

# Variational Inference

**Mark van der Wilk**

Department of Computing  
Imperial College London

@markvanderwilk  
m.vdwilk@imperial.ac.uk

February 27, 2023

# Introduction and Background

# Approximate Inference Methods

- ▶ Laplace approximation
  - ▶ Procedure to give Gaussian
  - ▶ Fixed and limited approximation quality
  - ▶ No way to use better approximating distributions
  - ▶ No measure of quality of approximation
- ▶ Markov Chain Monte Carlo (to sample from the posterior)
  - ▶ Would always converge to the right answer
  - ▶ No idea about how long it takes to converge
- ▶ **Variational inference** (Jordan et al., 1999)
  - ▶ Somewhere in between
  - ▶ Can (in principle) use complicated approximating distributions
  - ▶ Has measure of approximation quality

## Further Reading

- ▶ Pattern Recognition and Machine Learning, Chapter 10 (Bishop, 2006)
- ▶ Machine Learning: A Probabilistic Perspective, Chapter 21 (Murphy, 2012)
- ▶ Variational Inference: A Review for Statisticians (Blei et al., 2017)
- ▶ NIPS-2016 Tutorial by Blei, Ranganath, Mohamed  
<https://nips.cc/Conferences/2016/Schedule?showEvent=6199>
- ▶ Tutorials by S. Mohamed  
<http://shakirm.com/papers/VITutorial.pdf>  
<http://shakirm.com/slides/MLSS2018-Madrid-ProbThinking.pdf>

# Variational Inference

- ▶ Variational inference is the most **scalable approximate inference method** available (at the moment)
- ▶ Can handle (arbitrarily) large datasets
- ▶ Applications include:
  - ▶ Topic modeling (Hoffman et al., 2013)
  - ▶ Community detection (Gopalan & Blei, 2013)
  - ▶ Genetic analysis (Gopalan et al., 2016)
  - ▶ Reinforcement learning (e.g., Eslami et al., 2016)
  - ▶ Neuroscience analysis (Manning et al., 2014)
  - ▶ Compression and content generation (Gregor et al., 2016)
  - ▶ Traffic analysis (Kucukelbir et al., 2016; Salimbeni & Deisenroth, 2017)

# Key Idea: Approximation by Optimization

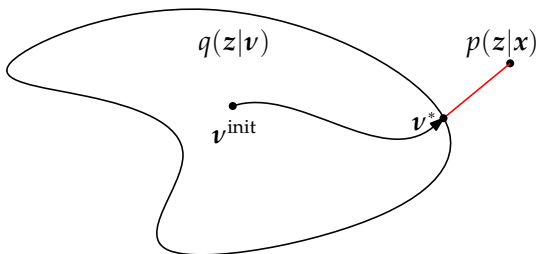


Figure adopted from Blei et al.'s NIPS-2016 tutorial

- ▶ Find approximation of a probability distribution (e.g., posterior) by **optimization**:
  1. Define a (parametrized) family of approximating distributions  $q_v$
  2. Define a measure of similarity of distributions to the true posterior
  3. Optimize objective function w.r.t. **variational parameters**  $v$
- ▶ Inference ►► Optimization

# From importance sampling to variational inference

# Problem setting

- ▶ We have the joint  $p(\mathbf{x}, \mathbf{z})$ .
- ▶ We are interested in posterior  $p(\mathbf{z}|\mathbf{x})$ .
- ▶ Marginal likelihood is  $p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{z}) d\mathbf{z}$ .

This is a very general formulation, as  $\mathbf{z}$  can be a vector containing many random variables. We will consider variational bounds for more structured graphical models later.



# Importance sampling

In Q34 we saw a connection between the **variance of importance sampling** and the **proposal being the posterior**.

$$I = \int p(\mathbf{x} | \mathbf{z}) p(\mathbf{z}) d\mathbf{z} \quad (1)$$

$$\hat{I} = \frac{1}{S} \sum_{s=1}^S \frac{p(\mathbf{x} | \mathbf{z}^{[s]}) p(\mathbf{z}^{[s]})}{q(\mathbf{z}^{[s]})}, \quad \mathbf{z}^{[s]} \sim q(\mathbf{z}). \quad (2)$$

$$\mathbb{V}_{q(\mathbf{z})}[\hat{I}] = 0 \quad \text{iff} \quad q(\mathbf{z}) = p(\mathbf{z} | \mathbf{x}) = \frac{p(\mathbf{x} | \mathbf{z}) p(\mathbf{z})}{p(\mathbf{x})} \quad (3)$$

# Importance sampling

Importance sampling gave an **unbiased** approximation of the marginal likelihood.

- ▶ View  $q(\mathbf{z})$  as an approximation to  $p(\mathbf{z} | \mathbf{x})$
- ▶ Estimator variance is a measure of quality of  $q(\mathbf{z}) \approx p(\mathbf{z} | \mathbf{x})$

By comparing the variance of approximations we could compare different  $q(\mathbf{z})$  as approximations to  $p(\mathbf{z} | \mathbf{x})$ .

How to compare approximations?

- ▶ Draw many samples
- ▶ Estimate variance using samples

Problem: High variance makes it hard to compare

# Lower bounds

Instead of **unbiased** estimates where we try to **minimise the variance**, we can have a **biased** estimate, where we try to **minimise the bias**.

Lower bound

$$\log p(\mathbf{x}) \geq \mathcal{L}(q(\mathbf{z})) \quad (4)$$

Wishlist of properties:

- ▶ The posterior recovers the marginal likelihood  $\mathcal{L}(p(\mathbf{z} | \mathbf{x})) = p(\mathbf{y})$
- ▶ Continuous in  $q(\mathbf{z})$
- ▶ Easily computable estimate

Procedure: **Adjust  $q(\mathbf{z})$  to maximise  $\mathcal{L}$ .**

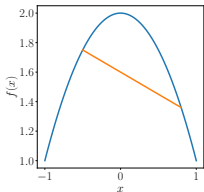
# Jensen's Inequality

An important result from convex analysis:

## Jensen's Inequality

For concave functions  $f$ :

$$f(\mathbb{E}[z]) \geq \mathbb{E}[f(z)]$$



Logarithms are concave. Therefore:

$$\log \mathbb{E}[g(z)] = \log \int g(z)p(z)dz \geq \int p(z) \log g(z)dz = \mathbb{E}[\log g(z)]$$

Idea: For estimating the log marginal likelihood, use Jensen's inequality instead of Monte Carlo.

# Deriving the Variational Lower Bound

Look at log-marginal likelihood (log-evidence):

$$\begin{aligned}\log p(\mathbf{x}) &= \log \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z} \\ &= \log \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})\frac{q(\mathbf{z})}{q(\mathbf{z})}d\mathbf{z} \\ &= \log \int p(\mathbf{x}|\mathbf{z})\frac{p(\mathbf{z})}{q(\mathbf{z})}q(\mathbf{z})d\mathbf{z} \\ &= \log \mathbb{E}_q \left[ p(\mathbf{x}|\mathbf{z})\frac{p(\mathbf{z})}{q(\mathbf{z})} \right] \\ &\geq \mathbb{E}_q \log \left( p(\mathbf{x}|\mathbf{z})\frac{p(\mathbf{z})}{q(\mathbf{z})} \right) \\ &= \mathbb{E}_q [\log p(\mathbf{x}|\mathbf{z})] - \mathbb{E}_q \left[ \log \left( \frac{q(\mathbf{z})}{p(\mathbf{z})} \right) \right] \\ &= \mathbb{E}_q [\log p(\mathbf{x}|\mathbf{z})] - \text{KL}[q(\mathbf{z})||p(\mathbf{z})]\end{aligned}$$

# What have we gained?

Marginal likelihood bound<sup>1</sup>:

$$\mathcal{L}(q) = \mathbb{E}_q[\log p(\mathbf{x}|\mathbf{z})] - \text{KL}[q(\mathbf{z})||p(\mathbf{z})] \quad (5)$$

- ▶ Objective function that can be optimised to find  $q(\mathbf{z})$ 
  - ▶ Terms only include prior and likelihood (can evaluate)
  - ▶ Often, integrals **can** be found in closed form!
- ▶ Bound allows us to compare approximations! Higher is better.
  - ▶ Compare to importance sampling: Two estimates with unknown variances. Don't know which one to believe!

With parameterised  $q_{\mathbf{v}}(\mathbf{z})$ , use gradient-based optimisation to find  $\mathbf{v}$ .

---

<sup>1</sup>Also called **negative variational free energy**, or **Evidence Lower BOund** (ELBO).

A different derivation:

# Minimising the KL

# What is the measure of similarity?

- ▶ So far, the justification for VI came from that if  $q(\mathbf{z}) = p(\mathbf{z} | \mathbf{x})$ , then  $\mathcal{L} = \log p(\mathbf{x})$ .
- ▶ Measure of similarity to  $p(\mathbf{z} | \mathbf{x})$  was defined simply as “how good a bound” does the  $q(\mathbf{z})$  give.

**Can we understand more about the measure of similarity?**



# What is the measure of similarity?

We can find an equation for the measure of similarity by investigating the difference between  $\mathcal{L}$  and  $\log p(\mathbf{x})$ :

$$\begin{aligned}\log p(\mathbf{x}) - \mathcal{L} &= \log p(\mathbf{x}) - \int q(\mathbf{z}) \log \frac{p(\mathbf{x} | \mathbf{z}) p(\mathbf{z})}{q(\mathbf{z})} d\mathbf{z} \\ &= \int q(\mathbf{z}) \log p(\mathbf{x}) d\mathbf{z} - \int q(\mathbf{z}) \log \frac{p(\mathbf{x} | \mathbf{z}) p(\mathbf{z})}{q(\mathbf{z})} d\mathbf{z} \\ &= \int q(\mathbf{z}) \log \frac{p(\mathbf{x}) q(\mathbf{z})}{p(\mathbf{x} | \mathbf{z}) p(\mathbf{z})} d\mathbf{z} \\ &= \int q(\mathbf{z}) \log \frac{q(\mathbf{z})}{p(\mathbf{z} | \mathbf{x})} d\mathbf{z} \\ &= \text{KL}[q(\mathbf{z}) || p(\mathbf{z} | \mathbf{x})]\end{aligned}$$

**VI minimises the KL from the true posterior!**

# Properties of Variational Inference

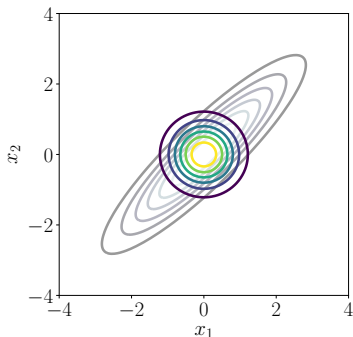
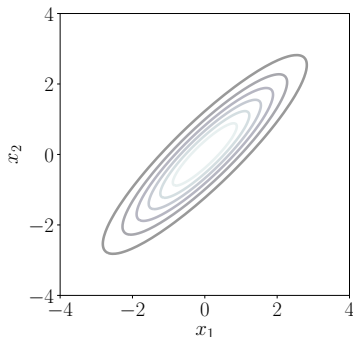
# Properties of the KL divergence

The KL divergence is a **measure of difference** between probability distributions.

$$\text{KL} = \text{KL}[q(\mathbf{z})||p(\mathbf{z})] = \int q(\mathbf{z}) \log \frac{q(\mathbf{z})}{p(\mathbf{z})} d\mathbf{z} \quad (6)$$

- ▶  $\text{KL} \geq 0$
- ▶  $\text{KL} = 0$  iff  $q(\mathbf{z}) = p(\mathbf{z})$
- ▶ Related to information theory and code lengths
- ▶ Related to decision theory and betting returns
- ▶ Intuitively:
  - ▶ Strong penalty for  $q(\mathbf{z})$  for placing mass where  $p(\mathbf{z})$  doesn't
  - ▶ Weak penalty for  $q(\mathbf{z})$  for placing too much mass compared to  $p(\mathbf{z})$

## Example: Gaussian KL divergence



$$\text{KL}[\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \parallel \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)] = \frac{1}{2} \left[ \text{Tr}(\boldsymbol{\Sigma}_1^{-1} \boldsymbol{\Sigma}_0) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) - D + \log \frac{\det \boldsymbol{\Sigma}_1}{\det \boldsymbol{\Sigma}_0} \right]$$

$$\blacktriangleright \boldsymbol{\Sigma}_0 \rightarrow \mathbf{0} \quad \implies \quad \text{KL} \rightarrow \infty$$

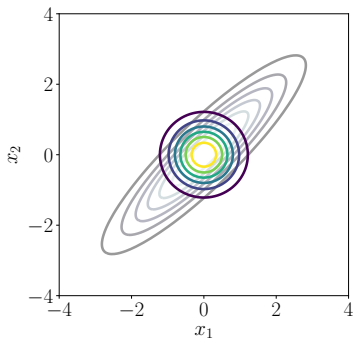
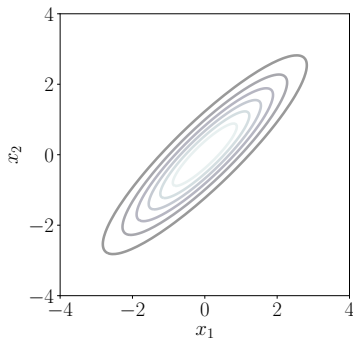
# Approximating Distributions



## Trade-off

- ▶ More expressive gets closer to the true posterior
- ▶ Less expressive is easier to handle
- ▶ Expressive distributions may not allow integrals in ELBO to be computed

# Mean-Field Approximation: Limitation



- ▶ Mean-field VI to approximate a correlated Gaussian with a factorized Gaussian
- ▶ Generally, mean-field VI tends to yield an approximation that is **too compact** ➡ Need better classes of posterior approximations

# Interpretation of terms

$$\log p(\mathbf{x}) \geq \mathbb{E}_q[\log p(\mathbf{x}|\mathbf{z})] - \text{KL}[q(\mathbf{z})||p(\mathbf{z})] =: \text{ELBO}$$

- ▶ **Data-fit term** (expected log-likelihood): Measures how well samples from  $q(\mathbf{z})$  explain the data (“reconstruction cost”).
  - ▶▶ Place  $q$ ’s mass on the MAP estimate.
- ▶ **Regularizer**: Variational posterior  $q(\mathbf{z})$  should not differ much from the prior  $p(\mathbf{z})$

# Alternative form of ELBO

$$\begin{aligned}\mathcal{L}(q_v) &= \int q_v(\mathbf{z}) \log p(\mathbf{x} | \mathbf{z}) d\mathbf{z} && - \underbrace{\int q_v(\mathbf{z}) \log \frac{q_v(\mathbf{z})}{p(\mathbf{z})} d\mathbf{z}}_{\text{KL}} \\ &= \int q_v(\mathbf{z}) \log p(\mathbf{x} | \mathbf{z}) p(\mathbf{z}) d\mathbf{z} && - \int q_v(\mathbf{z}) \log q_v(\mathbf{z}) d\mathbf{z} \\ &= \int q_v(\mathbf{z}) \log p(\mathbf{x} | \mathbf{z}) p(\mathbf{z}) d\mathbf{z} && + \mathcal{H}(q_v(\mathbf{z}))\end{aligned}$$



# Comparison to MAP

$$\mathcal{L}(q_{\mathbf{v}}) = \int q_{\mathbf{v}}(\mathbf{z}) \log p(\mathbf{x} | \mathbf{z}) p(\mathbf{z}) d\mathbf{z} + \mathcal{H}(q_{\mathbf{v}}(\mathbf{z})) \quad (7)$$

$$L_{\text{MAP}}(\mathbf{z}) = \log p(\mathbf{x} | \mathbf{z}) + \log p(\mathbf{z}) \quad (8)$$

- ▶ Fit the data like MAP
- ▶ but also be as **uncertain** as possible (entropy)

# Properties of the differential entropy

$$\mathcal{H}[q(\mathbf{z})] = - \int q(\mathbf{z}) \log q(\mathbf{z}) d\mathbf{z} \quad (9)$$

- ▶ Generalises entropy to continuous variables
- ▶ Limit of: Entropy of quantised  $q(\mathbf{z})$  minus uniform distribution
- ▶ Can be negative! (i.e. more certain than a uniform)

# Summary

- ▶ Variational turns inference into optimisation
- ▶ Two ways to derive:
  - ▶ We minimise the KL divergence to the posterior
  - ▶ Lower bound marginal likelihood with Jensen's inequality
- ▶ Constrained approximation families (e.g. mean-field) tend to underestimate uncertainty

Next time:

- ▶ How to compute ELBOs
- ▶ How to optimise ELBOs

# References I

- [1] C. M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer-Verlag, 2006.