Optimally adjusted mixture sampling and locally weighted histogram analysis

Zhiqiang Tan[1]

February 2014, revised October 2015

Abstract. Consider the two problems of simulating observations and estimating expectations and normalizing constants for multiple distributions. First, we present a self-adjusted mixture sampling method, which accommodates both adaptive serial tempering and a generalized Wang–Landau algorithm. The set of distributions are combined into a labeled mixture, with the mixture weights depending on the initial estimates of log normalizing constants (or free energies). Then observations are generated by Markov transitions, and free energy estimates are adjusted online by stochastic approximation. We propose two stochastic approximation schemes by Rao–Blackwellization of the scheme commonly used, and derive the optimal choice of a gain matrix, resulting in the minimum asymptotic variance for free energy estimation, in a simple and feasible form. Second, we develop an offline method, locally weighted histogram analysis, for estimating free energies and expectations, using all the simulated data from multiple distributions by either self-adjusted mixture sampling or other sampling algorithms. This method can be computationally much faster, with little sacrifice of statistical efficiency, than a global method currently used, especially when a large number of distributions are involved. We provide both theoretical results and numerical studies to demonstrate the advantages of the proposed methods.

Key words and phrases. Free energy; Markov chain Monte Carlo; Normalizing constant; Parallel tempering; Potts model; Serial tempering; Stochastic approximation; Wang–Landau algorithm; Weighted histogram analysis method.

[1]Department of Statistics, Rutgers University. Address: 110 Frelinghuysen Road, Piscataway, NJ 08854. E-mail: ztan@stat.rutgers.edu. An earlier version of the work was presented in a statistics seminar at Columbia University in November 2013. The author thanks the Editor, an Associate Editor, and two referees for helpful comments.

1 Introduction

Monte Carlo computation often involves simulating observations and estimating expectations and normalizing constants for multiple distributions. Consider a set of m probability distributions on a state space $\mathcal{X}$ by

$$\mathrm{d}P_j = \frac{q_j(x)}{Z_j}\,\mathrm{d}\mu, \quad j = 1, \ldots, m,$$

where μ is a baseline measure, $q_j(x)$ is an unnormalized density function, and $Z_j = \int q_j(x)\,\mathrm{d}\mu$ is the normalizing constant. In addition, let P_0 be another distribution with an unnormalized density function $q_0(x)$. Assume that $q_j(x)$ can be directly evaluated, but Z_j is analytically intractable ($j = 0, 1, \ldots, m$). There are several types of settings, where $(P_1, \ldots, P_m)$ and P_0 can be specified for different purposes.

For the first type of settings, both $(P_1, \ldots, P_m)$ and P_0 are directly taken from a family of distributions under study, and the objective is to sample from $(P_1, \ldots, P_m)$ and estimate expectations and normalizing constants for $(P_1, \ldots, P_m)$ and P_0. For example, a Boltzmann distribution in statistical physics is of the form P_j with $q_j(x) = \exp\{-u(x)/T_j\}$, where $u(x)$ is a potential function and T_j is a temperature. The samples from Boltzmann distributions $(P_1, \ldots, P_m)$ can be reweighted to a nearby temperature P_0 for which no observations are simulated. The Potts model is studied at multiple temperatures near phase transition in Section 6.1.

For the second type of settings, the primary problem is to sample from and estimate expectations for only one of the distributions, P_m (e.g., Geyer & Thompson 1995), or to estimate the log ratio of normalizing constants (i.e., the free energy difference in physics) between two distributions, P_1 and P_m (e.g., Meng & Wong 1996; Tan et al. 2012). The remaining distributions are introduced to facilitate solving the primary problem. See, for example, Gelman & Meng (1998) and Jasra et al. (2007) for discussions on the construction of such auxiliary distributions.

The third type of settings involve partitioning the state space $\mathcal{X}$ along, for example, the energy function $-\log q_0(x)$ for importance sampling (e.g., Wang & Landau 2001; Liang et al. 2007; Atchadé & Liu 2010). Effectively, $(P_1, \ldots, P_m)$ are defined as the restrictions of P_0 to the individual regions of the partition, but then sampled with uniform proportions over time. The observations obtained from such a mixture distribution can be reweighted to distribution P_0 by importance sampling. See Appendix V for a numerical study of the Potts model in this approach.

There are at least three computational problems of interest: (i) to simulate observations from $(P_1, \ldots, P_m)$, (ii) to estimate the expectations $E_j(\phi) = \int \phi(x)\, \mathrm{d}P_j$ $(j = 0, 1, \ldots, m)$ for a function $\phi(x)$, and (iii) to estimate the normalizing constants $Z_j = \int q_j(x)\, \mathrm{d}\mu$ $(j = 0, 1, \ldots, m)$ up to a multiplicative constant or, equivalently, to estimate the log ratios of normalizing constants (i.e., the free energy differences),

$$\zeta_j^* = \log(Z_j/Z_1), \quad j = 0, 1, \ldots, m,$$

where, without loss of generality, Z_1 is chosen to be a reference value. In the following, we provide a brief discussion of existing methods.

For the sampling problem, it is possible to run separate simulations for $(P_1, \ldots, P_m)$ by Markov chain Monte Carlo (MCMC) (e.g., Liu 2001). However, this direct approach tends to be ineffective in the presence of multi-modality and irregular contours in $(P_1, \ldots, P_m)$. See, for example, bimodal energy histograms for the Potts model in Figure 1. To address these difficulties, various methods have been proposed by creating interactions between samples from different distributions. Such methods can be divided into at least two categories. On one hand, overlap-based algorithms, including parallel tempering (Geyer 1991) and its extensions (Liang & Wong 2001), serial tempering (Geyer & Thompson 1995), resample-move (Gilks & Berzuini 2001) and its extensions (De Moral et al. 2006; Tan 2015), require that there exist considerable overlaps between $(P_1, \ldots, P_m)$, in order to exchange or transfer information across distributions. On the other hand, the Wang–Landau (2001) algorithm and its extensions (Liang et al. 2007; Atchadé & Liu 2010) are typically based on partitioning of the state space $\mathcal{X}$, and hence there is no overlap between $(P_1, \ldots, P_m)$.

For the estimation problem, the expectations $\{E_1(\phi), \ldots, E_m(\phi)\}$ can be directly estimated by sample averages from $(P_1, \ldots, P_m)$. However, additional considerations are generally required for estimating $(\zeta_2^*, \ldots, \zeta_m^*, \zeta_0^*)$ and $E_0(\phi)$, depending on the type of sampling algorithms. For Wang–Landau type algorithms based on partitioning of $\mathcal{X}$, $(\zeta_2^*, \ldots, \zeta_m^*)$ are estimated during the sampling process, and ζ_0^* and $E_0(\phi)$ can then be estimated by importance sampling techniques (Liang 2009). For overlap-based settings, both $(\zeta_2^*, \ldots, \zeta_m^*, \zeta_0^*)$ and $\{E_1(\phi), \ldots, E_m(\phi), E_0(\phi)\}$ can be estimated after sampling by a methodology known in physics and statistics as the (binless) weighted histogram analysis method (WHAM) (Ferrenberg & Swendsen 1989; Tan et al. 2012), the multi-state Bennett acceptance ratio method (Bennett 1976; Shirts & Chodera

2008), reverse logistic regression (Geyer 1994), bridge sampling (Meng & Wong 1996) and the global likelihood method (Kong et al. 2003; Tan 2004). See Gelman & Meng (1998), Tan (2013a), and Cameron & Pettitt (2014) for reviews on this method and others such as thermodynamic integration or equivalently path sampling.

The purpose of this article is twofold, dealing with sampling and estimation respectively. First, we present a self-adjusted mixture sampling method, which not only accommodates adaptive serial tempering and the generalized Wang–Landau algorithm in Liang et al. (2007), but also facilitates further methodological development. The sampling method employs stochastic approximation to estimate the log normalizing constants (or free energies) online, while generating observations by Markov transitions. We propose two stochastic approximation schemes by Rao–Blackwellization of the scheme used in Liang et al. (2007) and Atchadé & Liu (2010). For all the three schemes, we derive the optimal choice of a gain matrix, resulting in the minimum asymptotic variance for free energy estimation, in a simple and feasible form. In practice, we suggest a two-stage implementation that uses a slow-decaying gain factor during burn-in before switching to the optimal gain factor.

Second, we make novel connections between self-adjusted mixture sampling and the global method of estimation (e.g., Kong et al. 2003). Based on this understanding, we develop a new offline method, locally weighted histogram analysis, for estimating free energies and expectations using all the simulated data by either self-adjusted mixture sampling or other sampling algorithms, subject to suitable overlaps between $(P_1, \ldots, P_m)$. The local method is expected to be computationally much faster, with little sacrifice of statistical efficiency, than the global method, because individual samples are locally pooled from neighboring distributions, which typically overlap more with each other than with other distributions. The computational savings from using the local method are important, especially when a large number of distributions are involved (i.e., m is large, in hundreds or more), for example, in physical and chemical simulations (Chodera & Shirts 2011), likelihood inference (Tan 2013a, 2013b), and Bayesian model selection and sensitivity analysis (Doss 2010).

We provide new theoretical results on the effect of Rao–Blackwellization and the comparison between online and offline estimation. We also provide several numerical studies to demonstrate the advantages of the proposed methods.

2 Labeled mixture sampling

We describe a sampling method, labeled mixture sampling, which is the non-adaptive version of self-adjusted mixture sampling. The ideas are recast from several existing methods, including serial tempering (Geyer & Thompson 1995), the Wang–Landau (2001) algorithm and its extensions (Liang et al. 2007; Atchadé & Liu 2010). However, we make explicit the relationship between mixture weights and hypothesized normalizing constants, which is crucial to the new development of adaptive schemes in Section 3.2 and offline estimation in Sections 4–5.

The basic idea of labeled mixture sampling is to combine $(P_1, \dots, P_m)$ into a joint distribution on the space $\{1, \dots, m\} \times \mathcal{X}$:

$$(L, X) \sim p(j; x; \zeta) \propto \frac{\pi_j}{\mathrm{e}^{\zeta_j}} q_j(x), \tag{1}$$

where $\pi = (\pi_1, \dots, \pi_m)^{\mathrm{T}}$ are fixed mixture weights with $\sum_{j=1}^m \pi_j = 1$ (typically, $\pi_1 = \cdots = \pi_m = m^{-1}$), and $\zeta = (\zeta_1, \dots, \zeta_m)^{\mathrm{T}}$ with $\zeta_1 = 0$ are *hypothesized* values of the true log ratios of normalizing constants $\zeta^* = (\zeta_1^*, \dots, \zeta_m^*)^{\mathrm{T}}$ with $\zeta_1^* = 0$.

The marginal distribution of L under (1) is

$$p(L = j; \zeta) = \frac{\pi_j \mathrm{e}^{-\zeta_j + \zeta_j^*}}{\sum_{l=1}^m \pi_l \mathrm{e}^{-\zeta_l + \zeta_l^*}}, \quad j = 1, \dots, m. \tag{2}$$

The marginal distribution of X under (1) is

$$p(x; \zeta) = \sum_{j=1}^m \pi_j \mathrm{e}^{-\zeta_j} q_j(x), \quad x \in \mathcal{X},$$

which is a mixture distribution with the weight $p(L = j; \zeta) \propto \pi_j \mathrm{e}^{-\zeta_j + \zeta_j^*}$ for P_j. We refer to (1) as a labeled mixture, because L is a label, indicating from which distribution P_j an observation X is drawn. By Eq. (2), there is a one-to-one relationship between hypothesized free energies ζ and mixture weights $p(L = \cdot; \zeta)$. In particular, π gives the *target* mixture weights that would be obtained by setting $\zeta = \zeta^*$.

For any fixed choice of ζ (in addition to π), the labeled mixture (1) can be sampled by standard MCMC, with the unnormalized density $\pi_j \mathrm{e}^{-\zeta_j} q_j(x)$. For some initial values (L_0, X_0), a general Metropolis–Hastings (MH) algorithm is as follows.

MH labeled mixture sampling:

- Generate (j, x) from a proposal distribution $Q\{(L_{t-1}, X_{t-1}), \cdot; \zeta\}$.

- Set $(L_t, X_t) = (j, x)$ with probability

$$\min\left[1,\ \frac{Q\{(j,x),(L_{t-1},X_{t-1});\zeta\}}{Q\{(L_{t-1},X_{t-1}),(j,x);\zeta\}}\frac{p(j,x;\zeta)}{p(L_{t-1},X_{t-1};\zeta)}\right],$$

  and, with the remaining probability, set $(L_t, X_t) = (L_{t-1}, X_{t-1})$.

At this point, it is important to distinguish two different settings as discussed in the Introduction. For overlap-based settings, $(P_1, \ldots, P_m)$ are required to be overlapped with each other (e.g., Geyer 1994; Geyer & Thompson 1995). In contrast, for partition-based settings, $(P_1, \ldots, P_m)$ are supported on mutually exclusive regions of a partition of $\mathcal{X}$ (Wang & Landau 2001). For concreteness, we focus on overlap-based settings, until Appendix IV in Supplementary Materials on partition-based settings.

For $j = 1, \ldots, m$, assume that a Markov transition kernel, $\Psi_j(x; \cdot)$ is constructed by MCMC with P_j as the stationary distribution. Then a particular choice of $Q(\cdot, \cdot; \zeta)$ is to update L_t and X_t one at a time, leading to a two-block MH algorithm using $(\Psi_1, \ldots, \Psi_m)$. In fact, the conditional distributions under (1) are

$$\begin{aligned} p(x|L=j) &\propto q_j(x), \\ p(L=j|x;\zeta) &= \frac{\pi_j \mathrm{e}^{-\zeta_j} q_j(x)}{\sum_{l=1}^m \pi_l \mathrm{e}^{-\zeta_l} q_l(x)} \propto \frac{\pi_j}{\mathrm{e}^{\zeta_j}} q_j(x). \end{aligned} \tag{3}$$

That is, $p(x|L = j)$ corresponds to the jth target distribution P_j, regardless of ζ, whereas $p(L = \cdot|x; \zeta)$ is a discrete distribution on $\{1, \ldots, m\}$, depending on ζ. Sampling directly from $p(L = \cdot|x; \zeta)$ leads to a global-jump algorithm.

Global-jump labeled mixture sampling:

- *Global jump*: Generate $L_t \sim p(L = \cdot|X_{t-1}; \zeta)$.
- *Markov move*: Generate $X_t \sim \Psi_{L_t}(X_{t-1}, \cdot)$.

Alternatively, an MH transition can be used for sampling from $p(L = \cdot|X_{t-1}; \zeta)$. For $k = 1, \ldots, m$, let $\mathcal{N}(k)$ be a neighborhood of labels to k and $\Gamma(k, \cdot)$ be a proposal distribution for jumping from k to another label. A typical example is to set $\Gamma(k, j) = 1/s(k)$ if $j \in \mathcal{N}(k)$ and 0 otherwise, where $s(k)$ is the size of $\mathcal{N}(k)$. The resulting local-jump algorithm gives serial tempering (Geyer & Thompson 1995).

Local-jump labeled mixture sampling (i.e., serial tempering):

- *Local jump*: Generate $j \sim \Gamma(L_{t-1}, \cdot)$ and then set $L_t = j$ with probability

$$\min\left[1,\ \frac{\Gamma(j, L_{t-1})}{\Gamma(L_{t-1}, j)}\frac{p(j|X_{t-1};\zeta)}{p(L_{t-1}|X_{t-1};\zeta)}\right],$$

  and, with the remaining probability, set $L_t = L_{t-1}$.

- *Markov move*: Generate $X_t \sim \Psi_{L_t}(X_{t-1}, \cdot)$.

The local-jump algorithm is computationally less costly than the global-jump algorithm. At each iteration, the m unnormalized densities, $q_1(X_{t-1}, \ldots, q_m(X_{t-1})$, are evaluated in the global-jump algorithm, whereas only 2 unnormalized densities, $q_{L_{t-1}}(X_{t-1})$ and $q_j(X_{t-1})$, are evaluated in the local-jump algorithm. On the other hand, the relative statistical efficiency seems to be problem-dependent between the global-jump and local-jump algorithms. Chodera & Shirts (2011) presented examples where the global-jump algorithm leads to more rapid mixing then the local-jump algorithm. But the two algorithms perform similarly to each other in our simulation study of the Potts model near phase transition in Section 6.1.

3 Self-adjusted mixture sampling

A crucial issue for labeled mixture sampling is that, to paraphrase Geyer (2011, Section 11.2.4) on serial tempering, the choice ζ must be specified reasonably close to ζ^* in order for the algorithm to work. By Eq. (2), if ζ is not close to ζ^*, then the marginal probability of one label j may be orders of magnitude smaller than those of other labels, indicating that P_j is not adequately sampled.

Recently, serial tempering have been extended for sampling and estimating ζ adaptively by Liang et al. (2007) and Atchadé & Liu (2010), both motivated by the Wang–Landau (2001) algorithm. In particular, the model selection sampler in Liang et al. (2007, Section 5) can be modified as follows in our setting of labeled mixture sampling. Let $\zeta^{(0)}$ be some initial choice of ζ, for example, the $m \times 1$ vector of zeros. Denote by $\zeta^{(t)} = (\zeta_1^{(t)}, \ldots, \zeta_m^{(t)})^{\mathrm{T}}$ a choice of ζ at iteration t.

Stochastic approximation Monte Carlo (SAMC):

- *Local jump & Markov move*: Same as local-jump labeled mixture sampling.
- *Free energy update*: Set $\delta^{(t)} = (1\{L_t = 1\}, \ldots, q\{L_t = m\})^{\mathrm{T}}$ and

$$\zeta^{(t-\frac{1}{2})} = \zeta^{(t-1)} + \gamma_t(\delta^{(t)} - \pi), \quad \zeta^{(t)} = \zeta^{(t-\frac{1}{2})} - \zeta_1^{(t-\frac{1}{2})}, \tag{4}$$

  where $\zeta_1^{(t-\frac{1}{2})}$ is the first element of $\zeta^{(t-\frac{1}{2})}$ and $\gamma_t = t_0/\max(t_0, t)$ for some fixed value $t_0 > 1$. Liang et al. (2007) suggested setting t_0 between $2m$ and $100m$.

The free energy update in the SAMC algorithm is an application of stochastic approximation to find ζ^* as a unique solution to $p(L = j; \zeta) = \pi_j$ $(j = 1, \ldots, m)$.

Informally, there is a self-adjusting mechanism as follows. If the jth element $\zeta_j^{(t-1)}$ is smaller (or greater) than ζ_j^*, then the label L_t will, on average over time, take value j more likely (or less likely) than with probability π_j by Eq. (2). By the update rule (4), $\zeta_j^{(t)}$ will then be increased (or decreased) from $\zeta_j^{(t-1)}$.

For the rest of this section, we provide a brief review of stochastic approximation in Section 3.1 and then further develop the use of stochastic approximation for labeled mixture sampling in Section 3.2.

3.1 Stochastic approximation

There is a vast literature on theory, methods, and applications of stochastic approximation since Robbins & Monro (1951). In addition to the SAMC algorithm (Liang et al. 2007) mentioned above, examples of using stochastic approximation for Monte Carlo computation include maximum likelihood estimation for missing-data problems and spatial models (e.g., Delyon et al. 1999; Gu & Zhu 2001) and adaptive MCMC (e.g., Roberts & Rosenthal 2009).

Suppose that the objective is to find a solution θ^* to $h(\theta) = 0$ with

$$h(\theta) = E_\theta\{H(Y;\theta)\},$$

where θ is a r-dimensional parameter in Θ, $H(\cdot;\theta)$ is a r-dimensional vector of functions, and $E_\theta(\cdot)$ denotes the expectation with $Y \sim f(\cdot;\theta)$, a probability density function depending on θ. Informally, it is of interest to find the value of θ such that the expectation of a "noisy observation" $H(Y;\theta)$ is 0. For some initial values θ_0 and Y_0, a general stochastic approximation algorithm is as follows.

Stochastic approximation (SA):

- Generate $Y_t \sim K_{\theta_{t-1}}(Y_{t-1}, \cdot)$, a Markov transition kernel that admits $f(\cdot;\theta_{t-1})$ as the invariant distribution.
- Set $\theta_t = \theta_{t-1} + A_t H(Y_t;\theta_{t-1})$, where A_t is a $r \times r$ matrix, called a gain matrix.

In the Appendix I of Supplementary Materials, we provide a summary of Theorems 1–2 in Song et al. (2014) on the convergence of $\{\theta_t : t \geq 1\}$, with an extension to the case where A_t is a $r \times r$ matrix, similarly as in Corollary 3.3.2 in Chen (2002). In the following, we discuss the relevant results in an informal manner.

Assume that $A_t = \gamma_t A$ for an invertible $r \times r$ matrix A, and $\gamma_t = t_0/t^\beta$, called the gain factor, for $t_0 > 0$ and $1/2 < \beta \leq 1$. Then under certain regularity conditions,

$\gamma_t^{-1/2}(\theta_t - \theta^*)$ converges in distribution to a multivariate normal distribution with mean 0 and variance matrix Σ depending on (t_0, β, A). The maximal rate of variance reduction is reached with $\beta = 1$. Moreover, if $\beta = 1$, then Σ achieves a minimum, $t_0^{-1}C^{-1}VC^{-1^{\mathrm{T}}}$, at $A = t_0^{-1}C^{-1}$, where $C = -\partial h(\theta^*)/\partial\theta^{\mathrm{T}}$ and V is defined in Appendix I in the Supplementary Materials.

For fixed $h(\theta)$ and $f(\cdot;\theta)$, the optimal choice of A_t does not depend on the specification of the "noisy observation" $H(Y;\theta)$ or the transition kernel K_θ, as long as $E_\theta\{H(Y;\theta)\} = h(\theta)$ with $Y \sim f(\cdot;\theta)$ and $f(\cdot;\theta)$ is the invariant distribution under K_θ. The resulting optimal SA recursion is

$$\theta_t = \theta_{t-1} + t^{-1}C^{-1}H(Y_t;\theta_{t-1}), \tag{5}$$

and the minimum asymptotic variance matrix for $t^{1/2}(\theta_t - \theta^*)$, normalized by $t^{1/2}$ instead of $\gamma_t^{-1/2}$, is $C^{-1}VC^{-1^{\mathrm{T}}}$. However, the optimal SA recursion is, in general, infeasible because $C = -\partial h(\theta^*)/\partial\theta^{\mathrm{T}}$ depends on unknown θ^*.

3.2 SA for labeled mixture sampling

The SAMC algorithm is an application of the general SA algorithm to local-jump labeled mixture sampling by the following choices. Let $Y = (L, X)$, $f(y;\theta) = p(j, x;\zeta)$, $\theta = (\zeta_2, \ldots, \zeta_m)^{\mathrm{T}}$, $\theta^* = (\zeta_2^*, \ldots, \zeta_m^*)^{\mathrm{T}}$, with the first element $\zeta_1 = \zeta_1^* = 0$ excluded from ζ and ζ^*, and

$$h(\theta) = \{p(L = 2;\zeta) - \pi_2, \ldots, p(L = m;\zeta) - \pi_m\}^{\mathrm{T}}, \tag{6}$$

$$H(Y;\theta) = (1\{L = 2\} - \pi_2, \ldots, 1\{L = m\} - \pi_m)^{\mathrm{T}}. \tag{7}$$

By Eq. (2), θ^* is a unique solution to $h(\theta) = 0$. Moreover, let

$$K_\theta(y_{t-1}, y_t) = p_{\mathrm{LJ}}(l_t|l_{t-1}, x_{t-1};\zeta)p(x_t|l_t, x_{t-1}), \tag{8}$$

where $p_{\mathrm{LJ}}(l_t|l_{t-1}, x_{t-1};\zeta)$ is the probability density function of L_t given (L_{t-1}, X_{t-1}) under local jump, and $p(x_t|l_t, x_{t-1})$ is the probability density function under the transition kernel $\Phi_{l_t}(x_{t-1}, x_t)$. The sequence of variables generated by the SA algorithm reduce to $Y_t = (L_t, X_t)$ and $\theta_t = (\zeta_2^{(t)}, \ldots, \zeta_m^{(t)})^{\mathrm{T}}$.

We further develop stochastic approximation for local-jump or global-jump labeled mixture sampling, with two alternative choices of $H(Y;\theta)$ and use of the optimal SA recursion (5). First, we show that for $H(Y,\theta)$ defined by (7), the optimal SA recursion (5) is simple and feasible, independent of unknown ζ^*. See the Appendix III of Supplementary Materials for proofs of Theorems 1–3.

Theorem 1. For $H(Y;\theta)$ defined by (7), the optimal SA recursion (5) reduces to $\zeta^{(t)} = \zeta^{(t-\frac{1}{2})} - \zeta_1^{(t-\frac{1}{2})}$ with

$$\zeta^{(t-\frac{1}{2})} = \zeta^{(t-1)} + t^{-1}\{\delta_1(L_t)/\pi_1, \ldots, \delta_m(L_t)/\pi_m\}^{\mathrm{T}}, \tag{9}$$

where $\delta_j(L_t) = 1\{L_t = j\}$ for $j = 1, \ldots, m$.

This result is remarkable because the optimal SA recursion is, in general, infeasible, as mentioned in Section 3.1. Evidently, labeled mixture sampling constitutes a special case where $C = -\partial h(\theta^*)/\partial\theta^{\mathrm{T}}$ is known, even though $\theta^* = (\zeta_2^*, \ldots, \zeta_m^*)^{\mathrm{T}}$ is unknown, so that the optimal SA recursion becomes feasible. A possible explanation is that the Rao–Blackwellized scheme (14) obtained later from (9) matches the offline estimating equations (16) and (17), as discussed in Section 4.

As mentioned in Section 3.1, the SA recursion (5) is optimal regardless of how the transition kernel K_θ is constructed such that $f(\cdot;\theta)$ is the invariant distribution. Therefore, the SA recursion (9) is optimal not only for local-jump labeled mixture sampling with the transition kernel (8), but also for global-jump labeled mixture sampling with the transition kernel

$$K_\theta(y_{t-1}, y_t) = p_{\mathrm{GJ}}(l_t|x_{t-1};\zeta)p(x_t|l_t, x_{t-1}), \tag{10}$$

where $p_{\mathrm{GJ}}(l_t|x_{t-1};\zeta)$ is the probability density function of L_t given (L_{t-1}, X_{t-1}) under global jump. Then $p_{\mathrm{GJ}}(l_t|x_{t-1};\zeta) = p(l_t|x_{t-1};\zeta)$ by Eq. (3).

The SA recursion (5) is also optimal regardless of how the "noisy observation" $H(Y;\theta)$ is specified such that $E_\theta\{H(Y;\theta)\} = h(\theta)$ with $Y \sim f(\cdot;\theta)$. In fact, it is possible to derive two alternative choices of $H(Y;\theta)$ from (7) by taking conditional expectations, known as Rao–Blackwellization (e.g., Gelfand & Smith 1990).

Theorem 2. Redefine

$$H(Y;\theta) = (w_2(X;\zeta) - \pi_2, \ldots, w_m(X;\zeta) - \pi_m)^{\mathrm{T}}, \tag{11}$$

where $w_j(X;\zeta) = p(L = j|X;\zeta)$ in Eq. (3). Then $h(\theta) = E_\theta\{H(Y;\theta)\}$ with $Y \sim f(\cdot;\theta)$ for each θ. The optimal SA recursion (5) reduces to $\zeta^{(t)} = \zeta^{(t-\frac{1}{2})} - \zeta_1^{(t-\frac{1}{2})}$ with

$$\zeta^{(t-\frac{1}{2})} = \zeta^{(t-1)} + t^{-1}\{w_1(X_t;\zeta^{(t-1)})/\pi_1, \ldots, w_m(X_t;\zeta^{(t-1)})/\pi_m\}^{\mathrm{T}}. \tag{12}$$

The choice (11) for $H(Y;\theta)$ is a conditional expectation (or Rao–Blackwellization) of the earlier choice (7). Equation $h(\theta) = E_\theta\{H(Y;\theta)\}$ holds because $w_j(X;\zeta) =$

$E(1\{L = j\}|X;\zeta)$ by definition and hence $E\{w_j(X;\zeta)\} = p(L = j;\zeta)$ by the rule of iterated expectations, where $(L, X) \sim p(j, x;\zeta)$. Moreover, the corresponding optimal SA recursion (12) is similar to (9), with $\delta_j(L_t)$ replaced by $w_j(X_t;\zeta^{(t-1)})$.

The Rao–Blackwellization is performed above directly with respect to the invariant distribution $p(j, x;\zeta)$ for each fixed ζ. Alternatively, it is more informative to perform Rao–Blackwellization with respect to a Markov transition kernel. For transition kernel (10) under global jump, Rao–Blackwellization gives

$$E(1\{L_t = j\}|L_{t-1}, X_{t-1};\zeta) = E(1\{L_t = j\}|X_{t-1};\zeta) = w_j(X_{t-1};\zeta),$$

leading again to the choice (11) and the update scheme (12). On the other hand, Rao–Blackwellization under local jump with transition kernel (8) gives

$$E(1\{L_t = j\}|L_{t-1}, X_{t-1};\zeta) = u_j(L_{t-1}, X_{t-1};\zeta),$$

where

$$u_j(L, X;\zeta) = \begin{cases} \Gamma(L, j) \min\left\{1, \frac{\Gamma(j,L)}{\Gamma(L,j)} \frac{p(j|X;\zeta)}{p(L|X;\zeta)}\right\}, & \text{if } j \in \mathcal{N}(L), \\ 1 - \sum_{l\in\mathcal{N}(L)} u_l(L, X;\zeta), & \text{if } j = L, \end{cases}$$

and $u_j(L, X;\zeta) = 0$ if $j \notin \{L\} \cup \mathcal{N}(L)$. This leads to a new choice for $H(Y;\theta)$, different from (7) or (11), and the following result.

Theorem 3. Redefine

$$H(Y;\theta) = (u_2(L, X;\zeta) - \pi_2, \ldots, u_m(L, X;\zeta) - \pi_m)^{\mathrm{T}}, \tag{13}$$

where $u_j(L, X;\zeta)$ is defined as above. Then $h(\theta) = E_\theta\{H(Y;\theta)\}$ with $Y \sim f(\cdot;\theta)$ for each θ. The optimal SA recursion (5) reduces to $\zeta^{(t)} = \zeta^{(t-\frac{1}{2})} - \zeta_1^{(t-\frac{1}{2})}$ with

$$\zeta^{(t-\frac{1}{2})} = \zeta^{(t-1)} + t^{-1}\{u_1(L_t, X_t;\zeta^{(t-1)})/\pi_1, \ldots, u_m(L_t, X_t;\zeta^{(t-1)})/\pi_m\}^{\mathrm{T}}. \tag{14}$$

As a summary, there are two choices of transition kernel K_θ and three choices of "noisy observation" $H(Y;\theta)$. Our development gives a class of SA algorithms for labeled mixture sampling, which we call stochastic approximation mixture sampling or, more descriptively, self-adjusted mixture sampling.

Self-adjusted mixture sampling:

- *Labeled mixture sampling*: Generate (L_t, X_t) from transition kernel (8) under local jump or from (10) under global jump, with ζ set to $\zeta^{(t-1)}$.
- *Free energy update*: Compute $\zeta^{(t)}$ by (9), (12), or (14), referred to as a binary, global, or local update scheme respectively.

In principle, each of the update schemes (9), (12), and (14) can be combined with either local-jump or global-jump mixture sampling. For example, the update scheme (12), although derived by Rao–Blackwellization under the local-jump transition kernel, can be used when (L_t, X_t) are generated by global-jump mixture sampling. In practice, these choices should be decided from considerations of both statistical efficiency and computational cost. The local-jump and global-jump mixture sampling are briefly compared in Section 2. See Appendix II of Supplementary Materials for comparisons between the update schemes (9), (12), and (14), including a theoretical result showing that the global update scheme (12) is statistically more efficient than the binary scheme (9), when both used with global-jump mixture sampling.

Finally, we provide several remarks on implementation issues of self-adjusted mixture sampling. First, the initial choice $\zeta^{(0)}$ can be set as naively as to the vector of zeros, as done in all our simulation studies. Second, convergence of $\zeta^{(t)}$ to the target ζ^* can be slow if the initial value $\zeta^{(0)}$ is far away from ζ^* and if the dimension m is large, as in the numerical examples in Section 6.2 and Appendix V. In general, there are differences between the rate of convergence to stationarity and statistical efficiency, determined by the amount of random fluctuation once in stationarity (e.g, Liu 2001, Section 13.3.2). To overcome this issue, we suggest a two-stage implementation of self-adjusted mixture sampling by replacing the gain factor t^{-1} in the update scheme (9), (12), or (14), with the diagonal matrix with the jth diagonal element

$$\begin{cases} \min(\pi_j, t^{-\beta}), & \text{if } t \le t_0, \\ \min\{\pi_j, (t - t_0 + t_0^{\beta})^{-1}\}, & \text{if } t > t_0, \end{cases} \tag{15}$$

where $1/2 < \beta < 1$ and t_0 is the burn-in size. For example, β is set to 0.6 or 0.8, and t_0 is set such that the proportions of $L_t = j$ are within 50% – 20% of π_j in our numerical work. The minimum with π_j is taken to ensure that the adjustment term in the resulting scheme (9), (12), or (14) is never greater than 1 for all $t \ge 1$, even when m is large and some π_j is small. A slow-decaying gain factor $t^{-\beta}$ is used in the first stage, to introduce larger adjustments than with the factor t^{-1} and hence force $\zeta^{(t)}$ to fall faster into a neighborhood of ζ^*. See Gu & Zhu (2001) for a related two-stage SA algorithm, but a slow-decaying factor $t^{-\beta}$ is used at both stages, with β close to 0 or to 1/2 at the first or second stage respectively.

4 Offline estimation

Stochastic approximation (or self-adjusted) mixture sampling, after n iterations, provides not only a consistent estimator $\zeta^{(n)}$ for free energies, but also a sequence of draws $\{(L_i, X_i) : i = 1, \ldots, n\}$, which are expected to be ergodic with respect to the joint distribution $p(j, x; \zeta^*)$. The ergodicity of the pairs (L_i, X_i) can be decomposed into that of the labels L_i and that of the observations, $S_j = \{X_i : L_i = j,\, i = 1, \ldots, n\}$, with label j. For ease of discussion, assume that for $j = 1, \ldots, m$,

(A1) the observed proportion $\hat{\pi}_j = n_j/n$ converges to the target π_j almost surely, and $\alpha_n(\hat{\pi}_j - \pi_j)$ converges to a non-degenerate distribution, where n_j is the size of S_j, $\alpha_n \to \infty$ and possibly $n^{-1/2}\alpha_n \to 0$ as $n \to \infty$, and

(A2) the sample average $\tilde{E}_j(\phi) = n_j^{-1} \sum_{1 \le i \le n : L_i = j} \phi(X_i)$ converges to $E_j(\phi)$ almost surely, and $n_j^{1/2}\{\tilde{E}_j(\phi) - E_j(\phi)\}$ converges to a non-degenerate distribution.

Then S_j forms an approximate sample from P_j. By (A1) and (A2) jointly, the pooled sample $(X_1, \ldots, X_n)$ forms an approximate sample from the mixture $p(x; \zeta^*)$, although the convergence rate of empirical averages might be slower than $n^{-1/2}$.

Asymptotic theory justifying the almost-sure convergence of $\hat{\pi}_j$ and $\tilde{E}_j(\phi)$ can be obtained from, for example, Liang et al. (2015). However, formal results remain to be developed for central limit theorems. We *postulate* in (A2) that the average $\tilde{E}_j(\phi)$ converges at the usual rate $n^{-1/2}$, because for any fixed ζ, the conditional distribution of X given $L = j$ is always P_j, where (L, X) is drawn from the invariant distribution $p(j, x; \zeta)$. This is similar to related theory on adaptive MCMC in Andrieu & Moulines (2006), where the invariant distribution $f(\cdot; \theta)$ does not depend on the choice of θ, and the central limit theorem holds at the rate $n^{-1/2}$ for empirical averages.

The preceding properties can be exploited to construct offline estimators of free energies ζ^*, different from the SA estimator $\zeta^{(n)}$ or the average $\bar{\zeta}^{(n)} = n^{-1} \sum_{i=1}^{n} \zeta_i$. Throughout, an estimator is said to be online if, after n iterations, it can be determined from the estimator after $n-1$ iterations and the variables generated at the nth iteration. An estimator is said to be offline if it is not online. The estimator $\bar{\zeta}^{(n)}$ can be considered online jointly with $\zeta^{(n)}$, because $\bar{\zeta}^{(n)} = \bar{\zeta}^{(n-1)} + n^{-1}(\zeta^{(n)} - \bar{\zeta}^{(n-1)})$. Roughly, an online estimator is sequentially updated during a sampling process, whereas an offline estimator is computed after the sampling process is completed.

The development of the global choice (11) of $H(Y;\theta)$ for stochastic approximation shows that ζ^* satisfies $E_{\zeta^*}\{w_j(X;\zeta^*)\} = \pi_j$ for $j = 1,\ldots,m$, where $X \sim p(x;\zeta^*)$. This relationship and the fact that $(X_1,\ldots,X_n)$ forms an approximate sample from $p(x;\zeta^*)$ lead to the following estimator for ζ^*. Let $\tilde{\zeta}^{(n)} = (\tilde{\zeta}_1^{(n)},\ldots,\tilde{\zeta}_m^{(n)})^{\mathrm{T}}$ with $\tilde{\zeta}_1^{(1)} = 0$ be a solution to $n^{-1}\sum_{i=1}^n w_j(X_i;\zeta) = \pi_j$ or equivalently

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\mathrm{e}^{-\zeta_j} q_j(X_i)}{\sum_{l=1}^m \pi_l \mathrm{e}^{-\zeta_l} q_l(X_i)} = 1, \quad j = 1,2,\ldots,m. \tag{16}$$

The sums of both sides of (16) multiplied by π_j over $j = 1,2,\ldots,m$ are equal to 1. Therefore, Eq. (16) needs only to be solved for $j = 2,\ldots,m$. Remarkably, the ith term in the summation in (16) corresponds to the ratio $w_j(X_i;\zeta)/\pi_j$, which appears exactly in the optimal SA recursion (12).

The form of (16) is reminiscent of a related offline method for estimating free energies. Under Assumption (A2), the set of observations, S_j, with the same label j forms a proper sample of size n_j from P_j . Then $(X1,\ldots,X_n)$ can be treated as a sample from the stratified density $\sum_{j=1}^m \hat{\pi}_j \mathrm{e}^{-\zeta_j^*} q_j(x)$, where the observed proportion $\hat{\pi}_j$ is used instead of the target weight π_j. Replacing π_j by $\hat{\pi}_j$ in (16) yields the following estimator. Let $\hat{\zeta}^{(n)} = (\hat{\zeta}_1^{(n)},\ldots,\hat{\zeta}_m^{(n)})^{\mathrm{T}}$ with $\hat{\zeta}_1^{(n)} = 0$ be a solution to

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\mathrm{e}^{-\zeta_j} q_j(X_i)}{\sum_{l=1}^m \hat{\pi}_l \mathrm{e}^{-\zeta_l} q_l(X_i)} = 1, \quad j = 1,2,\ldots,m. \tag{17}$$

We refer to $\tilde{\zeta}^{(n)}$ or $\hat{\zeta}^{(n)}$ as the unstratified or stratified estimator. The estimator $\tilde{\zeta}^{(n)}$ relies on the fact that $(X_1,\ldots,X_n)$ is a proper sample from $p(x;\zeta^*)$. In contrast, $\hat{\zeta}^{(n)}$ is based on the fact that S_j is a proper sample form P_j, and would remain consistent even if $\hat{\pi}_j$ did not converge to π_j, for $j = 1,\ldots,m$. As mentioned in Section 1, the estimator $\hat{\zeta}^{(n)}$ has been widely used in physics and statistics (e.g., Kong et al. 2003; Tan et al. 2012). Our development adds a new understanding of the methodology, by making connections from the stochastic approximation schemes (9) and (12) to the estimators $\tilde{\zeta}^{(n)}$ and $\hat{\zeta}^{(n)}$ through Rao–Blackwellization and stratification.

An important feature of the existing methodology behind (17) is that the baseline measure μ is estimated by a discrete measure $\hat{\mu}$, which is supported on the pooled sample $(X_1,\ldots,X_n)$ with weights determined up to a positive multiple by

$$\hat{\mu}(\{X_i\}) \propto \left\{ \sum_{l=1}^{m} n_l \, \mathrm{e}^{-\hat{\zeta}_l^{(n)}} q_l(X_i) \right\}, \quad i = 1,\ldots,n.$$

For $j = 1, \ldots, m$, the free energy ζ_j^* is estimated by $\exp(\hat{\zeta}_j^{(n)}) = \int q_j(x)\, \mathrm{d}\hat{\mu}$ as in (17). For an unsampled distribution P_0, the free energy ζ_0^* is estimated by

$$\mathrm{e}^{\hat{\zeta}_0^{(n)}} = \sum_{i=1}^{n} \frac{q_0(X_i)}{\sum_{l=1}^{m} n_l \mathrm{e}^{-\hat{\zeta}_l^{(n)}} q_l(X_i)}.$$

The expectation $E_j(\phi) = \int \phi(x)\, \mathrm{d}P_j$ is estimated by

$$\hat{E}_j(\phi) = \sum_{i=1}^{n} \phi(X_i) \frac{\mathrm{e}^{-\hat{\zeta}_j^{(n)}} q_j(X_i)}{\sum_{l=1}^{m} n_l \mathrm{e}^{-\hat{\zeta}_l^{(n)}} q_l(X_i)}, \quad j = 0, 1, \ldots, m.$$

This estimator $\hat{E}_j(\phi)$, unlike the sample average $\tilde{E}_j(\phi)$, depends on the pooled sample $(X_1, \ldots, X_n)$ and is applicable even for $j = 0$.

In Appendix II, we provide additional theoretical results on comparison of statistical efficiency between the online estimator $\zeta^{(n)}$ and offline estimators $\tilde{\zeta}^{(n)}$ and $\hat{\zeta}^{(n)}$. The unstratified and stratified estimators differ from dynamically weighted estimators, adapted from partition-based settings (Liang 2009; Liang et al. 2015). See Appendices IV–VI in Supplementary Materials for further discussion.

5 Locally weighted histogram analysis

The offline estimator $\hat{\zeta}^{(n)}$ is known to be statistically efficient, at least in the case where $(X_1, \ldots, X_n)$ are independent (Tan 2004). However, the estimator $\hat{\zeta}^{(n)}$ requires evaluating m unnormalized densities $q_1(X_i), \ldots, q_m(X_i)$ for each X_i, similarly to the global-jump sampling scheme (10) and the global update scheme (12). Such computational cost can outweigh efficiency gains, compared with, for example, the SA estimator $\zeta^{(n)}$ obtained by the binary update scheme (9) under self-adjusted local-jump mixture sampling. In this section, we propose a local method for offline estimation, to reduce computational cost while preserving statistical efficiency.

First, we derive a local unstratified estimator for ζ^*, corresponding to the global unstratified estimator solved from (16). By Theorem 3 on the local choice (13) of $H(Y; \theta)$ for stochastic approximation, ζ^* satisfies $E_{\zeta^*}\{u_j(L, X; \zeta^*)\} = \pi_j$ for $j = 1, \ldots, n$, where $(L, X) \sim p(j, x; \zeta^*)$. This relationship and the fact that $\{(L_i, X_i) : i = 1, \ldots, n\}$ forms an approximate sample from $p(j, x; \zeta)$ suggests the following estimator. Let $\tilde{\zeta}^{(n)} = (\tilde{\zeta}_2^{(n)}, \ldots, \tilde{\zeta}_m^{(n)})^{\mathrm{T}}$ with $\tilde{\zeta}_1^{(n)} = 0$ be a solution to $n^{-1} \sum_{i=1}^{n} u_j(L_i, X_i; \zeta) = \pi_j$ for $j = 1, \ldots, m$. However, $u_j(L, X; \zeta)$ is not everywhere differentiable in ζ and hence computing $\tilde{\zeta}^{(n)}$ can be complicated. To address this issue, we replace the Metropolis–Hastings acceptance probability, $\min[1, \{\Gamma(j, L)p(j|X; \zeta)\}/\{\Gamma(L, j)p(L|X; \zeta)\}]$, by

Barker's (1965) acceptance probability, $\{\Gamma(j,L)p(j|X;\zeta)\}/\{\Gamma(L,j)p(L|X;\zeta)+\Gamma(j,L)p(j|X;\zeta)\}$, and redefine $u_j(L,X;\zeta)$ as

$$u_j(L,X;\zeta)=\begin{cases}\Gamma(L,j)\frac{\Gamma(j,L)p(j|X;\zeta)}{\Gamma(L,j)p(L|X;\zeta)+\Gamma(j,L)p(j|X;\zeta)}, & \text{if } j\in\mathcal{N}(L),\\ 1-\sum_{l\in\mathcal{N}(L)}u_l(L,X;\zeta), & \text{if } j=L,\end{cases}$$

and $u_j(L,X;\zeta)=0$ if $j\notin\{L\}\cup\mathcal{N}(L)$. Then Theorem 3 is easily shown to remain valid with redefined $u_j(L,X;\zeta)$, because Barker's (1965) acceptance probability ensures detailed balance (Liu 2001, Section 5.2).

There are several consequences of using Barker's (1965) acceptance probability instead of the Metropolis–Hastings acceptance probability. By direct calculation, the equation $n^{-1}\sum_{i=1}^n u_j(L_i,X_i;\zeta)=\pi_j$ for $\tilde{\zeta}^{(n)}$ can be equivalently expressed as

$$\frac{1}{n}\sum_{i=1}^n\sum_{l\in\mathcal{N}(j)}\Gamma(j,l)\left[\frac{1\{L_i=l\}\Gamma(l,j)\mathrm{e}^{-\zeta_j}q_j(X_i)}{\Gamma(l,j)\pi_l\mathrm{e}^{-\zeta_l}q_l(X_i)+\Gamma(j,l)\pi_j\mathrm{e}^{-\zeta_j}q_j(X_i)}\right.$$
$$\left.+\frac{1\{L_i=j\}\Gamma(j,l)\mathrm{e}^{-\zeta_j}q_j(X_i)}{\Gamma(l,j)\pi_l\mathrm{e}^{-\zeta_l}q_l(X_i)+\Gamma(j,l)\pi_j\mathrm{e}^{-\zeta_j}q_j(X_i)}\right]=1. \qquad (18)$$

Moreover, $\tilde{\zeta}^{(n)}$ solved from (18) can be shown to, equivalently, minimize the function

$$\kappa(\zeta)=\frac{1}{n}\sum_{i=1}^n\sum_{j\in\mathcal{N}(L_i)}\Gamma(L_i,j)\log\left\{\Gamma(j,L_i)\frac{\pi_j q_j(X_i)}{\mathrm{e}^{\zeta_j}}+\Gamma(L_i,j)\frac{\pi_{L_i}q_{L_i}(X_i)}{\mathrm{e}^{\zeta_{L_i}}}\right\}+\sum_{j=1}^m\pi_j\zeta_j,$$

which is convex and twice differentiable in ζ. Therefore, $\tilde{\zeta}^{(n)}$ can be computed effectively by using globally convergent optimization algorithms. This is similar to the fact that the global unstratified or stratified estimator, based on (16) or (17), can be equivalently obtained by convex minimization (Tan et al. 2012).

Eq. (18) can also be seen as combining estimating equations over all pairs of neighboring samples by Bennett's acceptance ratio method or two-sample bridge sampling (Meng & Wong 1996). In fact, if $m=2$, then Eq. (18) with $j=2$ yields

$$\frac{1}{n}\sum_{i=1}^n\left[\frac{1\{L_i=1\}\Gamma(1,2)\mathrm{e}^{-\zeta_2}q_2(X_i)}{\Gamma(1,2)\pi_1q_1(X_i)+\Gamma(2,1)\pi_2\mathrm{e}^{-\zeta_2}q_2(X_i)}\right.$$
$$\left.+\frac{1\{L_i=2\}\Gamma(2,1)\mathrm{e}^{-\zeta_2}q_2(X_i)}{\Gamma(1,2)\pi_1q_1(X_i)+\Gamma(2,1)\pi_2\mathrm{e}^{-\zeta_2}q_2(X_i)}\right]=1. \qquad (19)$$

Eq. (19) is somehow more general in allowing $\Gamma(1,2)\neq\Gamma(2,1)$ than the global estimating equation (16) with $m=2$, i.e.,

$$\frac{1}{n}\sum_{i=1}^n\frac{\mathrm{e}^{-\zeta_2}q_2(X_i)}{\pi_1q_1(X_i)+\pi_2\mathrm{e}^{-\zeta_2}q_2(X_i)}=1.$$

For a general m, Eq. (18) is a weighted average with weight $\Gamma(j,l)$ over $l \in \mathcal{N}(j)$ of equations in the form (19), depending on only the two samples S_j and S_l. In other words, S_j is pooled separately with S_l for $l \in \mathcal{N}(j)$ to obtain a two-sample estimating equation in the form (19). Then such two-sample equations with $l \in \mathcal{N}(j)$ are linearly combined with weights $\Gamma(j,l)$ to yield Eq. (18), in a dynamic manner determined by Rao–Blackwellization in labeled mixture sampling.

Next, we propose a local stratified estimator for ζ^*, corresponding to the global stratified estimator solved from (17). Let $\hat{\zeta}^{(n)} = (\hat{\zeta}_1^{(n)}, \ldots, \hat{\zeta}_m^{(n)})^{\mathrm{T}}$ with $\hat{\zeta}_1^{(n)} = 0$ be a solution to (18) or, equivalently, be a minimizer to $\kappa(\zeta)$, with $(\pi_1, \ldots, \pi_m)$ replaced by $(\hat{\pi}_1, \ldots, \hat{\pi}_m)$. The effect of such stratification can be seen as follows. With (π_1, π_2) replaced by $(\hat{\pi}_1, \hat{\pi}_2)$, the estimating equation (19) would remain asymptotically unbiased provided that (S_1, S_2) are proper samples from (P_1, P_2) respectively, even if $(\hat{\pi}_1, \hat{\pi}_2)$ converged to some constants different from (π_1, π_2). Therefore, similarly as in global estimation, the validity of $\hat{\zeta}^{(n)}$ requires that S_j is a proper sample from P_j for $j = 1, \ldots, m$, but not that $\{(L_i, X_i) : i = 1, \ldots, n\}$ is a proper sample from $p(j, x; \zeta)$. The stratified estimator $\hat{\zeta}^{(n)}$ is more robust than the unstratified one $\tilde{\zeta}^{(n)}$ to random deviations of $(\hat{\pi}_1, \ldots, \hat{\pi}_m)$ from $(\pi_1, \ldots, \pi_m)$.

The local method can be recast and used for estimating free energies and expectations, from the perspective of estimating the baseline measure (Kong et al. 2003; Tan et al. 2012). By the stratified version of (18), $\hat{\zeta}_j^{(n)}$ can be equivalently expressed as $\exp(\hat{\zeta}_j^{(n)}) = \int q_j(x)\, \mathrm{d}\hat{\mu}_j(x)$, where $\hat{\mu}_j$ is a discrete measure supported on the locally pooled sample, $S_j \cup (\cup_{l \in \mathcal{N}(j)} S_l)$, from P_j and its neighboring distributions $\{P_l : l \in \mathcal{N}(j)\}$, with weights determined by

$$\hat{\mu}_j(\{X_i\}) \propto \frac{1}{n} \sum_{l \in \mathcal{N}(j)} \Gamma(j,l) \left[\frac{1\{L_i = l\}\Gamma(l,j)}{\Gamma(l,j)\hat{\pi}_l \mathrm{e}^{-\hat{\zeta}_l^{(n)}} q_l(X_i) + \Gamma(j,l)\hat{\pi}_j \mathrm{e}^{-\hat{\zeta}_j^{(n)}} q_j(X_i)} + \frac{1\{L_i = j\}\Gamma(j,l)}{\Gamma(l,j)\hat{\pi}_l \mathrm{e}^{-\hat{\zeta}_l^{(n)}} q_l(X_i) + \Gamma(j,l)\hat{\pi}_j \mathrm{e}^{-\hat{\zeta}_j^{(n)}} q_j(X_i)} \right].$$

In contrast with global estimation, the baseline measure μ is estimated by $\hat{\mu}_j$, supported on a different subset of simulated data, depending on which free energy ζ_j^* is computed. Nevertheless, integrals of interest can be estimated by substituting $\hat{\mu}_j$ for μ with a suitable choice j, similarly as in global estimation. The free energy ζ_0^* for an unsampled distribution P_0 can be estimated by setting $\exp(\hat{\zeta}_0^{(n)}) = \int q_0(x)\, \mathrm{d}\hat{\mu}_0$, where $\hat{\mu}_0 = \hat{\mu}_{j_0}$ for some $1 \le j_0 \le m$ chosen such that P_0 is considered close to P_{j_0} and its

neighboring distributions $\{P_l : l \in \mathcal{N}(j_0)\}$. The expectation $E_j(\phi) = \int \phi(x)\, \mathrm{d}P_j$ can be estimated by $\int \phi(x) \exp(-\hat{\zeta}_j^{(n)}) q_j(x)\, \mathrm{d}\hat{\mu}_j$ for $j = 1, \ldots, m$ and $j = 0$.

To highlight the fact that the baseline measure is estimated using locally pooled samples, we refer to the local method as locally weighted histogram analysis (L-WHAM), in parallel to globally weighted histogram analysis (Tan et al. 2012). By design, the local method is computationally far less costly than the global method, which can be impractical for a large m. The local stratified estimator $\hat{\zeta}^{(n)}$ requires evaluating only $\{1 + s(L_i)\}$ un-normalized densities $\{q_j(X_i) : j = L_i \text{ or } j \in \mathcal{N}(L_i)\}$, which are the same as needed by the local update scheme (14). Moreover, statistical efficiency of the local method can be similar to that of the global method, because the accuracy of estimating free energies ζ^* is, to a large extent, affected by the degree of overlaps between the distributions $(P_1, \ldots, P_m)$ (e.g., Meng & Wong 1996), and each distribution P_j typically overlaps more with the neighboring distributions $\{P_l : l \in \mathcal{N}(j)\}$ than with other distributions. See Tan (2013a, 2013b) for related local methods, where individual samples are grouped into clusters and then global estimators are combined from different clusters in a static manner.

The global and local methods are discussed above for offline estimation when self-adjusted mixture sampling is used. However, these methods are broadly applicable with other sampling algorithms. Similarly as in Geyer (1994) and Tan (2004), the global or local stratified estimator $\hat{\zeta}^{(n)}$ can be shown to be valid under suitable conditions on the supports of $(P_1, \ldots, P_m)$, provided that S_j is a proper sample from P_j, satisfying usual asymptotic properties as in Assumption (A2), for $j = 1, \ldots, m$. For example, the samples from $(P_1, \ldots, P_m)$ can be simulated, with pre-specified sample sizes, by running m Markov chain simulations independently, parallel tempering (Geyer 1991), or resampling MCMC (Tan 2015) including resample-move (Gilks & Berzuini 2001) and equi-energy sampling (Kou et al. 2006).

6 Simulation studies

6.1 Potts model: Canonical ensemble simulation

The Potts model is important in statistical physics, with various applications. We study canonical ensemble simulation in this section, and generalized ensemble simulation based on partitioning of the state space in Appendix V of Supplementary

Materials. Consider a 10-state Potts model on a 20×20 lattice with periodic boundary conditions in the absence of a magnetic field. Each observation x corresponds to a collection of $K = 20^2$ spins $(s_1, \ldots, s_K)$ on the lattice, where s_i takes $q = 10$ possible values. At a temperature T, the density function of the Potts distribution is $Z^{-1}\mathrm{e}^{-u(x)/T}$, where $u(x) = -\sum_{i\sim j} 1\{s_i = s_j\}$, with $i \sim j$ indicating that sites i and j are nearest neighbors, and $Z = \sum_x \exp\{-u(x)/T\}$ is the normalizing constant. Statistically, the Potts distribution belongs to an exponential family, with canonical statistic $-u(x)$ and natural parameter $\theta = T^{-1}$. Let $U = E\{u(x)\}$ and $C = \mathrm{var}\{u(x)\}$ under the Potts distribution. For simplicity, the dependency of Z, U, and C on θ is suppressed in the notation. Then $U = -(\mathrm{d}/\mathrm{d}\theta)\log Z$ and $C = (\mathrm{d}^2/\mathrm{d}\theta^2)\log Z$ by theory of exponential family. In statistical physics, Z is called the partition function, U is internal energy, and C/T^2 is specific heat (Newman & Barkema 1999).

A special case of the Potts model with two states ($q = 2$) is equivalent to the Ising model, where $u(x) = -\sum_{i\sim j} s_i s_j$ and each s_i is either -1 or 1. Like the Ising model, the Potts model on an infinite lattice exhibits a phase transition at the inverse temperature $\theta_c = T_c^{-1} = \log(1+\sqrt{q})$, about 1.426 for $q = 10$. But the critical behavior is richer and more general than that of the Ising model (Wu 1982). For example, the histograms of $u(x)$, known as the energy histograms, are bimodal near the critical temperature T_c, as shown later in Figure 1. In contrast, the energy histograms are unimodal, centered at different locations for different temperatures under the Ising model (e.g., Newman & Barkema 1999, Figure 8.3).

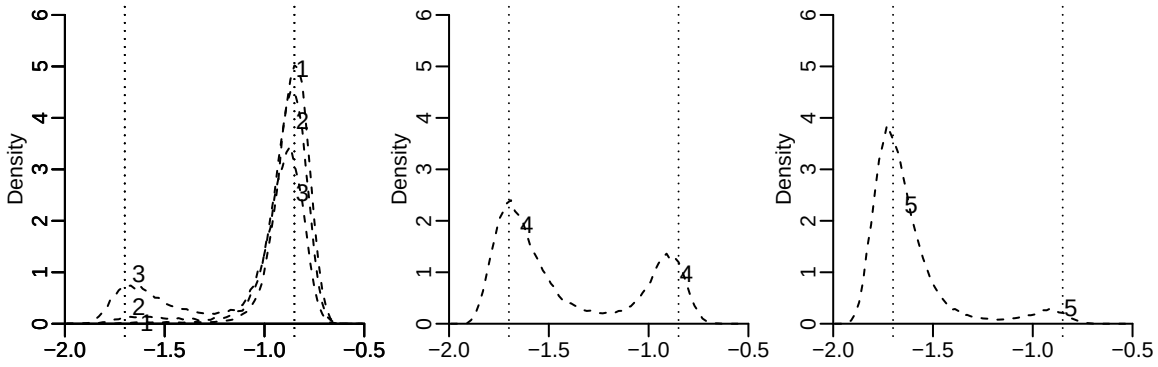
6.1.1 Simulation details

For $m = 5$, we take $(P_1, \ldots, P_5)$ as the Potts distributions at inverse temperatures $(T_1^{-1}, \ldots, T_5^{-1}) = (1.4, 1.4065, 1.413, 1.4195, 1.426)$, evenly spaced between 1.4 and 1.426. The Markov transition kernel Ψ_j for P_j is defined as a random-scan sweep using the single-spin-flip Metropolis algorithm at temperature T_j (Newman & Barkema 1999, Section 4.5.1). Each sweep consists of K iterations, where each iteration involves randomly picking a spin s_i, choosing a new value from the $q - 1$ remaining values, and then accepting or rejecting the new value by the Metropolis rule.

We compare the following four algorithms, where the neighborhood $\mathcal{N}(j)$ is defined as $\{1 \le l \le 5 : l = j - 1 \text{ or } j + 1\}$, of size 1 or 2.

- Parallel tempering (Geyer 1991), implemented as in Tan (2015) to ensure that

Figure 1: Histogram of $u(x)/K$ at the temperatures $(T_1, T_2, \ldots, T_5)$ labeled as $1, 2, \ldots, 5$ under the Potts model. Two vertical lines are placed at -1.7 and -0.85.



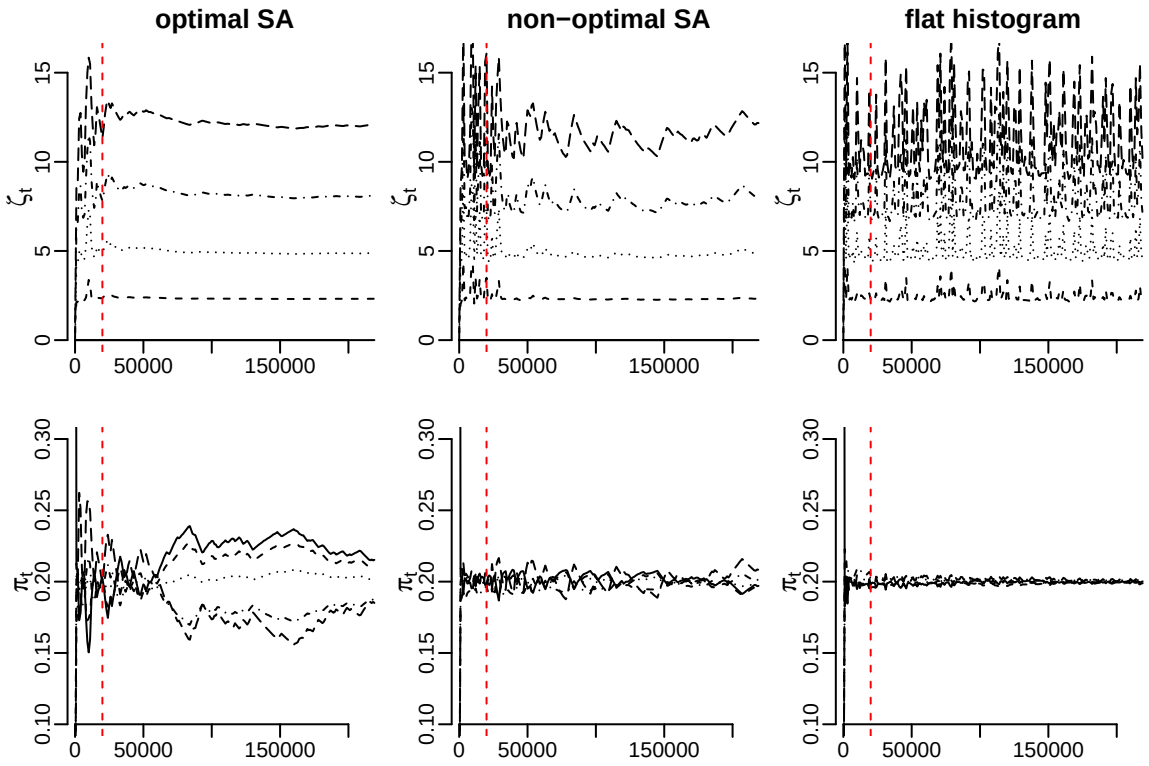
there is a Markov move per iteration and, on average, an exchange attempt per Markov move, similarly as in self-adjusted mixture sampling.

- Self-adjusted local-jump mixture sampling, with two-stage modification (15) of the optimal binary update scheme (9), where β is set to 0.8.
- Self-adjusted local-jump mixture sampling, with the gain factor t^{-1} in (9) replaced in two stages by $\min(1/m, 10/t^{\beta})$ if $t \le t_0$ or by $\min\{1/m, 10/(t-t_0+t_0^{\beta})\}$ if $t > t_0$, with $\beta = 0.8$. This is comparable to (4) with $\gamma_t = t_0/\max(t_0, t)$ and $t_0 = 10m$ for the SAMC algorithm (Liang et al. 2007).
- Self-adjusted local-jump mixture sampling, with the gain factor t^{-1} in (9) replaced by $1/(mk_t)$, where k_t increases by 1 only when a flat-histogram criterion is met: the observed proportions of labels, $L_t = j$, are all within 20% of the target $1/m$ since the last time the criterion was met (e.g., Atchadé & Liu 2010).

See Tan (2015) for a simulation study in the same setup of Potts distributions, where parallel tempering was found to perform better than several resampling MCMC algorithms, including resample-move and equi-energy sampling.

The initial value L_0 is set to 1, corresponding to temperature T_1, and X_0 is generated by randomly setting each spin. The same X_0 is used in parallel tempering for each of the 5 chains. For parallel tempering, the total number of iterations is set to 4.4×10^5 per chain, with the first 4×10^4 iterations as burn-in. For self-adjusted mixture sampling with comparable cost, the total number of iterations is set to 2.2×10^6, with the first 2×10^5 treated as burn-in. The data are recorded (or subsampled) every 10 iterations, yielding 5 chains each of length 4.4×10^4 with the first 4×10^3 iterations

Figure 2: Trace plots for self-adjusted mixture sampling with $t_0 = 2 \times 10^5$. The number of iterations is shown after subsampling. A vertical line is placed at the burn-in size.



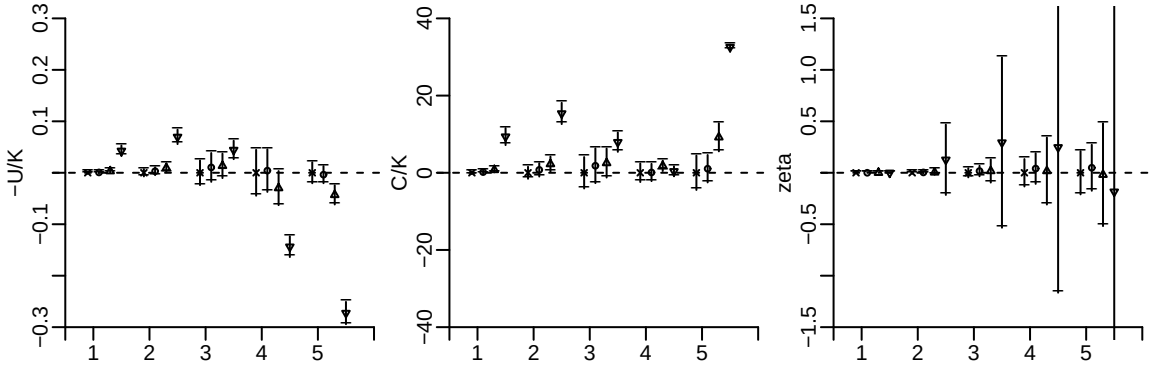
as burn-in for parallel tempering, and a single chain of length 2.2×10^5 with the first 2×10^4 iterations as burn-in for self-adjusted mixture sampling.

6.1.2 Simulation results

Figure 1 shows the histograms of $u(x)/K$ at the 5 temperatures, based on a single run of optimally adjusted mixture sampling with t_0 set to 2×10^5 (the burn-in size before subsampling). There are two modes in these energy histograms. As the temperature decreases from T_1 to T_5, the mode located at about -1.7 grows in its weight, from being a negligible one, to a minor one, and eventually to a major one, so that the spin system moves from the disordered phase to the ordered one.

Figure 2 shows the trace plots of free energy estimates $\zeta^{(t)}$ and observed proportions $\hat{\pi}$ for three algorithms of self-adjusted mixture sampling. There are striking differences between these algorithms. For optimally adjusted mixture sampling, the free energy estimates $\zeta^{(t)}$ fall quickly toward the truth (as indicated by the final estimates) in the first stage, with large fluctuations due to the gain factor of order $t^{-0.8}$. The estimates stay stable in the second stage, due to the optimal gain factor of order t^{-1}. The observed proportions $\hat{\pi}_j$ also fall quickly toward the target $\pi_j = 20\%$. But there are

Figure 3: Summary of estimates at the temperatures $(T1, \ldots, T_5)$ labeled as $1, \ldots, 5$, based on 100 repeated simulations. For each vertical bar, the center indicates Monte Carlo mean minus that obtained from parallel tempering ($\times$: parallel tempering, $\circ$: optimal SA scheme, $\triangle$: non-optimal SA scheme, ∇: flat-histogram scheme), and the radius indicates Monte Carlo standard deviation of the 100 estimates from repeated simulations. For $-U/K$ and C/K, all the estimates are directly based on sample averages $\tilde{E}_j(\phi)$. For free energies, offline estimates $\tilde{\zeta}_j^{(n)}$ are shown for parallel tempering ($\times$), whereas online estimates $\zeta_j^{(n)}$ are shown for self-adjusted mixture sampling ($\circ$, $\triangle$, or ∇).



considerable deviations of $\hat{\pi}_j$ from π_j over time, which reflects the presence of strong autocorrelations in the label sequence L_t.

For the second algorithm, the use of a gain factor about 10 times the optimal one forces the observed proportions $\hat{\pi}_j$ to stay closer to the target ones, but leads to greater fluctuations in the free energy estimates $\zeta^{(t)}$ than when the optimal SA scheme is used. For the third algorithm using the flat-histogram adaptive scheme, the observed proportions $\hat{\pi}_j$ are forced to be even closer to π_j, and the free energy estimates $\zeta^{(t)}$ are associated with even greater fluctuations than when the optimal SA scheme. These non-optimal algorithms seem to control the observed proportions $\hat{\pi}_j$ tightly about π_j, but potentially increase variances for free energy estimates and, as shown below, introduce biases for estimates of expectations.

Figure 3 shows the Monte Carlo means and standard deviations for the estimates of $-U/K$, the internal energy per spin, C/K, the specific heat times T^2 per spin, and free energies ζ^*, based on 100 repeated simulations. Similar results are obtained by parallel tempering and optimally adjusted mixture sampling. But the latter algorithm achieves noticeable variance reduction for the estimates of $-U/K$ and C/K at temperatures T_4 and T_5. The algorithm using a non-optimal SA scheme performs

much worse than the first two algorithms: not only the estimates of $-U/K$ and C/K are noticeably biased at temperature T_5, but also the online estimates of free energies have greater variances at temperatures T_2 to T_5. The algorithm using the flat-histogram scheme performs even poorly, with serious biases for the estimates of $-U/K$ and C/K and large variances for the online estimates of free energies. These illustrate advantages of using optimally adjusted mixture sampling.

In the Appendix VI we provide additional results on offline estimates of free energies and expectations and other versions of self-adjusted mixture sampling. For optimal or non-optimal SA, the single-stage algorithm perform similarly to the corresponding two-stage algorithm, due to the small number ($m = 5$) of distributions involved. The version with local jump and update scheme (14) or with global jump and update scheme (12) yields similar results to those of the basic version.

6.2 Censored Gaussian random field

Consider a Gaussian random field measured on a regular 6 grid in $[0,1]^2$ but right-censored at 0 in Stein (1992). Let $(u_1,\ldots,u_K)$ be the $K = 36$ locations of the grid, $\xi = (\xi_1,\ldots,\xi_K)$ be the uncensored data, and $y = (y_1,\ldots,y_K)$ be the observed data such that $y_j = \max(\xi_j, 0)$. Assume that ξ is multivariate Gaussian with $E(\xi_j) = \beta$ and $\mathrm{cov}(\xi_j,\xi_j) = c\,\mathrm{e}^{-\|u_j-u_{j'}\|}$ for $j,j' = 1,\ldots,K$, where $\|\cdot\|$ is the Euclidean norm. The density function of ξ is $p(\xi;\theta) = (2\pi c)^{-K/2}\det^{-1/2}(\Sigma)\exp\{-(\xi-\beta)^{\mathrm{T}}\Sigma^{-1}(\xi-\beta)/(2c)\}$, where $\theta = (\beta, \log c)$ and Σ is the correlation matrix of ξ. The likelihood of the observed data can be decomposed as $L(\theta) = p(\xi_{\mathrm{obs}};\theta) \times L_{\mathrm{mis}}(\theta)$ with

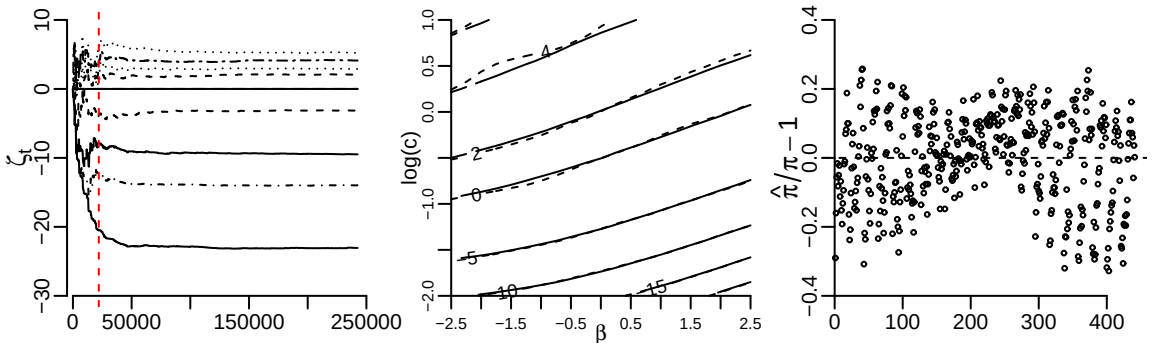
$$L_{\mathrm{mis}}(\theta) = \int_{-\infty}^{0}\cdots\int_{-\infty}^{0} p(\xi_{\mathrm{mis}}|\xi_{\mathrm{obs}};\theta)\prod_{j:y_j=0}\mathrm{d}\xi_j,$$

where ξ_{obs} or ξ_{mis} denotes the observed or censored subvector of ξ. Then $L_{\mathrm{mis}}(\theta)$ is the normalizing constant for the unnormalized density function $p(\xi_{\mathrm{mis}}|\xi_{\mathrm{obs}};\theta)$ in ξ_{mis}. For the dataset in Figure 1 of Stein (1992), it is of interest to compute $\{L(\theta) : \theta \in \Theta\}$, where Θ is a 21×21 regular grid in $[-2.5, 2.5] \times [-2, 1]$. There are 17 censored observations in Stein's dataset and hence $L_{\mathrm{mis}}(\theta)$ is a 17-dimensional integral.

6.2.1 Simulation details

We take $q_j(x) = p(\xi_{\mathrm{mis}}|\xi_{\mathrm{obs}};\theta)$ for $j = 1,\ldots,m(= 441)$, where $\theta_{j_1+21\times(j_2-1)}$ denotes the grid point $(\theta^1_{j_1},\theta^2_{j_2})$ for $j_1,j_2 = 1,\ldots,21$, and $(\theta^1_1,\ldots,\theta^1_{21})$ are evenly spaced in

Figure 4: Trace plots of online estimates of ζ_j^* for 9 points θ_j that form a 3×3 subgrid of Θ, the contour plots of online (dashed) and offline (solid) estimates of $\{\zeta_j^* : j = 1, \ldots, m\}$ and the plot of $\hat{\pi}_j/\pi_j - 1$ over j.



$[-2.5, 2.5]$ and $(\theta_1^2, \ldots, \theta_{21}^2)$ are evenly spaced in $[-2, 1]$. The transition kernel Ψ_j is defined as a systematic scan of Gibbs sampling for the target distribution P_j (Liu 2001). In this example, Gibbs sampling seems to work reasonably well for each P_j. Previously, Gelman & Meng (1998) and Tan (2013a, 2013b) studied the problem of computing $\{L(\theta) : \theta \in \Theta\}$, up to a multiplicative constant, using Gibbs sampling to simulate m Markov chains independently for $(P_1, \ldots, P_m)$.

We investigate self-adjusted local-jump mixture sampling, with two-stage modification (15) of the local update scheme (14), and locally weighted histogram analysis for estimating $\theta_j^* = \log\{L_{\text{mis}}(\theta_j)/L_{\text{mis}}(\theta_{11}^1, \theta_{11}^2)\}$ for $j = 1, \ldots, m$. The use of self-adjusted mixture sampling is mainly to provide online estimates of ζ_j^*, to be compared with offline estimates, rather than to improve sampling as in the usual use of serial tempering (Geyer & Thompson 1995). As discussed in Sections 2–5, global-jump mixture sampling, the global update scheme (12), and globally weighted histogram analysis are too costly to be implemented for a large m.

The neighborhood $\mathcal{N}(j)$ is defined as the set of 2, 3, or 4 indices l such that θ_l lies within Θ and next to θ_j in one of the four directions. That is, if $j = j_1 + 21 \times (j_2 - 1)$, then $\mathcal{N}(j) = \{l_1 + 21 \times (l_2 - 1) : l_1 = j_1 \pm 1 (1 \le l_1 \le 21) \text{ and } l_2 = j_2 \pm 1 (1 \le l_2 \le 21)\}$. Additional simulations using larger neighborhoods (e.g., $l_1 = j_1 \pm 2$ and $l_2 = j_2 \pm 2$) lead to similar results to those reported in Section 6.2.2.

The initial value L_0 is set to $(\theta_{11}^1, \theta_{11}^2)$ corresponding to the center of Θ, and X_0 is generated by independently drawing each censored component ξ_j from the conditional distribution of ξ_j, truncated to $(-\infty, 0]$, given the observed components of ξ, with

Table 1: Log-likelihood ratios for a censored Gaussian model

		Single path		Averaged path		Stochastic approx		
	Chib	β-first	$\log c$-first	β-first	$\log c$-first	Simple	Ave	L-WHAM
CPU	$1 + 1.6$	$1 + 0.25$		$1 + 0.46$		$1 + 0$		$1 + 0.12$
10^3MSE	0.208	3.40	3.55	0.326	3.42	13.6	28.9	0.304

Note: Simple and Ave are the online estimators $\zeta^{(n)}$ and $\bar{\zeta}^{(n)}$, and L-WHAM is the locally weighted estimator $\hat{\zeta}^{(n)}$. Results are reproduced from Tan (2013b) for Chib's (1995) estimator and for Gelman & Meng's (1998) single-path and averaged-path estimators, depending on the type of paths, labeled as β-first or $\log c$-first. The CPU time, $a + b$, consists of a for simulating data and b for evaluating estimators, both divided by a for standardization. MSE $= \sum_{j=1}^{441} \mathrm{MSE}_{\theta_j}/441$, where MSE_{θ_j} is the Monte Carlo mean squared error for estimating ζ_j^*. The true values are approximated by the specialized method of Genz (1992) with estimated errors $< .001$ for all $\theta \in \Theta$.

$\theta = (\theta_{11}^1, \theta_{11}^2)$. The total number of iterations is set to 441×550 with the first 441×50 iterations as burn-in, corresponding to the cost in Gelman & Meng (1998) and Tan (2013a, 2013b), which involve simulating a Markov chain of length 550 per distribution, with the first 50 iterations as burn-in.

6.2.2 Simulation results

Figure 4 shows the output from a single run of self-adjusted mixture sampling with $\beta = 0.8$ and t_0 set to 441×50 (the burn-in size). There are a number of interesting features. First, the estimates $\zeta^{(t)}$ fall steadily toward the truth, with noticeable fluctuations, during the first stage, and then stay stable and close to the truth in the second stage, similarly as in Figure 2 for the Potts model. Second, the locally weighted offline estimates yield a more smooth contour than the online estimates. In fact, as shown later in Table 1 from repeated simulations, the offline estimates are orders of magnitude more accurate than the online estimates. Third, some of the observed proportions $\hat{\pi}_j$ differ from $\pi_j = 1/441$ by as much as 30%, even at the end of simulation. As discussed in Section 5, offline stratified estimation is robust to possible large deviations of $\hat{\pi}_j$ from π_j, which might explain why the offline estimates are much more accurate than the online estimates in this example.

Table 1 summarizes the results based on 100 repeated simulations using self-adjusted mixture sampling, and reproduces the corresponding results for related methods in Tan (2013b, Table 1), where samples are simulated separately from $(P_1, \ldots, P_m)$ by Gibbs sampling. Locally weighted estimation is also applied to the latter setting, and essentially the same results are obtained as in Table 1.

By Table 1, the offline locally weighted method yields mean squared errors about 45 times smaller than those of online estimation, with only a 12% increase in computational time. Moreover, there are computational advantages of locally weighted estimation over Chib's (1995) and Gelman & Meng's (1998) methods. Chib's method yields small mean squared errors, but is computationally costly, due to repeated evaluations of the transition kernel, a product of 17 conditional densities. Such evaluations are not needed in path sampling or locally weighted estimation. The performance of path sampling depends on the choice of paths: the averaged-path estimator along β-first paths is much more accurate than along $\log c$-first paths. But it may be difficult, in general, to distinguish between such implementation choices.

In the Appendix VI, we present additional results to illustrate the impact of using a non-optimal SA scheme, with gain factor t^{-1} replaced by $\min(1/m, 10/t)$ in (14), and that of using a single-stage algorithm, with $t_0 = 1$ in (15).

7 Conclusion

We develop not only a sampling method, self-adjusted mixture sampling, for simulation from multiple distributions and online estimation of expectations and normalizing constants, but also an offline method, locally weighted histogram analysis, for estimating expectations and normalizing constants.

Various topics can be further studied, in addition to those mentioned earlier. Labeled mixture sampling can be generalized to handle multiple distributions on state spaces of different dimensions, leading to a reversible jump algorithm (Green 1995). Then it is possible to incorporate stochastic approximation in reversible jump MCMC for adjusting pseudo priors (e.g., Atchadé & Liu 2010). Moreover, locally weighted histogram analysis can be generalized to trans-dimensional settings, similarly as bridge sampling for reversible jump MCMC (e.g., Bartolucci et al. 2006).

Supplementary Materials

Appendices: (I) Asymptotic theory of SA, (II) Additional theoretical results, (III) Technical details, (IV) Partition-based settings, (V) Potts model: Generalized ensemble simulation, (VI) Additional simulation results (SAMS-appendix.pdf)

Computer codes: C codes for simulations on the Potts model in Section 6.1 and Appendix IV and R codes for simulations on the censored Gaussian random

field in Section 6.2. Documentations of the codes are also provided. (SAMS-codes.tar.gz, GNU zipped tar file)

References

Andrieu, C. and Moulines, É. (2006) "On the ergodicity properties of some adaptive MCMC algorithms," *Annals of Applied Probability*, 16, 1462–1505.

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

Barker, A. A. (1965) "Monte Carlo calculations of the radial distribution functions for a proton-electron plasma," *Australian Journal of Physics*, 18, 119–134.

Bartolucci, F., Scaccia, L., and Mira, A. (2006) "Efficient Bayes factor estimation from the reversible jump output," *Biometrika*, 93, 41–52.

Bennett, C.H. (1976) "Efficient estimation of free energy differences from Monte Carlo data," *Journal of Computational Physics*, 22, 245–268.

Cameron, E. and Pettitt, A. (2014) "Recursive pathways to marginal likelihood estimation with prior-sensitivity analysis," *Statistical Science*, 29, 397–419.

Chen, H.-F. (2002) *Stochastic Approximation and Its Applications*, Dordrecht: Kluwer Academic Publishers.

Chib, S. (1995) "Marginal likelihood from the Gibbs output," *Journal of the American Statistical Association*, 90, 1313–1321.

Chodera, J.D. and Shirts, M.R. (2011) "Replica exchange and expanded ensemble simulations as Gibbs sampling: Simple improvements for enhanced sampling," *Journal of Chemical Physics*, 135, 194110.

Del Moral, P., Doucet, A., and Jasra, A. (2006) "Sequential Monte Carlo samplers," *Journal of the Royal Statistical Society*, Ser. B, 68, 411–436.

Delyon, B., Lavielle, M., and Moulines, É. (1999) "Convergence of a stochastic approximation version of the EM algorithm," *Annals of Statistics*, 27, 94–128.

Doss, H. (2010) "Estimation of large families of Bayes factors from Markov chain output", *Statistica Sinica*, 20 537–560.

Ferrenberg, A.M. and Swendsen, R.H. (1989) "Optimized Monte Carlo data analysis," *Physics Review Letters*, 63, 1195–1198.

Gelfand, A.E. and Smith, A.F.M. (1990) "Sampling-based approaches to calculating marginal densities," *Journal of American Statistical Association*, 85, 398–409.

Gelman, A. and Meng, X.-L. (1998) "Simulating normalizing constants: From importance sampling to bridge sampling to path sampling," *Statistical Science*, 13, 163–185.

Genz, A. (1992) "Numerical computation of multivariate normal probabilities," *Journal of Computational and Graphical Statistics*, 1, 141–150.

Geyer, C.J. (1991) "Markov chain Monte Carlo maximum likelihood," in *Computing Science and Statistics: Proceedings of 23rd Symposium on the Interface*, ed, E.M. Keramidas, Fairfax, VA: Interface Foundation, 156–163.

Geyer, C.J. (1994) "Estimating normalizing constants and reweighing mixtures in Markov chain Monte Carlo," *Technical Report 568*, School of Statistics, University of Minnesota.

Geyer, C.J. (2011) "Importance sampling, simulated tempering, and umbrella sampling," in *Handbook of Markov Chain Monte Carlo*, eds. S. Brooks, A. Gelman, G.L. Jones, and X.-L. Meng, Boca Raton, FL: Chapman & Hall, 295–311.

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

Gilks, W.R. and Berzuini, C. (2001) "Following a moving target — Monte Carlo inference for dynamic Bayesian models," *Journal of the Royal Statistical Society*, Ser. B, 63, 127-146.

Green, P.J. (1995) "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, 82, 711–732.

Gu, M.G., and Zhu, H.T. (2001) "Maximum likelihood estimation for spatial models by Markov chain Monte Carlo stochastic approximation,", *Journal of the Royal Statistical Society*, B, 63, 339–355.

Jasra, A., Stephens, D., and Holmes, C. (2007) "On Population-based simulation for static inference," *Statistics and Computing*, 17, 263–279.

Kong, A., McCullagh, P., Meng, X.-L., Nicolae, D., and Tan, Z. (2003) "A theory of statistical models for Monte Carlo integration" (with discussion), *Journal of the Royal Statistical Society*, B, 65, 585–618.

Kou, S.C., Zhou, Q., and Wong, W.H. (2006) "Equi-energy sampler with applications in statistical inference and statistical mechanics" (with discussion), *Annals of Statistics*, 34, 1581-1619.

Liang, F. (2009) "On the use of stochastic approximation Monte Carlo for Monte Carlo integration," *Statistics and Probability Letters*, 79, 581–587.

Liang, F. and Wong, W.H. (2001) "Real-parameter evolutionary Monte Carlo with applications to Bayesian mixture models," *Journal of the American Statistical Association*, 96, 653-666.

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

Liang, F., Jin, I.H., Song, Q., and Liu, J.S. (2015) "An adaptive exchange algorithm for sampling from distributions with intractable normalizing constants," *Journal of the American Statistical Association*, to appear.

Liu, J.S. (2001) *Monte Carlo Strategies in Scientific Computing*, New York: Springer.

Meng, X.-L. and Wong, W.H. (1996) "Simulating ratios of normalizing constants via a simple identity: A theoretical explanation," *Statistica Sinica*, 6, 831–860.

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

Robbins, H. and Monro, S. (1951) "A stochastic approximation method," *Annals of Mathematical Statistics*, 22 400–407.

Roberts, G.O. and Rosenthal, J.S. (2009) "Examples of adaptive MCMC," *Journal of Computational and Graphical Statistics*, 18, 349–367.

Shirts, M.R. and Chodera, J.D. (2008) "Statistically optimal analysis of samples from multiple equilibrium states," *Journal of Chemical Physics*, 129, 124105.

Song, Q., Wu, M., and Liang, F. (2014) "Weak convergence rates of population versus single-chain stochastic approximation MCMC algorithms," *Advances in Applied Probability*, 46, 1059–1083.

Stein, M. (1992) "Prediction and inference for truncated spatial data," *Journal of Computational and Graphical Statistics*, 1, 91–110.

Tan, Z. (2004) "On a likelihood approach for Monte Carlo integration," *Journal of the American Statistical Association*, 99, 1027–1036.

Tan, Z. (2013a) "A cluster-sample approach for Monte Carlo integration using multiple samplers," *Canadian Journal of Statistics*, 41, 151–173.

Tan, Z. (2013b) "Calibrated path sampling and stepwise bridge sampling," *Journal of Statistical Planning and Inference*, 143, 675–900.

Tan, Z. (2015) "Resampling Markov chain Monte Carlo algorithms: Basic analysis and empirical comparisons," *Journal of Computational and Graphical Statistics*, 24, 328–356.

Tan, Z., Gallicchio, E., Lapelosa, M., and Levy, R.M. (2012) "Theory of binless multi-state free energy estimation with applications to protein-ligand binding," *Journal of Chemical Physics*, 136, 144102.

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

Wu, F.Y. (1982) "The Potts model," *Reviews of Modern Physics*, 54, 235–268.