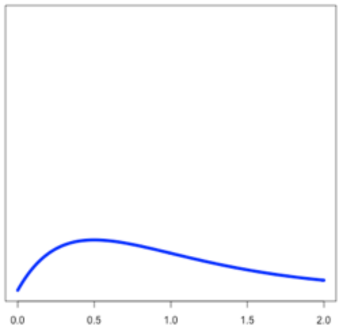
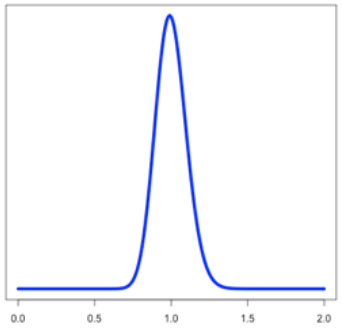
Variational Bayes/Inference

Bayesian Inferencing

- Posterior distribution: $p(\theta|x) \propto p(x|\theta)\, p(\theta)$



X likelihood $\propto$



Prior

Posterior

Both prior and likelihood distributions are user choices/assumptions

Tractability of Bayesian Inferencing

- $p(\theta|x) = p(x|\theta)\, p(\theta) / p(x) = p(x|\theta)\, p(\theta) / \int p(x|\theta)\, p(\theta) d\theta$

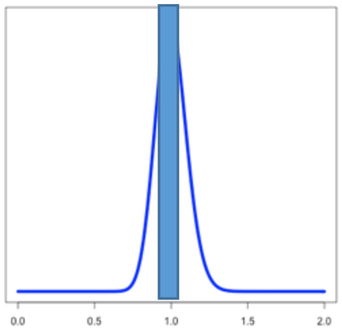
  data belief evidence

- The (normalizing) denominator is difficult to compute in closed form
  - Except for the case of conjugate priors

- Need for Approximate Bayesian methods

Approximate Bayesian Inferencing

- MAP point estimate
  - All probability mass concentrated at maxima
  - Finding MAP doesn't need evidence computation

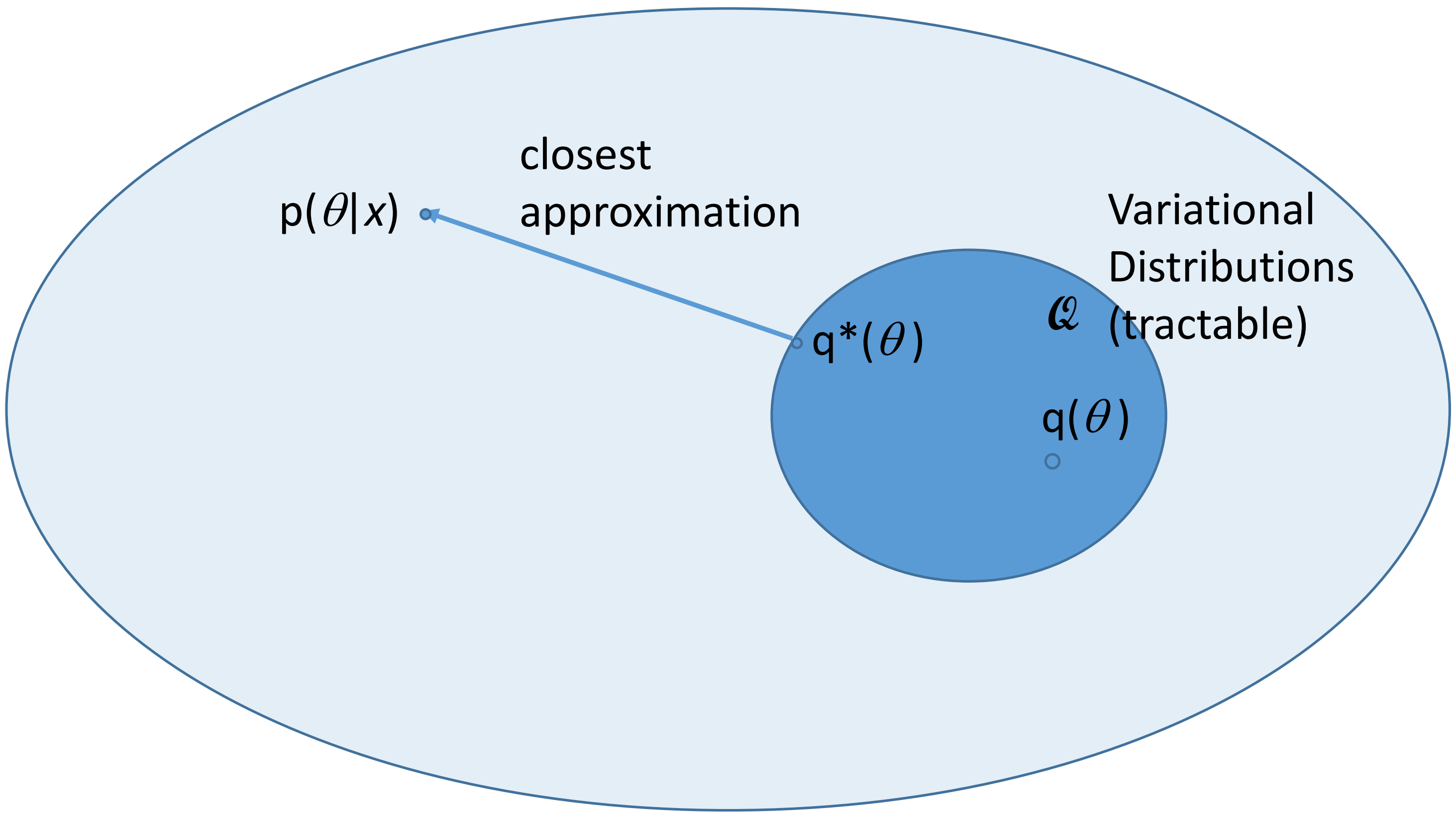


- Laplace approximation

- Variational inferencing

- Sampling based methods

Variational Inference



Optimization problem

Kullback-Leibler Divergence

- KL divergence is a measure of closeness between two distributions

- $KL(q||p) = \int q(x) \log (q(x)/p(x))dx = E_q[\log(q(x)/p(x))]$
    - q and p should have same domain
    - KL is always greater than or equal to zero
    - KL = 0, if q and p are equal almost everywhere
    - Not symmetric
    - Works well in practice

Variational Inference

- Optimization problem over distribution functions $q(\theta)$
  - Functional: KLD wrt $p(\theta|x)$

$$\text{Minimize}_{\,q(\theta) \in \mathcal{Q}} \, KL(q(\theta)\,||\,p(\theta|x))$$

Solving the optimization problem

- Posterior distribution p(θ|x) is not tractable/known
  - Difficult to compute KL(q(x)|| p(θ|x))

- We can transform the VI optimization problem terms of p($x|\theta$), p(θ)
  - p($x|\theta$), p(θ) are known and usually tractable
  - We will away with the normalizing denominator "evidence" in Bayes rule which is usually difficult to compute

Variational Inference: Expanding the Evidence

$$\log p(x) = \int q(\theta) \log p(x) d\theta$$

$$= \int q(\theta) \log \frac{p(x,\theta)}{p(\theta|x)} d\theta$$

$$= \int q(\theta) \log \frac{p(x,\theta) q(\theta)}{p(\theta|x) q(\theta)} d\theta$$

$$= \int q(\theta) \log \frac{p(x,\theta)}{q(\theta)} d\theta + \int q(\theta) \log \frac{q(\theta)}{p(\theta|x)} d\theta$$

$$= \mathcal{L}(q(\theta)) + \mathrm{KL}(q(\theta)||p(\theta|x))$$

Since KLD is positive -

$\log p(x) \geq \mathcal{L}(q(\theta))$ Evidence Lower Bound (ELBO)

If there is no restriction on q, then KLD is zero, and the lower bound is exact

Variational Inference: Optimization

$$\log p(x) = \mathcal{L}(q(\theta)) + \mathrm{KL}(q(\theta)||p(\theta|x))$$

Since, p(x) is independent of q(θ) it can be considered as a constant in the optimization problem -

$$\text{Minimize}_{\,q(\theta) \in \mathcal{Q}} \ \mathrm{KL}(q(\theta)||p(\theta|x))$$

Is equivalent to

$$\text{Maximize}_{\,q(\theta) \in \mathcal{Q}} \ \mathcal{L}(q(\theta))$$

ELBO - Variational lower bound

ELBO

$$\mathcal{L}(q(\theta)) = \int q(\theta) \log \frac{p(x,\theta)}{q(\theta)} d\theta$$
$$= \int q(\theta) \log \frac{p(x|\theta)p(\theta)}{q(\theta)} d\theta$$
$$= \int q(\theta) \log p(x|\theta) d\theta + \int q(\theta) \log \frac{p(\theta)}{q(\theta)} d\theta$$
$$= \mathbb{E}_{q(\theta)}[\log p(x|\theta)] - \mathrm{KL}(q(\theta)||p(\theta))$$

- The first term is maximized when q(θ) is a concentrated delta function at MLE – data term
- The second term is maximized when q(θ) is same as the prior – regularization term
- A combination of both is maximized

Statistical Physics Interpretation of ELBO

$$\mathcal{L}(q(\theta)) = \int q(\theta) \log \frac{p(x,\theta)}{q(\theta)} d\theta$$

$$= \mathbb{E}_{q(\theta)}[\log p(x,\theta)] - \int q(\theta) \log q(\theta) d\theta$$

$\mathcal{L}(q(\theta)) =$ Energy of $p(x,\theta) +$ Entropy of $q(\theta)$
$\mathcal{L}(q(\theta))$ is known as variational energy of Helmholtz free energy in statistical physics

Generalizing KL Divergence

One can create a family of divergence measures indexed by a parameter $\alpha \in \mathbb{R}$ by defining the **alpha divergence** as follows:

$$D_\alpha(p||q) \triangleq \frac{4}{1-\alpha^2}\left(1 - \int p(x)^{(1+\alpha)/2} q(x)^{(1-\alpha)/2} dx\right) \tag{21.21}$$

This measure satisfies $D_\alpha(p||q) = 0$ iff $p = q$, but is obviously not symmetric, and hence is not a metric. $\mathbb{KL}(p||q)$ corresponds to the limit $\alpha \to 1$, whereas $\mathbb{KL}(q||p)$ corresponds to the limit $\alpha \to -1$. When $\alpha = 0$, we get a symmetric divergence measure that is linearly related to the **Hellinger distance**, defined by

$$D_H(p||q) \triangleq \int \left(p(x)^{\frac{1}{2}} - q(x)^{\frac{1}{2}}\right)^2 dx \tag{21.22}$$

See Murphy book for details

Variational Inference: Summary

$$\log p(x) = \mathcal{L}(q(\theta)) + \mathrm{KL}(q(\theta)||p(\theta|x))$$

Since, p(x) is independent of q(θ) it can be considered as a constant in the optimization problem -

$$\text{Minimize}_{\,q(\theta) \in \mathcal{Q}} \mathrm{KL}(q(\theta)||p(\theta|x))$$

Is equivalent to

$$\text{Maximize}_{\,q(\theta) \in \mathcal{Q}} \mathcal{L}(q(\theta))$$

ELBO - Variational lower bound

The above derivation for θ can be generalized to latent variables + parameters
We call all of them as latent variables Z in subsequent discussion

What is the class of $\mathcal{Q}$ (the approximating distributions)?

Factorized Distributions

Let Z partition into non-overlapping groups Z_i

Assume:

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

(Recall naïve Bayes)

We do not make any assumption of the functional form of the individual factors

Note that it is a restriction on q and not on p

Mean Field Theory

- In physics and probability theory, mean-field theory studies the behavior of high-dimensional random models by studying a simpler model that approximates the original by averaging over degrees of freedom.
- Such models consider many individual components that interact with each other. In MFT, the effect of all the other individuals on any given individual is approximated by a single averaged effect, thus reducing a many-body problem to a one-body problem.
- The main idea of MFT is to replace all interactions to any one body with an average or effective interaction, sometimes called a molecular field.

Mean Field Approximation

- Do not consider interaction among all variable
- Cluster variables into groups
- Assume interaction within cluster – locally joint distribution
- Assume independence across cluster – factorization

- Mean field of the clusters are considered

Mean Field VI

- One of the simplest ways of doing VB
- In mean-field VB, we define a partition of the latent variables $\mathbf{Z}$ into M groups $\mathbf{Z}_1, \ldots, \mathbf{Z}_M$
- Assume our approximation $q(\mathbf{Z})$ factorizes over these groups

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

- As a short-hand, sometimes we write $q = \prod_{i=1}^{M} q_i$ where $q_i = q(\mathbf{Z}_i|\phi_i)$
- In mean-field VB, learning the optimal q reduces to learning the optimal $q_1, \ldots, q_M$

Deriving Mean Field VI

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

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

- Suppose we wish to find the optimal q_j given all other q_i $(i \neq j)$. Let's re-express $\mathcal{L}(q)$ as

$$\begin{aligned} \mathcal{L}(q) &= \int q_j \left[\int \log p(\mathbf{X}, \mathbf{Z}) \prod_{i \neq j} q_i \, d\mathbf{Z}_i \right] d\mathbf{Z}_j - \int q_j \log q_j \, d\mathbf{Z}_j \quad + \quad \text{consts w.r.t. } q_j \\ &= \int q_j \log \tilde{p}(\mathbf{X}, \mathbf{Z}_j) d\mathbf{Z}_j - \int q_j \log q_j \, d\mathbf{Z}_j \end{aligned}$$

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

- Note that $\mathcal{L}(q) = -KL(q_j || \tilde{p}) + \text{const}$. Which q_j will maximize it?

$$\boxed{q_j = \tilde{p}(\mathbf{X}, \mathbf{Z}_j)}$$

Contd..

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

$$q_j^*(\mathbf{Z}_j) = \frac{\exp(\mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})])}{\int \exp(\mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})]) d\mathbf{Z}_j} \qquad \forall j$$

- For locally-conjugate models, $q_j^*(\mathbf{Z}_j)$ will have the same form as the prior $p(\mathbf{Z}_j)$
- Important: For estimating q_j, the required expectation depends on other $\{q_i\}_{i \neq j}$
- Thus we need to cycle through updating each q_j in turn   co-ordinate ascent
- Guaranteed to converge (to a local optima)

Coordinate Ascent Algorithm

- Also known as Co-ordinate Ascent Variational Inference (CAVI) Algorithm
- Input: Model $p(\mathbf{X}, \mathbf{Z})$, Data $\mathbf{X}$
- Output: A variational distribution $q(\mathbf{Z}) = \prod_{j=1}^{M} q_j(\mathbf{Z}_j)$
- Initialize: Variational distributions $q_j(\mathbf{Z}_j)$, $j = 1, \ldots, M$
- While the ELBO has not converged
  - For each $j = 1, \ldots, M$, set

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

  - Compute ELBO $\mathcal{L}(q) = \mathbb{E}_q[\log p(\mathbf{X}, \mathbf{Z})] - \mathbb{E}_q[\log q(\mathbf{Z})]$

Nature of Approximation in VI

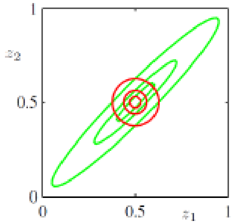
Recall that VB is equivalent to finding q by minimizing $\mathrm{KL}(q||p)$

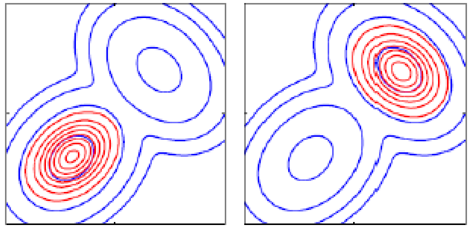
$$\mathrm{KL}(q||p) = \int q(\mathbf{Z}) \log \left[\frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X})} \right]$$

If the true posterior $p(\mathbf{Z}|\mathbf{X})$ is very small in some region then, to minimize $\mathrm{KL}(q||p)$, the approx. dist. q will also have to be very small (otherwise KL will be very large)

This has two key consequences for VB

- Underestimates the variances of the true posterior
- For multimodal posteriors, VB locks onto one of the modes





Simple Example: Univariate Gaussian

- Consider data $\mathbf{X} = \{x_1, \ldots, x_N\}$ from a 1-D Gaussian $\mathcal{N}(x|\mu, \tau^{-1})$ with mean μ, precision τ
- Assume the following normal-gamma prior on μ and τ

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

- Note: Here posterior is straightforward (normal-gamma due to the jointly conjugate prior)
- Let's try mean-field VI nevertheless to illustrate the idea
- 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 $\log p(\mathbf{X}, \mu, \tau) = \log p(\mathbf{X}|\mu, \tau) + \log p(\mu|\tau) + \log p(\tau)$. Therefore

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

Example Contd..

- Substituting the expressions $p(\mathbf{X}|\mu,\tau) = \prod_{n=1}^{N} p(x_n|\mu,\tau)$ and $\log p(\mu|\tau)$, 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. Thus $q_\mu^*(\mu) = \mathcal{N}(\mu|\mu_N, \tau_N)$ with

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

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

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

Updates of $q_\mu^*(\mu)$ and $q_\tau^*(\tau)$ depend on each-other

Mean Field Approximation: Univariate Gaussian

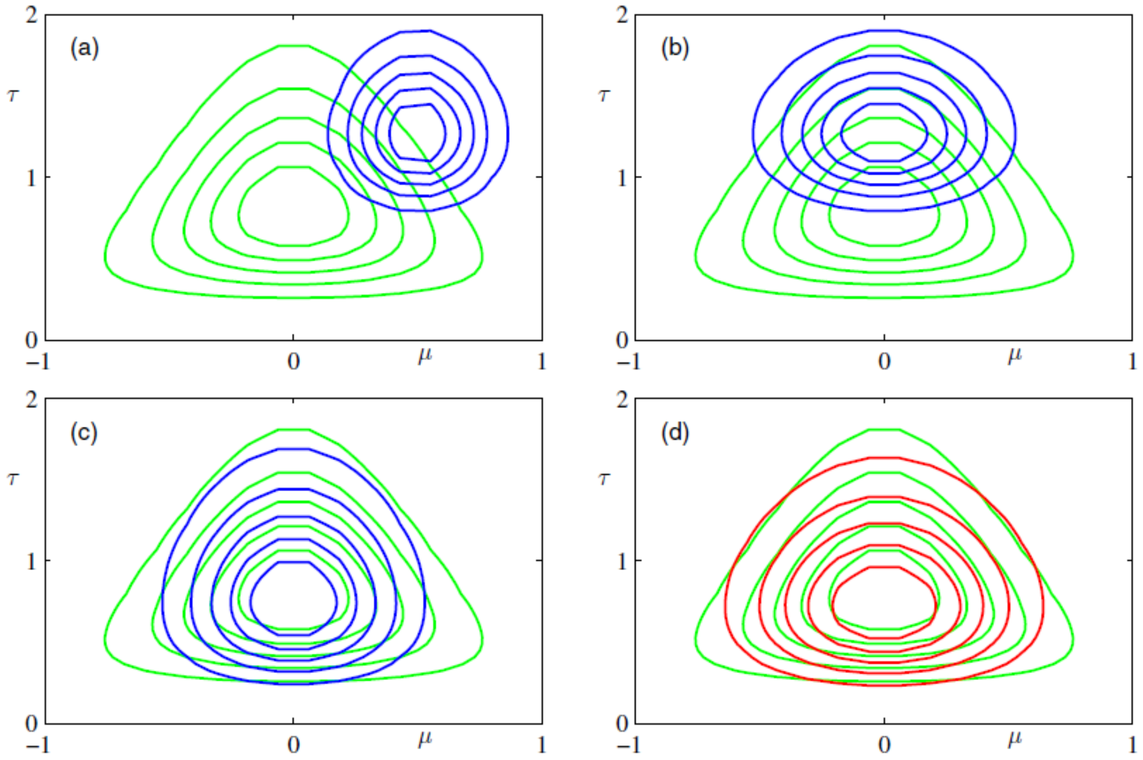


Figure 10.4 Illustration of variational inference for the mean μ and precision τ of a univariate Gaussian distribution. Contours of the true posterior distribution $p(\mu, \tau|D)$ are shown in green. (a) Contours of the initial factorized approximation $q_\mu(\mu)q_\tau(\tau)$ are shown in blue. (b) After re-estimating the factor $q_\mu(\mu)$. (c) After re-estimating the factor $q_\tau(\tau)$. (d) Contours of the optimal factorized approximation, to which the iterative scheme converges, are shown in red.

Locally Conjugate Models

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

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

- This is interesting: The <u>form</u> of optimal $q_j(\mathbf{Z}_j)$ will be the same as the conditional posterior of $\mathbf{Z}_j$
- For locally conjugate models, $p(\mathbf{Z}_j|\mathbf{X},\mathbf{Z}_{-j})$ is easy to find, and usually an exp-fam dist.

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

where $\eta()$ denotes the natural params of this exp-fam distribution (would depends on $\mathbf{X}$ and $\mathbf{Z}_{-j}$)

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

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

ELBO Gradient

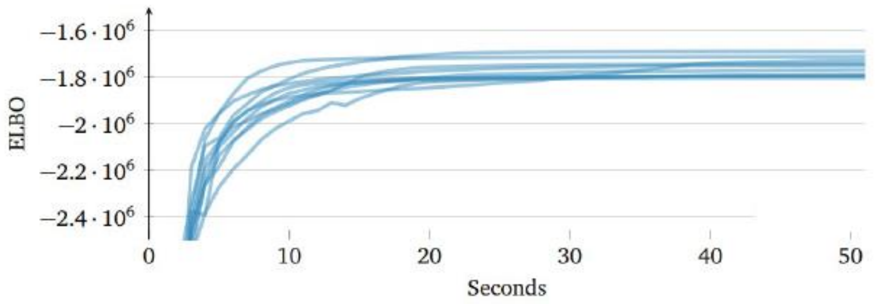
- More general way of doing VI is by computing ELBO's gradient and doing gradient ascent/descent
- The gradient based approach is broadly applicable, not just for mean-field VI. Works as follows
  1. Assume $q(\mathbf{Z})$ to be from some family of distributions with variational parameters ϕ
  2. Write down the **full ELBO** expression (this will give us a function of variational params ϕ)

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

  3. Compute **ELBO gradients**, i.e., $\nabla_\phi \mathcal{L}(\phi)$ and use gradient methods to find optimal ϕ
- Note: Step 2 may be simplified due to the problem structure or assumptions on the form of $q(\mathbf{Z})$
  - i.i.d. observations simplify $\log p(\mathbf{X}|\mathbf{Z})$; conditionally independent priors simplify $\log p(\mathbf{Z})$
  - Locally-conjugate models
  - The mean-field assumption simplifies $q(\mathbf{Z})$ as $q(\mathbf{Z}) = \prod_{i=1}^{M} q_i(\mathbf{Z}_i)$

Convergence of VI

- VI is guaranteed to converge but only to a local optima (just like EM)
- Therefore proper initialization is important (just like EM)



Different initializations may lead to different optima

- ELBO increases monotonically with iterations, so we can monitor the ELBO to assess convergence

Modern VI

- Moving beyond locally conjugate models
- Moving beyond the mean-field assumption
- More scalable variational inference
- General-purpose VI (that doesn't require model-specific derivations)
  - Posing VI as a general gradient based optimization problem

$$\phi^{new} = \phi^{old} + \eta \times \nabla_\phi \left[\mathbb{E}_{q_\phi}[\log p(\mathbf{X}, \mathbf{Z})] - \mathbb{E}_{q_\phi}[\log q(\mathbf{Z}|\phi)]\right]$$

  - A lot of recent research on approximating the gradient of an expectation

Questions