Function-Specific Mixing Times and Concentration Away from Equilibrium

Maxim Rabinovich, Aaditya Ramdas
Michael I. Jordan, Martin J. Wainwright
{rabinovich,aramdas,jordan,wainwrig}@berkeley.edu
University of California, Berkeley

September 30, 2016

Abstract

Slow mixing is the central hurdle when working with Markov chains, especially those used for Monte Carlo approximations (MCMC). In many applications, it is only of interest to estimate the stationary expectations of a small set of functions, and so the usual definition of mixing based on total variation convergence may be too conservative. Accordingly, we introduce function-specific analogs of mixing times and spectral gaps, and use them to prove Hoeffding-like function-specific concentration inequalities. These results show that it is possible for empirical expectations of functions to concentrate long before the underlying chain has mixed in the classical sense, and we show that the concentration rates we achieve are optimal up to constants. We use our techniques to derive confidence intervals that are sharper than those implied by both classical Markov chain Hoeffding bounds and Berry-Esseen-corrected CLT bounds. For applications that require testing, rather than point estimation, we show similar improvements over recent sequential testing results for MCMC. We conclude by applying our framework to real data examples of MCMC, providing evidence that our theory is both accurate and relevant to practice.

1 Introduction

Methods based on Markov chains play a critical role in statistical inference, where they form the basis of Markov chain Monte Carlo (MCMC) procedures for estimating intractable expectations [see, e.g., 9, 28]. In MCMC procedures, it is the stationary distribution of the Markov chain that typically encodes the information of interest. Thus, MCMC estimates are asymptotically exact, but their accuracy at finite times is limited by the convergence rate of the chain.

The usual measures of convergence rates of Markov chains—namely, the total variation mixing time or the absolute spectral gap of the transition matrix [21]—correspond to very strong notions of convergence and depend on global properties of the chain. Indeed, convergence of a Markov chain in total variation corresponds to uniform convergence of the expectations of all unit-bounded function to their equilibrium values. The resulting uniform bounds on the accuracy of expectations [4, 11, 18, 19, 20, 22, 27, 29] may be overly pessimistic—not indicative of the mixing times of specific expectations such as means and variances that are likely to be of interest in an inferential setting.

Another limitation of the uniform bounds is that they typically assume that the chain has arrived at the equilibrium distribution, at least approximately. Consequently, applying such bounds requires either assuming that the chain is started in equilibrium—impossible in practical applications of MCMC—or that the burn-in period is proportional to the mixing time of the chain, which is also unrealistic, if not impossible, in practical settings.

Given that the goal of MCMC is often to estimate specific expectations, as opposed to obtaining the stationary distribution, in the current paper we develop a function-specific notion of convergence with application to problems in Bayesian inference. We define a notion of "function-specific

mixing time," and we develop function-specific concentration bounds for Markov chains, as well as spectrum-based bounds on function-specific mixing times. We demonstrate the utility of both our overall framework and our particular concentration bounds by applying them to examples of MCMC-based data analysis from the literature and by using them to derive sharper confidence intervals and faster sequential testing procedures for MCMC.

1.1 Preliminaries

We focus on discrete time Markov chains on d states given by a $d \times d$ transition matrix P that satisfies the conditions of irreducibility, aperiodicity, and reversibility. These conditions guarantee the existence of a unique stationary distribution π. The issue is then to understand how quickly empirical averages of functions of the Markov chain, of the form $f : [d] \to [0, 1]$, approach the stationary average, denoted by

$$\mu := \mathbb{E}_{X\sim\pi}[f(X)].$$

The classical analysis of mixing defines convergence rate in terms of the total variation distance:

$$d_{\mathrm{TV}}\big(p,\ q\big) = \sup_{f\colon \Omega\to[0,\ 1]} \Big|\mathbb{E}_{X\sim p}\big[f(X)\big] - \mathbb{E}_{Y\sim q}\big[f(Y)\big]\Big|, \tag{1}$$

where the supremum ranges over all unit-bounded functions. The mixing time is then defined as the number of steps required to ensure that the chain is within total-variation distance δ of the stationary distribution—that is

$$T(\delta) := \min\Big\{n \in \mathbb{N} \ \mid\ \max_{i\in[d]} d_{\mathrm{TV}}\big(\pi_n^{(i)},\ \pi\big) \le \delta\Big\}, \tag{2}$$

where $\mathbb{N} = \{1, 2, \ldots\}$ denotes the natural numbers, and $\pi_n^{(i)}$ is the distribution of the chain state X_n given the starting state $X_0 = i$.

Total variation is a worst-case measure of distance, and the resulting notion of mixing time can therefore be overly conservative when the Markov chain is being used to approximate the expectation of a fixed function, or expectations over some relatively limited class of functions. Accordingly, it is of interest to consider the following function-specific discrepancy measure:

Definition 1 (f-discrepancy)**.** For a given function f, the f-discrepancy is

$$d_f\big(p,\ q\big) = \big|\mathbb{E}_{X\sim p}\big[f\big(X\big)\big] - \mathbb{E}_{Y\sim q}\big[f\big(Y\big)\big]\big|. \tag{3}$$

The f-discrepancy leads naturally to a function-specific notion of mixing time:

Definition 2 (f-mixing time)**.** For a given function f, the f-mixing time is

$$T_f\big(\delta\big) = \min\Big\{n \in \mathbb{N} \ \mid\ \max_{i\in[d]} d_f\big(\pi_n^{(i)},\ \pi\big) \le \delta\Big\}. \tag{4}$$

In the sequel, we also define function-specific notions of the spectral gap of a Markov chain, which can be used to bound the f-mixing time and to obtain function-specific concentration inequalities.

1.2 Related work

Mixing times are a classical topic of study in Markov chain theory, and there is a large collection of techniques for their analysis [see, e.g., 1, 6, 21, 23, 26, 30]. These tools and the results based on them, however, generally apply only to worst-case mixing times. Outside of specific examples [5, 7], relatively little is known about mixing with respect to individual functions or limited classes of functions. Similar limitations exist in studies of concentration of measure and studies of confidence intervals and other statistical functionals that depend on tail probability bounds. Existing bounds are generally uniform, or non-adaptive, and the rates that are reported include a factor that encodes the global mixing properties of the chain and does not adapt to the function [4, 11, 18, 19, 20, 22, 27, 29]. These factors, which do not appear in classic bounds for independent random variables, are generally either some variant of the spectral gap γ of the transition matrix, or else a mixing time of the chain $T\big(\delta_0\big)$ for some absolute constant $\delta_0 > 0$. For example, the main theorem from [20] shows that for a function $f\colon [d] \to [0,\ 1]$ and a sample $X_0 \sim \pi$ from the stationary distribution, we have

$$\mathbb{P}\big(\big|\frac{1}{N}\sum_{n=1}^{N} f\big(X_n\big) - \mu\big| \geq \epsilon\big) \leq 2\exp\Big\{-\frac{\gamma_0}{2\big(2-\gamma_0\big)}\cdot\epsilon^2 N\Big\}, \tag{5}$$

where the eigenvalues of P are given in decreasing order as $1 > \lambda_2(P) \geq \cdots \geq \lambda_d(P)$, and we denote the spectral gap of P by

$$\gamma_0 := \min\big\{1 - \lambda_2(P),\ 1\big\}.$$

The requirement that the chain start in equilibrium can be relaxed by adding a correction for the burn-in time [27]. Extensions of this and related bounds, including bounded-differences-type inequalities and generalizations to continuous Markov chains and non-Markov mixing processes have also appeared in the literature (e.g., [19, 29]).

The concentration result has an alternative formulation in terms of the mixing time instead of the spectral gap [4]. This version and its variants are weaker, since the mixing time can be lower bounded as

$$T\big(\delta\big) \geq \big(\frac{1}{\gamma_*} - 1\big)\log\big(\frac{1}{2\delta}\big) \geq \big(\frac{1}{\gamma_0} - 1\big)\log\big(\frac{1}{2\delta}\big), \tag{6}$$

where we denote the absolute spectral gap [21] by

$$\gamma_* := \min\big(1 - \lambda_2,\ 1 - \big|\lambda_d\big|\big)\ \leq\ \gamma_0.$$

In terms of the minimum probability $\pi_{\min} := \min_i \pi_i$, the corresponding upper bound is an extra factor of $\log\big(\frac{1}{\pi_{\min}}\big)$ larger, which potentially leads to a significant gap between $\frac{1}{\gamma_0}$ and $T\big(\delta_0\big)$, even for a moderate constant such as $\delta_0 = \frac{1}{8}$. Similar distinctions arise in our analysis, and we elaborate on them at the appropriate junctures.

1.3 Organization of the paper

In the remainder of the paper, we elaborate on these ideas and apply them to MCMC. In Section 2, we state some concentration guarantees based on function-specific mixing times, as well as some spectrum-based bounds on f-mixing times, and the spectrum-based Hoeffding bounds they imply. Section 3 is devoted to further development of these results in the context of several statistical models. More specifically, in Section 3.1, we show how our concentration guarantees can be used

to derive confidence intervals that are superior to those based on uniform Hoeffding bounds and CLT-type bounds, whereas in Section 3.2, we analyze the consequences for sequential testing. In Section 4, we show that our mixing time and concentration bounds improve over the non-adaptive bounds in real examples of MCMC from the literature. Finally, the bulk of our proofs are given in Section 5, with some more technical aspects of the arguments deferred to the appendices.

2 Main results

We now present our main technical contributions, starting with a set of "master" Hoeffding bounds with exponents given in terms of f-mixing times. As we explain in Section 2.3, these mixing time bounds can be converted to spectral bounds bounding the f-mixing time in terms of the spectrum. (We give some techniques for the latter in Section 2.2).

Recall that we use $\mu := \mathbb{E}_\pi[f]$ to denote the mean. Moreover, we follow standard conventions in setting

$$\lambda_* := \max\big\{\lambda_2(P),\ \big|\lambda_d(P)\big|\big\}, \quad \text{and} \quad \lambda_0 := \max\big\{\lambda_2(P),\ 0\big\}.$$

so that the absolute spectral gap and the (truncated) spectral gap introduced earlier are given by $\gamma_* := 1 - \lambda_*$, and $\gamma_0 := 1 - \lambda_0$. In Section 2.2, we define and analyze corresponding function-specific quantities, which we introduce as necessary.

2.1 Master Hoeffding bound

In this section, we present a master Hoeffding bound that provides concentration rates that depend on the mixing properties of the chain only through the f-mixing time T_f. The only hypotheses on burn-in time needed for the bounds to hold are that the chain has been run for at least $N \geq T_f$ steps—basically, so that thinning is possible—and that the chain was started from a distribution π_0 whose f-discrepancy distance from π is small—so that the expectation of each $f\big(X_n\big)$ iterate is close to μ—even if its total-variation discrepancy from π is large. Note that the latter requirement imposes only a very mild restriction, since it can always be satisfied by first running the chain for a burn-in period of T_f steps and then beginning to record samples.

Theorem 1. *Given any fixed $\epsilon > 0$ such that $d_f\big(\pi_0,\ \pi\big) \leq \frac{\epsilon}{2}$ and $N \geq T_f\big(\frac{\epsilon}{2}\big)$, we have*

$$\mathbb{P}\Big[\frac{1}{N}\sum_{n=1}^{N} f\big(X_n\big) \geq \mu + \epsilon\Big] \leq \exp\left\{-\frac{\epsilon^2 N}{8T_f\big(\frac{\epsilon}{2}\big)}\right\}. \tag{7}$$

Compared to the bounds in earlier work [e.g., 20], the bound (7) has several distinguishing features. The primary difference is that the "effective" sample size

$$N_{\text{eff}} := \frac{N}{T_f(\epsilon/2)}, \tag{8a}$$

is a function of f, which can lead to significantly sharper bounds on the deviations of the empirical means than the earlier uniform bounds can deliver. Further, unlike the uniform results, we do not require that the chain has reached equilibrium, or even approximate equilibrium, in a total variation sense. Instead, the result applies provided that the chain has equilibrated only approximately, and only with respect to f.

The reader might note that if one actually has access to a distribution π_0 that is $\epsilon/2$-close to π in f-discrepancy, then an estimator of μ with tail bounds similar to those guaranteed by Theorem 1 can

be obtained as follows: first, draw N i.i.d. samples from π_0, and second, apply the usual Hoeffding inequality for i.i.d. variables. However, it is essential to realize that Theorem 1 does not require that such a π_0 be available to the practitioner. Instead, the theorem statement is meant to apply in the following way: suppose that—starting from *any* initial distribution—we run an algorithm for $N \geq T_f(\epsilon/2)$ steps, and then use the last of $N - T_f(\epsilon/2)$ samples to form an empirical average. Our concentration result then holds with an effective sample size of

$$N_{\text{eff}}^{\text{burnin}} := \frac{N - T_f(\epsilon/2)}{T_f(\epsilon/2)} = \frac{N}{T_f(\epsilon/2)} - 1. \tag{8b}$$

In other words, the result can be applied with an arbitrary initial π_0, and accounting for burn-in merely reduces the effective sample size by one. By contrast, such an interpretation does not actually hold for the original result of [20]: it requires an initial sample $X_1 \sim \pi$, but such an exact sample is not attainable after any finite burn-in period.

The appearance of the function-specific mixing time T_f in the bounds comes with both advantages and disadvantages. A notable disadvantage, shared with the mixing time versions of the uniform bounds, is that spectrum-based bounds on the mixing time (including our f-specific ones) introduce a $\log\left(\frac{1}{\pi_{\min}}\right)$ term that can be a significant source of looseness. On the other hand, obtaining rates in terms of mixing times comes with the advantage that any bound on the mixing time translates directly into a version of the concentration bound (with the mixing time replaced by its upper bound). Moreover, since the $\pi_{\min}^{-1}$ term is likely to be an artifact of the spectrum-based approach, and possibly even just of the proof method, it may be possible to turn the mixing time based bound into a stronger spectrum-based bound with a more sophisticated analysis. We go part of the way toward doing this, albeit without completely removing the $\pi_{\min}^{-1}$ term.

An analysis based on mixing time also has the virtue of better capturing the non-asymptotic behavior of the rate. Indeed, as a consequence of the link (6) between mixing and spectral graph (as well as matching upper bounds [21]), for any fixed function f, there exists a function-specific spectral-gap $\gamma_f > 0$ such that

$$T_f\left(\frac{\epsilon}{2}\right) \approx \frac{1}{\gamma_f} \log\left(\frac{1}{\epsilon}\right) + O(1), \quad \text{for} \quad \epsilon \ll 1. \tag{8c}$$

These asymptotics can be used to turn our aforementioned theorem into a variant of the results of Léon and Perron [20], in which γ_0 is replaced by a value γ_f that (under mild conditions) is at least as large as γ_0. However, as we explore in Section 4, such an asymptotic spectrum-based view loses a great deal of information needed to deal with practical cases, where often $\gamma_f = \gamma_0$ and yet $T_f(\delta) \ll T(\delta)$ even for very small values of $\delta > 0$. For this reason, part of our work is devoted to deriving more fine-grained concentration inequalities that capture this non-asymptotic behavior.

By combining our definition (8a) of the effective sample size N_{eff} with the asymptotic expansion (8c), we arrive at an intuitive interpretation of Theorem 1: it dictates that the effective sample size scales as $N_{\text{eff}} \approx \frac{\gamma_f N}{\log(1/\epsilon)}$ in terms of the function-specific gap γ_f and tolerance ϵ. This interpretation is backed by the Hoeffding bound derived in Corollary 1 and it is useful as a simple mental model of these bounds. On the other hand, interpreting the theorem this way effectively plugs in the asymptotic behavior of T_f and does not account for the non-asymptotic properties of the mixing time; the latter may actually be more favorable and lead to substantially smaller effective sample sizes than the naive asymptotic interpretation predicts. From this perspective, the master bound has the advantage that any bound on T_f that takes advantage of favorable non-asymptotics translates directly into a stronger version of the Hoeffding bound. We investigate these issues empirically in Section 4.

Based on the worst-case Markov Hoeffding bound (5), we might hope that the $T_f(\frac{\epsilon}{2})$ term in Theorem 1 is spurious and removable using improved techniques. Unfortunately, it is fundamental. This conclusion becomes less surprising if one notes that even if we start the chain in its stationary distribution and run it for $N < T_f(\epsilon)$ steps, it may still be the case that there is a large set Ω_0 such that for $i \in \Omega_0$ and $1 \le n \le N$,

$$|f(X_n) - \mu| \gg \epsilon \quad \text{a.s. if } X_0 = i. \tag{9}$$

This behavior is made possible by the fact that large positive and negative deviations associated with different values in Ω_0 can cancel out to ensure that $\mathbb{E}[f(X_n)] = \mu$ marginally. However, the lower bound (9) guarantees that

$$\begin{aligned}\mathbb{P}\left(\frac{1}{N}\sum_{n=1}^{N} f(X_n) \ge \mu + \epsilon\right) &\ge \sum_{i\in\Omega_0} \pi_i \cdot \mathbb{P}\left(\frac{1}{N}\sum_{n=1}^{N} f(X_n) \ge \mu + \epsilon \mid X_0 = i\right) \\ &\ge \pi(\Omega_0),\end{aligned}$$

so that if $\pi(\Omega_0) \gg 0$, we have no hope of controlling the large deviation probability unless $N \gtrsim T_f(\epsilon)$. We make this intuitive argument precise in Section 2.5.

2.2 Bounds on f-mixing times

We generally do not have direct access either to the mixing time $T(\delta)$ or the f-mixing time $T_f(\delta)$. Fortunately, any bound on T_f translates directly into a variant of the tail bound (7). Accordingly, this section is devoted to methods for bounding these quantities. Since mixing time bounds are equivalent to bounds on d_{TV} and d_f, we frame the results in terms of distances rather than times. These results can then be inverted in order to obtain mixing-time bounds in applications.

The simplest bound is simply a uniform bound on total variation distance, which also yields a bound on the f-discrepancy. In particular, if the chain is started with distribution π_0, then we have

$$d_{\mathrm{TV}}(\pi_n, \pi) \le \frac{1}{\sqrt{\pi_{\min}}} \cdot \lambda_*^n \cdot d_{\mathrm{TV}}(\pi_0, \pi). \tag{10}$$

In order to improve upon this bound, we need to develop function-specific notions of spectrum and spectral gaps. The simplest way to do this is simply to consider the (left) eigenvectors to which the function is not orthogonal and define a spectral gap restricted only to the corresponding eigenvectors.

Definition 3 (f-eigenvalues and spectral gaps)**.** For a function $f\colon [d] \to \mathbb{R}$, we define

$$J_f := \left\{ j \in [d] \mid \lambda_j \ne 1 \text{ and } q_j^T f \ne 0 \right\}, \tag{11a}$$

where q_j denotes a left eigenvector associated with λ_j. Similarly, we define

$$\lambda_f = \max_{j \in J_f} |\lambda_j|, \quad \text{and} \quad \gamma_f = 1 - \lambda_f. \tag{11b}$$

Using this notation, it is straightforward to show that if the chain is started with the distribution π_0, then

$$d_f(\pi_n, \pi) \le \sqrt{\frac{\mathbb{E}_\pi[f^2]}{\pi_{\min}}} \cdot \lambda_f^n \cdot d_f(\pi_0, \pi). \tag{12}$$

This bound, though useful in many cases, is also rather brittle: it requires f to be exactly orthogonal to the eigenfunctions of the transition matrix. For example, a function f_0 with a good value of λ_f can be perturbed by an arbitrarily small amount in a way that makes the resulting perturbed function f_1 have $\lambda_f = \lambda_*$. More broadly, the bound is of little value for functions with a small but nonzero inner product with the eigenfunctions corresponding to large eigenvalues (which is likely to occur in practice; cf. Section 4), or in scenarios where f lacks symmetry (cf. the random function example in Section 2.4).

In order to address these issues, we now derive a more fine-grained bound on d_f. The basic idea is to split the lower f-spectrum J_f into a "bad" piece J, whose eigenvalues are close to 1 but whose eigenvectors are approximately orthogonal to f, and a "good" piece $J_f \setminus J$, whose eigenvalues are far from 1 and which therefore do not require control on the inner products of their eigenvectors with f. More precisely, for a given set $J \subset J_f$, let us define

$$\Delta_J^* := 2\big|J\big| \times \max_{j\in J} \|h_j\|_\infty \times \max_{j\in J} \big|q_j^T f\big|, \qquad \lambda_J := \max\Big\{\big|\lambda_j\big| \;\Big|\; j \in J\Big\}, \quad \text{and}$$
$$\lambda_{-J} := \max\Big\{\big|\lambda_j\big| \;\Big|\; j \in J_f \setminus J\Big\}.$$

We obtain the following bound, expressed in terms of λ_{-J} and λ_J, which we generally expect to obey the relation $1 - \lambda_{-J} \ll 1 - \lambda_J$.

Lemma 1 (Sharper f-discrepancy bound)**.** *Given $f\colon [d] \to [0,\ 1]$ and a subset $J \subset J_f$, we have*

$$d_f\big(\pi_n,\ \pi\big) \le \Delta_J^*\, \lambda_J^n \cdot d_{\mathrm{TV}}(\pi_0,\ \pi) + \sqrt{\frac{\mathbb{E}_\pi\big[f^2\big]}{\pi_{\min}}} \cdot \lambda_{-J}^n\, d_f(\pi_0,\ \pi). \tag{13}$$

The above bound, while easy to apply and comparatively easy to estimate, can be loose when the first term is a poor estimate of the part of the discrepancy that comes from the J part of the spectrum. We can get a still sharper estimate by instead making use of the following vector quantity that more precisely summarizes the interactions between f and J:

$$h_J\big(n\big) := \sum_{j\in J} \big(q_j^T f \cdot \lambda_j^n\big) h_j.$$

This quantity leads to what we refer to as an *oracle adaptive bound*, because it uses the exact value of the part of the discrepancy coming from the J eigenspaces, while using the same bound as above for the part of the discrepancy coming from $J_f \setminus J$.

Lemma 2 (Oracle f-discrepancy bound)**.** *Given $f\colon [d] \to [0,\ 1]$ and a subset $J \subset J_f$, we have*

$$d_f\big(\pi_n,\ \pi\big) \le \big|\big(\pi_0 - \pi\big)^T h_J\big(n\big)\big| + \sqrt{\frac{\mathbb{E}_\pi\big[f^2\big]}{\pi_{\min}}} \cdot \lambda_{-J}^n \cdot d_f\big(\pi_0,\ \pi\big). \tag{14}$$

We emphasize that, although Lemma 2 is stated in terms of the initial distribution π_0, when we apply the bound in the real examples we consider, we replace all quantities that depend on π_0 by their worst cases values, in order to avoid dependence on initialization; this results in a $\|h_J\big(n\big)\|_\infty$ term instead of the dot product in the lemma.

2.3 Concentration bounds

The mixing time bounds from Section 2.2 allow us to translate the master Hoeffding bound into a weaker but more interpretable—and in some instances, more directly applicable—concentration bound. The first result we prove along these lines applies meaningfully only to functions f whose absolute f-spectral gap γ_f is larger than the absolute spectral gap γ_*. It is a direct consequence of the master Hoeffding bound and the simple spectral mixing bound (12), and it delivers the asymptotics in N and ϵ promised in Section 2.1.

Corollary 1. *Given any $\epsilon > 0$ such that $d_f(\pi_0,\ \pi) \leq \frac{\epsilon}{2}$ and $N \geq T_f\left(\frac{\epsilon}{2}\right)$, we have*

$$\mathbb{P}\left[\frac{1}{N}\sum_{n=1}^{N} f(X_n) \geq \mu + \epsilon\right] \leq \begin{cases} \exp\left(-\frac{\epsilon^2}{8}\,\frac{\gamma_f N}{\log\left(\frac{2}{\epsilon\sqrt{\pi_{\min}}}\right)}\right) & \text{if } \epsilon \leq \frac{2\lambda_f}{\sqrt{\pi_{\min}}}, \\ \exp\left(-\frac{\epsilon^2 N}{8}\right) & \text{otherwise.} \end{cases}$$

Deriving a Hoeffding bound using the sharper f-mixing bound given in Lemma 1 requires more care, both because of the added complexity of managing two terms in the bound and because one of those terms does not decay, meaning that the bound only holds for sufficiently large deviations $\epsilon > 0$.

The following result represents one way of articulating the bound implied by Lemma 1; it leads to improvements over the previous two results when the contribution from the bad part of the spectrum J—that is, the part of the spectrum that brings γ_f closer to 1 than we would like—is negligible at the scale of interest. Recall that Lemma 1 expresses the contribution of J via the quantity Δ_J^*.

Corollary 2. *Given a triple of positive numbers $(\Delta, \Delta_J, \Delta_J^*)$ such that $\Delta_J \geq \Delta_J^*$ and $N \geq T_f(\Delta_J + \Delta)$, we have*

$$\mathbb{P}\left[\frac{1}{N}\sum_{n=1}^{N} f(X_n) \geq \mu + 2\left(\Delta_J + \Delta\right)\right] \leq \begin{cases} \exp\left(-\frac{\left(\Delta_J+\Delta\right)^2}{2}\,\frac{\left(1-\lambda_{-J}\right)N}{\log\left(\frac{1}{\Delta\sqrt{\pi_{\min}}}\right)}\right) & \text{if } \Delta \leq \frac{\lambda_{-J}}{\sqrt{\pi_{\min}}}, \\ \exp\left(-\frac{\left(\Delta_J+\Delta\right)^2 N}{2}\right) & \text{if } \Delta > \frac{\lambda_{-J}}{\sqrt{\pi_{\min}}}. \end{cases} \tag{15}$$

Similar arguments can be applied to combine the master Hoeffding bounds with the oracle f-mixing bound Lemma 2, but we omit the corresponding result for the sake of brevity. The proofs for both aforementioned corollaries are in Section 5.2.

2.4 Example: Lazy random walk on C_{2d}

In order to illustrate the mixing time and Hoeffding bounds from Section 2.2, we analyze their predictions for various classes of functions on the $2d$-cycle C_{2d}, identified with the integers modulo $2d$. In particular, consider the Markov chain corresponding to a lazy random walk on C_{2d}; it has transition matrix

$$P_{uv} = \begin{cases} \frac{1}{2} & \text{if } v = u, \\ \frac{1}{4} & \text{if } v = u+1 \mod 2d, \\ \frac{1}{4} & \text{if } v = u-1 \mod 2d, \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

It is easy to see that the chain is irreducible, aperiodic, and reversible, and its stationary distribution is uniform. It can be shown [21] that its mixing time scales proportionally to d^2. However, as we now show, several interesting classes of functions mix much faster, and in fact, a "typical" function, meaning a randomly chosen one, mixes much faster than the naive mixing bound would predict.

Parity function. The epitome of a rapidly mixing function is the parity function:

$$f_{\text{parity}}(u)) := \begin{cases} 1 & \text{if } u \text{ is odd,} \\ 0 & \text{otherwise.} \end{cases} \tag{17}$$

It is easy to see that no matter what the choice of initial distribution π_0 is, we have $\mathbb{E}\big[f_{\text{parity}}(X_1)\big] = \frac{1}{2}$, and thus f_{parity} mixes in a single step.

Periodic functions. A more general class of examples arises from considering the eigenfunctions of P, which are given by $g_j\big(u\big) = \cos\big(\frac{\pi j u}{d}\big)$; [see, e.g., 21]. We define a class of functions of varying regularity by setting

$$f_j = \frac{1 + g_j}{2}, \quad \text{for each } j = 0, 1, \ldots, d.$$

Here we have limited j to $0 \le j \le d$ because f_j and f_{2d-j} behave analogously. Note that the parity function f_{parity} corresponds to f_d.

Intuitively, one might expect that some of these functions mix well before d^2 steps have elapsed—both because the vectors $\{f_j,\ j \neq 1\}$ are orthogonal to the non-top eigenvectors with eigenvalues close to 1 and because as j gets larger, the periods of f_j become smaller and smaller, meaning that their global behavior can increasingly be well determined by looking at local snapshots, which can be seen in few steps.

Our mixing bounds allow us to make this intuition precise, and our Hoeffding bounds allow us to prove correspondingly improved concentration bounds for the estimation of $\mu = \mathbb{E}_\pi\big[f_j\big] = 1/2$. Indeed, we have

$$\gamma_{f_j} = \frac{1 - \cos\big(\frac{\pi j}{d}\big)}{2} \ge \begin{cases} \frac{\pi^2 j^2}{24 d^2} & \text{if } j \le \frac{d}{2}, \\ \frac{1}{2} & \text{if } \frac{d}{2} < j \le d. \end{cases} \tag{18}$$

Consequently, equation (12) predicts that

$$T_{f_j}\big(\delta\big) \le \tilde{T}_{f_j}\big(\delta\big) = \begin{cases} \frac{24}{\pi^2}\big[\frac{1}{2}\log 2d + \log\big(\frac{1}{\delta}\big)\big] \cdot \frac{d^2}{j^2} & \text{if } j \le \frac{d}{2}, \\ \log 2d + 2\log\big(\frac{1}{\delta}\big) & \text{if } \frac{d}{2} < j \le d, \end{cases} \tag{19}$$

where we have used the trivial bound $\mathbb{E}_\pi\big[f^2\big] \le 1$ to simplify the inequalities. Note that this yields an improvement over $\asymp d^2$ for $j \gtrsim \log d$. Moreover, the bound (19) can itself be improved, since each f_j is orthogonal to all eigenfunctions other than $\mathbf{1}$ and g_j, so that the $\log d$ factors can all be removed by a more carefully argued form of Lemma 1. It thus follows directly from the bound (18) that if we draw $N + \tilde{T}_{f_j}\big(\frac{\epsilon}{2}\big)$ samples, we obtain the tail bound

$$\mathbb{P}\Big[\frac{1}{N_0}\sum_{n=N_{\text{b}}}^{N+N_{\text{b}}} f_j\big(X_n\big) \ge \frac{1}{2} + \epsilon\Big] \le \begin{cases} \exp\big(-\frac{3d^2}{\pi^2 j^2} \cdot \frac{\epsilon^2 N}{\log\big(2\sqrt{2}d/\epsilon\big)}\big) & \text{if } j \le \frac{d}{2}, \\ \exp\big(-\frac{\epsilon^2 N}{16\log\big(2\sqrt{2}d/\epsilon\big)}\big) & \frac{d}{2} < j \le d, \end{cases} \tag{20}$$

where the burn-in time is given by $N_{\text{b}} = \tilde{T}_{f_j}\big(\epsilon/2\big)$. Note again that the sharper analysis mentioned above would allow us to remove the $\log 2d$ factors.

Random functions. A more interesting example comes from considering a randomly chosen function $f\colon C_{2d} \to [0,\ 1]$. Indeed, suppose that the function values are sampled iid from some distribution ν on $[0,\ 1]$ whose mean μ^* is $1/2$:

$$\{f(u),\ u \in C_{2d}\} \overset{\text{iid}}{\sim} \nu. \tag{21}$$

We can then show that for any fixed $\delta^* > 0$, with high probability over the randomness of f, have

$$T_f(\delta) \lesssim \frac{d \log d \big[\log d + \log\left(\frac{1}{\delta}\right)\big]}{\delta^2}, \qquad \text{for all } \delta \in (0, \delta^*]. \tag{22}$$

For $\delta \gg \frac{\log d}{\sqrt{d}}$, this scaling is an improvement over the global mixing time of order $d^2 \log(1/\delta)$.

The core idea behind the proof of equation (22) is to apply Lemma 1 with

$$J_\delta := \left\{ j \in \mathbb{N} \cap [1, 2d-1] \ \mid \ j \leq 4\delta\sqrt{\frac{d}{\log d}} \ \text{ or } \ j \geq 2d - 4\delta\sqrt{\frac{d}{\log d}} \right\}. \tag{23}$$

It can be shown that $\|h_j\|_\infty = 1$ for all $0 \leq j < 2d$ and that with high probability over f, $|q_j^T f| \lesssim \sqrt{\frac{\log d}{d}}$ simultaneously for all $j \in J_\delta$, which suffices to reduce the first part of the sharper f-discrepancy bound to order δ.

In order to estimate the rate of concentration, we proceed as follows. Taking $\delta = c_0\epsilon$ for a suitably chosen universal constant $c_0 > 0$, we show that $\Delta_J := \frac{\epsilon}{4} \geq \Delta_J^*$. We can then set $\Delta = \frac{\epsilon}{4}$ and observe that with high probability over f, the deviation in Corollary 2 satisfies the bound $2\left(\Delta_J + \Delta\right) \leq \epsilon$. With δ as above, we have $1 - \lambda_{-J} \geq \frac{c_1\epsilon^2}{d \log d}$ for another universal constant $c_1 > 0$. Thus, if we are given $N + T_f\big(\epsilon/2\big)$ samples for some $N \geq T_f\big(\frac{\epsilon}{2}\big)$, then we have

$$\mathbb{P}\left[\frac{1}{N}\sum_{n=T_f(\epsilon/2)}^{N+T_f(\epsilon/2)} f(X_n) \geq \mu + \epsilon\right] \leq \exp\left\{-\frac{c_2\epsilon^4 N}{d \log d\big[\log\left(\frac{4}{\epsilon}\right) + \log 2d\big]}\right\}, \tag{24}$$

for some $c_2 > 0$. Consequently, it suffices for the sample size to be lower bounded by

$$N \gtrsim \frac{d \log d\big[\log\left(1/\epsilon\right) + \log d\big]}{\epsilon^4},$$

in order to achieve an estimation accuracy of ϵ. Notice that this requirement is an improvement over the $\frac{d^2}{\epsilon^2}$ from the uniform Hoeffding bound provided that $\epsilon \gg (\frac{\log^2 d}{d})^{1/2}$. Proofs of all these claims can be found in Appendix B.

2.5 Lower bounds

Let us now make precise the intuitive argument set forth at the end of Section 2.1. The basic idea is to start with an arbitrary candidate function $\delta : (0,\ 1) \to (0,\ 1)$ such that $T_f\left(\frac{\epsilon}{2}\right)$ in the denominator of the function-specific Hoeffding bound (7) can be replaced by $T_f\left(\delta\left(\epsilon\right)\right)$ and show that if $\delta\left(\epsilon\right) \geq \epsilon$, the replacement is not actually possible. We prove this fact by constructing a Markov chain (which is independent of ϵ) and a function (which depends on both ϵ and δ) such that the Hoeffding bound is violated for the Markov chain-function pair for some value of N (which in general depends on the chain and ϵ).

As the following precise result shows, our lower bound continues to hold for an arbitrary constant in the exponent of the Hoeffding bound, meaning that Theorem 1 is optimal up to constants. We give the proof in Section 5.3.

Proposition 1. *For every constant $c_1 > 0$ and $\epsilon \in (0,1)$, there exists a Markov chain P_{c_1}, a number of steps $N = N(c_1, \epsilon)$ and a function $f = f_\epsilon$ such that*

$$\mathbb{P}_\pi\left(\left|\frac{1}{N}\sum_{n=1}^{N} f(X_n) - \frac{1}{2}\right| \geq \epsilon\right) > 2 \cdot \exp\left(-\frac{c_1 N \epsilon^2}{T_f\left(\delta(\epsilon)\right)}\right). \tag{25}$$

3 Statistical applications

We now consider how our results apply to Markov chain Monte Carlo (MCMC) in various statistical settings. Our investigation proceeds along three connected avenues. We begin by showing, in Section 3.1, how our concentration bounds can be used to provide confidence intervals for stationary expectations that avoid the over-optimism of pure CLT predictions without incurring the prohibitive penalty of the Berry-Esseen correction—or the global mixing rate penalty associated with spectral-gap-based confidence intervals. Then, in Section 3.2, we show how our results allow us to improve on recent sequential hypothesis testing methodologies for MCMC, again replacing the dependence on the spectral gap by a dependence on the f-mixing time. Later, in Section 4, we illustrate the practical significance of function-specific mixing properties by using our framework to analyze three real-world instances of MCMC, basing both the models and datasets chosen on real examples from the literature.

3.1 Confidence intervals for posterior expectations

In many applications, a point estimate of $\mathbb{E}_\pi\big[f\big]$ does not suffice; the uncertainty in the estimate must be quantified, for instance by providing $(1-\alpha)$ confidence intervals for some pre-specified constant α. In this section, we discuss how improved concentration bounds can be used to obtain sharper confidence intervals. In all cases, we assume the Markov chain is started from some distribution π_0 that need not be the stationary distribution, meaning that the confidence intervals must account for the burn-in time required to get close to equilibrium.

We first consider a bound that is an immediate consequence of the uniform Hoeffding bound given by [20]. As one would expect, it gives contraction at the usual Hoeffding rate but with an effective sample size of $N_{\text{eff}} \approx \gamma_0(N - T_0)$, where T_0 is the tuneable burn-in parameter. Note that this means that no matter how small T_f is compared to the global mixing time T, the effective size incurs the penalty for a global burn-in and the effective sample size is determined by the global spectral parameter γ_0. In order to make this precise, for a fixed burn-in level $\alpha_0 \in (0, \alpha)$, define

$$\epsilon_N(\alpha,\ \alpha_0) := \sqrt{2\big(2-\gamma_0\big)} \cdot \sqrt{\frac{\log\big(2/\big[\alpha - \alpha_0\big]\big)}{\gamma_0\big[N - T\big(\alpha_0\big)\big]}}. \tag{26a}$$

Then the uniform Markov Hoeffding bound [20, Thm. 1] implies that the set

$$I_N^{\text{unif}}\big(\alpha,\ \alpha_0\big) = \left[\frac{1}{N - T\big(\alpha_0/2\big)} \sum_{n=T\left(\alpha_0/2\right)+1}^{N} f\big(X_n\big) \pm \epsilon_N\big(\alpha,\ \alpha_0\big)\right] \tag{26b}$$

is a $1-\alpha$ confidence interval. Full details of the proof are given in Appendix C.1.

Moreover, given that we have a family of confidence intervals—one for each choice of $\alpha_0 \in (0, \alpha)$—we can obtain the sharpest confidence interval by computing the infimum $\epsilon_N^*\big(\alpha\big) := \inf_{0<\alpha_0<\alpha} \epsilon_N\big(\alpha,\ \alpha_0\big)$.

Equation (26b) then implies that

$$I_N^{\mathrm{unif}}(\alpha) = \Big[\frac{1}{N - T(\alpha_0)} \sum_{n=T(\alpha_0/2)+1}^{N} f(X_n) \pm \epsilon_N^*(\alpha)\Big]$$

is a $1-\alpha$ confidence interval for μ.

We now consider one particular application of our Hoeffding bounds to confidence intervals, and find that the resulting interval adapts to the function, both in terms of burn-in time required, which now falls from a global mixing time to an f-specific mixing time, and in terms of rate, which falls from $\frac{1}{\gamma_0}$ to $T_f(\delta)$ for an appropriately chosen $\delta > 0$. We first note that the one-sided tail bound of Theorem 1 can be written as $e^{-r_N(\epsilon)/8}$, where

$$r_N(\epsilon) := \epsilon^2 \left[\frac{N}{T_f\left(\frac{\epsilon}{2}\right)} - 1\right]. \tag{27}$$

If we wish for each tail to have probability mass that is at most $\alpha/2$, we need to choose $\epsilon > 0$ so that $r_N(\epsilon) \geq 8 \log \frac{2}{\alpha}$, and conversely any such ϵ corresponds to a valid two-sided $(1-\alpha)$ confidence interval. Let us summarize our conclusions:

Theorem 2. *For any width $\epsilon_N \in r_N^{-1}\big([8\log(2/\alpha),\ \infty)\big)$, the set*

$$I_N^{\mathrm{func}} := \left[\frac{1}{N - T_f\left(\frac{\epsilon}{2}\right)} \sum_{n=T_f\left(\frac{\epsilon}{2}\right)}^{N} f(X_n) \pm \epsilon_N\right]$$

is a $1-\alpha$ confidence interval for the mean $\mu = \mathbb{E}_\pi[f]$.

In order to make the result more amenable to interpretation, first note that for any $0 < \eta < 1$, we have

$$r_N(\epsilon) \geq \underbrace{\epsilon^2 \left[\frac{N}{T_f\left(\frac{\eta}{2}\right)} - 1\right]}_{r_{N,\eta}(\epsilon)} \quad \text{valid for all } \epsilon \geq \eta. \tag{28}$$

Consequently, whenever $r_{N,\eta}(\epsilon_N) \geq 8 \log \frac{2}{\alpha}$ and $\epsilon_N \geq \eta$, we are guaranteed that a symmetric interval of half-width ϵ_N is a valid $(1-\alpha)$-confidence interval. Summarizing more precisely, we have:

Corollary 3. *Fix $\eta > 0$ and let*

$$\epsilon_N = r_{N,\eta}^{-1}\big(8 \log \frac{2}{\alpha}\big) = 2\sqrt{2}\sqrt{\frac{T_f\left(\frac{\eta}{2}\right) \cdot \log(2/\alpha)}{N - T_f\left(\frac{\eta}{2}\right)}}.$$

If $N \geq T_f\left(\frac{\eta}{2}\right)$, then I_N^{func} is a $1-\alpha$ confidence interval for $\mu = \mathbb{E}_\pi[f]$.

Often, we do not have direct access to $T_f(\delta)$, but we can often obtain an upper bound $\tilde{T}_f(\delta)$ that is valid for all $\delta > 0$. In Section 5.4, therefore, which contains the proofs for this section, we prove a strengthened form of Theorem 2 and its corollary in that setting.

A popular alternative strategy for building confidence intervals using MCMC depends on the Markov central limit theorem (e.g., [8, 17, 12, 28]). If the Markov CLT held exactly, it would lead to appealingly simple confidence intervals of width

$$\tilde{\epsilon}_N = \sigma_{f,\text{asym}} \sqrt{\frac{\log(2/\alpha)}{N}},$$

where $\sigma^2_{f,\text{asym}} := \lim_{N\to\infty} \frac{1}{N} \text{Var}_{X_0\sim\pi}\big[\sum_{n=1}^N f(X_n)\big]$ is the asymptotic variance of f.

Unfortunately, the CLT does not hold exactly, even after the burn-in period. The amount by which it fails to hold can be quantified using a Berry-Esseen bound for Markov chains, as we now discuss. Let us adopt the compact notation $\tilde{S}_N = \sum_{n=1}^N \big[f(X_n) - \mu\big]$. We then have the bound [22]

$$\big|\mathbb{P}\big(\frac{\tilde{S}_N}{\sigma_{f,\text{asym}}\sqrt{N}} \le s\big) - \Phi(s)\big| \le \frac{e^{-\gamma_0 N}}{3\sqrt{\pi_{\min}}} + \frac{13}{\sigma_{f,\text{asym}}\sqrt{\pi_{\min}}} \cdot \frac{1}{\gamma_0\sqrt{N}}, \tag{29}$$

where Φ is the standard normal CDF. Note that this bound accounts for both the non-stationarity error and for non-normality error at stationarity. The former decays rapidly at the rate $e^{-\gamma_0 N}$, while the latter decays far more slowly, at the rate $\frac{1}{\gamma_0\sqrt{N}}$.

While the bound (29) makes it possible to prove a corrected CLT confidence interval, the resulting bound has two significant drawbacks. The first is that it only holds for extremely large sample sizes, on the order of $\frac{1}{\pi_{\min}\gamma_0^2}$, compared to the order $\frac{\log\left(1/\pi_{\min}\right)}{\gamma_0}$ required by the uniform Hoeffding bound. The second, shared by the uniform Hoeffding bound, is that it is non-adaptive and therefore bottlenecked by the global mixing properties of the chain. For instance, if the sample size is bounded below as

$$N \ge \max\Big(\frac{1}{\gamma_0}\log\Big(\frac{2}{\sqrt{\pi_{\min}}\alpha}\Big)\ ,\ \frac{1}{\gamma_0^2}\frac{6084}{\sigma^2_{f,\text{asym}}\pi_{\min}\alpha^2}\Big),$$

then both terms of equation (26b) are bounded by $1/6$, and the confidence intervals take the form

$$I_N^{\text{BE}} = \left[\frac{1}{N}\sum_{n=1}^N f(X_n) \pm \sigma_{f,\text{asym}}\sqrt{\frac{2\log\left(6/\alpha\right)}{N}}\right]. \tag{30}$$

See Appendix C.2 for the justification of this claim.

It is important to note that the width of this confidence interval involves a hidden form of mixing penalty. Indeed, defining the variance $\sigma_f^2 = \text{Var}_\pi\big[f(X)\big]$ and $\rho_f := \frac{\sigma_f^2}{\sigma^2_{f,\text{asym}}}$, we can rewrite the width as

$$\epsilon_N = \sigma_f\sqrt{\frac{2\log\left(6/\alpha\right)}{\rho_f N}}.$$

Thus, for this bound, the quantity ρ_f captures the penalty due to non-independence, playing the role of γ_0 and γ_f in the other bounds. In this sense, the CLT bound adapts to the function f, but only when it applies, which is at a sample-size scale dictated by the global mixing properties of the chain (i.e., γ_0).

3.2 Sequential testing for MCMC

For some applications, full confidence intervals may be unnecessary; instead, a practitioner may merely want to know whether $\mu = \mathbb{E}_\pi[f]$ lies above or below some threshold $0 < r < 1$. In these cases, we would like to develop a procedure for distinguishing between the two possibilities, at a given tolerable level $0 < \alpha < 1$ of combined Type I and II error. The simplest approach is, of course, to choose N so large that the $1-\alpha$ confidence interval built from N MCMC samples lies entirely on one side of r, but it may be possible to do better by using a sequential test. This latter idea was recently investigated in Gyori and Paulin [15], and we consider the same problem settings that they did:

(a) Testing with (known) indifference region, involving a choice between

$$H_0 \colon \mu \geq r + \delta$$
$$H_1 \colon \mu \leq r - \delta;$$

(b) Testing with no indifference region—that is, the same as above but with $\delta = 0$.

For the first setting (a), we always assume $0 < \delta < \nu := \min(\mu, 1-\mu)$, and the algorithm is evaluated on its ability to correctly choose between H_0 and H_1 when one of them holds, but it incurs no penalty for either choice when μ falls in the indifference region $\big(r-\delta,\ r+\delta\big)$. The error of a procedure $\mathcal{A}$ can thus be defined as

$$\mathrm{err}\big(\mathcal{A},\ f\big) = \begin{cases} \mathbb{P}\big(\mathcal{A}\big(X_{1:\infty}\big) = H_1\big) & \text{if } \mu \in H_0, \\ \mathbb{P}\big(\mathcal{A}\big(X_{1:\infty}\big) = H_0\big) & \text{if } \mu \in H_1, \\ 0 & \text{otherwise.} \end{cases}$$

The rest of this subsection is organized as follows. For the first setting (a), we analyze a procedure $\mathcal{A}_{\text{fixed}}$ that makes a decision after a fixed number $N := N(\alpha)$ of samples. We also analyze a sequential procedure $\mathcal{A}_{\text{seq}}$ that chooses whether to reject at a sequence $N_0, \ldots, N_k, \ldots$ of decision times. For the second, more challenging, setting (b), we analyze $\mathcal{A}_{\text{hard}}$, which also rejects at a sequence of decision times. For both $\mathcal{A}_{\text{seq}}$ and $\mathcal{A}_{\text{hard}}$, we calculate the expected stopping times of the procedures.

As mentioned above, the simplest procedure $\mathcal{A}_{\text{fixed}}$ would choose a fixed number N of samples to be collected based on the target level α. After collecting N samples, it forms the empirical average $\hat{\mu}_N = \frac{1}{N}\sum_{n=1}^{N} f\big(X_n\big)$ and outputs H_0 if $\hat{\mu}_N \geq r + \delta$, H_1 if $\hat{\mu}_N \leq r - \delta$, and outputs a special indifference symbol, say I, otherwise.

The sequential algorithm $\mathcal{A}_{\text{seq}}$ makes decisions as to whether to output one of the hypotheses or continue testing at a fixed sequence of decision times, say N_k. These times are defined recursively by

$$N_0 = \big\lfloor M \cdot \min\big(\frac{1}{r},\ \frac{1}{1-r}\big)\big\rfloor, \tag{31}$$
$$N_k = \big\lfloor N_0\big(1+\xi\big)^k \big\rfloor, \tag{32}$$

where $M > 0$ and $0 < \xi < 2/5$ are parameters of the algorithm. At each time N_k for $k \geq 1$, the algorithm $\mathcal{A}_{\text{seq}}$ checks if

$$\hat{\mu}_{N_k} \in \big(r - \frac{M}{N_k}, r + \frac{M}{N_k}\big). \tag{33}$$

If the empirical average lies in this interval, then the algorithm continues sampling; otherwise, it outputs H_0 or H_1 accordingly in the natural way.

For the sequential algorithm $\mathcal{A}_{\text{hard}}$, let $N_0 > 0$ be chosen arbitrarily,[1] and let N_k be defined in terms of N_0 as in (32). It once again decides at each N_k for $k \geq 1$ whether to output an answer or to continue sampling, depending on whether

$$\hat{\mu}_{N_k} \in \big(r - \epsilon_k(\alpha),\ r + \epsilon_k(\alpha)\big).$$

When this inclusion holds, the algorithm continues; when it doesn't hold, the algorithm outputs H_0 or H_1 in the natural way. The following result is restricted to the stationary case; later in the section, we turn to the question of burn-in.

Theorem 3. *Assume that $\alpha \leq \frac{2}{5}$. For $\mathcal{A}_{\text{fixed}}, \mathcal{A}_{\text{seq}}, \mathcal{A}_{\text{hard}}$ to all satisfy $\text{err}\big(\mathcal{A},\ f\big) \leq \alpha$, it suffices to (respectively) choose*

$$N = \frac{2T_f\big(\delta\big)\log\big(\frac{1}{\alpha}\big)}{\delta^2}, \tag{34}$$

$$M = \frac{8T_f\big(\frac{\delta}{2}\big)\log\big(\frac{2}{\sqrt{\alpha\xi}}\big)}{\delta},\ \textit{and} \tag{35}$$

$$\epsilon_k(\alpha) = \inf\Big\{\epsilon > 0\colon \frac{\epsilon^2}{8T_f\big(\frac{\epsilon}{2}\big)} \geq \frac{\log\big(1/\alpha\big) + 1 + 2\log k}{N_k}\Big\}, \tag{36}$$

where we let $\inf \emptyset = \infty$.

Our results differ from those of [15] because the latter implicitly control the worst-case error of the algorithm

$$\text{err}\big(\mathcal{A}\big) = \sup_{f\colon \Omega \to [0,\ 1]} \text{err}\big(\mathcal{A},\ f\big),$$

while our analysis controls $\text{err}\big(\mathcal{A},\ f\big)$ directly. The corresponding choices made in [15] are

$$N = \frac{\log(1/\alpha)}{\gamma_0\delta^2}, M = \frac{\log(\frac{2}{\sqrt{\alpha\xi}})}{\gamma_0\delta},\ \text{and}\ \epsilon_k(\alpha) = \sqrt{\frac{\log(1/\alpha) + 1 + 2\log k}{\gamma_0 N_k}}.$$

Hence, the T_f parameter in our bounds plays the same role that $\frac{1}{\gamma_0}$ plays in their uniform bounds. As a result of this close correspondence, we easily see that our results improve on the uniform result for a fixed function f whenever it converges to its stationary expectation faster than the chain itself converges— i.e., whenever $T_f\big(\delta\big) \leq \frac{1}{2\gamma_0}$.

The value of the above tests depends substantially on their sample size requirements. In setting (a), algorithm $\mathcal{A}_{\text{seq}}$ is only valuable if it reduces the number of samples needed compared to $\mathcal{A}_{\text{fixed}}$. In setting (b), algorithm $\mathcal{A}_{\text{hard}}$ is valuable because of its ability to test between hypotheses separated only by a point, but its utility is limited if it takes too long to run. Therefore, we now turn to the question of bounding expected stopping times.

In order to o carry out the stopping time analysis, we introduce the true margin $\Delta = |r - \mu|$. First, let us introduce some useful notation. Let $N(\mathcal{A})$ be the number of sampled collected by $\mathcal{A}$. Given a margin schedule $\big(\epsilon_k\big)$, let

$$k_0^*\big(\epsilon_{1:\infty}\big) := \min\big\{k \geq 1\colon \epsilon_k \leq \frac{\Delta}{2}\big\},\ \text{and}\ N_0^*\big(\epsilon_{1:\infty}\big) := N_{k_0^*\left(\epsilon_{1:\infty}\right)}.$$

We can bound the expected stopping times of $\mathcal{A}_{\text{seq}}, \mathcal{A}_{\text{hard}}$ in terms of Δ as follows:

[1] In Gyori and Paulin [15], the authors set $N_0 = \big\lfloor\frac{100}{\gamma_0}\big\rfloor$, but this is inessential.

Theorem 4. *Assume either H_0 or H_1 holds. Then,*

$$\mathbb{E}\big[N(\mathcal{A}_{\text{seq}})\big] \le \big(1+\xi\big)\big[\frac{M}{\Delta} + \frac{4}{\Delta}\sqrt{\frac{2T_f\big(\delta/2\big)M}{\Delta} + 8T_f\big(\delta/2\big)} + 1\big]; \tag{37}$$

$$\mathbb{E}\big[N(\mathcal{A}_{\text{hard}})\big] \le \big(1+\xi\big)\big(N_0^* + 1\big) + \frac{32\alpha T_f\big(\Delta/4\big)}{\Delta^2}. \tag{38}$$

With minor modifications to the proofs in [15], we can bound the expected stopping times of their procedures as

$$\mathbb{E}[N(\mathcal{A}_{\text{seq}})] \le (1+\xi)\Big\{\frac{M}{\Delta} + \frac{2}{\Delta}\sqrt{\frac{M}{\gamma_0\Delta} + \frac{4}{\gamma_0}} + 1\Big\};$$

$$\mathbb{E}[N(\mathcal{A}_{\text{hard}})] \le (1+\xi)\big(N_0^* + 1\big) + \frac{4\alpha}{\gamma_0\Delta^2}.$$

In order to see how the uniform and adaptive bounds compare, it is helpful to first note that, under either H_0 or H_1, we have the lower bound $\Delta \ge \delta$. Thus, the dominant term in the expectations in both cases is $(1+\xi)M/\Delta$. Consequently, the ratio between the expected stopping times is approximately equal to the ratio between the M values—viz.,

$$\frac{M_{\text{adapt}}}{M_{\text{unif}}} \approx \gamma_0 T_f\big(\delta/2\big). \tag{39}$$

As a result, we should expect a significant improvement in terms of number of samples when the relaxation time $\frac{1}{\gamma_0}$ is significantly larger than the f-mixing time $T_f\big(\delta/2\big)$. Framed in absolute terms, we can write

$$\bar{N}_{\text{unif}}(\mathcal{A}_{\text{seq}}) \approx \frac{\log\big(2/\sqrt{\alpha\xi}\big)}{\gamma_0\delta\Delta} \quad \text{and} \quad \bar{N}_{\text{adapt}}(\mathcal{A}_{\text{seq}}) \approx \frac{T_f\big(\delta/2\big)\log\big(2/\sqrt{\alpha\xi}\big)}{\delta\Delta}.$$

Up to an additive term, the bound for $\mathcal{A}_{\text{hard}}$ is also qualitatively similar to earlier ones, with $\frac{1}{\delta\Delta}$ replaced by $\frac{1}{\Delta^2}$.

4 Analyzing mixing in practice

We analyze several examples of MCMC-based Bayesian analysis from our theoretical perspective. These examples demonstrate that convergence in discrepancy can in practice occur much faster than suggested by naive mixing time bounds and that our bounds help narrow the gap between theoretical predictions and observed behavior.

4.1 Bayesian logistic regression

Our first example is a Bayesian logistic regression problem introduced by Robert and Casella [28]. The data consists of 23 observations of temperatures (in Fahrenheit, but normalized by dividing by 100) and a corresponding binary outcome—failure ($y = 1$) or not ($y = 0$) of a certain component; the aim is to fit a logistic regressor, with parameters $\big(\alpha,\ \beta\big) \in \mathbb{R}^2$, to the data, incorporating a prior and integrating over the model uncertainty to obtain future predictions. More explicitly, following the analysis in Gyori and Paulin [14], we consider the following model:

$$p\big(\alpha,\ \beta \mid b\big) = \frac{1}{b}\cdot e^{\alpha}\exp\big(-e^{\alpha}/b\big)$$

$$p\big(y \mid \alpha,\ \beta,\ x\big) \propto \exp\big(\alpha + \beta x\big),$$

which corresponds to an exponential prior on e^{α}, an improper uniform prior on β and a logit link for prediction. As in Gyori and Paulin [14], we target the posterior by running a Metropolis-Hastings algorithm with a Gaussian proposal with covariance matrix $\Sigma = \begin{pmatrix} 4 & 0 \\ 0 & 10 \end{pmatrix}$. Unlike in their paper, however, we discretize the state space to facilitate exact analysis of the transition matrix and to make our theory directly applicable. The resulting state space is given by

$$\Omega = \Big\{ \big(\hat{\alpha} \pm i \cdot \Delta,\ \hat{\beta} \pm j \cdot \Delta\big) \mid 0 \le i,\ j \le 8 \Big\},$$

where $\Delta = 0.1$ and $(\hat{\alpha},\ \hat{\beta})$ is the MLE. This space has $d = 17^2 = 289$ elements, resulting in a 289×289 transition matrix that can easily be diagonalized.

Robert and Casella [28] analyze the probability of failure when the temperature x is 65°F; it is specified by the function

$$f_{65}\big(\alpha,\ \beta\big) = \frac{\exp\big(\alpha + 0.65\beta\big)}{1 + \exp\big(\alpha + 0.65\beta\big)}.$$

Note that this function fluctuates significantly under the posterior, as shown in Figure 2.

We find that this function also happens to exhibit rapid mixing. The discrepancy $d_{f_{65}}$, before entering an asymptotic regime in which it decays exponentially at a rate $1 - \gamma^* \approx 0.386$, first drops from about 0.3 to about 0.01 in just 2 iterations, compared to the predicted 10 iterations from the naive bound $d_f(n) \le d_{\mathrm{TV}}(n) \le \frac{1}{\sqrt{\pi_{\min}}} \cdot (1 - \gamma^*)^n$. Figure 3 demonstrates this on a log scale, comparing the naive bound to a version of the bound in Lemmas 1 and 2. Note that the oracle f-discrepancy bound improves significantly over the uniform baseline, even though the non-oracle version does not. In this calculation, we took $J = \{2, \ldots, 140\}$ to include the top half of the spectrum excluding 1 and computed $\|h_j\|_\infty$ directly from P for $j \in J$ and likewise for $q_j^T f_{65}$. The oracle bound is given by Lemma 2. As shown in panel (b) of Figure 3, this decay is also faster than that of the total variation distance.

An important point is that the quality of the f-discrepancy bound depends significantly on the choice of J. In the limiting case where J includes the whole spectrum below the top eigenvalue, the oracle bound becomes exact. Between that and $J = \emptyset$, the oracle bound becomes tighter and tighter, with the rate of tightening depending on how much power the function has in the higher versus lower eigenspaces. Figure 4 illustrates this for a few settings of J, showing that although for this function and this chain, a comparatively large J is needed to get a tight bound, the oracle bound is substantially tighter than the uniform and non-oracle f-discrepancy bounds even for small J.

4.2 Bayesian analysis of clinical trials

The problem of missing data often necessitates Bayesian analysis, particularly in settings where uncertainty quantification is important, as in clinical trials. We illustrate how our framework would apply in this context by considering a clinical trials dataset [3, 14].

The dataset consists of $n = 50$ patients, some of whom participated in a trial for a drug and exhibited early indicators (Y_i) of success/failure and final indicators (X_i) of success/failure. Among the 50 patients, both indicator values are available for $n_X = 20$ patients; early indicators are available for $n_Y = 20$ patients; and no indicators are available for $n_0 = 10$ patients. The analysis

depends on the following parameterization:

$$\begin{aligned}\mathbb{P}\big(X_i = 1 \;\big|\; Y_i = 0\big) &= \gamma_0,\\ \mathbb{P}\big(X_i = 1 \;\big|\; Y_i = 1\big) &= \gamma_1,\\ \mathbb{P}\big(X_i = 1 \;\big|\; Y_i \text{ missing}\big) &= p.\end{aligned}$$

Note that, in contrast to what one might expect, p is to be interpreted as the marginal probability that $X_i = 1$, so that in actuality $p = \mathbb{P}\big(X_i = 1\big)$ unconditionally; we keep the other notation, however, for the sake of consistency with past work [3, 14]. Conjugate uniform (i.e., $\text{Be}\big(1,\ 1\big)$) priors are placed on all the model parameters.

The unknown variables include the parameter triple $\big(\gamma_0,\ \gamma_1,\ p\big)$ and the unobserved X_i values for $n_Y + n_0 = 30$ patients, and the full sample space is therefore $\tilde{\Omega} = [0,\ 1]^3 \times \big\{0,\ 1\big\}^{30}$. We cannot estimate the transition matrix for this chain, even with a discretization with as coarse a mesh as $\Delta = 0.1$, since the number of states would be $d = 10^3 \times 2^{30} \sim 10^{12}$. We therefore make two changes to the original MCMC procedure. First, we collapse out the X_i variables to bring the state space down to $[0,\ 1]^3$; while analytically collapsing out the discrete variables is impossible, we can estimate the transition probabilities for the collapsed chain analytically by sampling the X_i variables conditional on the parameter values and forming a Monte Carlo estimate of the collapsed transition probabilities. Second, since the function of interest in the original work—namely, $f\big(\gamma_0,\ \gamma_1,\ p\big) = \mathbf{1}\big(p > 0.5\big)$—depends only on p, we fix γ_0 and γ_1 to their MLE values and sample only p, restricted to the unit interval discretized with mesh $\Delta = 0.01$.

As Figure 1 shows, eigenvalue decay occurs rapidly for this sampler, with $\gamma^* \approx 0.86$. Mixing thus occurs so quickly that none of the bounds—uniform or function-specific—get close to the truth, due to the presence of the constant terms (and specifically the large term $\frac{1}{\sqrt{\pi_{\min}}} \approx 2.14 \times 10^{33}$). Nonetheless, this example still illustrates how in actual fact, the choice of target function can make a big difference in the number of iterations required for accurate estimation; indeed, if we consider the two functions

$$f_1(p) := \mathbf{1}\big(p > 0.5\big), \quad \text{and} \quad f_2(p) := p,$$

we see in Figure 5 that the mixing behavior differs significantly between them: whereas the discrepancy for the second decays at the asymptotic exponential rate from the outset, the discrepancy for the first decreases faster (by about an order of magnitude) for the first few iterations, before reaching the asymptotic rate dictated by the spectral gap.

4.3 Collapsed Gibbs sampling for mixture models

Due to the ubiquity of clustering problems in applied statistics and machine learning, Bayesian inference for mixture models (and their generalizations) is a widespread application of MCMC [10, 13, 16, 24, 25]. We consider the mixture-of-Gaussians model, applying it to a subset of the schizophrenic reaction time data analyzed in Belin and Rubin [2]. The subset of the data we consider consists of 10 measurements, with 5 coming from healthy subjects and 5 from subjects diagnosed with schizophrenia. Since our interest is in contexts where uncertainty is high, we chose the 5 subjects from the healthy group whose reaction times were greatest and the 5 subjects from the schizophrenic group whose reaction times were smallest. We considered a mixture with $K = 2$

components, viz.:

$$\begin{aligned} \mu_b &\sim \mathcal{N}\big(0,\ \rho^2\big),\ b = 0,\ 1, \\ \omega &\sim \mathrm{Be}\big(\alpha_0,\ \alpha_1\big) \\ Z_i \mid \omega &\sim \mathrm{Bern}\big(\omega\big) \\ X_i \mid Z_i = b,\ \mu &\sim \mathcal{N}\big(\mu_b,\ \sigma^2\big). \end{aligned}$$

We chose relatively uninformative priors, setting $\alpha_0 = \alpha_1 = 1$ and $\rho = 237$. Increasing the value chosen in the original analysis [2], we set $\sigma \approx 70$; we found that this was necessary to prevent the posterior from being too highly concentrated, which would be an unrealistic setting for MCMC. We ran collapsed Gibbs on the indicator variables Z_i by analytically integrating out ω and $\mu_{0:1}$.

As Figure 1 illustrates, the spectral gap for this chain is small—namely, $\gamma_* \approx 3.83 \times 10^{-4}$—yet the eigenvalues fall off comparatively quickly after λ_2, opening up the possibility for improvement over the uniform γ_*-based bounds. In more detail, define

$$z_b^* := \big(b \quad b \quad b \quad b \quad b \quad 1-b \quad 1-b \quad 1-b \quad 1-b \quad 1-b\big),$$

corresponding to the cluster assignments in which the patient and control groups are perfectly separated (with the control group being assigned label b). We can then define the indicator for exact recovery of the ground truth by

$$f(z) = \mathbf{1}\big(z \in \big\{z_0^*,\ z_1^*\big\}\big).$$

As Figure 6 illustrates, convergence in terms of f-discrepancy occurs much faster than convergence in total variation, meaning that predictions of required burn-in times and sample size based on global metrics of convergence drastically overestimate the computational and statistical effort required to estimate the expectation of f accurately using the collapsed Gibbs sampler. This behavior can be explained in terms of the interaction between the function f and the eigenspaces of P. Although the pessimistic constants in the bounds from the uniform bound (10) and the non-oracle function-specific bound (Lemma 1) make their predictions overly conservative, the oracle version of the function-specific bound (Lemma 2) begins to make exact predictions after just a hundred iterations when applied with $J = \big\{1, \ldots, 25\big\}$; this corresponds to making exact predictions of $T_f\big(\delta\big)$ for $\delta \leq \delta_0 \approx 0.01$, which is a realistic tolerance for estimation of μ. Panel (b) of Figure 6 documents this by plotting the f-discrepancy oracle bound against the actual value of d_f on a log scale.

Bound type	$T_f\big(0.01\big)$	$T_f\big(10^{-6}\big)$
Uniform	31,253	55,312
Function-Specific	25,374	49,434
Function-Specific (Oracle)	98	409
Actual	96	409

Table 1: Comparison of bounds on $T_f\big(\delta\big)$ for different values of δ. The uniform bound corresponds to the bound $T_f\big(\delta\big) \leq T\big(\delta\big)$, the latter of which can be bounded by the total variation bound. The function-specific bounds correspond to Lemmas 1 and 2, respectively. Whereas the uniform and non-oracle f-discrepancy bounds make highly conservative predictions, the oracle f-discrepancy bound is nearly sharp even for δ as large as 0.01.

The mixture setting also provides a good illustration of how the function-specific Hoeffding bounds can substantially improve on the uniform Hoeffding bound. In particular, let us compare

the T_f-based Hoeffding bound (Theorem 1) to the uniform Hoeffding bound established by Léon and Perron [20]. At equilibrium, the penalty for non-independence in our bounds is $(2T_f(\epsilon/2))^{-1}$ compared to roughly γ_*^{-1} in the uniform bound. Importantly, however, our concentration bound applies unchanged even when the chain has not equilibrated, provided it has approximately equilibrated with respect to f. As a consequence, our bound only requires a burn-in of $T_f(\epsilon/2)$, whereas the uniform Hoeffding bound does not directly apply for any finite burn-in. Table 1 illustrates the size of these burn-in times in practice. This issue can be addressed using the method of Paulin [27], but at the cost of a burn-in dependent penalty $d_{\mathrm{TV}}(T_0) = \sup_{\pi_0} d_{\mathrm{TV}}(\pi_n,\ \pi)$:

$$\mathbb{P}\Big[\frac{1}{N-T_0}\sum_{n=T_0}^{N} f(X_n) \geq \mu + \epsilon\Big] \leq d_{\mathrm{TV}}\big(T_0\big) + \exp\Big\{-\frac{\gamma_0}{2\big(1-\gamma_0\big)}\cdot\epsilon^2\big[N-T_0\big]\Big\}, \tag{40}$$

where we have let T_0 denote the burn-in time. Note that a matching bound holds for the lower tail. For our experiments, we computed the tightest version of the bound (40), optimizing T_0 in the range $\big[0,\ 10^5\big]$ for each value of the deviation ϵ. Even given this generosity toward the uniform bound, the function-specific bound still outperforms it substantially, as Figure 7 shows.

For the function-specific bound, we used the function-specific oracle bound (Lemma 2) to bound $T_f\big(\frac{\epsilon}{2}\big)$; this nearly coincides with the true value when $\epsilon \approx 0.01$ but deviates slightly for larger values of ϵ.

5 Proofs of main results

This section is devoted to the proofs of the main results of this paper.

5.1 Proof of Theorem 1

We begin with the proof of the master Hoeffding bound from Theorem 1. At the heart of the proof is the following bound on the moment-generating function (MGF) for the sum of an appropriately thinned subsequence of the function values $\{f(X_n)\}_{n=1}^{\infty}$. In particular, let us introduce the shorthand notation $\widetilde{X}_{m,t} := X_{(m-1)T_f(\epsilon/2)+t}$ and $N_0 := N/T_f(\frac{\epsilon}{2})$. With this notation, we have the following auxiliary result:

Lemma 3 (Master MGF bound)**.** *For any scalars $\beta \in \mathbb{R}$, $\epsilon \in (0,1)$, and integer $t \in \big[0, T_f(\frac{\epsilon}{2})\big)$, we have*

$$\mathbb{E}\left[\exp\Big(\beta\sum_{m=1}^{N_0} f\big(\widetilde{X}_{m,t}\big)\Big)\right] \leq \exp\left\{\big[\frac{1}{2}\beta\epsilon + \beta\mu + \frac{1}{2}\beta^2\big]\cdot N_0\right\}. \tag{41}$$

See Section 5.1.1 for the proof of this claim. Recalling the definition of $\widetilde{X}_{m,t}$, we have

$$\begin{aligned}
\mathbb{E}\left[e^{\alpha\sum_{n=1}^{N} f(X_n)}\right] &= \mathbb{E}\left[\exp\Big\{\alpha\sum_{t=1}^{T_f(\epsilon/2)}\sum_{m=1}^{N_0} f(\widetilde{X}_{m,t})\Big\}\right] \\
&= \mathbb{E}\left[\exp\Big\{\alpha T_f(\epsilon/2)\big[\frac{1}{T_f(\epsilon/2)}\sum_{t=1}^{T_f(\epsilon/2)}\sum_{m=1}^{N_0} f(\widetilde{X}_{m,t})\big]\Big\}\right] \\
&\leq \frac{1}{T_f\big(\epsilon/2\big)}\sum_{t=0}^{T_f\left(\epsilon/2\right)-1}\mathbb{E}\left[\exp\Big\{\alpha T_f\big(\epsilon/2\big)\sum_{m=1}^{N_0} f\big(\widetilde{X}_{m,t}\big)\Big\}\right],
\end{aligned}$$

where the last inequality follows from Jensen's inequality, as applied to the exponential function. Applying Lemma 3 with $\beta = \alpha T_f(\epsilon/2)$, we conclude

$$\mathbb{E}\left[e^{\alpha \sum_{n=1}^{N} f(X_n)}\right] \leq \exp\left\{\left[\frac{1}{2}\alpha T_f\left(\frac{\epsilon}{2}\right)\epsilon + \alpha T_f\left(\frac{\epsilon}{2}\right)\mu + \frac{1}{2}\alpha^2 T_f^2\left(\frac{\epsilon}{2}\right)\right] \cdot N_0\right\},$$

valid for $\alpha > 0$. By exponentiating and applying Markov's inequality, it follows that

$$\begin{aligned}\mathbb{P}\left[\frac{1}{N}\sum_{n=1}^{N} f\left(X_n\right) \geq \mu + \epsilon\right] &\leq e^{-\alpha(\mu+\epsilon)}\mathbb{E}\left[e^{\alpha \sum_{n=1}^{N} f(X_n)}\right] \\ &\leq \exp\left\{\frac{1}{2} \cdot \left[-\alpha\, T_f(\frac{\epsilon}{2})\epsilon + \alpha^2\, T_f^2(\frac{\epsilon}{2})\right] N_0(\frac{\epsilon}{2})\right\}.\end{aligned}$$

The proof of Theorem 1 follows by taking $\alpha = \frac{\epsilon}{2T_f\left(\frac{\epsilon}{2}\right)}$ since

$$\begin{aligned}\mathbb{P}\left[\frac{1}{N}\sum_{n=1}^{N} f\left(X_n\right) \geq \mu + \epsilon\right] &\leq \exp\left\{\frac{1}{2} \cdot \left[-\frac{\epsilon^2}{2} + \frac{\epsilon^2}{4}\right] \cdot N_0\right\} \\ &\leq \exp\left\{-\frac{\epsilon^2 N_0}{8}\right\} \\ &= \exp\left\{-\frac{\epsilon^2 N}{8T_f\left(\frac{\epsilon}{2}\right)}\right\}.\end{aligned}$$

5.1.1 Proof of Lemma 3

For the purposes of the proof, fix t and let $W_m = \tilde{X}_{m,t}$. For convenience, also define a dummy constant random variable $W_0 := 0$. Now, by assumption, we have

$$\left|\mathbb{E}\left[f\left(W_1\right)\right] - \mu\right| \leq \frac{\epsilon}{2} \quad \text{and} \quad \left|\mathbb{E}\left[f\left(W_{m+1}\right) \mid W_m\right] - \mu\right| \leq \frac{\epsilon}{2}.$$

We therefore have the bound

$$\mathbb{E}\left[e^{\alpha \sum_m f(W_m)}\right] \leq \mathbb{E}\left[\prod_{m=1}^{N_0} e^{\alpha[f(W_m) - \mathbb{E}[f(W_m) \mid W_{m-1}]]}\right] \cdot e^{\alpha\mu N_0 + \frac{\alpha\epsilon N_0}{2}}. \tag{42}$$

But now observe that the random variables $\Delta_m = f\left(W_m\right) - \mathbb{E}\left[f\left(W_m\right) \mid W_{m-1}\right]$ are deterministically bounded in $[-1,\ 1]$ and zero mean conditional on W_{m-1}. Moreover, by the Markovian property, this implies that the same is true conditional on $W_{<m} := W_{0:(m-1)}$. It follows by standard MGF bounds that

$$\mathbb{E}\left[e^{\alpha\Delta_m} \mid W_{<m}\right] \leq e^{\frac{\alpha^2}{2}}.$$

Combining this bound with inequality (42), we conclude that

$$\mathbb{E}\left[e^{\alpha \sum_m f(W_m)}\right] \leq e^{\frac{\alpha^2}{2} \cdot N_0} \cdot e^{\alpha\mu N_0 + \frac{\alpha\epsilon N_0}{2}},$$

as claimed.

5.2 Proofs of Corollaries 1 and 2

In this section, we prove the derived Hoeffing bounds stated in Corollaries 1 and 2.

5.2.1 Proof of Corollary 1

The proof is a direct application of Theorem 1. Indeed, it suffices to note that if $\epsilon \leq \frac{2\lambda_f}{\sqrt{\pi_{\min}}}$, then

$$T_f\left(\frac{\epsilon}{2}\right) \leq \frac{\log\left(\frac{2}{\epsilon\sqrt{\pi_{\min}}}\right)}{\log\left(\frac{1}{\lambda_f}\right)} = \frac{\log\left(\frac{2}{\epsilon}\right) + \frac{1}{2}\log\left(\frac{1}{\pi_{\min}}\right)}{\log\left(\frac{1}{\lambda_f}\right)},$$

which yields the first bound. Turning to the second bound, note that if $\epsilon > \frac{2\lambda_f}{\sqrt{\pi_{\min}}}$, then equation (12) implies that $T_f\left(\frac{\epsilon}{2}\right) = 1$, which establishes the claim.

5.2.2 Proof of Corollary 2

The proof involves combining Theorem 1 with Lemma 1, using the setting $\epsilon = 2\left(\Delta + \Delta_J\right)$. We begin by combining the bounds $\lambda_J \leq 1$, $d_{\mathrm{TV}}\left(\pi_0,\ \pi_n\right) \leq 1$, $d_f\left(\pi_0,\ \pi_n\right) \leq 1$, and $\mathbb{E}_\pi\left[f^2\right] \leq 1$ with the claim of Lemma 1 so as to find that

$$d_f\left(\pi_0,\ \pi_n\right) \leq \Delta_J^* + \frac{\lambda_{-J}^n}{\sqrt{\pi_{\min}}} \ \leq\ \Delta_J + \frac{\lambda_{-J}^n}{\sqrt{\pi_{\min}}}.$$

It follows that

$$T_f(\frac{\epsilon}{2}) = T_f\left(\Delta_J + \Delta\right) \leq \frac{\log\left(\frac{1}{\Delta}\right) + \frac{1}{2}\log\left(\frac{1}{\pi_{\min}}\right)}{\log(\frac{1}{\lambda_{-J}})} \qquad \text{whenever } \Delta \leq \frac{\lambda_{-J}}{\sqrt{\pi_{\min}}}.$$

Plugging into Theorem 1 now yields the first part of the bound. On the other hand, if $\Delta > \frac{\lambda_{-J}}{\sqrt{\pi_{\min}}}$, then Lemma 1 implies that $T_f\left(\Delta_J + \Delta\right) = 1$, which proves the bound in the second case.

5.3 Proof of Proposition 1

In order to prove the lower bound in Proposition 1, we first require an auxiliary lemma:

Lemma 4. *Fix a function $\delta\colon\ (0,\ 1) \to (0,\ 1)$ with $\delta\left(\epsilon\right) > \epsilon$. For every constant $c_0 \geq 1$, there exists a Markov chain P_{c_0} and a function f_ϵ on it such that $\mu = \frac{1}{2}$, yet, for $N = c_0 T_f\left(\delta\left(\epsilon\right)\right)$, and starting the chain from stationarity,*

$$\mathbb{P}_\pi\left(\left|\frac{1}{N}\sum_{n=1}^{N} f_\epsilon\left(X_n\right) - \frac{1}{2}\right| \geq \epsilon\right) \geq \frac{1}{3}.$$

Using this lemma, let us now prove Proposition 1. Suppose that we make the choices

$$c_0 := \left\lceil \frac{\log 7}{c_1\epsilon^2} \right\rceil \geq 1, \quad \text{and} \quad N_{c_1,\epsilon} := c_0 T_f\left(\epsilon\right),$$

in Lemma 4. Letting P_{c_0} be the corresponding Markov chain and f_ϵ the function guaranteed by the lemma, we then have

$$\mathbb{P}_\pi\left(\left|\frac{1}{N_{c_1,\epsilon}}\sum_{n=1}^{N_{c_1,\ \epsilon}} f_\epsilon\left(X_n\right) - \frac{1}{2}\right| \geq \epsilon\right) \geq \frac{1}{3} \ >\ \frac{2}{7} \ \geq\ 2\cdot\exp\left(-\frac{c_1 N_{c_1,\epsilon}\epsilon^2}{T_f\left(\delta(\epsilon)\right)}\right).$$

Proof of Lemma 4: It only remains to prove Lemma 4, which we do by constructing pathological function on a chain graph, and letting our Markov chain be the lazy random walk on this graph. For the proof, fix $\epsilon > 0$, let $\delta = \delta(\epsilon)$ and let $T_f = T_f(\delta)$. Now choose an integer $d > 0$ such that $d > 2c_0$ and let the state space be Ω be the line graph with $2d$ elements with the standard lazy random walk defining P. We then set

$$f(i) = \begin{cases} \frac{1}{2} - \delta & 1 \le i \le d, \\ \frac{1}{2} + \delta & d+1 \le i \le 2d. \end{cases}$$

It is then clear that $T_f = 1$.

Define the bad event

$$\mathcal{E} = \left\{ X_1 \in \left[0,\ \frac{d}{2}\right] \cup \left[\frac{3d}{2},\ 2d\right] \right\}.$$

When this occurs, we have

$$\left| \frac{1}{N'} \sum_{n=1}^{N'} f(X_n) - \frac{1}{2} \right| \ge \delta > \epsilon \qquad \text{with probability one,}$$

for all $N' < \frac{d}{2}$. Since $N = c_0 < \frac{d}{2}$, we can set $N' = N$.

On the other hand, under π, the probability of $\mathcal{E}$ is $\ge \frac{1}{3}$. (It is actually about $\frac{1}{2}$, but we want to ignore edge cases.) The claim follows immediately.

5.4 Proofs of confidence interval results

Here we provide the proof of the confidence interval corresponding to our bound (Theorem 2). Proofs of the claims (26b) and (30) can be found in Appendix C.

As discussed in Section 3.1, we actually prove a somewhat stronger form of Theorem 2, in order to guarantee that the confidence interval can be straightforwardly built using an upper bound $\tilde{T}_f$ on the f-mixing time rather than the true value. Setting $\tilde{T}_f = T_f$ recovers the original theorem.

Specifically, suppose $\tilde{T}_f \colon \mathbb{N} \to \mathbb{R}_+$ is an upper bound on T_f and note that the corresponding tail bound becomes $e^{-\tilde{r}_N(\epsilon)/8}$, where

$$\tilde{r}_N(\epsilon) = \epsilon^2 \left[\frac{N}{\tilde{T}_f\left(\frac{\epsilon}{2}\right)} - 1 \right].$$

This means that, just as before we wanted to make the rate r_N in equation (27) at least as large as $8 \log \frac{2}{\alpha}$, we now wish to do the same with $\tilde{r}_N$, which means choosing ϵ_N with $\tilde{r}_N(\epsilon_N) \ge 8 \log \frac{2}{\alpha}$. We therefore have the following result.

Proposition 2. *For any width $\epsilon_N \in \tilde{r}_N^{-1}\big(\big[8 \log(2/\alpha),\ \infty\big)\big)$, the set*

$$I_N^{\text{func}} = \left[\frac{1}{N - T_f\left(\frac{\epsilon}{2}\right)} \sum_{n=\tilde{T}_f\left(\frac{\epsilon}{2}\right)}^{N} f(X_n) \pm \epsilon_N \right]$$

is a $1-\alpha$ confidence interval for $\mu = \mathbb{E}_\pi[f]$.

Proof. For notational economy, let us introduce the shorthands $\tau_f(\epsilon) = T_f\big(\frac{\epsilon}{2}\big)$ and $\tilde{\tau}_f(\epsilon) = \tilde{T}_f\big(\frac{\epsilon}{2}\big)$. Theorem 1 then implies

$$\begin{aligned}\mathbb{P}\left[\frac{1}{N-\tilde{\tau}_f}\sum_{n=\tilde{\tau}_f}^{N} f\big(X_n\big) \geq \mu + \epsilon\right] &\leq \exp\big(-\frac{N-\tau_f}{4\tau_f}\cdot\epsilon^2\big)\\ &\leq \exp\big(-\frac{N-\tilde{\tau}_f}{4\tilde{\tau}_f}\cdot\epsilon^2\big)\\ &= \exp\big(-\frac{\tilde{r}_N\big(\epsilon\big)}{4}\big).\end{aligned}$$

Setting $\epsilon = \epsilon_N$ yields

$$\mathbb{P}\left[\frac{1}{N-\tilde{\tau}_f}\sum_{n=\tilde{\tau}_f}^{N} f\big(X_n\big) \geq \mu + \epsilon_N\right] \leq \frac{\alpha}{2}.$$

The corresponding lower bound leads to an analogous bound on the lower tail. □

As we did with Corollary 3, we can derive a more concrete, though slightly weaker, form of this result that is more amenable to interpretation. We derive the corollary from the specialized bound by setting $\tilde{T}_f = T_f$.

To obtain this bound, define the following lower bound, in parallel with equation (28):

$$\tilde{r}_N\big(\epsilon\big) \geq \tilde{r}_{N,\eta}\big(\epsilon\big) := \epsilon^2\big[\frac{N}{\tilde{T}_f\big(\frac{\eta}{2}\big)} - 1\big],\ \epsilon \geq \eta.$$

Since this is a lower bound, we see that whenever $\epsilon_N \geq \eta$ and $\tilde{r}_{N,\eta}\big(\epsilon_N\big) \geq 8\log\frac{2}{\alpha}$, ϵ_N is a valid half-width for a $\big(1-\alpha\big)$-confidence interval for the stationary mean centered at the empirical mean. More formally, we have the following:

Proposition 3. *Fix $\eta > 0$ and let*

$$\epsilon_N = \tilde{r}^{-1}_{N,\eta}\big(8\log\frac{2}{\alpha}\big) = 2\sqrt{2}\sqrt{\frac{\tilde{T}_f\big(\frac{\eta}{2}\big)\cdot\log\big(2/\alpha\big)}{N-\tilde{T}_f\big(\frac{\eta}{2}\big)}}.$$

If $N \geq \tilde{T}_f\big(\frac{\eta}{2}\big)$, then I_N^{func} is a $1-\alpha$ confidence interval for $\mu = \mathbb{E}_\pi\big[f\big]$.

Proof. By assumption, we have

$$\eta \leq \epsilon_N\big(\eta\big) = 2\sqrt{\frac{\tilde{T}_f\big(\frac{\eta}{2}\big)\cdot\log\big(2/\alpha\big)}{N-\tilde{T}_f\big(\frac{\eta}{2}\big)}}.$$

This implies $\tilde{T}_f\big(\frac{\epsilon_N}{2}\big) \geq \tilde{T}_f\big(\frac{\eta}{2}\big)$, which yields

$$\tilde{r}_N\big(\epsilon_N\big) = \epsilon_N^2\big[\frac{N}{\tilde{T}_f\big(\frac{\epsilon_N}{2}\big)} - 1\big] \geq \epsilon_N^2\big[\frac{N}{\tilde{T}_f\big(\frac{\eta}{2}\big)} - 1\big] = 8\log\big(2/\alpha\big).$$

But now Proposition 2 applies, so that we are done. □

5.5 Proofs of sequential testing results

In this section, we collect various proofs associated with our analysis of the sequential testing problem.

5.5.1 Proof of Theorem 3 for $\mathcal{A}_{\text{fixed}}$

We provide a detailed proof when H_1 is true, in which case we have $\mu \leq r - \delta$; the proof for the other case is analogous. When H_1 is true, we need to control the probability $\mathbb{P}\big(\mathcal{A}_{\text{fixed}}\big(X_{1:N}\big) = H_0\big)$. In order to do so, note that Theorem 1 implies that

$$\begin{aligned}
\mathbb{P}\big(\mathcal{A}_{\text{fixed}}\big(X_{1:N}\big) = H_0\big) &= \mathbb{P}\big(\frac{1}{N}\sum_{n=1}^{N} f\big(X_n\big) \geq r + \delta\big) \\
&\leq \mathbb{P}\big(\frac{1}{N}\sum_{n=1}^{N} f\big(X_n\big) \geq \mu + 2\delta\big) \\
&\leq \exp\big(-\frac{\delta^2 N}{2T_f\big(\delta\big)}\big).
\end{aligned}$$

Setting $N = \frac{2T_f(\delta)\log\left(\frac{1}{\alpha}\right)}{\delta^2}$ yields the bound $\mathbb{P}\big(\mathcal{A}_{\text{fixed}}\big(X_{1:N}\big) = H_0\big) \leq \alpha$, as claimed.

5.5.2 Proof of Theorem 3 for $\mathcal{A}_{\text{seq}}$

The proof is nearly identical to that given by [15], with $T_f\big(\delta/2\big)$ replacing $\frac{1}{\gamma_0}$. We again assume that H_1 holds, so $\mu \leq r - \delta$. In this case, it is certainly true that

$$\begin{aligned}
\text{err}\big(\mathcal{A}_{\text{seq}},\ f\big) &= \mathbb{P}\big(\exists k\colon \mathcal{A}_{\text{seq}}\big(X_{1:N_k}\big) = H_0\big) \\
&= \mathbb{P}\big(\exists k\colon \frac{1}{N_k}\sum_{n=1}^{N_k} f\big(X_n\big) \geq r + \frac{M}{N_k}\big) \\
&\leq \sum_{k=1}^{\infty} \mathbb{P}\big(\frac{1}{N_k}\sum_{n=1}^{N_k} f\big(X_n\big) \geq r + \frac{M}{N_k}\big).
\end{aligned}$$

It follows by Theorem 1, with $\epsilon_k = \delta + \frac{M}{N_k}$, that

$$\begin{aligned}
\mathbb{P}\big(\frac{1}{N_k}\sum_{n=1}^{N_k} f\big(X_n\big) \geq r + \frac{M}{N_k}\big) &\leq \mathbb{P}\big(\frac{1}{N_k}\sum_{n=1}^{N_k} f\big(X_n\big) \geq \mu + \delta + \frac{M}{N_k}\big) \\
&\leq \exp\big(-\frac{\epsilon_k^2 N_k}{8T_f\big(\frac{\epsilon_k}{2}\big)}\big) \\
&\leq \exp\big(-\frac{\epsilon_k^2 N_k}{8T_f\big(\frac{\delta}{2}\big)}\big).
\end{aligned}$$

In order to simplify notation, for the remainder of the proof, we define $\tau := 4T_f(\delta/2)$, $\beta := \frac{\sqrt{\alpha\xi}}{2}$,

and $\zeta_k := \frac{\delta^2 N_k}{2\tau \log(1/\beta)}$. In terms of this notation, we have $M = \frac{2\tau \log(1/\beta)}{\delta}$, and hence that

$$\begin{aligned}\exp\big(-\frac{\epsilon_k^2 N_k}{2\tau}\big) &= \exp\big(-\frac{1}{2\tau}\cdot\big(\delta^2 N_k + 2\delta M + \frac{M^2}{N_k}\big)\big)\\ &= \exp\big(-\big[\frac{\delta^2 N_k}{2\tau} + \log\big(1/\beta\big) + \frac{2\tau\log^2\big(1/\beta\big)}{\delta^2 N_k}\big]\big)\\ &= \exp\big(-\log\big(1/\beta\big)\big[1+\zeta_k+\zeta_k^{-1}\big]\big)\\ &= \beta\cdot\exp\big(-\log\big(1/\beta\big)\big[\zeta_k+\zeta_k^{-1}\big]\big).\end{aligned}$$

It follows that the error probability is at most

$$\beta\sum_{k=1}^{\infty}\exp\big(-\log\big(1/\beta\big)\big[\zeta_k+\zeta_k^{-1}\big]\big).$$

We now finish the proof using two small technical lemmas, whose proofs we defer to Appendix D.

Lemma 5. *In the above notation, we have*

$$\sum_{k=1}^{\infty}\exp\Big\{-\log(1/\beta)\big[\zeta_k+\zeta_k^{-1}\big]\Big\} \le 4\sum_{\ell=0}^{\infty}\exp\Big\{-\log(1/\beta)\Big[\big(1+\xi\big)^{\ell}+\big(1+\xi\big)^{-\ell}\Big]\Big\}.$$

Lemma 6. *For any integer $c \ge 0$, we have*

$$(1+\xi)^{\ell}+(1+\xi)^{-\ell} \ge 2(c+1) \quad \textit{for all } \ell \in \Big[\tfrac{9c}{5\xi}, \tfrac{9(c+1)}{5\xi}\Big).$$

Using this bound, and grouping together terms in blocks of size $\frac{9}{5\xi}$, we find that the error is at most

$$4\sum_{\ell=0}^{\infty}\exp\big(-\log\big(1/\beta\big)\big[\big(1+\xi\big)^{\ell}+\big(1+\xi\big)^{-\ell}\big]\big) \le \frac{36}{5\xi}\cdot\sum_{c=0}^{\infty}\beta^{2(c+1)}.$$

Since both α and ξ are at most $\frac{2}{5}$, we have $\beta = \frac{\sqrt{\alpha\xi}}{2} \le \frac{1}{5}$, and hence the error probability is bounded as

$$\frac{36\beta}{5\xi}\sum_{c=0}^{\infty}\beta^{2(c+1)} \le \frac{36\beta^3}{5\xi(1-\beta^2)} \le \frac{36\beta^2}{25\xi(1-\beta^2)} \le \frac{3\beta^2}{2\xi} = \frac{3\alpha}{4} < \alpha.$$

5.5.3 Proof of Theorem 3 for $\mathcal{A}_{\text{hard}}$

We may assume that H_1 holds, as the other case is analogous. Under H_1, letting k_0 be the smallest k such that $\epsilon_k < \infty$, we have

$$\mathrm{err}(\mathcal{A}_{\text{hard}},\ f) \le \sum_{k=k_0}^{\infty}\mathbb{P}\left(\hat{\mu}_{N_k} \ge r+\epsilon_k\right) \ \le\ \sum_{k=k_0}^{\infty}\mathbb{P}(\hat{\mu}_{N_k} \ge \mu + 2\epsilon_k).$$

By Theorem 1, and the definition of ϵ_k, we thus have

$$\begin{aligned}\mathrm{err}\left(\mathcal{A}_{\text{hard}},\ f\right) \le \sum_{k=k_0}^{\infty}\exp\left(-\frac{N_k\epsilon_k^2}{8T_f\left(\frac{\epsilon_k}{2}\right)}\right) &\le \frac{\alpha}{2}\sum_{k=k_0}^{\infty}\frac{1}{k^2}\\ &= \frac{\pi^2}{12}\,\alpha \ <\ \alpha,\end{aligned}$$

as claimed.

5.5.4 Proof of Theorem 4 for $\mathcal{A}_{\text{seq}}$

We may assume H_1 holds; the other case is analogous. Note that

$$\begin{aligned}
\mathbb{E}[N] &\leq N_1 + \sum_{k=1}^{\infty} (N_{k+1} - N_k)\mathbb{P}(N > N_k) \\
&\leq N_1 + \sum_{k=1}^{\infty} (N_{k+1} - N_k)\mathbb{P}\Big(\frac{1}{N_k}\sum_{n=1}^{N_k} f(X_n) \in \big(r - \frac{M}{N_k},\ r + \frac{M}{N_k}\big)\Big) \\
&\leq N_1 + \sum_{k=1}^{\infty} (N_{k+1} - N_k)\mathbb{P}\Big(\frac{1}{N_k}\sum_{n=1}^{N_k} f(X_n) > r - \frac{M}{N_k}\Big) \\
&= N_1 + \sum_{k=1}^{\infty} (N_{k+1} - N_k)\mathbb{P}\Big(\frac{1}{N_k}\sum_{n=1}^{N_k} f(X_n) > \mu + \Delta - \frac{M}{N_k}\Big) \\
&\leq N_1 + \sum_{k=1}^{\infty} (N_{k+1} - N_k) \exp\left\{-\frac{(\Delta N_k - M)_+^2}{8T_f(\delta/2)N_k}\right\}.
\end{aligned}$$

Our proof depends on the following simple technical lemma, whose proof we defer to Appendix D.3.

Lemma 7. *Under the conditions of Theorem 4, we have*

$$\sum_{k=1}^{\infty} (N_{k+1} - N_k) \exp\left\{-\frac{(\Delta N_k - M)_+^2}{8T_f(\delta/2)N_k}\right\} \leq (1+\xi)\big[1 + \int_{N_1}^{\infty} h(s)\mathrm{d}s\big], \tag{43}$$

where $h(s) := \exp\Big\{-\frac{(\Delta s - M)_+^2}{8T_f(\delta/2)s}\Big\}$.

Given this lemma, we then follow the argument of Gyori and Paulin [15] in order to bound the integral. We have

$$\int_{N_1}^{\infty} h(s)\mathrm{d}s \leq \frac{4}{\Delta}\sqrt{\frac{2T_f(\delta/2)M}{\Delta} + 8T_f(\delta/2)}.$$

To conclude, note that either $r \geq \Delta$ or $1 - r \geq \Delta$, since $0 < \mu < 1$, so that $\min\big(\frac{1}{r},\ \frac{1}{1-r}\big) \leq \frac{1}{\Delta}$. It follows that

$$N_1 \leq (1+\xi)N_0 \leq \frac{(1+\xi)M}{\Delta}.$$

Combining the bounds yields the desired result.

5.5.5 Proof of Theorem 4 for $\mathcal{A}_{\text{hard}}$

For concreteness, we may assume H_1 holds, as the H_0 case is symmetric. We now have that

$$\mathbb{P}\left[N \geq N_k\right] \leq \mathbb{P}\left[\Big|\frac{1}{N_k}\sum_{n=1}^{N_k} f(X_n) - r\Big| \leq \epsilon_k\right] \ \leq\ \mathbb{P}\left[\frac{1}{N_k}\sum_{n=1}^{N_k} f(X_n) \geq \mu + \Delta - \epsilon_k\right].$$

For convenience, let us introduce the shorthand

$$T_{f,k}^{+} := \begin{cases} T_f\big(\frac{\Delta - \epsilon_k}{2}\big) & \text{if } \epsilon_k \leq \Delta, \\ 1 & \text{otherwise.} \end{cases}$$

Applying the Hoeffding bound from Theorem 1, we then find that

$$\mathbb{P}\left[N \geq N_k\right] \leq \exp\left\{-\frac{N_k}{8T_{f,k}^+} \cdot \left(\Delta - \epsilon_k\right)_+^2\right\}.$$

Observe further that

$$\begin{aligned}
\mathbb{E}[N] &= N_1 + \sum_{k=1}^{\infty} \left(N_{k+1} - N_k\right)\mathbb{P}\left(N > N_k\right) \\
&\leq N_{k_0^*+1} + \sum_{k=k_0^*+1}^{\infty} \left(N_{k+1} - N_k\right)\mathbb{P}\left(N > N_k\right) \\
&\leq \left(1+\xi\right)\left(N_0^* + 1\right) + \sum_{k=k_0^*+1}^{\infty} \left(N_{k+1} - N_k\right)\mathbb{P}\left(N > N_k\right).
\end{aligned}$$

Combining the pieces yields

$$\mathbb{E}[N] \leq \left(1+\xi\right)\left(N_0^* + 1\right) + \sum_{k=k_0^*+1}^{\infty} \left(N_{k+1} - N_k\right) \exp\left(-\frac{N_k}{8T_{f,k}^+} \cdot \left(\Delta - \epsilon_k\right)_+^2\right). \tag{44}$$

The crux of the proof is a bound on the infinite sum, which we pull out as a lemma for clarity.

Lemma 8. *The infinite sum (44) is upper bounded by*

$$\sum_{k=k_0^*+1}^{\infty} \left(N_{k+1} - N_k\right) \exp\left(-\frac{N_k}{8T_{f,k}^+} \cdot \left(\Delta - \epsilon_k\right)_+^2\right) \leq \alpha \cdot \sum_{m=1}^{\infty} \exp\left(-m \cdot \frac{\Delta^2}{32 T_f\left(\frac{\Delta}{4}\right)}\right).$$

See Appendix D.4 for the proof of this claim.

Lemma 8 then implies that

$$\begin{aligned}
\sum_{k=k_0^*+1}^{\infty} \left(N_{k+1} - N_k\right) \exp\left(-\frac{N_k}{T_f\left(\frac{\Delta}{4}\right)} \cdot \frac{\Delta^2}{32}\right) &\leq \alpha \cdot \sum_{m=1}^{\infty} \exp\left(-m \cdot \frac{\Delta^2}{32 T_f\left(\frac{\Delta}{4}\right)}\right) \\
&= \frac{\alpha \exp\left(-\frac{\Delta^2}{32 T_f\left(\frac{\Delta}{4}\right)}\right)}{1 - \exp\left(\frac{\Delta^2}{32 T_f\left(\frac{\Delta}{4}\right)}\right)} \\
&\leq \frac{32 \alpha T_f\left(\frac{\Delta}{4}\right)}{\Delta^2}.
\end{aligned}$$

The claim now follows from equation (44).

6 Discussion

A significant obstacle to successful application of statistical procedures based on Markov chains—especially MCMC—is the possibility of slow mixing. The usual formulation of mixing is in terms of convergence in a distribution-level metric, such as the total variation or Wasserstein distance. On the other hand, algorithms like MCMC are often used to estimate equilibrium expectations over a limited class of functions. For such uses, it is desirable to build a theory of mixing times

with respect to these limited classes of functions and to provide convergence and concentration guarantees analogous to those available in the classical setting, and our paper has made some steps in this direction.

In particular, we introduced the f-mixing time of a function, and showed that it can be characterized by the interaction between the function and the eigenspaces of the transition operator. Using these tools, we proved that the empirical averages of a function f concentrate around their equilibrium values at a rate characterized by the f-mixing time; in so doing, we replaced the worst-case dependence on the spectral gap of the chain, characteristic of previous Markov chain concentration bounds, by an adaptive dependence on the properties of the actual target function. Our methodology yields sharper confidence intervals, as well as better rates for sequential hypothesis tests in MCMC, and we have provided evidence that the predictions made by our theory are accurate in some real examples of MCMC and thus potentially of significant practical relevance.

Our investigation also suggests a number of further questions. Two important ones concern the continuous and non-reversible cases. Both arise frequently in statistical applications—for example, when sampling continuous parameters or when performing Gibbs sampling with systematic scan—and are therefore of considerable interest. As uniform Hoeffding bounds do exist for the continuous case and, more recently, have been established for the non-reversible case, we believe many of our conclusions should carry over to these settings, although somewhat different methods of analysis may be required.

Furthermore, in practical applications, it would be desirable to have methods for estimating or bounding the f-mixing time based on samples. It would also be interesting to study the f-mixing times of Markov chains that arise in probability theory, statistical physics, and applied statistics itself. While we have shown what can be done with spectral methods, the classical theory provides a much larger arsenal of techniques, some of which may generalize to yield sharper f-mixing time bounds. We leave these and other problems to future work.

Acknowledgments

The authors thank Alan Sly and Roberto Oliveira for helpful discussions about the lower bounds and the sharp function-specific Hoeffding bounds (respectively). This work was partially supported by NSF grant CIF-31712-23800, ONR-MURI grant DOD 002888, and AFOSR grant FA9550-14-1-0016. In addition, MR is supported by an NSF Graduate Research Fellowship and a Fannie and John Hertz Foundation Google Fellowship.

References

[1] David Aldous and Persi Diaconis. Shuffling cards and stopping times. *American Mathematical Monthly*, 93(5):333–348, 1986.

[2] Thomas R Belin and Donald B Rubin. The analysis of repeated-measures data on schizophrenic reaction times using mixture models. *Statistics in Medicine*, 14(8):747–768, 1995.

[3] Scott M Berry, Bradley P Carlin, J Jack Lee, and Peter Muller. *Bayesian Adaptive Methods for Clinical Trials*. CRC press, 2010.

[4] Kai-Min Chung, Henry Lam, Zhenming Liu, and Michael Mitzenmacher. Chernoff-hoeffding bounds for markov chains: Generalized and simplified. In *29th International Symposium on Theoretical Aspects of Computer Science, STACS 2012, February 29th - March 3rd, 2012,*

Paris, France, pages 124–135, 2012. doi: 10.4230/LIPIcs.STACS.2012.124. URL http://dx.doi.org/10.4230/LIPIcs.STACS.2012.124.

[5] Mark Conger and Divakar Viswanath. Riffle shuffles of decks with repeated cards. *The Annals of Probability*, pages 804–819, 2006.

[6] Persi Diaconis and James Allen Fill. Strong stationary times via a new form of duality. *Ann. Probab.*, 18(4):1483–1522, 10 1990. doi: 10.1214/aop/1176990628. URL http://dx.doi.org/10.1214/aop/1176990628.

[7] Persi Diaconis and Bob Hough. Random walk on unipotent matrix groups. *arXiv preprint arXiv:1512.06304*, 2015.

[8] James M Flegal, Murali Haran, and Galin L Jones. Markov chain monte carlo: Can we trust the third significant figure? *Statistical Science*, 23(2):250–260, 2008.

[9] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis, Third Edition*. Chapman and Hall/CRC, 2013.

[10] Zoubin Ghahramani and Thomas L Griffiths. Infinite latent feature models and the indian buffet process. In *Advances in Neural Information Processing Systems*, pages 475–482, 2005.

[11] David Gillman. A chernoff bound for random walks on expander graphs. *SIAM Journal on Computing*, 27(4):1203–1220, 1998.

[12] Peter W Glynn and Eunji Lim. Asymptotic validity of batch means steady-state confidence intervals. In *Advancing the Frontiers of Simulation*, pages 87–104. Springer, 2009.

[13] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235, 2004.

[14] Benjamin M Gyori and Daniel Paulin. Non-asymptotic confidence intervals for mcmc in practice. *arXiv preprint arXiv:1212.2016*, 2012.

[15] Benjamin M. Gyori and Daniel Paulin. Hypothesis testing for markov chain monte carlo. *Statistics and Computing*, pages 1–12, July 2015. ISSN 0960-3174. doi: 10.1007/s11222-015-9594-1. URL http://dx.doi.org/10.1007/s11222-015-9594-1.

[16] Sonia Jain, Radford M Neal, et al. Splitting and merging components of a nonconjugate dirichlet process mixture model. *Bayesian Analysis*, 2(3):445–472, 2007.

[17] Galin L Jones and James P Hobert. Honest exploration of intractable probability distributions via markov chain monte carlo. *Statistical Science*, 16(4):312–334, 2001.

[18] Aldéric Joulin, Yann Ollivier, et al. Curvature, concentration and error estimates for markov chain monte carlo. *The Annals of Probability*, 38(6):2418–2442, 2010.

[19] Aryeh Kontorovich, Roi Weiss, et al. Uniform chernoff and dvoretzky-kiefer-wolfowitz-type inequalities for markov chains and related processes. *Journal of Applied Probability*, 51(4): 1100–1113, 2014.

[20] Carlos A. Léon and François Perron. Optimal hoeffding bounds for discrete reversible markov chains. *Ann. Appl. Probab.*, 14(2):958–970, 05 2004. doi: 10.1214/105051604000000170.

[21] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov Chains and Mixing Times*. American Mathematical Society, 2008.

[22] Pascal Lezaud. Chernoff and berry–esséen inequalities for markov processes. *ESAIM: Probability and Statistics*, 5:183–201, 2001.

[23] Sean P Meyn and Richard L Tweedie. *Markov Chains and Stochastic Stability*. Springer Science & Business Media, 2012.

[24] David M. Mimno, Matthew D. Hoffman, and David M. Blei. Sparse stochastic inference for latent dirichlet allocation. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*, 2012. URL http://icml.cc/discuss/2012/784.html.

[25] Radford M Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265, 2000.

[26] Yann Ollivier. Ricci curvature of markov chains on metric spaces. *Journal of Functional Analysis*, 256(3):810–864, 2009.

[27] Daniel Paulin. Concentration inequalities for markov chains by marton couplings and spectral methods. *arXiv preprint arXiv:1212.2015*, 2012.

[28] Christian P. Robert and George Casella. *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005. ISBN 0387212396.

[29] Paul-Marie Samson et al. Concentration of measure inequalities for Markov chains and ϕ-mixing processes. *The Annals of Probability*, 28(1):416–461, 2000.

[30] Alistair Sinclair. Improved bounds for mixing rates of markov chains and multicommodity flow. *Combinatorics, Probability, and Computing*, 1(04):351–370, 1992.

A Proofs for Section 2.2

In this section, we gather the proofs of the mixing time bounds from Section 2.2, namely equations (10) and (12) and Lemmas 1 and 2.

A.1 Proof of the bound (10)

Recall that

$$d_f(p,\ q) = \sup_{f\colon [d]\to[0,\ 1]} \left|\mathbb{E}_p\big[f\big(X\big)\big] - \mathbb{E}_q\big[f(Y)\big]\right|.$$

It follows from equation (12) that

$$\begin{aligned} d_{\mathrm{TV}}\big(\pi_n,\ \pi\big) &= \sup_{f\colon [d]\to[0,\ 1]} d_f\big(\pi_n,\ \pi\big) \\ &\leq \sup_{f\colon [d]\to[0,\ 1]} \big[\frac{\lambda_f^n}{\sqrt{\pi_{\min}}} \cdot d_f\big(\pi_0,\ \pi\big)\big] \\ &= \frac{1}{\sqrt{\pi_{\min}}} \cdot \lambda_*^n \cdot d_{\mathrm{TV}}\big(\pi_0,\ \pi\big), \end{aligned}$$

as claimed.

A.2 Proof of equation (12)

Let $D = \mathrm{diag}(\sqrt{\pi})$. Then the matrix $A = DPD^{-1}$ is symmetric and so has an eigendecomposition of the form $A = \gamma_1\gamma_1^T + \sum_{j=2}^d \lambda_j\gamma_j\gamma_j^T$. Using this decomposition, we have

$$P = \mathbf{1}\pi^T + \sum_{j=2}^d \lambda_j h_j q_j^T,$$

where $h_j := D^{-1}\gamma_j$ and $q_j := D\gamma_j$. Note that the vectors $\{q_j\}_{j=2}^d$ correspond to the left eigenvectors associated with the eigenvalues $\{\lambda_j\}_{j=2}^d$.

Now, if we let π_0 be an arbitrary distribution over $[d]$, we have

$$d_f(\pi_n,\ \pi) = \left|\pi_0^T P^n f - \pi^T P^n f\right| \ \le\ \left|(\pi_0 - \pi)^T P^n f\right|.$$

Defining $P_f := \mathbf{1}\pi^T + \sum_{j\in J_f} \lambda_j h_j q_j^T$, we have $P^n f = P_f^n f$. Moreover, if we define $\tilde{P}_f := \sum_{j\in J_f} \lambda_j h_j q_j^T$, and correspondingly $\tilde{A}_f := D\tilde{P}_f D^{-1}$, we then have the relation $\big(\pi_0 - \pi\big)^T \tilde{P}_f = \big(\pi_0 - \pi\big)^T P_f$. Consequently, by the definition of the operator norm and sub-multiplicativity, we have

$$\begin{aligned} d_f\big(\pi_n,\ \pi\big) &\le \left|\big(\pi_0 - \pi\big)^T \tilde{P}_f^n f\right| \\ &\le |\!|\!| \tilde{A}_f |\!|\!|_{\mathrm{op}}^n \|Df\|_2 \big\|D^{-1}\big(\pi_0 - \pi\big)\big\|_2 \\ &= \sqrt{\mathbb{E}_\pi\big[f^2\big] \cdot \sum_{i\in[d]} \frac{\big(\pi_{0,i} - \pi_i\big)^2}{\pi_i}} \cdot \lambda_f^n d_f\big(\pi_0,\ \pi\big). \end{aligned}$$

In order to complete the proof, let $Z \in \{0,\ 1\}^d$ denote the indicator vector $Z_j = \mathbf{1}\big(X_0 = j\big)$. Observe that the function

$$r(z) := \sum_{i\in[d]} \frac{\big(z_i - \pi_i\big)^2}{\pi_i}$$

is convex in terms of z. Thus, Jensen's inequality implies that

$$\mathbb{E}_{\pi_0}\big[r(Z)\big] \ge r\big(\mathbb{E}_{\pi_0}\big[Z\big]\big) = r\big(\pi_0\big) = \sum_{i\in[d]} \frac{\big(\pi_{0,i} - \pi_i\big)^2}{\pi_i}.$$

On the other hand, for any fixed value $X_0 = j$, corresponding to $Z = e_j$, we have

$$r\big(z\big) = r\big(e_j\big) = \frac{\big(1 - \pi_j\big)^2}{\pi_j} + \sum_{i\neq j} \pi_i = \frac{1 - \pi_j}{\pi_j} \le \frac{1}{\pi_{\min}}.$$

We deduce that $d_f(\pi_n,\ \pi) \le \sqrt{\frac{\mathbb{E}_\pi[f^2]}{\pi_{\min}}} \cdot \lambda_f^n \cdot d_f(\pi_0,\ \pi)$, as claimed.

A.3 Proof of Lemma 1

We observe that

$$\begin{aligned} \left|(\pi_0 - \pi)^T h_J(n)\right| &\le \|\pi_0 - \pi\|_1\ \|h_J(n)\|_\infty \\ &= 2d_{\mathrm{TV}}\big(\pi_0,\ \pi\big)\ \|h_J(n)\|_\infty \\ &\le 2d_{\mathrm{TV}}\big(\pi_0,\ \pi\big)\ \Big\{\sum_{j\in J} \big|\lambda_j\big|^n \cdot \big|q_j^T f\big| \cdot \|h_j\|_\infty\Big\} \\ &\le 2d_{\mathrm{TV}}\big(\pi_0,\ \pi\big)\ \Big\{2\big|J\big| \cdot \max_{j\in J} \big|q_j^T f\big| \cdot \max_{j\in J} \|h_j\|_\infty\Big\}, \end{aligned}$$

as claimed.

A.4 Proof of Lemma 2

We proceed in a similar fashion as in the proof of equation (12). Begin with the identity proved there, viz.

$$d_f(\pi_n,\ \pi) = \left|(\pi_0 - \pi)^T \tilde{P}_f^n f\right|,$$

where $\tilde{P}_f = \sum_{j \in J_f} \lambda_j h_j q_j^T$. Now decompose $\tilde{P}_f$ further into

$$P_J = \sum_{j \in J} \lambda_j h_j q_j^T \quad \text{and} \quad P_{-J} = \sum_{j \in J_f \setminus J} \lambda_j h_j q_j^T.$$

Note also that $\tilde{P}_f^n = P_J^n + P_{-J}^n$. We thus find that

$$d_f(\pi_n,\ \pi) \leq \left|(\pi_0 - \pi)^T P_J^n f\right| + \left|(\pi_0 - \pi)^T P_{-J}^n f\right|.$$

Now observe that $P_J^n f = h_J(n)$, so $\left|(\pi_0 - \pi)^T P_J^n f\right| = \left|(\pi_0 - \pi)^T h_J(n)\right|$. On the other hand, the second term can be bounded using the argument from the proof of equation (12) to obtain

$$\left|(\pi_0 - \pi)^T P_{-J}^n f\right| \leq \sqrt{\frac{\mathbb{E}_\pi\left[f^2\right]}{\pi_{\min}}} \cdot \lambda_{-J_\delta}^n \cdot d_f(\pi_0,\ \pi),$$

as claimed.

B Proofs for Section 2.4

In this section, we provide detailed proofs of the bound (24), as well as the other claims about the random function example on C_{2d}.

Proposition 4. *Let $f\colon [d] \to [0,\ 1]$ with $f(i) \sim \tau$ iid from some distribution on $[0,\ 1]$. There exists a universal constant $c_0 > 0$ such that with probability $\geq 1 - \frac{\delta^*}{128\sqrt{d \log d}}$ over the randomness f, we have*

$$T_f(\delta) \leq \frac{c_0 d \log d \log \frac{128 d}{\delta}}{\delta^2} \qquad \text{for all } 0 < \delta \leq \delta^*.$$

Proof. We proceed by defining a "good event" $\mathcal{E}_\delta$, and then showing that the stated bound on $T_f(\delta)$ holds conditioned on this event. The final step is to show that $\mathbb{P}[\mathcal{E}_\delta]$ is suitably close to one, as claimed.

The event $\mathcal{E}_\delta$ is defined in terms of the interaction between f and the eigenspaces of P corresponding to eigenvalues close to 1. More precisely, denote the indices of these eigenvalues by

$$J_\delta := \left\{ j \in \{1, \ldots, 2d-1\} \ \mid \ j \leq 4\delta \sqrt{\frac{d}{\log d}} \ \text{ or } \ j \geq 2d - 4\delta \sqrt{\frac{d}{\log d}} \right\}.$$

The good event $\mathcal{E}_\delta$ occurs when f has small inner product with all the corresponding eigenfunctions—that is

$$\mathcal{E}_\delta := \left\{ \max_{j \in J_\delta} |q_j^T f| \leq 2\sqrt{\frac{10 \log d}{d}} \right\}.$$

Viewed as family of events indexed by δ, these events form a decreasing sequence. (In particular, the associated sequence of sets J_δ is increasing in δ, in that whenever $\delta \leq \delta^*$, we are guaranteed that $J_\delta \subset J_{\delta^*}$.)

Establishing the bound conditionally on $\mathcal{E}_\delta$: We now exploit the spectral properties of the transition matrix to bound T_f conditionally on the event $\mathcal{E}_\delta$. Recall that the lazy random walk on C_{2d} has eigenvalues $\lambda_j = \frac{1}{2}\big(1+\cos(\frac{\pi j}{d})\big)$ for $j \in [d]$, with corresponding unit eigenvectors

$$v_j^T = \frac{1}{\sqrt{2d}}\begin{pmatrix} 1 & \omega_j & \cdots & \omega_j^{2d-1}\end{pmatrix}, \quad \omega_j := e^{\frac{\pi i j}{d}}.$$

(See [21] for details.) We note that this diagonalization allows us to write $P = \mathbf{1}\pi^T + \sum_{j=1}^{2d-1}\lambda_j h_j q_j^T$, where $h_j = \sqrt{2d}\cdot v_j$ and $q_j = \frac{v_j}{\sqrt{2d}}$, where we have used the fact that $\operatorname{diag}\big(\sqrt{\pi}\big) = \frac{1}{\sqrt{2d}}\cdot I$. Note that $\|h_j\|_\infty = 1$.

Combining Lemma 1 with the bounds $\lambda_{J_\delta} \le 1$, $\|h_j\|_\infty \le 1$, and $\big|J_\delta\big| \le 8\delta\sqrt{\frac{d}{\log d}}$, we find that

$$d_f(\pi_n,\ \pi) \le 16\delta\sqrt{\frac{d}{\log d}}\cdot\max_{j\in J}\big|q_j^T f\big| + \sqrt{d}\cdot\lambda_{-J_\delta}^n.$$

Therefore, when the event $\mathcal{E}_\delta$ holds, we have

$$d_f(\pi_n,\ \pi) \le 32\sqrt{10}\cdot\delta + \sqrt{d}\cdot\lambda_{-J_\delta}^n. \tag{45}$$

In order to conclude the argument, we use the fact that

$$\lambda_{-J_\delta} = \frac{1+\max_{j\in J_f\setminus J_\delta}\cos\big(\frac{\pi j}{d}\big)}{2} \le \frac{1+\cos\big(\frac{\pi j_0}{d}\big)}{2},$$

where $j_0 = 4\delta\sqrt{\frac{d}{\log d}}$. On the other hand, we also have

$$\cos\big(\pi x\big) \le 1 - \frac{\pi^2 x^2}{2} + \frac{\pi^4 x^4}{24} \le 1 - \frac{\pi^2 x^2}{12}, \qquad \text{for all } |x| \le 1,$$

which implies that

$$\lambda_{-J_\delta} \le 1 - \frac{2\pi^2\delta^2}{3d\log d} \le \exp\big(-\frac{2\pi^2\delta^2}{3d\log d}\big).$$

Together with equation (45), this bound implies that for $n \ge \frac{3d\log d\log\frac{d}{\delta}}{2\pi^2\delta^2}$, we have $\sqrt{d}\lambda_{-J_\delta}^n \le \delta$, whence

$$d_f(\pi_n,\ \pi) \le (32\sqrt{10}+1)\,\delta \le 128\delta.$$

Replacing δ by $\frac{\delta}{128}$ throughout, we conclude that for

$$n \ge \frac{3\,(128)^2 d\log d\log\frac{128d}{\delta}}{2\pi^2\delta^2} = \frac{3\cdot 2^{13}}{\pi^2}\cdot\frac{d\log d\log\frac{128d}{\delta}}{\delta^2},$$

we have $d_f(\pi_n,\ \pi) \le \delta$ with probability at least $\mathbb{P}\big(\mathcal{E}_{\delta/128}\big)$.

Controlling the probability of $\mathcal{E}_\delta$: It now suffices to prove $\mathbb{P}(\mathcal{E}_\delta) \geq 1 - \frac{\delta}{\sqrt{d \log d}}$, since this implies that $\mathbb{P}\left(\mathcal{E}_{\delta/128}\right) \geq 1 - \frac{\delta}{128\sqrt{d \log d}}$, as required. In order to do so, observe that the vectors $\{q_j\}_{j=1}^d$ are rescaled versions of an orthonormal collection of eigenvectors, and hence

$$\mathbb{E}\big[q_j^T f\big] = \mathbb{E}_\nu\left[\mu\right] \cdot q_j^T \mathbf{1} = 0.$$

We can write the inner product as $q_j^T f = A_j + iB_j$, where (A_j, B_j) are a pair of real numbers. The triangle inequality then guarantees that $|q_j^T f| \leq |A_j| + |B_j|$, so that it suffices to control these two absolute values.

By definition, we have

$$A_j = \frac{1}{2d} \sum_{\ell=0}^{2d-1} f\big(\ell\big) \cdot \cos\big(\frac{\pi j \ell}{d}\big),$$

showing that it is the sum of sub-Gaussian random variables with parameters $\sigma_{\ell,j}^2 = \cos^2\left(\frac{\pi j \ell}{d}\right) \leq 1$. Thus, the variable A_j is sub-Gaussian with parameter at most $\sigma_j^2 \leq \frac{1}{2d}$. A parallel argument applies to the scalar B_j, showing that it is also sub-Gaussian with parameter at most σ_j^2.

By the triangle inequality, we have $|q_j^T f| \leq |A_j| + |B_j|$, so it suffices to bound $\big|A_j\big|$ and $|B_j|$ separately. In order to do so, we use sub-Gaussianity to obtain

$$\mathbb{P}\big(\max_{j \in J} |A_j| \geq r\big) \leq |J| \cdot e^{-\frac{r^2}{2}} \leq 8\delta \sqrt{\frac{d}{\log d}} \cdot e^{-\frac{dr^2}{2}}.$$

With $r := \sqrt{\frac{2 \log 16d}{d}}$, we have

$$\mathbb{P}\big(\max_{j \in J_\delta} \big|A_j\big| \geq \sqrt{\frac{2 \log 16d}{d}}\big) \leq \frac{\delta}{2\sqrt{d \log d}}.$$

Applying a similar argument to B_j and taking a union bound, we find that

$$\mathbb{P}\big(\max_{j \in J_\delta} \big|q_j^T f\big| \geq 2\sqrt{\frac{2 \log 16d}{d}}\big) \leq \frac{\delta}{\sqrt{d \log d}}.$$

Since $2\sqrt{\frac{2 \log 16d}{d}} \leq 2\sqrt{\frac{10 \log d}{d}}$ for $d \geq 2$, we deduce that

$$1 - \mathbb{P}\left(\mathcal{E}_\delta\right) = \mathbb{P}\left(\max_{j \in J_\delta} \big|q_j^T f\big| \geq 2\sqrt{\frac{10 \log d}{d}}\right) \leq \frac{\delta}{\sqrt{d \log d}},$$

as required. □

The concentration result now follows.

Proposition 5. *The random function f on C_{2d} defined in equation (21) satisfies the mixing time and tail bounds*

$$T_f\big(\frac{\epsilon}{2}\big) \leq \frac{c_0 d \log d\big[\log d + \log\left(\frac{1}{\epsilon^2}\right)\big]}{\epsilon^2},$$

and

$$\mathbb{P}\left[\frac{1}{N}\sum_{n=T_f(\epsilon/2)}^{N+T_f(\epsilon/2)} f(X_n) \geq \mu + \epsilon\right] \leq \exp\big(-\frac{c_1\epsilon^4 N}{d\log d\big[\log(\frac{1}{\epsilon}) + \log d\big]}\big).$$

with probability at least $1 - \frac{c_2\epsilon^2}{\sqrt{d\log d}}$ *over the randomness of* f *provided* $\epsilon \geq c_3\big(\frac{\log d}{d}\big)^{1/2}$, *where* $c_0, c_1, c_2, c_3 > 0$ *are universal constants.*

Proof. We first note that from the proof of Proposition 4, we have the lower bound $1 - \lambda_{-J_\delta} \geq \frac{c_4\delta^2}{d\log d}$, valid for all $\delta \in (0,1)$. The proof of the previous proposition guarantees that $\Delta_J^* \leq 32\sqrt{10}\delta$, so setting $\delta = \frac{\epsilon}{128\sqrt{10}}$ yields

$$\frac{\epsilon}{4} = 32\sqrt{10}\delta \geq \Delta_J^*, \quad \text{and} \quad 1 - \lambda_{-J_\delta} \geq \frac{c_4'\epsilon^2}{d\log d}.$$

Now, by Proposition 4, there is a universal constant $c_5 > 0$ such that, with probability at least $1 - \frac{\delta}{128\sqrt{d\log d}}$, we have

$$T_f(\delta') \leq \frac{c_5 d\log d\log d/\delta'}{(\delta')^2} \qquad \text{for all } \delta' \geq \delta.$$

In particular, we have

$$T_f\big(\frac{\epsilon}{2}\big) \leq \frac{c_2' d\log d\log d/\epsilon}{\epsilon^2}$$

with this same probability. Thus, we have this bound on T_f with the high probability claimed in the statement of the proposition.

We now finish by taking $\Delta = \frac{\epsilon}{4}$ in Corollary 2. Noting that $\Delta_J + \Delta = \frac{\epsilon}{2}$ and $1 - \lambda_{-J} \geq \frac{c_4'\epsilon^2}{d\log d}$ completes the proof. □

C Proofs for Section 3.1

We now prove correctness of the confidence intervals based on the uniform Hoeffding bound (5), and the Berry-Esseen bound (29).

C.1 Proof of claim (26b)

This claim follows directly from a modified uniform Hoeffding bound, due to [27]. In particular, for any integer $T_0 \geq 0$, let $d_{\mathrm{TV}}(T_0) = \sup_{\pi_0} d_{\mathrm{TV}}(\pi_0 P^{T_0},\ \pi)$ be the worst-case total variation distance from stationarity after T_0 steps. Using this notation, [27] shows that for any starting distribution π_0 and any bounded function $f\colon [d] \to [0,\ 1]$, we have

$$\mathbb{P}\big(\big|\frac{1}{N-T_0}\sum_{n=T_0+1}^{N} f(X_n) - \mu\big| \geq \epsilon\big) \leq 2\exp\big(-\frac{\gamma_0}{2(2-\gamma_0)}\cdot\epsilon^2 N\big) + 2d_{\mathrm{TV}}(T_0). \tag{46}$$

We now use the bound (46) to prove our claim (26b). Recall that we have chosen T_0 so that $d_{\mathrm{TV}}(T_0) \le \alpha_0/2$. Therefore, the bound (46) implies that

$$\begin{aligned}\mathbb{P}\left[\Big|\frac{1}{N-T_0}\sum_{n=T_0+1}^{N} f(X_n) - \mu\Big| \ge \epsilon_N\right] &\le 2\exp\Big\{-\frac{\gamma_0}{2(2-\gamma_0)}\cdot\epsilon_N^2 N\Big\} + \alpha_0 \\ &\le 2\cdot\frac{\alpha-\alpha_0}{2} + \alpha_0 = \alpha,\end{aligned}$$

as required.

C.2 Proof of the claim (30)

We now use the result (29) to prove the claim (30).

By the lower bound on N, we have

$$\frac{e^{-\gamma_0 N}}{3\sqrt{\pi_{\min}}} \le \frac{\alpha}{6} \quad \text{and} \quad \frac{13}{\sigma_{f,\mathrm{asym}}\sqrt{\pi_{\min}}}\cdot\frac{1}{\gamma_0\sqrt{N}} \le \frac{\alpha}{6}.$$

It follows from equation (29) that

$$\begin{aligned}\mathbb{P}\left[\frac{1}{\sigma_{f,\mathrm{asym}}N}\sum_{n=1}^{N} f(X_n) \ge \mu + \epsilon_N\right] &\le \Phi\big(\epsilon_N\sqrt{N}\big) + \frac{\alpha}{3} \\ &\le \exp\big(-\frac{N}{2}\cdot\epsilon_N^2\big) + \frac{\alpha}{3} \\ &= \frac{\alpha}{2},\end{aligned}$$

and since a matching bound holds for the lower tail, we get the desired result.

D Proofs for Section 5.5

In this section, we gather the proofs of Lemmas 5–8.

D.1 Proof of Lemma 5

Observe that the function

$$g(\zeta) := \exp\Big\{-\log(1/\beta)\,(\zeta+\zeta^{-1})\Big\}$$

is increasing on $(0,\ 1]$ and decreasing on $[1,\ \infty)$. Therefore, bringing ζ closer to 1 can only increase the value of the function.

Now, for fixed $k \ge 1$, define

$$\ell_k := \begin{cases}\min\big\{\ell\colon (1+\xi)^{\ell} \ge \zeta_k\big\} & \text{if } \zeta_k \le 1, \\ \max\big\{\ell\colon (1+\xi)^{\ell} \le \zeta_k\big\} & \text{otherwise.}\end{cases}$$

In words, the quantity ℓ_k is either the smallest integer such that $(1+\xi)^{\ell}$ is bigger than ζ_k (if $\zeta_k \le 1$) or the largest integer such that $(1+\xi)^{\ell}$ is smaller than ζ_k (if $\zeta_k \ge 1$).

With this definition, we see that $(1+\xi)^{\ell_k}$ always lies between ζ_k and 1, so that we are guaranteed that $g\big((1+\xi)^{\ell_k}\big) \geq g(\zeta_k)$, and hence

$$\sum_{k=1}^{\infty} g(\zeta_k) \leq \sum_{k=1}^{\infty} g\big((1+\xi)^{\ell_k}\big).$$

Thus, it suffices to show that at most two distinct values of k map to a single ℓ_k. Indeed, when this mapping condition holds, we have

$$\sum_{k=1}^{\infty} g(\zeta_k) \leq 2 \sum_{\ell=-\infty}^{\infty} g\big((1+\xi)^{\ell}\big) \leq 4 \sum_{\ell=0}^{\infty} g\big((1+\xi)^{\ell}\big).$$

In order to prove the stated mapping condition, note first that ℓ_k is clearly nondecreasing in k, so that we need to prove that $\ell_{k+2} > \ell_k$ for all $k \geq 1$. It is sufficent to show that $\zeta_{k+2} \geq (1+\xi)\zeta_k$, since this inequality implies that $\ell_{k+2} \geq \ell_k + 1$.

We now exploit the fact that $\zeta_k = a n_k$ for some absolute constant a, where $n_k = \lfloor n_0 (1+\xi)^k \rfloor$. For this, let $b = n_0(1+\xi)^k$, so that $n_k = \lfloor b \rfloor$. Since $n_{k+1} > n_k$, we have $(1+\xi)b \geq \lfloor (1+\xi)b \rfloor \geq \lfloor b \rfloor + 1$, and hence

$$\begin{aligned} \frac{n_{k+2}}{n_k} \;=\; \frac{\lfloor (1+\xi)^2 b \rfloor}{\lfloor b \rfloor} &\geq \frac{(1+\xi)^2 b - 1}{\lfloor b \rfloor} \\ &\geq \frac{(1+\xi)\big[\lfloor b \rfloor + 1\big] - 1}{\lfloor b \rfloor} \\ &\geq 1 + \xi, \end{aligned}$$

as required.[2]

D.2 Proof of Lemma 6

When $c = 0$ and $\ell = 0$, we note that the claim obviously holds with equality. On the other hand, the left hand side is increasing in ℓ, so that the $c = 0$ case follows immediately.

Turning to the case $c > 0$, we first note that it is equivalent to show that

$$(1+\xi)^{2\ell} - 2(c+1)(1+\xi)^{\ell} + 1 \geq 0 \qquad \text{for all } \ell \in \big(\tfrac{9c}{5\xi}, \tfrac{9(c+1)}{5\xi}\big).$$

It suffices to show that $(1+\xi)^{\ell}$ is at least as large as the largest root of the the quadratic equation $z^2 - 2(c+1)z + 1 = 0$. This largest root is given by

$$z^* = c + 1 + \sqrt{c\,(c+2)} \leq 2(c+1).$$

Consequently, it is enough to show that $\ell \geq \frac{\log 2(c+1)}{\log(1+\xi)}$. Since $\frac{9c}{5\xi}$ is a lower bound on ℓ, we need to verify that

$$\frac{9c}{5\xi} \geq \frac{\log 2\,(c+1)}{\log(1+\xi)}.$$

[2] We thank Daniel Paulin for suggesting this argument as an elaboration on the shorter proof in Gyori and Paulin [15].

In order to verify this claim, note first that since $\xi \leq \frac{2}{5}$, we have $\log(1+\xi) \geq \xi - \frac{1}{2}\xi^2 \geq \frac{4}{5}\xi$, whence

$$\frac{\log 2\,(c+1)}{\log(1+\xi)} \leq \frac{5\log 2\,(c+1)}{4\xi}.$$

Differentiating the upper bound in c, we find that its derivative is

$$\frac{5}{4\,(c+1)\,\xi} \leq \frac{5}{8\xi} \leq \frac{9}{5\xi},$$

so it actually suffices to verify the claim for $c = 1$, which can be done by checking numerically that $\frac{5\log 4}{4} \leq \frac{9}{5}$.

D.3 Proof of Lemma 7

Our strategy is to split the infinite sum into two parts: one corresponding to the range of s where h is constant and equal to 1 and the other to the range of s where h is decreasing. In terms of the N_k, these two parts are obtained by splitting the sum into terms with $k < k_0$ and $k \geq k_0$, where $k_0 \geq 1$ is minimal such that $M \leq \Delta N_k$ for $k \geq k_0$.

For convenience in what follows, let us introduce the convenient shorthand

$$T_k := \exp\big(-\frac{\big(\Delta N_k - M\big)_+^2}{2\tau_f\big(\delta/2\big)N_k}\big).$$

Now, if $k_0 = 1$, we note that h must then be decreasing for $s \geq N_1$, so that

$$\sum_{k=1}^{\infty}\big(N_{k+1} - N_k\big)T_k \leq \int_{N_1}^{\infty} h\big(s\big)\mathrm{d}s.$$

Otherwise, if $k_0 > 1$, we have

$$\sum_{k=k_0}^{\infty}\big(N_{k+1} - N_k\big)T_k \leq \int_{N_{k_0}}^{\infty} h\big(s\big)\mathrm{d}s.$$

For $k < k_0$, we have $T_k = 1$, so that when $k < k_0 - 1$, we have

$$\big(N_{k+1} - N_k\big)\exp\big(-\frac{\big(\Delta N_k - M\big)_+^2}{2\tau_f\big(\delta/2\big)N_k}\big) = \int_{N_k}^{N_{k+1}} h\big(s\big)\ \mathrm{d}s.$$

Thus

$$\sum_{k=1}^{k_0-1}\big(N_{k+1} - N_k\big)\exp\big(-\frac{\big(\Delta N_k - M\big)_+^2}{2\tau_f\big(\delta/2\big)N_k}\big) = \int_{N_1}^{N_{k_0-1}} h\big(s\big)\mathrm{d}s.$$

Note that this implies

$$\int_{N_1}^{\infty}\exp\big(-\frac{\big(\Delta s - M\big)_+^2}{2\tau_f\big(\delta/2\big)s}\big)\ \mathrm{d}s \geq N_{k_0-1}.$$

Finally, we observe that $N_{k+1} \le \big(1+\xi\big)N_k + 1 + \xi$, so that $N_{k_0} - N_{k_0-1} \le \xi N_{k_0-1} + 1 + \xi$. Putting together the pieces, we have

$$\big(N_{k_0} - N_{k_0-1}\big) \exp\big(-\frac{\big(\Delta N_{k_0-1} - M\big)_+^2}{2\tau_f\big(\delta/2\big)N_{k_0-1}}\big) \le 1 + \xi + \xi \int_{N_1}^{\infty} \exp\big(-\frac{\big(\Delta s - M\big)_+^2}{2\tau_f\big(\delta/2\big)s}\big)\, \mathrm{d}s,$$

and hence

$$\sum_{k=1}^{\infty} \big(N_{k+1} - N_k\big) \exp\big(-\frac{\big(\Delta N_k - M\big)_+^2}{2\tau_f\big(\delta/2\big)N_k}\big) \le 1 + \xi + \big(1+\xi\big) \int_{N_1}^{\infty} h\big(s\big)\mathrm{d}s.$$

D.4 Proof of Lemma 8

Observe that for $k > k_0^*$, we have $\Delta - \epsilon_k \ge \frac{\Delta}{2}$. It follows that for $k > k_0^*$, we have $T_{f,k}^+ \le T_f\big(\frac{\Delta}{4}\big)$. Thus, we can bound each term in the sum by

$$\big(N_{k+1} - N_k\big) \exp\big(-\frac{N_k}{8T_{f,k}^+} \cdot \big(\Delta - \epsilon_k\big)_+^2\big) \le \underbrace{\big(N_{k+1} - N_k\big) \exp\big(-\frac{N_k}{T_f\big(\frac{\Delta}{4}\big)} \cdot \frac{\Delta^2}{32}\big)}_{F_k}.$$

Furthermore, the exponential in the definition of F_k is a decreasing function of N_k, so we further bound the overall sum as

$$\begin{aligned}
\sum_{k=k_0^*+1}^{\infty} F_k &\le \sum_{n=N_0^*+1}^{\infty} \exp\big(-n \cdot \frac{\Delta^2}{32T_f\big(\frac{\Delta}{4}\big)}\big) \\
&= \exp\big(-N_0^* \cdot \frac{\Delta^2}{32T_f\big(\frac{\Delta}{4}\big)}\big) \times \sum_{m=1}^{\infty} \exp\big(-m \cdot \frac{\Delta^2}{32T_f\big(\frac{\Delta}{4}\big)}\big) \\
&= \exp\big(-\frac{N_0^*}{8T_f\big(\frac{(\Delta/2)}{2}\big)} \cdot \big(\frac{\Delta}{2}\big)^2\big) \times \sum_{m=1}^{\infty} \exp\big(-m \cdot \frac{\Delta^2}{32T_f\big(\frac{\Delta}{4}\big)}\big).
\end{aligned}$$

On the other hand, by the definition of N_0^*, $\epsilon_{k_0^*} \le \frac{\Delta}{2}$, so

$$T_f\big(\frac{(\Delta/2)}{2}\big) \le T_f\big(\frac{\epsilon_{k_0^*}}{2}\big).$$

By the definition of ϵ_k, however, we know that

$$\frac{\epsilon_k^2}{8T_f\big(\frac{\epsilon_k}{2}\big)} \ge \frac{\log\big(1/\alpha\big) + 1 + 2\log k}{N_k} \ge \frac{\log\big(1/\alpha\big)}{N_k},$$

which implies that $(\Delta/2)^2 N_0^* \ge \log\big(1/\alpha\big)\ 8T_f\big(\frac{(\Delta/2)}{2}\big)$. Re-arranging yields the claim.

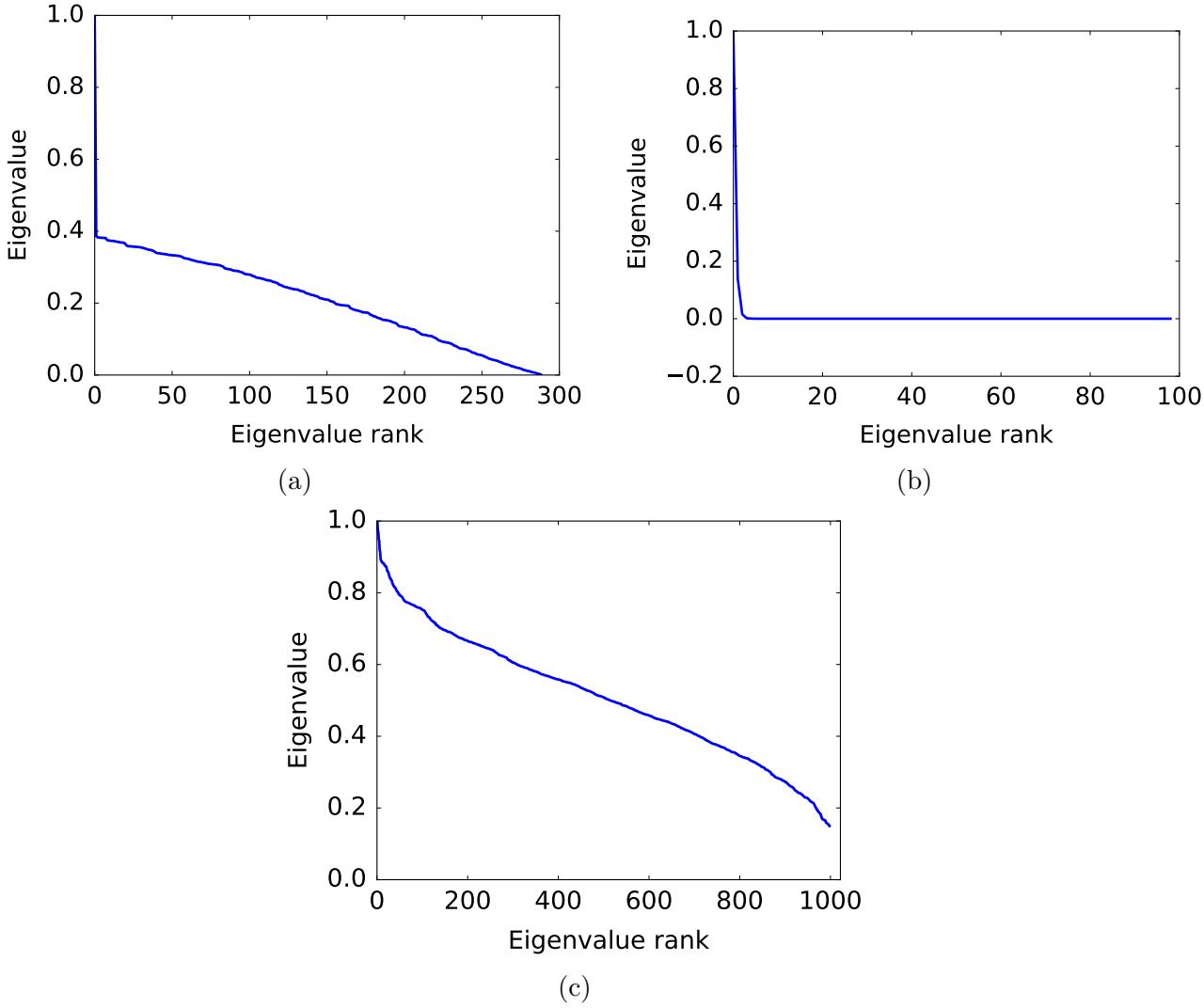


Figure 1: Spectra for three example chains: (a) Metropolis-Hastings for Bayesian logistic regression; (b) collapsed Gibbs sampler for missing data imputation; and (c) collapsed Gibbs sampler for a mixture model.

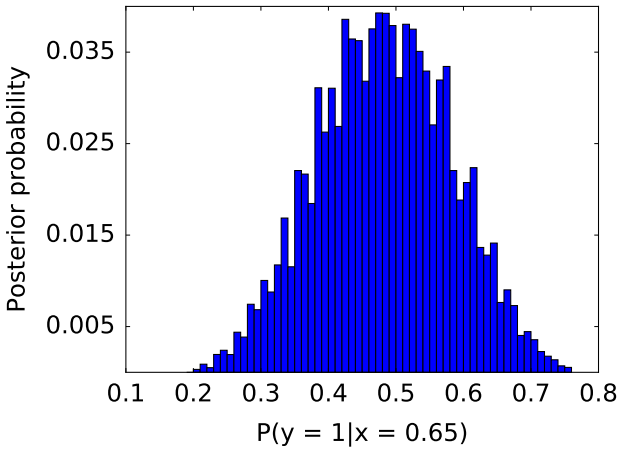
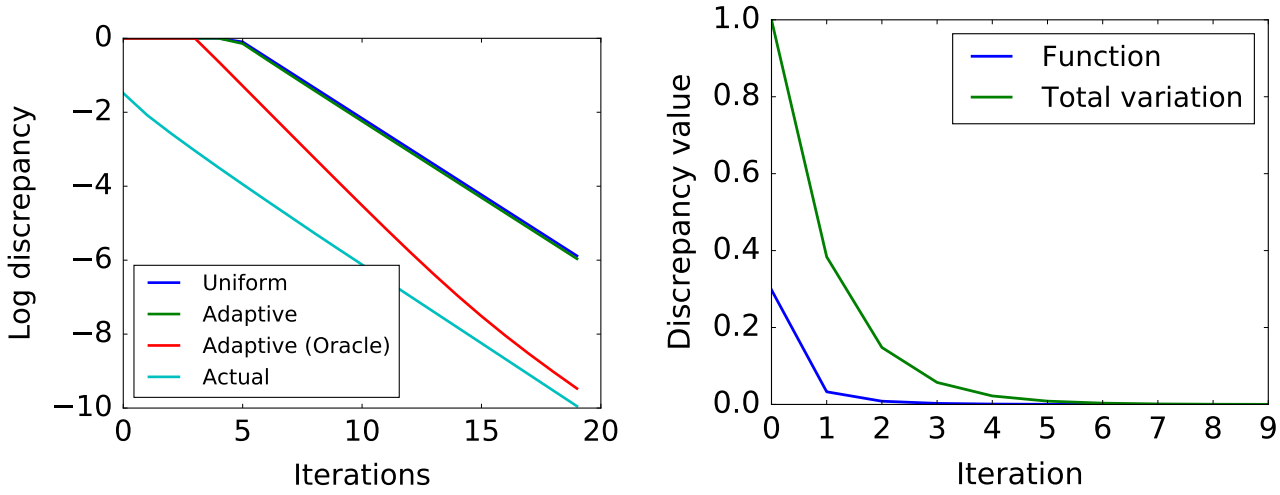


Figure 2: Distribution of f_{65} values under the posterior. Despite the discretization and truncation to a square, it generally matches the one displayed in Figure 1.2 in Robert and Casella [28].



(a) (b)

Figure 3: (a) Discrepancies (plotted on log-scale) for f_{65} as a function of iteration number. The prediction of the naive bound is highly pessimistic; the f-discrepancy bound goes part of the way toward closing the gap and the oracle version of the f-discrepancy bound nearly completely closes the gap in the limit and also gets much closer to the right answer for small iteration numbers. (b) Comparison of the function discrepancy $d_{f_{65}}$ and the total variation discrepancy d_{TV}. They both decay fairly quickly due to the large spectral gap, but the function discrepancy still falls much faster.

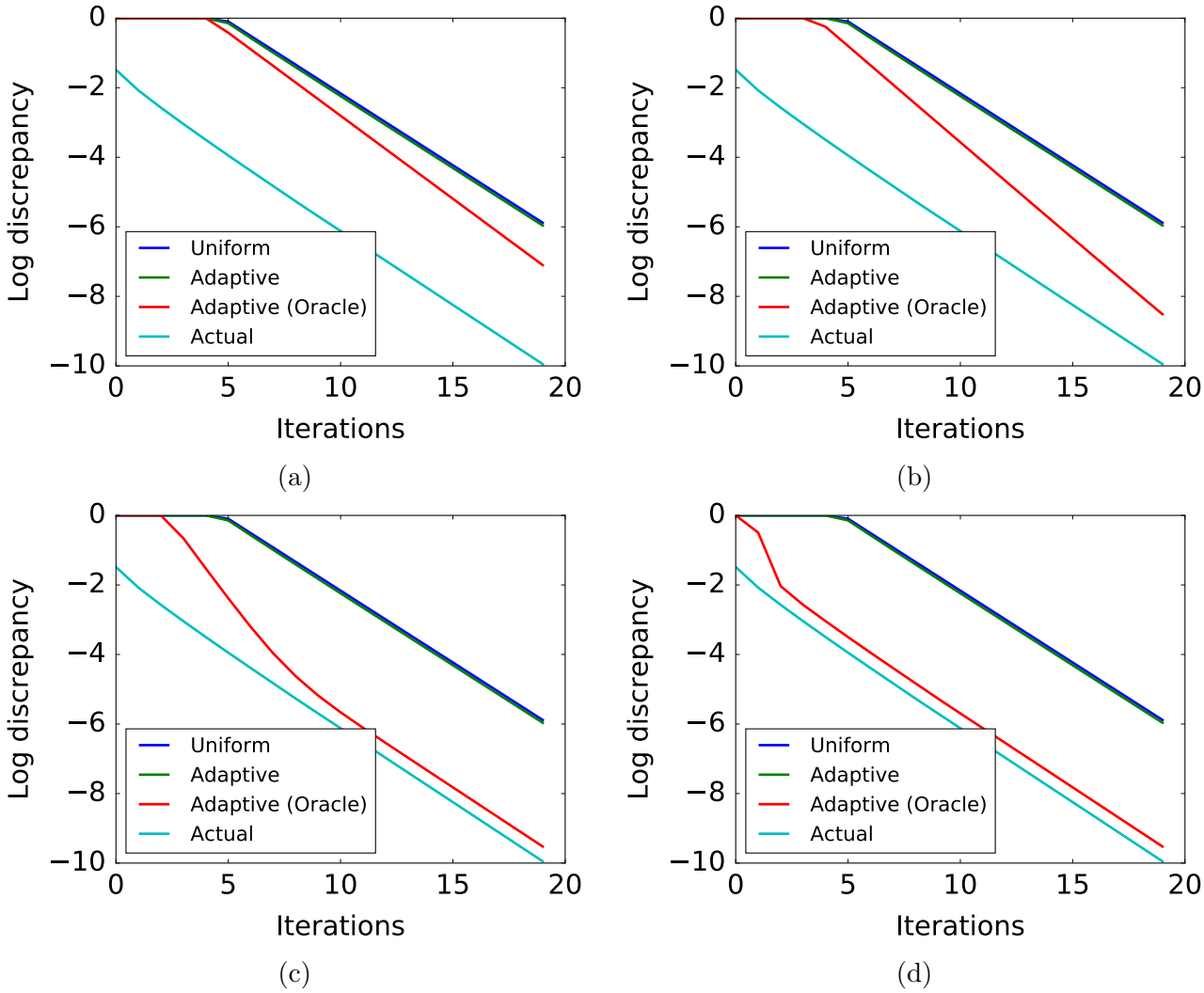


Figure 4: Comparisons of the uniform, non-oracle function-specific, and oracle function-specific bounds for various choices of J. In each case, $J = \{2, \ldots, J_{\max}\}$, with $J_{\max} = 50$ in panel (a), $J_{\max} = 100$ in panel (b), $J_{\max} = 200$ in panel (c), and $J_{\max} = 288$ in panel (d). The oracle bound becomes tight in the limit as $J_{\max}$ goes to $d = 289$, but it offers an improvement over the uniform bound across the board.

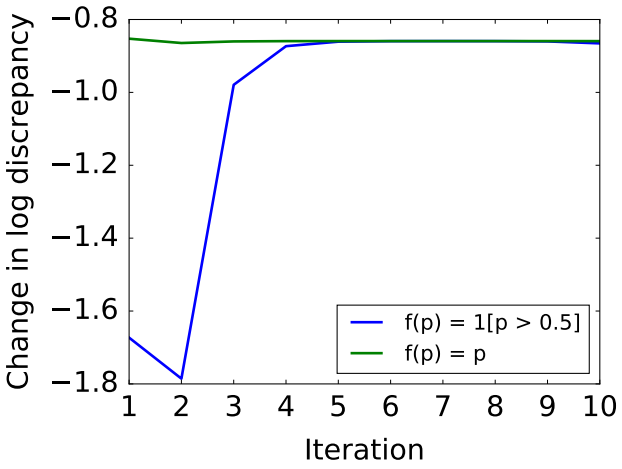
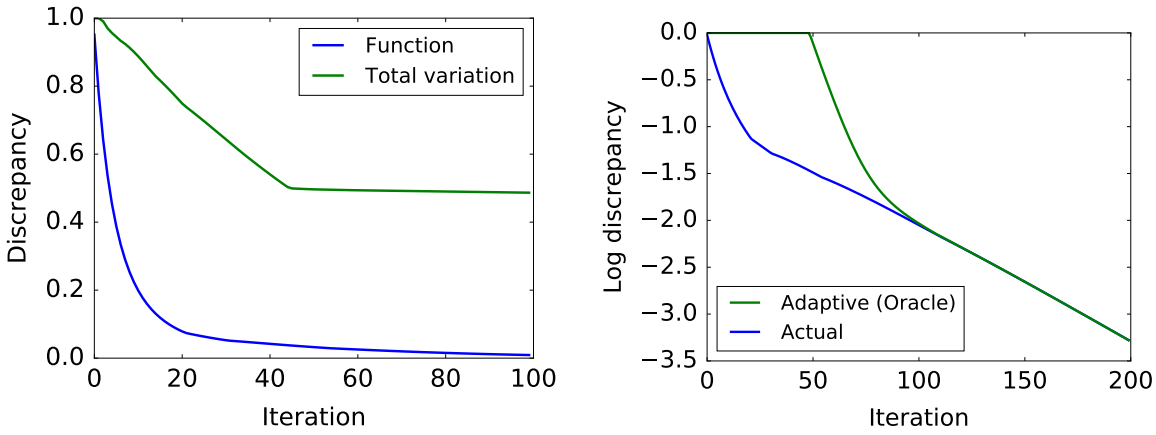


Figure 5: Change in log discrepancy for the two functions $f(p) = \mathbf{1}\big(p \geq 0.5\big)$ and $f(p) = p$ considered above. Whereas $f(p) = p$ always changes at the constant rate dictated by the spectral gap, the indicator discrepancy decays more quickly in the first few iterations.



(a) (b)

Figure 6: (a) Comparison of the f-discrepancy d_f and the total variation discrepancy d_{TV} over the first 100 iterations of MCMC. Clearly the function mixes much faster than the overall chain. (b) The predicted value of $\log d_f$ (according to the f-discrepancy oracle bound—Lemma 2) plotted against the true value. The predictions are close to sharp throughout and become sharp at around 100 iterations.

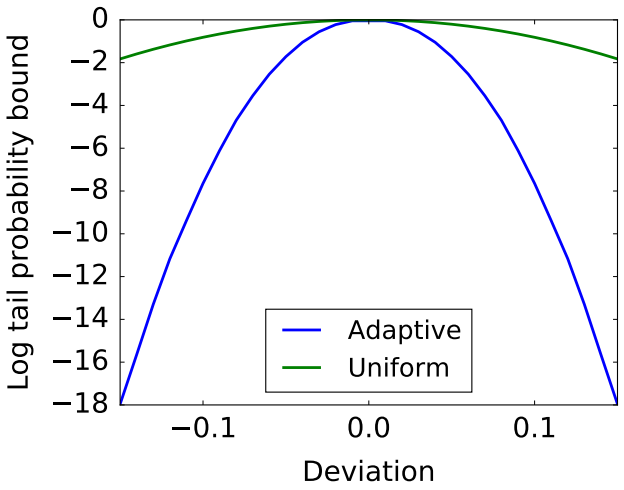


Figure 7: Comparison of the (log) tail probability bounds provided by the uniform Hoeffding bound due to [20] with one version of our function-specific Hoeffding bound (Theorem 1). Plots are based on $N = 10^6$ iterations, and choosing the optimal burn-in for the uniform bound and a fixed burn-in of $409 \geq T_f\left(10^{-6}\right)$ iterations for the function-specific bound. The function-specific bound improves over the uniform bound by orders of magnitude.