

確率的グラフィカルモデル の変分近似法

佐野 崇

東洋大学 情報連携学部

目次

1. 確率的グラフィカルモデルとは
2. 確率的グラフィカルモデルの推論
3. 確率的グラフィカルモデルの学習

今日の講義の目標

- 物理屋向けに、確率的グラフィカルモデルの面白いところ(統計力学が応用されているところ)を概観して紹介する
- ぜひとも物理屋に解決してもらいたい問題も紹介
- 逆に、この方法で物理の研究が進むのでは? という視点で機械学習を紹介するものではない

確率を用いたデータのモデル

- 収集したデータを説明するモデルがほしい
- 多くの場合、データは定まった値を持たず、確率的にゆらぐ
 - 例:理論的には定まった値が得られるはずでも、一般には測定誤差が乗る
- 測定すべき量を確率変数にとらえ、測定されたデータを確率変数の実現値と考える
- データから確率変数の従う確率分布を推定(学習)したい

- 例: ある物理量(長さ)を知りたい。しかし、測定誤差が存在する。誤差は正規分布(ガウス分布)すると仮定できる

$$P(X) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(X - \mu)^2}{2\sigma^2}\right) \equiv N(\mu, \sigma^2)$$

- 未知のパラメータは 平均 μ と分散 σ^2
- 複数回の測定値 x_1, x_2, \dots, x_n が得られたとき、ここから平均 μ と分散 σ^2 を推定する
- 測定値が独立に、同じ正規分布に従うのであれば、それらが同時に測定される確率は

$$P(x_1, x_2, \dots, x_n) = \prod_{i=1}^n P(X = x_i)$$

この尤度関数を最大化するようにパラメータを決定することを**最尤推定**という

- 尤度関数の代わりに対数尤度を最大化しても同じ

$$\log P(x_1, x_2, \dots, x_n) = \sum_{i=1}^n \ln P(X = x_i)$$

平均と分散について1階微分をゼロと置くと次を得る

$$\mu = \frac{1}{n} \sum_{i=1}^n x_i \quad \sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$$

最尤推定によって、パラメータを求めることができた
これは(広い意味で)学習である
(よく見ると不変推定量ではない)

多変数の場合

- 一般には、多数の測定値が同時に得られる。それらを確率変数の実現値だと思つと、多数の確率変数を持つ同時確率分布が背後にある、と考えることができる
- N 個の変数を考えるなら $P(X_1, X_2, \dots, X_N)$
- 仮に各変数が2値(0か1)を取りうるとしても、この確率分布は 2^N 個の確率値を保存しておく必要がある($O(2^N)$ のパラメータが必要)
- 一般的には非現実的

因子化の仮定

- 多くの場合は、お互いに関係のない(独立な)確率変数がある
- それらの同時確率分布は、単独の確率分布の単純な積
- 例: 2個の確率変数 X_1, X_2 が独立であれば

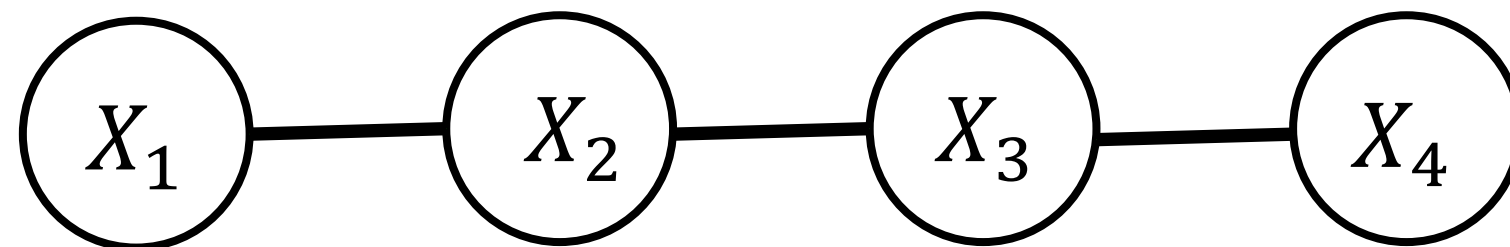
$$P(X_1, X_2) = P(X_1)P(X_2)$$

因子化の図形的表現

- 例えば4個の確率変数 X_1, X_2, X_3, X_4 があって、独立でない組み合わせは X_1, X_2 、 X_2, X_3 、 X_3, X_4 のみであるとする。同時確率分布は

$$P(X_1, X_2, X_3, X_4) = P(X_1, X_2)P(X_2, X_3)P(X_3, X_4)$$

と因子化できる。このとき、確率変数を頂点(ノード)とし、独立でない組み合わせの間に辺(エッジ)を書き入れてグラフとして表現すると...

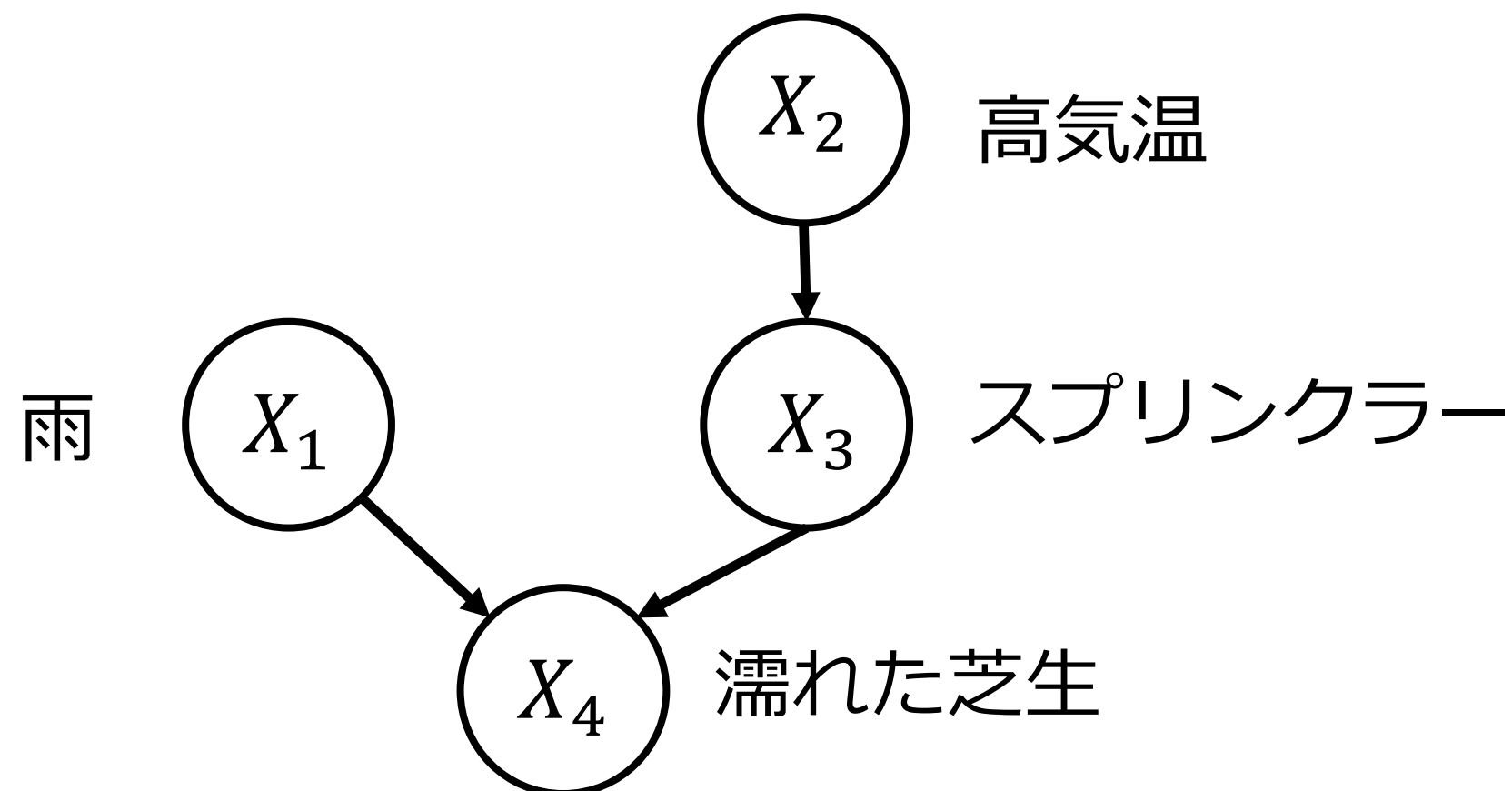


同時確率分布がどのように因子化されるかをグラフとして表したものを**確率的グラフィカルモデル**という。これは後述する**マルコフ確率場**の一例

ベイジアンネットワーク

- いろいろな確率変数が実現値を取るときに、背後の現象を反映していることがある
- 例：以下の4つの2値(真または偽)確率変数を考える
 - X_1 : 天気が雨である
 - X_2 : 気温が高い
 - X_3 : スプリンクラーが作動している
 - X_4 : 芝生が濡れている
- この4つには因果関係があるはず
 - 気温が高いとスプリンクラーが作動する(X_3 は X_2 に依存)
 - スプリンクラーが作動するか雨が降ると芝生が濡れる(X_4 は X_3 と X_1 に依存)
- この因果関係を図に書き表してみる

ベイジアンネットワーク

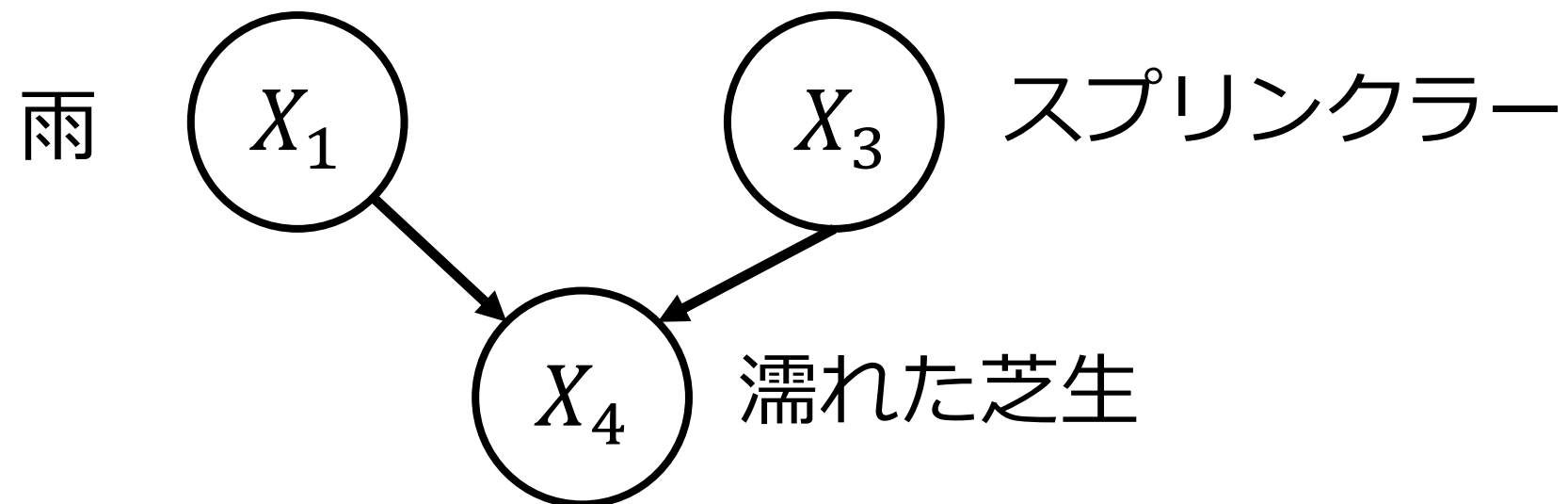


$$P(X_1, X_2, X_3, X_4) = P(X_4 | X_1, X_3) P(X_3 | X_2) P(X_1) P(X_2)$$

このような確率変数の因果関係(条件付き確率による同時確率分布の因子化)を図形的に表したものを**ベイジアンネットワーク**という

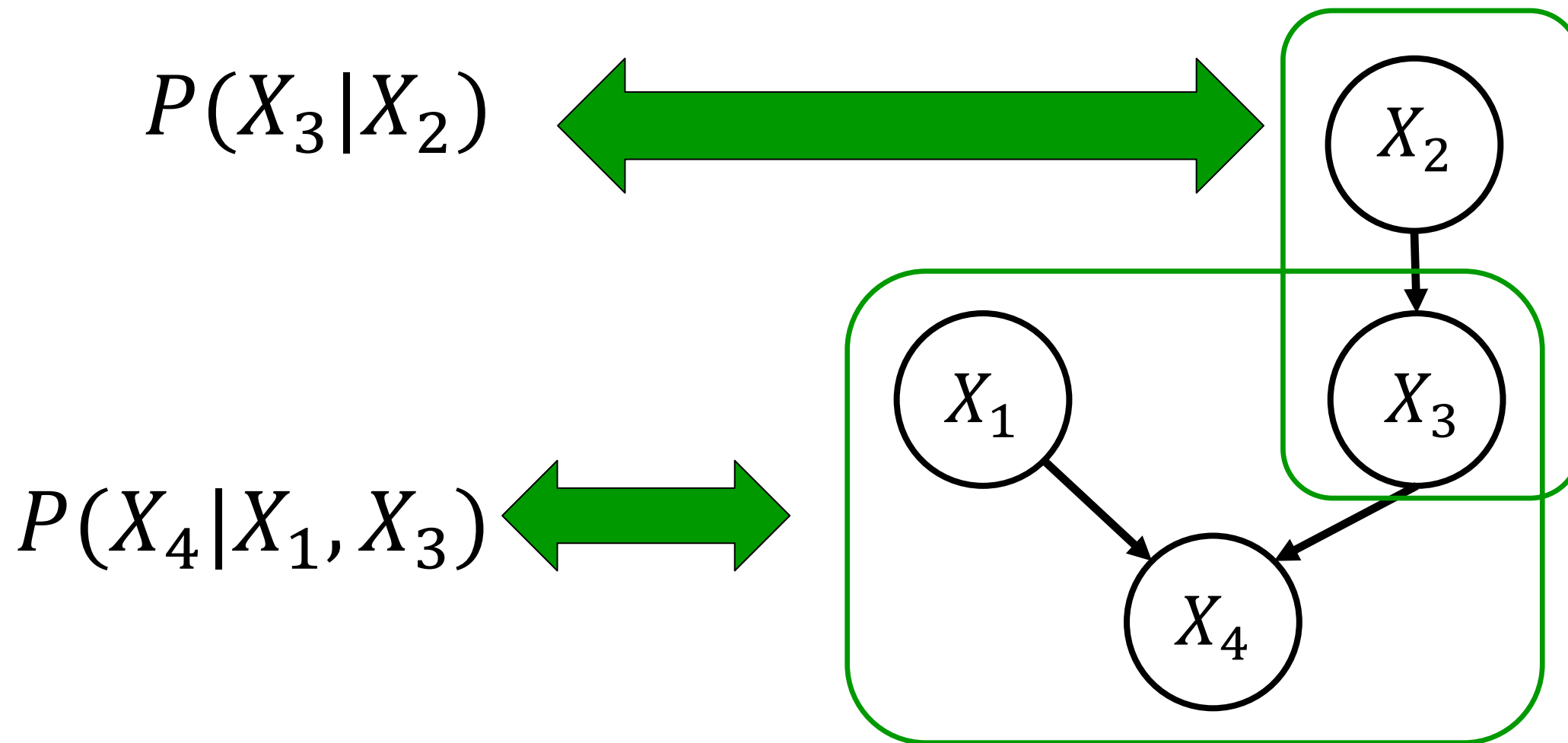
ベイジアンネットワークの見方

- ある変数に矢印を向けている変数を親変数という
- 子変数の確率は、親変数集合の条件付き確率になる



$$P(X_4 | X_1, X_3)$$

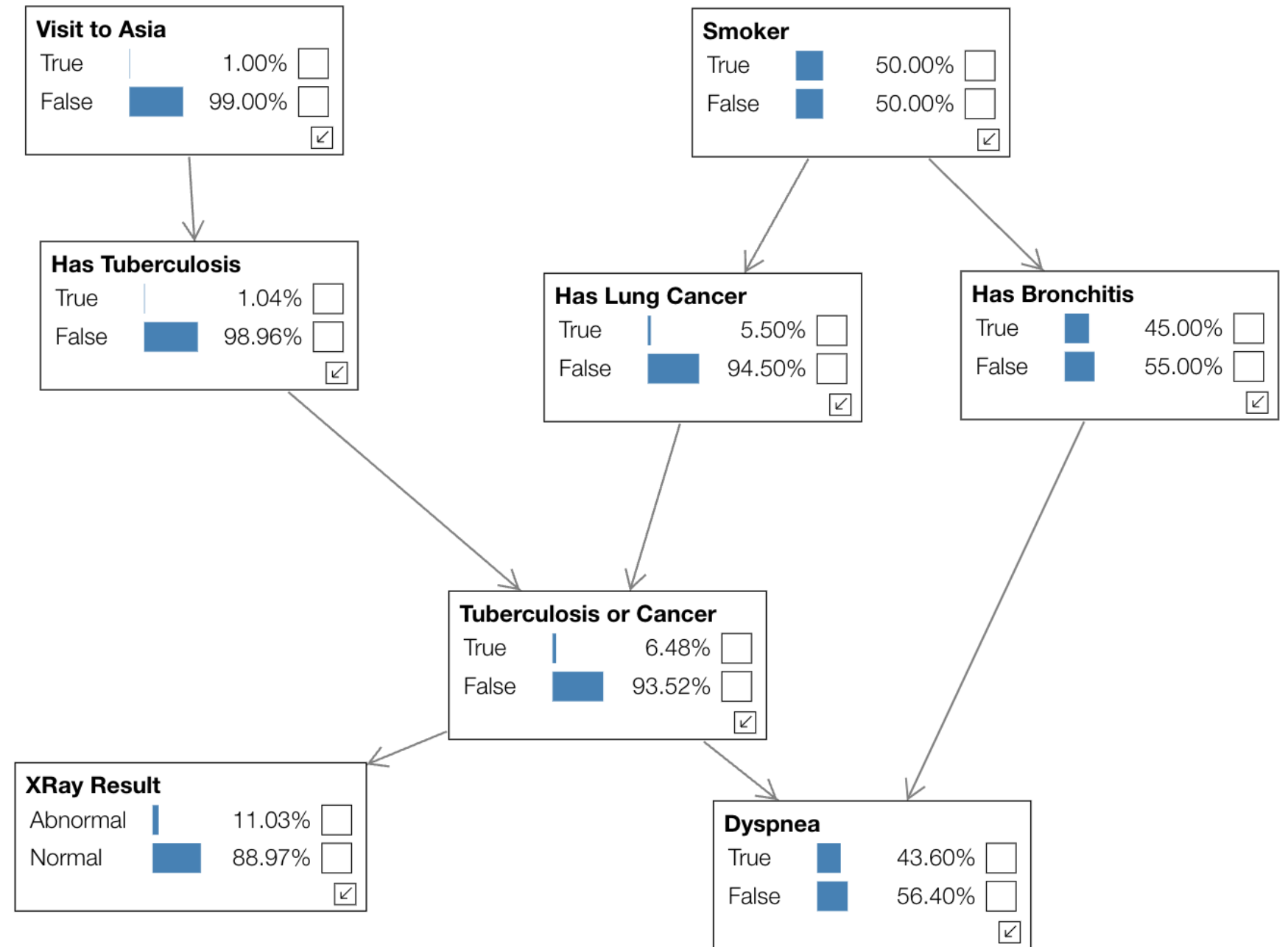
- すべての親子関係を条件付き確率に変換
- 親のない変数には事前分布を割り当てる
- それらの積が同時確率分布



$$P(X_1, X_2, X_3, X_4) = P(X_4|X_1, X_3)P(X_3|X_2)P(X_1)P(X_2)$$

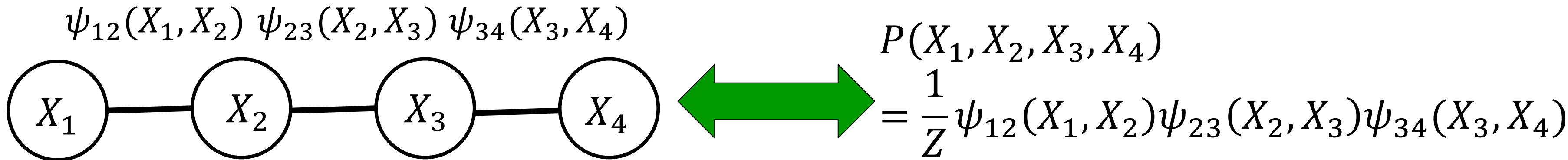
ASIA ネットワーク

- 肺疾患に関するベイジアンネットワーク
- 例えば dyspnea(呼吸困難)、XRy result(所見)のあるなしといった観測を与えたときに、肺がんである確率がどのくらいあるかを推論できる



無向グラフ: マルコフ確率場

- 辺で結ばれた変数は因子化できず、同じ関数(ポテンシャル関数)に所属するとする
- 同じポテンシャルに所属する変数同士を向きのない辺で結ぶ

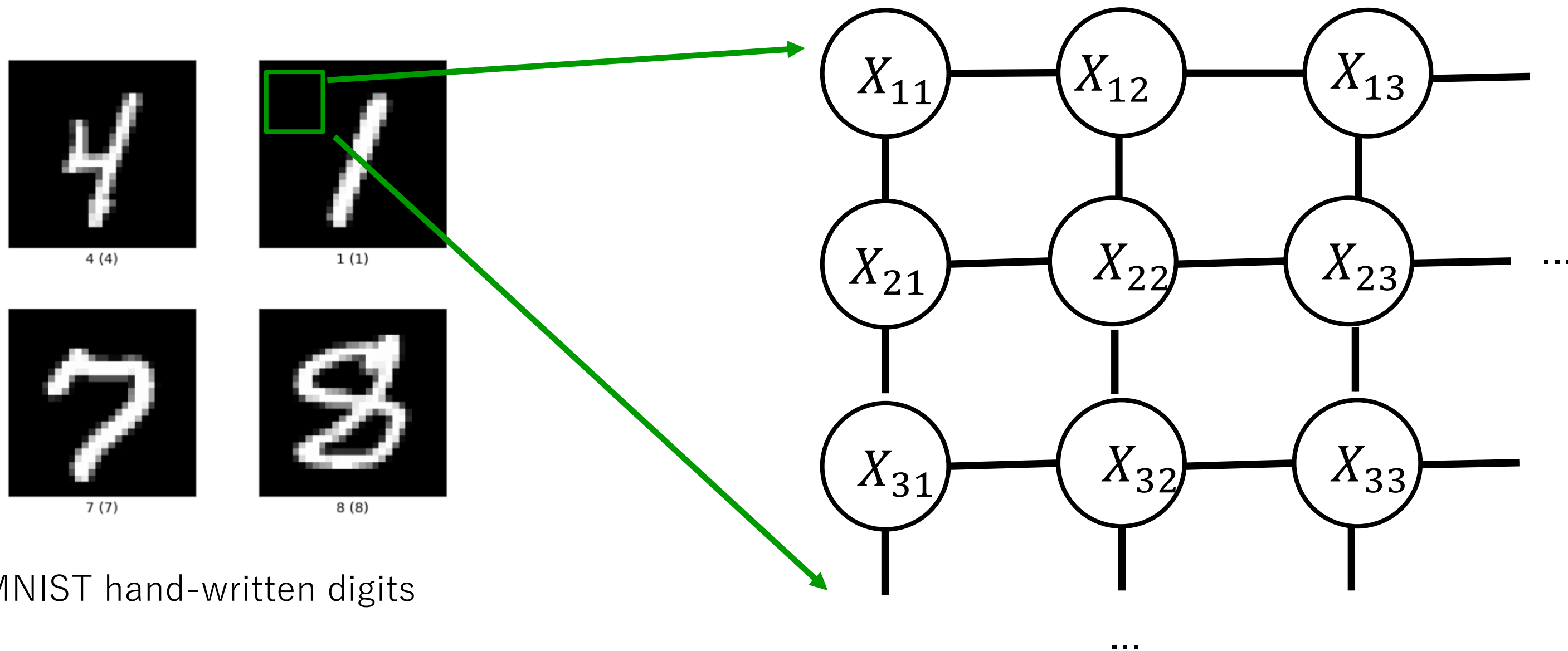


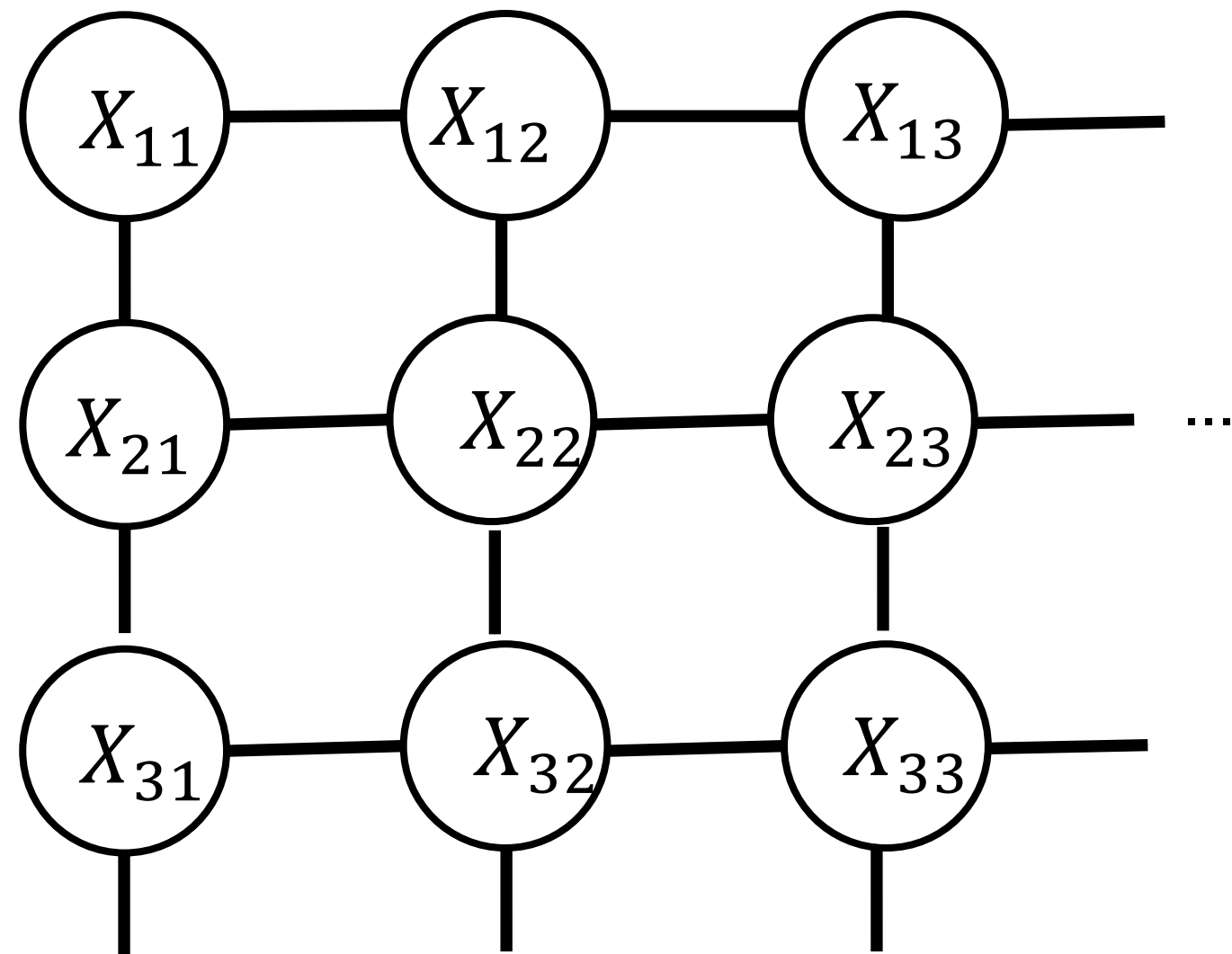
Zは確率を規格化するための**分配関数**

$$Z = \sum_{\mathbf{X}} \psi_{12}(X_1, X_2) \psi_{23}(X_2, X_3) \psi_{34}(X_3, X_4)$$

マルコフ確率場の例:ボルツマンマシン

- ホップフィールドネットワーク(連想記憶モデル)の確率版
- 例えば2次元の画像画像の各ピクセルが2値確率変数であり、その確率が隣接するピクセルの状態を決まるとする





$$\begin{aligned}
 P(\mathbf{X}) &= \frac{1}{Z} \prod_{(i,j) \in E} \psi_{ij}(X_i, X_j) \prod_{i \in V} \psi_i(X_i) \\
 &= \frac{1}{Z} \exp \left(\sum_{(i,j) \in E} W_{ij} X_i X_j + \sum_{i \in V} b_i X_i \right) \\
 &\equiv \frac{1}{Z} \exp(-E(\mathbf{X}))
 \end{aligned}$$

$\mathbf{X} = \{X_1, \dots, X_N\}$: 2値確率変数(0または1)

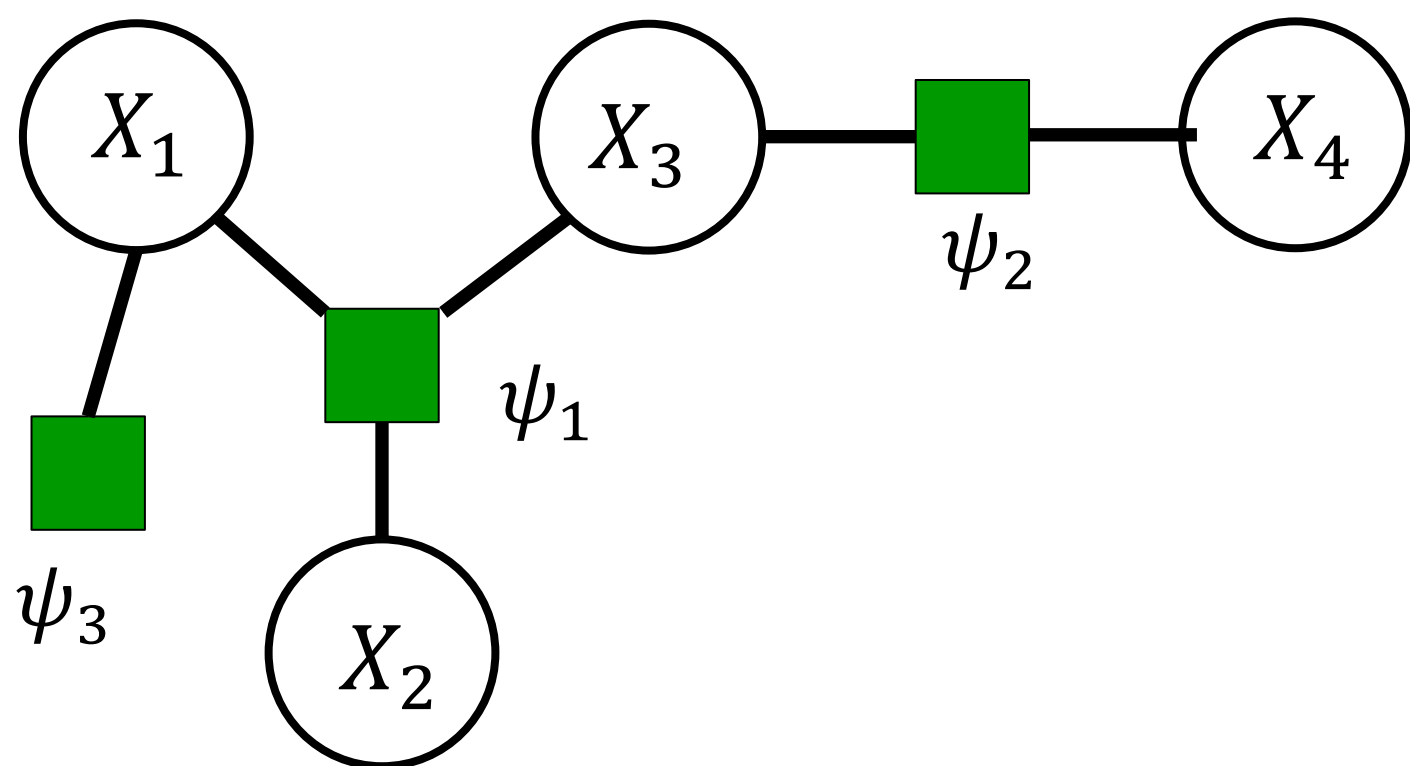
W_{ij}, b_i : パラメータ (重みとバイアス)

E : 辺(edge)集合、 V : 頂点(vertex)集合

- 画像の一部を復元したり、ノイズの除去が可能
- Ising (spin glass)モデルと等価

因子グラフ

- より一般には、3つ以上の確率変数が同じポテンシャル関数に所属しても良い
- その場合は変数頂点と因子頂点を持つ2部グラフとして表現できる(因子グラフ)



$$P(\mathbf{X}) = \frac{1}{Z} \prod_{f \in F} \psi_f(\mathbf{X}_f)$$

$$= \frac{1}{Z} \psi_1(X_1, X_2, X_3) \psi_2(X_3, X_4) \psi_3(X_1)$$

F : 因子の集合

\mathbf{X}_f : 因子 f に属する変数の集合

- マルコフ確率場やベイジアンネットワークは因子グラフの特別な場合
- 因子グラフで一般的な理論を作り、そこからマルコフ確率場やベイジアンネットワークに適用することもよく行われる

統計力学との対応

- 因子グラフの因子を指数の肩にあげてやると...

$$P(\mathbf{X}) = \frac{1}{Z} \prod_{f \in F} \psi_f(\mathbf{X}_f) = \frac{1}{Z} \exp(-E(\mathbf{X}))$$

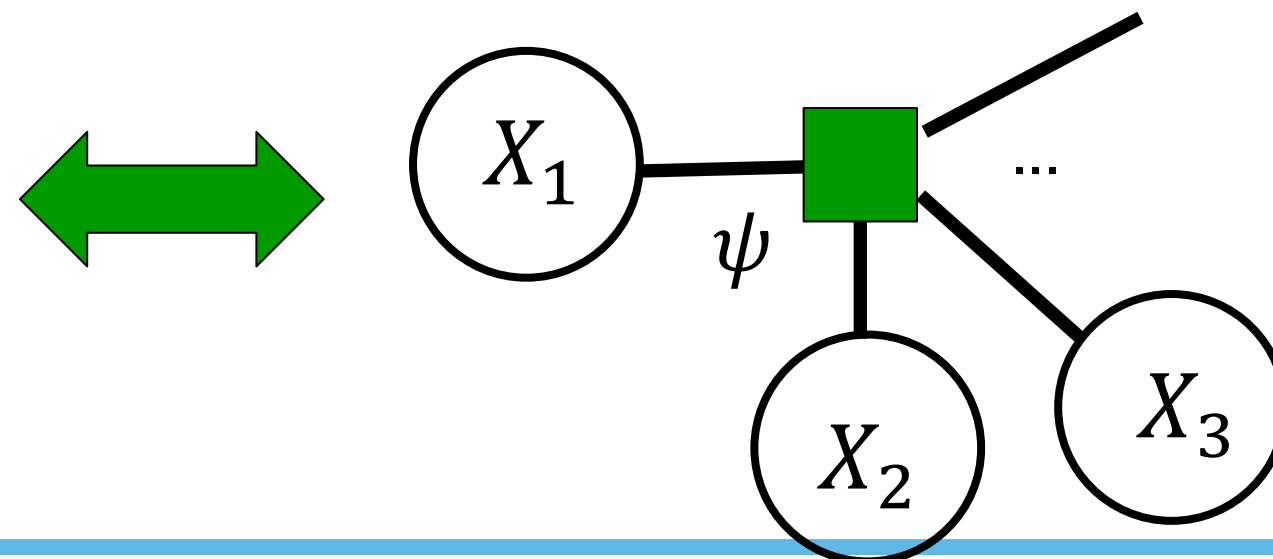
ここで $E(\mathbf{X}) = \sum_{f \in F} E_f(\mathbf{X}_f)$, $E_f(\mathbf{X}_f) = -\log \psi_f(\mathbf{X}_f)$

- すなわち、状態 X に対してエネルギー(ハミルトニアン)が $E(\mathbf{X})$ と与えられたときに、 X はボルツマン分布に従う(温度=1)
- すぐに分かるように、この対応から統計力学的手法をグラフィカルモデルに使うことができる(同じ困難を共有しているとも言える)

結局確率的グラフィカルモデルとは?

- 確率分布に何らかの因子化を仮定すると、それをグラフとして表現することができる
- 極論すれば、すべての確率分布(確率モデル) は確率的グラフィカルモデル、とも言える
- 因子化を仮定してから、それをグラフに表現して考察したり、逆にグラフを作ってから、対応する確率分布を書き下したり、といった使い方をする

$$P(X_1, X_2, \dots, X_N) \\ = \frac{1}{Z} \psi(X_1, X_2, \dots, X_N)$$



確率的グラフィカルモデル の推論

確率的グラフィカルモデルの推論

- 確率的グラフィカルモデルが与えられたとする(条件付き確率分布やポテンシャル関数が定まっているとする)
- このとき、しばしば「最も高い確率値を持つ確率変数の実現値」に興味がある

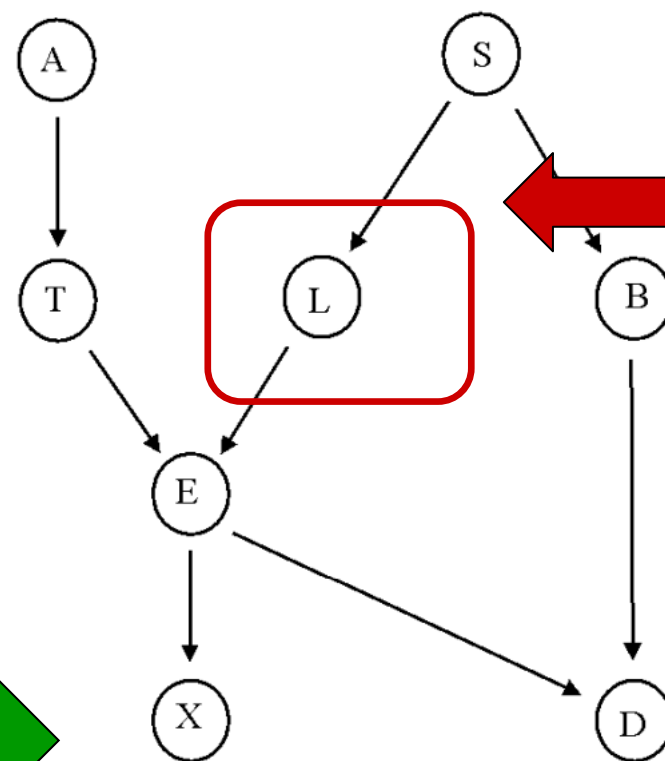
$$X = \operatorname{argmax}_X P(X)$$

- 実際上は、いくつかの確率変数が観測されたときに、観測されていない変数について推定を行いたい
 - 呼吸困難の所見があり、アジアへの訪問歴のある人は肺がんであるか?
 - 画像の半分が欠落したときに、残りの部分から復元を行う

$$X = \operatorname{argmax}_X P(X|\hat{X})$$

- **MAP**(Maximum a posteriori:最大事後確率)推定という
- 素朴にやると、2値変数がN個あるときに $O(2^N)$ の確率値の比較が必要であり、困難

X線所見(X)、アジアの渡航歴(A)が分かったとき



肺がんである(L)といえるか? あるいはその確率は?

ASIAネットワーク(Yedidia 2001より)

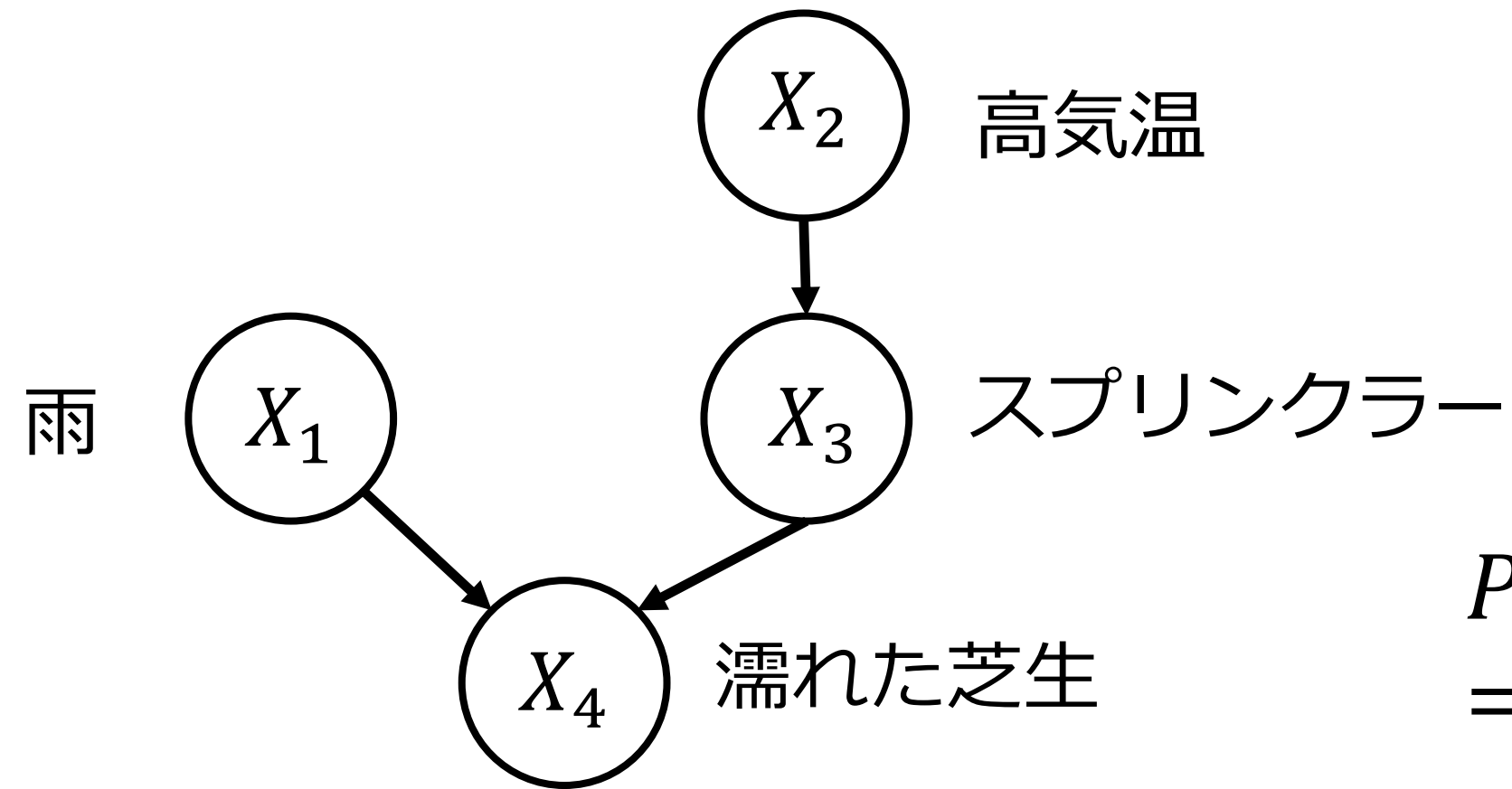
周辺確率

- MAP推定の他にも、注目する変数以外の変数を積分してしまった周辺確率に興味があることもある
 - 例：呼吸困難の所見があり、アジアへの訪問歴のある人が肺がんである確率は？

$$P_i(X_i) = \sum_{\mathbf{X} - \{X_i\}} P(\mathbf{X})$$

- こちらも素朴にやると、 $O(2^N)$ の和が必要
- MAP推定とよく似ている(maxと和の置き換え)が、周辺確率のargmaxを取ってもMAPが得られるとは限らない

ベイジアンネットワークの例



$$P(X_1, X_2, X_3, X_4) \\ = P(X_4 | X_1, X_3) P(X_3 | X_2) P(X_1) P(X_2)$$

- 実際の推論はいくつかの変数が観測されたときに、観測されていない変数の値や確率に興味がある
- 芝生が濡れている($X_4 = 1$)ときに、雨である確率 $P(X_1 = 1)$ は?
- 以下では簡単のため、観測された変数は陽には現れないが、観測があっても理論は同じ

統計力学との関係

- 統計力学(や場の理論)で問題になるのは、ある変数(例えばスピン)が、与えられた確率分布のもとでどのような期待値を持つかだった

$$\langle s_i \rangle = \sum_{\mathbf{s}} s_i P(\mathbf{s}) = \frac{1}{Z} \sum_{\mathbf{s}} \exp[-E(\mathbf{s})]$$

- これは周辺確率を求める問題とほぼ同じ
- よく知られているように、一般にはこの $O(2^N)$ の和を取ることはできない
- そのため、物理屋は色々な近似法を考案してきた
- **サンプリング法**(モンテカルロ法)や**変分近似法**(平均場近似)はどちらも統計的推論に応用されている
- 今回は特に、変分近似法について見ていく

変分近似法

- 手に負えない確率分布 $P(X)$ に対して、より扱いやすい確率分布 $Q(X)$ を導入する
- $Q(X)$ をできるだけ $P(X)$ に近づける
- 2つの確率分布の近さとして、Kullback-Leibler divergence (**KL情報量**)を用いる

$$\text{KL}(Q|P) \equiv \sum_X Q(X) \log \frac{Q(X)}{P(X)}$$

KL情報量は距離ではないが、距離に似た性質をもつ

- $\text{KL}(Q|P) \geq 0$
 - $\text{KL}(Q|P) = 0 \leftrightarrow Q(X) = P(X) \forall X$
- ただし一般に $\text{KL}(Q|P) \neq \text{KL}(P|Q)$

※ 以下では主に離散確率変数を扱うが、和を積分に置き換えれば連続確率変数でもほぼ同様のことが成立

変分分布

- Q を適当にパラメータ付けして、KL情報量を最小化するようにパラメータを求めたい
- KL情報量をパラメータで変分(微分)して最小値を探すので、変分近似法という。 Q を変分分布という
- よく使われる Q の形は、次の形の完全に因子化されたもの

$$Q(\mathbf{X}) = \prod_{i=1}^N q_i(X_i)$$

- これを**平均場近似**という
- KL情報量が最小になるように、パラメータ q_i を決定する。 q_i の満たす方程式は

$$q_i(X_i) \propto \exp\left[\sum_{\mathbf{X}-X_i} \prod_{j \neq i} q_j(X_j) \ln P\right] \quad X_i \text{ 以外で } \ln P \text{ の期待値を取ると、} \ln q_i \text{ に一致する}$$

平均場近似?

- 統計力学を勉強すると平均場近似が出てくるが、それとはかなり違うように見えるので、チェックしてみる
- 簡単のためIsingモデルを考えることにして、その教科書的な平均場近似を考察する
- 相互作用と外場が局所的に異なるspin glassモデルを考えてみる

$$E = - \sum_{\langle ij \rangle} J_{ij} S_i S_j - \sum_{i=1}^N h_i S_i$$

$\langle ij \rangle$ は接続のあるスピンの組み合わせの集合

- よくある説明は、スピン変数を、平均値 m_i とその周りのゆらぎに分解し、ゆらぎを小さいと考える

$$\begin{aligned}
 s_i &= m_i + \delta s_i & E &= - \sum_{\langle ij \rangle} J_{ij} (m_i + \delta s_i)(m_j + \delta s_j) - \sum_i h_i s_i \\
 & & &\cong - \sum_{\langle ij \rangle} J_{ij} (m_i m_j + \delta s_i m_j + m_i \delta s_j) - \sum_i h_i s_i && \delta s \text{ の 2 次 を 無 視} \\
 & & &= - \sum_{\langle ij \rangle} J_{ij} (s_i m_j + m_i s_j - m_i m_j) - \sum_i h_i s_i \equiv E_{MF}
 \end{aligned}$$

近似ハミルトニアンは s_i の1次のみになった。つまり

$$P(\mathbf{s}) \propto \exp[-E_{MF}] = \prod_i \exp[-E_{MF}^{(i)}(s_i)]$$

と全体の確率が1体の確率の積の形で書ける → Qの因子化と同じ!

- 次に、KL情報量の最小化の意味を考えてみる。一般の確率分布 Q に対して

$$\text{KL}(Q|P) = \sum_{\mathbf{X}} Q(\mathbf{X}) \ln \frac{Q(\mathbf{X})}{P(\mathbf{X})} = \sum_{\mathbf{X}} Q(\mathbf{X}) (\ln Q(\mathbf{X}) - \ln P(\mathbf{X}))$$

ここで、 $P = \frac{1}{Z} \exp(-E)$ であることを使うと

$$\begin{aligned} \text{KL}(Q|P) &= \sum_{\mathbf{X}} Q(\mathbf{X}) \ln Q(\mathbf{X}) + \sum_{\mathbf{X}} Q(\mathbf{X}) (E + \ln Z) \\ &= -H[Q] + \sum_{\mathbf{X}} Q(\mathbf{X}) E(\mathbf{X}) + \ln Z \end{aligned}$$

ただし $H[Q] = -\sum_{\mathbf{X}} Q(\mathbf{X}) \ln Q(\mathbf{X})$: エントロピー関数

まとめると

$$\begin{aligned} \text{KL}(Q|P) &= -H[Q] + \sum_{\mathbf{X}} Q(\mathbf{X})E(\mathbf{X}) + \ln Z \\ &= G[Q] - F \end{aligned}$$

$F = -\ln Z$: (真の)自由エネルギー

$G = \sum_{\mathbf{X}} Q(\mathbf{X})E(\mathbf{X}) - H[Q]$: 変分自由エネルギー

すなわち、KL情報量の最小化とは、変分分布 Q を動かして、**変分自由エネルギー G を最小化**し、できるだけ真の自由エネルギーに近づけようすることだった

- KL情報量が非負であることから、 $G \geq F$, つまり変分自由エネルギーは真の自由エネルギーの上界であることも分かる
- 等号が成立する $G = F$ のは Q が P に一致するときであるが、 Q の形に制限を置いてしまうと、一般に等号は成立しない

- Isingモデルの場合に、この平均場近似が、self-consistency 方程式を導くことをチェックする
- 変分分布を次の形に仮定

$$Q(\mathbf{s}) = \prod_{i=1}^N q_i(s_i) = \prod_{i=1}^N \frac{1}{2} (1 + s_i m_i)$$

このとき変分自由エネルギー G は

$$G = - \sum_{\langle ij \rangle} J_{ij} m_i m_j - \sum_i h_i m_i - \sum_i H[q_i]$$


m_i について微分して0とおくと

$$- \sum_{j \in N(i)} J_{ij} m_j - h_i + \frac{1}{2} \ln \frac{1 + m_i}{1 - m_i} = 0$$

$N(i)$: i と接続のある変数の集合

- 逆双曲線関数 $\tanh^{-1} x = \frac{1}{2} \ln \frac{1+x}{1-x}$ であるから

$$0 = - \sum_{j \in N(i)} J_{ij} m_j - h_i + \tanh^{-1} m_i$$


$$m_i = \tanh\left(\sum_{j \in N(i)} J_{ij} m_j + h_i\right)$$

特に、相互作用や外場が一様な場合 ($J_{ij} = J, h_i = h$)、期待値も一様 $m_i = m$ であるから

$$m = \tanh(zJm + h)$$

$$z = \sum_{j \in N(i)} 1: \text{変数 } s_i \text{ と相互作用を持つ変数の数}$$

有名なself-consistency 方程式が再現できた!

平均場近似を解くアルゴリズム

- 与えられた確率分布 P に対して、

$$q_i(X_i) \propto \exp\left[\sum_{X-X_i} \prod_{j \neq i} q_j(X_j) \ln P\right]$$

が成立するような $q_i(X_i)$ ($i = 1, \dots, N$)を見つける

- それには、適当な値で $\ln q_i(X_i)$ を初期化して

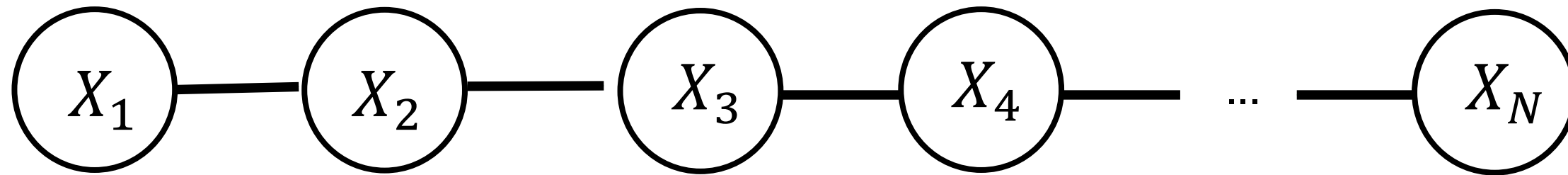
$$\ln q_i(X_i) \leftarrow \sum_{X-X_i} \prod_{j \neq i} q_j(X_j) \ln P$$

のように反覆更新していくのが簡便(規格化は適当に行う)

- 変分自由エネルギー G は $q_i(X_i)$ について凸であるため、この更新は収束性が非常に良い
- なお、平均場近似では、周辺確率を最大化する実現値は、この近似でのMAP推定値でもある

グラフ構造を利用する

- 平均場近似はどのようなモデルでも用いることができるが、もっと積極的にグラフ構造を用いた近似方法がないだろうか
- 例えば、以下のようなマルコフ確率場を考えてみる



$$P(X) \propto \psi_{12}(X_1, X_2) \psi_{23}(X_2, X_3) \dots \psi_{N-1N}(X_{N-1}, X_N)$$

いま、周辺確率 $P(X_n)$ を求めたいとする。

闇雲に計算すると、 $N-1$ 個の確率変数について和を取ることであり、 $O(2^N)$ の計算が必要になる

- 一方で、グラフ構造にそって端から計算を行うとどうなるか?
- まず X_1 について周辺化

$$\begin{aligned}
 P(X_2, \dots, X_N) &\propto \sum_{X_1} \psi_{12}(X_1, X_2) \psi_{23}(X_2, X_3) \dots \psi_{N-1N}(X_{N-1}, X_N) \\
 &= \mu_\alpha(X_2) \psi_{23}(X_2, X_3) \dots \psi_{N-1N}(X_{N-1}, X_N)
 \end{aligned}$$

ここで $\mu_\alpha(X_2) = \sum_{X_1} \psi_{12}(X_1, X_2)$ で、 $O(2)$ で計算できる

- 次に X_2 で周辺化

$$\begin{aligned}
 P(X_3, \dots, X_N) &\propto \sum_{X_2} \mu_\alpha(X_2) \psi_{23}(X_2, X_3) \dots \psi_{N-1N}(X_{N-1}, X_N) \\
 &= \mu_\alpha(X_3) \psi_{34}(X_3, X_4) \dots \psi_{N-1N}(X_{N-1}, X_N)
 \end{aligned}$$

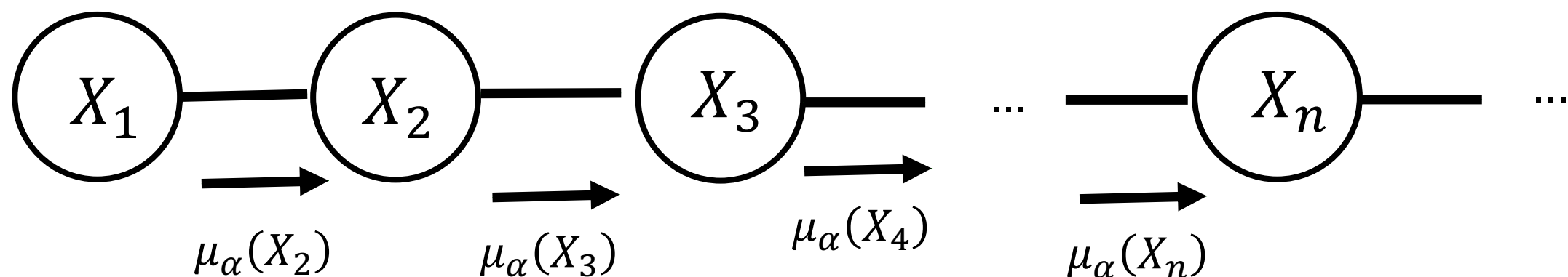
ここで $\mu_\alpha(X_3) = \sum_{X_2} \mu_\alpha(X_2) \psi_{23}(X_2, X_3)$ である。計算量は $O(2^2)$ (2x2行列とベクトルの積)

- これを繰り返すと

$\mu_\alpha(X_n) = \sum_{X_{n-1}} \mu_\alpha(X_{n-1}) \psi_{n-1,n}(X_{n-1}, X_n)$ で $\mu_\alpha(X_n)$ を順に計算していき

$P(X_n, \dots, X_N) \propto \mu_\alpha(X_n) \psi_{n,n+1}(X_n, X_{n+1}) \dots \psi_{N-1,N}(X_{N-1}, X_N)$ を得る

- μ_α は、グラフの左側からの「メッセージ」と解釈できる



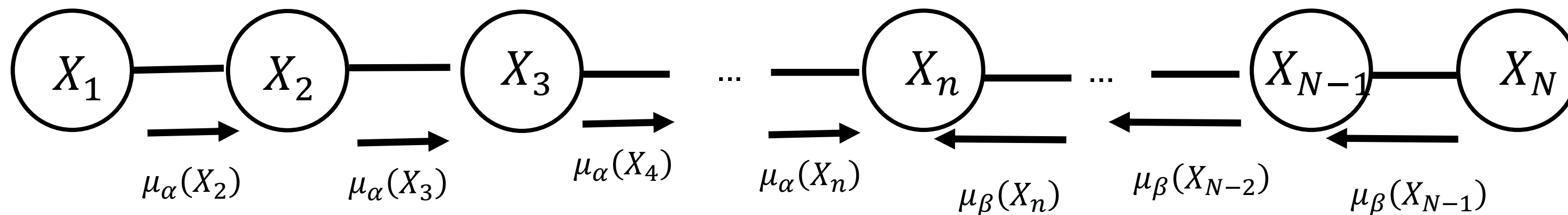
変数がメッセージを受け取ると、ポテンシャル関数をかけ、左側の変数について和をとって、自分のメッセージとして次の変数に送り出す

$$\mu_\alpha(X_n) = \sum_{X_{n-1}} \mu_\alpha(X_{n-1}) \psi_{n-1,n}(X_{n-1}, X_n)$$

メッセージを使って確率が表せる

$$P(X_n, \dots, X_N) \propto \mu_\alpha(X_n) \psi_{n,n+1}(X_n, X_{n+1}) \dots \psi_{N-1,N}(X_{N-1}, X_N)$$

- 同じことを、今度は X_N から逆向きに計算していく



メッセージの計算式:

$$\mu_\beta(X_n) = \sum_{X_{n+1}} \mu_\beta(X_{n+1}) \psi_{n,n+1}(X_n, X_{n+1})$$

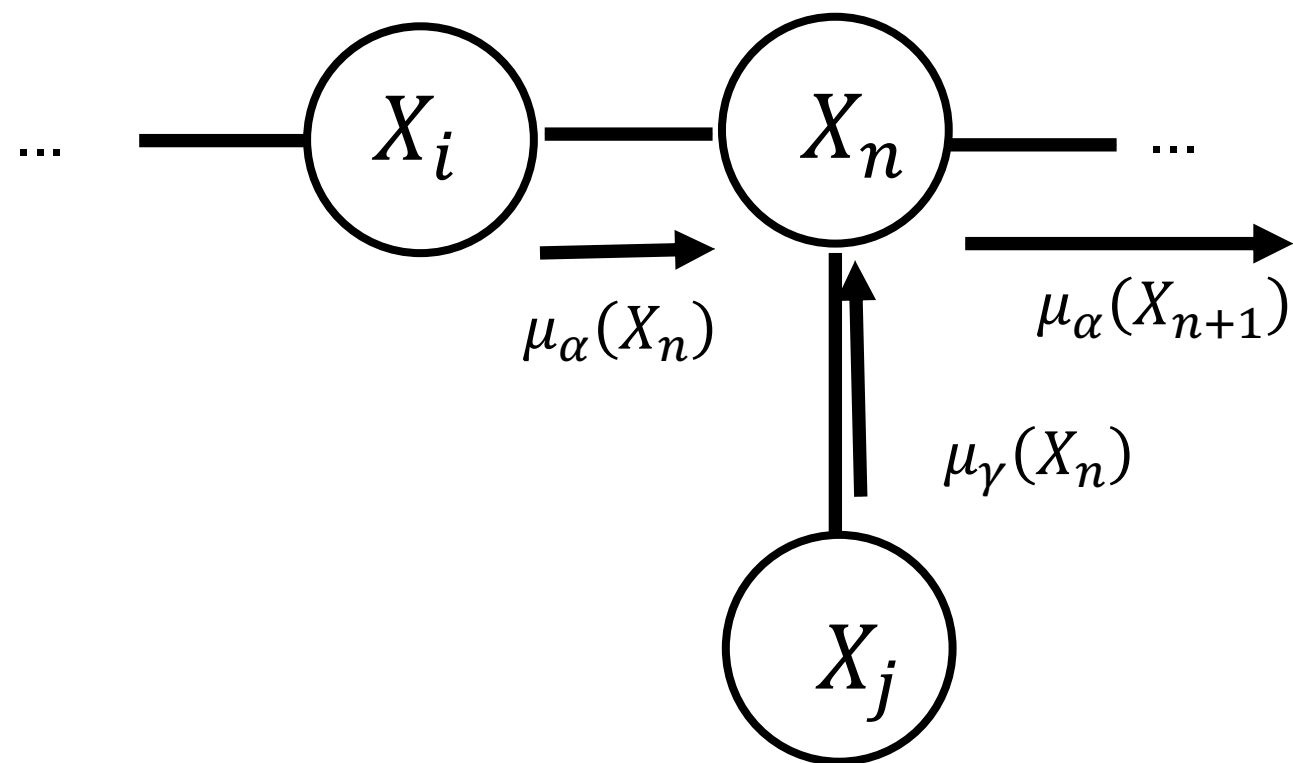
最終的に求めたい周辺確率は

$$P(X_n) \propto \mu_\alpha(X_n) \mu_\beta(X_n)$$

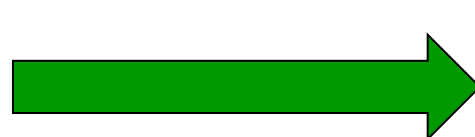
メッセージをすべて計算するのにかかる計算量は $O(N^2)$ 、
それらの積を取ることですべての変数の周辺確率を求められる!

確率伝播法という

- もしグラフに枝分かれがあっても、メッセージの集約を行えば新しいメッセージを計算できる

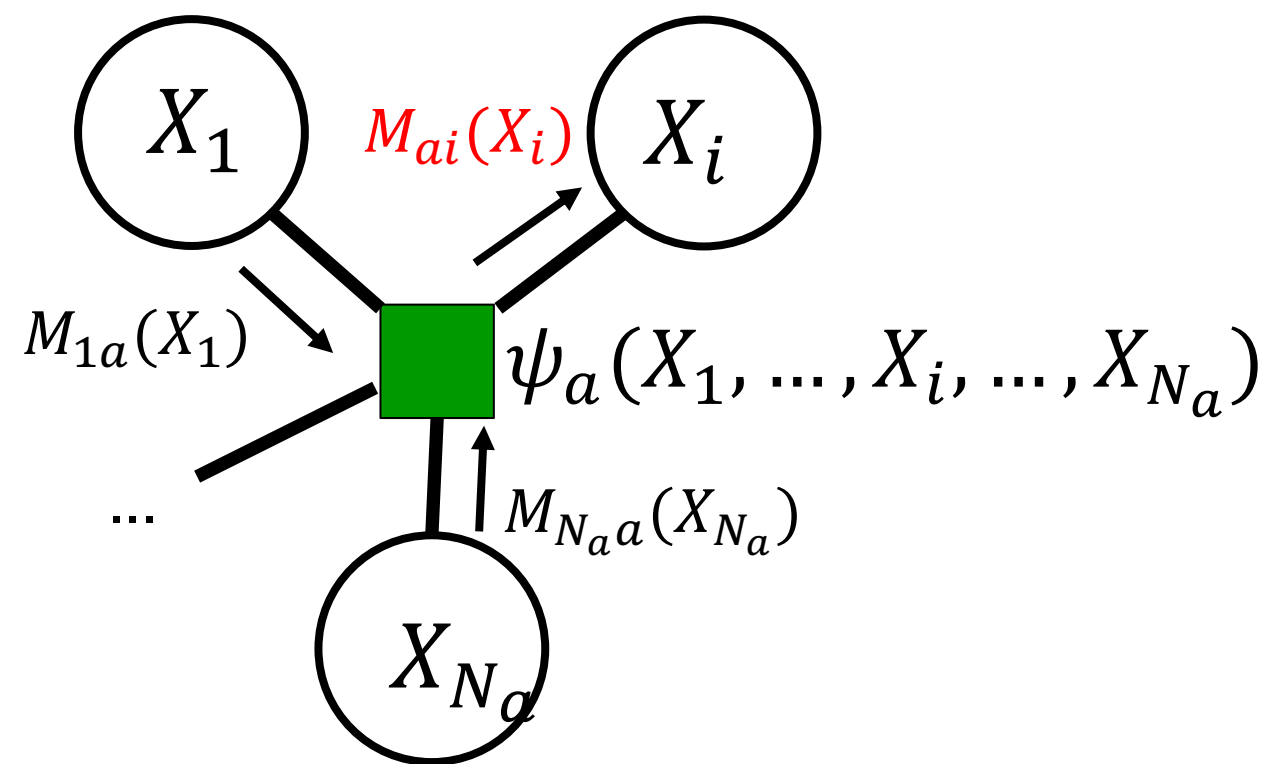


$$\begin{aligned} & \sum_{X_n} \mu_\alpha(X_n) \mu_\gamma(X_n) \psi_{n,n+1}(X_n, X_{n+1}) \psi_{n+1,n+2}(X_{n+1}, X_{n+2}) \dots \\ & = \mu_\alpha(X_{n+1}) \psi_{n+1,n+2}(X_{n+1}, X_{n+2}) \dots \end{aligned}$$



$$\mu_\alpha(X_{n+1}) = \sum_{X_n} \mu_\alpha(X_n) \mu_\gamma(X_n) \psi_{n,n+1}(X_n, X_{n+1})$$

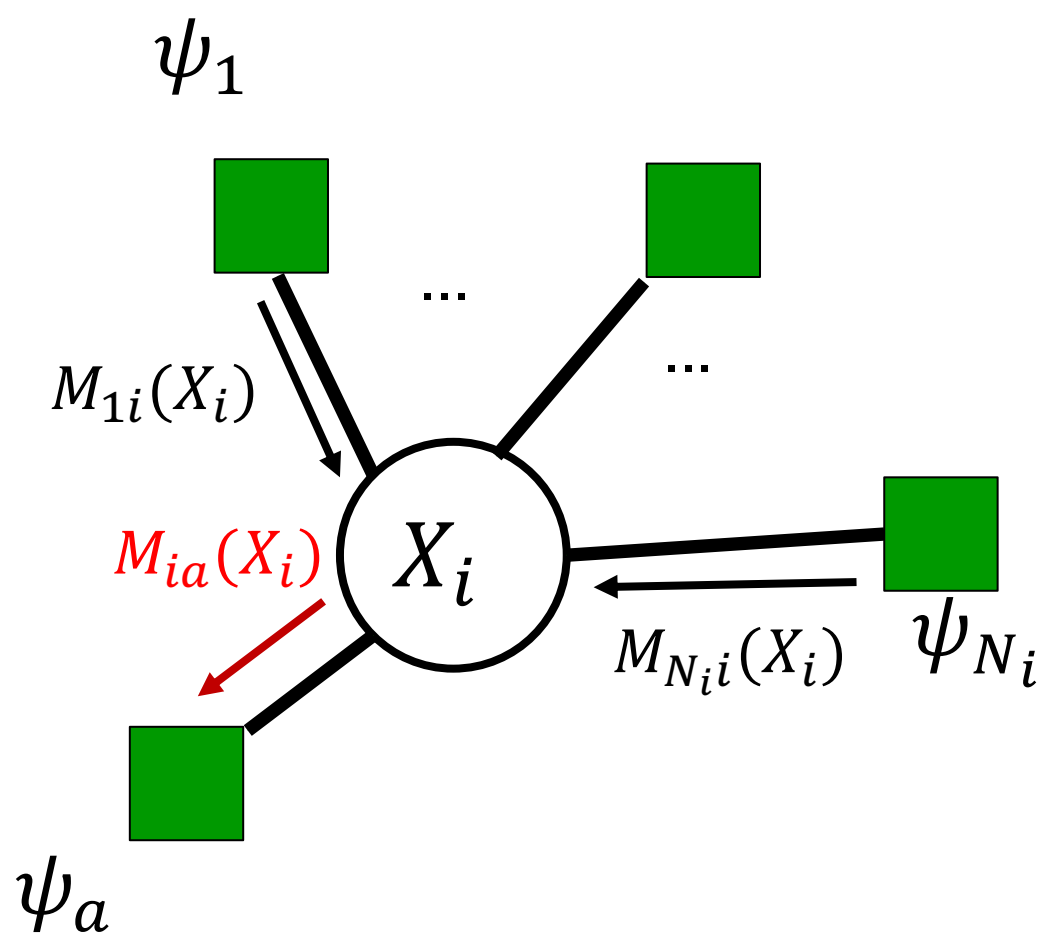
- 同じアイデアは、もっと複雑なグラフにも適用できる
- 一般に因子グラフの場合を紹介する
- E を因子の集合、 V を変数の集合とし、 a, b, c, \dots を因子の添字、 i, j, k, \dots を変数の添字に使う
- メッセージには、因子から変数へのメッセージ M_{ai} と変数から因子へのメッセージ M_{ia} がある。



$X_1, \dots, X_i, \dots, X_{N_a}$ が所属する因子 ψ_a から変数 X_i へ送られるメッセージ $M_{ai}(X_i)$ は、変数から因子へ送られるメッセージを用いて

$$M_{ai}(X_i) = \sum_{\{X_1, \dots, X_{N_a}\} - X_i} \psi_a(X_1, \dots, X_i, \dots, X_{N_a}) \prod_{j \neq i} M_{ja}(X_j)$$

- 変数 X_i から因子 ψ_a のメッセージは、 ψ_a 以外の因子からやってきたメッセージの積



$$M_{ia}(X_i) = \prod_{b \neq a} M_{bi}(X_i)$$

確率伝播法

- 最終的に、メッセージを用いて、周辺確率は次のように計算できる

$$P(X_i) \propto \prod_{a \in N(i)} M_{ai}(X_i)$$

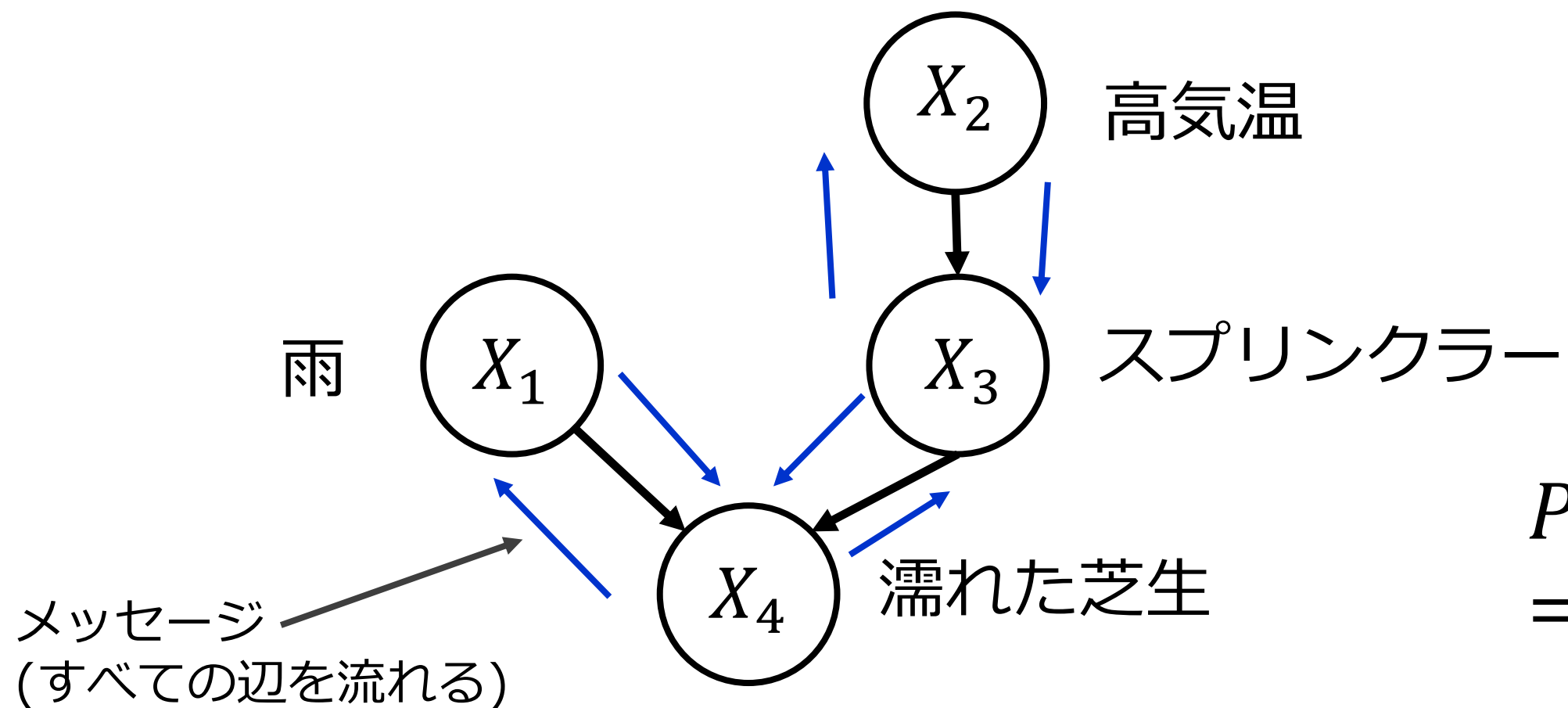
$$P(X_1, \dots, X_{N_a}) \propto \psi_a(X_1, \dots, X_{N_a}) \prod_{i \in N(a)} M_{ia}(X_i)$$

このような計算方法を**確率伝播法(Belief Propagation, BP)**や積和アルゴリズムと呼ぶ
特に1988年にPearlによってベイジアンネットワークの推論方法として導入され有名になった

命題: 確率伝播法は、グラフ構造が木のとき(閉路を含まないとき)、端のノードから順番にメッセージを計算していくことで、**厳密**な周辺確率を求めることができる

隠れマルコフモデルの推論アルゴリズムなども、実は確率伝播法と同等

ベイジアンネットワークの例



$$P(X_1, X_2, X_3, X_4) = P(X_4 | X_1, X_3) P(X_3 | X_2) P(X_1) P(X_2)$$

$X_4 = 1$ (濡れている) : 値が確定

- ベイジアンネットワークでも確率伝播法は(もちろん)実行できる
- 値が確定しているノードの情報がメッセージによって伝わり、その他の確率変数の周辺確率が求まる
- メッセージは矢印の順方向にも逆方向にも流れる

ループあり確率伝播法

- 確率伝播法で行う計算は局所的なものなので、グラフにループがあっても実行は可能
- もちろん収束するとは限らない
- 収束すれば近似解が得られると期待できる

ループあり確率伝播法 (Loopy Belief Propagation)

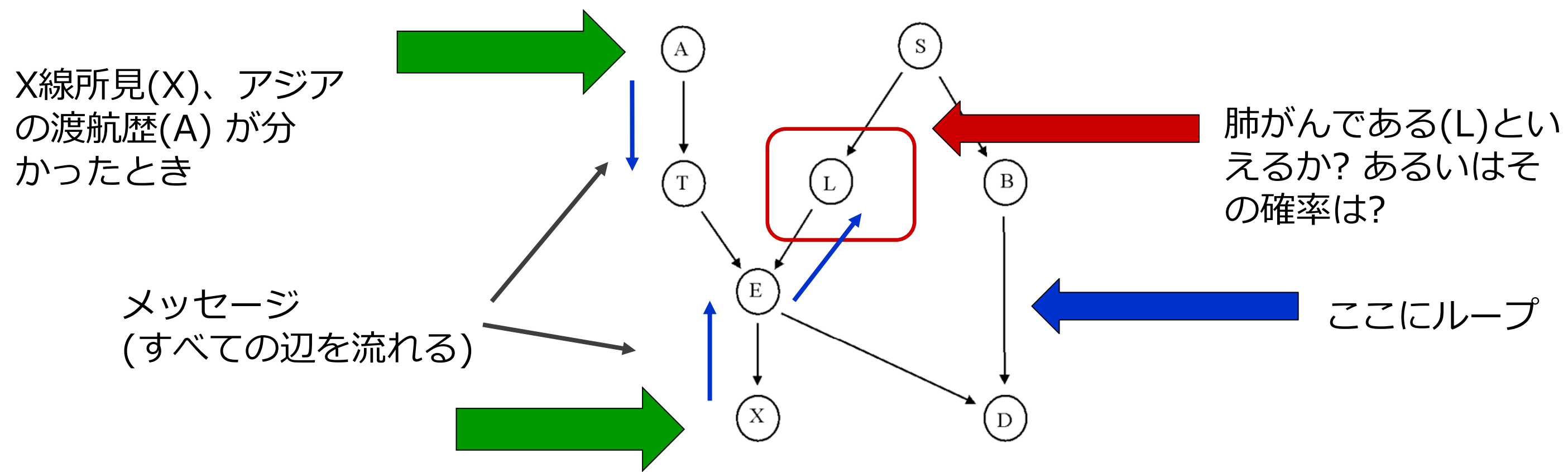
1. メッセージを適当に初期化する
2. 右の式でメッセージを収束するまで更新
3. 収束したメッセージを用いて近似周辺確率を計算

$$M_{ai}(X_i) = \sum_{\{X_1, \dots, X_{N_a}\} - X_i} \psi_a(X_1, \dots, X_i, \dots, X_{N_a}) \prod_{j \neq i} M_{ja}(X_j)$$

$$M_{ia}(X_i) = \prod_{b \neq a} M_{bi}(X_i)$$

$$P(X_i) \propto \prod_{a \in N(i)} M_{ai}(X_i)$$

$$P(X_1, \dots, X_{N_a}) \propto \psi_a(X_1, \dots, X_{N_a}) \prod_{i \in N(a)} M_{ia}(X_i)$$



- ループあり確率伝播法は、ループが少ないグラフィカルモデルでは良い近似を与える
- 例えばTurbo codeというエラー訂正符号化アルゴリズムは、実はループあり確率伝播法である

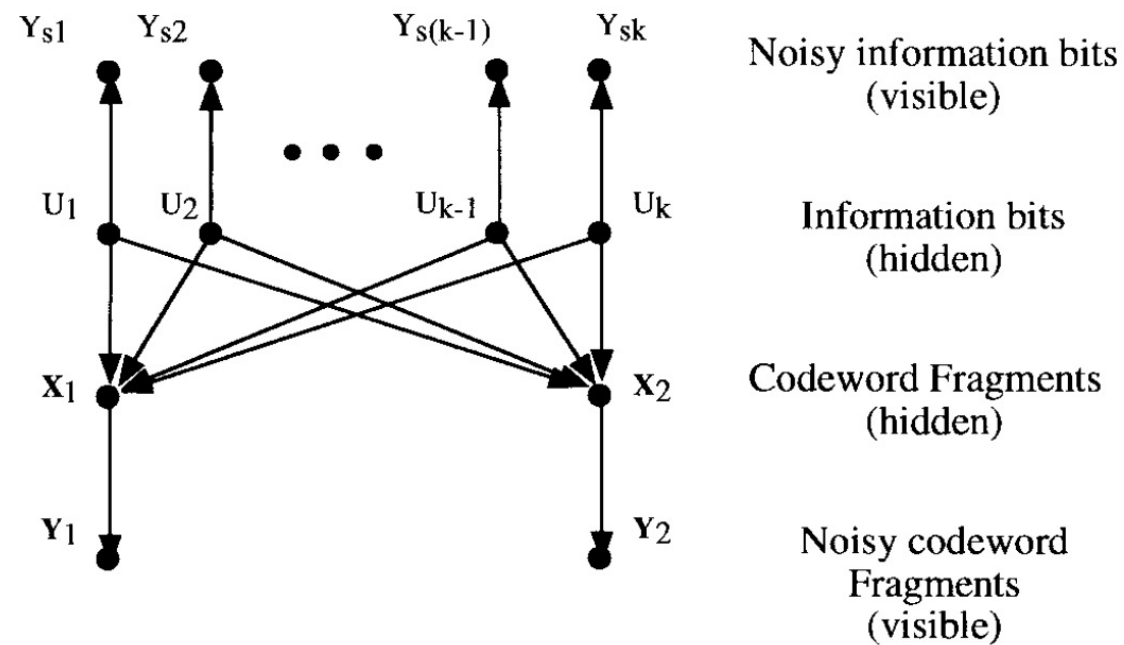


Fig. 7. Bayesian network interpretation of the turbo decoding problem. Note the presence of many loops, i.e., $U_1 \rightarrow X_2 \rightarrow U_2 \rightarrow X_1 \rightarrow U_1$.

McEliece+ 1998 より

- ループあり確率伝播法はヒューリスティックに導入されたが、すこし時代が下ると、理論的背景が色々と明らかにされた
- 驚くべきことに、ループあり確率伝播法とは、統計力学で有名なBethe近似と同じものであった
- Bethe近似は平均場近似の改良であり、ゆらぎの高次の項を取り込む

確率伝播法の理論

Yedidia+ 2001

- 表記を簡単にするため、マルコフ確率場の場合を考える
- 確率伝播法は、もし収束すれば、以下のBethe自由エネルギーの極値になる Q を与える

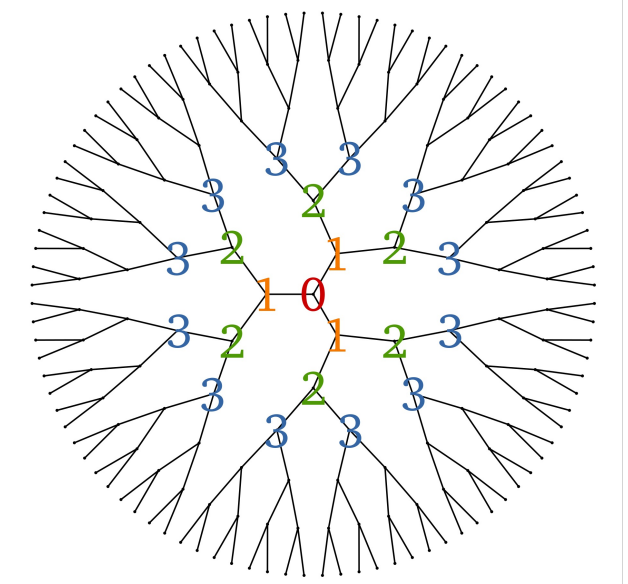
$$F_{\text{Bethe}} = \sum_{\mathbf{X}} Q(\mathbf{X}) E(\mathbf{X}) - H[Q] \quad \text{ただし} \quad Q(\mathbf{X}) = \frac{\prod_{(ij) \in E} q_{ij}(X_i, X_j)}{\prod_{i \in V} q_i(X_i)^{d_i-1}}$$

E : 辺の集合
 V : 変数の集合
 d_i : X_i に隣接する変数の数

$$q_i(X_i) = \sum_{X_j} q_{ij}(X_i, X_j)$$

cf. 平均場: $Q(\mathbf{X}) = \prod_{i=1}^N q_i(X_i)$

- 実は、これはBethe近似の目的関数に他ならない
- 平均場近似と比較すると、変分分布 $Q(\mathbf{X})$ の因子化が異なる。Bethe近似では隣接する変数の相関も取り込み、最適化を行う
- この因子化のせいで、Bethe自由エネルギーはもはや凸ではない
- Bethe近似はBethe格子(木構造)で厳密になるが、このことは確率伝播法が木構造のグラフ上で厳密であることと対応している


https://en.wikipedia.org/wiki/Bethe_lattice

平均場近似と確率伝播法

- 平均場近似も確率伝播法(Bethe近似)も、どちらも変分分布を仮定して、真の確率分布とのKL情報量が小さくなるように(=変分自由エネルギー G が小さくなるように)、変分パラメータ q_i, q_{ij} を決定している

$$G[Q] = \sum_{\mathbf{X}} Q(\mathbf{X}) E(\mathbf{X}) - H[Q]$$

$$Q(\mathbf{X}) = \prod_{i=1}^N q_i(X_i)$$

平均場

$$Q(\mathbf{X}) = \frac{\prod_{(ij) \in E} q_{ij}(X_i, X_j)}{\prod_{i \in V} q_i(X_i)^{d_i-1}}$$

確率伝播法

Welling & Teh 2003

- どちらも変分近似法的一种として統一的に理解できた
- すると、いろいろな変分分布 Q を用いることで、より良い近似ができそうである
- 実際、グラフ構造をクラスタに分けたKikuchi近似を元にした変分分布や、もっと自由に設定された一般化確率伝播法も提案されている

Yedidia+ 2000, 2001

確率伝播法の実際

- Bethe近似が平均場近似の改善であることから、常に確率伝播法を使えばよいのでは?と思われるかもしれない
- しかしながら、少しループの多いグラフィカルモデルでは確率伝播法は収束性が非常に悪く、また収束しても精度の悪い解しか得られないことが多い
 - 経験的には、変数が数百くらいだと厳しい(グラフ構造にもよるが)
- 平均場近似は自由エネルギーが凸であるため、収束性は非常に良い。そのために、複雑なモデルでは平均場近似が非常に多く用いられているのが現状である

確率的グラフィカルモデル の学習

グラフィカルモデルの学習とは

- 推論とは、グラフィカルモデルのパラメータ(ポテンシャル関数)が分かっているときに、確率変数の周辺確率やMAPを推定することだった
- 学習はその逆で、データとして確率変数の実現値(観測値)が与えられたときに、グラフィカルモデルのパラメータを求めることである
- Isingモデルだと、パラメータ J_{ij}, h_i が与えられたときに期待値 $\langle s_i \rangle, \langle s_i s_j \rangle$ を求めるのが推論であったが、逆にデータから計算される統計量 $\langle s_i \rangle_{\text{data}}, \langle s_i s_j \rangle_{\text{data}}$ を用いて J_{ij}, h_i を求めるのが学習である
- その意味で、学習のことを逆問題と呼ぶこともある

すべてのデータが与えられた場合

- 確率的グラフィカルモデルのすべての確率変数について、実現値の観測データが与えられる場合を考える
- 尤度関数を最大化することで、パラメータを学習(推定)することにしよう(最尤推定)

$$\ln \prod_{d=1}^D P(\mathbf{X}^{(d)}) = \sum_{d=1}^D \ln P(\mathbf{X}^{(d)})$$

$\mathbf{X}^{(d)}$: 確率変数の観測値。 $d = 1, \dots, D$

- マルコフ確率場や因子グラフには分配関数 Z が含まれる
- 分配関数の計算には指数的なコストがかかるため、尤度関数や、そのパラメータ微分を求めることは容易ではない
- ここで変分近似法を用いることができる

$$Z = \sum_{\mathbf{X}} \prod_a \psi_a(\mathbf{X}_a)$$

Isingモデルの学習

- Isingモデルについて、対数尤度関数をパラメータ J_{ij} 微分してみる

$$\begin{aligned}\frac{\partial}{\partial J_{ij}} \frac{1}{D} \sum_{d=1}^D \ln P(\mathbf{s}^{(d)}) &= \frac{\partial}{\partial J_{ij}} \frac{1}{D} \sum_{d=1}^D [E(\mathbf{s}^{(d)}) - \ln Z] \\ &= \frac{1}{D} \sum_{d=1}^D [s_i s_j] - \frac{\partial}{\partial J_{ij}} \ln Z = \langle s_i s_j \rangle_{\text{data}} - \langle s_i s_j \rangle\end{aligned}$$

この微分=0と置くと

$$\langle s_i s_j \rangle_{\text{data}} = \langle s_i s_j \rangle$$

moment matching 条件

h_i 微分からも同様に

$$\langle s_i \rangle_{\text{data}} = \langle s_i \rangle$$

- 左辺はデータを集計した平均値、右辺はIsingモデルで計算した期待値である
- 左辺は容易に計算できるが、右辺は一般には計算が大変である
- しかし、推論で用いた変分近似法を用いれば近似的に計算可能である

平均場近似を用いた学習

$$\langle s_i s_j \rangle_{\text{data}} = \langle s_i s_j \rangle$$

$$\langle s_i \rangle_{\text{data}} = \langle s_i \rangle$$

moment matching 条件

- アイデア: 右辺を平均場近似によってパラメータの関数として求めてしまい、これらを連立してパラメータについて解いてしまえば近似学習が可能である
- しかし、相関 $\langle s_i s_j \rangle$ は平均場近似で計算できるのか? $m_i m_j$ のようになってしまわないか?
- 実は、**線形応答関係式**を用いると、平均場近似でも $\langle s_i s_j \rangle$ が計算できる

$$\begin{aligned} \frac{\partial^2}{\partial h_i \partial h_j} \ln Z &= \frac{\partial}{\partial J_{ij}} \ln Z - \frac{\partial}{\partial h_i} \ln Z \frac{\partial}{\partial h_j} \ln Z \\ &= \langle s_i s_j \rangle - \langle s_i \rangle \langle s_j \rangle \end{aligned}$$

Kappen & Rodriguez 1997, Takana 1998

- この右辺は、moment matching 条件を用いて観測値の統計量と一致させる
- 左辺は、いわゆる感受率 $\partial m_j / \partial h_i$ であり、self-consistency 方程式を微分すると求められる

$$J_{ij} = \frac{\delta_{ij}}{1 - m_i^2} - [C^{-1}]_{ij}$$

$$\begin{aligned} C &= \langle s_i s_j \rangle_{\text{data}} - \langle s_i \rangle_{\text{data}} \langle s_j \rangle_{\text{data}} \\ m_i &= \langle s_i \rangle_{\text{data}} \end{aligned}$$

Bethe近似を用いた学習

- どんな近似方法でも、モデルで計算された相関(期待値)と、データの統計量を一致させる(moment matching)ことで学習が可能である
- Bethe近似(確率伝播法)を用いた学習法も、もちろん色々提案されていた
- 2012年、Bethe近似を用いたIsingモデルの学習について、非常に重要な論文が出る: [Ricci-Tersenghi 2012](#)
 - Bethe近似のもとで線形応答関係式とmoment matching条件を連立させたとき、その解として最適なパラメータ J_{ij} が閉形式で求まる

Bethe近似によるIsingモデルの学習方程式

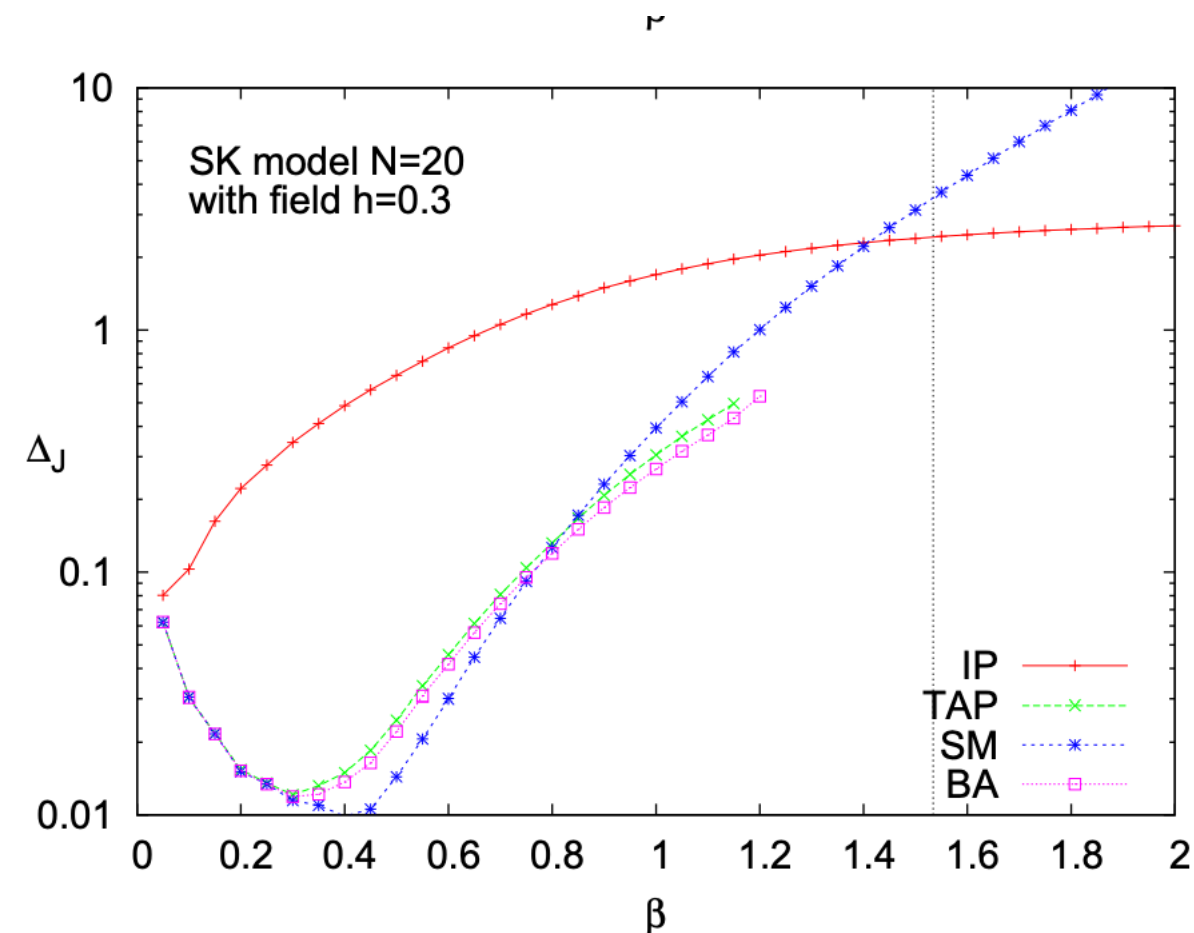
$$C = \langle S_i S_j \rangle_{\text{data}} - \langle S_i \rangle_{\text{data}} \langle S_j \rangle_{\text{data}}$$

$$m_i = \langle S_i \rangle_{\text{data}}$$

$$J_{ij}^{\text{BA}} = -\text{atanh} \left[\frac{1}{2(C^{-1})_{ij}} \sqrt{1 + 4(1 - m_i^2)(1 - m_j^2)(C^{-1})_{ij}^2} - m_i m_j - \frac{1}{2(C^{-1})_{ij}} \sqrt{\left(\sqrt{1 + 4(1 - m_i^2)(1 - m_j^2)(C^{-1})_{ij}^2} - 2m_i m_j (C^{-1})_{ij} \right)^2 - 4(C^{-1})_{ij}^2} \right]$$

Ricci-Tersenghi 2012

- Jについての4次方程式であるが、閉形式で求まる事が分かった
- グラフ構造にもよるが、単純な平均場と比較して精度が良かった
- 一方、平方根の中が負になってしまう場合がある



隠れ変数がある場合

- すべての変数の観測値が与えられておらず、観測されない隠れ(潜在)変数がある場合は多い
 - 可視変数集合を X 、隠れ変数集合を Z と書く
- 隠れ変数がある場合、それについて周辺化を行い、最尤推定を行う。尤度関数 L は

$$L = \frac{1}{D} \sum_{d=1}^D \ln P(\mathbf{X}^{(d)}) = \frac{1}{D} \sum_{d=1}^D \ln \sum_{\mathbf{Z}} P(\mathbf{X}^{(d)}, \mathbf{Z})$$

- 分配関数の時と同様に、確率変数 Z についての和に計算コストがかかる
- これについても、変分近似法による取り扱いがある

変分ベイズ法

- 尤度関数を次のように変形してみる

$$P(\mathbf{X}^{(d)}, \mathbf{Z}) = P(\mathbf{Z}|\mathbf{X}^{(d)})P(\mathbf{X}^{(d)}) \quad \longrightarrow \quad \ln P(\mathbf{X}^{(d)}) = \ln P(\mathbf{X}^{(d)}, \mathbf{Z}) - \ln P(\mathbf{Z}|\mathbf{X}^{(d)})$$

両辺に適当な確率分布 $Q(\mathbf{Z}|\mathbf{X}^{(d)})$ をかけ、 \mathbf{Z} について和を取ると、左辺はそのままなので、

$$\begin{aligned} \ln P(\mathbf{X}^{(d)}) &= \sum_{\mathbf{Z}} Q(\mathbf{Z}|\mathbf{X}^{(d)}) [\ln P(\mathbf{X}^{(d)}, \mathbf{Z}) - \ln P(\mathbf{Z}|\mathbf{X}^{(d)})] \\ &= \sum_{\mathbf{Z}} Q(\mathbf{Z}|\mathbf{X}^{(d)}) \left[\ln \frac{P(\mathbf{X}^{(d)}, \mathbf{Z})}{Q(\mathbf{Z}|\mathbf{X}^{(d)})} - \ln \frac{P(\mathbf{Z}|\mathbf{X}^{(d)})}{Q(\mathbf{Z}|\mathbf{X}^{(d)})} \right] \\ &= \sum_{\mathbf{Z}} Q(\mathbf{Z}|\mathbf{X}^{(d)}) \left[\ln \frac{P(\mathbf{X}^{(d)}, \mathbf{Z})}{Q(\mathbf{Z}|\mathbf{X}^{(d)})} \right] + \text{KL}(Q|P) \end{aligned} \quad \text{KL}(Q|P) = \sum_{\mathbf{Z}} Q(\mathbf{Z}|\mathbf{X}^{(d)}) \ln \frac{Q(\mathbf{Z}|\mathbf{X}^{(d)})}{P(\mathbf{Z}|\mathbf{X}^{(d)})}$$

KL情報量は非負であるから

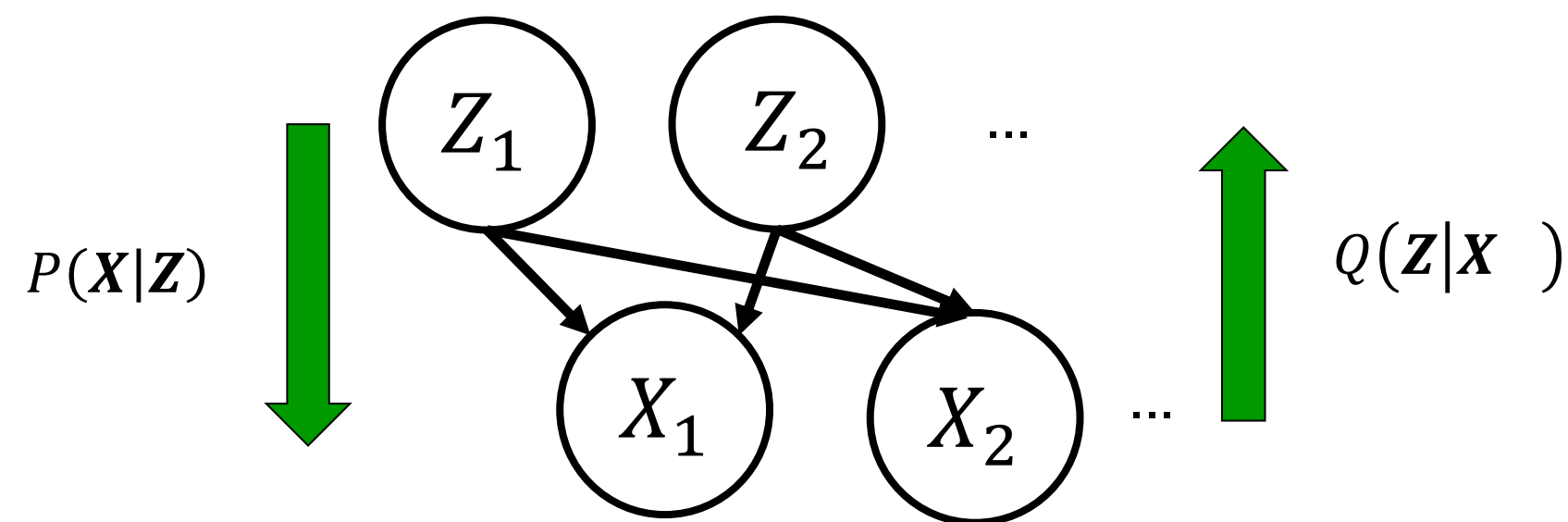
$$\ln P(\mathbf{X}^{(d)}) \geq \sum_{\mathbf{Z}} Q(\mathbf{Z}|\mathbf{X}^{(d)}) \left[\ln \frac{P(\mathbf{X}^{(d)}, \mathbf{Z})}{Q(\mathbf{Z}|\mathbf{X}^{(d)})} \right] \quad \text{が常に成立。等号は } Q(\mathbf{Z}|\mathbf{X}^{(d)}) = P(\mathbf{Z}|\mathbf{X}^{(d)}) \text{ のときのみ}$$

$$\ln P(\mathbf{X}^{(d)}) \geq \sum_{\mathbf{Z}} Q(\mathbf{Z}|\mathbf{X}^{(d)}) \left[\ln \frac{P(\mathbf{X}^{(d)}, \mathbf{Z})}{Q(\mathbf{Z}|\mathbf{X}^{(d)})} \right] = \sum_{\mathbf{Z}} Q(\mathbf{Z}|\mathbf{X}^{(d)}) \ln P(\mathbf{X}^{(d)}, \mathbf{Z}) - \sum_{\mathbf{Z}} Q(\mathbf{Z}|\mathbf{X}^{(d)}) \ln Q(\mathbf{Z}|\mathbf{X}^{(d)})$$

- 右辺をEvidence Lower Bound (ELBO)と呼ぶ
- ELBOはどんな Q についても、対数尤度の下界になっている
- 適当な Q を選べば、ELBOの和は評価できる場合がある(サンプリングを行うなど)
- そこで、 Q を適当にパラメータ付けし、 P のパラメータと同時に、ELBOを最大化するように勾配法(1階微分を使う)で学習を行う
- このような方法を**変分ベイズ法**という
- 変分ベイズ法で学習を行うと、 $Q(\mathbf{Z}|\mathbf{X}^{(d)})$ によって可視変数から隠れ変数を推定することもできる

多くの場合、 Q に**平均場近似**を用いる

$$Q(\mathbf{Z}|\mathbf{X}^{(d)}) = \prod_i q_i(Z_i|\mathbf{X}^{(d)})$$



トピックモデルとして有名なLatent Dirichlet allocation はベイジアンネットワークを平均場近似を用いた変分ベイズ法で学習する

Blei+ 2003

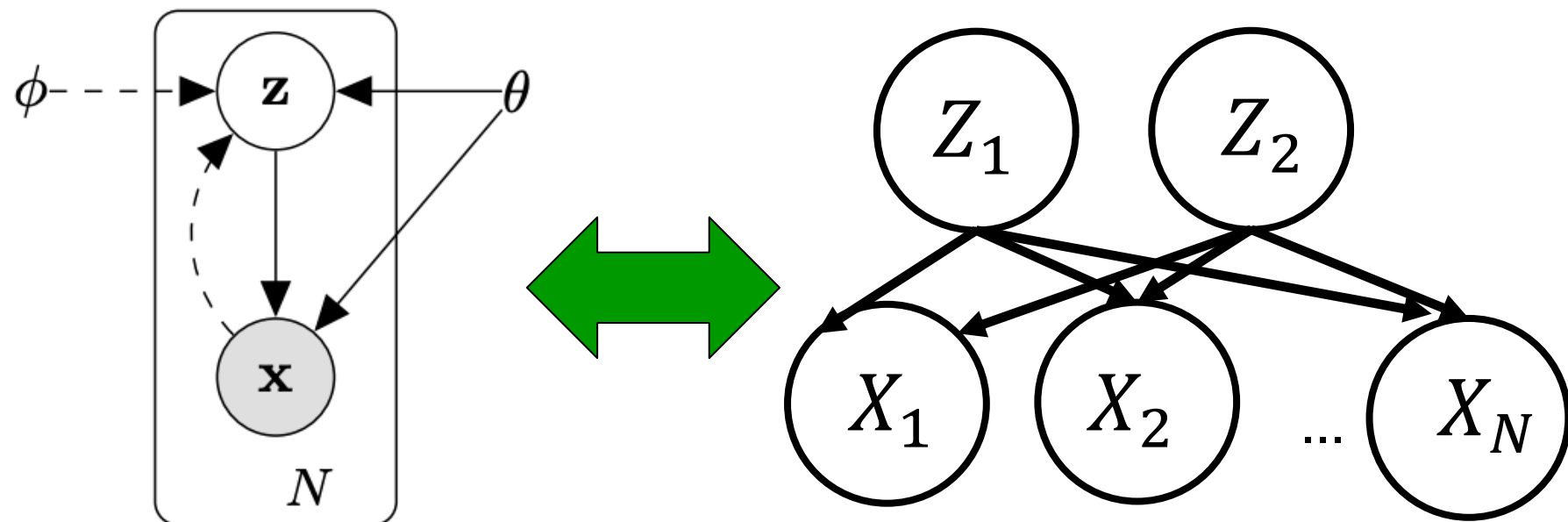
Stochastic Gradient Variational Bayes

- 変分分布は可視変数の関数として複雑なものでも良い
- 複雑で非線形な関数近似器としてニューラルネットワークがある
- 確率的グラフィカルモデル P と変分分布 Q のパラメータをニューラルネットワークの出力で表し、ELBOを最大化するようにニューラルネットワークを訓練する方法がある
 - 変分オートエンコーダ(VAE)
- この学習アルゴリズムはStochastic Gradient Variational Bayes と名付けられた

Kingma & Welling 2014

変分オートエンコーダ

- あまり変分オートエンコーダが確率的グラフィカルモデルであると認識されないが、元論文には書いてある
- 連続な隠れ確率変数への対応や、サンプリングを行った場合でも backpropagation を行えるような工夫など、非常に革新的な論文だった
- 変分オートエンコーダでも変分分布は平均場近似を用いている



可視変数は画素



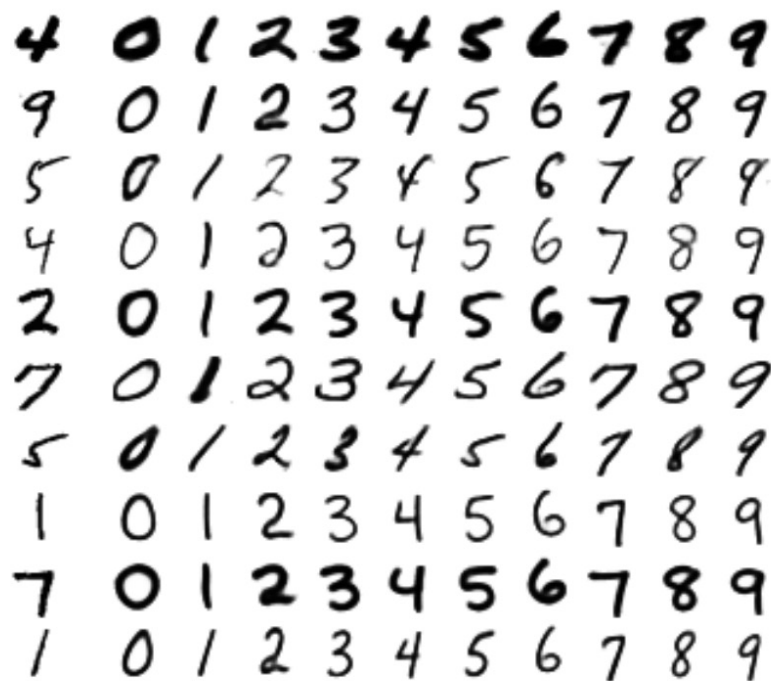
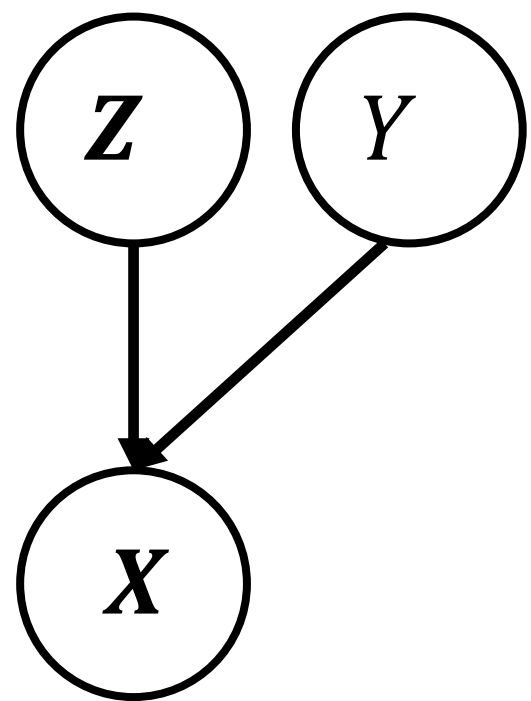
(a) Learned Frey Face manifold

Kingma & Welling 2014

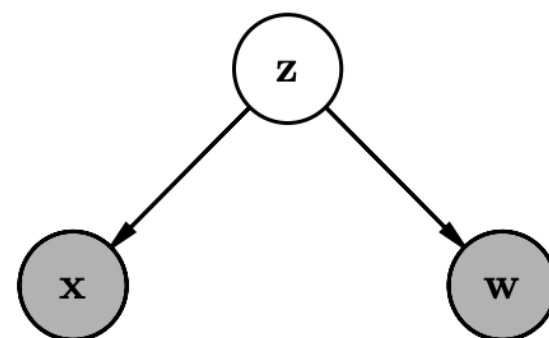
確率変数を少しずつ変えると、生成される可視変数も連続的に変わる

変分オートエンコーダの拡張

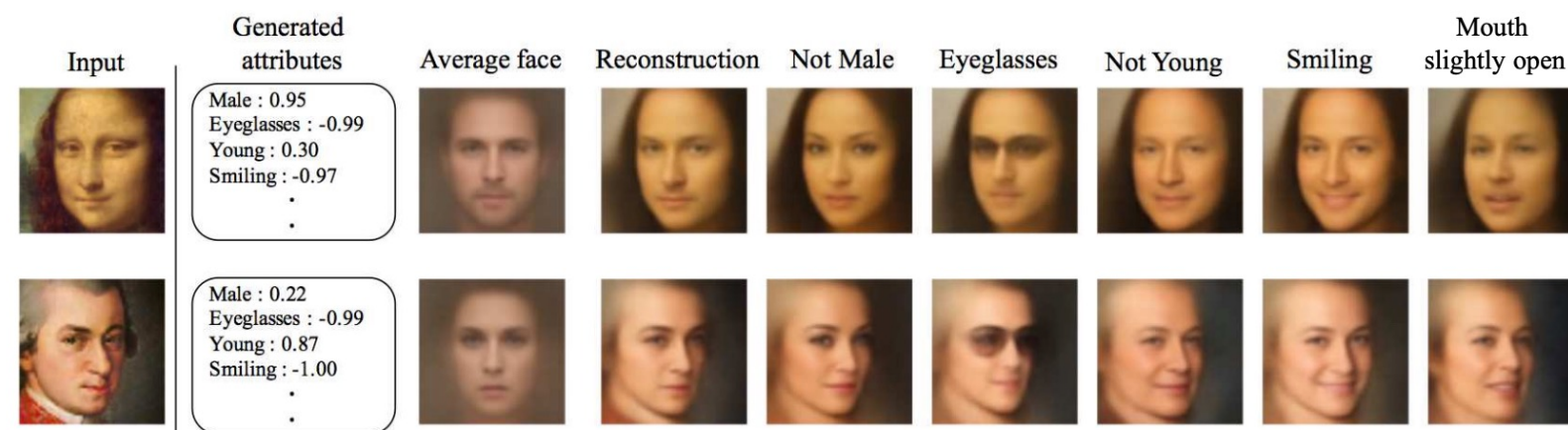
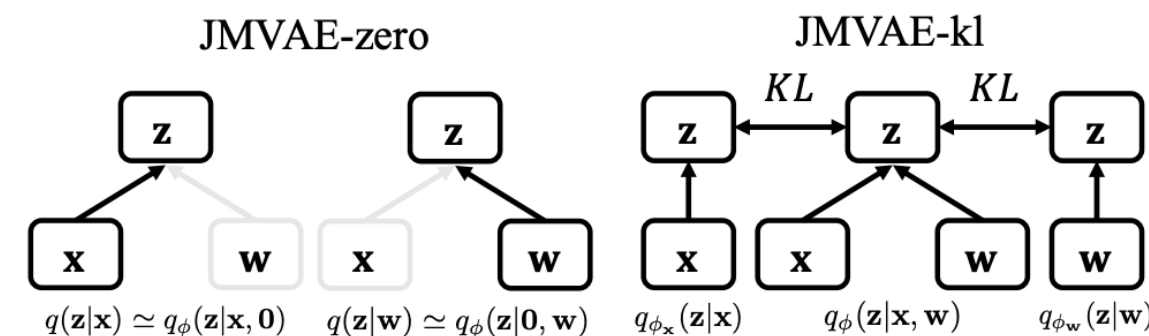
- 変分オートエンコーダが確率的グラフィカルモデルであることを意識すれば、その拡張は色々考えられる
- 例えば条件付き変分オートエンコーダは、隠れ変数 Z とラベル Y が協力して可視変数 X を生成する、というベイジアンネットワークとして理解できる
- Jointly Multimodal VAE(JMVAE)は、隠れ変数 Z から異なるモダリティ(画像とテキスト)が生成されるベイジアンネットワークを用いている



Kingma+ 2014



Suzuki+ 2017



平均場近似で十分なのか？

- ほとんどの変分ベイズ法による学習は、変分分布を平均場近似の形にしている
- 例えばBethe近似の形にしても良いはずである
- そのような研究もなくはないが、あまり流行っていない
- 平均場近似の簡単さ、収束性の良さのためと思われる。今後改善された手法が現れるだろうか？

まとめ

- 確率的グラフィカルモデルは視覚的で分かりやすい生成モデル
- しかし、推論にも学習にもコストがかかる
- 変分近似法として、平均場近似やBethe近似といった統計力学に現れた手法は強力であり、広く使われている
- 一方、このような近似の改善法は、存在はするがあまり応用されていない
- 今後より良い方法が生まれるか？ 理論的な研究に期待

参考文献(教科書)

- Bishop, 「パターン認識と機械学習」
- 西森, 「相転移・臨界現象の統計物理学」
- Koller & Freidman, "Probabilistic Graphical Models"