Making mean-estimation more efficient using an MCMC trace variance approach: DYNAMITE

Cyrus Cousins\textsuperscript{1}, Shahrzad Haddadan\textsuperscript{2}, and Eli Upfal\textsuperscript{3}

\begin{itemize}
\item \textsuperscript{1,2,3} Department of Computer Science, Brown University
\item \textsuperscript{2} Data Science Initiative, Brown University
\end{itemize}

\textsuperscript{1} cyrus\_cousins@brown.edu
\textsuperscript{2} shahrzad\_haddadan@brown.edu
\textsuperscript{3} eli\_upfal@brown.edu

Abstract

We introduce a novel statistical measure for MCMC-mean estimation, the \textit{inter-trace variance} $\text{trv}^{(\tau_{\text{rel}})}(M, f)$, which depends on a Markov chain $M$ and a function $f : S \rightarrow [a, b]$. We show that the inter-trace variance can be efficiently estimated from observed data, and that it leads to a more efficient MCMC-mean estimator, with complexity competitive with a lower-bound obtained from the central limit theorem of Markov chains. Most efficient MCMC mean-estimators receive, as input, upper-bounds on chain-dependent terms like \textit{mixing time} $\tau_{\text{mix}}$ or relaxation time $\tau_{\text{rel}}$, and often also function-dependent terms such as the \textit{stationary variance} $v_\pi$, and their performance is highly dependent to the sharpness of these bounds. In contrast, we introduce DYNAMITE, which dynamically adjusts the sample size using the observed data, and therefore it is less sensitive to the looseness of input upper-bounds on $\tau_{\text{rel}}$, and requires no bound on $v_\pi$.

Receiving only an upper-bound $T_{\text{rel}}$ on $\tau_{\text{rel}}$, DYNAMITE estimates w.h.p. the mean of $f$ to within $\varepsilon$ additive error in $\tilde{O}(\frac{T_{\text{rel}}}{\varepsilon^2} + \frac{T_{\text{rel}} \text{trv}^{(\tau_{\text{rel}})}}{\varepsilon^2})$ steps, without \textit{a priori} bounds on the stationary variance $v_\pi$ or the inter-trace variance $\text{trv}^{(\tau_{\text{rel}})}$. Thus we obtain minimal dependency on the tightness of $T_{\text{mix}}$, since the complexity is dominated by $\tau_{\text{rel}}\text{trv}^{(\tau_{\text{rel}})}$ as $\varepsilon \rightarrow 0$ (the high precision regime), even though the values of $\tau_{\text{rel}}$ and $\text{trv}^{(\tau_{\text{rel}})}$ are not known to the algorithm. Note that bounding $\tau_{\text{rel}}$ is known to be prohibitively difficult, however, DYNAMITE is able to reduce its principal dependence on $T_{\text{rel}}$ to $\tau_{\text{rel}}$, simply by exploiting properties of the inter-trace variance. To compare our method to known variance-aware bounds, we show $\text{trv}^{(\tau_{\text{rel}})}(M, f) \leq v_\pi$. Furthermore, we show when $f$’s image is distributed symmetrically on $M$’s traces, we have $\text{trv}^{(\tau_{\text{rel}})}(M, f) = o(v_\pi(f))$, thus DYNAMITE outperforms prior methods in these cases, even when tight bounds on variance and mixing or relaxation times are known. We demonstrate the advantage of our estimator through an interesting application, counting $k$-colorings of a class of graphs, e.g., graphs drawn from the planted partition model, wherein using DYNAMITE leads to significant improvement.

1 Introduction

Given a bounded, real-valued function $f : S \rightarrow [a, b]$, and a distribution $\pi$ defined on the domain $S$, our goal is to estimate the average of this function, $\mu = \mathbb{E}_\pi[f(x)]$. Having enough independent samples from $\pi$, one can estimate $\mu$ by the \textit{empirical mean} $\hat{\mu}$, the average value of $f(x)$ over the samples. In many important applications it is hard (or even impossible) to efficiently generate independent samples from $\pi$. Instead, one can often generate a sequence of (nonindependent)
samples by traversing a Markov chain with state space $S$ and stationary distribution $\pi$. MCMC-mean estimation is the process of generating an estimate of $\mu = \mathbb{E}_\pi[f(x)]$ from the Markov chain samples. The efficiency of an MCMC-mean estimator is measured by the accuracy of the estimate it generates, and its (sample) complexity - the (expected) number of Markov chain transitions it uses for generating the estimate. The number of Markov chain transitions clearly dominates the computational complexity of the estimator.

MCMC-mean estimation is a well-studied problem in probability theory, statistics, and computer science, with numerous applications in a variety of fields such as statistical physics [DMM07, MR01, Dia09, CA16], chemistry [Gil07], computational biology [VRU16, VCUR12, VUR11, EDO02, APO+18], statistical machine learning, image processing, etc [PLMT13, SAB+19, YK06, TZ02]. MCMC-Mean estimation is also an important component in solutions of related hard computational problems. For example, Jerrum, Valiant, and Vazirani [JVV86], and many others [SVV07, Hub15, Kol18] reduced estimating the size of a self-reducible set (often a \#hard problem) to solving a series of mean-estimation tasks.

A key concept in analyzing MCMC-estimates is the Markov chains’ mixing time (see section 1.1 for definition). Rigorously bounding mixing times needs potent analytical techniques and sophisticated problem-specific analysis [Gur00, ALG19, BD97, Vaz91] and it is an active area of research [Var18, HW17, BCC+21, CLV21, ABH19]. Unfortunately these bounds are often loose which hampers their applicability, thus, practitioners often ignore mixing time bounds, and instead run the Markov chain until a heuristic termination condition is observed [CC96, BC98, BMK14, DR17, BR98, R+02]. In particular, one popular method is to estimate the autocovariance between adjacent samples [CC96], and terminate when it is negligible. Unfortunately, correctness of such approaches is not supported by mathematical guarantees and in practice their applications are generally quite error-prone. Not surprisingly, there has been significant interest in providing mathematical tools to analyze the computation complexity of estimating mixing time from observations with no prior knowledge of it [BBM11, BHOP18, HKL+19, WK19, Wgd20]. These results are often negative, as they provide large lower bounds for this problem. Thus, some, even loose, upper bound on mixing time seems to be necessary for obtaining an efficient estimate.

Thus, we are motivated to focus on the following question:

*Can a dynamic adaptive algorithm significantly improve the state of the art sample complexity of MCMC-mean estimator, and, in particular, minimize the complexity dependence on loose mixing time upper bounds?*

This paper presents and analyzes a novel MCMC-estimation algorithm DYNAMITE, that provides an affirmative answer to the above question.

1.1 Preliminaries

Let $f : S \to [a, b]$ be a real valued function, and $\pi$ a probability distribution on $S$. We denote the mean and variance of $f$ by

$$
\mu \doteq \mathbb{E}_{x \sim \pi}[f] \doteq \int_S f(x) \, d\pi(x) \quad \text{and} \quad v_\pi \doteq \mathbb{V}_\pi(f) \doteq \mathbb{E}_{x \sim \pi}[(f(x) - \mu)^2].
$$

**MCMC Terminology** All the Markov chains we discuss are assumed to be ergodic and lazy\(^1\) (see [LP17]), with state space $S$ and stationary distribution $\pi$. We denote by $\mathcal{M}$ both the Markov chain and its transition matrix. For any $x, y \in S$, $\mathcal{M}(x, y) \doteq \mathbb{P}(X_i = y|X_{i-1} = x)$, and $\mathcal{M}^k(x, y) = \mathbb{P}(X_{i+k} = y|X_i = x)$, where $\mathcal{M}^k$ is the standard matrix power notation. Let $\nu$ be a probability

\(^1\)We explicitly note when reversibility is needed.
distribution on $S$, by $\mathcal{M}(\nu)$ we mean the distribution of $X_{i+1}$ conditioned on $X_i \sim \nu$, and the distribution of the chain after $k$ steps of $\mathcal{M}$, starting at $X_i \sim \nu$ is denoted by $X_{i+k} = \mathcal{M}^k(\nu)$.

A $\tau$-trace of Markov chain $\mathcal{M}$ is a sequence of $\tau$ consecutive states visited by the chain. We use vector notation to denote traces of Markov chains, e.g., $\vec{X}_{1:\tau} = (X_1, X_2, \ldots, X_\tau)$, (we drop the subscript $1: \tau$ when the length is clear from context).

We measure distance between distributions by the total variation distance (TVD). For a precision parameter $\epsilon > 0$, the $\epsilon$-mixing time, $\tau_{\text{mix}}(\epsilon)$, is the minimum $\tau$ satisfying $\text{TVD}(\mathcal{M}^\tau(\nu), \pi) \leq \epsilon$, for any initial starting distribution $\nu$. In particular, we denote $\tau_{\text{mix}} = \tau_{\text{mix}}(1/4)$, and note that $\tau_{\text{mix}}(\epsilon) \leq \frac{1}{4} \log(\epsilon^{-1}) \tau_{\text{mix}}$. Let the second largest absolute eigenvalue $\mathcal{M}$ be $\lambda$, and denote the relaxation time $\tau_{\text{rel}} = (1 - \lambda)^{-1}$. Note that $\tau_{\text{mix}}$ and $\tau_{\text{rel}}$ are closely related as $(\tau_{\text{rel}}(\mathcal{M}) - 1) \ln(2) \leq \tau_{\text{mix}}(\mathcal{M}) \leq \tau_{\text{rel}}(\mathcal{M}) \ln(\frac{1}{\sqrt{\tau_{\text{mix}}}})$, where $\tau_{\text{mix}} = \min_{x \in S} \pi(x)$, see [LP17].

Since exact parameters of the MCMC process are often unknown and hard to compute, most MCMC estimators rely on some (possibly loose) bounds given as input to the algorithm. We denote by capital letters $\tau_{\text{mix}}, \tau_{\text{rel}}, \Lambda$ and $V_\pi$ the upper bounds given to the algorithm for $\tau_{\text{mix}}, \tau_{\text{rel}}, \Lambda$ and $v_\pi$, respectively.

For an $m$-trace of $\mathcal{M}$, $\vec{X} = (X_1, X_2, \ldots, X_m)$ we define its empirical mean as $\hat{\mu}(\vec{X}) = \frac{1}{m} \sum_{i=1}^{m} f(X_i)$. We are interested in designing algorithms which run a Markov chain $\mathcal{M}$ for $m$ steps, and return $\hat{\mu}(\vec{X}_{1:m})$ as an estimate of $\mu$. We refer to any such algorithm as an MCMC-mean estimator, and measure their (sample) complexity using the following definition:

**Definition 1.1.** Assume algorithm $A$ runs a Markov chain $\mathcal{M}$ for $m$ steps and generates an $m$-trace $X_1, X_2, \ldots, X_m$. Algorithm $A$ uses this trace to find an estimate $\hat{\mu}$ of $\mu = \mathbb{E}_\pi[f]$. If there exists $m_A(\mathcal{M}, f, \epsilon, \delta)$ such that for any $m \geq m_A(\mathcal{M}, f, \epsilon, \delta)$, we have $\mathbb{P}(|\mu - \hat{\mu}| \geq \epsilon) \leq \delta$ (an $(\epsilon, \delta)$-approximation of $\mu$), we call $m_A(\mathcal{M}, f, \epsilon, \delta)$ the sample complexity of algorithm $A$.

### 1.2 Related Work

The complexity of any MCMC-mean estimator depends, directly or through a proxy, on two unrelated parameters: a function dependent parameter, such as $v_\pi$, the stationary variance of $f$; and chain dependent parameter, such as $\tau_{\text{mix}}$ or $\tau_{\text{rel}}$, the mixing or relaxation time of the Markov chain. Since these parameters are not known in general, efficient MCMC-mean estimator starts with some upper bound on these values, often through related parameters that are easier to estimate. For example, $v_\pi$ can be bounded by $R^2$, and $\tau_{\text{mix}}$ can be bounded by $\tau_{\text{rel}}$ and $\lambda$.

The simplest MCMC-mean estimator averages samples taken at least $\tau_{\text{mix}} \geq \tau_{\text{mix}}$ Markov chain transitions apart. It is not hard to verify that the sample complexity of this estimator is $\Theta\left(\tau_{\text{mix}}(\epsilon/2R)(R^2/\epsilon^2) \log \frac{1}{\epsilon} \right) = \Omega\left(\tau_{\text{rel}} \cdot (R^2/\epsilon^2) \log(\frac{R}{\epsilon}) \log(\frac{1}{\epsilon}) \right)$ chain transitions.

More efficient estimators compute the average over the entire trace of the Markov chain, and their complexity depends on a known upper-bound $\tau_{\text{rel}}$ on the relaxation time $\tau_{\text{rel}}$. Some of these bounds are variance agnostic [Lez98, Mia12, LP04, FJS18], and they imply MCMC-mean estimators with complexity $m_{\text{Hoff}}(\mathcal{M}, f, \epsilon, \delta) \leq \Theta\left(\tau_{\text{rel}} \ln(\frac{1}{\epsilon}) \frac{R^2}{\epsilon^2} \right)$. These bounds, also known as Hoeffding-type bounds, are used ubiquitously due to their simplicity and convenience. Unfortunately, they are generally much looser than variance-aware bounds [BC17, Lem20, GGW14, MSA08]. For example, having a known upper bound $V_\pi \geq v_\pi$, Bernstein type bounds imply MCMC mean estimators with complexity $m_{\text{Bern}}(\mathcal{M}, f, \epsilon, \delta) \leq \Theta\left(\tau_{\text{rel}} \ln(\frac{1}{\epsilon}) \left(R^2/\epsilon^2 + V_\pi/\epsilon^2 \right) \right)$.

In all of the above classic bounds, dependencies on function-specific terms are obtained separately from chain-specific terms. In contrast, Rabinovich et al. [RRJW20] analyze the sample complexity of MCMC-mean estimation using what they term the function-specific mixing
time, written $\tau_{\text{mix}}(f, M, \varepsilon)$ (parentheses omitted when clear from context), which always obeys $\tau_{\text{mix}}(f, M, \varepsilon) \leq \tau_{\text{mix}}(M, \varepsilon)$. They prove a Hoeffding like bound showing MCMC mean estimation can be done using $\tilde{O}(\tau_{\text{mix}} R^2 / \varepsilon^2 \log \frac{1}{\delta})$ samples where $\tau_{\text{mix}} \geq \tau_{\text{mix}}$ is an a priori known upper-bound. Another important bound was derived by Paulin [Pau15], who uses the (chain-specific) asymptotic variance, defined as $v_{\text{asy}}(f, M) \doteq \lim_{T \to \infty} E[\frac{1}{T} \sum_{i=1}^{T} (f(X_i) - \mu)^2]$, which, we prove (see lemma 2.1), is smaller than $2\tau_{\text{rel}} v_\pi$, and he shows sample complexity of $m_{\text{Paulin}}(\varepsilon, \delta) \in \Theta\left((\tau_{\text{rel}} \cdot \frac{R}{\varepsilon} + v_{\text{asy}} / \varepsilon^2) \ln(\frac{1}{\delta})\right)$.

While both Rabinovich et al’s [RRJW20] function-specific mixing time $\tau_{\text{mix}}$ and Paulin’s [Pau15] chain-specific asymptotic variance $v_{\text{asy}}$ beat prior bounds in theory, because of their more sophisticated definitions, obtaining tight upper bounds for them is harder than their classic counterparts e.g., $\tau_{\text{mix}}$, $\tau_{\text{rel}}$ or $v_\pi$. Indeed, a major flaw of all above bounds is their strict dependence on known upper bounds, thus their efficiency can highly be deteriorated by the looseness of said bounds.

In order to remove dependencies on known variance upper-bounds, MCMC-variance estimators, which estimate variance proxies by running the same chain as the mean estimator, have been used by practitioners [SS95, HJP10, JHCN06, FJ10, VFJ18, GV20, CBK19]. Among them, the most popular are the spectral variance (SV) [VFJ18, BIM+20, VF18] and batched means (BM) [FJ10, FJ10, JHCN06, HJP10, SS95, CBK19] estimators. The SV takes a weighted sum of covariances estimated from a single trace. The BM estimates the asymptotic variance by dividing the trace of the chain to batches, and estimating the empirical variance of the mean of each batch.

Since most of theoretical analysis of these methods only concentrate on asymptotic convergence, these estimators generally lack guarantees for finite samples which are necessary for rigorous bounds in applications. Furthermore, all of these estimators are biased (but consistent or asymptotically unbiased). To our knowledge no unbiased MCMC-variance estimator is known.

To circumvent dependency on known bounds on mixing parameters, some authors estimate the mixing or relaxation time of a Markov chain from a trace [HKL+15, HKL+19, WK19, Wol20, LP16, BHOP18]. Hsu et al [HKL+15, LP16, HKL+19] show how to estimate the relaxation time of a reversible Markov chain using $\tilde{O}(\tau_{\text{rel}} / \pi_{\text{mix}})$ steps, and show that $\tilde{O}(\tau_{\text{rel}} |S|)$ is necessary. Wolfer and Kontorovich [Wol20] and Wolfer [WK19] obtain similar results while removing the reversibility condition. These lower bounds eliminate the possibility of estimating $\tau_{\text{rel}}$ with no prior knowledge of it in any number of Markov chain steps that is sub-linear in $|S|$, which is generally prohibitive.

Our work complements all these results. DYNAMITE does not require function dependent parameter as input. It uses a novel parameter $\text{trv}^{(1)}(M, f)$, the inter-trace variance (see definition 2.1), which is estimated efficiently from the data. DYNAMITE requires a chain dependent parameter $\mathcal{T}_{\text{rel}} \geq \tau_{\text{rel}}$, an upper bound on the relaxation time, but the complexity of the algorithm is less dependent on the looseness of this bound compared to previously know algorithms.

1.3 Our Contributions

1. We introduce a novel statistical measure for MCMC-mean estimator, the inter-trace variance $\text{trv}^{(\tau_{\text{rel}})}(M, f)$ (definition 2.1). This tool yields improved estimator performance and analysis. We show that $\text{trv}^{(\tau_{\text{rel}})}(M, f) \leq v_\pi(f)$, and that $\text{trv}^{(\tau_{\text{rel}})}(M, f)$ can be estimated efficiently from the data.

2. We devise a new unbiased MCMC variance estimator $\hat{v}$ and show finite-sample bounds on its error. We use this estimator, together with other concepts that we develop here, to rigorously bound $\text{trv}^{(\tau_{\text{rel}})}$. (Section 2.1)

3. Leveraging $\text{trv}^{(\tau_{\text{rel}})}(M, f)$, we design DYNAMITE, a dynamic MCMC-mean estimator that adapts to the observed data. Given an upper bound on the relaxation time, $\mathcal{T}_{\text{rel}} \geq \tau_{\text{rel}}$, and a bound
R on the range of f, we prove that the complexity (measured in Markov chain steps) of DYNAMITE is $O\left(\frac{\tau_{rel}R}{\epsilon} + \tau_{rel}\text{trv}(\tau_{rel})/\epsilon^2\right)$, without a priori knowledge of $\tau_{rel}$ or $\text{trv}(\tau_{rel})$. (Section 3)

4. For small $\epsilon$ (high-precision case), the complexity of DYNAMITE is dominated by the term $\tau_{rel}(M)\text{trv}(\tau_{rel})(M,f)$, thus, in the high precision regime, the complexity of our algorithm is less dependent on a (possibly loose) bound $\tau_{rel}$ compared to previously known algorithms.

5. We show that $\tau_{rel}(M)\text{trv}(\tau_{rel})(M,f)$ (but not $\tau_{rel}(M)v_\pi(f)$) leads to bounds competitive with the lower bounds obtained from the central limit theorem of Markov chains [GL78]. (Section 3.1)

6. We show that when $f$’s image is distributed symmetrically on $M$’s traces, we have $\text{trv}(\tau_{rel})(M,f) = o(v_\pi(f))$. In those cases DYNAMITE outperforms prior methods, even when tight bounds on variance and mixing or relaxation times are known. We demonstrate this improvement in two applications: mean estimate on a cycle, and counting $k$-colorings. (Section 2 and Section 4).

7. We use DYNAMITE as a mean-estimation gadget in the FPRAS of Jerrum, Valiant and Vazirani [JVV86], and prove the sample complexity of counting proper colorings in planted partitions of $r$ communities and $|E| = E$ edges with our method is $O(\tau_{rel}E^3/\epsilon^2)$, improving the bound $O(\tau_{rel}E^3)$ of [Jer95], while assuming no knowledge of the planted-partition structure of the graph.

We chose the JVV FPRAS for simplicity; our methods apply, mutatis mutandis, to other MCMC-based telescoping mean algorithms, e.g., [SVV07, Hub15, Kol18, HK20]. (Section 4)

8. Our work bridges theory and practice by developing solid mathematical foundations for a combination of heuristics, e.g., batching, MCMC variance estimators and mixing diagnosis. These techniques are used by practitioners without proven convergence criteria, giving unreliable results.

## 2 The Inter-Trace Variance

In this section we define the inter-trace variance (definition 2.1) and prove its properties. In order to estimate the inter trace variance with no prior knowledge of it, we introduce an unbiased MCMC variance estimator (§ 2.1). The concept of the inter-trace variance is similar to batched mean variance estimators used by practitioners. Here we develop new mathematical concepts and prove rigorous finite-sample guarantees.

In its general form, the Inter-Trace variance, written $\text{trv}(\tau)$, is defined with respect to a length parameter $\tau$. In our analysis we use $\text{trv}(\tau_{rel})$, calling it the relaxed Inter-Trace variance.

For a parameter $\tau$, consider the set of all $\tau$-traces, and the probability distribution $\pi(\tau)$ on this set defined as $\pi(\tau)(X_1) = \pi(X_1)\prod_{i=1}^{\tau-1}M(X_i, X_{i+1})$, we call a $\tau$-trace sampled from $\pi(\tau)$ a stationary
trace. For an arbitrary $F$ defined on $S^\tau$, by $\mathbb{E}_{X_\sim\pi(\tau)}[F(X)]$, we mean the expectation of $F$ on any $\tau$-trace of the Markov chain when $X_1 \sim \pi$, and other consecutive $X_i$s follow the transition of the chain. Having $\tilde{X}_{1:}\tau = (X_1, X_2, \ldots, X_\tau)$, we define $f_{\text{avg}}^{(\tau)}$ as follows: $f_{\text{avg}}^{(\tau)}(\tilde{X}_{1:}\tau) = (\frac{1}{\tau}) \sum_{i=1}^\tau f(X_i)$.

**Definition 2.1 (Inter-Trace Variance).** Consider $f$, $M$ and $\pi$ as before. For arbitrary $\tau$ and a $\tau$-trace of $M$, $\tilde{X}_{1:}\tau$, let $f_{\text{avg}}^{(\tau)}(\tilde{X}_{1:}\tau)$ be as defined above. We denote the the inter-trace variance of $f$ by $\text{trv}^{(\tau)}(M, f)$, and define it as,

$$\text{trv}^{(\tau)}(M, f) = \mathbb{V}_{X_\sim\pi(\tau)}[f_{\text{avg}}^{(\tau)}(\tilde{X})].$$

Henceforward, we simply use $\text{trv}^{(\tau)}$ when removing $M$ and $f$ does not create ambiguity.

**Remark.** Note that by linearity of expectation, $\mathbb{E}_{X_\sim\pi(\tau)}[f_{\text{avg}}^{(\tau)}(\tilde{X})] = \mathbb{E}_{X \sim \pi}[f(X)] = \mu$, thus, $\text{trv}^{(\tau)}(M, f) = \mathbb{E}_{X_\sim\pi(\tau)}[(f_{\text{avg}}^{(\tau)}(\tilde{X}) - \mu)^2]$.

The following lemma (proved in appx. [A.1]), shows that $\tau\text{trv}^{(\tau)}$ is bounded above by $2\tau_{\text{rel}}v_\pi$. Which implies always equal or better sample complexity than classic variance aware bounds.

**Lemma 2.1.** Suppose a lazy reversible Markov chain $M$. The inter-trace variance obeys

$$v_\pi \geq \text{trv}^{(2)} \geq \text{trv}^{(3)} \geq \cdots,$$

and $\tau\text{trv}^{(\tau)}$ is nondecreasing and bounded as

$$v_\pi \leq 2\text{trv}^{(2)} \leq 3\text{trv}^{(3)} \leq \cdots \leq \lim_{i \to \infty} i\text{trv}^{(i)} \leq (2\tau_{\text{rel}} - 1)v_\pi.$$  

Furthermore, there exists some absolute constant $c > 0$, such that for any $M$, $f$, we have

$$\tau\text{trv}^{(\tau')} \geq c\tau'\text{trv}^{(\tau')} \geq c \lim_{i \to \infty} i\text{trv}^{(i)}, \quad \text{for any } \tau' \geq \tau \geq \tau_{\text{rel}}(M).$$

**Remark.** Paulin’s asymptotic variance of the inter-trace variance as $v_{\text{asy}} = \lim_{i \to \infty} i\text{trv}^{(i)}$, moreover he proves $\text{trv}^{(\tau)} \leq (2\tau_{\text{rel}}/\tau)v_\pi$ (see Thm. 3.1 of [Paul15]). Note that eq. (2) improves and extends this result. Furthermore, in the appendix we show corollary [A.2] which bounds the trace variance using, $\tau_{\text{rel}}$, the function specific relaxation time of Rabinovich et al [RRJW20].

An important consequence of lemma 2.1 is that for $T \geq \tau_{\text{rel}}$, we have

$$\text{trv}^{(T)} \in \Theta\left(\frac{\tau_{\text{rel}}}{T}\text{trv}^{(\tau_{\text{rel}})}\right),$$

which should remind the reader of $\mathbb{V}[\frac{1}{T}\sum_{i=1}^T f(X_i)] = \frac{\sigma^2}{T}$ which holds when $X_i$s are sampled independently from $\pi$. In fact, $\tau_{\text{rel}}\text{trv}^{(\tau_{\text{rel}})}$ captures the behavior of averages of Markovian random variables just as $v_\pi$ does for independent random variables. Furthermore the asymptotic term $\lim_{i \to \infty} i\text{trv}^{(i)}$ appears in the central limit theorem (see §[3.1]), where this average is Gaussian, and in both the finite and infinite cases, such variances are intimately tied to sample complexity. Thus, we express the sample complexity of DYNAMITE with respect to $\tau_{\text{rel}}\text{trv}^{(\tau_{\text{rel}})}$.

**The Trace Chain** The inter-trace variance can be thought of as the variance of $f_{\text{avg}}^{(\tau)}$ over stationary traces of length $\tau$, i.e., $\pi^{(\tau)}$. Letting $S^{(\tau)}$ be the space of all $\tau$-traces, we now define the $\tau$-trace chain, which naturally groups the output of $M$ into $\tau$-traces.

**Definition 2.2 (Trace chain).** For a Markov chain $M$ on state space $S$, we define the trace chain $M^{(\tau)}$ on state space $S^{(\tau)}$ as follows: given $\bar{a} = (a_1, a_2, \ldots, a_\tau)$ and $\bar{b} = (b_1, b_2, \ldots, b_\tau)$ in $S^{(\tau)}$, the probability of going from $\bar{a}$ to $\bar{b}$ is $M^{(\tau)}(\bar{a}, \bar{b}) = M(a_\tau, b_1) \prod_{i=1}^{\tau-1} M(b_i, b_{i+1})$. 


The definition of the trace chain in terms of transition probabilities is perhaps slightly unintuitive, but we note that an $m$-trace $X_{1,m}$ drawn from $\mathcal{M}$ is equivalently distributed to a $k$-trace $X_{1,k} = (X_1, \ldots, X_k)$ drawn from $\mathcal{M}^{(\tau)}$, where $k = m/\tau$ and each $X_i$ is a contiguous disjoint sub-trace of $X$, i.e., $X_1 = (X_1, \ldots, X_\tau)$, $X_2 = (X_{\tau+1}, \ldots, X_{2\tau}), \ldots, X_k = (X_{(k-1)\tau+1} \ldots X_m)$. We show in lemma A.3 that $\pi^{(\tau)}$ is the stationary distribution of $\mathcal{M}^{(\tau)}$, i.e., $\text{trv}^{(\tau)}(f, \mathcal{M})$ is the variance of $f^{\tau}$ on stationary distribution of $\mathcal{M}^{(\tau)}$.

When is $\text{trv}^{\tau_{\text{rel}}} = o(v_\pi)$? Plugging in $\tau = \tau_{\text{rel}}$ in Equation 2 we obtain $\tau_{\text{rel}} \text{trv}^{(\tau_{\text{rel}})} \leq 2\tau_{\text{rel}} v_\pi$, i.e., $\text{trv}^{(\tau_{\text{rel}})} = O(v_\pi)$. Using projection chains (see definition 2.3), we can show that the intertrace variance becomes smaller when $f$ is symmetrically projected on traces of $\mathcal{M}$. Example 2.2 illustrates this observation. In section 4 we present an application to a counting problem and show approximate-symmetry suffices.

Definition 2.3 (Projection chain [LP17]). Having an equivalence relationship $\simeq$ on $S$ and classes $\hat{S} = \{[x]; x \in S\}$ such that if $x \simeq x'$, then $\mathcal{M}(x, [y]) = \mathcal{M}(x', [y])$, we call the Markov chain $\hat{\mathcal{M}}$ with state space $\hat{S}$ and transition probabilities $\hat{\mathcal{M}}([x], [y]) = \mathcal{M}(x, [y])$ a projection chain (see section 2.3 of [LP17] for full discussion).

We now introduce a class of functions having equal means and stationary variances, but projecting differently on traces of a fixed chain. While they all have equal $v_\pi$, in lemma 2.3 (proof in appx. A.1.1), we prove that under appropriate parameterization, $\text{trv}^{(\tau_{\text{rel}})}$ takes any arbitrary value.

Figure 1: The image of $f_i$s on a cycle of length $n = 16$, and corresponding projection chains with the same trace variance.

Example 2.2. Consider the Markov chain $\mathcal{C}$, known as the cycle, defined on $[n] = \{1, 2, \ldots, n\}$ with transition probabilities: $C(i, i) = 1/2$, $C(i, i+1) = 1/4$ and $C(i, i-1) = 1/4$, where $i-1$ and $i+1$ are taken mod $n$. Clearly the stationary distribution on this chain is uniform, and it is known that the mixing time and relaxation time are both $\Theta(n^2)$. The following class of functions defined on $[n]$ all satisfy $E_\pi(f_i) = 1/2$ and $v_\pi(f_i) = 1/4$. However, as shown in lemma 2.3, as the image of $f_i$s distribute more evenly on $\mathcal{M}$’s traces $\text{trv}^{(\tau_{\text{rel}})}(f_i, \mathcal{M})$ becomes smaller.

For $1 \leq i < \frac{n}{2}$, let $f_i : [n] \rightarrow \{0, 1\}$ be $f_i(x) = 0$ if and only if $x \text{ mod } 2i < i$, so $f_i$’s image on the cycle is consecutive length-$i$ runs of 0s and 1s (see fig. 1). Note that in the two extreme cases, we have, (1) $f_1 : [n] \rightarrow \{0, 1\}, f_1(x) = 0$ if and only if $x \text{ mod } 2 = 0$, and (2) $f_{n/2} : [n] \rightarrow \{0, 1\}, f_{n/2}(x) = 0$ if and only if $x \leq \frac{n}{2}$.

On equivalence classes of mod 2, we find the corresponding projection chain, denoted by $\mathcal{C}_{f_1}$, which has transition probabilities: $C_{f_1}(0, 1) = C_{f_1}(1, 0) = 1/4$ and $C_{f_1}(0, 0) = C_{f_1}(1, 1) = 1/2$. Similarly, for arbitrary $i$, we denote the projection chain on equivalence classes of mod $2i$ by $\mathcal{C}_{f_i}$. Clearly for each $i$, the trace variance of $f_i$ on $\mathcal{C}$ and on $\mathcal{C}_{f_i}$ are distributed identically (see fig. 1).

Lemma 2.3. For $1 \leq i < \frac{n}{2}$, let $f_i$ be defined as above, we have $\text{trv}^{(\tau_{\text{rel}})}(\mathcal{C}, f_i) = O(i/\sqrt{n^2})$ e.g., $\text{trv}^{(\tau_{\text{rel}})}(\mathcal{C}, f_{n/2}) = \Theta(1)$, and $\text{trv}^{(\tau_{\text{rel}})}(\mathcal{C}, f_1) = \Theta(1/n^2)$. 


The reader may rightfully remark that the above scenario, in which the image of a function cleanly partitions the state space is rare. In Section 4, we develop new concepts to identify scenarios for more complex chains, demonstrating a case of functions whose image is approximately symmetric on traces of the chain, showing the above projection trick works with some modifications.

2.1 An unbiased MCMC-variances estimator: A tale of two chains

In the following we introduce an unbiased MCMC estimator for the variance, and prove its concentration guarantees for finite samples. To our knowledge, all existing MCMC variance estimators are biased but consistent and asymptotically unbiased, and finite sample guarantees are not provided.

Definition 2.4 (Two chain variance estimator). Suppose $\vec{X}_1 = (X_{11}, X_{12}, \ldots, X_{1m})$ and $\vec{X}_2 = (X_{21}, X_{22}, \ldots, X_{2m})$ are two independent $m$-traces drawn from $\mathcal{M}$ at stationarity. The estimator is defined as

$$\hat{\nu}(\mathcal{F}, \mathcal{M}) = \frac{1}{2m} \sum_{i=1}^{m} (\mathcal{F}(X_{1i}) - \mathcal{F}(X_{2i}))^2,$$

where we drop the parentheses when $\mathcal{M}$ and $\mathcal{F}$ are clear from the context.

The following lemma shows finite sample concentration of $\hat{\nu}$ and it is proved in appx. A.2.

Lemma 2.4 (A Tail of Two Chains). Let $\hat{\nu}$ be as defined above. It holds that $\hat{\nu}$ is an unbiased estimator for variance, i.e., $E[\hat{\nu}(\mathcal{F}, \mathcal{M})] = \mathbb{V}[\mathcal{F}]$. Furthermore, letting $R$ be range of $\mathcal{F}$ we have

$$\mathbb{P}(|\mathbb{V}[\mathcal{F}] - \hat{\nu}| > \varepsilon) \leq \delta \quad \text{for} \quad \varepsilon \in \Theta\left(\frac{\tau_{\text{rel}}R^2 \ln \frac{1}{\delta}}{m} + R\sqrt{\frac{\tau_{\text{rel}}\hat{\nu} \ln \frac{1}{\delta}}{m}}\right),$$

$$\Rightarrow m\hat{\nu}(\mathcal{M}, f, \delta, \varepsilon) \in \Theta\left(\frac{R^2}{\varepsilon \delta} + \frac{\mathbb{V}[\mathcal{F}]R^2}{\varepsilon^2} \tau_{\text{rel}} \log \frac{1}{\delta}\right).$$

Inter-trace variance estimation. In order to avoid assuming prior knowledge of variance or inter-trace variance in DynaMITE, we employ the above estimator to the the trace chain, $\mathcal{M}(\tau_{\text{rel}})$. In lemma A.4 we show $\mathcal{M}(T)$ has constant relaxation time when $T \geq \tau_{\text{rel}}$ (e.g., $T = \tau_{\text{rel}}$). Thus, we can estimate $\text{trv}(\tau_{\text{rel}})$ by employing lemma 2.4 and using $\hat{\nu}(f_{\text{avg}}, \mathcal{M}(\tau_{\text{rel}}))$.

3 DynaMITE

In this section we present the DynaMITE: DYNAMIC MCMC INTER-TRACE VARIANCE ESTIMATION method. We show that its sample complexity is dependent on the apriori unknown relaxed trace variance $\text{trv}(\tau_{\text{rel}})$. This section is prelude to the themes and techniques used in DynaMITE, which are fully developed in appx. A.3. In § 3.1 we compare our results with prior work.
A Prelude  Algorithm 1 shows a pseudocode of DYNAMITE. There are two main techniques constituting DYNAMITE: trace averaging and progressive sampling. To simplify the presentation, we separated the pseudocode to two subroutines: MCMCPro employs progressive sampling to dynamically estimate variance (with no prior knowledge of it), thus it is itself a dynamic MCMC mean estimator. DYNAMITE calls MCMCPro while using trace averaging, which improves sample complexity dependence from \( v_x \) to \( \text{trv}(\tau_{rel}) \). While application of progressive sampling is common in algorithm design [PJO99], here we introduce trace averaging.

**Definition 3.1 (Trace averaging).** Consider function \( f \), a trace of length \( m \) of \( M \) \( \bar{X}_{1:m} = (X_1, X_2, \ldots, X_m) \), and an upper bound on \( M \)’s relaxation time \( T_{rel} \). For any \( \bar{X} \)’s contiguous \( T_{rel} \)-subtrace \( \mathcal{X} = (X_{j+1}, X_{j+2}, \ldots, X_{j+T_{rel}}) \), let \( f^{(T_{rel})}(\mathcal{X}) = \left( \frac{1}{T_{rel}} \right) \sum_{i=1}^{T_{rel}} f(X_{j+i}) \). Trace averaging is then the process of estimating \( \mathbb{E}_\pi[f] \) by employing an MCMC mean estimator for \( f^{(T_{rel})} \) on the trace chain \( \mathcal{M}(T_{rel}) \).

We also employ progressive sampling. Beginning with a small number of chain transitions, we progressively increase the sample size until a stopping condition is met.

In each round, we calculate the unbiased variance estimator as developed in §2.4.1 (see lines 8–12 in alg. 1), using it we find a high probability upper-bound on \( \text{trv}(\tau_{rel}) \) (line 13 of alg. 1). From it and by employing an MCMC Bernstein bound e.g., thm. B.2 we obtain a suitable stopping condition guaranteeing that the empirical mean is sufficiently accurate (line 14 and 15 of alg. 1). Using \( T_{rel} \text{trv}(\tau_{rel}) \leq c \tau_{rel} \text{trv}(\tau_{rel}) \) (for universal constant \( c \)) as proved in lemma 2.1 we show our bounds in terms of the relaxed inter-trace variance \( \text{trv}(\tau_{rel}) \).

The following theorems, proved in appx. A.3, guarantee correctness and efficiency of DYNAMITE. Since many of the lemmas use stationary traces, in line 4 of alg. 1 we use the standard warm start trick which is running the chain from arbitrary starting points until proximity to stationarity is reached (see appx. A.3.3).

**Theorem 3.1 (Correctness of DYNAMITE).** Consider a Markov chain \( M \) and its relaxation time upper bound \( T_{rel} \), function \( f \) and its range \( R \). For arbitrary \( \varepsilon \) and \( \delta \), and starting points \( x_0, x_1 \in S \) taking \( \hat{\mu} \) as either of

1. \( \hat{\mu} \leftarrow \text{MCMCPro}((x_0, x_1), M, T_{rel}, f, \varepsilon, \delta) \); or
2. \( \hat{\mu} \leftarrow \text{DYNAMITE}((x_0, x_1), M, T_{rel}, f, \varepsilon, \delta) \) for lazy \( M \).

we will have a \((\varepsilon, \delta)\) estimator for \( \mu = \mathbb{E}_\pi[f] \), i.e, \( P(\left| \hat{\mu} - \mu \right| \geq \varepsilon) \leq \delta \)

**Theorem 3.2 (Efficiency of DYNAMITE).** Suppose as in thm. 3.1 With probability at least \( 1 - \delta \), it holds that total sample complexity of MCMCPro obeys

\[
m_{\text{MCMCPro}}(M, f, \varepsilon, \delta) \in \mathcal{O}\left( T_{rel} \log \left( \frac{\log(R/\varepsilon)}{\delta} \right) \left( \frac{R}{\varepsilon} + \frac{v_x}{\varepsilon^2} \right) \right),
\]

and that of DYNAMITE obeys

\[
m_{\text{DYNAMITE}}(M, f, \varepsilon, \delta) \in \mathcal{O}\left( \log \left( \frac{\log(R/\varepsilon)}{\delta} \right) \left( \frac{T_{rel}R}{\varepsilon} + \frac{\tau_{rel} \text{trv}(\tau_{rel})}{\varepsilon^2} \right) \right)
\]

### 3.1 Discussion and comparison with prior work

We now contrast the DYNAMITE sample complexity bound with those of prior art. In order to focus on the salient differences between methods, we consider the asymptotic high-precision regime

---

9
(i.e., sample complexity as $\varepsilon \to 0$), and to divide out the $\frac{1}{\varepsilon^2} \ln \frac{1}{\delta}$ terms, which necessarily appear in all bounds of this ilk, and we then report (in asymptotic notation) the quantity
\[
\text{HIPERM}(\mathcal{M}, f) \doteq \lim_{\varepsilon \to 0} \varepsilon^2 \ln \frac{1}{\delta} m(\mathcal{M}, f, \varepsilon, \delta).
\] (5)

where $m(\mathcal{M}, f, \varepsilon, \delta)$ is the sample complexity as used before.

The Central Limit Theorem for Markov chains  We first contrast our bound with bounds in terms of the 

asymptotic variance of $\mathcal{M}$, defined as
\[
v_{\text{asy}} = \lim_{\tau \to \infty} \tau \text{trv}(\tau).
\]

This quantity is fundamental to mean estimation, as the Markov chain central limit theorem [GL78] states
\[
\lim_{\tau \to \infty} \frac{1}{\sqrt{\tau}} f_{\text{avg}}^{(\tau)}(\mathcal{X}) \sim \mathcal{N}([\mathcal{E}[f]], v_{\text{asy}}),
\]

thus asymptotically, via the CLT we get mean estimation sample complexity
\[
\text{HIPERM}_{\text{CLT}}(\mathcal{M}, f) = \Theta(v_{\text{asy}}) = \Theta(\tau_{\text{rel}} \text{trv}(\tau_{\text{rel}})).
\]

There are two main factors which prevent finite-sample bounds from achieving $\Theta(v_{\text{asy}})$. The first is that, like in the i.i.d. case, the CLT is only an asymptotic result, and for finite samples, $f_{\text{avg}}^{(\tau)}(\mathcal{X})$ may be far from Gaussian. The second is that a priori knowledge of $v_{\text{asy}}$ requires highly sophisticated analysis. Thus, finite-sample guarantees are often stated in terms of mixing times or the stationary variance, instead of these fundamental quantities and they fail to match the CLT.

DynaMITE’s high precision complexity  From Thm. 3.2, we conclude that
\[
\text{HIPERM}_{\text{DynaMITE}} = \tau_{\text{rel}} \text{trv}(\tau_{\text{rel}}) \cdot (1 + \ln \ln \left(\frac{R}{\varepsilon}\right)) = \Theta(v_{\text{asy}} \cdot \ln \ln \left(\frac{R}{\varepsilon}\right)).
\]

Note that our bound matches the CLT, except for the $(1 + \ln \ln \frac{R}{\varepsilon})$ term, which is due to the progressive sampling union bound, and is what allows us to improve dependence on $v_{\text{asy}}$ to $\Theta(v_{\text{asy}})$.

Comparison with asymptotic variance bounds [Pau15]  Paulin [Pau15] (Thm. 3., Eq. 3.20) accounts for finite-sample approximation error with a finite sample bound of
\[
m_{\text{Paulin}}(\mathcal{M}, f, \varepsilon, \delta) = \Theta \left( \ln \frac{1}{\delta} \left( \frac{R \tau_{\text{rel}}}{\varepsilon} + \frac{V_{\text{asy}}}{\varepsilon^2} \right) \right),
\]

which implies \(\text{HIPERM}(\mathcal{M}, f) = \Theta(V_{\text{asy}}).\)

Comparison with MCMC Bernstein bound (thm. 3.2)  Most Bernstein-type bounds depend on (loose) a priori bounds on the stationary variance $v_{\pi}$ of the chain and its relaxation time. Thus they are inferior to our method which depends on $v_{\text{asy}}$.
\[
\text{HIPERM}_{\text{Bern}}(\mathcal{M}, f) = \Theta(\tau_{\text{rel}} V_{\pi}) \implies \text{HIPERM}_{\text{Bern}}(\mathcal{M}, f) = \Omega(v_{\text{asy}}).
\]
Comparison with function specific mixing Hoeffding bound \[\text{RRJW20}\] Rabinovich et al. \[\text{RRJW20}\] show a bound of the form 

\[
m_{\text{Rabi}}(\mathcal{M}, f, \varepsilon, \delta) \leq \Theta \left( \frac{\tau_{\text{mix}}(f, \mathcal{M}, R)}{\varepsilon^2} R^2 \log \frac{1}{\delta} \right) \implies \text{HIPERM}_{\text{Rabi}}(\mathcal{M}, f) \in \Theta \left( \frac{\tau_{\text{mix}}(f, \mathcal{M}, R)}{\varepsilon} R^2 \right)
\]

\[
\implies \text{HIPERM}_{\text{Rabi}}(\mathcal{M}, f) \in \Omega \left( v_{\text{asy}} \log \frac{R}{\varepsilon} \right).
\]

The above bound is loose because it depends on an upper-bound of \(\tau_{\text{mix}}\), rather than \(\tau_{\text{rel}}\) (see definition \[A.2\]), and because, as a Hoeffding like-bound, it uses \(R^2\) instead of any variance proxy. More importantly, as prior work has shown difficulty of estimating the standard relaxation time from observations, a serious obstacle in employing it in practice is to find \(\tau_{\text{mix}} \geq \tau_{\text{mix}}\) which needs full spectral decomposition of \(\mathcal{M}\)'s transition matrix.

4 Application to a counting problem

In section \[2\] we found \(\text{trv}^{(\tau_{\text{rel}})} = o(v_\pi)\) for functions whose images partition the state space of \(\mathcal{M}\), producing simpler projection chains (see example \[2.2\]). In this section, our goal is to obtain similar results for the JVV counting to sampling reduction when traces of a Markov chain are hardly distinguishable from a simpler projection chain. Developing some mathematical tools, we show that using DYNAMITE as a mean estimator not only makes the complexity of JVV algorithm less dependent on loose mixing (relaxation) time bounds, but more importantly, we show it significantly reduces computation cost in some instances like counting number of \(k\)-colorings in planted partitions. While we present results for the number of proper \(k\)-colorings of a graph \(G = (V, E)\), similar techniques can be used for other similar counting problems like independent sets.

Exact counting of the number of proper \(k\)-colorings is known to be \#P-hard \[Val79\], thus research has been conducted to find a fully polynomial time randomized approximation scheme (FPRAS) for this problem.

Jerrum \[2\] proved that a Glauber dynamics chain on \(k\)-colorings (see definition \[A.3\]) is rapidly mixing when \(k > 2d_{\text{max}}\). Furthermore, based on the seminal work of Jerrum, Valienta and Vazirani (JVV) on counting self-reducible structures \[JVV86\], he showed a FPRAS for counting \(k\)-colorings using a telescoping sum of MCMC-mean estimation sub-problems (see Section \[B.2\] or \[Jer95\] for more details). We plug-in DYNAMITE as a mean estimation subroutine in Jerrum's FPRAS. For an arbitrary graph \(G\) and fixed \(k\), let \(\mathcal{M}_G\) be the Glauber dynamics chain on \(k\)-colorings, and \(\{f_i\}_{i=1}^\mathcal{E}\) be the functions appearing in Jerrum's reduction whose mean is to be estimated, where \(\mathcal{E} = |E|\), the number of edges in \(G\). We identify a class of graphs, which we call loosely connected (see definition \[A.6\]), and we prove for any \(G\) in this class that \(\text{trv}^{(\tau_{\text{rel}})}(\mathcal{M}_G, f_i) = o(v_\pi)\) (see lemma \[A.8\]).

Using these results, we then prove that for \(G\) sampled from the planted partition model with \(r\) communities we have: \(\text{trv}^{(\tau_{\text{rel}})} = O(v_\pi)/r\). Since JVV reduction needs high precision mean estimations, as it needs a union bound over all intermediate steps, using DYNAMITE is indeed impactful. In particular it reduces the complexity of the \(k\)-coloring fpras of \[Jer95\] for planted partition from \(\tilde{O}(n^2 \mathcal{E}^3/\varepsilon^2)\) down to \(\tilde{O} \left( n^2 \left( \frac{\mathcal{E}^2}{\varepsilon} + \frac{\mathcal{E}^3}{r} \right) \right)\), where \(n = |V|\) and \(\mathcal{E} = |E|\) and \(r\) is the number of communities (see thm. \[A.6\]).

\[2\] This result was improved by many, most importantly \[Vig99\]. See e.g., \[FV07\] for a survey and recent result e.g., \[CGSV21\].
Figure 2: Consider a loosely connected graph $G = (V_r \cup V_b \cup V_c, E)$, and for any $e = (u, v) \in E$, let $f_e$ be as defined in the JVV reduction, and $u, v \in V_r$. Let $G'$ be the graph obtained by removing the edges connecting $V_r$ to the rest of the graph. We show that only a negligible mass of stationary traces of $M_{G'}$ have zero probability in $M_G$, and the remaining traces of $M_{G'}$ can be perfectly coupled to identical traces in $M_G$. We then use projection chain of $M_{G'}$ onto equivalence classes defined by the image of $f_e$. Note that the transition probabilities of this projection chain are identical to projection of $\left(\frac{|V_r|}{n}\right)M_{V_r}$ onto equivalence classes defined by the image of $f_e$, where $\left(\frac{|V_r|}{n}\right)M_{V_r}$ is a Markov chain looping w.p. $(1 - \frac{|V_r|}{n})$ and otherwise transitioning through $M_{V_r}$. Using this observation, we show $\text{trv}(\tau)_{M_G, f_e} \leq \left(\frac{\tau}{\tau_{\text{rel}}(M_{G'})}\right) \text{trv}(\tau)_{M_G', f_e} = o(\pi)$. 

A A Compendium of Complementary Material

We now present all the details backing our results including proofs and rigorous definitions for newly developed concepts. Some of these proofs are involved and are broken to a proof sketch describing ideas and intuitions accompanied by a full proof with rigorous details.

In appx. B we state the bounds, theorems, and algorithms which exist in the literature and are used in our proofs.

We have provided a table of contents at the end of this paper.

A.1 Intra-Trace variance properties

Definition A.1. Consider $\bar{X}_{1:T} : X_1, X_2, \ldots X_T$. We define $C_i$ to be the lag $i$-autocovariance, i.e., the autocovariance of two steps of $M$ (at stationarity) being $i$ apart. i.e., $C_i \triangleq \mathbb{C}(X_1, X_{1+i})$.

Note that for i.i.d samples the autocovariance of any pair of samples $X_i$ and $X_j$ is zero. We do not have independence here nevertheless for reversible Markov chains (See for example Equation 12.9 from [LP17]) we have $C_i \leq \lambda^i \sqrt{\mathbb{V}_i[f] \mathbb{V}_i[f]}$. The following lemmas will be used throughout:

Lemma A.1. Suppose $\bar{X}_{1:T} : X_1, X_2, \ldots X_T$ is a trace of length $T$ of $M$. using the above definition for $C_i$, the trace variance are related as

$$\text{trv}(T) = \frac{1}{T} \pi + \frac{2}{T^2} \sum_{i=1}^{T-1} (T - i) C_i .$$

(6)
Algorithm 1 McmcPro and DYNAMITE routines

1: procedure McmcPro:\((x_0, x_1), \mathcal{M}, \Lambda, f, \varepsilon, \delta) \mapsto \hat{\mu} \\
2: \textbf{Input:} arbitrary initial states \((x_0, x_1) \in S \times S\), Markov chain \(\mathcal{M}\) over \(S\), second absolute eigenvalue upper-bound \(\Lambda\), function \(f : S \to [a, b]\) with \(R = b - a\), confidence interval radius \(\varepsilon\), and failure probability \(\delta \in (0, 1)\). \\
3: \textbf{Output:} Additive \(\varepsilon, \delta\) approximation \(\hat{\mu}\) of \(\mu = \mathbb{E}_\pi[f]\).

4: \((X_{0,1}, X_{0,2}) \sim (\mathcal{M} \otimes \mathcal{M})^{\tau_{\text{unif}}}(x_0, x_1)\) \quad \triangleright \text{Before collecting samples for mean estimation run two independent copies of } \mathcal{M}, \text{ each } \tau_{\text{unif}} \left\lceil \frac{\ln \frac{1}{\min \{\varepsilon, \delta\}}}{\ln \frac{1}{1-\Lambda}} \right\rceil \text{ steps}.

5: \(I \leftarrow \max\left(1, \left\lceil \log_2 \left(\frac{R}{\varepsilon} \right) \right\rceil \right); \quad \alpha \leftarrow \frac{(1 + \Lambda)R \ln \frac{3\varepsilon}{\delta}}{(1 - \Lambda)\varepsilon}; \quad m_0 \leftarrow 0 \quad \triangleright \text{Initialize sampling schedule}

6: \textbf{for } i \in 1, 2, \ldots, I \textbf{ do}

7: \quad m_i \leftarrow \lceil \alpha 2^i \rceil \quad \triangleright \text{Total sample count at iteration } i

8: \quad \textbf{for } j \in (m_{i-1} + 1), \ldots, m_i \textbf{ do}

9: \quad \quad (X_{j,1}, X_{j,2}) \sim (\mathcal{M} \otimes \mathcal{M})(X_{j-1,1}, X_{j-1,2}) \quad \triangleright \text{Run both chains up to step } m_i

10: \quad \textbf{end for}

11: \quad \hat{\mu}_i \leftarrow \frac{1}{2m_i} \sum_{j=1}^{m_i} (f(X_{j,1}) + f(X_{j,2})) \quad \triangleright \text{Empirical mean}

12: \quad \hat{v}_i \leftarrow \frac{1}{2m_i} \sum_{j=1}^{m_i} (f(X_{j,1}) - f(X_{j,2}))^2 \quad \triangleright \text{Empirical variance}

13: \quad u_i \leftarrow \hat{v}_i + \frac{11(1 + \sqrt{2}) \frac{1 + \Lambda}{2} R^2 \ln \frac{3\varepsilon}{\delta}}{(1 - \Lambda)m_i} + \sqrt{\frac{1 + \Lambda)R^2 \hat{v}_i \ln \frac{3\varepsilon}{\delta}}{(1 - \Lambda)m_i}} \quad \triangleright \text{Variance upper bound}

14: \quad \hat{\epsilon}_i \leftarrow \frac{10 R \ln \frac{3\varepsilon}{\delta}}{(1 - \Lambda)m_i} + \sqrt{\frac{(1 + \Lambda)u_i \ln \frac{3\varepsilon}{\delta}}{(1 - \Lambda)m_i}} \quad \triangleright \text{Apply Bernstein bound}

15: \quad \textbf{if } (i = I) \lor (\hat{\epsilon}_i \leq \varepsilon) \textbf{ then} \quad \triangleright \text{Terminate if accuracy guarantee is met}

16: \quad \textbf{return } \hat{\mu}_i

17: \quad \textbf{end if}

18: \quad \textbf{end for}

19: \textbf{end procedure}

20: procedure DYNAMITE:\((x_0, x_1), \mathcal{M}, \Lambda, f, \varepsilon, \delta) \mapsto \hat{\mu} \\
21: \textbf{Input:} Initial state \((x_0, x_1) \in S \times S\), lazy Markov chain \(\mathcal{M}\) over \(S\), second absolute eigenvalue upper-bound \(\Lambda\), function \(f : S \to [a, b]\), confidence interval radius \(\varepsilon\), and failure probability \(\delta \in (0, 1)\).

22: \textbf{Output:} Additive \(\varepsilon, \delta\) approximation \(\hat{\mu}\) of \(\mu = \mathbb{E}_\pi[f]\).

23: \(T \leftarrow \left\lceil \frac{1 + \Lambda}{1 - \Lambda} \ln \sqrt{2} \right\rceil \quad \triangleright \text{Select } T \text{ s.t. } \tau_{\text{rel}}(\mathcal{M}^{(T)}) \leq 2 \)

24: \((X_0, X_1) \leftarrow \left(\underset{\text{Arbitrary } s_1, \ldots, s_{T-1} \in \bar{S}^{T-1}}{(s_1, s_2, s_3, \ldots, s_{T-1}, x_0), (s_1, s_2, s_3, \ldots, s_{T-1}, x_1)}\right) \quad \triangleright \text{Initialize trace chain state}

25: \textbf{return } \text{McmcPro}((X_0, X_1), \mathcal{M}^{(T)}, \Lambda^T, f_{avg}, \varepsilon, \delta) \quad \triangleright \text{Run McmcPro on trace chain}

26: \textbf{end procedure}

\textsuperscript{*}Starting at Nonstationarity. The MCMC Bernstein bound and McDiarmid bound we use here assume samples from stationary distribution. Often we can't assume even a single (perfectly) stationary sample may be efficiently drawn, and realistically can only start with an arbitrary \(x \in \text{Support}(\pi)\). Thus we use standard warm start technique (widely used in MCMC algorithms), where the chain is run from an arbitrary point for its uniform mixing time, and then DYNAMITE is run, applying a standard nonstationarity correction (see appx. \textsuperscript{A.3.3}) to account for the nonstationary start.
Proof. The trace variance is

\[ \text{trv}^{(T)} = \mathbb{E} \left[ \left( \frac{1}{T} \sum_{i=1}^{T} (f(X_i) - \mu) \right)^2 \right] = \mathbb{E} \left[ \frac{1}{T^2} \sum_{i=1}^{T} \sum_{j=1}^{T} (f(X_i) - \mu)(f(X_j) - \mu) \right] \]

\[ = \frac{1}{T} v_\pi + \frac{2}{T^2} \sum_{i=1}^{T-1} (T - i) C_i. \]

We now show Lemma 2.1.

Lemma 2.1. Suppose a lazy reversible Markov chain \( \mathcal{M} \). The inter-trace variance obeys

\[ v_\pi \geq \text{trv}^{(2)} \geq \text{trv}^{(3)} \geq \cdots, \tag{1} \]

and \( \tau \text{trv}^{(\tau)} \) is nondecreasing and bounded as

\[ v_\pi \leq 2 \text{trv}^{(2)} \leq 3 \text{trv}^{(3)} \leq \cdots \leq \lim_{i \to \infty} i \text{trv}^{(i)} \leq (2 \tau_{\text{rel}} - 1) v_\pi. \tag{2} \]

Furthermore, there exists some absolute constant \( c > 0 \), such that for any \( \mathcal{M} \), \( f \), we have

\[ \tau \text{trv}^{(\tau)} \geq c \tau' \text{trv}^{(\tau')} \geq c \lim_{i \to \infty} i \text{trv}^{(i)}, \quad \text{for any } \tau' \geq \tau \geq \tau_{\text{rel}}(\mathcal{M}). \tag{3} \]

Proof. We first show (1). First note that

\[ T \text{trv}^{(T)} = v_\pi + 2 \sum_{i=1}^{T-1} \frac{T - i}{T} C_i. \]

Laziness implies nonnegativity of each covariance term \( C_i \), and the weight \( \max(0, \frac{T - i}{T}) = \max(0, 1 - \frac{i}{T}) \) assigned to each \( C_i \) in \( T \text{trv}^{(T)} \) is monotonically increasing in \( T \).

We now show (2). We first show that laziness and reversibility implies \( C_i \) is monotonically decreasing \( \forall i \geq 0 \). To see this, consider again the spectral decomposition

\[ C_t = \int \lambda_i^{[t]} \, dE_f(\lambda_i), \]

and note that each \( \lambda_i^{[t]} \) in the integral is decreasing in \( t \) (as laziness implies \( \lambda_i \in [0, 1] \), and \( \forall x \in [0, 1] : x^{t+1} \leq x^t \)), thus by linearity of the integral, \( C_i \) is monotonically decreasing.

Now, as we have shown that \( C_i \) is monotonically decreasing, and for all \( t \), the cumulative weight of \( C_{1:t} \) in \( \text{trv}^{(T)} \) always exceeds the corresponding weight in \( \text{trv}^{(T+1)} \), we may conclude

\[ \text{trv}^{(T+1)} - \text{trv}^{(T)} \leq 0, \]

which implies (2).

We now show (3).
The Bernstein’s inequality applied to \( f_{\text{avg}} \) on blocks of size \( \tau_{\text{rel}} \), imply that (see thm. B.2)

\[
m \leq m_1 = \left\lceil \tau_{\text{rel}}(\mathcal{M}(T)) \right\rceil T \left( \frac{10R}{\varepsilon} + \frac{4\text{trv}(T)}{\varepsilon^2} \right) \ln \frac{2}{\delta} \leq C_1 T \left( \frac{10R}{\varepsilon} + \frac{4\text{trv}(T)}{\varepsilon^2} \right) \ln \frac{2}{\delta},
\]

where \( C_1 \) is the (constant) upper-bound to \( \left\lceil \tau_{\text{rel}}(\mathcal{M}(T)) \right\rceil \), are sufficient to \( \varepsilon, \delta \) estimate \( \mathbb{E}[f] \) with the empirical mean.

The Markov-chain central limit theorem, together with standard Gaussian anticoncentration inequalities, imply that

\[
m \geq m_2 = \frac{C_2 \text{trv}_\text{asy}}{\varepsilon^2} \ln \frac{1}{\delta},
\]

where \( C_2 \geq 1 \) is an absolute constant, are necessary (asymptotically) to \( \varepsilon, \delta \) estimate \( \mathbb{E}[f] \) with the empirical mean.

We thus have that, asymptotically, \( m_1 \geq m \geq m_2 \), from which we derive

\[
1 \leq \lim_{\varepsilon \to 0} \lim_{\delta \to 0} \frac{m_1}{m_2} = \frac{4C_1 T \text{trv}(T)}{C_2 \text{trv}_\text{asy}} \implies \frac{C_2}{4C_1} \text{trv}_\text{asy} \leq T \text{trv}(T),
\]

which concludes the proof.

Rabinovich et al. [RRJW20] define function specific mixing and relaxation times using the spectral decomposition of transition matrix of \( \mathcal{M} \). We restate this definition here and using it we extend lemma 2.1.

**Definition A.2** (Function specific relaxation time, \( \tau_{\text{rel}}(\mathcal{M}, f) \)). Given \( f \) and Markov chain \( \mathcal{M} \) defined on \( S \), Let \( P \) be the transition matrix of \( \mathcal{M} \) and consider all eigenvectors of \( P \) which are orthogonal to \( f \). Let \( \lambda_1 = 1, \lambda_2, \lambda_3, \ldots, \lambda_k \) be the eigenvalues corresponding to the remaining eigenvectors. We define \( \tau_{\text{rel}}(\mathcal{M}, f) = \frac{1}{1-\lambda_*} \), where \( \lambda_* = \max_{i=2:k} |\lambda_i| \).

It is now difficult to see the following corollary, which extends lemma 2.1 using the function specific relaxation time.

**Corollary A.2.** Suppose as in lemma 2.1. Then

\[
v_\pi \leq 2\text{trv}^{(2)} \leq 3\text{trv}^{(3)} \leq \cdots \leq \lim_{i \to \infty} \text{itrv}^{(i)} \leq (2\tau_{\text{rel}} - 1)v_\pi.
\]

Furthermore, there exists some absolute constant \( c > 0 \), such that for any \( \mathcal{M}, f \), we have

\[
\tau \text{trv}(\tau') \geq c \tau' \text{trv}(\tau') \geq c \lim_{i \to \infty} \text{itrv}^{(i)}, \quad \text{for any } \tau' \geq \tau \geq \tau_{\text{rel}}(\mathcal{M}).
\]

**Proof.** Proof of the first two inequalities is essentially identical to that of lemma 2.1 where we use the fact that in the spectral decomposition, eigenvalues orthogonal to \( f \) have 0 weight in the integral. We thus ignore all eigenvalues orthogonal to \( f \).

In particular, for reversible \( \mathcal{M} \), the spectral decomposition allows us to bound the autocovariance as

\[
|C_1(f)| = \left| \int \lambda^{|t|} dE_f(\lambda) \right| \leq \lambda^{|t|}(f) v_\pi,
\]

where \( \lambda^{|t|}(f) \) is the second largest absolute eigenvalue not orthogonal to \( f \). From here, all bounds on inter-trace variances follow as before, now using this inequality in place of the standard spectral decomposition inequality for covariances.
A.1.1 Bounding trace variance using projection chains: Cycle

We now seek to show lemma 2.3 restated below. Recall here that for \(1 \leq i \leq \frac{n}{2}\), we take \(f_i : [n] \rightarrow \{0,1\}\) be \(f_i(x) = 0\) if and only if \(x \mod 2i < i\), so \(f_i\)'s image on the cycle is periodic length-\(i\) runs of 0s and 1s (see fig. 1).

**Lemma 2.3.** For \(1 \leq i \leq \frac{n}{2}\), let \(f_i\) be defined as above, we have \(\text{trv}_{\text{rel}}(\mathcal{C}, f_i) = \Theta(i^2/n^2)\) e.g., \(\text{trv}_{\text{rel}}(\mathcal{C}, f_{n/2}) = \Theta(1)\), and \(\text{trv}_{\text{rel}}(\mathcal{C}, f_1) = \Theta(1/n^2)\).

**Proof.** We first prove that \(\text{trv}_{\text{rel}}(f_i) \leq \Theta(i^2/n^2)\). Note that \(x \mapsto x \mod 2i\) partitions the set \([n]\) into \(2i\) partitions (see fig. 1), and the relaxation time of the projection chain \(\mathcal{C}_{f_i}\) is \(\Theta(i^2)\). Since \(f_i\)'s image on traces of \(f_i\) on \(\mathcal{C}\) and \(\mathcal{C}_{f_i}\) is distributed identically, for \(\tau = \tau_{\text{rel}}(\mathcal{C})\) by applying lemma 2.1 we get \(\text{trv}_{\text{rel}}(\mathcal{C}_{f_i}, f_i) \leq \frac{\tau_{\text{rel}}(\mathcal{C}_{f_i})}{\tau} \text{trv}_{\text{rel}}(\mathcal{C})\). Replacing the known we get: \(\text{trv}_{\text{rel}}(\mathcal{C}, f_i) = \text{trv}_{\text{rel}}(\mathcal{C}_{f_i}, f_i) \leq \Theta(i^2/n^2)\).

A Sketch. Let \(S_0 = \{i/3 + 1, i/3 + 2, \ldots, 2i/3\} \subseteq [n]\), we show that any trace of length at most \(i^2/9\) starting at \(S_0\), or any other middle point of other monochromatic regions, have low probability of escaping from it (see Figure 3).

Using this we show that the \(i^2/9\)-trace variance conditioned on starting at one of these middle sub-regions is \(\Theta(1)\). Having a bound on \(\text{trv}_{\text{rel}}(i^2/9)\), we use lemma 2.1 and conclude the premise. The next paragraph presents proof details.

**Figure 3:** Image of \(f_{n/6}\) on a cycle for various values of \(n\), zero values are colored in red, and one values in blue. The trace variance for trace length shorter than \((n/54)^2\), conditioned starting at \(S = \bigcup_{k=1}^{6} S_k\) (circled regions), is at least constant.

**Proving \(\text{trv}_{\text{rel}}(i^2/9)(\mathcal{C}, f_i) \geq \Theta(1)\).** We assume WLOG that \(i\) divides 3: this is so we may analyze an convenient integer-length trace, but the remaining cases hold with constant-factor differences. Note that \(\tau_{\text{rel}} \in \Theta(n^2)\), and our proof operates by analyzing \(I\)-traces, for \(I = i^2/9\), and then applying trace-variance inequalities to draw the desired conclusions.

Assume \(X_1, X_2, \ldots, X_{i^2/9}\) is a trace of the unbiased walk on the cycle, and define \(Y_k = X_{k+1} - X_k\). Thus, \(X_k = \sum_{j=1}^{k-1} Y_j + s_0\) where \(s_0\) is the starting point. Note that \(Y_j\) is a symmetric random variable (thus has equal mean and median), thus, we shall use to apply Lévy’s inequality. The proof strategy here is to lower-bound \(\text{trv}_{\text{rel}}(i^2/9)\) by showing that a constant fraction of stationary traces in \(\mathcal{C}(i^2/9)\) see only one color. We call such traces *homogeneous*, and note that for such traces \(\vec{X}\), we have \((f_{\text{avg}}(\vec{X}) - \mu)^2 = \frac{1}{4}\) (as \(f_{\text{avg}}(\vec{X}) \in \{0,1\}\), and \(\mu = \frac{1}{2}\)), and from there we bound trace variances as appropriate.

We begin with a key step in deriving a lower-bound on the proportion of such homogeneous traces. In particular, take \(S_0 = \{i/3 + 1, i/3 + 2, \ldots, 2i/3\} \subseteq [n]\), and similarly take \(S_k = \{ki + i/3 + 1, ki + i/3 + 2, \ldots, ki + 2i/3\} \subseteq [n]\), in other words each \(S_k\) is the middle third of the \(k\)th contiguous color region. Finally, take \(S = \bigcup_{k=1}^{n/3} S_k\). These regions are depicted graphically in fig. 3

Let \(\text{SD}(\vec{Y}) = \max_{k_1, \ldots, i^2/9} \left| \sum_{j=1}^{I} Y_k \right|\), i.e., \(\text{SD}(\vec{Y}) = \max_{k_1, \ldots, i^2/9} \Delta_\diamond(X_1, X_k)\), for \(\Delta_\diamond\) the shortest-path
distance on the cycle, and observe

\[ \mathbb{P}\left( \text{SD}(\vec{Y}) \geq \frac{i}{3} \right) \leq 2 \mathbb{P}\left( \left| \sum_{j=1}^{i^2/9} Y_j \right| \geq \frac{i}{3} \right) \]

\[ \leq 4 \exp\left( -\frac{2}{\left( \frac{i}{3} \right)^2} \right) \]

\[ = 4 \exp(-2) . \]

From here, we decompose the trace variance and bound it as

\[ \text{trv}(I) = \mathbb{E}_{\vec{Y}} \left( \frac{i^2/9}{2} \sum_{j=1}^{i^2/9} Y_j - \mu \right)^2 \]

\[ \geq \mathbb{P}(X_1 \in S) \mathbb{P}\left( \text{SD}(\vec{Y}) \leq \frac{i}{3} \mid X_1 \in S \right) \mathbb{E}_{\vec{Y}|\text{SD}(\vec{Y})} \left( \frac{i^2/9}{2} \sum_{j=1}^{i^2/9} Y_j - \mu \right)^2 \]

\[ \geq \frac{1}{3} \left( 1 - 4 \exp(-2) \right) \frac{1}{4} > 0.03822 . \]

We thus conclude \( \text{trv}(I) \in \Theta(1) \), therefore via lemma 2.1 \( \text{trv}(\tau_{rel}) \in \Theta(i^2/n^2) \).

**A.2 An unbiased variance estimator**

Recall from definition 2.4 that

\[ \hat{v}(\mathcal{F}, \mathcal{M}) = \frac{1}{2m} \sum_{i=1}^{m} \left( \mathcal{F}(X_{1,i}) - \mathcal{F}(X_{2,i}) \right)^2 , \]

**Proof of lemma 2.4** We first note that by the tensor product chain rule (see Ex. 12.6 of [LP17]), the spectral gap \((1-\lambda)\) of \( \mathcal{M} \otimes \mathcal{M} \) equals that of \( \mathcal{M} \). We now define the function \( g : S \times S \mapsto [0, \frac{1}{2} R^2] \) as \( g(x_1, x_2) = \frac{1}{2} (x_1 - x_2)^2 \).

We first show that \( g \) is an unbiased estimator of the variance of \( f \), i.e., \( \mathbb{E}_{\mathcal{M} \times \mathcal{M}}[g] = \text{Var}_\pi[\mathcal{F}] \):

\[ \mathbb{E}[g] = \mathbb{E}\left[ \frac{1}{2} (X_1 - X_2)^2 \right] \]

\[ = \frac{1}{2} \left( \mathbb{E}[X_1^2] + \mathbb{E}[X_2^2] - 2 \mathbb{E}[X_1 X_2] \right) \]

\[ = \mathbb{E}[X_1^2] - (\mathbb{E}[X_1])^2 \]

\[ = \text{Var}[X_1] = \text{Var}_\pi[\mathcal{F}] \]
Note that \( \hat{v}(X_1, X_2) = \frac{1}{m} \sum_{i=1}^{m} (g(X_{1,i}, X_{2,i})) \), thus it is immediate that \( \mathbb{E}[\hat{v}(X_1, X_2)] = v_\pi \).

We now seek to apply the Bernstein bound to \( g \) on \( \mathcal{M} \otimes \mathcal{M} \). Note that the range of \( g \) is \([0, \frac{1}{2} R^2]\). We require also a bound on the variance of the variance \( \mathbb{V}[g] \). We break the infinite regress here by noting that

\[
\mathbb{V}[g] = \mathbb{E}[g^2] - (\mathbb{E}[g])^2 \\
\leq \mathbb{E}[g \cdot g] \\
\leq \mathbb{E}[g^2 R^2] \\
= \frac{1}{2} R^2 \mathbb{E}[g] \\
= \frac{1}{2} R^2 \mathbb{V}[\mathcal{F}]
\]

Note that this is effectively the argument of Bernstein’s inequality that removes higher moments (beyond the second) from the exponential-sum decomposition in the Chernoff-MGF bound.

Applying the Bernstein inequality (thm. \([B.2]\)) then yields

\[
P(\mathbb{V}[\mathcal{F}] - \hat{v} \geq \varepsilon) \leq e^{\varepsilon^2/(2 \mathbb{V}[\mathcal{F}])}
\]

for

\[
\varepsilon \leq \frac{10 R^2 \ln \frac{1}{\delta}}{2(1 - \lambda)m} + \sqrt{\frac{2(1 + \lambda) \mathbb{V}[g] \ln \frac{1}{\delta}}{(1 - \lambda)m}} \quad \text{Range}(g) = \frac{1}{2} R^2 \\
\leq \frac{5 R^2 \ln \frac{1}{\delta}}{(1 - \lambda)m} + \sqrt{\frac{(1 + \lambda) R^2 \mathbb{V}[\mathcal{F}] \ln \frac{1}{\delta}}{(1 - \lambda)m}} \quad \mathbb{V}[g] \leq \frac{1}{2} R^2 \mathbb{V}[\mathcal{F}]
\]

We now seek a form that depends only on the empirical variance, which we derive via the quadratic formula.

\[
\varepsilon \leq \frac{(11 + \lambda) R^2 \ln \frac{1}{\delta}}{2(1 - \lambda)m} + \sqrt{\left( \frac{(1 + \lambda) R^2 \ln \frac{1}{\delta}}{2(1 - \lambda)m} \right)^2 + \frac{5 R^2 \ln \frac{1}{\delta}}{(1 - \lambda)m} \cdot \frac{(1 + \lambda) R^2 \ln \frac{1}{\delta}}{(1 - \lambda)m} + \frac{(1 + \lambda) R^2 \hat{v} \ln \frac{1}{\delta}}{(1 - \lambda)m}} \quad \text{QUADRATIC FORMULA}
\]

\[
= \frac{(11 + \lambda) R^2 \ln \frac{1}{\delta}}{2(1 - \lambda)m} + \sqrt{\left( \frac{(1 + \lambda) R^2 \ln \frac{1}{\delta}}{2(1 - \lambda)m} \right)^2 + \left( \frac{5(1 + \lambda) R^2 \ln \frac{1}{\delta}}{(1 - \lambda)m} \right)^2 + \frac{(1 + \lambda) R^2 \hat{v} \ln \frac{1}{\delta}}{(1 - \lambda)m}} \quad \text{ALGEBRA}
\]

\[
= \frac{(11 + \lambda) R^2 \ln \frac{1}{\delta}}{2(1 - \lambda)m} + \sqrt{\left( \frac{(2\sqrt{5})^2 + (1 + \lambda)}{2(1 - \lambda)m} \right)^2 + \frac{(1 + \lambda) R^2 \hat{v} \ln \frac{1}{\delta}}{(1 - \lambda)m}} \quad ((2\sqrt{5})^2 + (1 + \lambda)) = 21 + \lambda
\]

\[
\leq \frac{(11 + \lambda) R^2 \ln \frac{1}{\delta}}{2(1 - \lambda)m} + \sqrt{\left( \frac{(\sqrt{21} + \frac{11}{\sqrt{21}} \lambda) R^2 \ln \frac{1}{\delta}}{2(1 - \lambda)m} \right)^2 + \frac{(1 + \lambda) R^2 \hat{v} \ln \frac{1}{\delta}}{(1 - \lambda)m}} \quad \sqrt{21 + \lambda \sqrt{1 + \lambda}} \leq \sqrt{21 + \frac{11}{\sqrt{21}} \lambda}
\]

\[
\leq \frac{(11 + \sqrt{21}) (1 + \frac{4}{21} \lambda) R^2 \ln \frac{1}{\delta}}{2(1 - \lambda)m} + \sqrt{\frac{(1 + \lambda) R^2 \hat{v} \ln \frac{1}{\delta}}{(1 - \lambda)m}} \quad \text{ALGEBRA}
\]

\[
= \frac{((11 + \sqrt{21})(1 + \frac{4}{21} \lambda)) R^2 \ln \frac{1}{\delta}}{2(1 - \lambda)m} + \sqrt{\frac{(1 + \lambda) R^2 \hat{v} \ln \frac{1}{\delta}}{(1 - \lambda)m}}
\]
From which we derive
\[
P \left( |V[F] - \hat{v}| \geq \frac{(11 + \sqrt{21})(1 + \frac{1}{\sqrt{m}})R^2 \ln \frac{1}{\delta}}{(1 - \lambda)m} + \sqrt{\frac{(1 + \lambda)R^2 \ln \frac{1}{\delta}}{(1 - \lambda)m}} \right) \leq \delta .
\]

\section*{A.3 Analysis of DynaMITE}

\subsection*{Prelude to a Proof}

We first note that correctness of DynaMITE follows from that of MmcPro, as they simply apply MmcPro to $f_{\text{avg}}$ and $M(T)$, the trace chain.

Our proof is divided to the following parts: proving trace chain properties, trace variance estimation, and progressive sampling.

\textit{The Trace Chain.} In appx. \ref{appx:A.3.1} we first obtain the stationary distribution of the trace chain and then we bound its mixing and relaxation time in term of the original chain $M$.

\textit{Estimating the Inter-Trace Variance.} In order to estimate the trace variance we use $\hat{v}$ whose correctness is proved in appx. \ref{appx:A.2}. We employ this estimator to $M(T)$, which we proved has constant relaxation time for $T \geq T_{\text{mix}}$. The termination condition which is used in the progressive sampling is based on the variance estimation and using the Bernstein bound (thm. \ref{thm:B.2}) to ensure its accuracy.

\textit{Sketch of the Sampling Schedule.} The key to MmcPro is an \textit{a priori} fixed sampling schedule, which determines the sizes of progressively larger samples. To ensure correctness, a sequence of tail bounds must all hold simultaneously w.h.p. by union bound. The main difficulty is that the schedule length $I$ and the probability concentration bounds are codependent. A shorter schedule is more statistically efficient, but can overshoot the sufficient sample size, and the opposite holds for longer schedules. We explain how to resolve this cyclic dependence in the next paragraph.

Over a run of DynaMITE, we take (up to) $3I$ probability concentration bounds ($I$ bounds for variance, line \ref{line:13} and $I$ bounds each for upper and lower mean bounds, line \ref{line:14}). We first establish the \textit{worst-case} Hoeffding sample complexity $m_I$ (thm. \ref{thm:B.1}) and \textit{best-case} Bernstein sample complexity $\alpha$ (thm. \ref{thm:B.2}) by taking $v = 0$.

\[
m_I \doteq m_H(\Lambda, R, \varepsilon, \frac{2\delta}{m_I}) = \frac{(1 + \Lambda)R^2 \ln \frac{3I}{\delta}}{2(1 - \Lambda)\varepsilon^2}, \quad \& \quad \alpha \doteq \frac{(1 + \Lambda)R \ln \frac{3I}{\delta}}{(1 - \Lambda)\varepsilon} \approx m_B(\Lambda, R, 0, \varepsilon, \frac{2\delta}{m_I}) .
\]

We know from these bounds that once a sample of size $m_I$ is drawn, the desired guarantee has been met, and similarly, before a sample of size $\approx \alpha$ is drawn, the desired Bernstein bound \textit{can not} be met.

We now select the minimal $I$ such that each sample size $m_i$ obeys $m_i \leq 2m_{i-1}$. Observe that in the ratio

\[
\frac{m_I}{\alpha} = \frac{(1 + \Lambda)R}{4 \cdot 5\varepsilon} = \frac{9R}{100\varepsilon} ,
\]

all dependence on $I$ is divided out ($\ln \frac{3I}{\delta}$ terms cancel). We initially run the chain for $m_1$ to be $[2\alpha]$, and in iteration $i$, we run up to $[2\alpha^i]$ steps, thus we conclude $I \doteq \lceil \log_2(\frac{R}{2\varepsilon}) \rceil \geq \lceil \log_2(\frac{R}{2\varepsilon}) \rceil - 1$ (doubling) iterations are sufficient. These computations are repeated verbatim by MmcPro (lines \ref{line:5} & \ref{line:7}) to compute the sampling schedule.

Full details of the schedule and details of putting the above pieces together are presented in thms. \ref{thm:B.1} and \ref{thm:B.2}. 

19
A.3.1 Stationarity, Relaxation, and Mixing Times of the Trace Chain

**Lemma A.3** (Trace Chain Mixing). Having a Markov chain $\mathcal{M}$ with stationary distribution $\pi$ and mixing time $\tau_{\text{mix}}$, the trace chain $\mathcal{M}(T)$’s stationary distribution is $\pi(T)$ where $\pi(T)(\vec{a}) = \pi(a_1) \prod_{i=1}^{k-1} \mathcal{M}(a_i, a_{i+1})$ and its mixing time is bounded as $\tau_{\text{mix}}(\mathcal{M}(T)) \leq 1 + \frac{1}{T} \tau_{\text{mix}}(\mathcal{M})$. In particular for $T \geq \tau_{\text{mix}}(\mathcal{M})$, we have that the second largest eigenvalue of $\mathcal{M}(T)$ is at most $4/5$.

**Proof of Lemma A.3** We first remark that it is not difficult to prove by induction that

$$\left(\mathcal{M}(T)\right)^j(\vec{a}, \vec{b}) = \mathcal{M}^{(j-1)T+1}(a_T, b_1) \prod_{i=1}^{T-1} \mathcal{M}(b_i, b_{i+1}).$$

It is not difficult to see that with the above definitions, for any $\vec{a}$ a trace of length $T$ we have:

$$\pi(T)(\vec{a}) = \sum_{\vec{b} \in \Omega^T} \pi(T)(\vec{a}, \vec{b}) \mathcal{M}(T)(\vec{a}, \vec{b}).$$

We now show that $\mathcal{M}(T)$ is close to $\pi(T)$ after $(\tau(\epsilon)/T) + 1$ steps. Let $X_0, X_1, X_2, \ldots$ be the trace of the original random walk, thus each $X_t$ is a random variable having distribution $X_t \sim \mathcal{M}^t \mu$. We partition this trace into blocks of length $T$ as follows: $B^j = (X_j, X_{j+1}, \ldots X_{j+T-1})$. Thus the trace of $\mathcal{M}(T)$ will be $B^0, B^1, B^2, \ldots$.

Since the mixing time of $\mathcal{M}$ is $\tau_{\text{mix}}$, we have: for any starting distribution $\nu$ over $\Omega$, and $\tau \geq \tau_{\text{mix}} \log(\epsilon)$, TVD($\mathcal{M}^\tau(\nu), \pi$) $\leq \epsilon$. Assume $\mathcal{M}(T)$ has started at initial distribution $\nu'$, looking at the distribution of $B_{\tau_{\text{rel}}+1}$ we will have:

$$\text{TVD}(B_{\tau_{\text{rel}}+1}, \pi(T)) = \frac{1}{2} \sum_{a \in \Omega^T} \left| \left(\mathcal{M}(T)\right)^{\tau_{\text{rel}}+1}\nu'(a) - \pi(T)(a)\right|$$

$$= \frac{1}{2} \sum_{a \in \Omega^T} \left| (\mathcal{M}^T)(\nu', a_1) \prod_{i=1}^{T-1} \mathcal{M}(a_i, a_{i+1}) - \pi(T)(a) \right|$$

$$= \frac{1}{2} \sum_{a \in \Omega^T} \left| (\mathcal{M}^T)(\nu', a_1) \prod_{i=1}^{T-1} \mathcal{M}(a_i, a_{i+1}) - \pi(a) \prod_{i=1}^{T-1} \mathcal{M}(a_i, a_{i+1}) \right|$$

$$\leq \epsilon \prod_{i=1}^{T-1} \mathcal{M}(a_i, a_{i+1}) \leq \epsilon.$$

The bound on $\lambda$ then follows since $(\tau_{\text{rel}}(\mathcal{M}) - 1) \ln(2) \leq \tau_{\text{mix}}(\mathcal{M}) \leq \tau_{\text{rel}}(\mathcal{M}) \ln\left(\frac{2}{\sqrt{\text{mix}}(\mathcal{M})}\right).$ □

**Lemma A.4** (Relaxation Times of Trace Chains). Suppose a reversible chain $\mathcal{M}$. Then for any $T \in \mathbb{N}$, we have

$$\lambda(\mathcal{M}(T)) = \lambda(\mathcal{M}) = \lambda^T(\mathcal{M}) \Leftrightarrow \tau_{\text{rel}}(\mathcal{M}(T)) = \tau_{\text{rel}}(\mathcal{M})^T.$$

**Proof.** We show this result in three movements. We first introduce the $T$th-root $\sqrt{T}\mathcal{M}(T)$ of the trace chain $\mathcal{M}(T)$. We then relate eigenvalues of $\sqrt{T}\mathcal{M}(T)$ to eigenvalues of $\mathcal{M}(T)$, which illuminates the proof structure. Finally, we show that the second eigenvalue of $\sqrt{T}\mathcal{M}(T)$ matches that of $\mathcal{M}$.

**First Movement:** We introduce the $T$th-root of the $T$-trace chain, i.e., $\sqrt{T}\mathcal{M}(T)$, which corresponds to 1-step overlapping windows of length $T$ on the chain $\mathcal{M}$ (as opposed to the nonoverlapping
windows of \( \mathcal{M}(T) \). In particular, \( \sqrt{T} \mathcal{M}(T) \) has transition matrix

\[
\sqrt{T} \mathcal{M}(T)(\vec{a}, \vec{b}) = \begin{cases} 
\vec{a}_{2:T} = \vec{b}_{1:T-1} & : \mathcal{M}(a_T, b_T) \\
\text{otherwise} & : 0
\end{cases}.
\]

Since \( b_{T-1} = a_T \) for all nonzero transition probabilities, we may alternatively express these transition probabilities as

\[
\sqrt{T} \mathcal{M}(T)(\vec{a}, \vec{b}) = \mathbb{1}_{\vec{a}_{2:T}}(\vec{b}_{1:T-1}) \mathcal{M}(b_{T-1}, b_T) = \mathbb{1}_{\vec{a}_{2:T}}(\vec{b}_{1:T-1}) \mathcal{M}(a_T, b_T)
\]

Here transitions between incompatible \( \vec{a} \) and \( \vec{b} \), i.e., those that don’t overlap in a \( T-1 \) window, thus \( \vec{a}_{2:T} \neq \vec{b}_{1:T-1} \), can not occur. Therefore, the transition matrix is always extremely sparse; only \(|S|\) of the total \(|S|^T|\) possible states are ever 1-step reachable.

Note that, in general, matrix roots may be non-unique, and stochastic matrices may not have roots that are also stochastic matrices. However, the \( T^{th} \)-root of \( \mathcal{M}(T) \) described above always exists and is convenient for the analysis, though we do not claim it is unique.

**Second Movement:** We now relate the eigenvalues of \( \sqrt{T} \mathcal{M}(T) \) and \( \mathcal{M}(T) \).

First, note that both \( \sqrt{T} \mathcal{M}(T) \) and \( \mathcal{M}(T) \) are chains over state space \( S^T \). Now, note that from the transition matrix, it is clear that \((\sqrt{T} \mathcal{M}(T))^T = \mathcal{M}(T)\), and thus \( \sqrt{T} \mathcal{M}(T) \) has the same stationary distribution as \( \mathcal{M}(T) \), i.e., \( \pi(\sqrt{T} \mathcal{M}(T)) = \pi(\mathcal{M}(T)) = \pi(T) \).

We now observe that \( \lambda(\mathcal{M}(T)) = \lambda(T(\sqrt{T} \mathcal{M}(T))) \), as is the case with any chain derived via transition-matrix powering (since powering is repeated multiplication, it preserves eigenvectors, and powers eigenvalues). It thus suffices to show that \( \lambda(\sqrt{T} \mathcal{M}(T)) = \lambda(\mathcal{M}) \), as this immediately implies \( \lambda(\mathcal{M}(T)) = \lambda(T(\mathcal{M})) = \lambda(T) \).

**Third Movement:** Here we show that \( \lambda(\sqrt{T} \mathcal{M}(T)) = \lambda(\mathcal{M}) \). This movement is rather more subtle than the prior two, and it is shown via direct computation of the second absolute eigenvalue.

In particular, we decompose the second absolute eigenvalue of \( \sqrt{T} \mathcal{M}(T) \), i.e., \( \lambda(\sqrt{T} \mathcal{M}(T)) \), into a supremum over vector-matrix-vector products (the quadratic form eigenvalue characterization). As we need only consider the second eigenvector, and the first is always the stationary distribution, we consider in the supremum only the unit eigenvectors that are orthogonal to \( \pi(T) \), written \( \vec{x} \perp \pi(T) \).

Note that here we take \( \vec{x} \in \mathbb{R}^{|S|^T} \) to denote a vector over states in \( S^T \), and \( s \in S^T \) denotes a single state in the state space of \( \sqrt{T} \mathcal{M}(T) \) (i.e., \( T \)-traces over \( S \)), which is used to index into \( x \). Likewise, w.r.t. \( \mathcal{M} \) and \( S \), we use \( \vec{x} \in \mathbb{R}^{|S|} \) and \( s \in S \). Observe now that

\[
\lambda'(\sqrt{T} \mathcal{M}(T)) = \sup_{\vec{x} \perp \pi(T):\|\vec{x}\|_2 = 1} \vec{x}^T \sqrt{T} \mathcal{M}(T) \vec{x}
\]

**Eigenvalue Characterization**

| Matrix Multiplication |
|------------------------|
| \( \mathcal{M}(T) \)   |
| \( x_s \mathcal{M}(T)(s, s') \) |
| Dot Product            |
| \( \sum x_s \mathcal{M}(s', s) \) |
| Summation Consolidation |
| \( \sum x_s \mathcal{M}(s', s) \) |

\[= \lambda(\mathcal{M}).\]
A.3.2 Correctness and efficiency proof

Theorem 3.1 (Correctness of DYNAMITE). Consider a Markov chain \( \mathcal{M} \) and its relaxation time upper bound \( T_{\text{rel}} \), function \( f \) and its range \( R \). For arbitrary \( \varepsilon \) and \( \delta \), and starting points \( x_0, x_1 \in S \) taking \( \hat{\mu} \) as either of

1. \( \hat{\mu} \leftarrow \text{McmcPro}((x_0, x_1), \mathcal{M}, T_{\text{rel}}, f, \varepsilon, \delta) \); or
2. \( \hat{\mu} \leftarrow \text{DynaMITE}((x_0, x_1), \mathcal{M}, T_{\text{rel}}, f, \varepsilon, \delta) \) for lazy \( \mathcal{M} \).

we will have a \((\varepsilon, \delta)\) estimator for \( \mu \).

Proof. We first show (1), i.e., correctness of McmcPro, and then show that correctness of Warm-StartDynaMITE easily follows from that of DynaMITE, which follows from that of McmcPro.

We now show claim (1). First note that in initialization (independent of any sampling), McmcPro computes iteration count \( I \) and initial sample size \( \alpha \) (line 5), which determine the schedule of sample sizes and probabilistic bounds. Over the course of the algorithm, at each of \( I \) timesteps, a 1-tail (upper) bound on variance \( u_i \) is computed (line 13), and a 2-tail bound on mean \( \mu_i \) is computed (line 14), for a total of 3 tail bounds. Each tail is bounded with probability \( 1 - \frac{\delta}{3I} \), thus by union bound, all hold simultaneously with probability \( 1 - \delta \). We assume henceforth that all tail bounds hold, thus all conclusions are thus qualified as holding with probability \( \geq 1 - \delta \).

In case 1, we have termination at \( i < I \), i.e., \( \hat{\varepsilon}_i \leq \varepsilon \). As assumed above, all tail bounds at each iteration hold by union bound. Thus for the sample drawn at iteration \( i \), in particular the variance bounds (line 13) hold, as \( \hat{\varepsilon}_i \) is an unbiased estimate of \( \text{trv}^{(T)} \), and by lemma 2.4 w.h.p., \( \text{trv}^{(T)} \leq u_i \). Similarly, the mean bounds (14) hold via thm. B.2 (noting that, by averaging over a pair of independent chains, the variance proxy of interest is \( \frac{1}{2} \text{trv}^{(T)} \)). Note that while both tail bounds are taken over the tensor-product chain \( \mathcal{M} \otimes \mathcal{M} \), it holds that \( \lambda(\mathcal{M} \otimes \mathcal{M}) = \lambda(\mathcal{M}) \), so the bound remains valid.

Consequently, when it holds that \( \hat{\varepsilon}_i \leq \varepsilon \), the algorithm returns the estimate \( \hat{\mu}_i \) (line 16), which by the above is sufficiently accurate to satisfy the stated guarantees.

We now consider case 2, wherein we have termination at step \( i = I \). When this occurs, it holds that

\[
m_i = m_I = \left\lceil \alpha 2^I \right\rceil \geq m_H(\lambda, R, \varepsilon, \frac{\delta}{3I}) ,
\]

and thus by Hoeffding’s inequality for mixing processes (thm. 3.1), we have \( |\mu - \hat{\mu}_I| \leq \varepsilon \), (i.e., the schedule was selected exactly to ensure \( m_I \) samples would be sufficient, regardless of early termination and variance.

We now proceed to show claim (2). To see this result, note that

\[
\lambda(\mathcal{M}^{(T)}) \leq \lambda^T(\mathcal{M}) ,
\]

and thus claim (2) follows directly from claim (1).

Finally, note that claim (3) follows from claim (2), paired with Eq. 7. In particular, note that the uniform mixing time is selected such that we have

\[
\forall \omega : \left\| \frac{\partial \mathcal{M}^{\text{unif}}(\omega)}{\partial \pi} \right\|_{\pi, \infty} \leq 2 ,
\]
i.e., \( M \) is uniformly mixed. This further implies that the trace chain is uniformly mixed, as both \( \pi_{\text{min}} \) and \( \Lambda \) are exponential in the trace-length \( T \), which cancels out in the uniform-mixing bound (see Eq. 12.13 of \[LP17\]). However, we still apply nonstationarity correction \( \frac{\delta}{4} \) instead of \( \frac{\delta}{2} \), to account for the fact that the tensor product chain \( (M \otimes M) \) uniformly mixes slightly slower than \( M \), even though it relaxes and mixes as quickly as \( M \).

We now present and prove an extended statement of thm. \[3.2\] which provides finite-sample and asymptotic sample complexity bounds to \textsc{McmcPro}, \textsc{DynaMITE}.

**Theorem A.5** (Efficiency of \textsc{DynaMITE}). Suppose as in thm. \[3.1\] and take \( I = \frac{1}{\ln 2} \left( \frac{R}{\varepsilon} \right) \) and \( T = \left\lceil \frac{1+\Lambda}{1-\Lambda} \ln \sqrt{2} \right\rceil \). Then with probability at least \( 1 - \delta \), each mean-estimation algorithm runs for no more than \( \hat{m} \) steps (individually), where \( \hat{m} \) is

1. for \textsc{McmcPro}:

   \[
   \hat{m} \leq 4T \ln \frac{3I}{\delta} \left( \frac{5(6+\Lambda)R}{2(1-\Lambda)\varepsilon} + \frac{(1+\Lambda)\text{trv}(T)}{(1-\Lambda)\varepsilon^2} \right) \\
   \in \mathcal{O}\left( \frac{1}{1-\Lambda} \log \left( \frac{\log(R/\varepsilon)}{\delta} \right) \left( \frac{R}{\varepsilon} + \frac{v_{\pi}}{\varepsilon^2} \right) \right); \\
   \]

2. for \textsc{DynaMITE}:

   \[
   \hat{m} \leq 2T \ln \frac{3I}{\delta} \left( \frac{65R}{\varepsilon} + \frac{12\text{trv}(T)}{\varepsilon^2} \right) \\
   \in \mathcal{O}\left( T \log \left( \frac{\log(R/\varepsilon)}{\delta} \right) \left( \frac{R}{\varepsilon} + \frac{\text{trv}(T)}{\varepsilon^2} \right) \right) \\
   = \mathcal{O}\left( \log \left( \frac{\log(R/\varepsilon)}{\delta} \right) \left( \frac{R}{(1-\Lambda)\varepsilon} + \frac{\tau_{\text{rel}}\text{trv}(\tau_{\text{rel}})}{\varepsilon^2} \right) \right); \&
   \]

3. Adding also warm start complexity (line 4):

   \[
   \hat{m} \leq 2 \left[ \frac{\ln \pi_{\text{min}}}{\ln \frac{1}{\Lambda}} \right] + 2T \ln \frac{12I}{\delta} \left( \frac{65R}{\varepsilon} + \frac{12\text{trv}(T)}{\varepsilon^2} \right) \\
   \in \mathcal{O}\left( \frac{\ln \pi_{\text{min}}}{\ln \frac{1}{\Lambda}} + T \log \left( \frac{\log(R/\varepsilon)}{\delta} \right) \left( \frac{R}{\varepsilon} + \frac{\text{trv}(T)}{\varepsilon^2} \right) \right) \\
   = \mathcal{O}\left( \frac{\ln \pi_{\text{min}}}{\ln \frac{1}{\Lambda}} + \log \left( \frac{\log(R/\varepsilon)}{\delta} \right) \left( \frac{R}{(1-\Lambda)\varepsilon} + \frac{\tau_{\text{rel}}\text{trv}(\tau_{\text{rel}})}{\varepsilon^2} \right) \right). 
   \]
Proof. The strategy here is to derive a sample size $m'$, dependent on $R, \text{trv}^T, \varepsilon, \delta$, s.t. w.h.p., each algorithm will terminate after drawing a sample of at least $m'$ traces. We then bound the total number of samples drawn over the course of this process, and make some substitutions to derive the result. For brevity, throughout this result we take $\eta \equiv \ln \frac{M}{\delta}$.

We show the result for McmcPro, using second absolute eigenvalue bound $\lambda$, as it immediately implies the corresponding results for DynaMITE and WarmStartDynaMITE.

We first show that, with high probability, the empirical variance is not much larger than the true variance, and thus with high probability, the variance-bounds used by McmcPro are not loose. Let

$$
\varepsilon_{v,1} \equiv \frac{5R^2\eta}{(1-\Lambda)m}, \quad \varepsilon_{v,2} \equiv \sqrt{(1+\Lambda)R^2\varepsilon_v\eta \over (1-\Lambda)m}.
$$

Now, note that by lemma 2.3[4] we have for any sample size $m$ that

$$
P\left(\bar{v}_t \geq v_\pi + \varepsilon_{v,1} + \varepsilon_{v,2}\right) \leq \frac{\delta}{3I}.
$$

We now consider the first iteration $i$ such that $m_i \geq m'$, letting $m \equiv m_i$ and $\hat{v} \equiv \hat{v}_t$. On line[13] of McmcPro, we have (w.h.p.)

$$
\begin{align*}
\mathbf{u}_i &\leq \hat{v} + \frac{(11 + \sqrt{21} + (1 + \frac{11}{21})\Lambda)R^2\eta}{(1-\Lambda)m} + \sqrt{\frac{(1+\Lambda)R^2\hat{v}\eta}{(1-\Lambda)m}} \\
&\leq v_\pi + \varepsilon_{v,1} + \varepsilon_{v,2} + \sqrt{\frac{(1+\Lambda)R^2(v_\pi + \varepsilon_{v,1} + \varepsilon_{v,2})\eta}{(1-\Lambda)m}} \\
&= v_\pi + \varepsilon_{v,2} + \sqrt{\frac{(1+\Lambda)R^2v_\pi\eta}{(1-\Lambda)m}} \\
&= v_\pi + \varepsilon_{v,2} + \sqrt{\frac{(1+\Lambda)R^2v_\pi\eta}{(1-\Lambda)m}} \\
&< v_\pi + \varepsilon_{v,2} + \sqrt{\frac{(1+\Lambda)R^2v_\pi\eta}{(1-\Lambda)m}} \quad \sqrt{a+b+2\sqrt{ab}} = \sqrt{a+b} \\
&\leq v_\pi + \varepsilon_{v,2} + 2\sqrt{\frac{(1+\Lambda)R^2v_\pi\eta}{(1-\Lambda)m}} \\
&= v_\pi + \sqrt{\frac{(1+\Lambda)R^2v_\pi\eta}{(1-\Lambda)m}} \equiv \varepsilon_{v,3}\end{align*}
$$

Note that here we take a lower-tail bound, rather than an upper-tail bound; the constants are identical and the result similarly follows from the Bernstein inequality.
Substitution into the Bernstein bound (line 14), and similar algebra, gives us

\[
\hat{\epsilon}_i = \frac{10R\eta}{(1-\Lambda)m_i} + \sqrt{\frac{(1+\Lambda)u_i\eta}{(1-\Lambda)m_i}} \leq \frac{10R\eta}{(1-\Lambda)m_i} + \sqrt{\frac{(1+\Lambda)(v_\pi + \epsilon_{v,3} + \epsilon_{v,4})\eta}{(1-\Lambda)m_i}} \leq \frac{5(6+\Lambda)R\eta}{2(1-\Lambda)m_i} + \sqrt{\frac{(1+\Lambda)v_\pi\eta}{(1-\Lambda)m_i}} \leq \frac{5(6+\Lambda)R\eta}{2(1-\Lambda)m_i} + \sqrt{\frac{(1+\Lambda)v_\pi\eta}{(1-\Lambda)m_i}} \leq \frac{5(6+\Lambda)R\eta}{2(1-\Lambda)m_i} + \sqrt{\frac{(1+\Lambda)(v_\pi + \epsilon_{v,3} + \epsilon_{v,4})\eta}{(1-\Lambda)m_i}} \leq \frac{5(6+\Lambda)R\eta}{2(1-\Lambda)m_i} + \sqrt{\frac{(1+\Lambda)v_\pi\eta}{(1-\Lambda)m_i}}.
\]

We terminate when \(\hat{\epsilon}_i \leq \varepsilon\), thus this implies sufficient sample size

\[
m' \leq +\eta \left( \frac{5(6+\Lambda)R}{2(1-\Lambda)\varepsilon} + \frac{(1+\Lambda)v_\pi}{(1-\Lambda)\varepsilon^2} \right).
\]

Now, due to the doubling geometric grid, we must have \(m_i \in [m', 2m']\), thus we have

\[
m_i \leq 2\eta \left( \frac{5(6+\Lambda)R}{2(1-\Lambda)\varepsilon} + \frac{(1+\Lambda)v_\pi}{(1-\Lambda)\varepsilon^2} \right).
\]

Now, each step of the tensor-product chain \((\mathcal{M} \otimes \mathcal{M})\) requires two steps of \(\mathcal{M}\), so we conclude (1) by noting that 4\(m'\) samples suffice.

Now, to get (2), note that in DYNAMITE, we take \(T = \left\lceil \frac{1+\Lambda}{1-\Lambda} \ln \sqrt{2} \right\rceil\), and thus \(\lambda(\mathcal{M}^{(T)}) \leq \lambda^T \leq \frac{1}{2}\). The finite-sample bound then follows from (1) as DYNAMITE simply calls MCMCPro (see line 25) multiplying total sample complexity by \(T\), as each step in \((\mathcal{M}^{(T)} \otimes \mathcal{M}^{(T)})\) (i.e., the tensor-product trace-chain) takes \(T\) steps in \(\mathcal{M}\) for every step in the \(\mathcal{M} \otimes \mathcal{M}\) chain of MCMCPro. Finally, applying lemma 2.1 yields the result.

\[\square\]

A.3.3 A Note on Nonstationarity

Note that theorems B.1 and B.2 assume stationarity i.e. they consider traces \(\bar{X}_{1:m} : X_1, X_2, \ldots X_m\) assuming \(X_1 \sim \pi\). This assumption is often prohibitive, as drawing even a single such sample can be NP-hard. We overcome this problem using the following equation for \(\bar{X}_{1:m} : X_1, X_2, \ldots X_m, X_1 \sim \nu\).

\[
\mathbb{P}_{X_{1\sim\nu}} (|\hat{\mu} - \mu| \geq \varepsilon) \leq \left\| \frac{\partial \nu}{\partial \pi} \right\|_{\pi,\infty} \mathbb{P}_{\bar{X}} (|\hat{\mu} - \mu| \geq \varepsilon) ; \tag{7}
\]

see, e.g., [FJS18], proof of thm 2.3. Note that by eq. 7 it is sufficient to have \(\left\| \frac{\partial \nu}{\partial \pi} \right\|_{\pi,\infty} = \text{ess sup}_{\omega \in \Omega} \frac{\nu(\omega)}{\pi(\omega)} \in \mathcal{O}(1)\). This can generally be accomplished straightforwardly with a warm-start by selecting an arbitrary fixed \(\omega \in \Omega\), taking \(\nu\) to be the distribution reached after running \(\mathcal{M}\) for \(\tau_{\text{rel}}(\mathcal{M}) \ln (1/\pi_{\text{min}})\) steps (see any standard MCMC text book, e.g. [LP17]). With this in mind, for simplicity we assume stationarity, knowing that our proofs and algorithm generalize with trivial modifications.
A.4 Missing proofs from [3.1]

The following equations compare and contrast Rabinovich et al.’s bound with the central limit theorem and this work:

\[
\frac{T_{\text{mix}}(f, M, \frac{R}{\varepsilon})}{\varepsilon^2} R^2 \log \frac{1}{\delta} \geq \frac{T_{\text{mix}}(f, M)}{\tau_{\text{mix}}(f, M)} \cdot \frac{R^2}{v_\pi} \cdot \frac{v_{\text{asy}}(f, M) \log(\frac{1}{\delta}) \log(\frac{R}{\varepsilon})}{l} \tag{8}
\]

From which we can conclude

\[
\text{HIPERM}_{\text{Rabi}}(\mathcal{M}, f) \in \Omega \left( \frac{T_{\text{mix}}(f)}{\tau_{\text{mix}}(f)} \cdot \frac{R^2}{v_\pi} \cdot \frac{v_{\text{asy}}(f, M) \log(\frac{R}{\varepsilon})}{l} \right),
\]

which implies

\[
\text{HIPERM}_{\text{Rabi}}(\mathcal{M}, f) \in \Omega \left( v_{\text{asy}}(f, M) \log(\frac{R}{\varepsilon}) \right).
\]

Note that the gap between our bound and the central limit theorem is \(\Theta(1 + \log \log(\frac{R}{\varepsilon}))\) which is exponentially smaller compare to the \(\log(\frac{R}{\varepsilon})\) appearing here.

Proof of eq. (8).

\[
T_{\text{mix}}(f, M, \frac{R}{\varepsilon}) R^2 = \frac{T_{\text{mix}}(f, M, \frac{R}{\varepsilon})}{\tau_{\text{mix}}(f, M, \frac{R}{\varepsilon})} \cdot \frac{R^2}{v_\pi} \cdot \tau_{\text{mix}}(f, M, \frac{R}{\varepsilon}) v_\pi \geq \frac{T_{\text{mix}}(f, M)}{\tau_{\text{mix}}(f, M)} \cdot \frac{R^2}{v_\pi} \cdot \tau_{\text{rel}}(f, M) v_\pi \log(\frac{R}{\varepsilon}) \geq \frac{T_{\text{mix}}(f, M)}{\tau_{\text{mix}}(f, M)} \cdot \frac{R^2}{v_\pi} \cdot \frac{v_{\text{asy}} \log(\frac{R}{\varepsilon})}{l} \tag{9c} \text{ [RRJW20]}
\]

\[\text{lemma 2.1}: v_{\text{asy}} \leq 2\tau_{\text{rel}}(f) v_\pi(f)\]

A.5 Application to counting k colorings

Consider a graph \(G = (V, E)\), and some number of colors \(k\). A coloring of \(G\) is a mapping \(\gamma: V \rightarrow \{1, 2, \ldots k\}\), where \(\gamma(v)\) denotes the color of \(v \in V\), and a proper coloring is any coloring \(\gamma\) s.t. \(\forall u, v \in V: \gamma(u) = \gamma(v) \implies (u, v) \notin E\). For a subset \(V' \subseteq V\), by \(\gamma(V')\), we mean the restriction of the mapping \(\gamma\) to domain \(V'\), i.e., a proper coloring on the induced subgraph on \(V'\). Furthermore we define \(\Gamma(G, k)\) to be the set containing all the proper \(k\)-colorings of \(G\). We denote the size of a set \(X\) by \(#X\), and the uniform distribution on it by \(U(X)\). For example \(#\Gamma(G, k)\) is the number of proper \(k\)-colorings of \(G\), and \(U(\Gamma(G, k))\) is the uniform distribution on it. For any graph \(G\), the Glauber dynamics\(^4\) chain is defined on \(\Gamma(G, k)\) as follows and it known that it converges to stationary distribution \(U(\Gamma(G))\):

**Definition A.3** (Glauber dynamics chain for proper \(k\)-colorings [Jer95]). We define the Markov chain \(\mathcal{M}_G\) on \(\Gamma(G, k)\) as follows: At each time step \(t\), let \(X_t\) be a proper \(k\)-coloring of \(G\),

1. Pick \(c \in \{1, 2, \ldots k\}\), and \(u \in V\) uniformly at random.
2. If changing \(u\)'s color to \(c\) is still a valid proper coloring, let \(X_{t+1}\) be this new coloring.
3. Else, \(X_{t+1} = X_t\).

\^4Also known as the zero temperature Pott’s model, or the single site update
The planted partition, or stochastic block, model generalizes the Erdős-Rényi model, allowing for communities in graphs, and it has the following distribution:

**Definition A.4** (Planted Partition Model). Given the following parameters: (i) number of vertices \( n \), (ii) a partitioning of \( \{1, 2, \ldots, n\} \) to \( r \) subsets \( C_1, C_2, \ldots, C_r \), and (iii) an edge placement matrix \( P \in [0, 1]^{r \times r} \). A graph \( G \) with \( \#V = n \) is generated as follows: for any two vertices \( u \in C_i \) and \( v \in C_j \), the edge \( (u, v) \) is in \( E \) with probability \( P_{i,j} \).

A simplified version of the planted partition model is when each \( C_i \) has size \( n/r \), the diagonal elements of \( P \) are all \( p \), and the other elements are \( q/r-1 \). Since the probability of having edges inside a community is often more than having edges between two communities we assume \( q/p \in o(1) \). We denote this model by \( P(n, r, p, q) \). We show bounds for this simplified model, though they may easily be extended to arbitrary planted partitions graphs or similar network models having small clusters.

**Theorem A.6** (Application of DynaMITE to Planted Partitions). Consider \( G \sim P(n, r, p, q) \). Assume \( p \geq qn \), and let \( k > qn^2/r^2 \). With probability \( \geq 1 - o(1) \), Jerrum’s counting algorithm on \( G \), with \( \delta = \epsilon = 1/4 \), equipped with DynaMITE as a mean-estimation gadget, has sample complexity

\[
\hat{Q} \left( n^2 \left( E^2 + \frac{E^3}{r} \right) \right).
\]

**Prelude to a Proof**. The proof is based on developing two new notions: loosely connectedness (definition A.6) and restriction of a chain (definition A.5) to a subset of its vertices. We use these definitions to prove lemma A.8 which then constitutes the proof of thm. A.6. Here we provide a road map and intuition, and the proofs are presented in full detail in appx. A.5.1.

Consider a subset of graph vertices, \( V' \subseteq V \), and let \( G' \) be a graph obtained from \( G \) after removing cutting the edges between \( V' \) and \( V \setminus V' \). When \( U(\Gamma(G, k)) \) is only negligibly different from \( U(\Gamma(G', k)) \) we say \( V' \) is loosely connected to the rest of the graph (see definition A.6). In lemma A.8 we show that when it occurs, \( \text{trv}^{(\text{rel}(M_G))}(M_G, f_e) = o(v_\pi(f_e)) \) where \( f_e \) is intermediate phase showing up in JVV reduction (see appx. B.2) and \( M_G \) is the Glauber dynamic chain. The proof is based on coupling the probability spaces of the two Markov chain’s traces (\( M_G \) and \( M_G' \)). Then we show it is sufficient that a simpler and faster mixing variant of \( M_G' \) mixes (see definition A.5).

Finally, we show that w.h.p., graphs distributed according to the planted partition model, and their subgraphs, have loosely connected parts.

\[\square\]

**A.5.1 Definitions and proof sketches from A.5**

Consider a subset of \( G \)’s vertices \( V' \subseteq V \) and assume that \( f : \Gamma(k, G) \to [0, 1] \) is a function whose value only depends on a coloring restricted to some \( V' \), i.e. \( f(\gamma(V)) = f(\gamma(V')) \) for any \( \gamma \). We define \( M'_{G} \) being restriction of \( M_G \) to \( V' \) as follows:

**Definition A.5** (Restricted Glauber dynamic chain). We define Markov chain \( M'_{G} \) as follows: At each time step \( t \), let \( X_t \) be a proper \( k \)-coloring of the induced subgraph of \( V' \), and then

1. Pick \( c \in \{1, 2, \ldots, k\} \), and \( u \in V \) uniformly at random.
2. If \( u \in V_i \) then follow \( M_G \) to get to \( X_{t+1} \) (Definition A.3).
3. Else, \( X_{t+1} = X_t \).
If there is no edge between \( V' \) and \( V \setminus V' \), then the image of \( f(\gamma) \) under traces of \( \mathcal{M}_G \) and restricted chain \( \mathcal{M}'_G \) will have the same distribution. From this, lemma \ref{lem:trv_bound} and \( v_r \leq 1/4 \), we bound the length-\( T \) trace variance as 

\[
\text{trv}^{(\tau_{\text{rel}})}(\mathcal{M}) \leq (1/4)(\tau_{\text{rel}}(\mathcal{M}'))/2\tau_{\text{rel}}(\mathcal{M}) = \tau_{\text{rel}}(\mathcal{M}')/2\tau_{\text{rel}}(\mathcal{M}) .
\]

By defining loosely connectedness our goal is to consider the case some when the cut between \( V' \) and \( V \) does not make the image of \( f(\gamma) \) in traces of \( \mathcal{M}' \) and under \( \mathcal{M} \) noticeably different, i.e., we can couple the images of \( f \) on a trace of \( \mathcal{M} \) and on a trace of \( \mathcal{M}' \) so that the coupled states are with high probability identical. Thus, we can largely ignore the connecting edges between \( V' \) and \( V \setminus V' \).

**Definition A.6** (Connectedness parameters \( \zeta \) and looseness parameter \( c \)). For a subset \( V' \) of \( V \), we define the cut set \( B(V') \subseteq E \) as \( B(V') = \{(u,v) \in E : u \in V', v \in V \setminus V'\} \). Let \( G' = (V, E \setminus B(V')) \). We say \( \zeta \) is the connectedness parameter of \( V' \) if

\[
P( \bigvee_{(u,v) \in B(V')} \gamma(u) = \gamma(v) ) \leq \zeta , \quad \text{for } \gamma \sim U(\Gamma(G', k)) .
\]

If there exists \( c > 1 \) such that \( 1/\zeta \geq \tau_{\text{rel}}(\mathcal{M}_G) \log(c) \), we say that \( V' \) is loosely connected to \( \mathcal{M}_G \) and \( c \) is the looseness parameter.

**Remark A.7.** The reader may postulate that with the above definition, verifying whether a subset is loosely connected needs calculating complex probabilities. This is correct, however, note that knowledge of \( \zeta \) or \( c \) is not required by DYNAMITE; it is merely used to characterize performance.

The following lemma whose proof in appx. \ref{app:proofs} shows that loosely connecteness implies small relaxed trace variance.

**Lemma A.8.** Let \( 0 < \zeta < 1, c > 1 \) and \( V' \subseteq V \) and \( G' \) be as defined in definition A.6. Let \( \mathcal{M} \) and \( \mathcal{M}' \) be the Glauber dynamics chains on \( G \) and \( G' \) and \( \tau_{\text{rel}}(\mathcal{M}) \) and \( \tau_{\text{rel}}(\mathcal{M}') \) respectively their relaxation times. Assume \( f : \Gamma(G, k) \rightarrow [0, 1] \) such that for a coloring \( \gamma \), \( f(\gamma) = f(\gamma(V')) \). We have

\[
\text{trv}^{(\tau_{\text{rel}}(\mathcal{M}'))}(\mathcal{M}) \leq \frac{2\tau_{\text{rel}}(\mathcal{M})}{\tau_{\text{rel}}(\mathcal{M}')} + \frac{\log c}{c} .
\]

The last step of the proof will be to show that having \( k > qn^4/r^2 \), in the planted partition model each \( C_i \) is with high probability loosely connected to \( \mathcal{M}_G \). This is proved in the following section.

**A.5.2 Detailed Proofs from A.5**

We now restate the definition of the planted partition graph model (definition A.4), and prove the result related to it.

**Definition A.4** (Planted Partition Model). Given the following parameters: (i) number of vertices \( n \), (ii) a partitioning of \( \{1, 2, \ldots, n\} \) to \( r \) subsets \( C_1, C_2, \ldots, C_r \), and (iii) an edge placement matrix \( P \in [0, 1]^{r \times r} \). A graph \( G \) with \( \#V = n \) is generated as follows: for any two vertices \( u \in C_i \) and \( v \in C_j \), the edge \( (u, v) \) is in \( E \) with probability \( P_{i,j} \).

**Theorem A.6** (Application of DYNAMITE to Planted Partitions). Consider \( G \sim \mathcal{P}(n, r, p, q) \). Assume \( p \geq qn \), and let \( k > qn^4/r^2 \). With probability \( \geq 1 - o(1) \), Jerrum’s counting algorithm on \( G \), with \( \delta = \epsilon = 1/4 \), equipped with DYNAMITE as a mean-estimation gadget, has sample complexity

\[
\tilde{O}\left(n^2 \left((\#E)^2 + \frac{(\#E)^3}{r}\right)\right).
\]
Proof of Theorem A.6. Let $G \sim \mathcal{P}(n, r, p, q)$. We first show that with probability at least $1 - re^{-\frac{2nq^2}{r^2}}$, and taking $\zeta = \frac{2nq^2}{kq^4}$, any partitions $C_j$, is $\zeta$–loosely connected to the rest of the graph $G$.

For any $j$, and $v \in C_j$, let $d_{out}(v)$ be number of edges having one end outside $C_j$. Remember definition of $B(C_j)$ from Definition A.6 we have:

$$B(C_j) = \sum_{v \in C_j} \#d_{out}(v),$$

Taking $B_{\text{max}} = \max_{j \in 1, r} B(C_j)$, note that by symmetry and the fact that for each two end points of an edge in $B(C_j)$ we can recolor one end point to make their colors different, we will have that

$$\mathbb{P}\left( \bigvee_{(u,v) \in B(C_j)} \gamma(u) = \gamma(v) \right) \leq \frac{B_{\text{max}}}{k}.$$

Now, note that $\#C_i = \frac{n}{r}$, and furthermore, for each $v \in C_j$,

$$d_{out}(v) \sim \mathcal{B}(n - \frac{n}{r}, \frac{q}{r - 1}) = \mathcal{B}(\frac{n(r - 1)}{r}, \frac{q}{r - 1}),$$

where $\mathcal{B}(n, p)$ denotes the binomial distribution. Thus by independence, we have

$$B(C_j) \sim \mathcal{B}\left(\frac{n^2(r - 1)}{r^2}, \frac{q}{r - 1}\right), \mathbb{E}[B(C_j)] = \frac{n^2q}{r^2}, \mathbb{V}[B(C_j)] = \frac{n^2q}{r^2}(1 - \frac{q}{r - 1}) < \frac{n^2q}{r^2}.$$

By the Gaussian CLT Chernoff bound, we have approximately

$$\mathbb{P}\left( \frac{n^2q}{r^2} + \sqrt{2\frac{n^2q}{r^2} \ln(\frac{1}{\delta})} \right) \leq \delta.$$

Furthermore, by the union bound, we have approximately

$$\mathbb{P}\left( B_{\text{max}} > \frac{n^2q}{r^2} + \sqrt{2\frac{n^2q}{r^2} \ln(\frac{1}{\delta})} \right) \leq \delta.$$

Take $\delta = \exp\left(\frac{-n^2q \ln(\frac{1}{r})}{2r^2}\right)$, and we get

$$\mathbb{P}\left( B_{\text{max}} > \frac{2n^2q}{r^2} \right) \leq \exp\left(\frac{-n^2q \ln(\frac{1}{r})}{2r^2}\right).$$

Taking $\zeta \leq \frac{2n^2q}{kq^4}$ with probability at least $1 - r \exp\left(\frac{-n^2q}{r^2}\right)$, we have that all $C_j$s are $\zeta$–loosely connected to the rest of the graph.

Thus, by taking $k \geq qn^4/r^2 \cdot \log c$ we will have that $\zeta \leq 1/(n^2 \log c)$.

We now show that with probability at least $1 - e^{-(np)/(8r)}$ in a graph generated from $G \sim \mathcal{P}(n, r, p, q)$ we have: $d_{\text{max}}(G) \geq (np)/(2r)$. This is easily derived from the Chernoff bound and having $\mathbb{E}[d(v)] \geq np/r$, where $d(v)$ is the degree of each vertex.

Having this bound, if $k > \frac{(11/6)np}{2} Vigoda’s result will be applicable to $G \sim \mathcal{P}(n, r, p, q)$ with probability at least $1 - o(1)$. Thus, we can conclude that for $k \geq qn^4/r^2 \cdot \log c$ the relaxation time of $\mathcal{M}$ it bounded by $n^2$. 

29
We now employ Lemma A.8 to $G$ and $C_j$s, we have $\text{trv}(\tau_{rel}) \leq \tau_{rel}(M_j)/2\tau_{rel}(M) + \log(c)/c$. Employing Vigoda’s bound on the relaxation time we know $\tau_{rel}(M) \leq \tau_{mix} \leq n^2$, and $M_j$ is Jerrum’s chain with a slowdown by a factor of $\frac{r}{n}$. Thus, $\tau_{rel}(M_j) \leq \frac{r}{n}(\#C_j)^2 \leq nr$. Thus, we will have: $\text{trv}(\tau_{rel}) \leq 1/r + \frac{\log c}{c} \leq 2/r$.

Note that the above analysis holds for all $G_i$s in intermediate steps. We approximate each $f_i$ with precision $\varepsilon/\#E$, setting $R \leq 1$ the sample complexity of each intermediate step will be:

$$T \left( 36 + 2\ln \left( \frac{3\log(\frac{\#E}{\delta})}{\delta/\#E} \right) \left( \frac{50}{\varepsilon} + \frac{(\frac{252}{3})^{2/3}(\#E)^{4/3}}{\varepsilon^{4/3}} + \frac{18(\#E)^2(\frac{1}{c} + \frac{22}{n})}{\varepsilon^2} \right) \right)$$

Setting $T = n^2$ by Vigoda’s result, letting $\delta, \varepsilon$ be constants, and ignoring vanishing terms, the complexity of Jerrum’s algorithm equipped with DYNAMITE is

$$O \left( n^2 \log(n \log(n)) \left( (\#E)^2 + \frac{(\#E)^3}{r} \right) \right).$$

Having this, the last piece of puzzle is to prove lemma A.8 so we restate and prove it.

**Lemma A.8.** Let $0 < \zeta < 1$, $c > 1$ and $V' \subseteq V$ and $G'$ be as defined in definition A.6. Let $M$ and $M'$ be the Glauber dynamics chains on $G$ and $G'$ and $\tau_{rel}(M)$ and $\tau_{rel}(M')$ respectively their relaxation times. Assume $f : \Gamma(G, k) \rightarrow [0, 1]$ such that for a coloring $\gamma$, $f(\gamma) = f(\gamma(V'))$. We have

$$\text{trv}(\tau_{rel}(M))(M) \leq \frac{2\tau_{rel}(M')}{\tau_{rel}(M)} + \frac{\log c}{c}.$$

**Proof of lemma A.8.** We bound the trace variance of $M$ by relating it to the trace variance of $M'$.

For a positive $c$, assume $1/\zeta \geq \tau_{rel}(M) \log(c)$. let $\tau = 1/\zeta$ and $n_\tau$ be the number of steps in a trace of length $\tau$ in which some edge in $B(V')$ has both end points colored the same. Note that $\mathbb{E}_{U(\Gamma(G, k))}[n_\tau] = \tau \zeta = 1/2$, applying the concentration bound of thm. B.1 we will have

$$\mathbb{P}(n_\tau \geq 1) \leq \mathbb{P}(|n_\tau - 1/2| \geq 1/\sqrt{2}) \leq \exp \left( -2\tau(1/\sqrt{2})^2/\tau_{rel}(M) \right) \leq \exp \left( -\tau/\tau_{rel}(M) \right) \leq \frac{1}{c}.$$

Let $\mathcal{E}(\gamma)$ be the event that $\gamma \sim U(\Gamma(G', k))$ colors none of the two end points of an edge in $B(V')$ to the same color. Let $\tilde{\mu}$ be the expectation of $f(\gamma)$ when $\gamma \sim U(\Gamma(G', k))$. We have $\mu = \mathbb{E}[f(\gamma)|\mathcal{E}(\gamma)]$, $\gamma \sim \Gamma(G', k)$. Furthermore, we can rewrite $\tilde{\mu}$ as $\tilde{\mu} = \mathbb{E}[f(\gamma)|\mathcal{E}(\gamma)]\mathbb{P}(\mathcal{E}(\gamma)) + \mathbb{E}[f(\gamma)|\neg\mathcal{E}(\gamma)](1 - \mathbb{P}(\mathcal{E}(\gamma)))$. Thus, $\mu(1 - \zeta) \leq \tilde{\mu} \leq \mu(1 - \zeta) + \zeta$, hence, $|\mu - \tilde{\mu}| \leq \zeta$.

Furthermore, under the condition that $n_\tau < 1$ the trace in $M$ and the traces in $M'$ are dis-
tributed identically. The following inequalities are thus concluded:

\[
\begin{align*}
\mathbb{E}_{X} \left[ (S_{X} - \mu)^2 \right] &= \mathbb{E}_{X} \left[ (S_{X} - \mu)^2 \bigg| n_{\tau} < 1 \right] \mathbb{P}(n_{\tau} < 1) + \mathbb{E}_{X} \left[ (S_{X} - \mu)^2 \bigg| n_{\tau} \geq 1 \right] (1 - \mathbb{P}(n_{\tau} < 1)) \\
&= \mathbb{E}_{X} \left[ (S_{X} - \tilde{\mu} - (\mu - \tilde{\mu}))^2 \bigg| n_{\tau} < 1 \right] \mathbb{P}(n_{\tau} < 1) + R^2 \cdot (1 - \mathbb{P}(n_{\tau} < 1)) \\
&\leq \left( \mathbb{E}_{X} \left[ (S_{X} - \tilde{\mu})^2 \bigg| n_{\tau} < 1 \right] + \zeta^2 \right) \mathbb{P}(n_{\tau} < 1) + (1 - \mathbb{P}(n_{\tau} < 1)) \\
&\leq \frac{2\tau_{\text{rel}}(M')}{\tau} + 2\zeta \mathbb{P}(n_{\tau} < 1) + (1 - \mathbb{P}(n_{\tau} < 1)).
\end{align*}
\]

Since \( \tau = 1/\zeta \) we have \( (1/\tau)\tau_{\text{rel}}(M') = \zeta \tau_{\text{rel}}(M') \geq \zeta \). Thus for \( \tau = \tau_{\text{rel}}(M) \log c \) we have \( \text{trv}(\tau)(M) \leq \frac{2\tau_{\text{rel}}(M')}{\tau} + \frac{1}{c} \). Using lemma 2.1 we have

\[
\text{trv}(\tau_{\text{rel}}(M))(M) \leq \log(c) \cdot \left( \frac{2\tau_{\text{rel}}(M')}{\tau} + \frac{1}{c} \right) = \frac{2\tau_{\text{rel}}(M')}{\tau_{\text{rel}}(M)} + \frac{\log c}{c}.
\]

\[ \square \]

B A Compendium of theorems and definitions used from the literature

B.1 MCMC concentration bounds

All of these bounds are static; they receive a fixed \( m \) as input and run the chain for \( m \) steps to generate the trace \( \vec{X}_{1:m} : X_1, X_2, \ldots, X_m \).

In the following theorems \( \vec{X}_{1:m} : X_1, X_2, \ldots, X_m \) is a length \( m \) stationary trace of \( M \), (i.e., \( \vec{X}_{1:m} \sim \pi^{(T)} \)), with mixing time \( \tau_{\text{mix}} \), relaxation time \( \tau_{\text{rel}} \), and second largest eigenvalue \( \lambda \).

**Theorem B.1** (Hoeffding-Type Bounds for Mixing Processes, [FJS18 Thm. 2.1]). For any \( \delta \in (0, 1) \), we have

\[
\mathbb{P} \left( |\hat{\mu} - \mu| \geq \sqrt{\frac{2(1 + \lambda)(\frac{4^2}{\tau}) \ln(\frac{2}{\delta})}{(1 - \lambda)m}} \right) \leq \delta.
\]

**Theorem B.2** (Bernstein-Type Bound for Mixing Process [JSF18 Thm. 1.2]). For any \( \delta \in (0, 1) \), we have

\[
\mathbb{P} \left( |\hat{\mu} - \mu| \geq \frac{10R \ln(\frac{2}{\delta})}{(1 - \lambda)m} + \sqrt{\frac{2(1 + \lambda)v_{\pi} \ln(\frac{2}{\delta})}{(1 - \lambda)m}} \right) \leq \delta.
\]

To compare the sample complexity of algorithms derived from the above bounds note that:
Sample complexity of the static variance-agnostic algorithm: This implies sample complexity
\[ m_H(\mathcal{M}, f, \varepsilon, \delta) = \frac{1 + \Lambda}{1 - \Lambda} \ln(\frac{2}{\delta}) \frac{R^2}{2\varepsilon^2} \in \Theta \left( \frac{1}{\varepsilon^2} \ln(\frac{1}{\varepsilon}) \right). \]

Sample complexity of the static variance-aware algorithm: This implies sample complexity
\[ m_B(\mathcal{M}, f, v, \varepsilon, \delta) = \frac{2}{1 - \Lambda} \ln(\frac{2}{\delta}) \left( \frac{5R}{\varepsilon} + \frac{(1 + \Lambda)V_{\varepsilon}}{\varepsilon^2} \right) \in \Theta \left( \frac{1}{\varepsilon^2} \ln(\frac{1}{\varepsilon}) \left( \frac{R}{\varepsilon} + \frac{V_{\varepsilon}}{\varepsilon^2} \right) \right). \]

Theorem B.3 (McDiarmid inequality for Markov chains [Pan15]). Let \( \mathcal{M} \) be a Markov chain on state space \( \Omega \) and mixing time \( \tau_{\text{mix}} \). Consider \( F : \Omega^m \to \mathbb{R} \) a \( c \)-Lipschitz function. Then for any trace \( \vec{X}_{1:m} : X_1, X_2, \ldots, X_m \) of \( \mathcal{M} \) and any \( \varepsilon > 0 \) we have that
\[ \mathbb{P}_{\vec{X} \sim \pi(T)} \left( F(\vec{X}) \geq \mathbb{E}_{\pi(T)}[F] + \varepsilon \right) \leq \exp \left( \frac{-2\varepsilon^2}{9c^2m\tau_{\text{mix}}} \right). \] (11)

Equivalently, for any \( \delta \in (0, 1) \), we have that
\[ \mathbb{P}_{\vec{X} \sim \pi(T)} \left( F(\vec{X}) \geq \mathbb{E}_{\pi(T)}[F] + \sqrt{\frac{m\tau_{\text{mix}}c^2(\frac{1}{2})^2 \ln(\frac{1}{\delta})}{2}} \right) \leq \delta. \] (12)

### B.2 Background on k-coloring problem

Let \( S \) be a set whose cardinality \( \#S \) is unknown. Assume there is a rapidly mixing Markov chain \( \mathcal{M} \) whose stationary distribution is the uniform distribution of \( S \) denoted by \( U(S) \). Now, note that with \( \mathcal{M} \), we can generate approximately uniform samples from \( S \). Jerrum, Valiente and Vazirani introduced the first FPRAS (henceforth denoted by JVV) for counting any self-reducible, \( \mathcal{M} \), using series of MCMC-mean estimations (employing \( \mathcal{M} \)). JVV’s reduction was applied to different counting problems [BW91, Jer95, JSV04], here we show it for counting k-colorings.

Consider an arbitrary graph \( G = (V, E) \), let \( E \) ve the number of edges, and \( e_1, e_2, \ldots, e_E \) be an ordering of its edges. Based on this ordering we define \( G_{1}, G_{5} = \cdots, G_0 \) a sequence of \( G \)'s subgraphs with the same vertex set \( V \), such that for each \( i \), \( G_i \) is obtained from \( G_i \) by removing \( e_i \), \( G_{i-1} = (V, E) \), and \( G_0 = (V, \emptyset) \).

For each \( 1 \leq i \leq E \) we define \( f_{e_i} : \Gamma(G_i) \to \{0, 1\} \) as \( f_{e_i}(\gamma) = 1 \) if \( \gamma(u) \neq \gamma(v) \), and \( f_{e_i} = 0 \) otherwise, where \( u \) and \( v \) are endpoints of \( e_i \).

Note that
\[ \mathbb{E}[f_{e_i}] = \frac{\#\Gamma(G_i, k)}{\#\Gamma(G_{i-1}, k)}. \]

Thus, to estimate \( \#S \) we can use the following telescoping sum, for a graph with \( n \) vertices:
\[ \#S = \#\Gamma(G_0, k) \prod_{i=1}^{E-1} \mathbb{E}[f_{e_i}] = \left( \prod_{i=1}^{E-1} \frac{\#\Gamma(G_i, k)}{\#\Gamma(G_{i-1}, k)} \right) \#\Gamma(G_0, k) = \left( \prod_{i=1}^{E-1} \frac{\#\Gamma(G_i, k)}{\#\Gamma(G_{i-1}, k)} \right) n^k. \] (13)

Given these intermediate steps, size of \( S \) can be estimated with \( \varepsilon \) precision, by estimating all \( \mathbb{E}[f_{e_i}] \)s with precision \( \varepsilon/\varepsilon \). Using the classic static approaches listed in appx. B.1 the complexity of abstaining a \( \varepsilon, \delta \) approximation for a graph with \( E \) edges will be \( \mathcal{O}(E \cdot \tau_{\text{rel}} \frac{\varepsilon^2}{\varepsilon^2} \ln(\frac{1}{\varepsilon})) = \tilde{\mathcal{O}} \left( E^3 \tau_{\text{rel}} \right) \).
Contents

1 Introduction
   1.1 Preliminaries .......................................................................................... 1
   1.2 Related Work .................................................... 2
   1.3 Our Contributions .......................................................... 3

2 The Inter-Trace Variance
   2.1 An unbiased MCMC-variances estimator: A tale of two chains ......................... 5

3 DynaMITE
   3.1 Discussion and comparison with prior work ......................................................... 8

4 Application to a counting problem ...................................................... 11

A A Compendium of Complementary Material ............................................. 12
   A.1 Intra-Trace variance properties ................................................................. 12
      A.1.1 Bounding trace variance using projection chains: Cycle ....................... 16
   A.2 An unbiased variance estimator .................................................................. 17
   A.3 Analysis of DynaMITE ............................................................................... 18
      A.3.1 Stationarity, Relaxation, and Mixing Times of the Trace Chain ........... 20
      A.3.2 Correctness and efficiency proof ......................................................... 22
      A.3.3 A Note on Nonstationarity ................................................................. 25
   A.4 Missing proofs from 3.1 ........................................................... 26
   A.5 Application to counting k colorings .......................................................... 26
      A.5.1 Definitions and proof sketches from A.5 .......................................... 27
      A.5.2 Detailed Proofs from A.5 ................................................................. 28

B A Compendium of theorems and definitions used from the literature .......... 31
   B.1 MCMC concentration bounds ................................................................. 31
   B.2 Background on k-coloring problem ......................................................... 32

References

[ABH19] Matteo Agostini, Marco Bressan, and Shahrzad Haddadan. Mixing time bounds for graphlet random walks. Information Processing Letters, 152:105851, 2019.

[ALG20] Nima Anari, Kuikui Liu, and S. Gharan. Spectral independence in high-dimensional expanders and applications to the hardcore model. 2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS), pages 1319–1330, 2020.

[ALW97] D. Aldous, L. Lovasz, and P. Winkler. Mixing times for uniformly ergodic markov chains. Stochastic Processes and their Applications, 71:165–185, 1997.

[APO+18] Ariful Azad, Georgios Pavlopoulos, Christos Ouzounis, Nikos Kyrpides, and Aydin Buluç. Hipmc1: a high-performance parallel implementation of the markov clustering algorithm for large-scale networks. Nucleic Acids Research, 46:1–11, 01 2018.
Zongchen Chen, Kuikui Liu, and Eric Vigoda. Optimal mixing of glauber dynamics: Entropy factorization via high-dimensional expansion. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1537–1550, 2021.

Persi Diaconis. The markov chain monte carlo revolution. *Bulletin of the American Mathematical Society*, 46:179–205, 04 2009.

Pierre Del Moral and L. Miclo. Branching and interacting particle systems approximations of feynman–kac formulae with applications to non-linear filtering. *Sem. Probab. Stras.*, 34:1–145, 05 2007.

Anand Dixit and Vivekananda Roy. Mcmc diagnostics for higher dimensions using kullback leibler divergence. *Journal of Statistical Computation and Simulation*, 87(13):2622–2638, 2017.

Anton Enright, S Dongen, and C.A. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic acids research*, 30:1575–84, 05 2002.

James M Flegal and Galin L Jones. Batch means and spectral variance estimators in markov chain monte carlo. *The Annals of Statistics*, 38(2):1034–1070, 2010.

Jianqing Fan, Bai Jiang, and Qiang Sun. Hoeffding’s lemma for markov chains and its applications to statistical learning. *arXiv preprint arXiv:1802.00211*, 2018.

Jianqing Fan, Yuan Liao, and Martina Mincheva. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 75, 09 2013.

A. Frieze and E. Vigoda. Survey of markov chains for randomly sampling colorings. *Combinatorics, Complexity and Chance: A Tribute to Dominic Welsh*, pages 53–71, 2007.

Fuqing Gao, Arnaud Guillin, and Liming Wu. Bernstein-type concentration inequalities for symmetric markov processes. *Theory of Probability & Its Applications*, 58(3):358–382, 2014.

Daniel T. Gillespie. Stochastic simulation of chemical kinetics. *Annual Review of Physical Chemistry*, 58(1):35–55, 2007.

Mikhail Iosifovich Gordin and Boris A Lifšic. The central limit theorem for stationary markov processes. In *Doklady Akademii Nauk*, volume 239(4), pages 766–767. Russian Academy of Sciences, 1978.

Venkatesan Guruswami. Rapidly mixing markov chains: A comparison of techniques. *Available: cs. washington. edu/homes/venkat/pubs/papers. html*, 2000.

Kushagra Gupta and Dootika Vats. Estimating monte carlo variance from multiple markov chains. *arXiv preprint arXiv:2007.04229*, 2020.

J. Hobert, G. Jones, B. Presnell, and J. Rosenthal. On the applicability of regenerative simulation in markov chain monte carlo. *Biometrika*, 89(4):731–743, 2002.

David G Harris and Vladimir Kolmogorov. Parameter estimation for gibbs distributions. *arXiv preprint arXiv:2007.10824*, 2020.
[HKL+15] Daniel J. Hsu, A. Kontorovich, D. Levin, Y. Peres, Csaba Szepesvari, and Geoffrey Wolfer. Mixing time estimation in reversible markov chains from a single sample path. In NIPS, 2015.

[HKL+19] Daniel Hsu, Aryeh Kontorovich, David A. Levin, Yuval Peres, Csaba Szepesvári, and Geoffrey Wolfer. Mixing time estimation in reversible Markov chains from a single sample path. The Annals of Applied Probability, 29(4):2439 – 2480, 2019.

[Hub15] Mark Huber. Approximation algorithms for the normalizing constant of gibbs distributions. The Annals of Applied Probability, 25(2):974–985, 2015.

[HW17] Shahriar Haddadan and Peter Winkler. Mixing of permutations by biased transposition. In Proceedings of the 34th Symposium on Theoretical Aspects of Computer Science, STACS 2017, 2017.

[Jer95] Mark Jerrum. A very simple algorithm for estimating the number of k-colorings of a low-degree graph. Random Struct. Algorithms, 7(2):157–165, September 1995.

[JHCN06] Galin Jones, Murali Haran, Brian Caffo, and Ronald Neath. Fixed-width output analysis for markov chain monte carlo. Journal of the American Statistical Association, 101:1537–1547, 02 2006.

[JSF18] Bai Jiang, Qiang Sun, and Jianqing Fan. Bernstein’s inequality for general markov chains. arXiv preprint arXiv:1805.10721, 2018.

[JSV04] Mark Jerrum, Alistair Sinclair, and Eric Vigoda. A polynomial-time approximation algorithm for the permanent of a matrix with nonnegative entries. J. ACM, 51(4):671–697, 2004.

[JVV86] Mark Jerrum, Leslie G. Valiant, and Vijay V. Vazirani. Random generation of combinatorial structures from a uniform distribution. Theor. Comput. Sci., 43:169–188, 1986.

[Kol18] Vladimir Kolmogorov. A faster approximation algorithm for the gibbs partition function. In Conference On Learning Theory, pages 228–249. PMLR, 2018.

[Lem20] Michal Lemańczyk. General bernstein-like inequality for additive functionals of markov chains. Journal of Theoretical Probability, 2020.

[Lez98] Pascal Lezaud. Chernoff-type bound for finite markov chains. The Annals of Applied Probability, 8, 08 1998.

[LP04] Carlos Leon and François Perron. Optimal hoeffding bounds for discrete reversible markov chains. Annals of Applied Probability, 14, 05 2004.

[LP16] David A Levin and Yuval Peres. Estimating the spectral gap of a reversible markov chain from a short trajectory. arXiv preprint arXiv:1612.05330, 2016.

[LP17] David A Levin and Yuval Peres. Markov chains and mixing times, volume 107. American Mathematical Soc., 2017.

[Mia12] Blazej Miasojedow. Hoeffding’s inequalities for geometrically ergodic markov chains on general state space. Statistics & Probability Letters, 87, 01 2012.
[MR01] Stefania Mignani and Rosa Rodolfo. Markov chain monte carlo in statistical mechanics: The problem of accuracy. Technometrics, 43:347–55, 2001.

[MSA08] Volodymyr Mnih, Csaba Szepesvári, and Jean-Yves Audibert. Empirical bernstein stopping. In Proceedings of the 25th international conference on Machine learning, pages 672–679. ACM, 2008.

[Pau15] Daniel Paulin. Concentration inequalities for markov chains by marton couplings and spectral methods. Electron. J. Probab., 20, 2015.

[PJO99] Foster Provost, David Jensen, and Tim Oates. Efficient progressive sampling. In Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’99, page 23–32, New York, NY, USA, 1999. Association for Computing Machinery.

[R+02] Jeffrey Rosenthal et al. Quantitative convergence rates of markov chains: A simple account. Electronic Communications in Probability, 7:123–128, 2002.

[Ran6] Dana Randall. Rapidly mixing markov chains with applications in computer science and physics. IEEE Educational Activities Department, 8(2):30–41, March 2006.

[RRJW20] Maxim Rabinovich, Aaditya Ramdas, Michael I Jordan, and Martin J Wainwright. Function-specific mixing times and concentration away from equilibrium. Bayesian Analysis, 15(2):505–532, 2020.

[SAB+19] M. Salatino, J. Austermann, J. A. Beall, S. Choi, K. T. Crowley, S. Duff, S. W. Henderson, G. Hilton, S. P. Ho, J. Hubmayr, Y. Li, M. D. Niemack, S. M. Simon, S. T. Staggs, and E. J. Wollack. Machine learning, markov chain monte carlo, and optimal algorithms to characterize the advact kilopixel transition-edge sensor arrays. IEEE Transactions on Applied Superconductivity, 29(5):1–5, 2019.

[SS95] W. Song and B. Schmeiser. Optimal mean-squared-error batch sizes. Management Science, 41:110–123, 1995.

[SVV07] D. Stefankovic, S. Vempala, and E. Vigoda. Adaptive simulated annealing: A near-optimal connection between sampling and counting. In 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS’07), pages 183–193, 2007.

[TZ02] Z. Tu and Song Zhu. Image segmentation by data-driven markov chain monte carlo. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 24:657–673, 06 2002.

[Val79] Leslie G. Valiant. The complexity of enumeration and reliability problems. SIAM J. Comput., 8:410–421, 1979.

[Var18] Shai Vardi. Randomly Coloring Graphs of Logarithmically Bounded Pathwidth. In Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2018), volume 116 of Leibniz International Proceedings in Informatics (LIPIcs), pages 57:1–57:19, 2018.

[Vaz91] Umesh Vazirani. Rapidly mixing markov chains. In Proceedings of symposia in applied mathematics, volume 44, pages 99–121, 1991.
[VCUR12] Fabio Vandin, Patrick Clay, Eli Upfal, and Benjamin J Raphael. Discovery of mutated subnetworks associated with clinical data in cancer. In Biocomputing 2012, pages 55–66. World Scientific, 2012.

[VF18] Dootika Vats and James M Flegal. Lugsail lag windows and their application to mcmc. arXiv preprint arXiv:1809.04541, 2018.

[VFJ18] Dootika Vats, James M. Flegal, and Galin L. Jones. Strong consistency of multivariate spectral variance estimators in markov chain monte carlo. Bernoulli, 24(3):1860–1909, 2018.

[Vig99] E. Vigoda. Improved bounds for sampling colorings. In 40th Annual Symposium on Foundations of Computer Science (Cat. No.99CB37039), volume 40, pages 51–59, 1999.

[VRU16] Fabio Vandin, Benjamin J Raphael, and Eli Upfal. On the sample complexity of cancer pathways identification. Journal of Computational Biology, 23(1):30–41, 2016.

[VUR11] Fabio Vandin, Eli Upfal, and Ben Raphael. Algorithms for detecting significantly mutated pathways in cancer. Journal of computational biology : a journal of computational molecular cell biology, 18:507–22, 03 2011.

[VUR12] Fabio Vandin, Eli Upfal, and Benjamin J Raphael. De novo discovery of mutated driver pathways in cancer. Genome research, 22(2):375–385, 2012.

[Wil04] David Bruce Wilson. Mixing times of lozenge tiling and card shuffling markov chains. The Annals of Applied Probability, 14(1):274–325, 2004.

[WK19] Geoffrey Wolfer and Aryeh Kontorovich. Estimating the mixing time of ergodic markov chains. In Alina Beygelzimer and Daniel Hsu, editors, Proceedings of the Thirty-Second Conference on Learning Theory, volume 99 of Proceedings of Machine Learning Research, pages 3120–3159, Phoenix, USA, 25–28 Jun 2019. PMLR.

[Wol20] Geoffrey Wolfer. Mixing time estimation in ergodic markov chains from a single trajectory with contraction methods. In Aryeh Kontorovich and Gergely Neu, editors, Proceedings of the 31st International Conference on Algorithmic Learning Theory, volume 117 of Proceedings of Machine Learning Research, pages 890–905. PMLR, 08 Feb–11 Feb 2020.

[YK06] Ming Yuan and Christina Kendziorski. Hidden markov models for microarray time course data in multiple biological conditions. Journal of the American Statistical Association, 101:1323–1332, 02 2006.