概率图模型理论及应用

Theory and Applications of Probabilistic Graphical Models

(Lesson 6 - variational)

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

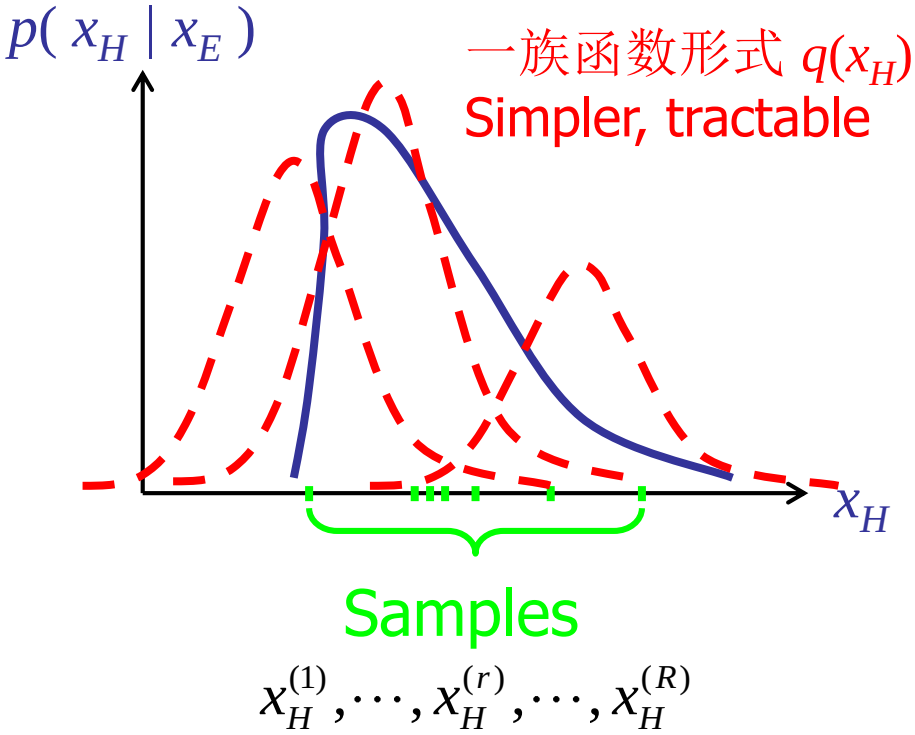
课程章节

- 第一章 图模型的表示理论（2）
  - **Semantics (DGM, UGM)**
  - **HMM, CRF**
- 第二章 图模型的推理理论（4）
  - 精确推理：**variable-elimination，cluster-tree，triangulate**
  - 连续变量：**Kalman**
  - 采样近似：**sampling**
  - 变分近似：**variational**
- 第三章 图模型的学习理论（2）
  - 参数学习：**maxlikelihoodEstimate，RFLearning, BayesEstimate**
  - 结构学习：**StructureLearning**

	pgm-1 semantics √		pgm-2 hmm-crf √ pgm-3 exact √	pgm-4 kalman √ pgm-5 sampling √
pgm-6 variational pgm-7 ML	pgm-8 Bayesian			

近似求解分布函数 $p(x_H \mid x_E)$

- 采样近似
  - 样本数足够多可任意近似
  - 适用任意分布函数
  - 速度慢，不适应大规模问题

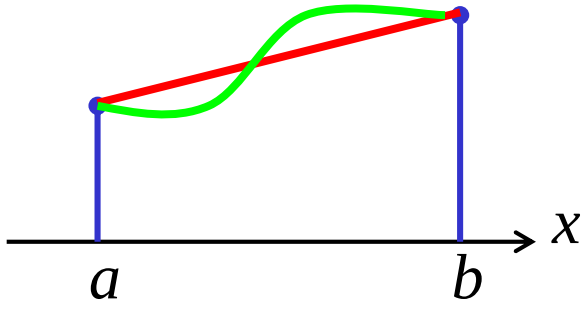


- 变分近似
  - 速度较快
  - 可应用于大规模问题
  - 较难分析近似误差
  - 将条件分布的求解 形式化成 一个最优化问题

$$\hat{q}(x_H) = \arg\min_{q} \underbrace{KL\left(q(x_H) \middle\| p(x_H \mid x_E)\right)}_{J(q(x_H))}$$

变分方法是一种经典的泛函最优化方法

- 泛函最优化
  - 例：求平面内两点之间具有最短长度的曲线



$$\max_{f} J(f)=\int_a^b \sqrt{1+\dot{f}^2}\,dx$$

$$f \in \{[a,b]\text{上的连续函数集合，}f(a),f(b)\text{固定}\}$$

- 泛函微分(Frechet微分)

$$\delta J(f;h)=\lim_{\alpha\to 0}\frac{J(f+\alpha h)-J(f)}{\alpha}=\frac{\partial J}{\partial f}\circ h$$

D. G. Luenberger，“最优化的矢量方法”，O244 35

变分近似推理

- 变分方法是一种经典的泛函最优化方法
- 变分近似推理：变分优化方法用于推理问题
- Block approach
  - 变分均值场方法（Variational mean field）
  - 结构变分方法（Structured variational approach）
  - 变分贝叶斯方法（Variational Bayesian）用于贝叶斯参数估计
- Sequential approach
  - Local variational method

Variational Inference for $p(x_H \mid x_E)$

用一个简单的好操作的函数 $q(x_H)$ 去近似真实函数 $p(x_H \mid x_E)$

$$\hat{q}(x_H) = \arg\min_q KL\left(q(x_H) \,\|\, p(x_H \mid x_E)\right)$$

Three steps …

① Use Kullback-Leibler distance $KL(q||p)$ as a measure of 'difference' between $p(x_H|x_E)$ and $q(x_H)$.

② Choose a family of variational distributions (变分分布) $q(x_H)$.

③ Find $q(x_H)$ which minimises KL distance.

① Minimise the KL distance

$$KL(q \| p) = \sum_{x_H} q(x_H) \log \frac{q(x_H)}{p(x_H \mid x_E)}$$

fixed maximise minimise

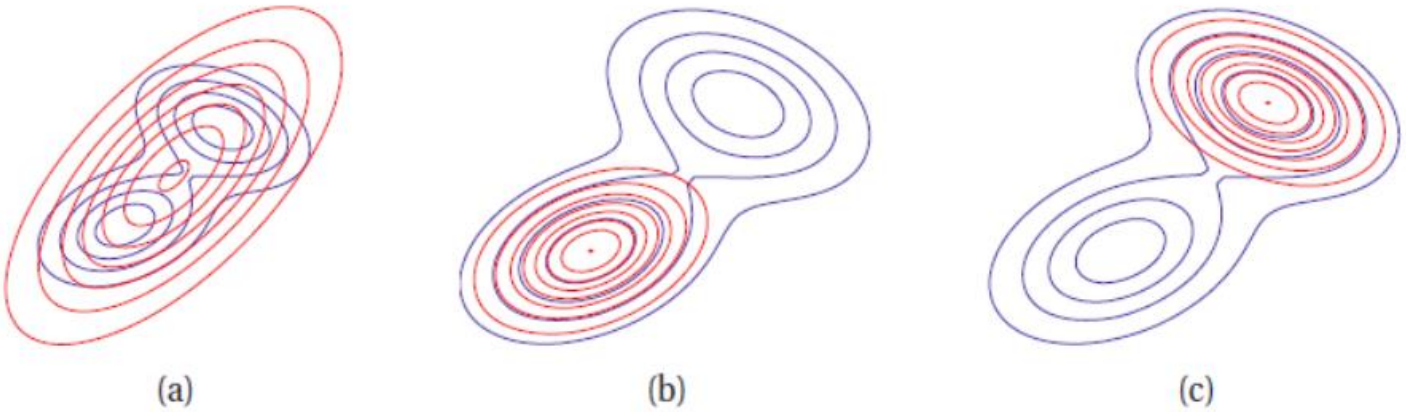
$$\log p(x_E) = L(q) + KL(q \| p)$$

$$L(q) = \sum_{x_H} q(x_H) \log \frac{p(x_H, x_E)}{q(x_H)}$$ Minus Free Energy

$$-L(q) = -\sum_{x_H} q(x_H) \log p(x_H, x_E) - \sum_{x_H} q(x_H) \log \frac{1}{q(x_H)}$$

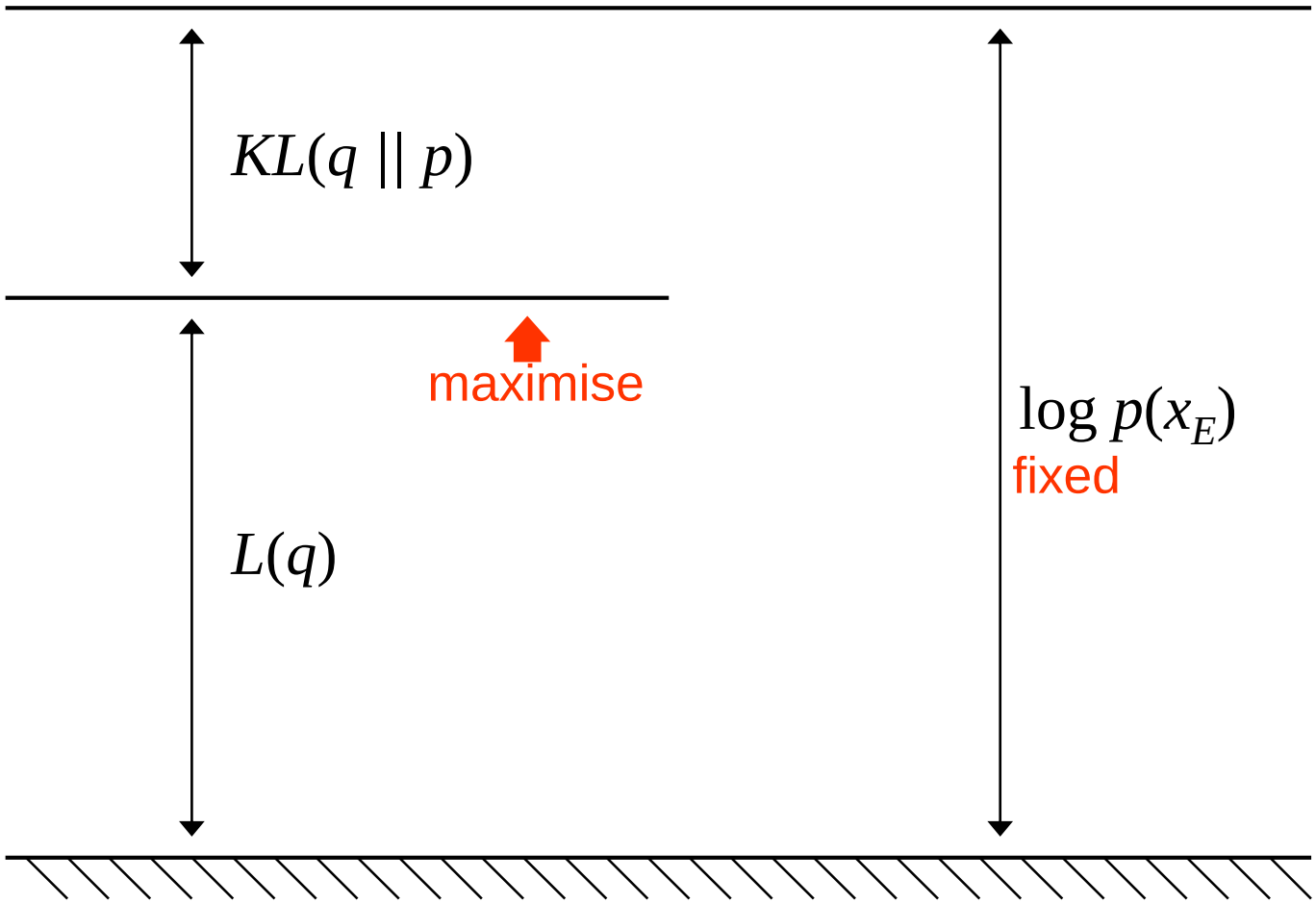
$KL(p \| q)$ Expectation Propagation (Minka, 2001), PRML 10.7

Discussion (Mackay book / Murphy book)

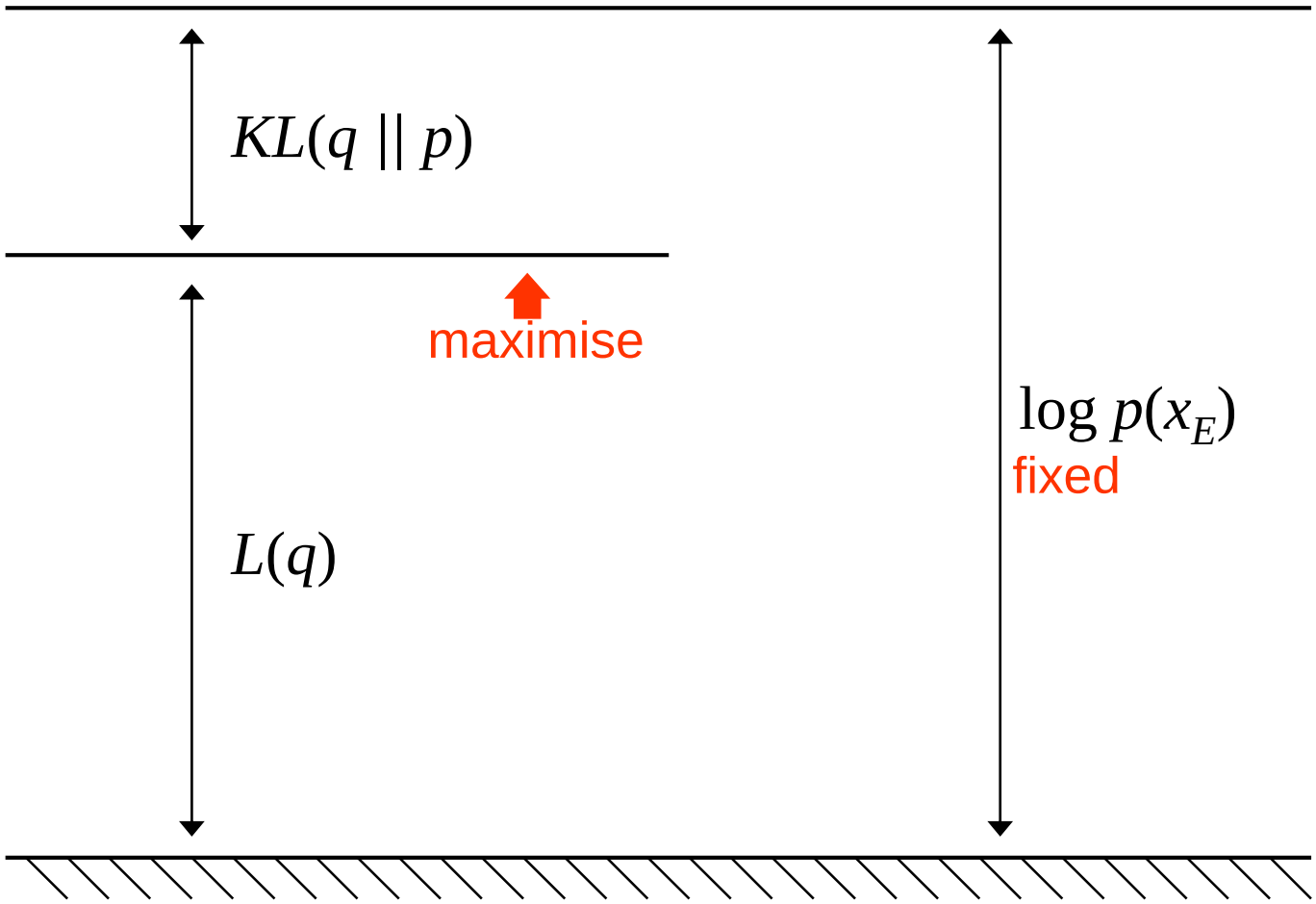


- Exclusive KL / Reverse KL: $\mathrm{KL}(\mathrm{q}||\mathrm{p}) = \int q log \frac{q}{p}$
  - Zero forcing (迫零) for q: if p=0 we must ensure q=0.
  - q will typically under-estimate the support of p.
  - q locks on to one of the two modes.
- Inclusive KL / Forwards KL: $\mathrm{KL}(\mathrm{p}||\mathrm{q}) = \int p log \frac{p}{q}$
  - Zero avoiding (避零) for q: if p>0 we must ensure q>0.
  - tends to find q that has higher entropy than the original
  - q tends to "cover" p

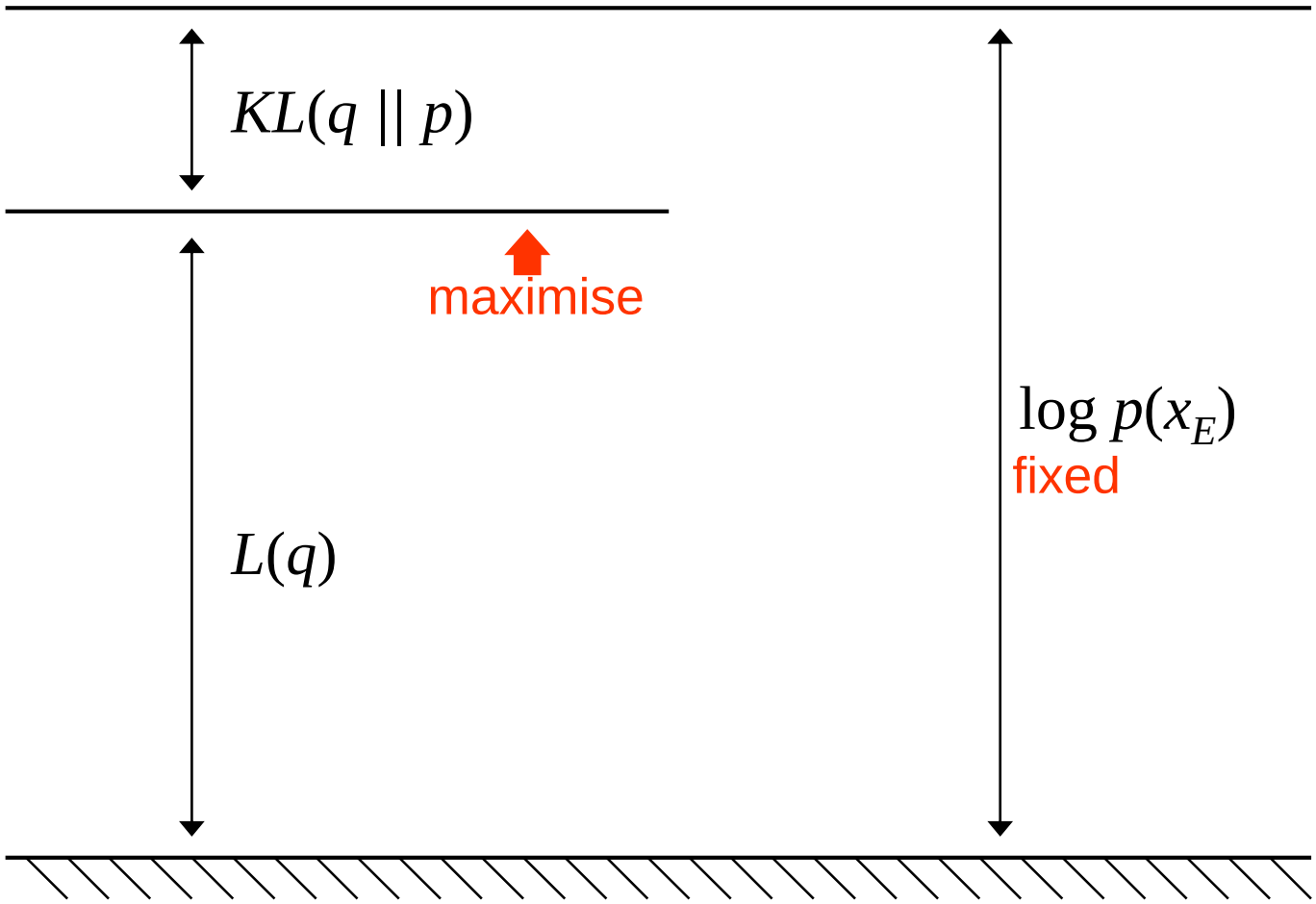
① Minimise the KL distance



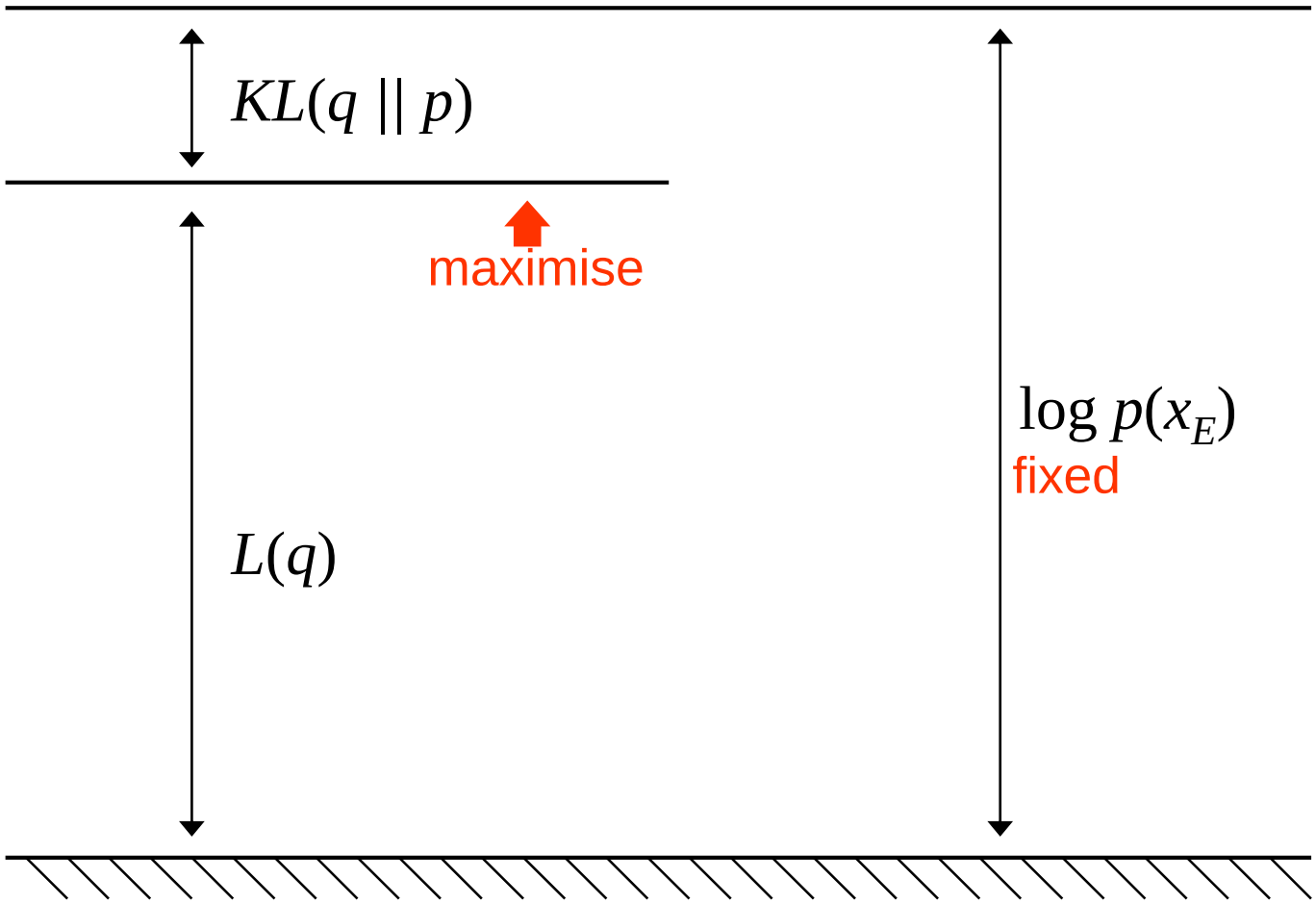
① Minimise the KL distance



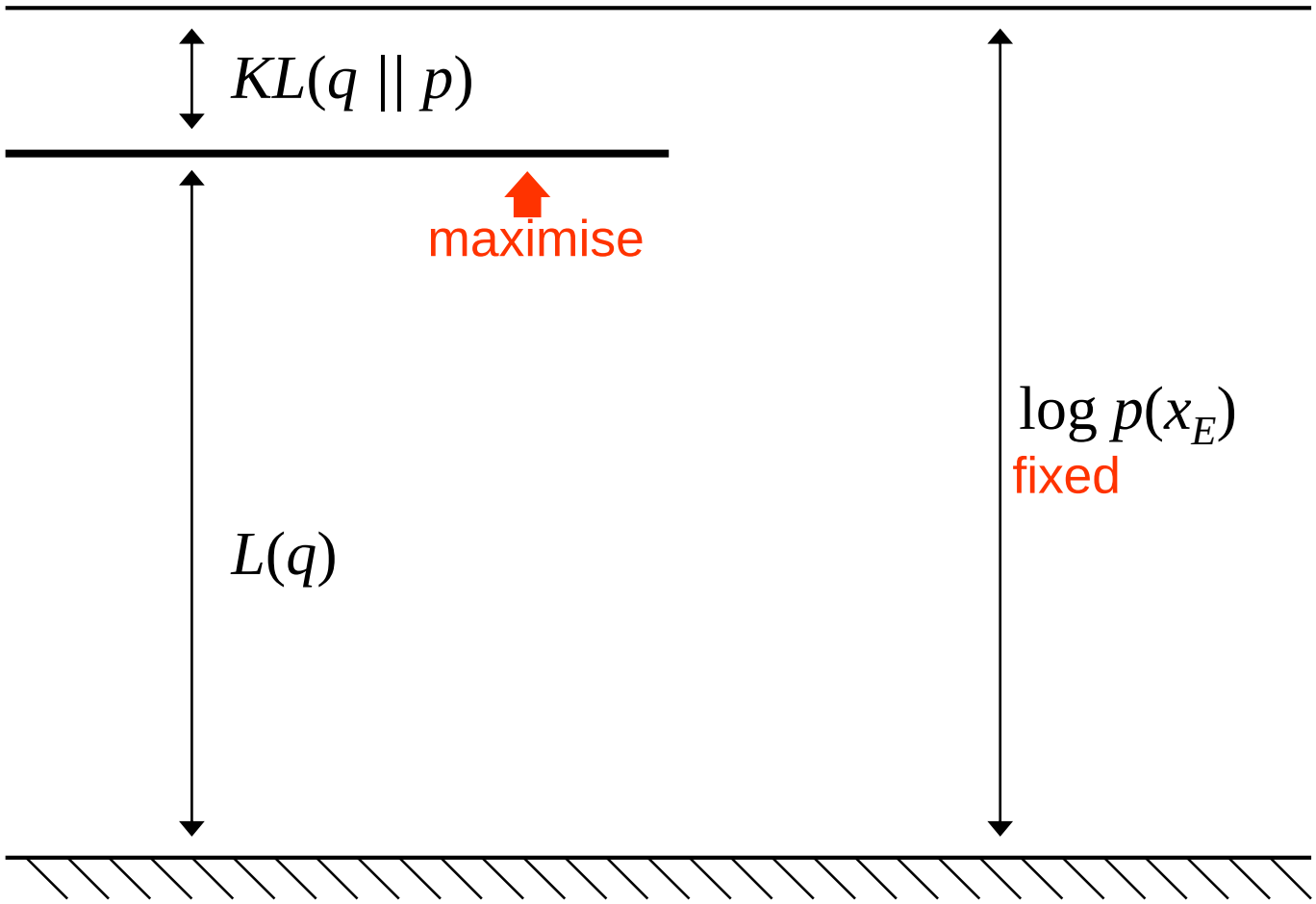
① Minimise the KL distance



① Minimise the KL distance



① Minimise the KL distance



① Minimise the KL distance

$$KL(q \| p) = \sum_{x_H} q(x_H) \log \frac{q(x_H)}{p(x_H \mid x_E)}$$

fixed　maximise　minimise

$$\log p(x_E) = L(q) + KL(q \| p)$$

$$L(q) = \sum_{x_H} q(x_H) \log \frac{p(x_H, x_E)}{q(x_H)} = H(q) + \underbrace{\sum_{x_H} q(x_H) \log p(x_H, x_E)}_{E_q\left[\log p(x_H, x_E)\right]}$$

$$\arg\max_{q} L(q) = ?$$

q: 任意

$$\arg\max_{q} L(q) = ?$$

q: 形式受限

$$\left\langle \log p(x_H, x_E) \right\rangle_q$$

选择一族形式受限的 q 分布函数

② Choose a family of variational distributions $q(x_H)$

- **变分均值场方法**假设 $q(x_H)=\prod_{i\in H} q(x_i)$
  - 假设变分分布下 x_H 的各分量统计独立
  - $q(x_H)$ 称为均值场变分分布（mean field distribution）
  - 变分边缘分布函数—variational marginal $q(x_i)$ 的函数形式无约束

② Choose a family of variational distributions $q(x_H)$

- 图像去噪
  - 给定带噪观测图像 y，恢复原始干净图像 x

$$p(x,y) \propto \exp\left\{\beta\sum_{i-j} x_i x_j + \gamma\sum_i x_i y_i\right\} \quad \beta>0, \gamma>0$$

x_i \ x_j

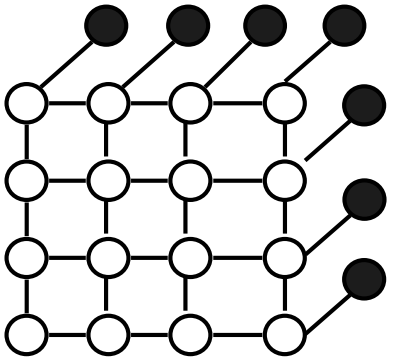
	-1	1
-1	e^{β}	$e^{-\beta}$
1	$e^{-\beta}$	e^{β}

x_i \ y_i

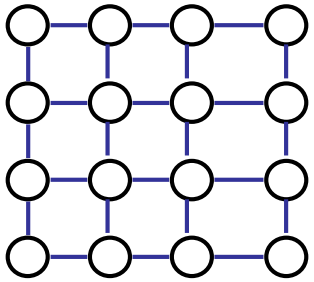
	-1	1
-1	e^{γ}	$e^{-\gamma}$
1	$e^{-\gamma}$	e^{γ}

$$p(x_i \mid y_{1:16}) = ?$$

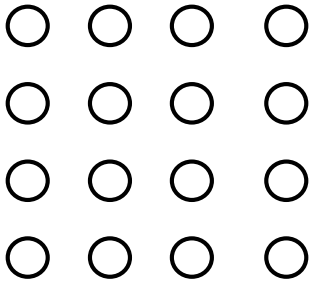
$$p(x_{1:16} \mid y_{1:16}) \approx q(x_{1:16}) = \prod_{i=1}^{16} q(x_i)$$



$p(x_{1:16}, y_{1:16})$的无向图表示



真实后验分布$p(x_{1:16}|y_{1:16})$的无向图表示



变分分布$q(x_{1:16})$的无向图表示

③ Find $q(x_H)$ which minimises KL distance.

$$\arg\max L(q) = ?$$

q: 形式受限

泛函最优化，求泛函微分

- **变分均值场方法**假设 $q(x_H) = \prod_{i \in H} q(x_i)$
  - $q(x_i)$ 无约束，可分别独立变动/调整，逐个优化

针对$q(x_k)$的最优化：将 $L(q)$ 视为 $q(x_k)$ 的函数，与 $q(x_k)$ 无关项并入常数

$$L(q) = H(q) + \sum_{x_H} q(x_H) \log p(x_H, x_E)$$

$$= H(q(x_k)) + \sum_{x_k} \sum_{x_{H \setminus k}} q(x_k) \prod_{i \neq k} q(x_i) \log p(x_H, x_E) + \text{常数}$$

定义一个新的分布 $\log \tilde{p}(x_k) = \sum_{x_{H \setminus k}} \prod_{i \neq k} q(x_i) \log p(x_H, x_E) + \text{常数}$

$$\max_{q(x_k)} L(q) = H(q(x_k)) + \sum_{x_k} q(x_k) \log \tilde{p}(x_k) + \text{常数}$$

$$= \min_{q(x_k)} KL(q(x_k) \| \tilde{p}(x_k)) + \text{常数}$$

③ Find $q(x_H)$ which minimises KL distance.

- 变分均值场方法 $q(x_H)=\prod_{i\in H} q(x_i)$

单个边缘分布的最优解 $\log q(x_k)=\log \tilde{p}(x_k)$

$$=\sum_{x_{H\setminus k}}\prod_{i\neq k} q(x_i)\log p(x_H,x_E)+\text{常数}$$

$$=\sum_{x_{H\setminus k}} q(x_{H\setminus k}\mid x_k)\times\log p(x_H,x_E)=E_q\left[\log p(x_H,x_E)\middle| x_k\right]+\text{常数}$$

均值场更新公式： $\log q(x_k)=E_q\left[\log p(x_H,x_E)\mid x_k\right]+const,\ k\in H$

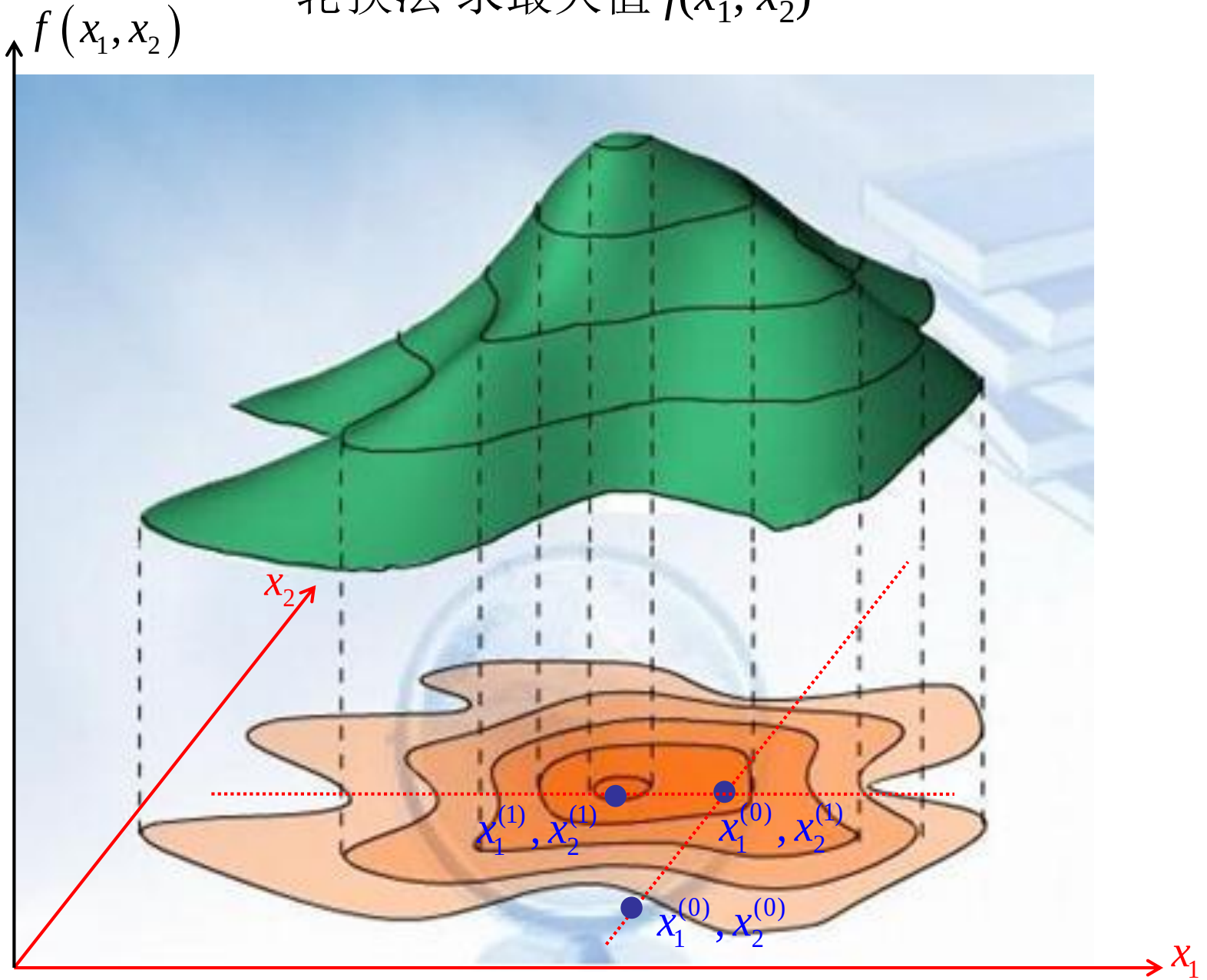
$$q(x_k)\propto \exp\left\{E_q\left[\log p(x_H,x_E)\mid x_k\right]\right\}$$

轮换法求解： $q^{(0)}(x_1),q^{(0)}(x_2),q^{(0)}(x_3),\cdots,q^{(0)}(x_K)$

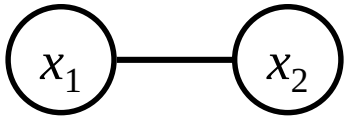
$$q^{(1)}(x_1),q^{(0)}(x_2),q^{(0)}(x_3),\cdots,q^{(0)}(x_K)$$

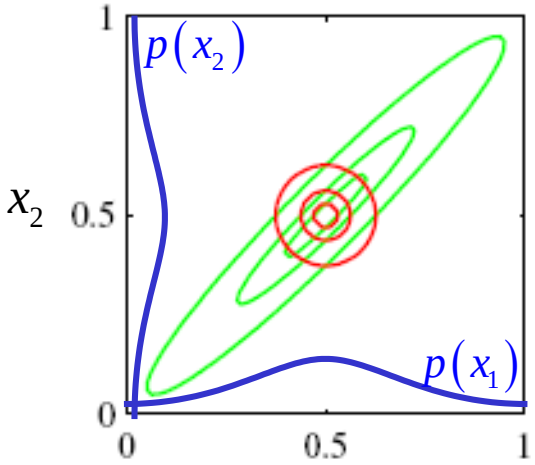
$$q^{(1)}(x_1),q^{(1)}(x_2),q^{(0)}(x_3),\cdots,q^{(0)}(x_K)$$

轮换法 求最大值 $f(x_1, x_2)$



A simple example





$$p(x_1,x_2)=N\left(\begin{pmatrix}x_1\\x_2\end{pmatrix}\middle|\begin{pmatrix}\mu_1\\\mu_2\end{pmatrix},\begin{pmatrix}K_{11}&K_{21}\\K_{21}&K_{22}\end{pmatrix}^{-1}\right)$$

$q(x_1)$ $q(x_2)$

$$\log p(x_1,x_2)=-\frac{D}{2}\log 2\pi+\frac{1}{2}\log|K|-\frac{1}{2}\begin{pmatrix}x_1-\mu_1\\x_2-\mu_2\end{pmatrix}^T\begin{pmatrix}K_{11}&K_{21}\\K_{21}&K_{22}\end{pmatrix}\begin{pmatrix}x_1-\mu_1\\x_2-\mu_2\end{pmatrix}$$

套公式 $\log q(x_k)=E_q\left[\log p(x_H,x_E)\,|\,x_k\right]+const$ 最小化$KL\left(p(x_1,x_2),q(x_1)q(x_2)\right)$

$$\log q(x_1)=\sum_{x_2}q(x_2)\log p(x_1,x_2)+const$$

$$=-\frac{1}{2}(x_1-\mu_1)^2K_{11}-(x_1-\mu_1)K_{12}\left(\langle x_2\rangle_q-\mu_2\right)+const$$

x_1的二次项： $-\frac{1}{2}K_{11}\cdot x_1^2$

x_1的一次项： $x_1\cdot\mu_1K_{11}-x_1\cdot K_{12}\left(\langle x_2\rangle_q-\mu_2\right)$

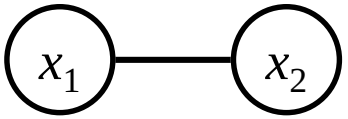
$$N(x\,|\,g,h,K)=\exp\left\{g+h\cdot x-\frac{1}{2}K\cdot x^2\right\}$$

$\mu=K^{-1}h$

$\Sigma=K^{-1}$

$$q(x_1)=N\left(x_1\,|\,\mu_1-K_{11}^{-1}K_{12}\left(\langle x_2\rangle_q-\mu_2\right),K_{11}^{-1}\right)$$

A simple example



$$p(x_1,x_2)=N\left(\begin{pmatrix}x_1\\x_2\end{pmatrix}\middle|\begin{pmatrix}\mu_1\\\mu_2\end{pmatrix},\begin{pmatrix}K_{11}&K_{21}\\K_{21}&K_{22}\end{pmatrix}^{-1}\right)$$

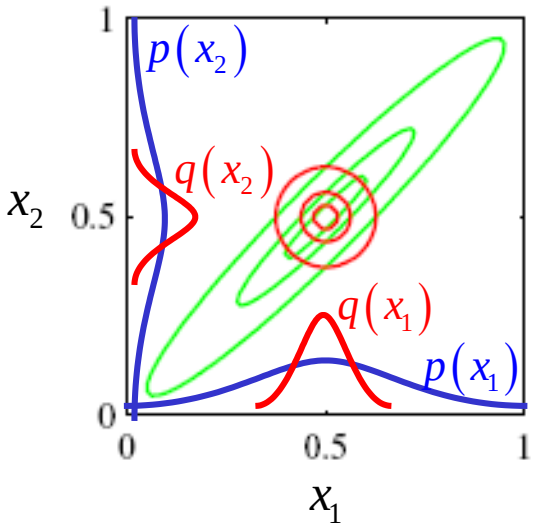
$q(x_1)$ $q(x_2)$

$$q^{(0)}(x_1),q^{(0)}(x_2)$$
$$q^{(1)}(x_1),q^{(0)}(x_2)$$
$$q^{(1)}(x_1),q^{(1)}(x_2)$$

最小化$KL\left(p(x_1,x_2),q(x_1)q(x_2)\right)$，最优解是：

$$\log q(x_1)=\sum_{x_2}q(x_2)\log p(x_1,x_2)+const$$
$$\log q(x_2)=\sum_{x_1}q(x_1)\log p(x_1,x_2)+const$$

$$\begin{cases}q(x_1)=N\left(x_1\mid\mu_1-K_{11}^{-1}K_{12}\left(\langle x_2\rangle_q-\mu_2\right),K_{11}^{-1}\right)\\q(x_2)=N\left(x_2\mid\mu_2-K_{22}^{-1}K_{21}\left(\langle x_1\rangle_q-\mu_1\right),K_{22}^{-1}\right)\end{cases}$$



$$q(x_1,x_2)=q(x_1)q(x_2)$$

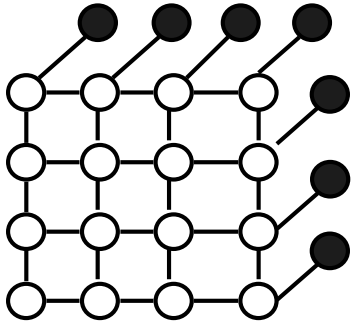
位于正确的均值位置，但是方差低估了

一般来说，均值场方法给出偏紧凑的近似分布

$$x_1\mid x_2\sim N\left(\mu_{1|2},\Sigma_{1|2}\right)$$
$$=N\left(\mu_1+\Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2),\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)$$

图像去噪

- 给定带噪观测图像 y，恢复原始干净图像 x

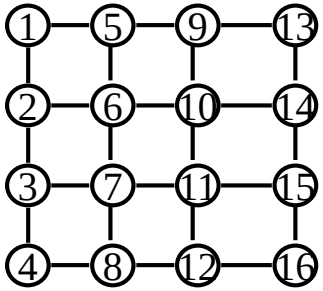


$$p\left(x_{1:16}, y_{1:16}\right) \propto \exp\left\{\beta \sum_{i-j} x_i x_j + \gamma \sum_i x_i y_i\right\} \quad \beta > 0, \gamma > 0$$

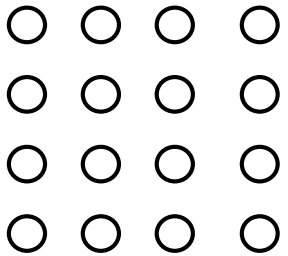
x_i \ x_j	-1	1
-1	e^{β}	$e^{-\beta}$
1	$e^{-\beta}$	e^{β}

x_i \ y_i	-1	1
-1	e^{γ}	$e^{-\gamma}$
1	$e^{-\gamma}$	e^{γ}

$$p\left(x_{1:16} \mid y_{1:16}\right) = ?$$



$$q\left(x_{1:16}\right) = \prod_{i=1}^{16} q\left(x_i\right)$$



套公式 $\log q\left(x_k\right) = E_q\left[\log p\left(x_H, x_E\right) \mid x_k\right] + const$

$$\log q\left(x_i\right) = E_q\left[\beta \sum_{i-j} x_i x_j + \gamma \sum_i x_i y_i \,\middle|\, x_i\right] + const$$

$$\log q\left(x_1\right) = ?$$

$$\log q\left(x_2\right) = ?$$

均值场变分推理 vs Gibbs采样

均值场变分分布计算公式：

$$\log q(x_k) = E_q\left[\log p(x_H, x_E) \mid x_k\right] + const$$

轮换求解：

$$q(x_1), q(x_2), q(x_3), \cdots, q(x_K)$$

$$\hat{q}(x_1), q(x_2), q(x_3), \cdots, q(x_K)$$

$$\hat{q}(x_1), \hat{q}(x_2), q(x_3), \cdots, q(x_K)$$

Gibbs采样公式：

$$x_k - sampling\ from\ p\left(x_k \mid x_{H\setminus\{k\}}, x_E\right)$$

$$x_k - sampling\ from\ p(x_H, x_E)$$

轮换采样：

$$x_1,\ x_2,\ x_3,\ \cdots,\ x_K$$

$$\hat{x}_1,\ x_2,\ x_3,\ \cdots,\ x_K$$

$$x_1,\ \hat{x}_2,\ x_3,\ \cdots,\ x_K$$

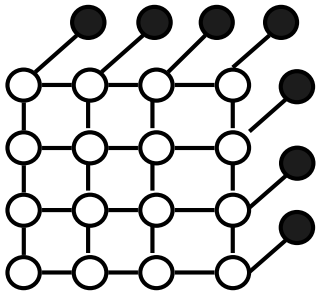
Mean field equations

均值场方程（k=1, …, K）的计算与两种图结构有关

$$\log q_k(x_k) = E_q\left[\log p(x_H, x_E) \mid x_k\right] + const = E_q\left[\sum_C \log \phi_C(x_C) \middle| x_k\right] + const$$

$$= \sum_C E_q\left[\log \phi_C(x_C) \middle| x_k\right] + const$$

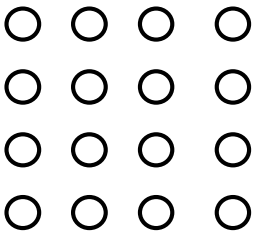
原概率分布 $p(x_H, x_E)$ 的结构



$$= \sum_C \sum_{x_{C\cap(H\setminus k)}} q\left(x_{C\cap(H\setminus k)} \mid x_k\right) \log \phi_C(x_C) + const$$

变分分布 $q(x_H)$ 的结构

C∩(*H**k*): 簇 *C* 中除 *k* 外的所有隐变量



变分近似推理

- 变分方法是一种经典的泛函最优化方法
- 变分近似推理：变分优化方法用于推理问题
- Block approach
  - 变分均值场方法（Variational mean field）
  - **结构变分方法（Structured variational approach）**
  - 变分贝叶斯方法（Variational Bayesian）用于贝叶斯参数估计
- Sequential approach
  - Local variational method

Structured variational approach

找出一些子结构（substructure），
子结构内部方便做精确推理，子结构之间做均值场近似

- 假设 $q(x_H)=\prod_i q(x_{h_i})$ $\bigcup_i h_i = H$, $h_i \cap h_j = 0$ for $i \neq j$
  - 变分边缘分布函数—variational marginal $q(x_{hi})$ 的函数形式无约束
  - 可分别独立变动/调整，逐个优化

$$\log q(x_{h_i}) = E_q\left[\log p(x_H, x_E) \mid x_{h_i}\right] + const$$

$$= \sum_C E_q\left[\log \phi_C(x_C) \mid x_{h_i}\right] + const$$

原概率分布 $p(x_H, x_E)$ 的结构

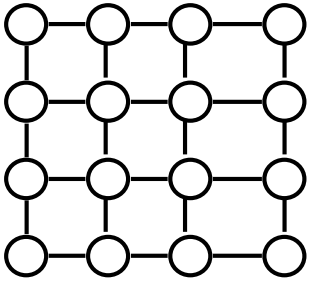
$$= \sum_C \sum_{x_{C\cap(H\setminus h_i)}} q\left(x_{C\cap(H\setminus h_i)} \mid x_{h_i}\right) \log \phi_C(x_C) + const$$

$C\cap(H\setminus h_i)$: 簇 C 中除 h_i 外的所有隐变量

变分分布 $q(x_H)$ 的结构

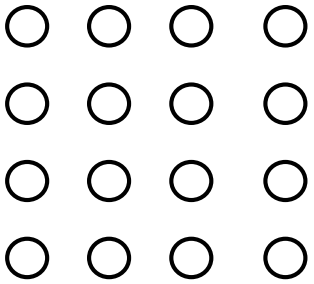
- 图像去噪
  - 给定带噪观测图像 y，恢复原始干净图像 x

$$p(x,y) \propto \exp\left\{\beta\sum_{i-j} x_i x_j + \gamma\sum_i x_i y_i\right\} \quad \beta>0, \gamma>0$$



真实后验分布$p(x_{1:16}|y_{1:16})$的无向图表示

- 均值场变分近似分布 $q(x_H)=\prod_{i\in H} q(x_i)$



均值场变分分布$q(x_{1:16})$的无向图表示

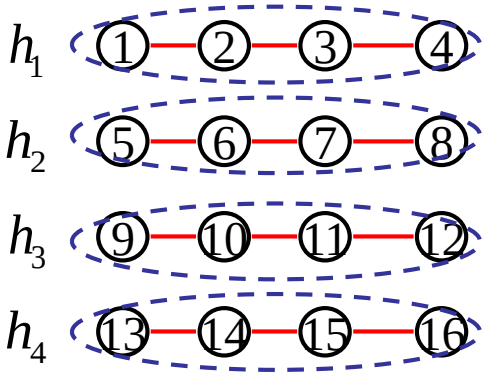
- 结构变分近似分布 $q(x_H)=\prod_{i=1}^{4} q(x_{h_i})$

套公式 $\log q(x_{h_i}) = E_q\left[\log p(x_H, x_E)\,|\,x_{h_i}\right] + const$

$$\log q(x_{h_i}) = E_q\left\{\beta\sum_{i-j} x_i x_j + \gamma\sum_i x_i y_i \,\middle|\, x_{h_i}\right\} + const$$

$\log q(x_{h_1}) = ?$

$\log q(x_{h_2}) = ?$



结构变分分布的无向图表示

变分近似推理

- 变分方法是一种经典的泛函最优化方法
- 变分近似推理：变分优化方法用于推理问题
- Block approach
  - 变分均值场方法（Variational mean field）
  - 结构变分方法（Structured variational approach）
  - 变分贝叶斯方法（Variational Bayesian）用于贝叶斯参数估计
- Sequential approach
  - Local variational method

M.J. Wainwright, M.I. Jordan.
"Graphical models, exponential families, and variational inference",
Foundations and Trends in Machine Learning, vol.1, pp.1–305, 2008.

Connect to Deep Learning

Auto-Encoding Variational Bayes

Diederik P. Kingma
Machine Learning Group
Universiteit van Amsterdam
dpkingma@gmail.com

Max Welling
Machine Learning Group
Universiteit van Amsterdam
welling.max@gmail.com

A Review of Learning with Deep Generative Models from Perspective of Graphical Modeling

Zhijian Ou

(Submitted on 5 Aug 2018 (v1), last revised 27 Mar 2019 (this version, v4))

This document aims to provide a review on learning with deep generative models (DGMs), which is an highly-active area in machine learning and more generally, artificial intelligence. This review is not meant to be a tutorial, but when necessary, we provide self-contained derivations for completeness. This review has two features. First, though there are different perspectives to classify DGMs, we choose to organize this review from the perspective of graphical modeling, because the learning methods for directed DGMs and undirected DGMs are fundamentally different. Second, we differentiate model definitions from model learning algorithms, since different learning algorithms can be applied to solve the learning problem on the same model, and an algorithm can be applied to learn different models. We thus separate model definition and model learning, with more emphasis on reviewing, differentiating and connecting different learning algorithms. We also discuss promising future research directions.

课程章节

- 第一章 图模型的表示理论（2）
  - **Semantics (DGM, UGM)**
  - **HMM, CRF**
- 第二章 图模型的推理理论（4）
  - 精确推理：**variable-elimination，cluster-tree，triangulate**
  - 连续变量：**Kalman**
  - 采样近似：**sampling**
  - 变分近似：**variational**
- 第三章 图模型的学习理论（2）
  - 参数学习：**maxlikelihoodEstimate，RFLearning, BayesEstimate**
  - 结构学习：**StructureLearning**

	pgm-1 semantics √		pgm-2 hmm-crf √ pgm-3 exact √	pgm-4 kalman √ pgm-5 sampling √
pgm-6 variational √ pgm-7 ML	pgm-8 Bayesian			