Big Bayes without sub-sampling bias:
Paths of Partial Posteriors

Heiko Strathmann

Gatsby Unit, University College London

22nd May 2015

Joint work

Being Bayesian: Averaging beliefs of the unknown

$$\phi = \int d\theta \varphi(\theta) \underbrace{p(\theta|\mathcal{D})}_{\text{posterior}}$$

where $p(\theta|\mathcal{D}) \propto \underbrace{p(\mathcal{D}|\theta)}_{\text{likelihood data}} \underbrace{p(\theta)}_{\text{prior}}$

Markov Chains

- Problem: Need iid $\theta^{(j)} \sim p(\theta|\mathcal{D})$. Hard!
- But can construct a Markov chain

$$\theta^{(0)} \to \theta^{(1)} \to \theta^{(2)} \to \dots$$

whose stationary distribution is $p(\theta|\mathcal{D})$, *i.e.*,

$$\lim_{j\to\infty} \theta^{(j)} \sim p(\theta|\mathcal{D})$$

and break dependence of the $\theta^{(j)}$ by thinning.

Metropolis Hastings Transition Kernel

Target $\pi(\theta) \propto p(\theta|\mathcal{D})$

- At iteration $j+1$, state $\theta^{(j)}$
- Propose $\theta' \sim q\left(\theta|\theta^{(j)}\right)$
- Accept $\theta^{(j+1)} \leftarrow \theta'$ with probability

$$\min\left(\frac{\pi(\theta')}{\pi(\theta^{(j)})} \times \frac{q(\theta^{(j)}|\theta')}{q(\theta'|\theta^{(j)})}, 1\right)$$

- Reject $\theta^{(j+1)} \leftarrow \theta^{(j)}$ otherwise.

Big $\mathcal{D}$ & MCMC

- Need to evaluate

$$\pi(\theta) \propto p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta)$$

  in every iteration.
- For example, for $\mathcal{D} = \{x_1, \ldots, x_N\}$,

$$p(\mathcal{D}|\theta) = \prod_{i=1}^{N} p(x_i|\theta)$$

- Infeasible for growing N
- Lots of current research: Can we use subsets of $\mathcal{D}$?

Alternative transition kernels

Existing methods construct alternative transition kernels.

(Welling & Teh 2011), (Korattikara, Chen, Welling 2014), (Bardenet, Doucet, Holmes 2014)
(Maclaurin & Adams 2014), (Chen, Fox, Guestrin 2014).

They

- use mini-batches
- inject noise
- augment the state space
- make clever use of approximations

Problem: Most methods

- are biased (in asymptotic sense)
- have no convergence guarantees
- mix badly

Desiderata for Bayesian estimators in Big Data

1. Computational costs sub-linear in N
2. No bias. Hard! No additional bias (compared to MCMC)
3. Finite & controllable variance

Reminder: Where we came from – expectations

$$\mathbb{E}_{p(\theta|\mathcal{D})}\{\varphi(\theta)\} \qquad \varphi : \Theta \to \mathbb{R}$$

Idea: Assuming the goal is estimation, give up on simulation.

Outline

Partial Posterior Path Estimators

Experiments & Extensions

Discussion

Idea

1. Construct partial posterior distributions
2. Compute partial expectations (biased)
3. Remove sub-sampling bias

Note:

- No access to $p(\theta|\mathcal{D})$
- Partial posterior inference less challenging
- Exploit existing methodology & engineering
- Not restricted to MCMC

Partial Posterior Paths

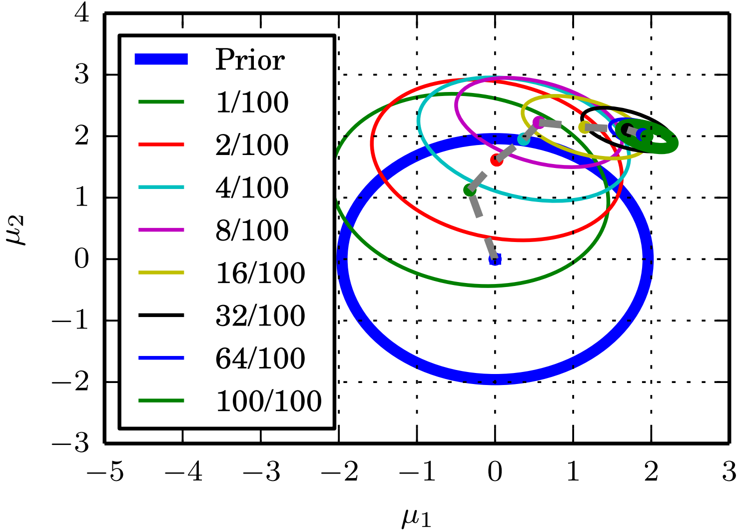
- Model $p(x,\theta) = p(x|\theta)p(\theta)$, data $\mathcal{D} = \{x_1, \ldots, x_N\}$
- Full posterior $\pi_N := p(\theta|\mathcal{D}) \propto p(x_1, \ldots, x_N|\theta)p(\theta)$

- L subsets $\mathcal{D}_l$ of sizes $|\mathcal{D}_l| = n_l$
- Here: $n_1 = a, n_2 = 2^1 a, n_3 = 2^2 a, \ldots, n_L = 2^{L-1} a$
- Partial posterior $\tilde{\pi}_l := p(\mathcal{D}_l|\theta) \propto p(\mathcal{D}_l|\theta)p(\theta)$

- Path from prior to full posterior

$$p(\theta) = \tilde{\pi}_0 \to \tilde{\pi}_1 \to \tilde{\pi}_2 \to \cdots \to \tilde{\pi}_L = \pi_N = p(\mathcal{D}|\theta)$$

Gaussian Mean, Conjugate Prior



Partial posterior path statistics

For partial posterior paths

$$p(\theta) = \tilde{\pi}_0 \to \tilde{\pi}_1 \to \tilde{\pi}_2 \to \cdots \to \tilde{\pi}_L = \pi_N = p(\mathcal{D}|\theta)$$

define a sequence $\{\phi_t\}_{t=1}^{\infty}$ as

$$\begin{aligned} \phi_t &:= \hat{\mathbb{E}}_{\tilde{\pi}_t}\{\varphi(\theta)\} \qquad & t < L \\ \phi_t := \phi &:= \hat{\mathbb{E}}_{\pi_N}\{\varphi(\theta)\} \qquad & t \geq L \end{aligned}$$

This gives

$$\phi_1 \to \phi_2 \to \cdots \to \phi_L = \phi$$

$\hat{\mathbb{E}}_{\tilde{\pi}_t}\{\varphi(\theta)\}$ is empirical estimate. Not necessarily MCMC.

Debiasing Lemma (Rhee & Glynn 2012, 2014)

- ϕ and $\{\phi_t\}_{t=1}^{\infty}$ real-valued random variables. Assume

$$\lim_{t\to\infty} \mathbb{E}\left\{|\phi_t - \phi|^2\right\} = 0$$

- T integer rv with $\mathbb{P}[T \geq t] > 0$ for $t \in \mathbb{N}$
- Assume

$$\sum_{t=1}^{\infty} \frac{\mathbb{E}\left\{|\phi_{t-1} - \phi|^2\right\}}{\mathbb{P}[T \geq t]} < \infty$$

- Unbiased estimator of $\mathbb{E}\{\phi\}$

$$\phi = \phi_\infty = \sum_{t=1}^{\infty} \phi_t - \phi_{t-1} \qquad\qquad \phi_T^* = \sum_{t=1}^{T} \frac{\phi_t - \phi_{t-1}}{\mathbb{P}[T \geq t]}$$

Computational complexity

- Recall for posterior paths: $\phi_{t+1} = \phi_t = \phi$ for $t \geq L$
- Assume geometric batch size increase n_t and truncation probabilities for $1 \leq t \leq L$

$$\Lambda_t := \mathbb{P}(T = t) = 2^{-\alpha t} \qquad \alpha \in (0, 1)$$

- Average computational cost sub-linear in N

$$\mathcal{O}\left(a \left(\frac{N}{a}\right)^{1-\alpha}\right)$$

Variance-computation tradeoffs in Big Data

Fixed N: Variance finite by construction

$$\mathbb{E}\left\{(\phi_T^*)^2\right\} = \sum_{t=1}^{\infty} \frac{\mathbb{E}\left\{|\phi_{t-1} - \phi|^2\right\} - \mathbb{E}\left\{|\phi_t - \phi|^2\right\}}{\mathbb{P}\left[T \geq t\right]}$$

If we assume $\forall t \leq L$, there is a constant c and $\beta > 0$ s.t.

$$\mathbb{E}\left\{|\phi_{t-1} - \phi|^2\right\} \leq \frac{c}{n_t^\beta}$$

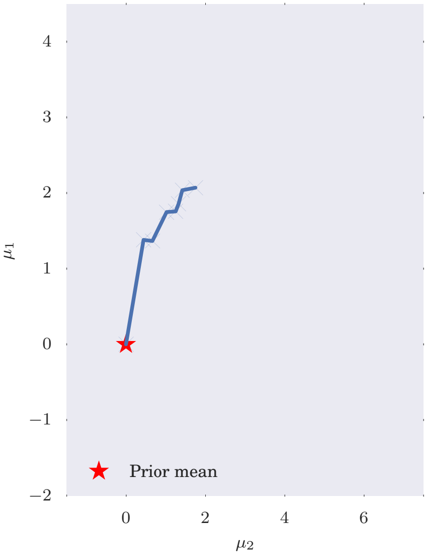
and furthermore $\alpha < \beta$, then

$$\sum_{t=1}^{L} \frac{\mathbb{E}\left\{|\phi_{t-1} - \phi|^2\right\}}{\mathbb{P}\left[T \geq t\right]} = \mathcal{O}(1)$$

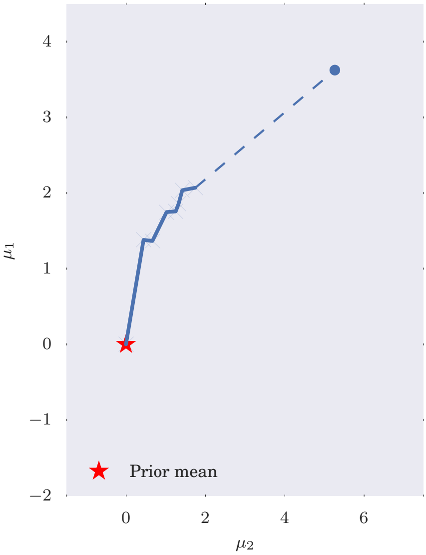
and variance stays bounded as $N \to \infty$.

Algorithm illustration

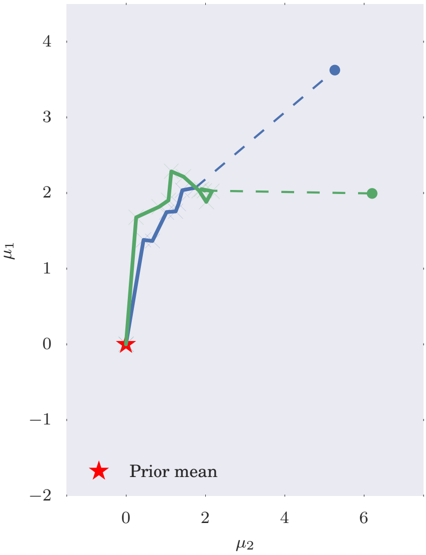
Algorithm illustration



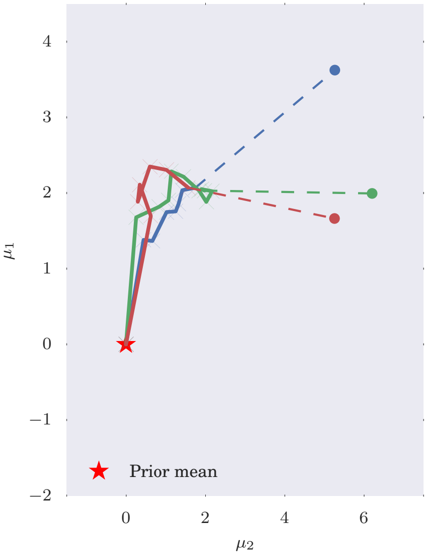
Algorithm illustration



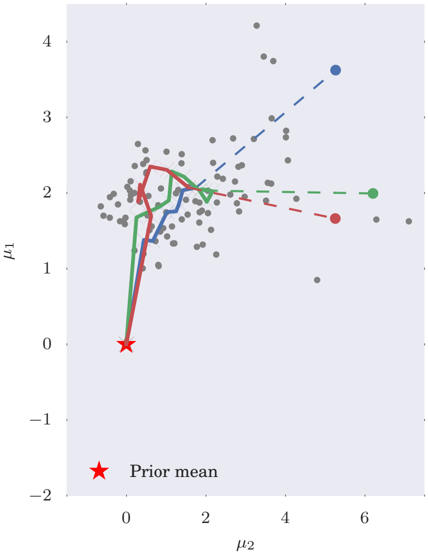
Algorithm illustration



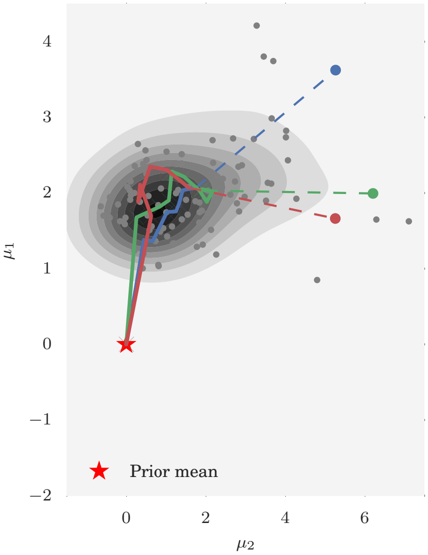
Algorithm illustration



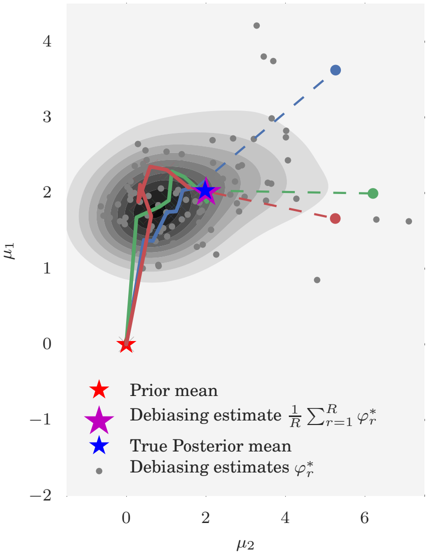
Algorithm illustration



Algorithm illustration

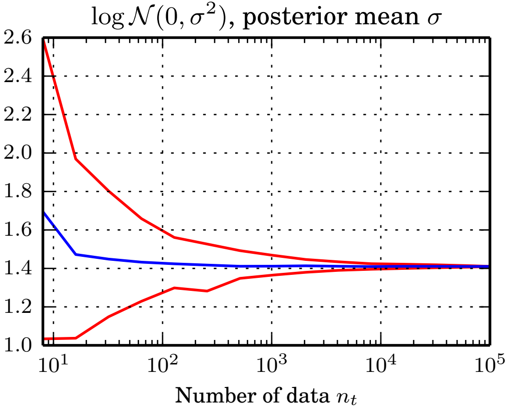


Algorithm illustration



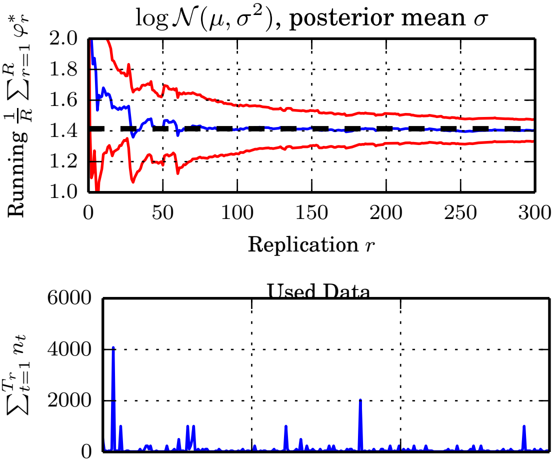
Outline

Synthetic log-Gaussian



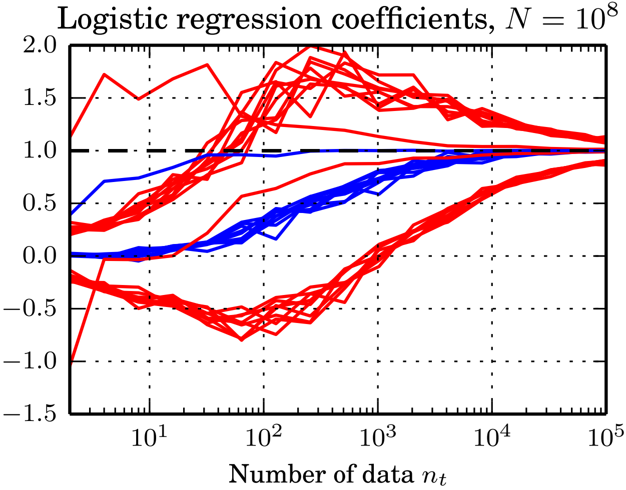
- (Bardenet, Doucet, Holmes 2014) – all data
- (Korattikara, Chen, Welling 2014) – wrong result

Synthetic log-Gaussian – debiasing

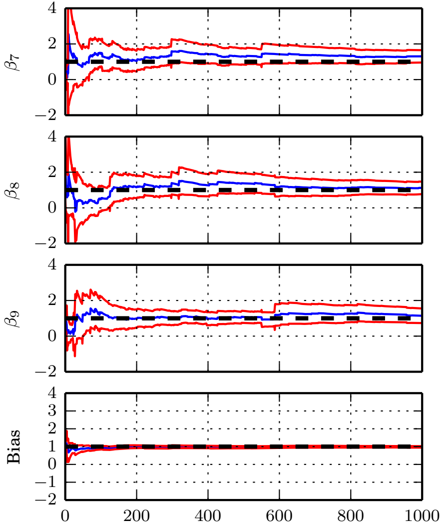


- Truly large-scale version: $N \approx 10^8$
- Sum of likelihood evaluations: $\approx 0.25N$

Large-scale synthetic logistic regression



Large-scale synthetic logistic regression



- Sum of likelihood evaluations: $\approx 9N$

Non-factorising likelihoods

No need for

$$p(\mathcal{D}|\theta) = \prod_{i=1}^{N} p(x_i|\theta)$$

Example: Approximate Gaussian Process regression

- Estimate predictive mean

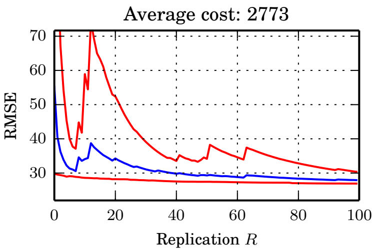
$$k_*^\top (K + \lambda I)^{-1} y$$

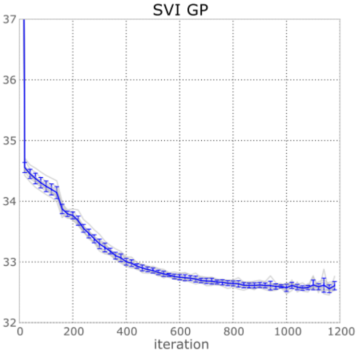
- Vanilla computational costs: $\mathcal{O}(N^3)$
- Finite rank kernel expansion: $\mathcal{O}(m^2 N)$
- Combined with debiasing $\mathcal{O}\left(m^2 N^{1-\alpha}\right)$, sub-linear
- No MCMC (!)

Gaussian Processes for Big Data

(Hensman, Fusi, Lawrence, 2013): SVI & inducing variables

- Airtime delays, $N = 700,000$, $D = 8$
- $m = 1000$ random Fourier features (Rahimi, Recht, 2007)
- Estimate predictive mean on $100,000$ test data





Outline

Conclusions

If goal is estimation rather than simulation, we arrive at

1. Data complexity sub-linear in N
2. No sub-sampling bias (in addition to MCMC)
3. Finite & controllable variance

Practical:

- Not limited to MCMC
- Not limited to factorising likelihoods
- Competitive initial results
- Parallelisable, re-uses existing engineering effort

Still biased?

MCMC and finite time

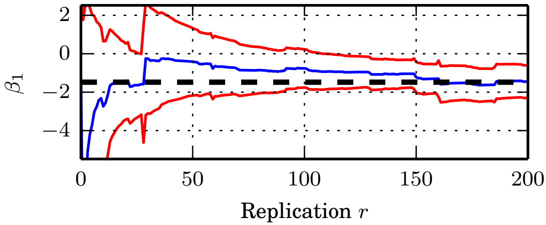
- MCMC estimator $\hat{\mathbb{E}}_{\tilde{\pi}_t}\{\varphi(\theta)\}$ is not unbiased
- Could imagine two-stage process
  - Apply debiasing to MC estimator
  - Use to debias partial posterior path
- Need conditions on MC convergence to control variance, (Rhee & Glynn 2012, Agapiou, Roberts, Vollmer, 2014)
- Hard! Even possible?

Memory restrictions

- Partial posterior expectations need be computable
- Memory limitations cause bias
- e.g. large-scale GMRF (Lyne et al, 2014)

Free lunch? Not uniformly better than MCMC

- Need $\mathbb{P}[T \geq t] > 0$ for all $t \leq L$
- This includes the whole dataset
- Negative example: a9a dataset (Welling & Teh, 2011)
- $N \approx 32,000$
- Converges, but full posterior sampling likely



The two extremes of Big Data

Xi'an's og, Feb 2015: Discussion of M. Betancourt's note on HMC and subsampling.

"...the information provided by the whole data is only available when looking at the whole data."

See `http://goo.gl/bFQvd6`

We claim:

The transition from highly redundant data on trivial models to sparse data on complex models is continuous!

Thank you

Questions?