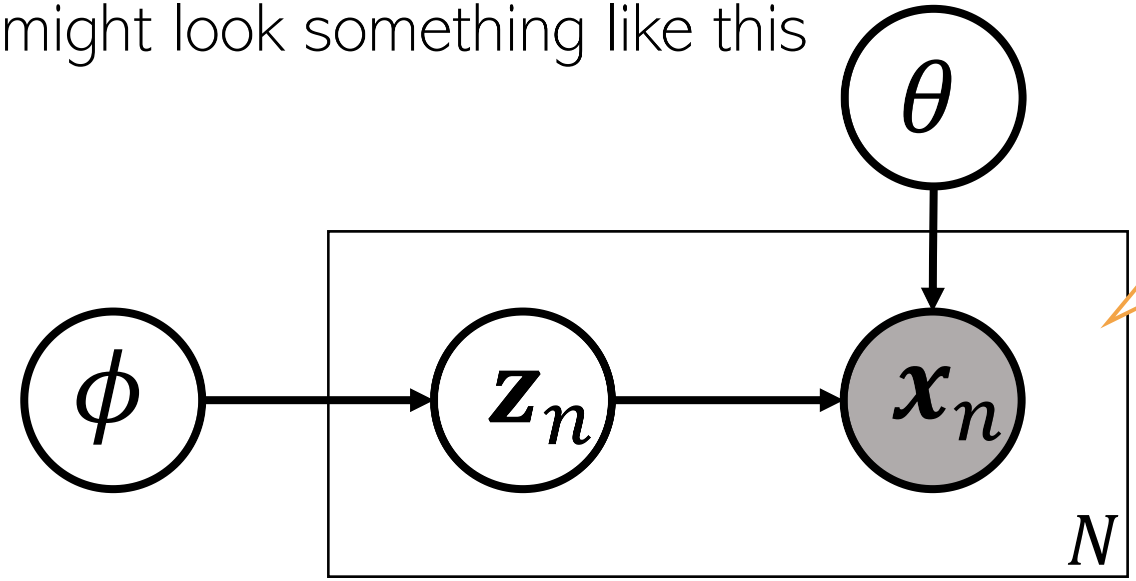
Variational Inference

CS772A: Probabilistic Machine Learning
Piyush Rai

Variational Inference (VI)

- Assume a latent variable model with data $\mathcal{D}$ and latent variables $\boldsymbol{Z}$
- A simple setting might look something like this



This setting is just one example. VI is applicable in more general and more complex probabilistic models with and without latent variables

- Assume the likelihood is $p(\mathcal{D}|\boldsymbol{Z},\Theta)$ and prior is $p(\boldsymbol{Z}|\Theta)$. Want posterior over $\boldsymbol{Z}$
- $\Theta = (\theta, \phi)$ denotes the other parameters that define the likelihood and the prior
- For now, assume Θ is known and only $\boldsymbol{Z}$ is unknown (the Θ unknown case later)
- Assume CP $p(\boldsymbol{Z}|\mathcal{D},\Theta)$ is intractable

Variational Inference (VI)

- Assuming $p(\boldsymbol{Z}|\boldsymbol{\mathcal{D}},\Theta)$ is intractable, VI approximates it by a distr $q(\boldsymbol{Z}|\phi)$ or $q_\phi(\boldsymbol{Z})$

Find the optimal ϕ which makes our approximation $q(\boldsymbol{Z}|\phi)$ as closed to the true as possible to the true posterior $p(\boldsymbol{Z}|\boldsymbol{\mathcal{D}})$

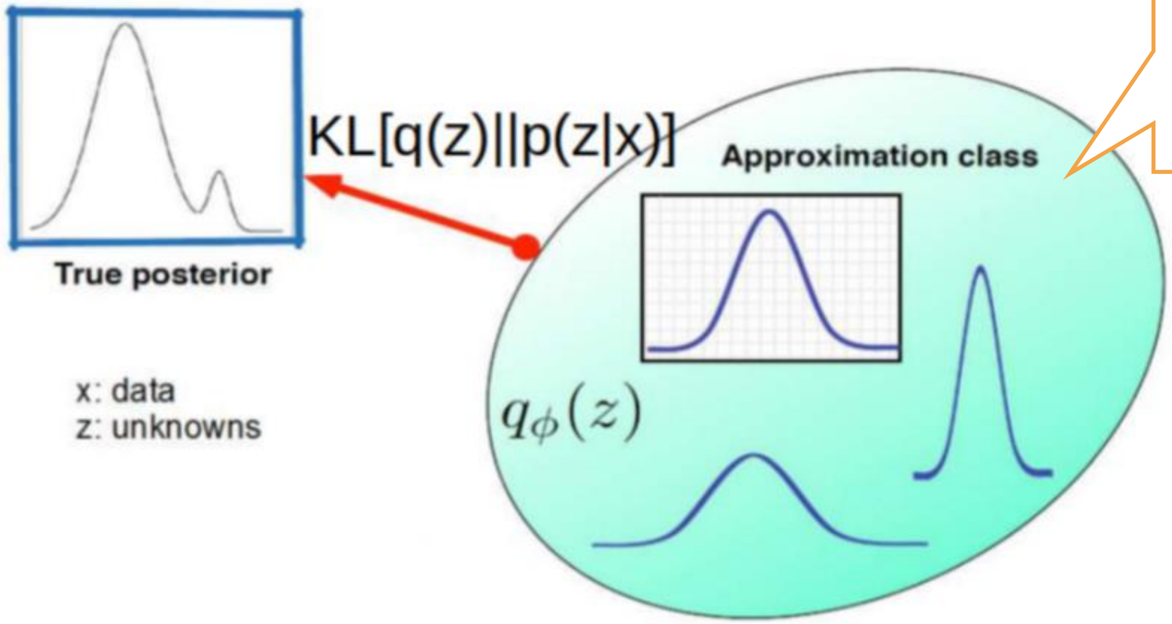
Kullback Leibler divergence $\mathrm{KL}[q||p]$ between q and p

Also possible to use $\mathrm{KL}[p||q]$ or divergences other than KL

$$\phi^* = \operatorname{argmin}_\phi \mathrm{KL}[q_\phi(\boldsymbol{Z})||p(\boldsymbol{Z}|\boldsymbol{\mathcal{D}},\Theta)]$$

q_ϕ defines a class of distributions parametrized by ϕ sometimes called "variational parameters"

Name "variational" comes from Physics and refers to problems where we are optimizing functions of distributions (here the function is the KL divergence)



Variational Inference (VI)

- The optimization problem

$$\phi^* = \text{argmin}_\phi \, \text{KL}[q_\phi(\boldsymbol{Z}) || p(\boldsymbol{Z}|\boldsymbol{\mathcal{D}}, \Theta)]$$

$$= \text{argmin}_\phi \, \mathbb{E}_{q_\phi(\boldsymbol{Z})}\left[\log q_\phi(\boldsymbol{Z}) - \log \frac{p(\boldsymbol{\mathcal{D}}|\boldsymbol{Z}, \Theta) p(\boldsymbol{Z}|\Theta)}{p(\boldsymbol{\mathcal{D}}|\Theta)}\right]$$

$$= \text{argmin}_\phi \, \mathbb{E}_{q_\phi(\boldsymbol{Z})}[\log q_\phi(\boldsymbol{Z}) - \log p(\boldsymbol{\mathcal{D}}|\boldsymbol{Z}, \Theta) - \log p(\boldsymbol{Z}|\Theta)] + \log p(\boldsymbol{\mathcal{D}}|\Theta)$$

- Since $\log p(\boldsymbol{\mathcal{D}}|\Theta)$ is independent of ϕ, the optimization problem becomes

$$\phi^* = \text{argmin}_\phi \, \mathbb{E}_{q_\phi(\boldsymbol{Z})}[\log q_\phi(\boldsymbol{Z}) - \log p(\boldsymbol{\mathcal{D}}|\boldsymbol{Z}, \Theta) - \log p(\boldsymbol{Z}|\Theta)]$$

$$\phi^* = \text{argmin}_\phi \, \mathbb{E}_{q_\phi(\boldsymbol{Z})}[\log q_\phi(\boldsymbol{Z}) - \log p(\boldsymbol{\mathcal{D}}, \boldsymbol{Z}|\Theta)]$$

$$\phi^* = \text{argmax}_\phi \, \mathbb{E}_{q_\phi(\boldsymbol{Z})}[\log p(\boldsymbol{\mathcal{D}}, \boldsymbol{Z}|\Theta) - \log q_\phi(\boldsymbol{Z})] = \text{argmax} \, \mathcal{L}(\phi, \Theta)$$

- Note that $\mathcal{L}(\phi, \Theta) \leq \log p(\boldsymbol{\mathcal{D}}|\Theta)$ and is called "Evidence Lower Bound" (ELBO)

The ELBO

- The ELBO is defined as

$$\mathcal{L}(\phi, \Theta) = \mathbb{E}_{q_\phi(\boldsymbol{Z})}[\log p(\boldsymbol{\mathcal{D}}, \boldsymbol{Z}|\Theta) - \log q_\phi(\boldsymbol{Z})]$$

$$= \mathbb{E}_{q_\phi(\boldsymbol{Z})}[\log p(\boldsymbol{\mathcal{D}}, \boldsymbol{Z}|\Theta)] + \mathrm{H}[q_\phi(\boldsymbol{Z})]$$

- Thus maximizing the ELBO w.r.t. ϕ gives us a $q_\phi(\boldsymbol{Z})$ which
  - Maximizes the expected joint probability of data and latent variables
  - Has a high entropy
- We can also write the ELBO as follows

$$\boxed{\mathcal{L}(\phi, \Theta) = \mathbb{E}_{q_\phi(\boldsymbol{Z})}[\log p(\boldsymbol{\mathcal{D}}|\boldsymbol{Z}, \Theta)] - \mathrm{KL}[q_\phi(\boldsymbol{Z})||p(\boldsymbol{Z}|\Theta)]}$$

- Thus maximizing the ELBO w.r.t. ϕ will give us a $q_\phi(\boldsymbol{Z})$ which
  - Explains the data $\boldsymbol{\mathcal{D}}$ well, i.e., gives it large expected probability $\mathbb{E}_q[\log p(\boldsymbol{\mathcal{D}}|\boldsymbol{Z}, \Theta)]$
  - Is close to the prior $p(\boldsymbol{Z})$, i.e. is simple/regularized (small $\mathrm{KL}[q_\phi(\boldsymbol{Z})||p(\boldsymbol{Z}|\Theta)]$)

Maximizing the ELBO

Unknown Θ case later

- We need to maximize the ELBO w.r.t. ϕ (for now, assuming Θ is known)

$$\boxed{\mathcal{L}(\phi, \Theta) = \mathbb{E}_{q_\phi(\boldsymbol{Z})}[\log p(\boldsymbol{\mathcal{D}}|\boldsymbol{Z}, \Theta)] - \text{KL}[q_\phi(\boldsymbol{Z})||p(\boldsymbol{Z}|\Theta)]}$$

- The general approach to maximize ELBO is based on gradient-based methods
  - Assume some suitable/convenient form for $q_\phi(\boldsymbol{Z})$, e.g., $\mathcal{N}(\boldsymbol{Z}|\mu, \Sigma)$ so $\phi = (\mu, \Sigma)$
  - Maximize the ELBO w.r.t. ϕ using gradient ascent

$$\phi_{t+1} = \phi_t + \eta_t \nabla_{\phi_t} \mathcal{L}(\phi, \Theta)$$

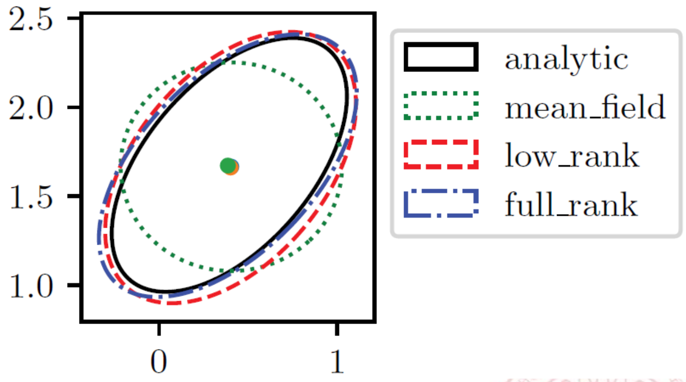
- Note: Expectations in ELBO and ELBO's gradients w.r.t. ϕ may not be easy
  - Will see methods to handle such issues later
  - Assuming simple forms for $q_\phi(\boldsymbol{Z})$ also helps (we can use random variable transformation methods to transform the simple form to more expressive ones – will see later)

A Simple Illustration for VI

- Assume a simple likelihood model

$$p(\boldsymbol{\mathcal{D}}|\boldsymbol{z}) = \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{z},\boldsymbol{\Sigma}) \propto \mathcal{N}(\bar{\boldsymbol{x}}|\boldsymbol{z},\frac{1}{N}\boldsymbol{\Sigma})$$

- Suppose we want to estimate the posterior of the mean $\boldsymbol{z}$
- Assuming a Gaussian prior on $\boldsymbol{z}$ and assuming $\boldsymbol{\Sigma}$ is known, the posterior can be computed analytically (because of conjugacy)
- Let's still try VI to see how well it does
- Figure shows VI result for three Gaussian forms for $\boldsymbol{q(z)}$
  - Low-rank: $q(z) = \mathcal{N}(z|\mu_z,\Sigma_z)$ where $\Sigma_z = LL^\top$
  - Full-rank: $q(z) = \mathcal{N}(z|\mu_z,\Sigma_z)$ with no constraint on Σ_z
  - Mean-field: $q(z) = q(z_1)q(z_2) = \mathcal{N}\left(z_1|\mu_{z_1},\sigma_{z_1}^2\right)\mathcal{N}\left(z_2|\mu_{z_2},\sigma_{z_2}^2\right)$



(Example courtesy: PML-2 (Murphy))

Detour

Transformed random variable

A one-to-one transformation function

- Consider a scalar transformation of a scalar random variable $\boldsymbol{u}$ as $\theta = T(u)$
- Probability distributions of random variables $\boldsymbol{u}$ and $\boldsymbol{\theta}$ are related as

$$p(\theta) = p(u)\left|\frac{du}{d\theta}\right|$$

- Similarly, for multivariate random variables (of same size) related as $\boldsymbol{\theta} = T(\boldsymbol{u})$

$$p(\boldsymbol{\theta}) = p(\boldsymbol{u})\left|\det\left(\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{\theta}}\right)\right|$$

Absolute value of the determinant of the Jacobian (note that $\boldsymbol{u} = T^{-1}(\boldsymbol{\theta})$)

- We can use such transformations for VI by using a simple distribution for $q(\boldsymbol{Z})$ and then transform it to a more expressive/appropriate distribution (more on this later)

Mean-Field VI

- A special way to maximize the ELBO is via the mean-field approximation
- Doesn't require specifying the form of $q(\boldsymbol{Z}|\phi)$ or computing ELBO's gradients
- The idea: Assumes unknowns $\boldsymbol{Z}$ can be partitioned into M groups $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_M$, s.t.,

As a shorthand, often written as $q = \prod_{i=1}^{M} q_i$ where $q_i = q(Z_i|\phi_i)$

$$q(\boldsymbol{Z}|\phi) = \prod_{i=1}^{M} q(\boldsymbol{Z}_i|\phi_i)$$

For models with local conjugacy, it becomes super easy!

- Learning the optimal $q(\boldsymbol{Z}|\phi)$reduces to learning the optimal $q_1, q_2, \ldots, q_M$

- Can select groupsbased on model's structure, e.g., in Bayesian neural net for regression

$$p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{y}, \lambda, \beta) \approx q(\boldsymbol{w}|\phi) = \prod_{\ell=1}^{L} q(w^{(\ell)}|\phi_\ell)$$

Assuming a network with L layers, mean-field across layers

- Mean-field has limitations. Factorized form ignores the correlations among unknowns
  - Variants such as "structured mean-field" exist where some correlations can be modeled

Deriving Mean-Field VI Updates

Writing this is the same as $\arg\max_{\phi} \mathcal{L}(\phi, \Theta)$. We are just writing optimization w.r.t. q directly

- With $q = \prod_{i=1}^{M} q_i$, what's the optimal q_i when we do $\arg\max_q \mathcal{L}(q)$?
- Note that under this mean-field assumption, the ELBO simplifies to

$$\mathcal{L}(q) = \int q(\boldsymbol{Z}) \log \left[\frac{p(\mathcal{D}, \boldsymbol{Z}|\Theta)}{q(\boldsymbol{Z})} \right] d\boldsymbol{Z} = \int \prod_i q_i \left[\log p(\mathcal{D}, \boldsymbol{Z}|\Theta) - \sum_i \log q_i \right] d\boldsymbol{Z}$$

- Suppose we wish to find the optimal q_j given all other q_i's $(i \neq j)$ as fixed, then

$$\mathcal{L}(q) = \int q_j \left[\int \log p(\mathcal{D}, \boldsymbol{Z}|\Theta) \prod_{i \neq j} q_i \, dZ_i \right] dZ_j - \int q_j \log q_j dZ_j + \text{const w.r.t. } q_j$$

$$= \int q_j \log \hat{p}(\mathcal{D}, Z_j|\Theta) \, dZ_j - \int q_j \log q_j Z_j$$

$$= -\text{KL}(q_j || \hat{p})$$

$\log \hat{p}(\mathcal{D}, Z_j|\Theta) = \mathbb{E}_{i \neq j}[\log p(\mathcal{D}, \boldsymbol{Z}|\Theta)] + \text{const}$

$$\boxed{q_j^* = \frac{\exp(\mathbb{E}_{i \neq j}[\log p(\mathcal{D}, \boldsymbol{Z}|\Theta)])}{\int \exp(\mathbb{E}_{i \neq j}[\log p(\mathcal{D}, \boldsymbol{Z}|\Theta)] \, d\boldsymbol{Z}_j}}$$

- Thus $q_j^* = \arg\max_{q_j} \mathcal{L}(q) = \arg\min_{q_j} \text{KL}(q_j || \hat{p}) = \hat{p}(\mathcal{D}, Z_j|\Theta)$

Deriving Mean-Field VI Updates

- So we saw that the optimal q_j when doing mean-field VI is

$$q_j^*(\boldsymbol{Z}_j) = \frac{\exp(\mathbb{E}_{i\neq j}[\log p(\boldsymbol{\mathcal{D}}, \boldsymbol{Z}|\Theta)])}{\int \exp(\mathbb{E}_{i\neq j}[\log p(\boldsymbol{\mathcal{D}}, \boldsymbol{Z}|\Theta)]\, d\boldsymbol{Z}_j}$$

- Note: Can often just compute the numerator and recognize denominator by inspection

- Important: For locally conj models, $q_j^*(\boldsymbol{Z}_j)$ will have the same form as prior $p(Z_j|\Theta)$
  - Only the distribution parameters will be different

- Important: For estimating q_j the required expectation depends on other $\{q_i\}_{i\neq j}$
  - Thus we use an alternating update scheme for these
- Guaranteed to converge (to a local optima)
  - We are basically solving a sequence of concave maximization problems
  - Reason: $\mathcal{L}(q) = \int q_j \log \hat{p}(\boldsymbol{\mathcal{D}}, Z_j|\Theta)\, Z_j - \int q_j \log q_j Z_j$ is concave in q_j

The Mean-Field VI Algorithm

- Also known as Co-ordinate Ascent Variational Inference (CAVI) Algorithm
- Input: Model in form of priors and likelihood, or joint $p(\mathcal{D}, \boldsymbol{Z}|\Theta)$, Data $\mathcal{D}$
- Output: A variational distribution $q(\boldsymbol{Z}) = \prod_{j=1}^{M} q_j(\boldsymbol{Z}_j)$
- Initialize: Variational distributions $q_j(\boldsymbol{Z}_j)$, $j = 1,2, \dots M$
- While the ELBO has not converged
  - For each $j = 1,2, \dots M$, set

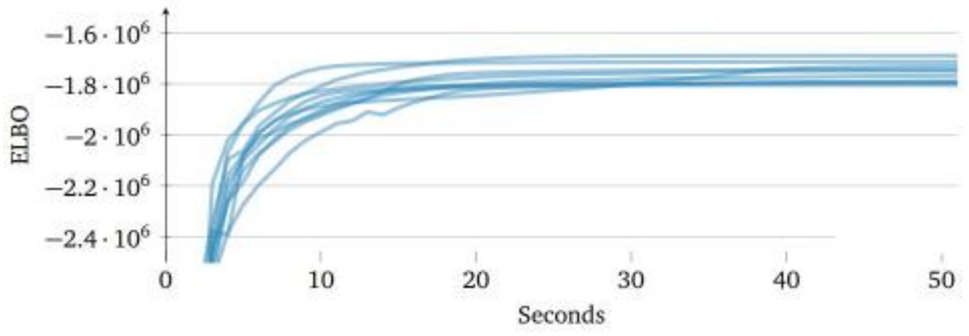
$$q_j(\boldsymbol{Z}_j) \propto \exp(\mathbb{E}_{i \neq j}[\log p(\mathcal{D}, \boldsymbol{Z}|\boldsymbol{\Theta})])$$

  - Compute ELBO $\mathcal{L}(q) = \mathbb{E}_q[\log p(\mathcal{D}, \boldsymbol{Z}|\Theta)] - \mathbb{E}_q[\log q(\boldsymbol{Z})]$

- NOTE: We can also use mean-field assumption for $q(\boldsymbol{Z})$ and optimize the ELBO using gradient based methods if we don't have local conjugacy

VI and Convergence

- VI is guaranteed to converge to a local optima (just like EM)
- Therefore proper initialization is important (just like EM)
  - Can sometimes run multiple times with different initializations and choose the best run



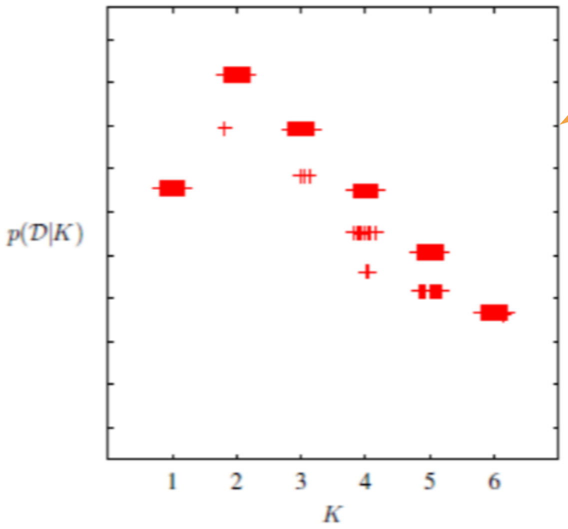
Different initializations may lead to different optima

- ELBO increases monotonically with iterations
  - Can thus monitor the ELBO to assess convergence

ELBO for Model Selection

- Recall that ELBO is a lower bound on log of model evidence $\log p(\boldsymbol{X}|m)$
- Can compute ELBO for each model m and choose the one with largest ELBO

Plot of the variational lower bound $\mathcal{L}$ versus the number K of components in the Gaussian mixture model, for the Old Faithful data, showing a distinct peak at $K = 2$ components. For each value of K, the model is trained from 100 different random starts, and the results shown as '+' symbols plotted with small random horizontal perturbations so that they can be distinguished. Note that some solutions find suboptimal local maxima, but that this happens infrequently.



Each value of K represents a different model

- Some criticism since we are using a lower-bound but often works well in practice

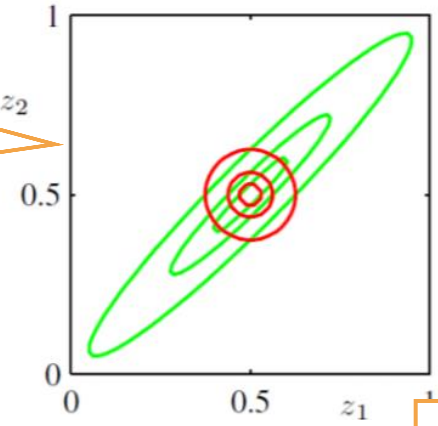
VI might under-estimate posterior's variance

- Recall that VI approximates a posterior $\boldsymbol{p}$ by finding $\boldsymbol{q}$ that minimizes $\mathrm{KL}(q||p)$

$$\mathrm{KL}(q||p) = -\int q(\boldsymbol{Z})\log\left\{\frac{p(\boldsymbol{Z}|\boldsymbol{\mathcal{D}})}{q(\boldsymbol{Z})}\right\}d\boldsymbol{Z}$$

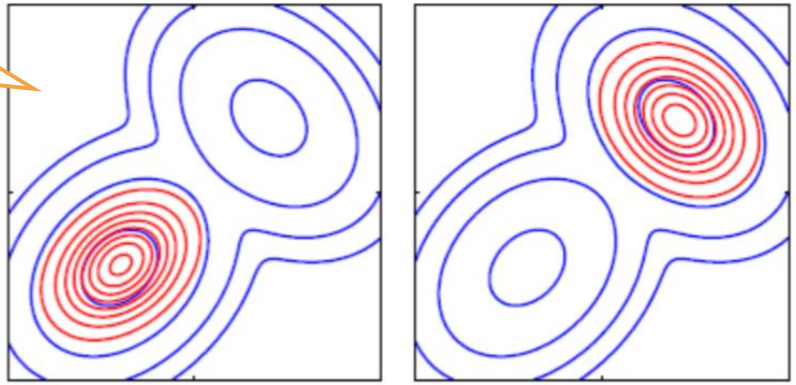
- $\boldsymbol{q(Z)}$ will be small where $\boldsymbol{p(Z|\mathcal{D})}$ is small otherwise KL will blow up
- Thus $\boldsymbol{q(Z)}$ avoids low-probability regions of the true posterior

$\boldsymbol{q}$ (red) avoids regions of $\boldsymbol{p}$ (green) where the latter has low values



$\boldsymbol{q}$ (red) concentrated on one of the modes of $\boldsymbol{p}$ (blue)

For $\boldsymbol{q}$ to also capture the other mode, it will require crossing the low-probability region of $\boldsymbol{p}$, thereby blowing up the KL



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

- Some methods, e.g., Expectation Propagation (EP), can avoid this behavior

Variational EM

- If the parameters Θ are also unknown then we can use variational EM (VEM)
- VEM is the same as EM except the E step uses VI to approximate the CP of $\boldsymbol{Z}$
- VEM alternates between the following two steps
  - Maximize the ELBO w.r.t. ϕ (gives the variational approximation $q(\boldsymbol{Z})$ of CP of $\boldsymbol{Z}$)

$$\phi^{(t)} = \text{argmax}_{\phi}\ \mathbb{E}_{q_\phi(\boldsymbol{Z})}\left[\log p\left(\mathcal{D}, \boldsymbol{Z} \middle| \Theta^{(t-1)}\right) - \log q_\phi(\boldsymbol{Z})\right]$$

  - Maximize the ELBO w.r.t. Θ (gives us point estimate of Θ)

$$\Theta^{(t)} = \text{argmax}_{\Theta}\ \mathbb{E}_{q_{\phi^{(t)}}(\boldsymbol{Z})}\left[\log p(\mathcal{D}, \boldsymbol{Z} | \Theta) - \log q_{\phi^{(t)}}(\boldsymbol{Z})\right]$$

$$= \text{argmax}_{\Theta}\ \mathbb{E}_{q_{\phi^{(t)}}(\boldsymbol{Z})}[\log p(\mathcal{D}, \boldsymbol{Z} | \Theta)]$$

This looks very similar to the expected CLL with the CP replaced by its variational approximation

- Note: If we want posterior for Θ as well, treat it similar to $\boldsymbol{Z}$ and apply variational approximation (instead of using VEM) if the posterior isn't tractable

Extra Slides - Mean-Field VI: A Simple Example

- Consider data $\mathbf{X} = \{x_1, x_2, \ldots, x_N\}$ from a one-dim Gaussian $\mathcal{N}(\mu, \tau^{-1})$
- Assume the following normal-gamma prior on μ and τ

$$p(\mu|\tau) = \mathcal{N}(\mu|\mu_0, (\lambda_0\tau)^{-1}) \quad p(\tau) = \text{Gamma}(\tau|a_0, b_0)$$

- Posterior is also normal-gamma due to the jointly conjugate prior
- Let's anyway verify this by trying mean-field VI for this model
- With mean-field assumption on the variational posterior $q(\mu, \tau) = q_\mu(\mu)q_\tau(\tau)$

$$\begin{aligned}\log q_\mu^*(\mu) &= \mathbb{E}_{q_\tau}[\log p(\mathbf{X}, \mu, \tau)] + \text{const} \\ \log q_\tau^*(\tau) &= \mathbb{E}_{q_\mu}[\log p(\mathbf{X}, \mu, \tau)] + \text{const}\end{aligned}$$

- In this example, the log-joint $\log p(\mathbf{X}, \mu, \tau) = \log p(\mathbf{X}|\mu, \tau) + \log p(\mu|\tau) + \log p(\tau)$. Thus

$$\begin{aligned}\log q_\mu^*(\mu) &= \mathbb{E}_{q_\tau}[\log p(\mathbf{X}|\mu, \tau) + \log p(\mu|\tau)] + \text{const} \qquad \text{(only keeping terms that involve } \mu\text{)} \\ \log q_\tau^*(\tau) &= \mathbb{E}_{q_\mu}[\log p(\mathbf{X}|\mu, \tau) + \log p(\mu|\tau) + \log p(\tau)] + \text{const}\end{aligned}$$

Extra Slides - Mean-Field VI: A Simple Example

- Substituting $p(\mathbf{X}|\mu,\tau) = \prod_{n=1}^{N} p(x_n|\mu,\tau)$ and $p(\mu|\tau)$, we get

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

- (Verify) The above is log of a Gaussian. This $q_\mu^* = \mathcal{N}(\mu|\mu_N,\lambda_N)$ with

$$\mu_N = \frac{\lambda_0\mu_0 + N\bar{x}}{\lambda_0 + N} \quad \text{and} \quad \lambda_N = (\lambda_0 + N)\mathbb{E}_{q_\tau}[\tau]$$

This update depends on q_τ

- Proceeding in a similar way (verify), we can show that $q_\tau^* = \text{Gamma}(\tau|a_N, b_N)$

$$a_N = a_0 + \frac{N+1}{2} \quad \text{and} \quad b_N = b_0 + \frac{1}{2}\mathbb{E}_{q_\mu}\left[\sum_{n=1}^{N}(x_n-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right]$$

This update depends on q_μ

- Note: Updates of q_μ^* and q_τ^* depend on each other (hence alternating updates needed)