Inference in the Presence of Intractable Normalizing Functions

Joint work with Jaewoo Park (Yonsei University)

Joint Statistical Meetings, Denver, Colorado.

July 2019

Murali Haran

Department of Statistics, Penn State University

Models with Intractable Normalizing Functions

- Data: $\mathbf{x} \in \chi$, parameter: $\theta \in \Theta$
- Model: $h(\mathbf{x}|\theta)/Z(\theta)$
- $Z(\theta) = \int_{\chi} h(\mathbf{x}|\theta) d\mathbf{x}$ is intractable
- Popular examples
  - Social network models: exponential random graph models (Robins et al., 2002; Hunter et al., 2008)
  - Models for lattice data (Besag, 1972, 1974)
  - Spatial point process models: interaction models (Strauss, 1975, Goldstein, Haran et al., 2015)

Inference

- $Z(\theta)$ makes inference difficult
- Maximum likelihood:

$$\hat{\theta}_{MLE} = \arg\max_{\theta \in \Theta} h(\mathbf{x}|\theta)/Z(\theta)$$

- Bayesian inference
  - Prior : $p(\theta)$
  - Posterior: $\pi(\theta|\mathbf{x}) \propto p(\theta)h(\mathbf{x}|\theta)/Z(\theta)$

This Talk

- New algorithm for Bayesian inference for models with intractable normalizing functions
- Works well for some problems where existing algorithms are computationally infeasible

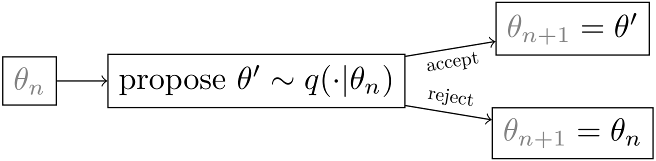
Markov chain Monte Carlo Basics

- Construct Harris-ergodic Markov chain $\theta_1, \theta_2, \ldots$ with stationary distribution $\pi(\theta \mid \mathbf{x})$
- Treat $\theta_1, \theta_2, \theta_3, ...$ as samples from $\pi(\theta|\mathbf{x})$
- For any real-valued $g(\cdot)$, approximate $E_\pi(g(\theta))$ by

$$\hat{\mu}_n = \frac{\sum_{i=1}^{n} g(\theta_i)}{n}$$

The Metropolis-Hastings Algorithm

Recipe for constructing Markov chain: given θ_n, obtain θ_{n+1}



θ_n

propose $\theta' \sim q(\cdot|\theta_n)$

accept

$\theta_{n+1} = \theta'$

reject

$\theta_{n+1} = \theta_n$

Accept-reject ratio

$$\frac{\pi(\theta'|\mathbf{x})q(\theta_n|\theta')}{\pi(\theta_n|\mathbf{x})q(\theta'|\theta_n)} = \frac{p(\theta')Z(\theta_n)h(\mathbf{x}|\theta')q(\theta_n|\theta')}{p(\theta_n)Z(\theta')h(\mathbf{x}|\theta_n)q(\theta'|\theta_n)}$$

Cannot evaluate because of $Z(\cdot)$

Previous Work

Two classes of algorithms (with some overlap)

1. Auxiliary variable algorithms (cf. Moller et al., 2006; Murray et al., 2007; Liang et al., 2010, 2016)
2. Likelihood approximation algorithms (Atchade et al., 2015; Andrieu and Roberts, 2009; Lyne et al., 2015; Alqueir et al., 2016)

I. Auxiliary Variable Algorithms

(Moller et al., 2016; Murray et al., 2017)

The acceptance ratio

$$\frac{\pi(\theta'|\mathbf{x})q(\theta_n|\theta')}{\pi(\theta_n|\mathbf{x})q(\theta'|\theta_n)} = \frac{p(\theta')Z(\theta_n)h(\mathbf{x}|\theta')q(\theta_n|\theta')}{p(\theta_n)Z(\theta')h(\mathbf{x}|\theta_n)q(\theta'|\theta_n)}$$

- Generate an auxiliary random variate from model $h(\mathbf{x}|\theta')$
  - $X' \sim h(\mathbf{x}|\theta')$
  - M-H algorithm: accept/reject both X' and θ'
- New acceptance ratio: $Z(\theta)$ gets canceled

Problem: $X \sim h(\mathbf{x}|\theta')$ **is difficult/expensive**

Most practical and general (asymptotically inexact): Double Metropolis algorithm (Liang, 2010)

2. Likelihood Approximation

The acceptance ratio

$$\frac{\pi(\theta'|\mathbf{x})q(\theta_n|\theta')}{\pi(\theta_n|\mathbf{x})q(\theta'|\theta_n)} = \frac{p(\theta')Z(\theta_n)h(\mathbf{x}|\theta')q(\theta_n|\theta')}{p(\theta_n)Z(\theta')h(\mathbf{x}|\theta_n)q(\theta'|\theta_n)}$$

Likelihood approximation methods

1. Approximate $Z(\theta)$ using importance sampling
   - Requires its own MCMC algorithm
2. Use approximation $\widehat{Z}(\theta)$ in acceptance ratio

Problem: Step 1 is computationally expensive

Emulation-Based Algorithm

Park and Haran (2019)

- Likelihood approximation approach with a two-step approximation
  1. Approximate $Z(\theta)$ using importance sampling on a set of θs
  2. Use Gaussian process “emulation” approach to interpolate this function at any new value
  3. Construct MCMC algorithm using this interpolation

Theory to justify this as number of design points and number of importance sampling draws increases

(Park and Haran, 2019)

Normalizing Function Emulation Algorithm

Part 1: Construct two-stage approximation

- Pre-MCMC

1 For each $\theta \in \{\boldsymbol{\theta}^{(1)}, ..., \boldsymbol{\theta}^{(d)}\}$, obtain importance sampling approximation $\widehat{Z}_{IMP}(\theta)$

2 Fit Gaussian process (GP) to $\{\widehat{Z}_{IMP}(\boldsymbol{\theta}^{(1)}), ..., \widehat{Z}_{IMP}(\boldsymbol{\theta}^{(d)})\}$
Now for each θ obtain GP approximation, $\widehat{Z}_{GP}(\theta)$

Part 2: MCMC algorithm with GP approximation

- Given $\boldsymbol{\theta}_n \in \boldsymbol{\Theta}$ at nth iteration.

3 Propose $\boldsymbol{\theta}' \sim q(\cdot|\boldsymbol{\theta}_n)$

4 Obtain $\widehat{Z}_{GP}(\boldsymbol{\theta}')$, accept $\boldsymbol{\theta}'$ with

$$\alpha = \min\left\{ \frac{p(\boldsymbol{\theta}')h(\mathbf{x}|\boldsymbol{\theta}')\widehat{Z}_{GP}(\boldsymbol{\theta})q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{p(\boldsymbol{\theta})h(\mathbf{x}|\boldsymbol{\theta})\widehat{Z}_{GP}(\boldsymbol{\theta}')q(\boldsymbol{\theta}'|\boldsymbol{\theta})}, 1 \right\}$$

Computational Benefits

- Can compute in parallel; much of this is done “offline”, before running the algorithm
- Two versions of our approach
  (i) NormEmul emulate $Z(\theta)$ with $\widehat{Z}_{GP}(\theta)$
  (ii) LikEmul emulate $\mathcal{L}(\theta) = h(\mathbf{x}|\theta)/Z(\theta)$ with $\widehat{\mathcal{L}}_{GP}(\theta)$

Theory

The Markov chain constructed by the function-emulation algorithm, with $n-$step transition kernel $P^n_{GP}(x, \cdot)$, converges in total variational distance to the target distribution π

$$\lim_{n\to\infty} \|P^n_{GP}(x, \cdot) - \pi(\cdot)\|_{TV} = 0, \forall x \in \Omega$$

- Key assumptions satisfied for all our examples
- Results as # samples for $\widehat{Z}_{IMP}$ and number of design points for $\widehat{Z}_{GP}$ both go to infinity. Hence, in practice asymptotically inexact (like Double Metropolis-Hastings)

Park and Haran (2019); also see Mitrophanov (2005); Alquier et al. (2016)

Examples

(1) Interaction point process model (Goldstein et al., 2015)

- Real data set, n =3,000 points
- Comparing fastest existing algorithm DMH (Double Metropolis-Hastings) with our two new algorithms
- HPD=highest posterior density region

θ_1	Mean	95%HPD	Time(hour)
DMH	1.34	(1.30,1.39)	18.99
NormEmul	1.34	(1.30,1.39)	3.60
LikEmul	1.34	(1.29, 1.39)	2.53

(2) 2,000 dimensional exponential random graph model for a network (Hunter et al, 2006)

- Reliable results from NormEmul, LikEmul within 2 hours
- All other algorithms are computationally infeasible

The Last Slide

- This methodology is widely applicable, e.g. network (ERGM) models
- LikeEmul algorithm is useful for intractable likelihood problems (not just intractable normalizing function problems): e.g. applied to disease model
- Open problems and caveats
  - Relies heavily on design points selected initially
    Either use Double Metropolis-Hastings or Approximate Bayesian Computing (ABC) to do this; not always efficient
  - Practical for low-dimensional (< 7) parameter space only
  - Automated tuning is difficult
  - Stopping rules are tricky
  - Theoretical challenges

References

All papers on `arxiv.org`

1. Framework, comparisons for current algorithms
   - Park and Haran (2018) "Bayesian Inference in the Presence of Intractable Normalizing Functions," *Journal of the American Statistical Association*
2. New algorithm
   - Park and Haran (2019) "A Function Emulation Approach for Doubly Intractable Distributions," *Journal of Computational and Graphical Statistics*

References

- Møller, J., Pettitt, A. N., Reeves, R., and Berthelsen, K. K. (2006) An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants, *Biometrika*
- Murray, I., Z. Ghahramani, and D. MacKay (2006) MCMC for doubly-intractable distributions. *Proc of 22nd Annual Conf on Uncertainty in Artificial Intelligence* UAI06
- Liang, F. (2010) A Double Metropolis-Hastings sampler for spatial models with intractable normalizing constants. *Journal of Statistical Computation and Simulation*

References

- Atchade, Y., Lartillot, N. and Robert, C. (2013) Bayesian computation for statistical models with intractable normalizing constants. *Brazilian Journal of Probability and Statistics*
- Liang, F., Jin, I. H., Song, Q., and Liu, J. S. (2015) An adaptive exchange algorithm for sampling from distributions with intractable normalising constants. *Journal of the American Statistical Association*
- Goldstein, J., Haran, M., Simeonov, I., Fricks, J., and Chiaromonte, F. (2015) An attraction-repulsion point process model for respiratory syncytial virus infections. *Biometrics*