FUNCTION-SPECIFIC MIXING TIMES AND CONCENTRATION AWAY FROM EQUILIBRIUM

By Maxim Rabinovich, Aaditya Ramdas, Michael I. Jordan, and Martin J. Wainwright

University of California, Berkeley

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. 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., 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, Peres and Wilmer, 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 to uniform


Keywords and phrases: Markov chains, Markov chain Monte Carlo, concentration inequalities, confidence intervals, sequential testing, statistics, probability
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; Joulín 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 functions 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 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 Markov chains that are discrete in space and time, and satisfy the usual conditions of irreducibility, aperiodicity, and reversibility. These conditions guarantee the existence of unique stationary distribution π. The issue is then to understand how quickly the distribution of the Markov chain approaches this stationary distribution.

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

\[
d_{TV}(p, q) = \sup_{f: \Omega \rightarrow [0, 1]} |\mathbb{E}_{X \sim p}[f(X)] - \mathbb{E}_{Y \sim q}[f(Y)]|,
\]

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; it is given by

\[
T(\delta) := \min \{ n > 0 : \max_{i \in [n]} d_{TV}(\pi_n^{(i)}, \pi) \leq \delta \},
\]
where $\pi_n^{(i)}$ denotes 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 measure, to which we refer as an $f$-discrepancy:

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

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

**Definition 1 ($f$-mixing time).** For a given function $f$, the $f$-mixing time is

$$T_f(\delta) = \min \left\{ n > 0 : \max_{i \in [d]} d_f(\pi_n^{(i)}, \pi) \leq \delta \right\}.$$

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., Aldous and Diaconis, 1986; Diaconis and Fill, 1990; Levin, Peres and Wilmer, 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. 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 that 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). These factors, which do not appear in classic bounds for independent random variables, are generally either some variant of the spectral gap $\gamma$ 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 and Perron (2004) shows...
that for a function $f : [d] \to [0, 1]$ and a sample $X_0 \sim \pi$ from the stationary distribution, we have

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

where $\gamma_0 := \min\left(1 - \lambda_2(P), 1\right)$. 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 (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

$$T(\delta) \geq \left( \frac{1}{\gamma_*} - 1 \right) \log\left( \frac{1}{2\delta} \right) \geq \left( \frac{1}{\gamma_0} - 1 \right) \log\left( \frac{1}{2\delta} \right),$$

where $\gamma_* := \min\left(1 - \lambda_2, 1 - |\lambda_d|\right) \leq \gamma_0$ is the absolute spectral gap (Levin, Peres and Wilmer, 2008).

In terms of the minimum probability $\pi_{\min} := \min_i \pi_i$, the corresponding upper bound is an extra factor of $\log\left( \frac{1}{\pi_{\min}} \right)$ larger, which potentially leads to a significant gap between $\frac{1}{\gamma_0}$ and $T(\delta_0)$, 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 study confidence intervals in the context of 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 by applying various bounds on the $f$-mixing time in terms of the spectrum, such as those we prove in Section 2.2.

Throughout the sequel, we use $\mu := \mathbb{E}_\pi[f]$ to denote the mean, and set $\nu := \min(\mu, 1 - \mu)$. Moreover, we follow standard conventions in setting

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

Correspondingly, we define the absolute spectral gap and the (truncated) spectral gap by

$$\gamma_* := 1 - \lambda_*, \quad \text{and} \quad \gamma_0 := 1 - \lambda_0.$$

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 $\pi_0$ whose $f$-discrepancy distance from $\pi$ is small—so that the expectation of each $f(X_n)$ iterate is close to $\mu$—even if its total-variation discrepancy from $\pi$ 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 $\nu := \min(\mu, 1 - \mu)$ and some fixed $\epsilon > 0$, assume $d_f(\pi_0, \pi) \leq \frac{\nu \epsilon^2}{4}$ and $N \geq T_f\left(\frac{\nu \epsilon^2}{4}\right)$. Then

$$\Pr\left[\frac{1}{N} \sum_{n=1}^{N} f(X_n) \geq \mu + \epsilon\right] \leq \exp\left\{-\frac{\epsilon^2 N}{4T_f\left(\frac{\nu \epsilon^2}{4}\right)}\right\}.$$

To gain intuition for this result, we consider a special case that applies whenever the true expectation $\mu$ lies in $\left[\frac{1}{4}, \frac{3}{4}\right]$. In this case, we have the lower bound $\frac{\nu \epsilon^2}{4} \geq \frac{\epsilon^2}{16}$. Therefore, provided that we draw the initial point
\( X_0 \sim \pi_0 \) from some distribution \( \pi_0 \) with \( |\mathbb{E}_{\pi_0}[f(X_0)] - \mu| \leq \frac{\epsilon^2}{16} \), Theorem 1 implies that

\[
\mathbb{P} \left[ \frac{1}{N} \sum_{n=1}^{N} f(X_n) \geq \mu + \epsilon \right] \leq \exp \left\{ - \frac{\epsilon^2 N}{4T_f(\frac{\epsilon^2}{16})} \right\} \quad \text{for all } N \geq T_f(\frac{\epsilon^2}{16}).
\]

Compared to the bounds in earlier work (e.g., Léon and Perron, 2004), the bounds in equations (7) and (8) have several distinguishing features. The primary difference is that the effective sample size \( N/T_f(\epsilon^2/16) \) is a function of \( f \), which can lead to significantly sharper bounds on the deviations of the empirical means than the uniform bounds can deliver. Further, unlike the uniform results, it does 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 \( f \). This condition can be guaranteed to hold simply by running the chain for a burn-in period of \( T_f(\epsilon^2/16) \) steps. Adding the burn-in period does not significantly alter the bound; in fact, it merely reduces the effective sample size by 1.

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 \( \log(\frac{1}{\pi_{\min}}) \) 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, for a fixed function \( f \), there exists a function-specific spectral-gap \( \gamma_f \geq 0 \) such that

\[
T_f(\frac{\epsilon^2}{16}) \approx \frac{1}{\gamma_f} \log \left( \frac{1}{\epsilon^2} \right) + O(1), \quad \text{for } \epsilon \ll 1.
\]

These asymptotics can be used to turn our aforementioned theorem into a variant of the results of Léon and Perron (2004) with \( \gamma_0 \) replaced by a value \( \gamma_f \) that under mild conditions is at least as large as \( \gamma_0 \). While (9) has \( \epsilon^2 \) inside
we note that this is only a small disadvantage as it only introduces at worst a \( \log \left( \frac{1}{\epsilon} \right) \) dependence in the denominator and therefore only loosens the bound by a log factor. 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 \( \delta > 0 \). There we derive more fine-grained concentration inequalities that capture this non-asymptotic behavior.

Intuitively we interpret Theorem 1 as establishing the “effective” sample size \( N_{\text{eff}} = \frac{\gamma_f^N}{\log \left( \frac{1}{\epsilon^2} \right)} \). This intuition 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 better effective sample sizes than the naive asymptotic interpretation predicts. Seen from this perspective, the master bounds come with 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.

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 \( \pi_0 \), then we have

\[
d_{TV}(\pi_n, \pi) \leq \frac{1}{\sqrt{n \lambda_{\min}}} \cdot d_{TV}(\pi_0, \pi).
\]

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

\[
J_f := \{ j : \lambda_j \neq 1 \text{ and } q_j^T f \neq 0 \},
\]
where \( q_j \) denotes a left eigenvector associated with \( \lambda_j \). Similarly, we define

\[
\lambda_f = \max_{j \in J_f} |\lambda_j|, \quad \text{and} \quad \gamma_f = 1 - \lambda_f.
\]

(11b)

Using this notation, it is straightforward to show that if the chain is started with the distribution \( \pi_0 \), then

\[
d_f(\pi_n, \pi) \leq \sqrt{\frac{E_{\pi}[f^2]}{\pi_{\min}}} \cdot \lambda_f^* \cdot d_f(\pi_0, \pi).
\]

(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 \( \lambda_f \) 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 \( J \subseteq J_f \), let us define

\[
\Delta^*_J := 2|J| \times \max_{j \in J} \|h_j\|_{\infty} \times \max_{j \in J} |q_j^T f|, \quad \lambda_J := \max \{|\lambda_j| : j \in J\}, \quad \text{and} \quad \lambda_{-J} := \max \{|\lambda_j| : j \in J_f \setminus J\}.
\]

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

**Lemma 1 (Sharper \( f \)-discrepancy bound).** Given \( f : \mathbb{R} \rightarrow [0, 1] \) and a subset \( J \subseteq J_f \), we have

\[
d_f(\pi_n, \pi) \leq \Delta^*_J \lambda_J^* \cdot d_{TV}(\pi_0, \pi) + \sqrt{\frac{E_{\pi}[f^2]}{\pi_{\min}}} \cdot \lambda_{-J}^* d_f(\pi_0, \pi),
\]

(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 $J$:

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

We call the bound based on this quantity the 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 $f : \mathbb{Z} \to [0, 1]$ and a subset $J \subset J_f$, we have

$$d_f(\pi_n, \pi) \leq |(\pi_0 - \pi)^T h_J(n)| + \sqrt{\frac{E_{\pi} [f^2]}{\pi_{\min}}} \cdot \lambda_{\pi}^n \cdot d_f(\pi_0, \pi),$$

(14)

We emphasize that, although Lemma 2 is stated in terms of the initial distribution $\pi_0$, when we apply the bound in the real examples we consider, we replace all quantities that depend on $\pi_0$ 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 $\gamma_f$ is larger than the absolute spectral gap $\gamma_*$. 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 $\epsilon$ promised in Section 2.1.

**Corollary 1.** Given any $\epsilon > 0$, suppose that $d_f(\pi_0, \pi) \leq \frac{\epsilon^2}{4}$ and $N \geq T_f(\frac{\epsilon^2}{4})$. Then

$$\mathbb{P} \left[ \frac{1}{N} \sum_{n=1}^{N} f(X_n) \geq \mu + \epsilon \right] \leq \begin{cases} \exp \left( -\frac{\epsilon^2}{4} \frac{\gamma_f N \log(\frac{1}{\epsilon^2 \epsilon^2 \pi_{\min}})}{4} \right) & \text{if } \epsilon^2 \leq \frac{4\lambda_f}{\nu \sqrt{\pi_{\min}}} \\ \exp \left( -\frac{\epsilon^2 N}{4} \right) & \text{otherwise.} \end{cases}$$
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 can yield a great improvement 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 $\gamma_f$ 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 $\Delta^*_J$.

**Corollary 2.** Consider constant $\Delta > 0$ and a pair $\Delta_J \geq \Delta^*_J$. Then if $N \geq T_f(\Delta_J + \Delta)$, we have

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

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 5.4.

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

In order to illustrate how 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 (Levin, Peres and Wilmer, 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.
The epitome of a rapidly mixing function is the parity function given by
\[
 f_{\text{parity}}(u) := \begin{cases} 
 1 & \text{if } u \text{ is odd}, \\
 0 & \text{otherwise}.
\end{cases}
\] (17)

It is easy to see that no matter what the choice of initial distribution \(\pi_0\) is, we have \(E[f_{\text{parity}}(X_1)] = \frac{1}{2}\), and thus \(f_{\text{parity}}\) mixes in a single step.

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, Peres and Wilmer, 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 \(0 \leq j \leq d\) because \(f_j\) and \(f_{2d-j}\) behave analogously. Note that the parity function \(f_{\text{parity}}\) corresponds to \(f_d\).

Intuitively, one might expect that some of these functions mix much 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 \(\mu = E_\pi[f_j] = 1/2\). Indeed, we have
\[
 \gamma_{f_j} = \frac{1 - \cos\left(\frac{\pi j}{d}\right)}{2} \geq \begin{cases} 
 \frac{\pi^2 j^2}{24d^2} & \text{if } j \leq \frac{d}{2}, \\
 \frac{1}{2} & \text{if } \frac{d}{2} < j \leq d.
\end{cases}
\] (18)

Consequently, equation (12) predicts that
\[
 T_{f_j}(\delta) \leq \tilde{T}_{f_j}(\delta) = \begin{cases} 
 \frac{24}{\pi^2} \left[\frac{1}{2} \log 2d + \log \left(\frac{1}{\delta}\right)\right] \cdot \frac{d}{2} & \text{if } j \leq \frac{d}{2}, \\
 \log 2d + 2 \log \left(\frac{1}{\delta}\right) & \text{if } \frac{d}{2} < j \leq d.
\end{cases}
\] (19)

where we have used the trivial bound \(E_\pi[f^2] \leq 1\) to simplify the inequalities. Note that this is a significant improvement over \(\asymp d^2\) as soon as \(j \gtrsim \log d\). Moreover, the bound (19) can itself be improved, since each \(f_j\) is orthogonal to all eigenfunctions other than \(1\) 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 $N + \tilde{T}_{f_j}(\epsilon^2/8)$ samples, we obtain the tail bound

\[
\Pr\left[ \frac{1}{N_0} \sum_{n=N_0}^{N+N_b} f_j(X_n) \geq \frac{1}{2} + \epsilon \right] \leq \begin{cases} 
\exp\left( -\frac{6d^2}{\pi^2 j^2} \cdot \frac{\epsilon^2 N}{\log(4\sqrt{2d}/\epsilon^2) + 1} \right) & \text{if } j \leq \frac{d}{2}, \\
\exp\left( -\frac{\epsilon^2 N}{8\left[ \log(4\sqrt{2d}/\epsilon^2) + 1 \right] \cdot \frac{\epsilon^2 N}{8\left[ \log(4\sqrt{2d}/\epsilon^2) + 1 \right]} \right) & \frac{d}{2} < j \leq d,
\end{cases}
\]

where the burn-in time is given by $N_b = \tilde{T}_{f_j}(\epsilon^2/8)$. We point out that the sharper analysis mentioned above would allow us to remove the log $2d$ factors.

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

\[
\{f(u), \ u \in C_{2d}\} \overset{\text{iid}}{\sim} \nu.
\]

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

\[
T_f(\delta) \lesssim \frac{d \log d}{\delta^2} \left[ \log d + \log \left( \frac{1}{\delta} \right) \right], \quad \delta \leq \delta^*.
\]

For $\delta \gg \frac{\log d}{\sqrt{d}}$, this reflects a drastic improvement over the global mixing time of order $d^2 \log \left( \frac{1}{\delta} \right)$.

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

\[
J_\delta = \left\{ 0 \leq j < 2d: j \leq 4\delta \sqrt{\frac{d}{\log d}} \text{ or } j \geq 2d - 4\delta \sqrt{\frac{d}{\log d}} \right\}.
\]

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 $\delta$.

In order to estimate the rate of concentration, we proceed as follows. Taking $\delta = c_0 \epsilon^2$ for a suitably chosen universal constant $c_0 > 0$, we show that $\Delta_j = \epsilon^2/32 \geq \Delta_j^*$. We can then set $\Delta = \epsilon^2/32$ and observe that with high probability over $f$, $\nu \geq \frac{1}{4}$, so that the deviation in Corollary 2 satisfies the bound $2\sqrt{\frac{\Delta_j + \Delta}{\nu}} \leq \epsilon$. With $\delta$ as above, we have $1 - \lambda_{-J} \geq \frac{c_1\epsilon^4}{d \log d}$ for another
universal constant $c_1 > 0$. Thus, if we are given $N + T_f(\epsilon^2/16)$ samples for some $N \geq T_f(\epsilon^2/16)$, then we have

\begin{equation}
\mathbb{P} \left[ \frac{1}{N} \sum_{n=1}^{N+T_f(\epsilon^2/16)} f(X_n) \geq \mu + \epsilon \right] \leq \exp \left\{ - \frac{c_2 \epsilon^6 N}{d \log d \left[ \log \left( \frac{32}{\epsilon^2} \right) + \log d \right]} \right\},
\end{equation}

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

$$N \gtrsim \frac{d \log d \left[ \log (1/\epsilon^2) + \log d \right]}{\epsilon^6},$$

in order to achieve an estimation accuracy of $\epsilon$. Notice that this requirement is an improvement over the $d^2 \epsilon^2$ from the uniform Hoeffding bound provided that $\epsilon \gg (\log d/d)^{1/4}$. Proofs of all these claims can be found in Appendix B.

2.5. Alternative Hoeffding bounds. Although we focus most of our attention in this paper on the master Hoeffding bound in Theorem 1 and its cousin Corollary 1, one can obtain a variation on this master Hoeffding bound, representing different tradeoffs, by making a small change in the argument used to prove Theorem 1. We believe that this alternate bound is of independent interest, and we thus state it here and briefly expound on its properties.

**Theorem 2.** If $d_f(\pi_0, \pi) \leq \frac{\nu}{N}$ and $N \geq T_f(\nu/N)$, then

\begin{equation}
\mathbb{P} \left[ \frac{1}{N} \sum_{n=1}^{N} f(X_n) \geq \mu + \epsilon \right] \leq 3 \exp \left\{ - \frac{\epsilon^2 N}{2T_f(\nu/N)} \right\} \quad \text{for any } \epsilon > 0.
\end{equation}

Intuitively, we interpret Theorem 2 as providing an effective sample size of $N_{\text{eff}} = \frac{\gamma f N}{\log N}$ instead of the analogous quantity $N_{\text{eff}} = \frac{\gamma f N}{\log (1/\epsilon^2)}$ that arises from Theorem 1.

Which version of the master bound to use depends on the application context. In terms of functional form, they represent a clear tradeoff: greater conceptual complexity as a function of $N$—in particular, $\frac{N}{T_f(\nu/N)}$ versus $N$—or of $\epsilon$, more precisely $\epsilon^2$ versus $\epsilon^2/T_f(\nu^2/N)$. These forms lead to very different arguments in theoretical analyses based on the bounds. In terms of concentration predictions, the tradeoff is less clear, since it depends on the relationship between $\epsilon$ and $N$. In the analyses presented in this paper, we
found that the bounds expressed in terms of $\epsilon^2$ were easier to apply than the bounds implied by Theorem 2, but it may be that in some scenarios the reverse will hold.

For completeness, we also state another version of the Hoeffding bound, whose relationship to Theorem 2 is the same as that of Corollary 1 to Theorem 1.

**Corollary 3.** If $d_f(\pi_0, \pi) \leq \frac{\nu}{N}$ and $N \geq T_f(\frac{\nu}{N})$, then

$$\mathbb{P} \left[ \frac{1}{N} \sum_{n=1}^{N} f(X_n) \geq \mu + \epsilon \right] \leq 3 \begin{cases} \exp \left\{ -\frac{\epsilon^2}{2} \cdot \frac{\gamma_f N}{\log \left( \frac{N}{\nu \sqrt{\pi \min}} \right)} \right\} & \text{if } N \geq \frac{\nu \sqrt{\pi \min}}{\lambda_f}, \\ \exp \left( -\frac{\epsilon^2 N}{2} \right) & \text{otherwise.} \end{cases}$$

for every $\epsilon > 0$.

The proofs of the aforementioned theorem and corollary are in Sections 5.1 and 5.4 respectively.

### 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 $\mathbb{E}_\pi [f]$ does not suffice; the uncertainty in the estimate must be quantified, for instance by providing confidence intervals. 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 $\pi_0$ 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, Thm. 1). As one would expect, it gives contraction at the usual Hoeffding rate but with an effective sample size of $N_{\text{eff}} = \gamma_0 (N - T_0)$, where $\gamma_0 = \min\{1 - \lambda_2, 1\}$ and $T_0$ 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; likewise, no matter how favorable the mixing rate of $f$ in the regime of interest, the effective sample size remains bounded by the global spectral parameter $\gamma_0$.

To make this precise, for a fixed burn-in level $\alpha_0 \in (0, \alpha)$, let us define $\epsilon_N(\alpha, \alpha_0) := \sqrt{2(2 - \gamma_0)} \cdot \sqrt{\frac{\log \left( \frac{2}{\alpha - \alpha_0} \right)}{\gamma_0 (N - T(\alpha_0))}}$. Then the uniform Markov Hoeffding bound (Léon and Perron, 2004, Thm. 1) implies that the interval

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

is a $1 - \alpha$ confidence interval.

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 (26) then implies that

$$I^*_N(\alpha) = \left[ \frac{1}{N - T(\alpha_0)} \sum_{n=T(\alpha_0/2)+1}^{N} f(X_n) \pm \epsilon^*_N(\alpha) \right]$$

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

Let us now state one result that can be derived using our Hoeffding bounds. The resulting confidence 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 $1/\gamma_0$ 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 $e^{-r_N(\epsilon)/4}$, where

$$r_N(\epsilon) := \epsilon^2 \left\{ \frac{N}{T_f(\frac{\mu^2}{4})} - 1 \right\}.$$

If we wish for each tail to have probability mass that is less than $\alpha/2$, we need to choose $\epsilon > 0$ so that $r_N(\epsilon) \geq 4 \log \frac{2}{\alpha}$, and conversely any such $\epsilon$ corresponds to a valid two-sided $(1 - \alpha)$ confidence interval. Let us summarize our conclusions:
Theorem 3. For any width $\epsilon_N \in r_N^{-1}\left(\left[4 \log \left(2/\alpha\right), \infty\right)\right)$, the set

$$I_N := \left[1/N - T_f\left(\frac{\nu^2}{4}\right) \sum_{n=T_f\left(\frac{\nu^2}{4}\right)}^N f(X_n) \pm \epsilon_N\right]$$

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

$$(28) \quad r_N(\epsilon) \geq \epsilon^2 \left[\frac{N}{\bar{T}_f\left(\frac{\nu^2}{4}\right)} - 1\right] \quad \text{valid for all } \epsilon \geq \eta.$$

Consequently, whenever $r_{N,\eta}(\epsilon_N) \geq 4 \log \frac{2}{\alpha}$ and $\epsilon_N \geq \eta$, then we are guaranteed that a symmetric interval of half-width $\epsilon_N$ is a valid $(1 - \alpha)$-confidence interval. Summarizing more precisely, we have:

Corollary 4. Fix $\eta > 0$ and let

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

If $N \geq T_f\left(\frac{\nu^2}{4}\right)$, then $I_N$ is a $1 - \alpha$ confidence interval for $\mu = \mathbb{E}_\pi[f]$.

In many situations we will 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 5.5, therefore, which contains the proofs for this section, we prove a strengthened form of Theorem 3 and its corollary that applies in that setting.

A popular alternative strategy for building confidence intervals using MCMC depends on the Markov central limit theorem (e.g., (Flegal, Haran and Jones, 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

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

where $\sigma_{f,\text{asym}}^2 := \lim_{N \to \infty} \frac{1}{N} \text{Var}_{X_0 \sim \pi} \left[\sum_{n=1}^N f(X_n)\right]$ is the so-called 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

\[ \tilde{S}_N = \sum_{n=1}^{N} [f(X_n) - \mu]. \]

We then have the bound

\[ \left| \mathbb{P}\left( \frac{\tilde{S}_N}{\sigma_{f,\text{asym}} \sqrt{N}} \leq s \right) - \Phi(s) \right| \leq \frac{e^{-\gamma_0 N}}{3\sqrt{\pi_{\text{min}}}} + \frac{13}{\sigma_{f,\text{asym}} \sqrt{\pi_{\text{min}}}} \cdot \frac{1}{\gamma_0 \sqrt{N}}, \]

where \( \Phi \) is the standard normal CDF. Note that this bound accounts for both the non-stationarity error and for non-normality error at stationarity. The non-normality term decays far more slowly than the non-stationarity term, going roughly at the rate \( \frac{1}{\gamma_0 \sqrt{N}} \) while the latter goes at the rate \( e^{-\gamma_0 N} \).

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

More precisely, to be valid, the interval requires a large sample size bounded below as

\[ N \geq \max \left( \log \left( \frac{2}{\sqrt{\pi_{\text{min}} \alpha}} \right) \cdot \frac{1}{\gamma_0}, \frac{6084}{\sigma_{f,\text{asym}}^2 \pi_{\text{min}} \alpha^2} \cdot \frac{1}{\gamma_0^2} \right), \]

and, when it is valid, it takes the form

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

It is worth pointing out that this confidence interval width contains a hidden mixing penalty. Indeed, defining the variance \( \sigma_f^2 = \text{Var}_\pi[f(X)] \) and \( \rho_f := \frac{\sigma_f^2}{\sigma_{f,\text{asym}}^2} \), we can rewrite the width as

\[ \epsilon_N = \sigma_f \sqrt{\frac{2 \log (6/\alpha)}{\rho_f N}}. \]
Thus, for this bound, the quantity $\rho_f$ captures the penalty due to non-independence, playing the role of $\gamma_0$ and $\gamma_f$ 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., $\gamma_0$).

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 (2015), and we consider the same problem settings that they did:

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

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

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

For the first setting, we always assume $0 < \delta < \nu$, 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 $\mu$ falls in the indifference region $(r - \delta, r + \delta)$. The error of a procedure $A$ can thus be defined as

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

The rest of this subsection is organized as follows. For the first setting above, we analyze a procedure $A_{\text{fixed}}$ that makes a decision after a fixed number $N := N(\alpha)$ of samples. We also analyze a sequential procedure $A_{\text{seq}}$ that chooses whether to reject at a sequence $N_0, \ldots, N_k, \ldots$ of decision times. For the second, more challenging, setting, we analyze $A_{\text{hard}}$, which also rejects at a sequence of decision times. For both $A_{\text{seq}}$ and $A_{\text{hard}}$, 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 $\alpha$. After collecting $N$ samples, it forms the empirical average $\hat{\mu}_N = \frac{1}{N} \sum_{n=1}^{N} f(X_n)$ 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

\begin{align}
N_0 &= \lceil M \cdot \min\left( \frac{1}{r}, \frac{1}{1-r} \right) \rceil, \\
N_k &= \lceil N_0 (1 + \xi)^k \rceil,
\end{align}

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

\[ \hat{\mu}_{N_k} \in \left( r - \frac{M}{N_k}, r + \frac{M}{N_k} \right). \]

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, 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

\[ \hat{\mu}_{N_k} \in \left( r - \epsilon_k(\alpha), r + \epsilon_k(\alpha) \right). \]

When this inclusion holds, the algorithm continues; when it doesn’t hold, the algorithm outputs $H_0$ or $H_1$ in the natural way.

In order to state our main result, first define

\[ \tau_f = \tau_f(\delta) = 2T_f(\nu \delta^2). \]

The following result is restricted to the stationary case; later in the section, we turn to the question of burn-in.

---

1In Gyori and Paulin (2015), the authors set $N_0 = \lceil \frac{100}{\gamma_0} \rceil$, but this is inessential.
Theorem 4. Assume that $\alpha \leq \frac{2}{5}$. For $A_{\text{fixed}}, A_{\text{seq}}, A_{\text{hard}}$ to all satisfy $\text{err}(A, f) \leq \alpha$, it suffices to (respectively) choose

\begin{align*}
N &= \frac{\tau_f(\delta) \log \left( \frac{1}{\alpha} \right)}{\delta^2}, \quad \text{(35)} \\
M &= \frac{2\tau_f\left( \frac{\delta}{2} \right) \log \left( \frac{2}{\sqrt{\alpha\epsilon}} \right)}{\delta}, \quad \text{and} \quad \text{(36)} \\
\epsilon_k(\alpha) &= \inf \left\{ \epsilon > 0 : \frac{\epsilon^2}{2\tau_f\left( \frac{\delta}{2} \right)} \geq \frac{\log \left( \frac{1}{\alpha} \right) + 1 + 2 \log k}{N_k} \right\}, \quad \text{(37)}
\end{align*}

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

Our results differ from those of Gyori and Paulin (2015) because the latter implicitly control the worst-case error of the algorithm

$\text{err}(A) = \sup_{f : \Omega \rightarrow [0, 1]} \text{err}(A, f),$

while our analysis controls $\text{err}(A, f)$ directly. For example, among other results, Gyori and Paulin (2015) showed that $A_{\text{fixed}}$ with $N = \frac{\log(1/\alpha)}{\gamma_0 \delta}$ has worst-case error at most $\text{err}(A_{\text{fixed}}) \leq \alpha$. The $\tau_f$ parameter in our bounds plays the same role in our bounds that $\frac{1}{\gamma_0}$ plays in their uniform bounds.

In particular, they showed that choosing $M = \frac{\log(\frac{\sqrt{\lambda}}{\delta})}{\gamma_0 \delta}$ suffices to guarantee $\text{err}(A_{\text{seq}}) \leq \alpha$. Similarly, for the no indifference case, they proved that the choice

\begin{equation}
\epsilon_k(\alpha) = \sqrt{\frac{\log(1/\alpha) + 1 + 2 \log k}{\gamma_0 N_k}} \quad \text{(38)}
\end{equation}

leads to a procedure with worst-case error bounded as $\text{err}(A_{\text{hard}}) \leq \alpha$. Once more, our improvements are of a similar nature.

As a result of this close correspondence, we easily see that our results improve on the uniform result for a fixed function $f$ whenever $f$ converges to its stationary expectation faster than the chain itself converges—more concretely, whenever $\frac{\tau_f}{\gamma_0} = T_f(\nu \delta^2) \leq \frac{1}{2\gamma_0}$.

The value of the above testing results depends substantially on their computational properties. In the indifference region case, for example, the sequential procedure will only be valuable if it can reduce the number of samples needed compared to the fixed sample size procedure. In the no indifference case, the procedure is valuable because of its ability to test between hypotheses separated only by a point, but its utility will be greatly
limited if it takes too long to run. We now turn, therefore, to the question
of bounding expected stopping times.

To carry out the analysis, we introduce the true margin \( \Delta = |r - \mu| \). First,
let us introduce the following notation. Let \( N(A) \) be the number of sampled
collected by \( A \). Given a margin schedule \((\epsilon_k)\), let
\[
k_0^*(\epsilon_{1:\infty}) = \min \{ k \geq 1 : \epsilon_k \leq \frac{\Delta}{2} \},
\]
\[
N_0^*(\epsilon_{1:\infty}) = N_{k_0^*}^{\epsilon_{1:\infty}}.
\]

We can bound the expected stopping times of \( A_{\text{seq}}, A_{\text{hard}} \) in terms of \( \Delta \) as follows:

**Theorem 5.** Assume either \( H_0 \) or \( H_1 \) holds. Then,
\[
\mathbb{E}[N(A_{\text{seq}})] \leq (1 + \xi) \left[ \frac{M}{\Delta} + \frac{2}{\Delta} \sqrt{\frac{2\tau_f(\delta/2)M}{\Delta}} + 8\tau_f(\delta/2) + 1 \right].
\]
\[
\mathbb{E}[N(A_{\text{hard}})] \leq (1 + \xi) \left( N_0^* + 1 \right) + \frac{16\alpha\tau_f(\nu\Delta^2/16)}{\Delta^2}.
\]

With minor modifications to the proofs in Gyori and Paulin (2015), we
can bound the expected stopping times as
\[
\mathbb{E}[N(A_{\text{seq}})] \leq (1 + \xi) \left\{ \frac{M}{\Delta} + \frac{2}{\Delta} \sqrt{\frac{M}{\gamma_0\Delta} + \frac{4}{\gamma_0} + 1} \right\},
\]
\[
\mathbb{E}[N(A_{\text{hard}})] \leq (1 + \xi) \left( N_0^* + 1 \right) + \frac{4\alpha}{\gamma_0\Delta^2}.
\]

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 \( \Delta \geq \delta \).
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 he ratio between the \( M \) values—viz.
\[
\frac{M_{\text{adapt}}}{M_{\text{unif}}} \approx \gamma_0\tau_f \approx \gamma_0 T_f(\nu\Delta^2/4).
\]

As a result, we should expect a significant improvement in terms of number of
samples when the relaxation time \( T_f(\nu\Delta^2/4) \) is significantly larger than the \( f \)-mixing
time \( T_f(\nu\Delta^2/4) \). Framed in absolute terms, we can write
\[
\tilde{N}_{\text{unif}}(A_{\text{seq}}) \approx \frac{\log (2/\sqrt{\alpha\xi})}{\gamma_0\Delta} \quad \text{and} \quad \tilde{N}_{\text{adapt}}(A_{\text{seq}}) \approx \frac{\tau_f(\delta/2)\log (2/\sqrt{\alpha\xi})}{\delta\Delta}.
\]
Up to an additive term, the bound for $A_{\text{hard}}$ is also qualitatively similar to earlier ones, with $\frac{1}{\Delta^2}$ replaced by $\frac{1}{\Delta}$.

![Graphs showing eigenvalue spectra for different MCMC chains](image)

**Fig 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.

### 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 go part of the way toward closing 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 (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 $(\alpha, \beta) \in \mathbb{R}^2$ 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:

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

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

which corresponds to an exponential prior on $e^\alpha$, an improper uniform prior on $\beta$ 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

$$\Sigma = \begin{pmatrix} 4 & 0 \\ 0 & 10 \end{pmatrix}.$$ 

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

$$\Omega = \left\{ (\hat{\alpha} \pm i \cdot \Delta, \hat{\beta} \pm j \cdot \Delta) \mid 0 \leq i, j \leq 8 \right\},$$

where $\Delta = 0.1$ and $(\hat{\alpha}, \hat{\beta})$ is the MLE. This space has $d = 17^2 = 289$ elements, resulting in a $289 \times 289$ 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

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

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

$$d_f(n) \leq d_{TV}(n) \leq \frac{1}{\sqrt{\min_n}} \cdot (1 - \gamma^*)^n.$$
Fig 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).

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 Figure 4, 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 5 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 (Berry et al., 2010; Gyori and Paulin, 2012).

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:

$$
P(X_i = 1 \mid Y_i = 0) = \gamma_0,
$$

$$
P(X_i = 1 \mid Y_i = 1) = \gamma_1,
$$

$$
P(X_i = 1 \mid Y_i \text{ missing}) = p.
$$

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 $p = P(X_i = 1)$
Fig 5. Comparisons of the uniform, non-oracle function-specific, and oracle function-specific bounds for various choices of \( J \). In each case, \( J = \{2, \ldots, J_{\text{max}}\} \), with \( J_{\text{max}} = 50 \) in panel (a), \( J_{\text{max}} = 100 \) in panel (b), \( J_{\text{max}} = 200 \) in panel (c), and \( J_{\text{max}} = 288 \) in panel (d). The oracle bound becomes tight in the limit as \( J_{\text{max}} \) goes to \( d = 289 \), but it offers an improvement over the uniform bound across the board. 

unconditionally; we keep the other notation, however, for the sake of consistency with past work (Berry et al., 2010; Gyori and Paulin, 2012). Conjugate uniform (i.e., \( \text{Be}(1, 1) \)) priors are placed on all the model parameters.

The unknown variables include the parameter triple \((\gamma_0, \gamma_1, p)\) and the unobserved \( X_i \) values for \( n_Y + n_0 = 30 \) patients, and the full sample space is therefore \( \Omega = [0, 1]^3 \times \{0, 1\}^{30} \). 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 tran-
sition 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, $f(\gamma_0, \gamma_1, p) = 1(p > 0.5)$—depends only on $p$, we fix $\gamma_0$ and $\gamma_1$ 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 $\frac{1}{\sqrt{\pi_{\text{min}}}} \approx 2.14 \times 10^{33}$). 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 $f_1(p) := 1(p > 0.5), \text{ and } f_2(p) := p,$

we see in Figure 6 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 than that (by about an order of magnitude) for the first few iterations, before reaching the asymptotic rate dictated by the value of 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 (Ghahramani and Griffiths, 2005; Griffiths and Steyvers,
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, viz.:

$$
\mu_b \sim \mathcal{N}(0, \rho^2), \quad b = 0, 1,
\omega \sim \text{Be}(\alpha_0, \alpha_1)
$$

$$
Z_i \mid \omega \sim \text{Bern}(\omega)
$$

$$
X_i \mid Z_i = b, \mu \sim \mathcal{N}(\mu_b, \sigma^2)
$$

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 $\omega$ 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 $\lambda_2$, opening up the possibility for improvement over the uniform $\gamma_*$-based bounds. In more detail, define the vector

$$
z^*_b := (b \ b \ b \ b \ 1-b \ 1-b \ 1-b \ 1-b \ 1-b \ 1-b),
$$

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

$$
f(z) = 1(z \in \{z^*_0, z^*_1\}).
$$

As Figure 7 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 $J = \{1, \ldots, 25\}$; this corresponds to making exact predictions of $T_f(\delta)$ for $\delta \leq \delta_0 \approx 0.01$, which is a realistic tolerance for estimation of $\mathbb{E}[f]$. Panel (b) of Figure 7 documents this by plotting the $f$-discrepancy oracle bound against the actual value of $d_f$ on a log scale.

Fig 7. (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.

![Comparison of bounds on $T_f(\delta)$ for different values of $\delta$. 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 $\delta$ as large as 0.01.](image)

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 Per-
Ron (2004). At equilibrium, the penalty for non-independence in our bounds is \((2T_f(\nu e^2/4))^{-1}\) compared to roughly \(\gamma^{-1}\) 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 \(T_f(\nu e^2/4)\), whereas the uniform Hoeffding bound does not directly apply for any finite burn-in. This issue can be addressed using the method of Paulin (2012), but at the cost of a burn-in dependent penalty \(d_{\text{TV}}(T_0) = \sup_{\pi_0} d_{\text{TV}}(\pi_n, \pi)\):

\[
\mathbb{P}\left[\frac{1}{N - T_0} \sum_{n=T_0}^N f(X_n) \geq \mathbb{E}_\pi[f] + \epsilon\right] \leq d_{\text{TV}}(T_0) + \exp\left\{ -\frac{\gamma_0}{2(1 - \gamma_0)} \cdot \epsilon^2 \left(\frac{N}{T_0} - 1\right)\right\},
\]

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 (42), optimizing \(T_0\) in the range \([0, 10^5]\) for each value of the deviation \(\epsilon\). Even given this generosity toward the uniform bound, the function-specific bound still outperforms it substantially, as Figure 8 shows. For the function-specific bound, we used the function-specific oracle bound (Lemma 2) to bound \(T_f(\nu e^2/4)\); this nearly coincides with the true value when \(\epsilon \approx 0.01\) but deviates slightly for larger values of \(\epsilon\).

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

5.1. Proofs of master Hoeffding bounds. The proofs of our master Hoeffding bounds (Theorems 1 and 2) are based on the following more general result:

**Proposition 1 (Master tail bound).** Suppose that \(d_f(\pi_0, \pi) \leq \delta\) and 
\(N_0(\delta) := \frac{N}{T_f(\delta)} \geq 1\). Then for every \(\alpha > 0\), we have

\[
\mathbb{P}\left[\frac{1}{N} \sum_{n=1}^N f(X_n) \geq \mu + \epsilon\right] \leq \exp\left\{ \left[\frac{\delta}{\nu} - \alpha T_f(\delta) \epsilon + \frac{1}{2} \alpha^2 T_f^2(\delta)\right] N_0(\delta)\right\},
\]
Fig 8. 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 $409 \geq T_f(10^{-6})$ iterations for the function-specific bound. The function-specific bound improves over the uniform bound by orders of magnitude.

5.1.1. Proof of Theorem 1. Take $\delta = \frac{\epsilon^2 \nu}{4}$ and $\alpha = \epsilon$. Proposition 1 then guarantees that

$$
P \left[ \frac{1}{N} \sum_{n=1}^{N} f(X_n) \geq \mu + \epsilon \right] \leq \exp \left\{ \frac{\epsilon^2}{4} \cdot \left[ - \epsilon^2 + \frac{1}{2} \epsilon^2 \right] \cdot N_0 \right\}
$$

$$
\leq \exp \left\{ - \frac{\epsilon^2 N_0}{4} \right\}
$$

$$
= \exp \left\{ - \frac{\epsilon^2 N}{4T_f(\frac{\epsilon^2 \nu}{4})} \right\}.
$$

5.1.2. Proof of Theorem 2. Take $\delta = \frac{N \nu}{N}$ and $\alpha = \epsilon$. The bound in Proposition 1 then yields

$$
P \left[ \frac{1}{N} \sum_{n=1}^{N} f(X_n) \geq \mu + \epsilon \right] \leq \exp \left\{ 1 - \frac{1}{2} \epsilon^2 N_0 \right\}
$$

$$
= \exp \left\{ 1 - \frac{\epsilon^2 N}{2T_f(\frac{\nu}{N})} \right\}
$$

$$
\leq 3 \exp \left\{ - \frac{\epsilon^2 N}{2T_f(\frac{\nu}{N})} \right\}.
$$

5.2. Proof of Proposition 1. At the heart of the proof is the following bound on the 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 \( \tilde{X}_{m,t} := X_{(m-1)T_f(\delta)+t} \). With this notation, we have

**Lemma 3 (Master MGF bound).** For \( 0 \leq t < T_f(\delta) \) and \( \alpha \in \mathbb{R} \),

\[
(44) \quad \mathbb{E} \left[ \exp \left( \alpha \sum_{m=1}^{N_0} f(\tilde{X}_{m,t}) \right) \right] \leq \exp \left\{ \left[ \frac{\delta}{\nu} + \alpha \mu + \frac{1}{2} \alpha^2 \right] \cdot N_0(\delta) \right\}.
\]

See Section 5.3 for the proof of this claim.

Using Lemma 3, we now prove the proposition. Recalling the definition of \( \tilde{X}_{m,t} \), we have

\[
\mathbb{E} \left[ e^{\alpha \sum_{n=1}^{N} f(X_n)} \right] = \mathbb{E} \left[ \exp \left( \alpha \sum_{m=1}^{N_0} f(\tilde{X}_{m,t}) \right) \right]
= \mathbb{E} \left[ \exp \left( \alpha T_f(\delta) \sum_{t=1}^{T_f(\delta)} \sum_{m=1}^{N_0} f(\tilde{X}_{m,t}) \right) \right]
\leq \frac{1}{T_f(\delta)} \sum_{t=0}^{T_f(\delta)-1} \mathbb{E} \left[ \exp \left( \alpha T_f(\delta) \sum_{m=1}^{N_0} f(\tilde{X}_{m,t}) \right) \right],
\]

where the last inequality follows from Jensen’s inequality, as applied to the exponential function.

Combined with Lemma 3, we conclude

\[
\mathbb{E} \left[ e^{\alpha \sum_{n=1}^{N} f(X_n)} \right] \leq \exp \left\{ \left[ \frac{\delta}{\nu} + \alpha T_f(\delta) \mu + \frac{1}{2} \alpha^2 T_f(\delta)^2 \right] \cdot N_0 \right\},
\]
valid for \( \alpha > 0 \). The claim then follows from the MGF Markov inequality.

5.3. **Proof of Lemma 3.** For any given scalar \( t \in (0, T_f(\delta)) \), we again make use of the shorthand \( \tilde{X}_{m,t} := X_{(m-1)T_f(\delta)+t} \). We then introduce a sequence of Bernoulli variables \( Y_m \sim \text{Bern}(f(\tilde{X}_{m,t})) \). With this notation, Jensen’s inequality implies that

\[
\mathbb{E} \left[ \exp \left( \alpha \sum_{m=1}^{N_0} f(\tilde{X}_{m,t}) \right) \right] \leq \mathbb{E} \left[ \exp \left( \alpha \sum_{m=1}^{N_0} Y_m \right) \right],
\]
so that it suffices to bound the latter MGF.
In order to do so, we first argue that for all integers \( m = 1, \ldots, N_0 - 1 \), we have

\[
\mathbb{P}[Y_{m+1} = b_{m+1} \mid Y_{1:m} = b_{1:m}] \leq (1 + \frac{\delta}{\nu}) \mu^{b_{m+1}}(1 - \mu)^{1-b_{m+1}}.
\]

In order to establish this fact, note that, if we let \( \rho(b_{1:m}) \) denote the distribution of \( \tilde{X}_{m+1,t} \) given that \( Y_{1:m} = b_{1:m} \),

\[
\mathbb{P}(Y_{m+1} = 1 \mid Y_{1:m} = b_{1:m}) = E_{\tilde{X}_{m+1,t} \sim \rho(b_{1:m})} \left[ f(\tilde{X}_{m+1,t}) \mid \tilde{X}_{m,t} \right] 
\leq \mu + \delta
\leq (1 + \frac{\delta}{\nu}) \mu
\]

by the definition of \( T_f(\delta) \). A similar argument can be given for \( b_{m+1} = 0 \). Finally note that

\[
\mathbb{P}[Y_1 = 1] \leq \mu + \delta \leq (1 + \frac{\delta}{\nu}) \mu.
\]

With this in hand, we observe

\[
\mathbb{E}\left[e^{\alpha \sum_{m=1}^{N_0} Y_m} \right] = \sum_{b_1,N_0} \mathbb{P}(Y_1 = b_1) \cdot \prod_{m=1}^{N_0-1} \mathbb{P}(Y_{m+1} = b_{m+1} \mid Y_{1:m} = b_{1:m}) \cdot e^{\alpha \sum_{m=1}^{N_0} b_m}
\leq (1 + \frac{\delta}{\nu})^{N_0} \cdot \sum_{b_1,N_0} \mathbb{P}(Y_{1:m} = b_{1:m}) \cdot e^{\alpha \sum_{m=1}^{N_0} b_m}
= (1 + \frac{\delta}{\nu})^{N_0} \cdot \mathbb{E}\left[e^{\alpha \sum_{m=1}^{N_0} Y'_m} \right],
\]

where the new sequence \( \{Y'_m\}_{m=1}^{N_0} \) consists of i.i.d. \( \text{Bern}(\mu) \) variables. By the sub-Gaussianity of the \( Y_m \)-variables, we deduce that

\[
\mathbb{E}\left[e^{\alpha \sum_{m=1}^{N_0} Y_m} \right] \leq (1 + \frac{\delta}{\nu})^{N_0} \cdot e^{\alpha \mu N_0 + \frac{1}{2} \alpha^2 N_0}
\leq \exp \left( \frac{\delta}{\nu} + \alpha \mu + \frac{1}{2} \alpha^2 \right) N_0,
\]

which completes the proof.

5.4. Proofs of derived Hoeffding bounds. In this section, we prove the derived Hoeffding bounds.
5.4.1. Proof of Corollary 1. The proof is a direct application of Theorem 1. Indeed, it suffices to note that if $\epsilon^2 \leq \frac{4\lambda_f}{\nu \sqrt{\pi_{\min}}}$, then
\[
T_f\left(\frac{\epsilon^2 \nu}{4}\right) \leq \frac{\log \left(\frac{4}{\epsilon^2 \nu \sqrt{\pi_{\min}}}\right)}{\log \left(\frac{1}{\lambda_f}\right)} = \frac{\log \left(\frac{4}{\epsilon^2 \nu}\right) + \log \left(\frac{1}{\nu}\right) + \frac{1}{2} \log \left(\frac{1}{\pi_{\min}}\right)}{\log \left(\frac{1}{\lambda_f}\right)},
\]
which yields the first bound. Turning to the second bound, note that if $\epsilon^2 > \frac{4\lambda_f}{\nu \sqrt{\pi_{\min}}}$, then equation (12) implies that $T_f\left(\frac{\epsilon^2 \nu}{4}\right) = 1$, which establishes the claim.

5.4.2. Proof of Corollary 2. The proof combines Theorem 1 with Lemma 1. Notice first that we have set $\epsilon = 2\sqrt{\frac{\Delta + \Delta^*}{\nu}}$. On the other hand, using the facts that $\lambda_f \leq 1$, $d_{TV}(\pi_0, \pi_n) \leq 1$, $d_f(\pi_0, \pi_n) \leq 1$, and that $E_{\pi} \left[f^2\right] \leq 1$, we have from Lemma 1 that
\[
d_f(\pi_0, \pi_n) \leq \Delta^* + \frac{\lambda_{o,f}}{\sqrt{\pi_{\min}}} \leq \Delta + \frac{\lambda_{o,f}}{\sqrt{\pi_{\min}}}.
\]
It follows that, provided $\Delta \leq \frac{\lambda_{o,f}}{\sqrt{\pi_{\min}}}$,
\[
T_f\left(\frac{\nu \epsilon^2}{4}\right) = T_f(\Delta_f + \Delta) \leq \frac{\log \left(\frac{1}{\Delta}\right) + \frac{1}{2} \log \left(\frac{1}{\pi_{\min}}\right)}{\log \left(\frac{1}{\lambda_{o,f}}\right)}.
\]
Plugging into Theorem 1 now gives the first part of the bound. On the other hand, if $\Delta > \frac{\lambda_{o,f}}{\sqrt{\pi_{\min}}}$, then Lemma 1 implies that $T_f(\Delta_f + \Delta) = 1$, giving the second case.

5.4.3. Proof of Corollary 3. The proof is a direct application of Theorem 2. Indeed, it suffices to note that if $N \geq \frac{\nu \sqrt{\pi_{\min}}}{\lambda_f}$, then
\[
T_f\left(\frac{\nu}{N}\right) \leq \frac{\log \left(\frac{1}{\nu}\right) + \log N + \frac{1}{2} \log \left(\frac{1}{\pi_{\min}}\right)}{\log \left(\frac{1}{\lambda_f}\right)}
\]
by equation (12). Plugging into the bound gives the first inequality, while the second follows from the fact that
\[
\log \left(\frac{1}{\lambda_f}\right) \geq 1 - \lambda_f = \gamma_f, \quad 0 \leq \lambda_f \leq 1.
\]
On the other hand, if $N \leq \frac{\nu \sqrt{\pi_{\min}}}{\lambda_f}$, then $T_f\left(\frac{\nu}{N}\right) = 1$, which leads to the improved concentration in the second statement.
5.5. Proofs of confidence interval results. Here we provide the proof of the confidence interval corresponding to our bound (Theorem 3). Proofs of the claims (26) and (30) can be found in Appendix C.

As discussed in Section 3.1, we actually prove a somewhat stronger form of Theorem 3, 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 $\tilde{T}_f : \mathbb{N} \to \mathbb{R}_+$ is an upper bound on $T_f$ and note that the corresponding tail bound becomes $e^{-\tilde{r}_N(\epsilon)/4}$, where

$$\tilde{r}_N(\epsilon) = \epsilon^2 \left[ \frac{N}{\tilde{T}_f(\nu^2/4)} - 1 \right].$$

This means that, just as before we wanted to make the rate $r_N$ in (27) at least as large as $4 \log \frac{2}{\alpha}$, we now wish to do the same with $\tilde{r}_N(\epsilon_N)$, which means choosing $\epsilon_N$ with $\tilde{r}_N(\epsilon_N) \geq 4 \log \frac{2}{\alpha}$. We therefore have the following result.

**Proposition 2.** For any width $\epsilon_N \in \tilde{r}_N^{-1}\left(\left[4 \log \left(\frac{2}{\alpha}\right), \infty\right]\right)$, the interval

$$I_N = \left[ \frac{1}{N - \tilde{T}_f(\nu^2/4)} \sum_{n=\tilde{T}_f(\nu^2/4)}^{N} f(X_n) \pm \epsilon_N \right]$$

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

**Proof.** For notational economy, let $\tau_f(\epsilon) = T_f(\nu^2/4)$ and $\tilde{\tau}_f(\epsilon) = \tilde{T}_f(\nu^2/4)$. Theorem 1 then implies

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

Setting $\epsilon = \epsilon_N$ yields

$$\mathbb{P} \left[ \frac{1}{N - \tilde{T}_f(\nu^2/4)} \sum_{n=\tilde{T}_f(\nu^2/4)}^{N} f(X_n) \geq \mu + \epsilon_N \right] \leq \frac{\alpha}{2}.$$

The corresponding lower bound leads to an analogous bound on the lower tail. \(\square\)
As we did with Corollary 4, 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):

$$\tilde{r}_N(\epsilon) \geq \tilde{r}_{N,\eta}(\epsilon) := \epsilon^2 \left[ \frac{N}{T_f(\frac{\nu \eta^2}{4})} - 1 \right], \epsilon \geq \eta.$$

Since this is a lower bound, we see that whenever $\epsilon_N \geq \eta$ and $\tilde{r}_{N,\eta}(\epsilon_N) \geq 4 \log \frac{2}{\alpha}$, $\epsilon_N$ is a valid half-width for a $(1 - \alpha)$-confidence interval for the stationary mean centered at the empirical mean. More formally, we have the following:

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

$$\epsilon_N = \tilde{r}_{N,\eta}^{-1}(4 \log \frac{2}{\alpha}) = 2 \sqrt{\frac{T_f(\frac{\nu \eta^2}{4}) \cdot \log \frac{2}{\alpha}}{N - T_f(\frac{\nu \eta^2}{4})}}.$$

If $N \geq \tilde{T}_f(\frac{\nu \eta^2}{4})$, then $I_N$ is a $1 - \alpha$ confidence interval for $\mu = \mathbb{E}_\pi [f]$.

**Proof.** By assumption, we have

$$\eta \leq \epsilon_N(\eta) = 2 \sqrt{\frac{T_f(\frac{\nu \eta^2}{4}) \cdot \log \frac{2}{\alpha}}{N - T_f(\frac{\nu \eta^2}{4})}}.$$

This implies $\tilde{T}_f(\frac{\nu \eta^2}{4}) \geq \tilde{T}_f(\frac{\nu \eta^2}{4})$, which yields

$$\tilde{r}_N(\epsilon_N) = \epsilon_N^2 \left[ \frac{N}{T_f(\frac{\nu \eta^2}{4})} - 1 \right] \geq \epsilon_N^2 \left[ \frac{N}{T_f(\frac{\nu \eta^2}{4})} - 1 \right] = 4 \log \frac{2}{\alpha}.$$

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

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

5.6.1. **Proof of Theorem 4 for $A_{\text{fixed}}.$** 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
\[ \mathbb{P}(A_{\text{fixed}}(X_{1:N}) = H_0) \]. In order to do so, note that Theorem 1 implies that

\[
\mathbb{P}(A_{\text{fixed}}(X_{1:N}) = H_0) = \mathbb{P} \left( \frac{1}{N} \sum_{n=1}^{N} f(X_n) \geq r + \delta \right)
\leq \mathbb{P} \left( \frac{1}{N} \sum_{n=1}^{N} f(X_n) \geq \mu + 2\delta \right)
\leq \exp \left( - \frac{\delta^2 N}{T_f(\nu \delta^2)} \right)
= \exp \left( - \frac{\delta^2 N}{\tau_f} \right).
\]

Finally, setting \( N = \frac{\tau_f \log \left( \frac{1}{\alpha} \right)}{\delta^2} \) yields the bound \( \mathbb{P}(A_{\text{fixed}}(X_{1:N}) = H_0) \leq \alpha \), as claimed.

5.6.2. Proof of Theorem 4 for \( A_{\text{seq}} \). The proof is nearly identical to that given by Gyori and Paulin (2015), with \( \tau_f(\delta/2) \) 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

\[
\text{err}(A_{\text{seq}}, f) = \mathbb{P}(\exists k: A_{\text{seq}}(X_{1:N_k}) = H_0)
= \mathbb{P}(\exists k: \frac{1}{N_k} \sum_{n=1}^{N_k} f(X_n) \geq r + \frac{M}{N_k})
\leq \sum_{k=1}^{\infty} \mathbb{P}(\frac{1}{N_k} \sum_{n=1}^{N_k} f(X_n) \geq r + \frac{M}{N_k}).
\]

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

\[
\mathbb{P} \left( \frac{1}{N_k} \sum_{n=1}^{N_k} f(X_n) \geq r + \frac{M}{N_k} \right) \leq \mathbb{P} \left( \frac{1}{N_k} \sum_{n=1}^{N_k} f(X_n) \geq \mu + \delta + \frac{M}{N_k} \right)
\leq \exp \left( - \frac{\epsilon_k^2 N_k}{4T_f(\nu \epsilon_k^2)} \right)
\leq \exp \left( - \frac{\epsilon_k^2 N_k}{4T_f(\nu \delta^2)} \right)
= \exp \left( - \frac{\epsilon_k^2 N_k}{2\tau_f(\frac{\delta}{2})} \right).
\]
In order to simplify notation, for the remainder of the proof, we define $\tau := \tau_f(\delta/2)$, $\beta := \sqrt{\alpha \xi}$, 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

$$
\exp\left(-\frac{\epsilon^2 N_k}{2\tau}\right) = \exp\left(-\frac{1}{2\tau} \cdot (\delta^2 N_k + 2\delta M + \frac{M^2}{N_k})\right)
$$

$$
= \exp\left(-\left[\frac{\delta^2 N_k}{2\tau} + \log(1/\beta) + \frac{2\tau \log^2(1/\beta)}{\delta^2 N_k}\right]\right)
$$

$$
= \exp\left(-\log(1/\beta) [1 + \zeta_k + \zeta_k^{-1}]\right)
$$

$$
= \beta \cdot \exp\left(-\log(1/\beta) [\zeta_k + \zeta_k^{-1}]\right).
$$

It follows that the error probability is at most

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

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

**Lemma 4.** In the above notation, we have

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

**Lemma 5.** For any integer $c \geq 0$, we have

$$
(1 + \xi)^\ell + (1 + \xi)^{-\ell} \geq 2(c + 1) \quad \text{for all } \ell \in \left[\frac{9c}{5\xi}, \frac{9(c+1)}{5\xi}\right).
$$

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

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

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

$$
\frac{36\beta}{5\xi} \sum_{c=0}^{\infty} \beta^{2(c+1)} \leq \frac{36\beta^3}{5\xi (1 - \beta^2)} \leq \frac{36\beta^2}{25\xi (1 - \beta^2)} \leq \frac{3\beta^2}{2\xi} = \frac{3\alpha}{4} < \alpha.
$$
5.6.3. Proof of Theorem 4 for $A_{\text{hard}}$. 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
\[
\text{err}(A_{\text{hard}}, f) \leq \sum_{k=k_0}^{\infty} \mathbb{P}(\hat{\mu}_{N_k} \geq r + \epsilon_k) \leq \sum_{k=k_0}^{\infty} \mathbb{P}(\hat{\mu}_{N_k} \geq \mu + 2\epsilon_k).
\]
By Theorem 1, and the definition of $\epsilon_k$, we thus have
\[
\text{err}(A_{\text{hard}}, f) \leq \sum_{k=k_0}^{\infty} \exp\left(-\frac{N_k\epsilon_k^2}{4T\left(\frac{\nu_k^2}{4}\right)}\right) \leq \frac{\alpha}{2} \sum_{k=k_0}^{\infty} \frac{1}{k^2}
\]
\[
= \frac{\pi^2}{12} \alpha < \alpha,
\]
as claimed.

5.6.4. Proof of Theorem 5 for $A_{\text{seq}}$. We may assume $H_1$ holds; the other case is analogous. Note that
\[
\mathbb{E}[N] \leq N_1 + \sum_{k=1}^{\infty} (N_{k+1} - N_k) \mathbb{P}(N > N_k)
\]
\[
\leq N_1 + \sum_{k=1}^{\infty} (N_{k+1} - N_k) \mathbb{P}\left(\frac{1}{N_k} \sum_{n=1}^{N_k} f(X_n) \in (r - M, r + M)\right)
\]
\[
\leq N_1 + \sum_{k=1}^{\infty} (N_{k+1} - N_k) \mathbb{P}\left(\frac{1}{N_k} \sum_{n=1}^{N_k} f(X_n) > r - \frac{M}{N_k}\right)
\]
\[
= N_1 + \sum_{k=1}^{\infty} (N_{k+1} - N_k) \mathbb{P}\left(\frac{1}{N_k} \sum_{n=1}^{N_k} f(X_n) > \mu + \Delta - \frac{M}{N_k}\right)
\]
\[
\leq N_1 + \sum_{k=1}^{\infty} (N_{k+1} - N_k) \exp\left\{-\frac{(\Delta N_k - M)^2}{2\tau_f(\delta/2)N_k}\right\}.
\]
Our proof depends on the following simple technical lemma, whose proof we defer to Appendix D.3.

**Lemma 6.** Under the conditions of Theorem 5, we have
\[
\sum_{k=1}^{\infty} (N_{k+1} - N_k) \exp\left\{-\frac{(\Delta N_k - M)^2}{2\tau_f(\delta/2)N_k}\right\} \leq (1 + \xi) \left[1 + \int_{N_1}^{\infty} h(s)ds\right],
\]
where $h(s) := \exp\left\{-\frac{(\Delta s - M)^2}{2\tau_f(\delta/2)s}\right\}$. 
Given this lemma, we then follow the argument of Gyori and Paulin (2015) in order to bound the integral: more precisely, we have

$$\int_{N_1}^{\infty} h(s) \, ds \leq 2\sqrt{\frac{2\tau_f(\delta/2)M}{\Delta}} + 8\tau_f(\delta/2).$$

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

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

Combining the bounds yields the desired result.

5.6.5. Proof of Theorem 5 for $A_{\text{hard}}$. For concreteness, we may assume $H_1$ holds, as the $H_0$ case is symmetric. We now have that

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

For convenience, let us introduce the shorthand

$$T_{f,k}^+ := \begin{cases} T_f\left(\frac{\mu^2}{4}\right) & \text{if } \epsilon_k \leq \Delta, \\ 1 & \text{otherwise}. \end{cases}$$

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

$$\mathbb{P}[N \geq N_k] \leq \exp\left\{-\frac{N_k}{4T_{f,k}^+} \cdot (\Delta - \epsilon_k)^2\right\}.$$ 

Observe further that

$$\mathbb{E}[N] = N_1 + \sum_{k=1}^{\infty} (N_{k+1} - N_k) \mathbb{P}(N > N_k)$$

$$\leq N_{k_0^*+1} + \sum_{k=k_0^*+1}^{\infty} (N_{k+1} - N_k) \mathbb{P}(N > N_k)$$

$$\leq (1 + \xi)(N_0^* + 1) + \sum_{k=k_0^*+1}^{\infty} (N_{k+1} - N_k) \mathbb{P}(N > N_k).$$
Combining the pieces yields

\[ E[N] \leq (1 + \xi) (N_0^* + 1) + \sum_{k=k_0^*+1}^{\infty} (N_{k+1} - N_k) \exp \left( -\frac{N_k}{4T_{f,k}} \cdot (\Delta - \epsilon_k)^2 \right). \]

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

**Lemma 7.** The infinite sum (46) is upper bounded by

\[ \sum_{k=k_0^*+1}^{\infty} (N_{k+1} - N_k) \exp \left( -\frac{N_k}{4T_{f,k}} \cdot (\Delta - \epsilon_k)^2 \right) \leq \alpha \cdot \sum_{m=1}^{\infty} \exp \left( -m \cdot \frac{\Delta^2}{16T_f \left( \frac{\nu \Delta^2}{16} \right)} \right). \]

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

Lemma 7 then implies that

\[ \sum_{k=k_0^*+1}^{\infty} (N_{k+1} - N_k) \exp \left( -\frac{N_k}{T_f \left( \frac{\nu \Delta^2}{16} \right)} \cdot \frac{\Delta^2}{16} \right) \leq \alpha \cdot \sum_{m=1}^{\infty} \exp \left( -m \cdot \frac{\Delta^2}{16T_f \left( \frac{\nu \Delta^2}{16} \right)} \right) \]

\[ = \frac{\alpha \exp \left( -\frac{\Delta^2}{16T_f \left( \frac{\nu \Delta^2}{16} \right)} \right)}{1 - \exp \left( \frac{\Delta^2}{16T_f \left( \frac{\nu \Delta^2}{16} \right)} \right)} \]

\[ \leq \frac{16\alpha T_f \left( \frac{\nu \Delta^2}{16} \right)}{\Delta^2}. \]

The claim now follows from equation (46).

**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. 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.

Acknowledgements. 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.


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

\[ d_f(p, q) = \sup_{f : [d] \to [0, 1]} |E_p[f(X)] - E_q[f(Y)]|. \]
It follows from equation (12) that
\[
\TV(\pi_n, \pi) = \sup_{f : [d] \to [0, 1]} d_f(\pi_n, \pi)
\]
\[
\leq \sup_{f : [d] \to [0, 1]} \left[ \frac{\lambda_n^f}{\sqrt{\pi_{\min}}} \cdot d_f(\pi_0, \pi) \right]
\]
\[
= \frac{1}{\sqrt{\pi_{\min}}} \cdot \lambda_n^f \cdot \TV(\pi_0, \pi),
\]
as claimed.

A.2. Proof of equation (12). Let \( D = \text{diag}(\sqrt{\pi}) \). 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
\[
P = \pi^T + \sum_{j=2}^d \lambda_j h_j q_j^T,
\]
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 \( \pi_0 \) be an arbitrary distribution over \([d]\), we have
\[
d_f(\pi_n, \pi) = |\pi_0^T P^n f - \pi^T P^n f| \leq |(\pi_0 - \pi)^T P^n f|.
\]
Defining \( P_f := \pi^T + \sum_{j \in J_f} \lambda_j h_j q_j^T \), 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 \( (\pi_0 - \pi)^T \tilde{P}_f = (\pi_0 - \pi)^T P_f \). Consequently, by the definition of the operator norm and sub-multiplicativity, we have
\[
d_f(\pi_n, \pi) \leq \left| (\pi_0 - \pi)^T \tilde{P}_f^n f \right|
\]
\[
\leq \| \tilde{A}_f \|_{op} \| Df \|_2 \| D^{-1} (\pi_0 - \pi) \|_2
\]
\[
= \sqrt{\E \| f^2 \|_2 \sum_{i \in [d]} \frac{(\pi_{0,i} - \pi_i)^2}{\pi_i} \cdot \lambda_i^f d_f(\pi_0, \pi)}.
\]

In order to complete the proof, let \( Z \in \{0, 1\}^d \) denote the indicator vector \( Z_j = 1(X_0 = j) \). Observe that the function
\[
r(z) := \sum_{i \in [d]} \frac{(z_i - \pi_i)^2}{\pi_i}
\]
is convex in terms of $z$. Thus, Jensen’s inequality implies that

$$E_{\pi_0}[r(Z)] \geq r(E_{\pi_0}[Z]) = r(\pi_0) = \sum_{i \in [d]} \left( \frac{\pi_{0,i} - \pi_i}{\pi_i} \right)^2.$$ 

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

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

We deduce that $d_f(\pi_n, \pi) \leq \sqrt{\frac{E_{\pi_0}[f^2]}{\pi_{\min}} \cdot \lambda_{f}^n \cdot d_f(\pi_0, \pi)}$, as claimed.

A.3. Proof of Lemma 1. We observe that

$$\left| (\pi_0 - \pi)^T h_J(n) \right| \leq \|\pi_0 - \pi\|_1 \cdot \|h_J(n)\|_{\infty}$$

$$= 2d_{TV}(\pi_0, \pi) \cdot \|h_J(n)\|_{\infty}$$

$$\leq 2d_{TV}(\pi_0, \pi) \left\{ \sum_{j \in J} |\lambda_j|^n \cdot |q_j^T f| \cdot \|h_j\|_{\infty} \right\}$$

$$\leq 2d_{TV}(\pi_0, \pi) \left\{ 2|J| \cdot \max_{j \in J} |q_j^T f| \cdot \max_{j \in J} \|h_j\|_{\infty} \right\},$$

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.

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

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

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

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

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

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

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

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 $f: [d] \rightarrow [0, 1]$ 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

$$T_f(\delta) \leq c_0 d \log d \log \frac{128d}{\delta}$$

for all $0 < \delta \leq \delta^*$.

**Proof.** We proceed by defining a “good event” $E_\delta$, and then showing that the stated bound on $T_f(\delta)$ holds conditioned on this event. The final step is to show that $P[E_\delta]$ is suitably close to one, as claimed.

The event $E_\delta$ 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

$$J_\delta := \{0 < j < 2d: j \leq 4\delta \sqrt{\frac{d}{\log d}} \text{ or } j \geq 2d - 4\delta \sqrt{\frac{d}{\log d}}\}.$$

Then a function is “good” if its inner product with all the corresponding eigenfunctions is small:

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

Note that these events form a decreasing sequence because the $J_\delta$ is increasing in $\delta$, in that whenever $\delta \leq \delta^*$, $J_\delta \subset J_{\delta^*}$. To obtain the desired result, it thus suffices to prove that the bound on $T_f(\delta)$ holds on $E_\delta$ for arbitrary $0 < \delta < 1$, then prove that $\delta > 0$, $P[E_\delta]$ is large enough.

We now exploit the spectral properties of the transition matrix to show that $E_\delta$ is in fact a good enough event. For this, recall (Levin, Peres and Wilmer, 2008) that the lazy random walk on $C_{2d}$ has eigenvalues $\lambda_j = \frac{1}{2}(1 + \cos(\frac{\pi j}{d}))$ for $j \in [d]$, with corresponding unit eigenvectors

$$v_j^T = \frac{1}{\sqrt{2d}} \left(1 \ \omega_j \cdots \omega_j^{2d-1}\right), \ \omega_j := e^{\frac{\pi i j}{d}}.$$

We note that this diagonalization allows us to write $P = \pi^T + \sum_{j=1}^{2d-1} \lambda_j h_j q_j^T$, where $h_j = \sqrt{2d} \cdot v_j$ and $q_j = \frac{v_j}{\sqrt{2d}}$, where we have used the fact that $\text{diag}(\sqrt{\pi}) = \frac{1}{\sqrt{2d}} \cdot I$. Note that $\|h_j\|_\infty = 1.$
Applying Lemma 1 and using the facts that $\lambda_{J_\delta} \leq 1$, that $\|h_j\|_\infty \leq 1$, and that $|J_\delta| \leq 8\delta \sqrt{\frac{d}{\log d}}$, we find

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

Therefore, when the event $E_\delta$ holds, we have

$$d_f(\pi_n, \pi) \leq 32\sqrt{10} \cdot \delta + \sqrt{d} \cdot \lambda_{-J_\delta}^n. \quad (47)$$

To conclude, we use the fact that

$$\lambda_{-J_\delta} = 1 + \max_{j \in J \setminus J_\delta} \cos \left( \frac{\pi j}{d} \right) \leq \frac{1 + \cos \left( \frac{\pi j_0}{d} \right)}{2},$$

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

$$\cos \left( \pi x \right) \leq 1 - \frac{\pi^2 x^2}{2} + \frac{\pi^4 x^4}{24} \leq 1 - \frac{\pi^2 x^2}{12}, \quad |x| \leq 1,$

which implies that

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

Together with equation (47), this bound implies that for $n \geq \frac{3d \log d \log d}{2\pi^2 \delta^2}$, we have $\sqrt{d} \lambda_{-J_\delta}^n \leq \delta$, whence

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

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

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

we have $d_f(\pi_n, \pi) \leq \delta$ with probability at least $\mathbb{P}(E_{\delta/128})$. It now suffices to prove $\mathbb{P}(E_{\delta}) \geq 1 - \frac{\delta}{\sqrt{d \log d}}$, since this will imply that $\mathbb{P}(E_{\delta/128}) \geq 1 - \frac{\delta}{128\sqrt{d \log d}}$, as required.

For this, notice that because the vectors $\{q_j\}_{j=1}^d$ are rescaled versions of an orthonormal collection of eigenvectors, we have

$$\mathbb{E}[q_j^T f] = \mathbb{E}_\nu [\mu] \cdot q_j^T 1 = 0$$
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

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

showing that it is the sum of sub-Gaussian random variables with parameters $\sigma^2 = \cos^2 \left( \frac{\pi j \ell}{d} \right) \leq 1$. 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 $\sigma_j^2$.

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

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

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

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

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

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

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

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

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

$$T_f \left( \frac{\epsilon^2}{16} \right) \leq c_0 d \log d \left[ \log d + \log \left( \frac{1}{\epsilon} \right) \right],$$
and

\[
\mathbb{h}\mathbb{P}\left[ \frac{1}{N} \sum_{n=T_f(\epsilon^2/16)}^{N+T_f(\epsilon^2/16)} f(X_n) \geq \mu + \epsilon \right] \leq \exp\left( - \frac{c_1 \delta N}{d \log d \left( \log \left( \frac{32}{\epsilon^2} \right) + \log 2d \right)} \right).
\]

with probability at least \(1 - \frac{c_2 \epsilon^2}{\sqrt{d \log d}} - 2e^{-\frac{d^2}{2}}\) over the randomness of \(f\) provided \(\epsilon \geq c_3 (\log d)^{1/4}\), where \(c_0, c_1, c_2, c_3 > 0\) are universal constants.

**Proof.** We first note that by the proof of Proposition 4, for \(0 < \delta < 1\), we have \(1 - \lambda_{-J} \delta \geq \frac{c_4 \delta^2}{d \log d}\). The proof of the previous proposition guarantees that \(\Delta^* J \leq 32 \sqrt{10} \delta\), so setting \(\delta = \frac{\epsilon^2}{1024 \sqrt{10}}\) yields

\[
\frac{\epsilon^2}{32} = 32 \sqrt{10} \delta \geq \Delta^* J, \quad \text{and} \quad 1 - \lambda_{-J} \delta \geq \frac{c_4' \epsilon^4}{d \log d}.
\]

Now, on the other hand, by the usual Hoeffding inequality for i.i.d. variates applied to \(\frac{1}{2d} \sum_{u=0}^{2d-1} f(u)\), we have that the probability that \(\nu < 1/4\) is bounded by \(2e^{-\frac{d^2}{2}}\). Moreover, by Proposition 4, we have

\[
T_f(\delta') \leq \frac{c_5 d \log d \log d / \delta'}{(\delta')^2}
\]

for a universal constant \(c_5 > 0\) and all \(\delta' \geq \delta\) with probability at least \(1 - \frac{\delta}{128 \sqrt{d \log d}}\). In particular,

\[
T_f\left( \frac{\epsilon^2}{16} \right) \leq \frac{c_5' d \log d \log d / \epsilon^2}{\epsilon^4}
\]

with this probability. Thus we have this bound on \(T_f\) and \(\nu \geq 1/4\) simultaneously with the high probability claimed in the statement of the proposition.

We now finish by taking \(\Delta = \frac{\epsilon^2}{16}\) in Corollary 2. Noting that \(\Delta J + \Delta = \frac{\epsilon^2}{16}\) and \(1 - \lambda_{-J} \geq \frac{c_4' \epsilon^4}{d \log d}\), we obtain the claimed tail bound.

**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 (26). This claim follows directly from a modified uniform Hoeffding bound, due to Paulin (2012). In particular, for any integer $T_0 \geq 0$, let $d_{\text{TV}}(T_0) = \sup_{\pi_0} d_{\text{TV}}(\pi_0 P_{T_0}, \pi)$ be the worst-case total variation distance from stationarity after $T_0$ steps. Using this notation, Paulin (2012) shows that for any starting distribution $\pi_0$ and any bounded function $f : [d] \to [0, 1]$, we have

\begin{equation}
\Pr(\left| \frac{1}{N - T_0} \sum_{n=T_0+1}^N f(X_n) - \mu \right| \geq \epsilon) \leq 2 \exp \left( -\frac{\gamma_0}{2(2 - \gamma_0)} \cdot \epsilon^2 N \right) + 2d_{\text{TV}}(T_0). \tag{48}
\end{equation}

We now use the bound (48) to prove our claim (26). Recall that we have chosen $T_0$ so that $d_{\text{TV}}(T_0) \leq \alpha_0 / 2$. Therefore, the bound (48) implies that

\begin{equation}
\Pr \left[ \left| \frac{1}{N - T_0} \sum_{n=T_0+1}^N f(X_n) - \mu \right| \geq \epsilon_N \right] \leq 2 \exp \left\{ -\frac{\gamma_0}{2(2 - \gamma_0)} \cdot \epsilon_N^2 N \right\} + \alpha_0
\leq 2 \cdot \frac{\alpha - \alpha_0}{2} + \alpha_0 = \alpha,
\end{equation}

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

\begin{equation}
\frac{e^{-\gamma_0 N}}{3\sqrt{\pi_{\text{min}}}} \leq \frac{\alpha}{6} \quad \text{and} \quad \frac{13}{\sigma_{f,\text{asym}} \sqrt{\pi_{\text{min}}}} \cdot \frac{1}{\gamma_0 \sqrt{N}} \leq \frac{\alpha}{6}.
\end{equation}

It follows from equation (29) that

\begin{equation}
\Pr \left[ \frac{1}{\sigma_{f,\text{asym}} N} \sum_{n=1}^N f(X_n) \geq \mu + \epsilon_N \right] \leq \Phi(\epsilon_N \sqrt{N}) + \frac{\alpha}{3}
\leq \exp \left( -\frac{N}{2} \cdot \epsilon_N^2 \right) + \frac{\alpha}{3}
= \frac{\alpha}{2},
\end{equation}

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

D. Proofs for Section 5.6. In this section, we gather the proofs of Lemmas 4–7.
D.1. *Proof of Lemma 4.* Observe that the function
\[ g(\zeta) := \exp \left\{ -\log(1/\beta) \left( \zeta + \zeta^{-1} \right) \right\} \]
is increasing on $(0, 1]$ and decreasing on $[1, \infty)$. Therefore, bringing \( \zeta \) closer to 1 can only increase the value of the function.

Now, for fixed \( k \geq 1 \), define
\[
\ell_k := \begin{cases} 
\min \{ \ell: (1 + \xi)\ell \geq \zeta_k \} & \text{if } \zeta_k \leq 1, \\
\max \{ \ell: (1 + \xi)\ell \leq \zeta_k \} & \text{otherwise.}
\end{cases}
\]

In words, the quantity \( \ell_k \) is either the smallest integer such that \( (1 + \xi)\ell \) is bigger than \( \zeta_k \) (if \( \zeta_k \leq 1 \)) or the largest integer such that \( (1 + \xi)\ell \) is smaller than \( \zeta_k \) (if \( \zeta_k \geq 1 \)).

With this definition, we see that \( (1 + \xi)\ell_k \) always lies between \( \zeta_k \) and 1, so that we are guaranteed that \( g((1 + \xi)\ell_k) \geq g(\zeta_k) \), and hence
\[
\sum_{k=1}^{\infty} g(\zeta_k) \leq \sum_{k=1}^{\infty} g((1 + \xi)\ell_k).
\]
The proof will thus be completed if we can show that at most two distinct values of \( k \) map to a single \( \ell_k \). Indeed, if we prove this, we will have
\[
\sum_{k=1}^{\infty} g(\zeta_k) \leq 2 \sum_{\ell=-\infty}^{\infty} g((1 + \xi)\ell) \leq 4 \sum_{\ell=0}^{\infty} g((1 + \xi)\ell).
\]

In order to prove the required claim, note first that \( \ell_k \) is clearly nondecreasing in \( k \), so that we need to prove that \( \ell_{k+2} > \ell_k \) for all \( k \geq 1 \). It is sufficient 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 \( \zeta_k = an_k \) 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
\[
\frac{n_{k+2}}{n_k} = \frac{(1 + \xi)^2 \lfloor b \rfloor}{\lfloor b \rfloor} \geq \frac{(1 + \xi)^2 b - 1}{\lfloor b \rfloor} \geq \frac{(1 + \xi)(\lfloor b \rfloor + 1) - 1}{\lfloor b \rfloor} \geq 1 + \xi,
\]
D.2. Proof of Lemma 5. 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 \( \ell \), so that the \( c = 0 \) case follows immediately.

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

\[
(1 + \xi)^{2\ell} - 2(c + 1)(1 + \xi)^\ell + 1 \geq 0 \quad \text{for all } \ell \in \left( \frac{9c}{8\xi}, \frac{9(c+1)}{5\xi} \right).
\]

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 \( \frac{9c}{8\xi} \) is a lower bound on \( \ell \), we need to verify that

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

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 6. Our strategy will be 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 \( k \geq k_0 \).

\(^2\)We thank Daniel Paulin for suggesting this argument as an elaboration on the shorter proof in Gyori and Paulin (2015).
For convenience in what follows, let us introduce the convenient shorthand

\[ T_k := \exp \left( -\frac{(\Delta N_k - M)^2}{2\tau_f (\delta/2) N_k} \right) \]

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

\[ \sum_{k=1}^{\infty} (N_{k+1} - N_k) T_k \leq \int_{N_1}^{\infty} h(s) \, ds. \]

Otherwise, if \( k_0 > 1 \), we have

\[ \sum_{k=k_0}^{\infty} (N_{k+1} - N_k) T_k \leq \int_{N_{k_0}}^{\infty} h(s) \, ds. \]

For \( k < k_0 \), we have \( T_k = 1 \), so that when \( k < k_0 - 1 \), we have

\[ (N_{k+1} - N_k) \exp \left( -\frac{(\Delta N_k - M)^2}{2\tau_f (\delta/2) N_k} \right) = \int_{N_{k_0} - 1}^{N_k} h(s) \, ds. \]

Thus

\[ \sum_{k=1}^{k_0-1} (N_{k+1} - N_k) \exp \left( -\frac{(\Delta N_k - M)^2}{2\tau_f (\delta/2) N_k} \right) = \int_{N_1}^{N_{k_0} - 1} h(s) \, ds. \]

Note that this implies

\[ \int_{N_1}^{\infty} \exp \left( -\frac{(\Delta s - M)^2}{2\tau_f (\delta/2) s} \right) \, ds \geq N_{k_0 - 1}. \]

Finally, we observe that \( N_{k+1} \leq (1+\xi) N_k + 1 + \xi \), so that \( N_{k_0} - N_{k_0 - 1} \leq \xi N_{k_0 - 1} + 1 + \xi \).

Putting together the pieces, we have

\[ (N_{k_0} - N_{k_0 - 1}) \exp \left( -\frac{(\Delta N_{k_0 - 1} - M)^2}{2\tau_f (\delta/2) N_{k_0 - 1}} \right) \leq 1 + \xi + \xi \int_{N_1}^{\infty} \exp \left( -\frac{(\Delta s - M)^2}{2\tau_f (\delta/2) s} \right) \, ds, \]

and hence

\[ \sum_{k=1}^{\infty} (N_{k+1} - N_k) \exp \left( -\frac{(\Delta N_k - M)^2}{2\tau_f (\delta/2) N_k} \right) \leq 1 + \xi + (1 + \xi) \int_{N_1}^{\infty} h(s) \, ds. \]
D.4. Proof of Lemma 7. Observe that for $k > k_0^\ast$, we have $\Delta - \epsilon_k \geq \frac{\Delta}{2}$. It follows that for $k > k_0^\ast$, we have $T_{f,k}^+ \leq T_f(\frac{\nu \Delta^2}{16})$. Thus, we can bound each term in the sum by

$$(N_{k+1} - N_k) \exp\left(-\frac{N_k}{4T_{f,k}^+} \cdot (\Delta - \epsilon_k)^2\right) \leq (N_{k+1} - N_k) \exp\left(-\frac{N_k}{T_f(\frac{\nu \Delta^2}{16})} \cdot \frac{\Delta^2}{16}\right)$$

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

$$\sum_{k=k_0^\ast + 1}^{\infty} F_k \leq \sum_{n=N_0^\ast + 1}^{\infty} \exp\left(-n \cdot \frac{\Delta^2}{16T_f(\frac{\nu \Delta^2}{16})}\right)$$

$$= \exp\left(-N_0^\ast \cdot \frac{\Delta^2}{16T_f(\frac{\nu \Delta^2}{16})}\right) \times \sum_{m=1}^{\infty} \exp\left(-m \cdot \frac{\Delta^2}{16T_f(\frac{\nu \Delta^2}{16})}\right)$$

$$= \exp\left(-\frac{N_0^\ast}{4T_f(\frac{\nu (\Delta/2)^2}{4})} \cdot (\frac{\Delta}{2})^2\right) \times \sum_{m=1}^{\infty} \exp\left(-m \cdot \frac{\Delta^2}{16T_f(\frac{\nu \Delta^2}{16})}\right).$$

On the other hand, by the definition of $N_0^\ast$, $\epsilon_k \leq \frac{\Delta}{2}$, so

$$T_f(\frac{\nu (\Delta/2)^2}{4}) \leq T_f(\frac{\nu \epsilon_k^2}{4}).$$

By the definition of $\epsilon_k$, however, we know that

$$\frac{\epsilon_k^2}{4T_f(\frac{\nu \epsilon_k^2}{4})} \geq \frac{\log (1/\alpha) + 1 + 2 \log k}{N_k} \geq \frac{\log (1/\alpha)}{N_k},$$

which implies that $(\Delta/2)^2 N_0^\ast \geq \log (1/\alpha) 4T_f(\frac{\nu (\Delta/2)^2}{4})$. Re-arranging yields the claim.