Variational Bayes and beyond: Foundations of scalable Bayesian inference (Part III)

Tamara Broderick
Associate Professor
MIT

http://www.tamarabroderick.com/tutorials.html
Bayesian inference

- Goals: good point estimates, uncertainty estimates
- Challenge: speed (compute, user), reliable inference
Criteo Online Ads Experiment

- MAP: **12 s**
- MFVB: **57 s**
- MCMC (5K samples): 21,066 s (**5.85 h**)

[Giordano, Broderick, Jordan 2018]
Criteo Online Ads Experiment

- MAP: 12 s
- MFVB: 57 s
- MCMC (5K samples): 21,066 s (5.85 h)

[Giordano, Broderick, Jordan 2018]
Criteo Online Ads Experiment

- MAP: **12 s**
- MFVB: **57 s**
- MCMC (5K samples): **21,066 s (5.85 h)**

[Giordano, Broderick, Jordan 2018]
Criteo Online Ads Experiment

- **MAP:** 12 s
- **MFVB:** 57 s
- **MCMC (5K samples):** 21,066 s (5.85 h)

[Giordano, Broderick, Jordan 2018]

[Turner & Sahani 2011; MacKay 2003; Bishop 2006; Wang, Titterington 2004]
Criteo Online Ads Experiment

- MAP: **12 s**
- MFVB: **57 s**
- MCMC (5K samples): 21,066 s (**5.85 h**)

[Giordano, Broderick, Jordan 2018]

- Also: microcredit, graph/network data, etc

[Turner & Sahani 2011; MacKay 2003; Bishop 2006; Wang, Titterington 2004]
### Criteo Online Ads Experiment

- **MAP:** 12 s
- **MFVB:** 57 s
- **MCMC (5K samples):** 21,066 s (5.85 h)

[Giordano, Broderick, Jordan 2018]

- **Also:**
  - microcredit,
  - graph/network data, etc
- **Posterior means**

[Turner & Sahani 2011; MacKay 2003; Bishop 2006; Wang, Titterington 2004]
Approximate Bayesian inference

Use $q^*$ to approximate $p(\cdot|y)$

Optimization

\[ q^* = \arg\min_{q \in Q} f(q(\cdot), p(\cdot|y)) \]

Variational Bayes

\[ q^* = \arg\min_{q \in Q} KL(q(\cdot)||p(\cdot|y)) \]

Mean-field variational Bayes

\[ q^* = \arg\min_{q \in Q_{MFVB}} KL(q(\cdot)||p(\cdot|y)) \]
Approximate Bayesian inference

Use $q^*$ to approximate $p(\cdot | y)$

Optimization
$$q^* = \underset{q \in Q}{\text{argmin}} f(q(\cdot), p(\cdot | y))$$

Variational Bayes
$$q^* = \underset{q \in Q}{\text{argmin}} KL(q(\cdot) || p(\cdot | y))$$

Mean-field variational Bayes
$$q^* = \underset{q \in Q_{MFVB}}{\text{argmin}} KL(q(\cdot) || p(\cdot | y))$$

How deep is the issue?
Approximate Bayesian inference

Use $q^*$ to approximate $p(\cdot|y)$

Optimization
\[
q^* = \arg\min_{q \in Q} f(q(\cdot), p(\cdot|y))
\]

Variational Bayes
\[
q^* = \arg\min_{q \in Q} KL(q(\cdot)||p(\cdot|y))
\]

Mean-field variational Bayes
\[
q^* = \arg\min_{q \in Q_{MFVB}} KL(q(\cdot)||p(\cdot|y))
\]

Algorithm

Implementation
Approximate Bayesian inference

Use $q^*$ to approximate $p(\cdot|y)$

**Optimization**

$$q^* = \arg\min_{q \in Q} f(q(\cdot), p(\cdot|y))$$

**Variational Bayes**

$$q^* = \arg\min_{q \in Q} KL(q(\cdot)||p(\cdot|y))$$

**Mean-field variational Bayes**

$$q^* = \arg\min_{q \in Q_{MFVB}} KL(q(\cdot)||p(\cdot|y))$$

**Algorithm**

**Implementation**

**Gaussian example was exact**

**How deep is the issue?**
Approximate Bayesian inference

Use $q^*$ to approximate $p(\cdot | y)$

Optimization

$$q^* = \arg\min_{q \in Q} f(q(\cdot), p(\cdot | y))$$

Variational Bayes

$$q^* = \arg\min_{q \in Q} KL(q(\cdot) \| p(\cdot | y))$$

Mean-field variational Bayes

$$q^* = \arg\min_{q \in Q_{MFVB}} KL(q(\cdot) \| p(\cdot | y))$$

Algorithm

Implementation

Gaussian example was exact
Approximate Bayesian inference

Use $q^*$ to approximate $p(\cdot \mid y)$

**Optimization**

$q^* = \text{argmin}_{q \in Q} f(q(\cdot), p(\cdot \mid y))$

**Variational Bayes**

$q^* = \text{argmin}_{q \in Q} KL(q(\cdot) \mid \mid p(\cdot \mid y))$

**Mean-field variational Bayes**

$q^* = \text{argmin}_{q \in Q_{MFVB}} KL(q(\cdot) \mid \mid p(\cdot \mid y))$

How deep is the issue?

Gaussian example was exact
Is it just MFVB?
Is it just MFVB?
Is it just MFVB?
Is it just MFVB?

\[
KL(q||p(\cdot|y)) = \int_\theta q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta
\]

\[
q(\theta) = \prod_{j=1}^{J} q_j(\theta_j)
\]
Is it just MFVB?
Is it just MFVB?

\[ p(\theta|y) \quad \rightarrow \quad q^*(\theta) \quad \text{NICE} \]
Is it just MFVB?

- Turner, Sahani (2011) showed (empirically) can have strictly larger NICE set but worse mean & variance estimates.
Is it just MFVB?

• Turner, Sahani (2011) showed (empirically) can have strictly larger NICE set but worse mean & variance estimates
Is it just MFVB?

- Turner, Sahani (2011) showed (empirically) that NICE’ can have strictly larger NICE set but worse mean & variance estimates.
Is it just MFVB?

- Turner, Sahani (2011) showed (empirically) can have strictly larger NICE set but worse mean & variance estimates

- Exercise: Show, with a simple example, that a smaller KL does not imply better mean and variance estimates
Is it just MFVB?

- Turner, Sahani (2011) showed (empirically) can have strictly larger NICE set but worse mean & variance estimates

- Exercise: Show, with a simple example, that a smaller KL does not imply better mean and variance estimates

- But how much worse can the estimates be? And could it have just been the implementation?
Is it just MFVB?
Is it just MFVB?

- Some KL values seen in practice: 
  \ (~1 to \sim 70, 0.5 to 3 \) \[ \text{[Baqué et al 2017; Huggins et al 2020]} \]
Is it just MFVB?

• Some KL values seen in practice: ~1 to ~70, 0.5 to 3  
  [Baqué et al 2017; Huggins et al 2020]

• Take any constant $c$
Is it just MFVB?

- Some KL values seen in practice: ~1 to ~70, 0.5 to 3
- Take any constant $c$

**Proposition.** Can have small KL ($<0.23$) & arbitrarily bad variance estimate

$$\sigma^2_p \geq c\sigma^2_q$$

[Baqué et al 2017; Huggins et al 2020]
Is it just MFVB?

- Some KL values seen in practice: ~1 to ~70, 0.5 to 3
  [Baqué et al 2017; Huggins et al 2020]
- Take any constant $c$

**Proposition.** Can have small KL (<0.23) & arbitrarily bad variance estimate

$$\sigma_p^2 \geq c\sigma_q^2$$
Is it just MFVB?

- Some KL values seen in practice: \( \sim 1 \) to \( \sim 70 \), 0.5 to 3  
  [Baqué et al 2017; Huggins et al 2020]
- Take any constant \( c \)

**Proposition.** Can have small KL (<0.23) & arbitrarily bad variance estimate

\[
\sigma_p^2 \geq c \sigma_q^2
\]

\( \rho: \) Student's t, variance \( \sigma_p^2 \)
Is it just MFVB?

- Some KL values seen in practice: ~1 to ~70, 0.5 to 3  
  [Baqué et al 2017; Huggins et al 2020]
- Take any constant $c$

**Proposition.** Can have small KL (<0.23) & arbitrarily bad variance estimate

$$\sigma_p^2 \geq c\sigma_q^2$$
Is it just MFVB?

- Some KL values seen in practice: \(\sim 1\) to \(\sim 70\), 0.5 to 3
- Take any constant \(c\)

**Proposition.** Can have small KL (<0.23) & arbitrarily bad variance estimate

\[
\sigma_p^2 \geq c \sigma_q^2
\]

\(\sigma^2\): Gaussian, variance \(\sigma^2_q\)

\(\sigma^2\): Student's t, variance \(\sigma^2_p\)
Is it just MFVB?

- Some KL values seen in practice: ~1 to ~70, 0.5 to 3  
  [[Baqué et al 2017; Huggins et al 2020]]
- Take any constant $c$

**Proposition.** Can have small KL ($<0.23$) & arbitrarily bad variance estimate  
\[ \sigma_p^2 \geq c\sigma_q^2 \]
Is it just MFVB?

- Some KL values seen in practice: ~1 to ~70, 0.5 to 3
- Take any constant $c$

**Proposition.** Can have small KL (<0.23) & arbitrarily bad variance estimate

$$\sigma_p^2 \geq c\sigma_q^2$$
Is it just MFVB?

- Some KL values seen in practice: ~1 to ~70, 0.5 to 3
  [Baqué et al 2017; Huggins et al 2020]
- Take any constant $c$

**Proposition.** Can have small KL ($<0.23$) & arbitrarily bad variance estimate

$$\sigma_p^2 \geq c\sigma_q^2$$
Is it just MFVB?

- Some KL values seen in practice: ~1 to ~70, 0.5 to 3
- Take any constant $c$

**Proposition.** Can have small KL (<0.23) & arbitrarily bad variance estimate

\[ \sigma_p^2 \geq c \sigma_q^2 \]

---

**Proposition.** Can have small KL (<0.9) and arbitrarily bad mean estimate

\[ (m_p - m_q)^2 \geq c \sigma_p^2 \]
Is it just MFVB?

- Some KL values seen in practice: ~1 to ~70, 0.5 to 3
- Take any constant $c$

**Proposition.** Can have small KL ($<0.23$) & arbitrarily bad variance estimate

\[ \sigma_p^2 \geq c\sigma_q^2 \]

$p$: Student's t.

$q$: Gaussian, variance $\sigma_q^2$. 

**Proposition.** Can have small KL ($<0.9$) and arbitrarily bad mean estimate

\[ (m_p - m_q)^2 \geq c\sigma_p^2 \]

[Baqué et al 2017; Huggins et al 2020]
Is it just MFVB?

- Some KL values seen in practice: ~1 to ~70, 0.5 to 3
- Take any constant $c$

**Proposition.** Can have small KL (<0.23) & arbitrarily bad variance estimate

$$\sigma_p^2 \geq c\sigma_q^2$$

$p$: Student's t.

$q$: Gaussian, variance $\sigma_q^2$.

$p$: Weibull, mean $m_p$

**Proposition.** Can have small KL (<0.9) and arbitrarily bad mean estimate

$$(m_p - m_q)^2 \geq c\sigma_p^2$$

[Baqué et al 2017; Huggins et al 2020]

[Huggins, Karsprzak, Campbell, Broderick 2020]
Is it just MFVB?

- Some KL values seen in practice: ~1 to ~70, 0.5 to 3
- Take any constant $c$

**Proposition.** Can have small KL (<0.23) & arbitrarily bad variance estimate

$$\sigma_p^2 \geq c\sigma_q^2$$

$p$: Weibull, mean $m_p$

$q$: Weibull, mean $m_q$

**Proposition.** Can have small KL (<0.9) and arbitrarily bad mean estimate

$$(m_p - m_q)^2 \geq c\sigma_p^2$$

[Huggins, Karsprzak, Campbell, Broderick 2020]
Is it just MFVB?

- Some KL values seen in practice: ~1 to ~70, 0.5 to 3
  
- Take any constant $c$

**Proposition.** Can have small KL ($<0.23$) & arbitrarily bad variance estimate

$$\sigma_p^2 \geq c\sigma_q^2$$

$p$: Weibull, mean $m_p$

$q$: Weibull, mean $m_q$

**Proposition.** Can have small KL ($<0.9$) and arbitrarily bad mean estimate

$$(m_p - m_q)^2 \geq c\sigma_p^2$$

[20]  

[Baqué et al. 2017; Huggins et al. 2020]

[Huggins, Karsprzak, Campbell, Broderick 2020]
Is it just MFVB?

- Some KL values seen in practice:
  \( \sim 1 \) to \( \sim 70 \), \( 0.5 \) to \( 3 \)
  
- Take any constant \( c \)

**Proposition.** Can have small KL \(<0.23\) & arbitrarily bad variance estimate

\[
\sigma_p^2 \geq c \sigma_q^2
\]

\[p: \text{Weibull, mean } m_p\]

\[q: \text{Weibull, mean } m_q\]

**Proposition.** Can have small KL \(<0.9\) and arbitrarily bad mean estimate

\[
(m_p - m_q)^2 \geq c \sigma_p^2
\]

[Huggins, Karsprzak, Campbell, Broderick 2020]
Approximate Bayesian inference

Use $q^*$ to approximate $p(\cdot | y)$

Optimization

$q^* = \arg\min_{q \in Q} f(q(\cdot), p(\cdot | y))$

Variational Bayes

$q^* = \arg\min_{q \in Q} KL(q(\cdot) || p(\cdot | y))$

Mean-field variational Bayes

$q^* = \arg\min_{q \in Q_{MFVB}} KL(q(\cdot) || p(\cdot | y))$

Algorithm

Implementation

How deep is the issue?

Gaussian example was exact
Approximate Bayesian inference

Use $q^*$ to approximate $p(\cdot|y)$

Optimization

$q^* = \arg\min_{q \in Q} f(q(\cdot), p(\cdot|y))$

Variational Bayes

$q^* = \arg\min_{q \in Q} KL(q(\cdot)||p(\cdot|y))$

Mean-field variational Bayes

$q^* = \arg\min_{q \in Q_{MFVB}} KL(q(\cdot)||p(\cdot|y))$

How deep is the issue?

Algorithm

Implementation

Gaussian example was exact
Roadmap

• Bayes & Approximate Bayes review
• What is:
  • Variational Bayes (VB)
  • Mean-field variational Bayes (MFVB)
• Why use VB?
• When can we trust VB?
• Where do we go from here?
Roadmap

- Bayes & Approximate Bayes review
- What is:
  - Variational Bayes (VB)
  - Mean-field variational Bayes (MFVB)
- Why use VB?
- When can we trust VB?
- Where do we go from here?
We can fix VB uncertainty
We can fix VB uncertainty

We can fix VB uncertainty

Exact posterior

MFVB

[Bishop 2006]
We can fix VB uncertainty

- We provide: linear response variational Bayes (LRVB) [see also Opper, Winther 2003]
We can fix VB uncertainty

- We provide: linear response variational Bayes (LRVB) [see also Opper, Winther 2003]
- Procedure: VB (e.g. MFVB), then LRVB

[Giordano, Broderick, Jordan 2015, 2018]
We can fix VB uncertainty

- We provide: linear response variational Bayes (LRVB) [see also Opper, Winther 2003]
- Procedure: VB (e.g. MFVB), then LRVB
  - Perturbation ideas, statistical physics
We can fix VB uncertainty

- We provide: linear response variational Bayes (LRVB) [see also Opper, Winther 2003]
- Procedure: VB (e.g. MFVB), then LRVB
  - Perturbation ideas, statistical physics
  - Correction to VB:

\[ \text{Exact posterior} \]

[Giordano, Broderick, Jordan 2015, 2018]
We can fix VB uncertainty

- We provide: linear response variational Bayes (LRVB) [see also Opper, Winther 2003]
- Procedure: VB (e.g. MFVB), then LRVB
  - Perturbation ideas, statistical physics
  - Correction to VB: \[ \hat{\Sigma} = (I - VH)^{-1}V \]

[Giordano, Broderick, Jordan 2015, 2018]
We can fix VB uncertainty

- We provide: linear response variational Bayes (LRVB) [see also Opper, Winther 2003]
- Procedure: VB (e.g. MFVB), then LRVB
  - Perturbation ideas, statistical physics
  - Correction to VB: $\hat{\Sigma} = (I - VH)^{-1}V$

Exact posterior

computable from model with autodiff

[Giordano, Broderick, Jordan 2015, 2018]
We can fix VB uncertainty

- We provide: linear response variational Bayes (LRVB) [see also Opper, Winther 2003]
- Procedure: VB (e.g. MFVB), then LRVB
  - Perturbation ideas, statistical physics
  - Correction to VB: \[ \hat{\Sigma} = (I - VH)^{-1}V \]
- Computable from model with autodiff [Giordano, Broderick, Jordan 2015, 2018]
We can fix VB uncertainty

- We provide: linear response variational Bayes (LRVB) [see also Opper, Winther 2003]
- Procedure: VB (e.g. MFVB), then LRVB
- Perturbation ideas, statistical physics
- Correction to VB: \[ \hat{\Sigma} = (I - VH)^{-1} V \]

Exact posterior

\[ \theta_2 \]

\[ \theta_1 \]

computable from model with autodiff

[Giordano, Broderick, Jordan 2015, 2018]
We can fix VB uncertainty

- We provide: linear response variational Bayes (LRVB) [see also Opper, Winther 2003]
- Procedure: VB (e.g. MFVB), then LRVB
  - Perturbation ideas, statistical physics
  - Correction to VB: $\sum = (I - VH)^{-1}V$
  - Exact posterior
  - Exact for Gaussians
  - Computable from model with autodiff

$\Sigma = (I - VH)^{-1}V$

$\theta_2$

$\theta_1$

 Exact posterior

MFVB

LRVB

[Bishop 2006]

Standard deviations

$\text{MCMC (ground truth)}$

$\text{VB and LRVB}$
We can fix VB uncertainty

- We provide: linear response variational Bayes (LRVB) [see also Opper, Winther 2003]
- Procedure: VB (e.g. MFVB), then LRVB
  - Perturbation ideas, statistical physics
  - Correction to VB: $\hat{\Sigma} = (I - VH)^{-1} V$

• Exact for Gaussians
• Needs good posterior mean approximation in practice

computable from model with autodiff

[Giordano, Broderick, Jordan 2015, 2018]
Focus on data “core” for guarantees
Focus on data “core” for guarantees

• Observe: redundancies can exist even if data isn’t “tall”
Focus on data “core” for guarantees

- Observe: redundancies can exist even if data isn’t “tall”
Focus on data “core” for guarantees

- Observe: redundancies can exist even if data isn’t “tall”
Focus on data “core” for guarantees

- Observe: redundancies can exist even if data isn’t “tall”
Focus on data “core” for guarantees

- Observe: redundancies can exist even if data isn’t “tall”
- Coresets: pre-process data to get a smaller, weighted data set

[Bădoiu, Har-Peled, Indyk 2002; Agarwal et al 2005; Feldman & Langberg 2011]
Focus on data “core” for guarantees

- Observe: redundancies can exist even if data isn’t “tall”
- Coresets: pre-process data to get a smaller, weighted data set
- Theoretical guarantees on quality

[Bădoiu, Har-Peled, Indyk 2002; Agarwal et al 2005; Feldman & Langberg 2011]
Focus on data “core” for guarantees

- Observe: redundancies can exist even if data isn’t “tall”
- Coresets: pre-process data to get a smaller, weighted data set
- Theoretical guarantees on quality
- How to develop **coresets for Bayes?**

[Bădoiu, Har-Peled, Indyk 2002; Agarwal et al 2005; Feldman & Langberg 2011; Huggins, Campbell, Broderick 2016; Campbell, Broderick 2019; Campbell, Broderick 2018; Agrawal, Campbell, Huggins, Broderick 2019]
Focus on data “core” for guarantees

- Observe: redundancies can exist even if data isn’t “tall”
- Coresets: pre-process data to get a smaller, weighted data set
- Theoretical guarantees on quality
- How to develop **coresets for diverse tasks/geometries**?

[Bădoiu, Har-Peled, Indyk 2002; Agarwal et al 2005; Feldman & Langberg 2011; Huggins, Campbell, Broderick 2016; Campbell, Broderick 2019; Campbell, Broderick 2018; Agrawal, Campbell, Huggins, Broderick 2019]
Focus on data “core” for guarantees

- Observe: redundancies can exist even if data isn’t “tall”
- Coresets: pre-process data to get a smaller, weighted data set
- Theoretical guarantees on quality
- How to develop **coresets for diverse tasks/geometries**?
- Previous heuristics: data squashing, big data GPs

[Bădoiu, Har-Peled, Indyk 2002; Agarwal et al 2005; Feldman & Langberg 2011; DuMouchel et al 1999; Madigan et al 2002; Huggins, Campbell, Broderick 2016; Campbell, Broderick 2019; Campbell, Broderick 2018; Agrawal, Campbell, Huggins, Broderick 2019]
Focus on data “core” for guarantees

- Observe: redundancies can exist even if data isn’t “tall”
- Coresets: pre-process data to get a smaller, weighted data set

- Theoretical guarantees on quality
- How to develop coresets for diverse tasks/geometries?
- Previous heuristics: data squashing, big data GPs
- Compare to subsampling
Focus on data “core” for guarantees

- Observe: redundancies can exist even if data isn’t “tall”
- Coresets: pre-process data to get a smaller, weighted data set
- Theoretical guarantees on quality
- How to develop **coresets for diverse tasks/geometries**?
- Previous heuristics: data squashing, big data GPs
- Compare to subsampling

[Bădoiu, Har-Peled, Indyk 2002; Agarwal et al 2005; Feldman & Langberg 2011; DuMouchel et al 1999; Madigan et al 2002; Huggins, Campbell, Broderick 2016; Campbell, Broderick 2019; Campbell, Broderick 2018; Agrawal, Campbell, Huggins, Broderick 2019]
Focus on data “core” for guarantees

- Observe: redundancies can exist even if data isn’t “tall”
- Coresets: pre-process data to get a smaller, weighted data set

- Theoretical guarantees on quality
- How to develop coresets for diverse tasks/geometries?
- Previous heuristics: data squashing, big data GPs
- Compare to subsampling

[Bădoiu, Har-Peled, Indyk 2002; Agarwal et al 2005; Feldman & Langberg 2011; DuMouchel et al 1999; Madigan et al 2002; Huggins, Campbell, Broderick 2016; Campbell, Broderick 2019; Campbell, Broderick 2018; Agrawal, Campbell, Huggins, Broderick 2019]
Uniform subsampling

Benign

Malicious
Uniform subsampling
Uniform subsampling
Uniform subsampling

Benign

Malicious
Uniform subsampling
Uniform subsampling

- Might miss important data
Uniform subsampling

- Might miss important data
Uniform subsampling

- Benign
- Malicious

- Might miss important data
Uniform subsampling

- Might miss important data
Uniform subsampling

- Might miss important data
Uniform subsampling

- Might miss important data
Uniform subsampling

- Might miss important data
Uniform subsampling

- Might miss important data
Uniform subsampling

- Might miss important data
Uniform subsampling

- Might miss important data
Uniform subsampling

- Benign
- Malicious

- Might miss important data
Uniform subsampling

- Might miss important data
Uniform subsampling

- Might miss important data
Uniform subsampling

- Might miss important data
- Noisy estimates
Uniform subsampling

- Might miss important data
- Noisy estimates

$M = 10$

[Benign, Malicious]

[Campbell, Broderick 2018, 2019]
Uniform subsampling

- Might miss important data
- Noisy estimates

\[ M = 10 \]
\[ M = 100 \]
\[ M = 1000 \]

[Campbell, Broderick 2018, 2019]
Data summarization alternatives

Uniform subsampling

$M = 10$

$M = 100$

$M = 1000$

[Campbell, Broderick 2018, 2019]
Data summarization alternatives

Uniform subsampling

Importance sampling

\[ M = 10 \]  \[ M = 100 \]  \[ M = 1000 \]

[Campbell, Broderick 2018, 2019]
Data summarization alternatives

Uniform subsampling

Importance sampling

Bayesian/Hilbert coresets

$M = 10$

$M = 100$

$M = 1000$

[Campbell, Broderick 2018, 2019]
Reliable diagnostics
Reliable diagnostics

- ELBO or KL alone isn’t enough
Reliable diagnostics

- ELBO or KL alone isn’t enough
Reliable diagnostics

- ELBO or KL alone isn’t enough

![Graph showing ELBO and KL over iterations]
Reliable diagnostics

- ELBO or KL alone isn’t enough
- Instead: easy-to-compute bound on Wasserstein
- Wasserstein bounds error in posterior mean and variance

[Huggins, Kasprzak, Campbell, Broderick, 2020]
Reliable diagnostics

- ELBO or KL alone isn’t enough

- Instead: easy-to-compute bound on Wasserstein
  - Wasserstein bounds error in posterior mean and variance
  - Part of a validated workflow for VB

[Huggins, Kasprzak, Campbell, Broderick, 2020]
Reliable diagnostics

- ELBO or KL alone isn’t enough

- Instead: easy-to-compute bound on Wasserstein
  - Wasserstein bounds error in posterior mean and variance
  - Part of a validated workflow for VB
  - Builds on e.g. [Dieng et al 2017; Yao et al 2018] [Huggins, Kasprzak, Campbell, Broderick, 2020]
Reliable diagnostics

- ELBO or KL alone isn’t enough

- Instead: easy-to-compute bound on Wasserstein
  - Wasserstein bounds error in posterior mean and variance

- Part of a validated workflow for VB
- Builds on e.g. [Dieng et al 2017; Yao et al 2018]

- See also [Gorham, Mackey 2015, 2017; Chwialkowski, Strathmann, Gretton 2016; Jitkrittum et al 2017; Talts et al 2018, etc.]
Roadmap

- Bayes & Approximate Bayes review
- What is:
  - Variational Bayes (VB)
  - Mean-field variational Bayes (MFVB)
- Why use VB?
- When can we trust VB?
- Where do we go from here?
Bayesian inference

- Goals: good point estimates, uncertainty estimates
- Challenge: speed (compute, user), reliable inference
What to read next

Textbooks and Reviews


Our Experiments


References (2/6)


ESO/L. Calçada/M. Kornmesser. 16 October 2017, 16:00:00. Obtained from: https://commons.wikimedia.org/wiki/File:Artist%E2%80%99s_impression_of_merging_neutron_stars.jpg || Source: https://www.eso.org/public/images/eso1733a/ (Creative Commons Attribution 4.0 International License)

J. Herzog. 3 June 2016, 17:17:30. Obtained from: https://commons.wikimedia.org/wiki/File:Airbus_A350-941_F-WWCF_MSN002_ILA_Berlin_2016_17.jpg (Creative Commons Attribution 4.0 International License)
