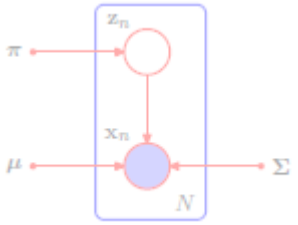
Deterministic (Variational) Approximate Inference

Reference:

Bayesian Reasoning and Machine Learning Ch. 28 (David Barber)
Probabilistic Graphical Model Ch. 11 (Koller & Friedman)
Pattern Recognition & Machine Learning Ch. 10. (Bishop)

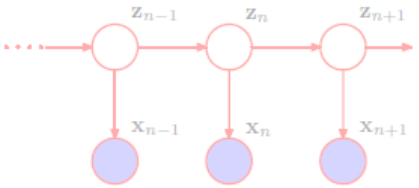
In terms of difficulty, there are 3 types of inference problem.

- Inference which is easily solved with Bayes rule.



- Inference which is tractable using some dynamic programming technique.

  (e.g. Variable Elimination or J-tree algorithm)

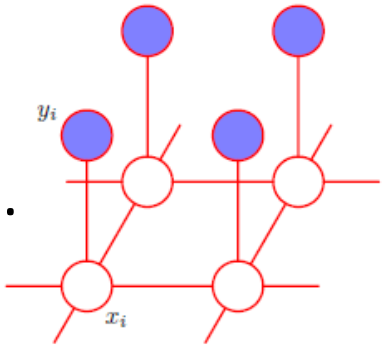


Today's focus

- Inference which is proved intractable

  & should be solved using some Approximate Method.

  (e.g. Approximation with Optimization or Sampling technique.)

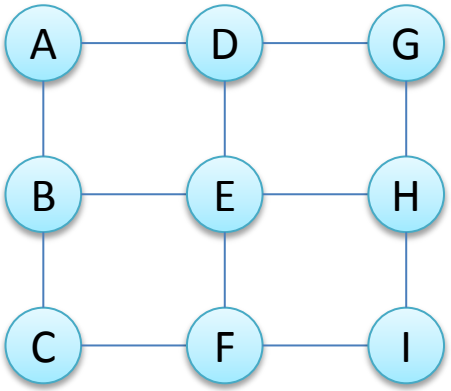


Agenda

- Principle of Variational Approximation
- Global Approximation
  (Mean Field Approximation)
- Message Approximation
  (Expectation Propagation)

Intractable Inference

Example: A N*N Grid MRF (N=3)



What we can solve: $\tilde{P}(X)$: *unnormalized distribution*

$$P(A,\ldots,I)=\frac{1}{Z}\tilde{P}(A,\ldots,I), \qquad \tilde{P}(A,\ldots,I)=\prod_{(X_1,X_2)}\phi(X_1,X_2)$$

(tractable)

What we cannot solve:

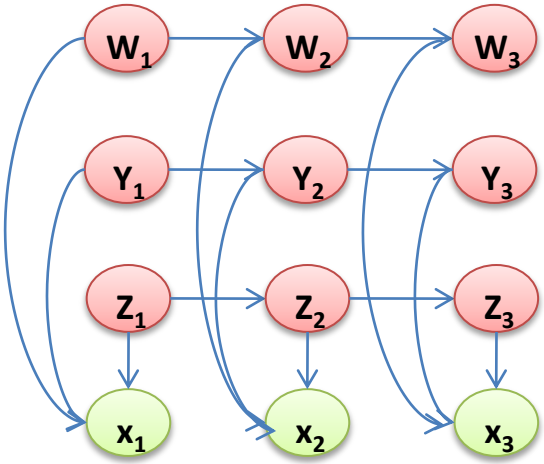
$$Z=\sum_{(A,\ldots,I)}\tilde{P}(A,\ldots,I)$$

$$\tilde{P}(A)=\sum_{(B,\ldots,I)}\tilde{P}(A,\ldots,I)$$

(intractable, as N increases)

Intractable Inference

Example:
N layers Factorial HMM



What we can solve:

$$P(W,Y,Z \mid X=x) = \frac{1}{Z(X=x)} P(W,Y,Z,X=x),$$

$$P(W,Y,Z,X=x) = \underbrace{P(W)*P(Y)*P(Z)*P(X=x \mid W,Y,Z)}_{\text{(easy)}}$$

What we cannot solve: $\tilde{P}(X)$: *unnormalized distribution*

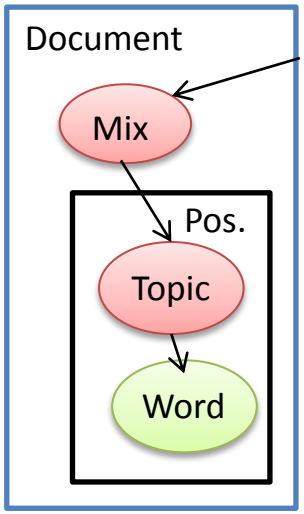
$$P(X=x) = \sum_{W,Y,Z} P(W,Y,Z,X=x) \quad \text{(hard)}$$

$$\tilde{P}(Z_1=z \mid X=x) = \sum_{(W,Y,Z_2 \ldots Z_3)} P(W,Y,Z_1 \ldots Z_3, X=x)$$

Intractable Inference

Example:
Latent Topic Model

Some intractability comes not from "Structure", but from passing message between **different type of distribution**.



$$Let\ \ Mix = \theta = (\theta_1, ..., \theta_K),\ \ K = number\ of\ \ topics$$

$$P(Topic_1 = Z_1 \mid Mix = \theta) = \theta_{Z_1} \qquad M_{Topic1 \to Mix}(\theta) = \sum_{Z_1=1}^{K} \theta_{Z_1} P(w_1 \mid Z_1)$$

$$P(Topic_2 = Z_2 \mid Mix = \theta) = \theta_{Z_2} \qquad M_{Topic2 \to Mix}(\theta) = \sum_{Z_2=1}^{K} \theta_{Z_2} P(w_2 \mid Z_2)$$

No compact representation for message:

$$\tilde{P}(Mix = \theta \mid w) = P(Mix = \theta) * M_{Topic_1 \to Mix}(\theta) * M_{Topic_2 \to Mix}(\theta)$$

$$= \left(\frac{1}{const} \prod_{k=1}^{K} \theta_k^{\alpha - 1} \right) \left(\sum_{Z_1} \theta_{Z_1} P(w_1 \mid Z_1) \right) \left(\sum_{Z_2} \theta_{Z_2} P(w_2 \mid Z_2) \right)$$

Summation is intractable. (Exponential to #variables)

$$\int_{\theta} \tilde{P}(Mix \mid w)\, d\theta = \frac{1}{const} \sum_{Z_1} \sum_{Z_2} \int_{\theta} \left(\prod_{k=1}^{K} \theta_k^{\alpha - 1 + 1[k=Z_1] + 1[k=Z_2]} \right) P(w_1 \mid Z_1) P(w_2 \mid Z_2)\ d\theta$$

Principle of Variational Approximation

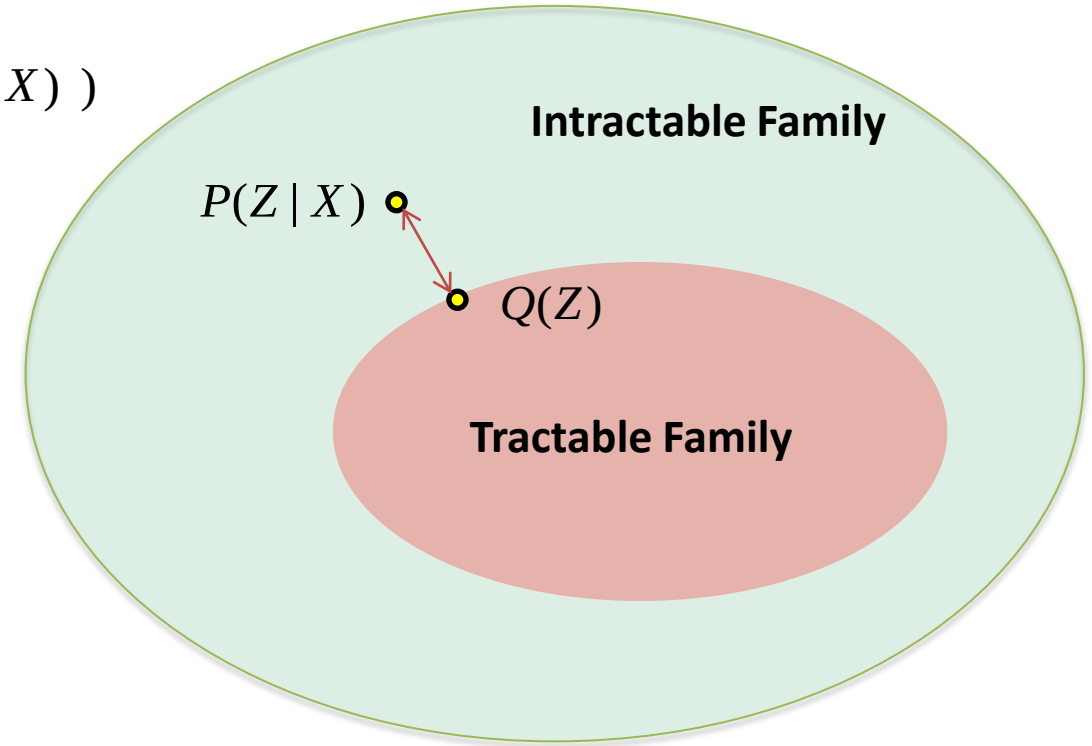
Let **X**: observation, **Z**: hidden variables.

Finds an approximate distribution **Q(Z)** from a "Tractable Family" that most similar to the target distribution **P(Z|X)** measured by some distance like KL divergence.

$$Q^*(Z) = \arg\min_{Q(Z)} \mathrm{KL}(\ P(Z \mid X) \| Q(Z)\ )$$

$$Q^*(Z) = \arg\min_{Q(Z)} \mathrm{KL}(\ Q(Z) \| P(Z \mid X)\ )$$

How can we optimize Q(Z) without computing P(Z|X) ?



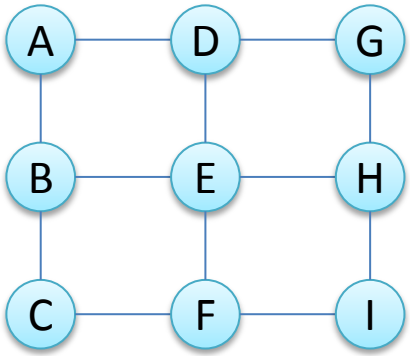
Agenda

- Principle of Variational Approximation
- Global Approximation
  (Mean Field Approximation)
- Message Approximation
  (Expectation Propagation)

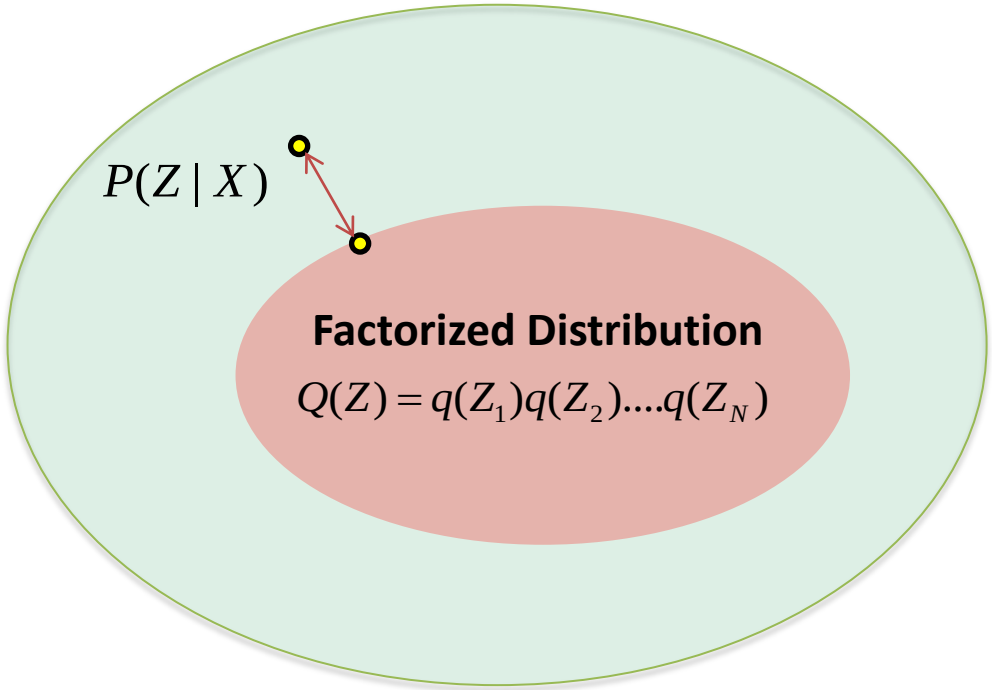
Global Approximation

One of the most popular tractable family is **Factorized Distribution**, which assumes the target (posterior) distribution **P(Z|X)** can be factorized into **$q(Z_1)*q(Z_2)…*q(Z_N)$**, that is, variables are independent to each other.

Example:



$$P(A,...,I) = \frac{1}{Z}\tilde{P}(A,...,I)$$
$$\approx q(A)q(B).....q(I)$$



How to Find $Q^*(Z)$?

$$Q^*(Z) = \underset{Q(Z) \in TractableFamily}{\arg\min} \; \mathrm{KL}(\; Q(Z) \,\|\, P(Z \mid X) \;)$$

$$\mathrm{KL}(\; Q(Z) \,\|\, P(Z \mid X) \;) = E_{Q(Z)}[\log \frac{1}{P(Z \mid X)} - \log \frac{1}{Q(Z)}]$$

$$= E_{Q(Z)}[\log Q(Z) - \log P(Z \mid X)]$$

$$= E_{Q(Z)}[\log Q(Z) - \log P(Z, X) + \log P(X)]$$

$$= E_{Q(Z)}[\log Q(Z)] - E_{Q(Z)}[\log P(Z, X)] + \log P(X)$$

(Tractable if Q(Z) is tractable)

(intractable….
but Independent of Q(Z))

The resulting problem is equivalent to:

$$Q^*(Z) = \underset{Q(Z) \in TractableFamily}{\arg\max} \; \mathrm{E}_{Q(Z)}[\log \frac{\mathrm{P(Z,X)}}{\mathrm{Q(Z)}}]$$

Find **Q(Z)** that put "similar weight" to **P(Z,X)** on which **Z=z** to happen.

How to Find $Q^*(Z)$?

$$Q^*(Z) = \underset{Q(Z)=q(Z_1)q(Z_2)\ldots q(Z_N)}{\arg\max} \; \mathrm{E}_{Q(Z)}[\log \frac{\mathrm{P}(Z,X)}{\mathrm{Q}(Z)}]$$

Find **Q(Z)** that put "similar weight" to **P(Z,X)** on which **Z=z** to happen.

We maximize w.r.t. one $q(Z_n)$, while fixing all the other.

$$\max_{q(Z_1)} \; \mathrm{E}_{Q(Z)}[\log P(Z,X)] - \mathrm{E}_{Q(Z)}[\log Q(Z)]$$

$$= \mathrm{E}_{q(Z_1)}\left[\mathrm{E}_{q(Z_2)\ldots q(Z_N)}[\log P(Z,X)] \right] - \mathrm{E}_{q(Z_1)}[\log q(Z_1)] - \sum_{k\neq 1} \mathrm{E}_{q(Z_k)}[\log q(Z_k)]$$

Expectation over other variables denote as $\log \hat{P}(Z_1, X) + const.$

Independent to $q(Z_1)$

$$= \mathrm{E}_{q(Z_1)}\left[\log \frac{\hat{\mathrm{P}}(Z_1,X)}{q(Z_1)} \right] + \mathrm{const.}$$

$$-KL(\, q(Z_1) \,\|\, \hat{P}(Z_1, X)\,)$$

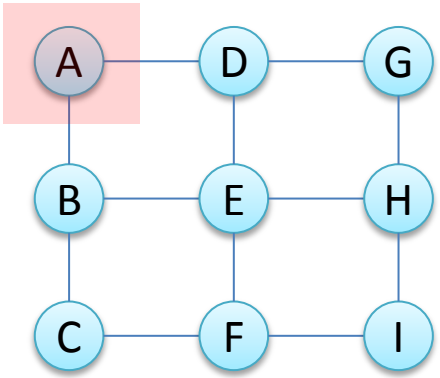
$$q^*(Z_1) = \hat{P}(Z_1, X) \;\Rightarrow\; \log q^*(Z_1) = \log \hat{P}(Z_1, X)$$

$$\log q^*(Z_1) = E_{q(Z_2)\ldots q(Z_N)}[\log P(Z, X)] + const.$$

$$= \sum_{\substack{f \in factors \\ related\ to\ Z_1}} E[\ \log f(Z_{k_1} \ldots Z_{k_m})\ ] + const.$$

How to Find $Q^*(Z)$?

Example:



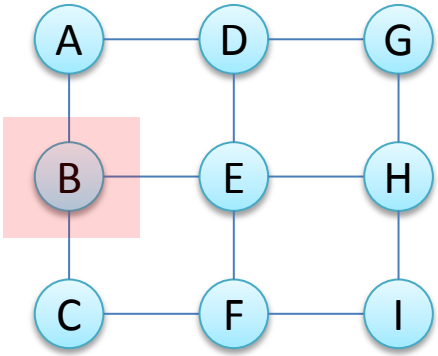
$$P(A,\ldots,I) = \frac{1}{Z}\tilde{P}(A,\ldots,I)$$
$$\approx q(A)q(B).....q(I)$$

Given other **q(B)…q(I)** fixed, maximize w.r.t. **q(A)**:

$$\log \tilde{q}^*(A) = E_{q(B)\ldots q(I)}[\log \tilde{P}(A,\ldots,I)]$$
$$= \mathrm{E}_{q(B)}[\log\phi(A,B)] + \mathrm{E}_{q(D)}[\log\phi(A,D)] + \text{const.}$$

How to Find $Q^*(Z)$?

Example:



Given other **q(B)…q(I)** fixed, maximize w.r.t. **q(A)**:

$$\log \tilde{q}^*(A) = E_{q(B)\ldots q(I)}[\log \tilde{P}(A,\ldots,I)]$$
$$= \mathrm{E}_{q(B)}[\log \phi(A,B)] + \mathrm{E}_{q(D)}[\log \phi(A,D)]$$

$$\log \tilde{q}^*(B) = E_{q(A)q(C)\ldots q(I)}[\log \tilde{P}(A,\ldots,I)]$$
$$= \mathrm{E}_{q(A)}[\log \phi(A,B)] + \mathrm{E}_{q(C)}[\log \phi(B,C)] + \mathrm{E}_{q(E)}[\log \phi(B,E)] + \text{const.}$$

$$P(A,\ldots,I) = \frac{1}{Z}\tilde{P}(A,\ldots,I)$$
$$\approx q(A)q(B)\ldots\ldots q(I)$$

Iterate over all variables ….. until convergence !!

Guarantee convergence to stationary point of $\max_{Q(Z)} \mathrm{E}_{Q(Z)}[\log \frac{P(Z,X)}{Q(Z)}]$ **(Why?)**

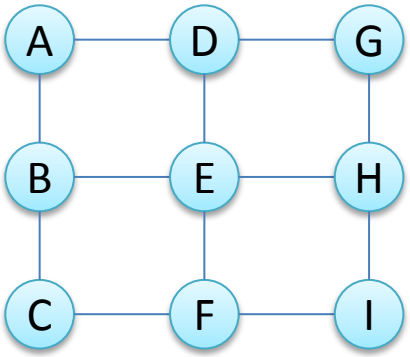
(Every update **strictly increase** objective function, since **KL(q || p)=0 only if $q(z_k)=p(z_k)$**.
Since the maximum is bounded, we are guaranteed to convergence.)

Agenda

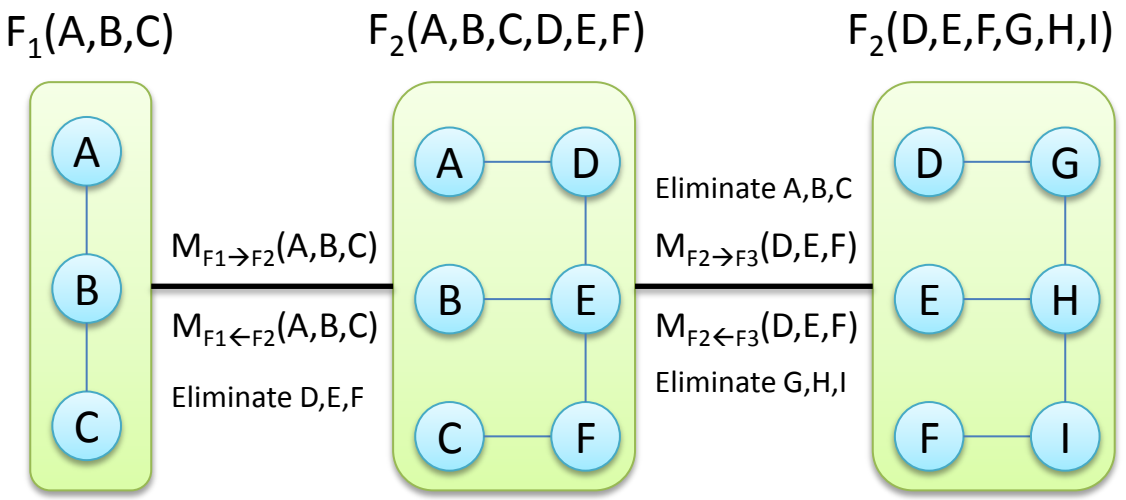
- Principle of Variational Approximation
- Global Approximation
  (Mean Field Approximation)
- Message Approximation
  (Expectation Propagation)

Message Approximation

Example: A N*N Grid MRF (N=3)



Variable Elimination ➔ Clique Tree



The Elimination:

$$M_{F_2 \to F_3}(D,E,F) = \sum_{A,B,C} M_{F_1 \to F_2}(A,B,C) F(A,B,C,D,E,F)$$

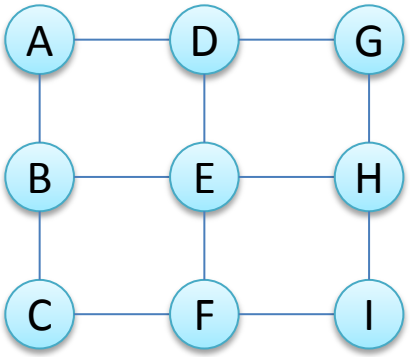
is intractable. (exponential in N)

However, can we approximate the message $\mathbf{M_{Fi \to Fj}(\ldots)}$ to make the elimination tractable ?

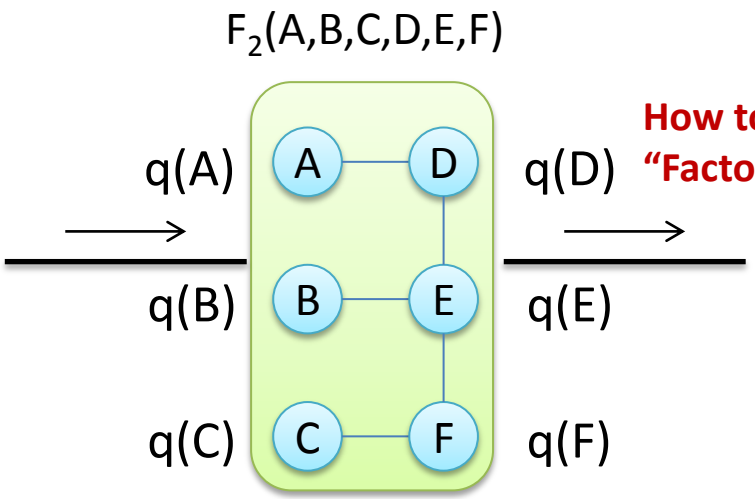
➔ **Assume it is factorized !!**

Message Approximation

Example: A N*N Grid MRF (N=3)



Variable Elimination ➔ Clique Tree



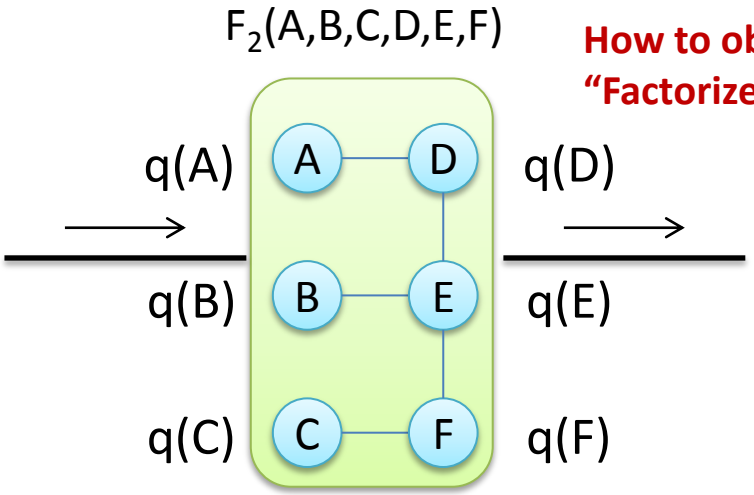
Approximate the message by a factorized distribution:

$$M_{F_1 \to F_2}(A,B,C) = q(A)q(B)q(C)$$

A, B, C not entangled !! $q(A)q(B)q(C)F_2(A,B,C,D,E,F)$ forms a tree.

➔ We can compute marginal by sum-product algorithm !!

How to obtain a Factorized Message ?



How to obtain the next "Factorized Message" ?

$$\min_{q(D)q(E)q(F)} \mathrm{KL}(\ q(D)q(E)q(F) \,\|\, M_{F_2 \to F_3}(\mathrm{D},\mathrm{E},\mathrm{F})\ ) \quad \textbf{?}$$

➔Need iteratively solve for **q(D), q(E), q(F)**.
(Is there more efficient way ?)

$$\min_{q(D)q(E)q(F)} \mathrm{KL}(\ M_{F_2 \to F_3}(\mathrm{D},\mathrm{E},\mathrm{F}) \,\|\, q(D)q(E)q(F)\ ) \quad \textbf{?}$$

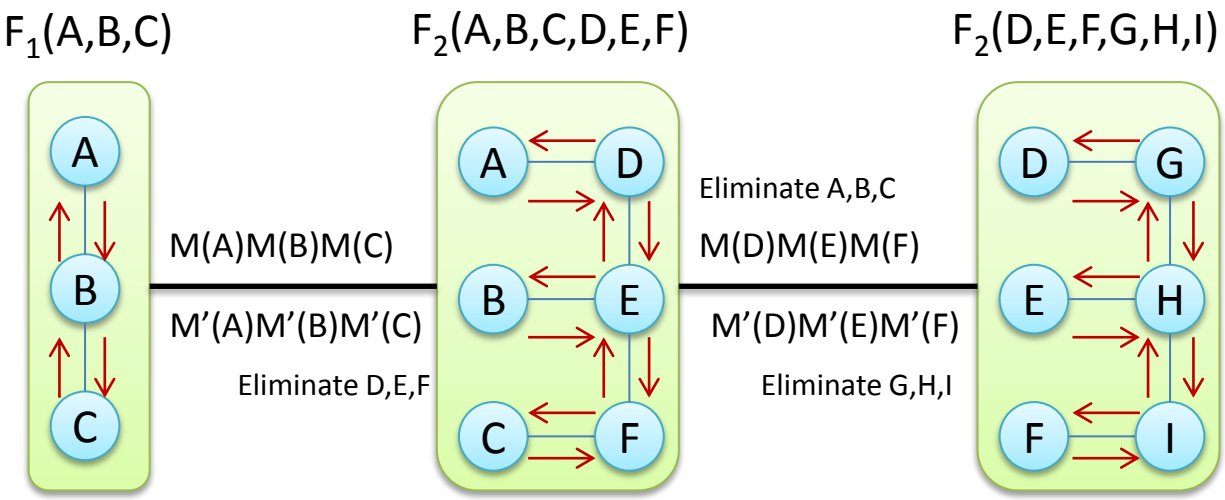
$$\mathrm{KL}(\ M(\mathrm{D},\mathrm{E},\mathrm{F}) \,\|\, q(D)q(E)q(F)\ ) = E_{M(\mathrm{D},\mathrm{E},\mathrm{F})}(\log \frac{M(D,E,F)}{q(D)q(E)q(F)})$$

$$= E_{M(\mathrm{D},\mathrm{E},\mathrm{F})}(\log \frac{M(D,E,F)}{M(D)M(E)M(F)}) + E_{M(D)}[\ \log \frac{M(D)}{q(D)}\ ] + E_{M(E)}[\ \log \frac{M(E)}{q(E)}\ ] + E_{M(F)}[\ \log \frac{M(F)}{q(F)}\ ]$$

$$= \underbrace{KL(M(D,E,F) \,\|\, M(D)M(E)M(F))}_{\textbf{const.}} + \underbrace{\sum_{X \in \{D,E,F\}} KL(M(X) \,\|\, q(X))}_{\textbf{Set q*(X) = M(X)}}$$

(set q(D), q(E), q(F) equal to the marginal.)

2-Layers Sum-Product Algorithm with Approximate Messages



Elimination is easy since **factors in every Clique form a "Tree"**.
Computing Marginal (ex. M(D), M(E), M(F)) can be done by **inner Sum-Product Algorithm**.

Approximate Message: Expectation Propagation

Previous example is a special case of "**Expectation Propagation**". General Expectation Propagation uses distribution come from **Log-linear model** (including **Gaussian**, **Multinomial**, **Poisson**, **Dirichlet** Distribution):

$$Q_\theta(\mathrm{X}) = \frac{1}{Z(\theta)} \exp\left\{\theta^T \mathrm{f}(\mathrm{X})\right\} \qquad Z(\theta) = \sum_X \exp\left\{\theta^T \mathrm{f}(\mathrm{X})\right\} \qquad \frac{\partial}{\partial\theta} \log Z(\theta) = E_{Q_\theta(\mathrm{X})}[f(X)]$$

where $f(X)^T = [f_1(X), f_2(X), \ldots, f_D(X)]^T$ are sufficient statistics(*features*) *derived from* X.

$$\min_\theta \ \mathrm{KL}(P(\mathrm{X}) \| Q_\theta(\mathrm{X})) = \underbrace{E_{P(\mathrm{X})}[\log P(X)]}_{\text{const.}} - E_{P(\mathrm{X})}[\log Q_\theta(X)]$$

$$\max_\theta \ E_{P(\mathrm{X})}[\log Q_\theta(X)] = E_{P(\mathrm{X})}[\theta^T f(X)] - \log Z(\theta)$$

$$\frac{\partial}{\partial\theta} E_{P(\mathrm{X})}[\theta^T f(X)] - \frac{\partial}{\partial\theta} \log Z(\theta) = 0 \quad \Rightarrow \quad E_{P(\mathrm{X})}[f(X)] = E_{Q_\theta(\mathrm{X})}[f(X)]$$

Moment Matching!!

Match the Expectation of feature to the original message.

Approximate Message: Expectation Propagation

Previous example is a special case of "**Expectation Propagation**". A more general version uses distribution come from **Log-linear model** (including **Gaussian**, **Multinomial**, **Poisson**, **Dirichlet** Distribution):

Moment Matching: $E_{P(X)}[f(X)] = E_{Q_\theta(X)}[f(X)]$

Example:

1. Q(X) is Multi(θ): $f_k(X) = 1[X = k]$

$$E_{Q_\theta(X)}[1[X = k]] = Q_\theta(X = k)$$

Moment Matching:
Set equal Marginal Probability

$$Q_\theta(X = k) = \theta_k = P(X = k)$$

As previous MRF example.

2. Q(X) is Gaussian(μ,Σ): $f_1(X) = X \quad f_2(X) = XX^T$

$$E_{Q_\theta(X)}[X] = \mu$$

$$E_{Q_\theta(X)}[XX^T] = \Sigma + \mu\mu^T$$

Moment Matching:
Set equal Mean, Variance.

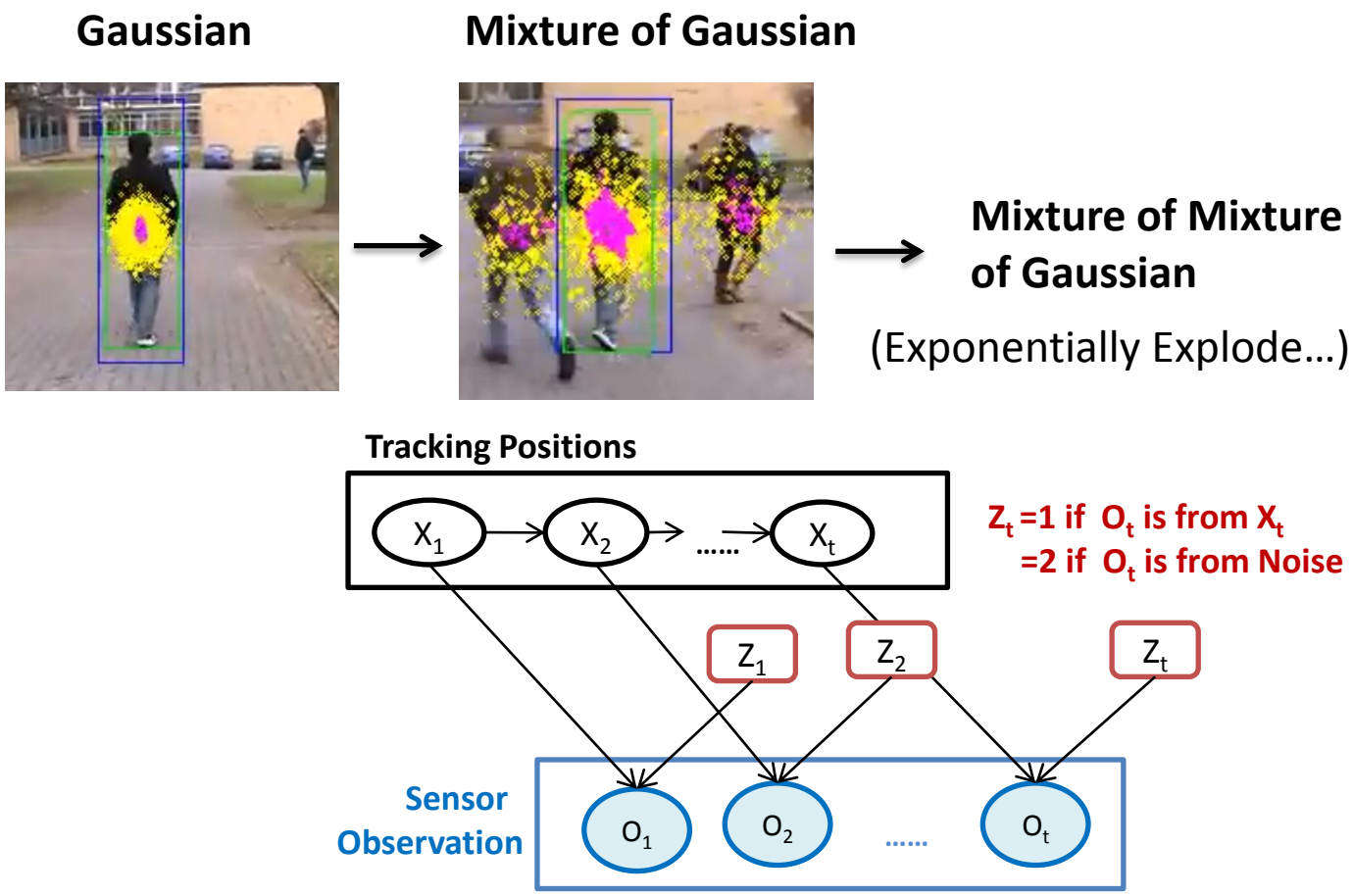
$$\mu = E_{P(X)}[X]$$

$$\Sigma = E_{P(X)}[XX^T] - \mu\mu^T$$

$$= Var_{P(X)}[X]$$

Example:
Use EP Handling Continuous / Discrete BN

When BN contains both Discrete / Continuous Variables, messages cannot have a compact representation…….



Example:
Use EP Handling Continuous / Discrete BN

When BN contains both Discrete / Continuous Variables, messages cannot have a compact distribution…….

To prevent message grows to exponentially many mixtures of Gaussian….

$$M_{\rightarrow}(X_1) = \sum_{Z_1} P(Z_1)P(X_1)P(O_1 | Z_1, X_1)$$

$$\frac{N(X; \mu_{Z1}, \Sigma_{Z1})}{w_1 * N(X; \mu_1, \Sigma_1) + w_2 * N(X; \mu_2, \Sigma_2)}$$

Approximate $M_{\rightarrow}(X_1)$ with
single Gaussian $Q(X_1)$ by "**Expectation Matching**":

$$\mu_Q = E_{M_{\rightarrow}(X_1)}[X_1] = w_1\mu_1 + w_2\mu_2$$

$$\Sigma_Q = Var_{M_{\rightarrow}(X_1)}[X_1]$$

$$= E_{P(Z_1)}[\ Var[X_1] \mid Z_1] + Var_{P(Z_1)}[\ E[X_1] \mid Z_1]$$

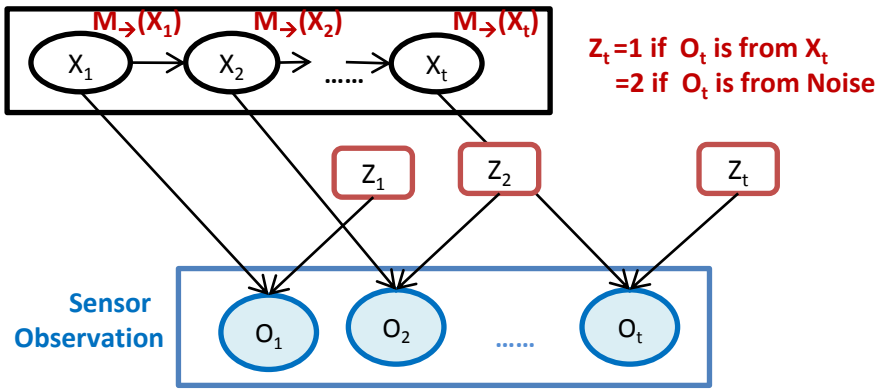
$$= E_{P(Z_1)}[\ \Sigma_{Z_1} \mid Z_1] + Var_{P(Z_1)}[\ \mu_{Z_1} \mid Z_1]$$

$$= w_1\Sigma_1 + w_1\Sigma_2$$

$$+ w_1(\mu_1 - \mu_Q)(\mu_1 - \mu_Q)^T$$

$$+ w_2(\mu_2 - \mu_Q)(\mu_2 - \mu_Q)^T$$

Approximate Message with $N(\mu_{Qt}, \Sigma_{Qt})$

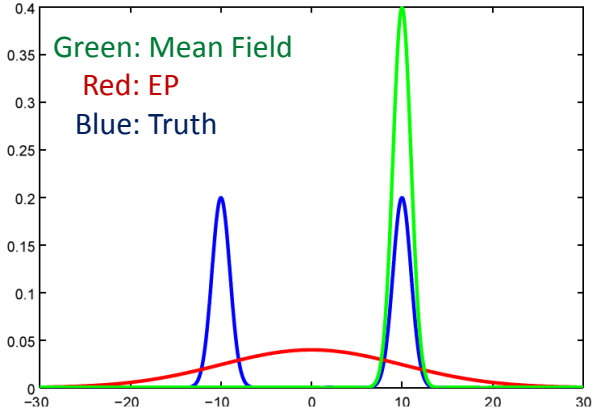


Agenda

- Principle of Variational Approximation
- Global Approximation
  (Mean Field Approximation)
- Message Approximation
  (Expectation Propagation)
- Comparison

Mean Field Approximation vs. Expectation Propagation

- Both of them find a tractable distribution (ex. Factorized distribution) Q(Z) to approximate the real distribution.
- Mean Field approximate joint posterior distribution P(Z|X), minimizing KL(Q||P). (Why not KL(P||Q) ? [1])
- Expectation Propagation approximate messages, minimizing KL(P||Q). (Why not KL(Q||P) ? [2])
- Expectation Propagation needs only one-pass Sum-Product, while Mean Field Approximation needs iterative maximization.
- min KL(Q||P) has more False Negative. (Why [3])
- min KL(P||Q) has more False Positive. (Why [4])



$$\frac{\partial}{\partial\theta}\log Z(\theta) = E_{Q_\theta(X)}[f(X)]$$

$$\frac{\partial}{\partial\theta}\log Z(\theta) = \frac{1}{Z(\theta)}\frac{\partial}{\partial\theta}\sum_X \exp\left\{\theta^T \mathrm{f(X)}\right\}$$

$$= \frac{1}{Z(\theta)}\sum_X \exp\left\{\theta^T \mathrm{f(X)}\right\} * \mathrm{f(X)} = \sum_X Q_\theta(X) f(X) = E_{Q_\theta(X)}[f(X)]$$

Back