Variational Bayes and beyond: Foundations of scalable Bayesian inference

Tamara Broderick
Associate Professor
MIT

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

[Nishiura et al. 2020; Flaxman et al. 2020; Dehning et al. 2020]
Bayesian inference
Bayesian inference

-Nishiura et al. 2020; Flaxman et al. 2020; Dehning et al. 2020-

-Gillon et al. 2017-
Grimm et al. 2018-

-Abbott et al. 2016a,b-

 [ESO/ L. Calçada/ M. Kornmesser 2017-]
Bayesian inference

[Stone et al 2014]

[Nishiura et al 2020; Flaxman et al 2020; Dehning et al 2020]

[Abbott et al 2016a,b]

[Gillon et al 2017; Grimm et al 2018]

[ESO/L. Calçada/M. Kornmesser 2017]
Bayesian inference

Nishiura et al. 2020; Flaxman et al. 2020; Dehning et al. 2020

Gillon et al. 2017; Grimm et al. 2018

Abbott et al. 2016a,b

ESO / L. Calçada / M. Kornmesser 2017

Stone et al. 2014

Woodard et al. 2017

[Abbott et al 2016a,b]

[ESO/
L. Calçada/
M. Kornmesser 2017]

[Stone et al 2014]
Bayesian inference
Bayesian inference

[Nishiura et al 2020; Flaxman et al 2020; Dehning et al 2020]

[Woodard et al 2017]

[amcharts.com 2016]

[Meager 2019,2020]

[Abbott et al 2016a,b]

[ESO/L. Calçada/M. Kornmesser 2017]

[Gillon et al 2017]

[Grimm et al 2018]

[Stone et al 2014]

[Chati, Balakrishnan et al 2020]
Bayesian inference
Bayesian inference

- Goals: good point estimates, uncertainty estimates
Bayesian inference

- Goals: good point estimates, uncertainty estimates
- More: interpretable, modular, expert info
Bayesian inference

- Goals: good point estimates, uncertainty estimates
- More: interpretable, modular, expert info
Bayesian inference

- Goals: good point estimates, uncertainty estimates
- More: interpretable, modular, expert info
- Challenge: speed (compute, user), reliable inference
Bayesian inference

- Goals: good point estimates, uncertainty estimates
- More: interpretable, modular, expert info
- Challenge: speed (compute, user), reliable inference
- Uncertainty doesn’t have to disappear in large data sets
Variational Bayes
Variational Bayes

- Modern problems: often large data, large dimensions
Variational Bayes

- Modern problems: often large data, large dimensions
- Variational Bayes can be very fast
Variational Bayes

- Modern problems: often large data, large dimensions
- Variational Bayes can be very fast

[Blei et al 2003]
Variational Bayes

- Modern problems: often large data, large dimensions
- Variational Bayes can be very fast

"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     MANAGERS
YORK      PLAN      WELFARE    NAMIPHY
OPERA     MONEY     MEN        STATE
THEATER   PROGRAMS  PERCENT    PRESIDENT
ACTRESS   GOVERNMENT CARE       ELEMENTARY
LOVE      CONGRESS  LIFE       HAITI

[Blei et al 2003]

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,” said President Randolph A. Hearst. “Every bit as important as our traditional areas of support in health, medical research, education and the social services,” he 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.
Variational Bayes

• Modern problems: often large data, large dimensions
• Variational Bayes can be very fast
Variational Bayes

- Modern problems: often large data, large dimensions
- Variational Bayes can be very fast

**Table:“Arts” “Budgets” “Children” “Education”**

<table>
<thead>
<tr>
<th></th>
<th>“Arts”</th>
<th>“Budgets”</th>
<th>“Children”</th>
<th>“Education”</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEW</td>
<td>MILLION</td>
<td>CHILDREN</td>
<td>SCHOOL</td>
<td></td>
</tr>
<tr>
<td>FILM</td>
<td>TAX</td>
<td>WOMEN</td>
<td>STUDENTS</td>
<td></td>
</tr>
<tr>
<td>SHOW</td>
<td>PROGRAM</td>
<td>PEOPLE</td>
<td>SCHOOLS</td>
<td></td>
</tr>
<tr>
<td>MUSIC</td>
<td>BUDGET</td>
<td>CHILD</td>
<td>EDUCATION</td>
<td></td>
</tr>
<tr>
<td>MOVIE</td>
<td>BILLION</td>
<td>YEARS</td>
<td>TEACHERS</td>
<td></td>
</tr>
<tr>
<td>PLAY</td>
<td>FEDERAL</td>
<td>FAMILIES</td>
<td>HIGH</td>
<td></td>
</tr>
<tr>
<td>MUSICAL</td>
<td>YEAR</td>
<td>WORK</td>
<td>PUBLIC</td>
<td></td>
</tr>
<tr>
<td>BEST</td>
<td>SPENDING</td>
<td>PARENTS</td>
<td>TEACHER</td>
<td></td>
</tr>
<tr>
<td>ACTOR</td>
<td>NEW</td>
<td>SAYS</td>
<td>BENNETT</td>
<td></td>
</tr>
<tr>
<td>FIRST</td>
<td>STATE</td>
<td>FAMILY</td>
<td>MANAGAT</td>
<td></td>
</tr>
<tr>
<td>YORK</td>
<td>PLAN</td>
<td>WELFARE</td>
<td>NAMPHYS</td>
<td></td>
</tr>
<tr>
<td>OPERA</td>
<td>MONEY</td>
<td>MEN</td>
<td>STATE</td>
<td></td>
</tr>
<tr>
<td>THEATER</td>
<td>PROGRAMS</td>
<td>PERCENT</td>
<td>PRESIDENT</td>
<td></td>
</tr>
<tr>
<td>ACTRESS</td>
<td>GOVERNMENT</td>
<td>CARE</td>
<td>ELEMENTARY</td>
<td></td>
</tr>
<tr>
<td>LOVE</td>
<td>CONGRESS</td>
<td>LIFE</td>
<td>HAITI</td>
<td></td>
</tr>
</tbody>
</table>

[Stegle et al 2010]

[Blei et al 2003]

[Airoldi et al 2008]

[Gershman et al 2014]

[Blei et al 2018]

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, after 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.
Variational Bayes

- Modern problems: often large data, large dimensions
- Variational Bayes can be very fast

```
<table>
<thead>
<tr>
<th>“Arts”</th>
<th>“Budgets”</th>
<th>“Children”</th>
<th>“Education”</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEW</td>
<td>MILLION</td>
<td>CHILDREN</td>
<td>SCHOOL</td>
</tr>
<tr>
<td>FILM</td>
<td>TAX</td>
<td>WOMEN</td>
<td>STUDENTS</td>
</tr>
<tr>
<td>SHOW</td>
<td>PROGRAM</td>
<td>PEOPLE</td>
<td>SCHOOLS</td>
</tr>
<tr>
<td>MUSIC</td>
<td>BUDGET</td>
<td>CHILD</td>
<td>EDUCATION</td>
</tr>
<tr>
<td>MOVIE</td>
<td>BILLION</td>
<td>YEARS</td>
<td>TEACHERS</td>
</tr>
<tr>
<td>PLAY</td>
<td>FEDERAL</td>
<td>FAMILIES</td>
<td>HIGH</td>
</tr>
<tr>
<td>MUSICAL</td>
<td>YEAR</td>
<td>WORK</td>
<td>PUBLIC</td>
</tr>
<tr>
<td>FIRST</td>
<td>SPENDING</td>
<td>PARENTS</td>
<td>TEACHER</td>
</tr>
<tr>
<td>ACTOR</td>
<td>NEW</td>
<td>SAYS</td>
<td>BENNETT</td>
</tr>
<tr>
<td>FIRST</td>
<td>STATE</td>
<td>FAMILY</td>
<td>MANAGERS</td>
</tr>
<tr>
<td>YORK</td>
<td>PLAN</td>
<td>WELFARE</td>
<td>NAMPHI</td>
</tr>
<tr>
<td>OPERA</td>
<td>MONEY</td>
<td>MEN</td>
<td>STATE</td>
</tr>
<tr>
<td>THEATER</td>
<td>PROGRAMS</td>
<td>PERCENT</td>
<td>PRESIDENT</td>
</tr>
<tr>
<td>ACTRESS</td>
<td>GOVERNMENT</td>
<td>CARE</td>
<td>ELEMENTARY</td>
</tr>
<tr>
<td>LOVE</td>
<td>CONGRESS</td>
<td>LIFE</td>
<td>HAITI</td>
</tr>
</tbody>
</table>
```

[Blei et al 2003]

[Stegle et al 2010]

[Xing et al 2004]

[Xing 2003]

[Airoldi et al 2008]

[Gershman et al 2014]

[Blei et al 2018]
Roadmap

• Bayes & Approximate Bayes review
Roadmap

• Bayes & Approximate Bayes review
• What is:
  • Variational Bayes (VB)
  • Mean-field variational Bayes (MFVB)
Roadmap

• Bayes & Approximate Bayes review
• What is:
  • Variational Bayes (VB)
  • Mean-field variational Bayes (MFVB)
• Why use VB?
Roadmap

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

$p(\theta)$
prior

parameters

Bayesian inference
Bayesian inference

\[ p(\theta) \]

prior

parameters
Bayesian inference

\[ p(y_{1:N} | \theta) p(\theta) \]

likelihood prior

parameters

\[ \theta \]
Bayesian inference

$$p(y_1:N|\theta)p(\theta)$$

likelihood prior

data parameters
Bayesian inference

\[ p(\theta | y_{1:N}) \propto p(y_{1:N} | \theta) p(\theta) \]

posterior  likelihood  prior

data  parameters
Bayesian inference

\[ p(\theta|y_{1:N}) \propto p(y_{1:N}|\theta)p(\theta) \]

posterior  likelihood  prior

Bayes Theorem
Bayesian inference

\[ p(\theta|y_{1:N}) \propto p(y_{1:N}|\theta)p(\theta) \]

posterior  likelihood  prior  

Bayes Theorem

data  parameters
Bayesian inference

\[ p(\theta \mid y_{1:N}) \propto \theta \cdot p(y_{1:N} \mid \theta) p(\theta) \]

posterior likelihood prior

1. Build a model: choose prior & choose likelihood
Bayesian inference

\[ p(\theta|y_{1:N}) \propto p(y_{1:N}|\theta)p(\theta) \]

posterior likelihood prior

1. Build a model: choose prior & choose likelihood
2. Compute the posterior
Bayesian inference

$p(\theta|y_{1:N}) \propto p(y_{1:N} | \theta)p(\theta)$

posterior \quad likelihood \quad prior

1. Build a model: choose prior & choose likelihood
2. Compute the posterior
3. Report a summary, e.g. posterior means and (co)variances
Bayesian inference

\[ p(\theta|y_{1:N}) \propto \theta \, p(y_{1:N}|\theta)p(\theta) \]

posterior likelihood prior

1. Build a model: choose prior & choose likelihood
2. Compute the posterior
3. Report a summary, e.g. posterior means and (co)variances
   - Why are steps 2 and 3 hard?
Bayesian inference

\[ p(\theta|y_{1:N}) \propto \theta \cdot p(y_{1:N} | \theta)p(\theta) \]

1. Build a model: choose prior & choose likelihood
2. Compute the posterior
3. Report a summary, e.g. posterior means and (co)variances
   - Why are steps 2 and 3 hard?
     - Typically no closed form
Bayesian inference

\[ p(\theta | y_{1:N}) \propto p(y_{1:N} | \theta) p(\theta) \]

posterior likelihood prior

1. Build a model: choose prior & choose likelihood
2. Compute the posterior
3. Report a summary, e.g. posterior means and (co)variances
   • Why are steps 2 and 3 hard?
   • Typically no closed form, high-dimensional integration
Bayesian inference

\[ p(\theta | y_{1:N}) = \frac{p(y_{1:N} | \theta) p(\theta)}{p(y_{1:N})} \]

1. Build a model: choose prior & choose likelihood
2. Compute the posterior
3. Report a summary, e.g. posterior means and (co)variances
   - Why are steps 2 and 3 hard?
     - Typically no closed form, high-dimensional integration
Bayesian inference

\[ p(\theta|y_{1:N}) = \frac{p(y_{1:N} | \theta)p(\theta)}{p(y_{1:N})} \]

posterior likelihood prior evidence

1. Build a model: choose prior & choose likelihood
2. Compute the posterior
3. Report a summary, e.g. posterior means and (co)variances
   • Why are steps 2 and 3 hard?
     • Typically no closed form, high-dimensional integration
Bayesian inference

\[ p(\theta|y_{1:N}) = \frac{p(y_{1:N}|\theta)p(\theta)}{\int p(y_{1:N}, \theta) d\theta} \]

1. Build a model: choose prior & choose likelihood
2. Compute the posterior
3. Report a summary, e.g. posterior means and (co)variances
   • Why are steps 2 and 3 hard?
     • Typically no closed form, high-dimensional integration
Approximate Bayesian Inference
Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC) [Bardenet, Doucet, Holmes 2017]
Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow

[Reference: Bardenet, Doucet, Holmes 2017]
Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow

Instead: an optimization approach

- Approximate posterior with $q^*$

[Bardenet, Doucet, Holmes 2017]
Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow

Instead: an optimization approach

- Approximate posterior with $q^*$

[Bardenet, Doucet, Holmes 2017]
Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow

Instead: an optimization approach

- Approximate posterior with $q^*$

[Bardenet, Doucet, Holmes 2017]
Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow

Instead: an optimization approach

- Approximate posterior with $q^*$

(Bardenet, Doucet, Holmes 2017)
Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow

Instead: an optimization approach

- Approximate posterior with $q^*$

[Bardenet, Doucet, Holmes 2017]
Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow

Instead: an optimization approach

- Approximate posterior with $q^*$
Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
- Eventually accurate but can be slow

Instead: an optimization approach

- Approximate posterior with \( q^* \)

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

[Bardenet, Doucet, Holmes 2017]
Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow

Instead: an optimization approach

- Approximate posterior with \( q^\star \)
  \[ q^\star = \arg\min_{q \in Q} f(q(\cdot), p(\cdot|y)) \]
Approximate Bayesian Inference

• Gold standard: Markov Chain Monte Carlo (MCMC)
  • Eventually accurate but can be slow

Instead: an optimization approach

• Approximate posterior with $q^*$

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

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow

Instead: an optimization approach

- Approximate posterior with $q^*$

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

[Bardenet, Doucet, Holmes 2017]
Approximate Bayesian Inference

• Gold standard: Markov Chain Monte Carlo (MCMC)
  • Eventually accurate but can be slow

Instead: an optimization approach

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

[Bardenet, Doucet, Holmes 2017]
Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow

Instead: an optimization approach

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

- Variational Bayes (VB): $f$ is Kullback-Leibler divergence
  \[ KL(q(\cdot)||p(\cdot|y)) \]

[Bardenet, Doucet, Holmes 2017]
Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
- Eventually accurate but can be slow

Instead: an optimization approach

- Approximate posterior with $q^*$
  
  $$q^* = \operatorname{argmin}_{q \in Q} f(q(\cdot), p(\cdot|y))$$

- Variational Bayes (VB): $f$ is Kullback-Leibler divergence
  
  $$KL(q(\cdot)||p(\cdot|y))$$

[Backen, Doucet, Holmes 2017]
Approximate Bayesian Inference

• Gold standard: Markov Chain Monte Carlo (MCMC)
  • Eventually accurate but can be slow

Instead: an optimization approach

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

• Variational Bayes (VB): \( f \) is Kullback-Leibler divergence
  \[
  KL(q(\cdot)||p(\cdot|y))
  \]

[Bardenet, Doucet, Holmes 2017]
Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow

Instead: an optimization approach

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

- Variational Bayes (VB): \( f \) is Kullback-Leibler divergence
  \[ KL(q(\cdot)||p(\cdot|y)) \]

[Bardenet, Doucet, Holmes 2017]
Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow

Instead: an optimization approach

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

- Variational Bayes (VB): $f$ is Kullback-Leibler divergence
  \[
  KL(q(\cdot)||p(\cdot|y))
  \]

- VB practical success

[Bardenet, Doucet, Holmes 2017]
Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow

Instead: an optimization approach

- Approximate posterior with $q^*$
  \[ q^* = \text{argmin}_{q \in Q} f(q(\cdot), p(\cdot | y)) \]

- Variational Bayes (VB): $f$ is Kullback-Leibler divergence
  \[ KL(q(\cdot) || p(\cdot | y)) \]

- VB practical success: point estimates and prediction

[Bardenet, Doucet, Holmes 2017]
Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow

Instead: an optimization approach

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

- Variational Bayes (VB): \( f \) is Kullback-Leibler divergence
  \[
  KL(q(\cdot) || p(\cdot | y))
  \]

- VB practical success: point estimates and prediction, fast

[4]

Bardenet, Doucet, Holmes 2017
Approximate Bayesian Inference

- Gold standard: Markov Chain Monte Carlo (MCMC)
  - Eventually accurate but can be slow

Instead: an optimization approach

- Approximate posterior with $q^*$
  $$q^* = \text{argmin}_{q \in Q} f(q(\cdot), p(\cdot | y))$$

- Variational Bayes (VB): $f$ is Kullback-Leibler divergence
  $$KL(q(\cdot) || p(\cdot | y))$$

- VB practical success: point estimates and prediction, fast, streaming, distributed (3.6M Wikipedia, 350K Nature)
Why KL?

- Variational Bayes

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

- Variational Bayes

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

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

- Variational Bayes

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

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

\[
= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta
\]
Why KL?

- Variational Bayes

\[ q^* = \operatorname{argmin}_{q \in \mathcal{Q}} \text{KL} \left( q(\cdot) \| p(\cdot | y) \right) \]

\[
\text{KL} \left( q(\cdot) \| p(\cdot | y) \right) := \int q(\theta) \log \frac{q(\theta)}{p(\theta | y)} d\theta
\]

\[ = \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta \]
Why KL?

- Variational Bayes

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

$$\text{KL} (q(\cdot) \| p(\cdot \mid y))$$

$$= \int q(\theta) \log \frac{q(\theta)}{p(\theta \mid y)} d\theta$$

$$= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta$$
Why KL?

- Variational Bayes

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

\[
KL (q(\cdot) || p(\cdot | y))
\]

\[ := \int q(\theta) \log \frac{q(\theta)}{p(\theta | y)} d\theta \]

\[ = \int q(\theta) \log \frac{q(\theta) p(y)}{p(\theta, y)} d\theta \]
Why KL?

- Variational Bayes
  
  \[ q^* = \text{argmin}_{q \in Q} \text{KL} (q(\cdot) \| p(\cdot|y)) \]

\[
\text{KL} (q(\cdot) \| p(\cdot|y)) := \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta \\
= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta,y)} d\theta = \int q(\theta) \left[ \log p(y) + \log \frac{q(\theta)}{p(\theta,y)} \right] d\theta
\]
Why KL?

- Variational Bayes

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

\[
\text{KL} (q(\cdot) \| p(\cdot \mid y)) := \int q(\theta) \log \frac{q(\theta)}{p(\theta \mid y)} d\theta
\]

\[
= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \int q(\theta) \left[ \log p(y) + \log \frac{q(\theta)}{p(\theta, y)} \right] d\theta
\]
Why KL?

- Variational Bayes

\[ q^* = \text{argmin}_{q \in Q} KL (q(\cdot) \| p(\cdot | y)) \]

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

\[
= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \int q(\theta) \left[ \log p(y) + \log \frac{q(\theta)}{p(\theta, y)} \right] d\theta
\]
Why KL?

- Variational Bayes

\[ q^* = \text{argmin}_{q \in Q} \text{KL} (q(\cdot) \| p(\cdot|y)) \]

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

\[
= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta,y)} d\theta = \int q(\theta) \left[ \log p(y) + \log \frac{q(\theta)}{p(\theta,y)} \right] d\theta
\]
Why KL?

- Variational Bayes

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

\[
\begin{align*}
\text{KL}(q(\cdot)\|p(\cdot|y)) &= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)} d\theta \\
&= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \int q(\theta) \left[ \log p(y) + \log \frac{q(\theta)}{p(\theta, y)} \right] d\theta
\end{align*}
\]
Why KL?

- Variational Bayes

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

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

\[
= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \int q(\theta) \left[ \log p(y) + \log \frac{q(\theta)}{p(\theta, y)} \right] d\theta
\]
Why KL?

- Variational Bayes

\[ q^* = \text{argmin}_{q \in Q} \text{KL} (q(\cdot) \| p(\cdot | y)) \]

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

\[
= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \int q(\theta) \left[ \log p(y) + \log \frac{q(\theta)}{p(\theta, y)} \right] d\theta
\]
Why KL?

- Variational Bayes

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

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

\[
= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \log p(y) + \int q(\theta) \log \frac{q(\theta)}{p(\theta, y)} d\theta
\]
Why KL?

- Variational Bayes

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

\[
\text{KL} (q(\cdot) \| p(\cdot | y)) := \int q(\theta) \log \frac{q(\theta)}{p(\theta | y)} d\theta \\
= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \log p(y) + \int q(\theta) \log \frac{q(\theta)}{p(\theta, y)} d\theta
\]
Why KL?

- Variational Bayes

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

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

\[
= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta, y)}{q(\theta)} d\theta
\]
Why KL?

- Variational Bayes

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

\[
\text{KL} (q(\cdot) \mid \mid p(\cdot | y)) := \int q(\theta) \log \frac{q(\theta)}{p(\theta | y)} d\theta
\]

\[
= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta, y)}{q(\theta)} d\theta
\]
Why KL?

- Variational Bayes

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

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

\[
= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta, y)}{q(\theta)} d\theta
\]
Why KL?

- Variational Bayes

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

\[
\text{KL} (q(\cdot) || p(\cdot | y)) := \int q(\theta) \log \frac{q(\theta)}{p(\theta | y)} d\theta \\
= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta, y)}{q(\theta)} d\theta
\]
Why KL?

- Variational Bayes

\[ q^* = \text{argmin}_{q \in Q} \text{KL} (q(\cdot) \| p(\cdot | y)) \]

\[
\text{KL}(q(\cdot) \| p(\cdot | y)) := \int q(\theta) \log \frac{q(\theta)}{p(\theta | y)} d\theta \\
= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta, y)}{q(\theta)} d\theta
\]

“Evidence lower bound” (ELBO)
Why KL?

- Variational Bayes

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

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

\[ = \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta, y)}{q(\theta)} d\theta \]

"Evidence lower bound" (ELBO)
Why KL?

- Variational Bayes
  
  \[ q^* = \text{argmin}_{q \in Q} \text{KL} \left( q(\cdot) \| p(\cdot | y) \right) \]

  \[
  \text{KL} \left( q(\cdot) \| p(\cdot | y) \right) \\
  \quad := \int q(\theta) \log \frac{q(\theta)}{p(\theta | y)} d\theta \\
  \quad = \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta, y)}{q(\theta)} d\theta 
  \]

- Exercise: Show \( \text{KL} \geq 0 \) \[ \text{[Bishop 2006, Sec 1.6.1]} \]

“Evidence lower bound” (ELBO)
Why KL?

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

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

  \[
  = \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta,y)} d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta,y)}{q(\theta)} d\theta
  \]

- Exercise: Show \( \text{KL} \geq 0 \) \[\text{Bishop 2006, Sec 1.6.1}\]
- \( \text{KL} \geq 0 \Rightarrow \log p(y) \geq \text{ELBO} \)
Why KL?

- Variational Bayes

\[ q^* = \text{argmin}_{q \in Q} \text{KL} (q(\cdot)||p(\cdot|y)) \]

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

\[
= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta,y)} d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta,y)}{q(\theta)} d\theta
\]

- Exercise: Show \( \text{KL} \geq 0 \) \[\text{Bishop 2006, Sec 1.6.1}\]

- \( \text{KL} \geq 0 \Rightarrow \log p(y) \geq \text{ELBO} \)

- \( q^* = \text{argmax}_{q \in Q} \text{ELBO}(q) \)
Why KL?

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

\[
\text{KL} (q(\cdot) \| p(\cdot | y)) = \int q(\theta) \log \frac{q(\theta)}{p(\theta | y)} d\theta = \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta, y)}{q(\theta)} d\theta
\]

- Exercise: Show \( \text{KL} \geq 0 \) [Bishop 2006, Sec 1.6.1]
- \( \text{KL} \geq 0 \Rightarrow \log p(y) \geq \text{ELBO} \)
- \( q^* = \arg\max_{q \in Q} \text{ELBO}(q) \)
- Why KL (in this direction)?
Variational Bayes

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

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

Choose “NICE” distributions
Variational Bayes

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

Choose “NICE” distributions
Variational Bayes

\[ q^* = \operatorname{argmin}_{q \in Q} \text{KL}(q(\cdot) \Vert p(\cdot | y)) \]

Choose “NICE” distributions
Variational Bayes

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

Choose “NICE” distributions

- Mean-field variational Bayes (MFVB)

\[ Q_{MFVB} := \left\{ q : q(\theta) = \prod_{j=1}^{J} q_j(\theta_j) \right\} \]
Variational Bayes

\[ q^* = \operatorname{argmin}_{q \in Q} \operatorname{KL}(q(\cdot)||p(\cdot|y)) \]

Choose “NICE” distributions

• Mean-field variational Bayes (MFVB)

\[ Q_{MFVB} := \left\{ q : q(\theta) = \prod_{j=1}^{J} q_j(\theta_j) \right\} \]

• Often also exponential family
Variational Bayes

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

Choose “NICE” distributions

- Mean-field variational Bayes (MFVB)

\[ Q_{MFVB} := \left\{ q : q(\theta) = \prod_{j=1}^{J} q_j(\theta_j) \right\} \]

- Often also exponential family
- *Not* a modeling assumption
Variational Bayes

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

Choose “NICE” distributions

- Mean-field variational Bayes (MFVB)

\[ Q_{MFVB} := \left\{ q : q(\theta) = \prod_{j=1}^{J} q_j(\theta_j) \right\} \]

- Often also exponential family
- Not a modeling assumption

[Bishop 2006]
Variational Bayes

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

Choose “NICE” distributions

- Mean-field variational Bayes (MFVB)

\[ Q_{MFVB} := \left\{ q : q(\theta) = \prod_{j=1}^{J} q_j(\theta_j) \right\} \]

- Often also exponential family
- Not a modeling assumption

Now we have an optimization problem; how to solve it?
Variational Bayes

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

Choose “NICE” distributions

- Mean-field variational Bayes (MFVB)
  
  \[ Q_{MFVB} := \left\{ q : q(\theta) = \prod_{j=1}^{J} q_j(\theta_j) \right\} \]

- Often also exponential family
- *Not* a modeling assumption

Now we have an optimization problem; how to solve it?

- *One* option: Coordinate descent in \(q_1, \ldots, q_J\)

[Bishop 2006]
Approximate Bayesian inference
Approximate Bayesian inference

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

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

Optimization

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

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

Optimization

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

Variational Bayes

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

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

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

Variational Bayes

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

Mean-field variational Bayes

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

• Coordinate descent
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

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

- Coordinate descent
- Stochastic variational inference (SVI) [Hoffman et al 2013]
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

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

Mean-field variational Bayes

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

- Coordinate descent
- Stochastic variational inference (SVI) [Hoffman et al. 2013]
- Automatic differentiation variational inference (ADVI) [Kucukelbir et al. 2015, 2017]
Approximate Bayesian inference

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

Optimization

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

Variational Bayes

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

Mean-field variational Bayes

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

- Coordinate descent
- Stochastic variational inference (SVI) [Hoffman et al 2013]
- Automatic differentiation variational inference (ADVI) [Kucukelbir et al 2015, 2017]
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

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

- Coordinate descent
- Stochastic variational inference (SVI) [Hoffman et al 2013]
- Automatic differentiation variational inference (ADVI) [Kucukelbir et al 2015, 2017]
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

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

- Coordinate descent
- Stochastic variational inference (SVI) [Hoffman et al 2013]
- Automatic differentiation variational inference (ADVI) [Kucukelbir et al 2015, 2017]
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?
**What to read next**

**Textbooks and Reviews**

**Our Experiments**
References

Full references at end of final slides