Optimally adjusted mixture sampling and locally weighted histogram analysis

Zhiqiang Tan[1]

February 2014, revised July 2014

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 schemes. This method can be computationally much cheaper, with little sacrifice of statistical efficiency, than a global method currently used, especially with a large number of distributions. 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. The author thanks the Editor, an associate editor, and two referees for helpful comments, which led to substantial improvements of the article.

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 $\mathcal{X}$ defined 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 in different manners.

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 7.1.

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., 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; Chopin et al. 2012; Bornn et al. 2013). 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 P_0 by importance sampling. See Section 7.2 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; Robert & Casella 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), serial tempering (Geyer & Thompson 1995), resample-move (Gilks & Berzuini 2001) and its extensions (De Moral et al. 2006; Tan 2014), 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 the corresponding 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 *online* as a part of the sampling process, and ζ_0^* and $E_0(\phi)$ are then estimated by standard 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 *offline* by a methodology known in physics and statistics as the (binless) weighted histogram analysis method (Tan et al. 2012), the multi-state Bennet acceptance ratio method (Shirts & Chodera 2008), reverse logistic regression (Geyer 1994), bridge sampling (Meng & Wong 1996) and the global likelihood method (Kong et al. 2003). 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 schemes, subject to suitable overlaps between $(P_1, \ldots, P_m)$. The local method is expected to be computationally much cheaper, with little sacrifice of statistical efficiency, than the global method, because individual samples are locally pooled from neighbor distributions, which are typically overlapped 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 for multiple distributions. The ideas are recast from several existing methods, including serial tempering (Geyer & Thompson 1995) and 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 identification of optimal adaptive schemes later in Section 3.2.

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* positive weights with $\sum_{j=1}^m \pi_j = 1$, for example, $\pi_1 = \cdots = \pi_m = m^{-1}$, and $\zeta = (\zeta_1, \dots, \zeta_m)^{\mathrm{T}}$ with $\zeta_1 = 0$ are *hypothesized* values of the true log 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

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

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 (2), there is a one-to-one relationship between hypothesized free energies ζ and mixture weights $p(L = \cdot; \zeta)$. In particular, π represents the mixture weights that would be obtained if $\zeta = \zeta^*$.

For each fixed choice of ζ, the labeled mixture (1) can be sampled by standard MCMC, with the unnormalized density $\pi_j e^{-\zeta_j} q_j(x)$. 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 Section 6 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

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

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 $j \in \mathcal{N}(j)$ 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 cheaper than the global-jump algorithm. At each time, 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 7.1.

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, ζ must be specified reasonably close to ζ^* in order for the algorithm to work. By equation (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 $\delta^{(t)} = (1\{L_t = 1\}, \ldots, 1\{L_t = m\})^{\mathrm{T}}$ 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 (2), so that $\zeta_j^{(t)}$ will increase (or decrease) from $\zeta_j^{(t-1)}$ by the update rule (4).

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 adapative MCMC (e.g., Harrio et al. 2001; 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 function, 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 value θ_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 Supplementary Material, we provide a summary of Theorems 1–2 in Song et al. (2013) 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 $C = -\partial h(\theta^*)/\partial\theta^{\mathrm{T}}$.

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 elements $\zeta_1 = \zeta_1^* = 0$ excluded from ζ and ζ^*, and

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

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

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

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

where $p_{\text{LJ}}(l_t|l_{t-1}, x_{t-1}; \zeta)$ 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 $\Psi_{l_t}(x_{t-1}, x_t)$. The sequence of variables generated by the SA algorithm reduce to $Y_t = (L_t, X_t)$ and $\theta_t = (\zeta_2^{(t)}, \ldots, \zeta_m^{(t)})^{\text{T}}$.

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 for proofs of Theorems 1–3 and the Supplementary Material for all other proofs.

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

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

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 $C = -\partial h(\theta^*)/\partial\theta^{\text{T}}$ is known, even though $\theta^* = (\zeta_2^*, \ldots, \zeta_m^*)^{\text{T}}$ is unknown, so that the optimal SA recursion becomes feasible.

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 for local-jump labeled mixture sampling with the transition kernel (8) as in SAMC and for global-jump labeled mixture sampling with the transition kernel

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

where $p_{\text{GJ}}(l_t|x_{t-1}; \zeta)$ is a probability density function of L_t given (L_{t-1}, X_{t-1}) or given X_{t-1} under global jump. Then $p_{\text{GJ}}(l_t|x_{t-1}; \zeta) = p(l_t|x_{t-1}; \zeta)$ by (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 $w_j(X;\zeta) = p(L = j|X;\zeta)$ by (3), i.e.,

$$w_j(X;\zeta) = \frac{\pi_j \mathrm{e}^{-\zeta_j} q_j(X)}{\sum_{l=1}^m \pi_l \mathrm{e}^{-\zeta_l} q_l(X)}, \quad j = 1, \ldots, m.$$

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) is a conditional expectation (or Rao–Blackwellization) of the earlier choice (7) for $H(Y;\theta)$. 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

$$\begin{aligned} &E(1\{L_t = j\}|L_{t-1} = k, X_{t-1};\zeta) \\ &= \begin{cases} \Gamma(k,j)\min\left\{1, \frac{\Gamma(j,k)}{\Gamma(k,j)}\frac{p(j|X_{t-1};\zeta)}{p(k|X_{t-1};\zeta)}\right\}, & \text{if } j \in \mathcal{N}(k), \\ 1 - \sum_{l\in\mathcal{N}(k)} E(1\{L_t = l\}|L_{t-1} = k, X_{t-1};\zeta), & \text{if } j = k, \end{cases} \end{aligned}$$

where $\mathcal{N}(k)$ denotes the neighborhood of k. 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 for $j = 1, 2, \ldots, m$,

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

and $u_j(L,X;\zeta) = 0$ if $j \notin \{L\} \cup \mathcal{N}(L)$. 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 scheme (14) for updating $\zeta^{(t)}$, 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. As discussed in Section 2, local-jump mixture sampling is computationally cheaper than global-jump mixture sampling, whereas the relative statistical efficiency seems to be problem-dependent between the two schemes. Next, we compare the update schemes (9), (12), and (14).

First, the update schemes (9), (14), and (12) are associated with increasing computational cost. At each time t, the binary scheme (9) requires no evaluation of

unnormalized densities, but the global scheme (12) requires evaluating m unnormalized densities $q_1(X_t), \ldots, q_m(X_t)$, which can be prohibitive when m is large (e.g., in hundreds). The local scheme (14) provides a useful compromise, with evaluation of only $\{1 + s(L_t)\}$ unnormalized densities $\{q_j(X_t) : j = L_t \text{ or } j \in \mathcal{N}(L_t)\}$, where $s(L_t)$ is the size of $\mathcal{N}(L_t)$ and can be made much smaller than m.

Second, the Rao–Blackwellized scheme (12) or (14) is expected to be statistically more efficient than scheme (9), at least when used with, respectively, global-jump or local-jump labeled mixture sampling. We provide a theoretical result (Theorem 4) for the global scheme (12) used with global-jump labeled mixture sampling. See Liu et al. (1994) and McKeague & Wefelmeyer (2000) for related results on two-block Gibbs sampling and reversible Markov chains, although these conditions are, in general, not satisfied by global-jump or local-jump labeled mixture sampling.

Theorem 4. Let $\{(Y_t, \theta_t) : t \geq 1\}$ be a sequence generated by the transition kernel (10) and the binary update scheme (9), and $\{(Y_t', \theta_t') : t \geq 1\}$ be a sequence generated by the transition kernel (10) and the global update scheme (12), each under self-adjusted global-jump mixture sampling. Then the asymptotic variance of $t^{1/2}(\theta_t' - \theta^*)$ is no greater than that of $t^{1/2}(\theta_t - \theta^*)$ as $t \to \infty$.

Third, as seen later in Sections 4–5, the global or local update scheme, (12) or (14), leads to, respectively, globally or locally weighted offline estimation after sampling is completed. The unnormalized densities evaluated in the scheme (12) or (14) are also needed in the corresponding offline estimation. But global or local offline estimation can be statistically more efficient than online estimation by any update scheme (9), (12) or (14). See our numerical example in Section 7.3, where, for a large m, locally weighted offline estimation is substantially more accurate than online estimation, with only limited increase of computational cost. Therefore, whether the cost of evaluating unnormalized densities in the Rao–Blackwellized scheme (12) or (14) is worthwhile should be compared with variance reduction achieved by the corresponding offline estimator over the binary SA scheme (9).

We now 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 Sections 6.2–6.3. 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 no 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. In the following, assume that for $j = 1, \ldots, m$,

(A1) the observed weight $\hat{\pi}_j = n_j/n$ 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 $n_j^{1/2}\{\tilde{E}_j(\phi) - E_j(\phi)\}$ is asymptotically normally distributed.

Therefore, 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 $n^{-1/2}$.

Formal theory remains to be developed to provide suitable regularity conditions for such results. The within-label average $\tilde{E}_j(\phi)$ is expected to converge at the usual rate $n^{-1/2}$ because the conditional distribution of X given $L = j$ is always P_j, where (L, X) is drawn from the invariant distribution $p(j, x; \zeta)$ for any fixed choice of ζ. See Andrieu & Moulines (2006) for related theory on adaptive MCMC, 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 $\tilde{\zeta}_1^{(n)} = 0$ 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, equation (16) needs only to be solved for $j = 2, \ldots, m$. Interestingly, 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 $(X_1, \ldots, X_n)$ 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 weight $\hat{\pi}_j$ 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 $j = 1, \ldots, m$. This difference underlies the efficiency differences between $\tilde{\zeta}^{(n)}$ and $\hat{\zeta}^{(n)}$ later in Theorem 5.

There are various developments leading to the same estimator $\hat{\zeta}^{(n)}$ in physics and statistics, including the (binless) weighted histogram analysis method (WHAM) (Ferrenberg & Swendsen 1989; Tan et al. 2012), the multi-state Bennet acceptance ratio method (Bennet 1976; Shirts & Chodera 2008), reverse logistic regression (Geyer 1994), multiple bridge sampling (Meng & Wong 1996), and the global likelihood method (Kong et al. 2003; Tan 2004). See Doss & Tan (2014) for variance estimation via regeneration. 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

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

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 $j = 0$.

Next, we present interesting results on comparison of statistical efficiency between the online estimator $\zeta^{(n)}$ and offline estimators $\tilde{\zeta}^{(n)}$ and $\hat{\zeta}^{(n)}$.

Theorem 5. Let θ_n be the SA estimator obtained from $\{(Y_i, \theta_i) : i = 1, \ldots, n\}$ generated by the transition kernel (10) and the global update scheme (12), under self-adjusted global-jump mixture sampling. Moreover, let $\tilde{\theta}_n = (\tilde{\zeta}_2^{(n)}, \ldots, \tilde{\zeta}_m^{(n)})^{\mathrm{T}}$ and $\hat{\theta}_n = (\hat{\zeta}_2^{(n)}, \ldots, \hat{\zeta}_m^{(n)})^{\mathrm{T}}$ be the unstratified and stratified estimators based on a Markov chain $\{Y_i^* = (L_i^*, X_i^*) : i = 1, \ldots, n\}$ generated by the transition kernel (10) under global-jump mixture sampling, with θ fixed at θ^*. Then the asymptotic variances, Σ_1, Σ_2, and Σ_3, of $n^{1/2}(\theta_n - \theta^*)$, $n^{1/2}(\tilde{\theta}_n - \theta^*)$, and $n^{1/2}(\hat{\theta}_n - \theta^*)$ satisfy

$$\Sigma_1 = C^{-1}VC^{-1}, \quad \Sigma_2 = D^{-1}VD^{-1},$$
$$\Sigma_3 = \Sigma_1 - (D^{-1} - C^{-1})D(D^{-1} - C^{-1}) \leq \Sigma_1.$$

where $C = -\partial h(\theta^*)/\partial\theta^{\mathrm{T}}$ and $D = -E\{\partial H(Y; \theta^*)/\partial\theta^{\mathrm{T}}\}$ with $Y = (L, X) \sim p(j, x; \zeta^*)$. As shown in Supplementary Material, C and D are symmetric, positive definite matrices, and $C \geq D$, i.e., $C - D$ is nonnegative definite.

A number of remarks are in order. First, the foregoing result does not address the asymptotic variances of the estimator $\tilde{\zeta}^{(n)}$ and $\hat{\zeta}^{(n)}$ under self-adjusted global-jump mixture sampling. Further study of these asymptotic variances is desirable.

Second, the stratified estimator $\hat{\zeta}^{(n)}$ or the unstratified one $\tilde{\zeta}^{(n)}$ seems to be statistically more or, respectively, less efficient than the SA estimator $\zeta^{(n)}$. The variance matrix Σ_3 is always no greater than Σ_1. On the other hand, $\Sigma_2 = D^{-1}VD^{-1}$ tends to be greater than $\Sigma_1 = C^{-1}VC^{-1}$, in view of the relationship $C \geq D$. The matrix inequality $D^{-1}VD^{-1} \geq C^{-1}VC^{-1}$ holds if C, D and V are scalars as in the case of $m = 2$, but may not, in general, be valid for $m \geq 3$.

Third, there is a clear comparison of Σ_1, Σ_2, and Σ_3 as follows, in the special case where each transition kernel Ψ_j reduces to independent sampling.

Corollary 1. Let θ_n, $\tilde{\theta}_n$, and $\hat{\theta}_n$ be as in Theorem 5, but assume that the transition kernel $\Psi_j(x, x')$ corresponds to drawing x' directly from P_j, independently of x, for $j = 1, \ldots, m$. Moreover, let $\tilde{\theta}_n^{\mathrm{ind}}$ and $\hat{\theta}_n^{\mathrm{ind}}$ be the unstratified and stratified estimators

based on n independent draws from $p(j, x; \zeta^*)$. Then

$$\Sigma_3^{\text{ind}} = \Sigma_3 \leq \Sigma_1 \leq \Sigma_2^{\text{ind}} \leq \Sigma_2,$$

where Σ_2^{ind} and Σ_3^{ind} are the asymptotic variances of $n^{1/2}(\tilde{\theta}_n^{\text{ind}} - \theta^*)$ and $n^{1/2}(\hat{\theta}_n^{\text{ind}} - \theta^*)$.

The estimators θ_n, $\tilde{\theta}_n^{\text{ind}}$, and $\hat{\theta}_n^{\text{ind}}$ are feasible, free of the truth θ^*, when independent sampling is feasible from each individual distribution P_j. The associated asymptotic variances satisfy $\Sigma_3^{\text{ind}} \leq \Sigma_1 \leq \Sigma_2^{\text{ind}}$, in the order discussed in the second remark above. Moreover, Σ_3^{ind} is easily seen to be the asymptotic variance of the stratified estimator obtained under stratified independent sampling, i.e., a sample of n_j observations are independently generated from P_j, with $n_j = n\pi_j$ pre-specified, for $j = 1, \ldots, m$. By Tan (2004), Σ_3^{ind} is the smallest asymptotic variance possible among a broad class of bridge-sampling estimators in Meng & Wong (1996). Therefore, Corollary 1 extends the optimality of the stratified estimator $\hat{\zeta}^{(n)}$, to the case where $\hat{\zeta}^{(n)}$ is compared with the SA estimator $\zeta^{(n)}$ obtained from a seemingly dependent process.

5 Locally weighted histogram analysis

The conjunction of Theorems 5(ii) and 6 suggests that the offline estimator $\hat{\zeta}^{(n)}$ is often statistically more efficient than the SA estimator $\zeta^{(n)}$. However, the estimator $\hat{\zeta}^{(n)}$ requires evaluating m unnormalized densities $q_1(X_i), \ldots, q_m(X_i)$, 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 $\tilde{\zeta}^{(n)} = (\tilde{\zeta}_1^{(n)}, \ldots, \tilde{\zeta}_m^{(n)})^{\mathrm{T}}$ with $\tilde{\zeta}_1^{(n)} = 0$ as 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 some algebra, 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).

Equation (18) can be interpreted as combining estimation over pairs of samples (Meng & Wong 1996). In fact, if $m = 2$, then equation (18) with $j = 2$ yields

$$\frac{1}{n}\sum_{i=1}^n \left[\frac{1\{L_i=1\}\Gamma(1,2)\mathrm{e}^{-\zeta_2}q_2(X_i)}{\Gamma(1,2)\pi_1 q_1(X_i)+\Gamma(2,1)\pi_2\mathrm{e}^{-\zeta_2}q_2(X_i)}\right.$$
$$\left. + \frac{1\{L_i=2\}\Gamma(2,1)\mathrm{e}^{-\zeta_2}q_2(X_i)}{\Gamma(1,2)\pi_1 q_1(X_i)+\Gamma(2,1)\pi_2\mathrm{e}^{-\zeta_2}q_2(X_i)}\right] = 1, \qquad (19)$$

Equation (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.,

$$\frac{1}{n}\sum_{i=1}^n \frac{\mathrm{e}^{-\zeta_2}q_2(X_i)}{\pi_1 q_j(X_i)+\pi_2\mathrm{e}^{-\zeta_2}q_2(X_i)} = 1.$$

For a general m, equation (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 equation (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, expectations, and even probability distributions 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 neighbor distributions $\{P_l : l \in \mathcal{N}(j)\}$, with weights determined by

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

In contrast with global estimation, the baseline measure μ is estimated by different $\hat{\mu}_j$, using a different subset of simulated data, depending on which free energy ζ_j^* is to be computed. Nevertheless, similarly as in global estimation, integrals of interest can be estimated by substituting $\hat{\mu}_j$ for μ with a suitable choice of j. The expectation $E_j(\phi) = \int \phi(x)\, \mathrm{d}P_j$ can be estimated by $\int \phi(x) \exp(-\zeta_j^{(n)}) q_j(x)\, \mathrm{d}\hat{\mu}_j$ for $j = 1, \ldots, m$.

As seen from the last expression, the probability distribution P_j is effectively estimated by a discrete distribution $\hat{P}_j$ supported on $S_j \cup (\cup_{l \in \mathcal{N}(j)} S_l)$, not just the single sample S_j, with probabilities $\hat{P}_j(\{X_i\}) = \exp(-\zeta_j^{(n)}) q_j(X_i) \hat{\mu}_j(\{X_i\})$.

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, in parallel to globally weighted histogram analysis (Tan et al. 2012). By design, the local method is computationally cheaper than the global method, which can be impractical for a large m. The local stratified estimator $\hat{\zeta}^{(n)}$ requires evaluating only $\{1 + s(L_t)\}$ 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 is typically overlapped more with the neighbor 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 Doss & Tan (2014), 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 2013c) including resample-move (Gilks & Berzuini 2001) and equi-energy sampling (Kou et al. 2006).

6 Partition-based settings

Sections 3–5 are focused on the overlap-based settings, where $(P_1, \ldots, P_m)$ are assumed to be overlapped with each other. Technically, this condition means that the

supports of $(P_1, \ldots, P_m)$ are inseparable (Vardi 1985; Geyer 1994), i.e., the union of the supports of any proper subset of $(P_1, \ldots, P_m)$ and that of the supports of the complement subset are *not* mutually exclusive. In this section, we discuss self-adjusted mixture sampling and estimation in possibly separable settings.

Self-adjusted mixture sampling can be directly extended by using MH labeled mixture sampling in Section 2 for Markov transitions, while using the same scheme for free energy updates. In fact, the optimal SA recursion (5) is not affected by the choice of the transition kernel K_θ as noted before, and still leads to the update scheme (9), (12), or (14). By this extension, self-adjusted mixture sampling is applicable even in separable settings, including partition-based settings for the Wang–Landau (2001) algorithm and its extensions (Liang et al. 2007; Atchadé & Liu 2010).

Let $(E_1, \ldots, E_m)$ be a partition of the state space $\mathcal{X}$, and define $q_j(x) = q_*(x)1\{x \in E_j\}$ for $j = 1, \ldots, m$, where $q_*(x)$ is an unnormalized density function. In this case, local-jump or global-jump labeled mixture sampling breaks down. Nevertheless, self-adjusted mixture sampling gives an optimally adjusted Wang–Landau algorithm as follows. The update schemes (12) and (14) both reduce to (9). In practice, we adopt a two-stage implementation with (15), especially when m is large.

Optimally adjusted Wang–Landau sampling:

- Labeled mixture sampling: Generate x from a proposal distribution $Q(X_{t-1}, \cdot)$, determine j such that $x \in E_j$, and set $(L_t, X_t) = (j, x)$ with probability $\min[1, \{Q(x, X_{t-1})p(x; \zeta^{(t-1)})\}/\{Q(X_{t-1}, x)p(X_{t-1}; \zeta^{(t-1)})\}]$.
- Free energy update: Compute $\zeta^{(t)}$ by (9).

For partition-based settings, the SA estimator $\zeta^{(n)}$ seems to be the only option for estimating free energies ζ^*. Equation (16) in general admits no solution, whereas equation (17) holds for an arbitrary vector ζ. For $q_0(x)$ possibly distinct from $q_*(x)$, standard importance sampling, with $\sum_{j=1}^m \pi_j \mathrm{e}^{-\zeta_j^{(n)}} q_*(x)1\{x \in E_j\}$ as a trial density, leads to the unstratified estimators for ζ_0^* and $E_0(\phi)$ (Liang 2009):

$$\mathrm{e}^{\tilde{\zeta}_0^{(n)}} = \sum_{j=1}^m \frac{n_j/n}{\pi_j} \mathrm{e}^{\zeta_j^{(n)}} \tilde{E}_j\left(\frac{q_0}{q_*}\right), \quad \tilde{E}_0(\phi) = \sum_{j=1}^m \frac{n_j/n}{\pi_j} \mathrm{e}^{\zeta_j^{(n)} - \tilde{\zeta}_0^{(n)}} \tilde{E}_j\left(\frac{q_0}{q_*}\phi\right), \tag{20}$$

where n_j and $\tilde{E}_j(\cdot)$ are defined as in Section 4. However, the argument in support of stratification behind (17) is also applicable here. Treating $\sum_{j=1}^m (n_j/n)\mathrm{e}^{-\zeta_j^{(n)}} q_*(x)1\{x \in$

$E_j\}$ as a trial density yields the stratified estimators

$$e^{\hat{\zeta}_0^{(n)}} = \sum_{j=1}^{m} e^{\zeta_j^{(n)}} \tilde{E}_j \left(\frac{q_0}{q_*}\right), \quad \hat{E}_0(\phi) = \sum_{j=1}^{m} e^{\zeta_j^{(n)} - \hat{\zeta}_0^{(n)}} \tilde{E}_j \left(\frac{q_0}{q_*}\phi\right). \tag{21}$$

This simple modification is found to achieve considerable variance reduction in our numerical study on the Potts model in Section 7.2.

7 Simulation studies

7.1 Potts model: Canonical ensemble simulation

The Potts model is important in statistical physics, with various applications. 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

$$Z^{-1} e^{-u(x)/T},$$

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 $\beta = T^{-1}$. 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 β is suppressed in the notation. Then $U = -(\mathrm{d}/\mathrm{d}\beta) \log Z$ and $C = (\mathrm{d}^2/\mathrm{d}\beta^2) \log Z$ 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 $\beta_c = T_c^{-1} = \log(1+\sqrt{q})$, 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).

7.1.1 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.

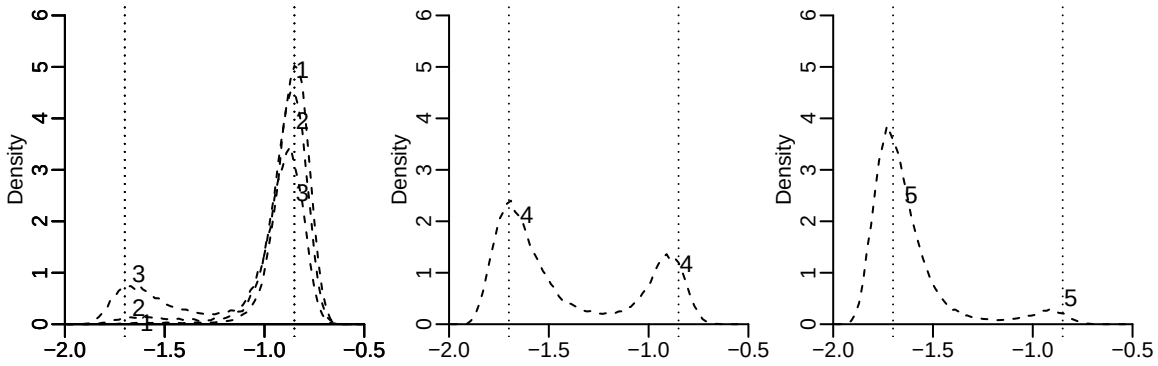
We compare the following four algorithms, where the neighborhood $\mathcal{N}(j)$ is defined as $\{1 \le l \le 5 : l = j - 1 \text{ or } j + 1\}$, of size 1 or 2.

- Parallel tempering (Geyer 1991), implemented as in Tan (2014) to ensure that 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 by $\min(1/m, 10/t)$, which 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 a gain factor that decreases only when a flat-histogram criterion is met: the observed weights $\hat{\pi}_j$ are all within 20% of the target weights $1/m$ (e.g., Atchadé & Liu 2010).

See Tan (2013c) 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 for each of the 5 chains in 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, in parallel tempering. For self-adjusted mixture

Figure 1: Histograms 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.



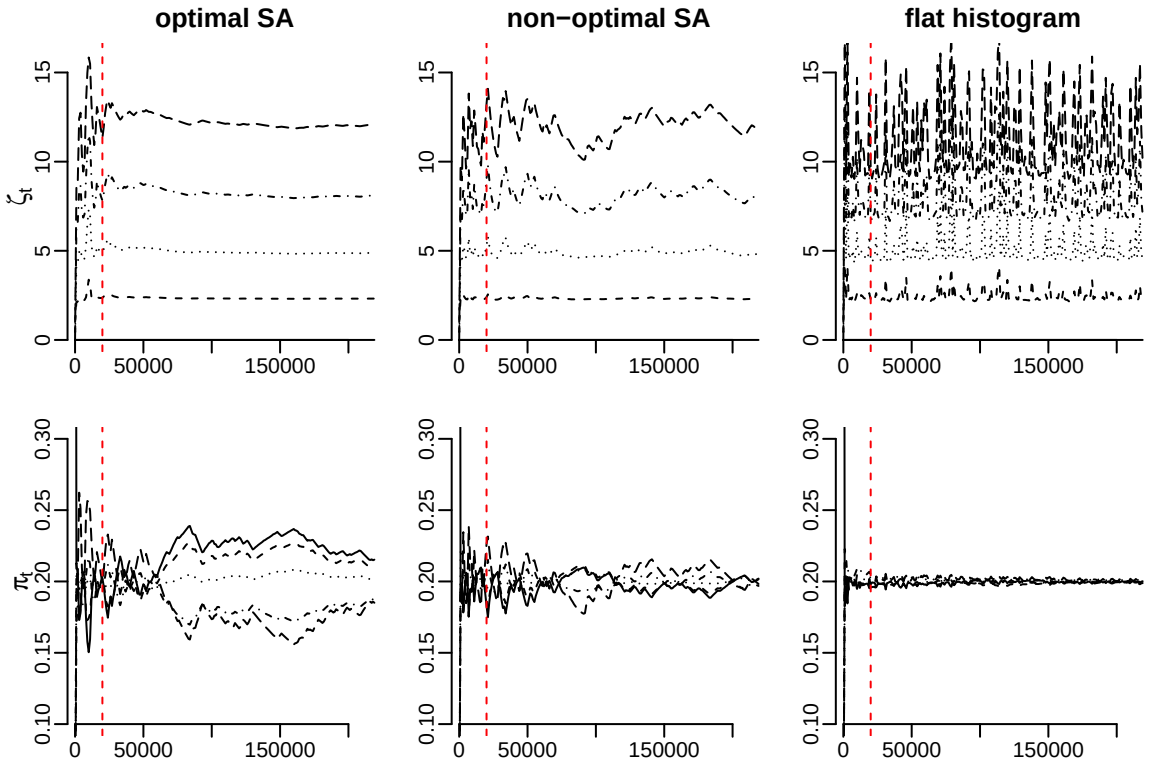
sampling with comparable cost, the total number of iterations is set to 2.2×10^6, with the first 2×10^5 treated as burn-in. The data are recorded (or subsampled) every 10 iterations, yielding 5 chains each of length 4.4×10^4 with the first 4×10^3 iterations as burn-in for parallel tempering, and a single chain of length 2.2×10^5 with the first 2×10^4 iterations as burn-in for self-adjusted mixture sampling.

7.1.2 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 considerable deviations of $\hat{\pi}_j$ from π_j over time, which reflects the presence

Figure 2: Trace plots for self-adjusted mixture sampling. The number of iterations is shown after subsampling. A vertical line is placed at the burn-in size.

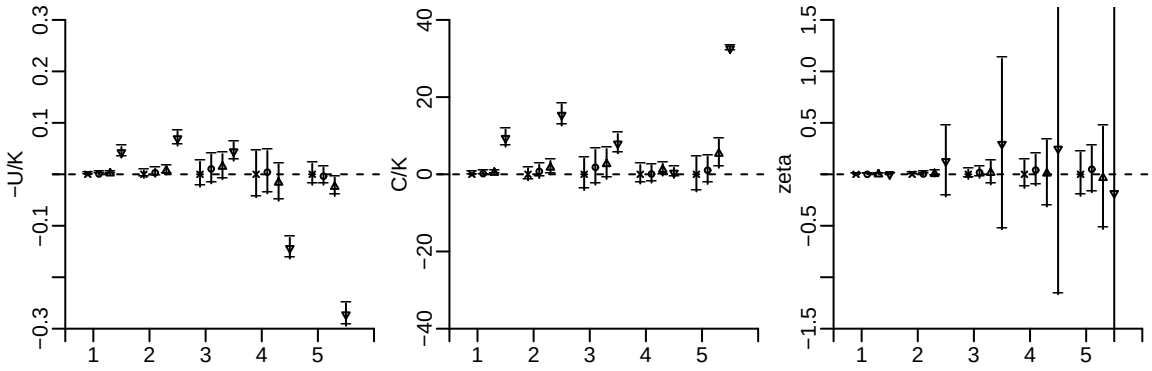


of strong autocorrelations in the label sequence L_t.

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, and 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 per spin multiplied by T^2, and free energies ζ^*, based on 100 repeated simulations. Similar results are obtained by parallel tempering and optimally-adjusted mixture sampling, although the latter algorithm achieves variance reduction for the estimates of $-U/K$ and C/K at the

Figure 3: Summary of estimates at the temperatures $(T_1, \ldots, T_5)$ 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, $\triangledown$: flat-histogram scheme), and the radius indicates Monte Carlo standard deviation of the 100 estimates from repeated simulations.



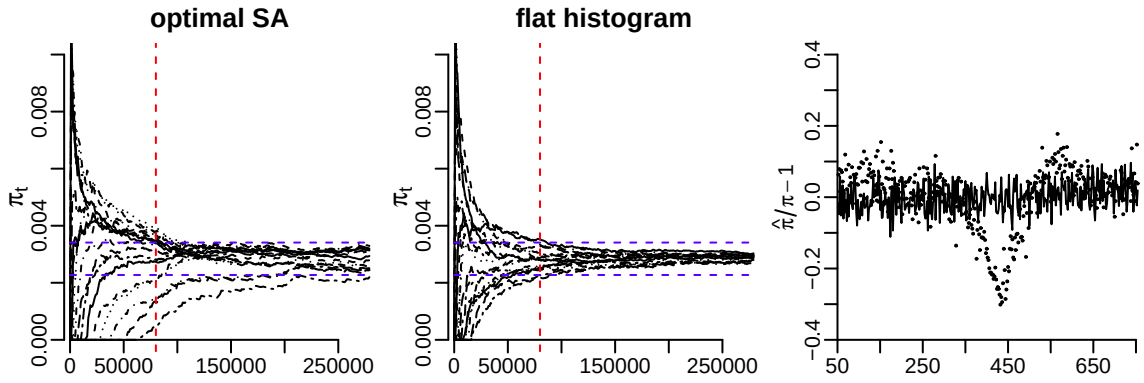
coldest temperature T_5. The algorithm using a non-optimal SA scheme performs worse than the first two algorithms: not only the estimates of $-U/K$ and C/K seem to be biased at temperature T_5, but also the online estimates of free energies have greater variances at all temperatures T_2 to T_5. The algorithm using the flat-histogram scheme performs poorly, with serious biases for the estimates of $-U/K$ and C/K and large variances for the online estimates of free energies. Therefore, it is important to use the optimal SA scheme for self-adjusted mixture sampling.

In the Supplementary Material, we provide additional results, including offline estimates of free energies, for other versions of self-adjusted mixture sampling. Compared with the two-stage version with $t_0 = 2 \times 10^5$ in (15), the single-stage version with $t_0 = 1$ yields similar results, but that with $t_0 = 2.2 \times 10^6$ (the length of entire simulation) and hence a non-optimal gain factor performs less satisfactorily. 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.

7.2 Potts model: Generalized ensemble simulation

Sampling directly from Potts distributions in Section 7.1 is often called canonical ensemble simulation. In this section, we study generalized ensemble simulation for the Potts model, based on partitioning of the state space

Figure 4: On the left are two trace plots of observed weights for Wang–Landau type sampling, where a vertical line is placed at the burn-in size and two horizontal lines are placed 20% away from the target weight $1/m$. On the right are the plots of observed weights at the end of simulation for the optimal SA (dot) or flat-histogram scheme (line).



7.2.1 Simulation details

The state space $\mathcal{X}$ is partitioned along $-u(x)$ into $m = 352$ regions, $E_1 = \{x : -u(x) \le 50\}$, $E_j = \{x : 50 + 2(j-2) < -u(x) \le 50 + 2(j-1)\}$ for $j = 2, \ldots, 351$, and $E_{352} = \{x : -u(x) > 750\}$. The distribution P_j is defined as the uniform distribution over E_j, with $q_j(x) = 1\{x \in E_j\}$, for $j = 1, \ldots, m$.

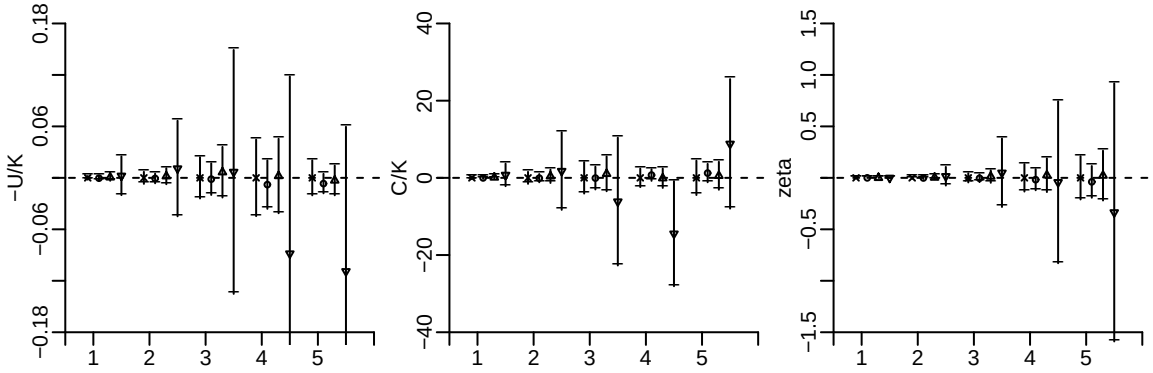
We compare the flat-histogram Wang–Landau (2001) algorithm and optimally adjusted Wang–Landau sampling with two-stage modification (15), where β is set to 0.6, to speed up adjustments of free energy estimates in the first stage because m is large. At time t, the sampling step consists of a random-scan sweep using a single-spin-flip Metropolis algorithm for the mixture $p(x; \zeta^{(t-1)})$.

The total number of iterations is set to 3×10^6, with the first 8×10^5 treated as burn-in. The data are recorded (or subsampled) every 10 iterations, yielding a chain of length 3×10^5 with the first 8×10^4 iterations as burn-in. The total sample size after burn-in is the same as in Section 7.1, but a larger burn-in size is used so that all observed proportions $\hat{\pi}_j$ are nonzero at the end of burn-in.

7.2.2 Simulation results

Figure 4 shows the observed proportions $\hat{\pi}_j$ from single simulations. See the Supplementary Material for trace plots of free energy estimates $\zeta^{(t)}$. Similarly as in Section 7.1, the flat-histogram scheme forces $\hat{\pi}_j$ to stay much closer to the target $\pi_j = 1/m$

Figure 5: Summary of estimates at the temperatures $(T_1, \ldots, T_5)$ labeled as $1, \ldots, 5$, based on 100 repeated simulations of Wang–Landau type algorithms. For each vertical bar, the center indicates Monte Carlo mean minus that obtained from parallel tempering in Section 7.1 ($\times$: parallel tempering, $\circ$ or $\triangle$: stratified or unstratified estimation with optimal SA scheme, $\triangledown$: unstratified estimation with flat-histogram scheme), and the radius indicates Monte Carlo standard deviation.



than the optimal SA scheme. This tight control of observed proportions, as shown below, is not necessarily desirable for sampling and estimation.

Figure 3 shows the Monte Carlo means and standard deviations, based on 100 repeated simulations, for the estimates of $-U/K$, C/K, and free energies for the 5 Potts distributions in Section 7.1 (each of which serves as P_0 in Section 6). The results obtained by unstratified estimation (20) with the optimal SA scheme are similar to those by parallel tempering and hence by optimally adjusted mixture sampling in Section 7.1. But stratified estimation (21) achieves considerable variance reduction over unstratified estimation. The flat-histogram scheme leads to large variances or biases for both stratified (not shown) and unstratified estimation.

7.3 Censored Gaussian random field

Consider a Gaussian random field measured on a regular 6×6 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 $L(\theta) = p(\xi_{\text{obs}}; \theta) \times L_{\text{mis}}(\theta)$ with

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

where ξ_{obs} or ξ_{mis} denotes the observed or censored subvector of ξ. Then $L_{\text{mis}}(\theta)$ is the normalizing constant for the unnormalized density function $p(\xi_{\text{mis}}|\xi_{\text{obs}}; \theta)$ 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 $L_{\text{mis}}(\theta)$ is a 17-dimensional integral.

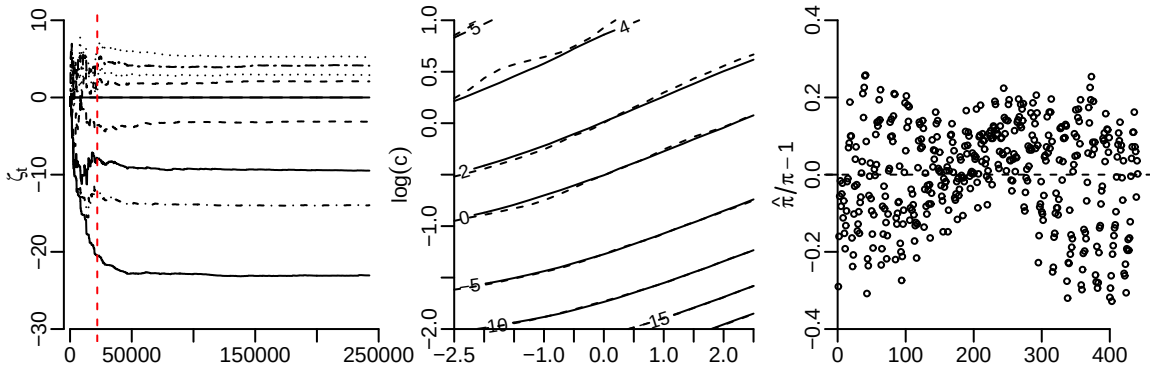
7.3.1 Simulation details

We take $q_j(x) = p(\xi_{\text{mis}}|\xi_{\text{obs}}; \theta_j)$ for $j = 1, \ldots, m\,(= 441)$, where $\theta_{j_1+21\times(j_2-1)}$ denotes the grid point $(\theta^1_{j_1}, \theta^2_{j_2}) \in \Theta$ for $j_1, j_2 = 1, \ldots, 21$, and $(\theta^1_1, \ldots, \theta^1_{21})$ are evenly spaced in $[-2.5, 2.5]$ and $(\theta^2_1, \ldots, \theta^2_{21})$ 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 $\zeta^*_j = \log\{L_{\text{mis}}(\theta_j)/L_{\text{mis}}(\theta^1_{11}, \theta^2_{11})\}$ 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 computationally impractical 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 $\mathcal{N}(j) = \{l_1 + 21 \times (l_2 - 1) : l_1 = j_1 \pm 1\,(1 \le l_1 \le 21)$ 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 7.3.2.

Figure 6: 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.



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 $\theta = (\theta_{11}^1, \theta_{11}^2)$. 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.

7.3.2 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: 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 (2016b) 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 $\theta \in \Theta$.

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 knowledge is not required 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 Supplementary Material, we present additional results to illustrate the impact of using a non-optimal SA scheme, with t^{-1} is replaced by $\min(1/m, 10/t)$ in (14), and that of using a single-stage algorithm, with $t_0 = 1$ in (15).

8 Conclusion

We develop not only a sampling method, self-adjusted mixture sampling, for simulation from multiple distributions and online estimation of expectations and normal-

izing constants, but also an offline method, locally weighted histogram analysis, for estimating expectations and normalizing constants.

There are various topics that can be further studied, in addition to those mentioned earlier. For example, labeled mixture sampling can be generalized to handle multiple distributions with different dimensions, leading to the reversible jump algorithm (Green 1995). Then it is possible to generalize the use of stochastic approximation in self-adjusted mixture sampling to reversible jump MCMC for adjusting pseudo priors (e.g., Atchadé & Liu 2010). Moreover, locally weighted histogram analysis can be generalized to the trans-dimensional setting where samples are generated from multiple distributions with different dimensions (e.g., Bartolucci et al. 2006; Chen & Shao 1997). These extensions are currently under investigation.

9 Appendix

Proof of Theorem 1. Recall from (6) that $h(\theta) = \{p(2;\zeta) - \pi_2, \ldots, p(m;\zeta) - \pi_m\}^{\mathrm{T}}$. For $j, k = 2, \ldots, m$, direct calculation yields

$$\frac{\partial}{\partial \zeta_k}\{p(j;\zeta) - \pi_j\} = \begin{cases} p(j;\zeta)p(k;\zeta) & \text{if } k \neq j, \\ -p(j;\zeta) + p^2(j;\zeta) & \text{if } k = j. \end{cases}$$

Therefore, $C = -\partial h(\theta^*)/\partial\theta = \Pi_{(1)} - \pi_{(1)}\pi_{(1)}^{\mathrm{T}}$ and hence $C^{-1} = \Pi_{(1)}^{-1} + \pi_1^{-1}\mathbf{1}\mathbf{1}^{\mathrm{T}}$, where $\Pi_{(1)} = \mathrm{diag}(\pi_{(1)})$, $\pi_{(1)} = (\pi_2, \ldots, \pi_m)^{\mathrm{T}}$, and $\mathbf{1} = (1, \ldots, 1)^{\mathrm{T}}$ of dimension $m - 1$. By direct calculation, $C^{-1}\{\delta_2(L_t) - \pi_2, \ldots, \delta_m(L_t) - \pi_m\}^{\mathrm{T}} = \{\delta_2(L_t)/\pi_2, \ldots, \delta_m(L_t)/\pi_m\}^{\mathrm{T}}$ $-\delta_1(L_t)/\pi_1$, leading to the update scheme (9).

Proof of Theorem 2. Equation $h(\theta) = E_\theta\{H(Y;\theta)\}$ holds as explained after Theorem 2. Because $h(\theta)$ and C remain the same, the update scheme (12) follows by similar calculation as in the proof of Theorem 1.

Proof of Theorem 3. Let (L_1, X_1) and (L_2, X_2) be two consecutive draws under local-jump labeled mixture sampling with the invariant distribution $p(j, x; \zeta)$. If $(L_1, X_1) \sim p(j, x; \zeta)$, then $(L_2, X_2) \sim p(j, x; \zeta)$ and hence $E\{u_j(L_1, X_1; \zeta)\} = E\{E(1\{L_2 = j\}|L_1, X_1; \zeta)\} = E(1\{L_2 = j\}) = \pi_j$. The update scheme (14) follows by similar calculation as in the proof of Theorem 1.

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, 119–133.

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.

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

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

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

Chen, M.-H. and Shao, Q.-M. (1997) "Estimating ratios of normalizing constants for densities with different dimensions," *Statistica Sinica*, 7, 607–630.

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.

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

Delyon, B., Lavielle, M., and Moulines, É. (1999) "Convergence of a stochastic approximation version of the EM algorithm," *Annals of Statistics*, 27, 94128.

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

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

Doss, H. and Tan, A. (2014) "Estimates and standard errors for ratios of normalizing constants from multiple Markov chains via regeneration," *Journal of the Royal Statistical Society*, B, to appear.

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 the 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, 141150.

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.

Haario, H., Saksman, E. and Tamminen, J. (2001) "An adaptive Metropolis algorithm," *Bernoulli*, 7, 223–242.

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

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

Liu, J.S., Wong, W.H., Kong, A. (1994) "Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes," *Biometrika*, 81, 27–40.

McKeague, I.W. and Wefelmeyer, W. (2000) "Markov chain Monte Carlo and Rao–Blackwellization," *Journal of Statistical Planning and Inference*, 85,171–182.

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.

Robert, C. and Casella, G. (2005) *Monte Carlo Statistical Methods*, New York: Springer.

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. (2013) "Weak convergence rates of population versus single-chain stochastic approximation MCMC algorithms," *Advances in Applied Probability*, in press.

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. (2014) "Resampling Markov chain Monte Carlo algorithms: Basic analysis and empirical comparisons," *Journal of Computational and Graphical Statistics*, to appear.

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.

Vardi, Y. (1985) "Empirical distributions in selection bias models," *Annals of Statistics*, 13, 178–203.

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.

Supplementary Material

for "Optimally adjusted mixture sampling and locally weighted histogram analysis" by Z. Tan

1 Asymptotic theory of SA

The convergence of stochastic approximation has been studied under various conditions on the noise ε_t (e.g., Benveniste et al. 1990; Chen 2002). For the Markovian setting, there are a collection of results on both the convergence of the sequence $\{\theta_t : t \geq 1\}$ and the law of large numbers and central limit theorem for the sequence $\{Y_t : t \geq 1\}$ (e.g., Andrieu et al. 2005; Andrieu & Moulines 2006, Liang 2010; Song et al. 2013). We provide a summary of Theorems 1–2 in Song et al. (2013) 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).

Theorem S1 (Song et al. 2013). Let $\gamma_t = t_0/t^\beta$ and $A_t = \gamma_t A$ for $t_0 > 0$, $1/2 < \beta \leq 1$, and an invertible $r \times r$ matrix A. Assume that Θ is compact and the Lyapunov condition on $h(\theta)$ and the drift condition on the transition kernel K_θ hold as in Song et al. (2013). Let $\hat{H}(y;\theta)$ be a solution to the Poisson equation

$$\hat{H}(y;\theta) - K_\theta\{\hat{H}(y;\theta)\} = H(y;\theta) - h(\theta), \tag{22}$$

where $K_\theta\{\hat{H}(y;\theta)\} = \int \hat{H}(y';\theta) K_\theta(y, y')\, \mathrm{d}y'$. Then the following results hold.

(i) $d(\theta_t, \mathcal{L}) \to 0$ almost surely as $t \to \infty$, where $\mathcal{L} = \{\theta : h(\theta) = 0\}$ and $d(\theta, \mathcal{L}) = \inf_{\theta' \in \mathcal{L}} \|\theta - \theta'\|$.

(ii) In addition, assume that $h(\theta)$ is differentiable, the stability condition on $h(\theta)$ holds as in Song et al. (2013), and $-AC + I/(2t_0)$ is stable (i.e., all the eigenvalues have negative real parts), where $C = -\partial h(\theta^*)/\partial \theta^{\mathrm{T}}$ and I is the identity matrix. Then $\gamma_t^{-1/2}(\theta_t - \theta^*) \to \mathrm{N}\{0, \Sigma(A)\}$ in distribution as $t \to \infty$, where if $\beta = 1$, then

$$\Sigma(A) = \int_0^\infty \mathrm{e}^{(-AC + \frac{1}{2t_0} I)t} \, (AVA^{\mathrm{T}}) \, \mathrm{e}^{(-AC + \frac{1}{2t_0} I)^{\mathrm{T}} t} \, \mathrm{d}t,$$

and if $\frac{1}{2} < \beta < 1$, then

$$\Sigma(A) = \int_0^\infty \mathrm{e}^{-ACt} \, (AVA^{\mathrm{T}}) \, \mathrm{e}^{-(AC)^{\mathrm{T}} t} \, \mathrm{d}t,$$

and $V = \lim_{t\to\infty} E(e_t e_t^{\mathrm{T}})$ with $e_t = \hat{H}(Y_t;\theta_{t-1}) - K_{\theta_{t-1}}\hat{H}(Y_{t-1};\theta_{t-1})$.

As mentioned in Song et al. (2013), the compactness of Θ can be relaxed, when the technique of varying truncations is incorporated in the SA algorithm (e.g., Chen 2002). Based on Theorem 1, we provide a simple interpretation for V and determine the minimum asymptotic variance matrix $\Sigma(A)$ for $\beta = 1$.

Corollary S1. Under the assumptions of Theorem 1, the following results hold.

(i) $t^{-1/2}\sum_{i=1}^{t} H(Y_i^*;\theta^*) \to \mathrm{N}(0, V)$ in distribution as $t \to \infty$, where $(Y_1^*, Y_2^*, \ldots)$ is a Markov chain such that $Y_t^* \sim K_{\theta^*}(Y_{t-1}^*, \cdot)$. That is, V is the asymptotic variance matrix for the normalized average of $H(\cdot;\theta^*)$ over a Markov chain generated by the SA algorithm with θ_t fixed at θ^* for all t.

(ii) If $\beta = 1$, then $\Sigma(A)$ achieves a minimum at $A = t_0^{-1}C^{-1}$, and $\Sigma(t_0^{-1}C^{-1}) = t_0^{-1}C^{-1}VC^{-1^{\mathrm{T}}}$.

As noted in Section 3.1, the optimal SA recursion (5) is, in general, infeasible because C is unknown. There are broadly two approaches available to achieving asymptotic efficiency. One approach involves estimating θ^* and C simultaneously by stochastic approximation (e.g., Gu & Zhu 2001). Alternatively, a popular approach is to use the trajectory average $\bar{\theta}_t = t^{-1}\sum_{i=1}^{t}\theta_i$, but set $\gamma_t = t_0/t^{\beta}$ decreasing more slowly than $O(t^{-1})$ for $1/2 < \beta < 1$ (e.g., Polyak & Juditsky 1992; Ruppert 1988). For the Markovian setting above, Liang (2010) showed that $\bar{\theta}_t$ is asymptotically efficient, i.e., $t^{1/2}(\bar{\theta}_t - \theta^*) \to \mathrm{N}(0, C^{-1}VC^{-1^{\mathrm{T}}})$ in distribution, even with $A = I$.

2 Technical details

Proof of Corollary S1. (i) Let

$$\begin{aligned} v(\theta_{t-1}, Y_{t-1}) &= E(e_t e_t^{\mathrm{T}}|\mathcal{F}_{t-1}) \\ &= E\{\hat{H}(Y_t;\theta_{t-1})\hat{H}^{\mathrm{T}}(Y_t;\theta_{t-1})|\mathcal{F}_{t-1}\} - K_{\theta_{t-1}}\{\hat{H}(Y_{t-1};\theta_{t-1})\}K_{\theta_{t-1}}^{\mathrm{T}}\{\hat{H}(Y_{t-1};\theta_{t-1})\}, \end{aligned}$$

where $\mathcal{F}_{t-1} = \sigma(\theta_0, Y_1, \ldots, \theta_{t-1}, Y_{t-1})$. By the proofs of Lemma A.5 in Liang (2010) and Lemma 1 in Song et al. (2013), we have

$$\frac{1}{t}\sum_{i=1}^{t} v(\theta_{i-1}, Y_{i-1}) \to E\{v(\theta^*, Y^*)\} = V,$$

almost surely, where $Y^* \sim f(\cdot;\theta^*)$. Direct calculation shows that

$$V = E\{\hat{H}(Y^*;\theta^*)\hat{H}^{\mathrm{T}}(Y^*;\theta^*)\} - E\big[K_{\theta^*}\{\hat{H}(Y^*;\theta^*)\}K_{\theta^*}^{\mathrm{T}}\{\hat{H}(Y^*;\theta^*)\}\big],$$

which is the asymptotic variance in the central limit theorem, $t^{-1/2}\sum_{i=1}^{t} H(Y_i^*;\theta^*) \to \mathrm{N}(0,V)$, by Theorem 17.4.4 in Meyn & Tweedie (1993).

(ii) The result follows directly from Theorem 3.4.1 in Chen (2002).

Proof of Theorem 4. Result (ii) follows from the discussion preceding Theorem 4 and application of result (i) with $\theta = \theta^*$. For result (ii), we prove a general result on Rao–Blackwellization for two-block MH sampling.

Let $\{(L_t, X_t) : t = 1, 2, \ldots\}$ be a two-block MH chain with an invariant distribution $p(l, x)$, such that L_t is drawn from the conditional distribution $p(\cdot|X_{t-1})$, but X_t is generated by an MH step, depending on (L_t, X_{t-1}). For a function $\delta(l)$, let $w(x) = E\{\delta(L)|X = x\}$, where $(L, X) \sim p(l, x)$. Suppose that $t^{-1/2}\sum_{i=1}^{t}\delta(L_i) \to \mathrm{N}(0, V^\delta)$ and $t^{-1/2}\sum_{i=1}^{t} w(X_i) \to \mathrm{N}(0, V^w)$ in distribution as $t \to \infty$. Then

$$V^\delta = V^w + \mathrm{var}\{\delta(L)\} + \mathrm{var}\{w(X)\}. \tag{23}$$

We provide two different proofs. Without loss of generality, assume that $E\{\delta(L)\} = 0$ and hence $E\{w(X)\} = 0$. First, if $(L_1, X_1) \sim p(l, x)$, then for $i \geq 2$,

$$\begin{aligned} E\{\delta(L_1)\delta^{\mathrm{T}}(L_i)\} &= E\big[E\{\delta(L_1)|X_1\}E\{\delta^{\mathrm{T}}(L_i)|X_1\}\big] \\ &= E\{w(X_1)\delta^{\mathrm{T}}(L_i)\} = E\big[E\{w(X_1)|X_{i-1}\}E\{\delta^{\mathrm{T}}(L_i)|X_{i-1}\}\big] \\ &= E\{w(X_1)w^{\mathrm{T}}(X_{i-1})\}. \end{aligned}$$

The first equation holds because L_1 and L_i are conditionally independent given X_1, and the third equation holds because X_1 and L_i are conditionally independent given X_{i-1}. Therefore, Rao–Blackwellization causes a one-lag delay for auto-covariances, similarly as in Liu et al. (1994) for two-block Gibbs sampling. Equation (23) then follows from the standard formula: $V^\delta = \mathrm{var}\{\delta(L_1)\} + 2\sum_{i=2}^{\infty} E\{\delta(L_1)\delta^{\mathrm{T}}(L_i)\}$ and $V^w = \mathrm{var}\{w(X_1)\} + 2\sum_{i=2}^{\infty} E\{w(X_1)w^{\mathrm{T}}(X_i)\}$, where $(L_1, X_1) \sim p(l, x)$.

Next, we give an alternative proof of (23) based on Poisson equations. Some of the results will be reused in the proof of Theorem 5. A special property of the two-block MH chain (L_t, X_t) is that the marginal sequence X_t is a Markov chain. Let

$\hat{w}(x)$ be a solution to the marginal Poisson equation $\hat{w}(x) - K\{\hat{w}(x)\} = w(x)$, where $K\{\hat{w}(x)\} = \int \hat{w}(x')K(x, x')\,\mathrm{d}x'$ and $K(x, \cdot)$ is the transition kernel for the marginal chain X_t. Then an interesting result is that $g(l, x) = \{\delta^{\mathrm{T}}(l) + \hat{w}^{\mathrm{T}}(x), \hat{w}^{\mathrm{T}}(x)\}^{\mathrm{T}}$ is a solution to the joint Poisson equation

$$g(l, x) - K^{\dagger}\{g(l, x)\} = \{\delta^{\mathrm{T}}(l), w^{\mathrm{T}}(x)\}^{\mathrm{T}},$$

where $K^{\dagger}\{g(l, x)\} = \int g(l', x')K^{\dagger}\{(l, x), (l', x')\}\,\mathrm{d}(l', x')$ and $K^{\dagger}\{(l, x), \cdot\}$ is the transition kernel for the joint chain (L_t, X_t). In fact, the desired equation follows from the fact that $K^{\dagger}\{\hat{w}(x)\} = K\{\hat{w}(x)\} = w(x)$ and $K^{\dagger}\{\delta(l)\} = w(x)$. By Meyn & Tweedie (1993, Theorem 17.4.4), the asymptotic variances V^{δ} and V^{w} are

$$\begin{aligned} V^{w} &= \mathrm{var}\{\hat{w}(X)\} - \mathrm{var}\{\hat{w}(X) - w(X)\} \\ &= -\mathrm{var}\{w(X)\} + E\{\hat{w}(X)w^{\mathrm{T}}(X)\} + E\{w(X)\hat{w}^{\mathrm{T}}(X)\}, \\ V^{\delta} &= \mathrm{var}\{\delta(L) + \hat{w}(X)\} - \mathrm{var}\{\hat{w}(X)\} \\ &= \mathrm{var}\{\delta(L)\} + E\{\hat{w}(X)w^{\mathrm{T}}(X)\} + E\{w(X)\hat{w}^{\mathrm{T}}(X)\}, \end{aligned}$$

where $(L, X) \sim p(l, x)$. The second equation for V^{δ} holds because $E\{\hat{w}(X)\delta^{\mathrm{T}}(L)\} = E\{\hat{w}(X)w^{\mathrm{T}}(X)\}$. From these expressions, equation (23) follows.

Proof of Theorem 5. First, by Corollary 1, $\Sigma_1 = C^{-1}VC^{-1^{\mathrm{T}}}$, where $C = \Pi_{(1)} - \pi_{(1)}\pi_{(1)}^{\mathrm{T}}$ and V is the asymptotic variance of $n^{-1/2}\sum_{i=1}^{n} H(Y_i^*; \theta^*)$. Write $w_{(1)}(x; \zeta) = \{w_2(x; \zeta), \ldots, w_m(x; \zeta)\}^{\mathrm{T}}$, $w_{(1)}(x) = w_{(1)}(x; \zeta^*)$, and $\bar{w}_{(1)}(x) = w_{(1)}(x) - \pi_{(1)}$. Then $H(l, x; \theta^*) = \bar{w}_{(1)}(x)$. By the proof of Theorem 4, the marginal sequence $(X_1^*, \ldots, X_n^*)$ is a Markov chain. Let $\hat{w}_{(1)}(x)$ be a solution to the Poisson equation $\hat{w}_{(1)}(x) - K_{\theta^*}\{\hat{w}_{(1)}(x)\} = \bar{w}_{(1)}(x)$. By Markov chain theory (Meyn & Tweedie 1993),

$$V = E(\hat{w}^{\otimes 2}) - E\{(\hat{w} - \bar{w})^{\otimes 2}\}, \tag{24}$$

where $g^{\otimes 2} = gg^{\mathrm{T}}$, $\bar{w} = \bar{w}_{(1)}(X)$, and $\hat{w} = \hat{w}_{(1)}(X)$ with $(L, X) \sim p(j, x; \zeta^*)$.

Second, $\tilde{\theta}_n = (\tilde{\zeta}_2^{(n)}, \ldots, \tilde{\zeta}_m^{(n)})^{\mathrm{T}}$ is a solution to $n^{-1}\sum_{i=1}^{n} w_{(1)}(X_i^*; \zeta) = 0$. By asymptotic theory of M-estimation with Markov chains (e.g., Geyer 1994; Doss & Tan 2013), $n^{1/2}(\tilde{\theta}_n - \theta^*) \to \mathrm{N}(0, \Sigma_2)$ in distribution as $n \to \infty$, where $\Sigma_2 = D^{-1}VD^{-1^{\mathrm{T}}}$, and V

and D are determined by

$$n^{-1/2}\sum_{i=1}^{n}\{w_{(1)}(X_i^*;\zeta^*)-\pi_{(1)}\}\to \mathrm{N}(0,V),$$
$$-n^{-1}\sum_{i=1}^{n}\frac{\partial}{\partial\theta^{\mathrm{T}}}w_{(1)}(X_i^*;\zeta^*)\to D=-E\left\{\frac{\partial}{\partial\theta^{\mathrm{T}}}w_{(1)}(X;\zeta^*)\right\}.$$

By direct calculation, $D=\Pi_{(1)}-\Pi_{(1)}O_{(1)}\Pi_{(1)}$, where $O_{(1)}$ is defined by

$$O=\begin{pmatrix} o_{11} & o_1^{\mathrm{T}} \\ o_1 & O_{(1)}\end{pmatrix}$$

and O is the matrix $(o_{jk})_{1\le j,k\le m}$ with

$$o_{jk}=\int\frac{\mathrm{e}^{-\zeta_j^*}q_j(x)\mathrm{e}^{-\zeta_k^*}q_k(x)}{\{\sum_{l=1}^m\pi_l\mathrm{e}^{-\zeta_l^*}q_l(x)\}^2}\,\mathrm{d}P_*,$$

and P_* the mixture distribution $\sum_{j=1}^m\pi_jP_j$. Notice that $C-D=\Pi_{(1)}O_{(1)}\Pi_{(1)}-\pi_{(1)}\pi_{(1)}^{\mathrm{T}}=\mathrm{var}\{\bar{w}_{(1)}(X)\}$ and hence $C-D$ is nonnegative definite.

Finally, $\hat{\theta}_n=(\hat{\zeta}_2^{(n)},\ldots,\hat{\zeta}_m^{(n)})^{\mathrm{T}}$ is a solution to $n^{-1}\sum_{i=1}^n w'_{(1)}(X_i^*;\zeta)=0$, where $w'_{(1)}(X_i^*;\zeta)=\{w'_2(X_i^*;\zeta),\ldots,w'_m(X_i^*;\zeta)\}^{\mathrm{T}}$ and

$$w'_j(X_i^*;\zeta)=\frac{\pi_j\mathrm{e}^{-\zeta_j}q_j(X_i^*)}{\sum_{l=1}^m\hat{\pi}_l\mathrm{e}^{-\zeta_l}q_l(X_i^*)}-\pi_j,\quad j=2,\ldots,m.$$

Then by standard asymptotic arguments, $n^{1/2}(\hat{\theta}_n-\theta^*)\to\mathrm{N}(0,\Sigma_3)$ in distribution as $n\to\infty$, where $\Sigma_3=D^{-1}V'D^{-1\mathrm{T}}$ and V' is determined by

$$n^{-1/2}\sum_{i=1}^{n}\{w'_{(1)}(X_i^*;\zeta^*)-\pi_{(1)}\}\to\mathrm{N}(0,V').$$

By Taylor expansion for $(\hat{\pi}_2,\ldots,\hat{\pi}_m)$ about $(\pi_2,\ldots,\pi_m)$, we have

$$\frac{1}{n}\sum_{i=1}^{n}w'_j(X_i^*;\zeta^*)=\frac{1}{n}\sum_{i=1}^{n}w_j(X_i^*;\zeta^*)$$
$$-\sum_{k=2}^{m}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\pi_j\mathrm{e}^{-\zeta_j^*}q_j(X_i^*)\{\mathrm{e}^{-\zeta_k^*}q_k(X_i^*)-\mathrm{e}^{-\zeta_1^*}q_1(X_i^*)\}}{\{\sum_{l=1}^m\pi_l\mathrm{e}^{-\zeta_l^*}q_l(X_i^*)\}^2}\right](\hat{\pi}_k-\pi_k)+o_p(n^{-1/2}),$$

where $\hat{\pi}_1=1-\sum_{j=2}^m\hat{\pi}_j$, treated as a function of $(\hat{\pi}_2,\ldots,\hat{\pi}_m)$. Applying the law of large numbers to the average in the square brackets above gives

$$\frac{1}{n}\sum_{i=1}^{n}w'_{(1)}(X_i^*;\zeta^*)=\frac{1}{n}\sum_{i=1}^{n}w_{(1)}(X_i^*;\zeta^*)$$
$$-\{\Pi_{(1)}(O_{(1)}-o_1\mathbf{1}^{\mathrm{T}})\}\times\frac{1}{n}\sum_{i=1}^{n}\{\delta_{(1)}(L_i^*)-\pi_{(1)}\}+o_p(n^{-1/2}),$$

where $\delta_{(1)}(L_i^*) = \{\delta_2(L_i^*), \ldots, \delta_m(L_i^*)\}^{\mathrm{T}}$. By the expression $C^{-1} = \Pi_{(1)}^{-1} + \pi_1^{-1}\mathbf{1}\mathbf{1}^{\mathrm{T}}$ from the proof of Theorem 1, we have $\Pi_{(1)}(O_{(1)} - o_1\mathbf{1}^{\mathrm{T}}) = I - DC^{-1}$, denoted by B. Therefore, V' is the asymptotic variance of $n^{-1/2}\sum_{i=1}^n [w_{(1)}(X_i^*;\zeta^*) - \pi_{(1)} - B\{\delta_{(1)}(L_i^*) - \pi_{(1)}\}]$. Now, by the proof of Theorem 4, $\{\bar{\delta}_{(1)}^{\mathrm{T}}(l) + \hat{w}_{(1)}^{\mathrm{T}}(x), \hat{w}_{(1)}^{\mathrm{T}}(x)\}$ is a solution to the joint Poisson equation for $\{\delta_{(1)}^{\mathrm{T}}(l), w_{(1)}(x)\}$, where $\bar{\delta}_{(1)}(l) = \delta_{(1)}(l) - \pi_{(1)}$ and, as before, $w_{(1)}(x) = w_{(1)}(x;\zeta^*)$. Then $\hat{w}_{(1)}(x) - B\{\bar{\delta}_{(1)}(l) + \hat{w}_{(1)}(x)\}$ is a solution to the Poisson equation for $w_{(1)}(x) - B\delta_{(1)}(l)$. By Markov chain theory (Meyn & Tweedie 1993) and direct calculation, we have

$$\begin{aligned}
V' &= E[\{\hat{w} - B(\bar{\delta} + \hat{w})\}^{\otimes 2}] - E[\{\hat{w} - B(\bar{\delta} + \hat{w}) - (\bar{w} - B\bar{\delta})\}^{\otimes 2}] \\
&= E(\hat{w}\bar{w}^{\mathrm{T}} + \bar{w}\hat{w}^{\mathrm{T}} - \bar{w}^{\otimes 2}) + E\{B(\bar{\delta}^{\otimes 2} + \hat{w}\bar{w}^{\mathrm{T}} + \bar{w}\hat{w}^{\mathrm{T}})B^{\mathrm{T}}\} \\
&\quad - E\{(\hat{w}\bar{w}^{\mathrm{T}} + \bar{w}\hat{w})B^{\mathrm{T}} + B(\bar{w}\hat{w}^{\mathrm{T}} + \hat{w}\bar{w}^{\mathrm{T}})\} \\
&= -E(\bar{w}^{\otimes 2} - B\bar{\delta}^{\otimes 2}B^{\mathrm{T}}) + DC^{-1}E(\hat{w}\bar{w}^{\mathrm{T}} + \bar{w}\hat{w})C^{-1}D \\
&= -(I - DC^{-1})D(I - C^{-1}D) + DC^{-1}VC^{-1}D,
\end{aligned}$$

where $\bar{\delta} = \bar{\delta}_{(1)}(L)$, the second equation holds because $E(\hat{w}\bar{\delta}^{\mathrm{T}}) = E(\hat{w}\bar{w}^{\mathrm{T}})$, and the fourth equation holds because $E(\hat{w}\bar{w}^{\mathrm{T}} + \bar{w}\hat{w}) = V + E(\bar{w}^{\otimes 2})$, $E(\bar{w}^{\otimes 2} - B\bar{\delta}^{\otimes 2}B^{\mathrm{T}}) = DC^{-1}E(\bar{w}^{\otimes 2})$ and $E(\bar{w}^{\otimes 2}) = C - D$. The expression for Σ_3 follows.

Proof of Corollary 1. First, we derive the solution $\hat{w}_{(1)}(x)$ to the Poisson equation for $w_{(1)}(x) = w_{(1)}(x;\zeta^*)$, if $\Psi_j(x, x')$ corresponds to drawing x' from P_j, independently of x, for $j = 1, \ldots, m$. By direct calculation, $K_{\theta^*}\{w_j(x;\zeta^*)\} = \sum_{k=1}^m w_k(x;\zeta^*)(\pi_j o_{jk})$, and hence $K_{\theta^*}\{w_j(x;\zeta^*) - \pi_j\} = \sum_{k=1}^m \{w_k(x;\zeta^*) - \pi_k\}(\pi_j o_{jk})$ because $\sum_{k=1}^m \pi_k o_{jk} = 1$. Arranging the expressions for $j = 1, \ldots, m$ gives

$$K_{\theta^*}\{\bar{w}_{(1)}(x)\} = (\Pi_{(1)}O_{(1)} - \Pi_{(1)}o_1\mathbf{1}^{\mathrm{T}})\bar{w}_{(1)}(x) = (I - DC^{-1})\bar{w}_{(1)}(x),$$

by the expression of $I - DC^{-1}$ in the proof of Theorem 5. Then $\hat{w}_{(1)}(x) = \{I - (I - DC^{-1})\}^{-1}\bar{w}_{(1)}(x) = CD^{-1}\bar{w}_{(1)}(x)$. The expression of V in (24) is

$$V = CD^{-1}(C - D)D^{-1}C - (CD^{-1} - I)(C - D)(D^{-1}C - I).$$

Second, similar calculations as in the proof of Theorem 5 yield

$$\begin{aligned}\Sigma_2^{\mathrm{ind}} &= D^{-1}E\{\bar{w}_{(1)}^{\otimes 2}(X)\}D^{-1} = D^{-1}(C-D)D^{-1},\\ \Sigma_3^{\mathrm{ind}} &= D^{-1}\big[E\{\bar{w}_{(1)}^{\otimes 2}(X)\} - BE\{\bar{\delta}_{(1)}^{\otimes 2}(L)\}B^{\mathrm{T}}\big]D^{-1}\\ &= \Sigma_2^{\mathrm{ind}} - (D^{-1}-C^{-1})C(D^{-1}-C^{-1}) = D^{-1}-C^{-1},\end{aligned}$$

where $(L, X) \sim p(j, x; \zeta^*)$.

Finally, by direct calculation, we have

$$\begin{aligned}\Sigma_1 &= C^{-1}VC^{-1} = \Sigma_2^{\mathrm{ind}} - (D^{-1}-C^{-1})(C-D)(D^{-1}-C^{-1})\\ &= \Sigma_3^{\mathrm{ind}} + (D^{-1}-C^{-1})D(D^{-1}-C^{-1}) = 2D^{-1} - 3C^{-1} + C^{-1}DC^{-1}.\end{aligned}$$

The third equation shows that $\Sigma_3 = \Sigma_3^{\mathrm{ind}}$ by Theorem 5. Moreover, we have

$$\begin{aligned}\Sigma_2 &= D^{-1}VD^{-1} = 2D^{-1}CD^{-1}CD^{-1} - 3D^{-1}CD^{-1} + D^{-1}\\ &= \Sigma_2^{\mathrm{ind}} + 2(D^{-1}C - I)D^{-1}(CD^{-1} - I).\end{aligned}$$

The inequalities in Corollary 1 follow easily.

3 Additional simulation results

Figures S1–S5 give trace plots up to different time points from single simulations and S6–S7 give the summary of estimates from repeated simulations, for self-adjusted mixture sampling in Section 7.1. Figure S8 and S9–S11 give additional plots of self-adjusted mixture sampling in Sections 7.2 and, respectively, 7.3.

There are interesting features in Figure S1, in addition to those discussed in Section 7.1.2. As seen from the first row of plots, the free energy estimates $\zeta^{(t)}$ fall remarkably close to the truth (as indicated by the final estimates), very quickly in about 200 iterations before subsampling (i.e., 20 iterations after subsampling). During this period, the energy per spin, u_t/K, has not even reached the modal area near -1.7 corresponding to the ordered phase, although the coldest temperature has been visited with $L_t = 5$. The energy sequence u_t/K has occasional jumps between the two modes (e.g., from -0.85 to -1.7), in which case the temperature label L_t is seen to move in the corresponding direction (e.g. from 1 to 5). Such jumps are crucial to the effectiveness of labeled mixture sampling.

Figure S4 or S5, similarly to Figure S1, shows the trace plots for a single run of self-adjusted mixture sampling, but with $t_0 = 1$ or $t_0 = 2.2 \times 10^6$ (the length of simulation before subsampling), i.e., the gain factor t^{-1} is replaced by $\min(1/m, t^{-1})$ or $\min(1/m, t^{-0.8})$, a slow-decaying gain factor, in the update scheme (9). There are more stable trajectories of free energy estimates $\zeta^{(t)}$ and less controlled deviations of observed proportions $\hat{\pi}_j$ from π_j in Figure S4 than in Figure S5. These differences are similar to those between the optimal scheme and non-optimal scheme (with a gain factor 10 times the optimal one) in Figures S1 and S2,

Figure S10 shows the output from a single run of self-adjusted mixture sampling with a non-optimal SA scheme as in Section 7.1, where the gain factor t^{-1} is replaced by $\min(1/m, 10/t)$ in the update scheme (14). Similarly to the differences between Figures S1 and S2, there are greater fluctuations along the trajectories of $\zeta^{(t)}$, but smaller discrepancies between observed and target proportions, $\hat{\pi}_j$ and π_j, than in Figure S9 for the optimal SA scheme. For this algorithm, the three estimators, $\zeta^{(n)}$, $\bar{\zeta}^{(n)}$, or $\hat{\zeta}^{(n)}$, are found to yield $10^3 \times \text{MSE} = 68.1$, 15.5, or 0.305 respectively, based on 100 repeated simulations as in Table 1.

Figure S11 shows the output from a single run of self-adjusted mixture sampling with $t_0 = 1$, hence a single-stage algorithm with the gain factor t^{-1} replaced by $\min(1/m, t^{-1})$ in (14). This algorithm fails to converge properly by the end of 441×550 iterations: some observed proportions $\hat{\pi}_j$ are zero or seriously below the target ones, and some estimates $\zeta_j^{(t)}$ stay far away from the truth for all t. As discussed at the end of Section 3.2, such failures, especially with a large m, can be effectively prevented by adopting the two-stage modification (15).

References

Andrieu, C., Moulines, É., and Priouret, P. (2005) "Stability of stochastic approximation under verifiable conditions," *SIAM Journal on Control and Optimization*, 44, 283–312.

Benveniste, A., Métivier, M., and Priouret, P. (1990) *Adaptive Algorithms and Stochastic Approximations*, New York: Springer.

Liang, F. (2010) "Trajectory averaging for stochastic approximation MCMC algorithms," *Annals of Statistics*, 38, 2823–2856.

Meyn, S.P. and Tweedie, R.L. (1993) *Markov Chains and Stochastic Stability*, London: Springer.

Polyak, B.T. and Juditsky, A.B. (1992) "Acceleration of stochastic approximation by averaging," *SIAM Journal on Control and Optimization*, 30, 838–855.

Ruppert, D. (1988) "Efficient estimators from a slowly convergent RobbinsMonro procedure," *Technical Report 781*, School of Operations Research and Industrial Engineering, Cornell University.

Figure S1: 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.

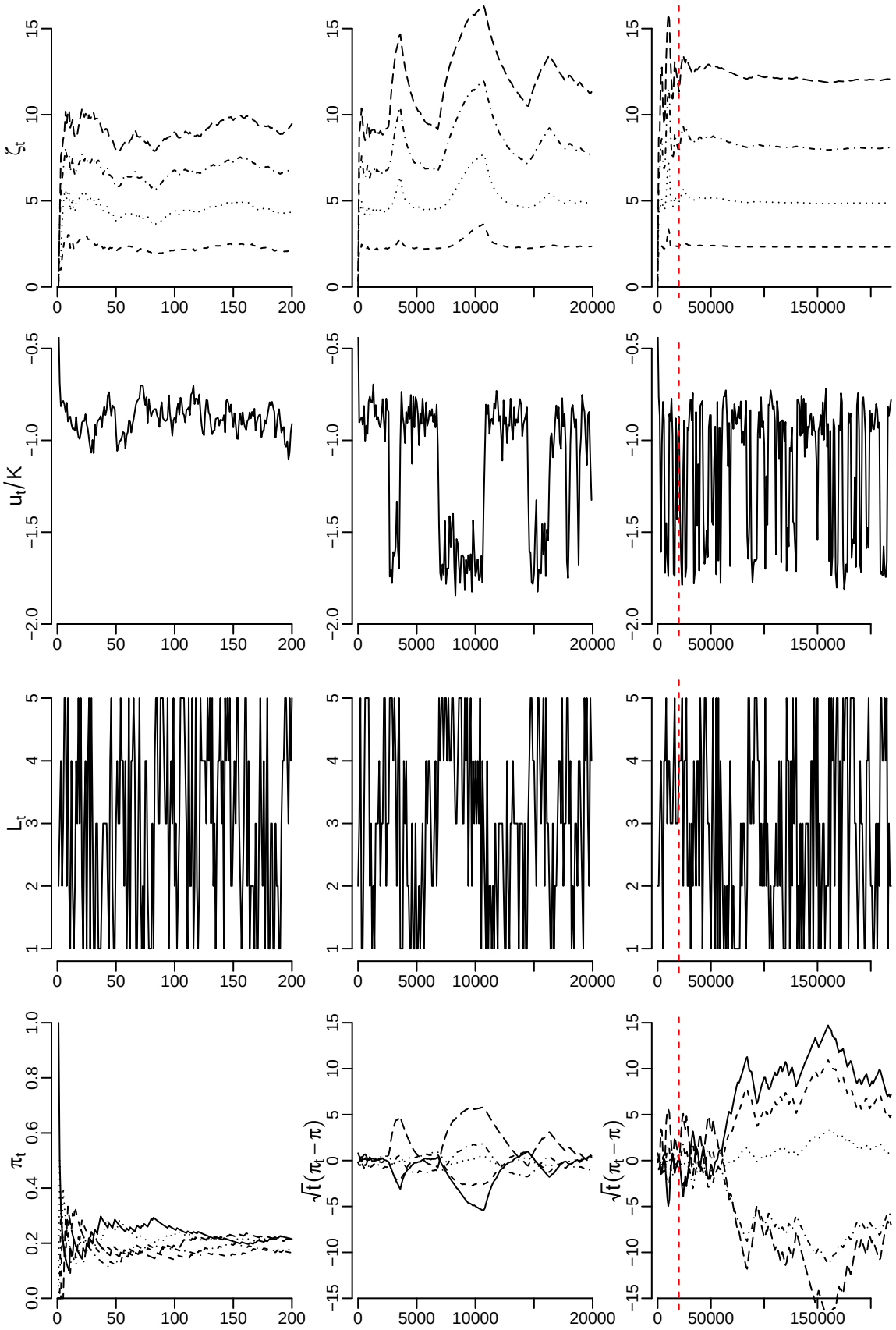


Figure S2: Trace plots for self-adjusted mixture sampling with a non-optimal SA scheme. The number of iterations is shown after subsampling. A vertical line is placed at burn-in.

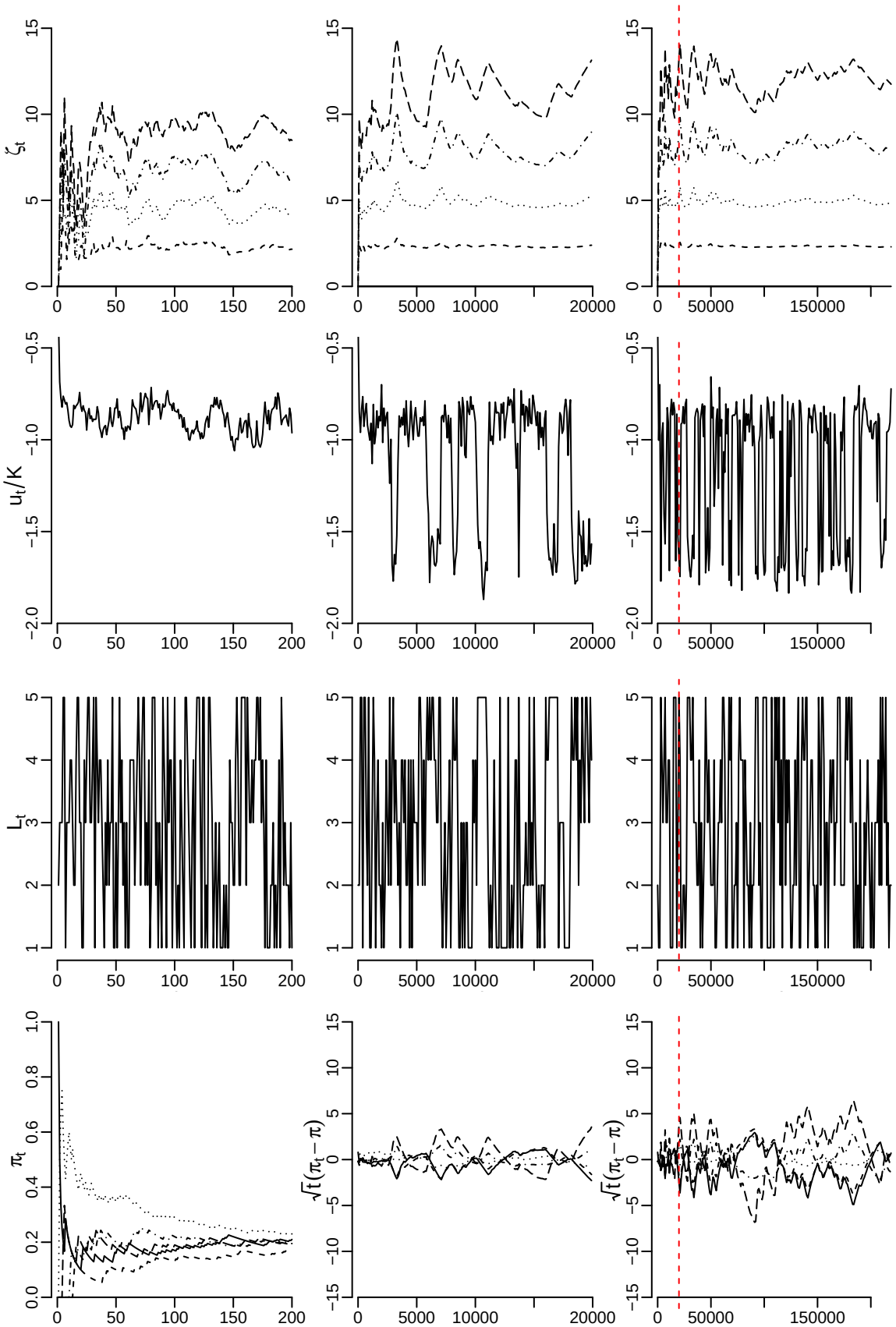


Figure S3: Trace plots for self-adjusted mixture sampling with the flat-histogram scheme. The number of iterations is shown after subsampling. A vertical line is placed at burn-in.

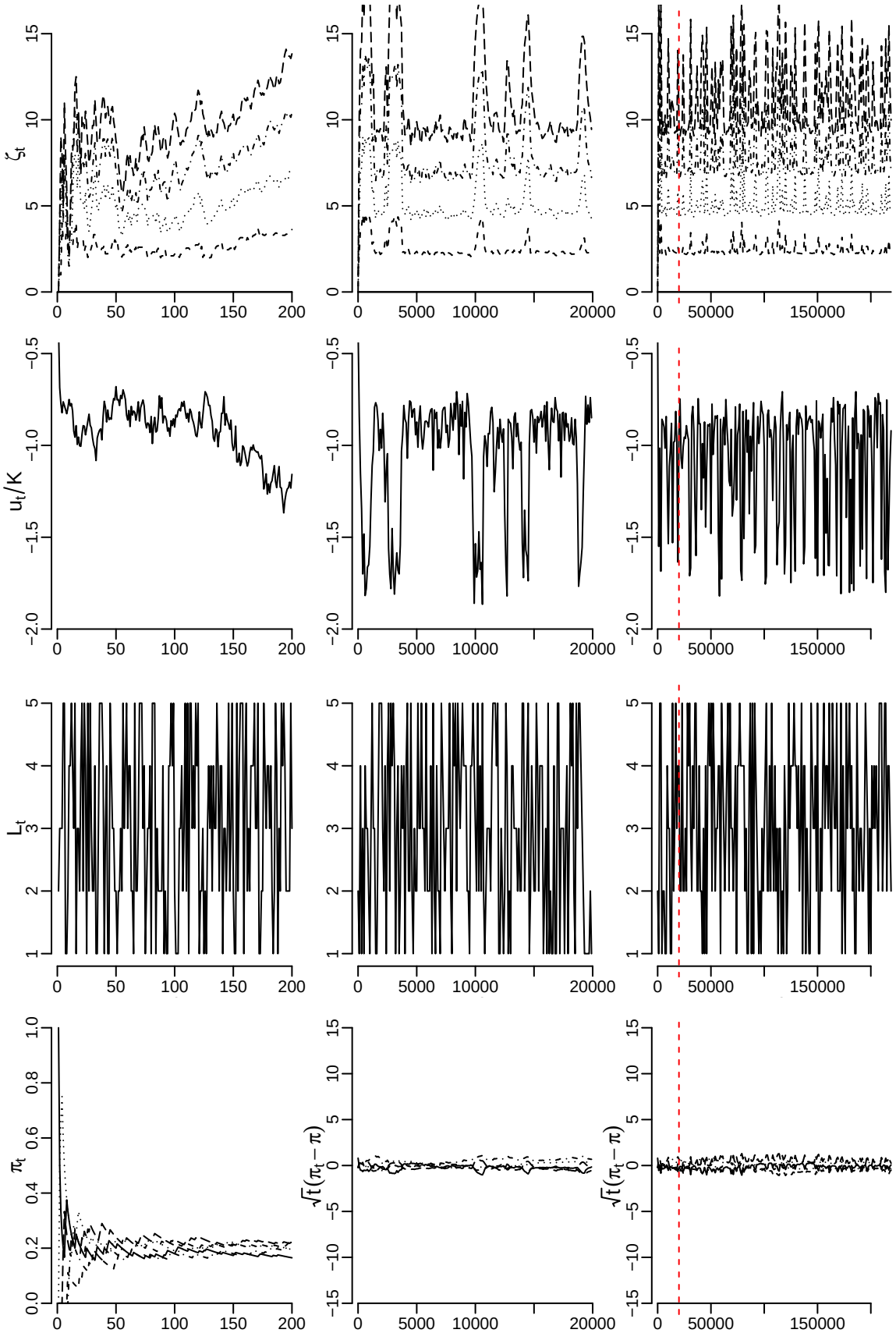


Figure S4: Trace plots for self-adjusted mixture sampling with $t_0 = 1$. The number of iterations is shown after subsampling. A vertical line is placed at the burn-in size.

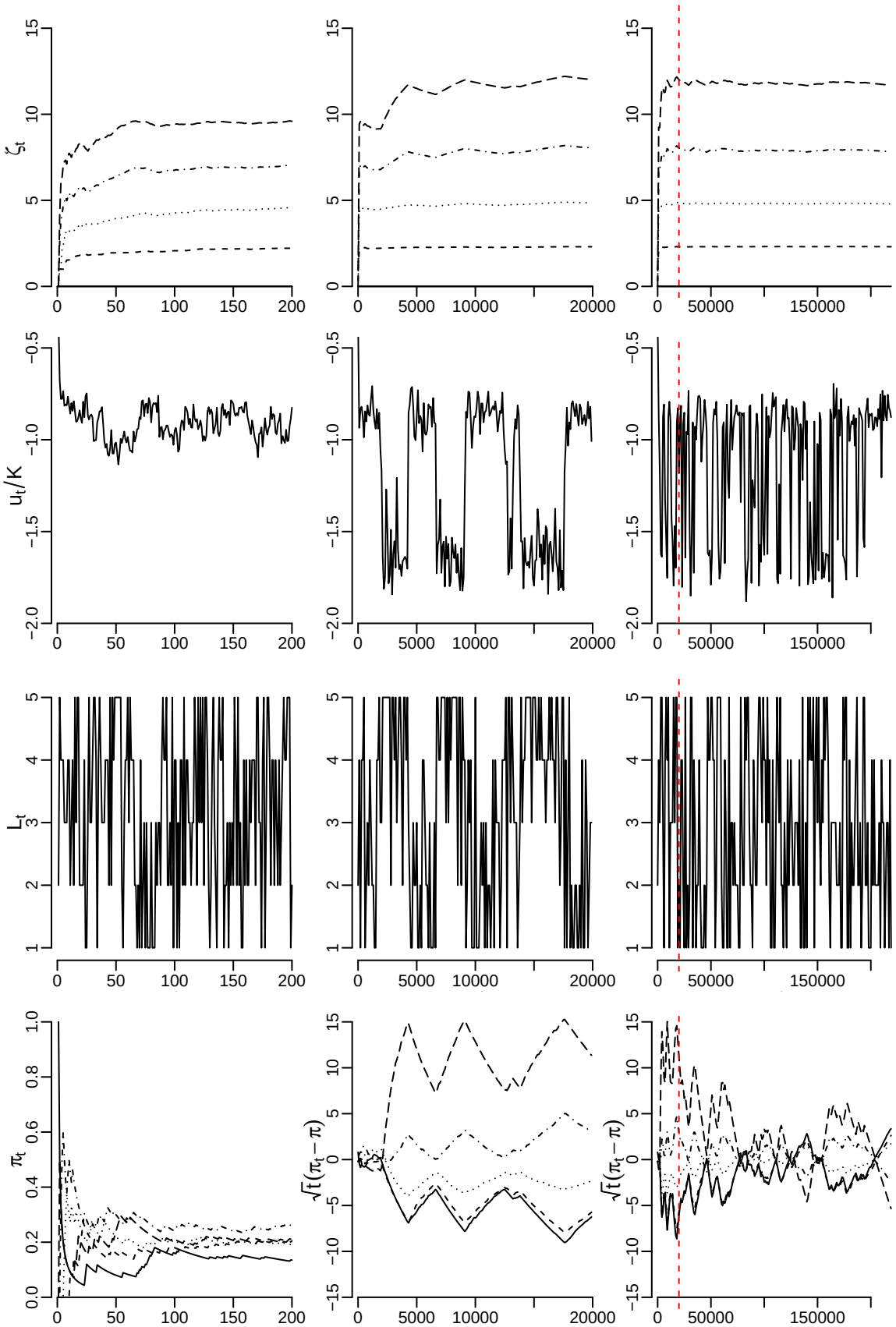


Figure S5: Trace plots for self-adjusted mixture sampling with $t_0 = 2.2 \times 10^6$. The number of iterations is shown after subsampling. A vertical line is placed at the burn-in size.

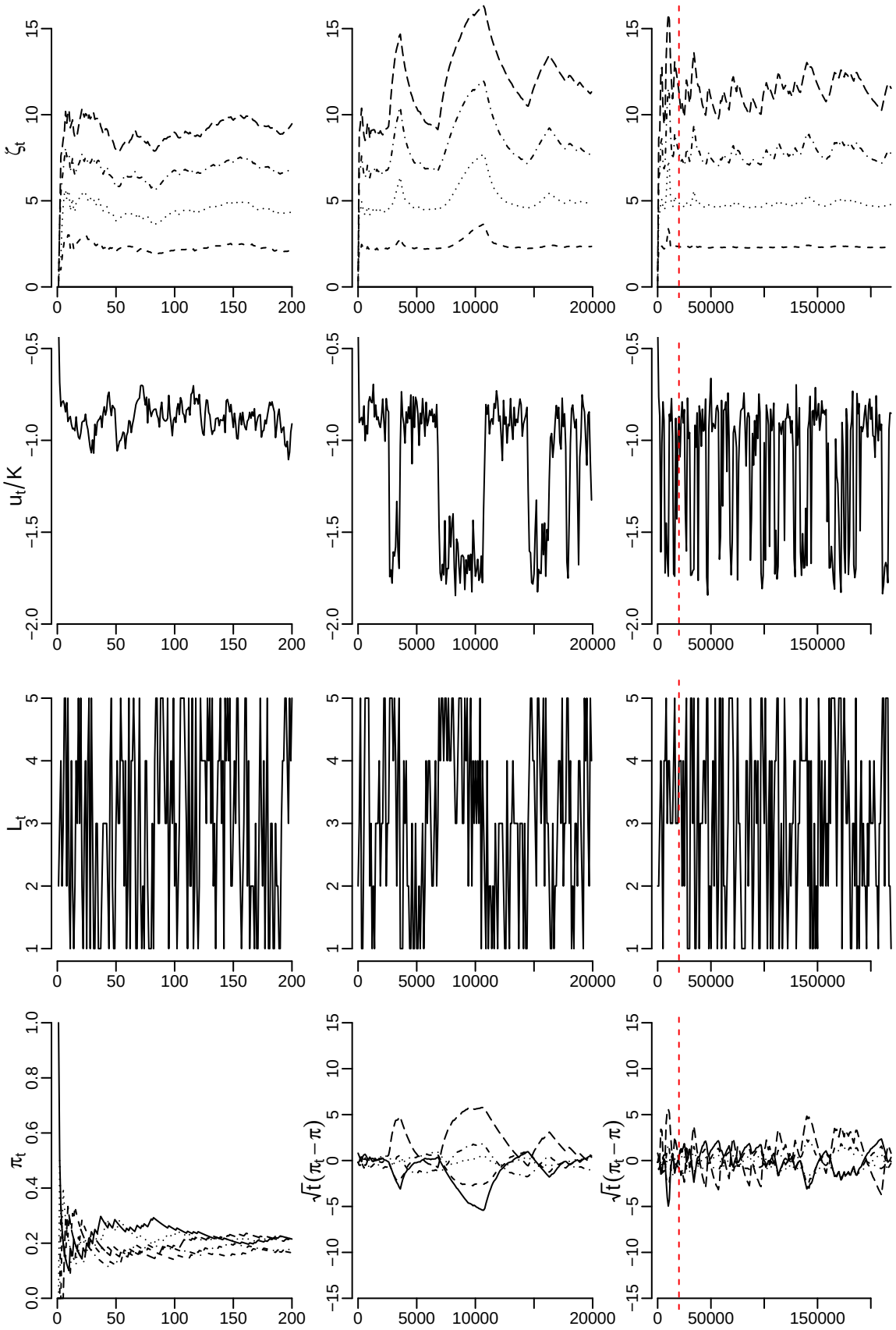


Figure S6: Summary of estimates at the temperatures $(T_1, \ldots, T_5)$ labeled as $1, \ldots, 5$, similarly to Figure 3 except for the following changes. For the top plots, the estimates are shown for self-adjusted mixture sampling with $t_0 = 2.2 \times 10^5$ in optimal scheme (9) ($\circ$) or non-optimal SA scheme ($\triangle$). For the bottom plots, the estimates are shown for self-adjusted mixture sampling with $t_0 = 2.2 \times 10^5$ in optimal scheme (9) ($\circ$) or flat-histogram scheme ($\triangle$). For free energies, locally weighted offline estimates are shown for parallel tempering ($\times$), and locally weighted offline (left) and online (right) estimates are shown for self-adjusted mixture sampling ($\circ$ or $\triangle$) in the top and bottom plots.

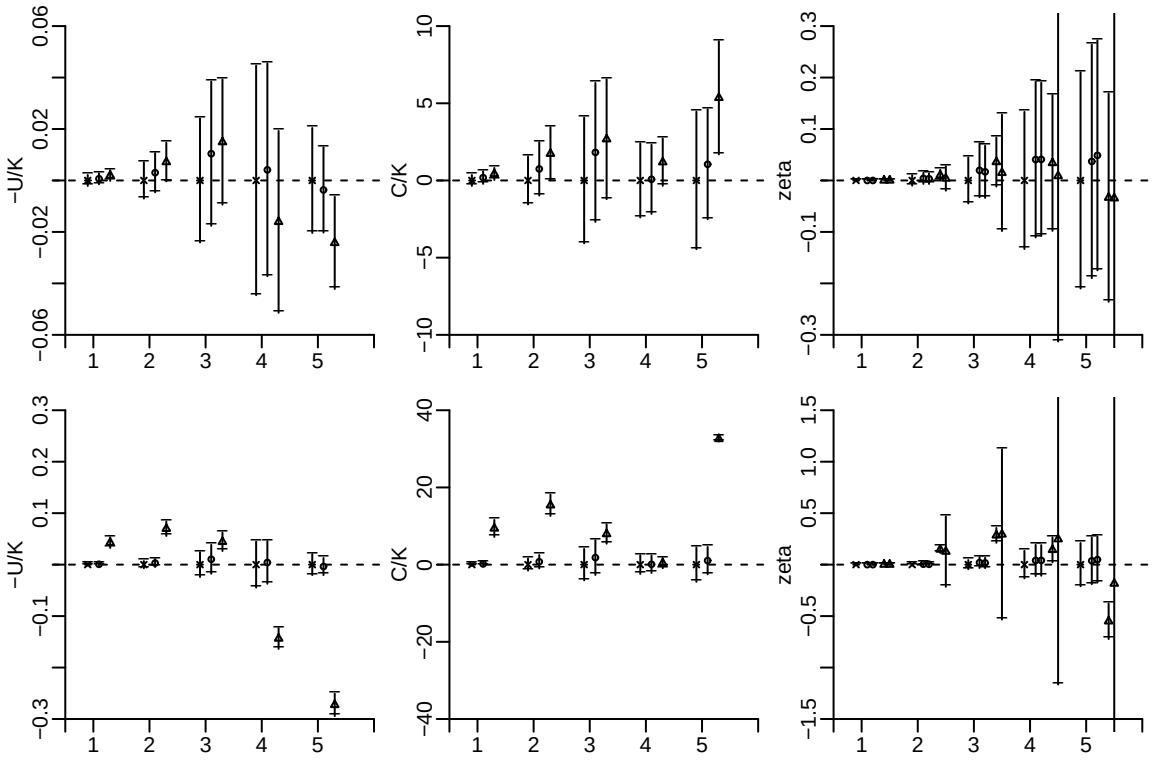


Figure S7: Summary of estimates at the temperatures $(T_1, \ldots, T_5)$ labeled as $1, \ldots, 5$, similarly to Figure 3 except for the following changes. For the top plots, the estimates are shown for self-adjusted mixture sampling with $t_0 = 1$ ($\circ$) or $t_0 = 2.2 \times 10^6$ ($\triangle$) in (9). For the bottom plots, the estimates are shown for self-adjusted mixture sampling with $t_0 = 2 \times 10^5$ and local jump and update scheme (14) ($\circ$) or global jump and update scheme (12) ($\triangle$). For free energies, globally weighted offline (left) and online (right) estimates are shown for the global version of self-adjusted mixture sampling ($\triangle$).

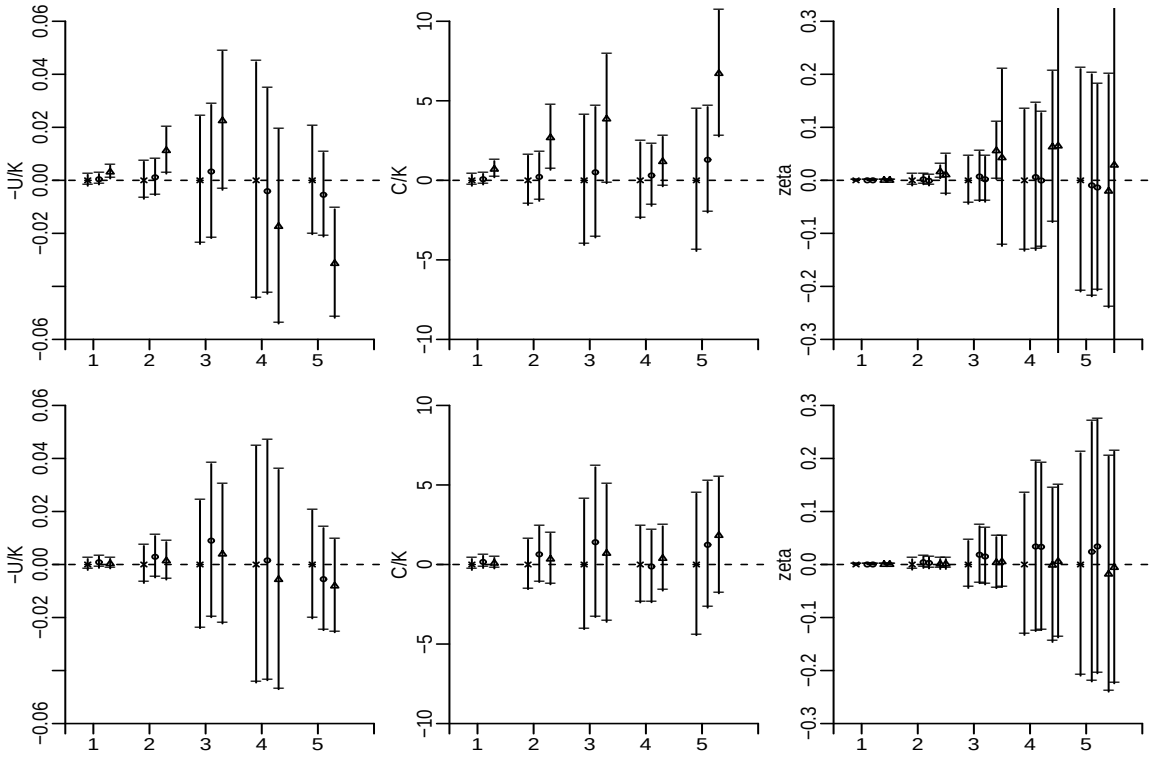


Figure S8: Trace plots and plots of observed weights at the end of simulation for optimally adjusted Wang–Landau algorithm (top) and flat-histogram Wang–Landau algorithm.

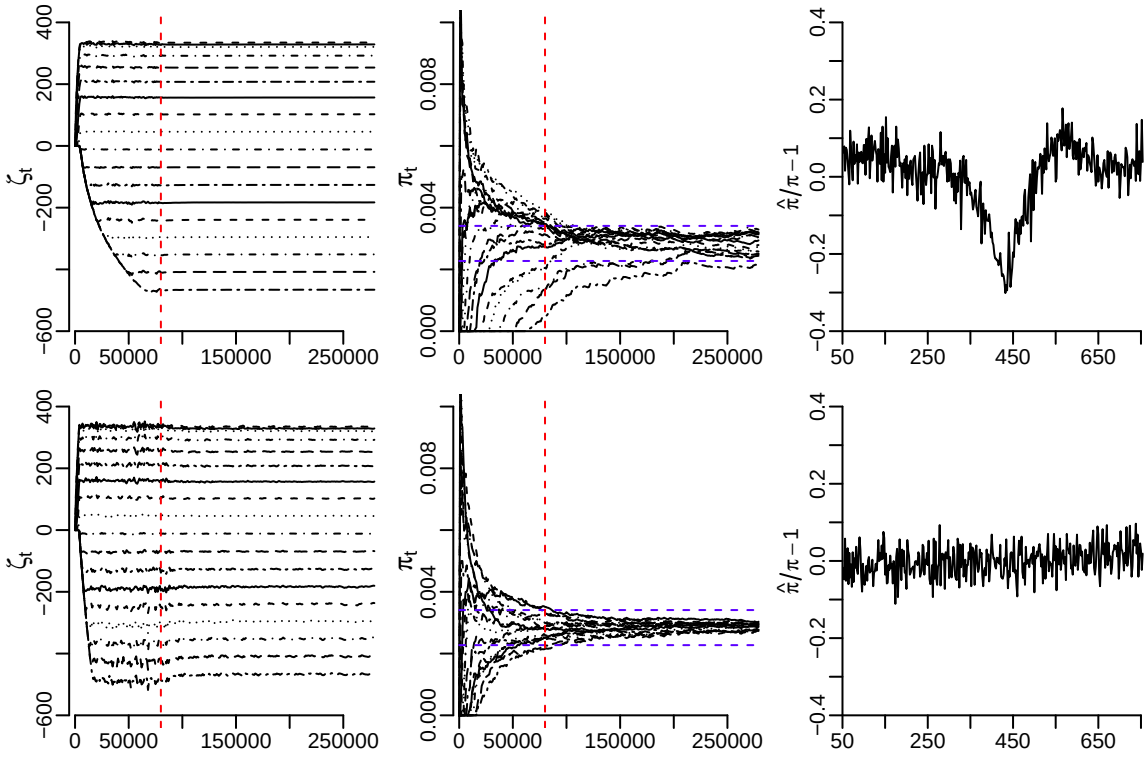


Figure S9: Output from self-adjusted mixture sampling with $t_0 = 441 \times 50$. On the top are the trace plots of online estimates of ζ_j^* and $\hat{\pi}_j$ for $\theta_j = (\theta_{j_1}^1, \theta_{j_2}^2)$ with $j_1, j_2 = 1, 11$, or 21. A vertical line is placed at burn-in and two horizontal lines are placed 20% away rom the target weight $1/m$. On the bottom are 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.

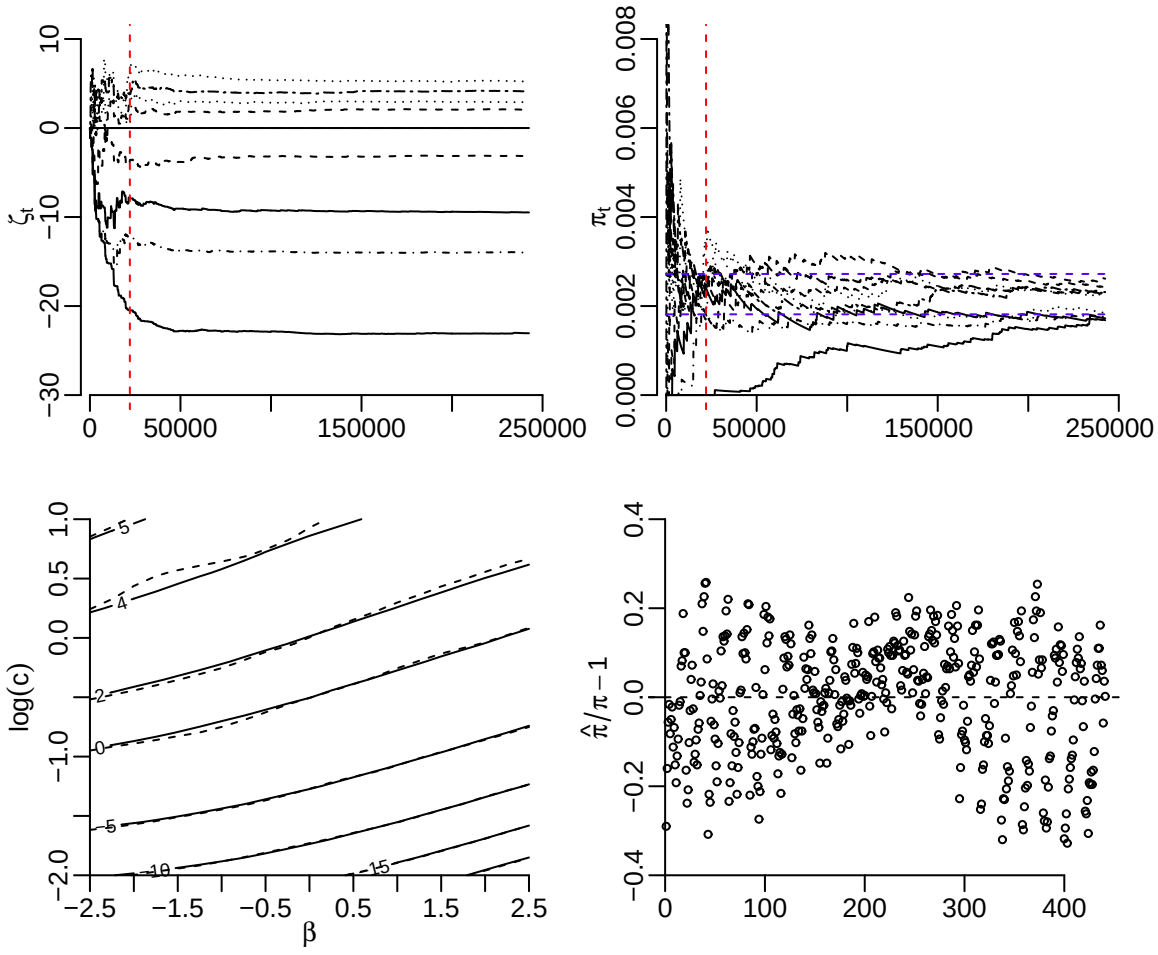


Figure S10: Output from self-adjusted mixture sampling with a non-optimal SA scheme.

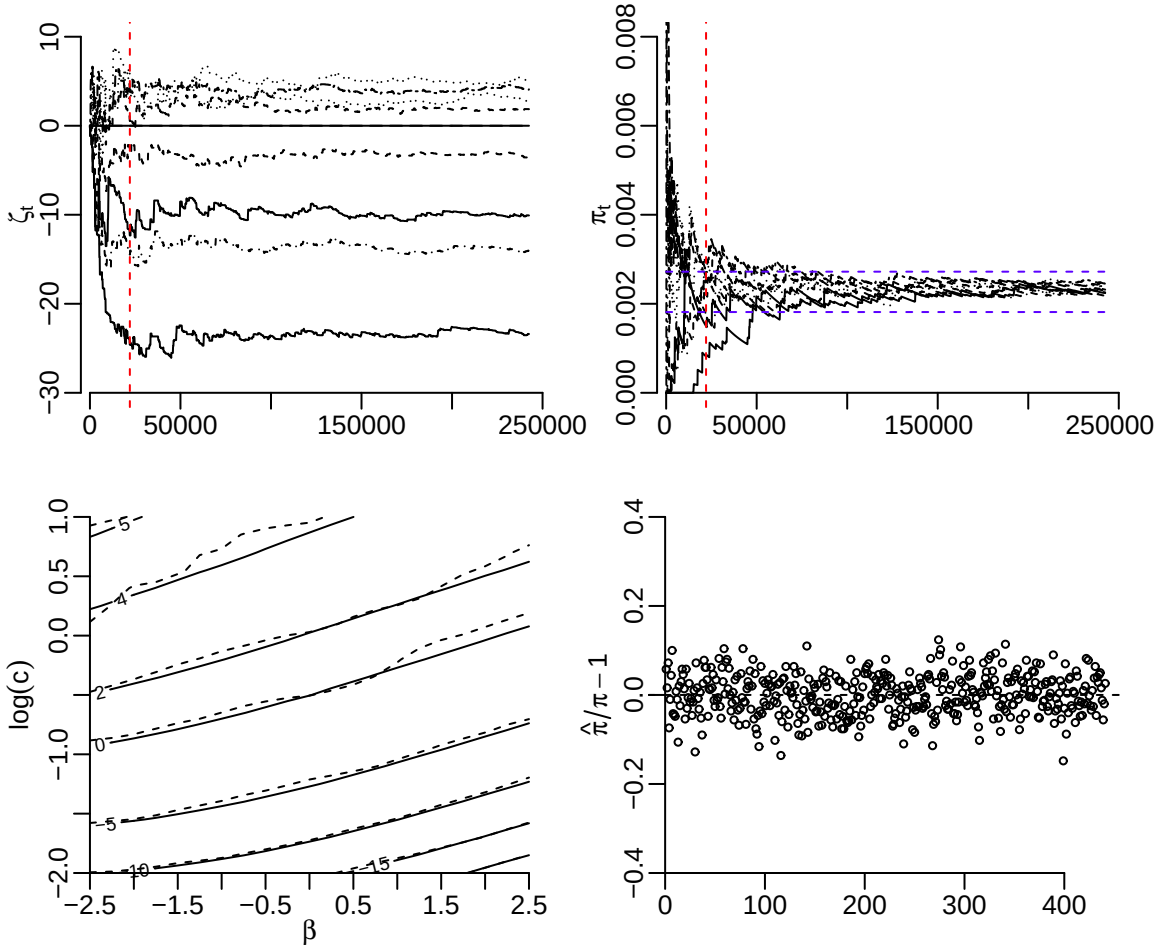


Figure S11: Output from self-adjusted mixture sampling with $t_0 = 1$.