Patterns of Scalable Bayesian Inference

Elaine Angelino
UC Berkeley
elaine@eecs.berkeley.edu

Matthew James Johnson
Harvard University
mattjj@csail.mit.edu

Ryan P. Adams
Harvard University and Twitter
rpa@seas.harvard.edu

the essence of knowledge
Boston — Delft

Foundations and Trends® in Machine Learning

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

E. Angelino, M. J. Johnson, and R. P. Adams. *Patterns of Scalable Bayesian Inference*. Foundations and Trends® in Machine Learning, vol. 9, no. 2-3, pp. 119–247, 2016.

This Foundations and Trends® issue was typeset in LaTeX using a class file designed by Neal Parikh. Printed on acid-free paper.

ISBN: 978-1-68083-219-8

Foundations and Trends® in Machine Learning
Volume 9, Issue 2-3, 2016

Editorial Scope

Topics

Foundations and Trends® in Machine Learning publishes survey and tutorial articles on the theory, algorithms and applications of machine learning, including the following topics:

- Adaptive control and signal processing
- Applications and case studies
- Behavioral, cognitive, and neural learning
- Bayesian learning
- Classification and prediction
- Clustering
- Data mining
- Dimensionality reduction
- Evaluation
- Game theoretic learning
- Graphical models
- Independent component analysis
- Inductive logic programming
- Kernel methods
- Markov chain Monte Carlo
- Model choice
- Nonparametric methods
- Online learning
- Optimization
- Reinforcement learning
- Relational learning
- Robustness
- Spectral methods
- Statistical learning theory
- Variational inference
- Visualization

Information for Librarians

Foundations and Trends® in Machine Learning, 2016, Volume 9, 6 issues. ISSN paper version 1935-8237. ISSN online version 1935-8245. Also available as a combined paper and online subscription.

Foundations and Trends® in Machine Learning
Vol. 9, No. 2-3 (2016) 119–247

DOI: 10.1561/2200000052

the essence of knowledge

Patterns of Scalable Bayesian Inference

Elaine Angelino*
UC Berkeley
elaine@eecs.berkeley.edu

Matthew James Johnson*
Harvard University
mattjj@csail.mit.edu

Ryan P. Adams
Harvard University and Twitter
rpa@seas.harvard.edu

* Authors contributed equally.

Contents

Abstract

Datasets are growing not just in size but in complexity, creating a demand for rich models and quantification of uncertainty. Bayesian methods are an excellent fit for this demand, but scaling Bayesian inference is a challenge. In response to this challenge, there has been considerable recent work based on varying assumptions about model structure, underlying computational resources, and the importance of asymptotic correctness. As a result, there is a zoo of ideas with a wide range of assumptions and applicability.

In this paper, we seek to identify unifying principles, patterns, and intuitions for scaling Bayesian inference. We review existing work on utilizing modern computing resources with both MCMC and variational approximation techniques. From this taxonomy of ideas, we characterize the general principles that have proven successful for designing scalable inference procedures and comment on the path forward.

E. Angelino, M. J. Johnson, and R. P. Adams. *Patterns of Scalable Bayesian Inference*. Foundations and Trends® in Machine Learning, vol. 9, no. 2-3, pp. 119–247, 2016.
DOI: 10.1561/2200000052.

1

Introduction

We have entered a new era of scientific discovery, in which computational insights are being integrated with large-scale statistical data analysis to enable researchers to ask both grander and more subtle questions about our natural world. This viewpoint asserts that we need not be limited to the narrow hypotheses that can be framed by traditional small-scale analysis techniques. Supporting new kinds of data-driven queries, however, requires that new methods be developed for statistical inference that can *scale up* along multiple axes — more samples, more dimensions, and greater model complexity — as well as *scale out* by taking advantage of modern parallel compute environments.

There are a variety of methodological frameworks for statistical inference; here we are concerned with the Bayesian formalism. In the Bayesian setting, inference queries are framed as interrogations of a posterior distribution over parameters, missing data, and other unknowns. By treating these unobserved quantities as random variables and conditioning on observed data, the Bayesian aims to make inferences and quantify uncertainty in a way that can coherently incorporate new data and other sources of information.

Coherently managing probabilistic uncertainty is central to Bayesian analysis, and so the computations associated with most inference tasks — estimation, prediction, hypothesis testing — are typically integrations. In some special situations it is possible to perform such integrations exactly, for example by taking advantage of tractable prior distributions and conjugacy in the prior-likelihood pair, or by using dynamic programming when the dependencies between random variables are relatively simple. Unfortunately, many inference problems are not amenable to these exact integration procedures, and so most of the interest in Bayesian computation focuses on methods of approximate inference.

There are two dominant paradigms for approximate inference in Bayesian models: Monte Carlo sampling methods and variational approximations. The Monte Carlo approach observes that integrations performed to query posterior distributions can be framed as expectations, and thus estimated with samples; such samples are most often generated via simulation from carefully designed Markov chains. Variational inference instead seeks to compute these integrals by approximating the posterior distribution with a more tractable alternative, finding the best approximation with powerful optimization algorithms.

In this paper, we examine how these techniques can be scaled up to larger problems and scaled out across parallel computational resources. This is not an exhaustive survey of a rapidly-evolving area of research; rather, we seek to identify the main ideas and themes that are emerging in this area, and articulate what we believe are some of the significant open questions and challenges.

1.1 Why be Bayesian with big data?

The Bayesian paradigm is fundamentally about integration: integration computes posterior estimates and measures of uncertainty, eliminates nuisance variables or missing data, and averages models to compute predictions or perform model comparison. While some statistical methods, such as MAP estimation, can be described from a Bayesian perspective, in which case the prior serves simply as a regularizer in an optimization problem, such methods are not inherently or exclusively Bayesian. Posterior integration is the distinguishing characteristic of Bayesian statistics, and so a defense of Bayesian ideas in the big data regime rests on the utility of integration.

The big data setting might seem to be precisely where integration isn't so important: as the dataset grows, shouldn't the posterior distribution concentrate towards a point mass? If big data means we end up making predictions using concentrated posteriors, why not focus on point estimation and avoid the specification of priors and the burden of approximate integration? These objections certainly apply to settings where the number of parameters is small and fixed ("tall data"). However, many models of interest have many parameters ("wide data"), or indeed have a number of parameters that grows along with the amount of data.

For example, an Internet company making inferences about its users' viewing and buying habits may have terabytes of data in total but only a few observations for its newest customers, the ones most important to impress with personalized recommendations. Moreover, it may wish to adapt its model in an online way as data arrive, a task that benefits from calibrated posterior uncertainties [Stern et al., 2009]. As another example, consider a healthcare company. As its dataset grows, it might hope to make more detailed and complex inferences about populations while also making careful predictions with calibrated uncertainty for each patient, even in the presence of massive missing data [Lawrence, 2015]. These scaling issues also arise in astronomy, where hundreds of billions of light sources, such as stars, galaxies, and quasars, each have latent variables that must be estimated from very weak observations, and are coupled in a large hierarchical model [Regier et al., 2015]. In Microsoft Bing's sponsored search advertising, predictive probabilities inform the pricing in the keyword auction mechanism. This problem nevertheless must be solved at scale, with tens of millions of impressions per hour [Graepel et al., 2010].

These are the regimes where big data can be small [Lawrence, 2015] and the number and complexity of statistical hypotheses grows with the data. The Bayesian inference methods we survey in this paper may provide solutions to these challenges.

1.2 The accuracy of approximate integration

Bayesian inference may be important in some modern big data regimes, but exact integration in general is computationally out of reach. While decades of research in Bayesian inference in both statistics and machine learning have produced many powerful approximate inference algorithms, the big data setting poses some new challenges. Iterative algorithms that read the entire dataset before making each update become prohibitively expensive. Sequential computation is at a significant and growing disadvantage compared to computation that can leverage parallel and distributed computing resources. Insisting on zero asymptotic bias from Monte Carlo estimates of expectations may leave us swamped in errors from high variance [Korattikara et al., 2014] or transient bias.

These challenges, and the tradeoffs that may be necessary to address them, can be viewed in terms of how accurate the integration in our approximate inference algorithms must be. Markov chain Monte Carlo (MCMC) algorithms that admit the exact posterior as a stationary distribution may be the gold standard for generically estimating posterior expectations, but if standard MCMC algorithms become intractable in the big data regime we must find alternatives and understand their tradeoffs. Indeed, someone using Bayesian methods for machine learning may be less constrained than a classical Bayesian statistician: if the ultimate goal is to form predictions that perform well according to a specific loss function, computational gains at the expense of the internal posterior representation may be worthwhile. The methods studied here cover a range of such approximate integration tradeoffs.

1.3 Outline

The remainder of this review is organized as five chapters. In Chapter 2, we provide relevant background material on exponential families, MCMC inference, mean field variational inference, and stochastic gradient optimization. The next three chapters survey recent algorithmic ideas for scaling Bayesian inference, highlighting theoretical results where possible. Each of these central technical chapters ends with a summary and discussion, identifying emergent themes and patterns as well as open questions. Chapters 3 and 4 focus on MCMC algorithms, which are inherently serial and often slow to converge; the algorithms in the first of these use various forms of data subsampling to scale up serial MCMC and in the second use a diverse array of strategies to scale out on parallel resources. In Chapter 5 we discuss two recent techniques for scaling variational mean field algorithms. Both process data in minibatches: the first applies stochastic gradient optimization methods and the second is based on incremental posterior updating. Finally, in Chapter 6 we provide an overarching discussion of the ideas we survey, focusing on challenges and open questions in large-scale Bayesian inference.

References

Sungjin Ahn, Anoop Korattikara Balan, and Max Welling. Bayesian posterior sampling via stochastic gradient Fisher scoring. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

Talal M. Alkhamis, Mohamed A. Ahmed, and Vu Kim Tuan. Simulated annealing for discrete optimization with estimation. *European Journal of Operational Research*, 116(3):530–544, 1999.

Shun-ichi Amari and Hiroshi Nagaoka. *Methods of Information Geometry*. American Mathematical Society, 2007.

Christophe Andrieu and Eric Moulines. On the ergodicity properties of some adaptive MCMC algorithms. *The Annals of Applied Probability*, 16(3): 1462–1505, 2006.

Christophe Andrieu and Gareth O. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *Annals of Statistics*, pages 697–725, 2009.

Christophe Andrieu and Johannes Thoms. A tutorial on adaptive MCMC. *Statistics and Computing*, 18(4):343–373, 2008.

Christophe Andrieu, Arnaud Doucet, and Roman Holenstein. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society Series B*, 72(3):269–342, 2010.

Elaine Angelino. *Accelerating Markov chain Monte Carlo via parallel predictive prefetching*. PhD thesis, School of Engineering and Applied Sciences, Harvard University, 2014.

Elaine Angelino, Eddie Kohler, Amos Waterland, Margo Seltzer, and Ryan P. Adams. Accelerating MCMC via parallel predictive prefetching. In *30th Conference on Uncertainty in Artificial Intelligence*, pages 22–31, 2014.

Kenneth J. Arrow, Leonid Hurwicz, Hirofumi Uzawa, H.B. Chenery, S.M. Johnson, S. Karlin, T. Marschak, and R.M. Solow. *Studies in linear and non-linear programming.* Stanford University Press, John Wiley & Sons, 1959.

Arthur U. Asuncion, Padhraic Smyth, and Max Welling. Asynchronous distributed learning of topic models. In *Advances in Neural Information Processing Systems 21*, pages 81–88, 2008.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

Rémi Bardenet and Odalric-Ambrym Maillard. Concentration inequalities for sampling without replacement. *Bernoulli*, 20(3):1361–1385, 2015.

Rémi Bardenet, Arnaud Doucet, and Chris Holmes. Towards scaling up Markov chain Monte Carlo: An adaptive subsampling approach. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.

Rémi Bardenet, Arnaud Doucet, and Chris Holmes. On Markov chain Monte Carlo methods for tall data. *arXiv preprint 1505.02827*, 2015.

Mark A. Beaumont. Estimation of population growth or decline in genetically monitored populations. *Genetics*, 164(3):1139–60, 2003.

Dimitri P. Bertsekas. *Nonlinear Programming.* Athena Scientific, 2016.

Dimitri P. Bertsekas and John N. Tsitsiklis. *Parallel and Distributed Computation: Numerical Methods.* Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1989.

Michael Betancourt. The fundamental incompatibility of scalable Hamiltonian Monte Carlo and naive data subsampling. In *Proceedings of The 32nd International Conference on Machine Learning*, pages 533–540, 2015.

Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics).* Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

J. Frederic Bonnans and Alexander Shapiro. *Perturbation Analysis of Optimization Problems.* Springer Science & Business Media, 2000.

Léon Bottou. On-line learning and stochastic approximations. In David Saad, editor, *On-line Learning in Neural Networks*, pages 9–42. Cambridge University Press, New York, NY, USA, 1998.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, January 2011.

A. E. Brockwell. Parallel Markov chain Monte Carlo simulation by prefetching. *Journal of Computational and Graphical Statistics*, 15(1):246–261, March 2006.

Tamara Broderick, Nicholas Boyd, Andre Wibisono, Ashia C. Wilson, and Michael I. Jordan. Streaming variational Bayes. In *Advances in Neural Information Processing Systems 26*, pages 1727–1735, 2013.

Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of Markov Chain Monte Carlo*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC press, 2011.

Akif Asil Bulgak and Jerry L. Sanders. Integrating a modified simulated annealing algorithm with the simulation of a manufacturing system to optimize buffer sizes in automatic assembly systems. In *Proceedings of the 20th Conference on Winter Simulation*, pages 684–690, New York, NY, USA, 1988. ACM.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *International Conference on Learning Representations*, 2016.

Jonathan M. R. Byrd, Stephen A. Jarvis, and Abhir H. Bhalerao. Reducing the run-time of MCMC programs by multithreading on SMP architectures. In *IEEE International Symposium on Parallel and Distributed Processing*, pages 1–8, 2008.

Jonathan M. R. Byrd, Stephen A. Jarvis, and Abhir H. Bhalerao. On the parallelisation of MCMC by speculative chain execution. In *IEEE International Symposium on Parallel and Distributed Processing - Workshop Proceedings*, pages 1–8, 2010.

Trevor Campbell and Jonathan P. How. Approximate decentralized Bayesian inference. In *30th Conference on Uncertainty in Artificial Intelligence*, pages 102–111, 2014.

Trevor Campbell, Julian Straub, John W. Fisher III, and Jonathan P. How. Streaming, distributed variational inference for Bayesian nonparametrics. In *Advances in Neural Information Processing Systems 28*, pages 280–288, 2015.

Tianqi Chen, Emily B. Fox, and Carlos Guestrin. Stochastic gradient Hamiltonian Monte Carlo. In *Proceedings of the 31st International Conference on Machine Learning*, June 2014.

John M. Danskin. *The Theory of Max-Min and its Application to Weapons Allocation Problems*. Springer-Verlag, New York, 1967.

Chris De Sa, Kunle Olukotun, and Christopher Ré. Ensuring rapid mixing and low bias for asynchronous Gibbs sampling. In *Proceedings of the 33rd International Conference on Machine Learning*, June 2016.

J. E. Dennis, Jr. and Robert B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice-Hall Series in Computational Mathematics, 1983.

Nan Ding, Youhan Fang, Ryan Babbush, Changyou Chen, Robert D. Skeel, and Hartmut Neven. Bayesian sampling using stochastic gradient thermostats. In *Advances in Neural Information Processing Systems 27*, pages 3203–3211, 2014.

Finale Doshi-Velez, David A. Knowles, Shakir Mohamed, and Zoubin Ghahramani. Large scale nonparametric Bayesian inference: Data parallelisation in the Indian buffet process. In *Advances in Neural Information Processing Systems 22*, pages 1294–1302, 2009.

Arnaud Doucet, Michael Pitt, Robert Kohn, and George Deligiannidis. Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika*, 102(2):295–313, 2015.

David Duvenaud and Ryan P. Adams. Black-box stochastic variational inference in five lines of python. *NIPS Workshop on Black-box Learning and Inference*, 2015.

Paul Fearnhead, Omiros Papaspiliopoulos, Gareth O. Roberts, and Andrew Stuart. Random-weight particle filtering of continuous time processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72 (4):497–512, 2010.

Anthony V. Fiacco. *Introduction to Sensitivity and Stability Analysis in Nonlinear Programming*. Academic Press, Inc., 1984.

Andrew Gelman and Xiao-Li Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical science*, pages 163–185, 1998.

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 3rd edition, 2014a.

Andrew Gelman, Aki Vehtari, Pasi Jylänki, Christian Robert, Nicolas Chopin, and John P Cunningham. Expectation propagation as a way of life. *arXiv preprint arXiv:1412.4869*, 2014b.

Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, pages 721–741, 1984.

Charles J. Geyer. Practical Markov chain Monte Carlo. *Statistical Science*, pages 473–483, 1992.

Ryan J. Giordano, Tamara Broderick, and Michael I. Jordan. Linear response methods for accurate covariance estimates from mean field variational Bayes. In *Advances in Neural Information Processing Systems 28*, pages 1441–1449, 2015.

Mark Girolami and Ben Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) (With Discussion)*, 73:123 – 214, 03 2011.

Joseph Gonzalez, Yucheng Low, Arthur Gretton, and Carlos Guestrin. Parallel Gibbs sampling: From colored fields to thin junction trees. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, pages 324–332, 2011.

Jackson Gorham and Lester Mackey. Measuring sample quality with Stein's method. In *Advances in Neural Information Processing Systems 28*, pages 226–234, 2015.

Thore Graepel, Joaquin Quñonero Candela, Thomas Borchert, and Ralf Herbrich. Web-scale Bayesian click-through rate prediction for sponsored search advertising in Microsoft's Bing search engine. In *Proceedings of the 27th International Conference on Machine Learning*, pages 13–20, 2010.

Roger B. Grosse, Zoubin Ghahramani, and Ryan P. Adams. Sandwiching the marginal likelihood using bidirectional Monte Carlo. *arXiv preprint arXiv:1511.02543*, 2015.

Heikki Haario, Eero Saksman, and Johanna Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223–242, 04 2001.

Leonard Hasenclever, Stefan Webb, Thibaut Lienart, Sebastian Vollmer, Balaji Lakshminarayanan, Charles Blundell, and Yee Whye Teh. Distributed Bayesian learning with stochastic natural-gradient expectation propagation and the posterior server. *arXiv preprint arXiv:1512.09327*, 2016.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, April 1970.

Tom Heskes and Onno Zoeter. Expectation propagation for approximate inference in dynamic Bayesian networks. In *18th Conference on Uncertainty in Artificial Intelligence*, pages 216–223, 2002.

Christian Hipp. Sufficient statistics and exponential families. *The Annals of Statistics*, 2(6):1283–1292, 1974.

Qirong Ho, James Cipar, Henggang Cui, Seunghak Lee, Jin Kyu Kim, Phillip B. Gibbons, Gregory R. Ganger, Garth Gibson, and Eric P. Xing. More effective distributed ML via a stale synchronous parallel parameter server. In *Advances in Neural Information Processing Systems 26*, pages 1223–1231, 2013.

Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1): 1303–1347, May 2013.

Zaiying Huang and Andrew Gelman. Sampling for Bayesian computation with large datasets. Technical report, Columbia University, 2005.

Michael C. Hughes and Erik B. Sudderth. Memoized online variational inference for Dirichlet process mixture models. In *Advances in Neural Information Processing Systems 26*, pages 1133–1141, 2013.

Michael C. Hughes, Dae Il Kim, and Erik B. Sudderth. Reliable and scalable variational inference for the hierarchical Dirichlet process. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, pages 370–378, 2015.

Alexander Ihler and David Newman. Understanding errors in approximate distributed latent Dirichlet allocation. *IEEE Transactions on Knowledge and Data Engineering*, 24(5):952–960, 2012.

Pierre E. Jacob and Alexandre H. Thiery. On nonnegative unbiased estimators. *Annals of Statistics*, 43(2):769–784, 04 2015.

Matthew J. Johnson, James Saunderson, and Alan S. Willsky. Analyzing Hogwild parallel Gaussian Gibbs sampling. In *Advances in Neural Information Processing Systems 26*, pages 2715–2723, 2013.

Matthew James Johnson. *Bayesian Time Series Models and Scalable Inference*. PhD thesis, Massachusetts Institute of Technology, 2014.

Robert W. Keener. *Theoretical Statistics: Topics for a Core Course*. Springer Texts in Statistics. Springer New York, 2010.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. *International Conference on Learning Representations*, 2014.

Jack P. C. Kleijnen and Reuven Y. Rubinstein. Optimization and sensitivity analysis of computer simulation models by the score function method. *European Journal of Operational Research*, 88(3):413–427, 1996.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

Augustine Kong, Jun S. Liu, and Wing Hung Wong. Sequential imputations and Bayesian missing data problems. *Journal of the American Statistical Association*, 89(425):278–288, 1994.

Anoop Korattikara, Yutian Chen, and Max Welling. Austerity in MCMC land: Cutting the Metropolis-Hastings budget. In *Proceedings of the 31th International Conference on Machine Learning*, volume 32, pages 181–189, 2014.

Alp Kucukelbir, Rajesh Ranganath, Andrew Gelman, and David Blei. Automatic variational inference in stan. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 568–576. Curran Associates, Inc., 2015.

Neil D. Lawrence. Modelling in the context of massively missing data, 2015. URL `http://staffwww.dcs.shef.ac.uk/people/N.Lawrence/talks/missingdata_tuebingen15.pdf`.

Benedict Leimkuhler and Xiaocheng Shang. Adaptive thermostats for noisy gradient systems. *SIAM Journal on Scientific Computing*, 38(2):A712–A736, 2016.

Yingzhen Li, José Miguel Hernández-Lobato, and Richard E. Turner. Stochastic expectation propagation. In *Advances in Neural Information Processing Systems 28*, pages 2323–2331, 2015.

L. Lin, K. F. Liu, and J. Sloan. A noisy Monte Carlo algorithm. *Physical Review D*, 61:074505, March 2000.

Zhiyuan Liu, Yuzhou Zhang, Edward Y. Chang, and Maosong Sun. PLDA+: Parallel latent Dirichlet allocation with data placement and pipeline processing. *ACM Transactions on Intelligent Systems and Technology*, 2(3): 26:1–26:18, May 2011.

Anne-Marie Lyne, Mark Girolami, Yves Atchaé, Heiko Strathmann, and Daniel Simpson. On Russian roulette estimates for Bayesian inference with doubly-intractable likelihoods. *Statistical Science*, 30(4):443–467, 11 2015.

Yi-An Ma, Tianqi Chen, and Emily Fox. A complete recipe for stochastic gradient MCMC. In *Advances in Neural Information Processing Systems 28*, pages 2917–2925, 2015.

David J. C. MacKay. *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, New York, NY, USA, 2003.

Dougal Maclaurin and Ryan P. Adams. Firefly Monte Carlo: Exact MCMC with subsets of data. In *30th Conference on Uncertainty in Artificial Intelligence*, pages 543–552, 2014.

Stephan Mandt, Matthew D. Hoffman, and David M. Blei. A variational analysis of stochastic gradient algorithms. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

James Martens. New insights and perspectives on the natural gradient method. *arXiv preprint arXiv:1412.1193*, 2015.

James Martens and Roger Grosse. Optimizing neural networks with Kronecker-factored approximate curvature. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

Peter S. Maybeck. *Stochastic Models, Estimation, and Control*, volume 3. Academic Press, 1982.

James McInerney, Rajesh Ranganath, and David M. Blei. The population posterior and bayesian inference on streams. In *Advances in Neural Information Processing Systems 28*, pages 1153–1161, 2015.

Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

Sean Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, New York, NY, USA, 2nd edition, 2009.

Thomas P. Minka. Expectation propagation for approximate Bayesian inference. In *17th Conference on Uncertainty in Artificial Intelligence*, pages 362–369, 2001.

Stanislav Minsker, Sanvesh Srivastava, Lizhen Lin, and David B. Dunson. Scalable and robust Bayesian inference via the median posterior. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1656–1664, 2014.

Andriy Mnih and Karol Gregor. Neural variational inference and learning in belief networks. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1791–1799, 2014.

Andriy Mnih and Danilo Jimenez Rezende. Variational inference for monte carlo objectives. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

Volodymyr Mnih, Csaba Szepesvári, and Jean-Yves Audibert. Empirical Bernstein stopping. In *Proceedings of the 25th International Conference on Machine Learning*, pages 672–679, 2008.

Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

Kevin P. Murphy, Yair Weiss, and Michael I. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *15th Conference on Uncertainty in Artificial Intelligence*, pages 467–475, 1999.

Radford M. Neal. An improved acceptance procedure for the hybrid Monte Carlo algorithm. *J. Comput. Phys.*, 111(1):194–203, March 1994.

Radford M. Neal. MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*, Chapman & Hall/CRC Handbooks of Modern Statistical Methods, pages 113–162. CRC Press, 2010.

Willie Neiswanger, Chong Wang, and Eric Xing. Asymptotically exact, embarrassingly parallel MCMC. In *30th Conference on Uncertainty in Artificial Intelligence*, pages 623–632, 2014.

David Newman, Padhraic Smyth, Max Welling, and Arthur U. Asuncion. Distributed inference for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems 20*, pages 1081–1088, 2007.

David Newman, Arthur U. Asuncion, Padhraic Smyth, and Max Welling. Distributed algorithms for topic models. *Journal of Machine Learning Research*, 10:1801–1828, 2009.

Robert Nishihara, Iain Murray, and Ryan P. Adams. Parallel MCMC with generalized elliptical slice sampling. *Journal of Machine Learning Research*, 15:2087–2112, 2014.

Manfred Opper and Ole Winther. A Bayesian approach to on-line learning. In David Saad, editor, *On-line Learning in Neural Networks*, pages 363–378. Cambridge University Press, New York, NY, USA, 1998.

John Paisley, David M. Blei, and Michael I. Jordan. Variational Bayesian inference with stochastic search. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

Omiros Papaspiliopoulos. A methodological framework for Monte Carlo probabilistic inference for diffusion processes. Technical report, Centre for Research in Statistical Methodology, University of Warwick, June 2009.

Omiros Papaspiliopoulos, Gareth O. Roberts, and Martin Sköld. A general framework for the parametrization of hierarchical models. *Statistical Science*, pages 59–73, 2007.

Sam Patterson and Yee Whye Teh. Stochastic gradient Riemannian Langevin dynamics on the probability simplex. In *Advances in Neural Information Processing Systems 26*, pages 3102–3110, 2013.

M. F. Pradier, P. G. Moreno, F. J. R. Ruiz, I. Valera, H. Mollina-Bulla, and F. Perez-Cruz. Map/reduce uncollapsed Gibbs sampling for Bayesian non parametric models. *Workshop in Software Engineering for Machine Learning at NIPS*, 2014.

Maxim Rabinovich, Elaine Angelino, and Michael I. Jordan. Variational Consensus Monte Carlo. In *Advances in Neural Information Processing Systems 28*, pages 1207–1215, 2015.

Rajesh Ranganath, Chong Wang, David M. Blei, and Eric P. Xing. An adaptive learning rate for stochastic variational inference. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pages 298–306, 2013.

Rajesh Ranganath, Sean Gerrish, and David M. Blei. Black box variational inference. In *17th International Conference on Artificial Intelligence and Statistics*, pages 814–822, 2014.

Benjamin Recht, Christopher Ré, Stephen J. Wright, and Feng Niu. Hogwild!: A lock-free approach to parallelizing stochastic gradient descent. In *Advances in Neural Information Processing Systems 24*, pages 693–701, 2011.

Jeffrey Regier, Andrew Miller, Jon McAuliffe, Ryan P. Adams, Matt Hoffman, Dustin Lang, David Schlegel, and Prabhat. Celeste: Variational inference for a generative model of astronomical images. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2095–2103, 2015.

Danilo J. Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1278–1286, 2014.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

Christian P. Robert and George Casella. *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Springer-Verlag New York, Inc., 2004.

Gareth O. Roberts and Richard L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4): 341–363, 12 1996.

Gareth O. Roberts, Andrew Gelman, and Walter R. Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *Annals of Applied Probability*, 7:110–120, 1997.

Tim Salimans, Diederik P. Kingma, and Max Welling. Markov chain Monte Carlo and variational inference: Bridging the gap. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1218–1226, 2015.

Issei Sato and Hiroshi Nakagawa. Approximation analysis of stochastic gradient Langevin dynamics by using Fokker-Planck equation and Ito process. In *Proceedings of the 31st International Conference on Machine Learning*, pages 982–990, 2014.

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. *International Journal of Management Science and Engineering Management*, 11(2):78–88, 2016.

R. J. Serfling. Probability inequalities for the sum in sampling without replacement. *The Annals of Statistics*, 2(1):39–48, 1974.

Xiaocheng Shang, Zhanxing Zhu, Benedict Leimkuhler, and Amos J. Storkey. Covariance-controlled adaptive Langevin thermostat for large-scale Bayesian sampling. In *Advances in Neural Information Processing Systems 28*, pages 37–45, 2015.

Sameer Singh, Michael L. Wick, and Andrew McCallum. Monte Carlo MCMC: Efficient inference by approximate sampling. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1104–1113, 2012.

Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. Practical Bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems 25*, pages 2951–2959, 2012.

David H. Stern, Ralf Herbrich, and Thore Graepel. Matchbox: Large scale online Bayesian recommendations. In *Proceedings of the 18th International Conference on World Wide Web*, pages 111–120, 2009.

Ingvar Strid. Efficient parallelisation of Metropolis-Hastings algorithms using a prefetching approach. *Computational Statistics & Data Analysis*, 54(11): 2814–2835, November 2010.

Alex Tank, Nicholas Foti, and Emily Fox. Streaming variational inference for Bayesian nonparametric mixture models. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, pages 968–976, 2015.

Yee Whye Teh, Alexandre H. Thiery, and Sebastian J. Vollmer. Consistency and fluctuations for stochastic gradient Langevin dynamics. *The Journal of Machine Learning Research*, 17(1):193–225, 2016.

Wolfgang Wagner. Unbiased Monte Carlo evaluation of certain functional integrals. *Journal of Computational Physics*, 71(1):21–33, 1987.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, November 2008.

Ling Wang and Liang Zhang. Stochastic optimization using simulated annealing with hypothesis test. *Applied Mathematics and Computation*, 174 (2):1329–1342, 2006.

Xiangyu Wang and David B. Dunson. Parallel MCMC via Weierstrass sampler. *arXiv preprint arXiv:1312.4605*, 2013.

Karl Weierstrass. Über die analytische darstellbarkeit sogenannter willkrlicher functionen einer reellen vernderlichen. *Sitzungsberichte der Königlich Preuischen Akademie der Wissenschaften zu Berlin*, 1885. (II). Erste Mitteilung (part 1) pp. 633–639, Zweite Mitteilung (part 2) pp. 789–805.

Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, 2011.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

E. E. Witte, R. D. Chamberlain, and M. A. Franklin. Parallel simulated annealing using speculative computation. *IEEE Transactions on Parallel and Distributed Systems*, 2(4):483–494, 1991.

Minjie Xu, Balaji Lakshminarayanan, Yee Whye Teh, Jun Zhu, and Bo Zhang. Distributed Bayesian posterior sampling via moment sharing. In *Advances in Neural Information Processing Systems 27*, pages 3356–3364, 2014.