace{q(\theta)}_{\text{老}} \propto \frac{p(x_i|\theta)\cdot q(\theta)}{Z}$$

$(Z=\int t_i(\theta)\,q(\theta)\,d\theta$ 是归一化常数$)$

变种 $\hat{p}(\theta) = \prod_{i=1}^{n} t_i(\theta)\cdot q(\theta)$

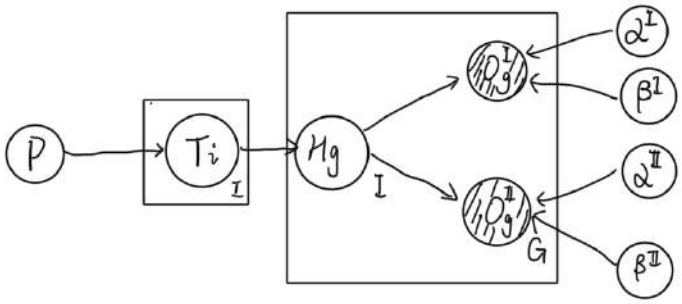
3


·EP与ADF的联系与区别
·EP与VI的联系与区别
·EP的科普博文观点：算法本质，应用情况
·EP在NAR与Nature BT上的两个应用
*VI中的KL
obs. Data: X
Hidden / Latent Variable: z
z~q(z)
log p(x) = E z~q(z) [log p(x)]
Jensen不等式
≥ Eq[log p(x,z)] − Eq log q(z)
ELBO
ELBO = Eq[log p(x) + log p(z|x)] − Eq[q(z)]
= Eq[log p(x)] + Eq[log p(z|x) / q(z)]
= Eq[log p(x)] − KL(q||p)
4
提高ELBO
减小 KL(q||p)
*ADF中的KL
先造 exact posterior: p̂(θ) = ti(θ) q(θ) / Z
新
旧
p: unknown
不需要知道
KL(p̂(θ)||q^new(θ))
q^new



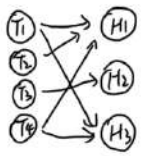
EP与ADF的联系与区别
EP与VI的联系与区别
EP的科普博文观点：算法实质，应用情况
EP在NAR与Nature BT上的两个应用
P
T
H
I
G
α
β
obs: O^I, O^II
Latent Variable: T
0 inactive
1 active
H
不关联
moder Parameters:
α, β
5
NAR 2013

8-23 NAR2013, Steffen



- Obs: O^I, O^{II} 组学数据
- 隐变量 { T: GO条目 / H: 基因 }
- Model parameter: P, α, β

例:



input: $\{T_1, T_4\}$

enriched gene $\{H_1, H_3\}$

获取富集基因集

模型应用: 通过 $P(H_g|T)$, T表示当前的所有GO条目

选中的条目 $T_i=1$ (否则为0), $\{g|H_g=1\}$

讨论点: 如何利用EP对该模型进行求解, 构成富集基因

$\Theta = \{P, T, H, \alpha^{I/II}, \beta^{I/II}\}$, $D = \{O^I, O^{II}\}$

$$P(\Theta|D) = \frac{P(T|P)\cdot P(H|T)\cdot P(D|H,\alpha,\beta)\cdot P(\alpha)\cdot P(\beta)\cdot P(P)}{P(D)}$$

(由图模型可以获得joint prob)

EP记号: $P(\theta|D) = \frac{\prod_{i}^{6} f_i(\theta)}{P(D)}$

ADF计算单位: $\hat{P} = \frac{1}{Z} t(\theta) q(\theta)$, $E_{\hat{P}}[T(\theta)] = E_{q^{new}}[T(\theta)]$

individual factors:

$$P \sim Beta(a,b) \quad (1)$$

$$P(H_g|T) = \begin{cases} 1, & \text{if } \exists T_j \in T(H_g),\ T_j = 1 \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

$$P(O_g^I = 1 | H_g) = \begin{cases} 1-\alpha^I, & \text{if } H_g = 1 \text{ (true positive)} \\ \alpha^I, & \text{if } H_g = 0 \text{ (false positive)} \end{cases}$$

$$P(T|P) = \begin{cases} 1, & \text{active} \\ 0, & \text{inactive} \end{cases}$$

注：查阅二项式分布 — 指数家族分布的 f(x)，未知的计算

？最终借助 2NFFR.NET （卡在此处）

8-23 NBT2014，Richard

任务：转录因子结合位点预测

μ：log-rates

Σ：协方差

$i,j \in \{1, \cdots, N\}$ 基因组上碱基

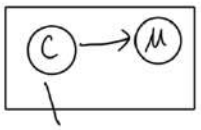
$k \in \{1, \cdots, K\}$ 实验次数

用EP估计 Latent rates μ

背景参数 $(\mu_0, \delta_0, k_{|i-j|}, \Sigma^{-1})$

binding identity : I_j

data(C_i)



DNase sequence后的 reads

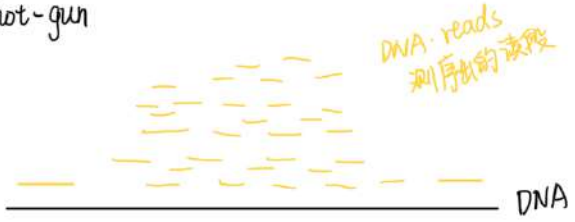
EP 通过背景参数和 data(C_i) 估计 μ 后验.

文章推断转录因子结合位点的步骤

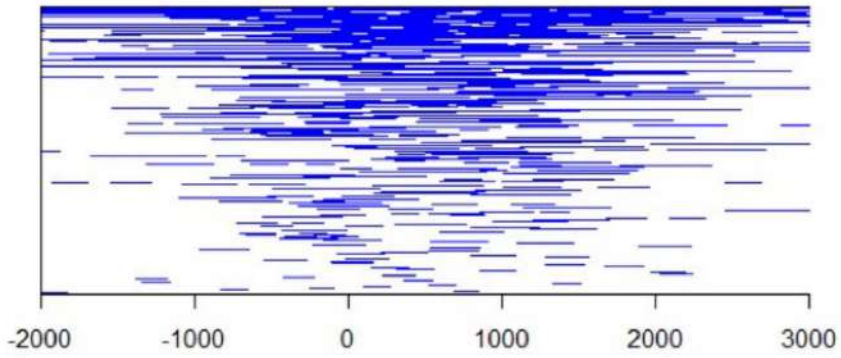
Step-1 把TF与DNA结合，再降解掉未结合的DNA部分，剩下的测出部分用来推测 TF binding site

Step-2 把剩下的DNA片断打散，匹配到 reference 序列，绘出 count 的峰图

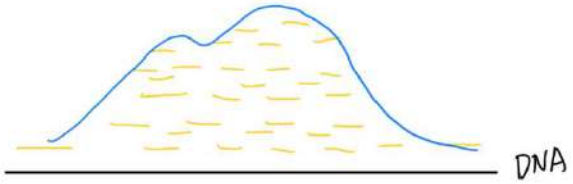
Shot-gun



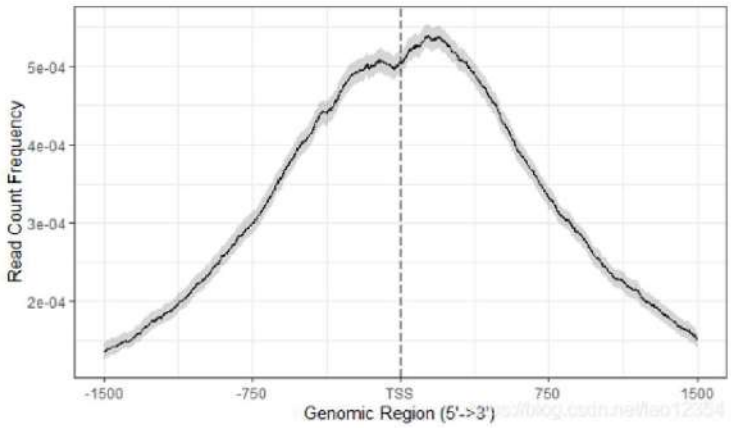
真实数据图例（来源CSDN）



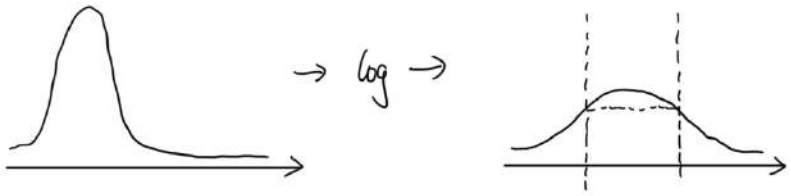
Step-3 count的峰图



真实数据图例（来源 CSDN）



Step-4 count的峰图往往是偏态信号，用log将其转为正态



Step-5 正态信号 μ, Σ （log-rate），再用EP估计μ，用以判断motif（binding site）

$q \xrightarrow{\text{估计}} p(u|c)$

$q = t_{pri} \cdot t_c \cdot t_{exp}$

概率图中 condition关系未知?

- t_{pri}: prior
- t_c: observed counts
- t_{exp}: intra-experiment

$u_{N\times K}$: N bases, K experiments 的 count数

$\Sigma_{K\times K}$: K experiments

$\Sigma_{N\times N}$: 协方差矩阵

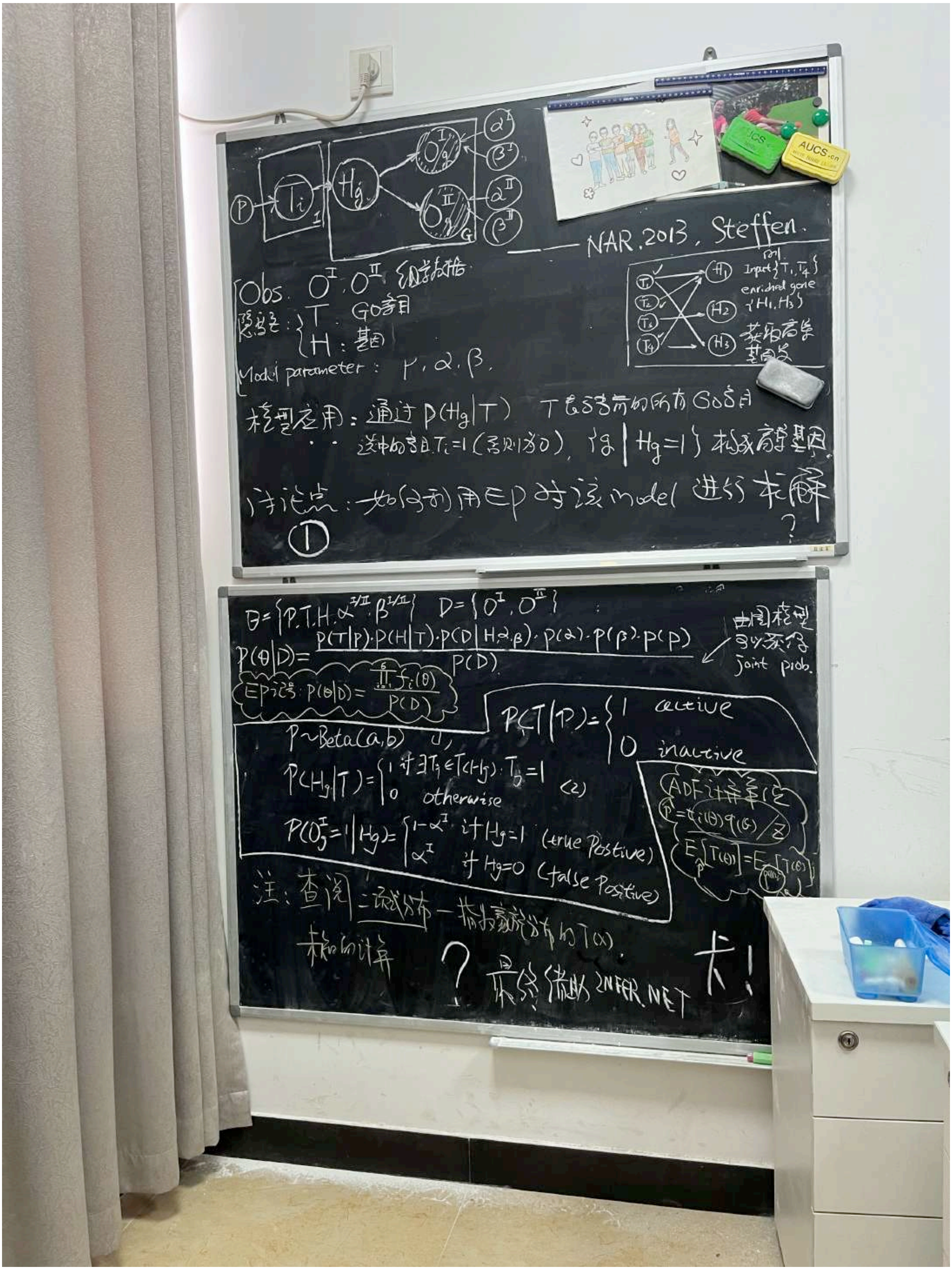
$$q(u|c) \propto t_{pri}(u|\Sigma, u_0)\, t_c(c_i|u)\, t_{exp}(u_k|u_{-k})$$

余味:

EP算景图模型的计算过程

ADF法号

$$q(\theta) \propto \prod t_i(\theta)$$
$$= p(\theta)\cdot p(x_1|\theta)\cdot p(x_2|\theta)$$
$$= t_0(\theta)\cdot t_1(\theta)\cdot t_2(\theta)$$


NAR.2013. Steffen.
Obs. O^I, O^II
Model parameter: p, α, β.
Input {T1, T4}
enriched gene {H1, H3}
D={O^I, O^II}
joint prob.
P~Beta(a,b)
active
inactive
otherwise
(true Postive)
(false Positive)
INFER.NET

NBT 2014, Richard,

任务：转录因子结合位点预测

μ：同一rates

Σ：协方差

$i \in \{1, \dots N\}$ 基因组上碱基

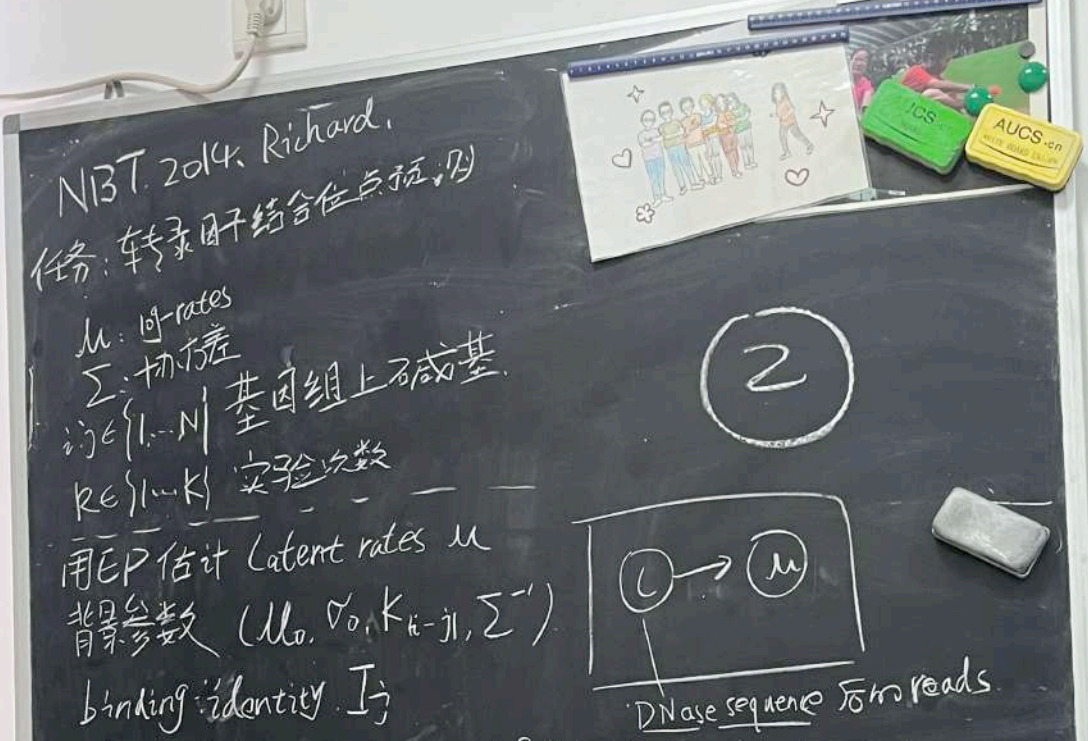
$k \in \{1, \dots K\}$ 实验次数

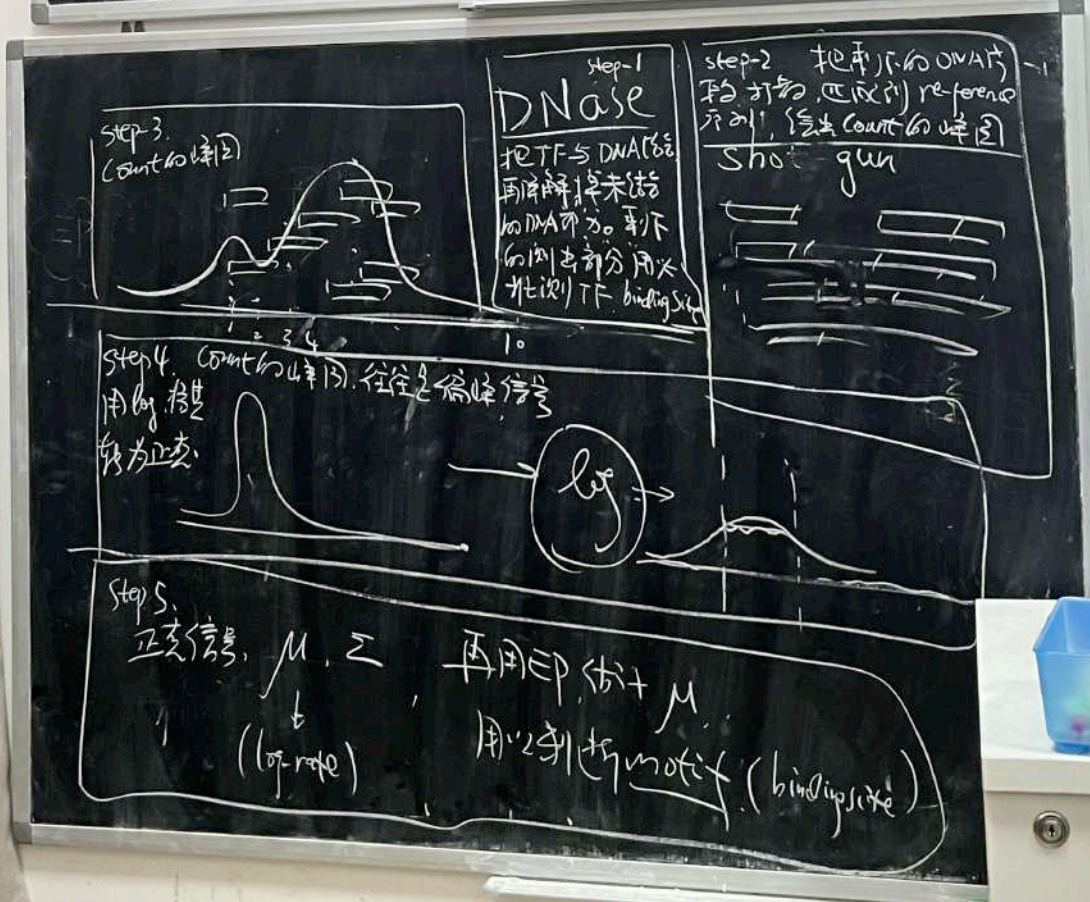
用EP估计 Latent rates μ

背景参数 $(\mu_0, \gamma_0, k_{k-1}, \Sigma^{-1})$

binding identity I_i

data (c_i) ，EP通过背景参数和data(c_i)估计从位点





NBT. 2014. Richard.

任务：转录因子结合位点预测

μ：motif rates

Σ：协方差

$i \in \{1 \cdots N\}$ 基因组上碱基

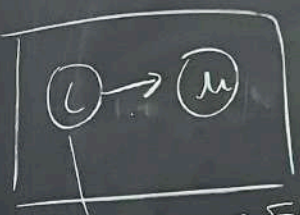
$k \in \{1 \cdots K\}$ 实验次数

用EP估计 Latent rates μ

背景参数 $(\mu_0, \sigma_0, k_{k-jl}, \Sigma^{-1})$

binding：identity. 与

data (c_i) ．EP通过背景参数和data(c_i)估计 μ 后验



DNase sequence 后的 reads

q 估计 $\rightarrow p(\mu|c)$

$q = t_{Pri} \cdot t_c \cdot t_{exp}$

t_{Pri}：prior

t_c：observed counts

t_{exp}：intra-experiment：texp

$\mu_{N \times k}$：N bases, k experiments 的 count 数

$\Sigma_{k \times k}$：k experiments

$\Sigma_{N \times N}$

$q(\mu|c) \propto t_{Pri}(\mu|\Sigma, \mu_0)\, t_c(c_i|\mu)\, t_{exp}(\mu_k|\mu_{-k})$

模型外 condition 导致未知？

ADF 论文

$q(\theta) \propto \prod t_i(\theta)$

$= p(\theta) \cdot p(x_1|\theta)\, p(x_2|\theta)$

$t_0(\theta) \cdot t_1(\theta) \cdot t_2(\theta)$

含义：

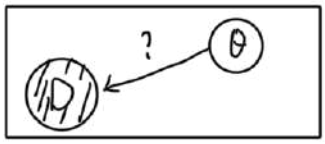
EP 算某 GM 后计算出来

关于EP的一些算法上的考虑.

—— 算法思想，必须的算法步骤

得益于她参考的"增量贝叶斯"的推法，看ADF

ADF：
(Mink 2001)

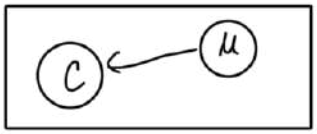


obs: $D=\{x_1, x_2, \cdots, x_n\}$

θ：待估计对象

clutter 问题，$p(x|\theta)=(1-w)N(\theta,I)+w\cdot N(0,10I)$ 杂波

NBT 2014-文



reads的counts，取log后（称为log-rate）

由(μ,Σ)生成

ADF-般性idea

$\hat{p}(\theta)$：exact porterior

$$\simeq \frac{t_i(\theta)q(\theta)}{Z} = \frac{t_i(\theta)q(\theta)}{\int t_i(\theta)q(\theta)}$$

$$\propto t_i(\theta)\,q(\theta) \qquad (t_i(\theta)=p(x_i|\theta))$$

$q^{new}(\theta) \xrightarrow{\text{逼近}} p(x_i|\theta)\cdot q(\theta)/Z$

$q^{(t+1)}(\theta) \xrightarrow{\text{逼近}} p(x_i|\theta)\,q^{(t)}(\theta)/Z$

$m_\theta^{new}, V_\theta^{new}$ ← 迭代 ← {贡献的x_i，贡献的(m_θ, V_θ)}

{贡献的x_i：online来的数据；贡献的(m_θ, V_θ)：旧数据}

$$\begin{cases} q^{new}(\theta) \sim pdf(\theta\text{-par}^{new}) \\ p(x_i|\theta) \sim pdf(\text{与}\theta\text{有关}) \\ q(\theta) \sim pdf(\theta\text{-par}) \end{cases}$$

迭代式的计算：$q^{new}(\theta) = \mathcal{N}(m_\theta^{new}, v_\theta^{new} I)$

$= \exp\left(\begin{pmatrix}\theta \\ \theta^T\theta\end{pmatrix}^T \eta - A(\eta)\right)$

$E_{q(\theta)}^{new}\begin{bmatrix}\theta \\ \theta^T\theta\end{bmatrix} = E_{\hat{p}}\begin{bmatrix}\theta \\ \theta^T\theta\end{bmatrix}$ .....

必须的计算：$q(\theta)$ - pdf (主动定义)

$p(x|\theta)$ - pdf

算 $q(\theta)\cdot p(x|\theta)$ pdf 的乘积 化为 pdf

就可以进行矩匹配

假设一：clutter. 但不是球形高斯，是其他分布.

也需要计算

假设二：能否把 log 之后的 NBT 数据看成一个 clutter 问题

EP 的一般性 idea

最开始 $q^{new} = \tilde{t}_1\ \tilde{t}_2 \cdots \tilde{t}_n$. 初始化 $\tilde{t}_i$

$q^{new}(\theta)$ ⟼ 逼近 → $p(x_i|\theta)\ q^{\setminus i}(\theta)/Z$ （增量）

新 $\tilde{t}_1\ \tilde{t}_2 \cdots t_i \cdots \tilde{t}_n$ ； 旧 $q^{\setminus i}(\theta) = \tilde{t}_1\ \tilde{t}_2 \cdots \tilde{t}_i \cdots \tilde{t}_n$

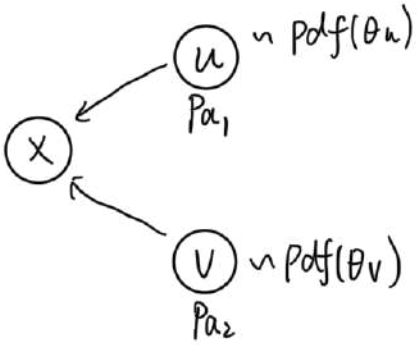
(设为) $= \mathcal{N}(m_\theta, v_\theta I)$

$q^{new}(\theta) / q^{\setminus i}(\theta)$ 取为新的 $\tilde{t}_i$

必须的计算环节：$\frac{q_1(\theta)}{q_2(\theta)}$ 的 pdf (θ) 见 (3.30)

代入到 (3.49) $q^{\setminus i}(\theta) \propto \frac{q^{new}(\theta)}{\tilde{t}_i(\theta)}$

EP和ADF的联系：都是增量贝叶斯



$$P(X,U,V) = P(U,V|X)\cdot P(X)$$
$$= P(X|U)P(X|V)P(U)\cdot P(V)$$

一般GM：joint prob

$$P(x, Pa(x)) = P(X|Pa_1)\cdot P(X|Pa_2)\cdot P(Pa_1)\cdot P(Pa_2)$$
$$= \prod_i \underbrace{p(X|Pa_i(X))}_{t_i(\theta)} \prod_i (Pa_i)$$

后验 $P(U,V|X) \propto P(X|U)P(X|V)P(U,V)$

$$\propto t_1(\theta_U)t_2(\theta_V)P(U,V)$$

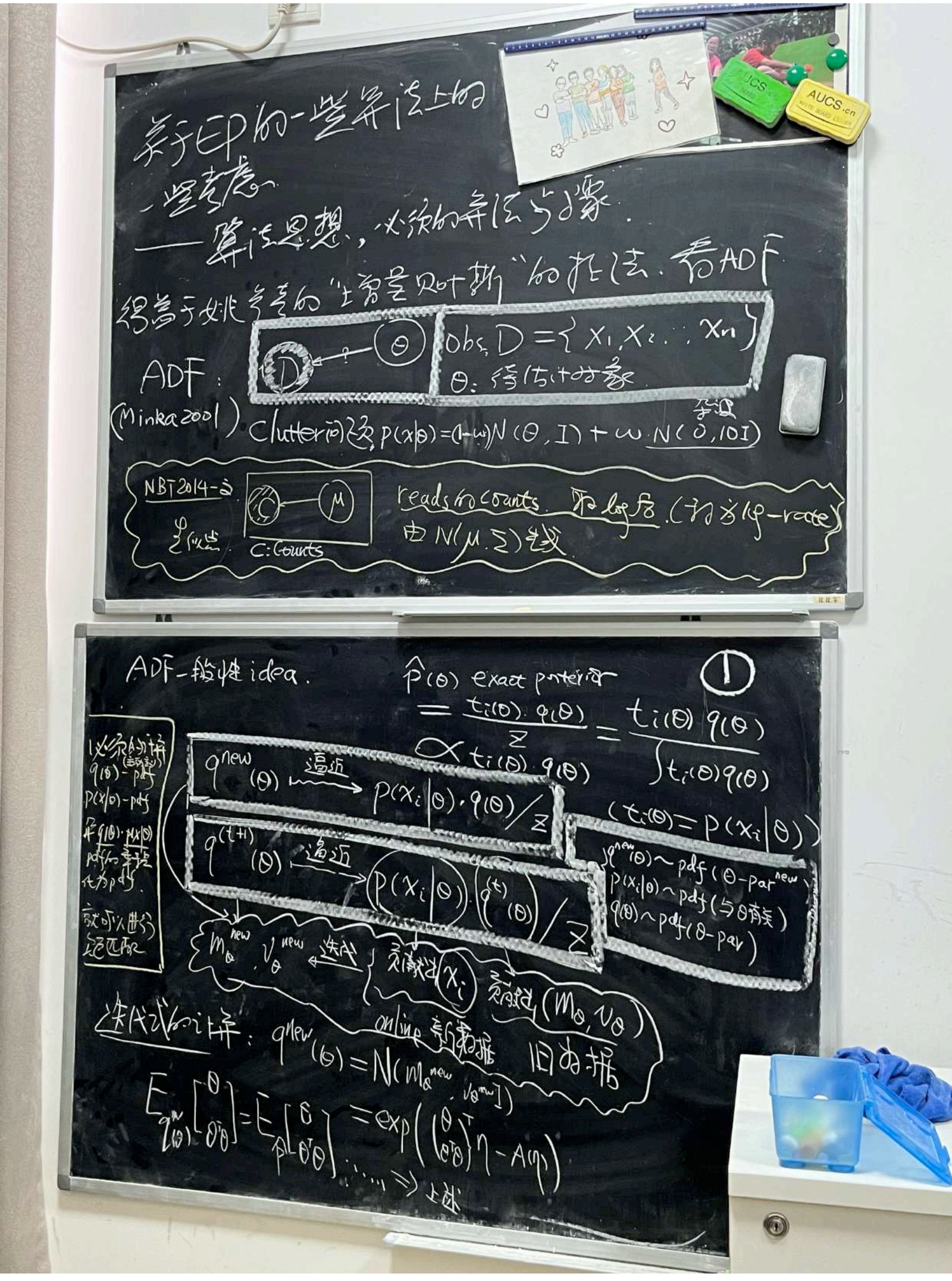
$q(U,V) \xrightarrow{\text{逼近}} P(U,V|X) \propto t_1(\theta_U)t_2(\theta_V)\cdot P(U,V)$

一般的 $q(X_{par}) \xrightarrow{\text{逼近}} P(X_{par}|X) \propto \prod_i t_i(\theta_{par_i})\cdot P(X_{par})$

X_{par}：X的所有父结点　理想后验　似然-增量　先验

ADF：D ← θ　$q^{new} \longmapsto t_i(\theta)\, q(\theta)$

EP：D ← θ　$q^{new} \longmapsto t_i(\theta)\, q^{\setminus i}(\theta)$


关于EP的一些算法上的
一些考虑
——算法思想，必须的算法步骤
得益于姚老师的"步步贝叶斯"的推法，看ADF
ADF:
(Minka 2001)
obs: D = {x1, x2 … xn}
θ: 待估计对象
Clutter问题 p(x|θ) = (1-w)N(θ, I) + w·N(0, 10I)
NBT2014
C: Counts
reads的counts 取log后（相当于log-rate）
由N(μ, Σ)生成
ADF-一般性idea.
p̂(θ) exact posterior
= ti(θ)·q(θ)/Z = ti(θ)q(θ)/∫ti(θ)q(θ)
∝ ti(θ)·q(θ)
q^new(θ) 逼近 p(xi|θ)·q(θ)/Z
q^(t+1)(θ) 逼近 p(xi|θ)·q^(t)(θ)/Z
(ti(θ) = p(xi|θ))
q^new(θ) ~ pdf(θ-par^new)
p(xi|θ) ~ pdf(与θ有关)
q(θ) ~ pdf(θ-par)
q(θ)-pdf
p(x|θ)-pdf
m_θ^new, v_θ^new 迭代
新数据(xi)
旧数据(m_θ, v_θ)
迭代式的计算: q^new(θ) = N(m_θ^new, v_θ^new)
online 新数据 旧数据
= exp((θ, θ'θ)^T η - A(η))
……⇒ 上述
AUCS.cn

关于EP的一些算法上的一些考虑

——算法思想，必须的算法与步骤.

得益于姚老师的"主要是贝叶斯"的指点. 看ADF

ADF: (Minka 2001) Clutter问题 $p(x|\theta)=(1-\omega)N(\theta,I)+\omega N(0,10I)$

$t_i(\theta)=p(x_i|\theta)$

obs: $D=\{x_1,x_2,\dots,x_n\}$

θ: 待估计的参数

(2)

EP一般性idea. 0) 初始化 $\tilde{t}_i$

(3.30)

看开始 $q^{new}=\tilde{t}_1\tilde{t}_2\cdots\tilde{t}_n$

$q^{new}(\theta)$ 逼近 $p(x_i|\theta)$ $q^{-i}(\theta)$

搞定!

$q^{-i}(\theta)\propto\frac{q^{new}(\theta)}{\tilde{t}_i(\theta)}$ (3.49)

$q^{-i}(\theta)=\tilde{t}_1\tilde{t}_2\cdots\cancel{\tilde{t}_i}\cdots\tilde{t}_n$

(设为) $=N(m_{\theta^{-i}},V_{\theta^{-i}}I)$

新 旧

$\tilde{t}_1\tilde{t}_2\cdots t_i\cdots\tilde{t}_n$

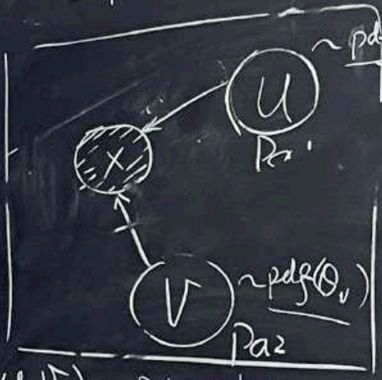
$\frac{q^{new}(\theta)}{q^{-i}(\theta)}$ 看为新的 $\tilde{t}_i$ (3.55) ← (3.30)

必须的计算环节. $\frac{q_1(\theta)}{q_2(\theta)}$ 的pdf(θ) 见(3.30)

EP. ADF: 均是贝叶斯

关于EP的一些算法上的一些考虑

—— 算法思想，必须的算法与对象



一般GM

joint prob.

$$P(X, Par(X)) = P(x|Pa1)\,P(x|Pa2)\,P(Pa1)\,P(Pa2)$$

$$= \prod_i P(x|pa_i(x)) \prod_i (pa_i)$$

$\prod_i P(x|pa_i(x))$: $t_i(\theta)$

$$p(x,u,v) = p(u,v|x)\cdot p(x) = p(x|u)\,p(x|v)\,p(u)\,p(v)$$

(3)

$$p(u,v|x) \propto p(x|u)\,p(x|v)\,p(u,v)$$

后验 $\propto t_1(\theta_u)\, t_2(\theta_v)\, p(u,v)$

$q(u,v)$ 逼近 $p(u,v|x) \propto t_1(\theta_u)\, t_2(\theta_v)\, p(u,v)$

一般的 $q(X_{par})$ 逼近 理想后验 $P(X_{parent}|X) \propto \prod_i t_i(\theta_{par_i}) \cdot P(X_{par})$

X_{par}: X的所有父结点

后验

似然 — 计算量

$P(X_{par})$ → 先验

ADF: (D)←(θ) $q^{new} \mapsto t_i(\theta)\, q(\theta)$

EP: (D)←(θ) $q^{new} \mapsto t_i(\theta)\, q^{-i}(\theta)$