11 - Approximation techniques

Bayesian Statistics

Spring 2022-2023

Josep Fortiana

Matemàtiques - Informàtica UB

Monday, May 22, 2023

11 - Approx. pdf's

Laplace approximation (uni- & multivariate)

Variational Bayes Inference

KL divergence and the ELBO

Candidate functions: Mean field VB (MFVB)

Computational VB

Approximate Bayesian inference

Target –posterior, post. predictive, . . . pdf,

is difficult to compute (that is, always).

Two paths:

- **Approximate:** Laplace, variational Bayes, EM, expectation propagation.
- **Simulate:** Independent or MC Monte Carlo.

Laplace approximation

Target posterior pdf $h(\theta|x) \approx$ A gaussian (normal) pdf.

Taylor expansion of $h(\theta|x)$ up to second order.

Variational Bayesian inference

A family of pdf's.

Find the pdf in this family closest to the target.

Closeness: Kullback-Leibler divergence.

11 - Approx. pdf's

Laplace approximation (uni- & multivariate)

Variational Bayes Inference

KL divergence and the ELBO

Candidate functions: Mean field VB (MFVB)

Computational VB

Univariate Laplace approximation

Target function $h(\theta)$, e.g. a pdf, with a maximum.

$$q(\theta) \stackrel{\text{def}}{=\!=} \log h(\theta).$$

Taylor expansion of $q(\theta)$ around a (global) maximum:

$$\theta_0 = \arg\max_\theta q(\theta).$$

For a posterior pdf, θ_0 is the MAP estimator.

Taylor expansion

$$q(\theta) = q(\theta_0) + (\theta - \theta_0) \cdot q'(\theta_0) + \tfrac{1}{2}(\theta - \theta_0)^2 \cdot q''(\theta_0) + \cdots$$

$$\approx q(\theta_0) - \tfrac{1}{2}(\theta - \theta_0)^2 \cdot |q''(\theta_0)|,$$

since $q'(\theta_0) = 0$ and $q''(\theta_0) < 0$ at the maximum.

Taylor expansion

$$h(\theta) = \exp(q(\theta)) \approx A \cdot \exp\left\{ -\frac{1}{2} \left(\frac{\theta - \theta_0}{\sigma} \right)^2 \right\},$$

Gaussian pdf, Normal(θ_0, σ^2), with $\sigma^2 = \dfrac{1}{|q''(\theta_0)|}$.

Multivariate Laplace approximation

Target $h(\boldsymbol{\theta})$, $\boldsymbol{\theta}$ p-dimensional,

$$q(\boldsymbol{\theta}) \stackrel{\text{def}}{=\!=} \log h(\boldsymbol{\theta}).$$

Taylor expansion of $q(\boldsymbol{\theta})$ around a (global) maximum:

$$\boldsymbol{\theta}_0 = \arg\max_{\boldsymbol{\theta}} q(\boldsymbol{\theta}).$$

When $h(\,\cdot\,)$ is a posterior pdf, $\boldsymbol{\theta}_0$ is the MAP estimator.

Multivariate Taylor expansion

$$q(\boldsymbol{\theta}) = q(\boldsymbol{\theta}_0) + \dot{q}(\boldsymbol{\theta}_0) \cdot (\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \tfrac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)' \cdot \ddot{q}(\boldsymbol{\theta}_0) \cdot (\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \cdots$$

$$\approx q(\boldsymbol{\theta}_0) - \tfrac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)' \cdot (-\ddot{q}(\boldsymbol{\theta}_0)) \cdot (\boldsymbol{\theta} - \boldsymbol{\theta}_0),$$

since the gradient $\dot{q}(\boldsymbol{\theta}_0) = 0$ and the Hessian $\ddot{q}(\boldsymbol{\theta}_0)$ is negative definite at the maximum.

Multivariate Taylor expansion - Notations

$\boldsymbol{\theta}$, etc., are $p \times 1$ column vectors.

Derivatives represented with dots
(prime to indicate matrix transposition).

Gradient $\dot{q}(\boldsymbol{\theta}_0)$ is a $1 \times p$ row vector,

Hessian $\ddot{q}(\boldsymbol{\theta}_0)$ is a $p \times p$ symmetric (neg. def.) matrix.

Approximate pdf

$$h(\boldsymbol{\theta}) = \exp(q(\boldsymbol{\theta}))$$

$$\approx A \cdot \exp\left\{-\tfrac{1}{2}\left(\boldsymbol{\theta} - \boldsymbol{\theta}_0\right)' \cdot \boldsymbol{\Sigma}^{-1} \cdot \left(\boldsymbol{\theta} - \boldsymbol{\theta}_0\right)\right\},$$

a Gaussian pdf,

Normal$(\boldsymbol{\theta}_0, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma} = -\ddot{q}(\boldsymbol{\theta}_0)^{-1}$.

The R-INLA project

Integrated Nested Laplace Approximations:

The R-INLA project website

```
install.packages("INLA",dependencies=TRUE,
repos="https://inla.r-inla-download.org/R/stable")
```

11 - Approx. pdf's

Laplace approximation (uni- & multivariate)

Variational Bayes Inference

KL divergence and the ELBO

Candidate functions: Mean field VB (MFVB)

Computational VB

What is Variational?

Calculus of variations:

Find a real function $g(s)$ in some set $\mathscr{G}$ of functions such that:

$$\mathbb{F}[g] = \int_a^b F(g(s), s)\, ds \qquad \text{is min (or max.)},$$

where $F(\,\cdot\,,\,\cdot\,)$ is a given function.

Maximum entropy pdf's

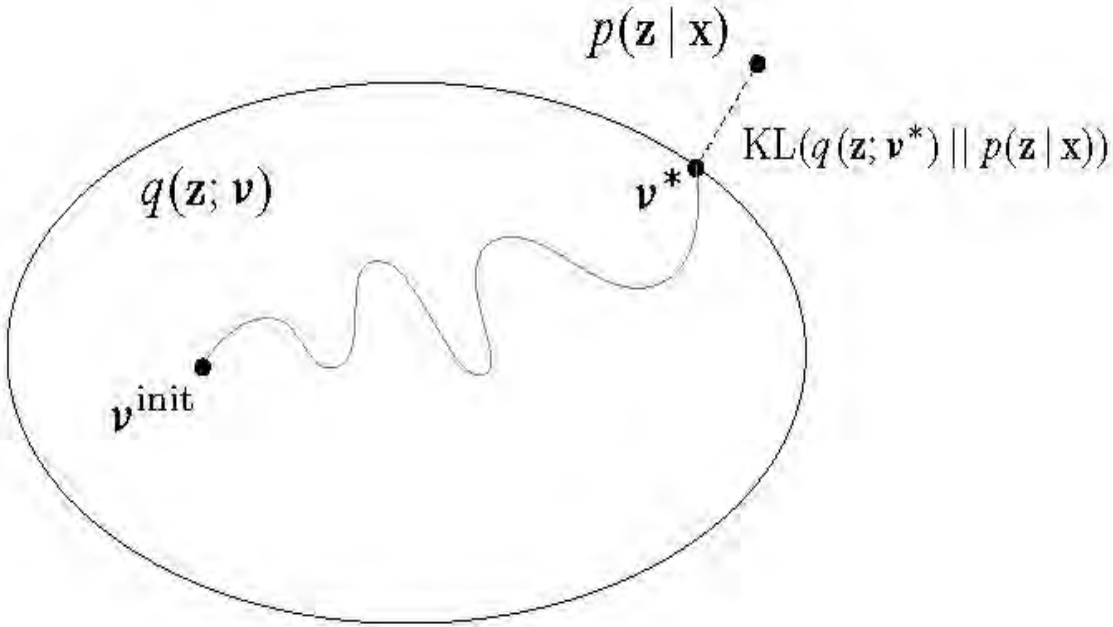
Example: pdf $f(s \mid \mu, \sigma^2)$ with mean μ and variance σ^2, maximizing:

$$\mathbb{H}[f(s \mid \mu, \sigma^2)] = -\int_a^b f(s \mid \mu, \sigma^2) \cdot \log f(s \mid \mu, \sigma^2)\, ds.$$

Solution with support on $\mathbb{R}$ is Normal(μ, σ^2).

Jaynes (2003), *Probability theory: the logic of science*.

What is VB?



[David Blei's 2011 lecture]

What is VB?

Target: $p(z \mid x)$ *(Posterior pdf: z = parameters)*.

A family of simple pdf's: $\{q(z;\nu)\}$

Find the closest q to the target p *(This is "variational")*.

"Distance" is $\mathrm{KL}(\cdot \,\|\, \cdot)$. *Why?*

Why VB?

- Acceptable approximation for posterior means and sd.

Why VB?

- Acceptable approximation for posterior means and sd.
- Quick for large data & many parameters:
  MCMC iterations take forever.

Why VB?

- Acceptable approximation for posterior means and sd.
- Quick for large data & many parameters:
  MCMC iterations take forever.

Microcredit Experiment (Rachael Meager, 2019),

Tamara Broderick, *Variational Bayes and Beyond: Bayesian Inference for Big Data (ICML 2018 tutorial)*

Why VB?

- Acceptable approximation for posterior means and sd.
- Quick for large data & many parameters:
  MCMC iterations take forever.

Microcredit Experiment (Rachael Meager, 2019),

Deep NN, e.g., Variational Autoencoders.

Tamara Broderick, *Variational Bayes and Beyond: Bayesian Inference for Big Data (ICML 2018 tutorial)*

Kingma DP, Welling M (2014), *Auto-Encoding Variational Bayes.*

A large hierarchical model

$K = 7$ microcredit trials

(Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)

N_k businesses in k-th site (900 to 17K)

Profit of n-th business at k-th site $(1 \leq n \leq N_k)$:

$$y_{kn} \sim \text{Normal}(\mu_k + T_{kn} \cdot \tau_k, \sigma_k^2),$$

Rachael Meager home page

11 - Approx. pdf's

Laplace approximation (uni- & multivariate)

Variational Bayes Inference

KL divergence and the ELBO

Candidate functions: Mean field VB (MFVB)

Computational VB

Kullback-Leibler (KL) divergence

$$\boxed{\mathrm{KL}(q \,\|\, p) = \mathrm{E}_q(\log \frac{q}{p})}$$

KL divergence properties

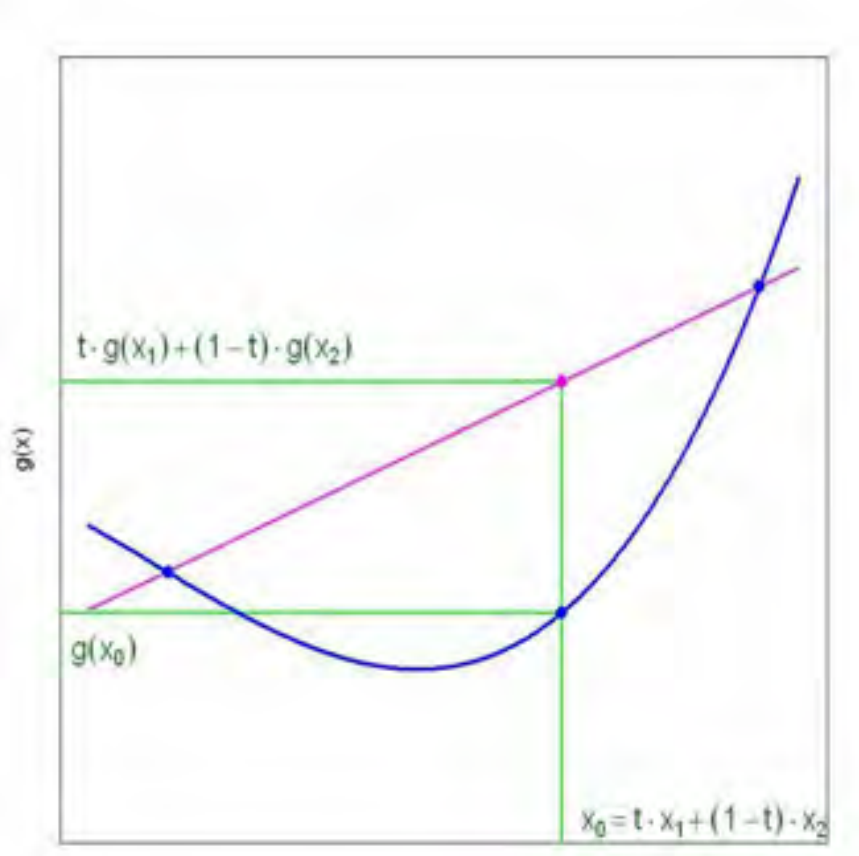
- $\mathrm{KL}(q \,\|\, p) \neq \mathrm{KL}(p \,\|\, q)$. Not a symmetric "distance".

  On regions where:

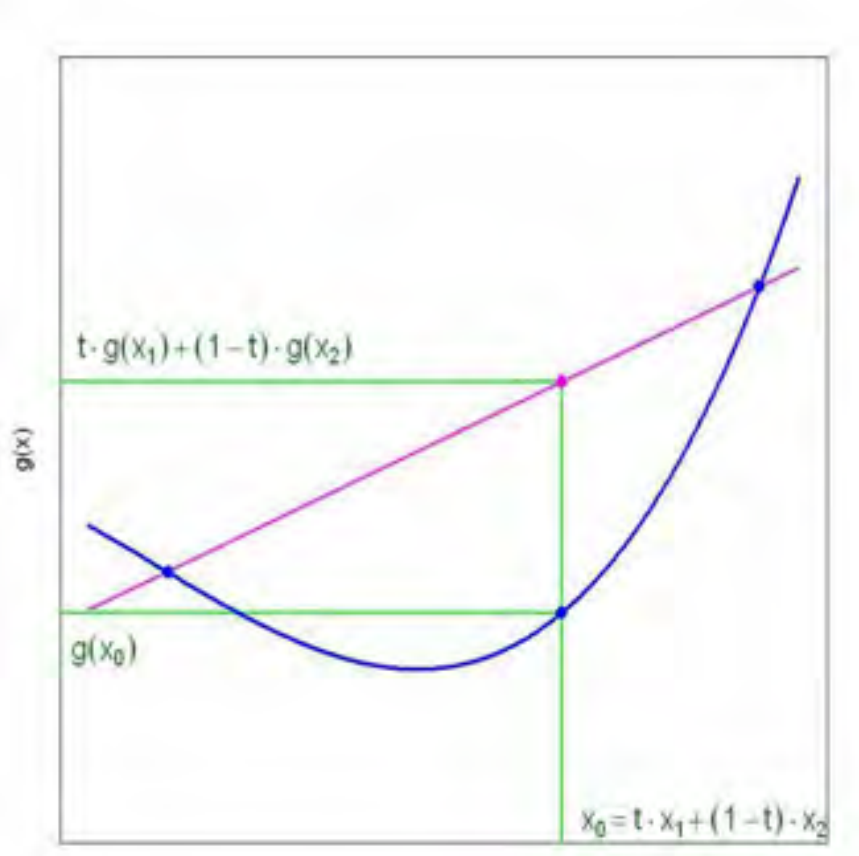
  - Both p and q are large or both small $\Rightarrow$ small contribution.
  - When q is large and p small $\Rightarrow$ big contribution.
  - When p is large and q small $\Rightarrow$ not so important.

- $\mathrm{KL}(q \,\|\, p) \geq 0$.

Convex functions



Convex functions



A *convex function* is "like x^2".

Convexity and Jensen inequality - Summary

φ is *convex* if it seems an upwards parabola. Positive 2nd derivative.

Its graph lies under any secant line:

$$\varphi(t \cdot a + (1-t) \cdot b) \leq t \cdot \varphi(a) + (1-t) \cdot \varphi(b).$$

Jensen's inequality:

$$\left[\varphi \text{ convex}\right] \implies \left[\varphi(\mathrm{E}(Z)) \leq \mathrm{E}(\varphi(Z))\right].$$

$\boxed{-\log}$ is a convex function.

The ELBO

$$p(x) = \int_z p(x,z)\,dz = x\text{-marginal} = \text{“the evidence”}.$$

$$\begin{aligned}
\log p(x) &= \log \int_z p(x,z)\,dz = \log \int_z p(x,z)\,\frac{q(z)}{q(z)}\,dz \\
&= \log\left(\mathrm{E}_q\left[\frac{p(x,Z)}{q(Z)}\right]\right) \\
\text{(Jensen)} \quad &\geq \underbrace{\mathrm{E}_q\left[\log p(x,Z)\right] - \mathrm{E}_q\left[\log q(Z)\right]}_{\text{ELBO}}.
\end{aligned}$$

Evidence Lower BOund.

KL divergence and the ELBO

$$\mathrm{KL}(\, q(z) \,\|\, p(z \,|\, x)\,)$$ Conditional pdf: $p(z \,|\, x) = p(z, x)/p(x)$.

$$\begin{aligned}
&= \mathrm{E}_q \left[\log \frac{q(Z)}{p(Z \,|\, x)} \right] \\
&= \mathrm{E}_q \left[\log q(Z) \right] - \mathrm{E}_q \left[\log p(Z \,|\, x) \right] \\
&= \mathrm{E}_q \left[\log q(Z) \right] - \mathrm{E}_q \left[\log p(Z, x) \right] + \log p(x) \\
&= -\left(\mathrm{E}_q \left[\log p(x, Z) \right] - \mathrm{E}_q \left[\log q(Z) \right] \right) + \log p(x) \geq 0.
\end{aligned}$$

Maximizing the ELBO

Since $\log p(x)$ does not depend on q,

minimizing $\mathrm{KL}(\,\cdot\,\|\,\cdot\,)$ (with respect to the q family)

is equivalent to maximizing the ELBO.

Procedure: choose a family of pdf's $\{q(z,\nu)\}$,
depending on parameters ν.

Then find the ELBO-maximizing ν.

11 - Approx. pdf's

Laplace approximation (uni- & multivariate)

Variational Bayes Inference

KL divergence and the ELBO

Candidate functions: Mean field VB (MFVB)

Computational VB

Choice of candidate functions

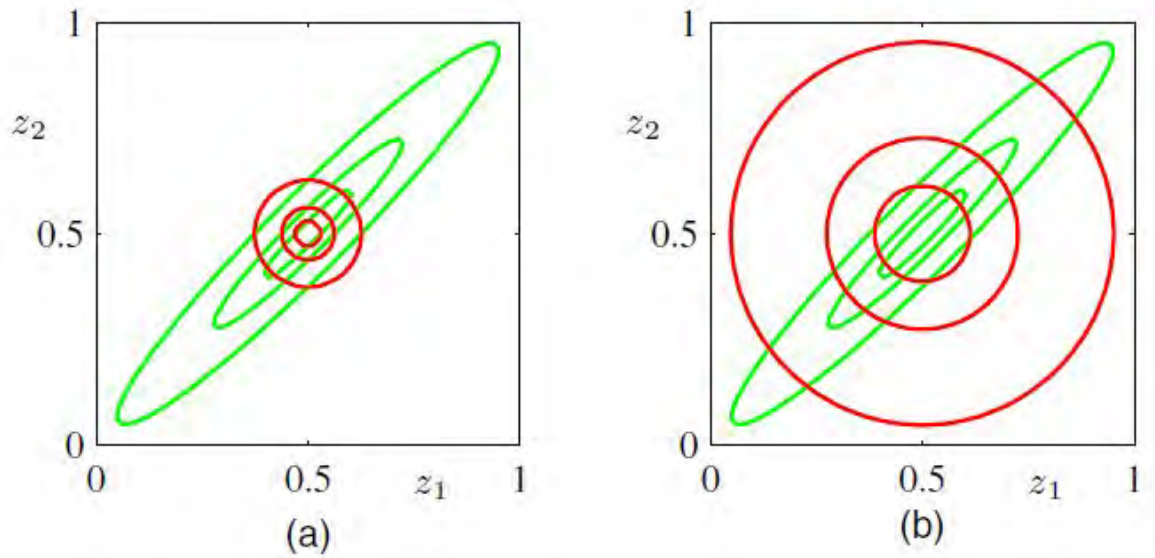
Mean field approximation: from Physics.

Assume the variational family q factorizes:

For $z = (z_1, \ldots, z_m)$

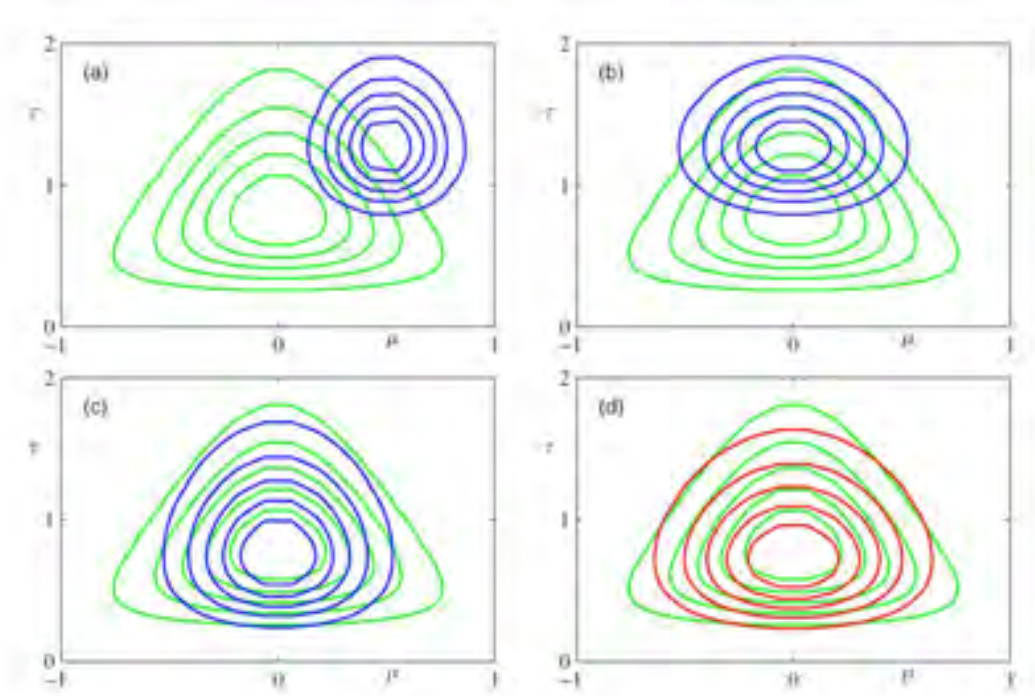
$$q(z) = q(z_1, \ldots, z_m) = \prod_{j=1}^{m} q(z_j).$$

Caution with a product of univariate factors



Bishop C (2006), PRML. Chap. 10, Figure 10.2.

VB for μ and τ (precision) of a 1D Gaussian pdf



Bishop C (2006), PRML. Chap. 10, Figure 10.4.

11 - Approx. pdf's

Laplace approximation (uni- & multivariate)

Variational Bayes Inference

KL divergence and the ELBO

Candidate functions: Mean field VB (MFVB)

Computational VB

Direct optimization

Direct computation of the minimum KL for a particular posterior pdf, for some given likelihood and prior.

Coordinate Ascent Variational Inference (CAVI):

Sequentially optimize each individual parameter, while holding the others fixed.

Stochastic Variational Inference (SVI)

The coordinate ascent algorithm is inefficient for large data sets because we must optimize the local variational parameters for each data point before re-estimating the global variational parameters.

In a nutshell: Stochastic variational inference uses stochastic optimization to fit the global variational parameters.

See Stackexchange: Difference between Stochastic VI and VI?

Hoffman M, Blei D, Wang C, Paisley J (2013),
Stochastic variational inference.

Black box VB inference (BBVI)

David Blei talk at 2018 PROBPROG conference:

Black Box Variational Inference [YouTube]

[Blei's 2011 lecture]

[SLIDES] [2014 TechReport] [and Paper]

Manousakas, D. et al (2022)

"Black-box Coreset Variational Inference"

Idea underlying BBVI

The key insight behind BBVI is that it's possible to write the gradient of the ELBO as an expectation:

$$\nabla_\nu \mathscr{L}(\nu) = \mathsf{E}_q[(\nabla_\nu \log q(z \mid \nu))(\log p(x, z) - \log q(z \mid \nu))].$$

So instead of evaluating a closed form expression for the gradient, we can use Monte Carlo samples and take the average to get a noisy estimate of the gradient.

See: Keyon Vafa (2017 blog entry),

Black Box Variational Inference for Logistic Regression

Automatic differentiation VI (ADVI)

This method is implemented in Stan.

Kucukelbir A, Tran D, Ranganath R, Gelman A, Blei D (2017),
Automatic differentiation variational inference.

Kucukelbir A, Ranganath R, Gelman A, Blei D (2017),
Automatic variational inference in Stan.