Variational Message Passing

By John Winn, Christopher M. Bishop

Presented by Andy Miller

Overview

- Background
  - Variational Inference
  - Conjugate-Exponential Models
- Variational Message Passing
  - Messages
  - Univariate Gaussian Example
  - Allowable Models and Constraints
- VIBES and Extensions

Variational Inference

- Our model (directed graphical model)
  - $\boldsymbol{X} = (\boldsymbol{V}, \boldsymbol{H})$
  - $\boldsymbol{V}$ are visible variables
  - $\boldsymbol{H}$ are latent variables – includes parameters
- Our dilemma
  - Exact inference algorithms are ‘computationally intractable for all but the simplest models.’
- Our goal
  - Find tractable approximate: $Q(\boldsymbol{H}) \approx P(\boldsymbol{H} \mid \boldsymbol{V})$

Variational Inference

- Note the natural decomposition of log likelihood:

$$\ln P(\mathbf{V}) = \ell(Q) + KL(Q \,||\, P)$$

where

$$\ell(Q) = \sum_{\mathbf{H}} Q(\mathbf{H}) \ln \frac{P(\mathbf{H},\mathbf{V})}{Q(\mathbf{H})}$$

and

$$KL(Q \,||\, P) = -\sum_{\mathbf{H}} Q(\mathbf{H}) \ln \frac{P(\mathbf{H} \,|\, \mathbf{V})}{Q(\mathbf{H})}$$

Variational Inference

- For arbitrary *Q*

$$\underset{\substack{\Uparrow\\ \text{fixed}}}{\ln P(\mathbf{V})} = \underset{\substack{\Uparrow\\ \text{maximize}}}{\ell(Q)} + \underset{\substack{\Uparrow\\ \text{minimize}}}{\mathrm{KL}(Q \,||\, P)}$$

- Minimize KL(Q || P) w.r.t unrestricted Q? We get:

$$Q(\mathbf{H}) = P(\mathbf{H} \mid \mathbf{V})$$

  - but this is what we're trying to avoid…

Variational Inference

- Family of distributions explored by Winn:

$$Q(\mathbf{H}) = \prod_i Q_i(\mathbf{H}_i)$$

  where {**H**$_i$} are disjoint groups of latent variables

- Vastly reduces space
  - e.g. assuming a fully disjoint set of discrete variables:

$$|\mathbf{H}| = N, \;\; H_i \in \{1,..,K\}$$

  - Q reduces P space:

$$K^N \rightarrow KN$$

Variational Inference

- Plug in factorized Q to lower bound equation:

$$\ell(Q) = \sum_{\mathbf{H}} \prod_i Q_i(\mathbf{H}_i) \ln P(\mathbf{H},\mathbf{V}) - \sum_i \sum_{\mathbf{H}_i} Q_i(\mathbf{H}_i) \ln Q_i(\mathbf{H}_i)$$

- Separate out all terms in one factor Q_j

$$\ell(Q) = -KL(Q_j \,||\, Q_j^*) + \text{ terms not in } Q_j$$

- Introduce some new distribution Q^*_j
- Minimize this KL divergence

Variational Inference

- Maximizing the lower bound w.r.t. some factor Q_j:

$$\ln Q_j^*(\mathbf{H}_j) = \langle \ln P(\mathbf{H},\mathbf{V}) \rangle_{\sim Q(\mathbf{H}_j)} + \text{const.}$$

$$\Rightarrow$$

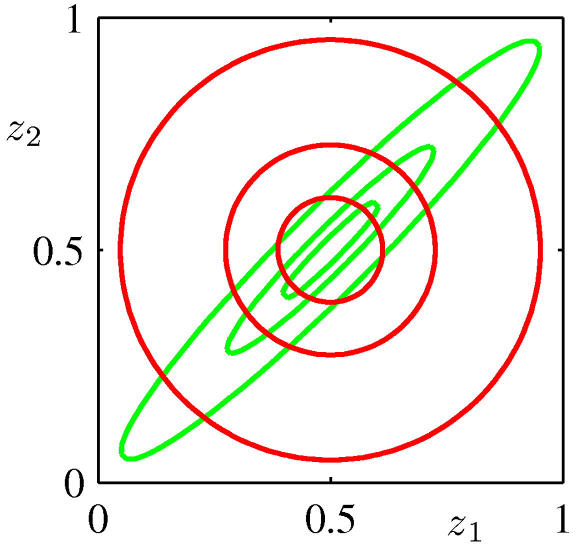
$$Q_j^*(\mathbf{H}_j) = \frac{1}{Z} \exp\left(\langle P(\mathbf{H},\mathbf{V}) \rangle_{\sim Q(\mathbf{H}_j)} \right)$$

- Can see that solutions are coupled, each Q_j depends on expectations w.r.t. factors $Q_{i \neq j}$
- Variational optimization proceeds by initializing each Q_i and then cycling through each factor

Variational Inference (recap)

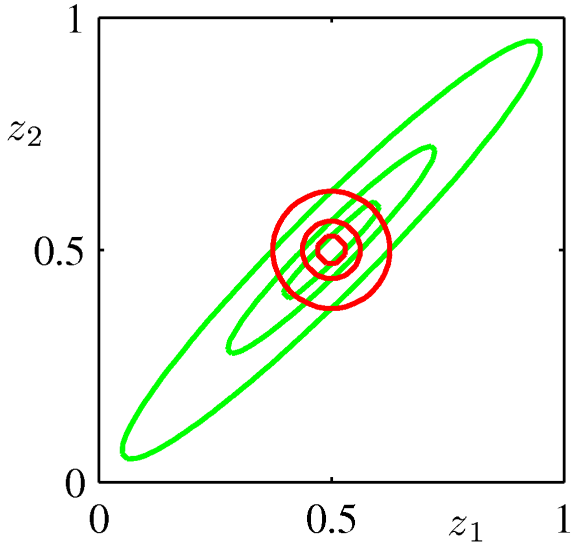
1. Choose a family, $Q(\mathbf{H})$ of variational distributions:
2. Use Kullback-Leibler divergence, $\mathrm{KL}(Q\,||\,P)$, as a measure of 'distance' between $P(H\,|\,V)$ and $Q(H)$.
3. Find Q that minimizes divergence (or equivalently, maximizes the lower bound).

KL Divergence



(b)

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

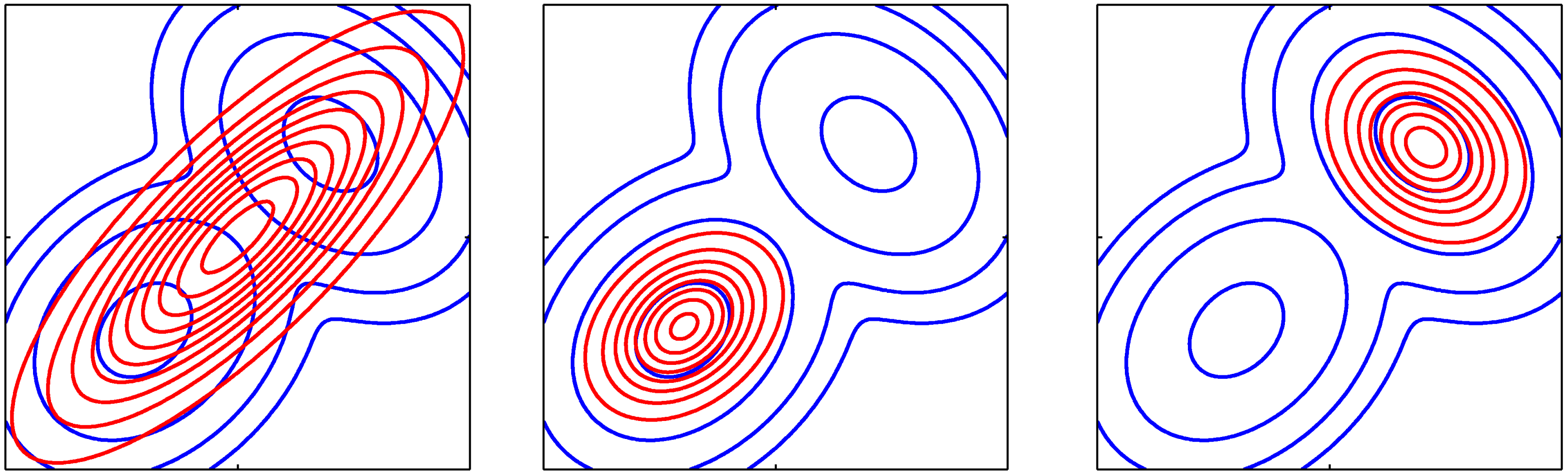


(a)

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

Figures from Pattern Recognition and Machine Learning. Bishop, 2006.

KL Divergence



$\min KL(p \| q)$

$\min KL(q \| p)$

Figures from <u>Pattern Recognition and Machine Learning</u>. Bishop, 2006.

Variational Inference in Directed Model

- Assuming a directed graphical model, full distribution:

$$P(\mathbf{X}) = \prod_i P(X_i \mid pa_i)$$

- Winn assumes fully factorized *Q*

$$Q(\mathbf{H}) = \prod_i Q_i(H_i)$$

Variational Inference in Directed Model

- Plugging factorized joint into optimized form of factor j:

$$\ln Q_j^* = \left\langle \sum_i \ln P(X_i \mid pa_i) \right\rangle_{\sim Q(H_j)} + \text{ const.}$$

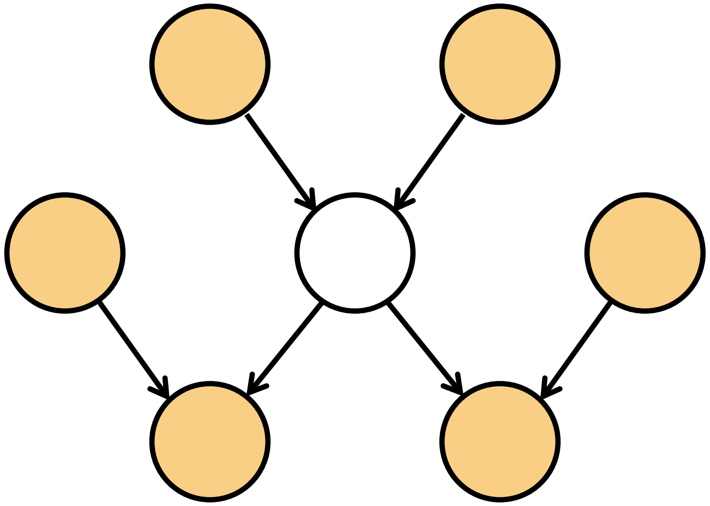
- Terms not depending on H_j will be constant, yielding:

$$\ln Q_j^*(H_j) = \left\langle \ln P(H_j \mid pa_j) \right\rangle_{\sim Q(H_j)} + \sum_{k \in ch_j} \left\langle \ln P(X_k \mid pa_k) \right\rangle_{\sim Q(H_j)} + \text{const.}$$

- Distribution only relies on parents, children and co-parents

Variational Inference in Directed Model

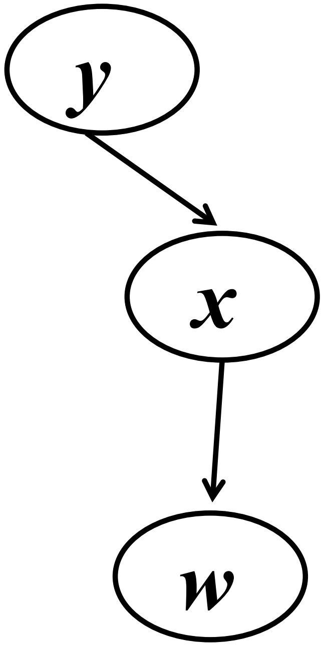
- For a factorized Q, each update equation relies only on variables in the Markov blanket



- Can decompose the overall optimization into a set of local computations

Conjugate-Exponential Models

- Simplify update equations
  - conditional distributions from the exponential family
  - conjugate w.r.t. distributions over parent variables



A parent distribution *p(x|y)* is said to be *conjugate to* child distribution *p(w|x)* if *p(x|y)* has the same functional form, with respect to *x*, as *p(w|x)*.

$$p(\underset{\uparrow}{x} \mid w,y) \propto \underset{\uparrow}{p}(w \mid x)\, \underset{\uparrow\uparrow}{p}(x \mid y)$$

same family

same functional form

Conjugate-Exponential Models

- Conditional distributions expressed in exponential family form:

$$\ln P(X \mid \boldsymbol{\theta}) = \underset{\Uparrow}{\boldsymbol{\theta}}^{\mathrm{T}} \underset{\Uparrow}{\boldsymbol{u}}(X) + g(\boldsymbol{\theta}) + f(X)$$

'natural' parameter vector

sufficient statistics vector

- E.g. univariate Gausian:

$$\ln P(X \mid \mu, \gamma) = \begin{bmatrix} \mu\gamma \\ -\gamma/2 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} X \\ X^2 \end{bmatrix} + \tfrac{1}{2} \ln \frac{\gamma}{2\pi} - \tfrac{1}{2} \gamma \mu^2 + 0$$

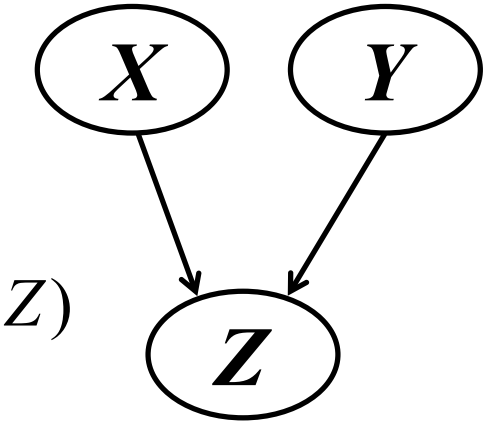
Conjugate-Exponential Models

- Parents and children are chosen to be conjugate, i.e. have the same functional form

$$\ln P(X \mid \boldsymbol{\theta}) = \boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{u}(X) + g(\boldsymbol{\theta}) + f(X)$$

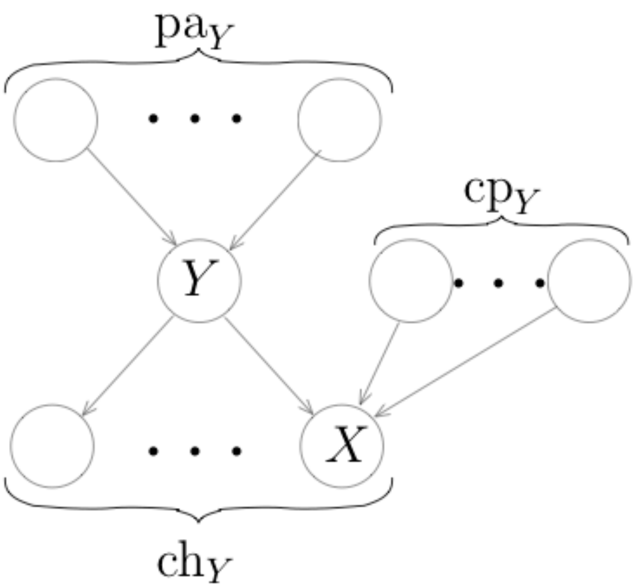
$$\updownarrow \text{ same}$$

$$\ln P(Z \mid X, Y) = \boldsymbol{\varphi}(Y, Z)^{\mathrm{T}} \boldsymbol{u}(X) + g'(X) + f'(Y, Z)$$



- E.g.
  - Gaussian for the mean of a Gaussian
  - Gamma for the precision of a Gaussian
  - Dirichlet for the parameters of a discrete distribution

Conjugate-Exponential Models

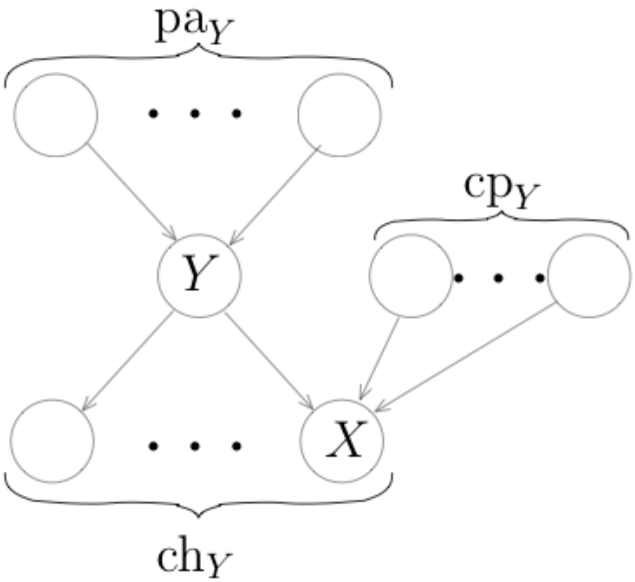


$$P(Y \mid X, pa_Y) = P(X \mid Y, cp_Y) P(Y \mid pa_Y)$$

same family

same form (wrt Y)

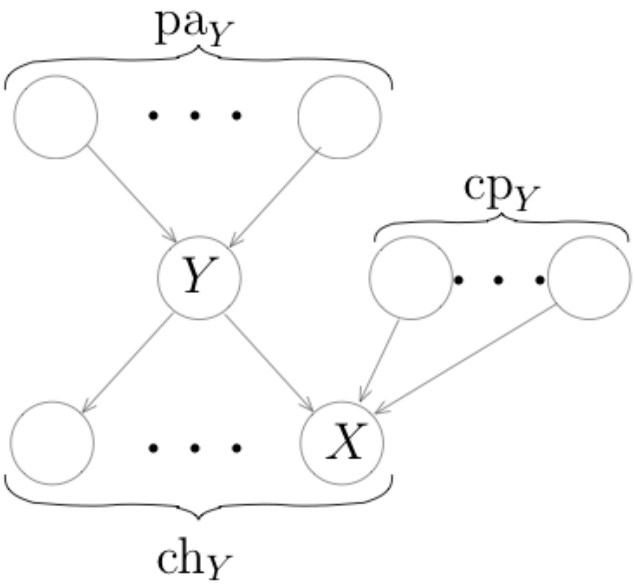
Conjugate-Exponential Models



$$\ln P(Y \mid pa_Y) = \phi_Y(pa_Y)^{\mathrm{T}}\mathbf{u}_Y(Y) + f_Y(Y) + g_Y(pa_Y)$$

$$\begin{aligned}\ln P(X \mid Y, cp_Y) &= \phi_X(Y, cp_Y)^{\mathrm{T}}\mathbf{u}_X(X) + f_X(X) + g_X(Y, cp_Y) \\ &= \phi_{XY}(X, cp_Y)^{\mathrm{T}}\mathbf{u}_Y(Y) + \lambda(X, cp_Y)\end{aligned}$$

Conjugate-Exponential Models

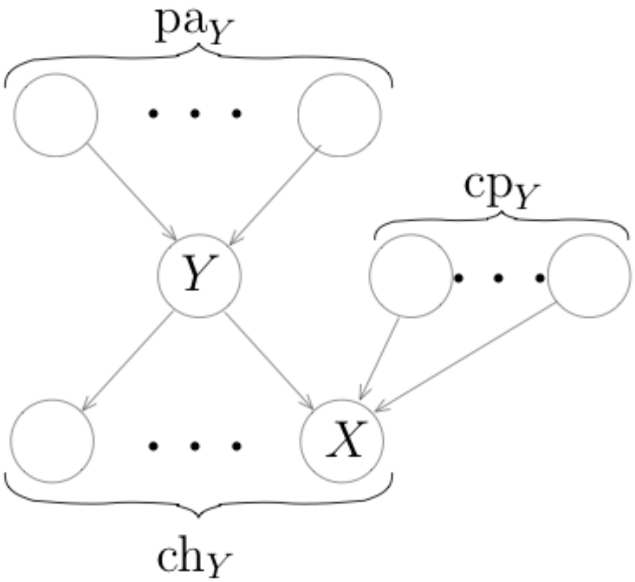


$$\ln Q_Y^*(Y) = \left\langle \phi_Y(pa_Y)^{\mathrm{T}} \mathbf{u}_Y(Y) + f_Y(Y) + g_Y(pa_Y) \right\rangle_{\sim Q(Y)}$$

$$+ \sum_{k \in ch_j} \left\langle \phi_{XY}(X_k, cp_k)^{\mathrm{T}} \mathbf{u}_Y(Y) + \lambda(X_k, cp_k) \right\rangle_{\sim Q(Y)} + \text{const.}$$

$$= \left[\left\langle \phi_{\mathrm{Y}}(pa_Y) \right\rangle_{\sim Q(Y)} + \sum_{k \in ch_Y} \left\langle \phi_{XY}(X_k, cp_k) \right\rangle_{\sim Q(Y)} \right]^{\mathrm{T}} \mathbf{u}_Y(Y) + f_Y(Y) + \text{const.}$$

Conjugate-Exponential Models



$$\phi_Y^* = \left\langle \phi_Y(pa_Y) \right\rangle_{\sim Q(Y)} + \sum_{k \in ch_Y} \left\langle \phi_{XY}(X_k, cp_k) \right\rangle_{\sim Q(Y)}$$

$$\tilde{\phi}_Y\left(\left\{ \left\langle \mathbf{u}_i \right\rangle \right\}_{i \in pa_Y} \right) = \left\langle \phi_Y(pa_Y) \right\rangle_{\sim Q(Y)}$$

$$\tilde{\phi}_{XY}\left(\left\langle \mathbf{u}_k \right\rangle, \left\{ \left\langle \mathbf{u}_j \right\rangle \right\}_{j \in cp_k} \right) = \left\langle \phi_{XY}(X_k, cp_k) \right\rangle_{\sim Q(Y)}$$

Variational Message Passing

- Conditional distributions:

$$\ln P(X \mid \boldsymbol{\theta}) = \boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{u}(X) + g(\boldsymbol{\theta}) + f(X)$$

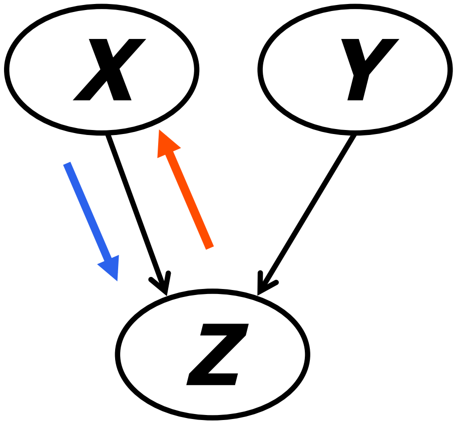
$$\ln P(Z \mid X, Y) = \boldsymbol{\varphi}(Y, Z)^{\mathrm{T}} \boldsymbol{u}(X) + g'(X) + f'(Y, Z)$$

- Messages:
  - Parent to child (X→Z)

$$m_{X \to Z} = \langle \boldsymbol{u}(X) \rangle_{Q(X)}$$

  - Child to parent (Z→X)

$$m_{Z \to X} = \langle \boldsymbol{\varphi}(Y, Z) \rangle_{Q(Y)Q(Z)}$$

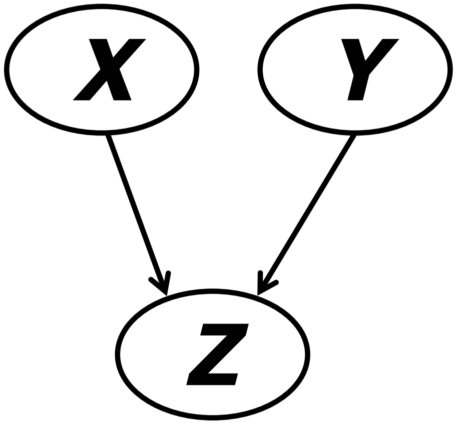


Varational Message Passing

- Optimal *Q(X)* has the same form as *P(X|θ)*, but with updated parameter vector *θ**

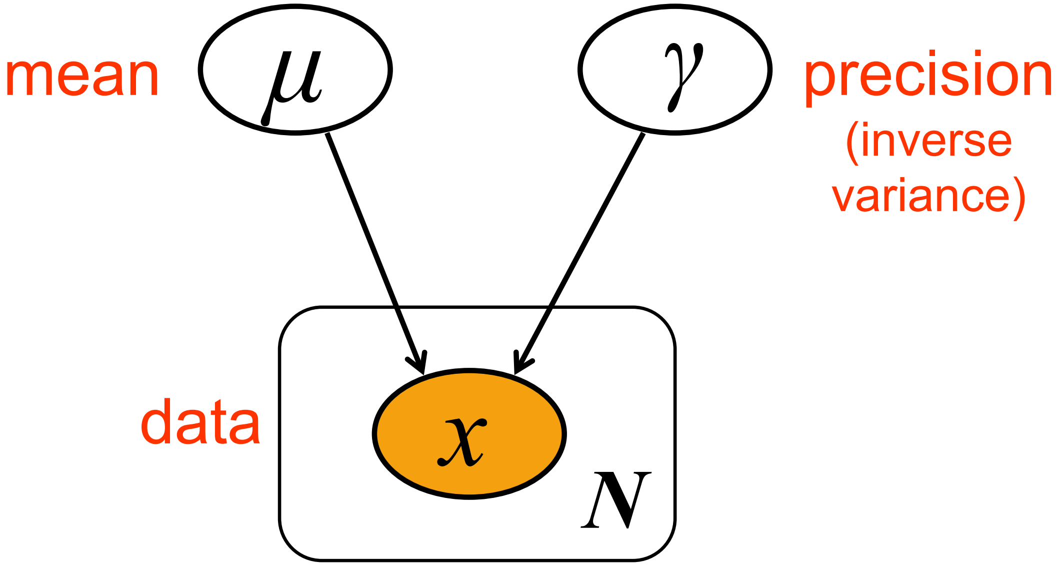
$$\boldsymbol{\theta}^* = \langle \boldsymbol{\theta} \rangle + \sum_{j \in ch(X)} m_{j \to X}$$

Computed from messages from parents



VMP Example: defining the model

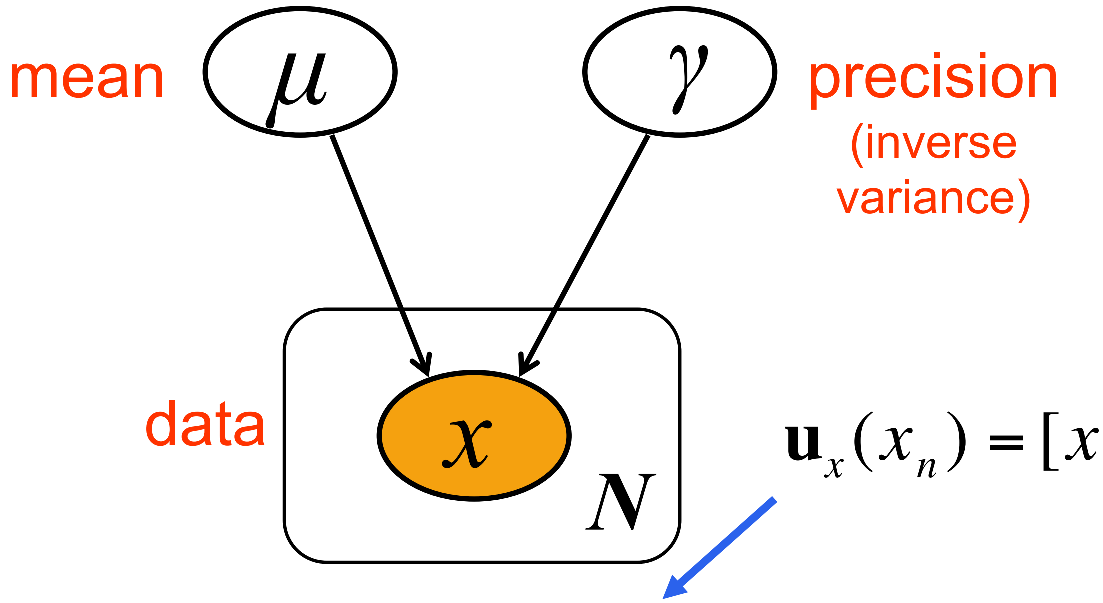
- Learning parameters of a Gaussian from N data points.



$$P(\mathbf{x} \mid \mu, \gamma^{-1}) = \prod_{n=1}^{N} N(x_n \mid \mu, \gamma^{-1})$$

VMP Example: defining the model

- Learning parameters of a Gaussian from N data points.

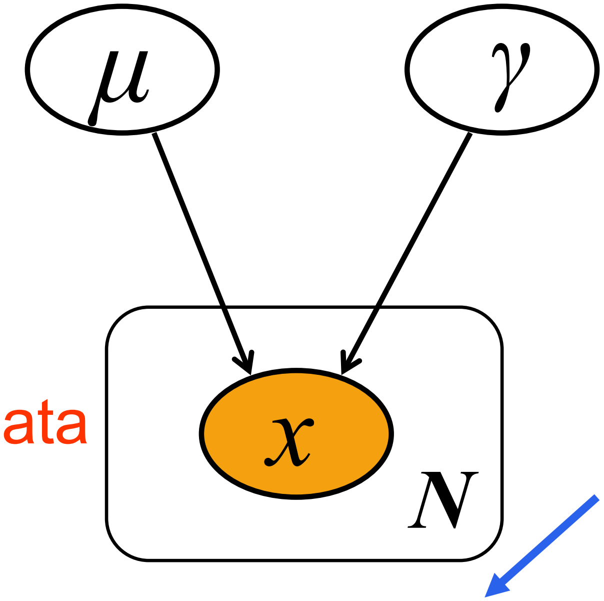


$$\mathbf{u}_x(x_n) = [x_n, x_n^2]^{\mathrm{T}}$$

$$\ln P(\mathbf{x} \mid \mu, \gamma^{-1}) = \begin{bmatrix} \gamma\mu \\ -\gamma/2 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} x_n \\ x_n^2 \end{bmatrix} + \frac{1}{2}(\ln\gamma - \gamma\mu^2 - \ln 2\pi)$$

VMP Example: defining the model

- Learning parameters of a Gaussian from N data points.



mean

precision
(inverse variance)

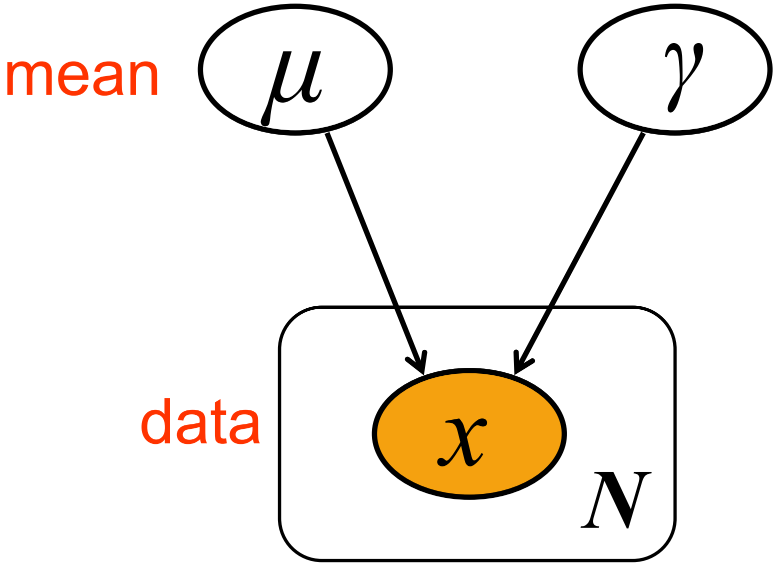
data

$$\mathbf{u}_\mu(\mu) = [\mu, \mu^2]^{\mathrm{T}}$$

$$\ln P(\mathbf{x} \mid \mu, \gamma^{-1}) = \begin{bmatrix} \gamma x_n \\ -\gamma/2 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \mu \\ \mu^2 \end{bmatrix} + \frac{1}{2}(\ln\gamma - \gamma x_n^2 - \ln 2\pi)$$

VMP Example: defining the model

- Learning parameters of a Gaussian from N data points.



$$\mathbf{u}_{\gamma}(\gamma) = [\gamma, \ln\gamma]^{\mathrm{T}}$$

$$\ln P(\mathbf{x} \mid \mu, \gamma^{-1}) = \begin{bmatrix} -\frac{1}{2}(x_n - \mu)^2 \\ \frac{1}{2} \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \gamma \\ \ln\gamma \end{bmatrix} - \ln 2\pi$$

VMP Example: defining the model

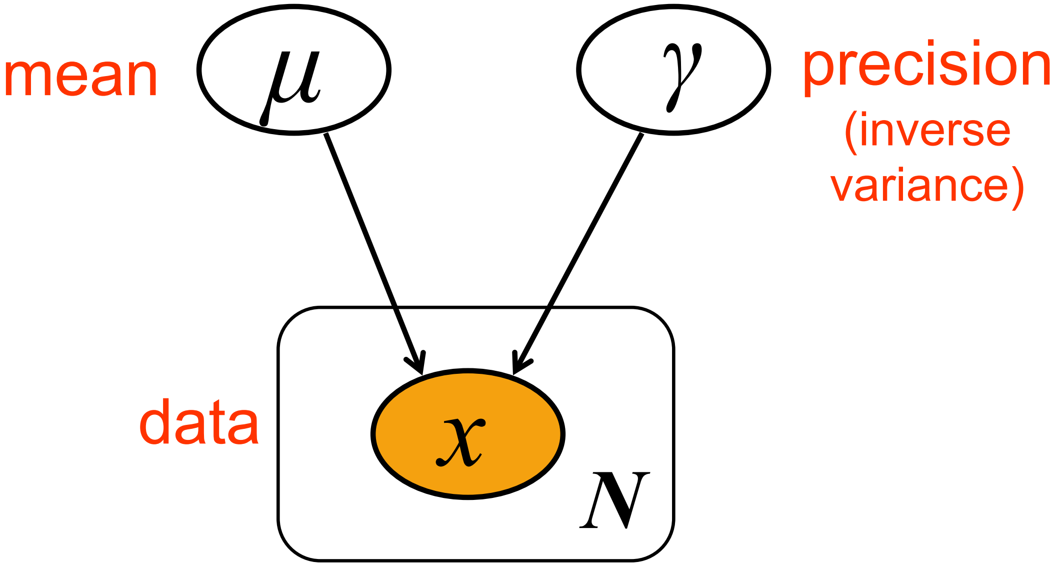
- Learning parameters of a Gaussian from N data points.

Gaussian distribution with hyper params (m, β)

$$\ln P(\mu \mid m, \beta^{-1}) = \begin{bmatrix} \beta m \\ -\beta/2 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \mu \\ \mu^2 \end{bmatrix} + \frac{1}{2}(\ln \beta - \beta m^2 - \ln 2\pi)$$

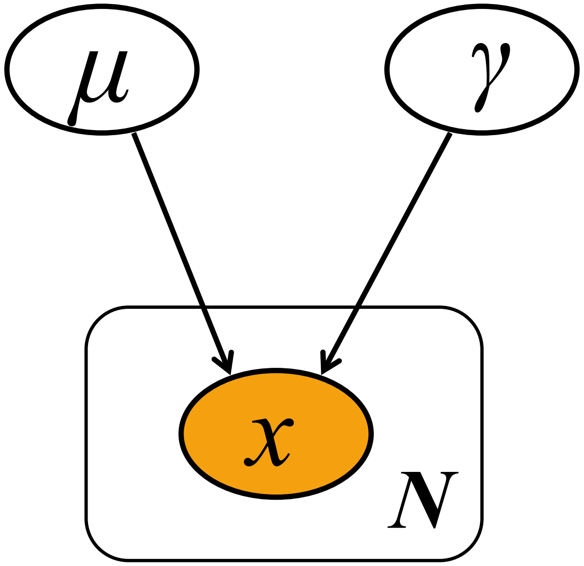
Gamma distribution with hyper params (a, b)

$$\ln P(\gamma \mid a, b) = \begin{bmatrix} -b \\ a-1 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \gamma \\ \ln \gamma \end{bmatrix} + a \ln b - \ln \Gamma(a)$$



VMP Example: passing messages

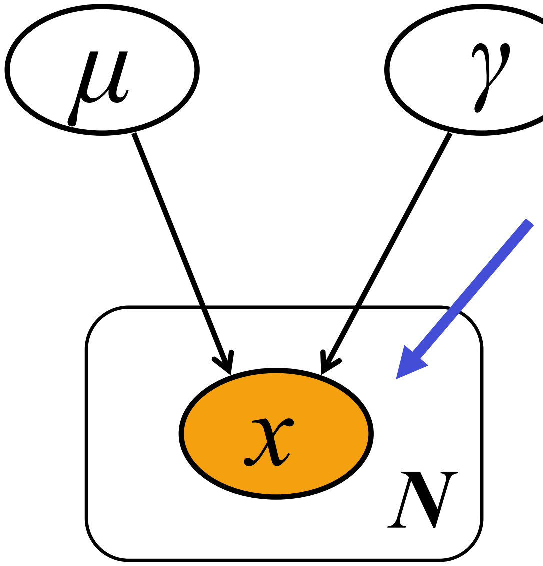
Variational Distribution: $Q(\mu,\gamma) = Q(\mu)Q(\gamma)$



Find initial values: $\langle \mathbf{u}_\mu(\mu) \rangle$ and $\langle \mathbf{u}_\gamma(\gamma) \rangle$

VMP Example: passing messages

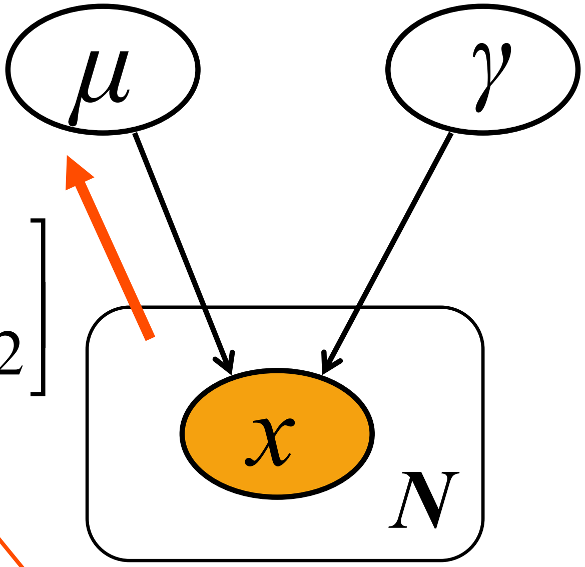
Message from γ to all x.



$$\mathbf{m}_{\gamma \to x_n} = \begin{bmatrix} \langle \gamma \rangle \\ \langle \ln \gamma \rangle \end{bmatrix}$$

VMP Example: passing messages

Messages from each x_n to μ.



$$\mathbf{m}_{x_n \to \mu} = \begin{bmatrix} \langle \gamma \rangle x_n \\ -\langle \gamma \rangle / 2 \end{bmatrix}$$

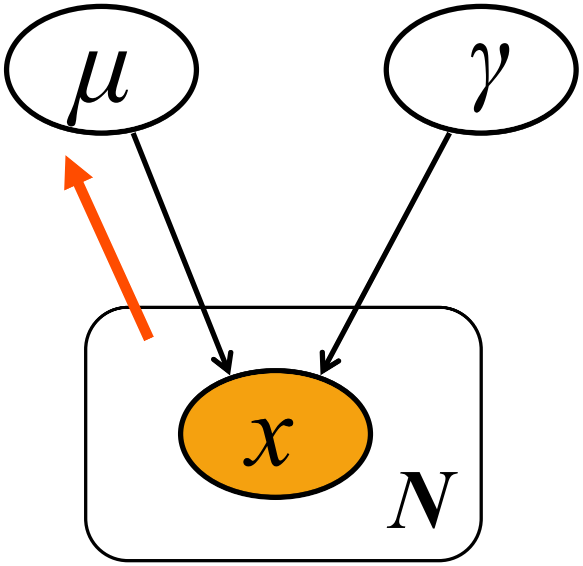
$$\ln P(\mathbf{x} \mid \mu, \gamma^{-1}) = \begin{bmatrix} \gamma x_n \\ -\gamma / 2 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \mu \\ \mu^2 \end{bmatrix} + \frac{1}{2}(\ln \gamma - \gamma x_n^2 - \ln 2\pi)$$

VMP Example: passing messages

Update $Q(\mu)$ parameter vector

$$\phi_\mu^* = \phi_\mu + \sum_{n=1}^{N} \mathbf{m}_{x_n \to \mu}$$

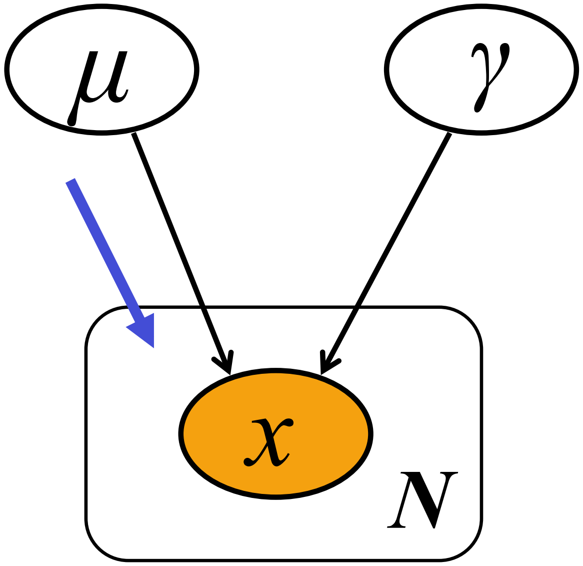
$$= \begin{bmatrix} \beta m \\ -\beta/2 \end{bmatrix} + \sum_{n=1}^{N} \mathbf{m}_{x_n \to \mu}$$



VMP Example: passing messages

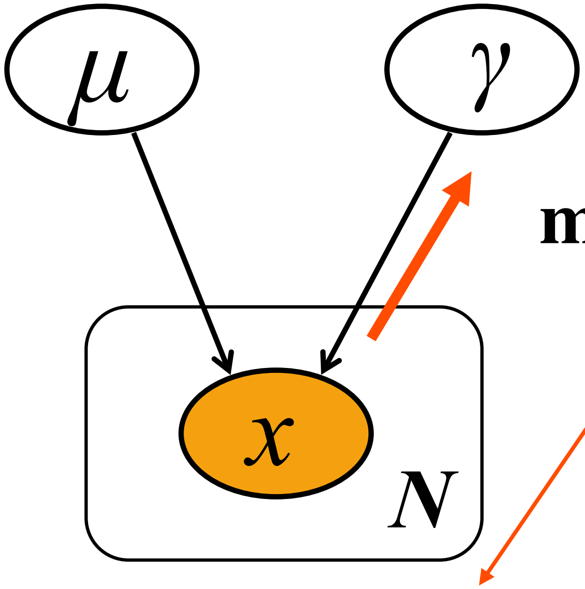
Message from updated μ to all x.

$$\mathbf{m}_{\mu \to x_n} = \begin{bmatrix} \langle \mu \rangle \\ \langle \mu^2 \rangle \end{bmatrix}$$



VMP Example: passing messages

Messages from each x_n to γ.

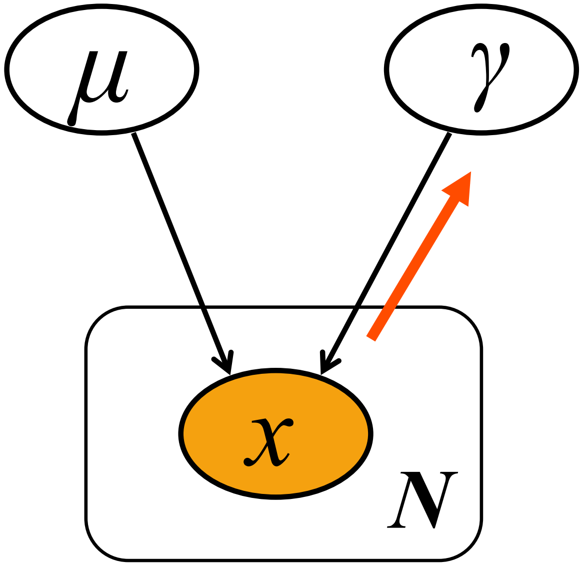


$$\mathbf{m}_{x_n \to \gamma} = \begin{bmatrix} -\tfrac{1}{2}\left\langle \left(x_n - \mu\right)^2 \right\rangle \\ \tfrac{1}{2} \end{bmatrix}$$

$$\ln P(\mathbf{x} \mid \mu, \gamma^{-1}) = \begin{bmatrix} -\tfrac{1}{2}(x_n - \mu)^2 \\ \tfrac{1}{2} \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \gamma \\ \ln \gamma \end{bmatrix} - \ln 2\pi$$

VMP Example: passing messages

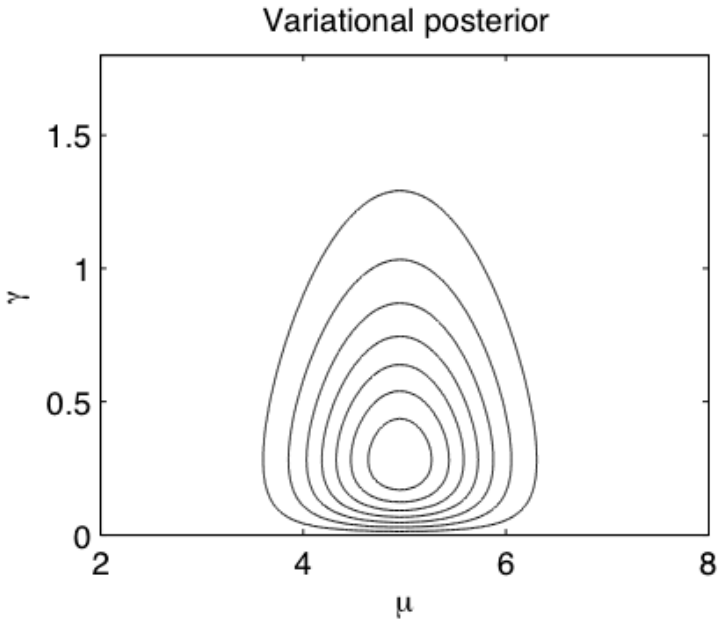
Update $Q(\gamma)$ parameter vector



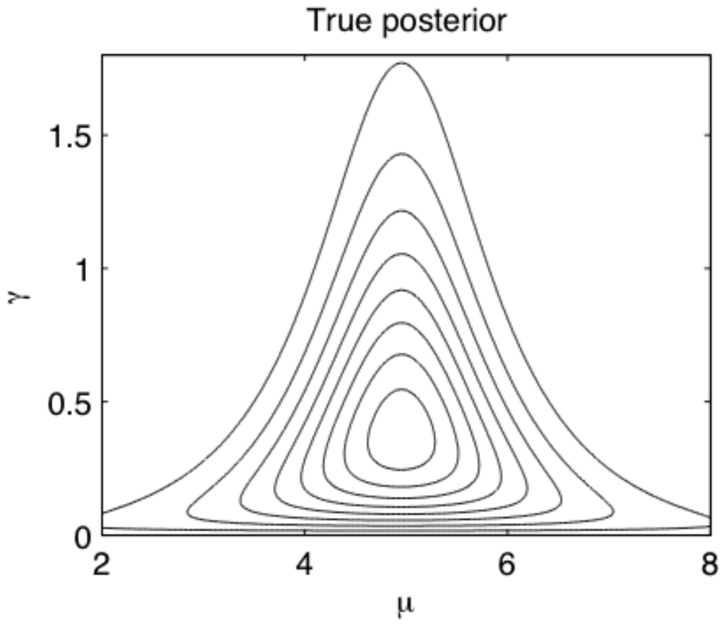
$$\phi_\gamma^* = \phi_\gamma + \sum_{n=1}^{N} m_{x_n \to \gamma}$$

$$= \begin{bmatrix} -b \\ a-1 \end{bmatrix} + \sum_{n=1}^{N} m_{x_n \to \gamma}$$

VMP Example: converged distribution



$$Q(\mu,\gamma) = Q_\mu(\mu)Q_\gamma(\gamma)$$



$$P(\mu,\gamma \mid \mathbf{V})$$

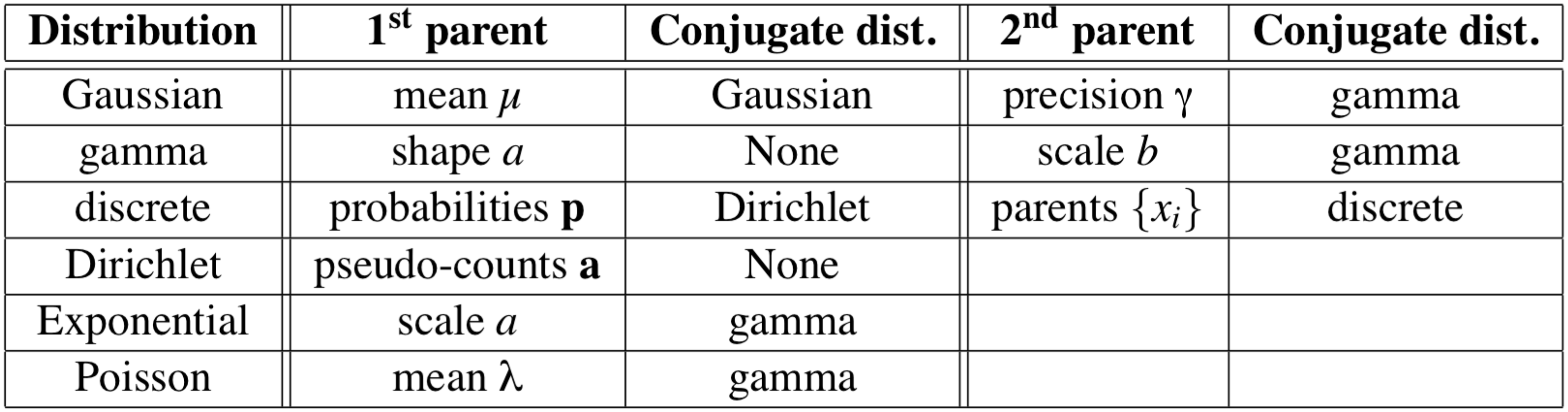
Features of VMP

- Guaranteed to converge to a local minimum of KL(Q||P)
- Flexible message passing schedule – factors can be updated in any order (thought it may alter convergence)
- Graph does not need to be a tree (needs to be acyclic)

Allowable Models and Constraints

- Parent-child edges must satisfy conjugacy
  - Gaussian variable:
    - Gaussian parent for its mean
    - Gamma parent for its precision
  - Gamma variable:
    - Gamma scale parameter b
  - Discrete Variable
    - Dirichlet prior

Allowable Models and Constraints



Distribution	1st parent	Conjugate dist.	2nd parent	Conjugate dist.
Gaussian	mean μ	Gaussian	precision γ	gamma
gamma	shape a	None	scale b	gamma
discrete	probabilities $\mathbf{p}$	Dirichlet	parents $\{x_i\}$	discrete
Dirichlet	pseudo-counts $\mathbf{a}$	None		
Exponential	scale a	gamma		
Poisson	mean λ	gamma		

Table 1: Distributions for each parameter of a number of exponential family distributions if the model is to satisfy conjugacy constraints. Conjugacy also holds if the distributions are replaced by their multivariate counterparts e.g. the distribution conjugate to the precision matrix of a multivariate Gaussian is a Wishart distribution. Where “None” is specified, no standard distribution satisfies conjugacy.

Allowable Models and Constraints

- Truncated Distributions
- Incorporates deterministic variables
- Mixture distributions
- Multivariate distributions

Allowable Models: Mixture Models

- Not in the exponential family:

$$P(X \mid \{\pi_k\},\{\theta_k\}) = \sum_{k=1}^{K} \pi_k P_k(X \mid \theta_k)$$

- Introduce latent variable, λ, which indicates component

$$P(X \mid \lambda,\{\theta_k\}) = \prod_{k=1}^{K} P_k(X \mid \theta_k)^{\delta_{\lambda k}}$$

$$\ln P(X \mid \lambda,\{\theta_k\}) = \sum_{k} \delta(\lambda,k)\left[\phi_k(\theta_k)^{\mathrm{T}}\mathbf{u}_k(X) + f_k(X) + g_k(\theta_k)\right]$$

Allowable Models: Mixture Models

- Require that all component distributions have the same natural statistic vector:

$$\mathbf{u}_X(X) \overset{\Delta}{=} \mathbf{u}_1(X) = \ldots = \mathbf{u}_K(X)$$

- Can rewrite log conditional:

$$\ln P(X \mid \lambda, \{\theta_k\}) = \phi_X(\lambda, \{\theta_k\})^{\mathrm{T}} \mathbf{u}_X(X) + f_X(X) + \tilde{g}(\phi_X(\lambda, \{\theta_k\}))$$

where $$\phi_X = \sum_k \delta(\lambda, k) \phi_k(\theta_k)$$

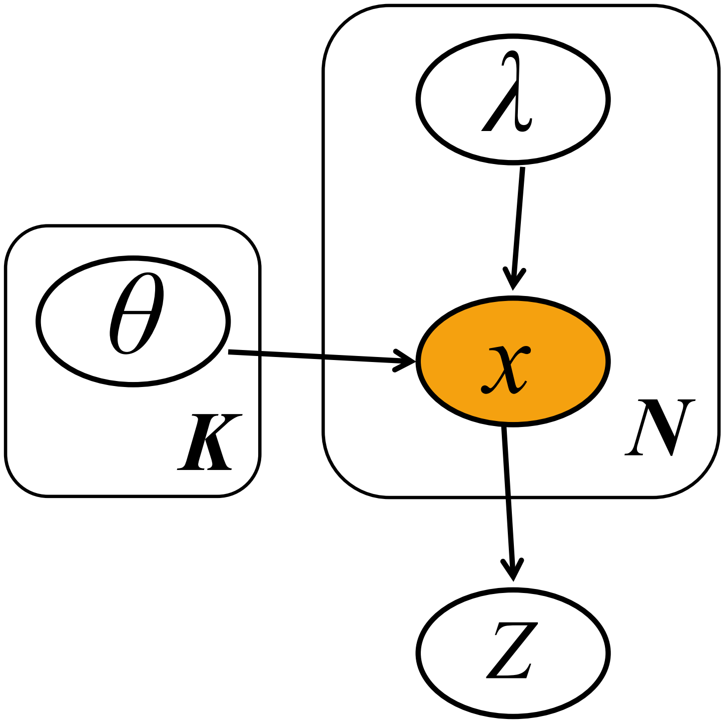
Allowable Models: Mixture Models

- Now the messages are defined as:

$$\mathbf{m}_{X\to\theta_k} = Q(\lambda = k)\left\langle \phi_{X\theta_k} \right\rangle$$

$$\mathbf{m}_{X\to Z\in ch_X} = \left\langle \mathbf{u}_X(X) \right\rangle$$

$$\mathbf{m}_{X\to\lambda} = \left\langle \ln P_k(X\,|\,\theta_k) \right\rangle$$

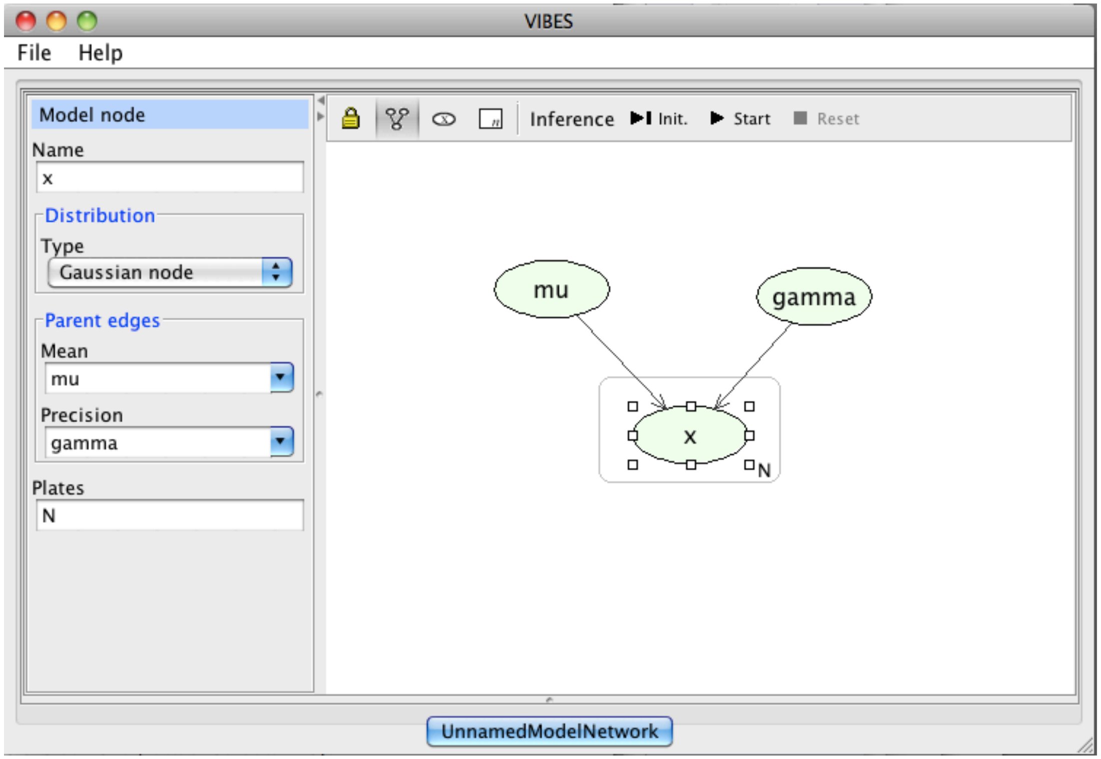


Allowable Models

- General architecture: arbitrary directed acyclic subgraphs of multinomial discrete variables (with Dirichlet priors)
- Arbitrary subgraphs of univariate and multivariate linear Gaussian nodes (having gamma and Wishart priors)
- Arbitrary mixture nodes providing connections from discrete to continuous subgraphs
- Can include deterministic nodes
- Any continuous distribution can be truncated to restrict range of allowable values
- Includes: HMMs, Kalman Filters, Factor Analysis, PCA, ICA

VIBES

- VIBES – inspired by WinBUGS
- Graphically specify models, and run inference



Model node

Name: x

Distribution

Type: Gaussian node

Parent edges

Mean: mu

Precision: gamma

Plates: N

Extensions

- Can introduce additional variational parameters to use non-conjugate distributions
- Logistic Sigmoid function can be estimated by Gaussian-like bound
- Next step would be achieve a posterior estimate with some dependency structure (i.e. structured variational inference)