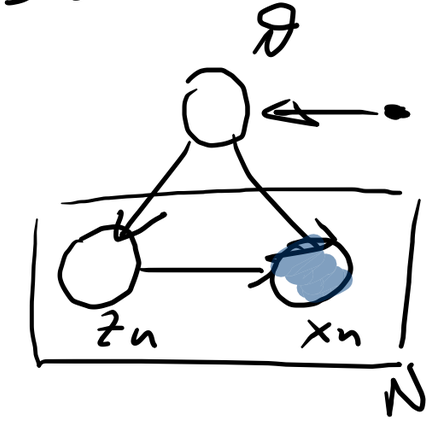
VARIATIONAL INFERENCE

$p(x, z)$, with x observable variables
with z latent variables $\begin{cases} \text{local latent variables } z_n \\ \text{global latent variables } \vartheta \end{cases}$

$z = (z_1, \dots, z_N, \vartheta)$



We have observation $x_1, \dots x_N = \underline{x}$ (sometimes x)

$p(z \mid \underline{x})$ posterior

$p(\underline{x}) = \int p(\underline{x}, t) dz$ evidence

ELBO

$$\underbrace{p(\underline{x})}_{\text{fixed}} = \underbrace{\mathcal{L}(q)}_{} + \overbrace{KL[q \| p]}^{KL[q(\underline{z}) \| p(\underline{z}|\underline{x})]}$$

≥ 0

$$\mathcal{L}(q) = \int q(\underline{z}) \left[\log p(\underline{x}, \underline{z}) - \log q(\underline{z}) \right] d\underline{z} = \mathbb{E}_q \left[\log p(\underline{x}, \underline{z}) - \log q(\underline{z}) \right]$$

$$KL[q(\underline{z}) \| p(\underline{z}|\underline{x})] = -\int q(\underline{z}) \left[\log p(\underline{z}|\underline{x}) - \log q(\underline{z}) \right] d\underline{z}$$

q - VARIATIONAL DISTRIBUTION

FIND THE BEST q THAT APPROXIMATES $p(\underline{z}|\underline{x})$

minimize $KL[q \| p] \Leftrightarrow$ maximize ELBO $\mathcal{L}(q)$

$\mathcal{L}(q)$ is maximum when $KL[q \| p] = 0 \Leftrightarrow q(\underline{z}) = p(\underline{z}|\underline{x})$

Point $p(z|x)$ is intractable

Solution: RESTRICT q to a TRACTABLE FAMILY OF DISTRIBUTIONS!

The optimal $q(z)$ likely is s.t. $KL[q||p] > 0$

VARIATIONAL INFERENCE => solve an optimization problem to maximize $f(q)$ in a suitable space of variational distributions $Q \ni q$

SCENARIO: Q = set of parametric distributions $q(z|\lambda)$, λ parameters in $\mathbb{R}^K$

Then we need to find $\hat{\lambda} = \arg\max f(\lambda)$

This is a highly non-linear non-convex optimization

$$q(z|\lambda) = q(\vartheta|\lambda_\vartheta) \prod_i q(z_i|\lambda_i, \vartheta) \qquad \lambda = (\lambda_\vartheta, \lambda_1, \ldots, \lambda_N)$$

VARIATIONAL MEAN FIELD

$p(z|x)$ by $q(z)$. $z = (z_1, \dots, z_M)$ decomposition

$$q(z) = \underbrace{\prod_{i=1}^{M} q_i(z_i)}_{\text{MEAN FIELD ASSUMPTION}} \qquad q_i \text{ for simplicity}$$

$$J(q) = \mathbb{E}_q\left[\log p(x,z) - \log q(z)\right]$$

$$= \int \prod_i q_i \left[\log p(x,z) - \sum_i \log q_i\right] dz$$

factorized terms q_j

$$= \int q_j \left[\int \log p(x,z) \prod_{i \neq j} q_i \, dz_i\right] dz_j - \int q_j \log q_j \, dz_j + \text{const}$$

$$\mathbb{E}_{i \neq j}\left[\log p(x,z)\right]$$

$$= \mathbb{E}_{q_j}\left[\log \tilde{p}(x, z_j) - \log q_j\right] + \text{const}$$

defines $\tilde{p}(x, z_j)$

$$\log \tilde{p}(x, z_j) = \mathbb{E}_{i \neq j}\left[p(x,z)\right] + \text{const}$$

$$f(q_j) = KL\left[q_j \,\|\, \tilde{p}(x, z_j)\right] + \text{const}$$

p

$\tilde{p}(x, z_j)$ as

$$\log \tilde{p}(x, z_j) = \mathbb{E}_{i \neq j}\left[\log p(x, z)\right] + \text{const}$$

This is maximised for q_j, with q_i, $i \neq j$ FIXED

by $q_j^*(z_j) = \tilde{p}(x, z_j)$

Hence $\log q_j^*(z_j) = \mathbb{E}_{i \neq j}\left[\log p(x, z)\right] + \text{const}$

$$\Rightarrow q_j^*(z_j) = \frac{\exp\left(\mathbb{E}_{i \neq j}\left[\log p(x, z)\right]\right)}{\int \exp\left(\mathbb{E}_{i \neq j}\left[\log p(x, z)\right]\right) dz_j}$$

IF we can compute (analytically) $\mathbb{E}_{i \neq j}\left[\log p(x, z)\right]$

we initialize q_i, then cycle through q_j, optimize them in turn

(COORDINATE GRADIENT ASCEND), until convergence

The lower bound is convex on q_j

A SIMPLE EXAMPLE

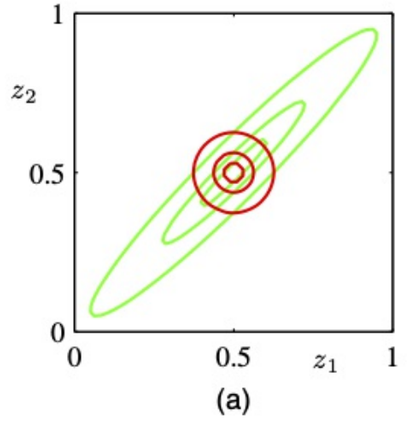
Gaussian to factorized gaussian $\quad p(z) = \mathcal{N}(z \mid \mu, \Lambda^{-1})$, $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$, Λ precision matrix

$z = (z_1, z_2)$ $\qquad$ MFVB $\quad q(z) = q(z_1) q(z_2)$

$$\ln q_1^*(z_1) = \mathbb{E}_{z_2}\left[\ln p(z)\right] + \text{const}$$

$$= \mathbb{E}_{z_2}\left[-\frac{1}{2}(z_1-\mu_1)^2 \Lambda_{11} - (z_1-\mu_1)\Lambda_{12}(z_2-\mu_2)\right] + \text{const}$$

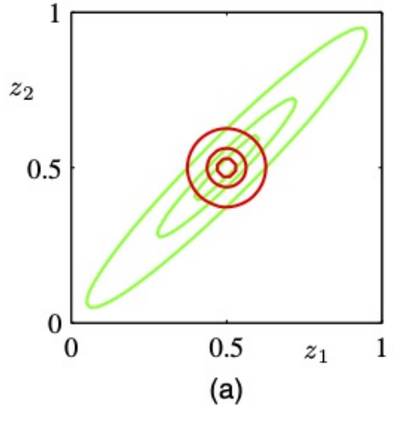
$$= -\frac{1}{2}(z_1-\mu_1)^2 \Lambda_{11} - (z_1-\mu_1)\Lambda_{12}\left(\mathbb{E}[z_2]-\mu_2\right) + \text{const}$$



$\Rightarrow q_1^*(z_1)$ is Gaussian $\mathcal{N}(z_1 \mid m_1, \Lambda_{11}^{-1})$ $\quad m_1 = \mu_1 - \Lambda_{11}^{-1}\Lambda_{12}\left(\mathbb{E}[z_2]-\mu_2\right)$

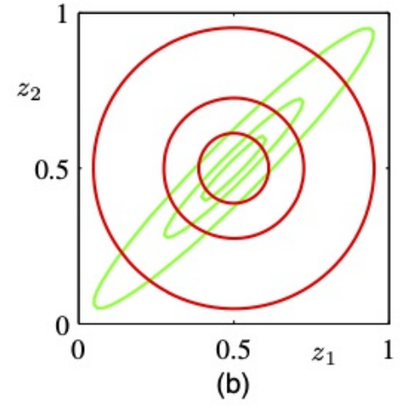
$\Rightarrow q_2^*(z_2) = \mathcal{N}(z_2 \mid m_2, \Lambda_{22}^{-1})$, $m_2 = \mu_2 - \Lambda_{22}^{-1}\Lambda_{21}\left(\mathbb{E}[z_1]-\mu_1\right)$

$m_1 : \mathbb{E}[z_1] \Rightarrow m_1 = \mu_1 \quad m_2 = \mu_2$



(a)

$q(z) = q(z_1)q(z_2)$



(b)

$KL[q\|p]$

ZERO FORCING

if $p(z) \simeq 0$ then $q(z) \simeq 0$

one mode

$KL[p\|q]$

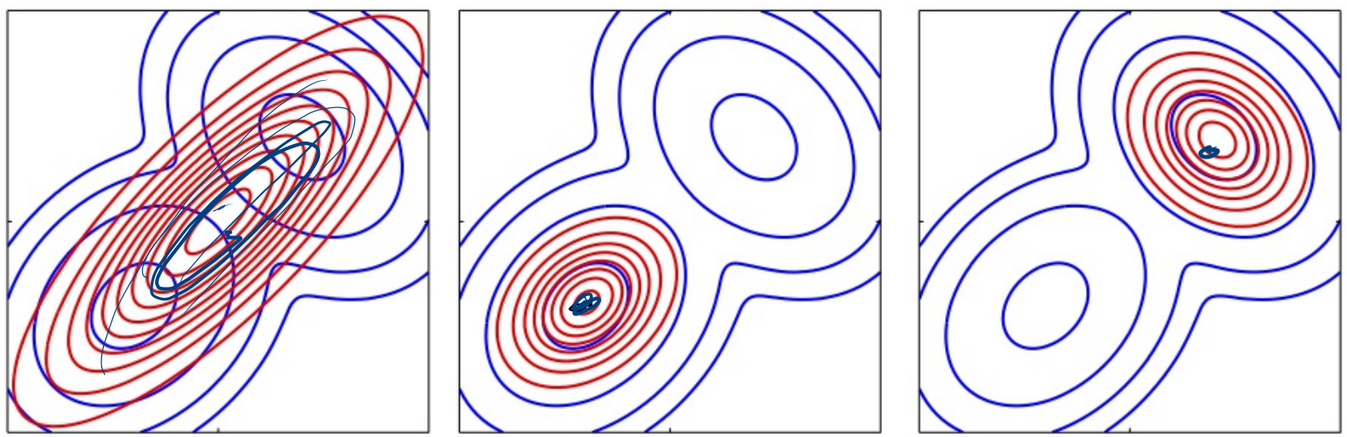
ZERO AVOIDING

$q(z)$ is non-zero where

$p(z)$ is non-zero

overlaps modes

$KL[p\|q]$



(a) (b) (c)

$KL[q\|p]$

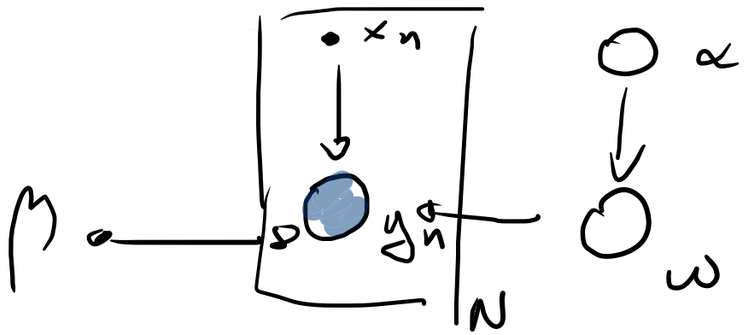
VARIATIONAL LINEAR REGRESSION

$$p(y, w, \alpha) = p(y|w)\, p(w|\alpha)\, p(\alpha)$$

$$p(y|w) = \prod_{n=1}^{N} \mathcal{N}(y_n | w^T \phi(x_n), \beta^{-1})$$

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

$$p(\alpha) = \text{GAMMA}(\alpha | a_0, b_0) = \frac{1}{\Gamma(a_0)} b_0^{a_0} \alpha^{a_0 - 1} e^{-b_0 \alpha}$$



GOAL

$p(w, \alpha | y)$ $\quad$ $q(w, \alpha) = q(w)\, q(\alpha)$

$$q^*(\alpha) = \mathbb{E}_w [\log p(y, w, \alpha)] + \text{const}$$
$$\overset{!}{=} \log p(\alpha) + \mathbb{E}_w [\log p(w|\alpha)] + \text{const}$$
$$\overset{!}{=} (a_0 - 1) \log \alpha - b_0 \alpha + \frac{M}{2} \log \alpha - \frac{\alpha}{2} \mathbb{E}[w^T w] + \text{const}$$

$$q^*(\alpha) = \text{GAMMA}(\alpha \mid a_N, b_N) \qquad a_N = a_0 + \frac{M}{2}$$

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

$$\log q^*(w) = \log p(y|w) + \mathbb{E}_\alpha[\log p(w|\alpha)] + \text{const}$$

$$\overset{!}{=} -\frac{\beta}{2}\sum_{n=1}^{N}\left[w^T\phi(x_n) - y_n\right]^2 - \frac{1}{2}\mathbb{E}[\alpha]\, w^T w + \text{const}$$

$$\Rightarrow q^*(w) = \mathcal{N}(w \mid m_N, S_N) \qquad m_N = \beta S_N \Phi^T y$$

$$S_N = \left(\mathbb{E}[\alpha] I + \beta \Phi^T \Phi\right)^{-1}$$

$$\mathbb{E}[\alpha] = a_N/b_N \,, \quad \mathbb{E}[w^T w] = m_N^T m_N + \text{Trace}(S_N)$$

We can compute $\mathcal{L}(q) \approx p(y)$ (approximates model evidence)

BLACK BOX VARIATIONAL INFERENCE

$q(z|\lambda)$ parametric distribution

$$f(\lambda) = \mathbb{E}_{q(z|\lambda)}\left[\log p(x,z) - \log q(z|\lambda)\right]$$

we want to compute $\nabla_\lambda f(\lambda)$

SOLUTION 1: Reparametrization TRICK

If $z = g_\nu(\varepsilon)$, $\varepsilon \sim \hat{q}(\varepsilon)$ (independ λ and fixed), $\lambda = (\lambda, \nu)$

$$f(\tilde{\lambda}) = \mathbb{E}_{\hat{q}(\varepsilon)}\left[\log p(x, g_\nu(\varepsilon)) - \log q(g_\nu(\varepsilon)|\lambda)\right]$$

$$\nabla_{\tilde{\lambda}} f(\tilde{\lambda}) = \mathbb{E}_{\hat{q}(\varepsilon)}\left[\underbrace{\nabla_{\tilde{\lambda}} \log p(x, g_\nu(\varepsilon)) - \nabla_{\tilde{\lambda}} q(g_\nu(\varepsilon)|\lambda)}_{G(\varepsilon)}\right]$$

$\varepsilon_s \sim \hat{q}(\varepsilon)$ $\quad \nabla_{\tilde{\lambda}} f(\tilde{\lambda}) = \frac{1}{S}\sum_{s=1}^{S} G(\varepsilon_s)$ ← Thousand SGA stock gradient ascent

case 2 : non-reparameterizable $q(z|\lambda)$

$$\nabla_\lambda \mathcal{L}(\lambda) = \nabla_\lambda \mathbb{E}_{q(z|\lambda)}\left[\log p(x,z) - \log q(z|\lambda)\right]$$

$$= \nabla_\lambda \int q(z|\lambda)\left[\log p(x,z) - \log q(z|\lambda)\right] dz$$

$$= \int \nabla_\lambda \left[\log p(x,z) - \log q(z|\lambda)\right] q(z|\lambda)\, dz +$$

$$+ \int \nabla_\lambda q(z|\lambda)\left[\log p(x,z) - \log q(z|\lambda)\right] dz$$

$\nabla_\lambda p(x,z) = 0$

$$= -\mathbb{E}_q\left[\nabla_\lambda \log q(z|\lambda)\right] +$$

$$\mathbb{E}_q\left[\frac{\nabla_\lambda q(z|\lambda)}{q(z|\lambda)}\right] = \int \nabla_\lambda q(z|\lambda)\, dz = \nabla_\lambda \int q(z|\lambda)\, dz = \nabla_\lambda 1 = 0$$

$$\nabla_\lambda \log q(z|\lambda) = \frac{\nabla_\lambda q(z|\lambda)}{q(z|\lambda)} \Rightarrow \nabla_\lambda q(z|\lambda) = \nabla_\lambda \log q(z|\lambda) \cdot q(z|\lambda)$$

$$\nabla_\lambda f(\lambda) = \int q(t|\lambda)\, \nabla_\lambda \log q(t|\lambda) \left[\log p(x,t) - \log q(t|\lambda)\right] dt$$

$$= E_q\left[\nabla_\lambda \log q(t|\lambda) \left[\log p(x,t) - \log q(t|\lambda)\right]\right]$$

$$t_s \sim q(t|\lambda) \Rightarrow \nabla_\lambda f_\lambda \approx \frac{1}{S}\sum_{s=1}^{S} \nabla_\lambda \log(t_s|\lambda)\left[\log p(x,t_s) - \log q(t_s|\lambda)\right] \quad (*)$$

VARIANCE REDUCTION

ISSUE: The estimator $(*)$ typically has HIGH VARIANCE

→ RAO-BLACKWELLIZATION

→ CONTROL VARIATES

RAO-BLACKWELLIZATION

x, y random variates $\quad f(x,y)$, $\mathbb{E}[f(x,y)]$ is our target

a) $\hat{f}(x) = \mathbb{E}_y[f(x,y)\,|\,x]$ s.t. $\mathbb{E}_x[\hat{f}(x)] = \mathbb{E}_{x,y}[f(x,y)]$

$$\mathrm{VAR}[\hat{f}(x)] = \mathrm{VAR}[f(x,y)] - \mathbb{E}[(f(x,y) - \hat{f}(x))^2] < \mathrm{VAR}[f(x,y)]$$

Mean field factorization of $\quad q(z|\lambda) = \prod_{i=1}^{N} q(z_i|\lambda_i)$

- $q_{(i)}$: marginal of $q(z|\lambda)$ on the variables blanketed : $z_{(i)}$ of z_i (in p)

$p_i(x, z_{(i)})$ the product of factors of $p(x,z)$ depending on $z_{(i)}$

$$\mathbb{E}_{q_{(i)}}\mathbb{E}_{q_i}\left[\nabla_{\lambda_i} \mathcal{L}\,|z_{(i)}\right] \quad \hat{\nabla}_{\lambda_i}\mathcal{L} = \mathbb{E}_{q_{(i)}}\left[\nabla_{\lambda_i} \log q(z_i|\lambda_i)\left(\log p_i(x, z_{(i)}) - \log q(z_i|\lambda_i)\right)\right]$$

$$z_s \sim q_{(i)}(z|\lambda)$$

CONTROL VARIATES

f: $E_q[f]$ our goal. Instead $E_q[\hat{f}]$ s.t. $E_q[\hat{f}] = E[f]$ and $VAR_q[\hat{f}] < VAR_q[f]$

we choose h: $E[h] < \infty$, define

$$\hat{f}_a(z) = f(z) - a\left(h(z) - E[h(z)]\right)$$

The a^* the minimizes variance is $a^* = \dfrac{cov(f,h)}{VAR(h)}$

$$f_i(z) = \nabla_{\lambda_i} \log q(z_i|\lambda_i)\left[\log p_i(x, z_{(i)}) - \log(z_i|\lambda_i)\right]$$

$$h_i(z) = \underbrace{\nabla_{\lambda_i} \log q(z_i|\lambda_i)}_{E[\cdot]=0},$$

$\#z_{(i)} = n_i$

$$\Rightarrow \hat{a}_i^* = \frac{\sum_{d=1}^{n_i} \hat{cov}(f_i^d, h_i^d)}{\sum_{d=1}^{n_i} \hat{VAR}(h_i^d)}$$

$$\hat{\nabla}_{\lambda_i} f \triangleq \frac{1}{S}\sum_{s=1}^{S} \nabla_{\lambda_i} \log q(z_s|\lambda_i)\left[\log p_i(x, z_s) - \log q_i(z_s|\lambda_i) - \hat{a}_i^*\right]$$

$$z_s \sim q_{(i)}(z|\lambda)$$