Chapter 8, Part 2: Variational Inference

Advanced Topics in Statistical Machine Learning

Tom Rainforth

Hilary 2024

rainforth@stats.ox.ac.uk

Variational Inference

- Variational inference (VI) methods are a class of ubiquitously used approaches for Bayesian inference wherein we try to directly learn an approximation for the posterior $p(\mathbf{Z}|\mathbf{X})$[1]
- Key idea: reformulate the inference problem to an **optimization** by learning parameters of a posterior approximation
- We do this through introducing a parameterized variational family $q_\phi(\mathbf{Z})$ then finding the ϕ that gives the "best" approximation
- VI is especially powerful for factorized LVMs as it will allow easy and effective exploitation of the factorization

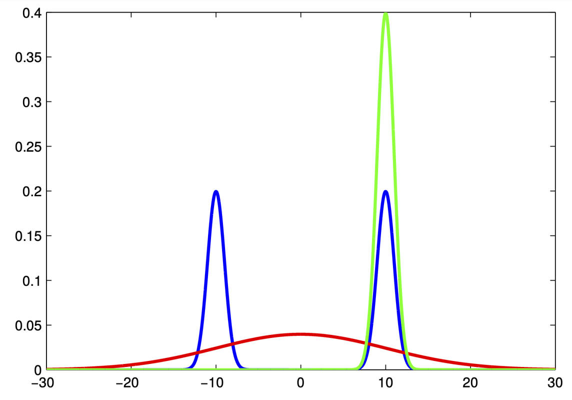
[1] Note we will differ from our earlier notation on Bayesian modeling to match that of LVMs: our data is $\mathbf{X}$ and our target for inference is $\mathbf{Z}$. However, the ideas introduced apply more generally and so, for now, we will omit any separate consideration of deterministic global parameters θ, while the form of $\mathbf{Z}$ is taken to be arbitrary: it need not be the case that $\mathbf{Z}$ is collection of latents associated with individual datapoints.

Divergences

- How do we quantitatively assess how similar two distributions P and Q are to one another?
- Similarity between distributions is much more subjective than you might expect, particularly for continuous variables
- A **divergence** $\mathbb{D}(P||Q)$ is a, typically asymmetric, way of measuring **dissimilarity** between two distributions P and Q
- We already came across an example divergence in the form of the MMD (which was special case of symmetric divergence, hence a proper distance metric)

Subjectivity of Divergences

Which is the best fitting Gaussian to our target blue distribution?



Either can be the best depending how we define our divergence

Image Credit: Barber, Section 28.3.4

The Kullback–Leibler (KL) Divergence

The Kullback–Leibler (KL) divergence is one of the most commonly used due to its simplicity, useful computational properties, and the fact that it naturally arises in a number of scenarios

$$\mathbb{D}_{\mathsf{KL}}(Q \parallel P) = \mathbb{E}_{X \sim Q}\left[\log\left(\frac{q(X)}{p(X)}\right)\right] \tag{1}$$

As we will mostly be dealing with densities, we will use the slightly imprecise notation

$$\mathbb{D}_{\mathsf{KL}}(q(x) \parallel p(x)) = \mathbb{E}_{q(x)}\left[\log\left(\frac{q(x)}{p(x)}\right)\right]$$

Important properties:

- $\mathbb{D}_{\mathsf{KL}}(q(x) \parallel p(x)) \geq 0, \ \forall q(x), p(x)$ **(Gibbs' inequality)**
- $\mathbb{D}_{\mathsf{KL}}(q(x) \parallel p(x)) = 0$ if and only if $p(x) = q(x) \, \forall x$
- In general, $\mathbb{D}_{\mathsf{KL}}(q(x) \parallel p(x)) \neq \mathbb{D}_{\mathsf{KL}}(p(x) \parallel q(x))$

Asymmetry of KL Divergence

Blue: target $p(x)$

Green: Gaussian $q(x)$ that minimizes $\mathbb{D}_{\mathsf{KL}}(q(x) \parallel p(x))$

Red: Gaussian $q(x)$ that minimizes $\mathbb{D}_{\mathsf{KL}}(p(x) \parallel q(x))$

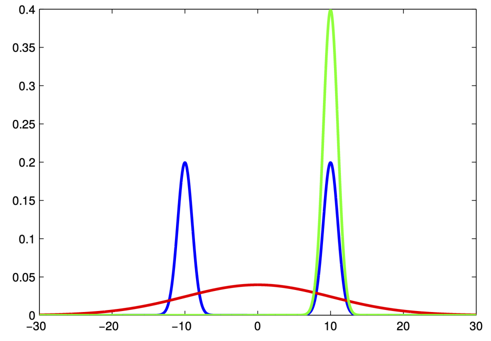


Image Credit: Barber, Section 28.3.4

Mode Covering KL

Let $p(x)$ again be our target and $q(x)$ our approximation

The "forward KL," $\mathbb{D}_{\mathsf{KL}}(p(x) \,\|\, q(x))$, is **mode covering**: $q(x)$ must place mass anywhere $p(x)$ does

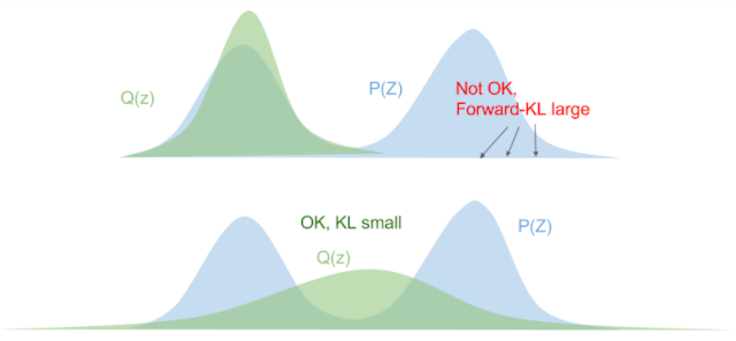


Image Credit: Eric Jang

Mode Seeking KL

The "reverse KL," $\mathbb{D}_{\mathsf{KL}}(q(x) \,\|\, p(x))$, is **mode seeking**: $q(x)$ must not place mass anywhere $p(x)$ does not

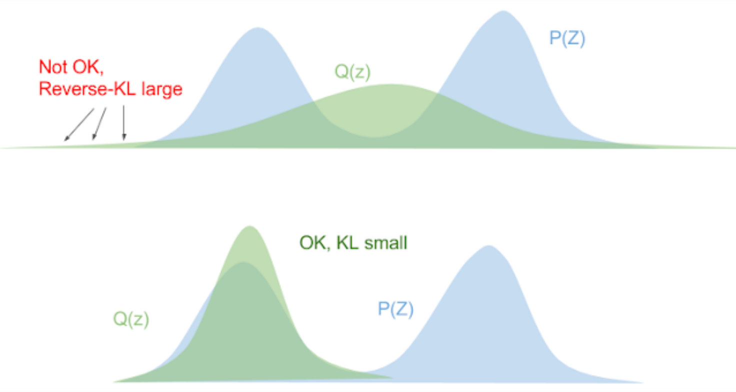


Image Credit: Eric Jang

Tail Behavior

- We can get insights into why this happens by considering the cases $q(x) \to 0$ and $p(x) \to 0$, noting that $\lim_{x \to 0} x \log x = 0$
- If $q(x) = 0$ when $p(x) > 0$, then $q(x) \log(q(x)/p(x)) = 0$ and $p(x) \log(p(x)/q(x)) = \infty$
    - $\mathbb{D}_{\mathsf{KL}}(p(x) \,\|\, q(x)) = \infty$ if $q(x) = 0$ anywhere $p(x) > 0$
    - $\mathbb{D}_{\mathsf{KL}}(q(x) \,\|\, p(x))$ is still fine when this happens
- By symmetry, $\mathbb{D}_{\mathsf{KL}}(q(x) \,\|\, p(x)) = \infty$ if $p(x) = 0$ anywhere $q(x) > 0$ anywhere, but now $\mathbb{D}_{\mathsf{KL}}(p(x) \,\|\, q(x))$ is fine

Variational Inference

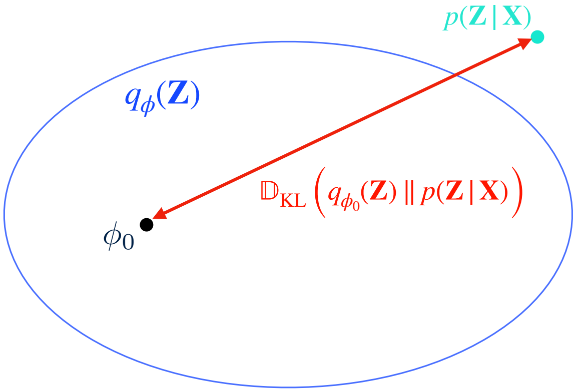
In variational inference we learn an approximation of a posterior $p(\mathbf{Z}|\mathbf{X})$ by introducing a parameterized **variational family** $q_\phi(\mathbf{Z})$ and then optimizing the **variational parameters** ϕ to minimize the KL divergence to $p(\mathbf{Z}|\mathbf{X})$. That is we find

$$\phi^* = \arg\min_\phi \mathbb{D}_{\mathsf{KL}}(q_\phi(\mathbf{Z}) \,\|\, p(\mathbf{Z}|\mathbf{X})) \tag{2}$$

This allows us to convert the original inference problem into an **optimization**

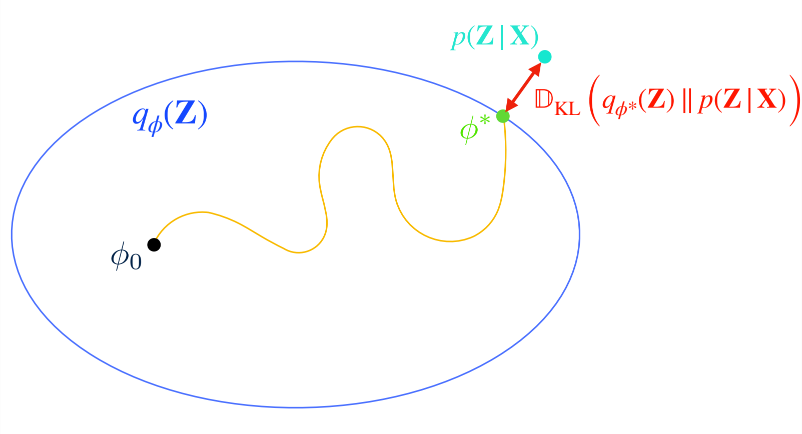
Critically, we will find that we only need the **unnormalized** form of this posterior, $p(\mathbf{X}, \mathbf{Z})$, to perform this optimization

Variational Inference (2)



Images based on example from David Blei

Variational Inference (2)



Images based on example from David Blei

Variational Inference (3)

We cannot work directly with $\mathbb{D}_{\mathsf{KL}}(q_\phi(\mathbf{Z}) \parallel p(\mathbf{Z}|\mathbf{X}))$ because we don't know the posterior density

However, by noting that the marginal likelihood $p(\mathbf{X})$ is independent of our variational parameters ϕ, we see that we can work with the joint instead

$$\begin{aligned}
\phi^* &= \arg\min_{\phi} \mathbb{D}_{\mathsf{KL}}(q_\phi(\mathbf{Z}) \parallel p(\mathbf{Z}|\mathbf{X})) \\
&= \arg\min_{\phi} \mathbb{E}_{q_\phi(\mathbf{Z})}\left[\log\left(\frac{q_\phi(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X})}\right)\right] - \log p(\mathbf{X}) \\
&= \arg\min_{\phi} \mathbb{E}_{q_\phi(\mathbf{Z})}\left[\log\left(\frac{q_\phi(\mathbf{Z})}{p(\mathbf{X},\mathbf{Z})}\right)\right]
\end{aligned}$$

This trick is a large part of why we work with $\mathbb{D}_{\mathsf{KL}}(q_\phi(\mathbf{Z}) \parallel p(\mathbf{Z}|\mathbf{X}))$ rather than $\mathbb{D}_{\mathsf{KL}}(p(\mathbf{Z}|\mathbf{X}) \parallel q_\phi(\mathbf{Z}))$

The ELBO

We can equivalently think about the optimization problem in VI as the maximization

$$\phi^* = \arg\max_{\phi} \mathcal{L}(\phi)$$

$$\text{where} \quad \mathcal{L}(\phi) := \mathbb{E}_{q_\phi(\mathbf{Z})}\left[\log\left(\frac{p(\mathbf{X},\mathbf{Z})}{q_\phi(\mathbf{Z})}\right)\right] = \log p(\mathbf{X}) - \mathbb{D}_{\mathsf{KL}}(q_\phi(\mathbf{Z}) \,\|\, p(\mathbf{Z}|\mathbf{X}))$$

$\mathcal{L}(\phi)$ is known as the **evidence lower bound (ELBO)** or occasionally as the **variational free energy**

Note that if our variational approximation is exact, that is $q_\phi(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X})$, then $\mathcal{L}(\phi) = \log p(\mathbf{X})$ such that it exactly equals the log evidence

The ELBO (2)

The name ELBO comes from the fact that it is a lower bound on the log evidence by Jensen's inequality using the concavity of log

$$\mathbb{E}_{q_\phi(\mathbf{Z})}\left[\log\left(\frac{p(\mathbf{X},\mathbf{Z})}{q_\phi(\mathbf{Z})}\right)\right] \leq \log\left(\mathbb{E}_{q_\phi(\mathbf{Z})}\left[\frac{p(\mathbf{X},\mathbf{Z})}{q_\phi(\mathbf{Z})}\right]\right) = \log p(\mathbf{X})$$

This bound is **tight** when $q_\phi(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X})$

$$\log(w_1 u_1 + w_2 u_2) \geq w_1 \log u_1 + w_2 \log u_2$$

for any $u_1, u_2 > 0$ and $w_1 + w_2 = 1$

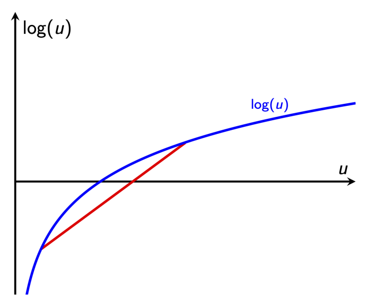
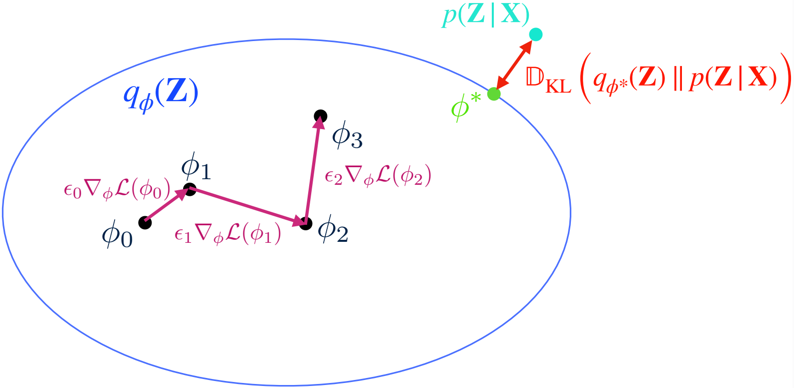
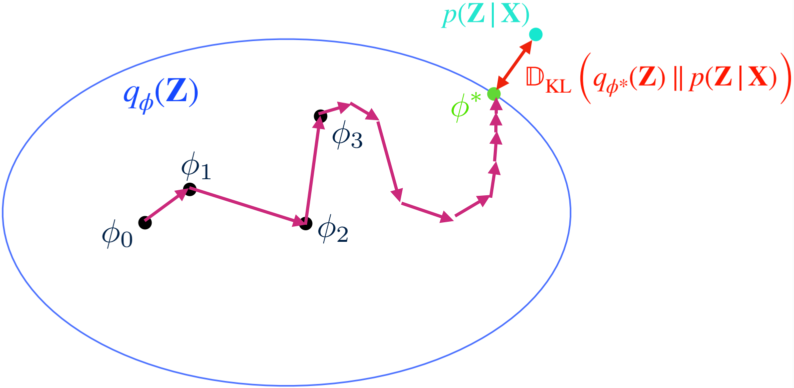


Image Credit: Michael Gutmann

Optimizing the ELBO



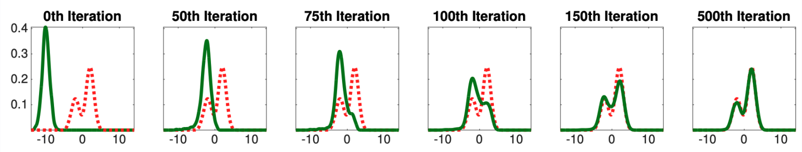
Optimizing the ELBO



Variational Approximation

When learned in this way, the variational approximation will improve with training, i.e. as we update ϕ to increase the ELBO, until reaching a local optimum (or saddle point) in $\mathcal{L}(\phi)$

Example convergence of a variational approximation (green) to a 1D target (dotted red):



Credit: Liu and Wang 2016 `https://arxiv.org/abs/1608.04471`

Worked Example—Gaussian with Unknown Mean and Variance

As a simple worked example (taken from Bishop 10.1.3), consider the following model where we are trying to infer the mean μ and precision τ of a Gaussian (such that $\mathbf{Z} = \{\mu, \tau\}$) given a set of observations $\mathbf{X} = \{x_i\}_{i=1}^n$.

Our full model is given by

$$
\begin{aligned}
p(\tau) &= \textsc{Gamma}(\tau; \alpha, \beta) \\
p(\mu|\tau) &= \mathcal{N}(\mu; \mu_0, (\lambda_0 \tau)^{-1}) \\
p(\mathbf{X}|\mu, \tau) &= \prod_{i=1}^{n} \mathcal{N}(x_i; \mu, \tau^{-1})
\end{aligned}
$$

Worked Example—Gaussian with Unknown Mean and Variance

We care about the posterior $p(\mu, \tau|\mathbf{X})$ and we are going to try and approximate this using variational inference

For our variational family we will take

$$\begin{aligned} q_\phi(\tau, \mu) &= q_{\phi_\tau}(\tau) q_{\phi_\mu}(\mu) \\ q_{\phi_\tau}(\tau) &= \textsc{Gamma}(\tau; \phi_{\tau,1}, \phi_{\tau,2}) \\ q_{\phi_\mu}(\mu) &= \mathcal{N}(\mu; \phi_{\mu,1}, \phi_{\mu,2}^{-1}) \end{aligned}$$

where $\phi = \{\phi_{\tau,1}, \phi_{\tau,2}, \phi_{\mu,1}, \phi_{\mu,2}\}$ and we note that the factorization is an approximation: the posterior itself does not factorize

Mean–Field Approximations

In this example we chose a factorized variational approximation:

$$q_\phi(\tau, \mu) = q_{\phi_\tau}(\tau) q_{\phi_\mu}(\mu)$$

This factorization is actually a special case of a common simplifying assumption known as a **mean–field** approximation

More generally we have

$$q_\phi(\mathbf{Z}) = \prod_i q_{\phi_{z_i}}(z_i)$$

where each $q_{\phi_{z_i}}$ is its is own separate[2] variational approximation for the subset of parameters $z_i \subset \mathbf{Z}$ (with $\cup_i z_i = \mathbf{Z}$)

There are a number of scenarios where this can help make maximizing the ELBO more tractable

[2] As we will see next time, one sometimes introduces parameter sharing mechanisms across these approximations

Coordinate Ascent Variational Inference

Using a mean–field approximation gives a closed form solution for the optimal $q_{\phi_{z_i}}$ given $\{q_{\phi_{z_j}}\}_{j\neq i}$ (examples sheet):

$$q_i^*(z_i) \propto \exp\left(\mathbb{E}_{\prod_{j\neq i} q_{\phi_{z_j}}(z_j)}\left[\log p(\mathbf{X}, \mathbf{Z})\right]\right)$$
$$\propto \exp\left(\mathbb{E}_{\prod_{j\neq i} q_{\phi_{z_j}}(z_j)}\left[\log p(z_i|\mathbf{X}, \mathbf{Z}\backslash z_i)\right]\right)$$

If the conditionals $p(z_i|\mathbf{X}, \mathbf{Z}\backslash z_i)$ are in the exponential family, this can be calculated analytically.

This allows gradientless **coordinate ascent variational inference** (CAVI), where one simply cycles over each i and directly calculates the corresponding q_i^* and sets $q_{\phi_{z_i}} \leftarrow q_i^*$ until the ELBO converges

Occasionally one also runs a CAVI approach even when these updates can only be done approximately, e.g. alternating between approximating global parameters and local latents

Worked Example—Gaussian with Unknown Mean and Variance

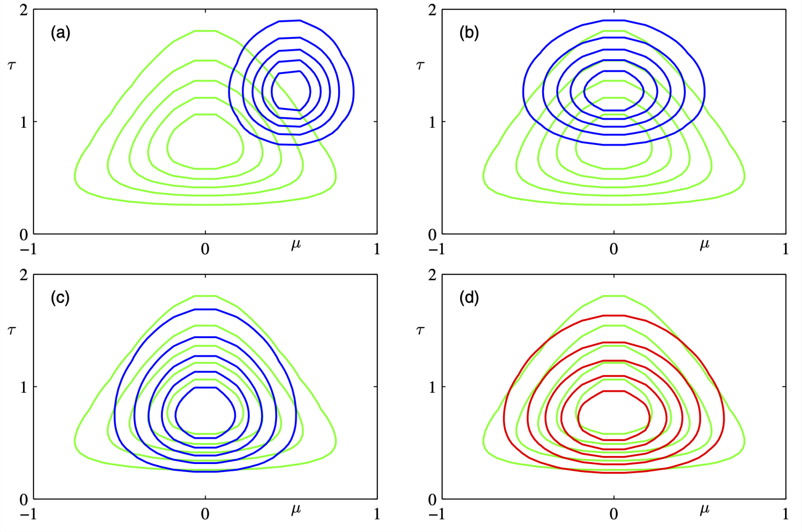
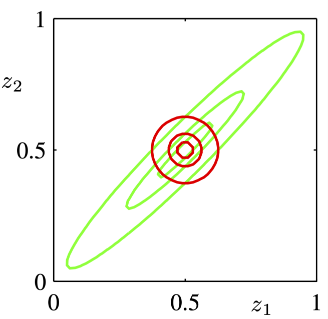


Figure 10.4 from Bishop

Effect of Mean–field Approximations

Mean–field assumptions negate dependencies between z_i; the reasonableness of this will depend on the problem and how we breakdown $\mathbf{Z}$

A secondary effect of mean–field approximations is that they tend to lead to underestimating the variance once coupled with the mode–seeking behavior of $\mathbb{D}_{\mathsf{KL}}(q_\phi(\mathbf{Z}) \,\|\, p(\mathbf{Z}|\mathbf{X}))$



Optimal variational approximation (red) for target in green when making a mean–field assumption of two dimensions of $\mathbf{Z}$. Taken from Figure 10.2 in Bishop

Variational Inference for Factorized LVMs

Consider a factorized LVM with fixed global parameters θ:

$$p_\theta(\mathbf{X}, \mathbf{Z}) = \prod_{i=1}^{n} p_\theta(z_i) p_\theta(x_i | z_i)$$

Here we have

$$p_\theta(z_i | \mathbf{X}, \mathbf{Z} \backslash z_i) = p_\theta(z_i | x_i)$$

such that the optimal posterior approximation is separable and making a mean–field approximation over different **datapoints** is not actually an approximation at all, with

$$q_i^*(z_i) = p_\theta(z_i | x_i)$$

We can thus think about doing inference separately for each individual datapoint

Variational Inference for Factorized LVMs (2)

The ELBO is also separable in this case, such that we have

$$\mathcal{L}(\theta, \phi) = \sum_{i=1}^{n} \mathcal{L}(x_i, \theta, \phi_{z_i})$$

where $\quad \mathcal{L}(x_i, \theta, \phi_{z_i}) := \mathbb{E}_{q_{\phi_{z_i}}(z_i)}\left[\log\left(\frac{p_\theta(x_i, z_i)}{q_{\phi_{z_i}}(z_i)}\right)\right] \leq \log p_\theta(x_i)$

We can thus think about independently running VI for each datapoint based on its individual ELBO $\mathcal{L}(x_i, \theta, \phi_{z_i})$

Though this in itself is not of that much direct interest (e.g. we could also separately run MCMC for each z_i), we will see next time that it becomes critical in two scenarios:

- Training (or performing inference for) the global parameters θ
- Performing **amortized** inference, where we learn an **inference network** $q_\phi(z|x)$ that maps from inputs to approximations

Pros and Cons of Variational Methods

Pros

- Typically more efficient than MCMC approaches, particularly in high dimensions once we exploit the stochastic variational approaches introduced in the next lecture
  - Can often provided effective inference for models where MCMC methods have impractically slow convergence
- Allows simultaneous optimization of model parameters as we will show in the next lecture

Pros and Cons of Variational Methods (2)

Cons

- It produces (potentially very) biased estimates and requires strong structural assumptions to be made about the form of the posterior
  - Unlike MCMC methods, this bias stays even in the limit of large computation
- Can require substantial tailoring to a particular problem
- Very difficult to estimate how much error their is in the approximation: subsequent estimates can be unreliable, particular in their uncertainty
- Tends to underestimate the variance of the posterior due to mode–seeking nature of reverse KL, particularly when using inexact mean–field approximations

Further Reading

- Chapters 9 and 10 of C M Bishop. **Pattern recognition and machine learning**. 2006
- David M Blei, Alp Kucukelbir, and Jon D McAuliffe. "Variational inference: A review for statisticians". In: **Journal of the American statistical Association** (2017)
- NeurIPS tutorial on variational inference that accompanies the previous paper: `https://www.youtube.com/watch?v=ogdv_6dbvVQ`
- Powerful modern deep variational families known as normalizing flows: `https://arxiv.org/pdf/1912.02762.pdf`