Variational Bayes and beyond: Bayesian inference for big data

Tamara Broderick
Associate Professor,
Electrical Engineering & Computer Science
MIT

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

[Gillon et al 2017]

[ESO/ L. Calçada/ M. Kornmesser 2017] [Abbott et al 2016a,b]

[Stone et al 2014]
Bayesian inference

[Image 1: Bayesian inference diagram

[Image 2: Image with labels

[ESO/L. Calçada/M. Kornmesser 2017] [Abbott et al 2016a,b]

[Image 3: Image with labels

[Stone et al 2014]

[Woodard et al 2017]
Bayesian inference

[Woodard et al 2017]

[ESO/L. Calçada/M. Kornmesser 2017] [Abbott et al 2016a,b]

[amcharts.com 2016] [Meager 2018a,b]

[Stone et al 2014]
Bayesian inference
Bayesian inference
Bayesian inference

- Goals: good point estimates, uncertainty estimates
Bayesian inference

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

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

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

- Goals: good point estimates, uncertainty estimates
- More: interpretable, flexible, 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

<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>BEST</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>MANIGAT</td>
</tr>
<tr>
<td>YORK</td>
<td>PLAN</td>
<td>WELFARE</td>
<td>NAMPHY</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>

[2]

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,” a director said. The Metropolitan Opera, Artistic Director of the New York Philharmonic and Juilliard School, 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

[Blei et al 2003]
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>BEST</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>MANIGAT</td>
</tr>
<tr>
<td>YORK</td>
<td>PLAN</td>
<td>WELFARE</td>
<td>NAMPHY</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]

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.

[Airoldi et al 2008]  [Gershman et al 2014]
Variational Bayes

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

Variational Bayes

<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>BEST</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>MANIGAT</td>
</tr>
<tr>
<td>YORK</td>
<td>PLAN</td>
<td>WELFARE</td>
<td>NAMPHY</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]

[2]

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.

[Airoldi et al 2008]

[Gershman et al 2014]

[Blei et al 2018]
Variational Bayes

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

[Variational Bayes  

[Blei et al 2003]  

Modern problems: often large data, large dimensions

Variational Bayes can be very fast

[Blei et al 2003]  

Variational Bayes can be very fast

[Blei et al 2003]  

Variational Bayes can be very fast

[Blei et al 2003]
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 MFVB?
Roadmap

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

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

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

Bayesian inference

$p(\theta)$

prior

parameters
Bayesian inference

\[ p(\theta) \]

prior

\[ \theta \]
Bayesian inference

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

likelihood prior
Bayesian inference

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

likelihood  prior

parameters  data
Bayesian inference

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

posterior likelihood prior
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
Bayesian inference

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

1. Build a model: choose prior & choose likelihood

Bayes Theorem
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

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

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

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

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

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

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

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

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 \hspace{1cm} likelihood \hspace{1cm} prior \hspace{1cm} evidence

Bayes Theorem

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

[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 $q^*$

[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 $q^*$

[Bardenet, Doucet, Holmes 2017]
Approximate Bayesian Inference

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

\[
p(\theta | y) \rightarrow q^{*}(\theta)
\]

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

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

[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 $q^*$

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

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

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

Instead: an optimization approach

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

Variational Bayes (VB): $f$ is Kullback-Leibler divergence

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

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

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

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

- VB practical success: point estimates and prediction

[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

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

- 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)

[Broderick, Boyd, Wibisono, Wilson, Jordan 2013]
[Bardenet, Doucet, Holmes 2017]
Why KL?

- Variational Bayes

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

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

\[
\text{KL} (q(\cdot) \| p(\cdot | y)) = \int q(\theta) \log \frac{q(\theta)}{p(\theta | y)} d\theta = \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta
\]
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
  \[ q^* = \operatorname{argmin}_{q \in Q} \text{KL} \left( q(\cdot) \| p(\cdot | y) \right) \]

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

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

- Variational Bayes

\[ q^* = \underset{q \in Q}{\text{argmin}} \, KL \left( q(\cdot) \left\| p(\cdot | y) \right. \right) \]

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

- Variational Bayes

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

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

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

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

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

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

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

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

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

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

- Exercise: Show \( \text{KL} \geq 0 \) \hspace{1em} [Bishop 2006, Sec 1.6.1]
- \( \text{KL} \geq 0 \Rightarrow \log p(y) \geq \text{ELBO} \)
- \( q^* = \arg \max_{q \in Q} \text{ELBO}(q) \)

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

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

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

Choose “NICE” distributions
Variational Bayes

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

Choose “NICE” distributions
Variational Bayes

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

Choose “NICE” distributions
Variational Bayes

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

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

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

Choose “NICE” distributions

• Mean-field variational Bayes (MFVB)

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

• 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
- \emph{Not} a modeling assumption
Variational Bayes

Choose “NICE” distributions

- Mean-field variational Bayes (MFVB)

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

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

- Often also exponential family
- Not a modeling assumption

[Bishop 2006]
Variational Bayes

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?

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

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

Choose “NICE” distributions

- Mean-field variational Bayes (MFVB)

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

- 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

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

Variational Bayes

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

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

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

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

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

**Optimization**

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

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

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

- 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 MFVB?
• When can we trust MFVB?
• Where do we go from here?
Roadmap

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

- Catalogued midge wing lengths (mm) \( y = (y_1, \ldots, y_N) \)
Midge wing length

• Catalogued midge wing lengths (mm) \( y = (y_1, \ldots, y_N) \)

• Model:

\[
p(y|\theta) : \quad y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \quad n = 1, \ldots, N
\]
Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$
- Parameters of interest: population mean and variance $\theta = (\mu, \sigma^2)$
- Model:
  $$p(y|\theta) : \quad y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \quad n = 1, \ldots, N$$

[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]
Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$
- Parameters of interest: population mean and variance $\theta = (\mu, \sigma^2)$
- Model:
  
  $$p(y|\theta) : \quad y_n \sim^{iid} \mathcal{N} (\mu, \sigma^2), \quad n = 1, \ldots, N$$
  
  $$p(\theta) : \quad (\sigma^2)^{-1} \sim \text{Gamma}(a_0, b_0)$$
  
  $$\mu|\sigma^2 \sim \mathcal{N} (\mu_0, \lambda_0 \sigma^2)$$
Midge wing length

- Catalogued midge wing lengths (mm) \( y = (y_1, \ldots, y_N) \)
- Parameters of interest: population mean and variance
- Model (conjugate prior):
  \[
  p(y|\theta) : \quad y_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \quad n = 1, \ldots, N
  \]
  \[
  p(\theta) : \quad (\sigma^2)^{-1} \sim \text{Gamma}(a_0, b_0)
  \]
  \[
  \mu|\sigma^2 \sim \mathcal{N}(\mu_0, \lambda_0\sigma^2)
  \]

\[\theta = (\mu, \sigma^2)\]
Midge wing length

• Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$

• Parameters of interest: population mean and variance

• Model (conjugate prior): [Exercise: find the posterior]

\[ p(y|\theta) : \ y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \quad n = 1, \ldots, N \]

\[ p(\theta) : \ (\sigma^2)^{-1} \sim \text{Gamma}(a_0, b_0) \]

\[ \mu | \sigma^2 \sim \mathcal{N}(\mu_0, \lambda_0 \sigma^2) \]
Midge wing length

- Catalogued midge wing lengths (mm)  \( y = (y_1, \ldots, y_N) \)
- Parameters of interest: population mean and variance
- Model (conjugate prior): \( p(y|\theta) \)  
  \[
  y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \quad n = 1, \ldots, N
  \]
  \[
  p(\theta) : \quad (\sigma^2)^{-1} \sim \text{Gamma}(a_0, b_0)
  \]
  \[
  \mu | \sigma^2 \sim \mathcal{N}(\mu_0, \lambda_0 \sigma^2)
  \]

[Exercise: find the posterior]
Midge wing length

- Catalogued midge wing lengths (mm) \( y = (y_1, \ldots, y_N) \)
- Parameters of interest: population mean and precision \( \theta = (\mu, \tau) \)
- Model (conjugate prior): [Exercise: find the posterior]

\[
p(y|\theta) : \quad y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2), \quad n = 1, \ldots, N
p(\theta) : \quad (\sigma^2)^{-1} \sim \text{Gamma}(a_0, b_0)
\mu|\sigma^2 \sim \mathcal{N}(\mu_0, \lambda_0\sigma^2)
\]

[CSIRO 2004]
Midge wing length

- Catalogued midge wing lengths (mm) \( y = (y_1, \ldots, y_N) \)
- Parameters of interest: population mean and precision
- Model (conjugate prior): [Exercise: find the posterior]

\[
p(y|\theta) : \quad y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2), \quad n = 1, \ldots, N
\]

\[
p(\theta) : \quad (\sigma^2)^{-1} \sim \text{Gamma}(a_0, b_0)
\]

\[
\mu | \sigma^2 \sim \mathcal{N}(\mu_0, \lambda_0 \sigma^2)
\]

[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]
Midge wing length

- Catalogued midge wing lengths (mm) \( y = (y_1, \ldots, y_N) \)
- Parameters of interest: population mean and precision
- Model (conjugate prior): [Exercise: find the posterior]
  \[
p(y|\theta) : \quad y_n \overset{iid}{\sim} \mathcal{N}(\mu, \tau^{-1}), \quad n = 1, \ldots, N
\]
  \[
p(\theta) : \quad \tau \sim \text{Gamma}(a_0, b_0)
\]
  \[
\mu | \tau \sim \mathcal{N}(\mu_0, (\rho_0 \tau)^{-1})
\]
Midge wing length

- Catalogued midge wing lengths (mm) \( y = (y_1, \ldots, y_N) \)
- Parameters of interest: population mean and precision \( \theta = (\mu, \tau) \)
- Model (conjugate prior): [Exercise: find the posterior]
  \[
  p(y|\theta) : \quad y_n \overset{\text{iid}}{\sim} N(\mu, \tau^{-1}), \quad n = 1, \ldots, N \\
  p(\theta) : \quad \tau \sim \text{Gamma}(a_0, b_0) \\
  \mu|\tau \sim N(\mu_0, (\rho_0\tau)^{-1})
  \]
Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$
- Parameters of interest: population mean and precision
- Model (conjugate prior): $\theta = (\mu, \tau)$

\[ p(y|\theta) : \ y_n \overset{iid}{\sim} \mathcal{N}(\mu, \tau^{-1}), \quad n = 1, \ldots, N \]
\[ p(\theta) : \ \tau \sim \text{Gamma}(a_0, b_0) \]
\[ \mu | \tau \sim \mathcal{N}(\mu_0, (\rho_0 \tau)^{-1}) \]

Exercise: find the posterior

Exercise: check $p(\mu, \tau|y) \neq f_1(\mu, y)f_2(\tau, y)$
Midge wing length

- Catalogued midge wing lengths (mm) \( y = (y_1, \ldots, y_N) \)
- Parameters of interest: population mean and precision \( \theta = (\mu, \tau) \)
- Model (conjugate prior): [Exercise: find the posterior]
  \[
p(y|\theta) : \quad y_n \sim_{\text{iid}} \mathcal{N}(\mu, \tau^{-1}), \quad n = 1, \ldots, N
\]
  \[
p(\theta) : \quad \tau \sim \text{Gamma}(a_0, b_0)
\]
  \[
\mu|\tau \sim \mathcal{N}(\mu_0, (\rho_0\tau)^{-1})
\]
- Exercise: check \( p(\mu, \tau|y) \neq f_1(\mu, y)f_2(\tau, y) \)
- MFVB approximation:
  \[
  q^*(\mu, \tau) = q^*_\mu(\mu)q^*_\tau(\tau) = \arg\min_{q \in Q_{\text{MFVB}}} KL(q(\cdot)||p(\cdot|y))
  \]

[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]
Midge wing length

- Catalogued midge wing lengths (mm) $y = (y_1, \ldots, y_N)$
- Parameters of interest: population mean and precision $\theta = (\mu, \tau)$
- Model (conjugate prior): [Exercise: find the posterior]
  
  $p(y|\theta) : \quad y_n \sim^{iid} \mathcal{N}(\mu, \tau^{-1}), \quad n = 1, \ldots, N$

  $p(\theta) : \quad \tau \sim \text{Gamma}(a_0, b_0)$

  $\mu | \tau \sim \mathcal{N}(\mu_0, (\rho_0 \tau)^{-1})$

- Exercise: check $p(\mu, \tau | y) \neq f_1(\mu, y) f_2(\tau, y)$
- MFVB approximation:

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

- Coordinate descent [Exercise: derive this] [Bishop 2006, Sec 10.1.3]
Midge wing length

- Catalogued midge wing lengths (mm) \( y = (y_1, \ldots, y_N) \)
- Parameters of interest: population mean and precision \( \theta = (\mu, \tau) \)
- Model (conjugate prior): [Exercise: find the posterior]
  \[
p(y|\theta) : \quad y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \tau^{-1}), \quad n = 1, \ldots, N
  
p(\theta) : \quad \tau \sim \text{Gamma}(a_0, b_0) \\
  \mu | \tau \sim \mathcal{N}(\mu_0, (\rho_0 \tau)^{-1})
\]
- Exercise: check \( p(\mu, \tau|y) \neq f_1(\mu, y) f_2(\tau, y) \)
- MFVB approximation:
  \[
  q^*(\mu, \tau) = q^*_\mu(\mu) q^*_\tau(\tau) = \text{argmin}_{q \in Q_{\text{MFVB}}} KL(q(\cdot) || p(\cdot|y))
  \]
  \[
  q^*_\mu(\mu) = \mathcal{N}(\mu|\mu_N, \rho_N^{-1}) \quad q^*_\tau(\tau) = \text{Gamma}(\tau|a_N, b_N)
  \]
- Coordinate descent [Exercise: derive this] [Bishop 2006, Sec 10.1.3]
Midge wing length

- Catalogued midge wing lengths (mm) \( y = (y_1, \ldots, y_N) \)
- Parameters of interest: population mean and precision \( \theta = (\mu, \tau) \)
- Model (conjugate prior): [Exercise: find the posterior]
  \[
p(y|\theta) : \quad y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \tau^{-1}), \quad n = 1, \ldots, N
\]
  \[
p(\theta) : \quad \tau \sim \text{Gamma}(a_0, b_0)
\]
  \[
\mu|\tau \sim \mathcal{N}(\mu_0, (\rho_0 \tau)^{-1})
\]
- Exercise: check \( p(\mu, \tau|y) \neq f_1(\mu, y)f_2(\tau, y) \)
- MFVB approximation:
  \[
  q^*(\mu, \tau) = q^*_\mu(\mu)q^*_\tau(\tau) = \arg\min_{q \in Q_{\text{MFVB}}} KL(q(\cdot)||p(\cdot|y))
\]
- Coordinate descent [Exercise: derive this] [Bishop 2006, Sec 10.1.3]
  \[
  q^*_\mu(\mu) = \mathcal{N}(\mu|\mu_N, \rho_N^{-1}) \quad q^*_\tau(\tau) = \text{Gamma}(\tau|a_N, b_N)
  \]
  “variational parameters”

[CSIRO 2004; Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]
Midge wing length

- Catalogued midge wing lengths (mm) \( y = (y_1, \ldots, y_N) \)
- Parameters of interest: population mean and precision \( \theta = (\mu, \tau) \)
- Model (conjugate prior): \[ \text{[Exercise: find the posterior]} \]
  \[
p(y|\theta) : \quad y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \tau^{-1}), \quad n = 1, \ldots, N
\]
  \[
p(\theta) : \quad \tau \sim \text{Gamma}(a_0, b_0)
\]
  \[
\mu | \tau \sim \mathcal{N}(\mu_0, (\rho_0 \tau)^{-1})
\]
- Exercise: check \( p(\mu, \tau|y) \neq f_1(\mu, y) f_2(\tau, y) \)
- MFVB approximation:
  \[
  q^*(\mu, \tau) = q^*_\mu(\mu)q^*_\tau(\tau) = \arg\min_{q \in Q_{MFVB}} KL(q(\cdot)||p(\cdot|y))
  \]
- Coordinate descent \[ \text{[Exercise: derive this]} \] \[ \text{[Bishop 2006, Sec 10.1.3]} \]
  \[
  q^*_\mu(\mu) = \mathcal{N}(\mu|\mu_N, \rho_N^{-1}) \quad q^*_\tau(\tau) = \text{Gamma}(\tau|a_N, b_N)
  \]
- Iterate: \( (\mu_N, \rho_N) = f(a_N, b_N) \)
  \[
  (a_N, b_N) = g(\mu_N, \rho_N)
  \]

[Hoff 2009; Grogan, Wirth 1981; MacKay 2003; Bishop 2006]
Midge wing length approximation

exact posterior

[Bishop 2006]
Midge wing length approximation

exact posterior

\[ \tau \]

\[ \mu \]
Midge wing length approximation

exact posterior

[Bishop 2006]
Midge wing length approximation

exact posterior

[Bishop 2006]
Microcredit Experiment
Microcredit Experiment

• Simplified from Meager (2018a)
Microcredit Experiment

• Simplified from Meager (2018a)

• $K = 7$ microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
Microcredit Experiment

• Simplified from Meager (2018a)

• $K = 7$ microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)

• $N_k$ businesses in $k$th site (~900 to ~17K)
Microcredit Experiment

• Simplified from Meager (2018a)

• $K = 7$ microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)

• $N_k$ businesses in $k$th site (~900 to ~17K)
Microcredit Experiment

- Simplified from Meager (2018a)
- $K = 7$ 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:
Microcredit Experiment

• Simplified from Meager (2018a)
• $K = 7$ microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
• $N_k$ businesses in $k$th site ($\sim$900 to $\sim$17K)
• Profit of $n$th business at $k$th site:

$$y_{kn} \sim \text{N} (\mu_k + T_{kn}, \sigma_k^2)$$
Microcredit Experiment

- Simplified from Meager (2018a)
- $K = 7$ 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:

$$y_{kn} \overset{indep}{\sim} N(\mu_k + T_{kn}, \sigma_k^2)$$
Microcredit Experiment

- Simplified from Meager (2018a)
- \( K = 7 \) 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:

\[ y_{kn} \overset{indep}{\sim} \mathcal{N}(\mu_k, \sigma_k^2) \]
Microcredit Experiment

• Simplified from Meager (2018a)
• $K = 7$ microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
• $N_k$ businesses in $k$th site ($\sim 900$ to $\sim 17K$)
• Profit of $n$th business at $k$th site:

$$y_{kn} \overset{indep}{\sim} \mathcal{N}(\mu_k + T_{kn} \tau_k, \sigma_k^2)$$
Microcredit Experiment

- Simplified from Meager (2018a)
- $K = 7$ 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:

$$y_{kn} \sim \text{indep } \mathcal{N}(\mu_k + T_{kn} \tau_k, \sigma_k^2)$$

1 if microcredit
Microcredit Experiment

- Simplified from Meager (2018a)
- $K = 7$ 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:

$$y_{kn} \overset{indep}{\sim} \mathcal{N}(\mu_k + T_{kn}\tau_k, \sigma_k^2)$$
Microcredit Experiment

- Simplified from Meager (2018a)
- $K = 7$ 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:

\[ y_{kn} \sim \text{indep } \mathcal{N}(\mu_k + T_{kn} \tau_k, \sigma_k^2) \]

1 if microcredit
Microcredit Experiment

- Simplified from Meager (2018a)
- $K = 7$ 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:

$$y_{kn} \sim \text{indep }\mathcal{N}(\mu_k + T_{kn}\tau_k, \sigma_k^2)$$
Microcredit Experiment

- Simplified from Meager (2018a)
- $K = 7$ microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- $N_k$ businesses in $k$th site ($\sim 900$ to $\sim 17K$)
- Profit of $n$th business at $k$th site:
  
  $$y_{kn} \overset{indep}{\sim} \mathcal{N}(\mu_k + T_{kn} \tau_k, \sigma_k^2)$$

- Priors and hyperpriors:
  
  $1$ if microcredit
Microcredit Experiment

- Simplified from Meager (2018a)
- $K = 7$ microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- $N_k$ businesses in $k$th site ($\sim$900 to $\sim$17K)
- Profit of $n$th business at $k$th site:
  \[ y_{kn} \sim \text{iid } \mathcal{N}(\mu_k + T_{kn}\tau_k, \sigma_k^2) \]

  - Priors and hyperpriors:
    \[
    \left( \begin{array}{c}
    \mu_k \\
    \tau_k
    \end{array} \right) \sim \mathcal{N}\left( \left( \begin{array}{c}
    \mu \\
    \tau
    \end{array} \right), C \right)
    \]

  1 if microcredit
Microcredit Experiment

- Simplified from Meager (2018a)
- $K = 7$ 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:
  \[ y_{kn} \overset{\text{iid}}{\sim} \mathcal{N}(\mu_k + T_{kn}\tau_k, \sigma_k^2) \]
- Priors and hyperpriors:
  \[
  \begin{pmatrix}
    \mu_k \\
    \tau_k
  \end{pmatrix}
  \overset{iid}{\sim}
  \mathcal{N}
  \left(
    \begin{pmatrix}
      \mu \\
      \tau
    \end{pmatrix},
    C
  \right)
  \]
  \[ \sigma_k^{-2} \overset{iid}{\sim} \Gamma(a, b) \]
Microcredit Experiment

- Simplified from Meager (2018a)
- $K = 7$ 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:

  \[ y_{kn} \overset{\text{iid}}{\sim} \mathcal{N}(\mu_k + T_{kn} \tau_k, \sigma_k^2) \]

- Priors and hyperpriors:

  \[
  \begin{aligned}
  \begin{pmatrix}
  \mu_k \\
  \tau_k
  \end{pmatrix} & \overset{\text{iid}}{\sim} \mathcal{N}\left(\begin{pmatrix}
  \mu \\
  \tau
  \end{pmatrix}, C\right) \\
  \mu & \overset{\text{iid}}{\sim} \mathcal{N}\left(\begin{pmatrix}
  \mu_0 \\
  \tau_0
  \end{pmatrix}, \Lambda^{-1}\right) \\
  \sigma_k^{-2} & \overset{\text{iid}}{\sim} \Gamma(a, b) \\
  C & \sim \text{Sep&LKJ} (\eta, c, d)
  \end{aligned}
  \]
Microcredit

MFVB: Do we need to check the output?
Microcredit

MFVB: How will we know if it’s working?
Microcredit

[Giordano, Broderick, Meager, Huggins, Jordan 2016]
Microcredit

- One set of 2500 MCMC draws: 45 minutes

[Giordano, Broderick, Meager, Huggins, Jordan 2016]
Microcredit

- **One set** of 2500 MCMC draws: **45 minutes**
- **MFVB** optimization: **<1 min**

[Giordano, Broderick, Meager, Huggins, Jordan 2016]
Microcredit

- **One set** of 2500 MCMC draws: 45 minutes
- MFVB optimization: <1 min

[Giordano, Broderick, Meager, Huggins, Jordan 2016]
**Microcredit**

- *One set of 2500 MCMC draws:* 45 minutes
- MFVB optimization: <1 min

**Criteo Online Ads Experiment**

- Click-through conversion prediction
- Q: Will a customer (e.g.) buy a product after clicking?

[Giordano, Broderick, Meager, Huggins, Jordan 2016]
Microcredit

- **One set of 2500 MCMC draws:** 45 minutes
- **MFVB optimization:** <1 min

Criteo Online Ads Experiment

- Click-through conversion prediction
- Q: Will a customer (e.g.) buy a product after clicking?
- Q: How predictive of conversion are different features?

[Giordano, Broderick, Meager, Huggins, Jordan 2016]
Microcredit

- One set of 2500 MCMC draws: 45 minutes
- MFVB optimization: <1 min

Criteo Online Ads Experiment

- Click-through conversion prediction
- Q: Will a customer (e.g.) buy a product after clicking?
- Q: How predictive of conversion are different features?
- Logistic GLMM

[Giordano, Broderick, Meager, Huggins, Jordan 2016; Giordano, Broderick, Jordan 2018]
Microcredit

• One set of 2500 MCMC draws: 45 minutes
• MFVB optimization: <1 min

Criteo Online Ads Experiment

• Click-through conversion prediction
• Q: Will a customer (e.g.) buy a product after clicking?
• Q: How predictive of conversion are different features?
• Logistic GLMM; \( N = 61,895 \) subset to compare to MCMC

[Giordano, Broderick, Meager, Huggins, Jordan 2016; Giordano, Broderick, Jordan 2018]
Criteo Online Ads Experiment
Criteo Online Ads Experiment

- **MAP:** 12 s
Criteo Online Ads Experiment

• MAP: 12 s

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

- **MAP:** 12 s
- **MFVB:** 57 s

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

- **MAP:** 12 s
- **MFVB:** 57 s
Criteo Online Ads Experiment

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

[Giordano, Broderick, Jordan 2018]
Why use MFVB?

- Topic discovery

<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>MANIGAT</td>
<td></td>
</tr>
<tr>
<td>YORK</td>
<td>PLAN</td>
<td>WELFARE</td>
<td>NAMPHY</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>

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.
Why use MFVB?

• Topic discovery
• Latent Dirichlet allocation (LDA)

<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>BEST</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>MANIGAT</td>
</tr>
<tr>
<td>YORK</td>
<td>PLAN</td>
<td>WELFARE</td>
<td>NAMPHY</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>

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.
Why use MFVB?

• Topic discovery

• Latent Dirichlet allocation (LDA): 27,700+ citations

<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>BEST</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>MANIGAT</td>
</tr>
<tr>
<td>YORK</td>
<td>PLAN</td>
<td>WELFARE</td>
<td>NAMPHY</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>

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.
Roadmap

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

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

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

\[ q(\theta) = \prod_{j=1}^{J} q_j(\theta_j) \]
What about uncertainty?

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

\[ q(\theta) = \prod_{j=1}^{J} q_j(\theta_j) \]

[Turner & Sahani 2011; MacKay 2003; Bishop 2006; Wang, Titterington 2004]
What about uncertainty?

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

\[
q(\theta) = \prod_{j=1}^{J} q_j(\theta_j)
\]

- Conjugate linear regression

[Turner & Sahani 2011; MacKay 2003; Bishop 2006; Wang, Titterington 2004]
What about uncertainty?

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

\[ q(\theta) = \prod_{j=1}^{J} q_j(\theta_j) \]

- Conjugate linear regression
- Bayesian central limit theorem

[Turner & Sahani 2011; MacKay 2003; Bishop 2006; Wang, Titterington 2004]
What about uncertainty?

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

\[ q(\theta) = \prod_{j=1}^{J} q_j(\theta_j) \]

- Conjugate linear regression
- Bayesian central limit theorem

[Exercise: derive the MFVB-CA steps. Hint: use precision matrix.]
What about uncertainty?

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

\[ q(\theta) = \prod_{j=1}^{J} q_j(\theta_j) \]

- Conjugate linear regression
- Bayesian central limit theorem

[Exercise: derive the MFVB-CA steps. Hint: use precision matrix.]
What about uncertainty?

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

\[ q(\theta) = \prod_{j=1}^{J} q_j(\theta_j) \]

- Underestimates variance (sometimes severely)
- Conjugate linear regression
- Bayesian central limit theorem

[Exercise: derive the MFVB-CA steps. Hint: use precision matrix.]
What about uncertainty?

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

$$q(\theta) = \prod_{j=1}^{J} q_j(\theta_j)$$

- Underestimates variance (sometimes severely)
- Conjugate linear regression
- Bayesian central limit theorem

[Exercise: derive the MFVB-CA steps. Hint: use precision matrix.]
What about uncertainty?

- Underestimates variance (sometimes severely)
- No covariance estimates
- Conjugate linear regression
- Bayesian central limit theorem

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

\[ q(\theta) = \prod_{j=1}^{J} q_j(\theta_j) \]

[Exercise: derive the MFVB-CA steps. Hint: use precision matrix.]
What about uncertainty?

- Microcredit
What about uncertainty?

- Microcredit

[Giordano, Broderick, Meager, Huggins, Jordan 2016]
What about uncertainty?

- Microcredit effect
- $\tau$ mean: 3.08 USD PPP
What about uncertainty?

- Microcredit effect
  - $\tau$ mean: 3.08 USD PPP
  - $\tau$ std dev: 1.83 USD PPP

[Giordano, Broderick, Meager, Huggins, Jordan 2016]
What about uncertainty?

- Microcredit effect
- $\tau$ mean: 3.08 USD PPP
- $\tau$ std dev: 1.83 USD PPP
- Mean is 1.68 std dev from 0

[Giordano, Broderick, Meager, Huggins, Jordan 2016]
What about uncertainty?

- Microcredit effect
  - $\tau$ mean: 3.08 USD PPP
  - $\tau$ std dev: 1.83 USD PPP
  - Mean is 1.68 std dev from 0

- Criteo online ads experiment

[Giordano, Broderick, Meager, Huggins, Jordan 2016; Giordano, Broderick, Jordan 2018]
What about means?

- Model for relational data with covariates
- When 1000+ nodes, MCMC > 1 day  [Fosdick 2013, Ch 4]
What about means?

- Model for relational data with covariates
- When 1000+ nodes, MCMC > 1 day  

[Fosdick 2013, Ch 4]
What about means?

• Model for relational data with covariates
• When 1000+ nodes, MCMC > 1 day  [Fosdick 2013, Ch 4]
Posterior means: revisited

• Want to predict college GPA $y_n$
Posterior means: revisited

- Want to predict college GPA $y_n$
- Collect: standardized test scores (e.g., SAT, ACT) $x_n$

[Giordano, Broderick, Jordan 2015]
Posterior means: revisited

- Want to predict college GPA \( y_n \)
- Collect: standardized test scores (e.g., SAT, ACT) \( x_n \)
- Collect: regional test scores \( r_n \)
Posterior means: revisited

- Want to predict college GPA $y_n$
- Collect: standardized test scores (e.g., SAT, ACT) $x_n$
- Collect: regional test scores $r_n$
- Model: $y_n | \beta, z, \sigma^2 \overset{\text{indep}}{\sim} \mathcal{N}(\beta^T x_n + z_{k(n)}r_n, \sigma^2)$
Posterior means: revisited

- Want to predict college GPA $y_n$
- Collect: standardized test scores (e.g., SAT, ACT) $x_n$
- Collect: regional test scores $r_n$

Model:

$$y_n | \beta, z, \sigma^2 \overset{iid}{\sim} \mathcal{N}(\beta^T x_n + z_{k(n)} r_n, \sigma^2)$$

$$z_k | \rho^2 \overset{iid}{\sim} \mathcal{N}(0, \rho^2)$$

$$\beta \sim \mathcal{N}(0, \Sigma)$$

$$(\sigma^2)^{-1} \sim \text{Gamma}(a_{\sigma^2}, b_{\sigma^2})$$

$$(\rho^2)^{-1} \sim \text{Gamma}(a_{\rho^2}, b_{\rho^2})$$

[Giordano, Broderick, Jordan 2015]
Posterior means: revisited

- Want to predict college GPA $y_n$
- Collect: standardized test scores (e.g., SAT, ACT) $x_n$
- Collect: regional test scores $r_n$
- Model: $y_n | \beta, z, \sigma^2 \sim \mathcal{N}(\beta^T x_n + z_k(n) r_n, \sigma^2)$

\[
\begin{align*}
    z_k &\sim \mathcal{N}(0, \rho^2) \\
    \beta &\sim \mathcal{N}(0, \Sigma) \\
    (\sigma^2)^{-1} &\sim \text{Gamma}(a_{\sigma^2}, b_{\sigma^2}) \\
    (\rho^2)^{-1} &\sim \text{Gamma}(a_{\rho^2}, b_{\rho^2})
\end{align*}
\]

- Data simulated from model (3 data sets, 300 data points):

[Giordano, Broderick, Jordan 2015]
Posterior means: revisited

- Want to predict college GPA $y_n$
- Collect: standardized test scores (e.g., SAT, ACT) $x_n$
- Collect: regional test scores $r_n$
- Model:
  $$y_n | \beta, z, \sigma^2 \overset{iid}{\sim} \mathcal{N}(\beta^T x_n + z_{k(n)} r_n, \sigma^2)$$
  $$z_k | \rho^2 \overset{iid}{\sim} \mathcal{N}(0, \rho^2)$$
  $$\beta \sim \mathcal{N}(0, \Sigma)$$
  $$(\sigma^2)^{-1} \sim \text{Gamma}(a_{\sigma^2}, b_{\sigma^2})$$
  $$(\rho^2)^{-1} \sim \text{Gamma}(a_{\rho^2}, b_{\rho^2})$$

- Data simulated from model (3 data sets, 300 data points):

[Giordano, Broderick, Jordan 2015]
Posterior means: revisited

- Want to predict college GPA $y_n$
- Collect: standardized test scores (e.g., SAT, ACT) $x_n$
- Collect: regional test scores $r_n$
- Model:
  $$y_n | \beta, z, \sigma^2 \sim \mathcal{N}(\beta^T x_n + z_{k(n)} r_n, \sigma^2)$$
  $$z_k | \rho^2 \sim \mathcal{N}(0, \rho^2) \quad (\sigma^2)^{-1} \sim \text{Gamma}(a_{\sigma^2}, b_{\sigma^2})$$
  $$\beta \sim \mathcal{N}(0, \Sigma) \quad (\rho^2)^{-1} \sim \text{Gamma}(a_{\rho^2}, b_{\rho^2})$$

- Data simulated from model (100 data sets, 300 data points):

[Giordano, Broderick, Jordan 2015]
Posterior means: revisited

- Want to predict college GPA \( y_n \)
- Collect: standardized test scores (e.g., SAT, ACT) \( x_n \)
- Collect: regional test scores \( r_n \)
- Model:
  \[
y_n | \beta, z, \sigma^2 \overset{iid}{\sim} \mathcal{N}(\beta^T x_n + z_k(n) r_n, \sigma^2)
\]
  \[
z_k | \rho^2 \overset{iid}{\sim} \mathcal{N}(0, \rho^2)
\]
  \[
\beta \sim \mathcal{N}(0, \Sigma)
\]
  \[
(\sigma^2)^{-1} \sim \text{Gamma}(a_{\sigma^2}, b_{\sigma^2})
\]
  \[
(\rho^2)^{-1} \sim \text{Gamma}(a_{\rho^2}, b_{\rho^2})
\]

- Data simulated from model (100 data sets, 300 data points):

[Giordano, Broderick, Jordan 2015]
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

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

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

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

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

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

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

How deep is the issue?
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

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

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

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?
Is it just MFVB?

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

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

- Takeaway: A smaller KL does not imply better mean and variance estimates

- Exercise: show this
Is it just MFVB?
Is it just MFVB?

[Huggins, Karsprzak, Campbell, Broderick 2018]
Is it just MFVB?

$p$: Student's t

[Huggins, Karsprzak, Campbell, Broderick 2018]
Is it just MFVB?

\[ q: \text{Gaussian} \]

\[ p: \text{Student's t} \]
Is it just MFVB?

$q$: Gaussian

$p$: Student's t

[Huggins, Karsprzak, Campbell, Broderick 2018]
Is it just MFVB?

\[ q: \text{Gaussian, variance } \sigma_q^2 \]

\[ p: \text{Student's t} \]
Is it just MFVB?

\[ q: \text{Gaussian, variance } \sigma_q^2 \]

\[ p: \text{Student's t} \]

[Huggins, Karsprzak, Campbell, Broderick 2018]
Is it just MFVB?

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

$p$: Student's t, variance $\sigma_p^2$
Is it just MFVB?

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

$p$: Student's t, variance $\sigma_p^2$
Is it just MFVB?

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

$p$: Student's t, variance $\sigma_p^2$
Is it just MFVB?

$q$: Gaussian, variance $\sigma^2_q$

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

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

$p$: Student's t, variance $\sigma_p^2$
Is it just MFVB?

$q$: Gaussian, variance $\sigma^2_q$

$p$: Student's t, variance $\sigma^2_p$

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

[Huggins, Karsprzak, Campbell, Broderick 2018]
Is it just MFVB?

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

$p$: Student's t, variance $\sigma_p^2$

$$KL(q||p) < 0.802 \text{ but also } \sigma_p^2 \geq c\sigma_q^2$$
Is it just MFVB?

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

$p$: Student's t, variance $\sigma_p^2$

Proposition (HKCB). For any $c > 1$, there exist zero-mean, unimodal distributions $q$ and $p$ such that

$$KL(q||p) < 0.802$$

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

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

$p$: Student's t, variance $\sigma_p^2$

**Proposition (HKCB).** For any $c > 1$, there exist zero-mean, unimodal distributions $q$ and $p$ such that

$$KL(q||p) < 0.802 \text{ but also } \sigma_p^2 \geq c\sigma_q^2$$

Can have small KL and arbitrarily bad variance estimate.

[Huggins, Karsprzak, Campbell, Broderick 2018]
Is it just MFVB?

**Proposition (HKCB).** For any $c > 1$, there exist zero-mean, unimodal distributions $q$ and $p$ such that $KL(q||p) < 0.802$ but also $\sigma_p^2 \geq c \sigma_q^2$.

Can have small KL and arbitrarily bad variance estimate.

[Huggins, Karsprzak, Campbell, Broderick 2018]
Is it just MFVB?

\[ q: \text{Gaussian, variance } \sigma_q^2 \]

\[ p: \text{Student's t, variance } \sigma_p^2 \]

**Conjecture (HKCB).** For any \( c > 1 \), there exist zero-mean, unimodal distributions \( q \) and \( p \) such that

\[ KL(q||p) < 0.802 \text{ but also } \sigma_p^2 \geq c\sigma_q^2 \]

Can have small KL and arbitrarily bad variance estimate

[Huggins, Karsprzak, Campbell, Broderick 2018]
Is it just MFVB?

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

$p$: Student's t, variance $\sigma_p^2$

**Conjecture (HKCB).** For any $c > 1$, there exist zero-mean, unimodal distributions $q$ and $p$ such that

$$KL(q||p) < 0.802$$

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

Can have small KL and arbitrarily bad variance estimate

[Huggins, Karsprzak, Campbell, Broderick 2018]
Is it just MFVB?

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

$p$: Student's t, variance $\sigma_p^2$

**Conjecture (HKCB).** For any $c > 1$, there exist zero-mean, unimodal distributions $q$ and $p$ such that

$$KL(q||p) < 0.12$$

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

Can have small KL and arbitrarily bad variance estimate

[Huggins, Karsprzak, Campbell, Broderick 2018]
How small is KL in practice?
How small is KL in practice?

\[ p(\theta | y) \rightarrow q^*(\theta) \]

CLOSE

NICE
How small is KL in practice?

$p(\theta|y)$

$q^*(\theta)$
How small is KL in practice?

- Often optimum has non-zero KL (MFVB, Gaussian VB)
How small is KL in practice?

- Often optimum has non-zero KL (MFVB, Gaussian VB)
How small is KL in practice?

- Often optimum has non-zero KL (MFVB, Gaussian VB)
How small is KL in practice?

- Often optimum has non-zero KL (MFVB, Gaussian VB)
How small is KL in practice?

- Often optimum has non-zero KL (MFVB, Gaussian VB)
How small is KL in practice?

- Often optimum has non-zero KL (MFVB, Gaussian VB)
How small is KL in practice?

- Often optimum has non-zero KL (MFVB, Gaussian VB)
How small is KL in practice?

- Often optimum has non-zero KL (MFVB, Gaussian VB)
- Even if optimum has zero KL, algorithm might not reach zero
How small is KL in practice?

- Often optimum has non-zero KL (MFVB, Gaussian VB)
- Even if optimum has zero KL, algorithm might not reach zero
How small is KL in practice?

- Often optimum has non-zero KL (MFVB, Gaussian VB)
- Even if optimum has zero KL, algorithm might not reach zero
How small is KL in practice?

- Often optimum has non-zero KL (MFVB, Gaussian VB)
- Even if optimum has zero KL, algorithm might not reach zero
How small is KL in practice?

- Often optimum has non-zero KL (MFVB, Gaussian VB)
- Even if optimum has zero KL, algorithm might not reach zero
How small is KL in practice?

- Often optimum has non-zero KL (MFVB, Gaussian VB)
- Even if optimum has zero KL, algorithm might not reach zero
How small is KL in practice?

- Often optimum has non-zero KL (MFVB, Gaussian VB)
- Even if optimum has zero KL, algorithm might not reach zero

\[ p(\theta|y) = q^*(\theta) \]
How small is KL in practice?

• Often optimum has non-zero KL (MFVB, Gaussian VB)
• Even if optimum has zero KL, algorithm might not reach zero
How small is KL in practice?

- Often optimum has non-zero KL (MFVB, Gaussian VB)
- Even if optimum has zero KL, algorithm might not reach zero
How small is KL in practice?

- Often optimum has non-zero KL (MFVB, Gaussian VB)
- Even if optimum has zero KL, algorithm might not reach zero
How small is KL in practice?

- Often optimum has non-zero KL (MFVB, Gaussian VB)
- Even if optimum has zero KL, algorithm might not reach zero

Recall: the KL value isn’t free
How small is KL in practice?

- Often optimum has non-zero KL (MFVB, Gaussian VB)
- Even if optimum has zero KL, algorithm might not reach zero

Recall: the KL value isn’t free
How small is KL in practice?

- Often optimum has non-zero KL (MFVB, Gaussian VB)
- Even if optimum has zero KL, algorithm might not reach zero

Recall: the KL value isn’t free
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
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 MFVB?
• When can we trust MFVB?
• Where do we go from here?
Roadmap

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

- Bayes & Approximate Bayes review
- What is:
  - Variational Bayes (VB)
  - Mean-field variational Bayes (MFVB)
- Why use MFVB?
- 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 MFVB?
• 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 MFVB?
- When can we trust VB?
- Where do we go from here?
<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>BEST</td>
<td>SPENDING</td>
<td>PARENTS</td>
<td>TEACHER</td>
</tr>
<tr>
<td>ACTOR</td>
<td>NEW</td>
<td>cały</td>
<td>BENNETT</td>
</tr>
<tr>
<td>FIRST</td>
<td>STATE</td>
<td>FAMILY</td>
<td>MANIGAT</td>
</tr>
<tr>
<td>YORK</td>
<td>PLAN</td>
<td>WELFARE</td>
<td>NAMPHY</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>

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.
What can we do?

• Reliable diagnostics
What can we do?

- Reliable diagnostics
- cf. KL, ELBO
What can we do?

- Reliable diagnostics
- cf. KL, ELBO

[ Gorham, Mackey 2015, 2017; Chwialkowski, Strathmann, Gretton 2016; Jitkrittum et al 2017; Talts et al 2018; Yao et al 2018, etc. ]

"Yes, but did it work? Evaluating variational inference" ICML 2018
What can we do?

• Reliable diagnostics
  • cf. KL, ELBO
    [Gorham, Mackey 2015, 2017; Chwialkowski, Strathmann, Gretton 2016; Jitkrittum et al 2017; Talts et al 2018; Yao et al 2018, etc.]

“Yes, but did it work? Evaluating variational inference” ICML 2018
What can we do?

• Reliable diagnostics
  • cf. KL, ELBO
    [Gorham, Mackey 2015, 2017; Chwialkowski, Strathmann, Gretton 2016; Jitkrittum et al 2017; Talts et al 2018; Yao et al 2018, etc.]

  “Yes, but did it work? Evaluating variational inference” ICML 2018

• Alternative divergences: Time & accuracy
  [Huggins, Kasprzak, Campbell, Broderick, 2018]
What can we do?

• Reliable diagnostics
  • cf. KL, ELBO
    [Gorham, Mackey 2015, 2017; Chwialkowski, Strathmann, Gretton 2016; Jitkrittum et al 2017; Talts et al 2018; Yao et al 2018, etc.]

  “Yes, but did it work? Evaluating variational inference” ICML 2018

• Alternative divergences: Time & accuracy
  [Huggins, Kasprzak, Campbell, Broderick, 2018]

• Corrections
  [Giordano, Broderick, Jordan 2018]
What can we do?

• Reliable diagnostics
  • cf. KL, ELBO
    [Gorham, Mackey 2015, 2017; Chwialkowski, Strathmann, Gretton 2016; Jitkrittum et al 2017; Talts et al 2018; Yao et al 2018, etc.]
  
  “Yes, but did it work? Evaluating variational inference” ICML 2018

• Alternative divergences:
  Time & accuracy
  [Huggins, Kasprzak, Campbell, Broderick, 2018]

• Corrections

• Theoretical guarantees on finite-data quality
  [Huggins, Campbell, Broderick 2016; Campbell, Broderick 2018, 2019]
What to read next

Textbooks and Reviews


Our Experiments

Automated, Scalable Bayesian Inference via Data Summarization

http://www.tamarabroderick.com/tutorials.html
Recap

[Woodard et al 2017]

[ESO/L. Calçada/M. Kornmesser 2017] [Abbott et al 2016a,b]

[Meager 2016a,b] [amcharts.com 2016]

[Chati, Balakrishnan 2017]

[Stone et al 2014]

[Kuikka et al 2014] [Baltic Salmon Fund]

[Mc-stan.org]
Recap

\[ p(\theta | y) \propto p(y | \theta) p(\theta) \]
Recap

\[ p(\theta | y) \propto p(y | \theta) p(\theta) \]

\((x_n, y_n)\)
Recap

\[ p(\theta | y) \propto p(y | \theta) p(\theta) \]

\((x_n, y_n)\)
Recap

\[ p(\theta | y) \propto p(y | \theta) p(\theta) \]

\( (x_n, y_n) \)

\( \theta \)

Benign

\( \bullet \)

Malicious

\( \bullet \)
Recap

\[ p(\theta | y) \propto p(y | \theta) p(\theta) \]

- Benign
- Malicious

\[(x_n, y_n)\]

\[ p(\theta | y) = p(y | \theta) p(\theta) \]

- Posterior
- Likelihood
- Prior

Woodard et al 2017
Gillon et al 2017
Grimm et al 2018
Meager 2016a, b
Baltic Salmon Fund
Kuikka et al 2014
mc-stan.org
Abbott et al 2016a, b
ESO/L. Calçada/M. Kornmesser 2017
amcharts.com 2016
Chati, Balakrishnan 2017
Julian Hertzog 2016
Stone et al 2014
Woodard et al 2017
Mc-Stan.org

\[ \theta \]

\[ \bullet \text{Benign} \]

\[ \bullet \text{Malicious} \]

\[ \frac{1}{n} \sum \frac{1}{y} \]

\[ \frac{1}{n} \sum \frac{1}{\frac{1}{y}} \]
Recap

\[ p(\theta | y) \propto p(y | \theta) p(\theta) \]

\( (x_n, y_n) \)

Benign

Malicious

\[ p(\theta | y) \]

\[ p(y | \theta) \]

\[ p(\theta) \]

\[ \theta \]

\[ \theta_1 \]

\[ \theta_2 \]

Exact posterior

\[ \text{posterior} \]

\[ \text{likelihood} \]

\[ \text{prior} \]
Recap

- Proposal: **efficient data summaries** for **fast, automated**, approximations with **error bounds** for finite data

\[
p(\theta | y) \propto p(y | \theta) p(\theta)
\]

\[(x_n, y_n)\]

Benign  \hspace{2cm} \bullet \hspace{2cm} Malicious

\[ p(\theta | y) \] posterior likelihood prior

\[ p(y | \theta) p(\theta) \]
Roadmap

• The “core” of the data set
Roadmap

• The “core” of the data set
• Uniform data subsampling isn’t enough
Roadmap

• The “core” of the data set
• Uniform data subsampling isn’t enough
• Importance sampling for “coresets”
Roadmap

• The “core” of the data set
• Uniform data subsampling isn’t enough
• Importance sampling for “coresets”
• Optimization for “coresets”
Roadmap

• The “core” of the data set
• Uniform data subsampling isn’t enough
• Importance sampling for “coresets”
• Optimization for “coresets”
• Approximate sufficient statistics
Roadmap

• The “core” of the data set
• Uniform data subsampling isn’t enough
• Importance sampling for “coresets”
• Optimization for “coresets”
• Approximate sufficient statistics
“Core” of the data set
“Core” of the data set

- Observe: redundancies can exist even if data isn’t “tall”
“Core” of the data set

- Observe: redundancies can exist even if data isn’t “tall”
“Core” of the data set

• Observe: redundancies can exist even if data isn’t “tall”
“Core” of the data set

- Observe: redundancies can exist even if data isn’t “tall”
“Core” of the data set

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

- 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
“Core” of the data set

- 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]
“Core” of the data set

- 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]
“Core” of the data set

- 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]
“Core” of the data set

• 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]
“Core” of the data set

- 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]
“Core” of the data set

- 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]
“Core” of the data set

- 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]
“Core” of the data set

• 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]
Bayesian coresets

- Posterior $p(\theta | y) \propto \theta \ p(y | \theta) p(\theta)$

[Huggins, Campbell, Broderick 2016]
Bayesian coresets

- Posterior \( p(\theta|y) \propto p(y|\theta)p(\theta) \)
- Log likelihood \( \mathcal{L}_n(\theta) := \log p(y_n|\theta) \), \( \mathcal{L}(\theta) := \sum_{n=1}^{N} \mathcal{L}_n(\theta) \)

[Huggins, Campbell, Broderick 2016]
Bayesian coresets

- Posterior \( p(\theta | y) \propto p(y | \theta) p(\theta) \)

- Log likelihood \( \mathcal{L}_n(\theta) := \log p(y_n | \theta), \quad \mathcal{L}(\theta) := \sum_{n=1}^{N} \mathcal{L}_n(\theta) \)

- Coreset log likelihood

[Huggins, Campbell, Broderick 2016]
Bayesian coresets

- Posterior \( p(\theta|y) \propto p(y|\theta)p(\theta) \)
- Log likelihood \( \mathcal{L}_n(\theta) := \log p(y_n|\theta), \quad \mathcal{L}(\theta) := \sum_{n=1}^{N} \mathcal{L}_n(\theta) \)
- Coreset log likelihood \( \|w\|_0 \ll N \)

[Huggins, Campbell, Broderick 2016]
Bayesian coresets

- Posterior \( p(\theta|y) \propto p(y|\theta)p(\theta) \)

- Log likelihood \( \mathcal{L}_n(\theta) := \log p(y_n|\theta) \), \( \mathcal{L}(\theta) := \sum_{n=1}^{N} \mathcal{L}_n(\theta) \)

- Coreset log likelihood \( \mathcal{L}(w, \theta) := \sum_{n=1}^{N} w_n \mathcal{L}_n(\theta) \) s.t. \( \|w\|_0 \ll N \)

[Huggins, Campbell, Broderick 2016]
Bayesian coresets

- Posterior \( p(\theta|y) \propto p(y|\theta)p(\theta) \)

- Log likelihood \( \mathcal{L}_n(\theta) := \log p(y_n|\theta) \), \( \mathcal{L}(\theta) := \sum_{n=1}^{N} \mathcal{L}_n(\theta) \)

- Coreset log likelihood \( \mathcal{L}(w, \theta) := \sum_{n=1}^{N} w_n \mathcal{L}_n(\theta) \) s.t. \( \|w\|_0 \ll N \)

[Huggins, Campbell, Broderick 2016]
Bayesian coresets

- Posterior: \( p(\theta|y) \propto p(y|\theta)p(\theta) \)

- Log likelihood: \( \mathcal{L}_n(\theta) := \log p(y_n|\theta), \quad \mathcal{L}(\theta) := \sum_{n=1}^{N} \mathcal{L}_n(\theta) \)

- Coreset log likelihood: \( \mathcal{L}(w, \theta) := \sum_{n=1}^{N} w_n \mathcal{L}_n(\theta) \quad \text{s.t.} \quad \|w\|_0 \ll N \)

- \( \varepsilon \)-coreset: \( \|\mathcal{L}(w) - \mathcal{L}\| \leq \varepsilon \)

[Huggins, Campbell, Broderick 2016; Campbell, Broderick 2019]
Bayesian coresets

• Posterior \( p(\theta | y) \propto p(y | \theta)p(\theta) \)

• Log likelihood \( \mathcal{L}_n(\theta) := \log p(y_n | \theta) \), \( \mathcal{L}(\theta) := \sum_{n=1}^{N} \mathcal{L}_n(\theta) \)

• Coreset log likelihood \( \mathcal{L}(w, \theta) := \sum_{n=1}^{N} w_n \mathcal{L}_n(\theta) \) s.t. \( \|w\|_0 \ll N \)

• \( \varepsilon \)-coreset: \( \|\mathcal{L}(w) - \mathcal{L}\| \leq \varepsilon \)

• Bound on Wasserstein distance to exact posterior \( \Rightarrow \) bound on posterior mean/uncertainty estimate quality

Benign \quad Malicious

\[ \text{[Huggins, Campbell, Broderick 2016; Campbell, Broderick 2019; Huggins, Campbell, Kasprzak, Broderick 2018]} \]
Roadmap

• The “core” of the data set
• Uniform data subsampling isn’t enough
• Importance sampling for “coresets”
• Optimization for “coresets”
• Approximate sufficient statistics
Roadmap

• The “core” of the data set
• Uniform data subsampling isn’t enough
  • Importance sampling for “coresets”
  • Optimization for “coresets”
• Approximate sufficient statistics
Uniform subsampling
Uniform subsampling
Uniform subsampling
Uniform subsampling

- Benign
- Malicious

• 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

- 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

- 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

\[ M = 10 \]
Uniform subsampling

- Might miss important data
- Noisy estimates

Benign

Malicious

$M = 10$

$M = 100$

$M = 1000$
Roadmap

• The “core” of the data set
• Uniform data subsampling isn’t enough
  • Importance sampling for “coresets”
  • Optimization for “coresets”
• Approximate sufficient statistics
Roadmap

• The “core” of the data set
• Uniform data subsampling isn’t enough
• Importance sampling for “coresets”
• Optimization for “coresets”
• Approximate sufficient statistics
Importance sampling
Importance sampling

Benign

Malicious
Importance sampling

Benign

Malicious
Importance sampling
Importance sampling
Importance sampling
Importance sampling

Benign

Malicious
Importance sampling

Benign

Malicious
Importance sampling
Importance sampling

Benign

Malicious
Importance sampling

\[ \sigma_n \propto \| \mathcal{L}_n \| \]
Importance sampling

\[ \sigma \ := \sum_{n=1}^{N} ||\mathcal{L}_n|| \]

\[ \sigma_n \ := \frac{||\mathcal{L}_n||}{\sigma} \]
Importance sampling

\[ \sigma := \sum_{n=1}^{N} \| \mathcal{L}_n \| \]

\[ \sigma_n := \| \mathcal{L}_n \| / \sigma \]

1. data

Benign

Malicious

6
Importance sampling

\[ \sigma := \sum_{n=1}^{N} \|\mathcal{L}_n\| \]

\[ \sigma_n := \|\mathcal{L}_n\| / \sigma \]

1. data
2. importance weights
Importance sampling

\[ \sigma := \sum_{n=1}^{N} \| \mathcal{L}_n \| \]

\[ \sigma_n := \| \mathcal{L}_n \| / \sigma \]

1. data  
2. importance weights  
3. importance sample
Importance sampling

\[ \sigma := \sum_{n=1}^{N} \| \mathcal{L}_n \| \]

\[ \sigma_n := \frac{\| \mathcal{L}_n \|}{\sigma} \]

1. data
2. importance weights
3. importance sample
4. invert weights
Importance sampling

**Thm (CB).** $\delta \in (0, 1)$. With probability $\geq 1 - \delta$, after $M$ iterations,

$$\|\mathcal{L}(w) - \mathcal{L}\| \leq \frac{\sigma \bar{\eta}}{\sqrt{M}} \left( 1 + \sqrt{2 \log \frac{1}{\delta}} \right)$$
Importance sampling

**Thm (CB).** $\delta \in (0,1)$. With probability $\geq 1 - \delta$, after $M$ iterations,

$$\|\mathcal{L}(w) - \mathcal{L}\| \leq \frac{\sigma \tilde{\eta}}{\sqrt{M}} \left( 1 + \sqrt{2 \log \frac{1}{\delta}} \right)$$

- Still noisy estimates

$M = 10$
Thm (CB). $\delta \in (0,1)$. With probability $\geq 1 - \delta$, after $M$ iterations,

$$\| \mathcal{L}(w) - \mathcal{L} \| \leq \frac{\sigma \bar{\eta}}{\sqrt{M}} \left(1 + \sqrt{2 \log \frac{1}{\delta}}\right)$$

- Still noisy estimates

$M = 10$  $M = 100$  $M = 1000$
Hilbert coresets

• Want a good coreset:

\[
\min_{w \in \mathbb{R}^N} \| \mathcal{L}(w) - \mathcal{L} \|^2 \\
\text{s.t. } w \geq 0, \|w\|_0 \leq M
\]
Hilbert coresets

- Want a good coreset:

\[
\min_{w \in \mathbb{R}^N} \| \mathcal{L}(w) - \mathcal{L} \|^2 \\
\text{s.t. } w \geq 0, \|w\|_0 \leq M
\]
Hilbert coresets

- Want a good coreset:
  \[
  \min_{w \in \mathbb{R}^N} \| \mathcal{L}(w) - \mathcal{L} \|^2 \\
  \text{s.t. } w \geq 0, \|w\|_0 \leq M
  \]

\[\exp(\mathcal{L}(\theta))\]

\[\exp(\mathcal{L}_n(\theta))\]

\[y_n\]
Hilbert coresets

• Want a good coreset:

$$\min_{w \in \mathbb{R}^N} \| \mathcal{L}(w) - \mathcal{L} \|^2$$

s.t. \( w \geq 0, \|w\|_0 \leq M \)

• need to consider (residual) error direction
Hilbert coresets

• Want a good coreset:
  $$\min_{w \in \mathbb{R}^N} \| \mathcal{L}(w) - \mathcal{L} \|^2$$
  s.t. $w \geq 0, \|w\|_0 \leq M$

• need to consider (residual) error direction
• sparse optimization
Roadmap

• The “core” of the data set
• Uniform data subsampling isn’t enough
• Importance sampling for “coresets”
• Optimization for “coresets”
• Approximate sufficient statistics
Roadmap

• The “core” of the data set
• Uniform data subsampling isn’t enough
• Importance sampling for “coresets”
• Optimization for “coresets”
• Approximate sufficient statistics
Frank-Wolfe

Convex optimization on a polytope $D$

[Frank, Wolfe 1956]

$\text{Jaggi 2013}$
Frank-Wolfe
Convex optimization on a polytope $D$

• Repeat:
  1. Find gradient
  2. Find argmin point on plane in $D$
  3. Do line search between current point and argmin point

[Frank, Wolfe 1956]

[Jaggi 2013]
Frank-Wolfe

Convex optimization on a polytope $D$ [Frank, Wolfe 1956]

- Repeat:
  1. Find gradient
  2. Find argmin point on plane in $D$
  3. Do line search between current point and argmin point

- Convex combination of $M$ vertices after $M-1$ steps

[Jaggi 2013]
Frank-Wolfe
Convex optimization on a polytope $D$  

- Repeat:
  1. Find gradient
  2. Find argmin point on plane in $D$
  3. Do line search between current point and argmin point

- Convex combination of $M$ vertices after $M - 1$ steps

- Our problem: $\min_{w \in \mathbb{R}^N} \| \mathcal{L}(w) - \mathcal{L} \|$
Frank-Wolfe

Convex optimization on a polytope $D$ [Frank, Wolfe 1956]

• Repeat:
  1. Find gradient
  2. Find argmin point on plane in $D$
  3. Do line search between current point and argmin point

• Convex combination of $M$ vertices after $M-1$ steps

• Our problem: $\min_{w \in \mathbb{R}^N} \| \mathcal{L}(w) - \mathcal{L} \|^2$

[Jaggi 2013]
Frank-Wolfe

Convex optimization on a polytope $D$  

[Frank, Wolfe 1956]

• Repeat:
  1. Find gradient
  2. Find argmin point on plane in $D$
  3. Do line search between current point and argmin point

• Convex combination of $M$ vertices after $M-1$ steps

• Our problem: 
  \[
  \min_{w \in \mathbb{R}^N} \| \mathcal{L}(w) - \mathcal{L} \|^2 \\
  \text{s.t. } w \geq 0, \|w\|_0 \leq M
  \]
Frank-Wolfe

Convex optimization on a polytope $D$ [Frank, Wolfe 1956]

• Repeat:
  1. Find gradient
  2. Find argmin point on plane in $D$
  3. Do line search between current point and argmin point

• Convex combination of $M$ vertices after $M-1$ steps

Our problem:

$$
\min_{w \in \mathbb{R}^N} \| \mathcal{L}(w) - \mathcal{L} \|^2_2
$$

$$
\Delta^{N-1} := \left\{ w \in \mathbb{R}^N : \sum_{n=1}^{N} \sigma_n w_n = \sigma, \, w \geq 0 \right\}
$$

[Jaggi 2013]
Frank-Wolfe
Convex optimization on a polytope $D$

- Repeat:
  1. Find gradient
  2. Find argmin point on plane in $D$
  3. Do line search between current point and argmin point

- Convex combination of $M$ vertices after $M-1$ steps

- Our problem:
  \[
  \min_{w \in \mathbb{R}^N} \| \mathcal{L}(w) - \mathcal{L} \|^2
  \]

\[
\Delta^{N-1} := \left\{ w \in \mathbb{R}^N : \sum_{n=1}^{N} \sigma_n w_n = \sigma, \ w \geq 0 \right\}
\]

**Thm (CB).** After $M$ iterations,

\[
\| \mathcal{L}(w) - \mathcal{L} \| \leq \frac{c}{\sqrt{\alpha^{2M} + c'M}}
\]
Gaussian model (simulated)

- 1K pts; norms, inference: closed-form

Uniform subsampling

\[ M = 5 \]
Gaussian model (simulated)

- 1K pts; norms, inference: closed-form

Uniform subsampling

\( M = 5 \)

\( M = 50 \)

\( M = 500 \)
Gaussian model (simulated)

- 1K pts; norms, inference: closed-form

Uniform subsampling

Importance sampling

$M = 5$

$M = 50$

$M = 500$
Gaussian model (simulated)

- 1K pts; norms, inference: closed-form

Uniform subsampling

Importance sampling

Frank-Wolfe

- $M = 5$
- $M = 50$
- $M = 500$
Gaussian model (simulated)

- 1K pts; norms, inference: closed-form
Logistic regression (simulated)

- 10K pts; general inference

Uniform subsampling

Importance sampling

Frank-Wolfe

\[ M = 10 \]

\[ M = 100 \]

\[ M = 1000 \]
Poisson regression (simulated)

- 10K pts; general inference
Real data experiments

Data sets include:
- Phishing
- Chemical reactivity
- Bicycle trips
- Airport delays

lower error

less total time

Uniform subsampling
Frank Wolfe coresets

[Campbell, Broderick 2019]
Real data experiments

Data sets include:
- Phishing
- Chemical reactivity
- Bicycle trips
- Airport delays

lower error

less total time

Uniform subsampling
Frank Wolfe coresets
GIGA coresets

[Campbell, Broderick 2019, 2018]
Roadmap

• The “core” of the data set
• Uniform data subsampling isn’t enough
• Importance sampling for “coresets”
• Optimization for “coresets”
• Approximate sufficient statistics
Roadmap

• The “core” of the data set
• Uniform data subsampling isn’t enough
• Importance sampling for “coresets”
• Optimization for “coresets”
• Approximate sufficient statistics
Data summarization
Data summarization

• Exponential family likelihood
Data summarization

- Exponential family likelihood

\[ p(y_{1:N} | x_{1:N}, \theta) = \prod_{n=1}^{N} \exp\left[ T(y_n, x_n) \cdot \eta(\theta) \right] \]
Data summarization

- Exponential family likelihood

\[
p(y_{1:N} | x_{1:N}, \theta) = \prod_{n=1}^{N} \exp \left[ T(y_n, x_n) \cdot \eta(\theta) \right]
\]

\[
= \exp \left\{ \sum_{n=1}^{N} T(y_n, x_n) \right\} \cdot \eta(\theta)
\]
Data summarization

- Exponential family likelihood

\[ p(y_{1:N} | x_{1:N}, \theta) = \prod_{n=1}^{N} \exp \left[ T(y_n, x_n) \cdot \eta(\theta) \right] \]

- Sufficient statistics

= \exp \left[ \sum_{n=1}^{N} T(y_n, x_n) \right] \cdot \eta(\theta)

- Scalable, single-pass, streaming, distributed, complementary to MCMC
Data summarization

- Exponential family likelihood
  \[
p(y_{1:N} | x_{1:N}, \theta) = \prod_{n=1}^{N} \exp \left[ T(y_n, x_n) \cdot \eta(\theta) \right]
  \]

- Sufficient statistics

- Scalable, single-pass, streaming, distributed, complementary to MCMC

- **But**: Often no simple sufficient statistics
Data summarization

- Exponential family likelihood

\[ p(y_{1:N}|x_{1:N}, \theta) = \prod_{n=1}^{N} \exp \left[ T(y_n, x_n) \cdot \eta(\theta) \right] \]

= \exp \left[ \sum_{n=1}^{N} T(y_n, x_n) \right] \cdot \eta(\theta)

- Sufficient statistics

• Scalable, single-pass, streaming, distributed, complementary to MCMC

• But: Often no simple sufficient statistics

  • E.g. Bayesian logistic regression; GLMs; “deeper” models

  - Likelihood

\[ p(y_{1:N}|x_{1:N}, \theta) = \prod_{n=1}^{N} \frac{1}{1 + \exp(-y_n x_n \cdot \theta)} \]
Data summarization

- Exponential family likelihood

\[
p(y_{1:N} | x_{1:N}, \theta) = \prod_{n=1}^{N} \exp \left[ \sum_{n=1}^{N} T(y_n, x_n) \cdot \eta(\theta) \right]
\]

- Sufficient statistics

- Scalable, single-pass, streaming, distributed, complementary to MCMC

- But: Often no simple sufficient statistics

  - E.g. Bayesian logistic regression; GLMs; “deeper” models

  - Likelihood

\[
p(y_{1:N} | x_{1:N}, \theta) = \prod_{n=1}^{N} \frac{1}{1 + \exp(-y_n x_n \cdot \theta)}
\]

- Our proposal: (polynomial) approximate sufficient statistics
Data summarization

Criteo Releases its New Dataset

By: CriteoLabs / 31 Mar 2015

• 6M data points, 1000 features
• Streaming, distributed; minimal communication
• 22 cores, 16 sec
• Finite-data guarantees on Wasserstein distance to exact posterior
Conclusions

• Data summarization for scalable, automated approximate Bayes algorithms with error bounds on quality for finite data

[Campbell, Broderick 2017, 2018; Huggins, Campbell, Broderick 2016; Huggins, Adams, Broderick 2017]
Conclusions

• *Data summarization for scalable, automated* approximate Bayes algorithms with *error bounds on quality for finite data*
• Coresets
• Approx. suff. statistics

[Campbell, Broderick 2017, 2018; Huggins, Campbell, Broderick 2016; Huggins, Adams, Broderick 2017]
Conclusions

- Data summarization for **scalable, automated** approximate Bayes algorithms with **error bounds on quality for finite data**
- Coresets
- Approx. suff. statistics

[Campbell, Broderick 2017, 2018; Huggins, Campbell, Broderick 2016; Huggins, Adams, Broderick 2017]
Conclusions

• *Data summarization* for **scalable, automated** approximate Bayes algorithms with **error bounds on quality for finite data**

• Coresets

• Approx. suff. statistics

• More accurate with more computation investment

[Campbell, Broderick 2017, 2018; Huggins, Campbell, Broderick 2016; Huggins, Adams, Broderick 2017]
Conclusions

• *Data summarization* for **scalable, automated** approximate Bayes algorithms with **error bounds on quality for finite data**

• Coresets

• Approx. suff. statistics

• More accurate with more computation investment

• A start

• Lots of potential improvements/directions

[Campbell, Broderick 2017, 2018; Huggins, Campbell, Broderick 2016; Huggins, Adams, Broderick 2017]
References


Bayesian inference

- [Gillin et al 2017]
- [Grimm et al 2018]
- [ESO/ L. Calçada M. Kornmesser 2017]
- [Abbott et al 2016a,b]
- [Kuikka et al 2014]
- [Baltic Salmon Fund]
- [Chati, Balakrishnan 2016 2017]
- [mc-stan.org]
Bayesian inference

- Challenge: fast (compute, user), reliable inference
Bayesian inference

- Challenge: fast (compute, user), reliable inference
- Today: variational Bayes and beyond
Bayesian inference

- Challenge: fast (compute, user), reliable inference
- Today: variational Bayes and beyond

**Fundamental questions**
- What is achievable in speed and accuracy?
References (1/6)


Application References (5/6)


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)