Unbiased Bayes for Big Data:
Paths of Partial Posteriors

Heiko Strathmann

Gatsby Unit, University College London

Oxford ML lunch, February 25, 2015

Joint work

Being Bayesian: Averaging beliefs of the unknown

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

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

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

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

  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 $\mathcal{D}$?

Desiderata for Bayesian estimators

1. No (additional) bias
2. Finite & controllable variance
3. Computational costs sub-linear in N
4. No problems with transition kernel design

Outline

Literature Overview

Partial Posterior Path Estimators

Experiments & Extensions

Discussion

Outline

Literature Overview

Partial Posterior Path Estimators

Experiments & Extensions

Discussion

Stochastic gradient Langevin (Welling & Teh 2011)

$$\theta' = \frac{\epsilon}{2}\left(\nabla_{\theta=\theta^{(j)}} \log p(\theta) + \nabla_{\theta=\theta^{(j)}} \sum_{i=1}^{N} \log p(x_i|\theta)\right) + \eta_j$$

Two changes:

1. Noisy gradients with mini-batches. Let $\mathcal{I} \subseteq \{1, \ldots, N\}$ and use log-likelihood gradient

$$\nabla_{\theta=\theta^{(j)}} \sum_{i \in \mathcal{I}} \log p(x_i|\theta)$$

2. Don't evaluate MH ratio, but always accept, decrease step-size/noise $\epsilon_j \to 0$ to compensate

$$\sum_{i=1}^{\infty} \epsilon_i = \infty \qquad \sum_{i=1}^{\infty} \epsilon_i^2 < \infty$$

Austerity (Korattikara, Chen, Welling 2014)

- Idea: rewrite MH ratio as hypothesis test
- At iteration j, draw $u \sim \texttt{Uniform}[0,1]$ and compute

$$\mu_0 = \frac{1}{N} \log \left[u \times \frac{p(\theta^{(j)})}{p(\theta')} \times \frac{q(\theta'|\theta^{(j)})}{q(\theta^{(j)}|\theta')} \right]$$

$$\mu = \frac{1}{N} \sum_{i=1}^{N} l_i \qquad l_i := \log p(x_i|\theta') - \log p(x_i|\theta^{(j)})$$

- Accept if $\mu > \mu_0$; reject otherwise
- Subsample the l_i, central limit theorem, t-test
- Increase data if no significance, multiple testing correction

Bardenet, Doucet, Holmes 2014

Similar to Austerity, but with analysis:

- Concentration bounds for MH (CLT might not hold)
- Bound for probability of wrong decision

For uniformly ergodic original kernel

- Approximate kernel converges
- Bound for TV distance of approximation and target

Limitations:

- Still approximate
- Only random walk
- Uses all data on hard (?) problems

Firefly MCMC (Maclaurin & Adams 2014)

- First asymptotically exact MCMC kernel using sub-sampling
- Augment state space with binary indcator variables
- Only few data "bright"
- Dark points approximated by a lower bound on likelihood

Limitations:

- Bound might not be available
- Loose bounds $\rightarrow$ worse than standard MCMC$\rightarrow$ need MAP estimate
- Linear in N. Likelihood evaluations at least $q_{\text{dark}\rightarrow\text{bright}} \cdot N$
- Mixing time cannot be better than $1/q_{\text{dark}\rightarrow\text{bright}}$

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
- have no convergence guarantees
- mix badly

Reminder: Where we came from – expectations

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

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

Outline

Idea Outline

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

Note:

- No simulation from $p(\theta|\mathcal{D})$
- Partial posterior expectations less challenging
- Exploit standard MCMC methodology & engineering
- But not restricted to MCMC

Disclaimer

Goal is not to replace posterior sampling, but to provide a ...

- different perspective when the goal is estimation

Method does not do uniformly better than MCMC, but ...

- we show cases where computational gains can be achieved

Partial Posterior Paths

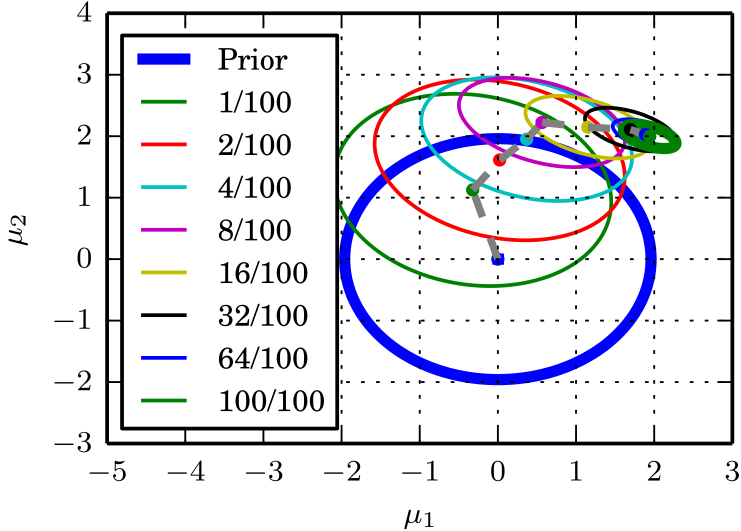
- Model $p(x, \theta) = p(x|\theta)p(\theta)$, data $\mathcal{D} = \{x_1, \ldots, x_N\}$
- Full posterior $\pi_N := p(\theta|\mathcal{D}) \propto p(x_1, \ldots, x_N|\theta)p(\theta)$

- 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

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

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

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

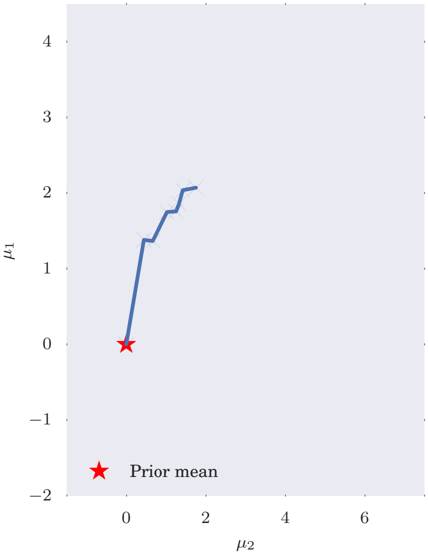
- Unbiased estimator of $\mathbb{E}\{\phi\}$

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

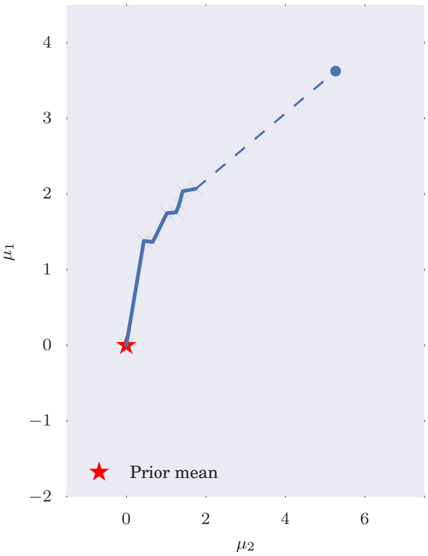
- Here: $\mathbb{P}[T \geq t] = 0$ for $t > L$ since $\phi_{t+1} - \phi_t = 0$

Algorithm illustration

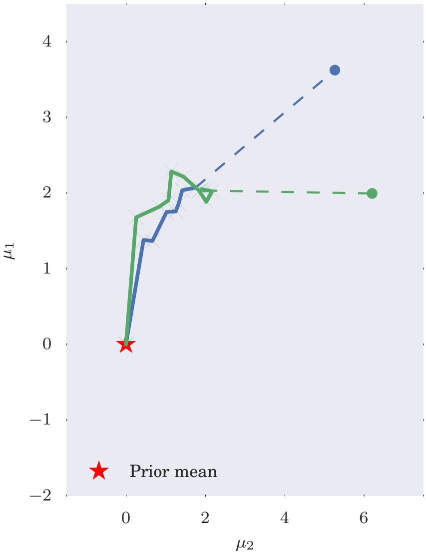
Algorithm illustration



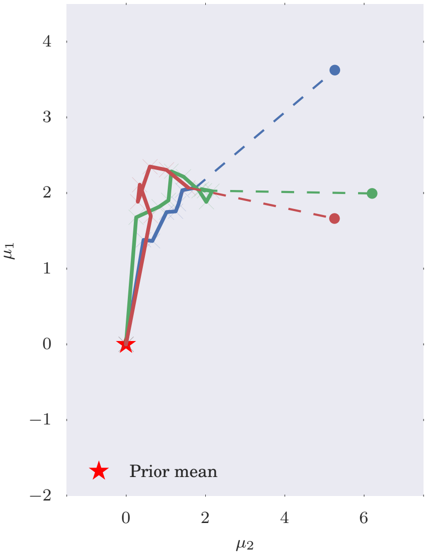
Algorithm illustration



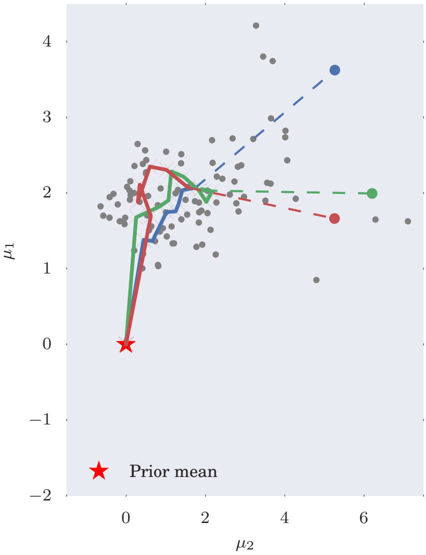
Algorithm illustration



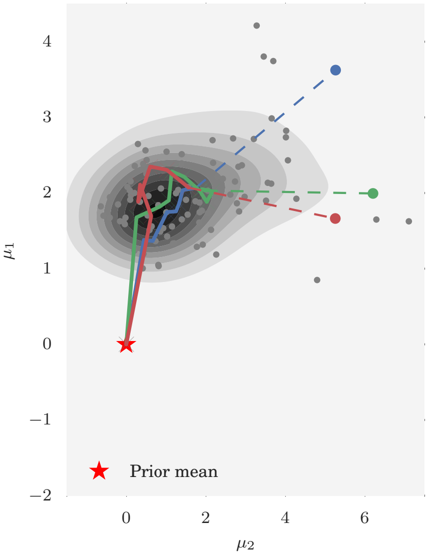
Algorithm illustration



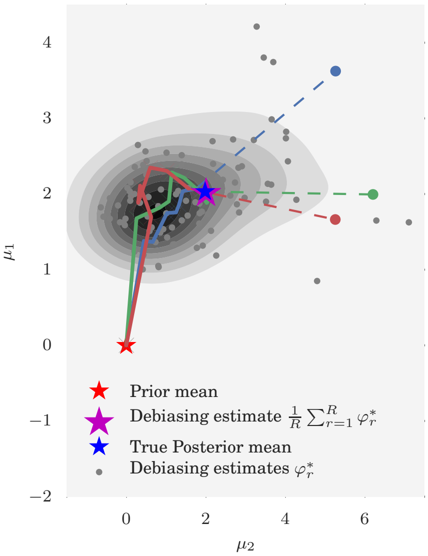
Algorithm illustration



Algorithm illustration



Algorithm illustration



Computational complexity

Assume geometric batch size increase n_t and truncation probabilities

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

Average computational cost sub-linear

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

Variance-computation tradeoffs in Big Data

Variance

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

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

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

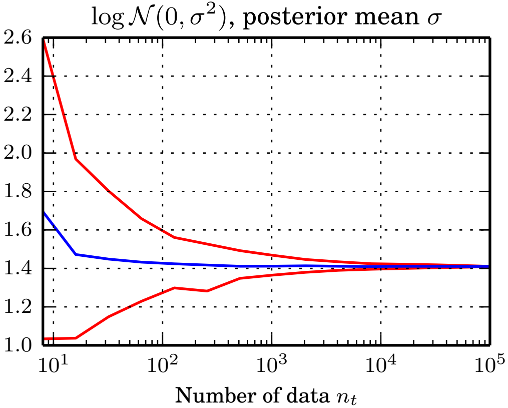
and furthermore $\alpha < \beta$, then

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

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

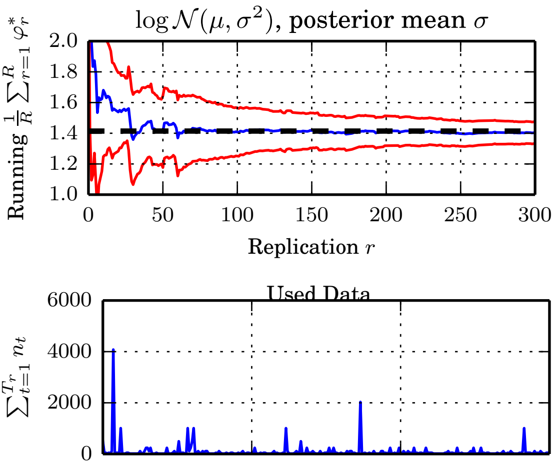
Outline

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: $\approx 0.25N$

Non-factorising likelihoods

No need for

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

Example: Approximate Gaussian Process regression

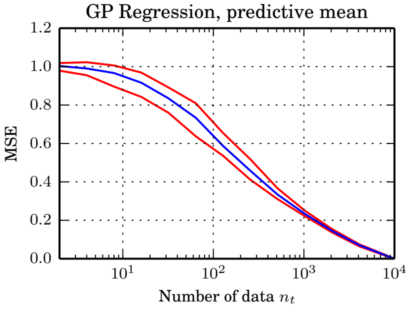
- Estimate predictive mean

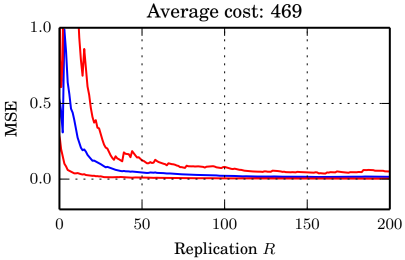
$$k_*^\top (K + \lambda I)^{-1} y$$

- No MCMC (!)

Toy example

- $N = 10^4, D = 1$
- $m = 100$ random Fourier features (Rahimi, Recht, 2007)
- Predictive mean on 1000 test data





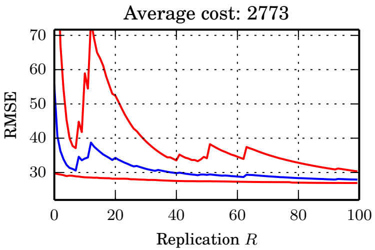
MSE Convergence

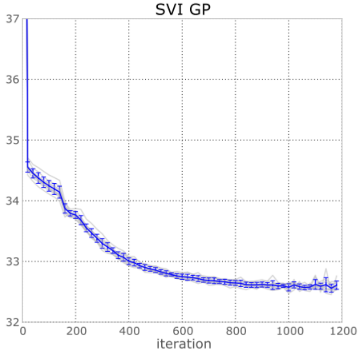
Debiasing

Gaussian Processes for Big Data

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

- Airtime delays, $N = 700,000$, $D = 8$
- Estimate predictive mean on $100,000$ test data





Outline

Conclusions

If goal is estimation rather than simulation, we arrive at

1. No bias
2. Finite & controllable variance
3. Data complexity sub-linear in N
4. No problems with transition kernel design

Practical:

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

Still biased?

MCMC and finite time

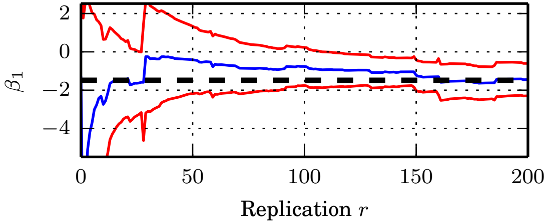
- MCMC estimator $\hat{\mathbb{E}}_{\tilde{\pi}_t}\{\varphi(\theta)\}$ is not unbiased
- Could imagine two-stage process
  - Apply debiasing to MC estimator
  - Use to debias partial posterior path
- Need conditions on MC convergence to control variance, (Agapiou, Roberts, Vollmer, 2014)

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
- Negative example: a9a dataset (Welling & Teh, 2011)
- $N \approx 32,000$
- Converges, but full posterior sampling likely



- Useful for very large (redundant) datasets

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`

Thank you

Questions?