Mean field approximation methods and information geometry

Shiro Ikeda

ISM, Tokyo, Japan

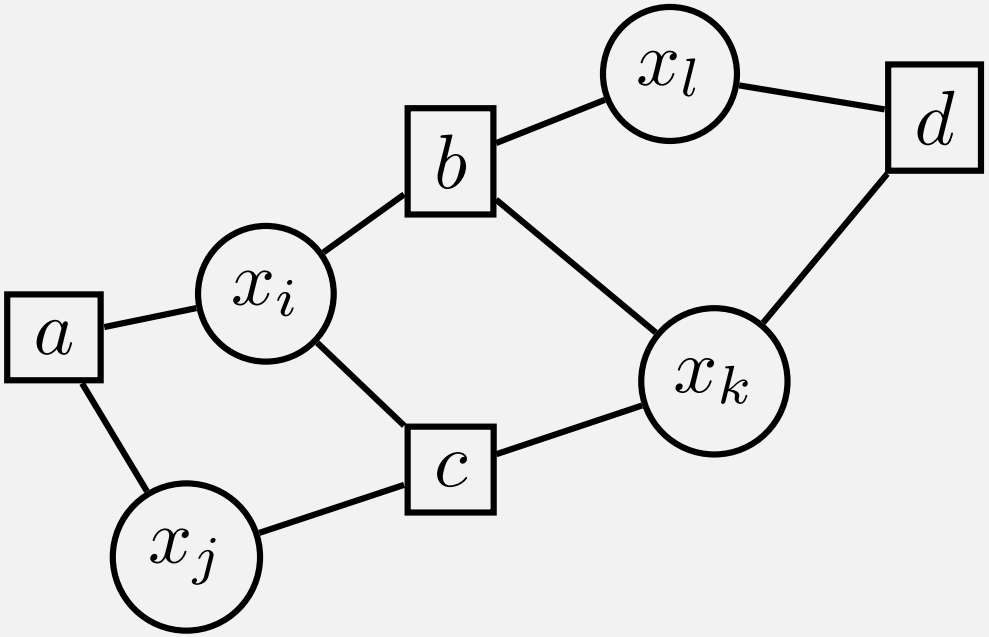
31 August 2009

Outline

1. Belief Propagation
2. Information Geometrical View
3. Survey Propagation (SP)
4. Conclusion

Graphical Model

Example



Stochastic Variable

$$\boldsymbol{x} = (x_i, x_j, x_k, x_l)^T$$

Clique

$$r \in \mathcal{L} = \{a, b, c, d\}$$

Graphical model

Joint distribution

$$q(\boldsymbol{x}) = \frac{1}{Z}\prod_{r\in\mathcal{L}}\psi_r(\boldsymbol{x}_r)$$
$$= \exp\Big(\sum_{r\in\mathcal{L}} c_r(\boldsymbol{x}_{\boldsymbol{r}}) - \varphi_q\Big).$$

$$\psi_r(\boldsymbol{x}_r) > 0, \quad Z = \sum_{\boldsymbol{x}}\Big(\prod_{r\in\mathcal{L}}\psi_r(\boldsymbol{x}_r)\Big)$$

$$c_r(\boldsymbol{x}_r) = \log\psi_r(\boldsymbol{x}_r), \quad \varphi_q = \log Z$$

$\boldsymbol{x}_r = \{x_i | i \in V(r)\}$, $V(r)$: member of clique r.

Belief Propagation

Message update

For each (r, i), update messages $\nu_{ri}(x_i)$, $\mu_{ri}(x_i)$.

1. initialize $t = 1$, $\nu_{ri}(x_i)^t = 1/2$ $\mu_{ri}(x_i)^t = 1/2$
2. update messages as follows

$$\nu_{ri}^{t+1}(x_i) \propto \sum_{\boldsymbol{x}_r \setminus x_i} \psi_r(\boldsymbol{x}_r) \prod_{j \in V(r) \setminus i} \mu_{rj}^t(x_j)$$

$$\mu_{ri}^{t+1}(x_i) \propto \prod_{s \in \mathcal{N}(i) \setminus r} \nu_{si}^{t+1}(x_i),$$

3. belief (marginal distribution) is

$$b_i^{t+1}(x_i) \propto \prod_{r \in \mathcal{N}(i)} \nu_{ri}^{t+1}(x_i)$$

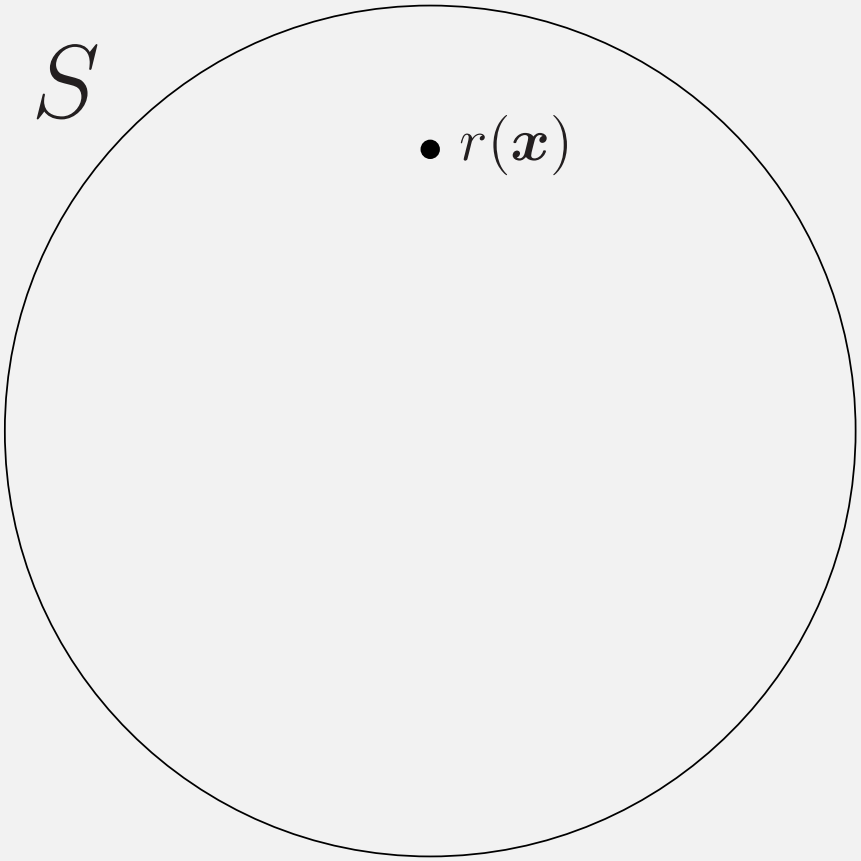
Information Geometrical View

Our results

Information geometry: Applied differential geometry to statistical/stochastic models.

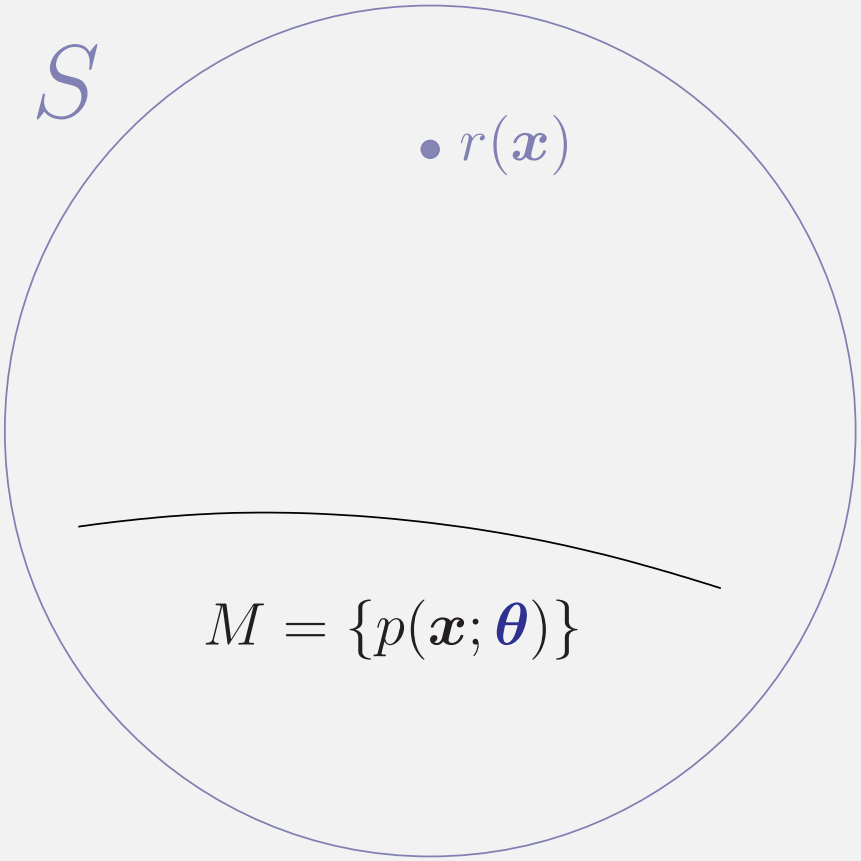
- Amari, (1985). *Springer-Verlag*.
- Murray & Rice, (1993). *Chipman & Hall*.
- Amari & Nagaoka, (2000). *AMS and Oxford University Press*.

Information Geometry



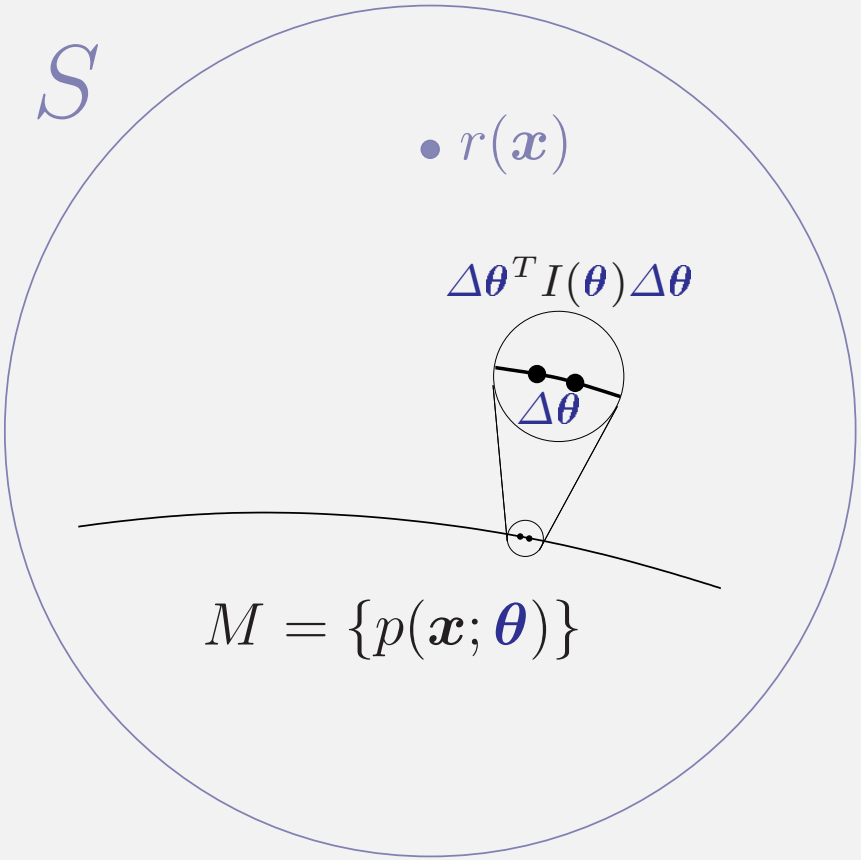
S: $\boldsymbol{x}$ Space of probability distributions.
Each point $r(\boldsymbol{x}) \in S$ is a probability distribution

Information Geometry



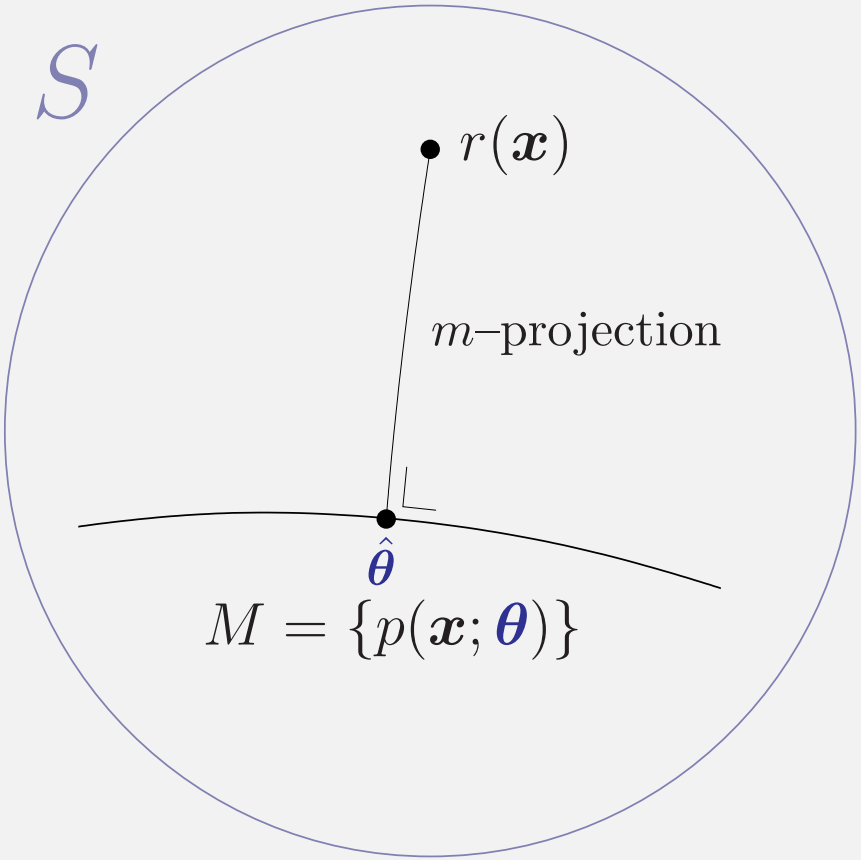
$M = \{p(\boldsymbol{x};\boldsymbol{\theta})\}$: distributions parametrized by $\boldsymbol{\theta}$.

Information Geometry



$M = \{p(\boldsymbol{x};\boldsymbol{\theta})\}$: Model manifold

Information Geometry

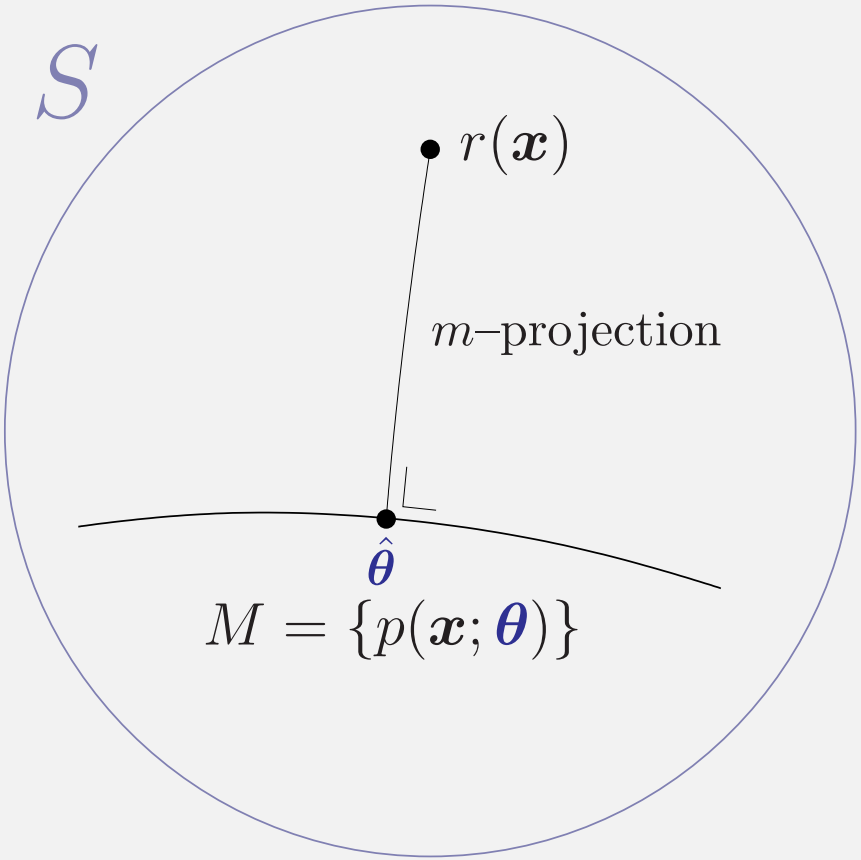


$$\hat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} KL(r(\boldsymbol{x}); p(\boldsymbol{x};\boldsymbol{\theta}))$$

Maximum Likelihood: m-projection to the model

If the model is an exponential family, projection is unique.

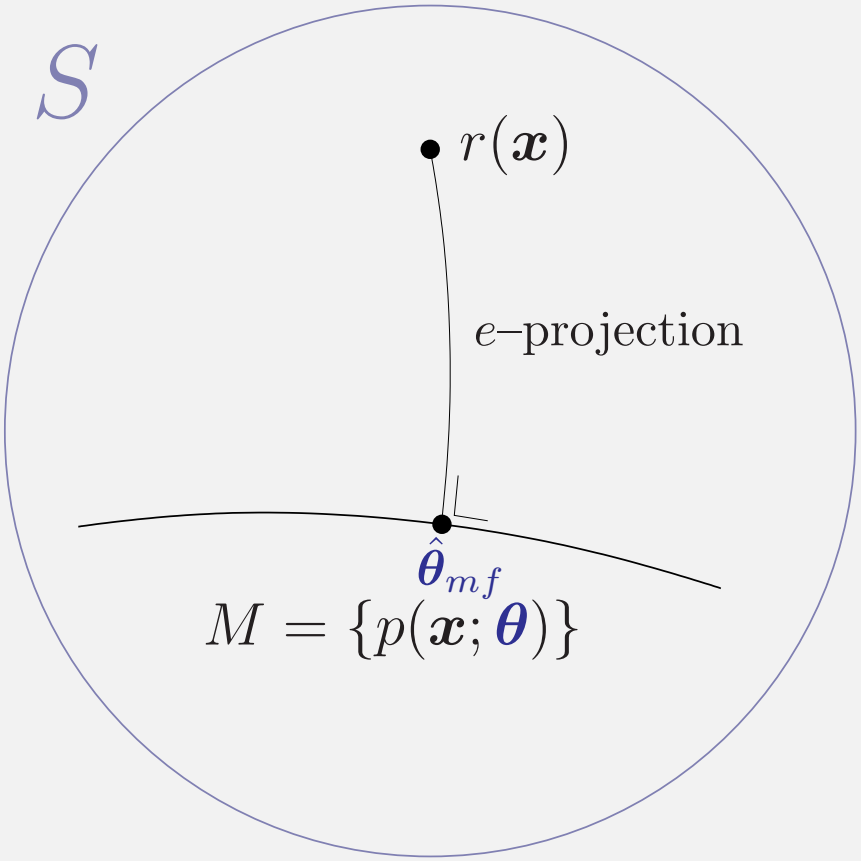
Information Geometry



If $p(\boldsymbol{x};\boldsymbol{\theta})$ is an exponential family, and its sufficient statistics is $\boldsymbol{t}(\boldsymbol{x})$, m-projection gives

$$E_{r(\boldsymbol{x})}[\boldsymbol{t}(\boldsymbol{x})] = E_{p(\boldsymbol{x};\boldsymbol{\theta})}[\boldsymbol{t}(\boldsymbol{x})]$$

Information Geometry

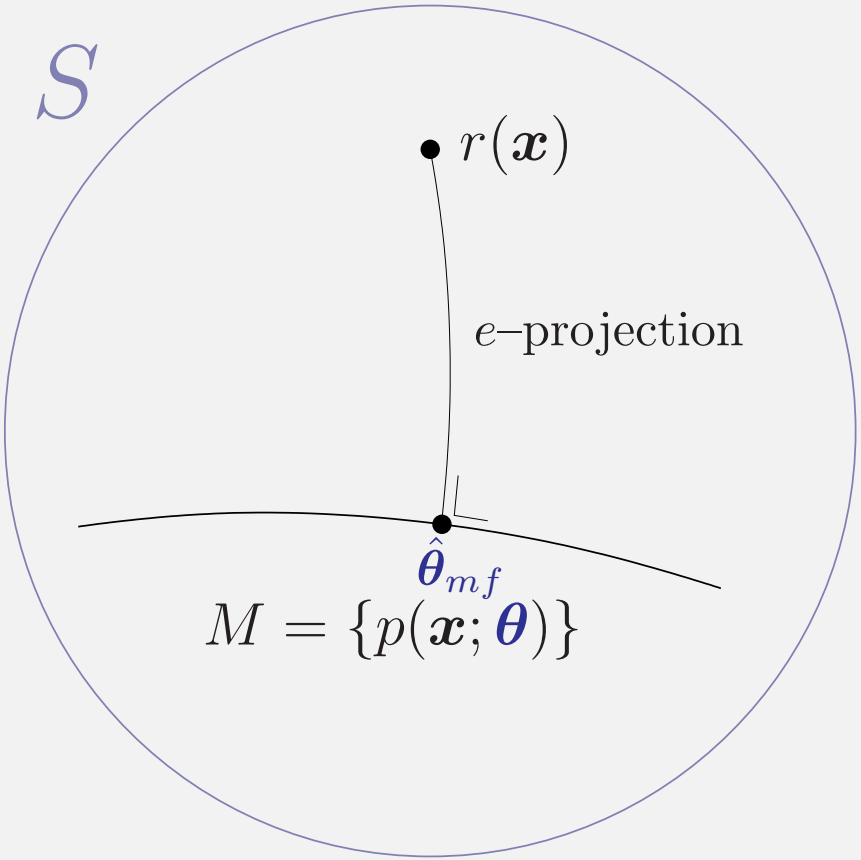


$$\hat{\boldsymbol{\theta}}_{mf} = \operatorname*{argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} KL(p(\boldsymbol{x};\boldsymbol{\theta}); r(\boldsymbol{x}))$$

Naive Mean Field approximation: e-projection

For an exponential family, there are a lot of local minima.

Information Geometry



In Statistical Physics, multiple local minima are important. They consider the multiple solutions corresponds to landscape of energy function and "phase transitions."

Information Geometrical View

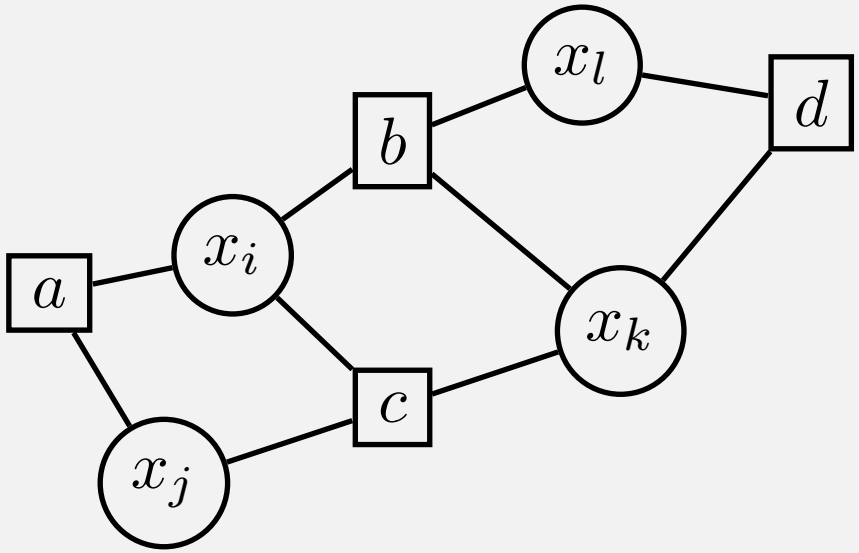
Our results

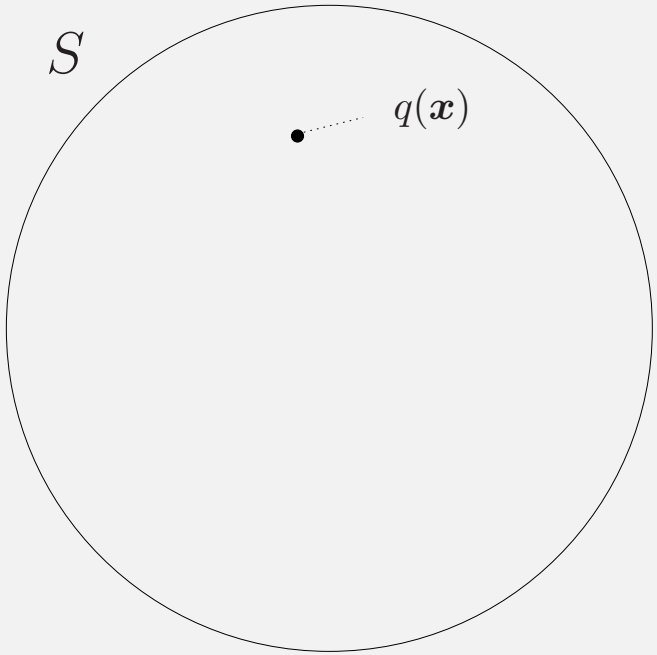
Discuss the accuracy, convergence of LBP with information geometry

- Ikeda, Tanaka, & Amari, (2004). *IEEE tr. on IT*, *50(6)*, 1097-1114.
- Ikeda, Tanaka, & Amari, (2004). *Neural Comput.*, *50(6)*, 1779-1810.

Belief Propagation

Joint distribution

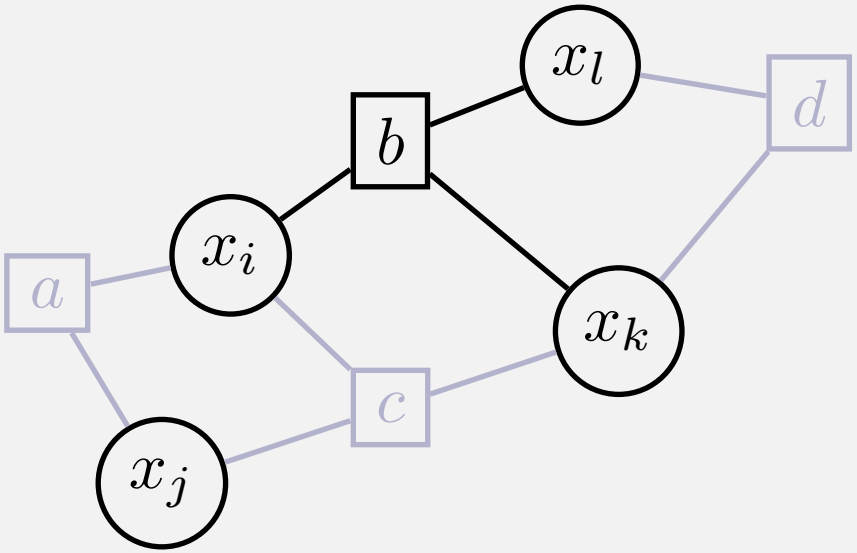


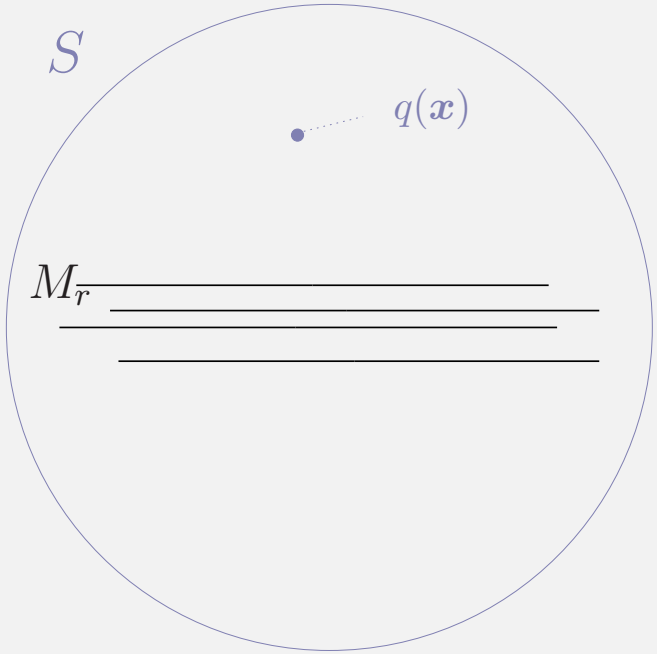


$$q(\boldsymbol{x}) = \exp\Big(\sum_r c_r(\boldsymbol{x}) - \psi_q\Big)$$

Belief Propagation

Single link models

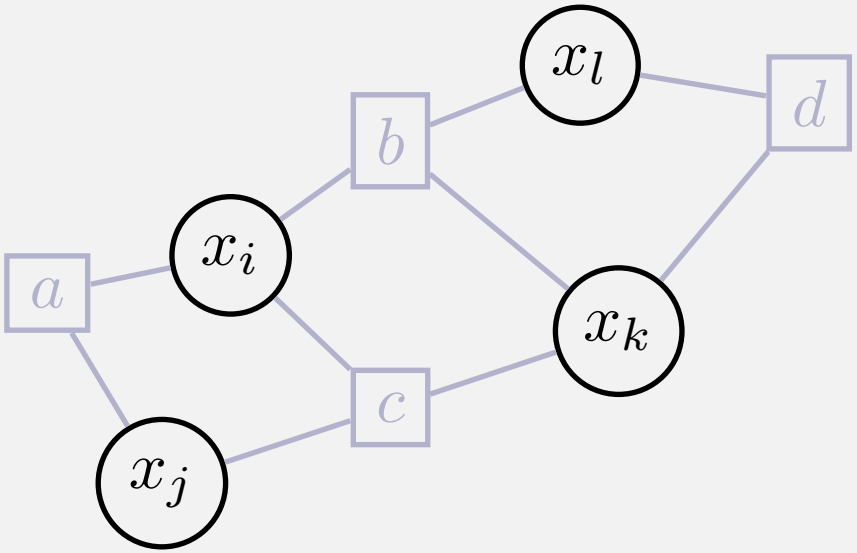


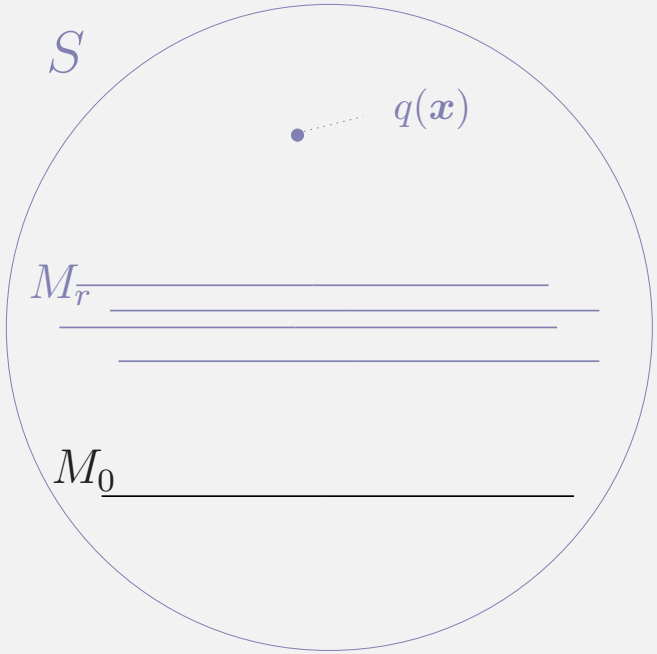


$$M_r = \left\{ p_r(\boldsymbol{x}; \boldsymbol{\zeta}_r) = \exp(c_r(\boldsymbol{x}) + \boldsymbol{\zeta}_r \cdot \boldsymbol{x} - \varphi_q(\boldsymbol{\zeta}_r)) \right\}, r = 1, \cdots, L.$$

Belief Propagation

Marginals

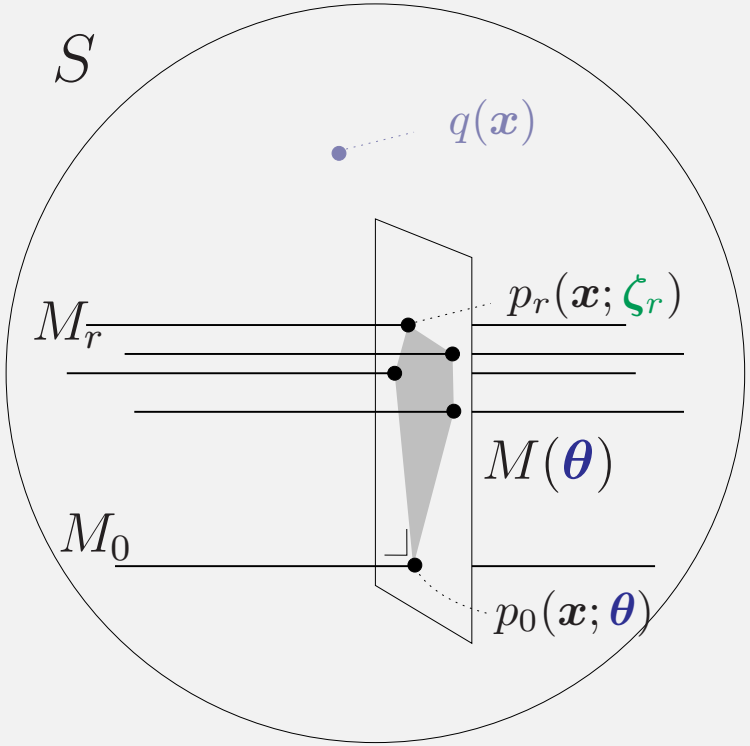




$$M_0 = \left\{ p_0(\boldsymbol{x}; \boldsymbol{\theta}) = \exp(\boldsymbol{\theta} \cdot \boldsymbol{x} - \varphi_0(\boldsymbol{\theta})) \right\}$$

Belief Propagation

Convergence



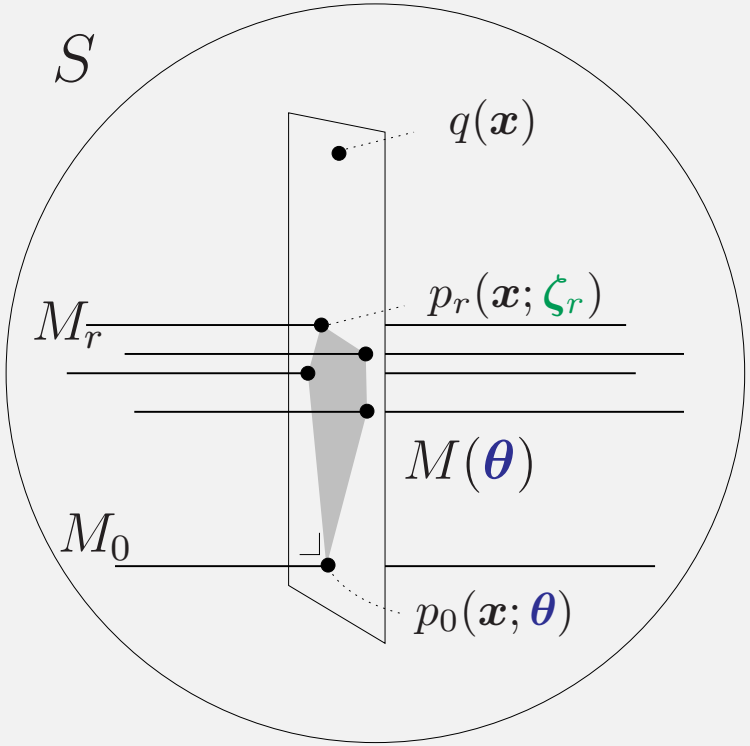
$$M(\boldsymbol{\theta}) = \{\text{Product of marginals} p_0(\boldsymbol{x};\boldsymbol{\theta})\}$$

Condition 1

$p_0(\boldsymbol{x};\boldsymbol{\theta}), p_r(\boldsymbol{x};\boldsymbol{\zeta}_r) \in M(\boldsymbol{\theta}),$

Belief Propagation

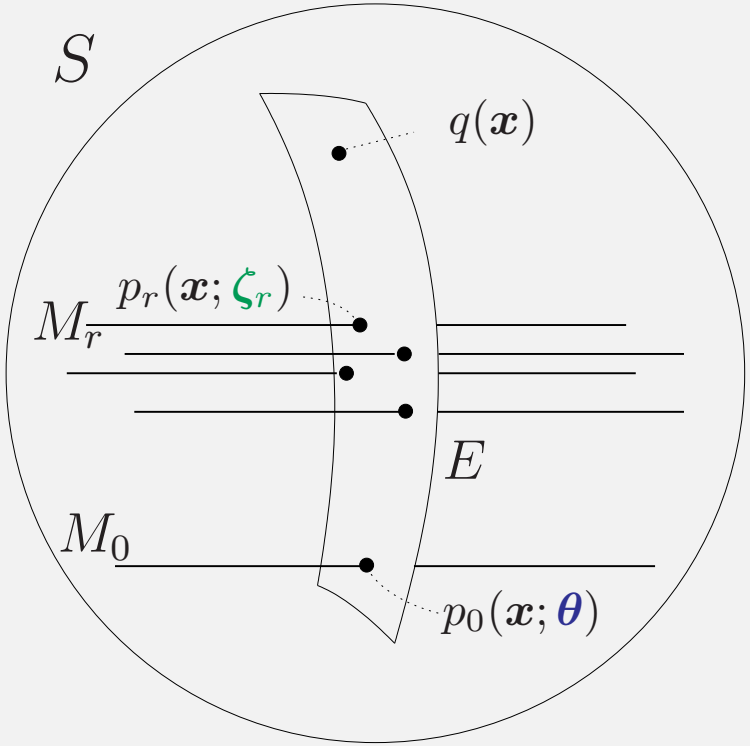
At convergent points



If $M(\boldsymbol{\theta})$ includes $q(\boldsymbol{x})$ $p_0(\boldsymbol{x};\boldsymbol{\theta})$ is the true marginals.

Belief Propagation

Convergence



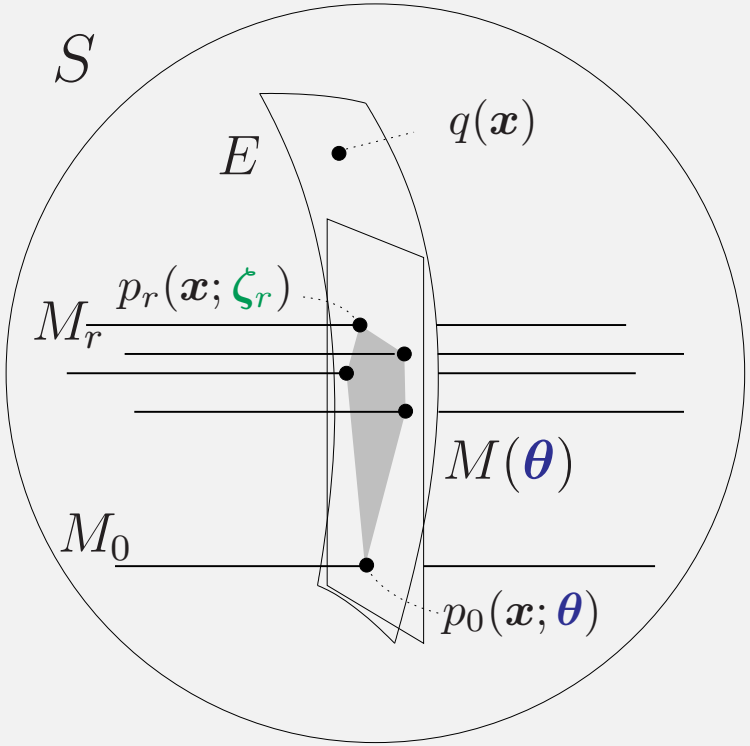
$$E = \{\text{log-linear mixture} of\, p_0, p_r\}$$
$$= \left\{ \frac{1}{Z_E(\boldsymbol{t})} p_0(\boldsymbol{x};\boldsymbol{\theta})^{t_0} \prod_r p_r(\boldsymbol{x};\boldsymbol{\zeta}_r)^{t_r} \;\middle|\; \sum t_r = 1 \right\}$$

Condition 2

$q(\boldsymbol{x}), p_0(\boldsymbol{x};\boldsymbol{\theta}), p_r(\boldsymbol{x};\boldsymbol{\zeta}_r) \in E$

Belief Propagation

Convergence



Theorem

When $p_0(\boldsymbol{x};\boldsymbol{\theta})$, *and* $p_r(\boldsymbol{x};\zeta_r)$ $r = 1, \cdots, L$ *satisfies*

Condition 1 *and* **Condition 2**

$$\longleftrightarrow$$

It is the convergent point of BP.

Approximate accuracy

Perturbation analysis

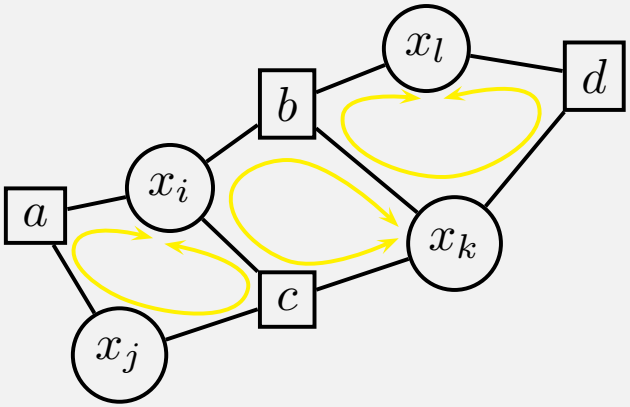
Difference between E and $M(\theta) \to$ Accuracy

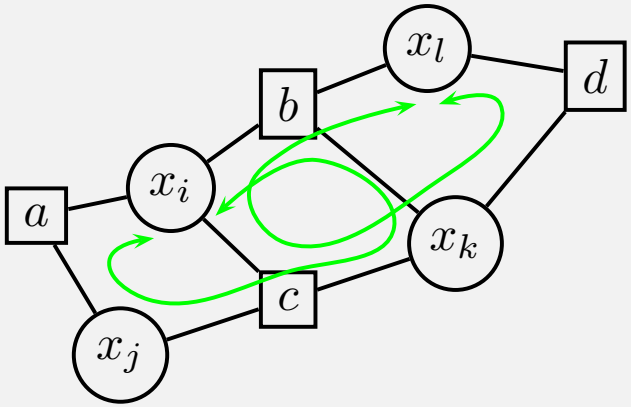
$\overline{\boldsymbol{x}}_q$: expectation of $\boldsymbol{x}$ w.r.t. $q(\boldsymbol{x})$. $\overline{\boldsymbol{x}}_{BP}$: convergent point of BP.

$$\overline{\boldsymbol{x}}_q \simeq \overline{\boldsymbol{x}}_{BP} + \frac{1}{2}\sum_{r\neq s} B_{rs}\overline{\boldsymbol{x}}_{BP} + \frac{1}{6}\Big(\sum_{rst} B_{rst} - \sum_{r} B_{rrr}\Big)\overline{\boldsymbol{x}}_{BP}.$$

$B_{rs}\overline{\boldsymbol{x}}_{BP}$: order 4 loop, embedded m–curvature of E

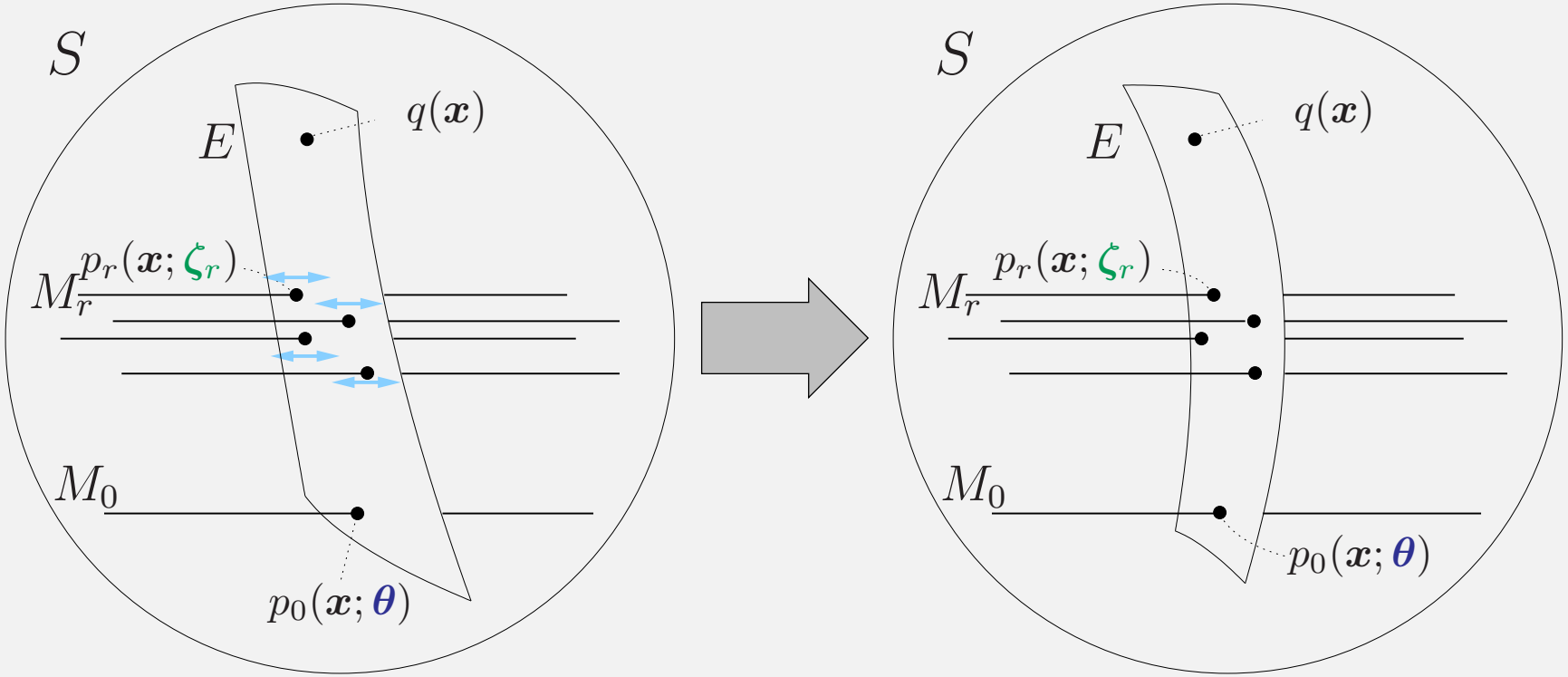
$B_{rst}\overline{\boldsymbol{x}}_{BP}$: order 6 loop, torsion of E.





Convergence

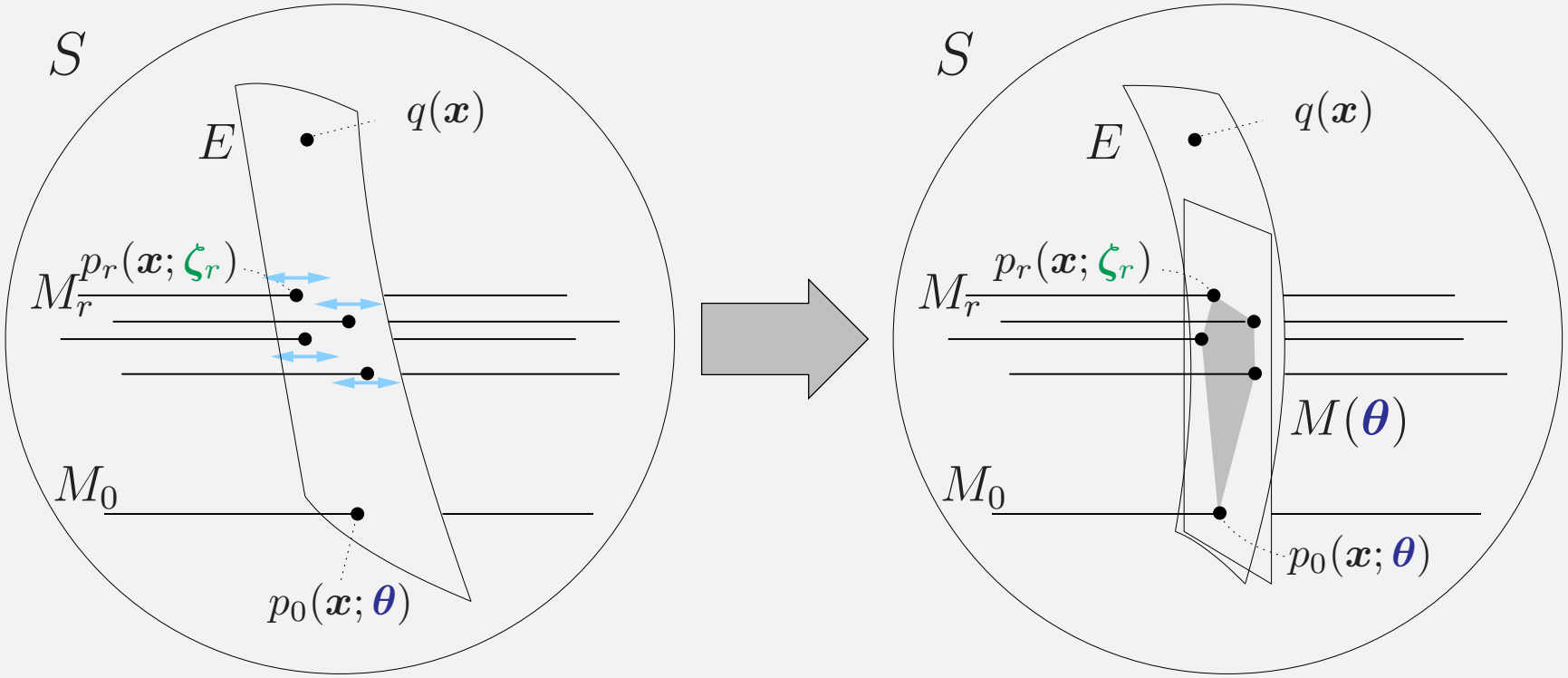
e–constraint algorithm



Condition 2 is always satisfied, update parameters to satisfy Condition 1

Convergence

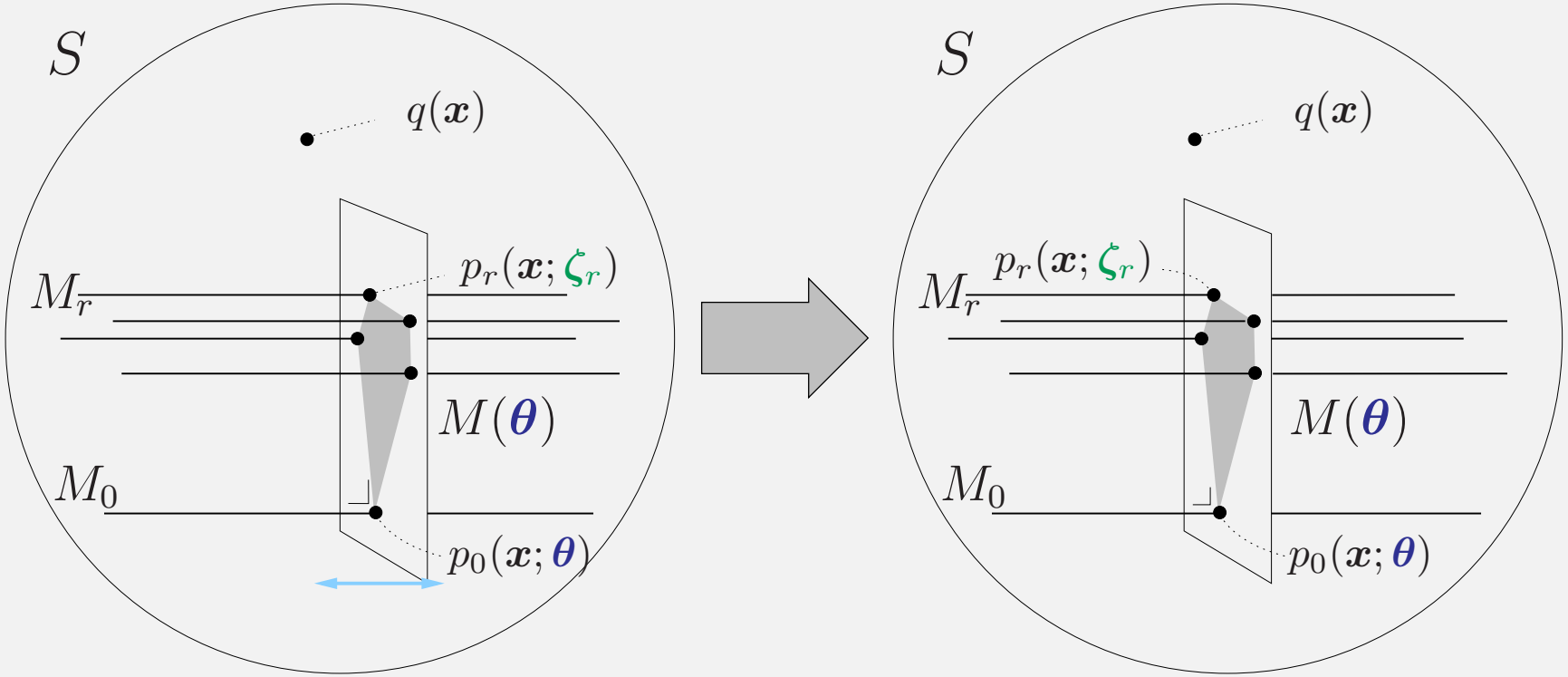
e–constraint algorithm



BP, TRP (Wainwright, et al. NIPS*14)

Convergence

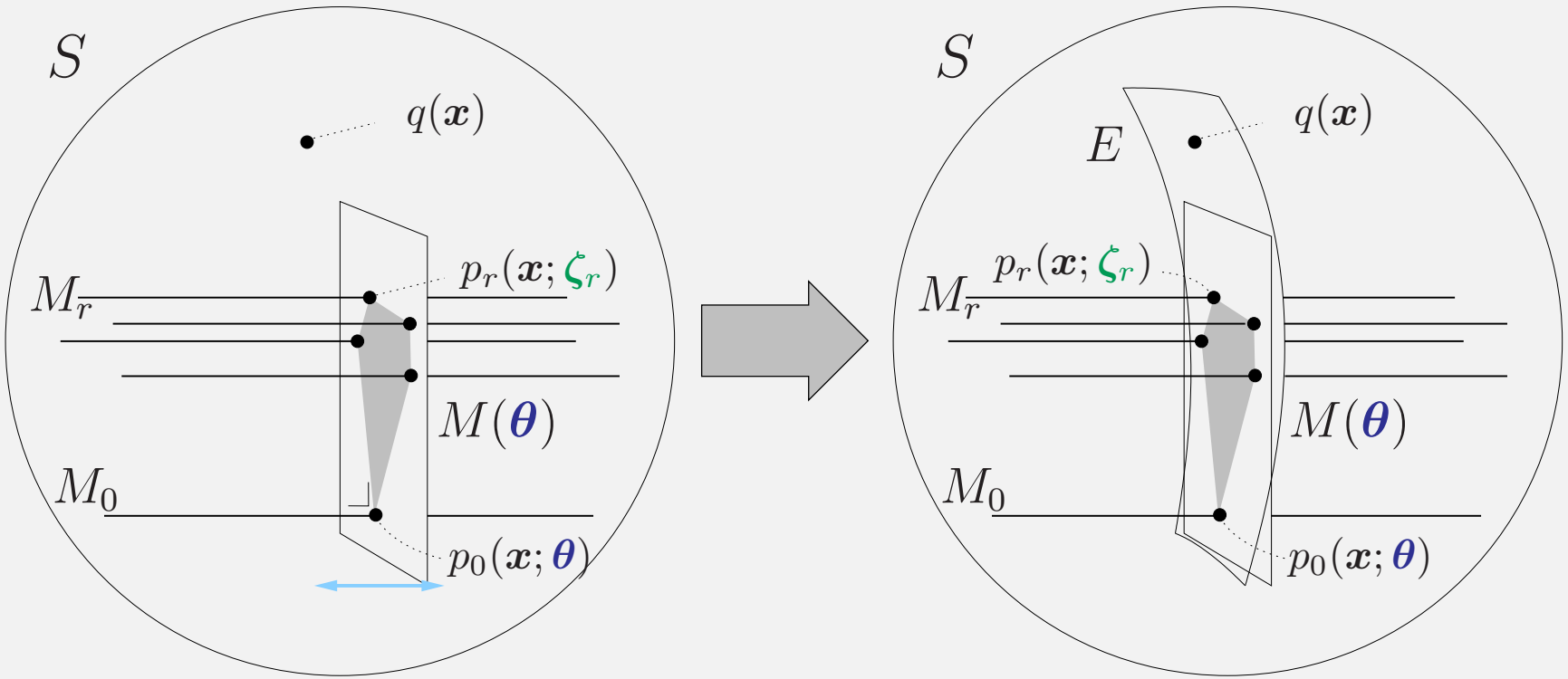
m–constraint algorithm



Condition 1 is always satisfied, update parameters to satisfy Condition 2

Convergence

m–constraint algorithm



CCCP (Yuille & Rangarajan, NIPS*15)

Survey Propagation

Background

Mézard, Parisi, & Zecchina, (2002). *Science*, *297*, 812–815.

Method to analyze K-sat problems.

3-sat problem

$$(x_1 \vee \bar{x}_2 \vee x_3) \wedge (\bar{x}_1 \vee \bar{x}_4 \vee x_5) \wedge (x_4 \vee x_2 \vee x_3)$$

3-sat is NP complete. For above example,
$(x_1, x_2, x_3, x_4, x_5) = (+, +, *, *, +)$ is a solution. Also $(*, *, +, -, *)$ is.

Survey Propagation

Notation

$$\psi_1(\boldsymbol{x}_1) = 1 - \frac{1}{8}(1 - x_1)(1 + x_2)(1 - x_3)$$

$\psi_1(\boldsymbol{x}_1)$ is 1 if $(x_1 \vee \bar{x}_2 \vee x_3)$ is True, otherwise 0, and

$$V = (x_1 \vee \bar{x}_2 \vee x_3) \wedge (\bar{x}_1 \vee \bar{x}_4 \vee x_5) \wedge (x_4 \vee x_2 \vee x_3) = \prod_r \psi_r(\boldsymbol{x}_r)$$

$$\psi_r(\boldsymbol{x}_r) = 1 - \frac{1}{8} \prod_{i \in V(r)} (1 + J_{ri} x_i), \quad J_{ri} \in \{-1, +1\}$$

if $V = 1$ is 1, it is SAT.

Probability Distribution

Model

$$q(\boldsymbol{x}) = \exp\Big(\beta \sum_r \psi_r(\boldsymbol{x}_r) - \varphi_q(\beta)\Big)$$

This distribution describes the SAT problem as $\beta \to \infty$.

Finite temperature

$\beta \to \infty$ is survey propagation.

Maneva, Mossel, & Wainwright, (2007). *JACM*, *54(4)*, No.17.

For finite temperature, the relation to BP is not clear.

Kabashima, (2005). *JPSJ*, *74(8)*, 2133–2136.

SP is based on BP

BP

Important distribution

$$p_r(\boldsymbol{x}_r;\zeta_r) = \exp(\beta\psi_r(\boldsymbol{x}_r) + \zeta_r \cdot \boldsymbol{x}_r - \varphi_r(\zeta_r))$$

$$p_0(x_i;\boldsymbol{\theta}) = \exp(\theta_i x_i - \varphi_0(\theta_i)) = \frac{\exp(\theta_i x_i)}{2\cosh(\theta_i x_i)}.$$

At the convergent point,

$$p_0(x_i;\theta_i^*) = \sum_{\boldsymbol{x}_r \setminus x_i} p_r(\boldsymbol{x}_r;\zeta_r^*)$$

$$q(\boldsymbol{x}) \propto p_0(\boldsymbol{x};\boldsymbol{\theta}^*) \prod_r \frac{p_r(\boldsymbol{x}_r;\zeta_r^*)}{p_0(\boldsymbol{x}_r;\boldsymbol{\theta}^*)} = \exp\Big(\beta \sum_r \psi(\boldsymbol{x}_r) + F_B(\boldsymbol{\theta}^*, \{\zeta_r^*\})\Big)$$

$F_B(\boldsymbol{\theta}^*, \{\zeta_r^*\}) = \sum_r(\varphi_0(\boldsymbol{\theta}^*) - \varphi_r(\zeta_r^*)) - \varphi_0(\boldsymbol{\theta}^*)$: Bethe free energy

Difference between SP and BP

BP

- There are many convergent points of BP
- SP utilizes all the BP convergent points

Let a be the index of and the convergent points be indexed as $\boldsymbol{\theta}^a$, $\{\zeta_r^a\}$. For each $a = 1, \cdots, A$ the following relation holds.

$$p_0(x_i; \theta_i^a) = \sum_{\boldsymbol{x}_r \setminus x_i} p_r(\boldsymbol{x}_r; \zeta_r^a)$$

$$q(\boldsymbol{x}) \propto p_0(\boldsymbol{x}; \boldsymbol{\theta}^a) \prod_r \frac{p_r(\boldsymbol{x}_r; \zeta_r^a)}{p_0(\boldsymbol{x}_r; \boldsymbol{\theta}^a)} = \exp\Big(\beta \sum_r \psi(\boldsymbol{x}_r) + F_B(\boldsymbol{\theta}^a, \{\zeta_r^a\})\Big)$$

$F_B(\boldsymbol{\theta}^a, \{\zeta_r^a\}) = \sum_r (\varphi_0(\boldsymbol{\theta}^a) - \varphi_r(\zeta_r^a)) - \varphi_0(\boldsymbol{\theta}^a)$: Bethe free energy

Idea of SP

Approximation with a mixture

Approximate $q(\boldsymbol{x})$ as follows

$$\sum_a \pi(a) p_0(\boldsymbol{x}; \boldsymbol{\theta}^a)$$

1-RSB ?

How to determine $\pi(a)$?

SP do not use $\pi(a)$ but distribution of ξ_r^a, ζ_r^a.

Update rule of SP

From Kabashima's result

If every ξ^a_{ri} and ζ^a_{ri} is different from each other depending on a, finite temperature SP is defined as follows.

1. Increase t by 1 and update $\pi(\xi^a_{ri})$ as follows,

$$\alpha\pi(\xi^a_{ri})\frac{\exp(\xi^a_{ri}x_i)}{\left(2\cosh(\xi^a_{ri})\right)^y} = \Big(\sum_{\boldsymbol{x}_r\setminus x_i} e^{\beta\psi_r(\boldsymbol{s}_r)}\Big)^y \prod_{j\in V(r)\setminus i}\frac{\pi(\zeta^a_{rj})\exp(\zeta^a_{rj}x_j)}{\left(2\cosh(\zeta^a_{rj})\right)^y}.$$

2. Update and $\pi(\zeta^a_{ri})$ as follows

$$\beta\pi(\zeta^a_{rj})\frac{\exp(\zeta^a_{ri}x_i)}{\left(2\cosh(\zeta^a_{ri})\right)^y} = \prod_{u\in\mathcal{N}(i)\setminus r}\frac{\pi(\xi^a_{ui})\exp(\xi^a_{ui}x_i)}{\left(2\cosh(\xi^a_{ui})\right)^y}.$$

Update rule of SP

$$\alpha \frac{\pi(\xi^a_{ri})}{\left(2\cosh(\xi^a_{ri})\right)^y} = e^{y(\varphi_r(\zeta^a_r)-\varphi_0(\theta^a_i))} \prod_{j\in V(r)\setminus i} \frac{\pi(\zeta^a_{rj})}{\left(2\cosh(\zeta^a_{rj})\right)^y}$$

$$\beta \frac{\pi(\zeta^a_{rj})}{\left(2\cosh(\zeta^a_{ri})\right)^y} = \prod_{u\in\mathcal{N}(i)\setminus r} \frac{\pi(\xi^a_{ui})}{\left(2\cosh(\xi^a_{ui})\right)^y},$$

Update rule of SP

If every ξ^a_{ri} and ζ^a_{ri} is different for each a $\pi(a) = \pi(\xi^a_{ri}) = \pi(\zeta^a_{ri})$ holds.

$$\frac{1}{\alpha}\pi(a)^{|V(r)|-2} = \left(e^{\varphi_0(\theta^a_i) - \varphi_r(\zeta^a_r)} \frac{\prod_{j\in V(r)} 2\cosh(\zeta^a_{rj})}{2\cosh(\xi^a_{ri}) 2\cosh(\zeta^a_{ri})} \right)^y$$

$$\frac{1}{\beta}\pi(a)^{|\mathcal{N}(i)|-2} = \left(\frac{\prod_{u\in\mathcal{N}(i)} 2\cosh(\xi^a_{ui})}{2\cosh(\zeta^a_{ri}) 2\cosh(\xi^a_{ri})} \right)^y,$$

Update rule of SP

Let E the number of edges and

$$\sum_r |V(r)| = E, \quad \sum_i |\mathcal{N}(i)| = E.$$

When $2\cosh(\zeta^a_{ri})2\cosh(\xi^a_{ri}) \simeq 4\cosh(\theta^a_i)$ holds. (since $\theta^a_i = \xi^a_{ri} + \zeta^a_{ri}$ holds, this means one of ξ^a_{ri} and ζ^a_{ri} is 0 or very large)

$$\alpha'\pi(a)^{E-2L} \simeq \frac{1}{2^L}\exp\Big(-y\sum_r \varphi_r(\zeta^a_r)\Big)\prod_r \prod_{j\in V(r)} \big(2\cosh(\zeta^a_{rj})\big)^y \tag{1}$$

$$\beta'\pi(a)^{E-2N} \simeq \frac{1}{2^N}\exp(-y\varphi_0(\theta^a_i))\prod_i \prod_{r\in\mathcal{N}(i)} \big(2\cosh(\xi^a_{ri})\big)^y, \tag{2}$$

Update rule of SP

$$C\pi(a)^{2(E-N-L)} \simeq \exp\Big(y\big[\sum_r(\varphi_0(\boldsymbol{\theta}_r^a) - \varphi_r(\boldsymbol{\zeta}_r^a)) - \varphi_0(\boldsymbol{\theta}^a)\big]\Big) \quad (3)$$

$$= \exp(yF_B(\boldsymbol{\theta}^a, \{\boldsymbol{\zeta}_r^a\})), \quad (4)$$

Conclusion

- Information geometrical view of BP
  - Approximation accuracy: perturbation analysis.
  - Convergence: e–m–constraint algorithms.
- New understanding of SP.