



Published in final edited form as:

SIAM/ASA J Uncertain Quantif. 2020 ; 8(3): 1139–1188. doi:10.1137/18M122964X.

Stratification as a general variance reduction method for Markov chain Monte Carlo

Aaron R. Dinner[†], Erik H. Thiede[‡], Brian Van Koten[§], Jonathan Weare[¶]

[†]Department of Chemistry and James Franck Institute, the University of Chicago, Chicago, IL 60637

[‡]Department of Chemistry and James Franck Institute, the University of Chicago, Chicago, IL 60637

[§]Department of Mathematics and Statistics, the University of Massachusetts, Amherst, MA 01003

[¶]Courant Institute of Mathematical Sciences, New York University, New York, NY 10012

Abstract

The Eigenvector Method for Umbrella Sampling (EMUS) [46] belongs to a popular class of methods in statistical mechanics which adapt the principle of stratified survey sampling to the computation of free energies. We develop a detailed theoretical analysis of EMUS. Based on this analysis, we show that EMUS is an efficient *general* method for computing averages over arbitrary target distributions. In particular, we show that EMUS can be dramatically more efficient than direct MCMC when the target distribution is multimodal or when the goal is to compute tail probabilities. To illustrate these theoretical results, we present a tutorial application of the method to a problem from Bayesian statistics.

1. Introduction.

Markov chain Monte Carlo (MCMC) methods have been widely used with great success throughout statistics, engineering, and the natural sciences. However, when estimating tail probabilities or when sampling from multimodal distributions, accurate MCMC estimates often require a prohibitively large number of samples. In this article, we analyze the Eigenvector Method for Umbrella Sampling (EMUS) [46]. We first proposed EMUS as a method for computing free energies, and we demonstrated that it was useful for treating the multimodality that typically arises in that context. Here, we demonstrate that EMUS is an effective *general* means of addressing the challenges posed not just by multimodality but also tail events, with potential applications to a broad range of problems in statistics, engineering, and the natural sciences.

EMUS was inspired by Umbrella Sampling [47] and other methods such as the Weighted Histogram Analysis Method (WHAM) [27] and the Multistate Bennett Acceptance Ratio (MBAR) [42] for computing potentials of mean force and free energies in statistical mechanics.¹ We call these *stratified MCMC methods* since they each adapt the principle of

stratified survey sampling to MCMC simulation. Stratified MCMC methods are among the most powerful, most successful, and most widely used tools in molecular simulation. (However, in contrast to our presentation here, they are not typically used in molecular simulation to compute averages of general observables.) WHAM, for example, has been instrumental for treating biomolecular processes ranging from protein folding [7] to conductance by ion channels [3].

While the practical utility of stratification has been established in many applications, the advantages and disadvantages of the method have remained poorly understood; cf. Remark 3.8 and [46]. Motivated by the substantial gap between theory and application of stratified MCMC within statistical mechanics, and also by the general challenges posed by multimodality and tail probabilities, the goal of this paper is to develop a clear theoretical explanation of the advantages of EMUS. Our theory suggests new applications of stratified MCMC (and EMUS in particular) to broad classes of sampling problems arising in statistics and statistical mechanics. For example, very recently EMUS was successfully applied to a parameter estimation problem in cosmology [34].

We now describe EMUS and its relationship to other MCMC methods. The EMUS algorithm proceeds roughly as follows:

1. We divide the support of the target distribution into regions called *strata*. Associated to each stratum, we define a *biased distribution* whose support lies within that stratum. For example, one might let the biased distribution corresponding to a stratum be the target distribution conditioned on the stratum.
2. We use MCMC to sample the biased distributions.
3. We weight the samples from each stratum to compute estimates of general averages with respect to the target distribution.

EMUS belongs to a large class of MCMC methods that by various mechanisms promote a more uniform sampling of space. For example, in parallel tempering [44, 16], one uses MCMC samples drawn from a distribution or sequence of distributions close to the uniform distribution to speed sampling of the target distribution. The bias introduced by the choice of distributions is corrected either by reweighting the samples or by a replica exchange strategy [16]. The Wang–Landau [50] and Metadynamics methods [28] adaptively construct a biased distribution to achieve uniform sampling in certain coordinates. The temperature accelerated molecular dynamics method [33] is also designed to achieve uniform sampling in a given coordinate, but it works by entirely different means. In EMUS and other stratified MCMC methods, one achieves more uniform sampling by ensuring that each stratum contains points from at least one MCMC simulation.

EMUS also resembles certain methods for computing normalization constants of families of probability densities [17, 49, 35, 26]. The resemblance arises because the weights in the third step of EMUS are the normalization constants of the biased distributions. These

¹A potential of mean force is the logarithm of a marginal density. A free energy is the logarithm of a normalization constant. Both quantities play fundamental roles in statistical mechanics, e.g. in theories of rates of chemical reactions.

methods have been used, for example, to compute Bayes factors in model selection problems [17] and for computations related to selection bias models [49]. However, despite a strong formal resemblance, EMUS has entirely different objectives from these methods: When computing normalization constants, the distributions analogous to our biased distributions are specified as part of the problem. By contrast, in EMUS and other stratified MCMC methods, the strata are chosen as in stratified survey sampling to maximize efficiency. EMUS is perhaps more similar in spirit to the parallel Markov chain Monte Carlo method [48]. Like EMUS, this method is designed to make MCMC more efficient.

Summary of Main Results.

Our most general results are a central limit theorem (CLT) for the EMUS method and a convenient upper bound on the asymptotic variance, cf. Theorem 3.3 and Theorem 3.5. We note that the proof of the upper bound relies on a new class of perturbation estimates for Markov chains which we derived in [45]. These estimates are substantially more detailed than previous results [10], cf. Remark 4.2. After proving the CLT, we address the dependence of the sampling error on the choice of strata. In particular, for a representative MCMC method, we estimate the asymptotic variances of trajectory averages sampling the biased distributions, cf. Theorem 3.7. Our estimate shows how factors such as the diameters of the strata influence the asymptotic variances.

In Section 4, we apply the general theory developed in Section 3 to case studies involving tail probabilities and multimodality. Our results concern two limits: a *small probability* limit and a *low-temperature* limit. In the small probability limit, we consider estimation of probabilities of the form

$$p_M := \mathbf{P}[X \geq M].$$

For a broad class of random variables X , we show that while the cost of computing p_M with relative precision by direct MCMC increases exponentially with M , the cost by EMUS increases only polynomially; cf. Section 4.2. In the low-temperature limit, a parameter of the target distribution decreases, intensifying the effects of multimodality on the efficiency of MCMC sampling. We show that the cost of computing an average to fixed precision by direct MCMC increases exponentially in this limit, whereas the cost by EMUS increases only polynomially; cf. Section 4.1. We conclude that EMUS may be dramatically more efficient than direct MCMC sampling when the target distribution is multimodal or when the goal is to compute a small tail probability.

To illustrate our theoretical results, we present a tutorial numerical study applying EMUS to a problem in Bayesian statistics, cf. Section 5. In addition to illustrating the theory, our numerical study demonstrates the problems that may occur when EMUS and other similar stratified MCMC methods are used carelessly. It also addresses practical issues such as the choice of strata and the computation of error bars for averages estimated by EMUS.

The results in this article significantly extend and generalize the ideas in [46]. We first proposed the EMUS method with the goal of analyzing and improving umbrella sampling

approaches in free energy calculations. Here, our goal is to establish EMUS as a *general* variance reduction technique, and we present many entirely new results, including an upper bound on the asymptotic variance of EMUS (Theorem 3.5), a condition to guide some aspects of the choice of strata (Remark 3.12), a theoretical argument demonstrating the benefits of EMUS for computing tail probabilities (Section 4.2), numerical results applying EMUS to Bayesian inference (Section 5), a method of correcting problems related to poorly chosen strata (Section 5.3), and a greatly improved numerical method for estimating the standard deviations of quantities computed by EMUS (Appendix F). In addition, we give complete justifications of some results that were stated without proof in [46], including Theorem 3.7 concerning the dependence of the sampling error on the choice of strata. Finally, we note that our results concerning multimodal distributions and the low-temperature limit generalize and clarify the results given in [46]; in particular, our Theorem 4.1 covers periodic boundary conditions and stratification in more than one variable.

2. The Eigenvector Method for Umbrella Sampling.

In this section, we derive the Eigenvector Method for Umbrella Sampling (EMUS), and we prove that it is consistent. We also derive a related method, iterative EMUS, and we compare iterative EMUS with the MBAR method from statistical mechanics [42].

2.1. Derivation of EMUS.

The objective of EMUS is to compute the average

$$\pi[g] := \int_{\Omega} g(x) \pi(dx),$$

of a function g with respect to a measure π defined on a set Ω . In EMUS, instead of sampling directly from π , we sample from biased distributions analogous to the strata in stratified survey sampling methods. We then weight the samples from the biased distributions to estimate $\pi[g]$.

We assume that the *biased distributions* take the form

$$\pi_i(dx) := \frac{\psi_i(x) \pi(dx)}{\pi[\psi_i]}$$

for some set $\{\psi_i\}_{i=1}^L$ of non-negative *bias functions* defined on Ω . We call the support of ψ_i the *i'th stratum* to make an analogy between the biased distributions of EMUS and the strata of stratified survey sampling.

The EMUS method is based on a formula expressing $\pi[g]$ as a function of expectations over the biased distributions. To derive this formula, we assume that

$$\sum_{i=1}^L \psi_i(x) > 0 \text{ for all } x \in \Omega.$$

We then define

$$g^*(x) := \frac{g(x)}{\sum_{k=1}^L \psi_k(x)}$$

for any function $g: \Omega \rightarrow \mathbb{R}$, and we observe that

$$\begin{aligned} \pi[g] &= \int_{\Omega} g(x) \pi(dx) \\ &= \left(\sum_{k=1}^L \pi[\psi_k] \right) \sum_{i=1}^L \frac{\pi[\psi_i]}{\sum_{k=1}^L \pi[\psi_k]} \int_{\Omega} \frac{g(x)}{\sum_{k=1}^L \psi_k(x)} \frac{\psi_i(x) \pi(dx)}{\pi[\psi_i]} \\ &= \Psi \sum_{i=1}^L z_i \pi_i[g^*], \end{aligned} \quad (2.1)$$

where we set

$$\Psi := \sum_{k=1}^L \pi[\psi_k] \text{ and } z_i := \frac{\pi[\psi_i]}{\Psi}. \quad (2.2)$$

We call the vector $z \in \mathbb{R}^L$ with entries z_i the *weight vector*. Now let $\mathbf{1}$ be the constant function with $\mathbf{1}(x) = 1$ for all $x \in \Omega$. Taking $g = \mathbf{1}$ in equation (2.1) yields

$$\Psi = \frac{1}{\sum_{i=1}^L z_i \pi_i[\mathbf{1}^*]},$$

and so

$$\Psi[g] = \frac{\sum_{i=1}^L z_i \pi_i[g^*]}{\sum_{i=1}^L z_i \pi_i[\mathbf{1}^*]}. \quad (2.3)$$

Therefore, to express $\pi[g]$ as a function of averages over the biased distributions, it will suffice to express the weight vector z as a function of averages over the biased distributions. Taking g equal to ψ_j in (2.1) yields

$$\Psi z_j = \pi[\psi_j] = \Psi \sum_{i=1}^L z_i \pi_i[\psi_j^*],$$

which is equivalent to the eigenvector equation

$$z^t F = z^t, \text{ where } F_{ij} = \pi_i[\psi_j^*]. \quad (2.4)$$

We call F the *overlap matrix*. We observe that F is stochastic (its entries are nonnegative and its rows each sum to one) and that z is a probability vector (its entries sum to one).

Therefore, by the Perron-Frobenius theorem, as long as F is irreducible, the eigenvector problem (2.4) determines z as a function of F , a matrix whose entries are averages over the biased distributions. We assume throughout the remainder of this work that F is irreducible. We give a simple condition on the bias functions which guarantees irreducibility of F in Lemma 2.1 below. In general, for any irreducible, stochastic matrix $G \in \mathbb{R}^{L \times L}$, we will let $w(G) \in \mathbb{R}^L$ denote the unique solution of

$$w(G)^t G = w(G)^t \text{ with } \sum_{i=1}^L w_i(G) = 1.$$

With this notation, $z = w(F)$, and by (2.3) and (2.4) we have

$$\pi[g] = \frac{\sum_{i=1}^L w_i(F) \pi_i[g^*]}{\sum_{i=1}^L w_i(F) \pi_i[\mathbf{1}^*]}, \quad (2.5)$$

expressing $\pi[g]$ as a function of averages over the biased distributions, as desired.

In EMUS, we substitute MCMC estimates for the averages over biased distributions on the right hand side of (2.5) to estimate $\pi[g]$. To be precise, let X_t^i be a Markov process ergodic for π_i . We call X_t^i the *biased process* sampling the biased distribution π_i . The EMUS algorithm proceeds as follows:

1. For each $i = 1, \dots, L$, compute N_i steps of the process X_t^i .
2. Compute the averages

$$\bar{g}_i^* = \frac{1}{N_i} \sum_{t=1}^{N_i} g^*(X_t^i), \bar{\mathbf{1}}_i^* = \frac{1}{N_i} \sum_{t=1}^{N_i} \mathbf{1}^*(X_t^i), \text{ and } \bar{F}_{ij} = \frac{1}{N_i} \sum_{t=1}^{N_i} \psi_j^*(X_t^i).$$

3. Compute $w(\bar{F})$ numerically, for example from the QR factorization of $I - \bar{F}$ [19].
4. Compute the estimate

$$\pi_{\text{US}}[g] = \frac{\sum_{i=1}^L w_i(\bar{F}) \bar{g}_i^*}{\sum_{i=1}^L w_i(\bar{F}) \bar{\mathbf{1}}_i^*}$$

of $\pi[g]$.

Note that $w(\bar{F})$ is defined only if \bar{F} is irreducible. In the following lemma, we state simple criteria for the irreducibility of F and \bar{F} :

Lemma 2.1. *The overlap matrix F is irreducible if and only if for every $A \subset \{1, 2, \dots, L\}$, we have*

$$\pi \left[\left(\sum_{i \in A} \psi_i \right) \left(\sum_{j \notin A} \psi_j \right) \right] > 0. \quad (2.6)$$

The approximate overlap matrix \bar{F} is irreducible if and only if for every $A \subset \{1, 2, \dots, L\}$, the set $\cup_{i \in A} \{x : \psi_i(x) > 0\}$ contains at least one sample point generated from one of the biased processes X_t^j with $j \notin A$.

Proof. We prove only the second statement; proof of the first is similar. By definition, a non-negative matrix $M \in \mathbb{R}^{L \times L}$ is irreducible if and only if for every subset $A \subset \{1, 2, \dots, L\}$ of the indices, there exist indices $i \in A$ and $j \notin A$ so that $M_{ji} > 0$. Now assume that for every $A \subset \{1, 2, \dots, L\}$, there exist $j \notin A$ and $t \geq 0$ so that

$$X_t^j \in \cup_{k \in A} \{x : \psi_k(x) > 0\}.$$

Then for some $i \in A$, $\psi_i(X_t^j) > 0$, so $\bar{F}_{ji} > 0$, hence \bar{F} is irreducible. ■

We claim that the EMUS estimator is consistent; that is, $\pi_{US}[g]$ converges almost surely to $\pi[g]$ as the total number of samples tends to infinity. To make this precise, we require the following assumption on the growth of N_i with the total number of samples:

Assumption 2.2. Let

$$N = \sum_{i=1}^L N_i$$

be the total number of samples from all biased distributions. Assume that for each i ,

$$\lim_{N \rightarrow \infty} N_i/N = \kappa_i > 0.$$

That is, assume that when N is large, the proportion of samples drawn from the i 'th biased distribution is fixed and greater than zero.

We now prove that EMUS is consistent:

Lemma 2.3. *Under Assumption 2.2 and the irreducibility condition (2.6), $\pi_{US}[g]$ converges almost surely to $\pi[g]$ as the total number of samples N tends to infinity.*

Proof. Since the processes X_t^i are ergodic,

$$\bar{F} \xrightarrow{\text{as}} F, \bar{g}_i^* \xrightarrow{\text{as}} g_i^*, \text{ and } \bar{\mathbf{1}}_i^* \xrightarrow{\text{as}} \mathbf{1}_i^* \text{ as } N \rightarrow \infty. \quad (2.7)$$

Moreover, by Lemma A.1 in Appendix A, $w(G)$ is continuous at F . (Technically, $w(G)$ admits an extension to the set of all $L \times L$ matrices, which is continuous at F .) Therefore, as a function of \bar{F} , \bar{g}_i^* , and $\bar{\mathbf{1}}_i^*$, $\pi_{\text{US}}[g]$ is continuous at F , $\pi_L[g^*]$, and $\pi_L[\mathbf{1}^*]$. It follows by the continuous mapping theorem and equation (2.5) that $\pi_{\text{US}}[g] \xrightarrow{\text{as}} \pi[g]$. ■

2.2. Iterative EMUS and comparison with Vardi's Estimator.

In this section, we explain how EMUS relates to Vardi's estimator for selection bias models [49] and its descendants such as the popular Multistate Bennett Acceptance Ratio (MBAR) method [42]. In addition to comparing EMUS with these methods, we review a method, iterative EMUS, for solving the nonlinear system of equations defining Vardi's estimator [46]. The first iterate of this method is exactly the EMUS estimator.

Vardi's estimator is similar to EMUS, except that it uses the identity

$$z_j = \sum_{i=1}^L z_i \pi_i \left[\frac{\psi_j N_i / z_i}{\sum_{k=1}^L \psi_k N_k / z_k} \right] \quad (2.8)$$

instead of our eigenvector problem (2.4). (This identity appears as equation (1.12) in [18].) That is, Vardi's estimate z^V of the weight vector is the solution of (2.8), but with trajectory averages replacing averages over the biased distributions:

$$z_j^V = \sum_{i=1}^L z_i^V \bar{G}_{ij}(z^V), \quad (2.9)$$

where for any positive $u \in \mathbb{R}^L$ we define

$$\bar{G}_{ij}(u) = \frac{1}{N_i} \sum_{n=1}^{N_i} \frac{\psi_j(X_n^i) N_i / u_i}{\sum_{k=1}^L \psi_k(X_n^i) N_k / u_k}. \quad (2.10)$$

By [49, Theorem 1], this nonlinear equation determines z^V uniquely up to a constant multiple whenever the irreducibility criterion of Lemma 2.1 holds.

Vardi's estimator was originally derived assuming that the samples X_i^i from the biased distributions were i.i.d. In that case, it is the nonparametric maximum likelihood estimator of the target distribution π given samples from the biased distributions π_i [49], and it has certain optimality properties [18]. Several adjustments to the estimator have been proposed for the case of samples from Markov processes. In the Multistate Bennett Acceptance Ratio (MBAR) method, one replaces the factors N_i appearing in the summand in (2.10) with *effective sample sizes* n_i , which are computed from estimates of the integrated autocovariance of a family of functions [42]. (In addition, in some versions of MBAR, the sample average over all N_i points on the right hand side of (2.10) is replaced with a sample

average over the n_i points obtained by including only every N_i/n_i 'th point along the trajectory X_t^i .) Another recent work proposes different effective sample sizes computed by minimizing an estimate of the asymptotic variance of the estimator [12]. In fact, the estimator is consistent with N_i replaced by any fixed positive number [12]. We have found that our numerical results do not depend sensitively on the choice of effective sample size, so we use N_i for simplicity.

We now review iterative EMUS, which we introduced in [46]. Iterative EMUS may be understood as a fixed point iteration for solving equation (2.9). The iteration proceeds as follows:

1. As an initial guess for z^V , choose a positive vector $z^0 \in \mathbb{R}^L$. Set $m = 0$. Choose a tolerance $\tau > 0$.
2. Compute $\bar{G}_{ij}(z^m)$ by (2.10). Solve the eigenvector equation

$$z_j^{m+1} = \sum_{i=1}^L z_i^{m+1} \bar{G}_{ij}(z^m) \quad (2.11)$$

for an updated estimate z^{m+1} of z^V .

3. If $\max_i |z_i^{m+1} - z_i^m| / z_i^m > \tau$, then increment m and repeat step 2.

Remark 2.4. In a related work, we show that the eigenvector equation (2.11) has a unique solution for every m , and we suggest a numerical method for finding the solution [46]. We also discuss the convergence of iterative EMUS, and we show that for every fixed m , z^m is a consistent estimator of the weight vector z .

Remark 2.5. If one chooses $z_i^0 = N_i/N$, then z^1 is the EMUS estimate of z , $w(\bar{F})$.

3. Error Analysis of EMUS.

Here, we develop tools for analyzing the error of EMUS. First, in Section 3.1, we prove a CLT for EMUS, and we derive a convenient upper bound on the asymptotic variance. Then, in Section 3.2, we analyze the dependence of the asymptotic variance of EMUS on the choice of biased distributions. We use these tools in Section 4 to prove limiting results demonstrating the advantages of EMUS for multimodal distributions and tail probabilities.

3.1. A CLT for EMUS and an Estimate of the Asymptotic Variance.

In this section, we prove a Central Limit Theorem (CLT) for EMUS, and we derive an upper bound on the asymptotic variance $\sigma_{\text{US}}^2(g)$ of $\pi_{\text{US}}[g]$. To prove the CLT for EMUS, we must assume that a CLT holds for trajectory averages over the biased processes:

Assumption 3.1. For any matrix H , let H_i denote the i 'th row of H . Define $\bar{G} \in \mathbb{R}^{L \times 2}$ by

$$\bar{G}_{i:} = (\bar{g}_{i:}^*, \bar{\mathbf{I}}_{i:}^*).$$

Assume that

$$\sqrt{N_i}((\bar{F}_{i:}, \bar{G}_{i:}) - (F_{i:}, \pi_i[g^*], \pi_i[\mathbf{1}^*])) \xrightarrow{d} N(0, \Sigma_i) \quad (3.1)$$

for some asymptotic covariance matrix $\Sigma_i \in \mathbb{R}^{(L+2) \times (L+2)}$ of the form

$$\Sigma_i = \begin{pmatrix} \sigma_i^i & \rho_i \\ \rho_i^t & \tau_i \end{pmatrix}, \quad (3.2)$$

where $\sigma_i^i \in \mathbb{R}^{L \times L}$ denotes the asymptotic covariance of $\bar{F}_{i:}$ with itself, $\rho_i \in \mathbb{R}^{L \times 2}$ denotes the asymptotic covariance of $\bar{F}_{i:}$ with $\bar{G}_{i:}$, and $\tau_i \in \mathbb{R}^{2 \times 2}$ denotes the asymptotic covariance of $\bar{G}_{i:}$ with itself.

We expect a CLT to hold for most MCMC methods, target distributions, and target functions of interest in statistics and statistical mechanics. We refer to [40] for a comprehensive review of conditions guaranteeing a CLT. In Theorem 3.7 of Section 3.2, we prove a CLT and an estimate of the asymptotic variance for a simple family of processes which one might use to sample the biased distributions in an application of EMUS.

We now prove a CLT for $\pi_{\text{US}}[g]$, and we give a formula expressing the asymptotic variance $\sigma_{\text{US}}^2(g)$ of $\pi_{\text{US}}[g]$ in terms of the asymptotic variances Σ_i of the trajectory averages. In this formula, $(I - F)^\#$ denotes the group generalized inverse of $I - F$; the group inverse $A^\#$ of a matrix A is characterized by the properties

$$AA^\#A = A, A^\#AA^\# = A^\#, \text{ and } AA^\# = A^\#A.$$

We refer to [19] for a detailed explanation of the properties of the group inverse, a proof that $(I - F)^\#$ exists whenever F is stochastic and irreducible, and an algorithm for computing $(I - F)^\#$.

In Theorem 3.3 and below we impose the following assumption:

Assumption 3.2. The processes X_t^i sampling the biased distributions are independent.

This assumption does not hold for all stratified MCMC methods. For example, in replica exchange umbrella sampling one periodically allows configuration exchanges between neighboring processes; see [32] for a general discussion of replica exchange strategies and [43] for an application of replica exchange in a method similar to EMUS. The result is a single process taking values in $\mathbb{R}^{L \times d}$ and sampling the product distribution $\Pi(x_1, x_2, \dots, x_L)$

$= \pi_1(x_1)\pi_2(x_2) \dots \pi_L(x_L)$. In this case, a CLT would still hold for EMUS, but the asymptotic variance would take a different form.

Theorem 3.3. *Let Assumptions 2.2, 3.1, and 3.2 and the irreducibility condition (2.6) hold. Let g be square integrable over π , so $\pi[g^2] < \infty$. Recall the definition of Ψ from (2.2), and define*

$$\ell := \Psi(1, -\pi[g])^t \in \mathbb{R}^2.$$

Let $\mathbf{g} \in \mathbb{R}^L$ be the vector with $\mathbf{g}_i := \ell \cdot (\pi_i[g^*], \pi_i[\mathbf{1}^*])$. We have

$$\sqrt{N}(\pi_{US}[g] - \pi[g]) \xrightarrow{d} N(0, \sigma_{US}^2(g)), \quad (3.3)$$

where

$$\sigma_{US}^2(g) = \sum_{i=1}^L \frac{z_i^2}{\kappa_i} \left\{ (I - F)^\# \mathbf{g} \cdot \sigma^i (I - F)^\# \mathbf{g} + 2(I - F)^\# \mathbf{g} \cdot \rho_i \ell + \ell^t \tau_i \ell \right\}. \quad (3.4)$$

Proof. The result follows using the delta method and a formula expressing $w'(F)$ in terms of $(I - F)^\#$; we give the details in Appendix A. ■

In Appendix F, we explain how to compute an estimate of $\sigma_{US}^2(g)$ given trajectories of the biased processes. We use this estimates in Section 5 when analyzing our computational experiments. We note that the numerical methods presented in Appendix F for estimating $\sigma^2(g)$ significantly improve upon our original proposal in [46]; see Figure 11.

We now derive a convenient upper bound on the asymptotic variance $\sigma_{US}^2(g)$. In Section 4, we use this bound to analyze the efficiency of EMUS in the low-temperature limit and in the limit of small tail probabilities. Our bound is based on the probability $\mathbf{P}_i[t_j < t_i]$ defined below:

Definition 3.4. *Let Y_n be the Markov chain with state space $\{1, 2, \dots, L\}$ and transition matrix F . Let $\mathbf{P}_i[t_j < t_i]$ denote the probability that Y_n hits j before returning to i , conditioned on $Y_0 = i$.*

Theorem 3.5. *Let Assumptions 2.2, 3.1, and 3.2 and the irreducibility condition (2.6) hold. Let g be square integrable over π , so $\pi[g^2] < \infty$. Let $\sigma^2(g)$ be the asymptotic variance of $\pi_{US}[g]$, and for any measure ν and function f let $\text{var}_\nu(f)$ be the variance of f over ν . Define the function*

$$h = g^* - \pi[g]\mathbf{1}^*,$$

and let $C(\bar{h}_i)$ be the asymptotic variance of the trajectory average of h over the biased process X_t^i . We have

$$\sigma_{US}^2(g) \leq 2 \sum_{i=1}^L \frac{1}{\kappa_i} \left[z_i^2 \Psi^2 C(\bar{h}_i) + \text{tr}(R^i) \pi[|h|]^2 \sum_{\substack{j \neq i \\ F_{ij} > 0}} \frac{\text{var}_{\pi_i}(\psi_j^*)}{\mathbf{P}_i[t_j < t_i]^2} \right], \quad (3.5)$$

where $R^i \in \mathbb{R}^{L \times L}$ with

$$R_{jk}^i = \frac{\sigma_{jk}^i}{\sqrt{\text{var}_{\pi_i}(\psi_j^*)} \sqrt{\text{var}_{\pi_i}(\psi_k^*)}}.$$

Proof. The result follows from Theorem 3.3, using the perturbation bounds which we derived in [45]. Details appear in Appendix A. ■

When the bias functions are a partition of unity, both the EMUS method and the statements of Theorems 3.3 and 3.5 simplify considerably. (The bias functions are a partition of unity if and only if $\sum_{i=1}^L \psi_i(x) = 1$ for all x .) In this case, $f^* = f$ for all functions f , and the EMUS method reduces to

$$\pi_{US}[g] = \sum_{i=1}^L w_i(\bar{F}) \bar{g}_i,$$

where

$$\bar{F}_{ij} = N_i^{-1} \sum_{t=1}^{N_i} \psi_j(X_t^i) \text{ and } \bar{g}_i = N_i^{-1} \sum_{t=1}^{N_i} g(X_t^i).$$

In the statement of Theorem 3.5, one can replace $\pi[|h|]^2$ with $\text{var}_{\pi}(g)$ or $\pi[g]^2$. We also have $\Psi = 1$, and one can replace $C(\bar{h}_i)$ with the asymptotic variance $C(\bar{g}_i)$ of \bar{g}_i . (We verify these claims in Appendix A as part of the proof of Theorem 3.5.) In our limiting results (Section 4) and computational experiments (Section 5), we choose the bias functions to be a partition of unity.

3.2. Dependence of the Asymptotic Variance on the Choice of Strata.

In this section, we consider how the choice of strata influences the factors in the upper bound (3.5) on $\sigma_{US}^2(g)$. Roughly, the asymptotic variances $C(\bar{h}_i)$ and $\text{tr}(R^i)$ characterize the sampling error, and for each i the factor

$$\sum_{\substack{j \neq i \\ F_{ij} > 0}} \frac{\text{var}(\pi_i(\psi_j^*))}{\mathbf{P}_i[\tau_j < \tau_i]^2},$$

measures the sensitivity of the EMUS estimator to sampling errors associated with π_i . We show in Section 3.2.1 that $C(\bar{h}_i)$ and $\text{tr}(R^i)$ may be controlled by decreasing the diameters of the strata, under some conditions. We show in Section 3.2.2 that that $\mathbf{P}_i[\tau_j < \tau_i]$ may be controlled by ensuring sufficient overlap between neighboring strata. This last observation leads to a practical condition guiding the choice of strata; see Remark 3.12 and (5.8).

To streamline our illustration of the benefits of stratification, we impose additional assumptions in this section. These assumptions are much stronger than those made above. We discuss how they relate to practical implementations of stratified MCMC in Remarks 3.9 and 3.12.

3.2.1. Asymptotic Variances of MCMC Averages.—Here, we consider the effect of the choice of strata on the asymptotic variances $C(\bar{h}_i)$ and $\text{tr}(R^i)$. Because such a diverse variety of biased processes and distributions could in principle be used, it is futile in our opinion to try for a completely general result. Instead, motivated by the efficiency analysis undertaken in Section 4, we introduce a simple parametric family of bias functions, and for this family we state Assumption 3.6 relating the diameters of the strata with the asymptotic variances. In Theorem 3.7, we verify Assumption 3.6 for one representative class of biased processes. Finally, at the end of this section, we explain why we expect the assumption to hold for other choices of biased processes and distributions; cf. Remark 3.9.

Consider the following representative class of bias functions: Given a family of sets $\{U_i: i = 1, \dots, L\}$ with $\cup_{i=1}^L U_i = \Omega$, define

$$\psi_i := \mathbf{1}_{U_i} \text{ and } \pi_i(dx) := \frac{\mathbf{1}_{U_i}(x)\pi(dx)}{\pi[\mathbf{1}_{U_i}]} \text{ for } i = 1, \dots, L, \quad (3.6)$$

where $\mathbf{1}_{U_i}$ denotes the characteristic function of U_i . Assume that the sets U_i are chosen so that the irreducibility criterion of Lemma 2.1 holds. For example, suppose that $\Omega = [0, 1]^d$ is the d -dimensional unit cube. One might choose $K \in \mathbb{N}$, set $h := 1/K$, and define

$$U_{\mathbf{i}} = (h[\mathbf{i} - 1, \mathbf{i}]^d + h\mathbf{i}) \cap \Omega \text{ for } \mathbf{i} \in \{0, 1, \dots, K\}^d, \quad (3.7)$$

covering Ω uniformly by a grid of strata having diameters proportional to h . We use this uniform grid as a device when analyzing the effect of the stratum size on the efficiency of EMUS in Section 4. However, while such a naïve choice may suffice for small d , it is not

practical for large d . We discuss appropriate bias functions for high-dimensional problems later in this section and again in Section 5.1.

Since we wish to study grids like (3.7) as $h = 1/K$ varies, we state our assumption on asymptotic variances in terms of the following parametric family of strata: Let $x_0 \in \Omega$, and let $Z \subset \mathbb{R}^d$ be a bounded set containing 0. For each $h > 0$, define a stratum and a biased distribution by

$$Z_h = x_0 + hZ \text{ and } \pi_h(dx) = \frac{\mathbf{1}_{Z_h}(x)\pi(dx)}{\pi[\mathbf{1}_{Z_h}]} . \quad (3.8)$$

Assumption 3.6 characterizes the dependence of the asymptotic variance of MCMC averages over π_h on the parameter h :

Assumption 3.6. Assume that $f: \Omega \rightarrow \mathbb{R}$ has finite variance $\text{var}_h(f)$ over π_h , and define $\sigma_h^2(f)$ to be the asymptotic variance of an MCMC trajectory average approximating $\pi_h[f]$. Write

$$\pi(x) = \frac{\exp(-\beta V(x))}{\int \exp(-\beta V(y))dy},$$

for some *potential* $V: \Omega \rightarrow \mathbb{R}$ and *inverse temperature* $\beta > 0$. We assume

$$\frac{\sigma_h^2(f)}{\text{var}_{\pi_h}(f)} \leq Ch^a \beta^b \exp\left(\beta \left(\max_{Z_h} V - \min_{Z_h} V\right)\right) \leq Ch^a \beta^b \exp(\beta h \text{diam}(Z) \|\nabla V\|_\infty)$$

for some $C, a, b \geq 0$ independent of h, Z , and f .

To motivate Assumption 3.6, we prove that a special case holds for a representative class of processes sampling the biased distributions, cf. Theorem 3.7. Assume that the potential V appearing in the assumption is continuously differentiable. Let $Z \subset \mathbb{R}^d$ be either a convex polyhedron or a set with C^3 boundary.² Now let X_t^h be the overdamped Langevin process with reflecting boundary conditions on Z_h . This process is defined by the Fokker–Planck equation

$$\begin{aligned} \frac{\partial u}{\partial t}(x, t) &= \text{div}(\beta^{-1} \nabla u(x, t) + u(x, t) \nabla V(x)) \quad \text{for } x \in U, t > 0, \\ (\beta^{-1} \nabla u(x, t) + u(x, t) \nabla V(x)) \cdot \mathbf{n}(x) &= 0 \quad \text{for } x \in \partial Z_h, t \geq 0, \text{ and} \\ u(x, 0) &= p(x) \quad \text{for } x \in Z_h, \end{aligned} \quad (3.9)$$

²The boundary of a set is C^3 if in a neighborhood of each point on the boundary, the boundary is the graph of a three times continuously differentiable function.

where β and V are the inverse temperature and potential defined in Assumption 3.6 and $\mathbf{n}(x)$ denotes the inward unit normal to Z_h at x . That is, X_t^h is the unique Markov process so that if X_0^h has density $p(x)$, then X_t^h has density $u(x, t)$. The existence of the reflected process is established in [51, 2] when Z is a convex polyhedron and in [13, Chapter 8] when Z has C^3 boundary. A simple introduction to the reflected process and its properties appears in [36, Chapter 4]. We show in Theorem 3.7 that X_t^h is ergodic for π_h , at least when Z is bounded.

The reflected process X_t^h shares many features with the processes used in practical stratified MCMC methods. In particular, it is closely related to the (unreflected) overdamped Langevin process Y_t which solves

$$dY_t = -\nabla V(Y_t)dt + \sqrt{2\beta^{-1}}dB_t. \quad (3.10)$$

In fact, the Fokker–Planck equation of the unreflected process is the same as the reflected process, except with no boundary condition and with \mathbb{R}^d in place of Z_h .

We now verify Assumption 3.6 for the reflected process:

Theorem 3.7. *Assume that $f: \Omega \rightarrow \mathbb{R}$ has finite variance $\text{var}_h(f)$ over π_h . Let Z either have C^3 boundary or be convex. Assume that V is continuously differentiable. Suppose that X_t^h is stationary; that is, X_0^h has distribution π_h . Let $\bar{f}^h := \frac{1}{T} \int_{t=0}^T f(X_t^h) dt$ be the continuous time trajectory average of f . We have*

$$\sqrt{T}(\bar{f}^h - \pi_h[f]) \xrightarrow{d} N(0, \sigma_h^2(f)),$$

where

$$\sigma_h^2(f) \leq \Lambda h^2 \beta \exp\left(\beta \left(\max_{Z_h} V - \min_{Z_h} V\right)\right) \text{var}_h(f). \quad (3.11)$$

The constant Λ depends only on Z , not on h , β , V , or f .

Sketch of proof. We give a detailed proof of Theorem 3.7 in Appendix C; here we present only an outline. Our proof is based on a formula expressing the asymptotic variance of trajectory averages in terms of the generator of X_t^h : Under certain conditions,

$$\sigma_h^2(g) = -2\pi_h[(g - \pi_h[g])L^{-1}(g - \pi_h[g])], \quad (3.12)$$

where L is the generator of X_t^h and $L^{-1}(g - \pi_h[g])$ is a function so that

$$L(L^{-1}(g - \pi_h[g])) = g - \pi_h[g],$$

and $\nabla L^{-1}(g - \pi_h[g]) \cdot \mathbf{n} = 0$ on Z_h . To estimate $\sigma_h^2(g)$, we prove a Poincaré inequality for π_h , which implies an upper bound on $L^{-1}(g - \pi_h[g])_{L^2(\pi_h)}$. This general approach is outlined in [31, Section 3], which treats the case of the overdamped Langevin dynamics on an unbounded domain. ■

Remark 3.8. In a related work, we summarize and refute a widely accepted argument in favor of umbrella sampling from the chemistry literature [9, Chapter 8]; see [46, Section VI.A]. Roughly speaking, the argument in the chemistry literature treated the dependence of the sampling error on the choice of strata correctly, but the sensitivity of the algorithm to those errors incorrectly. Our Theorem 3.7 verifies and extends the correct part of this argument.

Remark 3.9 (Practical Biased Processes and Assumption 3.6). As explained above, the reflected process X_t^h is quite similar to the processes used in practice. We now discuss practical methods, and we explain when we expect Assumption 3.6 to hold. We identify three major differences between X_t^h and practical methods: First, in molecular simulations, one typically chooses Gaussian bias functions instead of piecewise constant. Second, practical methods must be discrete in time, e.g. one might use a discretization of the continuous time process X_t^h . Third, for high-dimensional problems, one typically stratifies only a certain low-dimensional *reaction coordinate* or *collective variable*.

In the first case, for Gaussian bias functions, a version of Theorem 3.7 holds with minor adjustments; we omit the exact statement and proof for simplicity. In the second case, for discretizations of Langevin dynamics, the asymptotic variances of trajectory averages are closely related to the corresponding averages for the continuous time dynamics: In fact, under some conditions on the potential V ,

$$\lim_{\Delta t \rightarrow 0} \Delta t \zeta_{\Delta t}^2(f) = \zeta^2(f), \quad (3.13)$$

where $\zeta_{\Delta t}^2(f)$ is the asymptotic variance of the trajectory average of f for the discretization with time step Δt and $\zeta^2(f)$ is the asymptotic variance for the continuous time process [31, Section 3.2]. For other discrete time processes, we expect Assumption 3.6 to hold with different exponents a and b . For example, the affine invariance property of the affine invariant ensemble sampler [20] suggests $a = 0$.

The third case is subtle. When d is large, one typically stratifies only in a function $\theta: \Omega \subset \mathbb{R}^d \rightarrow \mathbb{R}^\ell$ with ℓ much smaller than d . To be precise, one might choose a uniform grid of nonnegative functions $\eta_i: \mathbb{R}^\ell \rightarrow \mathbb{R}$ defined as in (3.7), but with supports covering $\theta(\Omega) \subset \mathbb{R}^\ell$ instead of $\Omega \subset \mathbb{R}^d$. One would then define the bias functions

$$\psi_i(x) = \eta_i(\theta(x)). \quad (3.14)$$

(We make a similar choice in our calculations in Section 5, cf. the natural stratification (5.1).) For a clever choice of θ , these biased distribution may be much easier to sample than the target distribution. For example, suppose that the marginal π_θ of π in θ were multimodal, but that the conditional distributions $\pi(\cdot \mid \theta = \theta_0)$ were unimodal or otherwise easy to sample for each fixed θ_0 . In that case, for h sufficiently small, each biased distribution would be unimodal, hence easy to sample. (Recall that h sets the diameters of the strata for the grid of bias functions defined in (3.7), so h small means that the diameter of the support of η_i is small.) In free energy calculations, it is often possible to choose such a θ based on intuition or scientific principles; see [46, 30] for discussion. Also, when computing tails or marginals, the problem itself typically suggests a particular θ , cf. the *natural stratification* in Section 5.1.

The reader will notice that bias functions of the form (3.14) will typically have infinite support, rendering the bound in Assumption 3.6 useless. In this case, one might hope for a similar bound with the potential function V replaced by the free energy

$$F(\theta) := -\beta^{-1} \log(\pi_\theta(\theta)),$$

where π_θ is the marginal density of π in θ . Roughly, this replacement will be valid when, for MCMC processes sampling π , the θ variables equilibrate very slowly compared to other variables. This will occur, for example, when the marginal in θ is multimodal or otherwise difficult to sample, but the conditional distributions are easy to sample. More on the effective dynamics of low-dimensional variables can be found in [37] or [29].

3.2.2. Controlling the Probabilities $\mathbf{P}_i[\tau_j < \tau_i]$.—Here, we examine the effect of the choice of strata on the factor

$$\sum_{\substack{j \neq i \\ F_{ij} > 0}} \frac{\text{var}_{\pi_i}(\psi_j^*)}{\mathbf{P}_i[\tau_j < \tau_i]^2} \quad (3.15)$$

appearing in our upper bound (3.5) on $\sigma_{\text{US}}^2(g)$.

We begin with a lemma estimating $\text{var}_{\pi_i}(\psi_j^*)/\mathbf{P}_i[\tau_j < \tau_i]^2$ in terms of F_{ij} .

Lemma 3.10. *We have*

$$\frac{\text{var}_{\pi_i}(\psi_j^*)}{\mathbf{P}_i[\tau_j < \tau_i]^2} \leq \frac{1}{F_{ij}}.$$

Proof. We have

$$\mathbf{P}_i[t_k < t_i] \geq \mathbf{P}[X_1 = k \mid X_0 = i] = F_{ik},$$

where X_t denotes the Markov chain with transition matrix F . Therefore, since $\psi_j^*(x) \in [0, 1]$,

$$\frac{\text{var}_{\pi_i}(\psi_j^*)}{\mathbf{P}_i[\tau_j < \tau_i]^2} \leq \frac{\text{var}_{\pi_i}(\psi_j^*)}{F_{ij}^2} \leq \frac{\pi_i[(\psi_j^*)^2]}{F_{ij}^2} \leq \frac{\pi_i[\psi_j^*]}{F_{ij}^2} = \frac{1}{F_{ij}}.$$

■

We now estimate the size of F_{ij} for piecewise constant bias functions such as the uniform grid (3.7):

Lemma 3.11. *Assume as in (3.6) that the bias functions are piecewise constant, and write $\pi(x) \propto \exp(-\beta V(x))$. We have*

$$F_{ij} \geq \frac{|U_i \cap U_j|}{|U_i| \|\sum_{k=1}^L \mathbf{1}_{U_k}\|_\infty} \exp\left(\beta \left(\min_{U_i} V - \max_{U_i} V\right)\right)$$

In particular, for the uniform grid of strata (3.7), we have

$$F_{ij} \geq \frac{1}{4^d} \exp\left(\beta \left(\min_{U_i} V - \max_{U_i} V\right)\right) \geq \frac{1}{4^d} \exp(-2\beta h \sqrt{d} \|\nabla V\|_\infty) \quad (3.16)$$

for any $\mathbf{i}, \mathbf{j} \in \mathbb{Z}^d$ so that $F_{ij} > 0$.

Proof. We have

$$\begin{aligned} F_{ij} &= \pi_i[\psi_j^*] = \pi_i \left[\frac{\mathbf{1}_{U_j}}{\sum_{k=1}^L \mathbf{1}_{U_k}} \right] \geq \frac{\pi[U_i \cap U_j]}{\pi[U_i]} \frac{1}{\|\sum_{k=1}^L \mathbf{1}_{U_k}\|_\infty} \\ &\geq \frac{|U_i \cap U_j|}{|U_i| \|\sum_{k=1}^L \mathbf{1}_{U_k}\|_\infty} \exp\left(\beta \left(\min_{U_i} V - \max_{U_i} V\right)\right), \end{aligned}$$

which proves the first claim made in the statement of the lemma.

Now, for the uniform grid of strata (3.7), the minimum nonzero value of $|U_i \cap U_j|/|U_i|$ is $1/2^d$, attained when $\mathbf{j} = (1, 1, \dots, 1) + \mathbf{i}$. Moreover, except for a set of measure zero, each $x \in \mathbb{R}^d$ lies within 2^d strata, so $\|\sum_{k=1}^L \mathbf{1}_{U_k}\|_\infty = 2^d$. Finally, we have

$$\max_{U_i} V(x) - \min_{U_i} V(x) \leq \text{diam}(U_i) \|\nabla V\|_\infty = 2\sqrt{d}h \|\nabla V\|_\infty,$$

and the result follows. ■

Remark 3.12 (A Condition to Guide the Choice of Strata). Lemmas 3.10 and 3.11 suggest a practical constraint on the choice of strata: To ensure that the calculation of the weights is not too sensitive to sampling errors, it will suffice to choose strata so that nonzero entries of F are not too small. We let this condition guide the choice of strata in Section 5, cf. (5.8). However, the condition is only sufficient, not necessary. For example, consider a uniform grid of Gaussian bias functions similar to (3.7), but with Gaussian densities having mean $\mu = h[-1, 1]^d + h\mathbf{i}$ and variance $\sigma^2 = h^2$ replacing the characteristic functions $\mathbf{1}_{U_i}$. In that case, even though F will be dense and may have some extremely small nonzero entries, one can still control (3.15) by decreasing h , under some conditions on π . We omit the exact statement and proof for simplicity.

Despite the exponential dependence on d in (3.16), EMUS and other stratified MCMC methods are advantageous for high-dimensional problems because it often suffices to stratify only a low-dimensional collective variable. In such cases, the dimension of the grid of strata is much smaller than dimension of the state space Ω ; see our discussion of collective variables in Section 3.2.1 and our computations in Section 5. It is important to keep this in mind when reading our results below.

Remark 3.13. One may define a uniform grid of strata so that (3.15) increases only as d^2 with dimension, not exponentially: For any $\mathbf{i} \in \mathbb{Z}^d$, let $V'_i = h\mathbf{i} + h[-\frac{1}{2}, \frac{1}{2}]^d$. For $\mathbf{i} \neq \mathbf{j}$, define

$$W_{\mathbf{ij}} := \left\{ x \in V'_j : \min_{y \in V'_i} \|x - y\| \leq \min_{y \in V'_k} \|x - y\| \text{ for any } \mathbf{k} \in \mathbb{Z}^d \setminus \{\mathbf{j}\} \right\}$$

to be the d -dimensional pyramid consisting of all points in V'_j closer to V'_i than to any other cube V'_k . Now let e_n denote the n 'th standard basis vector in \mathbb{R}^d , and define

$$V_i := \bigcup_{n=1}^d \left(W_{\mathbf{i}, \mathbf{i} + e_n} \cup W_{\mathbf{i}, \mathbf{i} - e_n} \right) \cup V'_i$$

to be the cube V'_i enlarged by all the neighboring pyramids $W_{\mathbf{ij}}$. The strata V_i are convex, and the corresponding bias functions $\psi_i = \frac{1}{2}\mathbf{1}_{V_i}$ are a partition of unity. Each stratum V_i intersects only the $2d$ neighboring strata $V_{i \pm e_n}$ for $n = 1, \dots, d$. Moreover, each intersection between neighboring strata V_i and V_j consists of the pair of pyramids $W_{\mathbf{ij}}$ and $W_{\mathbf{ji}}$, and it has volume $1/d$. Therefore, by Lemma 3.11, for this choice of bias functions, the nonzero entries of F decrease as $1/d$. It follows that (3.15) increases as d^2 .

4. Limiting Results as a Rationale for EMUS.

In this section, we analyze the efficiency of EMUS in two limits: First, we consider a *low temperature limit*, where we write $\pi(x) \propto \exp(-\beta V(X))$ and let the inverse temperature β increase, concentrating the target distribution at its modes and intensifying the effects of multimodality on the efficiency of MCMC sampling. Second, we consider the estimation of increasingly small tail probabilities. Our goal in each case is to elucidate the advantages and disadvantages of EMUS for a broad class of problems, providing a rationale for the use of the method. We hope that others will use the tools of Section 3 in similar fashion to develop their own novel applications of EMUS.

4.1. Limit of Low Temperature.

Let the target distribution take the form

$$\pi_{\beta}(x) = \frac{\exp(-\beta V(x))}{\int \exp(-\beta V(x)) dx}$$

for some *potential* V and *inverse temperature* $\beta > 0$, as in Section 3.2. In this section, we analyze the efficiency of EMUS in the *low temperature limit* as β tends to infinity with V fixed. We observe that π_{β} concentrates at its modes (the minima of V) in this limit. As a consequence, MCMC methods for sampling π_{β} undergo transitions between modes only rarely, which makes direct MCMC sampling increasingly inefficient. To be precise, we show that the asymptotic variance of a trajectory average of the overdamped Langevin dynamics increases exponentially with β in the worst case. On the other hand, we show that the asymptotic variance of the EMUS estimate of the same average increases only polynomially. Therefore, EMUS is dramatically more efficient than direct sampling in the low temperature limit.

We consider the low temperature limit because it provides a convenient sequence of increasingly difficult to sample multimodal distributions: By analyzing EMUS in the low temperature limit, we hope to elucidate its advantages for multimodal problems in general. We have no other interest in low temperature.

We now examine the overdamped Langevin dynamics.

$$dX_t^{\beta} = -\nabla V(X_t^{\beta})dt + \sqrt{2\beta^{-1}}dB_t \quad (4.1)$$

in the low temperature limit. (The overdamped Langevin dynamics is ergodic for π_{β} under certain conditions on V ; see [41] for example.) For typical potentials V , the generator

$$\mathcal{L} := -\beta^{-1}\Delta + \nabla V \cdot \nabla$$

of (4.1) has a spectral gap that shrinks exponentially with β ; that is, for some $c > 0$,

$$-\exp(-c\beta) \leq \lambda_1 < 0, \quad (4.2)$$

where λ_1 is the greatest nonzero eigenvalue of \mathcal{L} . We refer to [31, Section 2.5] for a review of results on the spectrum of \mathcal{L} , and we refer to [21] for precise conditions on V which guarantee (4.2). Now let v_1 be an eigenfunction corresponding to λ_1 normalized so that $\pi_h[v_1^2] = 1$. By formula (3.12), the asymptotic variance $\sigma_\beta^2(v_1)$ of the trajectory average of v_1 satisfies

$$\sigma_\beta^2(v_1) = -\pi_\beta[v_1 L^{-1} v_1] = -\lambda_1^{-1} \pi_\beta[v_1^2] = -\lambda_1^{-1} \geq \exp(c\beta),$$

indicating that the cost of estimating $\pi[v_1]$ by direct MCMC grows exponentially with β .

Having analyzed the overdamped Langevin dynamics, we now examine EMUS in the low temperature limit. For convenience, we assume that Ω is the unit cube $[0, 1]^d \subset \mathbb{R}^d$ with periodic boundary conditions; to be more precise, we let $\Omega = \mathbb{R}^d / \mathbb{Z}^d$ be the set of all points in \mathbb{R}^d with x and y identified if and only if $x - y \in \mathbb{Z}^d$. Periodic boundary conditions are typical of problems in chemistry and computational statistical mechanics. We do not see any difficulties in generalizing our results to other types of domains.

As β increases, we must make the supports of the bias functions smaller. We accomplish this by adjusting the parameter h in a uniform grid of bias functions similar to those defined in (3.7). To be precise, we fix $K \in \mathbb{N}$, set $h := 1/K$, and define

$$\psi_{\mathbf{i}}(x) := \frac{1}{2^d} \mathbf{1}_{[-1, 1]^d}(K(x - h\mathbf{i})) \text{ for } \mathbf{i} \in \{0, 1, \dots, K-1\}^d. \quad (4.3)$$

This family of K^d bias functions is a partition of unity over Ω , and the support of the \mathbf{i} 'th bias function is

$$U_{\mathbf{i}} := h[-1, 1]^d + h\mathbf{i}.$$

For convenience, we treat the index \mathbf{i} as an element of $\mathbb{Z}^d / K\mathbb{Z}^d$; that is, we let \mathbf{i} be periodic with period K in each of its components, identifying $(0, i_2, \dots, i_d)$ with (K, i_2, \dots, i_d) , for example. Figure 1 illustrates such a family of bias functions, and it demonstrates the appropriate relationship between β and h .

We now show that the asymptotic variance of EMUS increases at most polynomially with β when K is chosen appropriately. In light of the above discussion, this means that EMUS may be dramatically more efficient than direct sampling for multimodal problems. We note that despite the exponential dependence on d in (4.4) below, EMUS and other stratified MCMC methods are often advantageous for high-dimensional multimodal problems; see our discussion of low-dimensional collective variables in Section 3.2 and also our computations in Section 5.

Theorem 4.1. For any bounded continuous function g , let $\sigma_{\beta, US}^2(g)$ denote the asymptotic variance of $\pi_{\beta, US}[g]$. Let the bias functions be defined by (4.3) with K equal to the least integer greater than β ; that is,

$$K = \lceil \beta \rceil.$$

Take $\kappa_1 = 1/K^d$. Let Assumption 3.6 hold. We have

$$\frac{\sigma_{\beta, US}^2(g)}{\text{var}_{\pi_{\beta}}(g)} \leq C(1 + \beta)^{qd} \quad (4.4)$$

for constants $C, q > 0$ independent of g and β , but depending on V and the constants in Assumption 3.6.

Proof. The proof is a straightforward application of the theory developed in Section 3; we present the details in Appendix D. ■

Remark 4.2. Our proof of Theorem 4.1 relies on the perturbation bounds which we derived in [45]. These bounds allow one to estimate the sensitivity of $w(F)$ to small perturbations of F . Most perturbation bounds in the literature predict that $w(F)$ is highly sensitive when the spectral gap of F is small, but ours show that this is not always the case. (The spectral gap is $1 - |\lambda_2|$, where λ_2 is the eigenvalue of F with second largest absolute value.) In the low-temperature limit, the spectral gap of F decreases exponentially with β ; see [45] for a simple example of this phenomenon. Nonetheless, using our bounds, we show that the cost to compute averages by EMUS increases only polynomially in β .

4.2. Limit of Small Probability.

In this section, we assess the performance of EMUS for computing tail probabilities. To be precise, we let $\Omega = [0, \infty)$, and we consider estimation of probabilities of the form

$$p_M := \pi([M, \infty)).$$

We show that for a broad class of distributions π , the cost of computing p_M with relative precision by direct MCMC increases exponentially with M , whereas the cost by EMUS increases only polynomially. Thus, EMUS is dramatically more efficient than direct sampling for computing the probabilities of tail events.

In Assumption 4.3 below, we state the conditions which we will impose on π in our analysis. These conditions specify a simple class of problems for which strong conclusions may be drawn. Similar results hold more generally. For example, in Section 5, we report the results of a computational experiment demonstrating the advantages of EMUS for computing tails of a marginal density.

Assumption 4.3. Write

$$\pi(x) = \exp(-V(x))$$

for some potential function $V: [0, \infty) \rightarrow \mathbb{R}$. Assume that for some $M_0 > 0$:

1. Whenever $x \leq M_0$,

$$0 \leq V''(x) \text{ and } 0 < V'(x). \quad (4.5)$$

2. For some $\alpha \in (0, 1)$ and $c > 0$, whenever $x \leq M_0$,

$$\alpha V'(x)^2 - V''(x) \geq c > 0. \quad (4.6)$$

For example, we might have

$$\pi(x) \propto \exp(-|x|^r) \text{ for any } r \geq 1.$$

Remark 4.4. Condition (4.6) in Assumption 4.3 implies geometric ergodicity of the overdamped Langevin dynamics with potential V [40]. We rely on this fact to motivate Assumption 4.5 concerning the convergence of MCMC processes sampling biased distributions with unbounded support. Interestingly, we use the same condition to prove lower bounds on some of the entries of the overlap matrix; cf. Lemma E.1.

Condition (4.5) in Assumption 4.3 implies

$$p_M \leq D \exp(-\gamma M)$$

whenever $M \leq M_0$ for some $D, \gamma > 0$. Therefore, the relative variance ρ_M^2 of $\mathbf{1}_{[M, \infty)}$ over π satisfies

$$\rho_M^2 = \frac{p_M - p_M^2}{p_M^2} \geq D^{-1} \exp(\gamma M) - 1.$$

We conclude that estimating p_M with relative accuracy by a direct MCMC method (or even Monte Carlo with independent samples) requires a number of samples increasing exponentially with M .

By contrast, we show that for an appropriate choice of bias functions, the cost to estimate p_M by EMUS increases only polynomially in M . For each $M > 0$ and $K \in \mathbb{N}$, let

$$h_i = \frac{M}{K},$$

and define the family of $K + 2$ bias functions

$$\psi_i(x) := \begin{cases} \frac{1}{2} \mathbf{1}_{[0,h]}(x) & \text{for } i = 0 \\ \frac{1}{2} \mathbf{1}_{[(i-1)h, (i+1)h]}(x) & \text{for } i = 1, \dots, K-1 \\ \frac{1}{2} \mathbf{1}_{[M-h, \infty)}(x) & \text{for } i = K, \text{ and} \\ \frac{1}{2} \mathbf{1}_{[M, \infty)}(x) & \text{for } i = K+1. \end{cases} \quad (4.7)$$

As in Section 4.1, let U_i denote the support of ψ_i . This family of bias functions is a partition of unity on $[0, \infty)$; see Figure 2.

We now address the cost of estimating p_M by EMUS. First, we observe that Assumption 3.6 on the asymptotic variances of MCMC averages does not cover the sampling of π_K and π_{K+1} , since the supports of these distributions are unbounded. We need to add to that assumption

Assumption 4.5. Let $f: [0, \infty) \rightarrow \mathbb{R}$, and define $\sigma_i^2(f)$ to be the asymptotic variance of an MCMC trajectory average approximating $\pi_i[f]$ for $i = K, K+1$. We assume

$$\frac{\sigma_i^2(f)}{\text{var}_{\pi_i}(f)} \leq D$$

for some D independent of M and f .

In fact, since Assumption 4.3 implies that the overdamped Langevin dynamics is ergodic for $\pi(x) = \exp(-V(x))$ on the unbounded domain $\Omega = \mathbb{R}$ (cf. Remark 4.4), we fully expect (but do not prove here) that under Assumption 4.3, Assumption 4.5 holds for overdamped Langevin constrained (by reflection as in (3.9)) to remain in the support of π_K or π_{K+1} . Alternatively the reader may simply assume that we draw i.i.d. samples from the biased distributions. All our results hold in that case.

We show in Theorem 4.6 that the relative asymptotic variance of the EMUS estimate of p_M grows only polynomially with M for a broad class of target distributions π . Therefore, EMUS may be dramatically more efficient than direct MCMC sampling when the goal is to compute tail probabilities. We observe that while the hypotheses of the theorem are somewhat restrictive, similar results hold more generally; for example, see Section 5 where we compute tails of a marginal density.

Theorem 4.6. *Let Assumptions 3.6, 4.3, and 4.5 hold. Set*

$$K = M \max_{x \leq M} [IV'(x)] .$$

Define a family of $K + 2$ bias functions ψ_i by (4.7). Take $\kappa_i = 1/(K + 2)$. Let $\sigma_{M,US}^2$ denote the asymptotic variance of the EMUS estimate of p_M . We have

$$\frac{\sigma_{M,US}^2}{p_M^2} \leq CK^2$$

for some constant $C > 0$ depending on V but not on M .

For example, suppose that

$$V(x) = \tilde{V}(x) + x^r,$$

where \tilde{V} has bounded support and $r \geq 1$. Then $|V'(x)| \leq C(1 + M^{r-1})$, and so

$$\frac{\sigma_{M,US}^2}{p_M^2} \leq CM^2(1 + M^{r-1})^2.$$

Proof. The proof is similar to the low temperature limit, Theorem 4.1, but with complications arising because not all strata are bounded and because here we consider the relative variance instead of the variance; see Appendix E. In particular, we require Assumption 4.3 to show that one can in fact choose h so that all nonzero entries of F are bounded above zero uniformly as M increases; cf. Lemma E.1. This is the only part of the proof relying on Assumption 4.3. ■

5. EMUS for tails: An example from Bayesian inference.

We demonstrate the use of EMUS for efficiently exploring and visualizing distributions. In particular, we show how EMUS may be used to efficiently compute both marginal densities and also tail probabilities of the form $\mathbf{P}[\eta(Z) \geq \varepsilon^{-1}]$ where $\eta(Z)$ is a real valued function of a high-dimensional random variable Z . For both tails and marginals, there is a natural and easy to implement choice of strata, which we describe in Section 5.1.

In Section 5.3, we calculate two different one-dimensional marginals of the posterior distribution of the hierarchical Bayesian mixture model described in Section 5.2. For one marginal, the natural stratification suffices. For the other, it does not, but a preliminary computation made with the natural stratification suggests a better choice of strata. We use this example to explain how to diagnose and correct problems related to poorly chosen strata: Our results will serve to guide the practice of stratified MCMC.

5.1. The natural stratification for tails and marginals.

Here, we briefly explain how EMUS can be used to estimate tail probabilities and low-dimensional marginals of high-dimensional distributions. Let $\Omega \subset \mathbb{R}^d$; let π be a probability distribution on Ω ; and let $\eta: \Omega \rightarrow \mathbb{R}$. Suppose that one wishes to estimate the very small tail

probability $\mathbf{P}[\eta(Z) \leq \varepsilon^{-1}]$. In this case, it is natural to stratify in η only. That is, one may choose a partition of unity $\{\phi_i\}_{i=1}^L$ on \mathbb{R} and define bias functions

$$\psi_i(x) = \phi_i(\eta(x)) \text{ for } i = 1, \dots, L \quad (5.1)$$

depending only on η . For a partition of unity, one might choose the regular grid of piecewise constant functions defined in Section 4.2. We refer to (5.1) as the *natural stratification*. To

compute the tail probability, one uses EMUS to estimate $\pi\left(\mathbf{1}_{\left[\varepsilon^{-1}, \infty\right)} \circ \eta\right)$.

Computing marginal densities is similar; in fact, computing tails may be understood as a special case of computing a marginal density. Suppose now that $\eta: \Omega \rightarrow \mathbb{R}^\ell$. To estimate the marginal π_η of π in η , one chooses a partition of unity $\{\phi_i\}_{i=1}^L$ on \mathbb{R}^ℓ , again defining bias functions by (5.1). One then uses EMUS to compute averages of *histogram bins*, which are functions of the form

$$b_{\eta_0}(\eta(x)) = \mathbf{1}_{\eta_0 + h[-1, 1]}(\eta(x)). \quad (5.2)$$

We have

$$\lim_{h \rightarrow 0} \frac{1}{\varepsilon^{2h}} \pi\left[b_{\eta_0}\right] = \pi_\eta(\eta_0),$$

so for small h the averages of the histogram bins approximate π_η .

By the argument in Section 4.2, EMUS with the natural stratification will be dramatically more efficient than direct sampling *as long the biased distributions are no harder to sample than the target distribution π* . Essentially, this is because with the natural stratification very small averages like $\mathbf{P}[\eta(Z) \leq \varepsilon^{-1}]$ over the target distribution π are expressed as functions of much larger averages over the biased distributions π_i . Unfortunately, however, for general functions η , the biased distributions of the natural stratification need not be easy to sample. In Section 5.3, we give one example where the natural stratification works and one where it does not. In the case where it does not, we explain how to make a better choice of strata.

5.2. A hierarchical Bayesian mixture model.

Here, we review the hierarchical Bayesian mixture model proposed in [39], and we discuss the difficulties which complicate inference under this model. As a tutorial in the use of EMUS, we present a numerical investigation of these difficulties in Section 5.

In the hierarchical mixture model, the data vector $\mathbf{y} = (y_1, \dots, y_n) \in \mathbb{R}^n$ consists of independent identically distributed samples drawn from a mixture distribution of the form

$$p(y_i | \phi) = \sum_{k=1}^K q_k \nu(y_i; \mu_k, \lambda_k^{-1}),$$

where K is the number of mixture components, q_k is the weight of the k 'th mixture component, $\nu(\cdot; \mu_k, \lambda_k^{-1})$ is the normal density with mean μ_k and variance λ_k^{-1} , and ϕ is the vector of parameters

$$\phi = (\mu_1, \dots, \mu_K, \lambda_1, \dots, \lambda_K, q_1, \dots, q_{K-1}).$$

(Since $p(y_i | \phi)$ is a probability distribution, $q_1 + \dots + q_K = 1$, and q_1, \dots, q_{K-1} determine q_K) The following prior distribution is imposed on ϕ :

$$\begin{aligned} \mu_i &\sim \mathcal{N}(m, \kappa^{-1}) \\ \lambda_k &\sim \text{Gamma}(\alpha, \beta) \\ \beta &\sim \text{Gamma}(g, h) \\ (q_1, \dots, q_{K-1}) &\sim \text{Dirichlet}_K(1, \dots, 1). \end{aligned}$$

As in [23, 11], we choose

$$m = M, \kappa = \frac{4}{R^2}, \alpha = 2, g = 0.2, \text{ and } h = \frac{100g}{\alpha R^2}$$

where R and M are the range and the mean of the observed data, respectively. The posterior density is

$$\begin{aligned} p(\theta | \mathbf{y}) &= \frac{\kappa^{K/2} g^h \beta^{K\alpha + g - 1}}{Z_K \Gamma(\alpha)^K \Gamma(g)(2\pi)^{\frac{n+K}{2}}} \left(\prod_{k=1}^K \lambda_k \right)^{\alpha-1} \\ &\times \exp \left[-\frac{\kappa}{2} \sum_{k=1}^K (\mu_k - M)^2 - \beta \left(h + \sum_{k=1}^K \lambda_k \right) \right] \\ &\times \prod_{i=1}^N \left(\sum_{k=1}^K q_k \lambda_k^{\frac{1}{2}} \exp \left[-\frac{\lambda_k}{2} (y_i - \mu_k)^2 \right] \right), \end{aligned}$$

where $\theta = (\phi, \beta)$ denotes the vector of all parameters to be inferred, including the hyperparameter β .

Several factors complicate inference based on this model: First, the mixture components are not identifiable; that is, the posterior distribution is invariant under permutation of the labels of the mixture components. Consequences of non-identifiability are discussed at length in [23, 11]. In our computations in Section 5.3, we impose the constraint

$$\mu_1 \leq \mu_2 \leq \dots \leq \mu_K$$

to ensure that the components are identifiable. Second, in Lemma 5.1, we show that the posterior density may be unbounded, introducing spurious modes with infinite density. Finally, even with identifiability constraints, the posterior distribution may have multiple modes of finite posterior density. For example, see the modes reported in [11]. In Section 5.3, we use EMUS to efficiently visualize the posterior, assessing the effects of multimodality and unboundedness.

We suspect that the unboundedness of the posterior for this model is well known. However, we are unable to find a reference, so we now explain. It is certainly well known that the likelihood of a Gaussian mixture model is unbounded: Roughly speaking, the likelihood is infinite when any mixture component is collapsed on a single data point [1]. Nonetheless, one might expect the posterior density $p(\theta|\mathbf{y})$ to be bounded, since the prior penalizes large values of the precisions λ_j . This is not always the case when the data vector contains repeated entries:

Lemma 5.1. If any datum y_i has frequency N_i greater than

$$2g + 2(K - 1)\alpha,$$

then the posterior density $p(\theta|\mathbf{y})$ is unbounded.

Proof. Take the limit of $p(\theta|\mathbf{y})$ as $\lambda_1 \rightarrow \infty$ with $\mu_1 = y_r$, $\beta = \lambda_1^{-1}$ and all other variables held fixed. ■

The reader will observe that under the model, the set of data vectors with repeated entries has probability zero. However, in practice, the data consist of measurements with finite precision, and therefore repeated entries occur commonly, cf. the Hidalgo stamp data used in Section 5.3.

5.3. Numerical experiments: Choosing strata, computing tails, diagnosis of problems.

In this section, we explain how to recognize and correct problems related to poor choices of strata, and we demonstrate the use of EMUS to investigate the multimodality and unboundedness of the posterior in the mixture model. We first compute two one-dimensional marginals of the high-dimensional posterior density $p(\theta|\mathbf{y})$ using the natural stratification (5.1). The natural stratification works in one case but not the other. In the case where the natural stratification does not work, preliminary calculations based on the natural stratification suggest a better choice of strata.

Here, we let \mathbf{y} be the Hidalgo stamp data set first studied in [22], consisting of the thicknesses of 485 stamps, ranging between 60 μm and 130 μm . We let there be three mixture components ($K = 3$), following previous computational studies [11, 23]. In our first calculation, we estimated the marginal in μ_2 using the natural stratification with a grid of 201

bias functions covering the range $[7, 11]$, with the support of the leftmost and rightmost bias functions reaching to $-\infty$ and ∞ , respectively. For the middle strata, define $\phi_1: \mathbb{R} \rightarrow \mathbb{R}$ by

$$\phi_1(x) := \max\{0, 1 - |x|\}. \quad (5.3)$$

We used the bias functions

$$\psi_i(\theta) = \phi_1\left(\frac{\mu_2 - (7 + (i-1)h)}{h}\right), \text{ where } h := 0.02 \quad (5.4)$$

for $i = 2, \dots, 200$. Now, define $\phi_2: \mathbb{R} \rightarrow \mathbb{R}$ by

$$\phi_2(x) := \min\{\max\{0, 1 - x\}, 1\} \quad (5.5)$$

The first and last bias functions were

$$\psi_1(\theta) = \phi_2\left(\frac{\mu_2 - 7}{h}\right) \quad (5.6)$$

$$\psi_{201}(\theta) = \phi_2\left(\frac{(7 + 200h) - \mu_2}{h}\right), \quad (5.7)$$

where $h = 0.02$ as before.

We chose the total number of bias functions based on the sizes of the off-diagonal entries in the overlap matrix. For any bias functions of the form (5.4), the overlap matrix is tridiagonal. Thus, by Remark 3.12, if the superdiagonal and subdiagonal entries $F_{i,i+1}$ and $F_{i,i-1}$ are sufficiently large, then the EMUS estimator is not too sensitive to statistical errors in \bar{F} . For our choice of bias functions,

$$\min\{F_{i,i+1}; i = 1, \dots, 200\} \geq 0.01 \text{ and } \min\{F_{i,i-1}; i = 2, \dots, 201\} \geq 0.004. \quad (5.8)$$

We sampled the biased distributions using the affine invariant ensemble sampler with 100 walkers, as implemented in the emcee package [15]. Due to computational restrictions on memory, only every tenth sample point was saved. As a check on the sampling, the average acceptance probability over all walkers in the ensemble sampler was calculated for each biased distribution. Averaging over biased distributions gave a total average acceptance probability of 0.31. The minimum acceptance probability over all distributions was 0.12.

To initialize sampling, we computed an unbiased test trajectory; that is, a trajectory having ergodic distribution π . We then started by sampling a single biased distribution π_k , initializing with points drawn randomly from the unbiased trajectory. We sampled the other biased distributions in sequence, initializing with points drawn randomly from samples of adjacent biased distributions. Thus, we sampled π_k first, then π_{k-1} and π_{k+1} , then π_{k-2} and π_{k+2} , etc. We equilibrated the sampler in each π_i for 3000 Monte Carlo steps, and collected

data for an additional 100000 Monte Carlo steps. Each step of the ensemble sampler involves perturbing the positions of each of the 100 walkers.

We computed the marginal in μ_2 using a grid of 200 histogram bins, covering the region [7, 11]; this corresponds to taking $h = 0.01$ in (5.2). The result is the curve labeled EMUS in Figure 3a. The marginal in μ_2 has two modes, labeled 1 and 2 in Figure 3a. We plot the mixture distributions corresponding to these modes in Figure 4. (To be precise, the distributions in Figure 4 correspond to means over histogram bins centered at the labeled points.)

For comparison, we also estimated the marginal in μ_2 from multiple long, unbiased trajectories. We computed 100 unbiased trajectories of the affine invariant ensemble sampler in parallel. For each trajectory, the ensembles were first equilibrated for 10000 Monte Carlo steps, and then data were collected for 100000 steps. These trajectories were combined and binned to produce the density labeled Unbiased in Figure 4. We estimated the relative asymptotic variance of the marginal density for the unbiased calculation using ACOR [14], and we estimated the relative asymptotic variance for the EMUS calculation using the method outlined in Appendix F. We present the results in Figure 3a. Note that near the mode, unbiased MCMC performs slightly better than EMUS, but in the tails, EMUS performs dramatically better.

After computing the marginal in μ_2 , we tried computing the marginal in $\log_{10} \lambda_1$. We used the natural stratification with a grid of 50 bias functions with maxima equally spaced between -1 and 3.2 constructed as

$$\psi_i(\theta) = \phi\left(\frac{-1 + h(i-1) - \log_{10} \lambda_1}{h}\right)$$

where

$$h = \frac{3.2 - (-1)}{49}.$$

We used the same initialization scheme as for the marginal in μ_2 , beginning with a single biased distribution initialized from an unbiased test trajectory. We call this the *center* sample. The result of this calculation was the density labeled “1D Center” in Figure 5a. When we tried to compute the asymptotic variance of this density estimate, we noticed very slow convergence of the sampler for some biased distributions. To investigate, we performed another EMUS calculation using a similar initialization procedure, but starting from π_1 , the biased distribution at the extreme left, covering the lowest values of λ_1 . We call this the *left* sample. The result of this second calculation was the density labeled “1D Left” in Figure 5a. For both the center and left samples, the strata were equilibrated for 3000 steps and sampled for another 200000. We observe that the two densities differ significantly in the region $-1 \leq \log_{10} \lambda_1 \leq 0.5$. They should be the same up to sampling errors; for example, we observe that different initializations have no effect on the calculation of the marginal in μ_2 , cf. Figure 3a.

Figure 6 explains the problem and suggests a solution: In the region $0.2 \leq \log_{10} \lambda_1 \leq 0.7$, the center and left samples cover entirely different ranges of $\log_{10} \lambda_2$. This suggests that the biased distributions corresponding to the range $0.2 \leq \log_{10} \lambda_1 \leq 0.7$ are multimodal, with barriers in λ_2 impeding sampling.

To confirm the hypothesis that barriers in λ_2 were responsible for the poor convergence observed in the center and left samples, we performed a third calculation, stratifying in both $\log_{10} \lambda_1$ and $\log_{10} \lambda_2$. We used a 50×50 grid of bilinear bias functions, with maxima equally spaced between -1 and 3.2 . To be precise, for $i, j = 1, \dots, 50$, we defined the bias functions

$$\psi_{ij}(\theta) = \phi\left(\frac{-1 + h(i-1) - \log_{10} \lambda_1}{h}\right) \times \phi\left(\frac{-1 + h(j-1) - \log_{10} \lambda_2}{h}\right),$$

with h as before. Let η_{ij} denote the biased distribution corresponding to ψ_{ij} .

We performed the two-dimensional EMUS calculation twice, initializing from the center and left samples drawn from the natural stratification in $\log_{10} \lambda_1$. For each $i = 1, \dots, L$, to sample the row $\{\eta_{ij} : j = 1, \dots, 50\}$ of biased distributions, we began by initializing sampling of a single biased distribution η_{ik} with points from the either the center or left sample of π_j . We then sampled the other distributions η_{ij} for $j \neq k$ in sequence, again initializing with points from samples of adjacent distributions, either $\eta_{i,j+1}$ or $\eta_{i,j-1}$ in this case. If no samples were found inside the support of a biased distribution, that distribution was ignored. For each biased distribution, sampling was burned in for 4500 steps, and samples were collected for an additional 2500 steps. Ultimately, 1397 of the 2500 biased distributions were sampled; the unsampled distributions correspond to the white space in Figure 7a.

We computed the marginal in $\log_{10} \lambda_1$ and $\log_{10} \lambda_2$ using a 200×200 grid of histogram bins, covering the region $-1 \leq \log_{10} \lambda_1 \leq 3.2$ and $-1 \leq \log_{10} \lambda_2 \leq 3.2$; this corresponds to taking $h = (3.2 - (-1))/200$ in (5.2); the result from the center calculation appears in Figure 7a. In Figure 8, we show the mixture distributions corresponding to the modes of the two-dimensional marginal in Figure 7a. The two-dimensional marginals were essentially the same for the center and left initializations; see Figure 9. We also estimated the one-dimensional marginal in $\log_{10} \lambda_1$ using the two-dimensional stratification; see the results labeled “2D Center” and “2D Left” in Figure 5a. Finally, we estimated the relative asymptotic variance of the marginal in $\log_{10} \lambda_1$ computed by two-dimensional stratification. Again, we observe that EMUS performs much better than unbiased sampling in the tails, cf. Figure 5b.

The marginal in $\log_{10} \lambda_1$ and $\log_{10} \lambda_2$ confirms that barriers in λ_2 caused the problems observed in calculating the marginal in $\log_{10} \lambda_1$ using the natural stratification. In fact, we see that computing the marginal in either λ_1 or λ_2 requires stratifying both variables, as stratifying only one leads to barriers that impede sampling in the other. In particular, there are barriers in λ_2 along the line $\log_{10} \lambda_1 = 0.45$ and a barrier in λ_1 along $\log_{10} \lambda_2 = 0.6$: In

Figure 10, we plot an estimate of the conditional distribution of $\log_{10} \lambda_2$ with $\log_{10} \lambda_1 = 0.45$ fixed. This distribution is multimodal with a region of very low probability separating the modes, which explains the poor sampling depicted in Figure 6.

To conclude, we have confirmed that EMUS can be extremely efficient for computing tails. However, one must exercise care in the choice of strata. The natural stratification often suffices, but in some cases, like computing the marginal in $\log_{10} \lambda_1$, the biased distributions of the natural stratification may be very difficult to sample. We propose the use of different initializations, like the center and left samples, as a method of identifying problems related to poorly chosen strata. Careful inspection of simulations performed with these different initializations can identify problems and suggest better strata.

6. Conclusions.

We have analyzed the Eigenvector Method for Umbrella Sampling (EMUS), an especially simple and effective stratified MCMC method sharing many features with the popular WHAM [27] and MBAR [42] methods of computational chemistry. We have demonstrated the advantages of EMUS for sampling from multimodal distributions and computing tail probabilities, and we have explained how to identify and resolve the problems which may occur if the method is implemented poorly. We have also given a tutorial intended to explain how to diagnose and correct problems related to poorly chosen strata.

Our purpose was to explain the benefits of stratified MCMC analytically, with the ultimate goal of introducing stratified MCMC to a diverse audience of statisticians, engineers, and scientists. Since stratified MCMC had previously been applied only to a particular class of statistical mechanics calculations without any general justification, we began by developing a general theory. We hope that our theory will serve as the basis for further developments. For example, it may now be possible to undertake a comparison of EMUS and other so-called reaction coordinate methods such as Wang–Landau sampling [50] or Metadynamics [28]. Despite some similarities with EMUS, these methods work by a substantially different mechanism and understanding the relative advantages of the two approaches is non-trivial.

Acknowledgments

BvK was supported by NSF RTG: Computational and Applied Mathematics in Statistical Science, number 1547396. ARD and EHT were supported by National Institutes of Health (NIH) Grant Number 5 R01 GM109455-02. We wish to thank Jonathan Mattingly, Jeremy Tempkin, and Charlie Matthews for helpful discussions.

Appendix A.: Proof of Theorem 3.3.

Our proof of Theorem 3.3 (the CLT for EMUS) is based on the delta method. To apply the delta method, we require the following result ensuring the differentiability of $w(G)$:

Lemma A.1. The function $w(G)$ admits an extension $\tilde{w}: \mathbb{R}^{L \times L} \rightarrow \mathbb{R}^L$ which is differentiable on the set of irreducible stochastic matrices.

Proof. By [45, Lemma 3.1], $w(G)$ admits a continuously differentiable extension to an open set $U \subset \mathbb{R}^{L \times L}$. We further extend the domain of $w(G)$ to $\mathbb{R}^{L \times L}$ by arbitrarily defining $w(G) = 0$ whenever $G \in \mathbb{R}^{L \times L} \setminus U$. ■

The extension in Lemma A.1 resolves two technicalities: First, the set of stochastic matrices is not a vector space but a compact, convex subset of $\mathbb{R}^{L \times L}$ with empty interior. Therefore, the derivative of w is undefined. Second, \bar{F} may be reducible for some values of N and some realizations of the processes sampling the biased distributions. In that case, the invariant distribution of \bar{F} is not unique, so $w(\bar{F})$ is undefined. Throughout the remainder of this work, $w(G)$ will denote the extension guaranteed by the lemma. We now prove the CLT for EMUS.

Proof of Theorem 3.3. The proof is based on the delta method [6, Proposition 6.2] and a formula for $w'(\bar{F})$ given in [19].

By Lemma A.1, $w(\bar{F})$ is differentiable at F , so the function

$$B(\bar{F}, \{\bar{g}_i^*\}_{i=1}^L, \{\bar{\mathbf{1}}_i^*\}_{i=1}^L) = \pi_{\text{US}}[g] = \frac{\sum_{i=1}^L w_i(\bar{F}) \bar{g}_i^*}{\sum_{i=1}^L w_i(\bar{F}) \bar{\mathbf{1}}_i^*}$$

is differentiable at $(F, \{\pi_i[g^*]\}_{i=1}^L, \{\pi_i[\mathbf{1}^*]\}_{i=1}^L)$. Let $\partial_i B \in \mathbb{R}^{L+2}$ be the derivative of B with respect to those quantities computed from X_t^i . That is,

$$\partial_i B = \left(\frac{\partial B}{\partial \bar{F}_{i:}}, \frac{\partial B}{\partial \bar{G}_{i:}} \right) \in \mathbb{R}^{L+2}, \quad (\text{A.1})$$

where $\frac{\partial B}{\partial \bar{F}_{i:}} \in \mathbb{R}^L$ denotes the partial derivative of B with respect to the i 'th row of \bar{F} and

$$\frac{\partial B}{\partial \bar{G}_{i:}} = \left(\frac{\partial B}{\partial \bar{g}_i^*}, \frac{\partial B}{\partial \bar{\mathbf{1}}_i^*} \right) \in \mathbb{R}^2.$$

To simplify notation, we will assume throughout the remainder of this argument that all derivatives are evaluated at $(F, \{\pi_i[g^*]\}_{i=1}^L, \{\pi_i[\mathbf{1}^*]\}_{i=1}^L)$. In formulas involving matrix multiplication, we will treat B , $\frac{\partial B}{\partial \bar{F}_{i:}}$ and $\frac{\partial B}{\partial \bar{G}_{i:}}$ as row vectors.

Since we assume that the processes X_t^i sampling the different measures π_i are independent, [5, Chapter 1, Theorem 2.8] implies that

$$\begin{aligned} & \sqrt{M}((\bar{F}_1, \bar{g}_1, \bar{\mathbf{1}}_1^*, \dots, \bar{F}_L, \bar{g}_L, \bar{\mathbf{1}}_L^*) - (F_1, \pi_1[g^*], \pi_1[\mathbf{1}^*], \dots, F_L, \pi_L[g^*], \pi_L[\mathbf{1}^*])) \\ & \xrightarrow{d} N(0, \Sigma), \end{aligned} \quad (\text{A.2})$$

where Σ is the covariance matrix of the product of the distributions $N(0, \kappa_i^{-1} \Sigma_i)$. (That is, $\Sigma \in \mathbb{R}^{L(L+2) \times L(L+2)}$ is the block diagonal matrix with the matrices $\kappa_i^{-1} \Sigma_i$ along the diagonal.) Therefore, by the delta method,

$$\sqrt{M}(\pi_{\text{US}}[g] - \pi[g]) \xrightarrow{d} N(0, \sigma^2),$$

where

$$\begin{aligned} \sigma^2 &= (\partial_1 B, \dots, \partial_L B) \Sigma (\partial_1 B, \dots, \partial_L B)^t \\ &= \sum_{i=1}^L \kappa_i^{-1} \partial_i B \Sigma_i \partial_i B^t. \end{aligned} \tag{A.3}$$

Now we observe that for any column vector $v \in \mathbb{R}^L$ having mean zero,

$$\left. \frac{d}{d\varepsilon} w_k(F + \varepsilon e_i v^t) \right|_{\varepsilon=0} = \frac{\partial w_k}{\partial F_{i:}} v = z_i v^t (I - F)^{\#} e_k,$$

by [19, Theorem 3.1]. (In the formula above, $e_i \in \mathbb{R}^L$ denotes the i th standard basis vector.)

Therefore, we have

$$\begin{aligned} \frac{\partial B}{\partial F_{i:}} v &= \frac{\sum_{k=1}^L \frac{\partial w_k}{\partial F_{i:}} v \pi_k[g^*]}{\sum_{k=1}^L z_k \pi_k[\mathbf{1}^*]} - \frac{\sum_{k=1}^L \frac{\partial w_k}{\partial F_{i:}} v \pi_k[\mathbf{1}^*]}{\sum_{k=1}^L z_k \pi_k[\mathbf{1}^*]} \frac{\sum_{i=1}^L z_k \pi_k[g^*]}{\sum_{k=1}^L z_k \pi_k[\mathbf{1}^*]} \\ &= \sum_{k=1}^L \frac{\partial w_k}{\partial F_{i:}} v \Psi(\pi_k[g^*] - \pi[g] \pi_k[\mathbf{1}^*]) \\ &= z_i v^t (I - F)^{\#} \mathbf{g}, \end{aligned} \tag{A.4}$$

where

$$\mathbf{g}_k = \Psi \pi_k[g^* - \pi[g] \mathbf{1}^*] = \ell \cdot (\pi_k[g^*], \pi_k[\mathbf{1}^*]).$$

(Equality (A.4) above follows from (2.3) and the definition (2.2) of Ψ .) Also,

$$\frac{\partial B}{\partial \bar{G}_i} = z_i \Psi(1, -\pi[g]) = z_i \ell. \tag{A.5}$$

Thus,

$$\begin{aligned}\partial_i B \Sigma_i \partial_i^t B^t &= \frac{\partial B}{\partial F_{ij}} \sigma^i \frac{\partial B^t}{\partial F_{ij}} + 2 \frac{\partial B}{\partial F_{ij}} \rho_i \frac{\partial B^t}{\partial G_{ij}} + \frac{\partial B}{\partial G_{ij}} \tau_i \frac{\partial B^t}{\partial G_{ij}} \\ &= z_i^2 \left\{ (I - F)^{\#} \mathbf{g} \cdot \sigma^i (I - F)^{\#} \mathbf{g} + 2(I - F)^{\#} \mathbf{g} \cdot \rho_i \ell + \ell^t \tau_i \ell \right\},\end{aligned}$$

and the result follows by (A.3). ■

Appendix B.: Proof of Theorem 3.5.

Definition B.1. Let $e_i \in \mathbb{R}^L$ denote the i 'th standard basis vector. For $i, j \in \{1, 2, \dots, L\}$ with $i \neq j$, define the logarithmic partial derivatives

$$\begin{aligned}\frac{\partial \log w_k}{\partial F_{ij}}(F) &:= \frac{\partial}{\partial F_{ij}} \log w_k \left(\sum_{i \neq j} I + F_{ij} (e_i e_j^t - e_i^t e_j) \right) \\ &= \frac{d}{d\varepsilon} \Big|_{\varepsilon=0} \log w_k \left(F + \varepsilon (e_i e_j^t - e_i^t e_j) \right).\end{aligned}\tag{B.1}$$

(These partial derivatives must be understood as derivatives of the extension guaranteed by Lemma A.1; otherwise, they are defined only when $F_{ij} > 0$ and $F_{ji} > 0$.)

Our definition of logarithmic partial derivatives in (B.1) is not standard. However, we observe that a version of the standard formula relating the total and partial derivatives of $\log w$ holds: For all matrices H whose rows sum to zero,

$$\frac{d}{d\varepsilon} \Big|_{\varepsilon=0} \log w_k(F + \varepsilon H) = \sum_{i \neq j} \frac{\partial \log w_k}{\partial F_{ij}}(F) H_{ij}.\tag{B.2}$$

We need only consider matrices whose rows sum to zero, since these are the only perturbations for which $F + \varepsilon H$ can be stochastic.

The following result appears in [45, Theorem 3.6]. It is crucial in our proof of Theorem 3.5.

Lemma B.2. Recall $\mathbf{P}_i[t_j < t_i]$ and $\frac{\partial \log w_k}{\partial F_{ij}}$ and from Definitions 3.4 and B.1. For all stochastic irreducible matrices F , we have

$$\frac{1}{2} \frac{1}{\mathbf{P}_i[t_j < t_i]} \leq \max_k \left| \frac{\partial \log w_k}{\partial F_{ij}}(F) \right| \leq \frac{1}{\mathbf{P}_i[t_j < t_i]}.$$

We also require the following lemma in the proof of Theorem 3.5.

Lemma B.3. The asymptotic covariance matrix σ^j has the properties:

1. The rows and columns of σ^j sum to zero. That is, for $e \in \mathbb{R}^L$ the vector of all ones,

$$\sigma^i e = 0 \text{ and } e^t \sigma^i = 0.$$

2. For all $j = 1, \dots, L$,

$$\sigma_{jk}^i = \sigma_{kj}^i = 0 \text{ whenever } F_{ik} = 0.$$

Proof. Since the rows of \bar{F} sum to one with probability one, we have

$$\text{var}(\bar{F}_{i:}, e) = 0$$

for any fixed number of samples N_i . Therefore, the asymptotic variance σ^i has $e^t \sigma^i e = 0$, and it follows that $e^t \sigma^i = \sigma^i e = 0$ since σ^i is symmetric and positive semidefinite.

Let k be such that $F_{ik} = 0$. Since $\bar{F}_{ik} = 0$ with probability one, we have

$$\text{cov}(\bar{F}_{ik}, \bar{F}_{ij}) = 0$$

for any $j = 1, \dots, L$, and therefore $\sigma_{jk}^i = 0$. ■

We now prove Theorem 3.5.

Theorem 3.5. We begin with formula (A.3):

$$\sigma^2 = \sum_{i=1}^L \kappa_i^{-1} \partial_i B^t \Sigma_i \partial_i B. \quad (\text{B.3})$$

Since the asymptotic covariance matrix Σ_i is symmetric and positive semidefinite, the Cauchy inequality holds:

$$a^t \Sigma_i b \leq \frac{1}{2} a^t \Sigma_i a + \frac{1}{2} b^t \Sigma_i b,$$

for all $a, b \in \mathbb{R}^{L+1}$. Therefore,

$$\begin{aligned} \partial_i B^t \Sigma_i \partial_i B &= \left(\frac{\partial B}{\partial \bar{F}_{i:}}, \frac{\partial B}{\partial \bar{G}_{i:}} \right) \Sigma_i \left(\frac{\partial B}{\partial \bar{F}_{i:}}, \frac{\partial B}{\partial \bar{G}_{i:}} \right)^t \\ &\leq 2 \left(\frac{\partial B}{\partial \bar{F}_{i:}}, 0 \right) \Sigma_i \left(\frac{\partial B}{\partial \bar{F}_{i:}}, 0 \right)^t + 2 \left(0^t, \frac{\partial B}{\partial \bar{G}_{i:}} \right) \Sigma_i \left(0^t, \frac{\partial B}{\partial \bar{G}_{i:}} \right)^t \\ &= 2 \frac{\partial B}{\partial \bar{F}_{i:}} \sigma^i \frac{\partial B}{\partial \bar{F}_{i:}}^t + 2 \frac{\partial B}{\partial \bar{G}_{i:}} \tau_i \frac{\partial B}{\partial \bar{G}_{i:}}^t. \\ &=: 2A_0 + 2A_1. \end{aligned} \quad (\text{B.4})$$

(Here, $\mathbf{0}$ denotes the zero vector in \mathbb{R}^L , interpreted as a column vector.)

We now estimate the term A_0 defined above. By (A.4), we have

$$\begin{aligned}
 A_0 &= \frac{\partial B^t}{\partial \bar{F}_i} \sigma^i \frac{\partial B}{\partial \bar{F}_i} \\
 &= \sum_{j,k,\ell,m=1}^L \mathfrak{g}_\ell \frac{\partial w_\ell}{\partial \bar{F}_{ij}} \sigma_{jk}^i \mathfrak{g}_m \frac{\partial w_m}{\partial \bar{F}_{ij}} \\
 &= \sum_{\ell,m=1}^L z_\ell \mathfrak{g}_\ell z_m \mathfrak{g}_m \sum_{j \neq i} \sum_{\substack{k \neq i \\ F_{ij} > 0 \ F_{ik} > 0}} \frac{\partial \log w_\ell}{\partial \bar{F}_{ij}} \sigma_{jk}^i \frac{\partial \log w_m}{\partial \bar{F}_{ik}}. \\
 &= \sum_{\ell,m=1}^L z_\ell \mathfrak{g}_\ell z_m \mathfrak{g}_m \sum_{j \neq i} \sum_{\substack{k \neq i \\ F_{ij} > 0 \ F_{ik} > 0}} \sqrt{\text{var}_{\pi_i}(\psi_j^*)} \frac{\partial \log w_\ell}{\partial \bar{F}_{ij}} R_{jk}^i \sqrt{\text{var}_{\pi_i}(\psi_k^*)} \frac{\partial \log w_m}{\partial \bar{F}_{ik}},
 \end{aligned} \tag{B.5}$$

where

$$R_{jk}^i = \frac{\sigma_{jk}^i}{\sqrt{\text{var}_{\pi_i}(\psi_j^*)} \sqrt{\text{var}_{\pi_i}(\psi_k^*)}}.$$

(The third equality above follows from formula (B.2) relating the total and partial derivatives of $\log w$, since the rows and columns of σ^i sum to zero by Lemma B.3.)

We claim that

$$\begin{aligned}
 &\sum_{\substack{j \neq i \\ F_{ij} > 0}} \sum_{\substack{k \neq i \\ F_{ik} > 0}} \sqrt{\text{var}_{\pi_i}(\psi_j^*)} \frac{\partial \log w_\ell}{\partial \bar{F}_{ij}} R_{jk}^i \sqrt{\text{var}_{\pi_i}(\psi_k^*)} \frac{\partial \log w_m}{\partial \bar{F}_{ik}} \\
 &\leq \text{tr}(R^i) \sum_{\substack{j \neq i \\ F_{ij} > 0}} \text{var}_{\pi_i}(\psi_j^*) \frac{\partial \log w_\ell}{\partial \bar{F}_{ij}}^2.
 \end{aligned} \tag{B.6}$$

To prove this, we observe that R^i is symmetric and positive semidefinite since σ^i is symmetric and positive semidefinite. Therefore, R^i has the spectral decomposition

$$R_i = \sum_{j=1}^L \lambda_{i,j} v^{i,j} (v^{i,j})^t$$

with eigenvalues $\lambda_{i,j} > 0$ and corresponding eigenvectors $v^{i,j}$ such that $\|v^{i,j}\| = 1$. Thus, for any $a \in \mathbb{R}^L$,

$$a^t R^i a = \sum_{j=1}^L \lambda_{i,j} |v^{i,j} \cdot a|^2 \leq \left(\sum_{j=1}^L \lambda_{i,j} \right) \|a\|^2 = \text{tr}(R^i) \|a\|^2. \quad (\text{B.7})$$

Inequality (B.6) follows from (B.7) by setting

$$a_j = \begin{cases} \sqrt{\text{var}_{\pi_i}(\psi_j^*)} \frac{\partial \log w_\ell}{\partial F_{ij}} & \text{if } j \neq i \text{ and } F_{ij} > 0, \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

Finally, combining (B.5), (B.6), and Lemma B.2 yields

$$A_0 \leq \text{tr}(R^i) \left(\sum_{\ell=1}^L z_{\ell} |q_{\ell}| \right)^2 \sum_{\substack{j \neq i \\ F_{ij} > 0}} \frac{\text{var}_{\pi_i}(\psi_j^*)}{\mathbf{P}_i[\tau_j < \tau_i]^2}.$$

Moreover, we have

$$\sum_{\ell=1}^L z_{\ell} |q_{\ell}| = \Psi \sum_{\ell=1}^L z_{\ell} |\pi_{\ell}[g^* - \pi[g] \mathbf{1}^*]| \leq \Psi \sum_{\ell=1}^L z_{\ell} \pi_{\ell}[|h|] = \pi[|h|],$$

by (2.3), and therefore

$$A_0 \leq \text{tr}(R^i) \pi[|h|]^2 \sum_{\substack{j \neq i \\ F_{ij} > 0}} \frac{\text{var}_{\pi_i}(\psi_j^*)}{\mathbf{P}_i[\tau_j < \tau_i]^2}. \quad (\text{B.8})$$

We now observe that by (A.5)

$$A_1 = z_i^2 \ell^t \tau_i \ell = \Psi^2 z_i^2 C(\tilde{h}_i).$$

$C(\tilde{h}_i)$ denotes the asymptotic covariance of the trajectory average \tilde{h}_i of h over the biased process X_t^i . Therefore, combining (B.3) and (B.8), we find

$$\sigma^2 \leq 2 \sum_{i=1}^L \frac{1}{\kappa_i} \left(z_i^2 \Psi^2 C(\tilde{h}_i) + \text{tr}(R^i) \pi[|h|]^2 \sum_{\substack{j \neq i \\ F_{ij} > 0}} \frac{\text{var}_{\pi_i}(\psi_j^*)}{\mathbf{P}_i[\tau_j < \tau_i]^2} \right),$$

as desired.

Finally, suppose that the bias functions are a partition of unity. In that case,

$$\pi[h]^2 = \pi[g - \pi[g]]^2 \leq \pi[g - \pi[g]]^2 = \text{var}_{\pi}(g),$$

and so we may replace $\pi[h]^2$ with $\text{var}_{\pi}(g)$. In addition, we observe that for a partition of unity, equation (A.4) holds with $g_k = \pi_k[g]$. Thus, following the argument above, one may verify that the result holds with $\pi[g]^2$ in place of $\pi[h]^2$.

Appendix C.: Proof of Theorem 3.7.

In the arguments below, for any probability measure ν on a set Ω , we let

$$L^2(\nu) := \left\{ u: \Omega \rightarrow \mathbb{R} : \int_{\Omega} u^2 d\nu < \infty \right\},$$

and we define the $L^2(\nu)$ inner product

$$\langle f, g \rangle_{\nu} = \int_{\Omega} fg d\nu$$

with the corresponding norm

$$\|f\|_{L^2(\nu)} := \sqrt{\langle f, f \rangle_{\nu}}.$$

Given a set $U \subset \mathbb{R}^d$, we define $L^2(U)$, $\|\cdot\|_{L^2(U)}$, $\langle \cdot, \cdot \rangle_U$ to be the analogous function space, norm, and inner product for Lebesgue measure on U .

Our proof of Theorem 3.7 requires a Poincaré inequality, Lemma C.1. We refer to [31, Section 3] for an introduction to Poincaré inequalities and their role in the theory of diffusion processes.

Lemma C.1. *Assume that the Poincaré inequality holds for U with constant Λ ; that is, assume that for all weakly differentiable $f: U \rightarrow \mathbb{R}$ so that $\nabla f \in L^2(U)$,*

$$\|f - \int_U f dx\|_{L^2(U)} \leq \Lambda(U) \|\nabla f\|_{L^2(U)}$$

We have a similar Poincaré inequality for π_h :

$$\|f - \pi_h(f)\|_{L^2(\pi_h)} \leq h\Lambda(U) \exp\left(\frac{\beta}{2} \left(\sup_{U_h} V - \inf_{U_h} V \right)\right) \|\nabla f\|_{L^2(\pi_h)}.$$

Proof. By a standard scaling argument, the Poincaré inequality holds for U_h with constant $h\Lambda$. To see this, let $A_h: U \rightarrow U_h$ be the affine transformation

$$A_h x = x_0 + h(x - x_0).$$

For any $f: U_h \rightarrow \mathbb{R}$ with $\nabla f \in L^2(U_h)$, using the change of variable formula and the chain rule, we have

$$\begin{aligned} \|f - \int_{U_h} f\|_{L^2(U_h)}^2 &= h^d \|f \circ A_h - \int_U f \circ A_h\|_{L^2(U)}^2 \\ &\leq h^d \Lambda^2 \|\nabla(f \circ A_h)\|_{L^2(U)}^2 \\ &= h^d h^2 \Lambda^2 \|(\nabla f) \circ A_h\|_{L^2(U)}^2 \\ &= h^2 \Lambda^2 \|\nabla f\|_{L^2(U_h)}^2. \end{aligned}$$

Now observe that for any $f \in L^2(\pi_h)$,

$$\|f - \pi_h[f]\|_{L^2(\pi_h)} = \min_{c \in \mathbb{R}} \|f - c\|_{L^2(\pi_h)},$$

since $\pi_h[f]$ is the $L^2(\pi_h)$ orthogonal projection of f onto the space of constant functions. Therefore, we have

$$\begin{aligned} \|f - \pi_h[f]\|_{L^2(\pi_h)}^2 &\leq \|f - \int_{U_h} f\|_{L^2(\pi_h)}^2 \\ &\leq \left(\sup_{x \in U_h} \pi_h(x) \right) \|f - \int_{U_h} f\|_{L^2(U_h)}^2 \\ &\leq h^2 \Lambda^2 \left(\sup_{U_h} \pi(x) \right) \|\nabla f\|_{L^2(U_h)}^2 \\ &\leq h^2 \Lambda^2 \frac{\sup_{x \in U_h} \pi_h(x)}{\inf_{x \in U_h} \pi_h(x)} \|\nabla f\|_{L^2(\pi_h)}^2, \end{aligned}$$

and the result follows. ■

Remark C.2. The Poincaré inequality for the Lebesgue measure on a set U holds under very weak conditions on U . For example, when U is convex, the Poincaré inequality holds with constant $\Lambda(U) = D/\pi$, where D is the diameter of the domain [38].

We now prove Theorem 3.7:

Proof of Theorem 3.7. We begin by stating a simple consequence of the functional central limit theorem for reversible, continuous time Markov processes: Let Y_t be a reversible, stationary Markov process with ergodic distribution π and generator L . Let $g \in L^2(\pi)$, and define

$$\bar{g} := T^{-1} \int_0^T g(Y_s) ds.$$

By [25, Corollary 1.9],

$$\sqrt{T}(\bar{g} - \pi[g]) \xrightarrow{d} N(0, \sigma^2(g)),$$

where

$$\sigma^2(g) = \langle g - \pi[g], L^{-1}(g - \pi[g]) \rangle_{\pi}. \quad (\text{C.1})$$

Here, $L^{-1}(g - \pi[g])$ denotes any function in the domain of L with

$$L(L^{-1}(g - \pi[g])) = g - \pi[g]$$

and $\pi[L^{-1}(g - \pi[g])] = 0$. Such a function must exist when $g \in L^2(\pi)$ and X_t is reversible [25].

We now show that the process X_t^h meets the conditions above for the central limit theorem.

First, we recall that the generator of X_t^h is the operator

$$L_h = \beta^{-1} \Delta - \nabla V \cdot \nabla$$

with domain

$$D(L_h) := \left\{ g \in C^2(U^h) : \nabla g(x) \cdot \mathbf{n}(x) = 0 \text{ for all } x \in \partial U^h \right\};$$

see [2, Proposition 3.2] for the case of a convex polyhedron or [13, Chapter 8] for a domain with C^3 boundary.

By [24, Theorem 4.3.3], a process Y_t with invariant distribution π is reversible if its generator is symmetric and it has the strong continuity property

$$\lim_{t \rightarrow 0^+} \|T_t f - f\|_{\pi} = 0 \text{ for all } f \in L^2(\pi), \quad (\text{C.2})$$

where $T_t f(x) := \mathbf{E}_x[f(Y_t)]$ denotes the backwards semigroup associated with Y_t . The generator L_h of X_t^h is symmetric, since for all $f, g \in D(L_h)$, using integration by parts, we have

$$\begin{aligned}
 -\beta^{-1} \left\langle \nabla f, \nabla g \right\rangle_\pi &= -\beta^{-1} \int_{U_h} \nabla f \cdot \nabla g z_h^{-1} \exp(-\beta V) dx \\
 &= \beta^{-1} \int_{U_h} f \operatorname{div}(z_h^{-1} \exp(-\beta V) \nabla g) dx \\
 &\quad - \beta^{-1} \int_{\partial U_h} f z_h^{-1} \exp(-\beta V) \nabla g \cdot n \, dS \\
 &= \int_{U_h} (\beta^{-1} \Delta g - \nabla V \cdot \nabla g) f z_h^{-1} \exp(-\beta V) dx \\
 &= \langle f, L_h g \rangle_\pi.
 \end{aligned} \tag{C.3}$$

(Here, $z_h^{-1} := \int_{U_h} \exp(-\beta V) dx$ is the normalizing constant for π_h .) Since $\langle \nabla f, \nabla g \rangle_\pi$ is invariant under exchanging f and g , $\langle f, L_h g \rangle_\pi = \langle L_h f, g \rangle_\pi$ and L_h is symmetric. We postpone discussion of the strong continuity of X_t^h to the end of the proof.

We now use the Poincaré inequality (Lemma C.1) and (C.3) to prove that X_t^h is ergodic and to estimate the term $L_h^{-1}(g - \pi_h[g])$ appearing in the formula for $\sigma_h^2(g)$; in essence, we adapt the approach outlined in [31, Section 3] to the family of reflected processes X_t^h . We prove ergodicity first. By [4, Proposition 2.2], a process is ergodic if and only if 0 is a simple eigenvalue of its generator. By the Poincaré inequality (Lemma C.1) and (C.3), for all $u \in D(L_h)$,

$$\begin{aligned}
 \|u - \pi_h[u]\|_{L^2(\pi_h)}^2 &\leq C_h^2 \|\nabla u\|_{L^2(\pi_h)}^2 = C_h^2 \beta \langle u, -Lu \rangle_{\pi_h} \leq C_h^2 \beta \|u\|_{L^2(\pi_h)} \\
 &\quad \|Lu\|_{L^2(\pi_h)},
 \end{aligned} \tag{C.4}$$

where

$$C_h = h\Lambda(U) \exp\left(\frac{\beta}{2} \left(\sup_{U_h} V - \inf_{U_h} V \right)\right).$$

Now if u is not constant, $\|u - \pi_h[u]\|_{L^2(\pi_h)}^2 > 0$, so $\|L_h u\|_{L^2(\pi_h)} > 0$ and u is not an eigenvector with eigenvalue 0. Hence, 0 is a simple eigenvalue of L_h , and X_t^h is ergodic.

Finally, we estimate $\sigma_h^2(g)$. We have

$$\|u\|_{L^2(\pi_h)} \leq C_h^2 \beta \|Lu\|_{L^2(\pi_h)}.$$

Taking $u = L_h^{-1}(g - \pi_h[g])$ in the above yields

$$\|L_h^{-1}(g - \pi_h[g])\|_{L^2(\pi_h)} \leq C_h^2 \beta \|g - \pi_h[g]\|_{L^2(\pi_h)},$$

which implies

$$\sigma_h^2(g) = \left\langle g - \pi_h[g], L^{-1}(g - \pi_h[g]) \right\rangle_{\pi_h} \leq C_h^2 \beta \text{var}_{\pi_h}(g),$$

using the Cauchy–Schwarz inequality.

It remains to show that the process X_t^h has the strong continuity property (C.2). We only sketch an argument, since the basic ideas are standard. First, one can use the Lipschitz continuity of strong solutions of the reflected process [2, Lemma 4.1] to show that X_t^h has the Feller property. (That is, one can show that $T_t u$ is continuous whenever u is continuous.) In addition, since the process X_t^h has an infinitesimal generator, we have the pointwise continuity property

$$\lim_{t \rightarrow 0^+} T_t u(x) = u(x) \tag{C.5}$$

for all $x \in U_h$ and all $u \in D(L_h)$. Now we have $\|T_t\|_\infty \leq 1$ for all $t \geq 0$, where $\|T_t\|_\infty$ is the operator norm of T_t on the space of continuous functions with the sup-norm, and therefore by a density argument the limit (C.5) holds for all continuous u . Hence, by [8, Lemma 1.4], we have

$$\lim_{t \rightarrow 0^+} \sup_{x \in U_h} |T_t u(x) - u(x)| = 0$$

for all continuous u . The strong continuity property (C.2) then follows by another density argument, using that $\|T_t\|_{L^2(\pi_h)} \leq 1$ for all $t \geq 0$. ■

Appendix D.: Proof of Theorem 4.1.

Proof of Theorem 4.1. By the remarks immediately following Theorem 3.5, since the bias functions are a partition of unity, we have

$$\sigma^2(g) \leq 2 \sum_{\mathbf{i} \in \mathbb{Z}^d/K\mathbb{Z}^d} \kappa_{\mathbf{i}}^{-1} \left\{ C(\bar{g}_{\mathbf{i}}) z_{\mathbf{i}}^2 + \text{var}_{\pi}(g) \text{tr}(R^{\mathbf{i}}) \sum_{\substack{\mathbf{j} \neq \mathbf{i} \\ F_{\mathbf{ij}} > 0}} \frac{1}{F_{\mathbf{ij}}} \right\}. \quad (\text{D.1})$$

To prove the desired upper bound, we substitute estimates of $C(\bar{g}_{\mathbf{i}})$, $R^{\mathbf{i}}$, and $F_{\mathbf{ij}}$ into the inequality above.

First, we consider the asymptotic covariances $R^{\mathbf{i}}$ and $C(\bar{g}_{\mathbf{i}})$. Let

$$h = 1/K.$$

The diameter of $U_{\mathbf{i}}$ is $2\sqrt{d}h$, so by Assumption 3.6

$$R_{\mathbf{jj}}^{\mathbf{i}} \leq Ch^a \beta^b \exp(2\sqrt{d}h\beta \|\nabla V\|_{L^\infty}) \leq Ch^{a-b} \exp(2\sqrt{d}\|\nabla V\|_{L^\infty}). \quad (\text{D.2})$$

(The second inequality follows since $h\beta \geq 1$ by definition.) Similarly,

$$C(\bar{g}_{\mathbf{i}}) \leq Ch^{a-b} \exp(2\sqrt{d}\|\nabla V\|_{L^\infty}) \text{var}_{\pi_{\mathbf{i}}}(g). \quad (\text{D.3})$$

Second, by Lemma 3.11, the nonzero entries of the overlap matrix F are bounded below as β tends to infinity:

$$F_{\mathbf{ij}} \geq \frac{\exp(-2\sqrt{d}\|\nabla V\|_{L^\infty})}{4^d} \quad (\text{D.4})$$

for all \mathbf{i}, \mathbf{j} so that $F_{\mathbf{ij}} > 0$. We also observe that each row of F has 3^d nonzero entries, since $F_{\mathbf{i}, \mathbf{i}+\mathbf{k}} > 0$ only when all entries of \mathbf{k} belong to $\{-1, 0, 1\}$.

We now estimate the term involving $C(\bar{g}_{\mathbf{i}})$ in (D.1). By (D.3), we have

$$\sum_{\mathbf{i} \in \mathbb{Z}^d/K\mathbb{Z}^d} z_{\mathbf{i}}^2 C(\bar{g}_{\mathbf{i}}) \leq Ch^{a-b} \exp(2\sqrt{d}\|\nabla V\|_{L^\infty}) \sum_{\mathbf{i} \in \mathbb{Z}^d/K\mathbb{Z}^d} z_{\mathbf{i}}^2 \text{var}_{\pi_{\mathbf{i}}}(g). \quad (\text{D.5})$$

Now we have

$$\text{var}_{\pi_{\mathbf{i}}}(g) = \pi_{\mathbf{i}}[|g - \pi_{\mathbf{i}}[g]|^2] \leq \pi_{\mathbf{i}}[|g - \pi[g]|^2].$$

Therefore,

$$\sum_{\mathbf{i} \in \mathbb{Z}^d/K\mathbb{Z}^d} z_{\mathbf{i}}^2 \text{var}_{\pi}(g) \leq \sum_{\mathbf{i} \in \mathbb{Z}^d/K\mathbb{Z}^d} z_{\mathbf{i}} \pi_{\mathbf{i}} [lg - \pi[g]]^2 = \pi[lg - \pi[g]]^2 = \text{var}_{\pi}(g). \quad (\text{D.6})$$

(The inequality follows since $0 \leq z_{\mathbf{i}} \leq 1$ for all \mathbf{i} ; the second to last equality follows using (2.5) and that $\{\psi_{\mathbf{i}}\}_{\mathbf{i} \in \mathbb{Z}^d/K\mathbb{Z}^d}$ is a partition of unity.) Thus,

$$\sum_{\mathbf{i} \in \mathbb{Z}^d/K\mathbb{Z}^d} z_{\mathbf{i}}^2 C(\bar{g}_{\mathbf{i}}) \leq Ch^{a-b} \exp\left(2\sqrt{d} \|\nabla V\|_{L^\infty}\right) \text{var}_{\pi}(g). \quad (\text{D.7})$$

It remains to address the term involving $R^{\mathbf{i}}$ in (D.1): Using (D.2), (D.4), and that each row of F has 3^d nonzero entries, we have

$$\text{tr}(R^{\mathbf{i}}) \sum_{\substack{j \neq \mathbf{i} \\ F_{ij} > 0}} \frac{1}{F_{ij}} \leq C 6^{2d} h^{a-b} \exp\left(4\sqrt{d} \|\nabla V\|_{L^\infty}\right) \quad (\text{D.8})$$

for every $\mathbf{i} \in \mathbb{Z}^d/K\mathbb{Z}^d$. Finally, using (D.7), (D.8), and $\kappa_{\mathbf{i}}^{-1} = K^d = \beta^d$, we conclude

$$\begin{aligned} \sigma^2(g) &\leq 2Ch^{a-b} \left\{ K^d \exp\left(2\sqrt{d} \|\nabla V\|_{L^\infty}\right) + 6^{2d} K^{2d} \exp\left(4\sqrt{d} \|\nabla V\|_{L^\infty}\right) \right\} \text{var}_{\pi}(g) \\ &\leq (D[\beta]^d + b - a + E[\beta]^{2d + b - a}) \text{var}_{\pi}(g), \end{aligned}$$

where the constants D and E depend on d and V , but not on g or β . ■

We note that if one uses the bias functions proposed in Remark 3.13, then the constants D and E in the proof of Theorem 4.1 grow only polynomially with the dimension d , not exponentially. However, we do not claim that those bias functions perform better than the uniform grid (3.7) or the bias functions of Section 5.3 in practice.

Appendix E.: Proof of Theorem 4.6.

Proof of Theorem 4.6. Take $g := \mathbf{1}_x \cdot M$. As explained in the remarks after the statement of Theorem 3.5, since the bias functions are a partition of unity, we have

$$\sigma_M^2 \leq 2 \sum_{i=0}^{K+1} \kappa_i^{-1} \left\{ C(\bar{g}_i) z_i^2 + p_M^2 \text{tr}(R^i) \sum_{\substack{j \neq i \\ F_{ij} > 0}} \frac{1}{F_{ij}} \right\}. \quad (\text{E.1})$$

First, we estimate

$$R_{jj}^i \leq Ch^a \exp\left(h \max_{x \leq M} |V'(x)|\right) \leq Ceh^a$$

for all $i = 1, \dots, K-1$ by Assumption 3.6. By Assumption 4.5,

$$R_{jj}^K \leq D \text{ for } j = K-1, K,$$

and $R_{jj}^K = 0$ for $j = K-1, K$, since ψ_{K+1} is constant over the support of π_K . In addition,

$$R_{jj}^{K+1} = 0 \text{ for all } j = 1, \dots, L,$$

since all bias functions ψ_i take a constant value over the support of π_{K+1} . Likewise,

$$C(\bar{g}_K) \leq Ceh^a \text{var}_{\pi_K}(g) \leq Ceh^a,$$

and $C(\bar{g}_i) = 0$ for all $i = K$.

We now show that the nonzero entries of the overlap matrix are bounded below independent of M . First, we estimate the entries which are averages over the biased distributions with bounded support. By Lemma 3.11, we have

$$F_{ij} \geq \frac{1}{2\exp(2)} > 0$$

for all $i = 0, \dots, K-1$ and j so that $F_{ij} > 0$. whenever $F_{ij} > 0$. It remains to address those entries related to biased distributions with unbounded support, so with $i = K, K+1$. By Lemma E.1, $F_{K,K+1}$ and $F_{K,K-1}$ are bounded below by some $\theta > 0$ independent of M , for this choice of bias functions. (Lemma E.1 and its proof appear in Appendix E. Lemma E.1 is the only part of the proof which relies on Assumption 4.3.) In addition, for any $i = 0, \dots, K+1$, we have $F_{ii} = \frac{1}{2}$, which implies $F_{K+1,K} = 1 - F_{K+1,K+1} = \frac{1}{2}$ since F is stochastic when the bias functions are a partition of unity.

Finally, we substitute the above estimates of the overlap matrix and the variances into (E.1). Let $c = \min\{\theta, 1/2\exp(2)\}$. Observe that h decreases with M , so $Ceh^a = E$ for some constant E , uniformly in M . Let $F = \max\{D, E\}$. We have

$$\frac{\sigma^2}{p_M^2} \leq \frac{2}{p_M^2} \sum_{i=0}^{K+1} (K+2) \left\{ C(\bar{g}_i) z_i^2 + p_M^2 (2F) \frac{2}{c^2} \right\} \leq 2(K+2)F \frac{z_K^2}{p_M^2} + \frac{4(K+2)^2}{c^2}.$$

We now observe that

$$\frac{z_K}{p_M} = \frac{z_K}{z_{K+1}} = \frac{F_{K+1,K}}{F_{K,K+1}} \leq \frac{1}{c}.$$

Therefore,

$$\frac{\sigma^2}{p_M^2} \leq \frac{2F(K+2) + 4F(K+2)^2}{c^2},$$

which proves the result. ■

We now prove Lemma E.1, which is used in the proof of Theorem 4.6.

Lemma E.1. *Under the hypotheses of Theorem 4.6, there exist constants M_1 , θ_+ , $\theta_- > 0$ depending on V but not on M so that*

$$F_{K,K+1} \geq \theta_+ > 0 \text{ and } F_{K,K-1} \geq \theta_- > 0$$

whenever $M \geq M_1$.

Proof. We consider $F_{K,K-1}$ first. We have

$$F_{K,K-1} = \frac{1}{2} \frac{\pi([M-h, M])}{\pi([M-h, \infty))} = \frac{1}{2} \frac{\int_{M-h}^M \exp(-V(x)) dx}{\int_{M-h}^{\infty} \exp(-V(x)) dx}.$$

By the integral mean value theorem,

$$\int_{M-h}^M \exp(-V(x)) dx = h \exp(-V(\xi_{M-h, M}))$$

for some $\xi_{M-h, M} \in [M-h, M]$. Moreover, by (4.5), we have

$$V(x) \leq V(M) + V'(M)(x - M) \text{ for all } x \geq M \geq M_0.$$

Therefore, when $M-h \geq M_0$,

$$\begin{aligned} \int_{M-h}^{\infty} \exp(-V(x)) dx &\leq \int_{M-h}^{\infty} \exp(-V(M-h) - V'(M-h)(x - M + h)) dx \\ &= \frac{\exp(-V(M-h))}{V'(M-h)}. \end{aligned}$$

It follows that

$$\begin{aligned}
F_{K,K-1} &\geq hV'(M-h)\exp(V(M-h) - V(\xi_{M-h,M})) \\
&\geq hV'(M-h)\exp\left(-h \max_{x \leq M} |V'(x)|\right) \\
&\geq hV'(M-h)\exp(-1) \\
&= \frac{V'(M-h)}{\lceil \max_{x \leq M} |V'(x)| \rceil} \exp(-1),
\end{aligned} \tag{E.2}$$

using the definition $h = M/K$.

To estimate the quotient in expression (E.2), we distinguish two cases: By (4.5), V' is nondecreasing on $[M_0, \infty)$, so either $\lim_{x \rightarrow \infty} V'(x) = C_2 < \infty$ or $\lim_{x \rightarrow \infty} V'(x) = \infty$. In the first case, V' is bounded, and we have

$$\frac{V'(M-h)}{\lceil \max_{x \leq M} |V'(x)| \rceil} \geq \frac{V'(M_0)}{\lceil \max_{x \in [0, \infty)} |V'(x)| \rceil} > 0, \tag{E.3}$$

whenever $M-h \geq M_0$. In the second case, for M sufficiently large,

$$\max_{x \leq M} |V'(x)| = V'(M).$$

Therefore, applying in succession the mean value theorem, the monotonicity of V' , assumption (4.6), and the hypothesis $\lim_{x \rightarrow \infty} V'(x) = \infty$, we have that for all M sufficiently large,

$$\begin{aligned}
\frac{V'(M-h)}{\lceil \max_{x \leq M} |V'(x)| \rceil} &= \frac{V'(M)}{\lceil V'(M) \rceil} - \frac{V'(M) - V'(M-h)}{\lceil V'(M) \rceil} \\
&\geq \frac{V'(M)}{\lceil V'(M) \rceil} - \frac{hV''(\eta_{M-h,M})}{\lceil V'(M) \rceil} \\
&\geq \frac{V'(M)}{\lceil V'(M) \rceil} - \frac{V''(\eta_{M-h,M})}{V'(\eta_{M-h,M})^2} \\
&\geq \frac{V'(M)}{\lceil V'(M) \rceil} - \alpha \\
&\geq \frac{1-\alpha}{2} \\
&> 0.
\end{aligned} \tag{E.4}$$

(In the second and third lines above, $\eta_{M-h,M} \in [M-h, M]$ denotes the point guaranteed by the mean value theorem so that $V'(M) - V'(M-h) = hV''(\eta_{M-h,M})$.) It follows from (E.2), (E.3), and (E.4) that there exist $M_-, \theta_- > 0$ so that

$$F_{K,K-1} \geq \theta_- > 0 \quad (\text{E.5})$$

whenever $M \geq M_-$.

Now we prove that $F_{K,K+1}$ is bounded below. We have

$$\begin{aligned} F_{K,K+1} &= \frac{1}{2} \frac{\int_M^\infty \exp(-V(x)) dx}{\int_{M-h}^\infty \exp(-V(x)) dx} \\ &= F_{K,K-1} \frac{\int_M^\infty \exp(-V(x)) dx}{\int_{M-h}^M \exp(-V(x)) dx} \\ &\geq \theta_- \frac{\int_M^{M+h} \exp(-V(x)) dx}{\int_{M-h}^M \exp(-V(x)) dx} \\ &\geq \theta_- \frac{\int_M^{M+h} \exp(V(x-h) - V(x)) \exp(-V(x-h)) dx}{\int_{M-h}^M \exp(-V(x)) dx} \\ &\geq \theta_- \exp\left(\min_{[M-h, M+h]} V - \max_{[M-h, M+h]} V\right) \\ &\geq \theta_- \exp\left(-2h \max_{[M-h, M+h]} |V'|\right). \end{aligned} \quad (\text{E.6})$$

As above, to bound the quantity appearing in the exponent in (E.6), we distinguish the two cases $\lim_{x \rightarrow \infty} V'(x) = C_1 < \infty$ and $\lim_{x \rightarrow \infty} V'(x) = \infty$. In the first case, for M sufficiently large that $2C_1 \leq |V'(x)| \leq C_1/2$ whenever $x \in [M-h, M+h]$, we have

$$h \max_{[M-h, M+h]} |V'| = \frac{\max_{[M-h, M+h]} |V'|}{\lceil \max_{[0, M]} |V'| \rceil} \leq \frac{2C_1}{C_1/2} = 4. \quad (\text{E.7})$$

In the second case, for M sufficiently large,

$$h \max_{[M-h, M+h]} |V'| = \frac{\max_{[M-h, M+h]} |V'|}{\lceil \max_{[0, M]} |V'| \rceil} \leq \frac{V'(M+h)}{V'(M)}. \quad (\text{E.8})$$

By (4.6), we have the differential inequality

$$V'' < \alpha |V'|^2.$$

This implies

$$V'(M+s) \leq y'(s)$$

for

$$y(s) = \frac{1}{V'(M)^{-1} - \alpha s}$$

the solution of the initial value problem

$$y' = \alpha y^2 \text{ and } y(0) = V'(M).$$

Therefore,

$$V'(M+h) \leq \frac{1}{V'(M)^{-1} - \alpha h} = \frac{1}{V'(M)^{-1} - \alpha [V'(M)]^{-1}} \leq \frac{V'(M)}{1 - \alpha},$$

so by (E.8),

$$h \max_{[M-h, M+h]} |V'| \leq \frac{1}{1 - \alpha}. \quad (\text{E.9})$$

It follows from (E.6), (E.7), and (E.9) that there exist M_+ , $\theta_+ > 0$ so that

$$F_{K, K+1} \geq \theta_+ > 0 \quad (\text{E.10})$$

whenever $M \geq M_+$. ■

Appendix F.: An improved method of computing error bars for EMUS.

In [46, Section VII.B.1], we proposed a practical method of estimating the asymptotic standard deviations (error bars) of averages computed by EMUS. Using the notation established in Appendix A, our method proceeds as follows:

1. Compute \bar{F} , $\{\bar{g}_i^*\}_{i=1}^L$ and $\{\bar{\mathbf{I}}_i^*\}_{i=1}^L$.
2. Compute $w(\bar{F})$ and the group inverse $(I - \bar{F})^\#$.
3. Evaluate $\partial_{\bar{F}} B$, and \bar{F} , $\{\bar{g}_i^*\}_{i=1}^L$, and $\{\bar{\mathbf{I}}_i^*\}_{i=1}^L$.
4. Compute the time series

$$\bar{\zeta}_t^i = \partial_{\bar{F}} B \cdot \left(\left(\psi_1(X_t^i), \dots, \psi_L(X_t^i), g^*(X_t^i), 1^*(X_t^i) \right) - (\bar{F}_{i1}, \dots, \bar{F}_{iL}, \bar{g}_i^*, \bar{\mathbf{I}}_i^*) \right).$$

5. Compute an estimate $\bar{\chi}_i^2$ of the integrated autocovariance of $\bar{\zeta}_t^i$ using an algorithm such as ACOR [14].
6. Compute as an estimate σ^2 the quantity

$$\bar{\sigma}^2 = \sum_{i=1}^L \frac{\bar{\chi}_i^2}{\kappa_i}. \quad (\text{F.1})$$

We originally proposed computing the group inverse $(I - \bar{F})^\#$ using the method of [19] based on the QR factorization. We have since discovered that this method does not always yield sufficiently accurate results. For example, when computing error bars for the marginal in μ_2 in Section 5.3, we observed a highly oscillatory numerical error affecting some entries of $(I - \bar{F})^\#$. That the sign pattern in Figure 11a fails to be symmetric is evidence of this numerical error. We note that since the exact overlap matrix F is in detailed balance with $w(F)$, we have $\text{diag}(w(F))F\text{diag}(w(F))^{-1} = F$. (Here, $\text{diag}(w(F))$ denotes the diagonal matrix with $w(F)$ along the diagonal.) Therefore,

$$\left((I - F)^\#\right)^t = \text{diag}(w(F))(I - F)^\# \text{diag}(w(F))^{-1},$$

which implies that the sign pattern of $(I - F)^\#$ is symmetric since $w(F)$ is positive. As a result of these numerical errors, we were unable to accurately compute error bars for the EMUS estimate of the marginal density.

We therefore propose computing the group inverse by a new method combining QR factorization with power iteration. We first compute an estimate G^0 of $(I - \bar{F})^\#$ by the method of [19]. We then iterate

$$G^{n+1} = \mathcal{J}(G^n) = \tilde{F}G^n + I - ew(\bar{F})^t, \quad (\text{F.2})$$

where $e \in \mathbb{R}^L$ denotes the column vector of all ones and $\tilde{F} = (I - ew(F)^t)\bar{F}$. We observe that $(I - \bar{F})^\#$ is a fixed point of this iteration, since

$$\begin{aligned} \mathcal{J}\left((I - \bar{F})^\#\right) &= (I - ew(\bar{F})^t)\bar{F}(I - \bar{F})^\# + (I - ew(\bar{F})^t) \\ &= (\bar{F} - I)(I - \bar{F})^\# + (I - ew(\bar{F})^t) + (I - \bar{F})^\# \\ &= (I - \bar{F})^\#. \end{aligned}$$

Above, we use well known properties of the group inverse, including that the spectral projector $I - ew(\bar{F})^t$ commutes with \bar{F} , that $(I - ew(\bar{F})^t)(I - \bar{F})^\# = (I - F)^\#$, and that $(I - \bar{F})(I - \bar{F})^\# = I - ew(\bar{F})^t$.

Moreover, when \bar{F} is irreducible, \mathcal{J}^K is a contraction for K sufficiently large. By the Perron-Frobenius theorem, the spectral radius of \tilde{F} is smaller than $1 - \varepsilon$, for some $\varepsilon > 0$. Therefore, by Gelfand's formula, for any matrix norm $\|\cdot\|$, we have $\lim_{k \rightarrow \infty} \|\tilde{F}\|^{k1/k} < 1 - \varepsilon/2$, and so for some K ,

$$\|\tilde{F}^k\| < (1 - \varepsilon/2)^k \text{ whenever } k \geq K.$$

Now

$$\mathcal{J}^K(G) = \tilde{F}^K G + \left(I - e^{\pi(F)}\right) \sum_{j=0}^{K-1} F^j.$$

Thus, assuming that the norm $\|\cdot\|$ is submultiplicative,

$$\|\mathcal{J}^K(G) - \mathcal{J}^K(H)\| = \|\tilde{F}^K(G - H)\| \leq \|\tilde{F}^K\| \|G - H\| \leq (1 - \varepsilon/2)^K \|G - H\|.$$

Therefore, the power iteration converges and its limit is the group inverse $(I - \bar{F})^\#$.

Using this new method, we computed $(I - \bar{F})^\#$ for \bar{F} the overlap matrix involved in estimating the marginal in μ_2 in Section 5.3. We performed 10^6 power method iterates. Observe that the sign pattern of the group inverse computed with power iteration is symmetric; see Figure 11b.

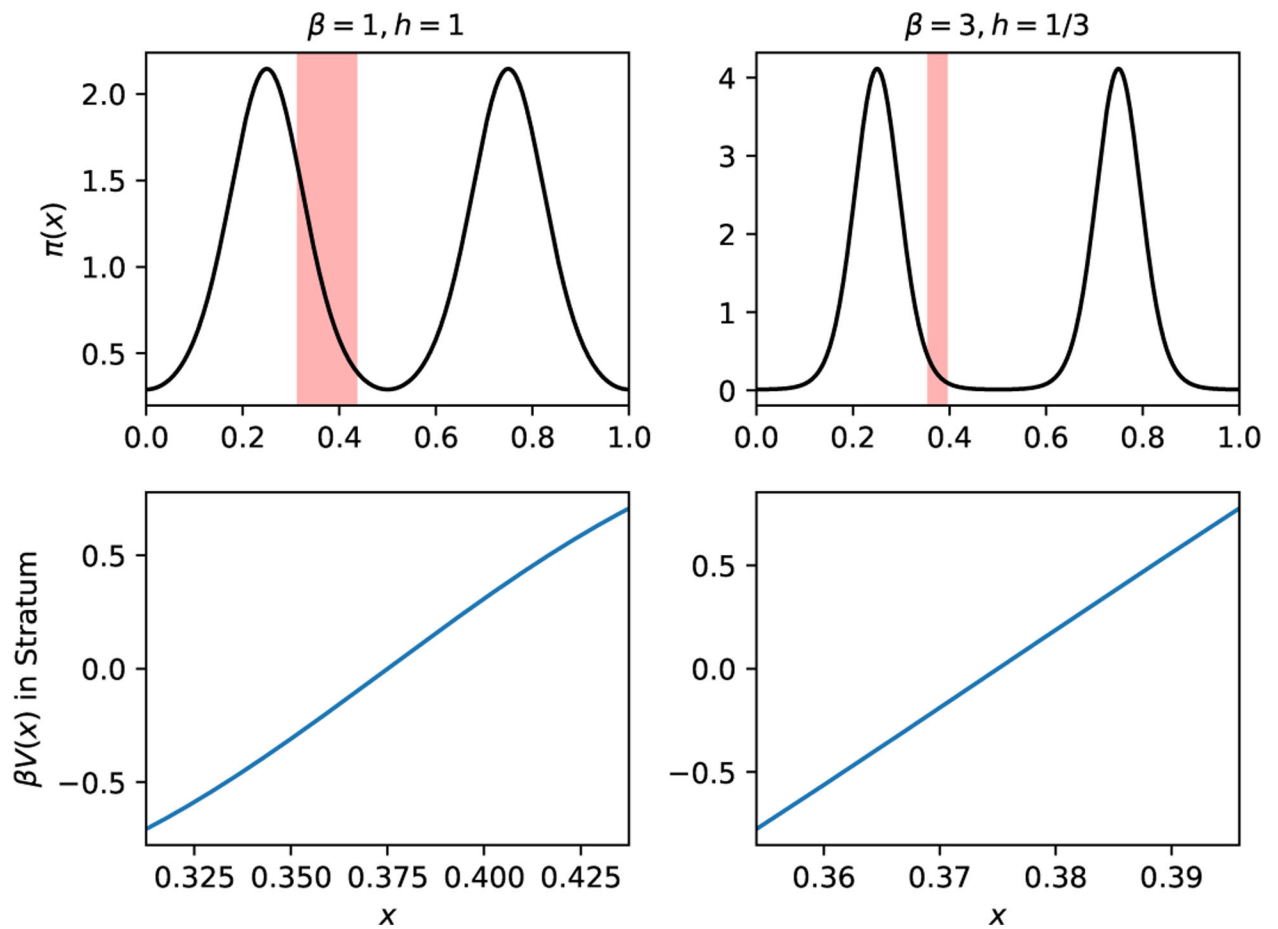
The power iteration (F.2) converges slowly when the spectral gap of \bar{F} is small. We have shown in [45] that the spectral gap may be very small: It decreases exponentially with a temperature parameter in a limit similar to the one analyzed in Section 4.1 above. However, even when the spectral gap is small, we conjecture that a modest number of power iterations will significantly reduce the numerical error in the group inverse, since the error in the initial calculation seems to be highly oscillatory and the power iteration has a smoothing effect.

REFERENCES

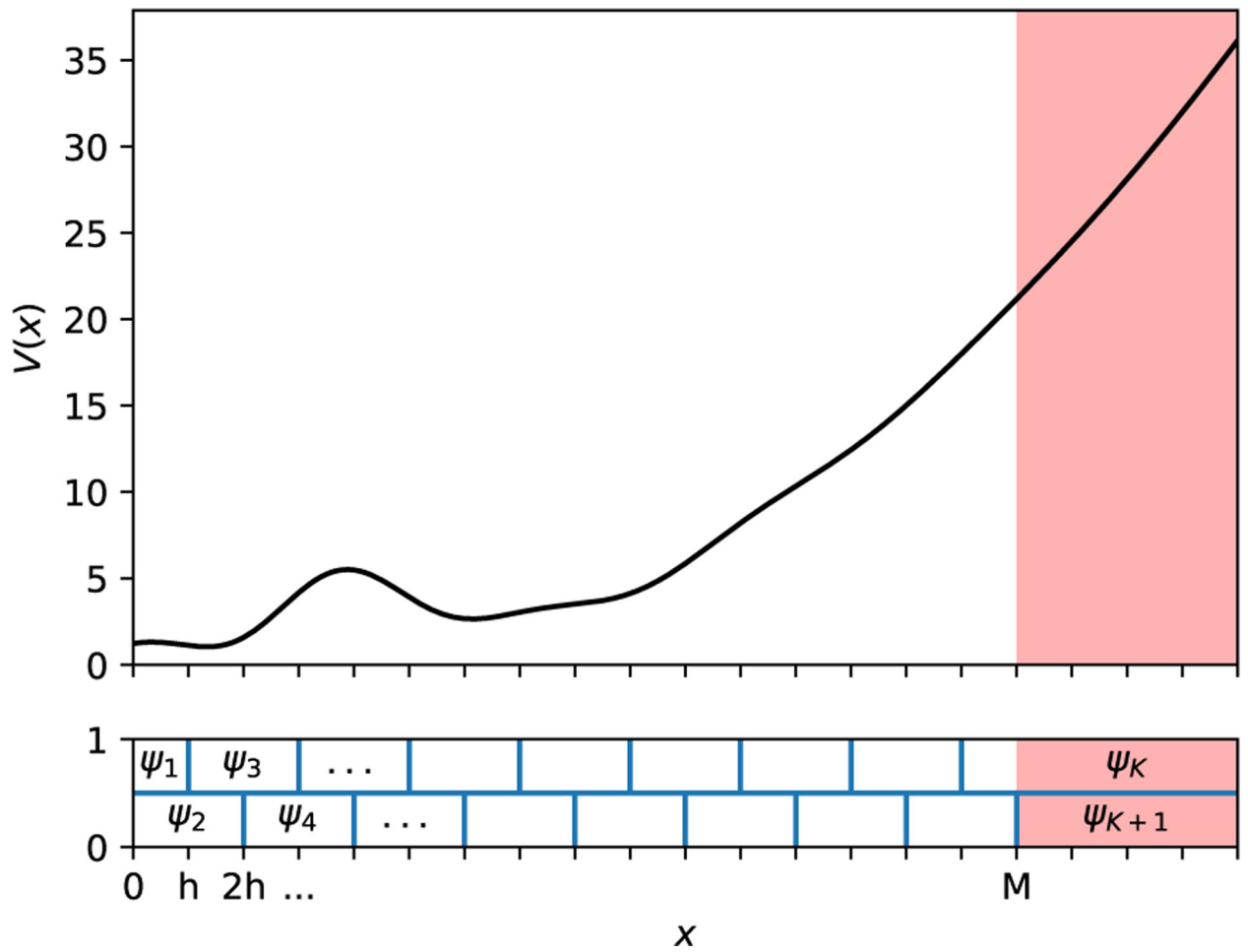
- [1]. Aitkin M: Likelihood and Bayesian analysis of mixtures. *Statistical Modelling* 1(4), 287–304 (2001)
- [2]. Andres S: Pathwise differentiability for SDEs in a convex polyhedron with oblique reflection. *Ann. Inst. Henri Poincaré Probab. Stat* 45(1), 104–116 (2009)
- [3]. Berneche S, Roux B: Energetics of ion conduction through the K⁺ channel. *Nature* 414(6859), 73 (2001) [PubMed: 11689945]
- [4]. Bhattacharya RN: On the functional central limit theorem and the law of the iterated logarithm for Markov processes. *Z. Wahrsch. Verw. Gebiete* 60(2), 185–201 (1982)
- [5]. Billingsley P: *Convergence of Probability Measures*, second edn. Wiley series in probability and statistics. Wiley-Interscience, New York (1999)
- [6]. Bilodeau M, Brenner D: *Theory of multivariate statistics*. Springer texts in statistics. Springer, New York (1999)
- [7]. Boczek EM, Brooks CL: First-principles calculation of the folding free energy of a three-helix bundle protein. *Science* 269(5222), 393–396 (1995) [PubMed: 7618103]
- [8]. Böttcher B, Schilling RL, Wang J: A primer on Feller processes. In: *Lévy Matters III : Lévy-type processes: construction, approximation and sample path properties*, Lecture notes in mathematics (Springer-Verlag), chap. 1, pp. 1–30. Springer (2013)
- [9]. Chandler D: *Introduction to modern statistical mechanics*. Oxford University Press, New York (1987)

- [10]. Cho GE, Meyer CD: Comparison of perturbation bounds for the stationary distribution of a Markov chain. *Linear Algebra Appl.* 335, 137–150 (2001)
- [11]. Chopin N, Lelièvre T, Stoltz G: Free energy methods for Bayesian inference: efficient exploration of univariate Gaussian mixture posteriors. *Statistics and Computing* 22(4), 897–916 (2012)
- [12]. Doss H, Tan A: Estimates and standard errors for ratios of normalizing constants from multiple Markov chains via regeneration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76(4), 683–712 (2014) [PubMed: 28706463]
- [13]. Ethier SN, Kurtz TG: Markov processes. *Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics.* John Wiley & Sons, Inc., New York (1986). Characterization and convergence
- [14]. Foreman-Mackey D, Goodman J: ACOR 1.1.1 <https://pypi.python.org/pypi/acor/1.1.1> (2014)
- [15]. Foreman-Mackey D, Hogg DW, Lang D, Goodman J: emcee: The MCMC hammer. *Publications of the Astronomical Society of the Pacific* 125(925), 306 (2013)
- [16]. Geyer CJ: Markov chain Monte Carlo maximum likelihood. In: *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface.* American Statistical Association (1991)
- [17]. Geyer CJ: Estimating normalizing constants and reweighting mixtures (1994). Technical Report No. 568. Retrieved from the University of Minnesota Digital Conservancy
- [18]. Gill RD, Vardi Y, Wellner JA: Large sample theory of empirical distributions in biased sampling models. *Ann. Statist* 16(3), 1069–1112 (1988)
- [19]. Golub GH, Meyer CD Jr.: Using the *QR* factorization and group inversion to compute, differentiate, and estimate the sensitivity of stationary probabilities for Markov chains. *SIAM J. Algebraic Discrete Methods* 7(2), 273–281 (1986)
- [20]. Goodman J, Weare J: Ensemble samplers with affine invariance. *Commun. Appl. Math. Comput. Sci* 5(1), 65–80 (2010)
- [21]. Helffer B, Klein M, Nier F: Quantitative analysis of metastability in reversible diffusion processes via a Witten complex approach. *Mat. Contemp* 26, 41–85 (2004)
- [22]. Izenman AJ, Sommer CJ: Philatelic mixtures and multimodal densities. *Journal of the American Statistical Association* 83(404), 941–953 (1988)
- [23]. Jasra A, Holmes CC, Stephens DA: Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statist. Sci* 20(1), 50–67 (2005)
- [24]. Jiang DQ, Qian M, Qian MP: Mathematical theory of nonequilibrium steady states : on the frontier of probability and dynamical systems. *Lecture notes in mathematics* (Springer-Verlag). Springer, Berlin; New York (2004)
- [25]. Kipnis C, Varadhan SRS: Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions. *Comm. Math. Phys* 104(1), 1–19 (1986)
- [26]. Kong A, McCullagh P, Meng XL, Nicolae D, Tan Z: A theory of statistical models for Monte Carlo integration. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 65(3), 585–604 (2003)
- [27]. Kumar S, Rosenberg JM, Bouzida D, Swendsen RH, Kollman PA: The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem* 13(8), 1011–1021 (1992)
- [28]. Laio A, Parrinello M: Escaping free-energy minima. *Proceedings of the National Academy of Sciences* 99(20), 12562–12566 (2002)
- [29]. Legoll F, Lelièvre T: Effective dynamics using conditional expectations. *Nonlinearity* 23(9), 2131–2163 (2010)
- [30]. Lelièvre T, Rousset M, Stoltz G: Free energy computations. Imperial College Press, London (2010). A mathematical perspective
- [31]. Lelièvre T, Stoltz G: Partial differential equations and stochastic methods in molecular dynamics. *Acta Numerica* 25, 681 (2016)
- [32]. Liu JS: Monte Carlo strategies in scientific computing. *Springer Series in Statistics.* Springer-Verlag, New York (2001)

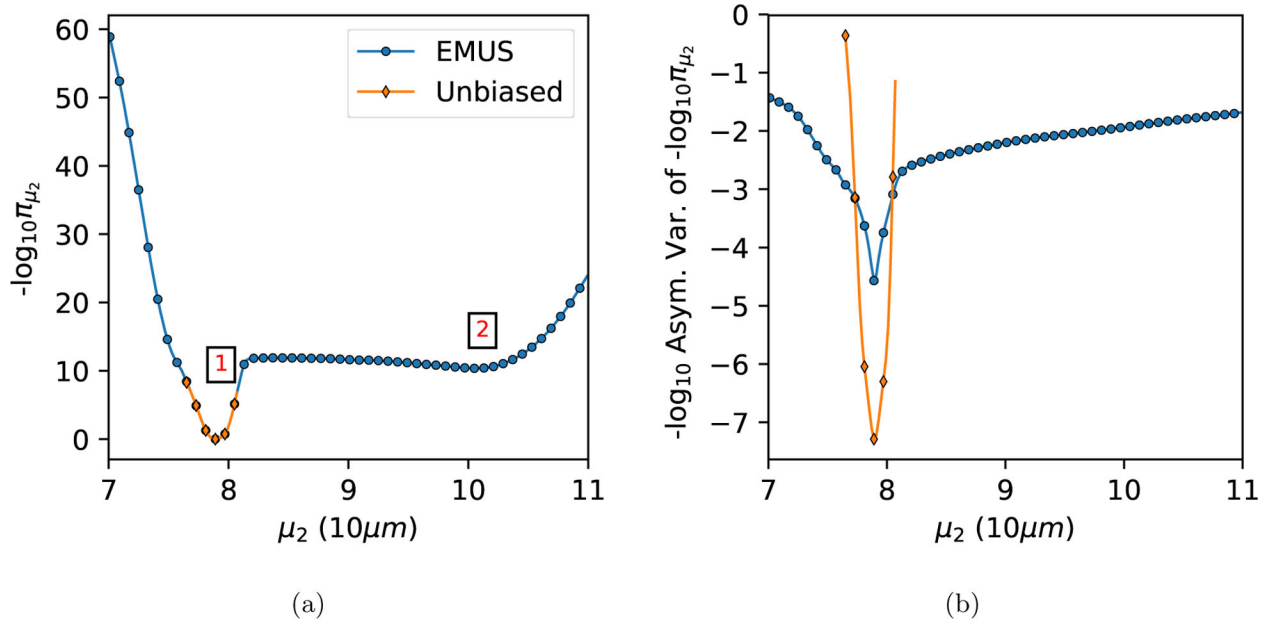
- [33]. Maragliano L, Vanden-Eijnden E: A temperature accelerated method for sampling free energy and determining reaction pathways in rare events simulations. *Chemical Physics Letters* 426(1–3), 168–175 (2006)
- [34]. Matthews C, Weare J, Kravtsov A, Jennings E: Umbrella sampling: a powerful method to sample tails of distributions (2017). *ArXiv:1712.05024*
- [35]. Meng XL, Wong WH: Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica* 6(4), 831–860 (1996)
- [36]. Pavliotis GA: *Stochastic processes and applications*, Texts in Applied Mathematics, vol. 60. Springer, New York (2014). *Diffusion processes, the Fokker-Planck and Langevin equations*
- [37]. Pavliotis GA, Stuart AM: *Multiscale methods*, Texts in Applied Mathematics, vol. 53. Springer, New York (2008). *Averaging and homogenization*
- [38]. Payne LE, Weinberger HF: An optimal Poincaré inequality for convex domains. *Archive for Rational Mechanics and Analysis* 5(1), 286–292 (1960)
- [39]. Richardson S, Green PJ: On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 59(4), 731–792 (1997)
- [40]. Roberts GO, Rosenthal JS: Geometric ergodicity and hybrid Markov chains. *Electron. Comm. Probab* 2, no. 2, 13–25 (1997)
- [41]. Roberts GO, Tweedie RL: Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli* 2(4), 341–363 (1996)
- [42]. Shirts MR, Chodera JD: Statistically optimal analysis of samples from multiple equilibrium states. *The Journal of chemical physics* 129(12), 124105 (2008) [PubMed: 19045004]
- [43]. Sugita Y, Kitao A, Okamoto Y: Multidimensional replica-exchange method for free-energy calculations. *J. Chem. Phys* 113(15), 11 (2000)
- [44]. Swendsen RH, Wang JS: Replica monte carlo simulation of spin-glasses. *Physical review letters* 57(21), 2607 (1986) [PubMed: 10033814]
- [45]. Thiede E, Van Kote B, Weare J: Sharp entrywise perturbation bounds for Markov chains. *SIAM Journal on Matrix Analysis and Applications* 36(3), 917–941 (2015) [PubMed: 26491218]
- [46]. Thiede EH, Van Kote B, Weare J, Dinner AR: Eigenvector method for umbrella sampling enables error analysis. *The Journal of Chemical Physics* 145(8), 084115 (2016) [PubMed: 27586912]
- [47]. Torrie GM, Valleau JP: Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics* 23(2), 187 (1977)
- [48]. VanDerwerken DN, Schmidler SC: Parallel Markov chain Monte Carlo (2013). *ArXiv:1312.7479*
- [49]. Vardi Y: Empirical distributions in selection bias models. *The Annals of Statistics* 13(1), 178–203 (1985)
- [50]. Wang F, Landau DP: Efficient, multiple-range random walk algorithm to calculate the density of states. *Phys. Rev. Lett* 86, 2050–2053 (2001) [PubMed: 11289852]
- [51]. Wang FY, Yan L: Gradient estimate on convex domains and applications. *Proc. Amer. Math. Soc* 141(3), 1067–1081 (2013)

**Figure 1:**

Bias functions and target distributions in the low temperature limit. In the upper two plots, the black curves are the densities of the target distributions for two different values of β . Observe that π concentrates at the minima of V as β increases. The red bands each lie above a single stratum chosen from a family of strata for which $h \propto \beta^{-1}$. In the lower two plots, the blue curve is $\beta V(x)$ and the x -axis covers the bottom of the red band in the plot immediately above. Observe that the range of $\beta V(x)$ over the red band is the same for each of the two values of β . By Theorem 3.7 and the ensuing discussion in Section 3.2, this implies that the cost of sampling a single biased distribution increases at most polynomially with β when $h \propto \beta^{-1}$.

**Figure 2:**

The bias functions $\{\psi_i: i = 0, \dots, K + 1\}$ defined in (4.7) and a potential function V satisfying Assumption 4.3. Observe that the bias functions ψ_K and ψ_{K+1} have unbounded support.

**Figure 3:**

Estimates of the logarithm of the marginal density in μ_2 and the asymptotic variances of those estimates. Figure 3a displays estimates of the marginal in μ_2 computed by EMUS and by an unbiased trajectory of the ensemble sampler. Figure 3b displays the asymptotic variances of these two estimates of the marginal density. We note that while the unbiased calculation has greater accuracy near the mode, the EMUS calculation has greater accuracy in the tails. The relative errors in this figure were estimated using the method described in Appendix F.

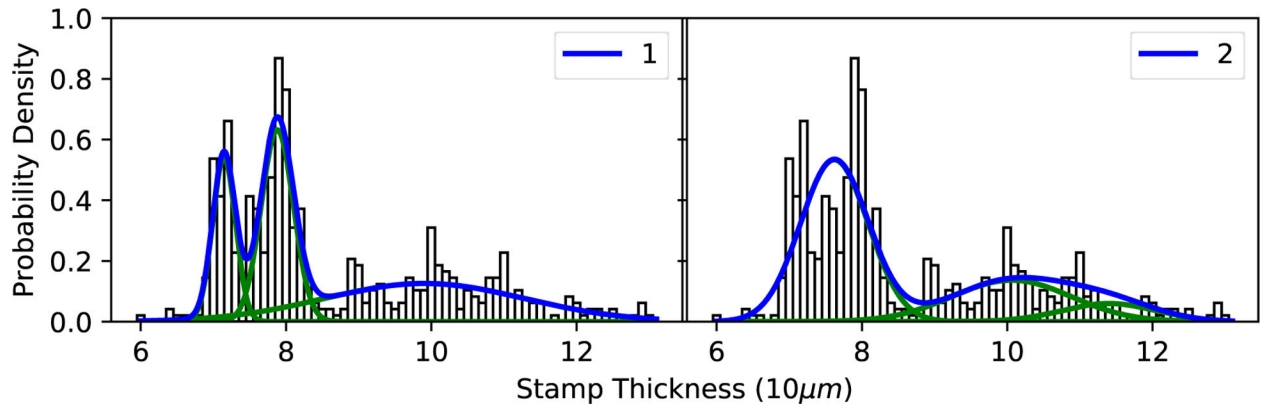
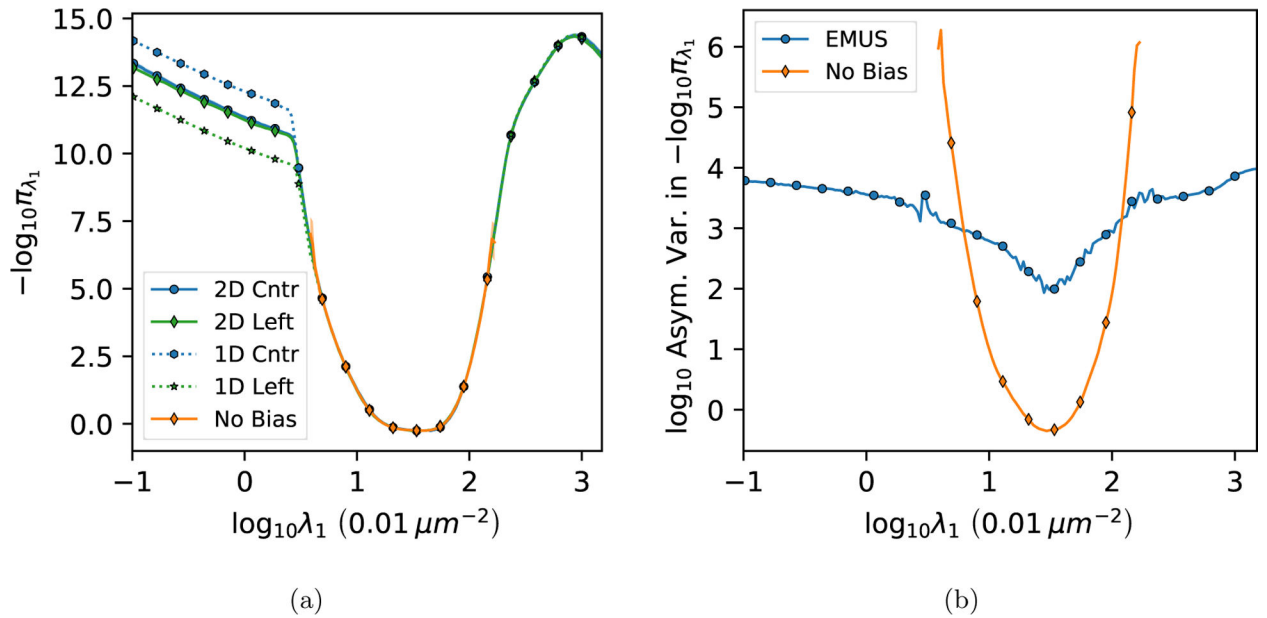


Figure 4:

Gaussian mixtures corresponding to modes of the marginal in μ_2 . Mixtures 1 and 2 correspond to the labeled points in Figure 3a. To be precise, the blue curve in each plot is the mixture distribution corresponding to the mean of a histogram bin centered at the point labeled in Figure 3a. The green curves are the individual mixture components. The black bars are a histogram of the Hidalgo stamp data.

**Figure 5:**

Estimates of the logarithm of the marginal density in $\log_{10} \lambda_1$ and the asymptotic variances of those estimates. Figure 5a displays the estimates of the marginal in $\log_{10} \lambda_1$ computed by various methods. The error bars are twice the estimated asymptotic standard deviation in each histogram bin. For the two-dimensional EMUS calculations, standard deviations were estimated using the method described in Appendix F. For both the unbiased calculation asymptotic variances were estimated using ACOR [14]. No error bars are given for the two one-dimensional calculations, as the barrier depicted in Figure 10 makes accurate estimation of the asymptotic variance impossible. A clear error is visible in the two one-dimensional umbrella sampling calculations, due to initialization along either side of the barrier in Figure 10. Figure 5b displays the asymptotic variance of the marginal density in $\log_{10} \lambda_1$ for the unbiased and the two-dimensional EMUS calculations. We note that while the unbiased calculation achieves greater accuracy near the mode, the EMUS calculation achieves greater accuracy in the tails.

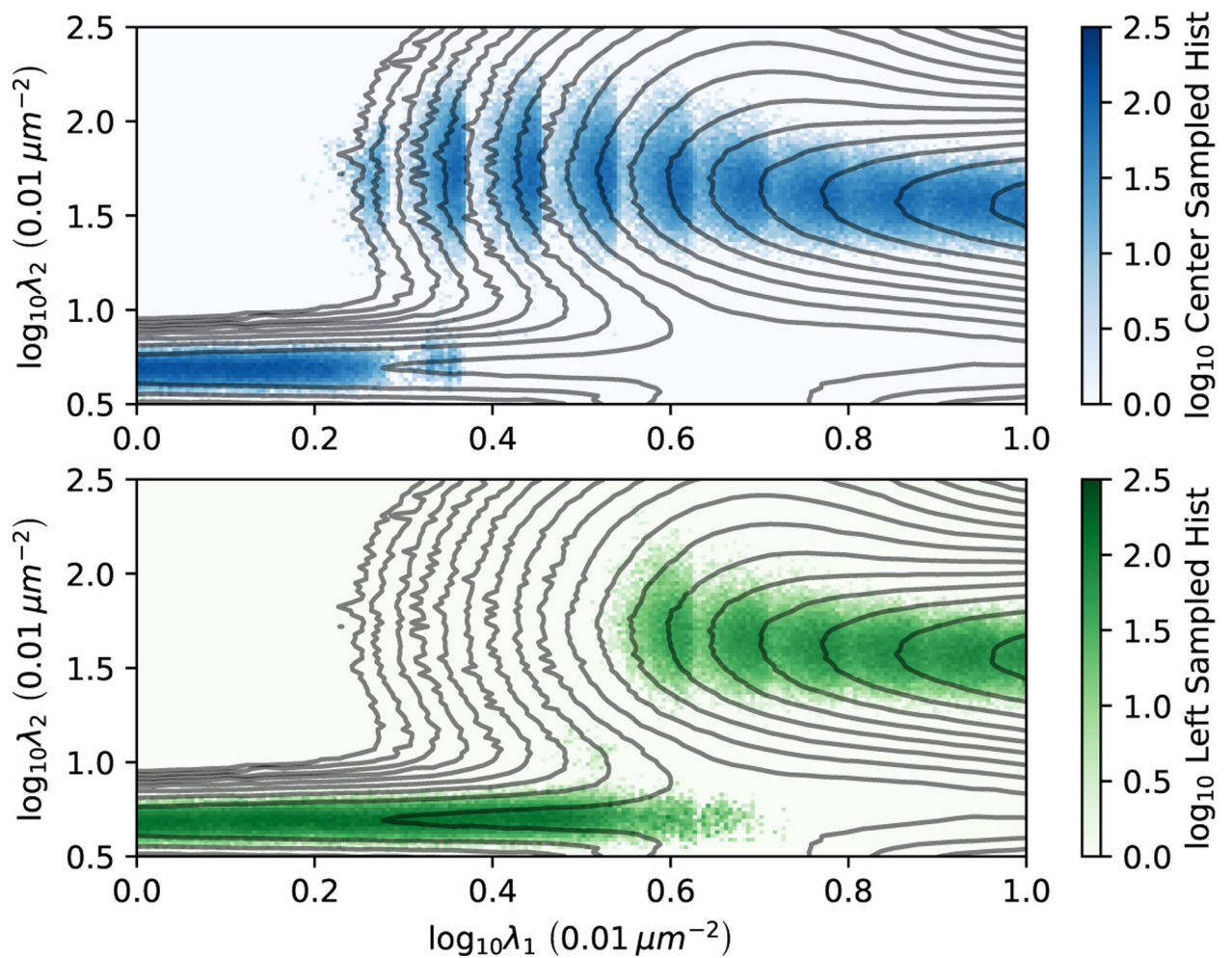
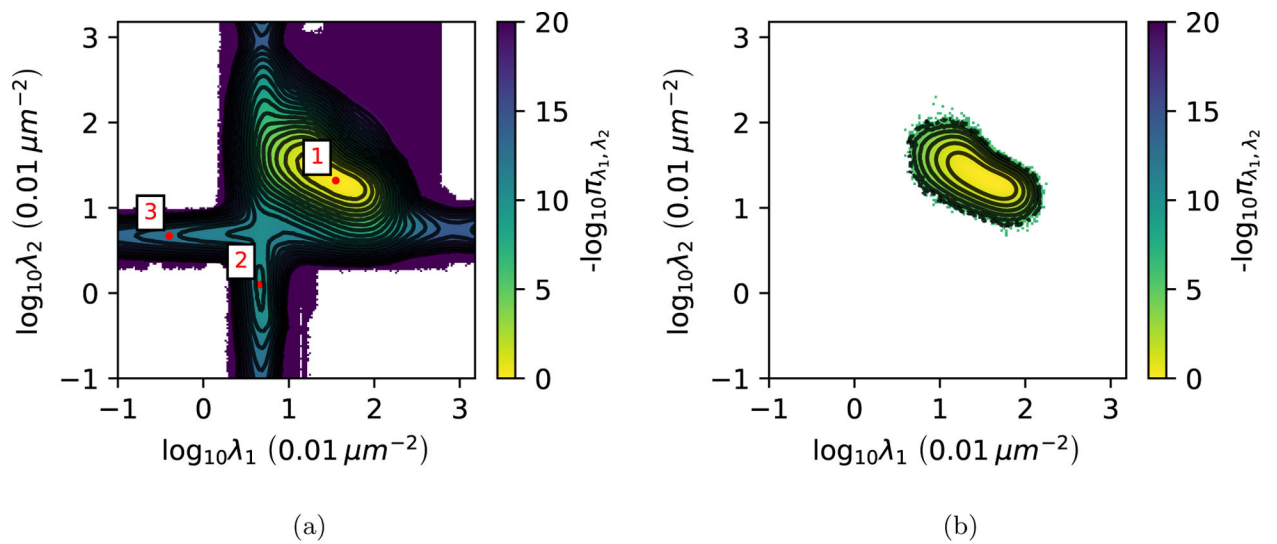


Figure 6:

To generate Figure 6, we binned the samples for the one-dimensional left and center EMUS calculations, and we plotted the difference in the histograms. The contour lines are contours of the log marginal density, as in Figure 7a. Figure 6 shows that while the two calculations largely sample the same regions, near $\log_{10}\lambda_1 = 0.45$ they become trapped on opposite sides of a barrier. This leads to poor sampling, causing a slowly decaying error in the estimates of the marginal density, cf. Figure 5a.

**Figure 7:**

Logarithm of marginal density in $\log_{10} \lambda_1$ and $\log_{10} \lambda_2$ as estimated by EMUS and unbiased MCMC. Contour lines in both figures are every unit change in the estimated \log_{10} marginal density. Figure 7a is the EMUS estimate. The numbers 1, 2, and 3 on this figure correspond to the mixture densities in Figure 8. Note that at values of $\log_{10} \lambda$ near 3.0 we begin to see the modes corresponding to singularities of the posterior. Figure 7b is the marginal density estimated from a long unbiased trajectory of the ensemble sampler. Note that the entire trajectory lies in a small neighborhood of the mode labeled 1 in Figure 7a.

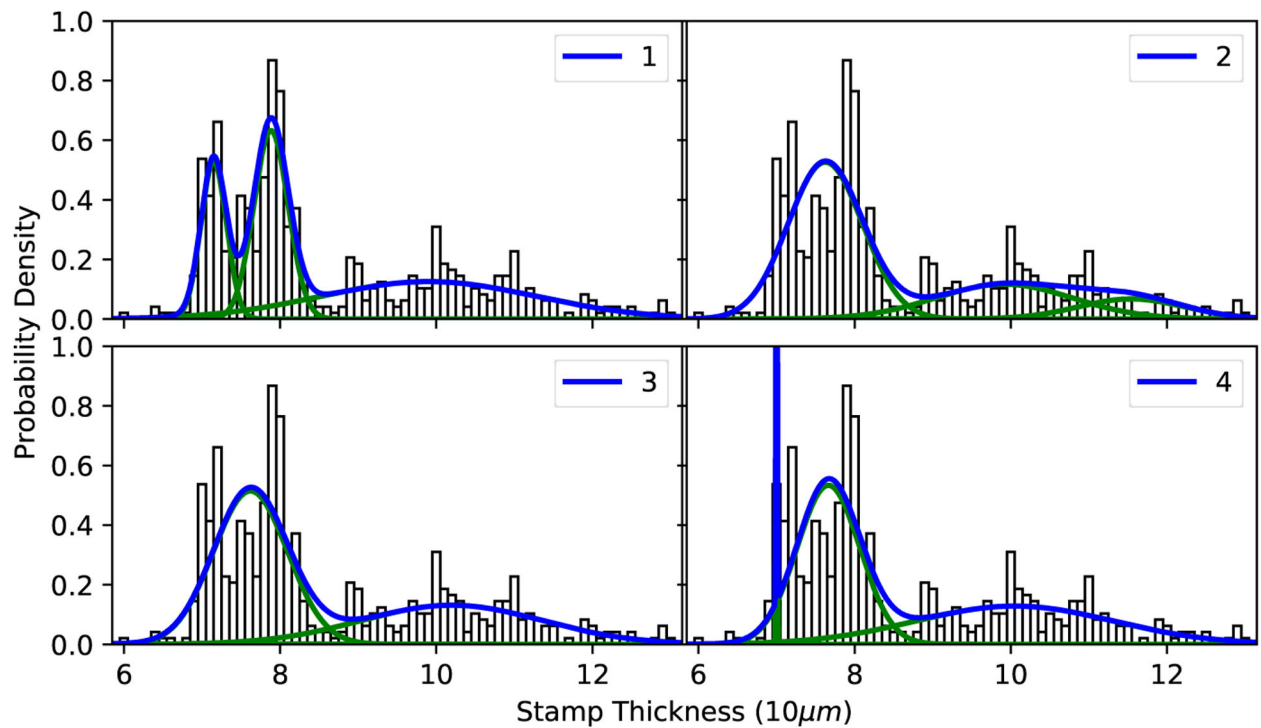


Figure 8:

Gaussian mixtures corresponding to means of histogram bins. Mixtures one through three correspond to the labeled points on Figure 7a, mixture four corresponds to a distribution near a singularity of the posterior, with $\log_{10} \lambda_1 = 4.34$ and $\log_{10} \lambda_2 = 0.79$. To be precise, the blue curve in each plot is the mixture distribution corresponding to the mean of a histogram bin centered at the point labeled in Figure 7a. The green curves are the individual mixture components. The black bars are a histogram of the Hidalgo stamp data.

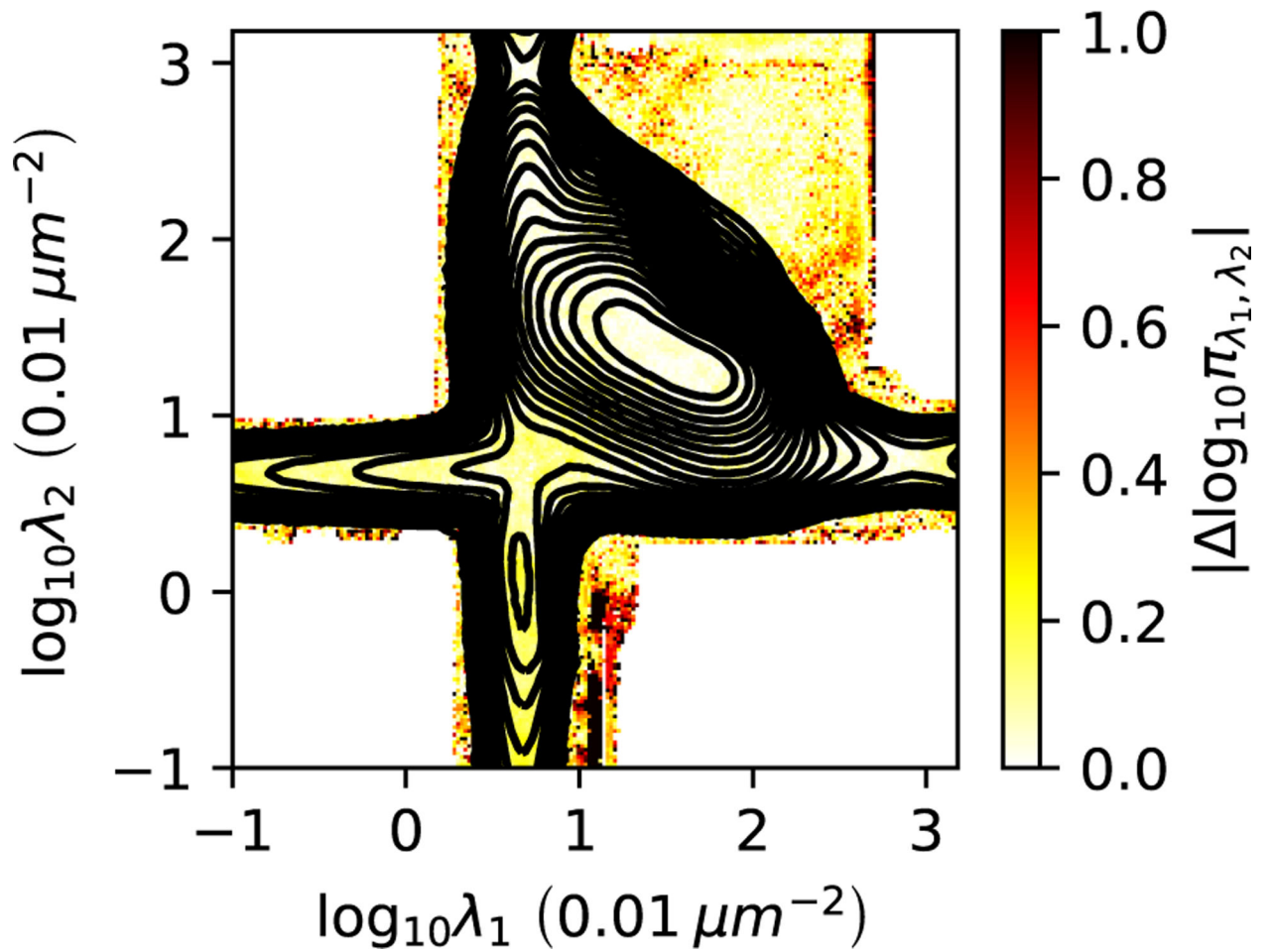


Figure 9:

The difference between the free energy surfaces of the two-dimensional umbrella sampling runs. The center calculation was initialized from the center one-dimensional calculation, and the left calculation from the left one-dimensional calculation. In general the difference is small, roughly a tenth of an order of magnitude in the log marginal.

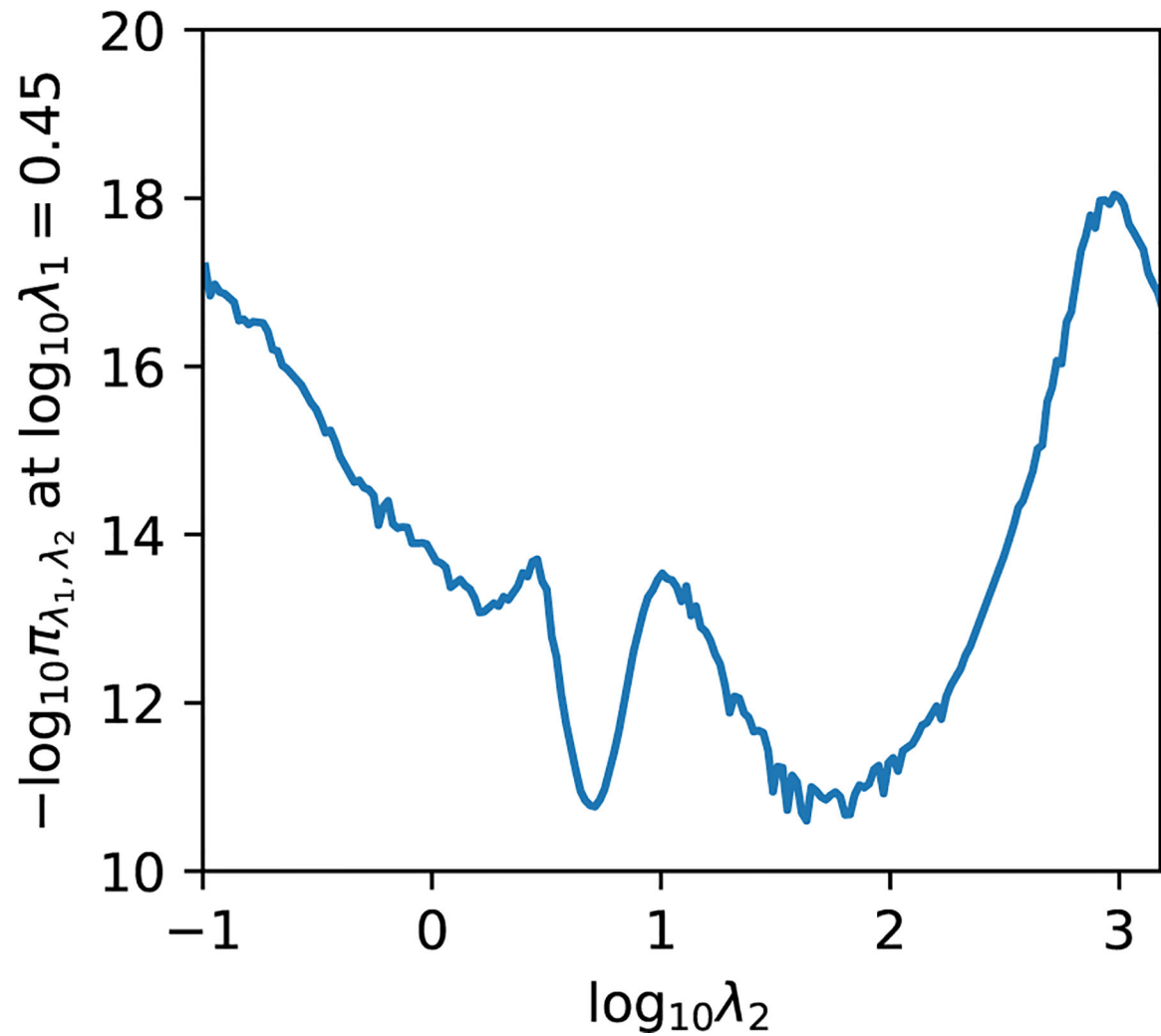
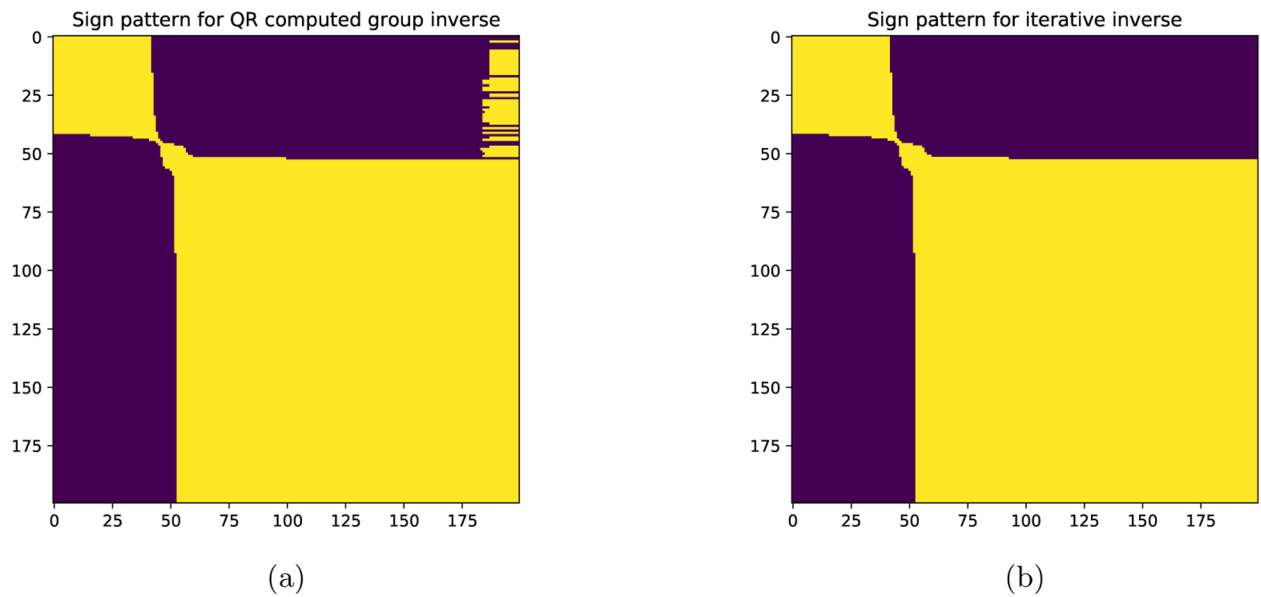


Figure 10:

Here we give an estimate of the conditional distribution of $\log_{10} \lambda_2$ with $\log_{10} \lambda_1 = 0.45$ calculated from the two-dimensional marginal seen in Figure 7a. The conditional distribution is multimodal. The mode on the left corresponds to mixtures with the data from thicknesses of 60 to 85 μm covered by a single Gaussian similar to mode 2 in Figure 8. The mode on the right corresponds to mixtures with these data covered by two Gaussians similar to mode 1 in Figure 8.

**Figure 11:**

Sign pattern of group inverse $(I - \bar{F})^\#$ computed by method of [19] (Figure 11a) and using power iteration (Figure 11b). Yellow indicates an entry with positive sign, blue a negative sign. Here, we consider the overlap matrix \bar{F} computed to estimate the marginal density of μ_2 in Section 5.3. The oscillations in sign observed in the upper right corner of Figure 11a are evidence of numerical error.