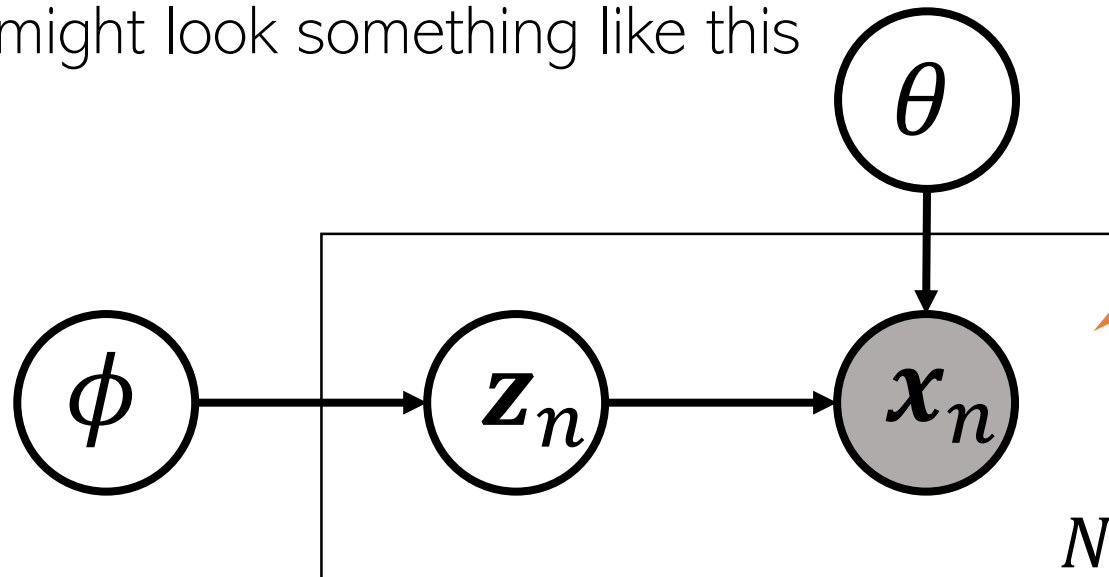


# Variational Inference

# Variational Inference (VI)

- Assume a latent variable model with data  $\mathcal{D}$  and latent variables  $\mathbf{Z}$
- 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  $p(\mathcal{D}|\mathbf{Z}, \Theta)$  and prior is  $p(\mathbf{Z}|\Theta)$ . **Want posterior over  $\mathbf{Z}$**
- $\Theta = (\theta, \phi)$  denotes the other parameters that define the likelihood and the prior
- For now, assume  $\Theta$  is known and only  $\mathbf{Z}$  is unknown (the  $\Theta$  unknown case later)
- Assume CP  **$p(\mathbf{Z}|\mathcal{D}, \Theta)$  is intractable**

# Variational Inference (VI)

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

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

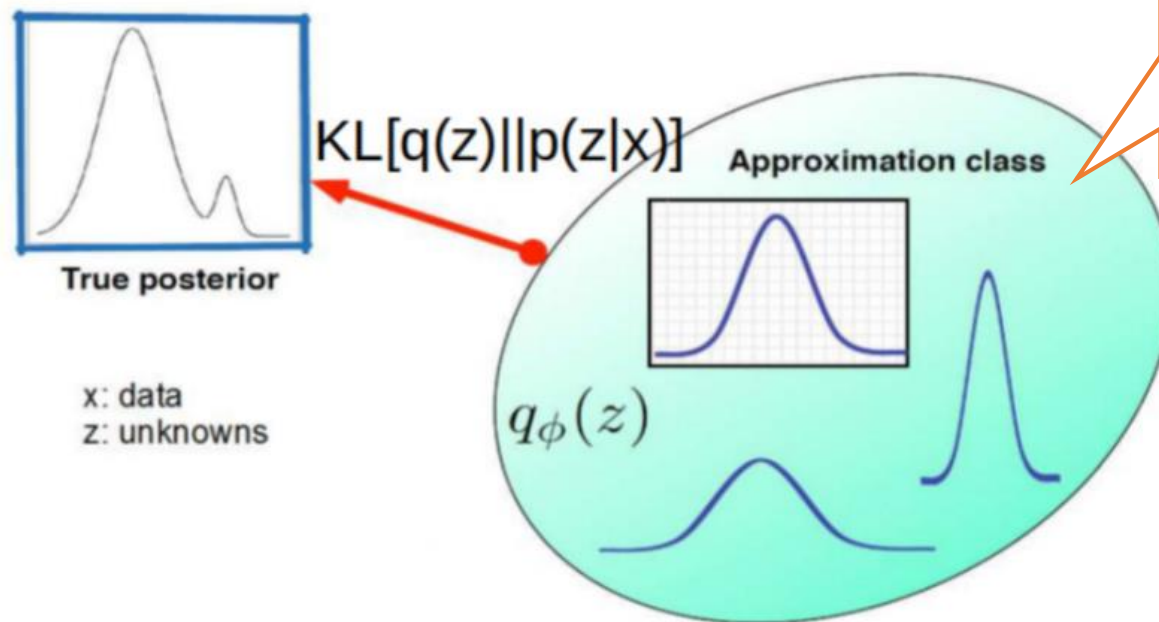
Kullback Leibler divergence  $KL[q||p]$  between  $q$  and  $p$

Also possible to use  $KL[p||q]$  or divergences other than KL

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

$q_{\phi}$  defines a class of distributions parametrized by  $\phi$  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

$$\begin{aligned}
 \phi^* &= \operatorname{argmin}_{\phi} \operatorname{KL}[q_{\phi}(\mathbf{Z}) || p(\mathbf{Z}|\mathcal{D}, \Theta)] \\
 &= \operatorname{argmin}_{\phi} \mathbb{E}_{q_{\phi}(\mathbf{Z})} \left[ \log q_{\phi}(\mathbf{Z}) - \log \frac{p(\mathcal{D}|\mathbf{Z}, \Theta)p(\mathbf{Z}|\Theta)}{p(\mathcal{D}|\Theta)} \right] \\
 &= \operatorname{argmin}_{\phi} \mathbb{E}_{q_{\phi}(\mathbf{Z})} [\log q_{\phi}(\mathbf{Z}) - \log p(\mathcal{D}|\mathbf{Z}, \Theta) - \log p(\mathbf{Z}|\Theta)] + \log p(\mathcal{D}|\Theta)
 \end{aligned}$$

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

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

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

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

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

# The ELBO

- The ELBO is defined as

$$\begin{aligned}\mathcal{L}(\phi, \Theta) &= \mathbb{E}_{q_{\phi}(\mathbf{Z})} [\log p(\mathcal{D}, \mathbf{Z} | \Theta) - \log q_{\phi}(\mathbf{Z})] \\ &= \mathbb{E}_{q_{\phi}(\mathbf{Z})} [\log p(\mathcal{D}, \mathbf{Z} | \Theta)] + H[q_{\phi}(\mathbf{Z})]\end{aligned}$$

- Thus maximizing the ELBO w.r.t.  $\phi$  gives us a  $q_{\phi}(\mathbf{Z})$  which
  - Maximizes the expected joint probability of data and latent variables
  - Has a high entropy

- We can also write the ELBO as follows

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

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

# Maximizing the ELBO

Unknown  $\Theta$  case later

- We need to maximize the ELBO w.r.t.  $\phi$  (for now, assuming  $\Theta$  is known)

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

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

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

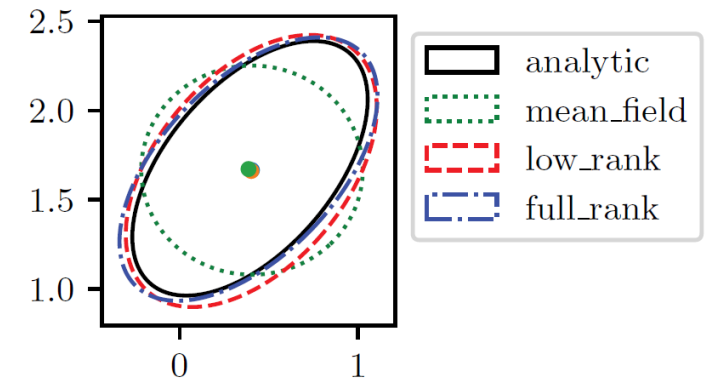
- Note: Expectations in ELBO and ELBO's gradients w.r.t.  $\phi$  may not be easy
  - Will see methods to handle such issues later
  - Assuming simple forms for  $q_{\phi}(\mathbf{Z})$  also helps (we can use random variable transformation methods to transform the simple form to more expressive ones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

# A Simple Illustration for VI

- Assume a simple likelihood model

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

- Suppose we want to estimate the posterior of the mean  $\mathbf{z}$
- Assuming a Gaussian prior on  $\mathbf{z}$  and assuming  $\mathbf{\Sigma}$  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  $q(\mathbf{z})$ 
  - Low-rank:  $q(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mu_{\mathbf{z}}, \mathbf{\Sigma}_{\mathbf{z}})$  where  $\mathbf{\Sigma}_{\mathbf{z}} = \mathbf{L}\mathbf{L}^\top$
  - Full-rank:  $q(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mu_{\mathbf{z}}, \mathbf{\Sigma}_{\mathbf{z}})$  with no constraint on  $\mathbf{\Sigma}_{\mathbf{z}}$
  - Mean-field:  $q(\mathbf{z}) = q(z_1)q(z_2) = \mathcal{N}(z_1|\mu_{z_1}, \sigma_{z_1}^2) \mathcal{N}(z_2|\mu_{z_2}, \sigma_{z_2}^2)$



# Detour

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

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

Transformed random variable

A one-to-one transformation function

If  $T$  stretches a small interval around  $\mathbf{u}$  by a factor  $s$ , the corresponding  $\boldsymbol{\theta}$ -interval is  $s$  times larger, so density must shrink by  $\frac{1}{s}$  to keep probabilities same.

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

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

Absolute value of the determinant of the Jacobian. It tells how a tiny volume element around  $\mathbf{u}$  is scaled when mapped to  $\boldsymbol{\theta}$ : densities scale inversely with that volume change.

- We can use such transformations for VI by using a simple distribution for  $q(\mathbf{Z})$  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  $q(\mathbf{Z}|\phi)$  or computing ELBO's gradients
- The idea: Assumes unknowns  $\mathbf{Z}$  can be partitioned into  $M$  groups  $\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_M$ , s.t.,

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

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

For models with **local conjugacy**,  
 it becomes super easy!

- Learning the optimal  $q(\mathbf{Z}|\phi)$  reduces to learning the optimal  $q_1, q_2, \dots, q_M$
- Can select groups based on model's structure, e.g., in Bayesian neural net for regression

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

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  $\operatorname{argmax}_{\phi} \mathcal{L}(\phi, \Theta)$ . 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  $\operatorname{argmax}_q \mathcal{L}(q)$ ?
- Note that under this mean-field assumption, the ELBO simplifies to

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

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

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

$$= \int q_j \log \hat{p}(\mathcal{D}, \mathbf{Z}_j | \Theta) d\mathbf{Z}_j - \int q_j \log q_j d\mathbf{Z}_j$$

$$= -\text{KL}(q_j || \hat{p}) \quad \log \hat{p}(\mathcal{D}, \mathbf{Z}_j | \Theta) = \mathbb{E}_{i \neq j} [\log p(\mathcal{D}, \mathbf{Z} | \Theta)] + \text{const}$$

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

- Thus  $q_j^* = \operatorname{argmax}_{q_j} \mathcal{L}(q) = \operatorname{argmin}_{q_j} \text{KL}(q_j || \hat{p}) = \hat{p}(\mathcal{D}, \mathbf{Z}_j | \Theta)$

# Separating integration over $Z_j$ and $Z_{-j}$

Write  $Z = (Z_j, Z_{-j})$  and integrate out  $Z_{-j}$  using the fixed factors  $q_{-j}$ . Define the expectation with respect to  $q_{-j}$ :

$$\mathbb{E}_{-j}[\cdot] \equiv \int \left( \prod_{i \neq j} q_i(Z_i) \right) (\cdot) dZ_{-j}.$$

Plugging into the ELBO and grouping terms that do and do not depend on  $q_j$ :

$$\mathcal{L}(q) = \int q_j(Z_j) \left\{ \mathbb{E}_{-j}[\log p(D, Z)] - \log q_j(Z_j) \right\} dZ_j + \underbrace{(\text{terms independent of } q_j)}_C,$$

where  $C$  is a constant w.r.t.  $q_j$  because it only involves  $q_{-j}$ .

So we can write the objective as a functional of  $q_j$ :

$$\mathcal{L}(q_j) = \int q_j(Z_j) \mathbb{E}_{-j}[\log p(D, Z)] dZ_j - \int q_j(Z_j) \log q_j(Z_j) dZ_j + C.$$

# Deriving Mean-Field VI Updates

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

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

- Note: Can often just compute the numerator and recognize denominator by inspection
- **Important:** For locally conjugate models,  $q_j^*(\mathbf{Z}_j)$  will have the same form as prior  $p(\mathbf{Z}_j|\Theta)$ 
  - 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:  $\mathcal{L}(q) = \int q_j \log \hat{p}(\mathcal{D}, \mathbf{Z}_j|\Theta) d\mathbf{Z}_j - \int q_j \log q_j d\mathbf{Z}_j$  is concave in  $q_j$

# The Mean-Field VI Algorithm

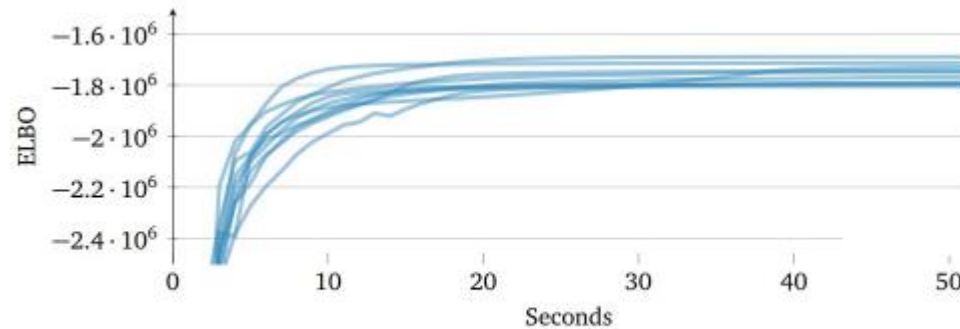
- Also known as **Co-ordinate Ascent Variational Inference** (CAVI) Algorithm
- Input: Model in form of priors and likelihood, or joint  $p(\mathcal{D}, \mathbf{Z}|\Theta)$ , Data  $\mathcal{D}$
- Output: A variational distribution  $q(\mathbf{Z}) = \prod_{j=1}^M q_j(\mathbf{Z}_j)$
- Initialize: Variational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

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

- Compute ELBO  $\mathcal{L}(q) = \mathbb{E}_q[\log p(\mathcal{D}, \mathbf{Z}|\Theta)] - \mathbb{E}_q[\log q(\mathbf{Z})]$
- NOTE: We can also use mean-field assumption for  $q(\mathbf{Z})$  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

# Recap: Variational Inference (VI)

Variational  
distribution

Variational  
parameters

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

KL minimization

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

ELBO  
maximization

$$\begin{aligned} \phi^* &= \operatorname{argmax}_\phi \mathbb{E}_{q_\phi(\mathbf{Z})} [\log p(\mathcal{D}|\mathbf{Z}, \Theta)] - \operatorname{KL}[q_\phi(\mathbf{Z}) || p(\mathbf{Z}|\Theta)] \\ &= \operatorname{argmax}_\phi \mathbb{E}_{q_\phi(\mathbf{Z})} [\log p(\mathcal{D}, \mathbf{Z}|\Theta) - \log q_\phi(\mathbf{Z})] = \operatorname{argmax}_\phi \mathcal{L}(\phi, \Theta) \end{aligned}$$

Can use gradient-based optimization to learn the parameters of the variational distribution

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

Case when  $\Theta$  is also unknown will be discussed later

Mean-field assumption on the variational distribution

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

$\mathbb{E}_{i \neq j}$  denotes expectations w.r.t.  $\prod_{i \neq j} q(\mathbf{Z}_i|\phi_i)$

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

This, for simple enough model, when using mean-field VI, we can get optimal  $q$  "directly" without taking ELBO derivatives

Equivalent to writing  $\log q_j^*(\mathbf{Z}_j) = \mathbb{E}_{i \neq j} [\log p(\mathcal{D}, \mathbf{Z}|\Theta)] + \text{const}$

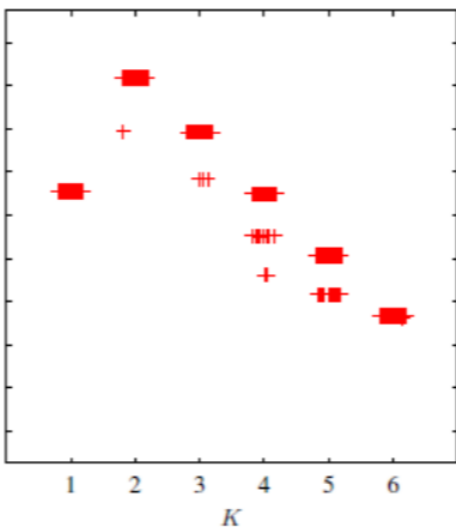
"simple enough" means the cases where these expectations can be analytically computed

# ELBO for Model Selection

- Recall that ELBO is a lower bound on log of model evidence  $\log p(\mathbf{X}|\mathbf{m})$
- Can compute ELBO for each model  $\mathbf{m}$  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.

$p(\mathcal{D}|K)$



K	Approximate ELBO values (from '+' symbols)
1	Low
2	High (Peak)
3	Medium-High
4	Medium
5	Low-Medium
6	Low

Each value of  $K$  represents a different model

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

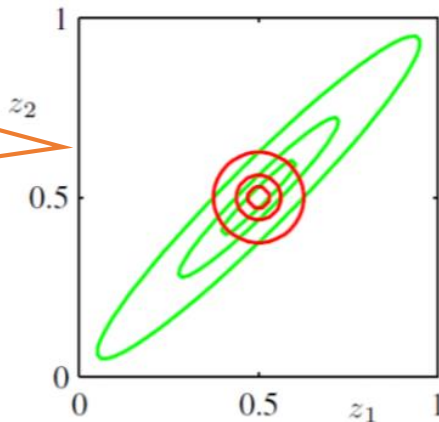
# VI might under-estimate posterior's variance

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

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

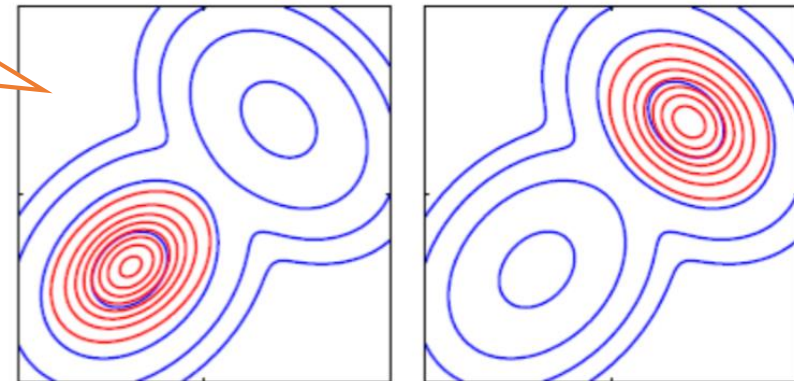
- $q(\mathbf{Z})$  will be small where  $p(\mathbf{Z}|\mathcal{D})$  is small otherwise KL will blow up
- Thus  $q(\mathbf{Z})$  avoids low-probability regions of the true posterior

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



$q$  (red) concentrated on one of the modes of  $p$  (blue)

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



# Variational EM

- If the parameters  $\Theta$  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  $\mathbf{Z}$
- VEM alternates between the following two steps
  - Maximize the ELBO w.r.t.  $\phi$  (gives the variational approximation  $q(\mathbf{Z})$  of CP of  $\mathbf{Z}$ )

$$\phi^{(t)} = \operatorname{argmax}_{\phi} \mathbb{E}_{q_{\phi}(\mathbf{Z})} [\log p(\mathcal{D}, \mathbf{Z} | \Theta^{(t-1)}) - \log q_{\phi}(\mathbf{Z})]$$

- Maximize the ELBO w.r.t.  $\Theta$  (gives us point estimate of  $\Theta$ )

$$\begin{aligned} \Theta^{(t)} &= \operatorname{argmax}_{\Theta} \mathbb{E}_{q_{\phi^{(t)}}(\mathbf{Z})} [\log p(\mathcal{D}, \mathbf{Z} | \Theta) - \log q_{\phi^{(t)}}(\mathbf{Z})] \\ &= \operatorname{argmax}_{\Theta} \mathbb{E}_{q_{\phi^{(t)}}(\mathbf{Z})} [\log p(\mathcal{D}, \mathbf{Z} | \Theta)] \end{aligned}$$

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

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

# Example: Mean-field VI without ELBO Derivatives

No “latent variables” here. Data  $\mathbf{X}$  is fully observed, and parameters  $\mu, \tau$  need to be estimated

- Consider data  $\mathbf{X} = \{x_1, x_2, \dots, x_N\}$  from a one-dim Gaussian  $\mathcal{N}(\mu, \tau^{-1})$

- Assume the following normal-gamma prior on  $\mu$  and  $\tau$

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

Assume the hyperparameters  $\mu_0, \lambda_0, a_0, b_0$  are known

- Posterior is also normal-gamma due to the jointly conjugate prior

- Let's still try mean-field VI for this model

Note that we aren't even specifying the forms of these two distributions! We'll be able identify the forms in a few steps after working with the expectations

- With mean-field assumption on the variational posterior  $q(\mu, \tau) = q_\mu(\mu)q_\tau(\tau)$

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

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

- 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

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

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

## Example: Mean-field VI without ELBO Derivatives

- 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  $q_\mu^* = \mathcal{N}(\mu|\mu_N, \lambda_N^{-1})$  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  $q_\tau$

- 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  $q_\mu$

- Note: Updates of  $q_\mu^*$  and  $q_\tau^*$  depend on each other (hence alternating updates needed)

# Mean-Field VI for Locally Conjugate Models

- Since  $\log q_j^*(\mathbf{Z}_j) = \mathbb{E}_{i \neq j}[\log p(\mathbf{X}, \mathbf{Z})] + \text{const} = \mathbb{E}_{i \neq j}[\log p(\mathbf{X}, \mathbf{Z}_j, \mathbf{Z}_{-j})] + \text{const}$

$$\log q_j^*(\mathbf{Z}_j) = \mathbb{E}_{i \neq j}[\log p(\mathbf{Z}_j | \mathbf{X}, \mathbf{Z}_{-j})] + \text{const} \quad \text{For any model}$$

- Thus finding optimal  $q_j^*(\mathbf{Z}_j)$  only requires expectations of params of CP  $p(\mathbf{Z}_j | \mathbf{X}, \mathbf{Z}_{-j})$
- For locally conjugate models, we know CP is easy and is an exp-fam distr of the form

$$p(\mathbf{Z}_j | \mathbf{X}, \mathbf{Z}_{-j}) = h(\mathbf{Z}_j) \exp \left[ \eta(\mathbf{X}, \mathbf{Z}_{-j})^\top \mathbf{Z}_j - A(\eta(\mathbf{X}, \mathbf{Z}_{-j})) \right]$$

- Using the above, we can rewrite the optimal variational distribution as follows

$$\begin{aligned} \log q_j^*(\mathbf{Z}_j) &= \mathbb{E}_{i \neq j} \left[ \log \left( h(\mathbf{Z}_j) \exp \left[ \eta(\mathbf{X}, \mathbf{Z}_{-j})^\top \mathbf{Z}_j - A(\eta(\mathbf{X}, \mathbf{Z}_{-j})) \right] \right) \right] + \text{const} \\ \implies q_j^*(\mathbf{Z}_j) &\propto h(\mathbf{Z}_j) \exp \left[ \mathbb{E}_{i \neq j}[\eta(\mathbf{X}, \mathbf{Z}_{-j})]^\top \mathbf{Z}_j \right] \quad (\text{verify}) \end{aligned}$$

- Thus, with local conj, we just require expectation of nat. params. of CP of  $\mathbf{Z}_j$

# VI for models without “latent variables”

22

Recall the Gaussian mean and variance estimation problem

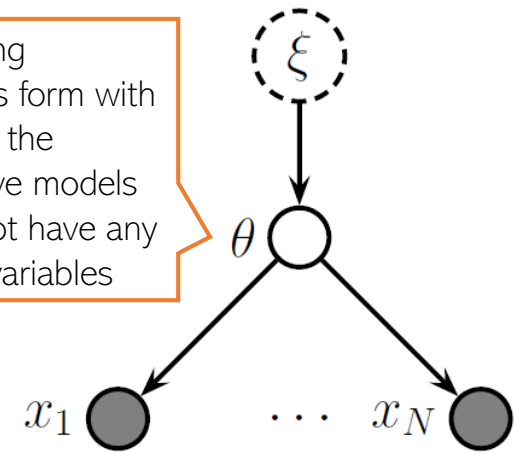
- Suppose we have a “fully observed” case (no missing data/latent variables but just some unknown global parameters  $\theta$  and known hyperparams  $\xi$ )
- A simple example of the model is shown in the figure below

$$p(\mathcal{D}, \theta | \xi) = p(\theta | \xi) \prod_{n=1}^N p(x_n | \theta)$$

If this CP is intractable, we can use VI to approximate this

$$p(\theta | \mathcal{D}, \xi) = \frac{p(\mathcal{D} | \theta) p(\theta | \xi)}{p(\mathcal{D} | \xi)}$$

Even supervised learning problems may have this form with  $\theta$  being the weights of the generative/discriminative models and the models may not have any missing data or latent variables



- If  $\xi$  are also unknown then one way would be to alternate like Variational EM
  - Approximating the CP  $p(\theta | \mathcal{D}, \xi)$  using VI
  - Using MLE-II to get point estimates of the hyperparameters  $\xi$

# VI using ELBO's gradients

- For simple locally conjugate models, VI updates are usually easy
  - Sometimes, can find the optimal  $q$  even without taking the ELBO's gradients
- For complex models, we have to use the more general gradient-based approach
- Consider the setting when we have latent variables  $\mathbf{Z}$  and parameters  $\Theta$
- The ELBO's gradient w.r.t.  $\Theta$

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

Monte-Carlo approximation using samples of  $q_{\phi}(\mathbf{Z})$  is straightforward here

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

Gradient can go inside expectation since  $q(\mathbf{Z})$  doesn't depend on  $\Theta$

- The ELBO's gradient w.r.t.  $\phi$

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

Monte-Carlo approximation using samples of  $q_{\phi}(\mathbf{Z})$  is NOT as straightforward

$$\neq \mathbb{E}_{q_{\phi}(\mathbf{Z})} [\nabla_{\phi} \{\log p(\mathcal{D}, \mathbf{Z} | \Theta) - \log q_{\phi}(\mathbf{Z})\}]$$

Gradient can't go inside expectation since  $q(\mathbf{Z})$  depends on  $\phi$

# Black-Box Variational Inference (BBVI)

- Black-box Var. Inference\* (BBVI) approximates ELBO derivatives using Monte-Carlo
- Uses the following identity for the ELBO's derivative

$$\begin{aligned}\nabla_{\phi} \mathcal{L}(q) &= \nabla_{\phi} \mathbb{E}_q[\log p(\mathbf{X}, \mathbf{Z}) - \log q(\mathbf{Z}|\phi)] \\ &= \mathbb{E}_q[\nabla_{\phi} \log q(\mathbf{Z}|\phi)(\log p(\mathbf{X}, \mathbf{Z}) - \log q(\mathbf{Z}|\phi))] \quad (\text{proof on next slide})\end{aligned}$$

- Thus ELBO gradient can be written solely in terms of expec. of gradient of  $\log q(\mathbf{Z}|\phi)$ 
  - Required gradients don't depend on the model; only on chosen var. distribution (hence “black-box”)
- Given  $S$  samples  $\{\mathbf{Z}_s\}_{s=1}^S$  from  $q(\mathbf{Z}|\phi)$ , we can get (noisy) gradient as follows

$$\nabla_{\phi} \mathcal{L}(q) \approx \frac{1}{S} \sum_{s=1}^S \nabla_{\phi} \log q(\mathbf{Z}_s|\phi)(\log p(\mathbf{X}, \mathbf{Z}_s) - \log q(\mathbf{Z}_s|\phi))$$

- Above is also called the “score function” based gradient (also REINFORCE method)

Gradient of a log-likelihood or log-probability function w.r.t. its params is called score function; hence the name

# Proof of BBVI Identity

- The ELBO gradient can be written as

$$\begin{aligned}
 \nabla_{\phi} \mathcal{L}(q) &= \nabla_{\phi} \int (\log p(\mathbf{X}, \mathbf{Z}) - \log q(\mathbf{Z}|\phi)) q(\mathbf{Z}|\phi) d\mathbf{Z} \\
 &= \int \nabla_{\phi} [(\log p(\mathbf{X}, \mathbf{Z}) - \log q(\mathbf{Z}|\phi)) q(\mathbf{Z}|\phi)] d\mathbf{Z} \quad (\nabla \text{ and } \int \text{ interchangeable; dominated convergence theorem}) \\
 &= \int \nabla_{\phi} [(\log p(\mathbf{X}, \mathbf{Z}) - \log q(\mathbf{Z}|\phi))] q(\mathbf{Z}|\phi) + \nabla_{\phi} q(\mathbf{Z}|\phi) [(\log p(\mathbf{X}, \mathbf{Z}) - \log q(\mathbf{Z}|\phi))] d\mathbf{Z} \\
 &= \mathbb{E}_q[-\nabla_{\phi} \log q(\mathbf{Z}|\phi)] + \int \nabla_{\phi} q(\mathbf{Z}|\phi) [(\log p(\mathbf{X}, \mathbf{Z}) - \log q(\mathbf{Z}|\phi))] d\mathbf{Z}
 \end{aligned}$$

$\exists g(z) \geq 0$  with  $\int g(z) dz < \infty$  such that  $|f_{\phi}(z)| \leq g(z) \forall \phi$  so we can apply dominated convergence. Holds for standard families (Gaussians, etc.).

- Note that  $\mathbb{E}_q[\nabla_{\phi} \log q(\mathbf{Z}|\phi)] = \mathbb{E}_q \left[ \frac{\nabla_{\phi} q(\mathbf{Z}|\phi)}{q(\mathbf{Z}|\phi)} \right] = \int \nabla_{\phi} q(\mathbf{Z}|\phi) d\mathbf{Z} = \nabla_{\phi} \int q(\mathbf{Z}|\phi) d\mathbf{Z} = \nabla_{\phi} 1 = 0$
- Also note that  $\nabla_{\phi} q(\mathbf{Z}|\phi) = \nabla_{\phi} [\log q(\mathbf{Z}|\phi)] q(\mathbf{Z}|\phi)$ , using which

$$\begin{aligned}
 \int \nabla_{\phi} q(\mathbf{Z}|\phi) [(\log p(\mathbf{X}, \mathbf{Z}) - \log q(\mathbf{Z}|\phi))] d\mathbf{Z} &= \int \nabla_{\phi} \log q(\mathbf{Z}|\phi) [(\log p(\mathbf{X}, \mathbf{Z}) - \log q(\mathbf{Z}|\phi))] q(\mathbf{Z}|\phi) d\mathbf{Z} \\
 &= \mathbb{E}_q[\nabla_{\phi} \log q(\mathbf{Z}|\phi) (\log p(\mathbf{X}, \mathbf{Z}) - \log q(\mathbf{Z}|\phi))]
 \end{aligned}$$

- Therefore  $\nabla_{\phi} \mathcal{L}(q) = \mathbb{E}_q[\nabla_{\phi} \log q(\mathbf{Z}|\phi) (\log p(\mathbf{X}, \mathbf{Z}) - \log q(\mathbf{Z}|\phi))]$

# Benefits of BBVI

- Recall that BBVI approximates the ELBO gradients by the Monte Carlo expectations

$$\nabla_{\phi} \mathcal{L}(q) \approx \frac{1}{S} \sum_{s=1}^S \nabla_{\phi} \log q(\mathbf{Z}_s | \phi) (\log p(\mathbf{X}, \mathbf{Z}_s) - \log q(\mathbf{Z}_s | \phi))$$

- Enables applying VI for a wide variety of probabilistic models
- Can also work with small minibatches of data rather than full data
- BBVI has very few requirements
  - Should be able to sample from  $q(\mathbf{Z} | \phi)$  (usually sampling routines exists!)
  - Should be able to compute  $\nabla_{\phi} \log q(\mathbf{Z} | \phi)$  (automatic differentiation methods exist!)
  - Should be able to evaluate  $\log p(\mathbf{X}, \mathbf{Z})$  and  $\log q(\mathbf{Z} | \phi)$  for any value of  $\mathbf{Z}$
- Some tricks needed to control the variance in the Monte Carlo estimate of the ELBO gradient (if interested in the details, please refer to the BBVI paper)

# Reparametrization Trick

- Monte-Carlo approx. of ELBO grad (with often lower var than BBVI gradient)
- Suppose we want to compute ELBO's gradient  $\nabla_{\phi} \mathbb{E}_{q_{\phi}(\mathbf{Z})} [\log p(\mathbf{X}, \mathbf{Z}) - \log q_{\phi}(\mathbf{Z})]$
- Assume a deterministic transformation  $g$

$$\mathbf{Z} = g(\epsilon, \phi) \quad \text{where} \quad \epsilon \sim p(\epsilon)$$

Assumed to not depend on  $\phi$

- With this reparametrization, and using LOTUS rule, the ELBO's gradient would be

$$\nabla_{\phi} \mathbb{E}_{p(\epsilon)} [\log p(\mathbf{X}, g(\epsilon, \phi)) - \log q_{\phi}(g(\epsilon, \phi))] = \mathbb{E}_{p(\epsilon)} \nabla_{\phi} [\log p(\mathbf{X}, g(\epsilon, \phi)) - \log q_{\phi}(g(\epsilon, \phi))]$$

- Given  $S$  i.i.d. random samples  $\{\epsilon_s\}_{s=1}^S$  from  $p(\epsilon)$ , we can get a Monte-Carlo approx.

$$\nabla_{\phi} \mathbb{E}_{q_{\phi}(\mathbf{Z})} [\log p(\mathbf{X}, \mathbf{Z}) - \log q_{\phi}(\mathbf{Z})] \approx \frac{1}{S} \sum_{s=1}^S [\nabla_{\phi} \log p(\mathbf{X}, g(\epsilon_s, \phi)) - \nabla_{\phi} \log q_{\phi}(g(\epsilon_s, \phi))]$$

- Such gradients are called **pathwise gradients**\* (since we took a “path” from  $\epsilon$  to  $\mathbf{Z}$ )

\*Autoencoding Variational Bayes - Kingma and Welling (2013), Stochastic Backpropagation and Approximate Inference in Deep Generative Models- Rezende et al (2014)

# Reparametrization Trick: An Example

- Suppose our variational distribution is  $q(\mathbf{w}|\phi) = \mathcal{N}(\mathbf{w}|\mu, \Sigma)$ , so  $\phi = \{\mu, \Sigma\}$
- Suppose our ELBO has a difficult expectation term  $\mathbb{E}_q[f(\mathbf{w})]$
- However, note that we need ELBO gradient, not ELBO itself. Let's use the trick
- Reparametrize  $\mathbf{w}$  as  $\mathbf{w} = \mu + \mathbf{L}\mathbf{v}$  where  $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

Or  $\phi = \{\mu, \mathbf{L}\}$   
where  $\mathbf{L} = \text{chol}(\Sigma)$

Note that we will still have  
 $q(\mathbf{w}|\phi) = \mathcal{N}(\mathbf{w}|\mu, \Sigma)$

$$\nabla_{\mu, \mathbf{L}} \mathbb{E}_{\mathcal{N}(\mathbf{w}|\mu, \Sigma)}[f(\mathbf{w})] = \nabla_{\mu, \mathbf{L}} \mathbb{E}_{\mathcal{N}(\mathbf{v}|\mathbf{0}, \mathbf{I})}[f(\mu + \mathbf{L}\mathbf{v})] = \mathbb{E}_{\mathcal{N}(\mathbf{v}|\mathbf{0}, \mathbf{I})}[\nabla_{\mu, \mathbf{L}} f(\mu + \mathbf{L}\mathbf{v})]$$

- The above is now straightforward
  - Easily take derivatives of  $f(\mathbf{w})$  w.r.t. variational params  $\mu, \mathbf{L}$
  - Replace exp. by Monte-Carlo averaging using samples of  $\mathbf{v}$  from  $\mathcal{N}(\mathbf{0}, \mathbf{I})$

Often even one or very few samples suffice

$$\begin{aligned} \nabla_{\mu} \mathbb{E}_{\mathcal{N}(\mathbf{w}|\mu, \Sigma)}[f(\mathbf{w})] &= \mathbb{E}_{\mathcal{N}(\mathbf{v}|\mathbf{0}, \mathbf{I})}[\nabla_{\mu} f(\mu + \mathbf{L}\mathbf{v})] \approx \nabla_{\mu} f(\mu + \mathbf{L}\mathbf{v}_s) \\ \nabla_{\mathbf{L}} \mathbb{E}_{\mathcal{N}(\mathbf{w}|\mu, \Sigma)}[f(\mathbf{w})] &= \mathbb{E}_{\mathcal{N}(\mathbf{v}|\mathbf{0}, \mathbf{I})}[\nabla_{\mathbf{L}} f(\mu + \mathbf{L}\mathbf{v})] \approx \nabla_{\mathbf{L}} f(\mu + \mathbf{L}\mathbf{v}_s) \end{aligned}$$

$\frac{\partial f}{\partial \mathbf{w}} \frac{\partial \mathbf{w}}{\partial \mu}$   
 Chain Rule  
 $\frac{\partial f}{\partial \mathbf{w}} \frac{\partial \mathbf{w}}{\partial \mathbf{L}}$

# Reparametrization Trick: Some Comments

- Standard Reparametrization Trick assumes the model to be differentiable

$$\nabla_{\phi} \mathbb{E}_{q_{\phi}(\mathbf{Z})} [\log p(\mathbf{X}, \mathbf{Z}) - \log q_{\phi}(\mathbf{Z})] = \mathbb{E}_{p(\epsilon)} [\nabla_{\phi} \log p(\mathbf{X}, g(\epsilon, \phi)) - \nabla_{\phi} \log q_{\phi}(g(\epsilon, \phi))]$$

- In contrast, BBVI (score function gradients) only required  $q(\mathbf{Z})$  to be differentiable
- Thus rep. trick often isn't applicable, e.g., when  $\mathbf{Z}$  is discrete (e.g., binary /categorical)
  - Recent work on continuous relaxation<sup>†</sup> of discrete variables<sup>†</sup> (e.g., Gumbel Softmax for categorical)
- Assumes that we can directly draw samples from  $p(\epsilon)$ . If not, then rep. trick isn't valid<sup>@</sup>

<sup>†</sup>Categorical Reparameterization with Gumbel-Softmax (Jang et al, 2017), <sup>@</sup> Reparameterization Gradients through Acceptance-Rejection Sampling Algorithms (Naesseth et al, 2016)