Part IV: Variational Bayes and beyond

Tamara Broderick
ITT Career Development
Assistant Professor,
MIT

Bayesian inference

- Desiderata:
  - Point estimates, coherent uncertainties
  - Interpretable, complex, modular; prior information
Bayesian inference

- Desiderata:
  - Point estimates, coherent uncertainties
  - Interpretable, complex, modular; prior information
  - Challenge: existing methods can be slow, tedious, unreliable
Bayesian inference

- Desiderata:
  - Point estimates, coherent uncertainties
  - Interpretable, complex, modular; prior information
- Challenge: existing methods can be slow, tedious, unreliable
- Our proposal: use efficient data summaries for scalable, automated algorithms with error bounds for finite data
Roadmap

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

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

\[ p(\theta) \]
Bayesian inference

\[ p(y|\theta)p(\theta) \]
Bayesian inference

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

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

\[(x_n, y_n)\]
Bayesian inference

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

Normal

\[(x_n, y_n)\]

Phishing
Bayesian inference

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

\((x_n, y_n)\)

Normal

\begin{align*}
\text{Phishing} & \quad \theta \\
& \quad \cdot \cdot \cdot 
\end{align*}
Bayesian inference

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

$(x_n, y_n)$

Normal

Phishing
Bayesian inference

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

Normal

\[(x_n, y_n)\]

Phishing

Exact posterior

[Bishop 2006]
Bayesian inference

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

- MCMC: Eventually accurate but can be slow  
[Bardenet, Doucet, Holmes 2017]

(Bishop 2006)
Bayesian inference

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

- MCMC: Eventually accurate but can be slow
- (Mean-field) variational Bayes: (MF)VB

\((x_n, y_n)\)

[Normal]

\[ \theta \]

[Phishing]

[Exact posterior]

[Bishop 2006]

[Bardenet, Doucet, Holmes 2017]
Bayesian inference

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

- MCMC: Eventually accurate but can be slow
- (Mean-field) variational Bayes: (MF)VB
  - Fast

\[(x_n, y_n)\]

Normal

Phishing

Exact posterior

[Bishop 2006]

[Bardenet, Doucet, Holmes 2017]
Bayesian inference

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

- MCMC: Eventually accurate but can be slow
- (Mean-field) variational Bayes: (MF)VB
  - Fast, streaming, distributed

(Bishop 2006) [Bardenet, Doucet, Holmes 2017]
Bayesian inference

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

- MCMC: Eventually accurate but can be slow
- (Mean-field) variational Bayes: (MF)VB
  - Fast, streaming, distributed
    - (3.6M Wikipedia, 32 cores, \( \sim \)hour)

\((x_n, y_n)\)

Normal

\(\theta\)

Phishing

\(\theta_2\)

\(\theta_1\)

[Broderick, Boyd, Wibisono, Wilson, Jordan 2013]

[Bardenet, Doucet, Holmes 2017]
Bayesian inference

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

- MCMC: Eventually accurate but can be slow
- (Mean-field) variational Bayes: (MF)VB
  - Fast, streaming, distributed [Broderick, Boyd, Wibisono, Wilson, Jordan 2013]
  - (3.6M Wikipedia, 32 cores, ~hour)
- Misestimation & lack of quality guarantees

[MacKay 2003; Bishop 2006; Wang, Titterington 2004; Turner, Sahani 2011; Fosdick 2013; Dunson 2014; Bardenet, Doucet, Holmes 2017; Opper, Winther 2003; Giordano, Broderick, Jordan 2015, 2017]
Bayesian inference

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

- MCMC: Eventually accurate but can be slow
- (Mean-field) variational Bayes: (MF)VB
  - Fast, streaming, distributed \([\text{Bishop 2006}]\) \((3.6\text{M Wikipedia, 32 cores, } \sim\text{hour})\)
  - Misestimation & lack of quality guarantees \([\text{MacKay 2003; Bishop 2006; Wang, Titterington 2004; Turner, Sahani 2011; Fosdick 2013; Dunson 2014; Bardenet, Doucet, Holmes 2017; Opper, Winther 2003; Giordano, Broderick, Jordan 2015, 2017}]\)
Bayesian inference

- MCMC: Eventually accurate but can be slow
- (Mean-field) variational Bayes: (MF)VB
  - Fast, streaming, distributed
    - (3.6M Wikipedia, 32 cores, ~hour)
  - Misestimation & lack of quality guarantees

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

\[
\begin{align*}
\theta_1 \\
\text{MFVB} \\
\text{LRVB}
\end{align*}
\]

[Bishop 2006; MacKay 2003; Wang, Titterington 2004; Turner, Sahani 2011; Fosdick 2013; Dunson 2014; Bardenet, Doucet, Holmes 2017; Opper, Winther 2003; Giordano, Broderick, Jordan 2015, 2017]
Bayesian inference

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

- \textbf{MCMC: Eventually accurate but can be slow}
- \textbf{(Mean-field) variational Bayes: (MF)VB}
  - Fast, streaming, distributed \cite{Broderick2013}
    (3.6M Wikipedia, 32 cores, \sim{} hour)
- \textbf{Automation: e.g. Stan, NUTS, ADVI} \cite{http://mc-stan.org/, Hoffman2014, Kucukelbir2017}
Roadmap

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

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

• Observe: redundancies can exist even if data isn’t “tall”
Bayesian coresets

• Observe: redundancies can exist even if data isn’t “tall”
Bayesian coresets

- Observe: redundancies can exist even if data isn’t “tall”
Bayesian coresets

- Observe: redundancies can exist even if data isn’t “tall”
Bayesian coresets

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

• 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

[Agarwal et al 2005; Feldman & Langberg 2011]
Bayesian coresets

• 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
• Previous heuristics: data squashing, big data GPs

[Agarwal et al 2005; Feldman & Langberg 2011; DuMouchel et al 1999; Madigan et al 2002]
Bayesian coresets

• 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
• Previous heuristics: data squashing, big data GPs
• Cf. subsampling

[Agarwal et al 2005; Feldman & Langberg 2011; DuMouchel et al 1999; Madigan et al 2002]
Bayesian coresets

- 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
- Previous heuristics: data squashing, big data GPs
- Cf. subsampling

[Agarwal et al 2005; Feldman & Langberg 2011; DuMouchel et al 1999; Madigan et al 2002]
Bayesian coresets

- 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
- Previous heuristics: data squashing, big data GPs
- Cf. subsampling

[Agarwal et al 2005; Feldman & Langberg 2011; DuMouchel et al 1999; Madigan et al 2002]
Bayesian coresets

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

How to develop coresets for Bayes?

• Theoretical guarantees on quality
• Previous heuristics: data squashing, big data GPs
• Cf. subsampling

[Agarwal et al 2005; Feldman & Langberg 2011; DuMouchel et al 1999; Madigan et al 2002; Huggins, Campbell, Broderick 2016; Campbell, Broderick 2017; Campbell, Broderick 2018]
Bayesian coresets

- Posterior \( p(\theta | y) \propto_{\theta} p(y | \theta)p(\theta) \)
Bayesian coresets

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

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

- 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

\[ \|w\|_0 \ll N \]
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. \( \|w\|_0 \ll N \)
Bayesian coresets

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

- 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. \( \|w\|_0 \ll N \)

Bayesian coresets Normal Phishing \( \mathcal{L}(w) \) \( \mathcal{L} \)
Bayesian coresets

- Posterior \( p(\theta|y) \propto \theta p(y|\theta)p(\theta) \)
- 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: \( \|\mathcal{L}(w) - \mathcal{L}\| \leq \varepsilon \)

\[ \text{Normal} \quad \text{Phishing} \]

\[ \mathcal{L}(w) \quad \mathcal{L} \]
Bayesian coresets

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

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

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

- Approximate posterior close in Wasserstein distance
  \[
  d_{W,j}(p_w(\cdot|y), p(\cdot|y)) \leq C_j \|\mathcal{L}(w) - \mathcal{L}\|_{WFID}, \quad j \in \{1, 2\}
  \]
Roadmap

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

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

- Normal
- Phishing

- Might miss important data
Uniform subsampling revisited

- Might miss important data
Uniform subsampling revisited

- Normal
- Phishing

• Might miss important data
Uniform subsampling revisited

- Normal
- Phishing

- Might miss important data
Uniform subsampling revisited

- Might miss important data
Uniform subsampling revisited

- Might miss important data
Uniform subsampling revisited

- Normal
- Phishing

- Might miss important data
Uniform subsampling revisited

- Might miss important data
Uniform subsampling revisited

- Normal
- Phishing

- Might miss important data
Uniform subsampling revisited

- Normal
- Phishing
- Might miss important data
Uniform subsampling revisited

- Normal
- Phishing

- Might miss important data
- Noisy estimates
Uniform subsampling revisited

- Might miss important data
- Noisy estimates

$M = 10$

Normal

Phishing
Uniform subsampling revisited

- Might miss important data
- Noisy estimates

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

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

• Approximate Bayes review
• 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
Importance sampling

- Normal
- Phishing
Importance sampling
Importance sampling
Importance sampling

Normal

Phishing
Importance sampling
Importance sampling

- Normal
- Phishing
Importance sampling
Importance sampling
Importance sampling

\[ \sigma_n \propto \| L_n \| \]
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
Importance sampling

\[ \sigma := \sum_{n=1}^{N} ||Ł_n|| \]

\[ \sigma_n := ||Ł_n|| / \sigma \]

1. data  
2. importance weights
Importance sampling

1. data
2. importance weights
3. importance sample

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

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

\[ \sigma := \sum_{n=1}^{N} ||L_n|| \]

\[ \sigma_n := ||L_n|| / \sigma \]

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

**Thm sketch (CB).** \( \delta \in (0,1) \). W.p. \( \geq 1 - \delta \), 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 sketch (CB).** \( \delta \in (0,1) \). W.p. \( \geq 1 - \delta \), 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

\[ M = 10 \]
Importance sampling

**Thm sketch (CB).** $\delta \in (0,1)$. W.p. \(\geq 1 - \delta\), 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
Hilbert coresets

- Want a good coreset: \[ \min_{w \in \mathbb{R}^N} \| \mathcal{L}(w) - \mathcal{L} \| \]

\[ \text{s.t. } w \geq 0, \|w\|_0 \leq M \]
Hilbert coresets

• Want a good coreset: \( \min_{w \in \mathbb{R}^N} \| \mathcal{L}(w) - \mathcal{L} \| \)

\[ \text{s.t. } w \geq 0, \|w\|_0 \leq M \]
Hilbert coresets

- Want a good coreset: \( \min_{w \in \mathbb{R}^N} ||\mathcal{L}(w) - \mathcal{L}|| \)

\[ \text{s.t. } w \geq 0, ||w||_0 \leq M \]
Hilbert coresets

• Want a good coreset: \( \min_{w \in \mathbb{R}^N} \| \mathcal{L}(w) - \mathcal{L} \| \)

\[ \text{s.t. } w \geq 0, \|w\|_0 \leq M \]

• need to consider (residual) error direction
Hilbert coresets

• Want a good coreset: \( \min_{w \in \mathbb{R}^N} \| \mathcal{L}(w) - \mathcal{L} \| \)

\[
\text{s.t. } w \geq 0, \|w\|_0 \leq M
\]

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

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

- Approximate Bayes review
- 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 $D$

[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

[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

[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 $D$

- 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 $D$

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

[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:

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

$$\Delta^{N-1} := \left\{ w \in \mathbb{R}^N : \sum_{n=1}^{N} \sigma_n w_n = \sigma, w \geq 0 \right\}$$
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 sketch (CB). After $M$ iterations,

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

• 10K pts; norms, inference: closed-form

Uniform subsampling

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

- 10K pts; norms, inference: closed-form

Uniform subsampling

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

- 10K pts; norms, inference: closed-form

Uniform subsampling

Importance sampling

\[ M = 5 \quad M = 50 \quad M = 500 \]
Gaussian model (simulated)

- 10K pts; norms, inference: closed-form

Uniform subsampling

Importance sampling

Frank-Wolfe

\[ M = 5 \quad M = 50 \quad M = 500 \]
Gaussian model (simulated)

- 10K pts; norms, inference: closed-form

![Graph showing KL divergence with M on the x-axis and natural log of KL divergence on the y-axis. The graph includes lines for Rand, Unif, IS, and FW, indicating lower error as M increases.]
Logistic regression (simulated)

- 10K pts; general inference

Uniform subsampling

Importance sampling

Frank-Wolfe

\[ M = 10 \quad M = 100 \quad M = 1000 \]
Poisson regression (simulated)

- 10K pts; general inference

Uniform subsampling

Importance sampling

Frank-Wolfe

\[ M = 10 \quad M = 100 \quad M = 1000 \]
Real data experiments

Data sets include:

- Phishing
- Chemical reactivity
- Bicycle trips
- Airport delays

Uniform subsampling
Frank Wolfe coresets

lower error
less total time

[Campbell, Broderick 2017]
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

10

[Campbell, Broderick 2017, 2018]
Roadmap

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

- Approximate Bayes review
- 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

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

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

- Sufficient statistics

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

- Exponential family likelihood

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

- 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

- *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) \]

• 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

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

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

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

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

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

[Huggins, Adams, Broderick 2017]
Conclusions

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

• Get more accurate with more computation investment
• Coresets
• Approx. suff. stats
Conclusions

- *Data summarization* for **scalable, automated** approx. Bayes algorithms with **error bounds on quality for finite data**

- Get more accurate with more computation investment
- Coresets
- Approx. suff. stats
- A start
  - Lots of potential improvements/ directions
Conclusions

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

• Get more accurate with more computation investment

• Coresets

• Approx. suff. stats

• A start

• Lots of potential improvements/directions

[Campbell, Broderick 2018]
References (1/5)


References (2/5)


D Dunson. Robust and scalable approach to Bayesian inference. Talk at *ISBA* 2014.


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)
Practicalities
Practicalities

• Choice of norm
Practicalities

- Choice of norm
  - E.g. (weighted) Fisher information distance

\[
\| \mathcal{L}(w) - \mathcal{L} \|_{\hat{\pi}, F}^2 := \mathbb{E}_{\hat{\pi}} \left[ \| \nabla \mathcal{L}(\theta) - \nabla \mathcal{L}(w, \theta) \|_2^2 \right]
\]
Practicalities

- Choice of norm
  - E.g. (weighted) Fisher information distance

\[
\| \mathcal{L}(w) - \mathcal{L} \|_{F,\hat{\pi}}^2 := \mathbb{E}_{\hat{\pi}} \left[ \| \nabla \mathcal{L}(\theta) - \nabla \mathcal{L}(w, \theta) \|_2^2 \right]
\]

- Associated inner product:

\[
\langle \mathcal{L}_n, \mathcal{L}_m \rangle_{\hat{\pi},F} := \mathbb{E}_{\hat{\pi}} \left[ \nabla \mathcal{L}_n(\theta)^T \nabla \mathcal{L}_m(\theta) \right]
\]
Practicalities

• Choice of norm
  • E.g. (weighted) Fisher information distance
    \[ \| \mathcal{L}(w) - \mathcal{L} \|_{\hat{\pi}, F}^2 := \mathbb{E}_{\hat{\pi}} \left[ \| \nabla \mathcal{L}(\theta) - \nabla \mathcal{L}(w, \theta) \|_2^2 \right] \]
  • Associated inner product:
    \[ \langle \mathcal{L}_n, \mathcal{L}_m \rangle_{\hat{\pi}, F} := \mathbb{E}_{\hat{\pi}} \left[ \nabla \mathcal{L}_n(\theta)^T \nabla \mathcal{L}_m(\theta) \right] \]
Practicalities

- Choice of norm
  - E.g. (weighted) Fisher information distance

\[
\| \mathcal{L}(w) - \mathcal{L} \|_{\hat{\pi}, F}^2 := \mathbb{E}_{\hat{\pi}} \left[ \| \nabla \mathcal{L}(\theta) - \nabla \mathcal{L}(w, \theta) \|_2^2 \right]
\]

- Associated inner product:

\[
\langle \mathcal{L}_n, \mathcal{L}_m \rangle_{\hat{\pi}, F} := \mathbb{E}_{\hat{\pi}} \left[ \nabla \mathcal{L}_n(\theta)^T \nabla \mathcal{L}_m(\theta) \right]
\]

- Random feature projection
Practicalities

- Choice of norm
  - E.g. (weighted) Fisher information distance
    \[
    \| \mathcal{L}(w) - \mathcal{L} \|_{\tilde{\pi}, F}^2 := \mathbb{E}_{\tilde{\pi}} \left[ \| \nabla \mathcal{L}(\theta) - \nabla \mathcal{L}(w, \theta) \|_2^2 \right]
    \]

- Associated inner product:
  \[
  \langle \mathcal{L}_n, \mathcal{L}_m \rangle_{\tilde{\pi}, F} := \mathbb{E}_{\tilde{\pi}} \left[ \nabla \mathcal{L}_n(\theta)^T \nabla \mathcal{L}_m(\theta) \right]
  \]

- Random feature projection
  \[
  \langle \mathcal{L}_n, \mathcal{L}_m \rangle_{\tilde{\pi}, F} \approx \frac{D}{J} \sum_{j=1}^{J} \left( \nabla \mathcal{L}_n(\theta_j) \right)_{d_j} \left( \nabla \mathcal{L}_m(\theta_j) \right)_{d_j},
  \]
  \[
  d_j \overset{iid}{\sim} \text{Unif}\{1, \ldots, D\}, \quad \theta_j \overset{iid}{\sim} \tilde{\pi}
  \]
Practicalities

- Choice of norm
  - E.g. (weighted) Fisher information distance
    \[ \| \mathcal{L}(w) - \mathcal{L} \|^2_{\hat{\pi}, F} := \mathbb{E}_{\hat{\pi}} \left[ \| \nabla \mathcal{L}(\theta) - \nabla \mathcal{L}(w, \theta) \|^2 \right] \]
  - Associated inner product:
    \[ \langle \mathcal{L}_n, \mathcal{L}_m \rangle_{\hat{\pi}, F} := \mathbb{E}_{\hat{\pi}} \left[ \nabla \mathcal{L}_n(\theta)^T \nabla \mathcal{L}_m(\theta) \right] \]
- Random feature projection
  \[ \langle \mathcal{L}_n, \mathcal{L}_m \rangle_{\hat{\pi}, F} \approx \frac{D}{J} \sum_{j=1}^{J} \langle \nabla \mathcal{L}_n(\theta_j) \rangle_{d_j} \langle \nabla \mathcal{L}_m(\theta_j) \rangle_{d_j}, \]
  \[ d_j \overset{iid}{\sim} \text{Unif}\{1, \ldots, D\}, \theta_j \overset{iid}{\sim} \hat{\pi} \]

**Thm sketch (CB).** With high probability and large enough J, a good coreset after random feat. proj. is a good coreset for \((\mathcal{L}_n)_{n=1}^N\).
Full pipeline

$N$

dataset size
Full pipeline

cost $\hat{\pi}$

$N$

dataset size
Full pipeline

random feature projection

$\text{cost } \hat{\pi}$

$N$

dataset size

$J$

projection dim
Full pipeline

random feature projection

\[ O(NJ) \]

+ cost \[ \hat{\pi} \]

\( N \)
dataset size

\( J \)
projection dim
Full pipeline

random feature projection

$O(NJ)$

Frank-Wolfe

$O(NJM)$

+ cost $\hat{\pi}$

$N$ dataset size

$M$ coreset size

$J$ projection dim
Full pipeline

random feature projection

$O(NJ)$

Frank-Wolfe

$O(NJM)$

MCMC

$O(MT)$

+ cost $\hat{\pi}$

$N$ dataset size

$M$ coreset size

$J$ projection dim

$T$ MCMC steps
Full pipeline

\[ O(NJ) + \text{cost } \hat{\pi} \]

- \( N \): dataset size
- \( M \): coreset size
- \( J \): projection dim
- \( T \): MCMC steps

\[ \text{vs. } O(NT) \]
Full pipeline

- vs. $O(NT)$
- Can make streaming, distributed