Chapter 1

Algorithms for Models with Intractable Normalizing Functions

Murali Haran, Bokgyeong Kang, Jaewoo Park

In this paper we discuss a well known computing problem – inference for models with intractable normalizing functions. Models with intractable normalizing functions arise in a wide variety of areas, for instance network models, models for spatial data on lattices, spatial point processes, flexible models for count data and gene expression, and models for permutations (Lu and Boutilier, 2014); for more examples see (Matsubara et al., 2022; Park and Haran, 2018). Simulating from these models for fixed parameter values is well studied, starting with work dating back seventy years to the origin of the Metropolis algorithm. On the other hand some of the most practical and theoretically justified algorithms for inference, particularly Bayesian inference, have only been developed within the past two decades. The most computationally efficient algorithms often do not have well developed theory and few if any approaches exist for assessing the quality of approximations based on them. For many problems even the best algorithms can be computationally infeasible. Hence, this is an exciting area of research with many open problems. We explain several key algorithms, providing connections and touching upon practical advantages and disadvantages of each,

with some discussion of theoretical properties where they impact practice. We discuss an approach for assessing the accuracy of approximations produced by these algorithms; this diagnostic is particularly valuable for algorithm tuning.

While our focus is largely on models with intractable normalizing functions, we also discuss algorithms that are more broadly applicable to models where the entire likelihood function is intractable; these methods are of course also applicable to intractable normalizing function problems. Intractable likelihood function problems are growing in importance as statistical models become increasingly sophisticated and data sets become larger and more complex. Inference for intractable likelihood models, also known as likelihood-free inference, therefore represents one of the most important computational challenges of modern statistics. The goals of this manuscript are to (i) provide an accessible introduction to the intractable normalizing function (INF) and intractable likelihood (IL) problems, as well as key ideas underpinning several algorithms used to solve them; (ii) describe an approach for assessing the sample quality of asymptotically inexact algorithms for INF problems; (iii) provide practical recommendations for solving INF problems based on a study of some challenging examples; (iv) suggest avenues for future research. The remainder of this paper is organized as follows. We provide some historical and technical background for INF and IL problems in Section 1.1. In Section 1.2 we provide a taxonomy for algorithms, along with explanations for several important algorithms, then explain an approach for assessing sample quality for both asymptotically exact and inexact algorithms in Section 1.3. We conclude with the application of several algorithms to challenging examples in Section 1.4, providing insights about the algorithms whenever possible, and then provide a summary and discussion of potential areas for future research in Section 1.5.

1.1 Background

The landmark Metropolis et al. (1953) paper that introduced the Metropolis algorithm and hence the beginnings of Markov chain Monte Carlo (MCMC), describes an algorithm to simulate a spin system in particle physics according to the Ising model (Ising, 1925; Lenz,

1920), which has an intractable normalizing function. Because the problem considered in the paper is simulating from the model for fixed parameters, the normalizing function (really a constant in this case) cancels out in the acceptance ratio of the Metropolis-Hastings algorithm. Hence, from the earliest days of MCMC, simulating realizations from a model with intractable normalizing functions is considered a solved problem. However, inference is an entirely different matter as the normalizing function does not cancel out. We explain below both the ease of simulation from the model and the challenge for inference.

Consider a random variable $x \in \mathcal{X}$, assumed to be a realization from the probability model $f(x \mid \theta)$ with parameter $\theta \in \Theta$. Bayesian inference for θ is based on the posterior $\pi(\theta \mid x) \propto f(x \mid \theta)p(\theta)$ where $p(\theta)$ is a prior density on θ. It is common to have $f(x \mid \theta) = h(x \mid \theta)/c(\theta)$ where $c(\theta)$ is an intractable normalizing function of θ. For instance, Ising, Potts, and exponential random graph models (ERGMs) can be expressed as exponential family models of the form $f(x \mid \theta) = \exp(\theta S(x))/c(\theta)$, where θ is a p-dimensional vector of parameters and $S(x)$ is the vector of jointly sufficient statistics for θ. The normalizing function $c(\theta) = \sum_{x \in \chi} \exp(\theta s(x))$ is intractable for any realistic problem as it involves summing over all possible $x \in \mathcal{X}$. For example for the Ising model which involves 2 potential spins $\{-1, 1\}$ for each of n particles, the number of configurations is 2^n, a large number even for relatively small n; for $n = 300$ it is more than 2×10^{90} configurations.

The posterior distribution for the kinds of models described above is $\pi(\theta \mid x) \propto h(x \mid \theta)p(\theta)/c(\theta)$. This reveals the source of the challenge with constructing a Metropolis-Hastings sampler for $\pi(\theta \mid x)$: if the current state of the Markov chain is θ^t then the proposed state $\theta^* \sim q(\theta^t, \cdot)$, for some proposal $q(\cdot, \cdot)$, is accepted with probability

$$\min\left(1, \frac{p(\theta^*)h(\theta^* \mid x)/c(\theta^*)}{p(\theta^t)h(\theta^t \mid x)/c(\theta^t)} \frac{q(\theta^*, \theta^t)}{q(\theta^t, \theta^*)}\right),$$

where the intractable normalizing function does not cancel out.

In contrast, simulating from the probability model is itself quite straightforward in principle. That is, for a given (fixed) θ, a Metropolis-Hastings update for sampling from the model $f(x \mid \theta)$ is as follows: if the current state of the Markov chain is x^t then the proposed

state $x^* \sim q(x^t, \cdot)$ for some proposal $q(\cdot, \cdot)$ is accepted with probability

$$\min\left(1, \frac{h(x^* \mid \theta)/c(\theta)}{h(x^t \mid \theta)/c(\theta)} \frac{q(x^*, x^t)}{q(x^t, x^*)}\right) = \min\left(1, \frac{h(x^* \mid \theta)}{h(x^t \mid \theta)} \frac{q(x^*, x^t)}{q(x^t, x^*)}\right),$$

which does not depend on the intractable normalizing function $c(\theta)$. This simple observation allows for easy simulation from the probability model, as first shown in Metropolis et al. (1953). As we will discuss below, simulation from $f(y \mid \theta)$ is key to most algorithms for simulating from $\pi(\theta \mid x)$. It is important to note that while the normalizing function poses no challenge to simulation, slow mixing is a common issue, inspiring many innovations, notably the Swendsen-Wang algorithm (Swendsen and Wang, 1987) and variants.

The earliest approaches, mostly focused on approximating maximum likelihood estimators, were based on pseudolikelihood approximations (Besag, 1975, Lindsay, 1988). These likelihood approximations are obtained by taking the product of the full conditional distributions of all variables, are simple and computationally expedient. However, they are limited to certain classes of spatial and network models. They do not apply, for instance, to the Conway-Maxwell Poisson (Conway, 1961) model or Mallows model (Lu and Boutilier, 2014; Mallows, 1957). In cases where one can apply it, pseudolikelihood is often a poor approximation. For example, it can work poorly when the dependence is moderately strong for the Potts model or the autologistic model (cf. Hughes et al., 2011; Okabayashi et al., 2011). In contrast, composite likelihoods, a general scheme for approximation derived by multiplying a collection of component likelihoods (cf. Varin et al., 2011), are much more flexible by allowing various kinds of marginal and conditional component likelihoods, and can be computationally expedient for many situations (cf. Okabayashi et al., 2011, for an applications to the Potts model). Markov chain Monte Carlo maximum likelihood is an elegant, theoretically justified approach to approximating MLEs (Geyer and Thompson, 1992b) using importance sampling to approximate the likelihood function as well as the curvature of the log likelihood. The challenges with MCMCLE are largely related to finding a good importance function that ensures that Monte Carlo errors do not balloon (see Hummel et al., 2012, for some strategies for ERGMs).

Our focus in this chapter is on algorithms for Bayesian inference. We do not discuss max-

imum likelihood estimation for intractable normalizing function models, nor do we discuss the large number of algorithms focused entirely on approximating normalizing functions or constants themselves; these are covered well in (REFS to other Handbook chapters).

1.2 Algorithms

Following the categories in Park and Haran (2018) we can broadly classify algorithms for Bayesian inference with intractable normalizing function models into the following overlapping categories: (i) auxiliary variable (AV) methods, and (ii) likelihood function approximation (LFA) methods. AV methods typically avoid the evaluation of normalizing functions by introducing a well chosen auxiliary variable into the sampling algorithm. LFA methods construct an approximation to the full likelihood function and use the approximation in place of the true likelihood when evaluating the posterior distribution. These approximations can vary widely, from composite likelihood or pseudo-likelihood approximations dating back to the 1970s to the use of Gaussian processes or a variety of new machine learning approaches to approximate the likelihood based on samples from $f(y \mid \theta)$ at various θ values. Likelihood function approximations include many methods that approximate the normalizing function. The categories clearly overlap heavily since most LFA algorithms require drawing samples from the data model $f(y \mid \theta)$, and hence may also be considered AV algorithms. In spite of the overlap, we find the above categorization to be helpful as a way to distinguish the fairly distinct thought processes behind constructing different algorithms for intractable normalizing function problems. The above algorithms may also be categorized as asymptotically exact or inexact: for asymptotically exact algorithms the asymptotic distribution of the stochastic process produced by the algorithm – often but not always a Markov chain – is exactly equal to $\pi(\theta \mid y)$; inexact algorithms do not have this property.

We provide an overview of algorithms for Bayesian inference for models with intractable normalizing functions (henceforth INF). We sprinkle in discussions of algorithms for models where the entirety of the likelihood is intractable (henceforth IL) because, of course, algorithms for IL also apply to INF problems. Note that there is a vast and fast growing

literature on IL methods that spans multiple disciplines; see Cranmer et al. (2020) for an authoritative review. As will become apparent, algorithms for sampling from $f(y \mid \theta)$ end up being crucial to almost all algorithms developed for INF or IL problems.

1.2.1 Auxiliary Variable Algorithms

Here we describe algorithms where an additional simulation is used in various ways to cancel out normalizing function evaluations in the Metropolis-Hastings acceptance ratio. Both the exchange algorithm and double Metropolis-Hastings (DMH) apply to INF problems. We also describe briefly the MCMC version of the Approximate Bayesian Computation (ABC) approach. This algorithm is widely applicable to IL problems, entirely eliminating the need for evaluating the likelihood function.

Exchange Algorithm

The exchange algorithm (Møller et al., 2006; Murray, 2007) relies on an exact sample from the probability model $f(\cdot \mid \theta)$ to produce a Markov chain with stationary distribution $\pi(\theta \mid x)$. Let the tth state of the Markov chain with augmented state space be $(\theta^t, x^t) \in \Theta \times \mathcal{X}$. In the exchange algorithm, each time a parameter value θ^* is proposed from $q(\theta^t, \cdot)$, an auxiliary sample from the probability model at that parameter value θ^* is drawn, $x^* \sim f(\cdot \mid \theta^*)$. The joint proposal (θ^*, x^*) is then accepted or rejected together to construct a Markov chain on an augmented state space. The Metropolis-Hastings acceptance ratio for the proposal (θ^*, x^*) is

$$\alpha(\theta^t, \theta^*) = \min\left\{1, \frac{p(\theta')h(x \mid \theta')/\cancel{c(\theta')}}{p(\theta^t)h(x \mid \theta^t)/\cancel{c(\theta^t)}} \; \frac{q(\theta', \theta^t)h(x^* \mid \theta^t)/\cancel{c(\theta^t)}}{q(\theta^t, \theta')h(x^* \mid \theta')/\cancel{c(\theta')}}\right\}, \tag{1.2.1}$$

which does not contain normalizing function evaluations. The distribution of the marginal chain, that is, just the θ component of the resulting Markov chain, has stationary distribution $\pi(\theta \mid x)$. This is a very elegant approach to constructing an asymptotically exact sampler for the target posterior. An important requirement for this algorithm is that we have an exact draw $x^* \sim f(\cdot \mid \theta^*)$, that is, using a draw from a Markov chain with stationary distribution

$f(x \mid \theta^*)$ does not suffice. For most models of interest, it is either impossible to construct an exact sampler for $f(x \mid \theta^*)$ or, in cases where it is possible, for example using perfect sampling techniques (Propp and Wilson, 1996), it is often computationally too expensive to be of value. In fact, in all the examples we provide in Section 1.4 perfect sampling is not a viable option.

Double Metropolis-Hastings and Variants

The Double Metropolis-Hastings algorithm (DMH) (Liang, 2010) simply takes the exchange algorithm and replaces the exact draw from $f(x \mid \theta^*)$ with a draw from a Markov chain with $f(x \mid \theta^*)$ as its stationary distribution. The algorithm gets its name from the fact that at each iteration of the Markov chain ("outer chain") the algorithm requires another Markov chain ("inner chain") to provide the auxiliary draw. This is obviously expensive, especially if the inner chain is long, but removing the requirement of having an exact draw makes it far more flexible and efficient than the exchange algorithm. There are no theoretical guarantees regarding the quality of DHM samples since the theory for DMH requires both inner and outer chain lengths get large simultaneously, which is impractical. Thus, its efficiency comes at a considerable cost, namely that DMH is asymptotically inexact so standard approaches for assessing convergence (Flegal et al., 2008) do not apply.

As we discuss in Section 1.4 DMH is both efficient and relatively easy to construct. Important ingredients for the construction of DMH include coming up with a good proposal for θ ($q(\theta, \cdot)$), and determining the length of the inner and outer chains. Improving proposals and determining a suitable outer chain length are standard issues in constructing MCMC algorithms but the inner chain length problem is specific to DMH. All three of these issues require the ability to assess the quality of samples produced by DMH; given the fact that the algorithm is asymptotically inexact usual MCMC diagnostics are not useful. This is an issue we will address more broadly in Section 1.3 and we discuss the implementation of DMH in challenging examples in Section 1.4.

Approximate Bayesian Computation (ABC)

Approximate Bayesian Computation (ABC) is a very widely used class of algorithms for Bayesian inference in the presence of intractable likelihood functions. Here we describe ABC-MCMC (Marjoram et al., 2003), a subset of ABC algorithms. In ABC-MCMC, for current state θ^t when θ^* is proposed the Metropolis-Hastings acceptance ratio is evaluated only if an auxiliary sample $x^* \sim f(\cdot \mid \theta^*)$ is close to the data according to some user-specified distance $d(x, x^*)$ and threshold $\epsilon > 0$; of course the distance could be defined on a statistics $S(x^*)$, which is a fruitful area to explore the use of dimension reduction methods. If $d(x, x^*) < \epsilon$, θ' is accepted with probability

$$\alpha(\theta^t, \theta^*) = \min\left\{1, \frac{p(\theta^*)}{p(\theta^t)} \; \frac{q(\theta^*, \theta^t)}{q(\theta^t, \theta^*)}\right\}, \tag{1.2.2}$$

For discrete data ($\mathcal{X}$ is a discrete space) when $\epsilon = 0$, Marjoram et al. (2003) use a simple detailed balance argument to prove that the resulting Markov chain converges to stationary distribution $\pi(\theta \mid x)$. Of course, this is an impractical requirement so the Markov chain with $\epsilon > 0$ is in reality asymptotically inexact. Because of the growing number of problems where the evaluation of $f(x \mid \theta)$ is impossible even while auxiliary simulation from $f(x \mid \theta)$ is relatively straightforward, ABC-MCMC and other "likelihood-free" algorithms have become very popular. ABC-MCMC is of course applicable to INF problems but its value in comparison to more specialized INF algorithms like DMH is still an open question.

1.2.2 Likelihood Approximation Algorithms

Some of the earliest approaches for inference with intractable normalizing function likelihoods involve replacing the entire likelihood function by an approximation that completely avoids computing the normalizing function. This idea remains potentially very useful and opens avenues for research, particularly the use of new machine learning methods for approximation. In the interest of brevity, and because likelihood approximations are major topics in their own right, we provide only brief notes on some of the key ideas in this class of algorithms.

Composite Likelihood

In composite likelihood (Lindsay, 1988), the likelihood function $\mathcal{L}(\theta \mid x)$ is replaced by an approximation $\hat{\mathcal{L}}_C(\theta \mid x)$ obtained by taking the product of a collection of component likelihoods. A maximum composite likelihood estimate (MCLE) is obtained by maximizing $\hat{\mathcal{L}}_C(\theta \mid x)$ with respect to θ. MCLEs are a good approximation to the MLE in certain settings (cf. Varin et al., 2011). Composite likelihood methods are also easy to implement and computationally expedient. In fact, they are the only class of algorithms discussed in this paper that do not require auxiliary simulations. Pseudolikelihood approaches, where the likelihood is approximated by a product of full conditional distributions of the random variables in the model, are somewhat limited in their flexibility but composite likelihood methods can be adapted well to particular problems. A nice example is provided in Okabayashi et al. (2011) for the Potts model for discrete spatial lattice data. Here, $\hat{\mathcal{L}}_C(\theta \mid x)$ is the product of the joint distributions of multiple pixels given the rest of the lattice. They find that the resulting MCLE is less statistically efficient but much faster and easier to compute than the MLE; the MLE can only be approximated using a combination of MCMC and importance sampling (Geyer and Thompson, 1992a). For Bayesian inference with composite likelihood, $\pi(\theta \mid x)$ is replaced with $\hat{\pi}(\theta \mid x) \propto \hat{\mathcal{L}}_C(\theta \mid x)p(\theta)$, as long as $\hat{\pi}(\theta \mid x)$ is proper. Composite likelihood for Bayes is a somewhat under-explored approach, though there are some studies for example for spatial extreme models (Ribatet et al., 2012) and for calibrating complex computer models (Chang et al., 2015).

Intractable Normalizing Function Approximations

There is a large literature on approximating the normalizing function using importance sampling, bridge sampling and path sampling methods Gelman and Meng (1998). These algorithms can all be used to approximate the Metropolis-Hastings acceptance ratio. For instance (Atchade et al., 2008) construct a stochastic process much like regular MCMC except the normalizing function is adaptively approximated at each step of the algorithm by a stochastic approximation using the entire sample path up to that point. In order to make the normalizing function approximation approach efficient Atchade et al. (2008) propose

using an umbrella sampling approach with multiple particles at each iteration (Torrie and Valleau, 1977).

Pseudo-marginal Algorithms

In the Metropolis-Hastings acceptance ratio, pseudo-marginal MCMC (Andrieu and Roberts, 2009) replaces an intractable $\mathcal{L}(\theta \mid x)$ with its positive and unbiased Monte Carlo approximation $\hat{\mathcal{L}}(\theta \mid x)$. The resulting algorithm is useful for IL problems, and has the advantage of being asymptotically exact.

For INF problems, pseudo-marginal MCMC requires an unbiased approximation of $1/c(\theta)$. Although we can easily obtain an unbiased approximation for $c(\theta)$, $\hat{c}(\theta)$, using importance sampling, obtaining an unbiased approximation of $1/c(\theta)$ is non-trivial. Note that $1/\hat{c}(\theta)$ is a consistent but biased approximation. To address this, Lyne et al. (2015) developed a geometric series correction method, called the Russian roulette algorithm. Under the pseudo-marginal framework, the algorithm is asymptotically exact, and assumptions are satisfied for general forms of $h(x \mid \theta)$. However, implementing the stochastic truncation of the series requires multiple $\hat{c}(\theta)$s; considering that obtaining each $\hat{c}(\theta)$ requires Monte Carlo samples from $h(x \mid \theta)$, the algorithm is computationally expensive.

To speed up the algorithm, one might consider adapting surrogate likelihood approximations. For IL problems, Drovandi et al. (2018) developed an approach to accelerate the pseudo-marginal algorithm by using a Gaussian process approximation of the log of an unbiased likelihood approximation. Developing asymptotically exact pseudo-marginal algorithms that are also computationally efficient is very challenging for INF problems.

Bayesian Synthetic Likelihood

Bayesian synthetic likelihood (Price et al., 2018) (BSL) is a likelihood-free algorithm that, like ABC, uses simulations from the probability model to modify the Metropolis-Hastings acceptance ratio. For each proposed θ^*, the algorithm generates m simulations $x_1^*, \ldots, x_m^*$

from $f(x \mid \theta^*)$, computes a summary statistic based on each simulation, $S(x_1^*), \ldots, S(x_m^*)$, and then constructs a multivariate normal based on these statistics by simply using the sample mean and sample covariance; this is treated as an approximation to the distribution of the summary statistics. This multivariate normal, evaluated at the real data, is then used as a replacement for the true likelihood function in the Metropolis-Hastings acceptance ratio. The idea behind this algorithm is that if the summary statistics are approximately normal, this is a reasonable approximation. The algorithm is asymptotically inexact, like ABC, but it seems to outperform ABC in terms of computational efficiency in some examples (Price et al., 2018), and so may be a useful addition to the toolkit for likelihood-free inference, and hence also intractable normalizing function problems.

Variational Bayes

Variational Bayes (VB) (Bishop and Nasrabadi, 2006; Jordan et al., 1999) approximates the posterior by minimizing the Kullback-Leibler divergence between $\pi(\theta \mid x)$ and a tractable distribution class, or equivalently maximizing the evidence lower bound. Tran et al. (2017) develop VB for intractable likelihood problems by replacing the likelihood with unbiased importance sampling estimates. For INF problems, Tan and Friel (2020) developed two classes of VB methods for ERGMs with Gaussian posterior approximation. The first approach replaces the intractable likelihood with the adjusted pseudolikelihood (Bouranis et al., 2018) in optimizing the evidence lower bound. These adjustments correct the mode, curvature, and magnitude of the pseudolikelihood based on an affine transformation. The second one is a stochastic gradient ascent approach to optimize the evidence lower bound. Here, the gradient term is approximated through importance sampling estimates.

VB approaches are attractive because they are potentially extremely fast, but both VB approaches in Tan and Friel (2020) need considerable tuning to be effective for a given distribution. First, we need to tune the algorithms for both approaches carefully. The adjusted pseudolikelihood approach uses MLE and covariance estimates of sufficient statistics for an affine transformation, which requires Monte Carlo simulation from the model. Therefore, the quality of the preliminary iteration of MCMC-MLE (Geyer and Thompson, 1992a;

Snijders et al., 2002) is crucial for the success of the algorithm. The stochastic gradient approach also uses the fitted results from the adjusted pseudolikelihood method as inputs for faster convergence; the performance of the algorithm can depend on them. Furthermore, both VB approaches require simulations from the model, for adjusting pseudolikelihood and estimating the gradient respectively. However, if carefully tuned, VB approaches can be computationally faster than other approaches, including DMH, while providing reasonable posterior approximations. For example VB requires fewer auxiliary model simulations than DMH and the simulations can be parallelized. It is important to note that the VB methods (Tan and Friel, 2020) have been developed for ERGMs; it is unclear how to extend them to other models.

Surrogate Likelihoods

A particularly exciting avenue for both INF and IL methods is exploring new machine learning methods that can be trained on samples to produce a surrogate model and likelihood function. The sampling can be run ahead of time, in parallel, and once the training is done an MCMC algorithm can use the surrogate likelihood. Hence, these algorithms can be very fast, though there are interesting challenges in terms of design – how to select the set of parameters for simulation – and architecture – how to construct the surrogate. For instance, Sainsbury-Dale et al. (2023) develops a neural network approximation to a Bayes estimator in the context of spatial data, and there are recent maximum likelihood approaches for expensive or intractable likelihoods (Walchessen et al., 2023). There is a long history of using Gaussian processes in the complex computer models framework, that is, for inference where the model is a complex simulation model with no closed-form expressions (cf. Gramacy, 2020; Santner et al., 2003). Surrogates based on Gaussian processes have been developed for INF problems (cf. Drovandi et al., 2018; Park and Haran, 2020). For instance, in the LikeEm algorithm (Park and Haran, 2020), the normalizing function is approximated for a set of parameter settings using importance sampling techniques, then a Gaussian process is used to interpolate the approximated likelihood function over the entire parameter space. Once this pre-computing step is completed, the Gaussian process emulator is used instead of the likelihood function in the subsequent MCMC algorithm. This algorithm is fast but

asymptotically approximate; it is studied as an example in Section 1.4. The review in Cranmer et al. (2020) is targeted at IL problems rather than INF problems, and contains helpful descriptions of several simulation-based surrogate likelihood methods.

1.3 Assessing Sample Quality

Assessing the quality of samples from MCMC algorithms has been an active area of research for over three decades, with many theoretically justified and practical approaches (cf. Roy (2020) and Chapters 4 and 5 of this book). However, there are relatively few if any attempts at finding ways to assess the quality of samples produced by asymptotically inexact algorithms. Such diagnostics are important for ensuring the reliability of our results as well as for guidance for tuning our algorithm for any given problem. For example, the double Metropolis-Hastings algorithm requires determining the length of the inner and outer Markov chains; standard MCMC diagnostics are not useful for tuning this algorithm or assessing the quality of samples. Heuristics for tuning are often based on simulated examples and hence can be of limited applicability to a particular problem. Determining which algorithm to prefer is also difficult without a good measure of sample quality, for example the exchange algorithm is asymptotically exact but computationally expensive while the double Metropolis-Hastings algorithm is faster but asymptotically inexact; which should we use for a given problem?

There are two measures of sample quality provided by Kang et al. (2023), the approximate curvature diagnostic (ACD) and the approximate inverse multiquadric kernel Stein discrepancy (AIKS). In the interest of brevity and because we find ACD to be more computationally expedient, we focus on ACD here. For more on AIKS and the kernel Stein discrepancy on which it is based, see Kang et al. (2023) and Gorham and Mackey (2017) respectively.

1.3.1 Curvature Diagnostic

The curvature diagnostic (Kang et al., 2023) is inspired by maximum likelihood theory. The second Bartlett identity (Bartlett, 1953a,b) can be used to assess whether a model is correctly specified, that is, whether the data we observe are compatible with a particular probability model. The curvature diagnostic considers whether the samples drawn from an algorithm are compatible with the target posterior distribution $\pi(\theta \mid x)$. Let $u_{\mathbf{x}}(\theta) = \nabla_\theta \log \pi(\theta \mid x), H(\theta) = \frac{\partial}{\partial \theta} u_{\mathbf{x}}(\theta)$, $J(\theta) = u_{\mathbf{x}}(\theta) u_{\mathbf{x}}(\theta)^\top$, and $d(\theta) = \text{vech}[J(\theta) + H(\theta)]$, where $\text{vech}(M)$ denotes the half-vectorization of the matrix M. If the samples from the algorithm are truly from the posterior, then $\mathrm{E}_\pi\{d(\theta)\} = 0$ by Bartlett's second identity. This leads to the following strategy: for samples $\{\theta^{(1)}, \ldots, \theta^{(n)}\}$ from the approximate algorithm, estimate the expectation as $d_n = \frac{1}{n} \sum_{i=1}^n d(\theta^{(i)})$. If the asymptotic distribution of the sample is the target distribution π, then $\sqrt{n} d_n \xrightarrow{d} \mathrm{N}(0, V)$ by the central limit theorem (CLT) for independent samples and MCMC CLT for samples from a Markov process. For independent samples, the unbiased and consistent approximation of V is calculated as $V_n = \frac{1}{n} \sum_{i=1}^n d(\theta^{(i)}) d(\theta^{(i)})^\top$. For samples from a Markov chain, we can estimate V using batch means estimator which is strongly consistent under some conditions (Damerdji, 1994; Jones et al., 2006; Vats et al., 2019). The curvature diagnostic is defined as $n d_n^\top V_n^{-1} d_n$. This has an asymptotic $\chi^2(r)$ distribution, where $r = p(p+1)/2$ and p is the dimension of θ, if the asymptotic distribution of the sample is equal to the target posterior. The $1 - \alpha$ quantile of the $\chi^2(r)$ can be used as a threshold for this diagnostic. A sample path for which the diagnostic value is below the threshold is considered to have an asymptotic distribution that is reasonably close to the target distribution.

1.3.2 Approximate Curvature Diagnostic (ACD)

In the context of intractable normalizing function problems, $H(\theta)$ and $J(\theta)$ are intractable and hence need to to be approximated, leading to the approximate curvature diagnostic (ACD). Kang et al. (2023) describes how ACD is computed efficiently, providing theoretical justification for using ACD in place of CD. The ACD is defined as $n \hat{d}_{n,N}^\top \hat{V}_{n,N}^{-1} \hat{d}_{n.,N}$ where $\hat{d}_{n,N}$ and $\hat{V}_{n,N}$ are the consistent estimates of d_n and V_n, respectively. The approximations are

obtained using auxiliary samples $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(N)}$ generated exactly from $f(\cdot \mid \theta)$ or generated by a Monte Carlo algorithm having $f(\cdot \mid \theta)$ as its stationary distribution. The self-normalized importance sampling (Tan and Friel, 2020) substantially speeds up the approximation step. Further details on the approximation are found in Kang et al. (2023). In order to control standard errors and computational costs, it is important to construct these approximations carefully. The theoretical justifications may be summarized as follows: if the asymptotic distribution of the sample is equal to the target distribution then $\hat{d}_{n,N} \overset{\text{a.s.}}{\to} d_n$ as $N \to \infty$ and $\hat{V}_{n,N} \overset{\text{a.s.}}{\to} V$ as $n, N \to \infty$, so that the ACD converges to the curvature diagnostic. These results hold under reasonable conditions on the prior and likelihood; for instance they are satisfied for the challenging examples provided later in this chapter. ACD has been shown to be effective in practice, as demonstrated on multiple challenging examples in Kang et al. (2023). Further details on the application of ACD are in Section 1.4.

1.4 Applications

We showcase the application of three algorithms – DMH (Liang, 2010), LikeEm (Park and Haran, 2020), and VB (Tan and Friel, 2020) – in the context of three challenging examples – the Potts model, an exponential random graph model (ERGM), and an Ising network model. The algorithms we have chosen are intentionally quite different from each other. DMH is an auxiliary variable algorithm, while LikeEm is a function approximation approach that uses sampling to produce a Gaussian process surrogate to the likelihood, and VB is a function approximation approach that uses optimization. The algorithms share one common feature – we find that they are of practical value in certain contexts, though not all the algorithms apply to all the examples we provide. We find that our examples provide insights about the algorithms and also demonstrate the value of the ACD measure of sample quality in tuning the algorithms, assessing whether the approximations are reasonable, and comparing the algorithms to each other.

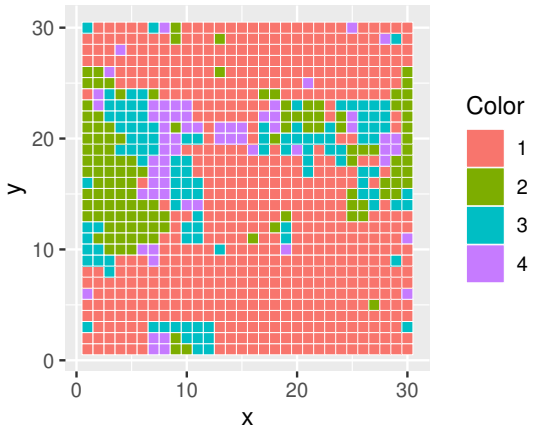

Figure 1.1: Multi-color image data simulated from the Potts model with $K = 4$ and $\theta = \log(1 + \sqrt{4})$.

1.4.1 Potts Model

The Potts model (Potts, 1952), a generalization of the Ising model (Ising, 1925; Lenz, 1920), provides an approach for modeling multi-colored images and hence discrete-valued spatial data on a lattice. For an $r \times s$ lattice x with discrete values $x_i \in \{1, \ldots, K\}$, the Potts model with $\theta > 0$ has probability model

$$f(x \mid \theta) = \frac{1}{c(\theta)} \exp\left\{\theta \sum_{i \sim j} I(x_i = x_j)\right\},$$

where $i \sim j$ indicates neighboring elements, and $I(\cdot)$ denotes the indicator function. A larger value for θ produces higher expected number of neighboring pairs that have the same color. Calculation of the normalizing function $c(\theta)$ requires summation over all K^{rs} possible outcomes for the model, which is computationally infeasible even for lattices of moderate size. We simulated a 30×30 lattice with $K = 4$ and $\theta = \log(1 + \sqrt{4})$ via 100,000 cycles of the Swendsen-Wang algorithm (Swendsen and Wang, 1987) using the R package `potts` (Geyer and Johnson, 2022). The simulated data are presented in Figure 1.1.

For this example we consider the DMH and LikeEm algorithms that are described in Section 1.2. Both algorithms are asymptotically inexact. We implement DMH with different numbers m of Swendsen-Wang (inner sampler) updates. The LikeEm algorithm has two

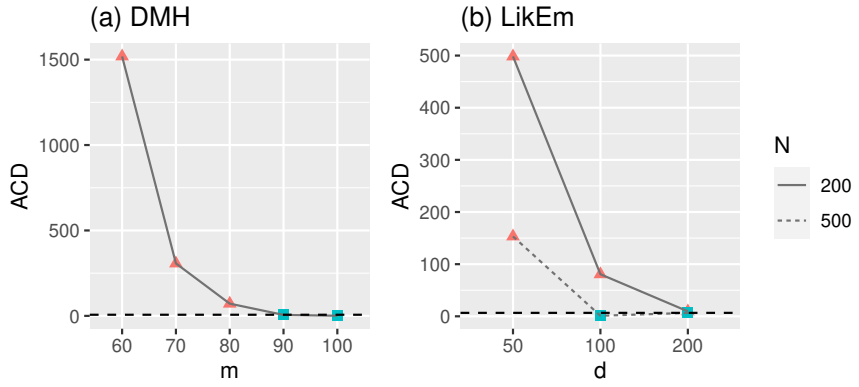

Figure 1.2: Results for the simulated data from the Potts model. (a) ACD applied to samples generated from DMH with different numbers m of (inner) Gibbs updates. (b) ACD applied to samples generated from LikeEm with different numbers d of particles and different sample sizes N for importance sampling estimates. The dashed horizontal lines represent the threshold value of ACD, and the red triangles and blue squares indicate poor sample quality and good sample quality, respectively. For DMH, ACD recommends m of at least 90. For LikeEm, ACD recommends d of at least 100 and $N = 500$.

tuning components, the number d of particles and the sample size N for importance sampling estimates. We implement LikeEm with different combination of d and L, where the particles are chosen from a short run of DMH with 90 cycles of inner updates. All algorithms were run for n = 300,000 iterations. We apply ACD to choose a suitable value of m for DMH and an appropriate combination of d and N for LikeEm. For ACD $N = 200{,}000$ auxiliary variables were generated via Swendsen-Wang sampling for approximating $d(\theta)$ at each unique posterior sample point. The threshold value of ACD is the 0.99 quantile of $\chi^2(1)$, which is 6.63.

Figure 1.2 (a) shows ACD for the DMH sample for a sequence of m values, and (b) presents ACD for the LikeEm sample for different combinations of d and N. Based on ACD, we choose $m = 90$ for DMH, and $d = 100$ and $N = 500$ for LikeEm. DMH with $m = 90$ takes approximately 2 hours but LikeEm with $d = 100$ and $N = 500$ only takes 50 seconds to run, including pre-computing time for particle selection. In summary, LikeEm can provide good-quality samples at a fraction of the computational cost compared to DMH.

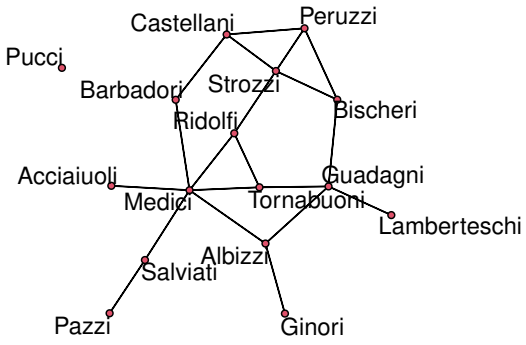

Figure 1.3: Florentine marriage network (Breiger and Pattison, 1986) consists of 16 vertices and 20 undirected edges.

1.4.2 An Exponential Random Graph Model

Exponential random graph models (ERGM) (Hunter et al., 2008; Robins et al., 2007b) have been widely used to study the relationships between network nodes. Consider we have undirected network data $x \in \mathbb{R}^{n\times n}$ with n nodes. For all $i \neq j$, $x_{i,j} = 1$ if the ith node and jth node are connected (neighbors), and $x_{i,j} = 0$ otherwise. Then, the probability model of ERGM is

$$f(x \mid \theta) = \frac{1}{c(\theta)} \exp\left\{\theta_1 S_1(x) + \theta_2 S_2(x)\right\},$$

$$S_1(x) = \sum_{i=1}^{n} \binom{x_{i+}}{1} \qquad S_2(x) = e^{0.2} \sum_{k=1}^{n-2} \left\{1 - (1 - e^{-0.2})^k\right\} ESP_k(x)$$

where $S_1(x)$ and $S_2(x)$ indicate edges and the geometrically weighted edge-wise shared partnership (GWESP) statistics (Hunter and Handcock, 2006a), respectively. The $ESP_k(\mathbf{x})$ term in the GWESP statistic indicates the number of connected i, j pairs, where i and j have k common neighbors. Therefore, GWESP can account for higher order transitivities with geometric weights. Evaluation of the normalizing function $c(\theta)$ is intractable because it requires summation over all $2^{n(n-1)}$ possible configurations in the network. We study the Florentine marriage dataset Breiger and Pattison (1986), which describes the marriage alliance networks among 16 Florentine families and is shown in Figure 1.3.

For this example we consider DMH and VB (Tan and Friel, 2020) algorithms that are described in Section 1.2. Both algorithms are asymptotically inexact. We implement DMH with different numbers m of Gibbs (inner) updates. For each choice of m, we run DMH for n = 200,000 iterations and apply ACD to the sample. We tune the VB algorithm according to Tan and Friel (2020) and obtain an approximate posterior distribution. We generate 100,000 samples independently from the approximate posterior and apply ACD to sample. For ACD N = 200,000 auxiliary variables were generated via Gibbs sampling for approximating $d(\theta)$ at each unique posterior sample point. The threshold value of ACD is the 0.99 quantile of $\chi^2(3)$, which is 11.34.

Figure 1.4 shows ACD for the DMH sample for a sequence of m values. ACD implies that DMH with $m = 3$ performs well in this example. ACD for VB is 956.31 which is much

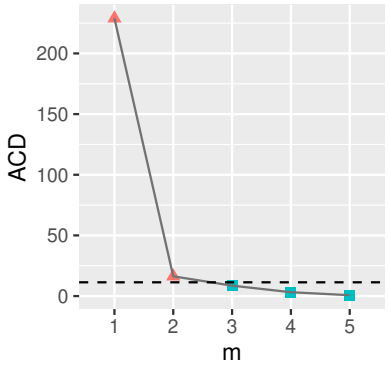

Figure 1.4: Results for the Florentine marriage dataset in the ERGM. ACD applied to samples generated from DMH with different numbers m of (inner) Gibbs updates. The dashed horizontal lines represent the threshold value of ACD, and the red triangles and blue squares indicate poor sample quality and good sample quality, respectively. ACD recommends m of at least 3.

Table 1.1: Summary statistics of posterior samples and a gold standard for parameters θ_1 and θ_2 in the ERGM for the Florentine marriage dataset.

Algorithm	θ_1				θ_2			
	Median	SD	LTP	RTP	Median	SD	LTP	RTP
VB	-1.73	0.37	0.06	0.05	0.08	0.29	0.03	0.06
DMH with $m = 3$	-1.70	0.37	0.05	0.05	0.05	0.29	0.05	0.05
Gold standard	-1.70	0.37	0.05	0.05	0.05	0.30	0.05	0.05

SD, standard deviation; LTP, left-tail probability; RTP, right-tail probability.

greater than the threshold value of 11.34. This implies that VB may produce poor sample quality even though we carefully tune the algorithm.

Table 1.1 presents summary statistics for the VB sample, DMH sample with $m = 3$, and a gold standard. We treat as the gold standard a run from the DMH algorithm with $m = 20$. Cutoff values for the left- and right-tail probabilities are the lower 5% and the upper 5% percentiles of the gold standard. We observe that the DMH sample with $m = 3$ provides almost the same values of the summary statistics as the gold standard. On the other hand, the VB sample does not match the gold standard well. For θ_1, it provides a small median and a slightly high left-tail probability compared to the gold standard. For θ_2, it provides a large median, a low left-tail probability, and a slightly high right-tail probability compared to the gold standard. DMH with $m = 3$ takes approximately 1.6 minutes to run. This shows that DMH not only permits fast computation but can generate high quality samples in this example.

1.4.3 An Ising Network Model: Applications to Verbal Aggression Data

Ising network models (Van Borkulo et al., 2014) are undirected graphical models that can describe interactions among binary responses. Consider binary item response data $x \in \mathbb{R}^{n\times p}$ with n responses to p items. For all i, j, $x_{i,j} = 1$ if the ith respondent answers the jth item correctly (or positively), and $x_{i,j} = 0$ otherwise. The Ising network model with parameters $\theta = (\beta, \gamma)$ has probability model

$$f(x \mid \theta) = \frac{1}{c(\theta)} \exp\left\{ \sum_{j=1}^{p} \beta_j \sum_{i=1}^{n} x_{ij} + \sum_{j<k} \gamma_{jk} \sum_{i=1}^{n} x_{ij} x_{ik} \right\}, \tag{1.4.1}$$

where β_j is an item easiness parameter, and γ_{jk} is a pairwise interaction parameter between item j and k. Calculation of the normalizing functions $c(\theta)$ require summation over all 2^{np} possible configurations, which is intractable. Furthermore, the model includes $p + \binom{p}{2}$ parameters, which are high-dimensional. Here we analyze the item responses to a questionnaire on verbal aggression (De Boeck, 2004). All items are about verbally aggressive reactions in a frustrating situation, and we focus on studying 12 *want* behavior mode items as follows:

(1) A bus fails to stop for me. I would want to curse.

(2) A bus fails to stop for me. I would want to scold.

(3) A bus fails to stop for me. I would want to shout.

(4) I miss a train because a clerk gave me faulty information. I would want to curse.

(5) I miss a train because a clerk gave me faulty information. I would want to scold.

(6) I miss a train because a clerk gave me faulty information. I would want to shout.

(7) The grocery store closes just as I am about to enter. I would want to curse.

(8) The grocery store closes just as I am about to enter. I would want to scold.

(9) The grocery store closes just as I am about to enter. I would want to shout.

(10) The operator disconnects me when I had used up my last 10 cents for a call. I would want to curse.

(11) The operator disconnects me when I had used up my last 10 cents for a call. I would want to scold.

(12) The operator disconnects me when I had used up my last 10 cents for a call. I would want to shout.

All responses were dichotomized to have binary values (either 1 ="yes" or 0 ="no"). The data include $n = 316$ respondents for the $p = 12$ items described above; the resulting Ising network model has $12 + \binom{12}{2} = 78$ parameters.

In order to detect significant interactions among items, we apply the spike and slab DMH sampler (Park et al., 2022) that poses spike and slab priors for θ as follows:

$$\begin{aligned} \theta_i \mid \lambda_i, \sigma^2, \omega &\overset{\text{ind}}{\sim} \lambda_i N(0, \omega^2\sigma^2) + (1-\lambda_i)N(0, \sigma^2), \\ \lambda_i &\overset{\text{iid}}{\sim} \text{Bernoulli}(1/2), \\ \frac{1}{\sigma^2} &\sim \text{Uniform}(4, 100), \\ \omega &\sim 1 + Y, \quad Y \sim \text{Gamma}(1, 1/100), \end{aligned} \tag{1.4.2}$$

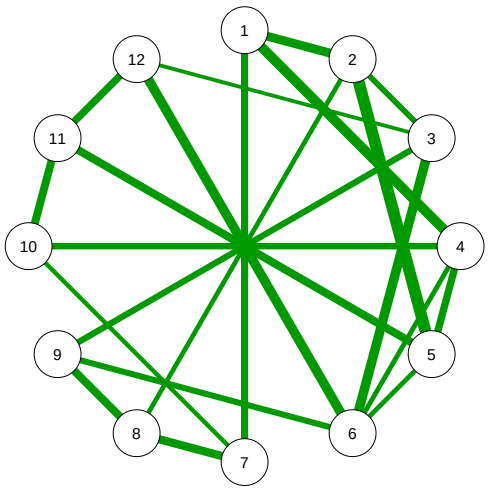

Figure 1.5: An estimated network structure for the verbal aggression data. The width of the lines indicates the strength of the connection between the relevant items; thicker lines indicate stronger interaction between items.

where λ_i is a latent variable indicating whether θ_i is included in the model ($\lambda_i = 1$) or not ($\lambda_i = 0$), and σ^2 and ω control the variances of spike and slab distributions, respectively. Inference for the parameters is carried out by the DMH algorithm. DMH is a practical option for such high-dimensional hierarchical models with intractable normalizing functions. DMH can be implemented in general cases once we have an inner sampler. Other algorithms like VB and LikeEm can be hard to tune. For DMH, the length m of the inner sampler should be carefully chosen. However, ACD is impractical as it requires approximating the 3,081 $\times$ 3,081 covariance matrix of $d(\theta)$ the dimension of which is . To provide a reliable approximation to the covariance matrix, we need long posterior sample paths, resulting in huge demands on memor. Park et al. (2022) study the performance of the spike and slab DMH sampler with different lengths m of inner sampler. Following the suggestion in Park et al. (2022), we run the algorithm with $m = 10n$, where n is the number of respondents.

Of the 66 pairwise interaction parameters γ_{jk}, the spike and slab DMH shrinks 45 interaction parameters toward 0 and provides positive values for the others. Figure 1.5 illustrates the resulting estimated network structure. We observe that $\gamma_{2,5}$, $\gamma_{1,4}$, and $\gamma_{3,6}$ have the greatest posterior means, indicating the strongest positive interactions. The strongest positive

interaction is observed between items (2) ("A bus fails to stop for me. I would want to scold") and (5) ("I miss a train because a clerk gave me faulty information. I would want to scold"), which makes sense because both items are about scold behavior. The second strongest positive interaction appears between items (1) ("A bus fails to stop for me. I would want to curse") and (4) ("I miss a train because a clerk gave me faulty information. I would want to curse"); both items are about curse behavior. Lastly, the third strongest positive interaction occurs between items (3) ("A bus fails to stop for me. I would want to shout") and (6) ("I miss a train because a clerk gave me faulty information. I would want to shout"), indicating that shout behaviors are also strongly connected. This example illustrates the shortcomings of the state-of-the-art in computing for INF problems. While DMH appears to be practical in this context, the closest we have to assurance about the quality of sample-based inference is based on a heuristic that relies on experience.

1.5 Summary

Inference in the presence of intractable normalizing functions is an exciting computing problem with lots of room for creativity. While it is difficult to provide an exhaustive review, we hope we have provided a reasonably broad perspective on key ideas for this problem. Because of the number of algorithms and the vast difference among them, as well as the scarcity of theory and heuristics for comparing them, it can be difficult to determine which algorithm to apply for any given situation. The recent measures of sample quality in Kang et al. (2023) may be helpful in this regard, as they have some potential for evaluating not only the quality of samples from a particular algorithm but for comparing algorithms from very different categories, including asymptotically exact and inexact algorithms. We find through our study, using the ACD diagnostic (Kang et al., 2023) to measure sample quality where possible, that the double Metropolis-Hastings algorithm is quite effective and broadly applicable when used in tandem with ACD, and also has the advantage of being easier to code than most algorithms for Bayesian inference with intractable normalizing functions. Of course we must add the caveat that our study of algorithms is necessarily limited to a few challenging examples and a small set of interesting algorithms and hence we do not claim

that our results will necessarily hold across the enormous spectrum of intractable normalizing function problems.

Potential Directions for Research

This chapter has provided a sampling of the many creative ideas that have emerged in recent years for inference with intractable normalizing functions and intractable likelihoods. There are clearly many open practical and theoretical issues. The computational complexity of algorithms for intractable likelihood and intractable normalizing function problems makes them computational expensive, and sometimes impractical, for many real applications. Examples of such problems include interaction point process models (cf. Goldstein et al., 2015), network psychometrics (cf. Van Borkulo et al., 2014), exponential random graph models (see Hunter and Handcock, 2006b; Robins et al., 2007a), mixed graphical models (cf. Cheng et al., 2017; Lauritzen and Wermuth, 1989; Lee and Hastie, 2015), and Conway–Maxwell-Poisson regression models (cf. Chanialidis et al., 2018; Conway and Maxwell, 1962; Shmueli et al., 2005). A major computational bottleneck for most algorithms is the need to generate a large number of expensive auxiliary samples from the probability model. Variational methods offer a promising alternative because the sampling can be parallelized and done in advance; they are potentially very fast but the amount of problem-specific tuning they require to make them work well can make them impractical for many settings; finding ways to make VB work well in these scenarios is an open area for research. While they have been around for a while, composite likelihood methods avoid the need for auxiliary simulation, can be relatively easy to implement, and seem to work well in certain contexts. It may be worth exploring composite likelihood approaches for a variety of INF problems. Finally, more generalized likelihood approximation approaches show much promise. We show in a real example in Section 1.4 that a Gaussian process-based likelihood approximation can be much more efficient while remaining just as reliable as other algorithms for inference. Approximations based on machine learning techniques have advanced a great deal beyond Gaussian processes. Hence, the application of fast machine learning techniques to likelihood function approximations is a promising avenue for future research as well, with many interesting challenges to address in terms of theory and applications.

An important problem is assessing the quality of approximations obtained from asymptotically inexact algorithms. The ACD and AIKS methods (Kang et al., 2023) provide a nice way to measure sample quality for asymptotically inexact algorithms, but do not actually provide a way to measure the error in the approximation of particular quantities of interest, typically taking the form of a specific expectation $E_\pi g(X)$ for a given target π and real-valued function g. As is well known to MCMC users, the requisite length of the chain varies depending on the required accuracy (MCMC standard error) and the specific targeted function g (Flegal et al., 2008). For instance approximating tail probabilities, $g(x) = I(x > c)$ for large c and higher moments, $g(x) = x^2$, tend to require more samples than simple expectations ($g(x) = x$). There are, to our knowledge, no methods for measuring the accuracy of approximations for these different quantities for an asymptotically inexact algorithm when the normalizing function or likelihood function is intractable. Another big open problem is measuring the quality of samples produced by inexact algorithms when the entire likelihood function is intractable, though we found some ABC-specific tools that may be useful (Prangle et al., 2014; Rendsburg et al., 2022).

Acknowledgments

JP was partially supported by the National Research Foundation of Korea (2020R1C1C1A0100386814, RS-2023-00217705) and the ICAN (ICT Challenge and Advanced Network of HRD) support program (RS-2023-00259934) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

Bibliography

Andrieu, C. and Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37(2):697 – 725.

Atchade, Y., Lartillot, N., and Robert, C. (2008). Bayesian computation for statistical models with intractable normalizing constants. Technical report, University of Michigan, Department of Statistics.

Bartlett, M. S. (1953a). Approximate confidence intervals. *Biometrika*, 40(1/2):12–19.

Bartlett, M. S. (1953b). Approximate confidence intervals. II. More than one unknown parameter. *Biometrika*, 40(3/4):306–317.

Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*. Springer.

Bouranis, L., Friel, N., and Maire, F. (2018). Bayesian model selection for exponential random graph models via adjusted pseudolikelihoods. *Journal of Computational and Graphical Statistics*, 27(3):516–528.

Breiger, R. L. and Pattison, P. E. (1986). Cumulated social roles: The duality of persons and their algebras. *Social networks*, 8(3):215–256.

Chang, W., Haran, M., Olson, R., and Keller, K. (2015). A composite likelihood approach to computer model calibration with high-dimensional spatial data. *Statistica Sinica*, pages 243–259.

Chanialidis, C., Evers, L., Neocleous, T., and Nobile, A. (2018). Efficient Bayesian inference for COM-Poisson regression models. *Statistics and Computing*, 28:595–608.

Cheng, J., Li, T., Levina, E., and Zhu, J. (2017). High-dimensional mixed graphical models. *Journal of Computational and Graphical Statistics*, 26:367–378.

Conway, R. (1961). A queueing model with state dependent service rate. *Journal of Industrial Engineering*, 12:132.

Conway, R. W. and Maxwell, W. L. (1962). Network dispatching by the shortest-operation discipline. *Operations Research*, 10:51–73.

Cranmer, K., Brehmer, J., and Louppe, G. (2020). The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062.

Damerdji, H. (1994). Strong consistency of the variance estimator in steady-state simulation output analysis. *Mathematics of Operations Research*, 19:494–512.

De Boeck, P. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. Springer Science & Business Media.

Drovandi, C. C., Moores, M. T., and Boys, R. J. (2018). Accelerating pseudo-marginal MCMC using Gaussian processes. *Computational Statistics & Data Analysis*, 118:1–17.

Flegal, J., Haran, M., and Jones, G. (2008). Markov chain Monte Carlo: Can we trust the third significant figure? *Statistical Science*, 23:250–260.

Gelman, A. and Meng, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical science*, pages 163–185.

Geyer, C. and Johnson, L. (2022). potts: Markov Chain Monte Carlo for Potts Models.

Geyer, C. J. and Thompson, E. A. (1992a). Constrained Monte Carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54(3):657–683.

Geyer, C. J. and Thompson, E. A. (1992b). Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 54:657–683.

Goldstein, J., Haran, M., Simeonov, I., Fricks, J., and Chiaromonte, F. (2015). An attraction–repulsion point process model for respiratory syncytial virus infections. *Biometrics*, 71(2):376–385.

Gorham, J. and Mackey, L. (2017). Measuring sample quality with kernels. In *International Conference on Machine Learning*, pages 1292–1301. PMLR.

Gramacy, R. B. (2020). *Surrogates: Gaussian process modeling, design, and optimization for the applied sciences*. CRC press.

Hughes, J., Haran, M., and Caragea, P. C. (2011). Autologistic models for binary data on a lattice. *Environmetrics*, 22(7):857–871.

Hummel, R. M., Hunter, D. R., and Handcock, M. S. (2012). Improving simulation-based algorithms for fitting ergms. *Journal of Computational and Graphical Statistics*, 21(4):920–939.

Hunter, D. R. and Handcock, M. S. (2006a). Inference in curved exponential family models for networks. *Journal of computational and graphical statistics*, 15(3):565–583.

Hunter, D. R. and Handcock, M. S. (2006b). Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics*, 15(3):565–583.

Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M., and Morris, M. (2008). ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of statistical software*, 24(3):nihpa54860.

Ising, E. (1925). Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik A Hadrons and Nuclei*, 31(1):253–258.

Jones, G. L., Haran, M., Caffo, B. S., and Neath, R. (2006). Fixed-width output analysis for Markov chain Monte Carlo. *Journal of the American Statistical Association*, 101:1537–1547.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine learning*, 37:183–233.

Kang, B., Hughes, J., and Haran, M. (2023). Measuring sample quality in algorithms for intractable normalizing function problems. *arXiv preprint arXiv:2109.05121*.

Lauritzen, S. L. and Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *The Annals of Statistics*, 17:31–57.

Lee, J. D. and Hastie, T. J. (2015). Learning the structure of mixed graphical models. *Journal of Computational and Graphical Statistics*, 24:230–253.

Lenz, W. (1920). Beitrag zum verständnis der magnetischen erscheinungen in festen körpern. *Physikalische Zeitschrift*, 21:613–615.

Liang, F. (2010). A double Metropolis-Hastings sampler for spatial models with intractable normalizing constants. *Journal of Statistical Computation and Simulation*, 80(9):1007–1022.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemp. Math.*, 80:221–39.

Lu, T. and Boutilier, C. (2014). Effective sampling and learning for mallows models with pairwise-preference data. *J. Mach. Learn. Res.*, 15(1):3783–3829.

Lyne, A.-M., Girolami, M., Atchadé, Y., Strathmann, H., and Simpson, D. (2015). On Russian Roulette Estimates for Bayesian Inference with Doubly-Intractable Likelihoods. *Statistical Science*, 30(4):443 – 467.

Mallows, C. L. (1957). Non-null ranking models. i. *Biometrika*, 44(1/2):114–130.

Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003). Markov chain monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328.

Matsubara, T., Knoblauch, J., Briol, F.-X., and Oates, C. J. (2022). Robust generalised bayesian inference for intractable likelihoods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):997–1022.

Metropolis, N., Rosenbluth, M., Rosenbluth, A., Teller, A., and Teller, E. (1953). Equations of state calculations by fast computing machines. *Chemical Physics*, 21:1087–1092.

Møller, J., Pettitt, A., Berthelsen, K., and Reeves, R. (2006). An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2):451–458.

Murray, I. (2007). *Advances in Markov Chain Monte Carlo Methods*. PhD thesis, University of London.

Okabayashi, S., Johnson, L., and Geyer, C. J. (2011). Extending pseudo-likelihood for potts models. *Statistica Sinica*, pages 331–347.

Park, J. and Haran, M. (2018). Bayesian inference in the presence of intractable normalizing functions. *Journal of the American Statistical Association*, 113(523):1372–1390.

Park, J. and Haran, M. (2020). A function emulation approach for doubly intractable distributions. *Journal of Computational and Graphical Statistics*, 29(1):66–77.

Park, J., Jin, I. H., and Schweinberger, M. (2022). Bayesian model selection for high-dimensional ising models, with applications to educational data. *Computational Statistics & Data Analysis*, 165:107325.

Potts, R. B. (1952). Some generalized order-disorder transformations. In *Mathematical proceedings of the cambridge philosophical society*, pages 106–109.

Prangle, D., Blum, M. G., Popovic, G., and Sisson, S. (2014). Diagnostic tools for approximate bayesian computation using the coverage property. *Australian & New Zealand Journal of Statistics*, 56(4):309–329.

Price, L. F., Drovandi, C. C., Lee, A., and Nott, D. J. (2018). Bayesian synthetic likelihood. *Journal of Computational and Graphical Statistics*, 27(1):1–11.

Propp, J. G. and Wilson, D. B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms*, 9(1-2):223–252.

Rendsburg, L., Kristiadi, A., Hennig, P., and Von Luxburg, U. (2022). Discovering inductive bias with gibbs priors: A diagnostic tool for approximate bayesian inference. In *International Conference on Artificial Intelligence and Statistics*, pages 1503–1526. PMLR.

Ribatet, M., Cooley, D., and Davison, A. (2012). Bayesian inference from composite likelihoods, with an application to spatial extremes. *Stat. Sinica*, 22:813–846.

Robins, G., Pattison, P., Kalish, Y., and Lusher, D. (2007a). An introduction to exponential random graph (p*) models for social networks. *Social Networks*, 29(2):173–191.

Robins, G., Pattison, P., Kalish, Y., and Lusher, D. (2007b). An introduction to exponential random graph (p*) models for social networks. *Social networks*, 29(2):173–191.

Roy, V. (2020). Convergence diagnostics for markov chain monte carlo. *Annual Review of Statistics and Its Application*, 7:387–412.

Sainsbury-Dale, M., Richards, J., Zammit-Mangion, A., and Huser, R. (2023). Neural bayes estimators for irregular spatial data using graph neural networks. *arXiv preprint arXiv:2310.02600*.

Santner, T. J., Williams, B. J., Notz, W. I., and Williams, B. J. (2003). *The design and analysis of computer experiments*, volume 1. Springer.

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., and Boatwright, P. (2005). A useful distribution for fitting discrete data: Revival of the Conway–Maxwell–Poisson distribution. *Journal of the Royal Statistical Society: Series C*, 54:127–142.

Snijders, T. A. et al. (2002). Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure*, 3(2):1–40.

Swendsen, R. H. and Wang, J.-S. (1987). Nonuniversal critical dynamics in Monte Carlo simulations. *Physical review letters*, 58(2):86–88.

Tan, L. S. and Friel, N. (2020). Bayesian variational inference for exponential random graph models. *Journal of Computational and Graphical Statistics*, 29(4):910–928.

Torrie, G. M. and Valleau, J. P. (1977). Nonphysical sampling distributions in Monte Carlo free-energy estimation - umbrella sampling. *Journal of Computational Physics*, 23:187–199.

Tran, M.-N., Nott, D. J., and Kohn, R. (2017). Variational bayes with intractable likelihood. *Journal of Computational and Graphical Statistics*, 26(4):873–882.

Van Borkulo, C. D., Borsboom, D., Epskamp, S., Blanken, T. F., Boschloo, L., Schoevers, R. A., and Waldorp, L. J. (2014). A new method for constructing networks from binary data. *Scientific reports*, 4(1):5918.

Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, pages 5–42.

Vats, D., Flegal, J. M., and Jones, G. L. (2019). Multivariate output analysis for Markov chain Monte Carlo. *Biometrika*, 106:321–337.

Walchessen, J., Lenzi, A., and Kuusela, M. (2023). Neural likelihood surfaces for spatial processes with computationally intensive or intractable likelihoods. *arXiv preprint arXiv:2305.04634*.