Distributed Bayesian Posterior Sampling via Moment Sharing

Minjie Xu[1], Balaji Lakshminarayanan[2], Yee Whye Teh[3], Jun Zhu[1], and Bo Zhang[1]

[1]Dept. of CST, Tsinghua University; [2]Gatsby Unit, University College London; [3]Dept. of Stats, University of Oxford

Neural Information Processing Systems Foundation

CONTRIBUTIONS

We proposed a **distributed MCMC inference** algorithm for large scale **Bayesian posterior simulation** that

- scales to "big data" and compute nodes,
- converges fast and is exempt from a final combination stage,
- supports flexible distributed schemes: synchronous, asynchronous, decentralized, etc.,
- incurs low communication costs,
- and achieves high approximation accuracy.

We **empirically** studied the performance of our algorithm and **compared** it with the state-of-the-art, on Bayesian logistic regression and sparse linear regression with a spike-and-slab prior.

MOTIVATION

Bayesian inference with "Big Data" $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^N$

$$p(\boldsymbol{\theta}|\mathcal{D}) \propto p(\mathcal{D}, \boldsymbol{\theta}) = p_0(\boldsymbol{\theta}) \cdot \prod_{i=1}^N p(\mathbf{x}_i|\boldsymbol{\theta})$$

- Variational inference: use $q(\boldsymbol{\theta}) \in \mathcal{Q}$.

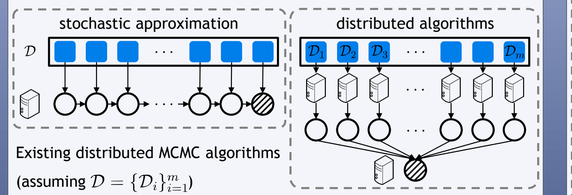
$$\mathcal{L}(q) = \mathbb{E}_q[\log p(\mathcal{D}, \boldsymbol{\theta})] + H(q) \quad \text{(evidence lower bound)}$$

- MCMC sampling: use $\Theta = \{\boldsymbol{\theta}^{(t)}\}_{t=1}^T$.

$$A(\boldsymbol{\theta}^{(t)} \to \boldsymbol{\theta}') = \min\left(1, \frac{p(\boldsymbol{\theta}'|\mathcal{D})}{p(\boldsymbol{\theta}^{(t)}|\mathcal{D})} \frac{q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}')}{q(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(t)})}\right) \quad \text{(acceptance rate)}$$

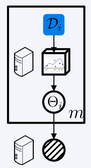
In both cases, the cost of one iteration (sample) is $O(N)$.

To scale up, people generally resort to two types of approaches, namely stochastic approximation (Welling'11, Hoffman'13, etc.) and distributed algorithms (Scott'13, Neiswanger'14, etc.)



Existing distributed MCMC algorithms (assuming $\mathcal{D} = \{\mathcal{D}_i\}_{i=1}^m$)

- partition the prior into $p_0(\boldsymbol{\theta})^{1/m}$, one for each node;
- let each node draw T samples from its local posterior:
$$\Theta_i = \{\boldsymbol{\theta}_i^{(t)}\}_{t=1}^T, \ \boldsymbol{\theta}_i^{(t)} \sim p_0(\boldsymbol{\theta})^{1/m} p(\mathcal{D}_i|\boldsymbol{\theta});$$
- and finally combines $\{\Theta_i\}_{i=1}^m$ into the full posterior samples.



Pros: *"embarrassingly parallel"* → simple and "efficient"

Cons: completely independent and local → **not very accurate**

SAMPLING VIA MOMENT SHARING (SMS)

Key idea: to encourage mutual awareness and agreement among the local samplers so as to improve inference quality, by enforcing **sharing** across them a small number of **moment statistics** of the local posteriors.

Local posterior: $\tilde{p}_i(\boldsymbol{\theta}|\mathcal{D}_i) \propto q(\boldsymbol{\theta}; \eta_{-i}) p(\mathcal{D}_i|\boldsymbol{\theta})$.

1. each local sampler independently draws samples from it;
2. moments of interest: $\mathbb{E}_{\tilde{p}_i(\boldsymbol{\theta}|\mathcal{D}_i)}[S(\boldsymbol{\theta})]$ for some sufficient statistics function $S(\boldsymbol{\theta})$;
3. the effective local prior $q(\boldsymbol{\theta}; \eta_{-i})$ is assumed a member of the exponential family with sufficient statistics $S(\boldsymbol{\theta})$ and natural parameter η_{-i};
4. **moment sharing**: find η_{-i} so that $\mathbb{E}_{\tilde{p}_i(\boldsymbol{\theta}|\mathcal{D}_i)}[S(\boldsymbol{\theta})] = \mu$ for some shared μ.

And that's where expectation propagation comes into play.

Expectation propagation (Minka'01): variational approximation via moment matching

1. approximate each $p(\mathcal{D}_i|\boldsymbol{\theta})$ with $q(\boldsymbol{\theta}; \eta_i)$ (use $q(\boldsymbol{\theta}; \eta_0)$ for $p_0(\boldsymbol{\theta})$)
2. posterior is thus approximated with $q(\boldsymbol{\theta}; \nu)$, where $\nu = \eta_0 + \sum_{i=1}^m \eta_i$
3. iteratively solve for each η_i as

$$\operatorname{argmin}_{\eta_i} \mathrm{KL}\big(q(\boldsymbol{\theta}; \eta_{-i}) p(\mathcal{D}_i|\boldsymbol{\theta}) \big\| q(\boldsymbol{\theta}; \eta_{-i}) q(\boldsymbol{\theta}; \eta_i)\big) \qquad (\eta_{-i} = \nu - \eta_i = \eta_0 + \textstyle\sum_{j \neq i} \eta_j)$$

which equates **matching the moments** of the two arguments in $\mathrm{KL}(\cdot\|\cdot)$.

Combining the above two ideas, we come up with our SMS algorithm.

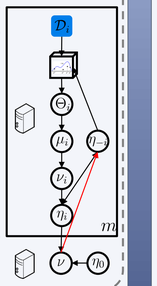
The following shows one iteration in the SMS algorithm. Upon convergence, $\nu_i = \nu$, $\mu_i = \mu$, and SMS outputs the collection of samples $\{\Theta_i\}_{i=1}^m$ from the last iteration as well as the EP estimated posterior $q(\boldsymbol{\theta}; \nu)$.

Algorithm 1: UPDATE (ON NODE i)

Input: current global natural parameter ν.
Output: updated local natural parameter η_i.

1 compute $\eta_{-i} \leftarrow \nu - \eta_i$
2 draw T samples $\Theta_i = \{\boldsymbol{\theta}_i^{(t)}\}_{t=1}^T$ from $\tilde{p}_i(\boldsymbol{\theta}|\mathcal{D})$ with any chosen sampler where $\tilde{p}_i(\boldsymbol{\theta}|\mathcal{D}) \propto q(\boldsymbol{\theta}; \eta_{-i}) p(\mathcal{D}_i|\boldsymbol{\theta})$
3 compute the empirical moment parameter $\mu_i \leftarrow \frac{1}{T}\sum_{t=1}^T S(\boldsymbol{\theta}_i^{(t)})$
4 transform μ_i into its natural parameter $\nu_i \leftarrow \nabla_\mu F^*(\mu_i)$ — **matching the moments!**
5 update $\eta_i \leftarrow \nu_i - \eta_{-i}$
6 **return** η_i

On master node: update $\nu = \eta_0 + \sum_{i=1}^m \eta_i$



Multivariate Gaussian family: motivated by the Bernstein-von Mises Theorem for big $\mathcal{D}_i$

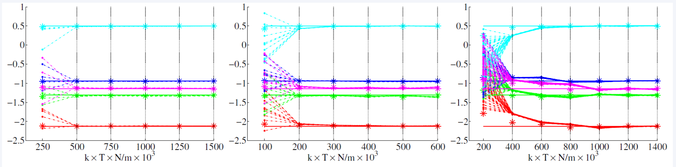
1. sufficient statistics: $S(\boldsymbol{\theta}) = (\boldsymbol{\theta}, \boldsymbol{\theta}\boldsymbol{\theta}^\top)$
2. moment parameter: $\mu_i = \frac{1}{T}\sum_{t=1}^T \boldsymbol{\theta}_i^{(t)}$, $\Sigma_i = \frac{1}{T-1}\sum_{t=1}^T (\boldsymbol{\theta}_i^{(t)} - \mu_i)(\boldsymbol{\theta}_i^{(t)} - \mu_i)^\top$
3. natural parameter: $\nu_i = (\Omega_i \mu_i, \Omega_i)$ where
$$\Omega_i = \tfrac{T-d-2}{T-1}\Sigma_i^{-1}$$

is an unbiased estimate of the precision matrix.

EMPIRICAL RESULTS

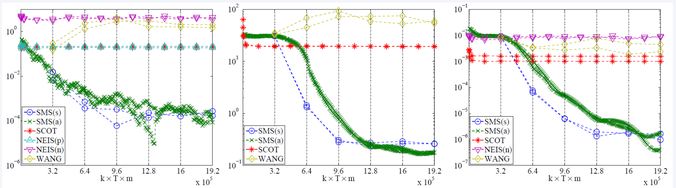
Bayesian logistic regression (synthetic dataset: $N = 4000,\ d = 20$)

- base sampler: NUTS (Hoffman'14), burnin = 400, thinning = 2.
- ground truth: 100,000 samples from $p(\boldsymbol{\theta}|\mathcal{D})$



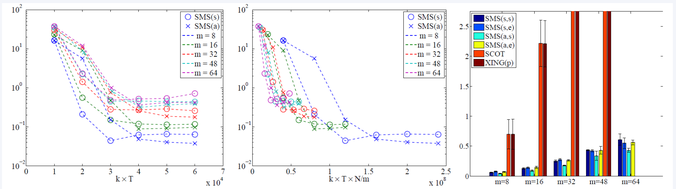
(a) $m = 4, T = 1000$ (b) $m = 10, T = 1000$ (c) $m = 50, T = 10000$

Figure 2: Convergence of local posterior means on a smaller Bayesian logistic regression dataset ($N = 1000, d = 5$). The x-axis indicates the number of likelihood evaluations, with vertical lines denoted EP iteration numbers. The y-axis indicates the estimated posterior means (dimensions indicated by different colours). We show ground truth with solid horizontal lines, the EP estimated mean with asterisks, and local sample estimated means dots connected with dash lines.



(a) MSE of posterior mean (b) Approximate KL-divergence (c) MSE of conditional prob. (5)

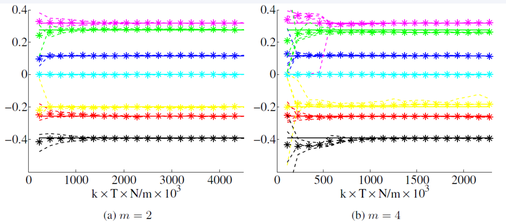
Figure 3: Errors (log-scale) against the cumulative number of samples drawn on all nodes (kTm). We tested two random splits of the dataset (hence 2 curves for each algorithm). Each complete EP iteration is highlighted by a vertical grid line.



(a) Approximate KL-divergence (b) Approximate KL-divergence (c) Approximate KL-divergence

Figure 4: Cross comparison with different numbers of nodes. Note that the x-axes have different meanings. In figure (a), it is the cumulative number of samples drawn locally on each node (kT). For the asynchronous **SMS(a)**, we only plot every m iterations so as to mimic the behaviour of **SMS(s)** for a more direct comparison. In figure (b) however, it is the cumulative number of likelihood evaluations on each node (kTN/m), which more accurately reflect computation time.

Bayesian sparse linear regression with a spike-and-slab prior



(a) $m = 2$ (b) $m = 4$

Figure 5: Results on Boston housing dataset for Bayesian sparse linear regression model with spike and slab prior. The x-axis plots the number of data points per node (equals the number of likelihood evaluations per sample) times the cumulative number of samples drawn per node, which is a surrogate for the computation times of the methods. The y-axis plots the ground truth (solid), local sample estimated means (dashed) and EP estimated mean (asterisks) at every iteration.

Algorithm costs

Algorithm	Local sampling (per node)	Communication (overall)
SCOT	KT samples $\boldsymbol{\theta} \sim p_i(\boldsymbol{\theta})$	$\{\Theta_i\}_{i=1}^m$ $(1 \cdot O(mdKT))$
NEIS(p)	KT samples $\boldsymbol{\theta} \sim p_i(\boldsymbol{\theta})$	$\{\eta_i\}_{i=1}^m$ $(1 \cdot O(m\|S(\boldsymbol{\theta})\|))$
NEIS(n)	KT samples $\boldsymbol{\theta} \sim p_i(\boldsymbol{\theta})$	$\{\Theta_i\}_{i=1}^m$ $(1 \cdot O(mdKT))$
WANG	K' iterations, each time KT/K' samples $\boldsymbol{t} \sim \mathcal{N}(\boldsymbol{t}\|\boldsymbol{\theta}, \mathbb{H}) p_i(\boldsymbol{t})$	all the local samples $(K' \cdot O(mdKT/K'))$
SMS	K iterations, each time T samples $\boldsymbol{\theta} \sim q(\boldsymbol{\theta}; \eta_{-i}) p(\mathcal{D}_i\|\boldsymbol{\theta})$	$\{\eta_i\}_{i=1}^m$ for K iterations $(K \cdot O(m\|S(\boldsymbol{\theta})\|))$

where $p_i(\boldsymbol{\theta}) \propto p(\mathcal{D}_i|\boldsymbol{\theta}) p_0(\boldsymbol{\theta})^{1/m}$, $\mathbb{H} = \mathrm{diag}(h_1, \ldots, h_d)$ and $|S(\boldsymbol{\theta})| = O(d^2)$ for multivariate Gaussians.

REFERENCES

1. **[Welling'11]** Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 681-688, 2011.

2. **[Hoffman'13]** Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303-1347, 2013.

3. **[Scott'13]** Steven L Scott, Alexander W Blocker, Fernando V Bonassi, Hugh A Chipman, Edward I George, and Robert E McCulloch. Bayes and big data: The consensus Monte Carlo algorithm. *EFaBBayes 250 conference*, 16, 2013.

4. **[Neiswanger'14]** Willie Neiswanger, Chong Wang, and Eric Xing. Asymptotically exact, embarrassingly parallel MCMC. In *Proceedings of the 30th International Conference on Uncertainty in Artificial Intelligence (UAI-14)*, pages 623-632, 2014.

5. **[Minka'01]** Thomas P Minka. A family of algorithms for approximate Bayesian inference. PhD thesis, Massachusetts Institute of Technology, 2001.

6. **[Hoffman'14]** Matthew D Hoffman and Andrew Gelman. The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15:1593-1623, 2014.

ACKNOWLEDGEMENT

We thank Willie Neiswanger for sharing his implementation of NEIS(n), and Michalis K Titsias for sharing the code used in spike-and-slab sampling. MX, JZ and BZ gratefully acknowledge funding from the National Basic Research Program of China (No. 2013CB329403) and National NSF of China (Nos. 61322308, 61332007). BL gratefully acknowledges generous funding from the Gatsby charitable foundation. YWT gratefully acknowledges EPSRC for research funding through grant EP/K009362/1.