Variational Bayes and beyond: Bayesian inference for big data

Tamara Broderick
ITT Career Development Assistant Professor, MIT
Bayesian inference
Bayesian inference

[Gillon et al 2017]

[Grimm et al 2018]
Bayesian inference

[Gillon et al 2017]

[Abbott et al 2016a,b]

[ESO/L. Calçada/M. Kornmesser 2017]

[Grimm et al 2018]
Bayesian inference

[Abbott et al 2016a,b]

[ESO/L. Calçada/M. Kornmesser 2017]

[Stone et al 2014]
Bayesian inference
Bayesian inference

Bayesian inference

Bayesian inference
Bayesian inference

- Analysis goals: Point estimates, coherent uncertainties
- Interpretable, complex, modular; expert information
Bayesian inference

- Analysis goals: Point estimates, coherent uncertainties
  - Interpretable, complex, modular; expert information

Bayesian inference

- Analysis goals: Point estimates, coherent uncertainties
- Interpretable, complex, modular; expert information

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

- Analysis goals: Point estimates, coherent uncertainties
- Interpretable, complex, modular; expert information
- Challenge: fast (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

---

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.
Variational Bayes

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

[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]
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>

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

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.
Variational Bayes

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

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

Variational Bayes has been applied in various fields, including gene regulatory sequences, standard eQTL mapping, and variational autoencoders. Refer to the following works:

- [Blei et al 2003](#)
- [Xing et al 2004](#)
- [Xing 2003](#)
- [Stegle et al 2010](#)
- [Airoldi et al 2008](#)
- [Gershman et al 2014](#)
- [Blei et al 2018](#)

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services,” Hearst Foundation President Randolph A. Hearst said Monday 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?
Bayesian inference
Bayesian inference

θ

parameters
Bayesian inference

\[ p(\theta) \]

prior

parameters
Bayesian inference

$p(\theta)$

prior

parameters
Bayesian inference

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

likelihood prior

parameters
Bayesian inference

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

likelihood prior

data parameters

\[ \theta \]
Bayesian inference

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

posterior likelihood prior

![Graph showing posterior distribution of \( \theta \) with data and parameters indicated.]
Bayesian inference

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

posterior likelihood prior

Bayes Theorem

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

\[ p(\theta|y_1:N) \propto \theta \cdot p(y_1:N|\theta) \cdot p(\theta) \]

posterior \quad likelihood \quad prior
Bayesian inference

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

posterior  likelihood  prior

data  parameters

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

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

posterior  likelihood  prior

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

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

posterior likelihood prior

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

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, 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, likelihood
2. Compute the posterior
3. Report a summary, e.g. posterior means and (co)variances
   - Why are steps 2 and 3 hard? 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, likelihood
2. Compute the posterior
3. Report a summary, e.g. posterior means and (co)variances
   - Why are steps 2 and 3 hard? High-dimensional integration
Bayesian inference

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

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

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

- Why are steps 2 and 3 hard? High-dimensional integration
1. Build a model: choose prior, likelihood
2. Compute the posterior
3. Report a summary, e.g. posterior means and (co)variances
   - Why are steps 2 and 3 hard? High-dimensional integration
   - Turn to approximation

Bayesian inference

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

posterior likelihood prior evidence

Bayes Theorem
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 $q^*$

[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

\[ p(\theta | y) \]

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 $q^*$

\[ p(\theta | y) \quad \text{NICE} \quad q(\theta) \]
Approximate Bayesian Inference

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

[Please see image for diagram showing $p(\theta|y)$, $q^*(\theta)$, and NICE]

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

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

Instead: an optimization approach

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

[Bardenet, Doucet, Holmes 2017]
Approximate Bayesian Inference

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

Instead: an optimization approach

- Approximate posterior with $q^*$

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

[Bardenet, Doucet, Holmes 2017]
Approximate Bayesian Inference

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

Instead: an optimization approach

- Approximate posterior with $q^*$

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

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

[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
Approximate Bayesian Inference

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

Instead: an optimization approach

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

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

- VB practical success: point estimates and prediction, fast

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

- 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

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

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

- Variational Bayes

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

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

- Variational Bayes

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

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

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

- Variational Bayes

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

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

“Evidence lower bound” (ELBO)
Why KL?

- Variational Bayes

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

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

“Evidence lower bound” (ELBO)
Why KL?

- Variational Bayes

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

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

\[ \begin{align*}
&:= \int q(\theta) \log \frac{q(\theta)}{p(\theta | y)} d\theta \\
&= \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta, y)}{q(\theta)} d\theta
\end{align*} \]

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

“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]
- \( \text{KL} \geq 0 \Rightarrow \log p(y) \geq \text{ELBO} \)
Why KL?

- Variational Bayes
  \[ q^* = \arg\min_{q \in Q} KL(q(\cdot) \| p(\cdot \mid y)) \]
  \[
  KL(q(\cdot) \| p(\cdot \mid y)) := \int q(\theta) \log \frac{q(\theta)}{p(\theta \mid y)} \, d\theta
  \]
  \[
  = \int q(\theta) \log \frac{q(\theta)p(y)}{p(\theta, y)} \, d\theta = \log p(y) - \int q(\theta) \log \frac{p(\theta, y)}{q(\theta)} \, d\theta
  \]

- Exercise: Show \( KL \geq 0 \) [Bishop 2006, Sec 1.6.1]
- \( KL \geq 0 \Rightarrow \log p(y) \geq \text{ELBO} \)
- \( q^* = \arg\max_{q \in Q} \text{ELBO}(q) \)
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 \( \text{KL} \geq 0 \) \citep{Bishop2006, Sec 1.6.1}
- \( \text{KL} \geq 0 \Rightarrow \log p(y) \geq \text{ELBO} \)
- \( q^* = \text{argmax}_{q \in Q} \text{ELBO}(q) \)
- Why KL (in this direction)?
Variational Bayes

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

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

Choose “NICE” distributions
Variational Bayes

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

Choose “NICE” distributions
Variational Bayes

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

Choose “NICE” distributions
Variational Bayes

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

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)

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

- Not a modeling assumption
Variational Bayes

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

Choose “NICE” distributions

- Mean-field variational Bayes (MFVB)

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

- *Not a modeling assumption*

[Bishop 2006]
Variational Bayes

Choose “NICE” distributions

- Mean-field variational Bayes (MFVB)

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

- Not a modeling assumption

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

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

\[ \begin{aligned}
q^* &= \arg\min_{q \in Q} \text{KL} (q(\cdot) || p(\cdot|y)) \\
\text{Choose “NICE” distributions} \\
\text{• Mean-field variational Bayes (MFVB)} \\
Q_{\text{MFVB}} &:= \left\{ q : q(\theta) = \prod_{j=1}^{J} q_j(\theta_j) \right\} \\
\text{• Not a modeling assumption}
\end{aligned} \]

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

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

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

Optimization

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

Variational Bayes

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

Mean-field variational Bayes

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

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

Optimization

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

Variational Bayes

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

Mean-field variational Bayes

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

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

Optimization

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

Variational Bayes

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

Mean-field variational Bayes

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

- 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

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

Variational Bayes

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

Mean-field variational Bayes

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

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

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

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

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

[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 variance \( \theta = (\mu, \sigma^2) \)
- Model:
  \[
  p(y|\theta) : \quad y_n \overset{iid}{\sim} 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 N(\mu_0, \lambda_0 \sigma^2)
  
- Exercise: check \( p(\mu, \sigma^2|y) \neq f_1(\mu, y)f_2(\sigma^2, y) \)

[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 variance \( \theta = (\mu, \sigma^2) \)
- Model:
  \[
p(y|\theta) : \quad y_n \overset{iid}{\sim} 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 N(\mu_0, \lambda_0 \sigma^2)
\]
- Exercise: check \( p(\mu, \sigma^2|y) \neq f_1(\mu, y)f_2(\sigma^2, y) \)
- MFVB approximation:
  \[
  q^*(\mu, \sigma) = q^*_\mu(\mu)q^*_\sigma^2(\sigma^2) = \arg\min_{q \in Q_{MFVB}} KL(q(\cdot)||p(\cdot|y))
  \]

[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 variance \( \theta = (\mu, \sigma^2) \)
- Model:
  \[
p(y|\theta) : \quad y_n \overset{iid}{\sim} 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 N(\mu_0, \lambda_0\sigma^2)
\]
- Exercise: check \( p(\mu, \sigma^2|y) \neq f_1(\mu, y)f_2(\sigma^2, y) \)
- MFVB approximation:
  \[
  q^*(\mu, \sigma) = q^*_\mu(\mu)q^*_\sigma^2(\sigma^2) = \arg\min_{q \in Q_{\text{MFVB}}} KL(q(\cdot)||p(\cdot|y))
  \]
- Coordinate descent [Exercise: derive this] \[\text{[Bishop 2006, Sec 10.1.3]}\]
Midge wing length

- Catalogued midge wing lengths (mm) \( y = (y_1, \ldots, y_N) \)
- Parameters of interest: population mean and variance \( \theta = (\mu, \sigma^2) \)
- Model:
  \[
  p(y|\theta): \quad y_n \overset{iid}{\sim} N(\mu, \sigma^2), \quad n = 1, \ldots, N \\
  p(\theta): \quad (\sigma^2)^{-1} \sim \text{Gamma}(a_0, b_0) \\
  \mu|\sigma^2 \sim N(\mu_0, \lambda_0\sigma^2)
  \]
- Exercise: check \( p(\mu, \sigma^2|y) \neq f_1(\mu, y)f_2(\sigma^2, y) \)
- MFVB approximation:
  \[
  q^*(\mu, \sigma) = q_{\mu}^*(\mu)q_{\sigma^2}^*(\sigma^2) = \arg\min_{q \in Q_{\text{MFVB}}} KL(q(\cdot)||p(\cdot|y))
  \]
  \[
  q^*(\mu) = N(\mu|m_\mu, \rho_\mu^2) \quad q^*((\sigma^2)^{-1}) = \text{Gamma}((\sigma^2)^{-1}|a_\sigma, b_\sigma)
  \]

[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 variance
- Model:
  \[
  p(y|\theta) : \quad y_n \overset{iid}{\sim} N(\mu, \sigma^2), \quad n = 1, \ldots, N
  \]
  \[
  p(\theta) : \quad (\sigma^2)^{-1} \sim \text{Gamma}(a_0, b_0)
  \]
  \[
  \mu|\sigma^2 \sim N(\mu_0, \lambda_0\sigma^2)
  \]
- Exercise: check \( p(\mu, \sigma^2|y) \neq f_1(\mu, y)f_2(\sigma^2, y) \)
- MFVB approximation:
  \[
  q^*(\mu, \sigma) = q^*_\mu(\mu)q^*_\sigma(\sigma^2) = \arg\min_{q\in Q_{MFVB}} KL(q(\cdot)||p(\cdot|y))
  \]
- Coordinate descent [Exercise: derive this] [Bishop 2006, Sec 10.1.3]
  \[
  q^*(\mu) = N(\mu|m_\mu, \rho_\mu^2) \quad q^*((\sigma^2)^{-1}) = \text{Gamma}((\sigma^2)^{-1}|a_\sigma, b_\sigma)
  \]
  "variational parameters"

[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
- Model:
  \[
p(y|\theta) : \quad y_n \overset{iid}{\sim} N(\mu, \sigma^2), \quad n = 1, \ldots, N
  \]
  \[
p(\theta) : \quad (\sigma^2)^{-1} \sim \text{Gamma}(a_0, b_0)
  \]
  \[
  \mu | \sigma^2 \sim N(\mu_0, \lambda_0 \sigma^2)
  \]
- Exercise: check \( p(\mu, \sigma^2 | y) \neq f_1(\mu, y) f_2(\sigma^2, y) \)
- MFVB approximation:
  \[
  q^*(\mu, \sigma) = q^*_\mu(\mu) q^*_\sigma(\sigma^2) = \text{argmin}_{q \in Q_{MFVB}} KL(q(\cdot) || p(\cdot | y))
  \]
- Coordinate descent [Exercise: derive this] [Bishop 2006, Sec 10.1.3]
  \[
  q^*(\mu) = N(\mu | m_\mu, \rho^2_\mu) \quad q^*(((\sigma^2)^{-1}) = \text{Gamma}((\sigma^2)^{-1} | a_\sigma, b_\sigma)
  \]
- Iterate:
  \[
  (m_\mu, \rho^2_\mu) = f(a_\sigma, b_\sigma)
  \]
  \[
  (a_\sigma, b_\sigma) = g(m_\mu, \rho^2_\mu)
  \]

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

\((\sigma^2)^{-1}\) approximation

exact posterior

\[\mu\]
Midge wing length

\[
(\sigma^2)^{-1}
\]

approximation

exact posterior

\[ \mu \]
Midge wing length approximation

$((\sigma^2)^{-1})$

exact posterior

$\mu$

[Bishop 2006]
Midge wing length approximation

\[(\sigma^2)^{-1}\]

exact posterior

\(\mu\)
Microcredit Experiment
Microcredit Experiment

- Simplified from Meager (2018a)
- \( K \) microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
Microcredit Experiment

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

\[
y_{kn} \sim N(\mu_k + T_{kn} \tau_k, \sigma_k^2)
\]
Microcredit Experiment

• Simplified from Meager (2018a)
• $K$ microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
• $N_k$ businesses in $k$th site ($\sim900$ to $\sim17K$)
• 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)
• $K$ 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)
- $K$ 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:

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

• Simplified from Meager (2018a)
• \( K \) 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)
- $K$ 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:

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

- Simplified from Meager (2018a)
- $K$ microcredit trials (Mexico, Mongolia, Bosnia, India, Morocco, Philippines, Ethiopia)
- $N_k$ businesses in $k$th site ($\sim900$ to $\sim17K$)
- Profit of $n$th business at $k$th site:

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

1 if microcredit
Microcredit Experiment

• Simplified from Meager (2018a)
• $K$ 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:

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

1 if microcredit

profit

$y_{kn}$
Microcredit Experiment

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

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

- Priors and hyperpriors:
Microcredit Experiment

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

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

$$\begin{pmatrix} \mu_k \\ \tau_k \end{pmatrix} \overset{\text{iid}}{\sim} \mathcal{N}\left(\begin{pmatrix} \mu \\ \tau \end{pmatrix}, C\right)$$

$$\sigma_k^{-2} \overset{\text{iid}}{\sim} \Gamma(a, b)$$
Microcredit Experiment

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

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

- Priors and hyperpriors:

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

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

- *One set of 2500 MCMC draws: 45 minutes*
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
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; \( N = 61,895 \) subset to compare to MCMC

[Giordano, Broderick, Meager, Huggins, Jordan 2016; Giordano, Broderick, Jordan 2017]
Criteo Online Ads Experiment

[Giordano, Broderick, Jordan 2017]
Criteo Online Ads Experiment

- MAP: 12 s

[Giordano, Broderick, Jordan 2017]
Criteo Online Ads Experiment

- **MAP:** 12 s

[Giordano, Broderick, Jordan 2017]
Criteo Online Ads Experiment

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

- MAP: 12 s
- VB: 57 s

[Giordano, Broderick, Jordan 2017]
Criteo Online Ads Experiment

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

[Giordano, Broderick, Jordan 2017]
How to optimize: MFVB
How to optimize: MFVB

• Conditionally conjugate model
• Coordinate ascent in $q_1, \ldots, q_J$ [MacKay 2003, Bishop 2006]
How to optimize: MFVB

- Conditionally conjugate model
- Coordinate ascent in $q_1, \ldots, q_J$ [MacKay 2003, Bishop 2006]
- ELBO in closed form
How to optimize: MFVB

- Conditionally conjugate model
- Coordinate ascent in $q_1, \ldots, q_J$ [MacKay 2003, Bishop 2006]
- ELBO in closed form
- E.g. $p(\cdot | y)$ nice enough (e.g. exponential family components) and familiar $q$ (e.g. exp fam)
How to optimize: MFVB

- Conditionally conjugate model
  - Coordinate ascent in $q_1, \ldots, q_J$ [MacKay 2003, Bishop 2006]
- ELBO in closed form
  - E.g. $p(\cdot | y)$ nice enough (e.g. exponential family components) and familiar $q$ (e.g. exp fam)
  - Optimize over variational params (method of choice)
How to optimize: MFVB

- Conditionally conjugate model
- Coordinate ascent in $q_1, \ldots, q_J$ [MacKay 2003, Bishop 2006]
- ELBO in closed form
- E.g. $p(\cdot|y)$ nice enough (e.g. exponential family components) and familiar $q$ (e.g. exp fam)
- Optimize over variational params (method of choice)
- Continuous parameters
How to optimize: MFVB

• Conditionally conjugate model
  • Coordinate ascent in $q_1, \ldots, q_J$ [MacKay 2003, Bishop 2006]
• ELBO in closed form
  • E.g. $p(\cdot|y)$ nice enough (e.g. exponential family components) and familiar $q$ (e.g. exp fam)
• Optimize over variational params (method of choice)
• Continuous parameters
  • Assume Gaussian $q$ (possibly in transformed space)
How to optimize: MFVB

• Conditionally conjugate model
  • Coordinate ascent in $q_1, \ldots, q_J$ [MacKay 2003, Bishop 2006]
• ELBO in closed form
  • E.g. $p(\cdot|y)$ nice enough (e.g. exponential family components) and familiar $q$ (e.g. exp fam)
  • Optimize over variational params (method of choice)
• Continuous parameters
  • Assume Gaussian $q$ (possibly in transformed space)
  • Automatic differentiation variational inference (ADVI) [Kucukelbir et al 2015, 2017]
    [Baydin et al 2018]
Stochastic gradient descent (SGD)

- MFVB: \[
\min_{\eta: q_\eta \in Q_{MFVB}} - \mathbb{E}_{q_\eta} \log \frac{p(\theta, y_{1:N})}{q_\eta(\theta)} d\theta
\]
Stochastic gradient descent (SGD)

- MFVB: \[ \min_{\eta:q_\eta \in Q_{MFVB}} -\mathbb{E}_{q_\eta} \log \frac{p(\theta, y_{1:N})}{q_\eta(\theta)} d\theta \]
- Recall: Stochastic gradient
Stochastic gradient descent (SGD)

- MFVB: \( \min_{\eta:q_\eta \in Q_{MFVB}} -\mathbb{E}_{q_\eta} \log \frac{p(\theta, y_{1:N})}{q_\eta(\theta)} \, d\theta \)
- Recall: Stochastic gradient
- Goal: \( \min_\eta f(\eta) := \mathbb{E}_Y f(Y, \eta) \)
Stochastic gradient descent (SGD)

- MFVB: \[ \min_{\eta: q_\eta \in Q_{MFVB}} -\mathbb{E}_{q_\eta} \log \frac{p(\theta, y_{1:N})}{q_\eta(\theta)} \, d\theta \]
- Recall: Stochastic gradient
- Goal: \[ \min_{\eta} f(\eta) := \mathbb{E}_Y f(Y, \eta) \]
- Observe \( y_n \sim iid Y \)
Stochastic gradient descent (SGD)

• MFVB: \[
\min_{\eta: q_\eta \in Q_{MFVB}} -\mathbb{E}_{q_\eta} \log \frac{p(\theta, y_{1:N})}{q_\eta(\theta)} d\theta
\]

• Recall: Stochastic gradient

• Goal: \[
\min_{\eta} f(\eta) := \mathbb{E}_Y f(Y, \eta)
\]

• One option: \[
\min_{\eta} N^{-1} \sum_{n=1}^{N} f(y_n, \eta)
\]

• Observe \( y_n \overset{iid}{\sim} Y \)
Stochastic gradient descent (SGD)

- MFVB: \( \min_{\eta: q_\eta \in Q_{\text{MFVB}}} -\mathbb{E}_{q_\eta} \log \frac{p(\theta, y_{1:N})}{q_\eta(\theta)} d\theta \)
- Recall: Stochastic gradient
- Goal: \( \min_{\eta} f(\eta) := \mathbb{E}_Y f(Y, \eta) \)
- One option: \( \min_{\eta} N^{-1} \sum_{n=1}^{N} f(y_n, \eta) \)
- GD: \( \eta^{(t)} = \eta^{(t-1)} - \rho_t \nabla_\eta \left[ N^{-1} \sum_{n=1}^{N} f(y_n, \eta) \right]_{\eta=\eta^{(t)}} \)
- Observe \( y_n \overset{iid}{\sim} Y \)
Stochastic gradient descent (SGD)

- MFVB: \( \min_{\eta: q_\eta \in Q_{\text{MFVB}}} -\mathbb{E}_{q_\eta} \log \frac{p(\theta, y_{1:N})}{q_\eta(\theta)} d\theta \)
- Recall: Stochastic gradient
- Goal: \( \min_\eta f(\eta) := \mathbb{E}_Y f(Y, \eta) \)
  - One option: \( \min_\eta N^{-1} \sum_{n=1}^{N} f(y_n, \eta) \)
  - GD: \( \eta^{(t)} = \eta^{(t-1)} - \rho_t \nabla_\eta \left[ N^{-1} \sum_{n=1}^{N} f(y_n, \eta) \right]_{\eta=\eta^{(t)}} \)
  - SGD: \( \eta^{(t)} = \eta^{(t-1)} - \rho_t \left[ \nabla_\eta f(y_t, \eta) \right]_{\eta=\eta^{(t)}} \)
- Observe \( y_n \overset{\text{iid}}{\sim} Y \)
Stochastic gradient descent (SGD)

• MFVB: \[ \min_{\eta:q_{\eta} \in Q_{\text{MFVB}}} -\mathbb{E}_{q_{\eta}} \log \frac{p(\theta, y_{1:N})}{q_{\eta}(\theta)} d\theta \]

• Recall: Stochastic gradient

• Goal: \[ \min_{\eta} f(\eta) := \mathbb{E}_{Y} f(Y, \eta) \]

• One option: \[ \min_{\eta} N^{-1} \sum_{n=1}^{N} f(y_{n}, \eta) \]

• GD: \[ \eta^{(t)} = \eta^{(t-1)} - \rho_{t} \nabla_{\eta} \left[ N^{-1} \sum_{n=1}^{N} f(y_{n}, \eta) \right]_{\eta=\eta^{(t)}} \]

• SGD: \[ \eta^{(t)} = \eta^{(t-1)} - \rho_{t} \left[ \nabla_{\eta} f(y_{t}, \eta) \right]_{\eta=\eta^{(t)}} \]

• Can we apply SGD to our problem? Assume indep data
Stochastic gradient descent (SGD)

- MFVB: \( \min_{\eta: q_\eta \in Q_{MFVB}} -\mathbb{E}_{q_\eta} \log \frac{p(\theta, y_{1:N})}{q_\eta(\theta)} \, d\theta \)
- Recall: Stochastic gradient
- Goal: \( \min_{\eta} f(\eta) := \mathbb{E}_Y f(Y, \eta) \)
- One option: \( \min_{\eta} N^{-1} \sum_{n=1}^{N} f(y_n, \eta) \)
- GD: \( \eta^{(t)} = \eta^{(t-1)} - \rho_t \nabla_\eta \left[ N^{-1} \sum_{n=1}^{N} f(y_n, \eta) \right]_{\eta=\eta^{(t)}} \)
- SGD: \( \eta^{(t)} = \eta^{(t-1)} - \rho_t \left[ \nabla_\eta f(y_t, \eta) \right]_{\eta=\eta^{(t)}} \)
- Can we apply SGD to our problem? Assume indep data

\( y_n \overset{iid}{\sim} Y \)
Stochastic gradient descent (SGD)

- MFVB: \[ \min_{\eta:q_\eta \in Q_{\text{MFVB}}} \mathbb{E}_{q_\eta} \log \prod_{n=1}^{N} \frac{p(y_n | \theta)}{q_\eta(\theta)} \]  
  \[ \int \ d\theta \]

- Recall: Stochastic gradient

- Goal: \[ \min_\eta f(\eta) := \mathbb{E}_Y f(Y, \eta) \]  
  \[ \text{• Observe } y_n \overset{iid}{\sim} Y \]

- One option: \[ \min_\eta N^{-1} \sum_{n=1}^{N} f(y_n, \eta) \]

- GD: \[ \eta(t) = \eta(t-1) - \rho_t \nabla_\eta \left[ N^{-1} \sum_{n=1}^{N} f(y_n, \eta) \right]_{\eta=\eta(t)} \]

- SGD: \[ \eta(t) = \eta(t-1) - \rho_t \left[ \nabla_\eta f(y_t, \eta) \right]_{\eta=\eta(t)} \]

- Can we apply SGD to our problem? Assume indep data
Stochastic gradient descent (SGD)

- MFVB: \[
\min_{\eta: q_{\eta} \in Q_{MFVB}} \mathbb{E}_{q_{\eta}} \log \prod_{n=1}^{N} \frac{p(y_n | \theta)}{q_{\eta}(\theta)} \]

- Recall: Stochastic gradient

- Goal: \[
\min_{\eta} f(\eta) := \mathbb{E}_{Y} f(Y, \eta)
\]

- One option: \[
\min_{\eta} N^{-1} \sum_{n=1}^{N} f(y_n, \eta)
\]

- GD: \[
\eta^{(t)} = \eta^{(t-1)} - \rho_t \nabla_{\eta} \left[ N^{-1} \sum_{n=1}^{N} f(y_n, \eta) \right]_{\eta=\eta^{(t)}}
\]

- SGD: \[
\eta^{(t)} = \eta^{(t-1)} - \rho_t \left[ \nabla_{\eta} f(y_t, \eta) \right]_{\eta=\eta^{(t)}}
\]

- Can we apply SGD to our problem? Assume indep data
Stochastic gradient descent (SGD)

- MFVB: \( \min_{\eta: q_\eta \in Q_{\text{MFVB}}} \mathbb{E}_{q_\eta} \log \left( \prod_{n=1}^{N} p(y_n | \theta) \frac{p(\theta)}{q_\eta(\theta)} \right) \) \( d\theta \)
- Recall: Stochastic gradient
- Goal: \( \min_{\eta} f(\eta) := \mathbb{E}_{Y} f(Y, \eta) \) • Observe \( y_n \overset{iid}{\sim} Y \)
- One option: \( \min_{\eta} N^{-1} \sum_{n=1}^{N} f(y_n, \eta) \)
- GD: \( \eta^{(t)} = \eta^{(t-1)} - \rho_t \nabla_{\eta} \left[ N^{-1} \sum_{n=1}^{N} f(y_n, \eta) \right]_{\eta=\eta^{(t)}} \)
- SGD: \( \eta^{(t)} = \eta^{(t-1)} - \rho_t \left[ \nabla_{\eta} f(y_t, \eta) \right]_{\eta=\eta^{(t)}} \)
- Can we apply SGD to our problem? Assume indep data
  \( \min_{\eta} N^{-1} \sum_{n=1}^{N} \left[ -N \mathbb{E}_{q_\eta} \log p(y_n | \theta) - \mathbb{E}_{q_\eta} \log (p(\theta)/q_\eta(\theta)) \right] \)
Stochastic gradient descent (SGD)

- MFVB: \[ \min_{\eta: q_\eta \in Q_{\text{MFVB}}} \mathbb{E}_{q_\eta} \log \left[ \prod_{n=1}^{N} p(y_n | \theta) \frac{p(\theta)}{q_\eta(\theta)} \right] \ d\theta \]
- Recall: Stochastic gradient
- Goal: \( \min_{\eta} f(\eta) := \mathbb{E}_Y f(Y, \eta) \)
  - One option: \( \min_{\eta} N^{-1} \sum_{n=1}^{N} f(y_n, \eta) \)
  - GD: \( \eta^{(t)} = \eta^{(t-1)} - \rho_t \nabla_{\eta} \left[ N^{-1} \sum_{n=1}^{N} f(y_n, \eta) \right]_{\eta=\eta^{(t)}} \)
  - SGD: \( \eta^{(t)} = \eta^{(t-1)} - \rho_t \left[ \nabla_{\eta} f(y_t, \eta) \right]_{\eta=\eta^{(t)}} \)
- Can we apply SGD to our problem? Assume indep data
  \[ \min_{\eta} N^{-1} \sum_{n=1}^{N} \left[ -N \mathbb{E}_{q_\eta} \log p(y_n | \theta) - \mathbb{E}_{q_\eta} \log (p(\theta) / q_\eta(\theta)) \right] \]

\[ p(\theta | y_1) \]

\[ \theta \]
Stochastic gradient descent (SGD)

- MFVB: \( \min_{\eta:q_\eta \in Q_{MFVB}} \mathbb{E}_{q_\eta} \log \left[ \prod_{n=1}^{N} p(y_n | \theta) \frac{p(\theta)}{q_\eta(\theta)} \right] d\theta \)
- Recall: Stochastic gradient
- Goal: \( \min_{\eta} f(\eta) := \mathbb{E}_Y f(Y, \eta) \)
- One option: \( \min_{\eta} N^{-1} \sum_{n=1}^{N} f(y_n, \eta) \)
- GD: \( \eta^{(t)} = \eta^{(t-1)} - \rho_t \nabla_{\eta} \left[ N^{-1} \sum_{n=1}^{N} f(y_n, \eta) \right]_{\eta=\eta^{(t)}} \)
- SGD: \( \eta^{(t)} = \eta^{(t-1)} - \rho_t \left[ \nabla_{\eta} f(y_t, \eta) \right]_{\eta=\eta^{(t)}} \)
- Can we apply SGD to our problem? Assume indep data
  \( \min_{\eta} N^{-1} \sum_{n=1}^{N} \left[ -N \mathbb{E}_{q_\eta} \log p(y_n | \theta) - \mathbb{E}_{q_\eta} \log (p(\theta)/q_\eta(\theta)) \right] \)
Stochastic gradient descent (SGD)

- **MFVB:** \[ \min_{\eta: q_\eta \in \mathcal{Q}_{MFVB}} -\mathbb{E}_{q_\eta} \log \prod_{n=1}^{N} p(y_n | \theta) \frac{p(\theta)}{q_\eta(\theta)} \] 
  \[ \frac{d\theta}{d\theta} \]

- **Recall:** Stochastic gradient

- **Goal:** \[ \min_\eta f(\eta) := \mathbb{E}_Y f(Y, \eta) \]
  - **Observe** \( y_n \overset{iid}{\sim} Y \)
  - **One option:** \( \min_\eta N^{-1} \sum_{n=1}^{N} f(y_n, \eta) \)
  - **GD:** \( \eta^{(t)} = \eta^{(t-1)} - \rho_t \nabla_\eta \left[ N^{-1} \sum_{n=1}^{N} f(y_n, \eta) \right]_{\eta=\eta^{(t)}} \)
  - **SGD:** \( \eta^{(t)} = \eta^{(t-1)} - \rho_t \left[ \nabla_\eta f(y_t, \eta) \right]_{\eta=\eta^{(t)}} \)

- Can we apply SGD to our problem? Assume indep data \( \min_\eta N^{-1} \sum_{n=1}^{N} \left[ -N \mathbb{E}_{q_\eta} \log p(y_n | \theta) - \mathbb{E}_{q_\eta} \log (p(\theta) / q_\eta(\theta)) \right] \)

\[ p(\theta | y_1) \] \[ p(\theta | y_{1:10}) \] \[ p(\theta | y_{1:9,000}) \]
Stochastic gradient descent (SGD)

- MFVB: \[ \min_{\eta: q_\eta \in Q_{MFVB}} - \mathbb{E}_{q_\eta} \log \prod_{n=1}^{N} p(y_n | \theta) \frac{p(\theta)}{q_\eta(\theta)} \] d\theta

- Recall: Stochastic gradient

- Goal: \[ \min_\eta f(\eta) := \mathbb{E}_Y f(Y, \eta) \] • Observe \( y_n \overset{iid}{\sim} Y \)

- One option: \( \min_\eta N^{-1} \sum_{n=1}^{N} f(y_n, \eta) \)

- GD: \( \eta^{(t)} = \eta^{(t-1)} - \rho_t \nabla_\eta \left[ N^{-1} \sum_{n=1}^{N} f(y_n, \eta) \right]_{\eta = \eta^{(t)}} \)

- SGD: \( \eta^{(t)} = \eta^{(t-1)} - \rho_t \left[ \nabla_\eta f(y_t, \eta) \right]_{\eta = \eta^{(t)}} \)

- Can we apply SGD to our problem? Assume indep data \( \min_\eta N^{-1} \sum_{n=1}^{N} \left[ -N \mathbb{E}_{q_\eta} \log p(y_n | \theta) - \mathbb{E}_{q_\eta} \log (p(\theta) / q_\eta(\theta)) \right] \)

- ELBO not an estimate of the objective; is the objective

\[
p(\theta | y_1)
p(\theta | y_{1:10})
p(\theta | y_{1:9,000})
\]
Stochastic gradient descent (SGD)

• MFVB: \( \min_{\eta} \min_{q_\eta \in Q_{MFVB}} -\mathbb{E}_{q_\eta} \log \prod_{n=1}^{N} p(y_n | \theta) \frac{p(\theta)}{q_\eta(\theta)} \) 

• Recall: Stochastic gradient

• Goal: \( \min_{\eta} f(\eta) := \mathbb{E}_Y f(Y, \eta) \) • Observe \( y_n \overset{iid}{\sim} Y \)

• One option: \( \min_{\eta} N^{-1} \sum_{n=1}^{N} f(y_n, \eta) \)

• GD: \( \eta^{(t)} = \eta^{(t-1)} - \rho_t \nabla_\eta \left[ N^{-1} \sum_{n=1}^{N} f(y_n, \eta) \right]_{\eta=\eta^{(t)}} \)

• SGD: \( \eta^{(t)} = \eta^{(t-1)} - \rho_t \left[ \nabla_\eta f(y_t, \eta) \right]_{\eta=\eta^{(t)}} \)

• Can we apply SGD to our problem? Assume indep data \( \min_{\eta} N^{-1} \sum_{n=1}^{N} \left[ -N \mathbb{E}_{q_\eta} \log p(y_n | \theta) - \mathbb{E}_{q_\eta} \log(p(\theta)/q_\eta(\theta)) \right] \)

• ELBO not an estimate of the objective; \( is \) the objective
Stochastic gradient descent (SGD)

- MFVB: \( \min_{\eta:q_\eta \in Q_{MFVB}} \mathbb{E}_{q_\eta} \log \prod_{n=1}^{N} p(y_n | \theta) \frac{p(\theta)}{q_\eta(\theta)} \) \( d\theta \)
- Recall: Stochastic gradient
- Goal: \( \min_{\eta} f(\eta) := \mathbb{E}_Y f(Y, \eta) \)  
  - Observe \( y_n \overset{iid}{\sim} Y \)
- One option: \( \min_{\eta} N^{-1} \sum_{n=1}^{N} f(y_n, \eta) \)
- GD: \( \eta^{(t)} = \eta^{(t-1)} - \rho_t \nabla_{\eta} \left[ N^{-1} \sum_{n=1}^{N} f(y_n, \eta) \right]_{\eta=\eta^{(t)}} \)
- SGD: \( \eta^{(t)} = \eta^{(t-1)} - \rho_t \left[ \nabla_{\eta} f(y_t, \eta) \right]_{\eta=\eta^{(t)}} \)
- Can we apply SGD to our problem? Assume indep data
  \( \min_{\eta} N^{-1} \sum_{n=1}^{N} \left[ -N \mathbb{E}_{q_\eta} \log p(y_n | \theta) - \mathbb{E}_{q_\eta} \log(p(\theta)/q_\eta(\theta)) \right] \)
- ELBO not an estimate of the objective; is the objective
  \( f(\eta) = \mathbb{E}_Y f(Y, \eta) = -\mathbb{E}_{q_\eta} \log(p(\theta, y_1:N)/q_\eta(\theta)) \)
Stochastic gradient descent (SGD)

- MFVB: \( \min_{\eta:q_\eta \in Q_{\text{MFVB}}} \mathbb{E}_{q_\eta} \log \left[ \prod_{n=1}^{N} p(y_n | \theta) \frac{p(\theta)}{q_\eta(\theta)} \right] \)
- Recall: Stochastic gradient
- Goal: \( \min_\eta f(\eta) := \mathbb{E}_Y f(Y, \eta) \)
  - One option: \( \min_\eta N^{-1} \sum_{n=1}^{N} f(y_n, \eta) \)
  - GD: \( \eta^{(t)} = \eta^{(t-1)} - \rho_t \nabla_\eta \left[ N^{-1} \sum_{n=1}^{N} f(y_n, \eta) \right]_{\eta=\eta^{(t)}} \)
  - SGD: \( \eta^{(t)} = \eta^{(t-1)} - \rho_t \left[ \nabla_\eta f(y_t, \eta) \right]_{\eta=\eta^{(t)}} \)
- Can we apply SGD to our problem? Assume indep data \( \min_\eta N^{-1} \sum_{n=1}^{N} \left[ -N \mathbb{E}_{q_\eta} \log p(y_n | \theta) - \mathbb{E}_{q_\eta} \log (p(\theta)/q_\eta(\theta)) \right] \)
- ELBO not an estimate of the objective; is the objective \( f(\eta) = \mathbb{E}_Y f(Y, \eta) = -\mathbb{E}_{q_\eta} \log (p(\theta, y_1:N)/q_\eta(\theta)) \)
- \( f(\cdot, \eta) = -N \mathbb{E}_{q_\eta} \log p(\cdot | \theta) - \mathbb{E}_{q_\eta} \log (p(\theta)/q_\eta(\theta)) \)
- \( Y = y_n \) w.p. \( 1/N, \quad n \in \{1, \ldots, N\} \)
Stochastic gradient descent (SGD)

- MFVB: \( \min_{\eta:q_{\eta} \in Q_{\text{MFVB}}} \mathbb{E}_{q_{\eta}} \log \left[ \prod_{n=1}^{N} p(y_{n}|\theta) \frac{p(\theta)}{q_{\eta}(\theta)} \right] d\theta \)

- Recall: Stochastic gradient

- Goal: \( \min_{\eta} f(\eta) := \mathbb{E}_{Y} f(Y, \eta) \)  
  - Observe \( y_{n} \overset{iid}{\sim} Y \)

- One option: \( \min_{\eta} N^{-1} \sum_{n=1}^{N} f(y_{n}, \eta) \)

- GD: \( \eta^{(t)} = \eta^{(t-1)} - \rho_{t} \nabla_{\eta} \left[ N^{-1} \sum_{n=1}^{N} f(y_{n}, \eta) \right]_{\eta=\eta^{(t)}} \)

- SGD: \( \eta^{(t)} = \eta^{(t-1)} - \rho_{t} \left[ \nabla_{\eta} f(y_{t}, \eta) \right]_{\eta=\eta^{(t)}} \)

- Can we apply SGD to our problem? Assume indep data \( \min_{\eta} N^{-1} \sum_{n=1}^{N} \left[ -N \mathbb{E}_{q_{\eta}} \log p(y_{n}|\theta) - \mathbb{E}_{q_{\eta}} \log (p(\theta)/q_{\eta}(\theta)) \right] \)

- ELBO not an estimate of the objective; is the objective
  \[
  f(\eta) = \mathbb{E}_{Y} f(Y, \eta) = -\mathbb{E}_{q_{\eta}} \log (p(\theta, y_{1:N})/q_{\eta}(\theta))
  \]
  \[
  f(\cdot, \eta) = -N \mathbb{E}_{q_{\eta}} \log p(\cdot|\theta) - \mathbb{E}_{q_{\eta}} \log (p(\theta)/q_{\eta}(\theta))
  \]

- \( Y = y_{n} \) w.p. \( 1/N \), \( n \in \{1, \ldots, N\} \)

- Stochastic variational inference [Hoffman et al 2013]
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?

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

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

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

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

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

[Turner & Sahani 2011]

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

- Conjugate linear regression
What about uncertainty?

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

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

- Conjugate linear regression
- Bayesian central limit theorem
What about uncertainty?

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

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

- Conjugate linear regression
- Bayesian central limit theorem
What about uncertainty?

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

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

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

[Turner & Sahani 2011]
What about uncertainty?

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

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

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

[Turner & Sahani 2011]
What about uncertainty?

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

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

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

[Turner & Sahani 2011]
What about uncertainty?

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

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

- Underestimates variance (sometimes severely)
- No covariance estimates
- Conjugate linear regression
- Bayesian central limit theorem
- Exercise: derive exact (closed) form of $q^*$

[Turner & Sahani 2011]
What about uncertainty?

- Microcredit

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

- Microcredit

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

- Microcredit effect
- \( \tau \) mean: 3.08 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

[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 2017]
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 \( r_n \)

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

[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:
  $$y_n | \beta, z, \sigma^2 \sim \mathcal{N}(\beta^T x_n + z_{k(n)} r_n, \sigma^2)$$
  $$z_k | \rho^2 \sim \mathcal{N}(0, \rho^2)$$
  $$(\sigma^2)^{-1} \sim \text{Gamma}(a_{\sigma^2}, b_{\sigma^2})$$
  $$(\rho^2)^{-1} \sim \text{Gamma}(a_{\rho^2}, b_{\rho^2})$$

[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: $y_n|\beta, z, \sigma^2 \overset{indep}{\sim} \mathcal{N}(\beta^T x_n + z_k(n)r_n, \sigma^2)$
  
  
  $z_k|\rho^2 \overset{iid}{\sim} \mathcal{N}(0, \rho^2)$ 
  
  $(\sigma^2)^{-1} \overset{}{\sim} \text{Gamma}(a_{\sigma^2}, b_{\sigma^2})$

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

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

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

\[\text{MCMC mean} \quad \text{MFVB mean}\]

[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: 
  \[ y_n \mid \beta, z, \sigma^2 \overset{indep}{\sim} \mathcal{N}(\beta^T x_n + z_k(n) r_n, \sigma^2) \]
  \[ z_k \mid \rho^2 \overset{iid}{\sim} \mathcal{N}(0, \rho^2) \]
  \[ (\sigma^2)^{-1} \sim \text{Gamma}(a_{\sigma^2}, b_{\sigma^2}) \]
  \[ (\rho^2)^{-1} \sim \text{Gamma}(a_{\rho^2}, b_{\rho^2}) \]

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

[Giordano, Broderick, Jordan 2015]
What can we do?

• Good evaluation methods
What can we do?

• Good evaluation methods
  [Gorham, Mackey 2015, 2017]
  [Talts et al 2018]
  [Yao et al 2018]
  etc.
What can we do?

- Good evaluation methods
  
  [Gorham, Mackey 2015, 2017]
  
  [Talts et al 2018]
  
  [Yao et al 2018]
  
  etc.

- A correction to MFVB (Part II)

  [Giordano, Broderick, Jordan 2015, 2017; Giordano, Broderick, Meager, Huggins, Jordan 2016]
What can we do?

- Good evaluation methods
  - [Gorham, Mackey 2015, 2017]
  - [Talts et al 2018]
  - [Yao et al 2018]
  - etc.

- A correction to MFVB (Part II)
  - Also VB & robustness quantification

  - [Giordano, Broderick, Jordan 2015, 2017; Giordano, Broderick, Meager, Huggins, Jordan 2016]
What can we do?

• Good evaluation methods
  [Gorham, Mackey 2015, 2017]
  [Talts et al 2018]
  [Yao et al 2018]
  etc.

• A correction to MFVB (Part II)
  • Also VB & robustness quantification
    [Giordano, Broderick, Jordan 2015, 2017; Giordano, Broderick, Meager, Huggins, Jordan 2016]

• Data summarization for scalability (Part III)
  [Campbell, Broderick 2017, 2018]
  [Huggins, Campbell, Broderick 2016; Huggins, Adams, Broderick 2017]
References (1/6)


References (2/6)


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

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