Lecture : Approximate Inference

Riashat Islam

Reasoning and Learning Lab
McGill University

October 29, 2018

Approximate Inference

Sampling and Variational Approximations

Inference Problem

Given a dataset $\mathcal{D} = \{x_1, ..., x_n\}$:

Bayes Rule:

$$P(\theta|\mathcal{D}) = \frac{P(D|\theta)P(\theta)}{P(\mathcal{D})}$$

$P(\mathcal{D}\|\theta)$	Likelihood function of θ
$P(\theta)$	Prior probability of θ
$P(\theta\|\mathcal{D})$	Posterior distribution over θ

Computing posterior distribution is known as the **inference** problem.
But:

$$P(\mathcal{D}) = \int P(\mathcal{D}, \theta) d\theta$$

This integral can be very high-dimensional and difficult to compute.

Prediction

$$P(\theta|\mathcal{D}) = \frac{P(D|\theta)P(\theta)}{P(\mathcal{D})}$$

$P(\mathcal{D}|\theta)$	Likelihood function of θ
$P(\theta)$	Prior probability of θ
$P(\theta|\mathcal{D})$	Posterior distribution over θ

Prediction: Given $\mathcal{D}$, computing conditional probability of x^* requires computing the following integral:

$$\begin{aligned} P(x^*|\mathcal{D}) &= \int P(x^*|\theta, \mathcal{D})P(\theta|\mathcal{D})d\theta \\ &= \mathbb{E}_{P(\theta|\mathcal{D})}[P(x^*|\theta, \mathcal{D})] \end{aligned}$$

which is sometimes called **predictive distribution**.

Computing predictive distribution requires posterior $P(\theta|\mathcal{D})$.

Inference

Observe data: $\mathcal{D} = \{\mathbf{x}^{(n)}, y^{(n)}\}$

Unknowns: $\theta = \{\mathbf{w}, \alpha, \epsilon, \Sigma, \{z^{(n)}\}, \ldots\}$

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

Marginalization

Interested in particular parameter θ_i

$$p(\theta_i \,|\, \mathcal{D}) = \int p(\theta \,|\, \mathcal{D}) \, \mathrm{d}\theta_{\setminus i}$$

Sampling solution:

— Sample everything: $\theta^{(s)} \sim p(\theta \,|\, \mathcal{D})$

— $\theta_i^{(s)}$ comes from marginal $p(\theta_i \,|\, \mathcal{D})$

Computational Challenges

- Computing marginal likelihoods often requires computing very high dimensional integrals
- Computing posterior distributions (and hence the predictive distribution) is often **analytically intractable**

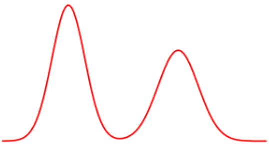
Approximation Methods for Posteriors and Marginal Likelihoods

Markov Chain Monte-Carlo Methods (MCMC)

Variational Approximations

Expectation Propagation (not covered here..)

Inference



For most situations we will be interested in evaluating the expectation:

$$\mathbb{E}[f] = \int f(\mathbf{z})p(\mathbf{z})dz$$

We will use the following notation: $p(\mathbf{z}) = \frac{\tilde{p}(\mathbf{z})}{\mathcal{Z}}$.

We can evaluate $\tilde{p}(\mathbf{z})$ pointwise, but cannot evaluate $\mathcal{Z}$.

- Posterior distribution: $P(\theta|\mathcal{D}) = \frac{1}{P(\mathcal{D})}P(\mathcal{D}|\theta)P(\theta)$
- Markov random fields: $P(z) = \frac{1}{\mathcal{Z}}\exp(-E(z))$

Markov Chain Monte-Carlo Methods (MCMC)

An Overview of Sampling Methods

Monte Carlo Methods (last lecture)

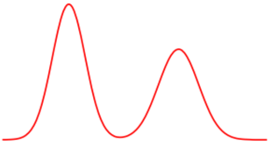
- Simple Monte Carlo
- Rejection Sampling
- Importance Samping

Markov Chain Monte Carlo Methods

- Gibbs Sampling
- Metropolis Algorithm

Recap : Importance Sampling

Suppose we have an easy-to-sample *proposal distribution* $q(z)$, such that $q(z) > 0$ if $p(z) > 0$.



$$\begin{aligned}
\mathbb{E}[f] &= \int f(z)p(z)dz \\
&= \int f(z)\frac{p(z)}{q(z)}q(z)dz \\
&\approx \frac{1}{N}\sum_n \frac{p(z^n)}{q(z^n)}f(z^n), \quad z^n \sim q(z)
\end{aligned}$$

The quantities $w^n = {}^{p(z^n)}/_{q(z^n)}$ are known as **importance weights**.
Unlike rejection sampling, all samples are retained.
But wait: we cannot compute $p(z)$, only $\tilde{p}(z)$.

Problems

If our proposal distribution $q(z)$ poorly matches our target distribution $p(z)$ then:

- Rejection Sampling: almost always rejects
- Importance Sampling: has large, possibly infinite, variance (unreliable estimator).

For high-dimensional problems, finding good proposal distributions is very hard. What can we do?

Markov Chain Monte Carlo.

Markov Chains

A first-order Markov chain: a series of random variables $\{z^1, ..., z^N\}$ such that the following conditional independence property holds for $n \in \{z^1, ..., z^{N-1}\}$:

$$p(z^{n+1}|z^1, ..., z^n) = p(z^{n+1}|z^n)$$

We can specify Markov chain:

- probability distribution for initial state $p(z^1)$.
- conditional probability for subsequent states in the form of transition probabilities $T(z^{n+1} \leftarrow z^n) \equiv p(z^{n+1}|z^n)$.

Remark: $T(z^{n+1} \leftarrow z^n)$ is sometimes called a **transition kernel**.

Markov Chains

A marginal probability of a particular state can be computed as:

$$p(z^{n+1}) = \sum_{z^n} T(z^{n+1} \leftarrow z^n) p(z^n)$$

A distribution $\pi(z)$ is said to be **invariant** or **stationary** with respect to a Markov chain if each step in the chain leaves $\pi(z)$ invariant:

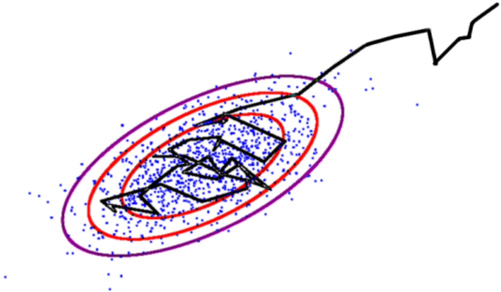
$$\pi(z) = \sum_{z'} T(z \leftarrow z') \pi(z')$$

A given Markov chain may have many stationary distributions. For example: $T(z \leftarrow z') = I\{z = z'\}$ is the identity transformation. Then any distribution is invariant.

Markov Chain Monte Carlo

Construction a random walk that explores $P(x)$

Markov steps $x_t \sim T(x_t \leftarrow x_{t-1})$



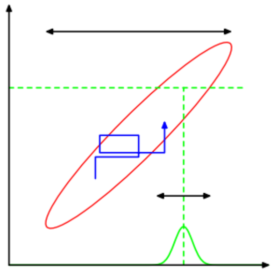
MCMC gives approximate, correlated samples from $P(x)$

Markov Chain Monte Carlo

- Markov chain Monte Carlo (MCMC) methods also use a proposal distribution to generate samples from another distribution
- Unlike the previous methods, we keep track of the samples generated $\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(\tau)}$
- The proposal distribution depends on the current state: $q(\boldsymbol{z}|\boldsymbol{z}^{(\tau)})$
  - Intuitively, walking around in state space, each step depends only on the current state

Gibbs Sampler

Consider sampling from $p(z_1, ..., z_N)$.



Initialize z_i, $i = 1, ..., N$

For t=1,...,T

Sample $z_1^{t+1} \sim p(z_1 | z_2^t, ..., z_N^t)$

Sample $z_2^{t+1} \sim p(z_2 | z_1^{t+1}, x_3^t, ..., z_N^t)$

...

Sample $z_N^{t+1} \sim p(z_N | z_1^{t+1}, ..., z_{N-1}^{t+1})$

Gibbs sampler is a particular instance of M-H algorithm with proposals $p(z_n|\mathbf{z}_{i \neq n}) \rightarrow$ accept with probability 1. Apply a series (component-wise) of these operators.

Advantages : MCMC

Powerful tool for high-dimensional integrals

Good proposals may require ingenuity

Sometimes simple and routine

But can be **very slow**!

Main Problems of MCMC

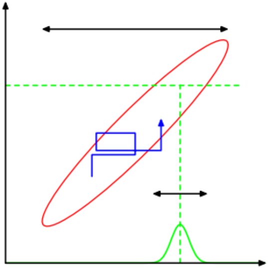
- Hard to diagnose convergence (burning in)
- Sampling from isolated modes

Hamiltonian Monte Carlo methods make use of gradient information (not covered here)

Variational Methods

Recap : EM Algorithm

Consider sampling from $p(z_1, ..., z_N)$.



Initialize z_i, $i = 1, ..., N$

For t=1,...,T

Sample $z_1^{t+1} \sim p(z_1 | z_2^t, ..., z_N^t)$

Sample $z_2^{t+1} \sim p(z_2 | z_1^{t+1}, x_3^t, ..., z_N^t)$

...

Sample $z_N^{t+1} \sim p(z_N | z_1^{t+1}, ..., z_{N-1}^{t+1})$

Gibbs sampler is a particular instance of M-H algorithm with proposals $p(z_n | \mathbf{z}_{i \neq n}) \rightarrow$ accept with probability 1. Apply a series (component-wise) of these operators.

Recap : EM Algorithm

Given observed/visible variables $\mathbf{y}$, unobserved/hidden/latent/missing variables $\mathbf{x}$, and model parameters θ, **maximize the likelihood** w.r.t. θ.

$$\mathcal{L}(\theta) = \log p(\mathbf{y}|\theta) = \log \int p(\mathbf{x}, \mathbf{y}|\theta) d\mathbf{x},$$

where we have written the marginal for the visibles in terms of an integral over the joint distribution for hidden and visible variables.

Using *Jensen's inequality*, any distribution[1] over hidden variables $q(\mathbf{x})$ gives:

$$\mathcal{L}(\theta) = \log \int q(\mathbf{x}) \frac{p(\mathbf{x}, \mathbf{y}|\theta)}{q(\mathbf{x})} d\mathbf{x} \geq \int q(\mathbf{x}) \log \frac{p(\mathbf{x}, \mathbf{y}|\theta)}{q(\mathbf{x})} d\mathbf{x} = \mathcal{F}(q, \theta),$$

defining the $\mathcal{F}(q, \theta)$ functional, which is a lower bound on the log likelihood.

In the EM algorithm, we alternately optimize $\mathcal{F}(q, \theta)$ wrt q and θ, and we can prove that this will never decrease $\mathcal{L}(\theta)$.

[1] s.t. $q(\mathbf{x}) > 0$ if $p(\mathbf{x}, \mathbf{y}|\theta) > 0$.

Recap : EM Algorithm

The lower bound on the log likelihood:

$$\mathcal{F}(q,\theta) = \int q(\mathbf{x}) \log \frac{p(\mathbf{x},\mathbf{y}|\theta)}{q(\mathbf{x})} d\mathbf{x} = \int q(\mathbf{x}) \log p(\mathbf{x},\mathbf{y}|\theta) d\mathbf{x} + \mathcal{H}(q),$$

where $\mathcal{H}(q) = -\int q(\mathbf{x}) \log q(\mathbf{x}) d\mathbf{x}$ is the entropy of q. We iteratively alternate:

E step: maximize $\mathcal{F}(q,\theta)$ wrt distribution over hidden variables given the parameters:

$$q^{(k)}(\mathbf{x}) := \underset{q(\mathbf{x})}{\operatorname{argmax}} \ \mathcal{F}\big(q(\mathbf{x}), \theta^{(k-1)}\big).$$

M step: maximize $\mathcal{F}(q,\theta)$ wrt the parameters given the hidden distribution:

$$\theta^{(k)} := \underset{\theta}{\operatorname{argmax}} \ \mathcal{F}\big(q^{(k)}(\mathbf{x}), \theta\big) = \underset{\theta}{\operatorname{argmax}} \int q^{(k)}(\mathbf{x}) \log p(\mathbf{x},\mathbf{y}|\theta) d\mathbf{x},$$

which is equivalent to optimizing the expected complete-data likelihood $p(\mathbf{x},\mathbf{y}|\theta)$, since the entropy of $q(\mathbf{x})$ does not depend on θ.

Variational Approximation

Assume your goal is to maximize likelihood $\ln p(\mathbf{y}|\theta)$.
Any distribution $q(\mathbf{x})$ over the hidden variables defines a lower bound on $\ln p(\mathbf{y}|\theta)$:

$$\ln p(\mathbf{y}|\theta) \geq \sum_{\mathbf{x}} q(\mathbf{x}) \ln \frac{p(\mathbf{x}, \mathbf{y}|\theta)}{q(\mathbf{x})} = \mathcal{F}(q, \theta)$$

Constrain $q(\mathbf{x})$ to be of a particular *tractable* form (e.g. factorised) and maximise $\mathcal{F}$ subject to this constraint

- **E-step:** Maximise $\mathcal{F}$ w.r.t. q with θ fixed, subject to the constraint on q, equivalently minimize:

  $$\ln p(\mathbf{y}|\theta) - \mathcal{F}(q, \theta) = \sum_{\mathbf{x}} q(\mathbf{x}) \ln \frac{q(\mathbf{x})}{p(\mathbf{x}|\mathbf{y}, \theta)} = \mathrm{KL}(q\|p)$$

  The inference step therefore tries to find q closest to the exact posterior distribution.

- **M-step:** Maximise $\mathcal{F}$ w.r.t. θ with q fixed

Variational Bayesian Learning

Let the latent variables be $\mathbf{x}$, observed data $\mathbf{y}$ and the parameters $\boldsymbol{\theta}$.
We can lower bound the marginal likelihood (Jensen's inequality):

$$
\begin{aligned}
\ln p(\mathbf{y}|m) &= \ln \int p(\mathbf{y},\mathbf{x},\boldsymbol{\theta}|m)\, d\mathbf{x}\, d\boldsymbol{\theta} \\
&= \ln \int q(\mathbf{x},\boldsymbol{\theta}) \frac{p(\mathbf{y},\mathbf{x},\boldsymbol{\theta}|m)}{q(\mathbf{x},\boldsymbol{\theta})}\, d\mathbf{x}\, d\boldsymbol{\theta} \\
&\geq \int q(\mathbf{x},\boldsymbol{\theta}) \ln \frac{p(\mathbf{y},\mathbf{x},\boldsymbol{\theta}|m)}{q(\mathbf{x},\boldsymbol{\theta})}\, d\mathbf{x}\, d\boldsymbol{\theta}.
\end{aligned}
$$

Use a simpler, factorised approximation for $q(\mathbf{x},\boldsymbol{\theta}) \approx q_{\mathbf{x}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$:

$$
\begin{aligned}
\ln p(\mathbf{y}|m) &\geq \int q_{\mathbf{x}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \ln \frac{p(\mathbf{y},\mathbf{x},\boldsymbol{\theta}|m)}{q_{\mathbf{x}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}\, d\mathbf{x}\, d\boldsymbol{\theta} \\
&\stackrel{\text{def}}{=} \mathcal{F}_m(q_{\mathbf{x}}(\mathbf{x}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta}), \mathbf{y}).
\end{aligned}
$$

Variational Bayesian Learning

Maximizing this lower bound, $\mathcal{F}_m$, leads to **EM-like** iterative updates:

$$q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) \propto \exp\left[\int \ln p(\mathbf{x},\mathbf{y}|\boldsymbol{\theta},m)\, q_{\boldsymbol{\theta}}^{(t)}(\boldsymbol{\theta})\, d\boldsymbol{\theta}\right] \qquad \text{E-like step}$$

$$q_{\boldsymbol{\theta}}^{(t+1)}(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta}|m)\, \exp\left[\int \ln p(\mathbf{x},\mathbf{y}|\boldsymbol{\theta},m)\, q_{\mathbf{x}}^{(t+1)}(\mathbf{x})\, d\mathbf{x}\right] \qquad \text{M-like step}$$

Maximizing $\mathcal{F}_m$ is equivalent to minimizing KL-divergence between the *approximate posterior*, $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})\, q_{\mathbf{x}}(\mathbf{x})$ and the *exact posterior*, $p(\boldsymbol{\theta},\mathbf{x}|\mathbf{y},m)$:

$$\ln p(\mathbf{y}|m) - \mathcal{F}_m(q_{\mathbf{x}}(\mathbf{x}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta}), \mathbf{y}) = \int q_{\mathbf{x}}(\mathbf{x})\, q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \ln \frac{q_{\mathbf{x}}(\mathbf{x})\, q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}{p(\boldsymbol{\theta},\mathbf{x}|\mathbf{y},m)}\, d\mathbf{x}\, d\boldsymbol{\theta} = \mathbf{KL}(q\|p)$$

In the limit as $n \to \infty$, for identifiable models, the variational lower bound approaches the BIC criterion.

Variational Bayesian Learning

EM for MAP estimation

Goal: maximize $p(\boldsymbol{\theta}|\mathbf{y},m)$ w.r.t. $\boldsymbol{\theta}$

E Step: compute

$$q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) = p(\mathbf{x}|\mathbf{y},\boldsymbol{\theta}^{(t)})$$

M Step:

$$\boldsymbol{\theta}^{(t+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \int q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) \ln p(\mathbf{x},\mathbf{y},\boldsymbol{\theta})\, d\mathbf{x}$$

Variational Bayesian EM

Goal: lower bound $p(\mathbf{y}|m)$

VB-E Step: compute

$$q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) = p(\mathbf{x}|\mathbf{y},\bar{\boldsymbol{\phi}}^{(t)})$$

VB-M Step:

$$q_{\boldsymbol{\theta}}^{(t+1)}(\boldsymbol{\theta}) \propto \exp\left[\int q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) \ln p(\mathbf{x},\mathbf{y},\boldsymbol{\theta})\, d\mathbf{x}\right]$$

Properties:

- Reduces to the EM algorithm if $q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta}-\boldsymbol{\theta}^*)$.
- $\mathcal{F}_m$ increases monotonically, and incorporates the model complexity penalty.
- Analytical parameter distributions (but not constrained to be Gaussian).
- VB-E step has same complexity as corresponding E step.
- We can use the junction tree, belief propagation, Kalman filter, etc, algorithms in the VB-E step of VB-EM, but **using expected natural parameters**, $\bar{\boldsymbol{\phi}}$.

Variational Bayesian Learning

The Variational Bayesian EM algorithm has been used to approximate Bayesian learning in a wide range of models such as:

- probabilistic PCA and factor analysis
- mixtures of Gaussians and mixtures of factor analysers
- hidden Markov models
- state-space models (linear dynamical systems)
- independent components analysis (ICA)
- discrete graphical models...

The main advantage is that it can be used to **automatically do model selection** and does not suffer from overfitting to the same extent as ML methods do.

Also it is about as computationally demanding as the usual EM algorithm.

See: `www.variational-bayes.org`

Variational Inference

Key Idea: Approximate intractable distribution $p(\theta|D)$ with simpler, tractable distribution $q(\theta)$.

We can lower bound the marginal likelihood using Jensen's inequality:

$$
\begin{aligned}
\ln p(\mathcal{D}) &= \ln \int p(\mathcal{D},\theta) d\theta = \ln \int q(\theta) \frac{P(\mathcal{D},\theta)}{q(\theta)} d\theta \\
&\geq \int q(\theta) \ln \frac{p(\mathcal{D},\theta)}{q(\theta)} d\theta = \underbrace{\int q(\theta) \ln p(\mathcal{D},\theta) d\theta + \underbrace{\int q(\theta) \ln \frac{1}{q(\theta)} d\theta}_{\text{Entropy functional}}}_{\text{Variational Lower-Bound}} \\
&= \ln p(\mathcal{D}) - \mathrm{KL}(q(\theta)||p(\theta|D)) = \mathcal{L}(q)
\end{aligned}
$$

where $\mathrm{KL}(q||p)$ is a Kullback–Leibler divergence. It is a non-symmetric measure of the difference between two probability distributions q and p.

The goal of variational inference is to maximize the variational lower-bound w.r.t. approximate q distribution, or minimize $\mathrm{KL}(q||p)$.

Variational Inference

Key Idea: Approximate intractable distribution $p(\theta|D)$ with simpler, tractable distribution $q(\theta)$ by minimizing $\mathrm{KL}(q(\theta)||p(\theta|D))$.

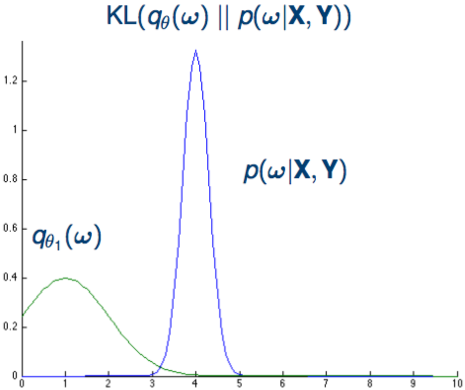
We can choose a fully factorized distribution: $q(\theta) = \prod_{i=1}^{D} q_i(\theta_i)$, also known as a mean-field approximation.

The variational lower-bound takes form:

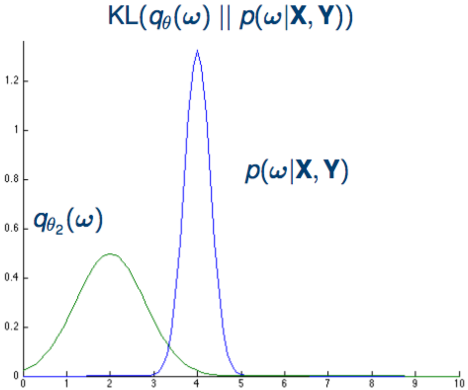
$$\begin{aligned}\mathcal{L}(q) &= \int q(\theta) \ln p(\mathcal{D}, \theta) d\theta + \int q(\theta) \ln \frac{1}{q(\theta)} d\theta \\ &= \int q_j(\theta_j) \underbrace{\left[\ln p(\mathcal{D}, \theta) \prod_{i \neq j} q_i(\theta_i) d\theta_i \right]}_{\mathbb{E}_{i \neq j}[\ln p(\mathcal{D}, \theta)]} d\theta_j + \sum_i \int q_i(\theta_i) \ln \frac{1}{q(\theta_i)} d\theta_i\end{aligned}$$

Suppose we keep $\{q_{i \neq j}\}$ fixed and maximize $\mathcal{L}(q)$ w.r.t. all possible forms for the distribution $q_j(\theta_j)$.

Variational Inference

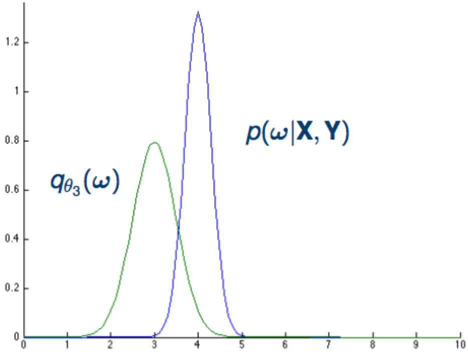


Variational Inference



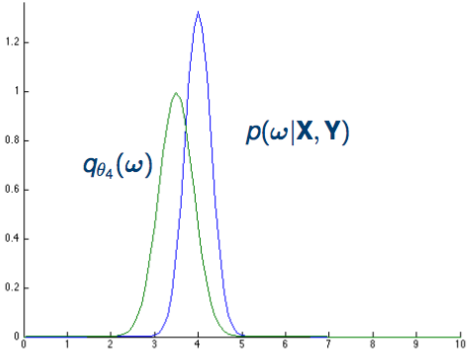
Variational Inference

$$\mathrm{KL}(q_\theta(\omega)\ ||\ p(\omega|\mathbf{X},\mathbf{Y}))$$



Variational Inference

$$\mathrm{KL}(q_\theta(\omega) \,||\, p(\omega|\mathbf{X},\mathbf{Y}))$$



Variational Inference