SPECIAL ISSUE: BAYESIAN INFERENCE AND STOCHASTIC COMPUTATION

Multicanonical MCMC for sampling rare events: an illustrative review

Yukito Iba · Nen Saito · Akimasa Kitajima

Abstract Multicanonical MCMC (Multicanonical Markov Chain Monte Carlo; Multicanonical Monte Carlo) is discussed as a method of *rare event sampling*. Starting from a review of the generic framework of importance sampling, multicanonical MCMC is introduced, followed by applications in random matrices, random graphs, and chaotic dynamical systems. Replica exchange MCMC (also known as parallel tempering or Metropolis-coupled MCMC) is also explained as an alternative to multicanonical MCMC. In the last section, multicanonical MCMC is applied to data surrogation; a successful implementation in surrogating time series is shown. In the appendix, calculation of averages and normalizing constant in an exponential family, phase coexistence, simulated tempering, parallelization, and multivariate extensions are discussed.

Y. Iba (✉)
The Institute of Statistical Mathematics and SOKENDAI, 10-3 Midori-cho,
Tachikawa, Tokyo 190-8562, Japan
e-mail: iba@ism.ac.jp

N. Saito
Research Center for Complex Systems Biology, The University of Tokyo,
3-8-1 Komaba, Meguro-ku, Tokyo 153-8902, Japan
e-mail: saito@complex.c.u-tokyo.ac.jp

A. Kitajima
Digital Information Services Division, Digital Information Department, National Diet Library,
1-10-1 Nagata-cho, Chiyoda-ku, Tokyo 100-8924, Japan
e-mail: a-kitaji@ndl.go.jp

1 Introduction

Multicanonical MCMC (Multicanonical Markov Chain Monte Carlo; Multicanonical Monte Carlo) was introduced in statistical physics in the early 1990s (Berg and Neuhaus 1991, 1992; Berg and Celik 1992); it can be viewed as a variant of umbrella sampling, whose origin can be traced back to the 1970s (Torrie and Valleau 1974).[1] The Wang–Landau algorithm developed in Wang and Landau (2001a,b) provides an effective realization of a similar idea and many current studies use this implementation. It, however, relies on step-by-step realization of "multicanonical weight" defined in Sect. 2.2.1 in this paper, which is an essential part of the original multicanonical algorithms by Berg and Neuhaus (1991, 1992) and Berg and Celik (1992). In this paper, we will use the term "multicanonical MCMC" for any method that uses the multicanonical weight.

In these studies, multicanonical MCMC is applied to simultaneous sampling from Gibbs distributions of different temperatures; in terms of statistics, it corresponds to sampling from an exponential family. From this viewpoint, a major advantage of multicanonical MCMC is fast mixing in multimodal problems. It often realizes an order of magnitude improvement in the speed of convergence over conventional MCMC. Some examples in statistical physics are provided by the references in Sect. 3.3.1; see also review articles (Berg 2000; Janke 1998; Landau et al. 2004; Higo et al. 2012; Iba 2001).

Recent studies, however, provide another look at this algorithm. Multicanonical MCMC enables an efficient way of sampling *rare events* under a given distribution. Suppose that rare events of x in a high-dimensional sample space are characterized by the value of statistics $\xi(x)$. Then, in some examples, rare events even with probabilities $P(\xi_0 \leq \xi(x)) \approx 10^{-100}$ are sampled within a reasonable computational time.[2] Further, these probabilities are precisely estimated without additional computation.

This novel viewpoint opens the door to a broad application field of multicanonical MCMC, while providing a more intuitive and easy understanding of the same algorithm. Even though some surveys have already introduced multicanonical MCMC as a method of rare event sampling (see Driscoll and Maki 2007; Bononi et al. 2009; Wolfsheimer et al. 2011),[3] it will be useful to conduct another survey with a broad perspective and novel applications. An aim of this paper is to provide such an introduction, including recent results by the authors.

Another aim of this paper is to apply multicanonical MCMC to *exact tests* in statistics. Multicanonical MCMC is useful for sampling from highly constrained systems, and this will be explained in this paper in connection with rare event sampling. Hence, it can be naturally applied to MCMC exact tests (Besag and Clifford 1989; Diaconis and Sturmfels 1998), where constraints among variables make it difficult to construct

---

[1] A rarely cited paper (Mezei 1987), already proposed an adaptive version of umbrella sampling, which uses a general "reaction coordinate" instead of total energy. Baumann (1987) is also referred to as a prototype of multicanonical MCMC.

[2] The constant ξ_0 controls rareness; see Sect. 2.1.1 for details.

[3] See also Birge et al. (2012); this paper introduced a related algorithm, split sampling, as a method of rare event sampling.

Markov chains for efficient sampling from null distributions. As an example, we will discuss surrogation of nonlinear time series; yet the proposed method can be generalized to the other MCMC exact tests such as sampling from tables with fixed marginals. The results discussed in Sect. 4.2 are published here for the first time in English.

The rest of this paper is organized as follows: In Sect. 2, multicanonical MCMC is surveyed as a rare event sampling technique. Starting from general issues on rare event sampling, the use of an exponential family with replica exchange MCMC is discussed as an alternative to multicanonical MCMC. Then, the key idea of multicanonical MCMC is introduced, and a concise description of the Wang–Landau algorithm is provided. Section 3 provides examples of multicanonical rare event sampling, focusing on the authors' recent studies on random matrices, random graphs, and dynamical systems. Section 4 begins with a multicanonical approach to highly constrained systems. Then, exact statistical tests and data surrogation are introduced as application fields. A numerical experiment is discussed, where surrogates of time series that maintain the values of correlation functions are generated. An appendix deals with several other issues on multicanonical MCMC, that is, calculating averages and normalizing constant in an exponential family, "phase coexistence," simulated tempering, parallel computation, and multivariate extensions.[4]

This paper is mainly intended to describe the possibility of multicanonical MCMC in various fields. Therefore, we focus on basic concepts and examples, omitting details such as mathematical proofs of convergence and practical issues of implementation. We assume the readers are familiar with standard algorithms of MCMC, but do not have specific knowledge on rare event sampling nor multicanonical MCMC. Thus, we begin with basics of rare event sampling and proceed to multicanonical MCMC, skipping details of the implementation of MCMC. In fact, we can combine almost any kind of MCMC algorithm to the idea of multicanonical MCMC. It is, however, essential to pay attention to the behavior of the sample path in the case of multimodal distributions, which we will discuss in detail in the paper.

Readers who are not familiar with MCMC will find necessary backgrounds, for example, in Gilks et al. (1996), Robert and Casella (2004), Brooks et al. (2011). See also books on MCMC by physicists, such as Newman and Barkema (1999), Frenkel and Smit (2002), Berg (2004), Landau and Binder (2009), Binder and Heermann (2012).

2 Multicanonical sampling of rare events

2.1 Rare event sampling

We first consider general issues in rare event sampling, namely, importance sampling and the use of exponential families; replica exchange MCMC is also explained. For further details on general frameworks and other approaches, see Bucklew (2004), Rubinstein and Kroese (2008), Rubino and Tuffin (2009).

---

[4] In this paper, double quotes (" … ") are used for marking technical terms in physics, non-technical expressions, and terms defined in this paper, whereas *italics* are utilized for emphasizing other terms.

2.1.1 Importance sampling

Let us assume that the value of a variable X is randomly sampled from the probability distribution P; throughout this paper, we assume that P is precisely known. Hereafter, for simplicity, we explain cases where variable X takes discrete values; however, generalization to a continuous X is not difficult.

When we specify target statistics ξ, "rare events" of X with a rare value $\xi(X)$ of ξ are defined as a set $\mathscr{A} = \{x \mid \xi_0 \leq \xi(x)\}$, where the probability $P(\xi_0 \leq \xi(X))$ takes a small value;[5] the constant ξ_0 controls the rareness of the events.

Our problem is to generate samples of X that satisfy $\xi_0 \leq \xi(X)$ and estimate their probability $P(\xi_0 \leq \xi(X))$. Given current hardware, we can still complete the task by a direct computation, even when the probability $P(\xi_0 \leq \xi(X))$ takes considerably smaller values such as 10^{-4} or 10^{-6}. However, when the probability of rare events is much smaller, say, 10^{-12} or even 10^{-100}, it is virtually impossible to deal with the problem by naive random sampling from the original distribution P.

A standard solution to this problem is the use of *importance sampling* techniques, that is, we generate samples of X from another distribution Q, which has a larger probability in the set $\mathscr{A}$. Hereafter, we assume that $Q(X = x) \neq 0$ for the value of x satisfying $P(X = x) \neq 0$. Using samples $X^{(i)}, i = 1, \ldots, M$ from Q, the probability under the original distribution P is estimated as

$$P(\xi_0 \leq \xi(X)) \simeq \frac{1}{M} \sum_{i=1}^{M} \left[\frac{P\left(X^{(i)}\right)}{Q\left(X^{(i)}\right)} I(\xi_0 \leq \xi(X^{(i)})) \right], \qquad (1)$$

where I is defined by

$$I(\xi_0 \leq \xi(X^{(i)})) = \begin{cases} 1, & \xi_0 \leq \xi(X^{(i)}) \\ 0, & \xi_0 > \xi(X^{(i)}) \end{cases}. \qquad (2)$$

By the law of large numbers, (1) becomes an equality as $M \rightarrow \infty$. An average of arbitrary statistics $A(X)$ in the set $\mathscr{A}$ with weights proportional to P is calculated as

$$\mathbb{E}[A(X) \mid \xi_0 \leq \xi(X)] \simeq \frac{\frac{1}{M} \sum_{i=1}^{M} \left[A(X^{(i)}) \frac{P\left(X^{(i)}\right)}{Q\left(X^{(i)}\right)} I(\xi_0 \leq \xi(X^{(i)})) \right]}{P(\xi_0 \leq \xi(X))}, \qquad (3)$$

which also becomes an equality as $M \rightarrow \infty$.

A critical issue in importance sampling is the choice of the distribution Q. Prior to the introduction of MCMC, there was a severe limitation on the choice of Q; this was because efficient generation of samples is possible only for a simple Q. In contrast,

---

[5] $P(\xi_0 \geq \xi(X))$ is reduced to the case $P(\xi_0 \leq \xi(X))$ by considering $-\xi$, and hence, it is not discussed separately. The probability $P(\xi_0 - \delta \leq \xi(X) \leq \xi_0 + \delta)$ is also considered. In this case, we should maintain an adequate value of δ and/or consider the relative probabilities using the same value of δ for a proper definition of "rareness."

MCMC provides much freedom in the selection of Q. On the other hand, samples from Q generated by MCMC are usually correlated, and such correlation can severely affect the convergence of the averages. Thus, we should pay attention to the mixing of MCMC in the choice of Q.

2.1.2 Exponential family and replica exchange MCMC

A strategy,[6] which we will discuss in this paper, is to choose Q in the form

$$Q(x) = \frac{G(\xi(x)) P(x)}{\sum_x G(\xi(x)) P(x)},$$

where $G(\xi)$ is an appropriate univariate function and $\sum_x$ indicates the sum over the domain of x. Multicanonical MCMC belongs to this class. Here, we will discuss a different choice $G(\xi) = \exp(\beta\xi)$ as an alternative to multicanonical approach; this leads to

$$Q_\beta(x) = \frac{\exp(\beta\xi(x)) P(x)}{\sum_x \exp(\beta\xi(x)) P(x)}. \tag{4}$$

Q_β is interpreted as an exponential family with sufficient statistics ξ and a canonical parameter β; it is also regarded as a Gibbs distribution with energy $-\xi$ and inverse temperature β, when the base measure P is uniform.

Assuming Q_β defined by (4), we can sample regions with larger values of ξ by increasing the value of β. Thus, in principle, MCMC sampling from Q_β with a large value of β can efficiently generate rare events defined by $\xi_0 \leq \xi(x)$. When β increases, however, the set $\mathscr{A}$ of x defined by $\xi_0 \leq \xi(x)$ often almost disconnects, that is, it consists of multiple "islands" of x separated by regions with tiny values of Q_β. Such a multimodal property of Q_β obviously leads to slow convergence of MCMC.

In many examples, this difficulty is reduced using *replica exchange MCMC*, which is also known as *parallel tempering* or *Metropolis-coupled MCMC* (Kimura and Taki 1991; Geyer 1991; Hukushima and Nemoto 1996; Iba 2001). In this algorithm, Markov chains with different values of β run in parallel; here, we assume K chains with $(\beta_1, \beta_2, \ldots, \beta_K)$. Selecting a pair i and j of chains in a regular interval of steps, the current values of the states X_i^* and X_j^* of chains are swapped with probability P_{swap} defined as

$$P_{\text{swap}} = \max\left\{1, \frac{Q_{\beta_i}(X_j^*)Q_{\beta_j}(X_i^*)}{Q_{\beta_i}(X_i^*)Q_{\beta_j}(X_j^*)}\right\} = \max\left\{1, \exp((\beta_i - \beta_j)(\xi(X_j^*) - \xi(X_i^*)))\right\}.$$

Note that the combined probability $\prod_{k=1}^{K} Q_{\beta_k}(x_k)$ is a stationary distribution of the Markov chain defined by a combination of the original MCMC and the exchange procedure defined above. This property ensures that replica exchange MCMC realizes a proper sampling procedure at each value of β.

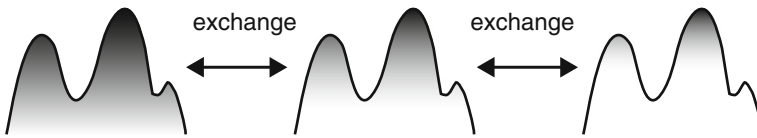[6] A different approach to combine importance sampling with MCMC is found in Botev et al. (2013).

Fig. 1 Exchange of states between the distribution P_β with different values of β. The *vertical axis* corresponds to the value of ξ, whereas the *horizontal axis* for each sub-chart schematically represents a high-dimensional space of X. Here, the values of β are assumed to increase from *left* to *right*; *shading* represents the changes in high-probability regions

Exchange of states between chains is introduced for facilitating mixing at large values of β. Owing to these exchanges, states generated at smaller values of β successively "propagate" to chains with larger β (Fig. 1). This mechanism is similar to that in the *simulated annealing* algorithm (Kirkpatrick et al. 1983) for optimization. An essential difference is that replica exchange MCMC utilizes a time-homogeneous Markov chain designed for sampling from each of the given distributions. In contrast, simulated annealing utilizes a time-inhomogeneous chain; at least in principle, it is not suitable for sampling.

The combination of replica exchange MCMC and Q_β given by (4) provides a powerful tool for rare event sampling, which is easy to implement on parallel hardware. However, the estimation of the probability of rare events $P(\xi_0 \leq \xi(X))$ under the original distribution P requires some additional consideration. Namely, samples at a single value of β are usually not enough for computing relative values of $P(\xi_0 \leq \xi(X))$ for all values of ξ_0. Hence, the normalizing constant $Z_\beta = \sum_x \exp(\beta \xi(x)) P(x)$ should be estimated for combining the results at different β.

These difficulties are well treated using samples at multiple values of β, which are most naturally obtained as outputs of replica exchange MCMC. Here, however, we omit details; essentially, the same problem in statistical physics is known as the estimation of "density of states." See, for example, an intuitive method used in Hartmann (2002) and a rather sophisticated approach, the *multiple histogram method*, explained in Newman and Barkema (1999).

2.2 Multicanonical MCMC

Here, we explain multicanonical MCMC, which is the main subject of this paper. First, we define a "multicanonical weight" and discuss the behavior of MCMC with this weight. Then, we introduce adaptive MCMC schemes for realizing the multicanonical weight. In this section, we explain the algorithm for cases where both X and ξ take discrete values. A simple way to treat a continuous ξ is introduction of a binning function $\tilde{\xi}$ defined in Sect. 2.2.3; for more sophisticated methods, see references in Sect. 2.2.4.

2.2.1 Multicanonical weight

As already explained, when Q_β given in (4) is used, some additional computation is required for estimating the probabilities of rare events. The situation can be worse in

some examples; a region of ξ is virtually not sampled for any choice of the canonical parameter β. This may not be typical but possible; see "First-order transition and Phase coexistence" in the appendix for further details.

In contrast, multicanonical MCMC has an advantage in that it provides probabilities such as $P(\xi_0 \le \xi(X))$ directly as outputs of the MCMC simulation and no additional computation is required. Further, the problem of the missing region of ξ can be avoided, at least in some examples. In addition to these nice properties, multicanonical MCMC enables fast convergence in multimodal problems, similar to replica exchange MCMC.

To realize these properties, multicanonical MCMC utilizes $G(\xi)$ defined in the following way. First, we assume that an approximation $\tilde{P}(\xi)$ of $P(\xi)$ is given, in which the marginal probability of ξ is defined as $P(\xi') = \sum_{\xi(x)=\xi'} P(X = x)$, where $\sum_{\xi(x)=\xi'}$ indicates the sum over x that satisfies $\xi(x) = \xi'$. Then, $G(\xi)$ is given by the inverse $1/\tilde{P}(\xi)$ of $\tilde{P}(\xi)$; more precisely, we define

$$G(\xi(x)) = \begin{cases} c\,\tilde{P}(\xi(x))^{-1} & \text{if } \xi \in [\xi_{\min}, \xi_{\max}] \\ 0 & \text{else} \end{cases}, \qquad (5)$$

where c is an arbitrary constant and $[\xi_{\min}, \xi_{\max}]$ is an interval ξ of interest. Note that the values of ξ that give $P(\xi) = 0$ should be excluded from the set $[\xi_{\min}, \xi_{\max}]$. Hereafter, we refer to $G(\xi)$ defined in (5) as a "multicanonical weight." The corresponding $Q(x)$ is defined as $Q(x) = G(\xi(x))P(x)/C$, where $C = \sum_x G(\xi(x))P(x)$ is the normalizing constant; hereafter, the constant c is absorbed in C and omitted from the expressions.[7]

At first sight, the choice of $G(\xi(x))$ shown in (5) does not make sense in practice since the distribution $P(\xi)$ is essentially the one that we want to calculate by the algorithm. In some cases, we guess a form of $P(\xi)$ and use it to approximate the multicanonical weight (Körner et al. 2006; Monthus and Garel 2006), but this is rather exceptional. Nevertheless, we leave this question for a while and discuss the properties of a multicanonical weight.

Let us tentatively assume an ideal case that $\tilde{P}(\xi)$, which appeared in the multicanonical weight defined in (5), is exactly equal to $P(\xi)$. Then, the marginal distribution $Q(\xi)$ defined by $Q(x) = G(\xi(x))P(x)/C$ is uniform in the interval $[\xi_{\min}, \xi_{\max}]$, excluding the values of ξ that give $P(\xi) = 0$. This is because the multicanonical weight is designed for canceling the factor $P(\xi)$, which is confirmed via direct calculation as

$$Q(\xi') = \frac{1}{C} \sum_x G(\xi(x)) P(x) I(\xi(x) = \xi')$$

$$= \frac{1}{C} G(\xi') \sum_x P(x)\, I(\xi(x) = \xi') = \frac{1}{C}\, P(\xi')^{-1}\, P(\xi') = \frac{1}{C},$$

where $\sum_x$ is the sum over the all possible values of x and $I(\zeta = \zeta')$ is defined as

---

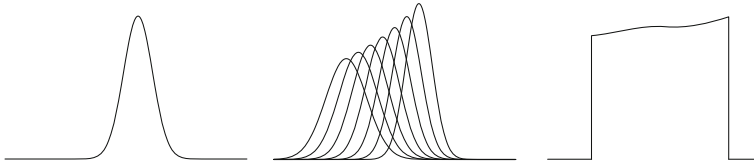[7] This $Q(x)$ will also be referred to as a "multicanonical weight" on the space of x.

Fig. 2 "Flat" marginal $Q(\xi)$ realized by (5) is compared to the marginal $Q_\beta(\xi)$ of the exponential family (4). *Left* $Q_\beta(\xi)$ with a fixed value of β. *Center* a series of $Q_\beta(\xi)$ with $(\beta_1, \beta_2, \ldots, \beta_7)$ are printed over one another. *Right* $Q(\xi)$ realized by a multicanonical weight defined in (5); a case where $\tilde{P}(\xi) \simeq P(\xi)$ is shown, while it becomes completely flat when $\tilde{P}(\xi) = P(\xi)$. In some cases, behaviors very different from these are observed; see Fig. 13

$$I(\zeta = \zeta') = \begin{cases} 1 & \zeta = \zeta' \\ 0 & \zeta \neq \zeta' \end{cases}.$$

This "flat" distribution $Q(\xi)$ of ξ realized by a multicanonical weight defined in (5) is illustrated in the rightmost panel of Fig. 2. For comparison, $Q(\xi)$ given by an exponential family (4) is shown in the other two panels of Fig. 2.

2.2.2 MCMC sampling with a multicanonical weight

So far, we discuss a rather obvious conclusion, but it is more interesting to consider an MCMC simulation that samples the corresponding distribution $Q(x) = G(\xi(x))P(x)/C$. To uniformly cover the region $[\xi_{\min}, \xi_{\max}]$, the sample path moves randomly in the region. In other words, the multicanonical weight realizes a random walk on the axis of the target statistics ξ; this walk has a memory because the value of X does not determined uniquely by $\xi(X)$.

This behavior enables us to obtain the desired properties using a single chain, as shown in Fig. 3. First, efficient sampling of a tail region with a large value of ξ is possible if we choose a sufficiently large $\xi_{\max}$. On the other hand, fast mixing of MCMC is attained if we choose $\xi_{\min}$ such that the set defined by $\xi_{\min} \leq \xi(x)$ is tightly connected and a sample path can easily move around in it.[8] Therefore, MCMC sampling with a multicanonical weight shares an "annealing" property with replica exchange MCMC.

Finally, we confirm how probabilities of rare events are computed under the original distribution P. We assume that $X^{(i)}, i = 1, \ldots, M$ are samples from Q defined by G of the equation (5). Then, the following expression is derived from (1):

$$P(\xi_0 \leq \xi(X) \leq \xi_{\max}) \simeq C \times \frac{1}{M} \sum_{i=1}^{M} \left[\tilde{P}(\xi(X^{(i)})) \, I(\xi_0 \leq \xi(X^{(i)})) \right],$$

---

[8] Such a region corresponds to a "high-temperature" region in statistical physics, whereas the tail region with rare events corresponds to a "low-temperature" region.
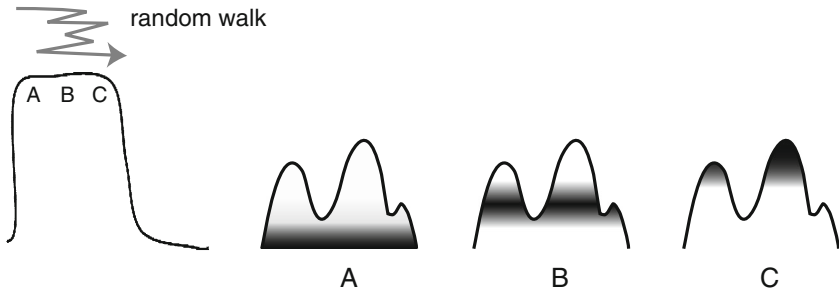
Fig. 3 Random walk of $\xi(X)$ realized by a multicanonical weight. *Left* $P(\xi)$ and a sample path of the random walk projected on the axis of the target statistics ξ. Note that the statistics ξ is a function of state X, and MCMC updates of X naturally cause such a walk of ξ; no separate procedure is required for changing ξ. *Right A*, *B*, and *C* show the distributions of $(X, \xi(X))$, each of which corresponds to the regions A, B, and C in the *left panel*. The *vertical axis* corresponds to the value of ξ, whereas the *horizontal axis* for each sub-chart schematically represents the high-dimensional space where X takes its value

where I is defined by (2). Because the values of ξ are limited in $\xi_{\min} \leq \xi \leq \xi_{\max}$ by our definition of the multicanonical weight,

$$P(\xi_{\min} \leq \xi(X) \leq \xi_{\max}) \simeq C \times \frac{1}{M} \sum_{i=1}^{M} \tilde{P}(\xi(X^{(i)})),$$

also holds. Hence, we arrive at

$$\frac{P(\xi_0 \leq \xi(X) \leq \xi_{\max})}{P(\xi_{\min} \leq \xi(X) \leq \xi_{\max})} \simeq \frac{\sum_{i=1}^{M} \left[\tilde{P}(\xi(X^{(i)})) \, I(\xi_0 \leq \xi(X^{(i)})) \right]}{\sum_{i=1}^{M} \tilde{P}(\xi(X^{(i)}))}. \tag{6}$$

The value of the denominator $P(\xi_{\min} \leq \xi(X) \leq \xi_{\max})$ becomes almost unity when the interval $[\xi_{\min}, \xi_{\max}]$ contains most of the probability mass; otherwise, in some cases, we are mainly interested in relative probabilities. The expectation of arbitrary statistics A in the tail region $\xi_0 \leq \xi(x) \leq \xi_{\max}$ is also derived from (3) in a similar manner as

$$\mathbb{E}[A(X) \,|\, \xi_0 \leq \xi(X) \leq \xi_{\max}] \simeq \frac{\sum_{i=1}^{M} \left[A(X^{(i)}) \, \tilde{P}(\xi(X^{(i)})) \, I(\xi_0 \leq \xi(X^{(i)})) \right]}{\sum_{i=1}^{M} \left[\tilde{P}(\xi(X^{(i)})) \, I(\xi_0 \leq \xi(X^{(i)})) \right]}. \tag{7}$$

2.2.3 Entropic sampling

Now, we return to the following problem. How to estimate the multicanonical weight $G(\xi)$ in (5) without prior knowledge? The key idea is to use adaptive Monte Carlo; "preliminary runs" of MCMC are repeated to tune the weight $G(\xi)$ until the marginal distribution $Q(\xi)$ becomes almost flat in the interval $[\xi_{\min}, \xi_{\max}]$. After tuning the weight, a "production run" is performed, where $G(\xi)$ is fixed; this run realizes MCMC sampling with a multicanonical weight. Note that virtually any type of MCMC can be used for sampling in both of these stages.

An important point is that $G(\xi)$ is a univariate function of a scalar variable ξ, while $Q(x) = G(\xi(x))P(x)/C$ is defined on a high-dimensional space of x; thus, tuning $G(\xi)$ is much easier than performing a direct adaptation of $Q(x)$ itself.

To illustrate the principle, we describe a simple method, sometimes known as *entropic sampling* (Lee 1993). First, we consider the histogram H of the values of ξ. It is convenient to introduce a discretized or binned version $\tilde{\xi}(x)$ of $\xi(x)$, which takes an integer value $\tilde{\xi} \in \{1, 2, \ldots, N_b\}$.[9]

Then, the histogram of the values of $\tilde{\xi}$ in the kth iteration of the preliminary runs is represented by $\{H^{(k)}(\tilde{\xi})\}, \tilde{\xi} = 1, \ldots, N_b$. We define $\bar{H}$ as expected counts in each bin of a flat histogram, which is the target of our adaptation.[10] Further, the weight in the kth iteration is represented by $\{G^{(k)}(\tilde{\xi})\}, \tilde{\xi} = 1, \ldots, N_b$. Now that the adaptation in the kth step is expressed as a recursion

$$G^{(k+1)}(\tilde{\xi}) = G^{(k)}(\tilde{\xi}) \times \frac{\bar{H} + \epsilon}{H^{(k)}(\tilde{\xi}) + \epsilon}. \tag{8}$$

Here, a constant ϵ is required for eliminating the divergence at $H^{(k)} = 0$, which is set to a small value, say, unity. The idea behind this recursion is simple—increase the weight if the counts are smaller than $\bar{H}$ and decrease the weight if the counts are larger than $\bar{H}$.

The tuning stage of the algorithm is formally described as follows. Here, we use $LG(\tilde{\xi}) = \log G(\tilde{\xi})$ instead of $G(\tilde{\xi})$.

1. Initialize LG and set parameters.
   – Set $LG(i) = 0$ for $i = 1, \ldots, N_b$.
   – Set the maximum number of iterations $K_{\max}$.
   – Set the number of MCMC steps $M_{\max}$ within each iteration.
   – Set the number of MCMC steps M_s between histogram updates.
   – Set a regularization parameter ϵ (e.g., $\epsilon = 1$).
   – Set $\bar{H} = (M_{\max}/M_s)/N_b$.
   – Set the counter of iterations K to 0.
2. Initialize H and X.
   – Set $H(i) = 0$ for $i = 1, \ldots, N_b$.
   – Initialize the state X.
   – Set the counter of MCMC trials M to 0.
3. Run MCMC.
   – Run M_s steps of MCMC with the weight $P(x) \exp[LG(\tilde{\xi}(x))]$.
4. Update the histogram H.
   – $H(\tilde{\xi}(x^*)) = H(\tilde{\xi}(x^*)) + 1$, where $X = x^*$ is the current state. ♠
   – $M = M + M_s$.

---

[9] Giving a partition $\mathscr{F}_i, i \in \{1, 2, \ldots, N_b\}$ of the interval $[\xi_{\min}, \xi_{\max}]$, it is defined by $\tilde{\xi}(x) = j \Leftrightarrow \xi(x) \in \mathscr{F}_j$. If $\xi_{\max} < \xi$ or $\xi < \xi_{\min}$, it is often convenient to define $\tilde{\xi} = N_b$ or $\tilde{\xi} = 1$, respectively. Another way is to reject the value of x that satisfies $\xi_{\max} < \xi(x)$ or $\xi(x) < \xi_{\min}$ within the Metropolis–Hastings algorithm (See also a remark in Schulz et al. 2003).

[10] The constant factor $\bar{H}$ is not essential in the following argument when we consider relative weights, but we retain it because it clarifies the meaning of formulae.

– If $M < M_{\max}$, go to Step 3.

5. Check whether H is "sufficiently flat."

    – If so, **end**.

    – If not and $K < K_{\max}$, modify LG.

        – $LG(i) = LG(i) + \log[(\epsilon + \bar{H})/(\epsilon + H(i))]$ for $i = 1, \ldots, N_b$. ♣

        – $K = K + 1$.

        – Go to Step 2.

    – If not and $K_{\max} \le K$, the algorithm **fails**.

Note that the update formula (8) is included as a step marked with ♣, while the histogram is incremented in the step marked with ♠.

After completing the above procedure, the production run is performed. If the above algorithm fails to converge, we can increase the numbers $M_{\max}$ and/or $K_{\max}$. Another choice is to reduce our requirement and decrease the value of ξ_0, which determines the rareness of the obtained events.

The construction of the histogram can be replaced by other density estimation techniques. In the original studies (Berg and Neuhaus 1991, 1992; Berg and Celik 1992), $\log G(\xi)$ is represented by a piecewise linear curve, instead of a piecewise constant curve used in entropic sampling; parametric curve fitting is also utilized. Another useful method is kernel density estimation, which is particularly convenient in continuous and/or multivariate ξ cases; it is also used with the Wang–Landau algorithm explained later, as seen in Zhou et al. (2006). Finally, we mention methods based on the *broad histogram equation*. In these methods, the number of transitions between states is used for optimizing the weight, instead of the number of visits to a state. Such an idea has a somewhat different origin (de Oliveira et al. 1998), but it can be interpreted as a way to realize a multicanonical weight; see Wang and Swendsen (2002).

2.2.4 Wang–Landau algorithm

Entropic sampling is already sufficient for realizing a multicanonical weight in many problems. In current studies, however, the Wang–Landau algorithm (Wang and Landau 2001a,b) is often utilized, which provides a more efficient strategy to construct a multicanonical weight.

An essential feature of the Wang–Landau algorithm is the use of a time-inhomogeneous chain in the preliminary runs; that is to say, we change the weights after each trial of MCMC moves instead of changing them only at the end of each iteration consisting of a fixed number of MCMC steps. This may lead to an "incorrect" MCMC sampling in the preliminary runs, but it causes no problem if we fix the weights in the final production run, where we compute the required probabilities and expectations.

In the actual implementation, whenever a state x with $\tilde{\xi}^* = \tilde{\xi}(x)$ appears, we multiply the value of weight $G(\tilde{\xi}^*)$ by a constant factor $0 < C < 1$;[11] it reduces the weights of the already visited values of $\tilde{\xi}$, whereas it effectively increases the relative weights of the other values of $\tilde{\xi}$. In parallel, we construct the histogram H of $\tilde{\xi}$ that

---

[11] Do not confuse this C with the normalization constant C in the previous sections.

appeared in MCMC sampling. After some steps of MCMC, we reach a "sufficiently flat" histogram;[12] then, a step of iterative tuning of the weights is completed.

When we rerun MCMC where the weight is fixed to the values obtained by this procedure, the run usually does not provide a sufficiently flat histogram of $\tilde{\tilde{\xi}}$. Then, an iterative method is introduced, that is, we increase the value of the constant C and repeat the procedure in the preceding paragraph. A heuristics proposed in the original papers (Wang and Landau 2001a,b) is to change C to $\sqrt{C}$. After each iteration step, the histogram H is cleared, whereas the values of G are retained.

Again, we stress that any type of MCMC can be used for sampling at each of these stages; we use the familiar Metropolis–Hasting algorithms in the examples considered in this paper. As shown in later sections, however, the choice of moves in the Metropolis–Hasting algorithms significantly affects the efficiency of the entire algorithm.

The tuning of the weight by the Wang–Landau algorithm is summarized as shown below. Again, we use $LG(\tilde{\xi}) = \log G(\tilde{\xi})$ in place of $G(\tilde{\xi})$; further, we define $LC = -\log C$ (i.e., with a minus sign).

1. Initialize LG and LC; set other parameters.
    – Set $LG(i) = 0$ for $i = 1, \ldots, N_b$.
    – Set $LC > 0$ (e.g., $LC = -\log(1/e) = 1$).
    – Set the maximum number of iterations $K_{\max}$ (e.g., $K_{\max} = 15$ or 18).
    – Set the maximum number of MCMC steps $M_{\max}$ within each iteration.
    – Set the counter of iterations K to 0.
2. Initialize H and X.
    – If $K > K_{\max}$, **end**.
    – Set $H(i) = 0$ for $i = 1, \ldots, N_b$.
    – Initialize the state X.
    – Set the counter of MCMC trials M to 0.
3. Run MCMC.
    – Run a step of MCMC with the weight $P(x) \exp(LG(\tilde{\xi}(x)))$.
4. Modify LG and update the histogram H.
    – $LG(\tilde{\xi}(x^*)) = LG(\tilde{\xi}(x^*)) - LC$, where $X = x^*$ is the current state. ♣
    – $H(\tilde{\xi}(x^*)) = H(\tilde{\xi}(x^*)) + 1$, where $X = x^*$ is the current state. ♠
5. Check whether H is "sufficiently flat."[13]
    – If so, $LC = LC/2$, $K = K + 1$ and go to Step 2
    – If not and $M < M_{\max}$, $M = M + 1$ and go to Step 3.
    – If not and $M_{\max} \leq M$, the algorithm **fails**.

Note that update ♣ of the weight G and increment ♠ of the histogram H are done simultaneously, in contrast to entropic sampling.

The criterion for a "sufficiently flat" histogram used in Sects. 3.1.1 and 3.1.2 is that counts in every bin of the histogram are larger than 92 % of the value expected in a

---

[12] Usually, in the Wang–Landau algorithm, this criterion for flatness should be severer than the requirement on the flatness of the histogram expected in the final production run.

[13] In actual implementation, this step need not to be performed after each step of MCMC; it can be done, for example, each time after trying to update all random variables.

perfectly flat histogram. In the cases of Sect. 3.1.2, we exclude "permanently" zero count bins from the criterion, where true probability seems zero; it is usually difficult to know a priori and some trial and error is required.

After completing the above procedure, the production run is performed. If this algorithm does not converge or the production run using the obtained weights does not give a flat histogram of ξ, what can we do? One possibility is to change the criterion that the histogram H is "sufficiently flat;" when we make it more strict and increase the value of M_{max}, convergence may be attained with increasing computational time. Increasing the value of K_{max} may not be effective when we use the original $\sqrt{C}$ rule for modifying C because the value of C becomes nearly unity for large K. Another possibility is to relax our requirement on the rareness and decrease the value of ξ_0.

The algorithm presented here still contains a number of ad hoc procedures and should be manually adapted to a specific problem. It, however, provides solutions to problems otherwise difficult to treat. On the other hand, many modifications of the algorithm are proposed. Examples of treating continuous variables are seen in Yan et al. (2002), Shell et al. (2002), Liang (2005), Zhou et al. (2006), Atchadé and Liu (2010). The following authors have criticized the $\sqrt{C}$ rule and have proposed modified algorithms: Belardinelli and Pereyra (2007a,b); Liang et al. (2007); Zhou and Su (2008); Atchadé and Liu (2010). The convergence of the algorithms is analyzed in Lee et al. (2006), Belardinelli and Pereyra (2007b), while rigorous mathematical proofs are discussed in Atchadé and Liu (2010), Jacob and Ryder (2011), Fort et al. (2012). Bornn et al. (2013) proposed an automatic procedure including the adaptation of step and bin size.

2.2.5 Variance of estimators

Finally, we will briefly discuss the variance of the estimators. Here, we restrict ourselves to the final production run with a fixed weight. An experimental study on convergence of estimates is shown in Sect. 3.1.1.

At first, we assume that all samples are independent, although it is not true for samples generated by MCMC. Then, variances of the numerator and denominator of the right-hand side of (6) are estimated as

$$\sigma_m^2 = \frac{1}{M} \left\{ \sum_{\xi=\xi_0}^{\xi_{max}} \tilde{P}(\xi) \, P(\xi) - [P(\xi_0 \leq \xi)]^2 \right\} \tag{9}$$

$$\sigma_{m0}^2 = \frac{1}{M} \left\{ \sum_{\xi=\xi_{min}}^{\xi_{max}} \tilde{P}(\xi) \, P(\xi) - [P(\xi_{min} \leq \xi \leq \xi_{max})]^2 \right\}. \tag{10}$$

From (9) and (10), the relative variance of the right-hand side of (6) is estimated as[14]

---

[14] Here, we apply the delta method using an approximation $\frac{a+\delta a}{b+\delta b} \big/ \frac{a}{b} \simeq 1 + \frac{\delta a}{a} - \frac{\delta b}{b}$; correlation between the denominator and the numerator is ignored.

$$\frac{\sigma_m^2}{[P(\xi_0 \leq \xi)]^2} + \frac{\sigma_{m0}^2}{[P(\xi_{\min} \leq \xi \leq \xi_{\max})]^2}. \tag{11}$$

In the case of MCMC, sample correlation becomes important and we should modify these formulae. Let us define integrated auto correlation of statistics $B(X)$ as

$$\mathcal{T}_B = \frac{1}{\sigma_B^2} \sum_{\tau=1}^{\infty} \left\{ \mathbb{E}_{\text{path}}[\, B(X^{(0)})\, B(X^{(\tau)})\,] - \mathbb{E}[\, B(X)\,]^2 \right\},$$

where the expectation $\mathbb{E}_{\text{path}}$ indicates an average over sample paths $X^{(0)}, X^{(1)}, \dots$ generated by MCMC, and σ_B^2 is the variance of independent samples from the same distribution. Then, the effective number of samples changes from M to $M/\mathcal{T}_B$, when we calculate the average of B. If we define $\mathcal{T}_m$ and $\mathcal{T}_{m0}$ as $\mathcal{T}_B$ with $B(X) = \tilde{P}(\xi(X))\, I(\xi_0 \leq \xi(X))$ and $\tilde{P}(\xi(X))$, respectively, (11) is substituted for

$$\frac{(1 + 2\mathcal{T}_m)\sigma_m^2}{[P(\xi_0 \leq \xi)]^2} + \frac{(1 + 2\mathcal{T}_{m0})\sigma_{m0}^2}{[P(\xi_{\min} \leq \xi \leq \xi_{\max})]^2}. \tag{12}$$

Unfortunately, it is rarely possible to estimate $\mathcal{T}_m$ and $\mathcal{T}_{m0}$ a priori. Expression (12), however, suggests that variances $\sigma_m^2, \sigma_{m0}^2$ of independent samples and integrated auto correlations $\mathcal{T}_m, \mathcal{T}_{m0}$ are both important in rare event sampling using MCMC. The multicanonical weight provides a practical method for balancing them.

If correlation among samples is ignored, a reasonable choice of Q for sampling from $\mathscr{A} = \{x | \xi_0 \leq \xi(x)\}$ is $Q_*(x) = \tilde{C} P(x) I(\xi_0 \leq \xi(x))$, which corresponds to the generation of samples using MCMC from the tail $\xi_0 \leq \xi(x)$ of the distribution P. It is, however, not useful in most practical problems, because it is difficult to design a Markov chain that efficiently samples from $Q_*(x)$.[15]

3 Examples of rare event sampling by multicanonical MCMC

Here, we discuss two applications of multicanonical MCMC, rare event sampling in random matrices and chaotic dynamical systems. Other applications in physics, engineering, and statistics are briefly surveyed.

3.1 Rare events in random matrices

A pioneering study on rare events in random matrices with multicanonical MCMC is Driscoll and Maki (2007), which computes large deviation in *growth ratio*, a quantity relevant to the numerical difficulty in treating matrices. The results in this subsection are discussed in detail in Saito et al. (2010) and Saito and Iba (2011). Kumar (2013) also applied the Wang-Landau algorithm to random matrices using coulomb gas formulation.

---

[15] In fact, even when conventional MCMC can produce samples of rare events from $Q_{opt}(x)$, calculation of the normalizing constant $\tilde{C}$ and the probability of rare events are not straightforward. An advantage of multicanonical MCMC is that it provides a way to calculate the probability using (6).

3.1.1 Largest eigenvalue

Distributions of the largest eigenvalue $\lambda^{\max}$ of random matrices are of considerable interest in statistics, ecology, cosmology, physics, and engineering. Small deviations have been studied in this problem, and have yielded the celebrated $N^{1/6}$ law by Tracy and Widom (1994, 1996). Here, we are interested in the numerical estimation of large deviations; the present analytical approach to large deviations is limited to specific types of distributions (Dean and Majumdar 2008; Majumdar and Vergassola 2009). Specifically, the probability $P(\lambda^{\max} < 0)$ that all eigenvalues are negative is important in many examples, because it is often related to the stability of the corresponding systems (May 1972; Aazami and Easther 2006).

In Saito et al. (2010), multicanonical MCMC is applied to this problem. Rare events whose probability $P(\lambda^{\max} < 0)$ is as small as 10^{-200} are successfully sampled for matrices of size $N \leq 30$ (or 40).[16]

Examples of the results in Saito et al. (2010) are shown in Figs. 4 and 5. In Fig. 4, the probability $P(\lambda_0^{\max} < \lambda^{\max} \leq \lambda_0^{\max} + \delta)$ is plotted against the values of $\lambda_0^{\max}$ with a small binsize δ for the case of Gaussian orthogonal ensemble (GOE). GOE is defined as an ensemble of random real symmetric matrices such that entries are independent Gaussian variables; hereafter, the variances of the diagonal and off-diagonal components are 1 and 0.5, respectively, while means are all zero.

Figure 5 shows the probability $P(\lambda^{\max} < 0)$ that all eigenvalues are negative. The results for GOE and an ensemble of real symmetric matrices whose components are uniformly distributed (hereafter "uniform") are shown.[17] For small Ns, the results from the proposed method reproduce those by simple random sampling.[18] On the other hand, for a large N for which simple random sampling hardly suffices, the obtained results match theoretical results in the case of GOE. The typical number of steps in preparing the multicanonical weight is $2 \sim 5 \times 10^9$, and the length of the final productive run ranges from 1×10^9 (GOE $N = 20$) to 2.5×10^9 (GOE $N = 40$, uniform $N = 30$).[19]

Examples of convergence of estimates are shown in Fig. 6. For an ensemble of matrices whose components are uniformly distributed, multicanonical weights for $N = 6, 12, 18$ and 24 are calculated by the Wang–Landau algorithm using at most 5.0×10^9 steps. Then, five independent production runs are performed for each N using the same weight obtained by this procedure. The results for an increasing length of the production run are shown in the figure. Noting that the vertical axis of Fig. 5

---

[16] The most time-consuming part of the proposed algorithm is the diagonalization procedure required for each step of MCMC; the Householder method is used here. It can be improved by the use of a more efficient method for calculating the eigenvalue $\lambda^{\max}$.

[17] The support of the uniform distributions is chosen as having the same variance as GOE.

[18] Hereafter, "simple random sampling" refers to the method wherein a large number of matrices are independently generated from the ensemble and the empirical proportion is used as an estimator.

[19] Hereafter, the length of MCMC runs is measured by the number of Metropolis–Hastings trials; we do *not* use physicists' "Monte Carlo steps (MCS)," which is defined as the number of trials divided by the number of random variables.
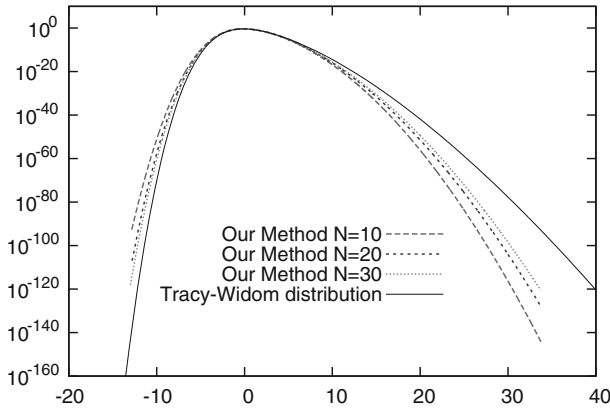
Fig. 4 The probability $P(\lambda_0^{\max} < \lambda^{\max} \leq \lambda_0^{\max} + \delta)$ estimated by the proposed method is plotted for $N = 10, 20$, and 30, where δ is a small binsize. The Tracy–Widom distribution for *small deviation* asymptotics is shown by the *solid curve*; systematic deviations from the obtained result for *large deviation* are observed as expected. The *horizontal axis* corresponds to the scaled variable $(\lambda_0^{\max} - \mathbb{E}(\lambda_0^{\max}))N^{1/6}$. Gaussian orthogonal ensemble (GOE) is assumed. From Saito et al. (2010), © 2010 American Physical Society

Fig. 5 Probabilities $P(\lambda^{\max} < 0)$ obtained by the proposed method and simple random sampling method are shown against N; the latter is available only for small N. *Left* GOE; *curve* indicates a quadratic fit to the results with Coulomb gas representation (Dean and Majumdar 2008). *Right* an ensemble of matrices whose components are uniformly distributed; *curve* indicates the probability for GOE with the same variance. From Saito et al. (2010), © 2010 American Physical Society

is log-scale, the variance of the estimates attained in Fig. 6 is reasonably small and is enough for providing an accurate test for asymptotics.

The proposed method is quite general and can be applied to random matrices whose components are sampled from an arbitrary distribution, or even random sparse matrices, to which no analytical solution is available. These results are discussed in detail in Saito et al. (2010), along with the detailed specifics of the proposed algorithm.

An important lesson from this example is that we should be careful while choosing the moves in the Metropolis–Hasting algorithm. If we generate candidates using conditional distributions of the original distribution, such as the Gaussian distribution for each component in GOE, the algorithm fails in some cases. This occurs because such a method cannot generate candidates with very large deviations in a component. This difficulty is avoided by the use of a random walk Metropolis algorithm with an
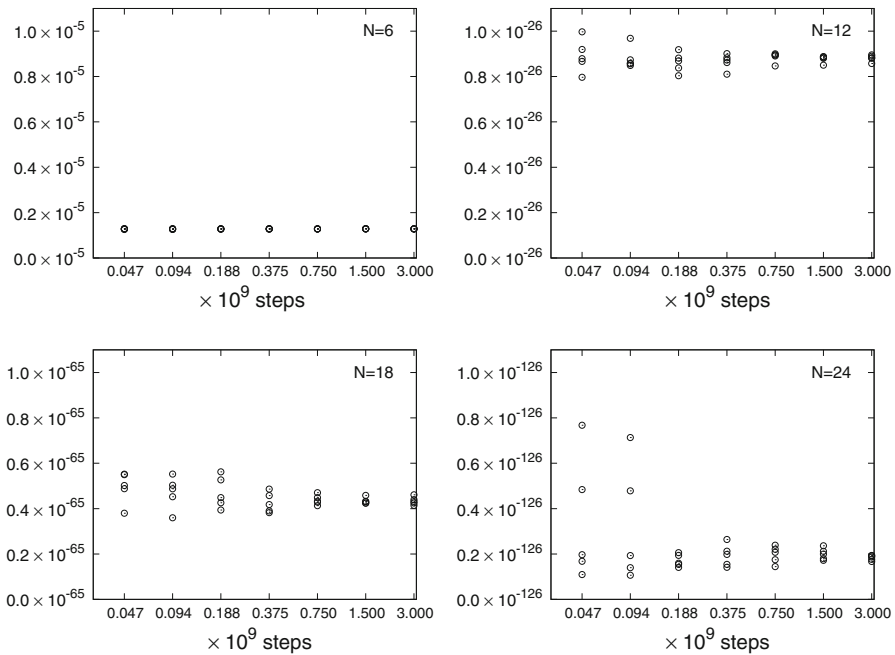
Fig. 6 Examples of the convergence of estimated probabilities. *Upper Left N* = 6, *Upper Right N* = 12, *Lower Left N* = 18, and *Lower Right N* = 24. The *vertical axis* corresponds to probabilities $P(\lambda^{\max} < 0)$, while the *horizontal axis* (log-scale) corresponds to the steps of the algorithm. Results of five production runs with the same weight and *different random numbers* are shown for each N; in the $N = 6$ case, *symbols* are almost overlapped one another. An ensemble of matrices whose components are uniformly distributed is assumed

adequate step size; in the example of GOE, we use Gaussian distributions as proposal distributions in the Metropolis algorithm (variances are unity for diagonals and 0.5 for non diagonals, respectively); see Saito et al. (2010).

3.1.2 Random graphs

The search for rare events in random graphs is also an interesting subject. An undirected graph is represented by the corresponding adjacency matrix, whose components take values in the set {0, 1}. For a k-regular graph, the maximum eigenvalue takes a fixed value equal to k, and hence it is not interesting. On the other hand, the spectral gap λ^{gap}, given as the difference between the maximum and the second-largest eigenvalue in the case of regular graphs, is related to many important properties of the corresponding graph. Specifically, graphs with larger values of the spectral gap are called *Ramanujan graphs* or *expanders*; Ramanujan graphs have interesting properties for communications and dynamics on networks (see references in Donetti et al. 2006; Saito and Iba 2011).

In earlier studies, Donetti et al. (2005, 2006) optimized the spectral gaps of graphs by simulated annealing; in their algorithm, a pair of edges of the graph is modified in
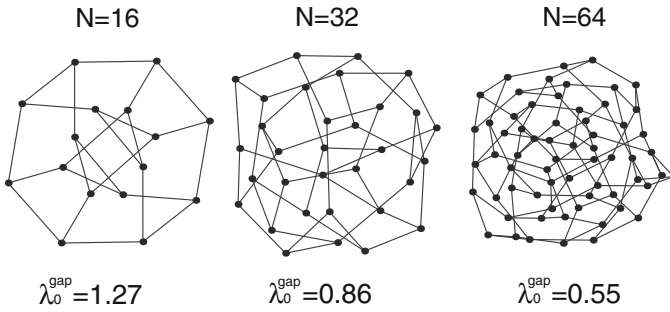
N=16 N=32 N=64

$\lambda_0^{\text{gap}}=1.27$ $\lambda_0^{\text{gap}}=0.86$ $\lambda_0^{\text{gap}}=0.55$

Fig. 7 Examples of 3-regular graphs with a large spectral gap found in the simulation. From Saito and Iba (2011), © 2011 Elsevier
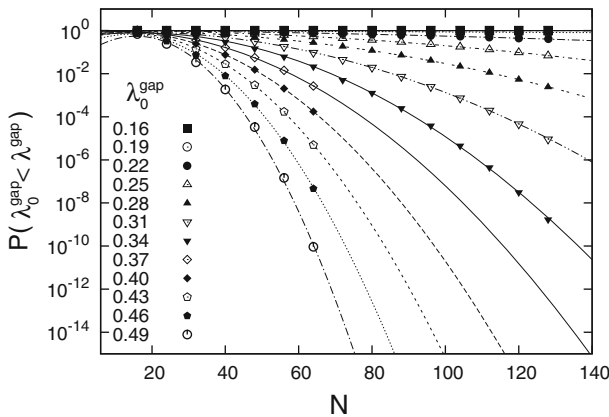

Fig. 8 Estimated $P(\lambda_0^{\text{gap}} < \lambda^{\text{gap}})$ are shown as functions of N. Each *curve* corresponds to *different values* of λ_0^{gap}. Data are well fitted by quadratic functions when $\lambda_0^{\text{gap}} \gtrsim 0.18$. From Saito and Iba (2011), © 2011 Elsevier

each Metropolis–Hasting step. Using this method, they showed that expanders with interesting structures automatically appear.

Saito and Iba (2011) applied multicanonical MCMC to this problem; they defined the Metropolis-Hasting update as in Donetti et al. (2005, 2006) and used the Wang–Landau algorithm for realizing multicanonical weights. Examples of the obtained graphs are shown in Fig. 7, while Fig. 8 gives probability $P(\lambda_0^{\text{gap}} < \lambda^{\text{gap}})$ as a function of λ_0^{gap} and the size N of matrices. The typical number of Metropolis steps used in preparing multicanonical weights is $0.5 \sim 1.0 \times 10^9$, while the length of the final production run is $0.25 \sim 0.5 \times 10^9$. See Saito and Iba (2011) for further details.

3.2 Rare events in dynamical systems

Rare events in deterministic dynamical systems are important both in theory and application (Ott 2002; Beck and Schlögl 1993). An example is a quantitative study on tiny tori embedded in a "chaotic sea" of Hamiltonian dynamical systems, which

is a familiar subject in this field. Numerical effort required for uncovering these tiny structures dramatically increases with the dimension of the system. Therefore, it is natural to introduce MCMC and other stochastic sampling methods to this field. Studies on MCMC search for unstable structures in dynamical systems are found in Sasa and Hayashi (2006), Yanagita and Iba (2009), Geiger and Dellago (2010), and references therein.[20]

Kitajima and Iba (2011) applied multicanonical MCMC to the study of dynamical systems. In the proposed algorithm, a measure of the chaoticity of a trajectory is defined as a function of the initial condition,[21] which corresponds to statistics representing rareness. Then, the Metropolis–Hastings update is defined as follows: (1) perturb the initial condition, (2) simulate a fragment of trajectory from the new initial condition, and (3) calculate the chaoticity of the trajectory and reject/accept the new initial condition using the current weight. Then, the entire algorithm is defined as multicanonical MCMC with the Wang–Landau algorithm for tuning the weight.

Again, the choice of moves in the Metropolis–Hastings algorithm is important; here, we sample a perturbation to the initial conditions from a mixture of uniform densities with different order of widths. This idea, taken from Sweet et al. (2001), seems essential for sampling from fractal-like densities; see Kitajima and Iba (2011) for details.

In Kitajima and Iba (2011), sampling of tiny tori in the chaotic sea of a four-dimensional map

$$u_{n+1} = u_n - \frac{K}{2\pi} \sin(2\pi v_n) + \frac{b}{2\pi} \sin(2\pi (v_n + y_n))$$
$$v_{n+1} = v_n + u_{n+1}$$
$$x_{n+1} = x_n - \frac{K}{2\pi} \sin(2\pi y_n) + \frac{b}{2\pi} \sin(2\pi (v_n + y_n))$$
$$y_{n+1} = y_n + x_{n+1}$$

is studied, where K and b are constants that characterize the map. An example of tiny tori found by the proposed method is shown in Fig. 9. In this case, the total number of initial conditions tested in the proposed algorithm is about 4×10^9, while the probability to find an initial configuration leading to a trajectory with the same degree of chaoticity is as small as 10^{-12}, assuming random sampling from the Lebesgue measure. In addition, the relative volume of initial conditions that lead to trajectories of the given order of "chaoticity" are successfully estimated by the algorithm; that is, the proposed method is not only useful for the search but also provides quantitative information on rare events in dynamical systems, see Fig. 2 of Kitajima and Iba (2011).

---

[20] Sequential Monte Carlo-like algorithms are also used (see Tailleur and Kurchan 2007; Laffargue et al. 2013, and references therein).

[21] Here, the chaoticity is defined as the number of iteration required for the divergence of perturbed trajectories; the algorithm according to this definition is stable on finite precision machines.
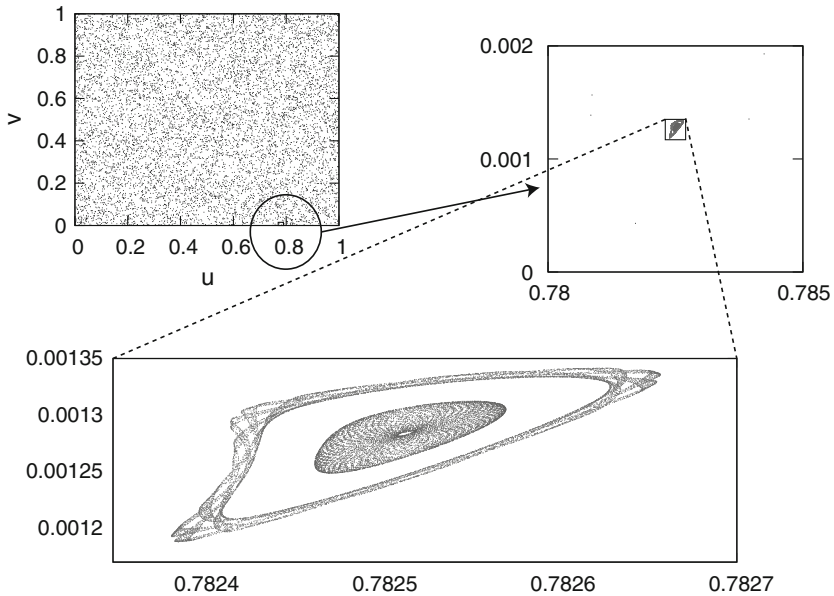
Fig. 9 Pair of tiny tori in the chaotic sea found by the proposed method. Projections on the (u_n, v_n)-plane are shown. Enlargement of a tiny area in the *small circle* in the *left panel* is given in the *right panel*; further enlargement is given in the *lower panel*. $K = 7.8$ and $b = 0.001$. From Kitajima and Iba (2011), © 2011 Elsevier

3.3 Other applications

The rest of this section briefly describes other fields of applications of multicanonical MCMC.

3.3.1 Statistical physics

Multicanonical MCMC was originally developed for sampling from Gibbs distributions in statistical physics. Hence, a number of studies in this field have successfully applied it to problems where simple MCMC is virtually disabled by slow mixing. Some typical examples are studies on the Potts and other classical spin models (Berg and Neuhaus 1992; Wang and Landau 2001a; Zhou et al. 2006), spin glass models (Berg and Celik 1992; Wang and Landau 2001a), and liquid models (Yan et al. 2002; Shell et al. 2002; Calvo 2002). Multicanonical MCMC is also used for the study of biomolecules (see Mitsutake et al. 2001; Higo et al. 2012 for full-atom protein models and Chikenji et al. 1999; Wüst and Landau 2012 for lattice protein models). Attempts to combine the idea of multicanonical weight with chain growth algorithms are found in Bachmann and Janke (2003), Prellberg and Krawczyk (2004). More examples are found in the review articles mentioned in Sect.1.

On the other hand, the use of multicanonical MCMC for other types of rare event sampling in physics is a recent challenge.[22] Hartmann (2002) introduced the idea of

---

[22] In terms of physics, it corresponds to the sampling of the "quenched disorder," whereas conventional applications in physics deal with sampling from the Gibbs distribution of thermal disorder.

rare event sampling by MCMC to the physics community. Körner et al. (2006) and Monthus and Garel (2006) applied MCMC to the sampling of disorder configurations that gives large deviations in ground state energies; in these studies, modifications of the Gumbel distribution are used for approximating multicanonical weights. Subsequently, Hukushima and Iba (2008) and Matsuda et al. (2008) applied the Wang–Landau algorithm to the study of Griffiths singularities in random magnets, which is known to be sensitive to rare configurations of impurities; Wolfsheimer and Hartmann (2010) discussed RNA secondary structures. In these studies, any prior knowledge on the functional form of $\tilde{P}(\xi)$ is assumed.

3.3.2 Optical telecommunication and related fields

Multicanonical MCMC is intensively used for rare event sampling in optical telecommunication and related fields. After a pioneering work by Yevick (2002), a number of applications appeared; see, for example, Holzlöhner and Menyuk (2003), and a recent review, Bononi et al. (2009). The sampling of rare noises that cause failures of error correction is discussed in Holzlöhner et al. (2005) and Iba and Hukushima (2008), which can be useful for predicting the performance of error-correcting codes.

3.3.3 Statistics

Algorithms based on the multicanonical weight, specifically, the Wang–Landau algorithm and its generalizations, increasingly attract the attention of statisticians. Liang (2005) introduced the Wang–Landau algorithm to statistics. Atchadé and Liu (2010) and Chopin et al. (2012) developed closely related algorithms and tested them in examples of Bayesian inference and model selection. Bornn et al. (2013) and Kastner et al. (2013) also discussed applications in Bayesian statistics; Kwon and Lee (2008) treated a target tracking problem. Yu et al. (2011) (also Liang et al. 2010) discussed hypothesis testing using stochastic approximation Monte Carlo. Wolfsheimer et al. (2011) extended the study of Hartmann (2002) and applied rare event sampling using the Wang-Landau method to the computation of p-values for local sequence alignment problems. In the following section, we will discuss exact tests and data surrogation as an application field of multicanonical MCMC for constrained systems.

4 Sampling from constrained systems and hypothesis testing

Sampling from highly constrained systems and combinatorial calculations are discussed here as a variation of the theme of rare event sampling. Exact tests and data surrogation are introduced as an application field of this idea, where efficient sampling from constrained systems is essential.

For general issues on Monte Carlo approximate counting, see Jerrum and Sinclair (1996); Rubinstein and Kroese (2008); Rubino and Tuffin (2009).

4.1 MCMC sampling from constrained systems

MCMC sampling is difficult when constraints exist among random variables. In such cases, it is often not easy to find a set of Metropolis–Hastings moves that realizes an

ergodic Markov chain without violating the constraints. For example, considerable effort is devoted to find ergodic moves for contingency tables with fixed margins and other constraints (Diaconis and Sturmfels 1998; Bunea and Besag 2000; Takemura and Aoki 2004).[23] Although partial success has been obtained using highly sophisticated mathematics, the problem becomes increasingly difficult when problem complexity increases.

Yet another general strategy for dealing with highly constrained systems is an introduction of "soft constraints." First, given constraints $f_i(x) = 0$, $i = 1, \ldots L$, we define statistics ξ of the state variables X that satisfy the following conditions: (1) $\xi(X) \geq 0$ and (2) $\xi(X) = 0$, if and only if X satisfy $f_i(X) = 0$ for all i. A simple example of such statistics is

$$\xi(x) = \sum_{i=1}^{L} c_i |f_i(x)|^{\alpha}.$$

Here, $\alpha > 0$ and $c_i > 0$ are arbitrary constants; $\alpha = 1$ is usually better than $\alpha = 2$ because ξ keeps small values when $|f_i(x)|$ increases in the case of $\alpha = 1$. Then, a finite value of ξ represents *soft* constraints, whereas $\xi = 0$ corresponds to the original *hard* constraints. Random sampling of the value of X usually gives a large value of $\xi(X)$; hence, $\xi(X) = 0$ can be regarded as a "rare event."

At this point, we introduce multicanonical MCMC with target statistics ξ and sample rare events X defined by $\xi(X) = 0$ (or, for a continuous variable X, $\xi(X) \approx 0$). Then, after tuning weights with the Wang–Landau algorithm, a production run provides samples of X that (nearly) satisfy the constraints $f_i(X) = 0$ (or $f_i(X) \approx 0$) for all i. Note that a similar strategy can be implemented using a combination of an exponential family with sufficient statistics ξ and replica exchange MCMC; in this case, a large value of β corresponds to hard constraints.

Some references are as follows.[24] Pinn and Wieczerkowski (1998) introduced replica exchange MCMC with soft constraints to this field, and the number of magic squares of size 6×6 is estimated in their paper. Kitajima and Kikuchi (private communication) extended it to 30×30 using multicanonical MCMC. Hukushima (2002) estimated the number of N-queen configurations by replica exchange MCMC, while Zhang and Ma (2009) treated N-queen and Latin squares using a hybrid of simulated tempering (see the appendix) and the Wang–Landau algorithm; they dealt with Latin squares up to size 100×100. Fishman (2012) proposed an approach based on soft constraints for counting contingency tables; conventional MCMC is used in his paper.

---

[23] See also Jacobson and Matthews (1996) for an algorithm specialized for Latin squares; it partially utilized a soft constraint strategy.

[24] "Self-avoidingness" of random walk is also well treated by the soft constraint strategy discussed here (see Vorontsov-Velyaminov et al. 1996, 2004; Iba et al. 1998; Chikenji et al. 1999; Shirai and Kikuchi 2013).

4.2 Application to hypothesis testing

Here, we discuss how multicanonical MCMC (and also replica exchange MCMC) can be useful for exact tests and data surrogation; the proposed method is tested with a simple example of time series.

4.2.1 MCMC exact tests

MCMC is useful for implementing statistical tests with a complicated null distribution. Particularly important cases occur when the null distribution is a distribution conditioned with a set of statistics ζ_i. In these cases, the null hypothesis is represented as the uniform distribution of X on the set defined by $\zeta_i(x) = \zeta_i^o$, $i = 1, \ldots L$, where X is a random variable and ζ_i^o is the value of statistics ζ_i corresponding to the observed data. For a continuous variable X, this condition can be relaxed as

$$|\zeta_i(x) - \zeta_i^o| < \epsilon_i, \ i = 1, \ldots L, \tag{13}$$

where ϵ_i is a constant with a small value.

A prototype of such a test is Fisher's exact test of contingency tables (Agresti 1992), where the marginals of the table correspond to ζ_i's; a number of extended versions exist and MCMC algorithms with complicated Metropolis moves have been developed for them, as mentioned in the previous section. Besag and Clifford (1989) described a test where an Ising model on the square lattice represents the null hypothesis.

In our view, it is natural to introduce the "soft constraint" strategy described in Sect. 4.1 to this problem. When we define the statistics ξ as $\xi(x) = \sum_{i=1}^{L} |\zeta_i(x) - \zeta_i^o|$, it is straightforward to apply multicanonical MCMC for sampling X that uniformly distributed on the set defined by $\zeta_i(x) = \zeta_i^o$, $i = 1, \ldots L$ or its generalization (13). This strategy is quite general and can be applied to a variety of MCMC hypothesis testing.[25]

4.2.2 Data surrogation

In nonlinear dynamics and neural science, statistical tests for time series based on (13) are well developed (Schreiber and Schmitz 2000). They are called as *surrogate data* methods, and samples from null distributions defined by (13) are called as *surrogates* of the original data. An example of the problem where surrogation is intensively used is testing of statistical properties of neural spike trains (Grün and Rotter 2010).

In conventional approaches, surrogates are generated by partial randomization of the original data. For example, if the phase of time series data $x^o(t), t = 1, \ldots, N$ is

---

[25] As mentioned in the previous section, Yu et al. (2011), Liang et al. (2010) also discussed hypothesis testing with stochastic approximation Monte Carlo, which can be regarded as a version of multicanonical MCMC in this case. They, however, focused on the problem of calculating small p-values; it differs from our idea of using multicanonical MCMC as a sampler from highly constrained systems.

randomized after the complex Fourier transform, then its inverse transform $x = \{x(t)\}$ has the same sets of correlation functions

$$C(x; \tau) = \sum_{t=1}^{N-\tau} x(t)x(t+\tau) \tag{14}$$

as the original time series[26] and is considered as a surrogate that maintains the value of sufficient statistics $\zeta_\tau(x) = C(x; \tau)$. Although a quick solution is provided in this case, solutions to general cases are only found on a case-by-case basis, and it becomes increasingly difficult as the complexity of the problems increases.

Therefore, Schreiber proposed a general idea of regarding data surrogation as an optimization problem (Schreiber 1998; Schreiber and Schmitz 2000). According to this idea, generating a surrogate is equivalent to finding a solution of (13), which can be treated by a general-purpose optimization algorithm, for example, simulated annealing. An application of this idea in neural science is found in Hirata et al. (2008).

This was an epoch-making idea in this field; randomization via a clever idea was no longer required, being replaced by a routine procedure at the cost of computational time. However, in data surrogation, we want to generate a sample (or a set of samples) unbiasedly selected from the null distribution defined by (13), and not obtain a sample that satisfies (13).

Therefore, applying multicanonical MCMC seems a better choice. Hence, we again arrive at the idea of exact testing with multicanonical MCMC.

4.2.3 Example

Let us illustrate the idea of "multicanonical surrogation" using an example from Schreiber (1998).[27,28] In this example, the problem is to generate artificial time series $x = \{x_1, x_2, \ldots, x_N\}$ by permuting the original time series $x^o = \{x_1^o, x_2^o, \ldots, x_N^o\}$ given as observed data. The constraint is to maintain the correlation functions $C(x; \tau)$, defined as (14), to be nearly equal to the original correlation functions $C(x^o; \tau)$ for $\tau = 1 \ldots T$; here, the constant $0 < T < N$ is the maximum of the delay τ, where we expect correlation coincidence.

Here, $\xi(X) = \sum_{\tau=1}^{T} |C(X; \tau) - C(x^o; \tau)|$ is used to define multicanonical MCMC that samples $X = \{X_1, X_2, \ldots X_N\}$. $\xi(X)$ is zero if and only if $C(X; \tau) = C(x^o; \tau)$ for all $1 \leq \tau \leq T$. Then, Metropolis–Hastings moves are defined by the swap of a randomly selected pair. In detail, a pair i and j is selected by a random number in each step and a new candidate x^{new} of X is generated by $x_i^{\text{new}} = x_j$ and $x_j^{\text{new}} = x_i$

---

[26] To be precise, we should assume a periodic boundary condition and change the upper limit of the summation from $N - \tau$ to N.

[27] The results in this subsection (including Figs. 10, 11) appeared in an IEICE Technical Report IBISML2011-7(2011-06) in Japanese, as a report without peer review. These have never been published in English.

[28] A quick practical solution is present for this problem, but it is not a perfect one; see Schreiber (1998).
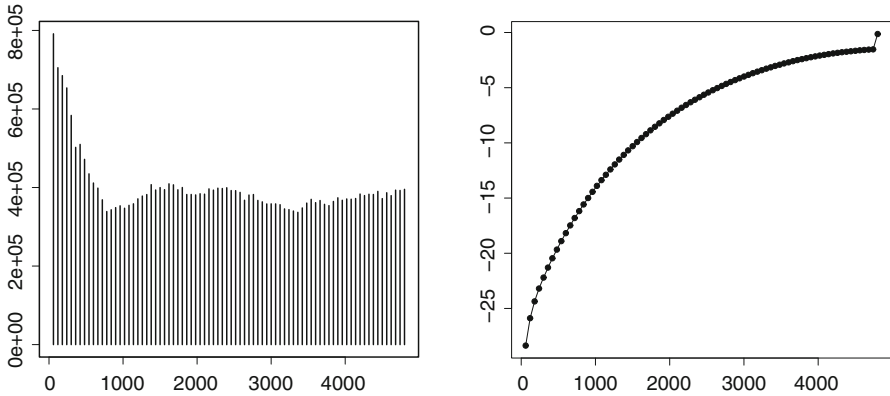
Fig. 10 *Left* frequency of the occurrence of ξ in the production run of multicanonical MCMC. The *horizontal* and *vertical axes* correspond to the value of ξ and the observed frequency in the given bins, respectively. *Right* probability density of ξ. The *horizontal* and *vertical axes* correspond to the value of ξ and the estimated log-probabilities ($\log_{10}$), respectively; a set of bins used in the *left panel* is also applied in the *right panel* for defining probabilities. The spike in the rightmost bin corresponds to a cumulated probability of larger values of ξ [from Y. Iba, IEICE Technical Report IBISML2011-7(2011-06), 43–50, in Japanese, © IEICE 2011]

without changing other components, using the current values $X_i = x_i$ and $X_j = x_j$. Here, the value of $\{X_i\}$ is initialized as a random permutation of $\{x_i^o\}$.

In the following experiment, we consider time series x^o of length $N = 400$ generated by nonlinear observations of a linear AR process y driven by uniform noise, that is,

$$x_t^o = y_t^3, \quad y_{t+1} = 0.3y_t + \eta_t, \quad \eta_t \sim U(-2, 2).$$

Here, we choose $T = 8$. Multicanonical MCMC is designed for realizing an approximately flat distribution of ξ in the interval [0, 4800.8], which is divided into 80 bins.[29] In this choice of the interval, we consider two conditions: (1) the interval contains a high entropy region where the values of ξ are readily realized by a random permutation of the original time series, and (2) the last bin $\xi \simeq 0$ corresponds to a tail region of ξ that we are interested in. The Wang–Landau algorithm with $K_{\max} = 15$ is used to tune the weight; the $\sqrt{C}$ rule is utilized. At each step of the iteration, we run MCMC until counts in each bin coincide with the value for the uniform histogram within 1 % accuracy. The total number of Metropolis trials is 2.5×10^8, of which 3.2×10^7 are used for the final production run.

The results of this experiment are shown in Figs. 10 and 11. In Fig. 10, the distribution of ξ realized in the production run and the estimated log-density of ξ are shown. The former is not quite flat in a non-logarithmic scale, but enough to ensure efficient production of the desired samples. According to the right panel of Fig. 10, the probability of obtaining a sample within the bin $\xi \approx 0$ is estimated to be as small as 10^{-25} or less, assuming a random permutation of x^o.

---

[29] Here, we round the value of ξ to $\xi^{\max}$ when it exceeds $\xi^{\max}$ instead of rejecting the candidate; this causes the spike at the right edge of the density in the right panel of Fig. 10.
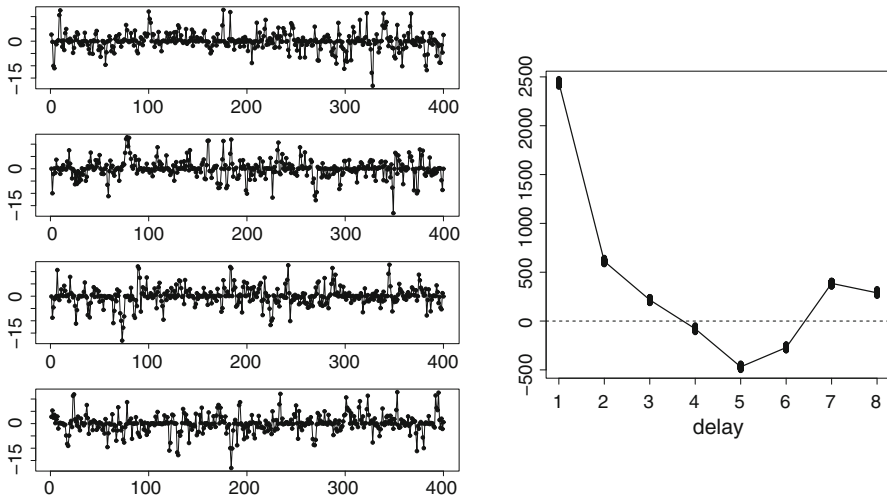
Fig. 11 *Left* surrogate data generated by the proposed method. Uppermost series correspond to the original data and other three are surrogates. The *horizontal* and *vertical axes* correspond to t and x_t, respectively. *Right* comparison of correlation functions. The *horizontal and vertical axes* correspond to the delay τ and the values of the correlation function, respectively. The *line* represents $C(x^o, \tau)$, which corresponds to the original data, while *black dots* represent sets of $C(x^{(k)}, \tau)$ obtained from surrogated data; $N = 400$ is the length of the time series. The results of 1976 samples are printed over each other; hence, *symbols* are almost overlapping [from Y. Iba, IEICE Technical Report IBISML2011-7(2011-06), 43–50, in Japanese, © IEICE 2011]

In Fig. 11, the quality of the obtained samples is examined. In the left panel, three samples in the last bin $\xi \simeq 0$ are shown, which are considerably different from one another. In the right panel, correlation functions $C(x^{(k)}, \tau)$ are calculated for each of the 1976 samples $X = x^{(k)}$, $k = 1, \ldots, 1976$, in the bin $\xi \approx 0$ and compared to the original $C(x^o, \tau)$, which indicate an extremely good agreement between them.[30]

5 Summary and discussions

In this paper, we discussed rare event sampling using multicanonical MCMC. Two different methods of tuning the weight, entropic sampling and the Wang–Landau algorithm, are explained. Then, examples for random matrices, random graphs, chaotic dynamical systems, and data surrogation are shown. We hope our exposition will be useful for the exploration of further novel applications of multicanonical MCMC.

Appendix

Multicanonical MCMC for exponential family

We begin this paper with a history of multicanonical MCMC; it was originally developed as a method for sampling from Gibbs distributions, or an exponential family. Here, we briefly discuss how to use multicanonical MCMC for this original purpose.

---

[30] Note that not all 1976 samples are independent; some additional test is needed for estimating the number of independent samples in our run.

Assume that we want to compute the expectation $\mathbb{E}_\beta[A(X)] = \sum_x A(x) \exp(\beta\xi(x))/Z_\beta$ of statistics A from the output $X^{(i)}, i = 1, \ldots, M$ obtained from multicanonical MCMC that realizes an almost flat marginal of ξ in a "sufficiently wide" interval $[\xi_{\min}, \xi_{\max}]$. Then, for $M \to \infty$, the desired expectation is computed by the *reweighting formula*[31]

$$\mathbb{E}_\beta[A(X)] \simeq \frac{\sum_{i=1}^M \left[A(X^{(i)}) \, \tilde{P}(\xi(X^{(i)})) \, \exp(\beta\xi(X^{(i)})) \right]}{\sum_{i=1}^M \left[\tilde{P}(\xi(X^{(i)})) \, \exp(\beta\xi(X^{(i)})) \right]}. \tag{15}$$

Further, we have an expression for the normalizing constant Z_β as

$$\frac{Z_\beta}{V} \simeq \frac{\sum_{i=1}^M \left[\tilde{P}(\xi(X^{(i)})) \, \exp(\beta\xi(X^{(i)})) \right]}{\sum_{i=1}^M \tilde{P}(\xi(X^{(i)}))}, \tag{16}$$

where V is the total number of states of the variable X that satisfy $\xi_{\min} < \xi(X) < \xi_{\max}$; it is useful for the calculation of marginal likelihood in statistics and free energy in physics.

It is easy to derive these expressions[32] considering that the multicanonical weight is proportional to $\tilde{P}(\xi(X^{(i)}))^{-1}$. Expressions (15) and (16), however, are quite unusual in the sense that we can use them for a broad range of β where the interval $[\xi_{\min}, \xi_{\max}]$ covers a necessary region. Using this property, multicanonical MCMC simultaneously gives the expectations $\mathbb{E}_\beta[A(X)]$ for all β, through a single production run of a single chain. This is because a multicanonical weight gives a flat distribution of ξ that has a considerable overlap with the distribution $\exp(\beta\xi(x))/Z_\beta$ for any value of β, which is intuitively understood from the left panel in Fig. 12.

If we consider a similar reweighing that uses outputs of MCMC at β' for computing the expectation at a different β, it is practically impossible for a high-dimensional X unless the difference $|\beta' - \beta|$ is very small. This is because the overlap of the distributions virtually vanishes as shown in the right panel of Fig. 12; in such cases, the variance of summands on the right-hand side of (15) drastically increases.

First-order transition and "Phase coexistence"

As already mentioned in the main text, there are examples in which a region of ξ is virtually not realized for any choice of the canonical parameter β of the exponential family with sufficient statistics ξ. The marginal distribution of ξ has multiple peaks in this region of β, as illustrated in Fig. 13. Such examples naturally appear in statistical physics, when we study the "phase coexistence" phenomena near first-order phase

---

[31] To use this formula for an off-line calculation of the average of A, the values of ξ and A should be recorded as pairs in the simulation, like $(\xi(X^{(i)}), A(X^{(i)})), i = 1, \ldots, M$.

[32] Note that (15) becomes (7), if we substitute $I(\xi_0 \le \xi(X^{(i)}))$ for $\exp(\beta\xi(X^{(i)}))$.
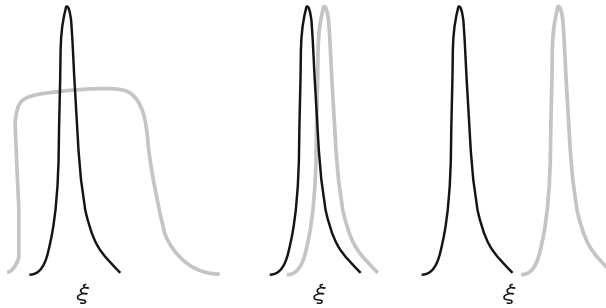
Fig. 12 Overlap of marginals of ξ. The *horizontal axis* corresponds to the sufficient statistics ξ. *Left* a multicanonical weight (*gray*) and a member with a given β of the exponential family (*black*). *Central and right panels* a pair of members with different values β and β' of the exponential family. The *center panel* corresponds to cases with a small $|\beta' - \beta|$, whereas the *right panel* corresponds to cases with a large $|\beta' - \beta|$
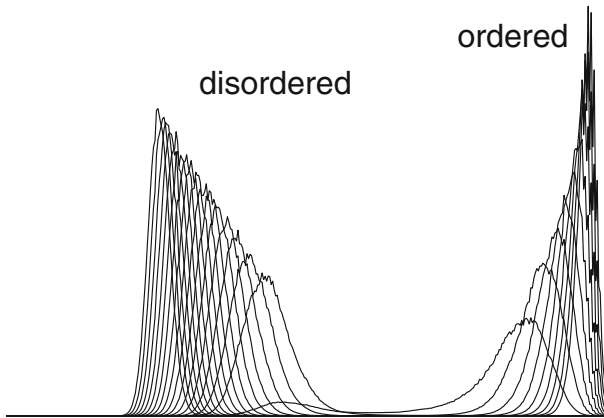

Fig. 13 Marginals of ξ with different values of β in the case of phase coexistence; the *horizontal axis* corresponds to the sufficient statistics ξ. These *curves* are obtained from the 10-states Potts model (see Berg and Neuhaus 1992), which consists of discrete variables $\{X_i\}$, $X_i \in \{1, \ldots, 10\}$ on a *square* lattice; they are computed by reweighting of the outputs of a single production run of multicanonical MCMC. If $\{X_i\}$ belongs to an "ordered" component, most X_is take the same value. In contrast, their values are almost random in the "disordered" component. In the disordered component, ξ takes smaller values, but the number of states of X that belongs to the component is large; hence, the total probability is comparable in both components

transitions.[33] On the other hand, it seems that the significance of such phenomena in statistics and engineering has not been fully explored.

In such cases, distributions defined by multicanonical weights are not well approximated by a mixture of the members of the corresponding exponential family; this is easily understood by considering Fig. 13. Hence, the advantage of replica exchange MCMC is limited because the sample path is blocked by the gap of ξ, while multi-

---

[33] Ice and water coexist at 0 °C; that is, both of them correspond to the same β but the values of average energy $-\mathbb{E}(\xi)$ are different.
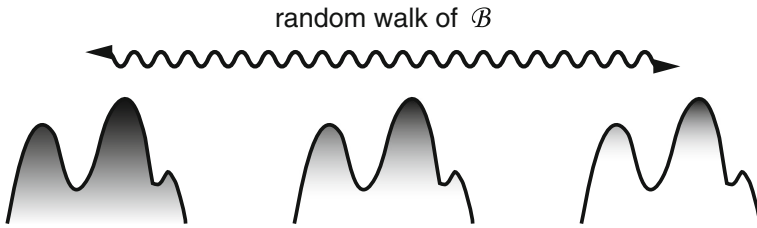
random walk of $\mathcal{B}$

Fig. 14 Random walk of inverse temperature variable $\mathcal{B}$ in simulated tempering. The *vertical axis* corresponds to the value of ξ, whereas the *horizontal axis* for each sub-chart schematically represents a high-dimensional space of X. The values of inverse temperature β are assumed to increase from *left to right*; the *shading* represents the corresponding changes in high-probability regions. Note that a random variable $\mathcal{B}$ is updated by MCMC, retaining the value of X at that time; in other words, a separated procedure for changing $\mathcal{B}$ is required for simulated tempering

canonical MCMC can, in principle, do better. Both methods, however, seem to fail in very difficult cases; see Iba and Takahashi (2005).

Simulated tempering

The "third" method, *simulated tempering* (Marinari and Parisi 1992; Geyer and Thompson 1995), or *expanded ensemble Monte Carlo* (Lyubartsev et al. 1992),[34] is briefly explained here. Practically, we recommend choosing between multicanonical MCMC and replica exchange MCMC. Simulated tempering, however, provides an idea that interpolates these two algorithms and is conceptually important. The idea is simple—inverse temperature β is regarded as a random variable (hereafter denoted by $\mathcal{B}$) and we consider MCMC sampling of $(X, \mathcal{B})$ from the combined distribution

$$P(x, \beta) = \frac{\exp(\beta\xi(x))}{Z_\beta} \pi(\beta) = \exp(\beta\xi(x) - \log Z_\beta + \log \pi(\beta)).$$

Hereafter, we choose a "pseudo prior" $\pi(\beta)$ as a uniform density on $[\beta_{\min}, \beta_{\max}]$, resulting in a random walk of $\mathcal{B}$ that uniformly covers the interval $[\beta_{\min}, \beta_{\max}]$; see Fig. 14. This behavior is similar to that of multicanonical MCMC, but here $\mathcal{B}$ is a variable updated in a separate step of MCMC; in contrast, a random walk of $\xi(X)$ is induced by the update of the state X in case of multicanonical MCMC.

Although this concept is simple, a difficulty arises because the MCMC update of $\mathcal{B}$ requests the value of Z_β as a function of β, which is unknown in most cases.[35] Hence, we should introduce the estimation of Z_β using repeated preliminary runs, which is similar to the weight tuning procedure in multicanonical MCMC. See the

---

[34] This paper introduced an idea similar to simulated tempering in an even more general framework.

[35] The normalizing constant (partition function) Z_β is not required for replica exchange MCMC, because it cancels in the Metropolis–Hastings ratio necessary for deciding whether to accept/reject the swap of the states between chains; this is an essential advantage of replica exchange MCMC.

above references, as well as Zhang and Ma (2007), which introduced a method like the Wang–Landau algorithm.

Implementation on parallel hardware

Replica exchange MCMC is naturally parallelizable. Then, how can we efficiently implement multicanonical MCMC on parallel hardware? A simple solution is parallelization of the weight tuning stage. That is, a set of preliminary runs is performed in parallel, each of which runs on a CPU; they share a histogram where total number of visits to each value of ξ is recorded. Some variants of this idea are discussed in, for example, Zhan (2008), Bornn et al. (2013). If we want to go beyond these schemes, something more intricate is required. For example, the range of ξ is divided into a set of intervals and a multicanonical weight is realized in each of them (see Wang and Landau 2001a; Mitsutake et al. 2001; Vogel et al. 2013). In the latter two studies, the exchange of states between neighboring intervals is incorporated.

Multivariate extensions

Multicanonical MCMC samples a high-dimensional X, while adaptation of the weight is performed in a one-dimensional space of ξ. It is possible to introduce a "multivariate multicanonical weight," which realizes an almost uniform density in a region of two-dimensional (ξ_1, ξ_2) or even three-dimensional (ξ_1, ξ_2, ξ_3) spaces, where ξ_k is a function of X. Examples of such extensions are found in Shteto et al. (1997), Iba et al. (1998), Higo et al. (1997), Chikenji et al. (1999), Chikenji and Kikuchi (2000), Yan et al. (2002), Zhou et al. (2006). Usually, the adaptation of weights in a multivariate case is more difficult than in a univariate case, because of the sparseness of the data collected in the preliminary run; Zhou et al. (2006) proposed the use of kernel density estimation for this problem.

References

Aazami, A., Easther, R. (2006). Cosmology from random multifield potentials. *Journal of Cosmology and Astroparticle Physics*, *3*(03), 013.

Agresti, A. (1992). A survey of exact inference for contingency tables. *Statistical Science*, *7*(1), 131–153.

Atchadé, Y. F., Liu, J.S. (2010). The Wang-Landau algorithm in general state spaces: Applications and convergence analysis. *Statistica Sinica*, *20*(1), 209–233.

Bachmann, M., Janke, W. (2003). Multicanonical chain-growth algorithm. *Physical Review Letters*, *91*(20), 208105.

Baumann, B. (1987). Noncanonical path and surface simulation. *Nuclear Physics B*, *285*, 391–409.

Beck, C., Schlögl, F. (1993). *Thermodynamics of Chaotic Systems: An Introduction*. Cambridge: Cambridge University Press.

Belardinelli, R., Pereyra, V. (2007a). Fast algorithm to calculate density of states. *Physical Review E*, *75*, 046701.

Belardinelli, R., Pereyra, V. (2007b). Wang-Landau algorithm: A theoretical analysis of the saturation of the error. *The Journal of Chemical Physics*, *127*, 184105.

Berg, B. A. (2000). Introduction to multicanonical Monte Carlo simulations. *Fields Institute Communications*, *26*, 1–24.

Berg, B. A. (2004). *Markov Chain Monte Carlo Simulations and Their Statistical Analysis*. Singapore: World Scientific.

Berg, B. A., Celik, T. (1992). New approach to spin-glass simulations. *Physical Review Letters*, *69*(15), 2292–2295.

Berg, B. A., Neuhaus, T. (1991). Multicanonical algorithms for first order phase transitions. *Physics Letters B*, *267*(2), 249–253.

Berg, B. A., Neuhaus, T. (1992). Multicanonical ensemble: A new approach to simulate first-order phase transitions. *Physical Review Letters*, *68*(1), 9–12.

Besag, J., Clifford, P. (1989). Generalized Monte Carlo significance tests. *Biometrika*, *76*(4), 633–642.

Binder, K., Heermann, D. (2012). *Monte Carlo Simulation in Statistical Physics: An Introduction*. Berlin: Springer.

Birge, J. R., Chang, C., Polson, N. G. (2012). Split sampling: Expectations, normalisation and rare events. *ArXiv e-prints*, *1212*, 0534.

Bononi, A., Rusch, L., Ghazisaeidi, A., Vacondio, F., Rossi, N. (2009). A fresh look at multicanonical Monte Carlo from a telecom perspective. In *Global Telecommunications Conference, 2009. GLOBECOM 2009*, IEEE, pp 1–8.

Bornn, L., Jacob, P. E., Del Moral, P., Doucet, A. (2013). An adaptive interacting Wang-Landau algorithm for automatic density exploration. *Journal of Computational and Graphical Statistics*, *22*(3), 749–773.

Botev, Z. I., L'Ecuyer, P., Tuffin, B. (2013). Markov chain importance sampling with applications to rare event probability estimation. *Statistics and Computing*, *23*(2), 271–285.

Brooks, S., Gelman, A., Jones, G. L., Meng, X. L. (Eds.). (2011). *Handbook of Markov Chain Monte Carlo*. New York: Chapman and Hall/CRC.

Bucklew, J. A. (2004). *Introduction to Rare Event Simulation (Springer Series in Statistics)*. New York: Springer.

Bunea, F., Besag, J. (2000). MCMC in $I \times J \times K$ contingency tables. *Fields Institute Communications*, *26*, 25–36.

Calvo, F. (2002). Sampling along reaction coordinates with the Wang-Landau method. *Molecular Physics*, *100*(21), 3421–3427.

Chikenji, G., Kikuchi, M. (2000). What is the role of non-native intermediates of β-lactoglobulin in protein folding? *Proceedings of the National Academy of Sciences, 97*(26), 14,273–14,277.

Chikenji, G., Kikuchi, M., Iba, Y. (1999). Multi-self-overlap ensemble for protein folding: ground state search and thermodynamics. *Physical Review Letters*, *83*(9), 1886–1889.

Chopin, N., Lelièvre, T., Stoltz, G. (2012). Free energy methods for Bayesian inference: Efficient exploration of univariate Gaussian mixture posteriors. *Statistics and Computing*, *22*(4), 897–916.

Dean, D. S., Majumdar, S. N. (2008). Extreme value statistics of eigenvalues of Gaussian random matrices. *Physical Review E*, *77*(4), 041108.

de Oliveira, P. M. C., Penna, T. J. P., Herrmann, H. J. (1998). Broad histogram Monte Carlo. *The European Physical Journal B-Condensed Matter and Complex Systems*, *1*(2), 205–208.

Diaconis, P., Sturmfels, B. (1998). Algebraic algorithms for sampling from conditional distributions. *The Annals of statistics*, *26*(1), 363–397.

Donetti, L., Hurtado, P. I., Muñoz, M. A. (2005). Entangled networks, synchronization, and optimal network topology. *Physical Review Letters*, *95*(18), 188701.

Donetti, L., Neri, F. (2006). Muñoz MA (2006) Optimal network topologies: Expanders, cages, Ramanujan graphs, entangled networks and all that. *Journal of Statistical Mechanics: Theory and Experiment*, *08*, P08007.

Driscoll, T. A., Maki, K. L. (2007). Searching for rare growth factors using multicanonical Monte Carlo methods. *SIAM Review*, *49*(4), 673–692.

Fishman, G. S. (2012). Counting contingency tables via multistage Markov chain Monte Carlo. *Journal of Computational and Graphical Statistics*, *21*(3), 713–738.

Fort, G., Jourdain, B., Kuhn, E., Lelièvre, T., Stoltz, G. (2012). Convergence and efficiency of the Wang-Landau algorithm. *ArXiv e-prints*, *1207*, 6880.

Frenkel, D., Smit, B. (2002). *Understanding Molecular Simulation, From Algorithms to Applications (Computational Science Series)* (2nd ed.). San Diego: Academic Press.

Geiger, P., Dellago, C. (2010). Identifying rare chaotic and regular trajectories in dynamical systems with Lyapunov weighted path sampling. *Chemical Physics*, *375*(2–3), 309–315.

Geyer, C. J. (1991). Markov chain Monte Carlo maximum likelihood. In E. Keramidas (Ed.), *Computing science and statistics: Proceedings of 23rd Symposium on the Interface* (pp. 156–163). Fairfax Station: Interface Foundation.

Geyer, C. J., Thompson, E. A. (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. *Journal of the American Statistical Association*, *90*(431), 909–920.

Gilks, W. R., Richardson, S., Spiegelhalter, D. J. (Eds.). (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

Grün, S., Rotter, S. (Eds.). (2010). *Analysis of Parallel Spike Trains (Springer Series in Computational Neuroscience)*. New York: Springer.

Hartmann, A. K. (2002). Sampling rare events: Statistics of local sequence alignments. *Physical Review E*, *65*(5), 056102.

Higo, J., Nakajima, N., Shirai, H., Kidera, A., Nakamura, H. (1997). Two-component multicanonical Monte Carlo method for effective conformation sampling. *Journal of computational chemistry*, *18*(16), 2086–2092.

Higo, J., Ikebe, J., Kamiya, N., Nakamura, H. (2012). Enhanced and effective conformational sampling of protein molecular systems for their free energy landscapes. *Biophysical Reviews*, *4*, 27–44.

Hirata, Y., Katori, Y., Shimokawa, H., Suzuki, H., Blenkinsop, T. A., Lang, E. J., et al. (2008). Testing a neural coding hypothesis using surrogate data. *Journal of Neuroscience Methods*, *172*(2), 312–322.

Holzlöhner, R., Menyuk, C. R. (2003). Use of multicanonical Monte Carlo simulations to obtain accurate bit error rates in optical communications systems. *Optics Letters*, *28*(20), 1894–1896.

Holzlöhner, R., Mahadevan, A., Menyuk, C. R., Morris, J. M., Zweck, J. (2005). Evaluation of the very low BER of FEC codes using dual adaptive importance sampling. *IEEE Communications Letters*, *9*(2), 163–165.

Hukushima, K. (2002). Extended ensemble Monte Carlo approach to hardly relaxing problems. *Computer Physics Communications*, *147*(1–2), 77–82.

Hukushima, K., Iba, Y. (2008). A Monte Carlo algorithm for sampling rare events: application to a search for the Griffiths singularity. *Journal of Physics: Conference Series*, *95*, 012005.

Hukushima, K., Nemoto, K. (1996). Exchange Monte Carlo method and application to spin glass simulations. *Journal of the Physical Society of Japan*, *65*(6), 1604–1608.

Iba, Y. (2001). Extended ensemble Monte Carlo. *International Journal of Modern Physics C*, *12*(05), 623–656.

Iba, Y., Hukushima, K. (2008). Testing error correcting codes by multicanonical sampling of rare events. *Journal of the Physical Society of Japan*, *77*(10), 103801.

Iba, Y., Takahashi, H. (2005). Exploration of multi-dimensional density of states by multicanonical Monte Carlo algorithm. *Progress of Theoretical Physics Supplements*, *157*, 345–348.

Iba, Y., Chikenji, G., Kikuchi, M. (1998). Simulation of lattice polymers with multi-self-overlap ensemble. *Journal of the Physical Society of Japan*, *67*, 3327–3330.

Jacob, P. E., Ryder, R. J. (2011). The Wang-Landau algorithm reaches the flat histogram criterion in finite time. *ArXiv e-prints*, *1110*, 4025.

Jacobson, M. T., Matthews, P. (1996). Generating uniformly distributed random Latin squares. *Journal of Combinatorial Designs*, *4*(6), 405–437.

Janke, W. (1998). Multicanonical Monte Carlo simulations. *Physica A: Statistical Mechanics and its Applications*, *254*(1–2), 164–178.

Jerrum, M., Sinclair, A. (1996). The Markov chain Monte Carlo method: An approach to approximate counting and integration. *Approximation algorithms for NP-hard problems*, pp. 482–520.

Kastner, C. A., Braumann, A., Man, P. L. W., Mosbach, S., Brownbridge, G. P. E., Akroyd, J., et al. (2013). Bayesian parameter estimation for a jet-milling model using Metropolis-Hastings and Wang-Landau sampling. *Chemical Engineering Science*, *89*, 244–257.

Kimura, K., Taki, K. (1991). Time-homogeneous parallel annealing algorithm. In *Proceedings of the 13th IMACS World Congress on Computation and Applied Mathematics (IMACS'91)*, Vol. 2, pp. 827–828.

Kirkpatrick, S., Gelatt, C. D., Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, *220*(4598), 671–680.

Kitajima, A., Iba, Y. (2011). Multicanonical sampling of rare trajectories in chaotic dynamical systems. *Computer Physics Communications*, *182*(1), 251–253.

Körner, M., Katzgraber, H. G., Hartmann, A. K. (2006). Probing tails of energy distributions using importance-sampling in the disorder with a guiding function. *Journal of Statistical Mechanics: Theory and Experiment*, *04*, P04005.

Kumar, S. (2013). Random matrix ensembles: Wang-Landau algorithm for spectral densities. *Europhysics Letters*, *101*(2), 20002.

Kwon, J., Lee, K.M. (2008). Tracking of abrupt motion using Wang-Landau Monte Carlo estimation. In *Proceedings of the 10th European Conference on Computer Vision: Part I*, ECCV '08, Heidelberg: Springer, pp. 387–400.

Laffargue, T., Lam, K. D. N. T., Kurchan, J., Tailleur, J. (2013). Large deviations of Lyapunov exponents. *Journal of Physics A: Mathematical and Theoretical*, *46*(25), 254002.

Landau, D.P., Binder, K. (2009). *A Guide to Monte Carlo Simulations in Statistical Physics* (3rd ed.). Cambridge University Press.

Landau, D. P., Tsai, S. H., Exler, M. (2004). A new approach to Monte Carlo simulations in statistical physics: Wang-Landau sampling. *American Journal of Physics*, *72*(10), 1294–1302.

Lee, H. K., Okabe, Y., Landau, D. P. (2006). Convergence and refinement of the Wang-Landau algorithm. *Computer Physics Communications*, *175*(1), 36–40.

Lee, J. (1993). New Monte Carlo algorithm: Entropic sampling. *Physical Review Letters*, *71*(2), 211–214.

Liang, F. (2005). A generalized Wang-Landau algorithm for Monte Carlo computation. *Journal of the American Statistical Association*, *100*(472), 1311–1327.

Liang, F., Liu, C., Carroll, R. J. (2007). Stochastic approximation in Monte Carlo computation. *Journal of the American Statistical Association*, *102*(477), 305–320.

Liang, F., Liu, C., Carroll, R. J. (2010). *Advanced Markov Chain Monte Carlo Methods: Learning from Past Samples (Wiley Series in Computational Statistics)*. West Sussex: Wiley.

Lyubartsev, A. P., Martsinovski, A. A., Shevkunov, S. V., Vorontsov-Velyaminov, P. N. (1992). New approach to Monte Carlo calculation of the free energy: Method of expanded ensembles. *The Journal of Chemical Physics*, *96*(3), 1776–1783.

Majumdar, S. N., Vergassola, M. (2009). Large deviations of the maximum eigenvalue for Wishart and Gaussian random matrices. *Physical Review Letters*, *102*(6), 060601.

Marinari, E., Parisi, G. (1992). Simulated tempering: A new Monte Carlo scheme. *Europhysics Letters*, *19*(6), 451–458.

Matsuda, Y., Nishimori, H., Hukushima, K. (2008). The distribution of Lee-Yang zeros and Griffiths singularities in the $\pm$ J model of spin glasses. *Journal of Physics A: Mathematical and Theoretical*, *41*(32), 324012.

May, R. M. (1972). Will a large complex system be stable? *Nature*, *238*, 413–414.

Mezei, M. (1987). Adaptive umbrella sampling: Self-consistent determination of the non-Boltzmann bias. *Journal of Computational Physics*, *68*(1), 237–248.

Mitsutake, A., Sugita, Y., Okamoto, Y. (2001). Generalized-ensemble algorithms for molecular simulations of biopolymers. *Biopolymers (Peptide Science)*, *60*(2), 96–123.

Monthus, C., Garel, T. (2006). Probing the tails of the ground-state energy distribution for the directed polymer in a random medium of dimension d = 1, 2, 3 via a Monte Carlo procedure in the disorder. *Physical Review E*, *74*(5), 051109.

Newman, M. E. J., Barkema, G. T. (1999). *Monte Carlo Methods in Statistical Physics*. New York: Clarendon Press.

Ott, E. (2002). *Chaos in Dynamical Systems*. Chambridge: Cambridge University Press.

Pinn, K., Wieczerkowski, C. (1998). Number of magic squares from parallel tempering Monte Carlo. *International Journal of Modern Physics C*, *09*(04), 541–546.

Prellberg, T., Krawczyk, J. (2004). Flat histogram version of the pruned and enriched Rosenbluth method. *Physical Review Letters*, *92*(12), 120602.

Robert, C. P., Casella, G. (2004). *Monte Carlo Statistical Methods* (2nd ed.). New York: Springer.

Rubino, G., Tuffin, B. (Eds.). (2009). *Rare Event Simulation using Monte Carlo Methods*. West Sussex: Wiley.

Rubinstein, R. Y., Kroese, D. P. (2008). *Simulation and the Monte Carlo Method (Wiley Series in Probability and Statistics)* (2nd ed.). Hoboken: Wiley-Interscience.

Saito, N., Iba, Y. (2011). Probability of graphs with large spectral gap by multicanonical Monte Carlo. *Computer Physics Communications*, *182*(1), 223–225.

Saito, N., Iba, Y., Hukushima, K. (2010). Multicanonical sampling of rare events in random matrices. *Physical Review E*, *82*(3), 031142.

Sasa, S., Hayashi, K. (2006). Computation of the Kolmogorov-Sinai entropy using statistical mechanics: Application of an exchange Monte Carlo method. *Europhysics Letters*, *74*(1), 156–162.

Schreiber, T. (1998). Constrained randomization of time series data. *Physical Review Letters*, *80*(10), 2105–2108.

Schreiber, T., Schmitz, A. (2000). Surrogate time series. *Physica D: Nonlinear Phenomena*, *142*(3–4), 346–382.

Schulz, B. J., Binder, K., Müller, M., Landau, D. P. (2003). Avoiding boundary effects in Wang-Landau sampling. *Physical Review E*, *67*(6), 067102.

Shell, M. S., Debenedetti, P. G., Panagiotopoulos, A. Z. (2002). Generalization of the Wang-Landau method for off-lattice simulations. *Physical Review E*, *66*(5), 056703.

Shirai, N. C., Kikuchi, M. (2013). Multicanonical simulation of the Domb-Joyce model and the Gō model: new enumeration methods for self-avoiding walks. *Journal of Physics: Conference Series*, *454*(1), 012039.

Shteto, I., Linares, J., Varret, F. (1997). Monte Carlo entropic sampling for the study of metastable states and relaxation paths. *Physical Review E*, *56*(5), 5128–5137.

Sweet, D., Nusse, H. E., Yorke, J. A. (2001). Stagger-and-step method: Detecting and computing chaotic saddles in higher dimensions. *Physical Review Letters*, *86*(11), 2261–2264.

Tailleur, J., Kurchan, J. (2007). Probing rare physical trajectories with Lyapunov weighted dynamics. *Nature Physics*, *3*(3), 203–207.

Takemura, A., Aoki, S. (2004). Some characterizations of minimal Markov basis for sampling from discrete conditional distributions. *Annals of the Institute of Statistical Mathematics*, *56*(1), 1–17.

Torrie, G. M., Valleau, J. P. (1974). Monte Carlo free energy estimates using non-Boltzmann sampling: Application to the sub-critical Lennard-Jones fluid. *Chemical Physics Letters*, *28*(4), 578–581.

Tracy, C. A., Widom, H. (1994). Level-spacing distributions and the Airy kernel. *Communications in Mathematical Physics*, *159*(1), 151–174.

Tracy, C. A., Widom, H. (1996). On orthogonal and symplectic matrix ensembles. *Communications in Mathematical Physics*, *177*(3), 727–754.

Vogel, T., Li, Y. W., Wüst, T., Landau, D. P. (2013). Generic, hierarchical framework for massively parallel Wang-Landau sampling. *Physical Review Letters*, *110*, 210603.

Vorontsov-Velyaminov, P. N., Broukhno, A. V., Kuznetsova, T. V., Lyubartsev, A. (1996). Free energy calculations by expanded ensemble method for lattice and continuous polymers. *The Journal of Physical Chemistry*, *100*(4), 1153–1158.

Vorontsov-Velyaminov, P. N., Volkov, N. A., Yurchenko, A. A. (2004). Entropic sampling of simple polymer models within Wang-Landau algorithm. *Journal of Physics A: Mathematical and General*, *37*(5), 1573–1588.

Wang, F., Landau, D. P. (2001a). Determining the density of states for classical statistical models: A random walk algorithm to produce a flat histogram. *Physical Review E*, *64*(5), 056101.

Wang, F., Landau, D. P. (2001b). Efficient, multiple-range random walk algorithm to calculate the density of states. *Physical Review Letters*, *86*(10), 2050–2053.

Wang, J. S., Swendsen, R. H. (2002). Transition matrix Monte Carlo method. *Journal of Statistical Physics*, *106*(1–2), 245–285.

Wolfsheimer, S., Hartmann, A. K. (2010). Minimum-free-energy distribution of RNA secondary structures: Entropic and thermodynamic properties of rare events. *Physical Review E*, *82*(2), 021902.

Wolfsheimer, S., Herms, I., Rahmann, S., Hartmann, A. K. (2011). Accurate statistics for local sequence alignment with position-dependent scoring by rare-event sampling. *BMC Bioinformatics, 12*(1), 47.

Wüst, T., Landau, D. P. (2012). Optimized Wang-Landau sampling of lattice polymers: Ground state search and folding thermodynamics of HP model proteins. *The Journal of Chemical Physics*, *137*(6), 064903.

Yan, Q., Faller, R., de Pablo, J. J. (2002). Density-of-states Monte Carlo method for simulation of fluids. *The Journal of Chemical Physics*, *116*(20), 8745–8749.

Yanagita, T., Iba, Y. (2009) Exploration of order in chaos using the replica exchange Monte Carlo method. *Journal of Statistical Mechanics: Theory and Experiment*, *02*, P02043.

Yevick, D. (2002). Multicanonical communication system modeling-Application to PMD statistics. *IEEE Photonics Technology Letters*, *14*(11), 1512–1514.

Yu, K., Liang, F., Ciampa, J., Chatterjee, N. (2011). Efficient p-value evaluation for resampling-based tests. *Biostatistics*, *12*(3), 582–593.

Zhan, L. (2008). A parallel implementation of the Wang-Landau algorithm. *Computer Physics Communications*, *179*(5), 339–344.

Zhang, C., Ma, J. (2007). Simulation via direct computation of partition functions. *Physical Review E*, *76*(3), 036708.

Zhang, C., Ma, J. (2009). Counting solutions for the N-queens and Latin-square problems by Monte Carlo simulations. *Physical Review E*, *79*(1), 016703.

Zhou, C., Su, J. (2008). Optimal modification factor and convergence of the Wang-Landau algorithm. *Physical Review E*, *78*(4), 046705.

Zhou, C., Schulthess, T. C., Torbrügge, S., Landau, D. P. (2006). Wang-Landau algorithm for continuous models and joint density of states. *Physical Review Letters*, *96*(12), 120201.