arXiv: arXiv:0000.0000

BERNSTEIN'S INEQUALITY FOR GENERAL MARKOV CHAINS

By Bai Jiang[†], Qiang Sun[‡] and Jianqing Fan[†,*]

Princeton University[†], *University of Toronto*[‡]

We prove a sharp Bernstein inequality for general-state-space and not necessarily reversible Markov chains. It is sharp in the sense that the variance proxy term is optimal. Our result covers the classical Bernstein's inequality for independent random variables as a special case.

1. Introduction. Concentration inequalities bounds the probability that the sum of random variables deviates from its mean and have found enormous applications in statistics, machine learning and information theory. Many of them can be derived by the Chernoff approach (Boucheron et al., 2013). Let $Z_1, \ldots, Z_n$ be n independent random variables with $\mathbb{E}Z_i = 0$. Suppose the moment generating function of their sum $\sum_{i=1}^n Z_i$ is bounded with a certain convex function g in the way

$$\mathbb{E}\left[e^{t\sum_{i=1}^n Z_i}\right] \le e^{ng(t)}. \tag{1.1}$$

Let $g^*(\epsilon) = \sup_t\{t\epsilon - g(t)\}$ be the Fenchel conjugate of g. The Chernoff approach implies that, for $\epsilon > 0$,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n Z_i > \epsilon\right) \le e^{-ng^*(\epsilon)}. \tag{1.2}$$

The conjugation and second-order properties of convex functions (Gorni, 1991) assert that if

$$g(t) = \frac{Vt^2}{2} + o(t^2) \quad \text{for small } t,$$

then

$$g^*(\epsilon) = \frac{\epsilon^2}{2V} + o(\epsilon^2) \quad \text{for small } \epsilon.$$

---

[*]Corresponding author.

MSC 2010 subject classifications: Primary 60J05; secondary 62J05, 65C05

Keywords and phrases: Markov chain, Bernstein's inequality, general state space

Thus a tighter bound of the moment generating function in (1.1) with a sharper coefficient $V := 2 \times \lim_{t\to 0} g(t)/t^2$ leads to a tighter concentration in (1.2).

We refer the quantity V as the variance proxy of $\sum_{i=1}^n Z_i/\sqrt{n}$, as it not only characterizes the shape of the sub-gaussian tail in (1.2) but also naturally upper bounds the variance of $\sum_{i=1}^n Z_i/\sqrt{n}$. Indeed, plugging expansions

$$\begin{aligned}
\mathbb{E}\left[e^{t\sum_{i=1}^n Z_i}\right] &= 1 + \mathbb{E}\left[\sum_{i=1}^n Z_i\right]\cdot t + \mathbb{E}\left[\left(\sum_{i=1}^n Z_i\right)^2\right]\cdot\frac{t^2}{2} + \dots \\
&= 1 + \mathrm{Var}\left[\sum_{i=1}^n Z_i\right]\cdot\frac{t^2}{2} + o(t^2), \quad \text{and} \\
e^{ng(t)} &= 1 + ng(t) + \frac{(ng(t))^2}{2} + \dots \\
&= 1 + nV\cdot\frac{t^2}{2} + o(t^2)
\end{aligned}$$

into (1.1) yields $\mathrm{Var}\left[\sum_{i=1}^n Z_i\right]\cdot t^2 \le nV\cdot t^2 + o(t^2)$. Dividing both sides by t^2 and taking $t \to 0$ yields $\mathrm{Var}\left[\sum_{i=1}^n Z_i\right] \le nV$.

Sergei Bernstein, George Bennett and Wassily Hoeffding pioneered in the study of concentration inequalities and named three well-known results. For n independent random variables $Z_1, \dots, Z_n$ such that $\mathbb{E}Z_i = 0$ and $|Z_i| < c$ almost surely, Hoeffding's inequality (Hoeffding, 1963) holds with

$$g(t) = \frac{c^2t^2}{2}, \quad g^*(\epsilon) = \frac{\epsilon^2}{2c^2}, \quad V = c^2.$$

If taking the variances of Z_i into account and letting $\sigma^2 = \sum_{i=1}^n \mathrm{Var}Z_i/n$, Bennett's inequality (Bennett, 1962) holds with

$$g(t) = \frac{\sigma^2}{c^2}(e^{tc} - 1 - tc), \quad g^*(\epsilon) = \frac{\sigma^2}{c^2}h\left(\frac{c\epsilon}{\sigma^2}\right), \quad V = \sigma^2, \tag{1.3}$$

where $h(u) = (1+u)\log(1+u) - u$. Bennett's inequality, together with the fact that $h(u) \ge u^2/2(1+u/3)$ for $u \ge 0$, further implies Bernstein's inequality (Bernstein, 1946):

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n Z_i > \epsilon\right) \le \exp\left(-\frac{n\epsilon^2}{2\left(\sigma^2 + \frac{c\epsilon}{3}\right)}\right). \tag{1.4}$$

For both theoretical interest in probability and practical needs from modern applications in statistics, machine learning and information theory, researchers have tried to study concentration inequalities for Markov-dependent random variables. These inequalities consider the cases in which $Z_i = f_i(X_i)$ with $\{X_i\}_{i\geq 1}$ being a Markov chain. A large body of literature utilize the operator theory in Hilbert spaces and produce inequalities involving the $\mathcal{L}_2$-spectral gap, a quantity measuring the converging speed of the Markov chain towards its invariant distribution.

Among these work, León and Perron (2004) established a sharp Hoeffding-type inequality for finite-state-space, reversible Markov chains, using convex majorizations of transition probability matrices. Miasojedow (2014) developed a discretization technique to prove a variant of Léon and Perron's result for general Markov chains. Throughout the paper we use general Markov chains to denote general-state-space and not necessarily reversible Markov chains. Fan et al. (2018) improved upon (Miasojedow, 2014) and established the exact counterpart of Hoeffding's inequality for general Markov chains.

Comparing with Hoeffding-type inequalities which only use the bound c, both Bennett's and Berstein's inequalities have incorporated variances and thus can be sharper, especially in the case $\sigma^2 \ll c^2$, as would be the case for a random variable that occasionally takes on large values but has relatively small variance. Lezaud (1998a) derive an inequality of this kind for finite-state-space, reversible Markov chains. This work has recently inspired Bernstein-type inequalities in (Paulin, 2015). Unfortunately, their inequalities do not hold for general-state-space Markov chains[1].

In this paper, we establish a sharp Bernstein-type inequality for general Markov chains. Throughout the paper, we denote by $\pi(f)$ the integral of function f with respect to a distribution π. We formally state our first main result.

THEOREM 1.1. *Let $\{X_i\}_{i\geq 1}$ be a stationary Markov chain with invariant distribution π and $\mathcal{L}_2$-spectral gap (see Definition 2.1) $1-\lambda \in (0,1]$. Let $f_i : \mathcal{X} \to [-c, +c]$ be a sequence of bounded functions with $\pi(f_i) = 0$. Let $\sigma^2 = \sum_{i=1}^n \pi(f_i^2)/n$. Then, for any $0 \leq t < (1-\lambda)/5c$,*

$$\mathbb{E}\left[e^{t\sum_{i=1}^n f_i(X_i)}\right] \leq \exp\left(\frac{n\sigma^2}{c^2}(e^{tc} - 1 - tc) + \frac{n\sigma^2\lambda t^2}{1-\lambda-5ct}\right). \tag{1.5}$$

[1] At the beginning of Section 4 and in the remarks after display (13) in (Lezaud, 1998a), Lezaud mentioned that his method based on Kato's perturbation theory (Kato, 2013) does not work for general-state-space Markov chains. However, on page 97 of his PhD thesis (Lezaud, 1998b) (in French), Lezaud wrote that his method has the possibility to be extended to general-state-space Markov chains but provided no proof. Paulin (2015) used Lezaud's method to establish Bernstein-type inequalities for Markov chains and claimed that his inequalities hold for general-state-space Markov chains. After attentively investigating into these literature, we concluded that Lezaud's method does not work for general-state-space Markov chains.

Moreover, we have, for any $\epsilon > 0$,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} f_i(X_i) > \epsilon\right) \le \exp\left(-\frac{n\epsilon^2}{2(A_1\sigma^2 + A_2 c\epsilon)}\right), \tag{1.6}$$

where

$$A_1 = \frac{1+\lambda}{1-\lambda}, \quad A_2 = \frac{1}{3}\mathbb{I}(\lambda = 0) + \frac{5}{1-\lambda}\mathbb{I}(\lambda > 0).$$

In case that $f_i = f$ are time-independent, we can sharpen coefficients in Theorem 1.1 by replacing λ with a smaller quantity $\lambda_+ \vee 0$ related to the right $\mathcal{L}_2$-spectral gap (see Definition 2.2) $1 - \lambda_+$. Here $a \vee b$ denotes the maximum of a and b. Note that $\lambda_+ \in [-1, 1)$, $\lambda \in [0, 1)$ and $|\lambda_+| \le \lambda$, thus $\lambda_+ \vee 0 \le \lambda$. We summarize this result in the following theorem.

THEOREM 1.2. *Let* $\{X_i\}_{i\ge 1}$ *be a stationary Markov chain with invariant distribution* π *and right* $\mathcal{L}_2$*-spectral gap (see Definition 2.2)* $1 - \lambda_+ \in (0, 2]$. *Let* $f : \mathcal{X} \to [-c, +c]$ *be a bounded function with* $\pi(f) = 0$ *and* $\pi(f^2) = \sigma^2$. *Then, for any* $0 \le t < (1 - \lambda_+ \vee 0)/5c$,

$$\mathbb{E}_\pi\left[e^{t\sum_{i=1}^{n} f(X_i)}\right] \le \exp\left(\frac{n\sigma^2}{c^2}(e^{tc} - 1 - tc) + \frac{n\sigma^2(\lambda_+ \vee 0)t^2}{1 - \lambda_+ \vee 0 - 5ct}\right). \tag{1.7}$$

Moreover, we have, for any $\epsilon > 0$,

$$\mathbb{P}_\pi\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) > \epsilon\right) \le \exp\left(-\frac{n\epsilon^2}{2(A_1\sigma^2 + A_2 c\epsilon)}\right), \tag{1.8}$$

where

$$A_1 = \frac{1 + \lambda_+ \vee 0}{1 - \lambda_+ \vee 0} \quad A_2 = \frac{1}{3}\mathbb{I}(\lambda_+ \le 0) + \frac{5}{1-\lambda_+}\mathbb{I}(\lambda_+ > 0).$$

Our results improve upon the literature on three folds. First, our inequality recovers Bernstein's inequality for independent random variables as a special case, whereas the previous inequalities do not. The exponent in the bound (1.5) has two terms. The leading term (scaled by $1/n$) is exactly the function $g(t)$ of Bennett's and Bernstein's inequalities in (1.3). The remainder is caused by the dependence across the Markov chain, and decreases to 0 linearly fast as λ decreases to zero. Correspondingly, the coefficients A_1 and A_2 in (1.6) are increasing with λ. When $\lambda = 0$, Theorem 1.1 recovers the classical Bernstein's inequality for independent random variables. Indeed, independent random variables $Z_1, \ldots, Z_n \in [-c, +c]$ can be seen as transformations of independently and identically distributed (i.i.d.) random variables $U_1, \dots, U_n \sim \text{Uniform}[0,1]$ via the inverse cumulative distribution function $F_{Z_i}^{-1}$, i.e. $Z_i = F_{Z_i}^{-1}(U_i)$; and, i.i.d. random variables $\{U_i\}_{i \geq 1}$ form a stationary Markov chain with $\lambda = \lambda_+ = 0$. Similarly, Theorem 1.2 reduces to the classical Bernstein's inequality for i.i.d. random variables with $\lambda_+ = 0$.

Second, our inequality holds for general-state-space Markov chains, in contrast that the Bernstein-type inequalities in (Paulin, 2015) do not. The latter, in a similar vein to (Lezaud, 1998a), used Kato's perturbation theory (Kato, 2013) to estimate the largest eigenvalue of the perturbed transition probability matrix. However, this estimation method is developed for finite-dimensional transition probability matrices, whereas the transition kernels of general-state-space Markov chains are infinite-dimensional. Such perturbed transition kernels may not have eigenvalues. This obstacle is the main focus of our paper. We overcome it by using techniques developed for the Hoeffding-type inequalities for general Markov chains (Fan et al., 2018).

Third, the variance proxies $\sigma^2 \cdot (1+\lambda)/(1-\lambda)$ for time-dependent functions or $\sigma^2 \cdot (1 + \lambda_+ \vee 0)/(1 - \lambda_+ \vee 0)$ for time-independent functions in our inequalities are optimal. For any $\lambda \in [0, 1)$, consider a stationary Markov chain $\{X_i\}_{i \geq 1}$ with transition probability $P(x, B) = \lambda \mathbb{I}(x \in B) + (1-\lambda)\pi(B)$ for state x and subset B of the state space. This chain has invariant distribution π, $\mathcal{L}_2$-spectral gap $1 - \lambda$ and right $\mathcal{L}_2$-spectral gap $1 - \lambda$, see definitions in Section 2. For function $f : x \mapsto [-c, +c]$ with $\pi(f) = 0$ and $\pi(f^2) = \sigma^2$, Theorem 2.1 in (Geyer, 1992) asserts that

$$\lim_{n\to\infty} \text{Var}\left(\frac{\sum_{i=1}^n f(X_i)}{\sqrt{n}}\right) = \sigma^2 \cdot \frac{1+\lambda}{1-\lambda}.$$

Recall that any variance proxy V bounds the variance of $\sum_{i=1}^n f(X_i)/\sqrt{n}$. Thus, $V \geq \sigma^2 \cdot (1+\lambda)/(1-\lambda)$. This lower bound is indeed achieved in our Theorem 1.1. Variance proxies of our inequalities and others' are compared in Table 1 (for time-independent functions f_i) and Table 2 (for time-dependent function $f_i = f$). Notations in the two tables are referred to Theorems 1.1 and 1.2.

The remaining of this paper is structured as follows. Section 2 presents the preliminaries about operator and spectral theories. Section 3 is devoted to the proofs of Theorems 1.1 and 1.2.

2. Preliminaries. Throughout the paper, we assume the state space $\mathcal{X}$ equipped with a sigma-algebra $\mathcal{B}$ is a standard Borel space[2]. This assumption holds in most practical examples in which $\mathcal{X}$ is a subset of a

[2]A measurable space $(\mathcal{X}, \mathcal{B})$ is standard Borel if it is isomorphic to a subset of $\mathbb{R}$. See Definition 4.33 in (Breiman, 1992).

TABLE 1
Variance proxy in inequalities for time-dependent functions f_i of Markov chains

type	reference	condition	variance proxy V
Hoeffding	(Hoeffding, 1963)	independent	c^2
Hoeffding	(Fan et al., 2018)	general-state-space	$\frac{1+\lambda}{1-\lambda} \cdot c^2$
Bernstein	(Bernstein, 1946)	independent	σ^2
Bennett	(Bennett, 1962)	independent	σ^2
Bernstein	(Paulin, 2015), (3.22)	finite-state-space, reversible	$\frac{4}{1-\lambda^2} \cdot \sigma^2$
Bernstein	Theorem 1.1	general-state-space	$\frac{1+\lambda}{1-\lambda} \cdot \sigma^2$

TABLE 2
Variance proxy in inequalities for time-independent function f of Markov chains

type	reference	condition	variance proxy V
Hoeffding	(Hoeffding, 1963)	independent	c^2
Hoeffding	(León and Perron, 2004)	finite-state-space, reversible	$\frac{1+\lambda_+ \vee 0}{1-\lambda_+ \vee 0} \cdot c^2$
Hoeffding	(Miasojedow, 2014)	general-state-space	$\frac{1+\lambda}{1-\lambda} \cdot c^2$
Hoeffding	(Fan et al., 2018)	general-state-space	$\frac{1+\lambda_+ \vee 0}{1-\lambda_+ \vee 0} \cdot c^2$
Bernstein	(Bernstein, 1946)	independent	σ^2
Bennett	(Bennett, 1962)	independent	σ^2
Chernoff	(Lezaud, 1998a), (1)	finite-state-space, reversible	$\frac{2}{1-\lambda_+ \vee 0} \cdot \sigma^2$
Chernoff	(Lezaud, 1998a), (13)	general-state-space	$\frac{4}{1-\lambda} \cdot \sigma^2$
Bernstein	(Paulin, 2015), (3.20)	finite-state-space, reversible	$\left(\frac{1+\lambda_+ \vee 0}{1-\lambda_+ \vee 0} + 0.8\right) \sigma^2$(*)
Bernstein	(Paulin, 2015), (3.21)	finite-state-space, reversible	$\frac{2}{1-\lambda_+ \vee 0} \cdot \sigma^2$
Bernstein	Theorem 1.2	general-state-space	$\frac{1+\lambda_+ \vee 0}{1-\lambda_+ \vee 0} \cdot \sigma^2$

(*) Inequality (3.20) in (Paulin, 2015) has variance proxy $\sigma^2_{\text{asy}} + 0.8\sigma^2$. Here σ^2_{asy} is the asymptotic variance of $\sum_{i=1}^n f(X_i)/\sqrt{n}$, which is unknown in practice and is $\frac{1+\lambda_+}{1-\lambda_+}\sigma^2$ in the worst case for reversible Markov chains (Rosenthal, 2003). And, the proof in (Paulin, 2015) implicitly requires $\lambda_+ \geq 0$. So we write $\left(\frac{1+\lambda_+ \vee 0}{1-\lambda_+ \vee 0} + 0.8\right) \sigma^2$ as the variance proxy for convenience of comparison.

multi-dimensional real space and $\mathcal{B}$ is the Borel sigma-algebra over $\mathcal{X}$. Let $\{X_i\}_{i \geq 1}$ be a Markov chain on the state space $(\mathcal{X}, \mathcal{B})$ with invariant distribution π.

2.1. *Hilbert space and $\mathcal{L}_2$-spectral gap.* Formally, let $\pi(h) := \int h(x)\pi(dx)$ for any real-valued, $\mathcal{B}$-measurable function $h : \mathcal{X} \to \mathbb{R}$. The set of all square-integrable functions

$$\mathcal{L}_2(\mathcal{X}, \mathcal{B}, \pi) := \{h : \pi(h^2) < \infty\}$$

is a Hilbert space endowed with the inner product

$$\langle h_1, h_2 \rangle_\pi = \int h_1(x) h_2(x) \pi(dx), \quad \forall h_1, h_2 \in \mathcal{L}_2(\mathcal{X}, \mathcal{B}, \pi).$$

We define the norm of a function $h \in \mathcal{L}_2(\mathcal{X}, \mathcal{B}, \pi)$ as

$$\|h\|_\pi = \sqrt{\langle h, h \rangle_\pi},$$

which induces the norm of a linear operator T on $\mathcal{L}_2(\mathcal{X}, \mathcal{B}, \pi)$ as

$$|||T|||_\pi = \sup\{\|Th\|_\pi : \|h\|_\pi = 1\}.$$

We write $\mathcal{L}_2$ in place of $\mathcal{L}_2(\mathcal{X}, \mathcal{B}, \pi)$, whenever the probability space $(\mathcal{X}, \mathcal{B}, \pi)$ is clear in the context.

Each transition probability kernel $P(x, B)$ with $x \in \mathcal{X}$ and $B \in \mathcal{B}$, if invariant with respect to π, corresponds to a bounded linear operator $h \mapsto \int h(y) P(\cdot, dy)$ on $\mathcal{L}_2$. We abuse P to denote this linear operator:

$$Ph(x) = \int h(y) P(x, dy), \quad \forall x \in \mathcal{X}, \ \forall h \in \mathcal{L}_2.$$

Let $\mathit{1} : x \in \mathcal{X} \mapsto 1$ denote the constant operator, and let Π denote the projection operator

$$\Pi : \ h \mapsto \langle h, \mathit{1} \rangle_\pi \mathit{1}.$$

The $\mathcal{L}_2$-spectral gap is defined as follows.

DEFINITION 2.1 ($\mathcal{L}_2$-spectral gap). *A Markov operator P has $\mathcal{L}_2$-spectral gap $1 - \lambda(P)$ if*

$$\lambda(P) := |||P - \mathit{\Pi}|||_\pi < 1.$$

An equivalent characterization of $\lambda(P)$ is given by

$$\lambda(P) = \sup\{\|Ph\|_\pi : \|h\|_\pi = 1, \pi(h) = 0\}.$$

Let P^* be the adjoint operator of Markov operator P. It corresponds to the time-reversal of the transition probability kernel. The real part or the additive reversiblization (Fill, 1991) of P is given by $(P + P^*)/2$, and is self-adjoint. Let $\mathcal{L}_2^0(\pi) := \{h \in \mathcal{L}_2 : \ \pi(h) = 0\}$. It is known that the spectrum of self-adjoint Markov operator acting on $\mathcal{L}_2^0(\pi) := \{h \in \mathcal{L}_2 : \ \pi(h) = 0\}$ is contained in $[-1, +1]$. We define the gap between 1 and the maximum of this spectrum as the right $\mathcal{L}_2$-spectral gap of P.

DEFINITION 2.2 (Right $\mathcal{L}_2$-spectral gap). *A Markov operator P has right $\mathcal{L}_2$-spectral gap $1-\lambda_+(R)$ if its additive reversiblization $R=(P+P^*)/2$ has*

$$\lambda_+(R) := \sup\left\{s: \; s \in \textit{spectrum of } R \textit{ acting on } \mathcal{L}_2^0(\pi)\right\} < 1.$$

2.2. *León-Perron operator and three lemmas.* Let I be the identity operator on $\mathcal{L}_2$. Note that every convex combination of Markov operators produces a Markov operator. We call a Markov operator *León-Perron* if it is a convex combination of I and Π.

DEFINITION 2.3 (León-Perron operator). *A Markov operator $\widehat{P}$ on $\mathcal{L}_2$ is said León-Perron if it is a convex combination of operators I and Π with some coefficient $\lambda \in [0,1)$, that is*

$$\widehat{P} = \lambda I + (1-\lambda)\Pi.$$

The associated transition kernel,

$$\widehat{P}(x,B) = \lambda \mathbb{I}(x \in B) + (1-\lambda)\pi(B), \quad \forall x \in \mathcal{X}, \; \forall B \in \mathcal{B},$$

characterizes a random-scan mechanism: the Markov chain either stays at the current state with probability λ or samples a new state from π with probability $1-\lambda$ at each step.

If a León-Perron operator $\widehat{P}$ shares the same $\mathcal{L}_2$-spectral gap and invariant measure π with a Markov operator P then we call it the León-Perron version of P, which is formally defined below.

DEFINITION 2.4 (León-Perron version). *For a Markov operator P with invariant measure π and $\mathcal{L}_2$-spectral gap $1-\lambda(P) = 1-\lambda$, we say $\widehat{P} = \lambda I + (1-\lambda)\Pi$ is the León-Perron version of P.*

The León-Perron version $\widehat{P}$ has played a central role in (León and Perron, 2004) on Hoeffding-type inequalities for Markov chains. This is why we call it León-Perron.

This paper uses three lemmas of León-Perron operators for general Markov chains, which are originally developed in (Fan et al., 2018). We list them as the following lemmas. In these results, E^{tf} denotes the multiplication operator of function $e^{tf(x)}$:

$$E^{tf}: h \in \mathcal{L}_2 \mapsto he^{tf}.$$

LEMMA 2.1 (Lemma 4.1 in (Fan et al., 2018)). *Let $\{X_i\}_{i\geq 1}$ be a Markov chain with invariant measure π and $\mathcal{L}_2$-spectral gap $1-\lambda \in (0,1]$. Let $\widehat{P} = \lambda I + (1-\lambda)\Pi$. For any bounded functions f_i and any $t \in \mathbb{R}$,*

$$\mathbb{E}\left[e^{t\sum_{i=1}^n f_i(X_i)}\right] \leq \prod_{i=1}^n |||E^{tf_i/2}\widehat{P}E^{tf_i/2}|||_\pi.$$

LEMMA 2.2 (Part of Theorem 2.2 in (Fan et al., 2018)). *Let $\{X_i\}_{i\geq 1}$ be a Markov chain with invariant measure π and right $\mathcal{L}_2$-spectral gap $1-\lambda_+ \in (0,2]$. Let $\widehat{R}_+ = (\lambda_+ \vee 0)I + (1-\lambda_+\vee 0)\Pi$. For any bounded function f and any $t \in \mathbb{R}$,*

$$\mathbb{E}\left[e^{t\sum_{i=1}^n f(X_i)}\right] \leq |||E^{tf/2}\widehat{R}_+E^{tf/2}|||_\pi^n.$$

LEMMA 2.3 (Lemma 4.3 in (Fan et al., 2018)). *Let $\{\widehat{X}_i\}_{i\geq 1}$ be a Markov chain driven by a León-Perron operator $\widehat{P} = \lambda I + (1-\lambda)\Pi$ with some $\lambda \in [0,1)$. For any bounded function f and any $t\in\mathbb{R}$,*

$$\lim_{n\to\infty}\frac{1}{n}\log\mathbb{E}\left[e^{t\sum_{i=1}^n f(\widehat{X}_i)}\right] = \log|||E^{tf/2}\widehat{P}E^{tf/2}|||_\pi.$$

Lemmas 2.1 and 2.2 upper bound the moment generating function of $\sum_{i=1}^n f_i(X_i)$ or $\sum_{i=1}^n f(X_i)$ by the product of the norms of operators $E^{tf_i}\widehat{P}E^{tf_i}$ or $E^{tf}\widehat{R}_+E^{tf}$, where $\widehat{P}$ and $\widehat{R}_+$ are León-Perron. Comparing them to (1.5) and (1.7) suggests that we only need to bound $|||E^{tf/2}\widehat{P}E^{tf/2}|||_\pi$ for a León-Perron operator $\widehat{P}$. Lemma 2.3 studies the large deviation behavior of a Markov chain driven by a León-Perron operator, and shows that the upper bound in Lemma 2.2 is asymptotically tight for such Markov chains.

2.3. *Asymptotic variance and reduced resolvent.* This subsection considers a reversible Markov chain $\{X_i\}_{i\geq 1}$ driven by a self-adjoint P. Suppose this Markov chain admits a right $\mathcal{L}_2$-spectral gap $1-\lambda_+$, i.e. the second largest eigenvalue of P is λ_+. Define the asymptotic variance of f with $\pi(f)=0$ and $\pi(f^2)=\sigma^2$ as

$$\sigma^2_{\text{asy}} := \lim_{n\to\infty}\text{Var}\left(\frac{\sum_{i=1}^n f(X_i)}{\sqrt{n}}\right) = \lim_{n\to\infty}\frac{1}{n}\mathbb{E}_\pi\left[\sum_{i=1}^n\sum_{j=1}^n f(X_i)f(X_j)\right].$$

By (Rosenthal, 2003, Proposition 1),

$$\sigma^2_{\text{asy}} \leq \frac{1+\lambda_+}{1-\lambda_+}\sigma^2.$$

Let S denote the reduced resolvent of P with respect to its largest eigenvalue 1, which can be equivalently defined through (2.11) in (Kato, 2013, Chapter 2):

$$S\mathit{\Pi} = \mathit{\Pi}S = 0, \quad (P - I)S = S(P - I) = I - \mathit{\Pi}.$$

Let $Z = -S$ then, with the convention $T^0 = I$ for any operator T,

$$\begin{aligned} Z = -S = (I - P)^{-1}(I - \mathit{\Pi}) &= \left(\sum_{n=0}^{\infty} P^n\right)(I - \mathit{\Pi}) \\ &= \sum_{n=0}^{\infty}(P^n - \mathit{\Pi}) = \sum_{n=0}^{\infty}(P - \mathit{\Pi})^n - \mathit{\Pi} \\ &= (I - P + \mathit{\Pi})^{-1} - \mathit{\Pi}. \end{aligned}$$

We now write

$$\begin{aligned} \mathbb{E}_\pi\left[\sum_{i=1}^{n}\sum_{j=1}^{n} f(X_i)f(X_j)\right] &= \left\langle \left(\sum_{i=1}^{n}\sum_{j=1}^{n}(P - \mathit{\Pi})^{|j-i|}\right) f, f \right\rangle_\pi \\ &= \left\langle \left[n(2Z - I) - 2Z^2P(I - P^n)\right] f, f \right\rangle_\pi, \end{aligned}$$

which implies that

$$\langle (2Z - I)f, f\rangle_\pi = \sigma_{\text{asy}}^2 \le \frac{1 + \lambda_+}{1 - \lambda_+}\sigma^2, \quad \langle Zf, f\rangle_\pi = \frac{\sigma_{\text{asy}}^2 + \sigma^2}{2} \le \frac{\sigma^2}{1 - \lambda_+}.$$

Because f could be an arbitrary function with mean 0, that Z is self-adjoint, that $Z\mathit{\Pi} = \mathit{\Pi}Z = 0$ and that $\langle Zf, f\rangle \ge 0$ for any f with mean 0, we obtain

$$\begin{aligned} |||Z|||_\pi &= \sup_{h:\ \|h\|_\pi = 1} |\langle Zh, h\rangle_\pi| \\ &= \sup_{h:\ \|h\|_\pi = 1} |\langle Z(I - \mathit{\Pi})h, (I - \mathit{\Pi})h\rangle_\pi| \\ &= \sup_{h:\ \|h\|_\pi = 1} \langle Z(I - \mathit{\Pi})h, (I - \mathit{\Pi})h\rangle_\pi \\ &\le \sup_{h:\ \|h\|_\pi = 1} \frac{\|(I - \mathit{\Pi})h\|_\pi^2}{1 - \lambda_+} = \frac{1}{1 - \lambda_+}. \end{aligned}$$

2.4. *Kato's perturbation theory.* This subsection uses Kato's perturbation theory (Kato, 2013) to expand the largest eigenvalue of PE^{tf} as a series in t. Here, P is a transition probability kernel (matrix) of a finite-state-space, irreducible Markov chain, and E^{tf} is the multiplication operator of function $e^{tf(x)}$, appearing as a diagonal matrix with elements $\{e^{tf(x)} : x \in \mathcal{X}\}$.

The irreducibility of P implies the uniqueness of the invariant distribution π. Let D be the diagonal matrix with elements $\{f(x) : x \in \mathcal{X}\}$. Then we can expand PE^{tf} as

$$PE^{tf} = P\left(\sum_{n=0}^{\infty} \frac{t^n D^n}{n!}\right) = \sum_{n=0}^{\infty} t^n T^{(n)},$$

where $T^{(n)} = PD^n/n!$ for $n \geq 0$. Let $\beta(t)$ be the largest eigenvalue of PE^{tf}. By (2.31) in (Kato, 2013, Chapter 2), for small $|t| < t_0$ (with t_0 being determined later),

$$\beta(t) = \beta^{(0)} + \beta^{(1)}t + \beta^{(2)}t^2 + \dots,$$

where $\beta^{(0)} = 1$ is the largest eigenvalue of P and

$$\beta^{(n)} = \sum_{p=1}^{n} \frac{(-1)^p}{p} \sum_{\substack{v_1 + \dots + v_p = n,\ v_i \geq 1 \\ k_1 + \dots + k_p = p-1,\ k_j \geq 0}} \mathrm{tr}\left(T^{(v_1)} S^{(k_1)} \dots T^{(v_p)} S^{(k_p)}\right)$$

with $S^{(0)} = -\mathit{\Pi}$, $S^{(n)} = S^n$ for $n \geq 1$, and S being the reduced resolvent of P with respect to its largest eigenvalue 1. It is more convenient to substitute S, $T^{(n)}$ with $-Z$ and $PD^n/n!$, respectively. Let $Z^{(n)} = (-1)^n S^{(n)}$ for $n \geq 0$, i.e. $Z^{(0)} = -\mathit{\Pi}$ and $Z^{(n)} = Z^n$ for $n \geq 1$.

$$\beta^{(n)} = \sum_{p=1}^{n} \frac{1}{p} \sum_{\substack{v_1 + \dots + v_p = n,\ v_i \geq 1 \\ k_1 + \dots + k_p = p-1,\ k_j \geq 0}} -\frac{\mathrm{tr}\left(PD^{v_1} Z^{(k_1)} \dots PD^{v_p} Z^{(k_p)}\right)}{v_1! \dots v_p!}$$

By a simple calculation,

$$\beta^{(1)} = -\,\mathrm{tr}\left(PD^1 Z^{(0)}\right) = \mathrm{tr}\left(PD\mathit{\Pi}\right) = \mathrm{tr}\left(D\mathit{\Pi}P\right) = \mathrm{tr}\left(D\mathit{\Pi}\right) = \langle f, \mathit{1}\rangle_\pi = 0.$$

Using the fact that $ZP = Z - (I - \mathit{\Pi})$ derived from the proceeding subsection,

$$\begin{aligned}\beta^{(2)} &= -\frac{1}{1}\frac{\mathrm{tr}\left(PD^2 Z^{(0)}\right)}{2!} - \frac{1}{2}\frac{\mathrm{tr}\left(PD^1 Z^{(0)} PD^1 Z^1\right) + \mathrm{tr}\left(PD^1 Z^1 PD^1 Z^{(0)}\right)}{1!1!} \\ &= \frac{\langle f, f\rangle_\pi}{2} + \langle ZPf, f\rangle_\pi = \langle Zf, f\rangle_\pi - \frac{\langle f, f\rangle_\pi}{2} = \frac{\sigma^2_{\mathrm{asy}}}{2}.\end{aligned}$$

For $n \geq 3$, $\beta^{(n)}$ is more complicated. Bounding $\beta^{(n)}$ is the main task in our proof of the main theorems, and will shown in the proof section.

The convergence radius t_0 is given by (6) in (Lezaud, 1998a)

$$t_0 = \frac{1}{2|||T^{(1)}|||_\pi (1-\lambda_+)^{-1} + c_0}$$

with any c_0 such that

$$|||T^{(n)}|||_\pi \leq |||T^{(1)}|||_\pi c_0^{n-1}.$$

It is easy to verify that $c_0 = c \geq |||D|||_\pi$ is a satisfactory choice. Indeed,

$$\begin{aligned} &|||T^{(1)}|||_\pi = |||PD|||_\pi \leq c, \\ &|||T^{(n)}|||_\pi = \frac{1}{n!}|||PD^n|||_\pi \leq |||PD|||_\pi |||D|||_\pi^{n-1} \leq |||T^{(1)}|||_\pi c^{n-1}. \end{aligned}$$

With this choice,

$$t_0 \geq \frac{1}{2c(1-\lambda_+)^{-1} + c} = \frac{1-\lambda_+}{(3-\lambda_+)c}.$$

3. Proof of Theorems 1.1 and 1.2.

Comparing Lemmas 2.1 and 2.2 to (1.5) and (1.7) suggests that we need to bound $|||E^{tf/2}\widehat{P}E^{tf/2}|||_\pi$ for a León-Perron operator $\widehat{P} = \lambda I + (1-\lambda)\Pi$. This task is attacked by Lemma 3.1.

3.1. *Bounding* $|||E^{tf/2}\widehat{P}E^{tf/2}|||_\pi$.

The proof of Lemma 3.1 consists of three steps. First, we discretize function f as f_k such that $\sup_{x\in\mathcal{X}}|f_k(x) - f(x)| \leq c/k$, $\pi(f_k) = 0$, $\sup_{x\in\mathcal{X}}|f_k(x)| \leq c$ and f_k takes finitely many possible values.

Denote by $\mathcal{Y}$ the range of f_k. Next, we show that $|||E^{tf_k/2}\widehat{P}E^{tf_k/2}|||_\pi = |||\boldsymbol{E}^{ty/2}\widehat{\boldsymbol{Q}}\boldsymbol{E}^{ty/2}|||_{\boldsymbol{\mu}}$, where $\widehat{\boldsymbol{Q}} = \lambda\boldsymbol{I} + (1-\lambda)\mathbf{1}\boldsymbol{\mu}'$ is a transition probability matrix of a finite-state-space Markov chain on $\mathcal{Y}$ with $\boldsymbol{\mu}(y) = \pi(\{x : f_k(x) = y\})$ for each $y \in \mathcal{Y}$, and $\boldsymbol{E}^{ty}$ is the diagonal matrix with elements $\{e^{ty} : y \in \mathcal{Y}\}$. This is formally proved in Lemma 3.2.

Finally, we bound $|||\boldsymbol{E}^{ty/2}\widehat{\boldsymbol{Q}}\boldsymbol{E}^{ty/2}|||_{\boldsymbol{\mu}}$ through Lemma 3.3, which is based on a refinement of the arguments in (Lezaud, 1998a). Lemma 3.1 is concluded by tending $k \to \infty$ such that $|||E^{tf_k/2}\widehat{P}E^{tf_k/2}|||_\pi \to |||E^{tf/2}\widehat{P}E^{tf/2}|||_\pi$.

LEMMA 3.1. *Let $\widehat{P} = \lambda I + (1-\lambda)\Pi$ with some $\lambda \in [0,1)$ be a León-Perron operator. Let $f : \mathcal{X} \to [-c, +c]$ be a bounded function with $\pi(f) = 0$ and $\pi(f^2) = \sigma^2$. Then, for any $0 \leq t < (1-\lambda)/5c$,*

$$|||E^{tf/2}\widehat{P}E^{tf/2}|||_\pi \leq \exp\left(\frac{\sigma^2}{c^2}(e^{tc} - 1 - tc) + \frac{\sigma^2\lambda t^2}{1-\lambda-5ct}\right).$$

PROOF OF LEMMA 3.1. Let a simple function f_k be the $(c/3k)$-approximation of f in the way

$$f_k(x) = \left\lceil \frac{f(x)+c}{c/3k} \right\rceil \times \frac{c}{3k} - c.$$

Then, $|\pi(f_k)| \leq c/3k$, and $\sup_x |f_k(x) - \pi(f_k)| \leq \sup_x |f_k(x)| + |\pi(f_k)| \leq c + c/3k$. Let $\widetilde{f}_k$ be the normalized version of f_k in the way

$$\widetilde{f}_k = \frac{f_k - \pi(f_k)}{1 + 1/3k}.$$

Then, $\sup_x |\widetilde{f}_k(x)| \leq c$, $\pi(\widetilde{f}_k) = 0$ and

$$\begin{aligned}
\sup_x |\widetilde{f}_k(x) - f(x)| &\leq \sup_x |\widetilde{f}_k(x) - [f_k(x) - \pi(f_k)]| + \sup_x |f_k(x) - f(x)| + |\pi(f_k)| \\
&= \sup_x \left| \frac{f_k(x) - \pi(f_k)}{1 + 1/3k} \right| + \sup_x |f_k(x) - f(x)| + |\pi(f_k)| \\
&\leq \frac{c/3k}{1 + 1/3k} + c/3k + c/3k < c/k.
\end{aligned}$$

We still write $\widetilde{f}_k$ as f_k. Note that f_k takes $k' \leq (6k+1)$ possible values, say $\{y_j : 1 \leq j \leq k'\} \subset [-c, +c]$. Let $\widehat{\boldsymbol{Q}} = \lambda \boldsymbol{I} + (1-\lambda)\boldsymbol{1}\boldsymbol{\mu}'$ be a $k' \times k'$ transition probability matrix with $\boldsymbol{\mu}_j = \pi(\{x : f_k(x) = y_j\})$ for $j = 1, \ldots, k'$, and $\boldsymbol{E}^{ty}$ be the diagonal matrix with elements $\{e^{ty_j} : 1 \leq j \leq k'\}$. By Lemma 3.2,

$$|||E^{tf_k/2}\widehat{P}E^{tf_k/2}|||_\pi = |||\boldsymbol{E}^{ty/2}\widehat{\boldsymbol{Q}}\boldsymbol{E}^{ty/2}|||_{\boldsymbol{\mu}}. \tag{3.1}$$

By Lemma 3.3, for any $0 \leq t < (1-\lambda)/5c$,

$$|||\boldsymbol{E}^{ty/2}\widehat{\boldsymbol{Q}}\boldsymbol{E}^{ty/2}|||_{\boldsymbol{\mu}} \leq \exp\left(\frac{\sum_{j=1}^{k'} \boldsymbol{\mu}_j y_j^2}{c^2}(e^{tc} - 1 - tc) + \frac{\left(\sum_{j=1}^{k'} \boldsymbol{\mu}_j y_j^2\right)\lambda t^2}{1 - \lambda - 5ct} \right).$$

Note that $\sum_{j=1}^{k'} \boldsymbol{\mu}_j y_j^2 = \pi(f_k^2)$. And, from the fact that $\sup_x |f_k(x) - f(x)| \leq c/k$, it follows that

$$|||E^{tf/2}\widehat{P}E^{tf/2}|||_\pi \leq e^{ct/k} \times |||E^{tf_k/2}\widehat{P}E^{tf_k/2}|||_\pi.$$

Collecting these pieces together yields that, for any $0 \leq t < (1-\lambda)/5c$,

$$|||E^{tf/2}\widehat{P}E^{tf/2}|||_\pi \leq \exp\left(\frac{ct}{k} + \frac{\pi(f_k^2)}{c^2}(e^{tc} - 1 - tc) + \frac{\pi(f_k^2)\lambda t^2}{1 - \lambda - 5ct} \right).$$

Tending $k \to \infty$ and $\pi(f_k^2) \to \sigma^2$ concludes the proof. $\square$

Lemma 3.2 details the proof of (3.1). For any function f taking finitely many values and any León-Perron operator $\widehat{P}$, we construct two stationary Markov chain $\{\widehat{X}_i\}_{i\geq 1}$ driven by $\widehat{P}$ and $\{\widehat{Y}_i\}_{i\geq 1}$ driven by $\widehat{\boldsymbol{Q}}$ such that $\widehat{Y}_i = f(\widehat{X}_i)$. Then we use the large deviation results of $\{\widehat{X}_i\}_{i\geq 1}$ and $\{\widehat{Y}_i\}_{i\geq 1}$ in Lemma 2.3 to establish the equivalence between the norms of $E^{tf/2}\widehat{P}E^{tf/2}$ and $\boldsymbol{E}^{ty/2}\widehat{\boldsymbol{Q}}\boldsymbol{E}^{ty/2}$.

Lemma 3.2. *Let $\widehat{P} = \lambda I + (1-\lambda)\Pi$ be a León-Perron operator on a general state space $\mathcal{X}$. Let $f : \mathcal{X} \to \{y_j : j = 1,\dots,k\}$ be a simple function taking k possible values. Define $\widehat{\boldsymbol{Q}} = \lambda \boldsymbol{I} + (1-\lambda)\mathbf{1}\boldsymbol{\mu}'$, with $\boldsymbol{\mu}_j = \pi(\{x : f(x) = y_j\})$, as a transition probability matrix on the finite state space $\mathcal{Y} = \{y_j : j = 1,\dots,k\}$. Let $\boldsymbol{E}^{ty}$ denote the diagonal matrix with elements $\{e^{ty_j} : j = 1,\dots,k\}$. Then*

$$|||E^{tf/2}\widehat{P}E^{tf/2}|||_\pi = |||\boldsymbol{E}^{ty/2}\widehat{\boldsymbol{Q}}\boldsymbol{E}^{ty/2}|||_{\boldsymbol{\mu}}.$$

Proofof Lemma 3.2. Construct two stationary Markov chains $\{\widehat{X}_i\}_{i\geq 1}$ and $\{\widehat{Y}_i\}_{i\geq 1}$ in the following way. Let $\{B_i\}_{i\geq 1}$ and $\{Z_i\}_{i\geq 1}$ be sequences of i.i.d. Bernoulli random variables with success probability λ and i.i.d. random variables following π, respectively. It is easy to verify that

$$\begin{aligned} \widehat{X}_1 &= Z_1, \quad \widehat{X}_i = B_i\widehat{X}_{i-1} + (1-B_i)Z_i, \quad \forall i \geq 2; \\ \widehat{Y}_1 &= f(Z_1), \quad \widehat{Y}_i = B_i\widehat{Y}_{i-1} + (1-B_i)f(Z_i), \quad \forall i \geq 2. \end{aligned}$$

are stationary Markov chains driven by $\widehat{P}$ and $\widehat{\boldsymbol{Q}}$, respectively; and that $\widehat{Y}_i = f(\widehat{X}_i)$ for $i \geq 1$. Putting them together with Lemma 2.3 yields

$$\begin{aligned} \log |||E^{tf/2}\widehat{P}E^{tf/2}|||_\pi &= \lim_{n\to\infty} \log \mathbb{E}\left[\exp\left(t\sum_{i=1}^n f(\widehat{X}_i)\right)\right] \\ &= \lim_{n\to\infty} \log \mathbb{E}\left[\exp\left(t\sum_{i=1}^n \widehat{Y}_i\right)\right] \\ &= \log |||\boldsymbol{E}^{ty/2}\widehat{\boldsymbol{Q}}\boldsymbol{E}^{ty/2}|||_{\boldsymbol{\mu}}. \end{aligned}$$

□

We now refine the arguments in Lezaud (1998a) to bound $|||\boldsymbol{E}^{ty/2}\widehat{\boldsymbol{Q}}\boldsymbol{E}^{ty/2}|||_{\boldsymbol{\mu}}$, and give a tighter bound for the coefficient $\beta^{(n)}$ in Kato's expansion than Lezaud (1998a) and Paulin (2015). This tighter bound is stated in Lemma 3.3 and eventually leads to the improvement over their inequalities.

LEMMA 3.3. *Let P be the transition probability matrix of a reversible, finite-state-space, irreducible Markov chain with invariant distribution π and right $\mathcal{L}_2$-spectral gap $1-\lambda_+$, i.e. the second largest eigenvalue of P is λ_+. Let $f : \mathcal{X} \to [-c, +c]$ be a bounded function with $\pi(f) = 0$ and $\pi(f^2) = \sigma^2$. If $\lambda_+ \in [0, 1)$ then, for any $0 \le t < (1-\lambda_+)/5c$,*

$$|||E^{tf/2}PE^{tf/2}|||_\pi \le \exp\left(\frac{\sigma^2}{c^2}(e^{tc} - 1 - tc) + \frac{\sigma^2 |||P - \Pi|||_\pi t^2}{1 - \lambda_+ - 5ct}\right).$$

PROOFOF THEOREM 3.3. Note that each element of matrix $E^{tf/2}PE^{tf/2}$ is non-negative. By Perron-Frobenius theorem, $|||E^{tf/2}PE^{tf/2}|||_\pi$ is equal to its largest eigenvalue. PE^{tf} is similar to $E^{tf/2}PE^{tf/2}$, and thus shares the same eigenvalues. It follows that

$$|||E^{tf/2}PE^{tf/2}|||_\pi = \beta(t) := \text{the largest eigenvalue of } PE^{tf}.$$

Recall that D is the diagonal matrix with elements $\{f(x) : x \in \mathcal{X}\}$, and $Z^{(0)} = -\mathit{\Pi}$ and $Z^{(n)} = Z^n$ for $n \ge 1$ with $Z = \sum_{n=0}^{\infty}(P^n - \mathit{\Pi})$. By Kato's perturbation theory,

$$\beta(t) = \beta^{(0)} + \beta^{(1)}t + \beta^{(2)}t^2 + \ldots, \quad \forall\ |t| \le \frac{1-\lambda_+}{(3-\lambda_+)c},$$

where $\beta^{(0)} = 1, \beta^{(1)} = 0, \beta^{(2)} = \sigma^2_{\text{asy}}/2$, and in general

$$\beta^{(n)} = \sum_{p=1}^{n} \frac{1}{p} \sum_{\substack{v_1 + \cdots + v_p = n,\ v_i \ge 1 \\ k_1 + \cdots + k_p = p-1,\ k_j \ge 0}} - \frac{\operatorname{tr}\left(PD^{v_1}Z^{(k_1)} \ldots PD^{v_p}Z^{(k_p)}\right)}{v_1! \ldots v_p!}.$$

Proceed to bound $\beta^{(n)}$ for $n \ge 3$. Consider two cases.
Case I: $p = 1$. Then $v_1 = n$, $k_1 = 0$, thus the above summand is equal to

$$\frac{\operatorname{tr}(PD^n \mathit{\Pi})}{n!} = \frac{\pi(f^n)}{n!}.$$

Case II: $p \ge 2$. Since $k_1 + \ldots k_p = p - 1$, there exists at least one index j such that k_j is zero. Suppose j_1 is the lowest of such indices. Let $(k'_1, \ldots, k'_p) = (k_{j_1+1}, \ldots, k_p, k_1, \ldots, k_{j_1})$ be the cyclic rotation of the indices. Correspondingly, let $(v'_1, \ldots, v'_p) = (v_{j_1+1}, \ldots, v_p, v_1, \ldots, v_{j_1})$. Then

$$\begin{aligned}
&- \operatorname{tr}\left(PD^{v_1} Z^{(k_1)} PD^{v_1} Z^{(k_2)} \ldots PD^{v_p} Z^{(k_p)}\right) \\
&= - \operatorname{tr}\left(PD^{v'_1} Z^{(k'_1)} PD^{v'_2} Z^{(k'_2)} \ldots PD^{v'_p} Z^{(k'_p)}\right) \\
&= \operatorname{tr}\left(PD^{v'_1} Z^{(k'_1)} PD^{v'_2} Z^{(k'_2)} \ldots PD^{v'_{p-1}} Z^{(k'_{p-1})} PD^{v'_p} \mathit{\Pi}\right) \\
&= \operatorname{tr}\left(D^{v'_1} Z^{(k'_1)} PD^{v'_2} Z^{(k'_2)} \ldots PD^{v'_{p-1}} Z^{(k'_{p-1})} PD^{v'_p} \mathit{\Pi} P\right) \\
&= \operatorname{tr}\left(D^{v'_1} Z^{(k'_1)} PD^{v'_2} Z^{(k'_2)} \ldots PD^{v'_{p-1}} Z^{(k'_{p-1})} PD^{v'_p} \mathit{\Pi}\right) \\
&= \langle f, D^{v'_1-1} Z^{(k'_1)} PD^{v'_2} Z^{(k'_2)} \ldots PD^{v'_{p-1}} Z^{(k'_{p-1})} PD^{v'_p-1} f\rangle_\pi.
\end{aligned}$$

On the other hand, $k'_1 + \cdots + k'_{p-1} = p - 1 \geq 1$, implying that at least one k'_{j_2} with index $j_2 \in \{1, \ldots, p-1\}$ is non-zero. From $Z\mathit{\Pi} = 0$, it follows that $Z^{(k'_{j_2})} P = Z^{(k'_{j_2})}(P - \mathit{\Pi})$. Thus,

$$\begin{aligned}
&|\langle f, D^{v'_1-1} Z^{(k'_1)} PD^{v'_2} Z^{(k'_2)} \ldots Z^{(k'_{j_2})} P \ldots PD^{v'_{p-1}} Z^{(k'_{p-1})} PD^{v'_p-1} f\rangle_\pi| \\
&= |\langle f, D^{v'_1-1} Z^{(k'_1)} PD^{v'_2} Z^{(k'_2)} \ldots Z^{(k'_{j_2})} (P - \mathit{\Pi}) \ldots PD^{v'_{p-1}} Z^{(k'_{p-1})} PD^{v'_p-1} f\rangle_\pi| \\
&\leq \sigma^2 |||P - \mathit{\Pi}|||_\pi |||D|||_\pi^{n-2} |||Z|||^{p-1} \leq \sigma^2 |||P - \mathit{\Pi}|||_\pi c^{n-2} (1 - \lambda_+)^{-(n-1)},
\end{aligned}$$

where the last step uses the facts that $|||D|||_\pi \leq c$, that $|||Z|||_\pi \leq (1 - \lambda_+)^{-1}$ and the assumption that $\lambda_+ \geq 0$ (thus $(1 - \lambda_+)^{-1} \geq 1$). Note that

$$\# \left\{ (v_1, \ldots, v_p) : \sum_{i=1}^p v_i = n,\ v_i \geq 1 \right\} = \binom{n-1}{p-1},$$
$$\# \left\{ (k_1, \ldots, k_p) : \sum_{j=1}^p k_j = p - 1,\ k_j \geq 0 \right\} = \binom{2p-2}{p-1}.$$

and that, by Lezaud (1998a), pp. 856,

$$\sum_{p=1}^n \frac{1}{p} \binom{n-1}{p-1} \binom{2p-2}{p-1} \leq 5^{n-2}, \quad \forall\ n \geq 3.$$

Collecting these pieces together yields

$$\begin{aligned}
|\beta^{(n)}| &\leq \frac{\pi(f^n)}{n!} + 5^{n-2} \times \frac{\sigma^2 |||P - \mathit{\Pi}|||_\pi c^{n-2}}{(1 - \lambda_+)^{n-1}} \\
&= \frac{\pi(f^n)}{n!} + \frac{\sigma^2 |||P - \mathit{\Pi}|||_\pi}{5c} \left(\frac{5c}{1 - \lambda_+} \right)^{n-1}, \ \forall\ n \geq 3.
\end{aligned}$$

This inequality also holds for $n = 2$, as

$$\beta^{(2)} = \frac{\sigma_{\rm asy}^2}{2} \le \sigma^2 \left(\frac{1}{2} + \frac{\lambda_+}{1-\lambda_+}\right) \le \frac{\sigma^2}{2} + |||P - \mathit{\Pi}|||_\pi \times \frac{\sigma^2}{1-\lambda_+}.$$

It follows that, for any $0 \le t < (1-\lambda_+)/5c < (1-\lambda_+)/(3-\lambda_+)c$, $\beta(t)$ is bounded as

$$\begin{aligned}
\beta(t) &= \beta^{(0)} + \beta^{(1)}t + \sum_{n=2}^{\infty} \beta^{(n)} t^n \\
&\le 1 + 0 + \sum_{n=2}^{\infty} \frac{\pi(f^n)t^n}{n!} + |||P - \Pi|||_\pi \times \sum_{n=2}^{\infty} \frac{\sigma^2 t}{5c}\left(\frac{5ct}{1-\lambda_+}\right)^{n-1} \\
&\le \exp\left(\sum_{n=2}^{\infty} \frac{\pi(f^n)t^n}{n!} + |||P - \Pi|||_\pi \times \sum_{n=2}^{\infty} \frac{\sigma^2 t}{5c}\left(\frac{5ct}{1-\lambda_+}\right)^{n-1}\right).
\end{aligned}$$

Putting it together with the facts that, for any $t \ge 0$

$$\sum_{n=2}^{\infty} \frac{\pi(f^n)t^n}{n!} \le \sum_{n=2}^{\infty} \frac{\pi(f^2)c^{n-2}t^n}{n!} = \frac{\sigma^2}{c^2}\sum_{n=2}^{\infty} \frac{c^n t^n}{n!} = \frac{\sigma^2}{c^2}\left(e^{tc} - 1 - tc\right),$$

and that, for any $0 \le t < (1-\lambda_+)/5c$

$$\sum_{n=2}^{\infty} \frac{\sigma^2 t}{5c}\left(\frac{5ct}{1-\lambda_+}\right)^{n-1} = \frac{\sigma^2 t^2}{1-\lambda_+ - 5ct}.$$

completes the proof. □

3.2. *Proofs of Theorems 1.1 and 1.2.* We present the proofs to Theorems 1.1 and 1.2.

Proofof Theorem 1.1. The first claim follows from Lemmas 2.1 and 3.1. For the second claim, we first consider the case of $\lambda > 0$. Let

$$g_1(t) = \begin{cases} 0 & \text{if } t \le 0 \\ \frac{\sigma^2}{c^2}\left(e^{tc} - 1 - tc\right) & \text{if } t > 0 \end{cases}$$

and

$$g_2(t) = \begin{cases} 0 & \text{if } t \le 0 \\ \frac{\sigma^2 \lambda t^2}{1-\lambda-5ct} & \text{if } 0 < t < \frac{1-\lambda}{5c} \\ \infty & \text{if } t \ge \frac{1-\lambda}{5c} \end{cases}.$$

Both are closed proper convex functions and admit Frechet conjugates

$$g_1^*(\epsilon_1) = \begin{cases} \frac{\sigma^2}{c^2} h_1(c\epsilon_1/\sigma^2) & \text{if } \epsilon_1 \geq 0, \\ +\infty & \text{if } \epsilon_1 < 0, \end{cases}$$

with $h_1(u) = (1+u)\log(1+u) - u$; and

$$g_2^*(\epsilon_2) = \begin{cases} \frac{(1-\lambda)\epsilon_2^2}{2\lambda\sigma^2 h_2(5c\epsilon_2/\lambda\sigma^2)} & \text{if } \epsilon_2 \geq 0, \\ +\infty & \text{if } \epsilon_2 < 0, \end{cases}$$

with $h_2(u) = \sqrt{1+u} + u/2 + 1$. By the Chernoff bound,

$$-\log \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n f_i(X_i) > \epsilon\right) \geq n \times \sup\{t\epsilon - g_1(t) - g_2(t) : \ t > 0\}.$$

Since $g_1(t) = O(t^2)$ and $g_2(t) = O(t^2)$ as $t \to 0$, $t\epsilon - g_1(t) - g_2(t) > 0$ for some $t > 0$; and, for $t \leq 0$, $t\epsilon - g_1(t) - g_2(t) \leq 0$. Hence

$$\sup\{t\epsilon - g_1(t) - g_2(t) : \ t > 0\} = \sup\{t\epsilon - g_1(t) - g_2(t) : \ t \in \mathbb{R}\} = (g_1 + g_2)^*(\epsilon).$$

Recall that both g_1 and g_2 are closed proper convex functions. Thus, the Fenchel conjugate of their sum is the infimal convolution of their conjugates g_1^* and g_2^*.

$$\begin{aligned}(g_1 + g_2)^*(\epsilon) &= \inf\{g_1^*(\epsilon_1) + g_2^*(\epsilon_2) : \ \epsilon_1 + \epsilon_2 = \epsilon\} \\ &= \inf\left\{\frac{\sigma^2}{c^2} h_1\left(\frac{c\epsilon_1}{\sigma^2}\right) + \frac{(1-\lambda)\epsilon_2^2}{2\lambda\sigma^2 h_2(\frac{5c\epsilon_2}{\lambda\sigma^2})} : \ \epsilon_1 + \epsilon_2 = \epsilon, \ \epsilon_1 \geq 0, \ \epsilon_2 \geq 0\right\}\end{aligned}$$

Using the facts that $h_1(u) \geq u^2/2(1+u/3)$ and $h_2(u) \leq 2 + u$ for $u \geq 0$, we lower bound this term as

$$(g_1 + g_2)^*(\epsilon) \geq \inf\left\{\frac{\epsilon_1^2}{2\left(\sigma^2 + \frac{c\epsilon_1}{3}\right)} + \frac{\epsilon_2^2}{2\left(\frac{2\lambda}{1-\lambda}\sigma^2 + \frac{5c\epsilon_2}{1-\lambda}\right)} : \ \epsilon_1 + \epsilon_2 = \epsilon, \ \epsilon_1 \geq 0, \ \epsilon_2 \geq 0\right\}$$

Using the fact that $\frac{\epsilon_1^2}{a} + \frac{\epsilon_2^2}{b} \geq \frac{(\epsilon_1+\epsilon_2)^2}{a+b}$ for any positive $\epsilon_1, \epsilon_2, a, b$, we further bound

$$\begin{aligned}(g_1 + g_2)^*(\epsilon) &\geq \inf\left\{\frac{(\epsilon_1 + \epsilon_2)^2}{2\left(\sigma^2 + \frac{c\epsilon_1}{3}\right) + 2\left(\frac{2\lambda}{1-\lambda}\sigma^2 + \frac{5c\epsilon_2}{1-\lambda}\right)} : \ \epsilon_1 + \epsilon_2 = \epsilon, \ \epsilon_1 \geq 0, \ \epsilon_2 \geq 0\right\} \\ &= \frac{\epsilon^2}{2\left(\frac{1+\lambda}{1-\lambda}\sigma^2 + \frac{5c\epsilon}{1-\lambda}\right)}.\end{aligned}$$

It remains to consider the case of $\lambda = 0$. To this end,

$$-\log \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} f_i(X_i) > \epsilon\right) \geq n \times \sup\{t\epsilon - g_1(t) : \ t \in \mathbb{R}\}$$

$$= n \times g_1^*(\epsilon) \geq \frac{n\epsilon^2}{2(\sigma^2 + c\epsilon/3)}.$$

□

PROOFOF THEOREM 1.2. The first claim follows from Lemmas 2.2 and 3.1. And the second claim follows from a similar argument used for the second claim of Theorem 1.1. □

References.

BENNETT, G. (1962). Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association* **57** 33–45.

BERNSTEIN, S. (1946). The theory of probabilities.

BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration inequalities: a nonasymptotic theory of independence*. Oxford University Press.

BREIMAN, L. (1992). *Probability, volume 7 of Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics.

FAN, J., JIANG, B. and SUN, Q. (2018). Hoeffding's lemma for markov chains and its applications to statistical learning. *arXiv preprint arXiv:1802.00211* .

FILL, J. A. (1991). Eigenvalue bounds on convergence to stationarity for non-reversible markov chains, with an application to the exclusion process. *The Annals of Applied Probability* 62–87.

GEYER, C. J. (1992). Practical markov chain monte carlo. *Statistical science* 473–483.

GORNI, G. (1991). Conjugation and second-order properties of convex functions. *Journal of Mathematical Analysis and Applications* **158** 293–315.

HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American statistical association* **58** 13–30.

KATO, T. (2013). *Perturbation theory for linear operators*, vol. 132. Springer Science & Business Media.

LEÓN, C. A. and PERRON, F. (2004). Optimal hoeffding bounds for discrete reversible markov chains. *Annals of Applied Probability* 958–970.

LEZAUD, P. (1998a). Chernoff-type bound for finite markov chains. *Annals of Applied Probability* 849–867.

LEZAUD, P. (1998b). *Quantitative study of Markov chains by perturbation of their kernels*. Ph.D. thesis, University Paul Sabatier (Toulouse).

MIASOJEDOW, B. (2014). Hoeffdings inequalities for geometrically ergodic markov chains on general state space. *Statistics & Probability Letters* **87** 115–120.

PAULIN, D. (2015). Concentration inequalities for markov chains by marton couplings and spectral methods. *Electronic Journal of Probability* **20**.

ROSENTHAL, J. S. (2003). Asymptotic variance and convergence rates of nearly-periodic markov chain monte carlo algorithms. *Journal of the American Statistical Association* **98** 169–177.

Department of ORFE
Princeton University
205 Sherred Hall
Princeton, NJ, USA, 08544
E-mail: jqfan@princeton.edu
baij@princeton.edu

Department of Statistical Sciences
University of Toronto
100 St. George Street
Toronto, Ontario, Canada, M5S 3G3
E-mail: qsun@utstat.toronto.edu