École des Ponts ParisTech
Department of Statistics, University of Oxford

2017
Master's Internship Report

Émile Mathieu
Élève ingénieur, Third year

Bayesian Nonparametric Inference

within Probabilistic Programming Languages

Internship carried out at Department of Statistics, University of Oxford
From the 22nd of May, to the 15th of September 2017.

Company tutor: Teh, Yee Whye
Training supervisor: Obozinski, Guillaume

Acknowledgments

First of all, I would like to express my indebtedness appreciation to my departmental supervisor Prof. Yee Whye Teh. His belief in me and his advices played a decisive role in making the execution of my work and thus the report.

I also express my deepest thanks to Benjamin Bloem-Reddy, who as a postdoc, oversaw me during this internship.

Moreover, my gratitude goes to Guillaume Obozinski, my school training supervisor, whose guidance has continually shaped my career path since I have been at Ecole des Ponts ParisTech.

Abstract

Keywords : Probabilistic Programming, Bayesian Non-parametric, Bayesian Inference, Sampling methods

Contents

List of Figures

Glossary

- **BNP**: *Bayesian Non-Parametric*, explained in Section 3.2.1.
- **PPL**: *Probabilistic Programming Language*, explained in Section 3.1.1.

Introduction

cf Research Proposal ?

Presentation of the Department of Statistics

2.1 Creation

The Department of Statistics [1] is part of the University of Oxford, along with the other departments and the 38 constituent colleges.

The University of Oxford was founded in the 11th century, which makes it the oldest university in the English-speaking world and the world's second-oldest university in continuous operation.

The Department of Statistics was officially created in 1988, even though first moves in the development of Oxford statistics can be dated to the 19th century.

Indeed, In the 1870s, Florence Nightingale – the pioneer of modern nursing – discussed the possibility of endowing a Professorship of Statistics in Oxford, but the proposal eventually foundered. However, Oxford did appoint a statistician to a chair in 1891, although not to a chair in statistics.

The next significant moves in the development of Oxford statistics were by economists, who were increasingly keen to build economic theory on a foundation of sound data analysis. This led to the creation in 1935 of an Institute of Statistic, swhich was then renamed as the Institute of Economics and Statistics in 1962.

The sequence of events which led directly to the establishment of the present Department of Statistics began with the appointment in 1945 of David Finney as the universitys first Lecturer in the Design and Analysis of Scientific Experiment (LIDASE).

Then in the1980s, after the Department of Biomathematics' head increasingly felt that Oxford was losing out in the face of developments in statistics, a working party appointed by the general board of the university to assess a careful analysis of the organisation of statistics in Oxford. They found fragmentation to be the dominant feature of Oxford statistics and concluded that fragmentation has serious disadvantages ..... The working partys report recommended the creation of a university statistics department, which were to include the former Department of Biomathematics, together with a new Professorship in Statistical Science and the two existing lecturerships in statistics within the Mathematical Institute.

[1] https://www.stats.ox.ac.uk

These major recommendations were all accepted by the university and the new Department of Statistics was created in 1988.

2.2 Activities

The Department of Statistics at Oxford is a world leader in research including computational statistics and statistical methodology, applied probability, bioinformatics and mathematical genetics. The main research groups in the Department are Computational statistics and machine learning, Probability, Statistical genetics and bioinformatics, Protein Informatics and Statistical Genetics.

I am part of the Computational Statistics and Machine Learning Group (OxCSML) [2], which have research interests spanning Statistical Machine Learning, Monte Carlo Methods and Computational Statistics, Statistical Methodology and Applied Statistics.

The department offers an undergraduate degree (BA or MMath) in Mathematics and Statistics, jointly with the Mathematical Institute. At postgraduate level there is an MSc course in Applied Statistics (MSc in Statistical Science from 2017), as well as a lively and stimulating environment for postgraduate research (DPhil or MSc by Research). The department also has a consulting activity called *Oxford University Statistical Consulting*.

[2] http://csml.stats.ox.ac.uk/people/mathieu/

Mission

3.1 Probabilistic programming

3.1 What is it ?

Using PL techniques to abstract inference algorithms from stats/ML such that they apply automatically and correctly to the broadest possible set of model-based reasoning applications

Probabilistic programs are usual functional or imperative programs with two added constructs: (1) the ability to draw values at random from distributions, and (2) the ability to condition values of variables in a program via observations. from Gordon, Henzinger, Nori, and Rajamani Probabilistic programming. In Proceedings of On The Future of Software Engineering (2014).

3.1 Why is it usefull ?

Increase productivity: savings to be found in the amount of code that needs to written in order to prototype and develop models.

remove the burden of having to develop inference code for each new model: which is error-prone and time consuming This is done by providing a modelling language abstraction layer in which developers can denote their models. If done, generic inference is provided for free.

3.1 History / Existing PPLs

Graphical models: BUGS [8], STAN [3] Factor graphs: Factorie [10], Infer.NET [11]

First-Order PPLs: bounded loops = loops can be desugared to nested lets

High Order PPLs: Recursion / Turing complete infinite dimensional parameter space easy to program in, natural to express certain models, hard to perform inference in Anglican, Venture (graph MCMC), Church

3.1 Inference schemes

Importance Sampling: IS, SMC

PMCMC: PG, PMMH, IPMCMC, PGAS

Hamiltonian: HMC, HMCDA, NUTS, SGLD, SGHMC

3.1 Contributions

During this internship I have taken the time to actually implement several inference algorithms, by contributing to two existing PPLs. First, I implemented [1] both the Stochastic Gradient Langevin Dynamics (SGLD) and Stochastic Gradient Hamiltonian Monte Carlo (SGHMC) inference algorithms in Turing.jl [6], a PPL based on Julia and developed at the University of Cambridge. Then, I implemented [2] the Dual Averaging extension of HMC [5] for Edward [13], a PPL built on top of Tensorflow [1] by Blei's group at Columbia University.

[1] See `https://github.com/yebai/Turing.jl/tree/master/src/samplers`

[2] See `https://github.com/blei-lab/edward/pull/728`

SGLD

mini-batch / online setting, scale to bug dataset [3] [14]

SGHMC

Same setting as SGLD Naive version is wrong (posterior is not the invariant distribution), see [9] friction term [4]

Dual Averaging

In [7], the authors address the issue of choosing the two hyperparameters of HMC: a step size ϵ and a desired number of steps L, since HMCs performance is highly sensitive on those. [12]

[3] See for instance, SGLD applied to a Bayesian logistic regression at `https://github.com/yebai/Turing.jl/blob/master/example-models/sgld-paper/lr_sgld.jl`

3.2 Bayesian nonparametric

3.2 Definition

3.2 Usefulness

3.2 Canonical models

??? DP, PYP, etc

3.2 MCMC Inference

Constructing MCMC schemes for models with one or more Bayesian nonparametric components is an active research area since dealing with the infinite dimensional component P forbids the direct use of standard simulation-based methods.These methods usually require a finite-dimensional representation. The general idea for designing inference schemes is to find finite dimensional representations to be able to store the model in a computer with finite capacity.

There are two main sampling approaches to facilitate simulation in the case of Bayesian nonparametric models: random truncation and marginalisation. These two schemes are known in the literature as conditional and marginal samplers.

Marginal Samplers

Marginal samplers bypass the need to represent the infinite-dimensional component by marginalising it out. These schemes have lower storage requirements than conditional samplers because they only store the induced partition, but could potentially have worse mixing properties.

Conditional Samplers

Conditional samplers replace the infinite-dimensional prior by a finite-dimensional representation chosen according to a truncation level. Since these samplers do not integrate out the infinite-dimensional component, their output provides a more comprehensive representation of the random probability measure. thinning vs stick-breaking

Hybrid Samplers

YW paper on PK ?

SMC

Review of SMC ? cf Maria Lomeli thesis

3.2 BNP sampling in PPL

Stochastic Memoization with DPmem: $\alpha = 0$, deterministic memoization, $\alpha =$ inf no memoization
https://probmods.org/chapters/12-non-parametric-models.html

3.2 Link between BNP and High order PPL

See Frank Wood meeting

3.3 Future Work

3.3 Learnable parameters in PPL

cf AESMC: PyTorch + PPL ? Turing + Knet ?

3.3 Variational Inference

3.3 Adversarial Inference

Conclusion and personal review

4.1 Conclusion

Ouverture

4.2 Personal review

Responsability: organization of the probabilistic inference reading group. Research environment Research way of working / thinking: Finding the good questions, etc

Appendices

A Appendix Section

Bibliography

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, *Tensorflow: A system for large-scale machine learning*, in 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), 2016, pp. 265–283.

[2] D. M. Blei and M. I. Jordan, *Variational inference for dirichlet process mixtures*, Bayesian Analysis, 1 (2005), pp. 121–144.

[3] B. Carpenter, D. Lee, M. A. Brubaker, A. Riddell, A. Gelman, B. Goodrich, J. Guo, M. Hoffman, M. Betancourt, and P. Li, *Stan: A probabilistic programming language.*

[4] T. D. Chen and C. Fox, Emily B.and Guestrin, *Stochastic gradient hamiltonian monte carlo*, in International Conference on Machine Learning, 2014.

[5] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth, *Hybrid monte carlo*, Physics Letters B, 195 (1987), pp. 216 – 222.

[6] H. Ge, A. Ścibior, K. Xu, and Z. Ghahramani, *Turing: A fast imperative probabilistic programming language.*, (2016).

[7] M. D. Homan and A. Gelman, *The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo*, J. Mach. Learn. Res., 15 (2014), pp. 1593–1623.

[8] D. J. Lunn, A. Thomas, N. Best, and D. Spiegelhalter, *Winbugs - a bayesian modelling framework: Concepts, structure, and extensibility*, Statistics and Computing, 10 (2000), pp. 325–337.

[9] Y.-A. Ma, T. Chen, and E. B. Fox, *A complete recipe for stochastic gradient mcmc*, in Proceedings of the 28th International Conference on Neural Information Processing Systems, NIPS'15, Cambridge, MA, USA, 2015, MIT Press, pp. 2917–2925.

[10] A. McCallum, K. Schultz, and S. Singh, *FACTORIE: Probabilistic programming via imperatively defined factor graphs*, in Neural Information Processing Systems (NIPS), 2009.

[11] T. Minka, J. Winn, J. Guiver, S. Webster, Y. Zaykov, B. Yangel, A. Spengler, and J. Bronskill, *Infer.NET 2.6*, 2014. Microsoft Research Cambridge. http://research.microsoft.com/infernet.

[12] Y. Nesterov, *Primal-dual subgradient methods for convex problems*, Math. Program., 120 (2009), pp. 221–259.

[13] D. Tran, M. D. Hoffman, R. A. Saurous, E. Brevdo, K. Murphy, and D. M. Blei, *Deep probabilistic programming*, in International Conference on Learning Representations, 2017.

[14] M. Welling and Y. W. Teh, *Bayesian learning via stochastic gradient langevin dynamics.*, in ICML, L. Getoor and T. Scheffer, eds., Omnipress, 2011, pp. 681–688.