Approximate Inference

E. Burnaev, Skoltech

Variational Inference

Evaluate the posterior $p(\mathbf{Z}|\mathbf{X})$ of latent variables $\mathbf{Z}$

$\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ **is a data set**

$\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$ **is a set of latent variables**

$p(\mathbf{X}, \mathbf{Z})$ **is probabilistic model**

We want to approximate $p(\mathbf{Z}|\mathbf{X})$ and $p(\mathbf{X})$

$$\ln p(\mathbf{X}) = \mathcal{L}(q) + \mathrm{KL}(q\|p)$$

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}$$

$$\mathrm{KL}(q\|p) = -\int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}.$$

Variational Inference

We want to approximate $p(\mathbf{Z}|\mathbf{X})$ and $p(\mathbf{X})$

$$\ln p(\mathbf{X}) = \mathcal{L}(q) + \mathrm{KL}(q\|p)$$

$$\begin{aligned} \mathcal{L}(q) &= \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z} \\ \mathrm{KL}(q\|p) &= -\int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}. \end{aligned}$$

$\mathcal{L}(q)$ is a lower bound to $\ln p(\mathbf{X})$

We maximize this lower bound w.r.t. $q(\mathbf{Z})$ which is equivalent to minimizing KL-divergence between

$p(\mathbf{Z}|\mathbf{X})$ and $q(\mathbf{Z})$

Variational Inference

We want to approximate $p(\mathbf{Z}|\mathbf{X})$ **and** $p(\mathbf{X})$

$$\ln p(\mathbf{X}) = \mathcal{L}(q) + \mathrm{KL}(q\|p)$$

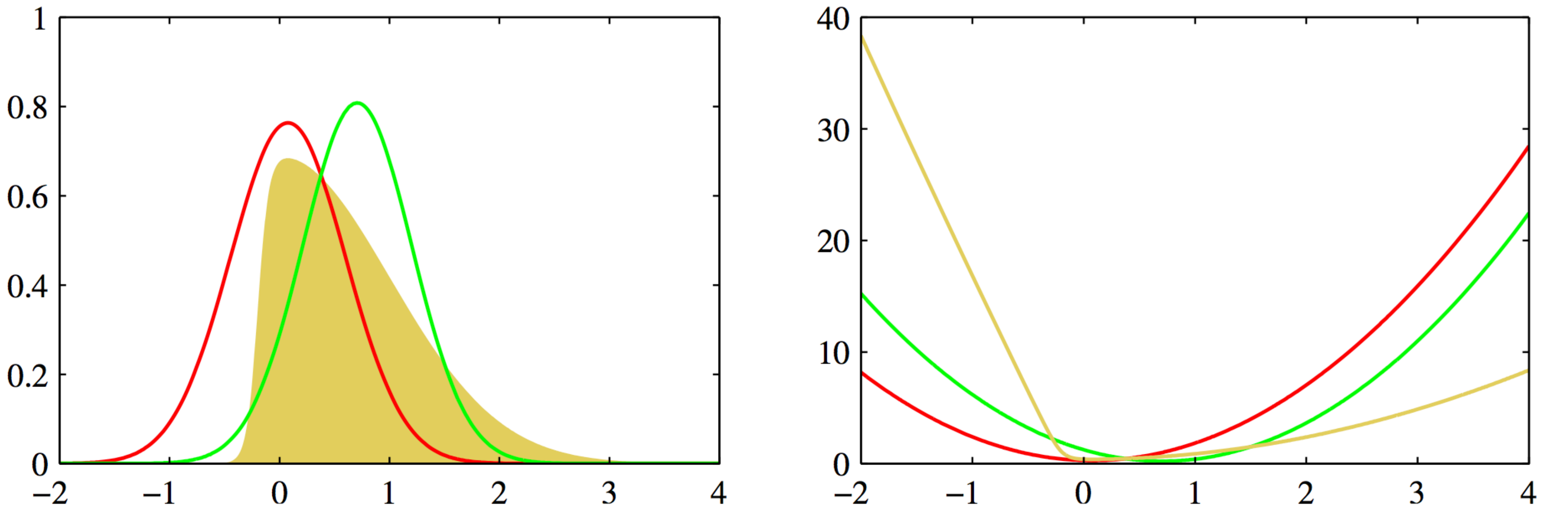
$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}$$

$$\mathrm{KL}(q\|p) = -\int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}.$$

$\mathcal{L}(q)$ **is a lower bound to** $\ln p(\mathbf{X})$

We maximize this lower bound w.r.t. $q(\mathbf{Z})$ **which is equivalent to minimizing KL-divergence between**

$p(\mathbf{Z}|\mathbf{X})$ **and** $q(\mathbf{Z})$



We can use parametric distributions $q(\mathbf{Z}|\omega)$

Variational distribution is Gaussian and we optimize w.r.t. its mean and variance

Laplace approximation is in red, Variational approximation is in green

Factorized Distributions

Let us assume that (we partitioned all latent variables into M disjoint groups)

$$q(\mathbf{Z}) = \prod_{i=1}^{M} q_i(\mathbf{Z}_i).$$

We want to find such factorization that the lower bound is maximal

$$\begin{aligned}
\mathcal{L}(q) &= \int \prod_i q_i \left\{ \ln p(\mathbf{X}, \mathbf{Z}) - \sum_i \ln q_i \right\} \mathrm{d}\mathbf{Z} \\
&= \int q_j \left\{ \int \ln p(\mathbf{X}, \mathbf{Z}) \prod_{i \neq j} q_i \, \mathrm{d}\mathbf{Z}_i \right\} \mathrm{d}\mathbf{Z}_j - \int q_j \ln q_j \, \mathrm{d}\mathbf{Z}_j + \text{const} \\
&= \int q_j \ln \widetilde{p}(\mathbf{X}, \mathbf{Z}_j) \, \mathrm{d}\mathbf{Z}_j - \int q_j \ln q_j \, \mathrm{d}\mathbf{Z}_j + \text{const}
\end{aligned}$$

Factorized Distributions

Let us introduce a new distribution $\widetilde{p}(\mathbf{X}, \mathbf{Z}_j)$ by the relation

$$\ln \widetilde{p}(\mathbf{X}, \mathbf{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] + \text{const.}$$

Here $\mathbb{E}_{i \neq j}[\cdots]$ is the expectation w.r.t. $\prod_{i \neq j} q_i$

$$\mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] = \int \ln p(\mathbf{X}, \mathbf{Z}) \prod_{i \neq j} q_i \, \mathrm{d}\mathbf{Z}_i.$$

Optimum of

$$\mathcal{L}(q) = \int q_j \ln \widetilde{p}(\mathbf{X}, \mathbf{Z}_j) \, \mathrm{d}\mathbf{Z}_j - \int q_j \ln q_j \, \mathrm{d}\mathbf{Z}_j + \text{const}$$

is obtained on $q_j(\mathbf{Z}_j) = \widetilde{p}(\mathbf{X}, \mathbf{Z}_j)$

Factorized Distributions

Let us introduce a new distribution $\widetilde{p}(\mathbf{X}, \mathbf{Z}_j)$ **by the relation**

$$\ln \widetilde{p}(\mathbf{X}, \mathbf{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] + \text{const.}$$

Here $\mathbb{E}_{i \neq j}[\cdots]$ **is the expectation w.r.t.** $\prod q_i$

$$\mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] = \int \ln p(\mathbf{X}, \mathbf{Z}) \prod_{i \neq j} q_i \, \mathrm{d}\mathbf{Z}_i.$$

Optimum of

$$\mathcal{L}(q) = \int q_j \ln \widetilde{p}(\mathbf{X}, \mathbf{Z}_j) \, \mathrm{d}\mathbf{Z}_j - \int q_j \ln q_j \, \mathrm{d}\mathbf{Z}_j + \text{const}$$

is obtained on $q_j(\mathbf{Z}_j) = \widetilde{p}(\mathbf{X}, \mathbf{Z}_j)$

Factorized Distributions

Thus we obtain optimal $q_j^\star(\mathbf{Z}_j)$

$$\ln q_j^\star(\mathbf{Z}_j) = \mathbb{E}_{i\neq j}[\ln p(\mathbf{X},\mathbf{Z})] + \text{const.}$$

$$q_j^\star(\mathbf{Z}_j) = \frac{\exp\left(\mathbb{E}_{i\neq j}[\ln p(\mathbf{X},\mathbf{Z})]\right)}{\int \exp\left(\mathbb{E}_{i\neq j}[\ln p(\mathbf{X},\mathbf{Z})]\right)\,\mathrm{d}\mathbf{Z}_j}.$$

1) First, we initialize all the factors $q_i(\mathbf{Z}_i)$
2) Cycle through the factors and replace each in turn with a revised estimate

Properties of Factorized Distributions

Let us consider a Gaussian distribution $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \qquad \boldsymbol{\Lambda} = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix}$$

$$q(\mathbf{z}) = q_1(z_1) q_2(z_2)$$

$$\begin{aligned} \ln q_1^\star(z_1) &= \mathbb{E}_{z_2}[\ln p(\mathbf{z})] + \text{const} \\ &= \mathbb{E}_{z_2}\left[-\frac{1}{2}(z_1 - \mu_1)^2 \Lambda_{11} - (z_1 - \mu_1)\Lambda_{12}(z_2 - \mu_2)\right] + \text{const} \\ &= -\frac{1}{2} z_1^2 \Lambda_{11} + z_1 \mu_1 \Lambda_{11} - z_1 \Lambda_{12} \left(\mathbb{E}[z_2] - \mu_2\right) + \text{const.} \end{aligned}$$

$$q^\star(z_1) = \mathcal{N}(z_1 | m_1, \Lambda_{11}^{-1})$$

$$m_1 = \mu_1 - \Lambda_{11}^{-1} \Lambda_{12} \left(\mathbb{E}[z_2] - \mu_2\right)$$

Properties of Factorized Distributions

Analogously $$q_2^\star(z_2) = \mathcal{N}(z_2 | m_2, \Lambda_{22}^{-1})$$

$$m_2 = \mu_2 - \Lambda_{22}^{-1}\Lambda_{21}\left(\mathbb{E}[z_1] - \mu_1\right)$$

Due to linearity everything is satisfied if $$\mathbb{E}[z_1] = \mu_1$$
$$\mathbb{E}[z_2] = \mu_2$$

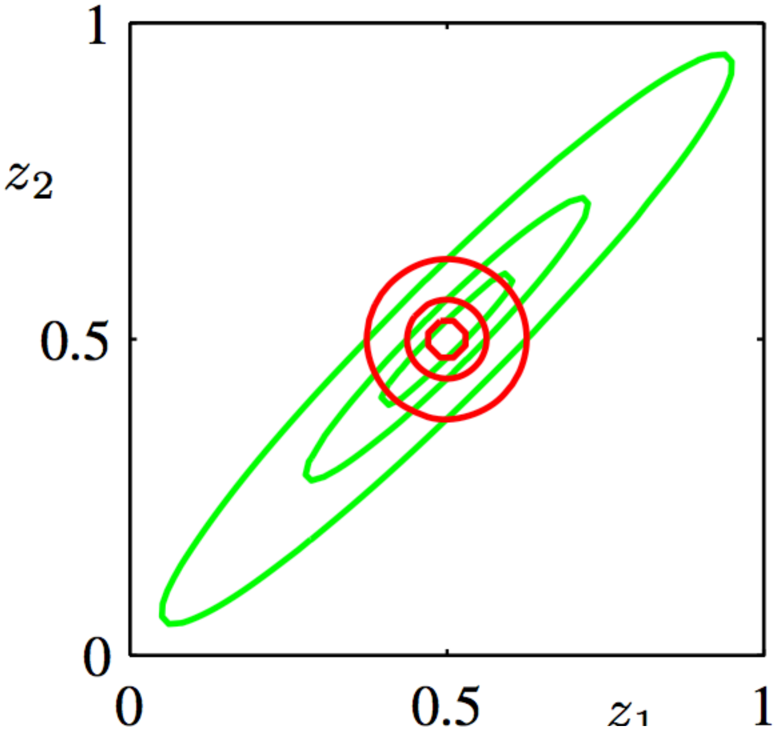
Here we were optimizing $\mathrm{KL}(q\|p)$

Let us optimize $\mathrm{KL}(p\|q)$ **!!!**

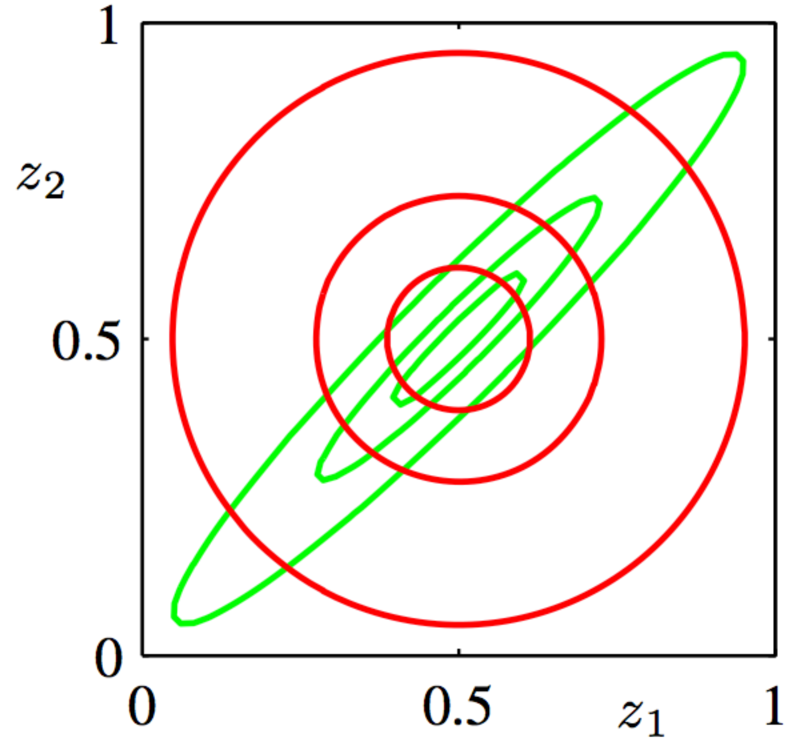
$$\mathrm{KL}(p\|q) = -\int p(\mathbf{Z})\left[\sum_{i=1}^{M} \ln q_i(\mathbf{Z}_i)\right] \mathrm{d}\mathbf{Z} + \mathrm{const}$$

Using Lagrange multipliers we get that

$$q_j^\star(\mathbf{Z}_j) = \int p(\mathbf{Z}) \prod_{i \neq j} \mathrm{d}\mathbf{Z}_i = p(\mathbf{Z}_j)$$



Optimization of $\mathrm{KL}(q\|p)$

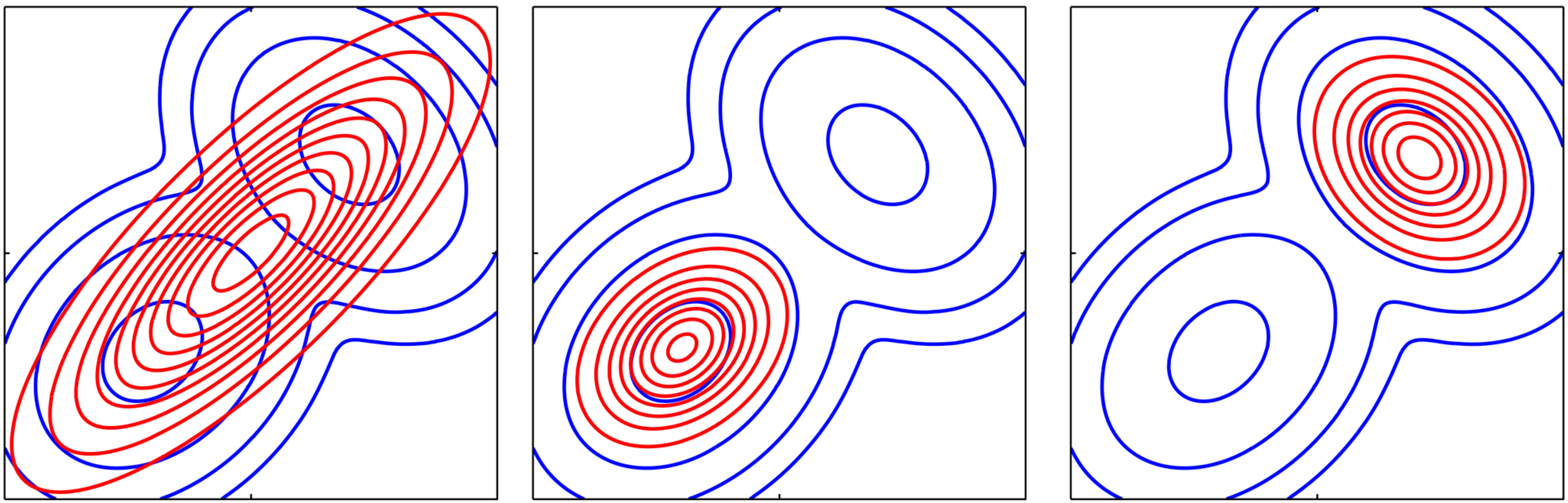


Optimization of $\mathrm{KL}(p\|q)$

$$\mathrm{KL}(q\|p) = -\int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}$$

Thus minimizing $\mathrm{KL}(q\|p)$ **w.r.t.** $q(\mathbf{Z})$ **we try to avoid regions where** $p(\mathbf{Z})$ **is small**

Conversely, $\mathrm{KL}(p\|q)$ **is minimized w.r.t.** $q(\mathbf{Z})$ **non-zero in those regions where** $p(\mathbf{Z})$ **is non-zero**



$\mathrm{KL}(p\|q)$ $\mathrm{KL}(q\|p)$ $\mathrm{KL}(q\|p)$

Example: 1d Gaussian

Data set $\mathcal{D} = \{x_1, \ldots, x_N\}$

$$p(\mathcal{D}|\mu, \tau) = \left(\frac{\tau}{2\pi}\right)^{N/2} \exp\left\{-\frac{\tau}{2}\sum_{n=1}^{N}(x_n - \mu)^2\right\}.$$

$$\begin{aligned} p(\mu|\tau) &= \mathcal{N}\left(\mu|\mu_0, (\lambda_0\tau)^{-1}\right) \\ p(\tau) &= \mathrm{Gam}(\tau|a_0, b_0) \end{aligned}$$

Variational Approximation $q(\mu, \tau) = q_\mu(\mu)q_\tau(\tau)$

$$\begin{aligned} \ln q_\mu^\star(\mu) &= \mathbb{E}_\tau\left[\ln p(\mathcal{D}|\mu, \tau) + \ln p(\mu|\tau)\right] + \text{const} \\ &= -\frac{\mathbb{E}[\tau]}{2}\left\{\lambda_0(\mu - \mu_0)^2 + \sum_{n=1}^{N}(x_n - \mu)^2\right\} + \text{const.} \end{aligned}$$

Example: 1d Gaussian

Variational Approximation $q(\mu, \tau) = q_\mu(\mu) q_\tau(\tau)$

$$
\begin{aligned}
\ln q_\mu^\star(\mu) &= \mathbb{E}_\tau \left[\ln p(\mathcal{D}|\mu, \tau) + \ln p(\mu|\tau) \right] + \text{const} \\
&= -\frac{\mathbb{E}[\tau]}{2} \left\{ \lambda_0 (\mu - \mu_0)^2 + \sum_{n=1}^{N} (x_n - \mu)^2 \right\} + \text{const.}
\end{aligned}
$$

$q_\mu(\mu)$ **=** $\mathcal{N}\left(\mu | \mu_N, \lambda_N^{-1}\right)$ **is a solution**

$$
\begin{aligned}
\mu_N &= \frac{\lambda_0 \mu_0 + N\overline{x}}{\lambda_0 + N} \\
\lambda_N &= (\lambda_0 + N)\mathbb{E}[\tau].
\end{aligned}
$$

Example: 1d Gaussian

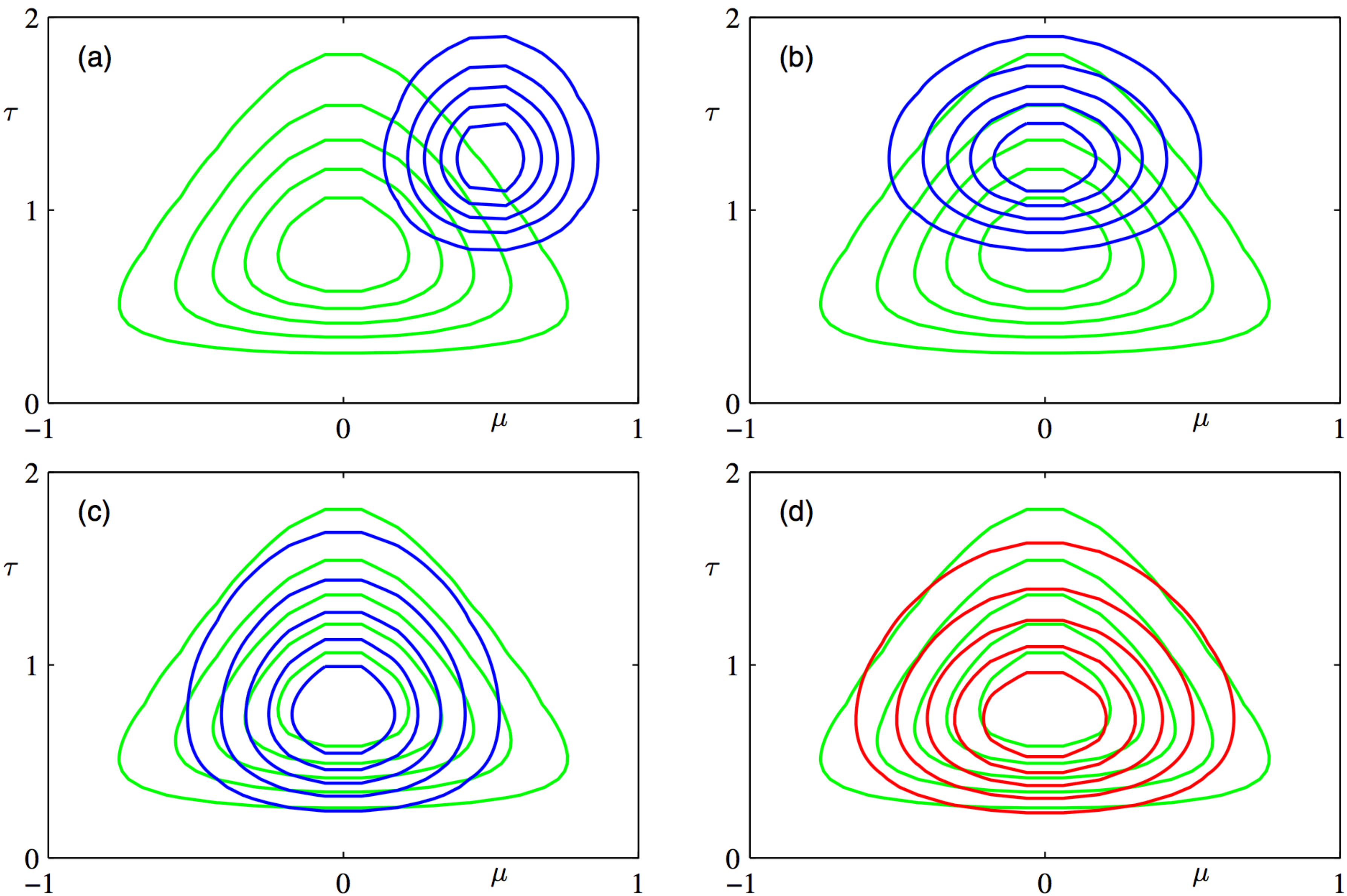
Analogously

$$\begin{aligned}\ln q_\tau^\star(\tau) &= \mathbb{E}_\mu\left[\ln p(\mathcal{D}|\mu,\tau) + \ln p(\mu|\tau)\right] + \ln p(\tau) + \text{const} \\ &= (a_0 - 1)\ln\tau - b_0\tau + \frac{N}{2}\ln\tau \\ &\quad -\frac{\tau}{2}\mathbb{E}_\mu\left[\sum_{n=1}^{N}(x_n-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right] + \text{const}\end{aligned}$$

$q_\tau(\tau)$ = $\mathrm{Gam}(\tau|a_N, b_N)$ **is a solution**

$$\begin{aligned}a_N &= a_0 + \frac{N}{2} \\ b_N &= b_0 + \frac{1}{2}\mathbb{E}_\mu\left[\sum_{n=1}^{N}(x_n-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right].\end{aligned}$$

Example: 1d Gaussian



Model Comparison

Hidden variables $\mathbf{Z}$.

Prior probabilities over models $p(m)$

Approximate $p(m|\mathbf{X})$

We do not consider factorized variational approximation since different models may have different dimension/ structure, so we consider

$$q(\mathbf{Z}, m) = q(\mathbf{Z}|m)q(m)$$

Model Comparison

Hidden variables $\mathbf{Z}$.

Prior probabilities over models $p(m)$ **and**

$q(\mathbf{Z}, m) = q(\mathbf{Z}|m)q(m)$

Approximate $p(m|\mathbf{X})$

$$\ln p(\mathbf{X}) = \mathcal{L}_m - \sum_m \sum_{\mathbf{Z}} q(\mathbf{Z}|m)q(m) \ln \left\{ \frac{p(\mathbf{Z}, m|\mathbf{X})}{q(\mathbf{Z}|m)q(m)} \right\}$$

$$\mathcal{L}_m = \sum_m \sum_{\mathbf{Z}} q(\mathbf{Z}|m)q(m) \ln \left\{ \frac{p(\mathbf{Z}, \mathbf{X}, m)}{q(\mathbf{Z}|m)q(m)} \right\}$$

First optimize $\mathcal{L}_m$ **w.r.t.** $q(\mathbf{Z}|m)$

Second optimizing w.r.t. $q(m)$ **we get** $q(m) \propto p(m) \exp\{\mathcal{L}_m\}$

Variational Linear Regression

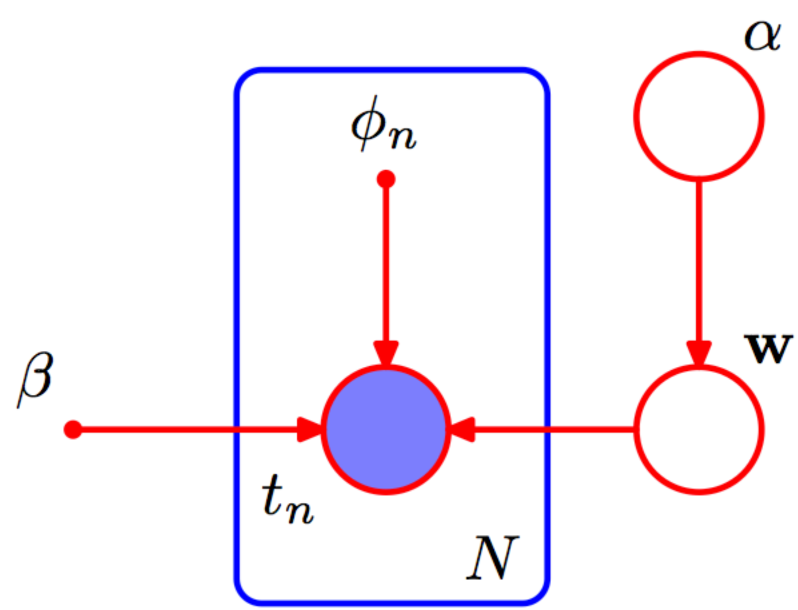
$$p(\mathbf{t}|\mathbf{w}) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n, \beta^{-1}) \ , \ \boldsymbol{\phi}_n = \boldsymbol{\phi}(\mathbf{x}_n)$$

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$$

We introduce prior over α

$$p(\alpha) = \mathrm{Gam}(\alpha|a_0, b_0)$$

$$p(\mathbf{t}, \mathbf{w}, \alpha) = p(\mathbf{t}|\mathbf{w})p(\mathbf{w}|\alpha)p(\alpha)$$



Let us approximate the posterior

$$p(\mathbf{w}, \alpha|\mathbf{t})$$

by the factorized distribution

$$q(\mathbf{w}, \alpha) = q(\mathbf{w})q(\alpha)$$

Variational Linear Regression

Let us approximate the posterior

$$p(\mathbf{w}, \alpha|\mathbf{t})$$

by the factorized distribution

$$q(\mathbf{w}, \alpha) = q(\mathbf{w})q(\alpha)$$

$$\ln q^{\star}(\alpha) = \ln p(\alpha) + \mathbb{E}_{\mathbf{w}}\left[\ln p(\mathbf{w}|\alpha)\right] + \text{const}$$

$$= \quad (a_0 - 1)\ln\alpha - b_0\alpha + \frac{M}{2}\ln\alpha - \frac{\alpha}{2}\mathbb{E}[\mathbf{w}^{\mathrm{T}}\mathbf{w}] + \text{const.}$$

It holds that

$$q^{\star}(\alpha) = \mathrm{Gam}(\alpha|a_N, b_N)$$

$$a_N = a_0 + \frac{M}{2}$$

$$b_N = b_0 + \frac{1}{2}\mathbb{E}[\mathbf{w}^{\mathrm{T}}\mathbf{w}]$$

Variational Linear Regression

Analogously we get that

$$\begin{aligned} \ln q^\star(\mathbf{w}) &= \ln p(\mathbf{t}|\mathbf{w}) + \mathbb{E}_\alpha\left[\ln p(\mathbf{w}|\alpha)\right] + \text{const} \\ &= -\frac{\beta}{2}\sum_{n=1}^{N}\{\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n - t_n\}^2 - \frac{1}{2}\mathbb{E}[\alpha]\mathbf{w}^{\mathrm{T}}\mathbf{w} + \text{const} \\ &= -\frac{1}{2}\mathbf{w}^{\mathrm{T}}\left(\mathbb{E}[\alpha]\mathbf{I} + \beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)\mathbf{w} + \beta\mathbf{w}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t} + \text{const.} \end{aligned}$$

It holds that

$q^\star(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$ **with**

$$\begin{aligned} \mathbf{m}_N &= \beta\mathbf{S}_N\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t} \\ \mathbf{S}_N &= \left(\mathbb{E}[\alpha]\mathbf{I} + \beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}. \end{aligned}$$

Variational Linear Regression

$$q^\star(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$$ **with** $$\begin{aligned} \mathbf{m}_N &= \beta \mathbf{S}_N \mathbf{\Phi}^{\mathrm{T}} \mathbf{t} \\ \mathbf{S}_N &= \left(\mathbb{E}[\alpha]\mathbf{I} + \beta \mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right)^{-1}. \end{aligned}$$

$$\begin{aligned} \mathbb{E}[\alpha] &= a_N / b_N \\ \mathbb{E}[\mathbf{w}\mathbf{w}^{\mathrm{T}}] &= \mathbf{m}_N \mathbf{m}_N^{\mathrm{T}} + \mathbf{S}_N \end{aligned}$$

$$\begin{aligned} a_N &= a_0 + \frac{M}{2} \\ b_N &= b_0 + \frac{1}{2}\mathbb{E}[\mathbf{w}^{\mathrm{T}}\mathbf{w}]. \end{aligned}$$

$$\mathbb{E}[\alpha] = \frac{a_N}{b_N} = \frac{M/2}{\mathbb{E}[\mathbf{w}^{\mathrm{T}}\mathbf{w}]/2} = \frac{M}{\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N + \mathrm{Tr}(\mathbf{S}_N)}.$$

Variational Linear Regression: Predictive Distribution

Predictive distribution

$$\begin{aligned}
p(t|\mathbf{x},\mathbf{t}) &= \int p(t|\mathbf{x},\mathbf{w})p(\mathbf{w}|\mathbf{t})\,\mathrm{d}\mathbf{w} \\
&\simeq \int p(t|\mathbf{x},\mathbf{w})q(\mathbf{w})\,\mathrm{d}\mathbf{w} \\
&= \int \mathcal{N}(t|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}),\beta^{-1})\mathcal{N}(\mathbf{w}|\mathbf{m}_N,\mathbf{S}_N)\,\mathrm{d}\mathbf{w} \\
&= \mathcal{N}(t|\mathbf{m}_N^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}),\sigma^2(\mathbf{x}))
\end{aligned}$$

$$\sigma^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x})$$

Variational Linear Regression: Lower Bound

Lower bound

$$
\begin{aligned}
\mathcal{L}(q) &= \mathbb{E}[\ln p(\mathbf{w}, \alpha, \mathbf{t})] - \mathbb{E}[\ln q(\mathbf{w}, \alpha)] \\
&= \mathbb{E}_{\mathbf{w}}[\ln p(\mathbf{t}|\mathbf{w})] + \mathbb{E}_{\mathbf{w},\alpha}[\ln p(\mathbf{w}|\alpha)] + \mathbb{E}_{\alpha}[\ln p(\alpha)] \\
&\quad -\mathbb{E}_{\alpha}[\ln q(\mathbf{w})]_{\mathbf{w}} - \mathbb{E}[\ln q(\alpha)].
\end{aligned}
$$

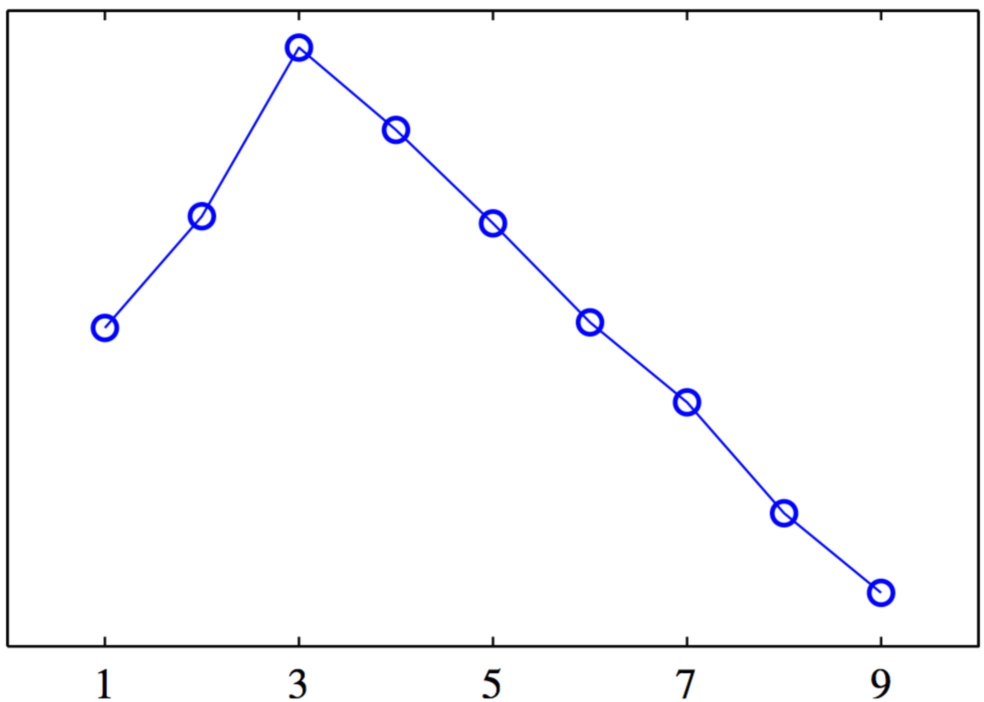
Variational Linear Regression: Lower Bound

$$
\begin{aligned}
\mathbb{E}[\ln p(\mathbf{t}|\mathbf{w})]_{\mathbf{w}} &= \frac{N}{2}\ln\left(\frac{\beta}{2\pi}\right) - \frac{\beta}{2}\mathbf{t}^{\mathrm{T}}\mathbf{t} + \beta\mathbf{m}_N^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\mathbf{t} \\
&\quad -\frac{\beta}{2}\mathrm{Tr}\left[\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}(\mathbf{m}_N\mathbf{m}_N^{\mathrm{T}} + \mathbf{S}_N)\right] \\
\mathbb{E}[\ln p(\mathbf{w}|\alpha)]_{\mathbf{w},\alpha} &= -\frac{M}{2}\ln(2\pi) + \frac{M}{2}(\psi(a_N) - \ln b_N) \\
&\quad -\frac{a_N}{2b_N}\left[\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N + \mathrm{Tr}(\mathbf{S}_N)\right] \\
\mathbb{E}[\ln p(\alpha)]_{\alpha} &= a_0 \ln b_0 + (a_0 - 1)\left[\psi(a_N) - \ln b_N\right] \\
&\quad -b_0\frac{a_N}{b_N} - \ln\Gamma(a_N) \\
-\mathbb{E}[\ln q(\mathbf{w})]_{\mathbf{w}} &= \frac{1}{2}\ln|\mathbf{S}_N| + \frac{M}{2}\left[1 + \ln(2\pi)\right] \\
-\mathbb{E}[\ln q(\alpha)]_{\alpha} &= \ln\Gamma(a_N) - (a_N - 1)\psi(a_N) - \ln b_N + a_N.
\end{aligned}
$$

Variational Linear Regression: Lower Bound

Lower bound versus degree M of a polynomial

We can interpret lower bound $\mathcal{L}$ **as an approximation to** $p(M|\mathbf{t})$



Expectation Propagation

Consider an exponential family

$$q(\mathbf{z}) = h(\mathbf{z})g(\boldsymbol{\eta})\exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{z})\right\}.$$

$$\mathrm{KL}(p\|q) = -\ln g(\boldsymbol{\eta}) - \boldsymbol{\eta}^{\mathrm{T}}\mathbb{E}_{p(\mathbf{z})}[\mathbf{u}(\mathbf{z})] + \mathrm{const}$$

$$-\nabla \ln g(\boldsymbol{\eta}) = \mathbb{E}_{p(\mathbf{z})}[\mathbf{u}(\mathbf{z})].$$

Thus we get that

$$\mathbb{E}_{q(\mathbf{z})}[\mathbf{u}(\mathbf{z})] = \mathbb{E}_{p(\mathbf{z})}[\mathbf{u}(\mathbf{z})]. \quad \textbf{(*)}$$

If $q(\mathbf{z})$ **is Gaussian** $\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$**, then (*) is equivalent to equating moments**

Expectation Propagation

General case. Data distribution $p(\mathcal{D}, \boldsymbol{\theta}) = \prod_i f_i(\boldsymbol{\theta}).$

E.g. $f_n(\boldsymbol{\theta}) = p(\mathbf{x}_n|\boldsymbol{\theta})$ **with a prior** $f_0(\boldsymbol{\theta}) = p(\boldsymbol{\theta})$

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{1}{p(\mathcal{D})} \prod_i f_i(\boldsymbol{\theta})$$

$$p(\mathcal{D}) = \int \prod_i f_i(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}.$$

$$q(\boldsymbol{\theta}) = \frac{1}{Z} \prod_i \widetilde{f}_i(\boldsymbol{\theta})$$

We constrain $\widetilde{f}_i(\boldsymbol{\theta})$ **in some way. In particular, we assume that they are from the exponential family**

Expectation Propagation

$$\mathrm{KL}\left(p\|q\right) = \mathrm{KL}\left(\frac{1}{p(\mathcal{D})}\prod_i f_i(\boldsymbol{\theta})\middle\|\frac{1}{Z}\prod_i \widetilde{f}_i(\boldsymbol{\theta})\right).$$

- Minimizing KL divergence is intractable as KL divergence involves averaging w.r.t. the true distribution
- Expectation propagation optimizes each factor while fixing others
- We initialize factors
- We would like to find $\widetilde{f}_i(\boldsymbol{\theta})$, such that

$$q^{\mathrm{new}}(\boldsymbol{\theta}) \propto \widetilde{f}_j(\boldsymbol{\theta})\prod_{i\neq j}\widetilde{f}_i(\boldsymbol{\theta})$$

is as close as possible to $f_j(\boldsymbol{\theta})\prod_{i\neq j}\widetilde{f}_i(\boldsymbol{\theta})$

- This ensures the approximation is most accurate in the regions of high posterior probability, defined by the remaining factors

Expectation Propagation

First remove the factor from the current approximation and get the unnormalized distribution

$$q^{\backslash j}(\boldsymbol{\theta}) = \frac{q(\boldsymbol{\theta})}{\widetilde{f}_j(\boldsymbol{\theta})}.$$

Combine with the real factor $\frac{1}{Z_j} f_j(\boldsymbol{\theta}) q^{\backslash j}(\boldsymbol{\theta})$

having the normalization constant

$$Z_j = \int f_j(\boldsymbol{\theta}) q^{\backslash j}(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}.$$

Revise the factor by minimizing

$$\mathrm{KL}\left(\frac{f_j(\boldsymbol{\theta}) q^{\backslash j}(\boldsymbol{\theta})}{Z_j} \middle\| q^{\mathrm{new}}(\boldsymbol{\theta}) \right).$$

Expectation Propagation

The revised factor can be obtained as

$$\widetilde{f}_j(\boldsymbol{\theta}) = K \frac{q^{\text{new}}(\boldsymbol{\theta})}{q^{\backslash j}(\boldsymbol{\theta})}$$

We can determine normalizing factor as

$$K = \int \widetilde{f}_j(\boldsymbol{\theta}) q^{\backslash j}(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}$$

Thus we can find *K* by matching zero-order moments

$$\int \widetilde{f}_j(\boldsymbol{\theta}) q^{\backslash j}(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} = \int f_j(\boldsymbol{\theta}) q^{\backslash j}(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}.$$

Expectation Propagation

Input: $p(\mathcal{D}, \boldsymbol{\theta}) = \prod_i f_i(\boldsymbol{\theta})$

Task: approximate by $q(\boldsymbol{\theta}) = \frac{1}{Z} \prod_i \widetilde{f}_i(\boldsymbol{\theta}).$

1. **Initialize** $\widetilde{f}_i(\boldsymbol{\theta}).$

2. **Initialize posterior calculation** $q(\boldsymbol{\theta}) \propto \prod_i \widetilde{f}_i(\boldsymbol{\theta}).$

   a. **Choose a** $\widetilde{f}_j(\boldsymbol{\theta})$ **factor to update**

   b. **Remove** $\widetilde{f}_j(\boldsymbol{\theta})$ **from the posterior** $q^{\backslash j}(\boldsymbol{\theta}) = \frac{q(\boldsymbol{\theta})}{\widetilde{f}_j(\boldsymbol{\theta})}.$

Expectation Propagation

Input: $p(\mathcal{D}, \boldsymbol{\theta}) = \prod_i f_i(\boldsymbol{\theta})$

Task: approximate by $q(\boldsymbol{\theta}) = \frac{1}{Z} \prod_i \widetilde{f}_i(\boldsymbol{\theta}).$

c. **Evaluate the new posterior by setting the moments of** $q^{\text{new}}(\boldsymbol{\theta})$ **equal to the moments of** $q^{\setminus j}(\boldsymbol{\theta}) f_j(\boldsymbol{\theta})$**, evaluate**

$$Z_j = \int q^{\setminus j}(\boldsymbol{\theta}) f_j(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}.$$

c. **Evaluate and store the new factor** $\widetilde{f}_j(\boldsymbol{\theta}) = Z_j \frac{q^{\text{new}}(\boldsymbol{\theta})}{q^{\setminus j}(\boldsymbol{\theta})}.$

4. **Calculate Model Evidence** $p(\mathcal{D}) \simeq \int \prod_i \widetilde{f}_i(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}.$