Variational Inference

Presented by: Weilin Wu (8140344) Huizhen Zhang (8858346)

Variational Inference

- What for?

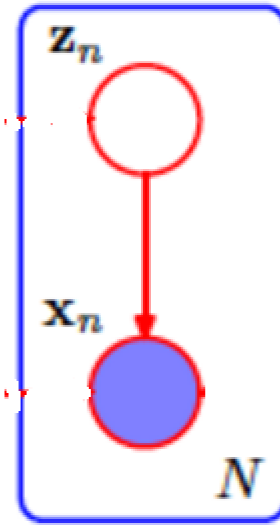
Use easy-to-compute probability distributions to approximate difficult-to-compute probability distributions.

Variational inference is widely used to approximate posterior distributions for Bayesian models.

Outline

- General problem and core idea
  - Kullback-Leibler divergence
- Approximate inference
  - ELBO - Evidence lower bound
  - Variational family - Mean-field approximation
- Optimization algorithm (CAVI)
- An example
- Summary
- References

General Problem



$$p(\mathbf{z}, \mathbf{x}) = p(\mathbf{z})\, p(\mathbf{x} \mid \mathbf{z})$$

$\mathbf{x}$: a set of observed variables
$\mathbf{z}$: a set of latent(hidden) variables

Postierior distribution:

$$p(\mathbf{z} \mid \mathbf{x}) = \frac{p(\mathbf{z}, \mathbf{x})}{p(\mathbf{x})}$$

Known

Difficult to compute

uOttawa

$$p(\mathbf{z} \mid \mathbf{x}) = \frac{p(\mathbf{z}, \mathbf{x})}{p(\mathbf{x})}$$

The marginal distribution of the observations, also called the ***evidence:***

$$p(\mathbf{x}) = \int_{\mathbf{z}} p(\mathbf{z}, \mathbf{x})$$

Difficult to compute:

1. this evidence integral is unavailable in closed form
2. requires exponential time to compute.

Variational Inference – core idea

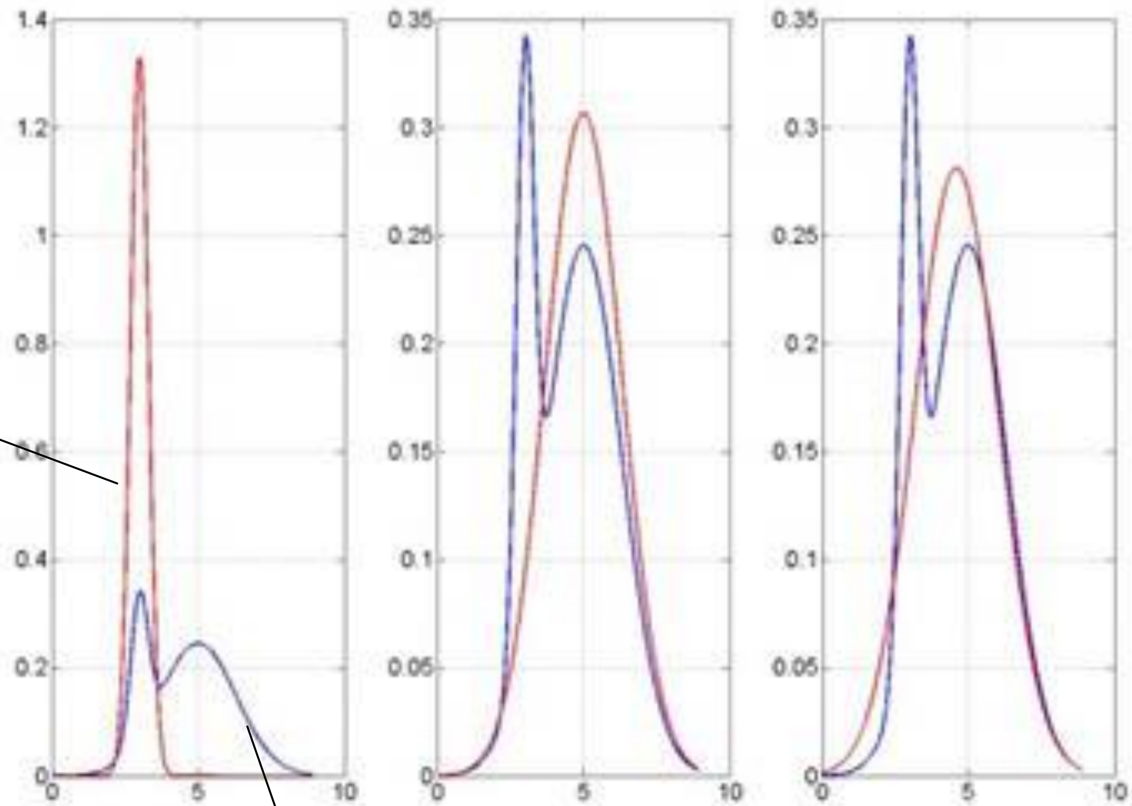
We restrict ourselves a family of approximate distributions D over the latent variables.

We then try to find the member of that family that minimizes the Kullback-Leibler divergence to the exact posterior.

$$q^*(\mathbf{z}) = \underset{q(\mathbf{z}) \in \mathrm{D}}{\arg\min}\, KL\left(q(\mathbf{z}) \,||\, p(\mathbf{z} \mid \mathbf{x})\right)$$

This reduces to solving an **optimization problem**.

$q(\mathbf{z})$



$p(\mathbf{z} \mid \mathbf{x})$

Kullback-Leibler divergence (KL divergence)

- an information-theoretic measure of proximity between two distributions.
- It is minimized to be 0 when two distributions are the same.

$$KL\,(q(\mathbf{z}) \,\|\, p(\mathbf{z} \mid \mathbf{x})) = E\,[\log q(\mathbf{z})] - E\,[\log p(\mathbf{z} \mid \mathbf{x})]$$

$$= E\,[\log q(\mathbf{z})] - E\,[\log p(\mathbf{z}, \mathbf{x})] + \log \boxed{p(\mathbf{x})}$$

Difficult to compute

* *Note: all the expectation in our presentation is taken with respect to q(z) unless otherwise specified.*

ELBO helps to get around this!

uOttawa

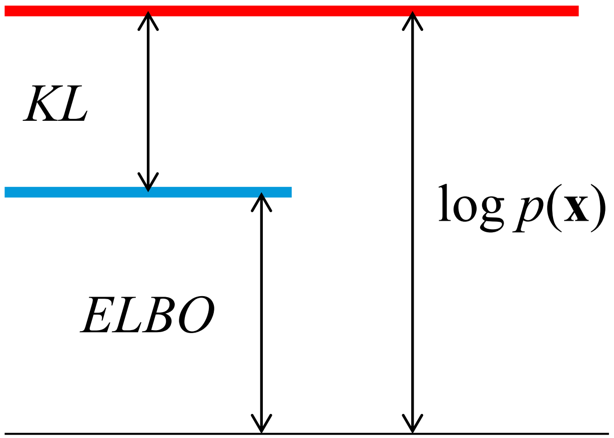
Outline

- General problem and core idea
  - Kullback-Leibler divergence
- Approximate inference
  - ELBO - Evidence lower bound
  - Mean-field variational family
- Optimization algorithm (CAVI)
- An example
- Summary
- References

ELBO - Evidence lower bound

$$KL + ELBO = \log p(\mathbf{x})$$

So minimizing KL is equivalent to maximizing ELBO



The objective function change to be ELBO:

$$q^*(\mathbf{z}) = \underset{q(\mathbf{z}) \in \mathrm{D}}{\arg\min}\, KL\,(q(\mathbf{z}) \,||\, p(\mathbf{z} \mid \mathbf{x}))$$

$$= \underset{q(\mathbf{z}) \in \mathrm{D}}{\arg\max}\, ELBO\,(q(\mathbf{z}))$$

ELBO - Evidence lower bound

Objective function:

$$ELBO\ (q(\mathbf{z})) = \log p(\mathbf{x}) - KL$$

$$= E\ [\log p(\mathbf{z}, \mathbf{x})] - E\ [\log q(\mathbf{z})]$$

Mean-field variational family

$$q(\mathbf{z}) \longrightarrow p(\mathbf{z} \mid \mathbf{x})$$

When we are picking the $q(\mathbf{z})$:

1. the complexity of the family of distributions from which we pick our approximate distribution determines the complexity of the optimization.

2. The more flexibility in the family of distributions, the closer the approximation and the harder the optimization.

Mean-field variational family

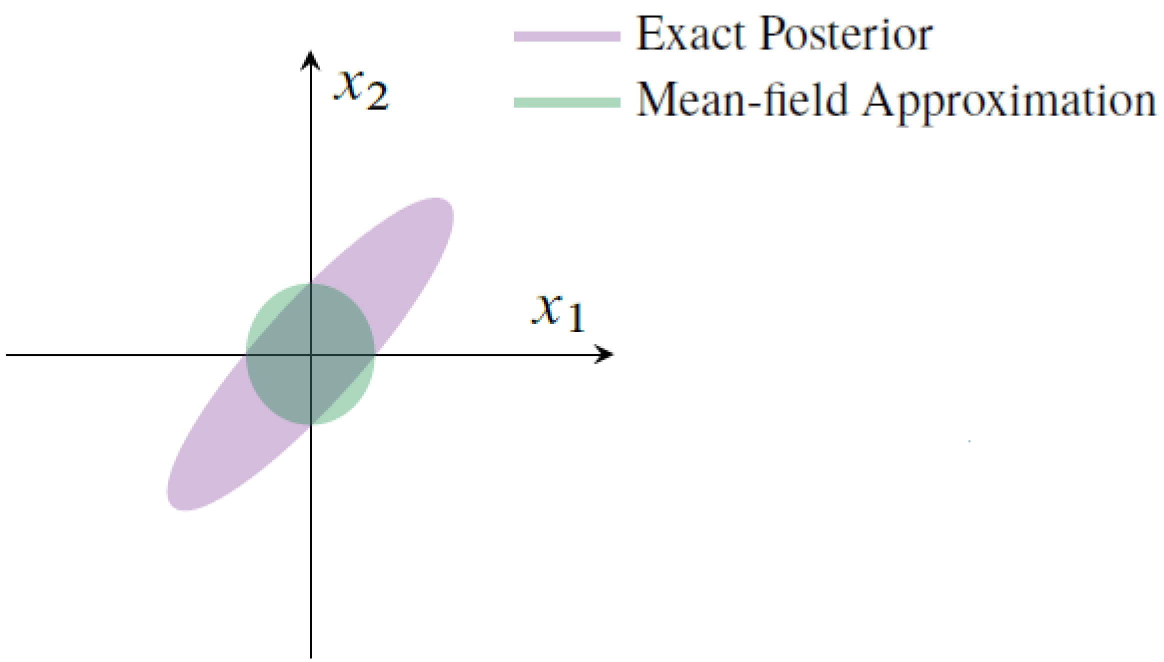
The Mean-field variational family is used in variatoinal inference:

$$q(z) = \prod_{j=1}^{m} q_j(z_j)$$

Properties:

1. the latent variables are mutually independent
2. each governed by a distinct factor in the variational distribution
3. Each can take on any paramteric form corresponding to the latent variable. For example, a continuous variable might have a Gaussian factor whereas a categorical variable will typically have a categorical factor.

An example where a 2D Gaussian Posterior is approximated by a mean-field variational structure with independent Gaussians in the 2 dimensions. The correlation could not be captured.



Outline

- General problem and core idea
  - Kullback-Leibler divergence
- Approximate inference
  - ELBO - Evidence lower bound
  - Mean-field variational family
- Optimization algorithm (CAVI)
- An example
- Summary
- References

Coordinate ascent variational inference (CAVI)

Algorithm 1: Coordinate ascent variational inference (CAVI)

Input: A model $p(\mathbf{x}, \mathbf{z})$, a data set $\mathbf{x}$

Output: A variational distribution $q(\mathbf{z}) = \prod_{j=1}^{m} q_j(z_j)$

Initialize: Variational factors $q_j(z_j)$

while *the* ELBO *has not converged* **do**

  for $j \in \{1, \ldots, m\}$ **do**

    Set $q_j(z_j) \propto \exp\{\mathbb{E}_{-j}[\log p(z_j \mid \mathbf{z}_{-j}, \mathbf{x})]\}$

  end

  Compute $\text{ELBO}(q) = \mathbb{E}[\log p(\mathbf{z}, \mathbf{x})] + \mathbb{E}[\log q(\mathbf{z})]$

end

return $q(\mathbf{z})$

Coordinate ascent variational inference (CAVI)

How to understand the fator update in each iteration?

$$q_j(z_j) \propto \exp\{E_{-j}[\log p(z_j \mid \mathbf{z}_{-j}, \mathbf{x})]\}$$

$$\Rightarrow q_j(z_j) \propto \exp\{E_{-j}[\log p(z_j, \mathbf{z}_{-j}, \mathbf{x})]\}$$

Note: $p(\mathrm{z}_j \mid \mathbf{z}_{-j}, \mathbf{x})$ is called complete conditional of z_j

Coordinate ascent variational inference (CAVI)

That is derived from ELBO:

$$ELBO\,(q(\mathbf{z})) = E\,[\log p(\mathbf{z}, \mathbf{x})] - E\,[\log q(\mathbf{z})]$$

$$\Rightarrow ELBO\,(q(z_j)) = E_j\,[E_{-j}\,[\log p(z_j\,, \mathbf{z}_{-j}\,, \mathbf{x})]] - E_j\,[\log q_j\,(z_j)]] + \text{constant}$$

$$= E_j\,[\mathrm{A}] - E_j\,[\log q_j\,(z_j)]] + \text{constant}$$

$$= -\,KL\,(q_j\,(z_j)\,||\,\mathrm{A}) + \text{constant}$$

Where A = $E_{-j}\,[\log p(z_j\,, \mathbf{z}_{-j}\,, \mathbf{x})]$.

ELBO is maximized when KL = 0. That is

$$\log q_j\,(z_j) = \mathrm{A} = E_{-j}\,[\log p(z_j\,, \mathbf{z}_{-j}\,, \mathbf{x})]$$

Coordinate ascent variational inference (CAVI)

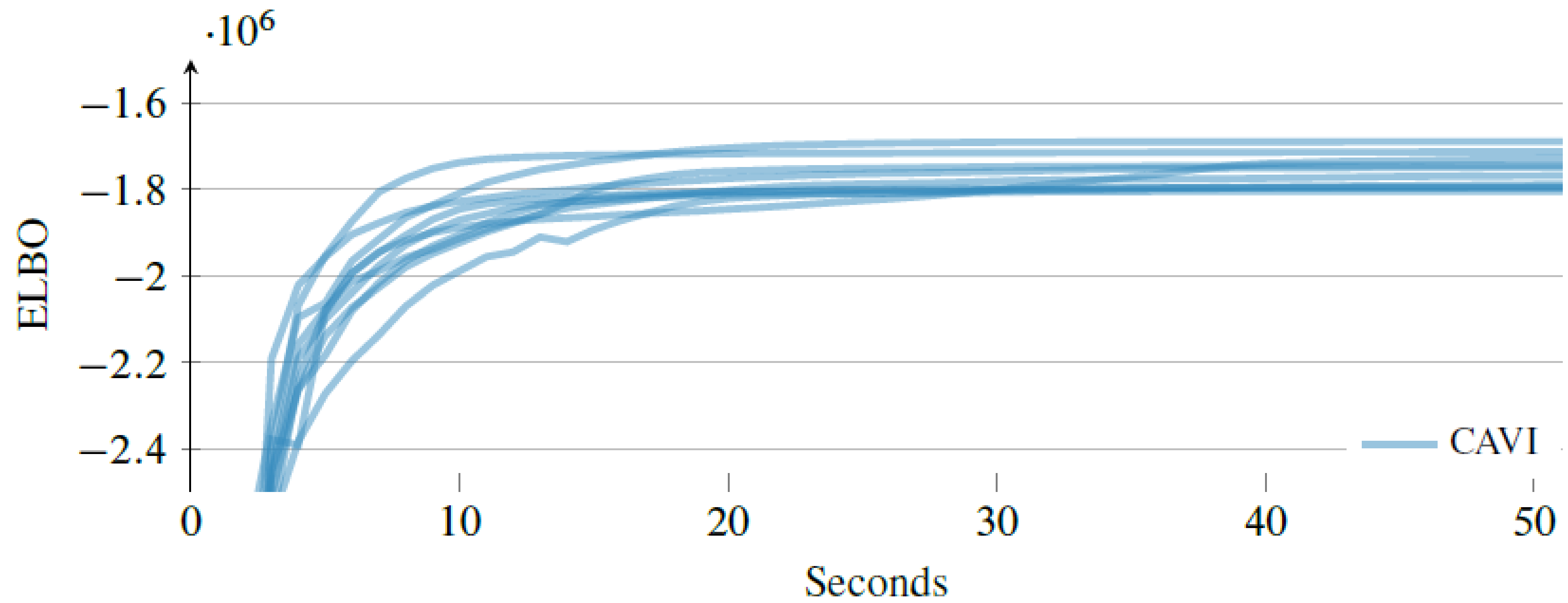
$$\boxed{q_j(z_j) = \exp\{E_{-j}[\log p(\mathbf{z}, \mathbf{x})]\}}$$

Normalize each factor:

$$q_j(z_j) \leftarrow \exp\{E_{-j}[\log p(\mathbf{z}, \mathbf{x})]\} / \int \exp\{E_{-j}[\log p(\mathbf{z}, \mathbf{x})]\}\, \mathrm{d}z_j$$

Coordinate ascent variational inference (CAVI)

Generally the ELBO is a non-convex objective function, CAVIonly guarantees convergence to a local optimum, which can be sensitive to initialization.

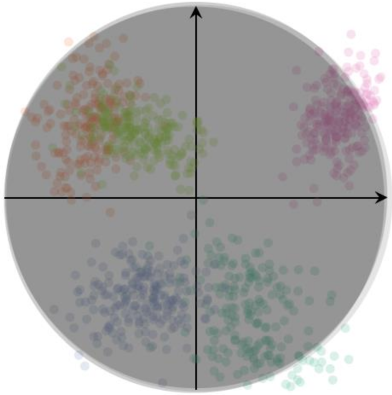


Outline

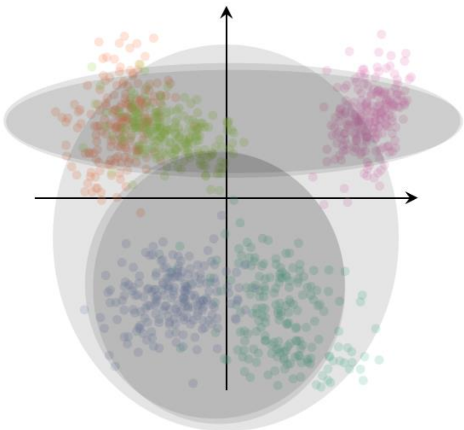
- General problem and core idea
  - Kullback-Leibler divergence
- Approximate inference
  - ELBO - Evidence lower bound
  - Mean-field variational family
- Optimization algorithm (CAVI)
- An example
- Summary
- References

Another example:

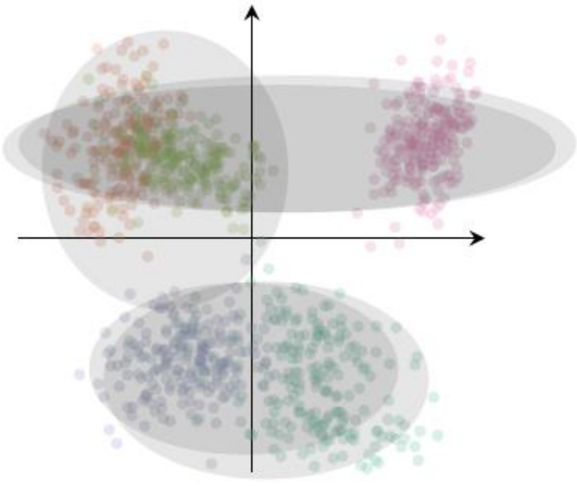
Mixture of 2-D Gaussians



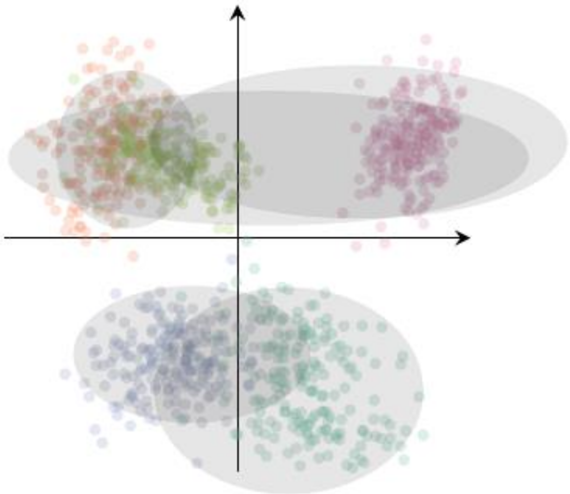
Initialization



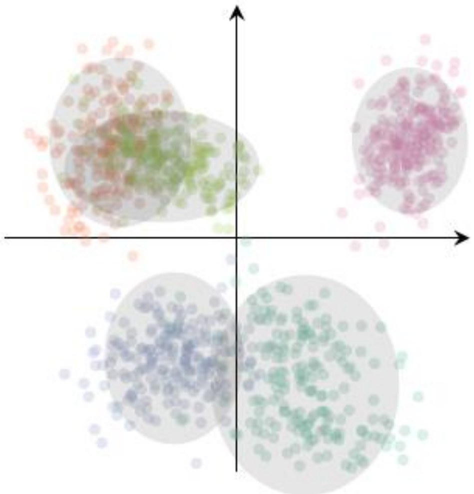
Iteration 20



Iteration 28



Iteration 35



Iteration 50

Quick Comparison: EM vs VI vs MCMC

EM:
- more accurate
- applied when integration of posterior distribution is computable

VI:
- applied when integration of posterior distribution is not computable
- compared to MCMC, much faster and easier to scale to large data set
- compared to MCMC, more manual work required

Markov chain Monte Carlo (MCMC) sampling:
- applied when integration of posterior distribution is not computable
- compared to VI, slower (more computationally expensive)
- compared to VI, less manual work

Summary

- **Variational inference** is an approach to approximate difficult-to-compute probability distributions.

- Key concepts:

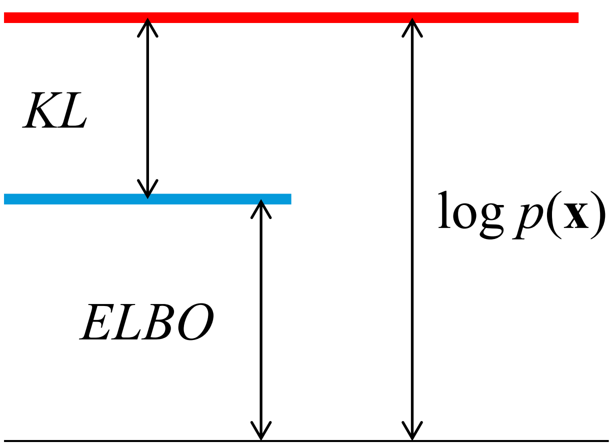
$$q^*(z) = \underset{q(z)\in D}{\arg\min} \quad KL(q(z)||p(z|x))$$

KL divergence

ELBO

Mean-field family

CAVI



An algorithm to optimize ELBO

$$q(z) = \prod_{j=1}^{m} q_j(z_j)$$

References

[1] David M. Blei, Alp Kucukelbir, and Jon D. McAulie. Variational inference: A review for statisticians. CoRR, abs/1601.00670, 2016.

[2] Variational Inference https://am207.github.io/2017/wiki/VI.html.

[3] PRML summary: Approximate Inference http://blog.csdn.net/nietzsche2015/article/details/43450853

[4] Notes of Variational Inference http://blog.csdn.net/guolinsen123/article/details/53241346?locationNum=2&fps=1