Bayesian Analysis (2020) **15**, Number 2, pp. 505–532

Function-Specific Mixing Times and Concentration Away from Equilibrium

Maxim Rabinovich*, Aaditya Ramdas†, Michael I. Jordan‡, and Martin J. Wainwright§

Abstract. Slow mixing is the central hurdle is applications of Markov chains, especially those used for Monte Carlo approximations (MCMC). In the setting of Bayesian inference, it is often 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 central limit theorem (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.

MSC 2010 subject classifications: Primary 60J10; secondary 62M05, 62M02.

Keywords: Markov chains, Markov chain Monte Carlo, concentration inequalities, confidence intervals, sequential testing, statistics, probability.

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., Gelman et al., 2013; Robert and Casella, 2005). 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 (Levin et al., 2008)—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

*University of California, Berkeley, rabinovich@cs.berkeley.edu
†University of California, Berkeley, aramdas@cmu.edu
‡University of California, Berkeley, jordan@cs.berkeley.edu
§University of California, Berkeley, wainwrig@berkeley.edu

 https://doi.org/10.1214/19-BA1151

to uniform convergence of the expectations of all unit-bounded function to their equilibrium values. The resulting uniform bounds on the accuracy of expectations (Chung et al., 2012; Gillman, 1998; Joulin et al., 2010; Kontorovich et al., 2014; Léon and Perron, 2004; Lezaud, 2001; Paulin, 2012; Samson et al., 2000) 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. Meanwhile, the few function-specific bounds available (Hayashi and Watanabe, 2016; Watanabe and Hayashi, 2017) are difficult to interpret, apply, and compute, and may not be optimal in finite samples and at finite precisions.

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:

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

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

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

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 $X_0 = i$.

Since we assume reversibility, the matrix $S = \operatorname{diag}(\pi)\, P$ is symmetric and has a spectral decomposition. Writing $P = \operatorname{diag}(\pi)^{-1} S$ then gives a corresponding decomposition

of P, which we denote by

$$P = \mathbf{1}\pi^T + \sum_{j=2}^{d} \lambda_j h_j q_j^T. \tag{3}$$

Here, in accordance with the decomposition at the top eigenvalue ($\lambda_1 = 1$), we should think of the h_j as functions, a view we revisit when they come into play below.

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

$$d_f(p,\ q) = \left|\mathbb{E}_{X\sim p}\left[f(X)\right] - \mathbb{E}_{Y\sim q}\left[f(Y)\right]\right|. \tag{4}$$

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

$$T_f(\delta) = \min\left\{n \in \mathbb{N} \ \mid \ \max_{i\in[d]} d_f\left(\pi_n^{(i)},\ \pi\right) \le \delta\right\}. \tag{5}$$

We sometimes use T_f without an argument either when the argument is obvious from context, or when we want to refer to the quantity generically rather than its evaluation at a specific δ. 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.

We also use some asymptotic notation, which we now clarify. If g_1, g_2 are two nonnegative functions of some variable x, we define the notations

$$\begin{aligned}
g_1 \approx g_2 &\iff \exists c,\ c' > 0, \quad cg_1 \le g_2 \le c'g_1,\\
g_1 \asymp g_2 &\iff g_1 \approx g_2,\\
g_1 \lesssim g_2 &\iff \exists c > 0, \quad g_1 \le cg_2,\\
g_1 \gg g_2 &\iff g_1 \not\lesssim g_2.
\end{aligned}$$

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., Aldous and Diaconis, 1986; Diaconis and Fill, 1990; Levin et al., 2008; Meyn and Tweedie, 2012; Ollivier, 2009; Sinclair, 1992). These tools and the results based on them, however, generally apply only to worst-case mixing times. Outside of specific examples (Conger and Viswanath, 2006; Diaconis and Hough, 2015), mixing with respect to individual functions or limited

classes of functions has received relatively little attention, and almost none at all in the statistics literature.

One important exception is the recent work by Hayashi and Watanabe (2016) and Watanabe and Hayashi (2017), who provide asymptotically sharp tail bounds on empirical averages of functions using methods from information geometry. These bounds are of the form

$$\mathbb{P}\left(\frac{1}{N}\sum_{n=1}^{N} f(X_n) \geq \mu + \epsilon\right) \leq \exp\big(NC(\epsilon) + D(\epsilon)\big),$$

where $D(\epsilon)$ is a constant defined explicitly (Watanabe and Hayashi, 2017) that tends to 0 as $\epsilon \to 0$, and $C(\epsilon)$ is the large-deviation rate, which is to say

$$\begin{aligned} C(\epsilon) &= (\mu+\epsilon)\phi^{'-1}(\mu+\epsilon) - \phi\big(\phi^{'-1}(\mu+\epsilon)\big) \\ &= \lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}\left(\frac{1}{N}\sum_{n=1}^{N} f(X_n) \geq \mu + \epsilon\right), \end{aligned}$$

where $\phi(t)$ denotes the largest eigenvalue of the matrix $P(t)$ defined by $(P(t))_{i,j} = P \cdot e^{tf(i)}$ for $1 \leq i, j \leq d$. (The largest eigenvalue is a nonnegative real number by the Perron-Frobenius Theorem.) Watanabe and Hayashi (2017) also provide lower bounds that are closely matching these upper bounds. Together these are used to derive classical results in probability theory (large deviations, moderate deviations, and CLT) for Markov chains.

While we do not claim that the inequalities in this paper are sharper than these results, they are stated in terms of f-mixing times which are much more intuitive and easier to use in practice than the large deviation rates. We provide several results based on spectral methods and coupling arguments that allow us to bound the f-mixing times, and illustrate the quality of our predictions in simulations, a task that appears to be more intensive computationally and algorithmically for the information-geometry bounds.

Other existing bounds are generally uniform over functions, 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 (Chung et al., 2012; Gillman, 1998; Joulin et al., 2010; Kontorovich et al., 2014; Léon and Perron, 2004; Lezaud, 2001; Paulin, 2012; Samson et al., 2000). (A degree of adaptation is possible in that the asymptotic variance of the function f can be accounted for in Bernstein-type bounds, but the key factor does not adapt—see for instance Lezaud (1998); Paulin (2012).) These factors, which do not appear in classic bounds for independent random variables,[1] are generally either some variant of the spectral gap γ of the transition matrix, or else a mixing time of the chain $T(\delta_0)$ for some absolute constant $\delta_0 > 0$. For example, the main theorem from Léon

[1]Technically, since independent random variables form a Markov chain with spectral gap 1, the spectral gap does appear, but it appears in a trivial way as a factor of unity.

and Perron (2004) shows that for a function $f\colon [d] \to [0,\ 1]$ and a sample $X_0 \sim \pi$ from the stationary distribution, we have

$$\mathbb{P}\big(\big|\frac{1}{N}\sum_{n=1}^{N} f\big(X_n\big) - \mu\big| \geq \epsilon\big) \leq 2\exp\Big\{-\frac{\gamma_0}{2\big(2-\gamma_0\big)}\cdot\epsilon^2 N\Big\}, \tag{6}$$

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 (Paulin, 2012). 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., Kontorovich et al. (2014); Samson et al. (2000)).

The concentration result has an alternative formulation in terms of the mixing time instead of the spectral gap (Chung et al., 2012). This version and its variants are weaker, since the mixing time can be lower bounded as

$$T\big(\delta\big) \geq \big(\frac{1}{\gamma_*} - 1\big)\log\big(\frac{1}{2\delta}\big) \geq \big(\frac{1}{\gamma_0} - 1\big)\log\big(\frac{1}{2\delta}\big), \tag{7}$$

where we denote the absolute spectral gap (Levin et al., 2008) 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.

We note that there remains a gap between theoretical work on the convergence of Markov chains, of the kind developed here, and practical applications of the theory to MCMC. Indeed, most current applications of MCMC do not use rigorous bounds of the Hoeffding type; rather, they build variance-based confidence intervals, either via CLT approximations or, more recently, via Chebyshev's inequality (Gyori and Paulin, 2012; Flegal et al., 2008). Such bounds are simple to compute and have good *asymptotic* theoretical properties, but they are not valid in a non-asymptotic setting; indeed, they may be over-optimistic and anti-conservative in finite samples. In contrast, Hoeffding-type bounds, including our own, sit at the other end of the spectrum; they come with finite-sample validity built in, but may be difficult to compute due to the dependence on the mixing time, either function-specific or uniform. There is a recent line of work aimed at estimating mixing times from individual sample trajectories (Hsu et al., 2015, 2017) that has begun to bridge the gap between the strong theory underlying Hoeffding bounds and their target applications, but this research direction is still nascent. We note that one other promising direction is the connection to analysis of specific statistically-relevant Markov chains (e.g., Choi and Hobert, 2013; Román and Hobert, 2015), which has the potential of yielding numerical bounds on mixing times (Jones and Hobert, 2001).

1.3 Organization of the paper

In the remainder of the paper, we develop a formal framework for bounding function-specific mixing times and we apply the framework to the analysis of MCMC algorithms. 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. (For reasons of space, we defer an analysis of sequential testing to Appendix E (Rabinovich et al., 2019).) 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 Appendix A (Rabinovich et al., 2019), with some more technical aspects of the arguments deferred to Appendices B, D, and F (Rabinovich et al., 2019).

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 quantities that bound 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 $N \geq T_f(\epsilon/2)$ 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 $f(X_n)$ 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. In fact, as we discuss below, it is not really necessary to explicitly discard the first T_f samples, so

knowing the (function-specific) mixing time is not actually necessary, as long as N is larger than T_f. The tacit assumption in this theorem and all our concentration results is that f is bounded in $[0,\ 1]$.

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

$$\mathbb{P}\Big[\frac{1}{N}\sum_{n=1}^{N} f(X_n) \geq \mu + \epsilon\Big] \leq \exp\left\{-\frac{\epsilon^2}{8} \cdot \left\lfloor \frac{N}{T_f(\frac{\epsilon}{2})} \right\rfloor\right\} \leq \exp\left\{-\frac{\epsilon^2 N}{16 T_f(\frac{\epsilon}{2})}\right\}. \tag{8}$$

Compared to the bounds in earlier work (e.g., Léon and Perron, 2004), the bound (8) has several distinguishing features. The primary difference is that the "effective" sample size, that is, the number of samples that would give an equivalent level of concentration if all the samples were i.i.d. from π,

$$N_{\text{eff}} := \Big\lfloor \frac{N}{T_f(\epsilon/2)} \Big\rfloor, \tag{9a}$$

is a function of f, which can lead to significantly sharper bounds on the deviations of empirical means than the earlier uniform bounds can deliver. Further, the result applies when 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

$$N_{\text{eff}}^{\text{burnin}} := \Big\lfloor \frac{N - T_f(\epsilon/2)}{T_f\,(\epsilon/2)} \Big\rfloor \;=\; \Big\lfloor \frac{N}{T_f\,(\epsilon/2)} \Big\rfloor - 1. \tag{9b}$$

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. One can take this reasoning further to determine that, as mentioned above, it is not even necessary to explicitly use a burn-in period. Indeed, in order for Theorem 1 to be meaningful, it must be that $\frac{N}{T_f(\epsilon/2)} \gtrsim \frac{1}{\epsilon^2}$, so that, up to constants, the burn-in period is at most an ϵ^2 fraction of the total number of samples. It follows that directly averaging the function values along the trajectory provides a good approximation of the average over the last $N - T_f(\epsilon/2)$ samples, up to an accuracy on the order of $\epsilon^2 \ll \epsilon$.

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 T_f-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 (7) between mixing and spectral graphs (as well as matching upper bounds (Levin et al., 2008)), for any fixed function f, there exists a function-specific spectral gap $\gamma_f > 0$ such that

$$T_f\big(\frac{\epsilon}{2}\big) \approx \frac{1}{\gamma_f} \log\Big(\frac{1}{\epsilon}\Big) + O\big(1\big), \quad \text{for} \quad \epsilon \ll 1. \tag{9c}$$

These asymptotics can be used to turn our aforementioned theorem into a variant of the results of Léon and Perron (2004), 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.

On the other hand, it is important to note that our bounds do not provide optimal rates in the asymptotic setting. Indeed, our results only imply convergence of the sample mean to the true mean at a rate of $\frac{\log N}{\sqrt{N}}$, due to the logarithmic dependence of T_f on the error ϵ. Our lower bound, Proposition 1, shows that the T_f factor cannot be removed in general, so that optimal asymptotic convergence rates do not seem attainable in general with a function-specific analysis. Nonetheless, by interpolating between the function-specific and global bounds, one can obtain the best of both worlds, so we do not believe the asymptotic sub-optimality to be a major concern. This point of view is supported by our experiments, which suggest that at the precisions one typically targets in practice, the seemingly extraneous $\log N$ factor is overcome by the gains of having a larger function-specific spectral gap, and the function-specific bounds end up being superior.

By combining our definition (9a) of the effective sample size N_{eff} with the asymptotic expansion (9c), 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 (6), 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 $1 \leq n \leq N$,

$$|f(X_n) - \mu| \gg \epsilon \quad \text{a.s. if } X_0 = i. \tag{10}$$

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 (10) guarantees that

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

so that if $\pi(\Omega_0) \gg 0$, we have no hope of controlling the large-deviation probability unless $N \gtrsim T_f(\epsilon)$.

To make this intuitive idea precise, the basic idea is to start with an arbitrary candidate function $\rho : (0, 1) \to (0, 1)$, such that $T_f\left(\frac{\epsilon}{2}\right)$ in the denominator of the function-specific Hoeffding bound (8) can putatively be replaced by $T_f(\rho(\epsilon))$. We then show that if $\rho(\epsilon) \geq \epsilon$, the replacement is not actually possible. That means that, up to a possible constant factor improvement in the argument to T_f, the dependence of the exponent in Theorem 1 on $T_f(\epsilon/2)$ cannot be eliminated—surprising in light of the fact that *uniform* Hoeffding bounds do not exhibit this behavior in their dependence on the mixing or relaxation times. In this sense, the rate we attain is improvable only in the constants in the exponent of the bound, as claimed above.

We prove Proposition 1 by constructing a Markov chain (which is independent of ϵ) and a function (which depends on both ϵ and ρ) 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). We defer the proof to Appendix A.3 (Rabinovich et al., 2019).

Proposition 1. *Fix a function ρ: $(0, 1) \to (0, 1)$ with $\rho(\epsilon) > \epsilon$. 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*

$$\mathbb{P}_\pi\left(\left|\frac{1}{N}\sum_{n=1}^{N} f(X_n) - \frac{1}{2}\right| \geq \epsilon\right) > 2 \cdot \exp\left(-\frac{c_1 N \epsilon^2}{T_f(\rho(\epsilon))}\right). \tag{11}$$

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 (8). 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

$$d_{\mathrm{TV}}\big(\pi_n,\ \pi\big) \le \frac{1}{\sqrt{\pi_{\min}}} \cdot \lambda_*^n \cdot d_{\mathrm{TV}}\big(\pi_0,\ \pi\big). \tag{12}$$

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

$$J_f := \Big\{ j \in [d] \ \mid\ \lambda_j \neq 1 \text{ and } q_j^T f \neq 0 \Big\}, \tag{13a}$$

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

$$\lambda_f = \max_{j \in J_f} \big|\lambda_j\big|, \quad \text{and} \quad \gamma_f = 1 - \lambda_f. \tag{13b}$$

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

$$d_f\big(\pi_n,\ \pi\big) \le \sqrt{\frac{\mathbb{E}_\pi\big[f^2\big]}{\pi_{\min}}} \cdot \lambda_f^n \cdot d_f\big(\pi_0,\ \pi\big). \tag{14}$$

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

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

Here the h_j are the functions defined in the decomposition of P in (3). 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*

$$d_f(\pi_n,\ \pi) \leq \Delta_J^*\, \lambda_J^n \cdot d_{\mathrm{TV}}(\pi_0,\ \pi) + \sqrt{\frac{\mathbb{E}_\pi\left[f^2\right]}{\pi_{\min}}} \cdot \lambda_{-J}^n\, d_f(\pi_0,\ \pi). \tag{15}$$

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:

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

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*

$$d_f(\pi_n,\ \pi) \leq \left|(\pi_0 - \pi)^T h_J(n)\right| + \sqrt{\frac{\mathbb{E}_\pi\left[f^2\right]}{\pi_{\min}}} \cdot \lambda_{-J}^n \cdot d_f(\pi_0,\ \pi). \tag{16}$$

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 $\|h_J(n)\|_\infty$ 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 (14), 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*

$$\mathbb{P}\left[\frac{1}{N}\sum_{n=1}^N f(X_n) \geq \mu + \epsilon\right] \leq \begin{cases} \exp\left(-\frac{\epsilon^2}{16}\,\frac{\gamma_f N}{\log\left(\frac{2}{\epsilon\sqrt{\pi_{\min}}}\right)}\right) & \text{if } \epsilon \leq \frac{2\lambda_f}{\sqrt{\pi_{\min}}}, \\ \exp\left(-\frac{\epsilon^2 N}{16}\right) & \text{otherwise.} \end{cases}$$

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* $\Delta_J \geq \Delta_J^*$, $d_f(\pi_0, \pi) \leq \Delta_J + \Delta$, *and* $N \geq T_f(\Delta_J + \Delta)$, *we have*

$$\mathbb{P}\left[\frac{1}{N}\sum_{n=1}^{N} f(X_n) \geq \mu + 2(\Delta_J + \Delta)\right] \leq \begin{cases} \exp\left(-\frac{(\Delta_J+\Delta)^2}{4}\,\frac{(1-\lambda_{-J})N}{\log\left(\frac{1}{\Delta\sqrt{\pi_{\min}}}\right)}\right) & \text{if } \Delta \leq \frac{\lambda_{-J}}{\sqrt{\pi_{\min}}}, \\ \exp\left(-\frac{(\Delta_J+\Delta)^2 N}{4}\right) & \text{if } \Delta > \frac{\lambda_{-J}}{\sqrt{\pi_{\min}}}. \end{cases} \tag{17}$$

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 Appendix A.2 (Rabinovich et al., 2019).

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

$$P_{uv} = \begin{cases} \frac{1}{2} & \text{if } v = u, \\ \frac{1}{4} & \text{if } v = u+1 \mod 2d, \\ \frac{1}{4} & \text{if } v = u-1 \mod 2d, \\ 0 & \text{otherwise.} \end{cases} \tag{18}$$

It is easy to see that the chain is irreducible, aperiodic, and reversible, and its stationary distribution is uniform. It can be shown (Levin et al., 2008) 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:

$$f_{\text{parity}}(u) := \begin{cases} 1 & \text{if } u \text{ is odd}, \\ 0 & \text{otherwise.} \end{cases} \tag{19}$$

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 $g_j(u) = \cos\left(\frac{\pi j u}{d}\right)$; (see, e.g., Levin et al., 2008). 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 $0 \leq j \leq d$ 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

$$\gamma_{f_j} = \frac{1 - \cos\left(\frac{\pi j}{d}\right)}{2} \geq \begin{cases} \frac{\pi^2 j^2}{24 d^2} & \text{if } j \leq \frac{d}{2}, \\ \frac{1}{2} & \text{if } \frac{d}{2} < j \leq d. \end{cases} \tag{20}$$

Consequently, equation (14) predicts that

$$T_{f_j}(\delta) \leq \tilde{T}_{f_j}(\delta) = \begin{cases} \frac{24}{\pi^2}\left[\frac{1}{2}\log 2d + \log\left(\frac{1}{\delta}\right)\right] \cdot \frac{d^2}{j^2} & \text{if } j \leq \frac{d}{2}, \\ \log 2d + 2\log\left(\frac{1}{\delta}\right) & \text{if } \frac{d}{2} < j \leq d, \end{cases} \tag{21}$$

where we have used the trivial bound $\mathbb{E}_\pi\big[f^2\big] \leq 1$ to simplify the inequalities. Note that this yields an improvement over $\asymp d^2$ for $j \gtrsim \log d$. Moreover, the bound (21) 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 (20) that if we draw $N + \tilde{T}_{f_j}\left(\frac{\epsilon}{2}\right)$ samples, we obtain the tail bound

$$\mathbb{P}\Big[\frac{1}{N_0}\sum_{n=N_{\text{b}}}^{N+N_{\text{b}}} f_j(X_n) \geq \frac{1}{2} + \epsilon\Big] \leq \begin{cases} \exp\Big(-\frac{3d^2}{2\pi^2 j^2} \cdot \frac{\epsilon^2 N}{\log\left(2\sqrt{2d}/\epsilon\right)}\Big) & \text{if } j \leq \frac{d}{2}, \\ \exp\Big(-\frac{\epsilon^2 N}{32\log\left(2\sqrt{2d}/\epsilon\right)}\Big) & \frac{d}{2} < j \leq d, \end{cases} \tag{22}$$

where the burn-in time is given by $N_{\text{b}} = \tilde{T}_{f_j}(\epsilon/2)$. 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

i.i.d. from some distribution ν on $[0,\ 1]$ whose mean μ^* is $1/2$:

$$\{f(u),\ u \in C_{2d}\} \overset{\text{i.i.d.}}{\sim} \nu. \tag{23}$$

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

$$T_f(\delta) \lesssim \frac{d \log d \big[\log d + \log\left(\frac{1}{\delta}\right)\big]}{\delta^2}, \qquad \text{for all } \delta \in (0, \delta^*]. \tag{24}$$

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 (24) is to apply Lemma 1 with

$$J_\delta := \left\{ j \in \mathbb{N} \cap [1, 2d-1] \ \mid\ j \leq 4\delta\sqrt{\frac{d}{\log d}} \ \text{ or } \ j \geq 2d - 4\delta\sqrt{\frac{d}{\log d}} \right\}. \tag{25}$$

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 $2\,(\Delta_J + \Delta) \leq \epsilon$. 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 $N + T_f\left(\epsilon/2\right)$ samples for some $N \geq T_f\left(\frac{\epsilon}{2}\right)$, then we have

$$\mathbb{P}\left[\frac{1}{N}\sum_{n=T_f(\epsilon/2)}^{N+T_f(\epsilon/2)} f(X_n) \geq \mu + \epsilon\right] \leq \exp\left\{-\frac{c_2\epsilon^4 N}{d \log d\big[\log\left(\frac{4}{\epsilon}\right) + \log 2d\big]}\right\}, \tag{26}$$

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 C (Rabinovich et al., 2019).

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. 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. In Appendix E (Rabinovich et al., 2019), 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.

3.1 Confidence intervals for posterior expectations

In many applications, a point estimate of $\mathbb{E}_\pi[f]$ 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 Léon and Perron (2004). 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

$$\epsilon_N(\alpha,\ \alpha_0) := \sqrt{2(2-\gamma_0)} \cdot \sqrt{\frac{\log\big(2/[\alpha-\alpha_0]\big)}{\gamma_0[N - T(\alpha_0)]}}. \tag{27a}$$

Then the uniform Markov Hoeffding bound (Léon and Perron, 2004, Theorem 1) implies that the set

$$I_N^{\text{unif}}(\alpha,\ \alpha_0) = \left[\frac{1}{N - T(\alpha_0/2)} \sum_{n=T(\alpha_0/2)+1}^{N} f(X_n) \pm \epsilon_N(\alpha,\ \alpha_0)\right] \tag{27b}$$

is a $1-\alpha$ confidence interval. Full details of the proof are given in Appendix D.2 (Rabinovich et al., 2019).

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 $\epsilon_N^*(\alpha) := \inf_{0<\alpha_0<\alpha} \epsilon_N(\alpha,\ \alpha_0)$. Equation (27b) then implies that

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

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 $e^{-r_N(\epsilon)/16}$, where

$$r_N(\epsilon) := \epsilon^2 \left[\frac{N}{T_f\left(\frac{\epsilon}{2}\right)} - 1 \right]. \tag{28}$$

If we wish for each tail to have probability mass that is at most $\alpha/2$, we need to choose $\epsilon > 0$ so that $r_N(\epsilon) \geq 16 \log \frac{2}{\alpha}$, 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* $\epsilon_N \in r_N^{-1}\big([16 \log(2/\alpha),\ \infty)\big)$, *the set*

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

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

$$r_N(\epsilon) \geq \underbrace{\epsilon^2 \left[\frac{N}{T_f\left(\frac{\eta}{2}\right)} - 1 \right]}_{r_{N,\eta}(\epsilon)} \quad \text{valid for all } \epsilon \geq \eta. \tag{29}$$

Consequently, whenever $r_{N,\eta}(\epsilon_N) \geq 16 \log \frac{2}{\alpha}$ 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*

$$\epsilon_N = r_{N,\eta}^{-1}\left(16 \log \frac{2}{\alpha}\right) = 4\sqrt{\frac{T_f\left(\frac{\eta}{2}\right) \cdot \log(2/\alpha)}{N - T_f\left(\frac{\eta}{2}\right)}}.$$

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 Appendix D (Rabinovich et al., 2019), 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., Flegal et al., 2008; Jones and Hobert,

2001; Glynn and Lim, 2009; Robert and Casella, 2005). If the Markov CLT held exactly, it would lead to appealingly simple confidence intervals of width

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

where $\sigma^2_{f,\mathrm{asym}} := \lim_{N\to\infty} \frac{1}{N} \mathrm{Var}_{X_0\sim\pi}\big[\sum_{n=1}^N f\big(X_n\big)\big]$ 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 $\tilde{S}_N = \sum_{n=1}^N \big[f\big(X_n\big) - \mu\big]$. We then have the bound (Lezaud, 2001):

$$\Big|\mathbb{P}\Big(\frac{\tilde{S}_N}{\sigma_{f,\mathrm{asym}}\sqrt{N}} \le s\Big) - \Phi\big(s\big)\Big| \le \frac{e^{-\gamma_0 N}}{3\sqrt{\pi_{\min}}} + \frac{13}{\sigma_{f,\mathrm{asym}}\sqrt{\pi_{\min}}} \cdot \frac{1}{\gamma_0\sqrt{N}}, \tag{30}$$

where Φ is the standard normal cumulative distribution function (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 (30) 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 $\frac{\log\big(1/\pi_{\min}\big)}{\gamma_0}$ 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

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

then both terms of equation (27b) are bounded by 1/6, and the confidence intervals take the form

$$I_N^{\mathrm{BE}} = \left[\frac{1}{N}\sum_{n=1}^N f\big(X_n\big) \pm \sigma_{f,\mathrm{asym}}\sqrt{\frac{2\log\big(6/\alpha\big)}{N}}\right]. \tag{31}$$

See Appendix D.3 (Rabinovich et al., 2019) 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 $\sigma_f^2 = \mathrm{Var}_\pi\big[f\big(X\big)\big]$ and $\rho_f := \frac{\sigma_f^2}{\sigma^2_{f,\mathrm{asym}}}$, we can rewrite the width as

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

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).

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. Two of these examples appear in the following sections, while a third is relegated to Appendix G (Rabinovich et al., 2019). Figure 1 shows the spectra of the transition matrix for all three examples.

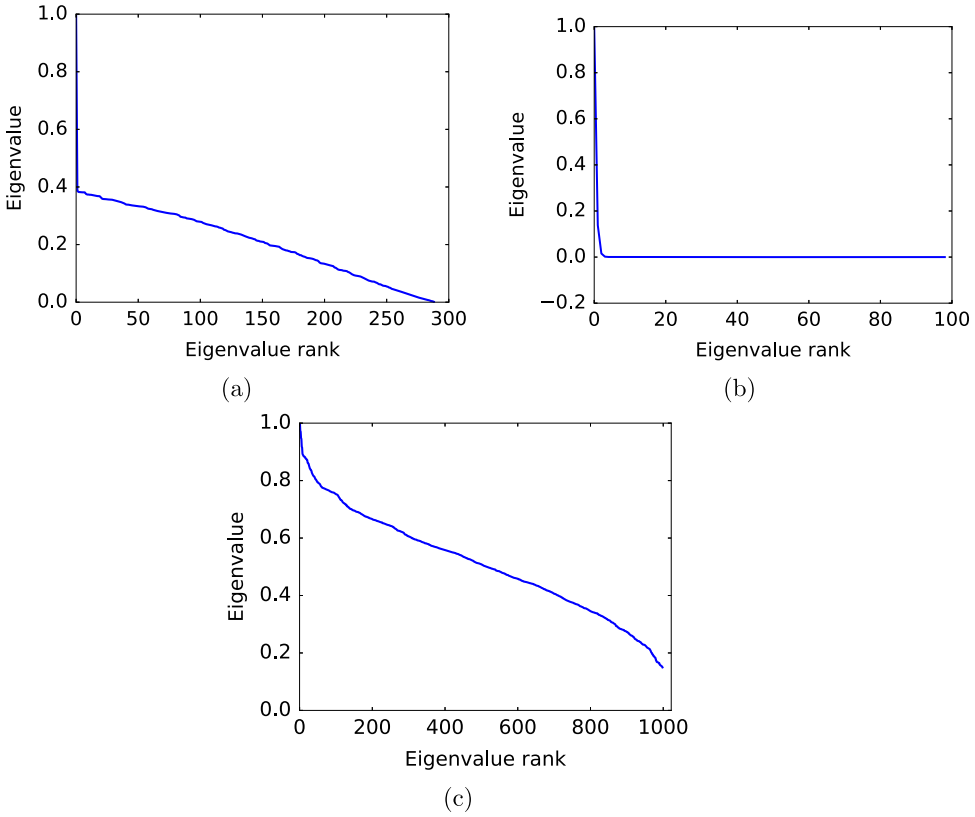


Figure 1: Spectra for three example chains: (a) Metropolis-Hastings for Bayesian logistic regression (Section 4.1); (b) collapsed Gibbs sampler for missing data imputation (Appendix G (Rabinovich et al., 2019)); and (c) collapsed Gibbs sampler for a mixture model (Section 4.2).

4.1 Bayesian logistic regression

Our first example is a Bayesian logistic regression problem introduced by Robert and Casella (2005). 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 $(\alpha,\ \beta) \in \mathbb{R}^2$, 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 (2012), we consider the following model:

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

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 (2012), we target the posterior by running a Metropolis-Hastings algorithm with a Gaussian proposal with covariance matrix $\Sigma = \left(\begin{smallmatrix} 4 & 0 \\ 0 & 10 \end{smallmatrix}\right)$. 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 maximum likelihood estimate (MLE). This space has $d = 17^2 = 289$ elements, resulting in a 289×289 transition matrix that can easily be diagonalized.

Robert and Casella (2005) analyze the probability of failure when the temperature x is 65°F; it is specified by the function

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

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

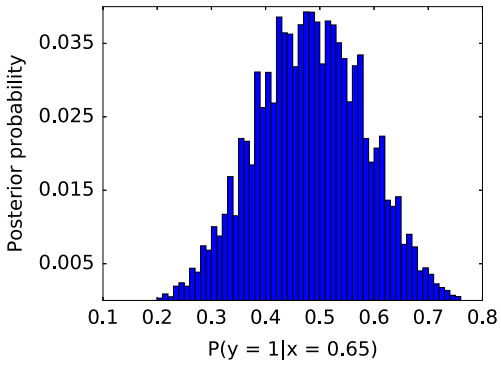
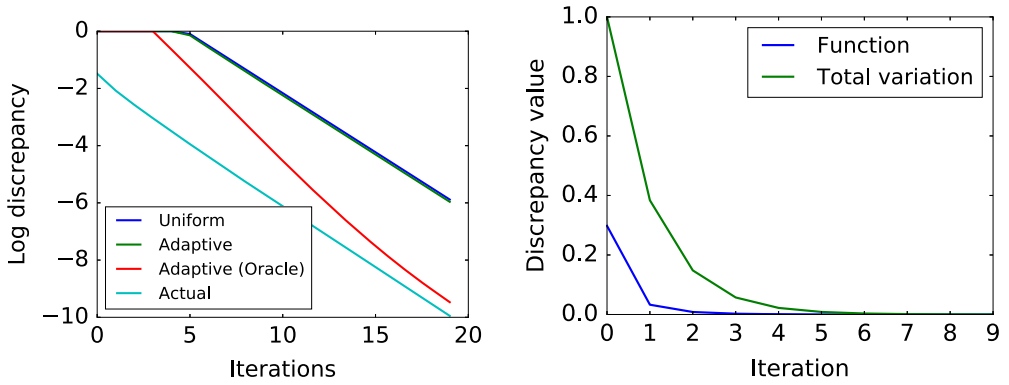


Figure 2: Distribution of f_{65} values under the posterior. Despite the discretization and truncation to a square, it generally matches the one displayed in Figure 1.2 in Robert and Casella (2005).

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 two iterations, compared to the predicted ten iterations from the naive bound $d_f(n) \leq d_{\mathrm{TV}}(n) \leq \frac{1}{\sqrt{\pi_{\min}}} \cdot (1-\gamma^*)^n$. 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.



(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.

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.

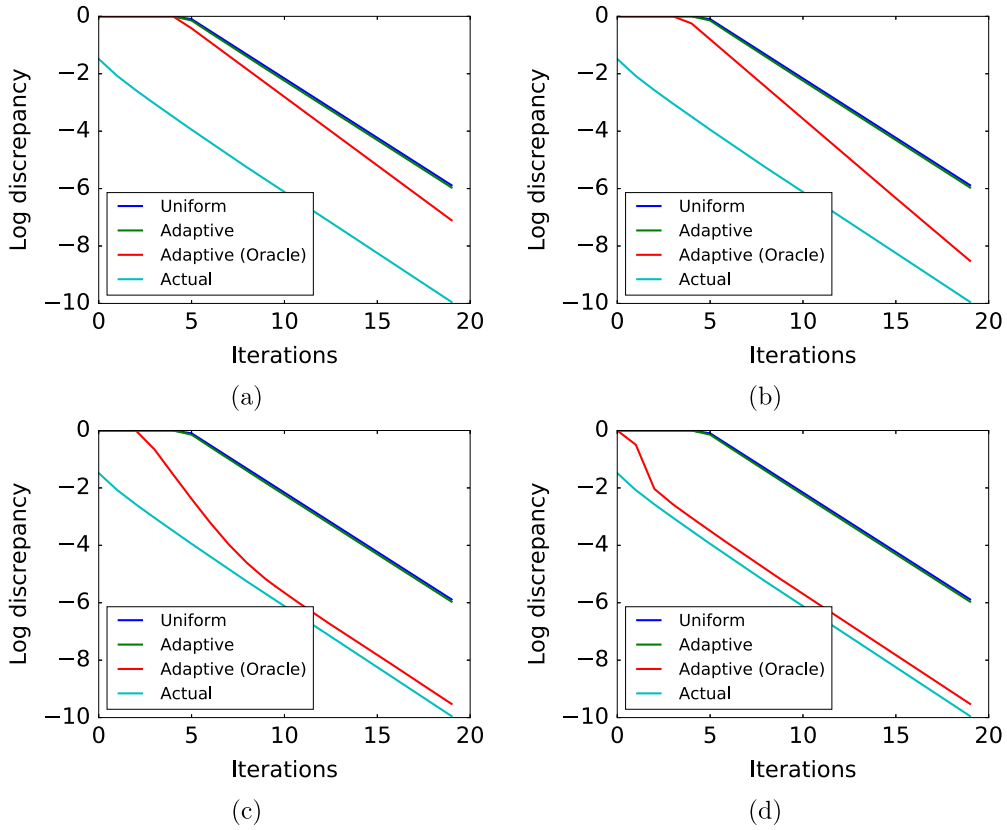


Figure 4: Comparisons of the uniform, non-oracle function-specific, and oracle function-specific bounds for various choices of J. In each case, $J = \{2, \ldots, J_{\max}\}$, with $J_{\max} = 50$ in panel (a), $J_{\max} = 100$ in panel (b), $J_{\max} = 200$ in panel (c), and $J_{\max} = 288$ in panel (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.

4.2 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 (Ghahramani and Griffiths, 2005; Griffiths and Steyvers, 2004; Jain et al., 2007; Mimno et al., 2012; Neal, 2000). We consider the mixture-of-Gaussians model, applying it to a subset of the schizophrenic reaction time data analyzed in Belin and Rubin (1995). 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:

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

We chose relatively uninformative priors, setting $\alpha_0 = \alpha_1 = 1$ and $\rho = 237$. Increasing the value chosen in the original analysis (Belin and Rubin, 1995), 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

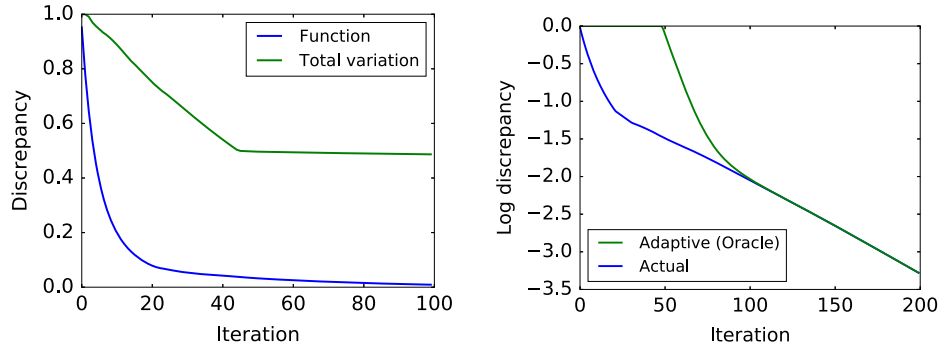
$$z_b^* := \begin{pmatrix} b & b & b & b & b & 1-b & 1-b & 1-b & 1-b & 1-b \end{pmatrix},$$

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 5 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 (12) 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 5 documents this by plotting the f-discrepancy oracle bound against the actual value of d_f on a log scale.

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 (2004). 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



(a) (b)

Figure 5: (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.

Bound type	$T_f(0.01)$	$T_f(10^{-6})$
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 $T_f(\delta)$ for different values of δ. The uniform bound corresponds to the bound $T_f(\delta) \leq T(\delta)$, 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.

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 (2012), but at the cost of a burn-in dependent penalty $d_{\mathrm{TV}}(T_0) = \sup_{\pi_0} d_{\mathrm{TV}}(\pi_n, \pi)$:

$$\mathbb{P}\Big[\frac{1}{N-T_0}\sum_{n=T_0}^{N} f(X_n) \geq \mu + \epsilon\Big] \leq d_{\mathrm{TV}}(T_0) + \exp\Big\{-\frac{\gamma_0}{2(1-\gamma_0)} \cdot \epsilon^2\big[N-T_0\big]\Big\}, \quad (32)$$

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 (32), optimizing T_0 in the range $\big[0,\ 10^5\big]$ for each value of the deviation ϵ. Even given this

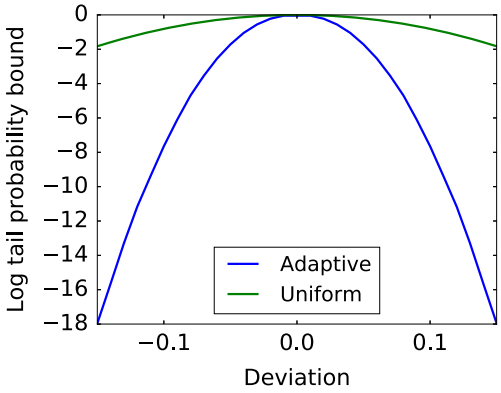


Figure 6: Comparison of the (log) tail probability bounds provided by the uniform Hoeffding bound due to Léon and Perron (2004) with one version of our function-specific Hoeffding bound (Theorem 1). Plots are based on $N = 10^6$ iterations, and choosing the optimal burn-in for the uniform bound and a 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.

generosity toward the uniform bound, the function-specific bound still outperforms it substantially, as Figure 6 shows.

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

5 Discussion

A significant obstacle to successful application of statistical procedures based on Markov chains—especially MCMC—is the possibility of slow mixing. Usually mixing means 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 eliminated the worst-case dependence on the spectral gap of the chain, characteristic of previous results. Our methodology yields sharper confidence intervals, and better rates for sequential hypothesis tests, and we have provided evidence

that our theory's predictions are accurate in some real instances of MCMC and therefore of potential practical interest.

Our investigation also suggests a several further questions, notably concerning 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. 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.

Furthermore, it would be desirable to have methods for estimating or bounding the f-mixing time based on samples. Likewise, 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.

Supplementary Material

Function-Specific Mixing Times and Concentration Away from Equilibrium (Supplementary Material) (DOI: 10.1214/19-BA1151SUPP; .pdf).

References

Aldous, D. and Diaconis, P. (1986). "Shuffling cards and stopping times." *American Mathematical Monthly*, 93(5): 333–348. MR0841111. doi: https://doi.org/10.2307/2323590. 507

Belin, T. R. and Rubin, D. B. (1995). "The analysis of repeated-measures data on schizophrenic reaction times using mixture models." *Statistics in Medicine*, 14(8): 747–768. 525, 526

Choi, H. M. and Hobert, J. P. (2013). "The Polya-Gamma Gibbs sampler for Bayesian logistic regression is uniformly ergodic." *Electronic Journal of Statistics*, 7: 2054–2064. MR3091616. doi: https://doi.org/10.1214/13-EJS837. 509

Chung, K., Lam, H., Liu, Z., and Mitzenmacher, M. (2012). "Chernoff-Hoeffding bounds for Markov chains: Generalized and simplified." In *29th International Symposium on Theoretical Aspects of Computer Science, STACS 2012*, 124–135. MR2909308. 506, 508, 509

Conger, M. and Viswanath, D. (2006). "Riffle shuffles of decks with repeated cards." *The Annals of Probability*, 34(2): 804–819. MR2223959. doi: https://doi.org/10.1214/009117905000000675. 507

Diaconis, P. and Fill, J. A. (1990). "Strong stationary times via a new form of duality." *The Annals of Probability*, 18(4): 1483–1522. MR1071805. 507

Diaconis, P. and Hough, B. (2015). "Random walk on unipotent matrix groups." *arXiv preprint* arXiv:1512.06304. 507

Flegal, J. M., Haran, M., and Jones, G. L. (2008). "Markov chain Monte Carlo: Can we trust the third significant figure?" *Statistical Science*, 23(2): 250–260. MR2516823. doi: https://doi.org/10.1214/08-STS257. 509, 520

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis, Third Edition*. Chapman and Hall/CRC. MR3235677. 505

Ghahramani, Z. and Griffiths, T. L. (2005). "Infinite latent feature models and the Indian buffet process." In *Advances in Neural Information Processing Systems 18: Annual Conference on Neural Information Processing Systems, NIPS 2005*, 475–482. 525

Gillman, D. (1998). "A Chernoff bound for random walks on expander graphs." *SIAM Journal on Computing*, 27(4): 1203–1220. MR1621958. doi: https://doi.org/10.1137/S0097539794268765. 506, 508

Glynn, P. W. and Lim, E. (2009). "Asymptotic validity of batch means steady-state confidence intervals." In *Advancing the Frontiers of Simulation*, 87–104. Springer. 520

Griffiths, T. L. and Steyvers, M. (2004). "Finding scientific topics." *Proceedings of the National Academy of Sciences*, 101(suppl 1): 5228–5235. 525

Gyori, B. M. and Paulin, D. (2012). "Non-asymptotic confidence intervals for MCMC in practice." *arXiv preprint* arXiv:1212.2016. 509, 523

Hayashi, M. and Watanabe, S. (2016). "Information geometry approach to parameter estimation in Markov chains." *The Annals of Statistics*, 44(4): 1495–1535. MR3519931. doi: https://doi.org/10.1214/15-AOS1420. 506, 508

Hsu, D., Kontorovich, A., Levin, D. A., Peres, Y., and Szepesvári, C. (2017). "Mixing Time Estimation in Reversible Markov Chains from A Single Sample Path." *arXiv preprint* arXiv:1708.07367. 509

Hsu, D. J., Kontorovich, A., and Szepesvári, C. (2015). "Mixing time estimation in reversible Markov chains from a single sample path." In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems, NIPS 2015*, 1459–1467. 509

Jain, S., Neal, R. M., et al. (2007). "Splitting and merging components of a nonconjugate Dirichlet process mixture model." *Bayesian Analysis*, 2(3): 445–472. MR2342168. doi: https://doi.org/10.1214/07-BA219. 525

Jones, G. L. and Hobert, J. P. (2001). "Honest exploration of intractable probability distributions via Markov chain Monte Carlo." *Statistical Science*, 16(4): 312–334. MR1888447. doi: https://doi.org/10.1214/ss/1015346317. 509, 520

Joulin, A., Ollivier, Y., et al. (2010). "Curvature, concentration and error estimates for Markov chain Monte Carlo." *The Annals of Probability*, 38(6): 2418–2442. MR2683634. doi: https://doi.org/10.1214/10-AOP541. 506, 508

Kontorovich, A., Weiss, R., et al. (2014). "Uniform Chernoff and Dvoretzky-Kiefer-Wolfowitz-type inequalities for Markov chains and related processes." *Journal of Applied Probability*, 51(4): 1100–1113. MR3301291. doi: https://doi.org/10.1239/jap/1421763330. 506, 508, 509

Léon, C. A. and Perron, F. (2004). "Optimal Hoeffding bounds for discrete reversible Markov chains." *The Annals of Applied Probability*, 14(2): 958–970. MR2052909. doi: https://doi.org/10.1214/105051604000000170. 506, 508, 511, 512, 519, 526, 528

Levin, D. A., Peres, Y., and Wilmer, E. L. (2008). *Markov Chains and Mixing Times*. American Mathematical Society. MR3726904. 505, 507, 509, 512, 516, 517

Lezaud, P. (1998). "Chernoff-type bound for finite Markov chains." *The Annals of Applied Probability*, 8(3): 849–867. MR1627795. doi: https://doi.org/10.1214/aoap/1028903453. 508

Lezaud, P. (2001). "Chernoff and Berry–Esséen inequalities for Markov processes." *ESAIM: Probability and Statistics*, 5: 183–201. MR1875670. doi: https://doi.org/10.1051/ps:2001108. 506, 508, 521

Meyn, S. P. and Tweedie, R. L. (2012). *Markov Chains and Stochastic Stability*. Springer Science & Business Media. MR1287609. doi: https://doi.org/10.1007/978-1-4471-3267-7. 507

Mimno, D. M., Hoffman, M. D., and Blei, D. M. (2012). "Sparse stochastic inference for latent Dirichlet allocation." In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*. 525

Neal, R. M. (2000). "Markov chain sampling methods for Dirichlet process mixture models." *Journal of Computational and Graphical Statistics*, 9(2): 249–265. MR1823804. doi: https://doi.org/10.2307/1390653. 525

Ollivier, Y. (2009). "Ricci curvature of Markov chains on metric spaces." *Journal of Functional Analysis*, 256(3): 810–864. MR2484937. doi: https://doi.org/10.1016/j.jfa.2008.11.001. 507

Paulin, D. (2012). "Concentration inequalities for Markov chains by Marton couplings and spectral methods." *arXiv preprint* arXiv:1212.2015. MR3383563. doi: https://doi.org/10.1214/EJP.v20-4039. 506, 508, 509, 527

Rabinovich, M., Ramdas, A., Jordan, M. I., and Wainwright, M. J. (2019). "Function-Specific Mixing Times and Concentration Away from Equilibrium (Supplementary Material)." *Bayesian Analysis*. doi: https://doi.org/10.1214/19-BA1151SUPP. 510, 513, 516, 518, 519, 520, 521, 522

Robert, C. P. and Casella, G. (2005). *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc. MR1707311. doi: https://doi.org/10.1007/978-1-4757-3071-5. 505, 520, 522, 523

Román, J. C. and Hobert, J. P. (2015). "Geometric ergodicity of Gibbs samplers for Bayesian general linear mixed models with proper priors." *Linear Algebra and its*

Applications, 473: 54–77. Special issue on Statistics. MR3338325. doi: https://doi.org/10.1016/j.laa.2013.12.013. 509

Samson, P.-M. et al. (2000). "Concentration of measure inequalities for Markov chains and Φ-mixing processes." *The Annals of Probability*, 28(1): 416–461. MR1756011. doi: https://doi.org/10.1214/aop/1019160125. 506, 508, 509

Sinclair, A. (1992). "Improved bounds for mixing rates of Markov chains and multicommodity flow." *Combinatorics, Probability, and Computing*, 1(4): 351–370. MR1211324. doi: https://doi.org/10.1017/S0963548300000390. 507

Watanabe, S. and Hayashi, M. (2017). "Finite-length analysis on tail probability for Markov chain and application to simple hypothesis testing." *The Annals of Applied Probability*, 27(2): 811–845. MR3655854. doi: https://doi.org/10.1214/16-AAP1216. 506, 508

Acknowledgments

The authors thank Allan 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.