Variational Inference

Cédric Archambeau
cedrica@amazon.com

OxWaSP, Machine Learning Module
Oxford, March 2019

Outline

1. Approximate Bayesian inference
2. Variational inference
    - Mean field
    - Relation to Expectation-Maximisation
    - Structured variational inference
3. Stochastic and extensions

Bayesian statistics

$$\underbrace{p(\boldsymbol{\Theta}|\mathbf{X})}_{\textit{posterior}} = \frac{\overbrace{p(\mathbf{X}|\boldsymbol{\Theta})}^{\textit{likelihood}}\,\overbrace{p(\boldsymbol{\Theta})}^{\textit{prior}}}{\underbrace{p(\mathbf{X})}_{\textit{evidence}}}, \qquad p(\mathbf{X}) = \int p(\mathbf{X}, \boldsymbol{\Theta})\; d\boldsymbol{\Theta}.$$

- The likelihood is the noise model.
- The prior encodes constraints (if any) on the parameters $\boldsymbol{\Theta}$.
- Structure is added to the model through latent variables $\mathbf{Z}$: $p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\Theta})$

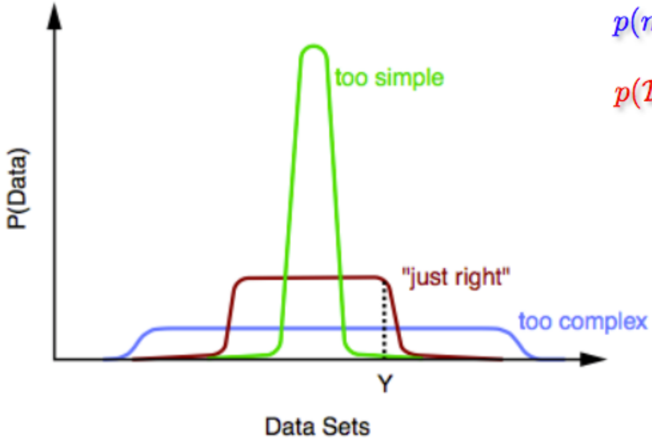
Bayesian statistics

$$\underbrace{p(\boldsymbol{\Theta}|\mathbf{X})}_{\textit{posterior}} = \frac{\overbrace{p(\mathbf{X}|\boldsymbol{\Theta})}^{\textit{likelihood}}\,\overbrace{p(\boldsymbol{\Theta})}^{\textit{prior}}}{\underbrace{p(\mathbf{X})}_{\textit{evidence}}}, \qquad p(\mathbf{X}) = \int p(\mathbf{X}, \boldsymbol{\Theta})\; d\boldsymbol{\Theta}.$$

- The likelihood is the noise model.
- The prior encodes constraints (if any) on the parameters $\boldsymbol{\Theta}$.
- Structure is added to the model through latent variables $\mathbf{Z}$: $p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\Theta})$

- Predictions are averaged over all possible models: $p(\mathbf{x}_*|\mathbf{X}) = \int p(\mathbf{X}_*|\boldsymbol{\Theta})\; p(\boldsymbol{\Theta}|\mathbf{X})\; d\boldsymbol{\Theta}$.
- The goal is to maximise the marginal likelihood or evidence $p(\mathbf{X}|m)$.

What is great about Bayesian inference?



$$p(m|\mathcal{D}) = \frac{p(\mathcal{D}|m)p(m)}{p(\mathcal{D})}$$

$$p(\mathcal{D}|m) = \sum_{\theta} p(\mathcal{D}|\theta, m)p(\theta|m)$$

What is not so great with Bayesian inference?

Posterior inference:

$$p(\boldsymbol{\Theta}|\mathbf{X}, m) \propto \int p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\Theta}|m)\; d\mathbf{Z}.$$

Model averaging:

$$p(\mathbf{x}_*|\mathbf{X}, m) = \int p(\mathbf{x}_*|\boldsymbol{\Theta}, m)\; p(\boldsymbol{\Theta}|\mathbf{X}, m)\; d\boldsymbol{\Theta}.$$

Evidence maximisation:

$$p(\mathbf{X}|m) = \int p(\mathbf{X}, \boldsymbol{\Theta}|m)\; d\boldsymbol{\Theta}.$$

Variational lower bound or evidence lower bound (ELBO)

$$\ln p(\mathbf{X}|m) \geqslant \ln p(\mathbf{X}|m) - \mathrm{KL}\left(q_{\mathbf{w}}(\mathbf{Z}, \boldsymbol{\Theta}) \| p(\mathbf{Z}, \boldsymbol{\Theta}|\mathbf{X}, m)\right) \triangleq -\mathcal{F}(\mathbf{w}).$$

Variational lower bound or evidence lower bound (ELBO)

$$\ln p(\mathbf{X}|m) \geqslant \ln p(\mathbf{X}|m) - \mathrm{KL}\left(q_{\mathbf{w}}(\mathbf{Z}, \boldsymbol{\Theta}) \| p(\mathbf{Z}, \boldsymbol{\Theta}|\mathbf{X}, m)\right) \triangleq -\mathcal{F}(\mathbf{w}).$$

- The lower bound to the log marginal likelihood is obtained by applying **Jensen**'s inequality:

$$\begin{aligned}
\ln p(\mathbf{X}|m) &= \ln \iint p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\Theta}|m)\, d\mathbf{Z}\, d\boldsymbol{\Theta} \\
&= \ln \iint q_{\mathbf{w}}(\mathbf{Z}, \boldsymbol{\Theta}) \frac{p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\Theta}|m)}{q_{\mathbf{w}}(\mathbf{Z}, \boldsymbol{\Theta})}\, d\mathbf{Z}\, d\boldsymbol{\Theta} \\
&\geqslant \iint q_{\mathbf{w}}(\mathbf{Z}, \boldsymbol{\Theta}) \ln \frac{p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\Theta}|m)}{q_{\mathbf{w}}(\mathbf{Z}, \boldsymbol{\Theta})}\, d\mathbf{Z}\, d\boldsymbol{\Theta} \\
&= \ln p(\mathbf{X}|m) + \iint q_{\mathbf{w}}(\mathbf{Z}, \boldsymbol{\Theta}) \ln \frac{p(\mathbf{Z}, \boldsymbol{\Theta}|\mathbf{X}, m)}{q_{\mathbf{w}}(\mathbf{Z}, \boldsymbol{\Theta})}\, d\mathbf{Z}\, d\boldsymbol{\Theta}.
\end{aligned}$$

Variational lower bound or evidence lower bound (ELBO)

$$\ln p(\mathbf{X}|m) \geqslant \ln p(\mathbf{X}|m) - \mathrm{KL}\left(q_{\mathbf{w}}(\mathbf{Z}, \boldsymbol{\Theta}) \| p(\mathbf{Z}, \boldsymbol{\Theta}|\mathbf{X}, m)\right) \triangleq -\mathcal{F}(\mathbf{w}).$$

- The lower bound to the log marginal likelihood is obtained by applying **Jensen**'s inequality:

$$\begin{aligned}
\ln p(\mathbf{X}|m) &= \ln \iint p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\Theta}|m)\, d\mathbf{Z}\, d\boldsymbol{\Theta} \\
&= \ln \iint q_{\mathbf{w}}(\mathbf{Z}, \boldsymbol{\Theta}) \frac{p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\Theta}|m)}{q_{\mathbf{w}}(\mathbf{Z}, \boldsymbol{\Theta})}\, d\mathbf{Z}\, d\boldsymbol{\Theta} \\
&\geqslant \iint q_{\mathbf{w}}(\mathbf{Z}, \boldsymbol{\Theta}) \ln \frac{p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\Theta}|m)}{q_{\mathbf{w}}(\mathbf{Z}, \boldsymbol{\Theta})}\, d\mathbf{Z}\, d\boldsymbol{\Theta} \\
&= \ln p(\mathbf{X}|m) + \iint q_{\mathbf{w}}(\mathbf{Z}, \boldsymbol{\Theta}) \ln \frac{p(\mathbf{Z}, \boldsymbol{\Theta}|\mathbf{X}, m)}{q_{\mathbf{w}}(\mathbf{Z}, \boldsymbol{\Theta})}\, d\mathbf{Z}\, d\boldsymbol{\Theta}.
\end{aligned}$$

- The analytically intractable integration problem is replaced by an optimisation problem!

Other forms of the ELBO

$$-\mathcal{F}(\mathbf{w}) = \iint q_{\mathbf{w}}(\mathbf{Z}, \boldsymbol{\Theta}) \ln \frac{p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\Theta} | m)}{q_{\mathbf{w}}(\mathbf{Z}, \boldsymbol{\Theta})} \, d\mathbf{Z} \, d\boldsymbol{\Theta}.$$

Other forms of the ELBO

$$-\mathcal{F}(\mathbf{w}) = \iint q_{\mathbf{w}}(\mathbf{Z}, \boldsymbol{\Theta}) \ln \frac{p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\Theta} | m)}{q_{\mathbf{w}}(\mathbf{Z}, \boldsymbol{\Theta})} \, d\mathbf{Z} \, d\boldsymbol{\Theta}.$$

- Free energy interpretation:

$$+\mathcal{F}(\mathbf{w}) = \underbrace{-\mathbb{E}\left(\ln p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\Theta} | m)\right)}_{\text{energy}} - \underbrace{\mathrm{H}\left(q_{\mathbf{w}}(\mathbf{Z}, \boldsymbol{\Theta})\right)}_{\text{entropy}}. \tag{1}$$

Other forms of the ELBO

$$-\mathcal{F}(\mathbf{w}) = \iint q_{\mathbf{w}}(\mathbf{Z}, \boldsymbol{\Theta}) \ln \frac{p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\Theta}|m)}{q_{\mathbf{w}}(\mathbf{Z}, \boldsymbol{\Theta})} \, d\mathbf{Z} \, d\boldsymbol{\Theta}.$$

- Free energy interpretation:

$$+\mathcal{F}(\mathbf{w}) = \underbrace{-\mathbb{E}\left(\ln p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\Theta}|m)\right)}_{energy} - \underbrace{\mathrm{H}\left(q_{\mathbf{w}}(\mathbf{Z}, \boldsymbol{\Theta})\right)}_{entropy}. \tag{1}$$

- Penalized model fit interpretation:

$$-\mathcal{F}(\mathbf{w}) = \underbrace{\mathbb{E}\left(\ln p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\Theta}, m)\right)}_{model\ \ fit} - \underbrace{\mathrm{KL}\left(q_{\mathbf{w}}(\mathbf{Z}, \boldsymbol{\Theta}) \| p(\mathbf{Z}, \boldsymbol{\Theta}|m)\right)}_{penalty}. \tag{2}$$

Definitions

The differential entropy measures the randomness of a random variable:

$$\mathrm{H}\left(p\right) = -\int p(\mathbf{x}) \ln p(\mathbf{x})\; d\mathbf{x}.$$

The Kullback-Leibler divergence or relative entropy measures how to probability densities differ:

$$\mathrm{KL}\left(q \| p\right) = -\int q(\mathbf{x}) \ln \frac{p(\mathbf{x})}{q(\mathbf{x})}\; d\mathbf{x} \;\geqslant 0.$$

The KL is asymmetric (thus not a distance) and only zero if $q(\mathbf{x}) = p(\mathbf{x})$ for all $\mathbf{x}$.

Variational Inference

Mean field variational inference [Bea03]

- A tractable solution is found by assuming $q_{\mathbf{w}}$ factorises given the data:

$$q_{\mathbf{w}}(\mathbf{Z}, \boldsymbol{\Theta}) = \prod_n q(\mathbf{z}_n; \mathbf{w}_n) \times \prod_m q(\boldsymbol{\theta}_m; \mathbf{w}_m).$$

Mean field variational inference [Bea03]

- A tractable solution is found by assuming $q_{\mathbf{w}}$ factorises given the data:

$$q_{\mathbf{w}}(\mathbf{Z}, \boldsymbol{\Theta}) = \prod_n q(\mathbf{z}_n; \mathbf{w}_n) \times \prod_m q(\boldsymbol{\theta}_m; \mathbf{w}_m).$$

- The ELBO is given by

$$-\mathcal{F}(\mathbf{w}) = \sum_n \mathbb{E}\left(\ln p(\mathbf{x}_n|\mathbf{z}_n, \boldsymbol{\Theta})\right) - \sum_n \mathrm{KL}\left(q(\mathbf{z}_n; \mathbf{w}_n) \| p(\mathbf{z}_n)\right) - \sum_m \mathrm{KL}\left(q(\boldsymbol{\theta}_m; \mathbf{w}_m) \| p(\boldsymbol{\theta}_m)\right).$$

- Variational inference (or variational Bayes or variational EM) alternates between the following two steps:

$$q(\mathbf{z}_n; \mathbf{w}_n) \propto e^{\mathbb{E}_{\neg \mathbf{z}_n}(\ln p(\mathbf{x}_n, \mathbf{z}_n|\boldsymbol{\Theta}))}, \qquad q(\boldsymbol{\theta}_m; \mathbf{w}_m) \propto e^{\mathbb{E}_{\neg \boldsymbol{\theta}_m}(\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\Theta}))} p(\boldsymbol{\theta}_m).$$

Mean field variational inference [Bea03]

- A tractable solution is found by assuming $q_{\mathbf{w}}$ factorises given the data:

$$q_{\mathbf{w}}(\mathbf{Z}, \boldsymbol{\Theta}) = \prod_n q(\mathbf{z}_n; \mathbf{w}_n) \times \prod_m q(\boldsymbol{\theta}_m; \mathbf{w}_m).$$

- The ELBO is given by

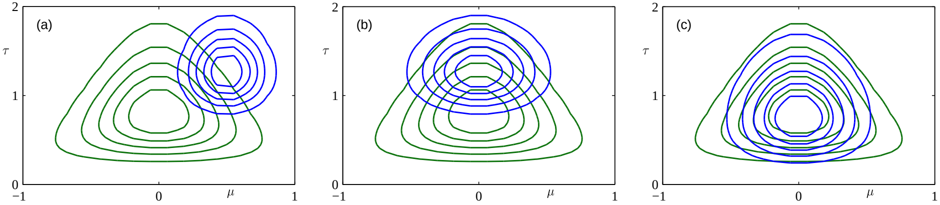
$$-\mathcal{F}(\mathbf{w}) = \sum_n \mathbb{E}\left(\ln p(\mathbf{x}_n|\mathbf{z}_n, \boldsymbol{\Theta})\right) - \sum_n \mathrm{KL}\left(q(\mathbf{z}_n; \mathbf{w}_n) \| p(\mathbf{z}_n)\right) - \sum_m \mathrm{KL}\left(q(\boldsymbol{\theta}_m; \mathbf{w}_m) \| p(\boldsymbol{\theta}_m)\right).$$

- Variational inference (or variational Bayes or variational EM) alternates between the following two steps:

$$q(\mathbf{z}_n; \mathbf{w}_n) \propto e^{\mathbb{E}_{\neg \mathbf{z}_n}(\ln p(\mathbf{x}_n, \mathbf{z}_n|\boldsymbol{\Theta}))}, \qquad q(\boldsymbol{\theta}_m; \mathbf{w}_m) \propto e^{\mathbb{E}_{\neg \boldsymbol{\theta}_m}(\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\Theta}))} p(\boldsymbol{\theta}_m).$$

- The algorithm iteratively and monotonically maximises the ELBO, converging to a local maximum of the bound (not the evidence!)

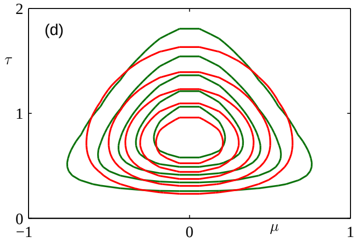
Variational inference in action



Iteration 1 Iteration 2 Iteration 3

(Image credit [Bis06])

What is lost?



Gaussian-Gamma

(Image credit [Bis06])

How to make predictions?

- The predictive distribution is approximated by plugging in the approximate posterior $q_{\mathbf{w}}$:

$$p(\mathbf{x}_*|\mathbf{X}) \approx \iint p(\mathbf{x}_*|\mathbf{z}_*, \boldsymbol{\Theta})\ q(\mathbf{z}_*; \mathbf{w}_*)\ q(\boldsymbol{\Theta}; \{\mathbf{w}_m\}_m)\ d\mathbf{z}_*\ d\boldsymbol{\Theta}.$$

- When analytically intractable, one can use Monte Carlo integration or heuristics based on statistics under the approximate posterior:

$$p(\mathbf{x}_*|\mathbf{X}) \approx p\left(\mathbf{x}_*|\mathbb{E}\left(\mathbf{z}_*\right), \mathbb{E}\left(\boldsymbol{\Theta}\right)\right).$$

Relation to expectation-maximisation (EM) [NH93]

$$-\mathcal{F}(\mathbf{w}) = \ln p(\mathbf{X}|\boldsymbol{\Theta}) - \mathrm{KL}\left(q(\mathbf{Z}) \| p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\Theta})\right),$$

$$-\mathcal{F}(\mathbf{w}) = \mathbb{E}\left(\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\Theta})\right) + \mathrm{H}\left(q(\mathbf{Z})\right).$$

Relation to expectation-maximisation (EM) [NH93]

$$-\mathcal{F}(\mathbf{w}) = \ln p(\mathbf{X}|\mathbf{\Theta}) - \mathrm{KL}\left(q(\mathbf{Z}) \| p(\mathbf{Z}|\mathbf{X}, \mathbf{\Theta})\right),$$
$$-\mathcal{F}(\mathbf{w}) = \mathbb{E}\left(\ln p(\mathbf{X}, \mathbf{Z}|\mathbf{\Theta})\right) + \mathrm{H}\left(q(\mathbf{Z})\right).$$

- Expectation step: $q(\mathbf{Z}) \leftarrow p(\mathbf{Z}|\mathbf{X}, \mathbf{\Theta}^{old})$.
- Maximisation step: $\mathbf{\Theta}^{new} = \arg\max_{\mathbf{\Theta}} \mathbb{E}\left(\ln p(\mathbf{X}, \mathbf{Z}|\mathbf{\Theta})\right)$.

Relation to expectation-maximisation (EM) [NH93]

$$-\mathcal{F}(\mathbf{w}) = \ln p(\mathbf{X}|\boldsymbol{\Theta}) - \mathrm{KL}\left(q(\mathbf{Z}) \| p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\Theta})\right),$$
$$-\mathcal{F}(\mathbf{w}) = \mathbb{E}\left(\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\Theta})\right) + \mathrm{H}\left(q(\mathbf{Z})\right).$$

- Expectation step: $q(\mathbf{Z}) \leftarrow p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\Theta}^{old})$.
- Maximisation step: $\boldsymbol{\Theta}^{new} = \arg\max_{\boldsymbol{\Theta}} \mathbb{E}\left(\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\Theta})\right)$.

- EM guarantees monotonic increase of the bound by construction.
- EM converges to local optimum of the log likelihood [Wu83].
- Approximate EM if q approximates the posterior [HZW03].

EM in pictures

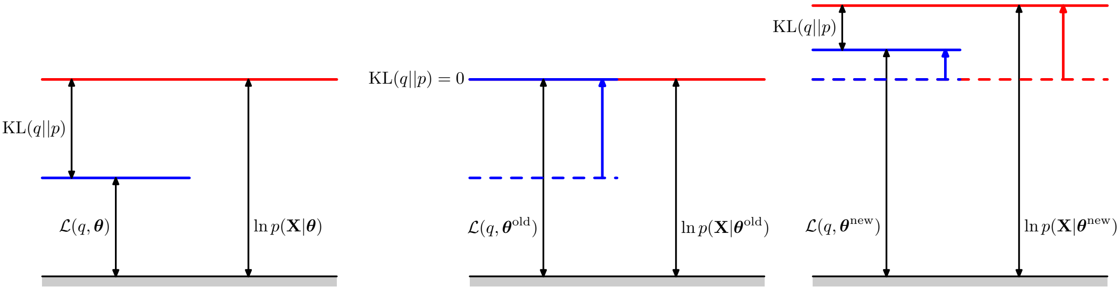


Image credit: [Bis06].

Structured variational inference [SJ95, Wie00]

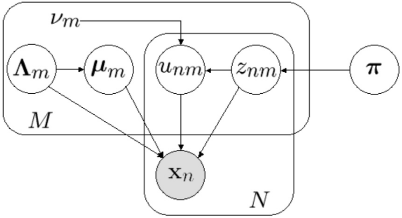
$$\arg\min_{\mathbf{w}} \mathrm{KL}\left(q_{\mathbf{w}}(\mathbf{Z}, \boldsymbol{\Theta}) \| p(\mathbf{Z}, \boldsymbol{\Theta} | \mathbf{X}, m)\right)$$

- Mean field considers a fully factorised form to find a tractable solution.
- Structured variational inference avoids factorising when possible or imposes an approximate posterior of a predefined specific form.

Example: mixture of Student-t distributions [AV07]

$$p(\mathbf{x}|\mathbf{\Theta}) = \sum_m \pi_m \,\mathrm{Student}\left(\mathbf{x}|\boldsymbol{\mu}_m, \mathbf{\Lambda}_m, \nu_m\right),$$

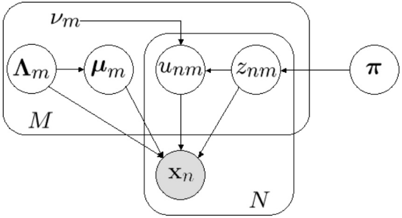
$$\mathrm{Student}\left(\mathbf{x}|\boldsymbol{\mu}_m, \mathbf{\Lambda}, \nu_m\right) = \int_{-\infty}^{+\infty} \mathrm{Gaussian}\left(\mathbf{x}|\boldsymbol{\mu}_m, \mathbf{\Lambda} u_m\right) \mathrm{Gamma}\left(u_m|\nu_m/2, \nu_m/2\right) du_m.$$



Example: mixture of Student-t distributions [AV07]

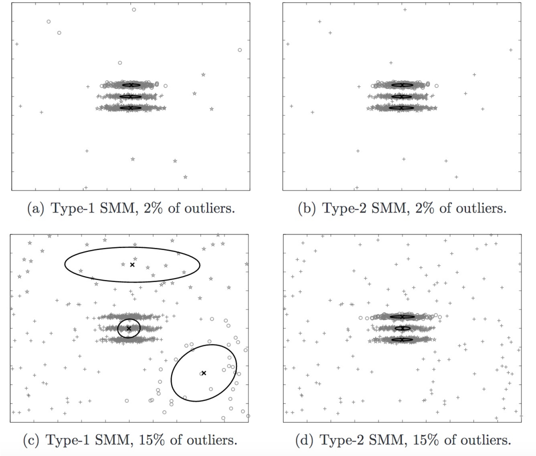
$$p(\mathbf{x}|\boldsymbol{\Theta}) = \sum_m \pi_m \,\mathrm{Student}\left(\mathbf{x}|\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m, \nu_m\right),$$

$$\mathrm{Student}\left(\mathbf{x}|\boldsymbol{\mu}_m, \boldsymbol{\Lambda}, \nu_m\right) = \int_{-\infty}^{+\infty} \mathrm{Gaussian}\left(\mathbf{x}|\boldsymbol{\mu}_m, \boldsymbol{\Lambda} u_m\right) \mathrm{Gamma}\left(u_m|\nu_m/2, \nu_m/2\right) du_m.$$

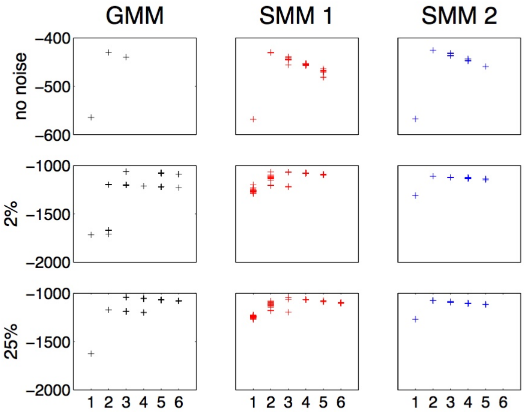


$$q(\mathbf{u}_n, \mathbf{z}_n) = \prod_m q(u_{nm}) q(z_{nm}) \quad \text{(SMM1)}$$

$$q(\mathbf{u}_n, \mathbf{z}_n) = \prod_m q(u_{nm}, z_{nm}) \quad \text{(SMM2)}$$



Robustness against outliers.



Model selection with ELBO.

Stochastic Variational Inference and Other Variants

Mean field variational inference (MVI)

$$-\mathcal{F}(\mathbf{w}) = \sum_n \underbrace{\mathbb{E}\left(\ln p(\mathbf{x}_n|\mathbf{z}_n, \boldsymbol{\Theta})\right)}_{=\ell_n(\mathbf{w})} - \sum_n \mathrm{KL}\left(q(\mathbf{z}_n; \mathbf{w}_n) \| p(\mathbf{z}_n)\right) - \sum_m \mathrm{KL}\left(q(\boldsymbol{\theta}_m; \mathbf{w}_m) \| p(\boldsymbol{\theta}_m)\right).$$

Mean field variational inference (MVI)

$$-\mathcal{F}(\mathbf{w}) = \sum_n \underbrace{\mathbb{E}\left(\ln p(\mathbf{x}_n|\mathbf{z}_n, \boldsymbol{\Theta})\right)}_{=\ell_n(\mathbf{w})} - \sum_n \mathrm{KL}\left(q(\mathbf{z}_n; \mathbf{w}_n) \| p(\mathbf{z}_n)\right) - \sum_m \mathrm{KL}\left(q(\boldsymbol{\theta}_m; \mathbf{w}_m) \| p(\boldsymbol{\theta}_m)\right).$$

MVI can be rewritten as **batch gradient ascent**:

$$\mathbf{w}_n \leftarrow \arg\max_{\mathbf{w}_n} \; \ell_n(\mathbf{w}) - \mathrm{KL}\left(q(\mathbf{z}_n; \mathbf{w}_n) \| p(\mathbf{z}_n)\right), \qquad (\mathrm{VE-step})$$

$$\mathbf{w}_m \leftarrow \arg\max_{\mathbf{w}_m} \; \sum_n \ell_n(\mathbf{w}) - \mathrm{KL}\left(q(\boldsymbol{\theta}_m; \mathbf{w}_m) \| p(\boldsymbol{\theta}_m)\right). \qquad (\mathrm{VM-step})$$

- Monotonic increase of the bound; converges to local maximum of ELBO
- Priors are conjugate to the likelihood; updates are similar to Gibbs sampling.
- Not suitable for large data sets!

Noisy, but unbiased estimates of the gradient wrt $\mathbf{w}_m$

$$-\mathcal{F}(\mathbf{w}) = \sum_n \ell_n(\mathbf{w}) - \sum_n \mathrm{KL}\left(q_{\mathbf{w}}(\mathbf{z}_n) \| p(\mathbf{z}_n)\right)) - \sum_m \mathrm{KL}\left(q_{\mathbf{w}}(\boldsymbol{\theta}_m) \| p(\boldsymbol{\theta}_m)\right).$$

Noisy, but unbiased estimates of the gradient wrt $\mathbf{w}_m$

$$-\mathcal{F}(\mathbf{w}) = \sum_n \ell_n(\mathbf{w}) - \sum_n \mathrm{KL}\left(q_{\mathbf{w}}(\mathbf{z}_n) \| p(\mathbf{z}_n)\right) - \sum_m \mathrm{KL}\left(q_{\mathbf{w}}(\boldsymbol{\theta}_m) \| p(\boldsymbol{\theta}_m)\right).$$

$$-\frac{\partial \mathcal{F}(\mathbf{w})}{\partial \mathbf{w}_m} = \frac{\partial}{\partial \mathbf{w}_m} \left(\sum_n \ell_n(\mathbf{w}) - \mathrm{KL}\left(q_{\mathbf{w}}(\boldsymbol{\theta}_m) \| p(\boldsymbol{\theta}_m)\right) \right)$$

Noisy, but unbiased estimates of the gradient wrt $\mathbf{w}_m$

$$-\mathcal{F}(\mathbf{w}) = \sum_n \ell_n(\mathbf{w}) - \sum_n \mathrm{KL}\left(q_{\mathbf{w}}(\mathbf{z}_n) \| p(\mathbf{z}_n)\right) - \sum_m \mathrm{KL}\left(q_{\mathbf{w}}(\boldsymbol{\theta}_m) \| p(\boldsymbol{\theta}_m)\right).$$

$$\begin{aligned}
-\frac{\partial \mathcal{F}(\mathbf{w})}{\partial \mathbf{w}_m} &= \frac{\partial}{\partial \mathbf{w}_m}\left(\sum_n \ell_n(\mathbf{w}) - \mathrm{KL}\left(q_{\mathbf{w}}(\boldsymbol{\theta}_m) \| p(\boldsymbol{\theta}_m)\right)\right) \\
&= \sum_n \frac{\partial}{\partial \mathbf{w}_m}\left(\ell_n(\mathbf{w}) - \frac{\mathrm{KL}\left(q_{\mathbf{w}}(\boldsymbol{\theta}_m) \| p(\boldsymbol{\theta}_m)\right)}{N}\right).
\end{aligned}$$

Stochastic variational inference (SVI) [HBB10]

We use stochastic gradient descent in the variational M-step:

$$\mathbf{w}_m \leftarrow \mathbf{w}_m + \rho_t N \frac{\partial}{\partial \mathbf{w}_m} \left(\ell_n(\mathbf{w}) - \frac{\mathrm{KL}\left(q(\boldsymbol{\theta}_m; \mathbf{w}_m) \| p(\boldsymbol{\theta}_m)\right)}{N} \right),$$

where $\sum_t \rho_t = \infty$ and $\sum_t \rho_t^2 < \infty$.

Stochastic variational inference (SVI) [HBB10]

We use stochastic gradient descent in the variational M-step:

$$\mathbf{w}_m \leftarrow \mathbf{w}_m + \rho_t N \frac{\partial}{\partial \mathbf{w}_m} \left(\ell_n(\mathbf{w}) - \frac{\mathrm{KL}\left(q(\boldsymbol{\theta}_m; \mathbf{w}_m) \| p(\boldsymbol{\theta}_m)\right)}{N} \right),$$

where $\sum_t \rho_t = \infty$ and $\sum_t \rho_t^2 < \infty$.

- Stochastic approximation of the gradient [RM51]:
  - Small memory footprint; sequential method.
  - Requires adjusting the learning rate ρ_t.
  - Monotonic increase of bound is lost (no sanity check)

Stochastic variational inference (SVI) [HBB10]

We use stochastic gradient descent in the variational M-step:

$$\mathbf{w}_m \leftarrow \mathbf{w}_m + \rho_t N \frac{\partial}{\partial \mathbf{w}_m} \left(\ell_n(\mathbf{w}) - \frac{\mathrm{KL}\left(q(\boldsymbol{\theta}_m; \mathbf{w}_m) \| p(\boldsymbol{\theta}_m)\right)}{N} \right),$$

where $\sum_t \rho_t = \infty$ and $\sum_t \rho_t^2 < \infty$.

- Stochastic approximation of the gradient [RM51]:
  - Small memory footprint; sequential method.
  - Requires adjusting the learning rate ρ_t.
  - Monotonic increase of bound is lost (no sanity check)
- SVI corresponds to stochastic natural gradients wrt $q_{\mathbf{w}_m}$ for exponential family distributions [HBWP13].

Incremental variational inference (IVI) [AE15]

$$-\mathcal{F}(\mathbf{w}) = \underbrace{\sum_n \ell_n(\mathbf{w})}_{=\ell_N(\mathbf{w})} - \sum_n \mathrm{KL}\left(q(\mathbf{z}_n; \mathbf{w}_n) \| p(\mathbf{z}_n)\right) - \sum_m \mathrm{KL}\left(q(\boldsymbol{\theta}_m; \mathbf{w}_m) \| p(\boldsymbol{\theta}_m)\right).$$

Incremental variational inference (IVI) [AE15]

$$-\mathcal{F}(\mathbf{w}) = \underbrace{\sum_n \ell_n(\mathbf{w})}_{=\ell_N(\mathbf{w})} - \sum_n \mathrm{KL}\left(q(\mathbf{z}_n;\mathbf{w}_n) \| p(\mathbf{z}_n)\right) - \sum_m \mathrm{KL}\left(q(\boldsymbol{\theta}_m;\mathbf{w}_m) \| p(\boldsymbol{\theta}_m)\right).$$

Let $\mathbf{s}(\mathbf{X},\mathbf{Z}) = \sum_n \mathbf{s}_n(\mathbf{x}_n,\mathbf{z}_n)$ be the vector of sufficient statistics:

$$\mathbf{w}_m \leftarrow \arg\max_{\mathbf{w}_m} \ \ell_N(\mathbf{s},\mathbf{w}) - \ell_n(\mathbf{s}_n,\mathbf{w}) + \ell_n(\mathbf{s}_n^*,\mathbf{w}) - \mathrm{KL}\left(q(\boldsymbol{\theta}_m;\mathbf{w}_m) \| p(\boldsymbol{\theta}_m)\right).$$

where $\mathbf{s}_n^*(\mathbf{x}_n,\mathbf{z}_n)$ is the new vector of sufficient statistics.

Incremental variational inference (IVI) [AE15]

$$-\mathcal{F}(\mathbf{w}) = \underbrace{\sum_n \ell_n(\mathbf{w})}_{=\ell_N(\mathbf{w})} - \sum_n \mathrm{KL}\left(q(\mathbf{z}_n; \mathbf{w}_n) \| p(\mathbf{z}_n)\right) - \sum_m \mathrm{KL}\left(q(\boldsymbol{\theta}_m; \mathbf{w}_m) \| p(\boldsymbol{\theta}_m)\right).$$

Let $\mathbf{s}(\mathbf{X}, \mathbf{Z}) = \sum_n \mathbf{s}_n(\mathbf{x}_n, \mathbf{z}_n)$ be the vector of sufficient statistics:

$$\mathbf{w}_m \leftarrow \arg\max_{\mathbf{w}_m} \; \ell_N(\mathbf{s}, \mathbf{w}) - \ell_n(\mathbf{s}_n, \mathbf{w}) + \ell_n(\mathbf{s}_n^*, \mathbf{w}) - \mathrm{KL}\left(q(\boldsymbol{\theta}_m; \mathbf{w}_m) \| p(\boldsymbol{\theta}_m)\right).$$

where $\mathbf{s}_n^*(\mathbf{x}_n, \mathbf{z}_n)$ is the new vector of sufficient statistics.

- Sequential like SVI, but maintains a batch estimate of $\mathbf{s}(\mathbf{X}, \mathbf{Z})$.
- Needs to store the sufficient statistics.
- No parameters to tune.
- Monotonic increase of bound is recovered!
- Can be interpreted as stochastic average gradient descent [SLB13].

Relation to incremental EM

- MVI updates can be re-written in terms of the sufficient statistics:

$$q(\mathbf{z}_n; \mathbf{w}_n) \propto e^{\mathbb{E}_{\neg \mathbf{z}_n}(\ln p(\mathbf{s}_n|\boldsymbol{\Theta}))}, \qquad q(\boldsymbol{\theta}_m; \mathbf{w}_m) \propto e^{\mathbb{E}_{\neg \boldsymbol{\theta}_m}(\ln p(\mathbf{s}|\boldsymbol{\Theta}))} p(\boldsymbol{\theta}_m).$$

Relation to incremental EM

- MVI updates can be re-written in terms of the sufficient statistics:

$$q(\mathbf{z}_n;\mathbf{w}_n) \propto e^{\mathbb{E}_{\neg\mathbf{z}_n}(\ln p(\mathbf{s}_n|\boldsymbol{\Theta}))}, \qquad q(\boldsymbol{\theta}_m;\mathbf{w}_m) \propto e^{\mathbb{E}_{\neg\boldsymbol{\theta}_m}(\ln p(\mathbf{s}|\boldsymbol{\Theta}))} p(\boldsymbol{\theta}_m).$$

- IVI updates can be re-written in a similar fashion as in incremental EM [NH93]:

$$q(\mathbf{z}_n;\mathbf{w}_n) \propto e^{\mathbb{E}_{\neg\mathbf{z}_n}(\ln p(\mathbf{s}_n^*|\boldsymbol{\Theta}))}, \qquad q(\boldsymbol{\theta}_m;\mathbf{w}_m) \propto e^{\mathbb{E}_{\neg\boldsymbol{\theta}_m}(\ln p(\mathbf{s}-\mathbf{s}_n+\mathbf{s}_n^*,\boldsymbol{\Theta}))} p(\boldsymbol{\theta}_m).$$

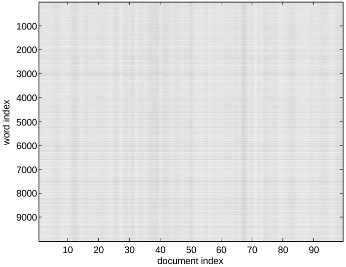
Topic models

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

- Organise and browse large document collections.
- Capture underlying semantic structure in an unsupervised way.
- Extremely popular (e.g., more than 22k citations in Google Scholar)

Latent Dirichlet allocation (LDA) [DMB03]



Observations are word counts per document. LDA assumes an admixture model:

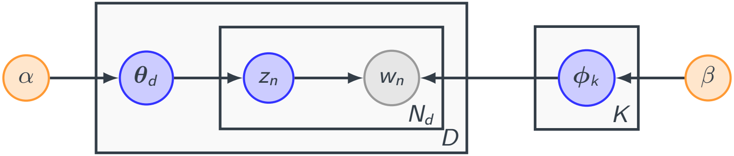
$$\mathbf{X} \in \mathbb{N}^{V \times D}.$$

LDA infers a low-rank approximation of the matrix of counts:

$$\mathrm{E}\left(\mathbf{X}\right) \approx \boldsymbol{\Phi}\boldsymbol{\Theta}^{\top}, \qquad \mathbf{x}_d \sim \mathrm{Multinomial}(\boldsymbol{\Phi}\boldsymbol{\theta}_d, N_d)$$

where $\boldsymbol{\Phi} \in \mathbb{R}_{+}^{V \times K}$, $\boldsymbol{\Theta} \in \mathbb{R}_{+}^{D \times K}$ and K is small.

Graphical model



$$\boldsymbol{\theta}_d \sim \text{Dirichlet}(\alpha \mathbf{1}_K), \qquad z_n | \boldsymbol{\theta}_d \sim \text{Categorical}(\boldsymbol{\theta}_d),$$

$$\boldsymbol{\phi}_k \sim \text{Dirichlet}(\beta \mathbf{1}_V), \qquad w_n | z_n, \{\boldsymbol{\phi}_k\}_{k=1}^{K} \sim \text{Categorical}(\boldsymbol{\phi}_{z_n}).$$

Log-predictive probability for LDA as a function of the number of processed documents

IVI converges faster and to a higher value on all considered datasets. (K=100, $\alpha_0 = 0.5$ and $\beta_0 = 0.05$)

Yet another form of the ELBO based on the score function

$$-\mathcal{F}(\mathbf{w}) = \mathbb{E}\left(\ln p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\Theta}|m)\right) + \mathrm{H}\left(q_{\mathbf{w}}(\mathbf{Z}, \boldsymbol{\Theta})\right).$$

Yet another form of the ELBO based on the score function

$$-\mathcal{F}(\mathbf{w}) = \mathbb{E}\left(\ln p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\Theta}|m)\right) + \mathrm{H}\left(q_{\mathbf{w}}(\mathbf{Z}, \boldsymbol{\Theta})\right).$$

Write the gradient in terms of the score function:

$$\begin{aligned}
-\frac{\partial \mathcal{F}(\mathbf{w})}{\partial \mathbf{w}_n} &= \mathbb{E}\left(\frac{\partial \ln q(\mathbf{z}_n; \mathbf{w}_n)}{\partial \mathbf{w}_n}\left(\ln p(\mathbf{x}_n, \mathbf{z}_n|\boldsymbol{\Theta}) - \ln q(\mathbf{z}_n; \mathbf{w}_n)\right)\right) \\
&\approx \frac{1}{K}\sum_{k=1}^{K}\left(\frac{\partial \ln q(\mathbf{z}_n^{(k)}; \mathbf{w}_n)}{\partial \mathbf{w}_n}\left(\ln p(\mathbf{x}_n, \mathbf{z}_n^{(k)}|\boldsymbol{\Theta}) - \ln q(\mathbf{z}_n^{(k)}; \mathbf{w}_n)\right)\right),
\end{aligned}$$

where $\mathbf{z}_n^{(k)} \sim q(\mathbf{z}_n^{(k)}; \mathbf{w}_n)$.

Black-box variational inference [RGB14]

$$\mathbf{w}_n \leftarrow \mathbf{w}_n + \frac{\lambda_t}{K} \sum_{k=1}^{K} \left(\frac{\partial \ln q(\mathbf{z}_n^{(k)}; \mathbf{w}_n)}{\partial \mathbf{w}_n} \left(\ln p(\mathbf{x}_n, \mathbf{z}_n^{(k)} | \boldsymbol{\Theta}) - \ln q(\mathbf{z}_n^{(k)}; \mathbf{w}_n) \right) \right),$$

where $\sum_t \lambda_t = \infty$ and $\sum_t \lambda_t^2 < \infty$.

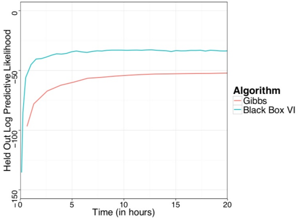


Figure 1: Comparison between Metropolis-Hastings within Gibbs and Black Box Variational Inference. In the x axis is time and in the y axis is the predictive likelihood of the test set. Black Box Variational Inference reaches better predictive likelihoods faster than Gibbs sampling. The Gibbs sampler's progress slows considerably after 5 hours.

Black-box variational inference [RGB14]

$$\mathbf{w}_n \leftarrow \mathbf{w}_n + \frac{\lambda_t}{K} \sum_{k=1}^{K} \left(\frac{\partial \ln q(\mathbf{z}_n^{(k)}; \mathbf{w}_n)}{\partial \mathbf{w}_n} \left(\ln p(\mathbf{x}_n, \mathbf{z}_n^{(k)} | \boldsymbol{\Theta}) - \ln q(\mathbf{z}_n^{(k)}; \mathbf{w}_n) \right) \right),$$

where $\sum_t \lambda_t = \infty$ and $\sum_t \lambda_t^2 < \infty$.

- Remove conjugacy requirement
- Variance reduction techniques:
  - Rao-Blackwellization
  - Control variates
- Can be scaled up with SVI

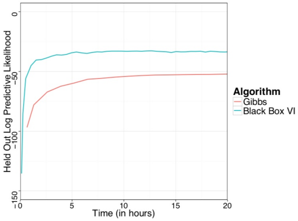
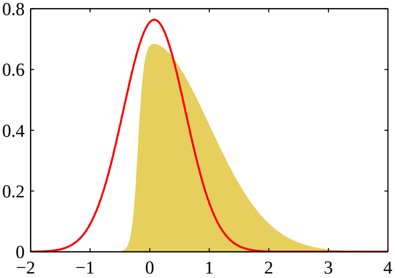
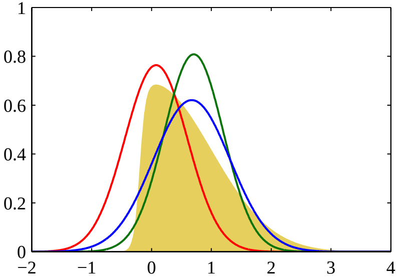


Figure 1: Comparison between Metropolis-Hastings within Gibbs and Black Box Variational Inference. In the x axis is time and in the y axis is the predictive likelihood of the test set. Black Box Variational Inference reaches better predictive likelihoods faster than Gibbs sampling. The Gibbs sampler's progress slows considerably after 5 hours.

Other approximate inference methods



Laplace approximation.



$\mathrm{KL}\,(q\|p)$ vs. $\mathrm{KL}\,(p\|q)$. [Min01]

(Image credit: [Bis06].

Further reading

Christopher Bishop (2006): Pattern Recognition and Machine Learning. [Bis06]

Kevin Murphy (2012): Machine Learning: a Probabilistic Perspective. [Mur12]

David Blei, et al. (2017): Variational Inference: a Review for Statisticians. [BKM17]

References I

C. Archambeau and B. Ermis.
Incremental Variational Inference for Latent Dirichlet Allocation.
ArXiv e-prints, 2015.

Cedric Archambeau and Michel Verleysen.
Robust bayesian clustering.
Neural Networks, 20(1):129 – 138, 2007.

M.J. Beal.
Variational Algorithms for Approximate Bayesian Inference.
PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

C. M. Bishop.
Pattern Recognition and Machine Learning.
Springer, 2006.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe.
Variational inference: A review for statisticians.
Journal of the American Statistical Association, 112(518):859–877, 2017.

References II

Michael I. Jordan David M. Blei, Andrew Y. Ng.
Latent dirichlet allocation.
Journal of Machine Learning Research, 3:993–1022, 2003.

Matthew Hoffman, Francis R. Bach, and David M. Blei.
Online learning for latent dirichlet allocation.
In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 856–864. 2010.

Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley.
Stochastic variational inference.
Journal of Machine Learning Research, 14:1303–1347, 2013.

Tom Heskes, Onno Zoeter, and Wim Wiegerinck.
Approximate expectation maximization.
In *Advances in Neural Information Processing Systems 16*, pages 353–360, 2003.

References III

Thomas Minka.
Expectation propagation for approximate bayesian inference.
In Daphne Koller Jack S. Breese, editor, *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, pages 362–369, 2001.

Kevin P Murphy.
Machine learning: a probabilistic perspective.
MIT press, 2012.

Radford M. Neal and Geoffrey E. Hinton.
A new view of the em algorithm that justifies incremental and other variants.
In *Learning in Graphical Models*, pages 355–368. Kluwer Academic Publishers, 1993.

Rajesh Ranganath, Sean Gerrish, and David Blei.
Black Box Variational Inference.
In Samuel Kaski and Jukka Corander, editors, *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 814–822, Reykjavik, Iceland, 22–25 Apr 2014. PMLR.

References IV

Herbert Robbins and Sutton Monro.
A stochastic approximation method.
The Annals of Mathematical Statistics, 22(3):400–407, 1951.

Lawrence Saul and Michael I. Jordan.
Exploiting tractable substructures in intractable networks.
In *Advances in Neural Information Processing Systems 8*, pages 486–492. MIT Press, 1995.

M. Schmidt, N. Le Roux, and F. Bach.
Minimizing Finite Sums with the Stochastic Average Gradient.
ArXiv e-prints, 2013.

Wim Wiegerinck.
Variational approximations between mean field theory and the junction tree algorithm.
In *Proceedings of the 16th Conference in Uncertainty in Artificial Intelligence*, pages 626–633, 2000.

C. F. Jeff Wu.
On the convergence properties of the em algorithm.
The Annals of Statistics, 11(1):95–103, 1983.