ANALYSIS OF A STOCHASTIC APPROXIMATION ALGORITHM FOR COMPUTING QUASI-STATIONARY DISTRIBUTIONS

J. BLANCHET, P. GLYNN, AND S. ZHENG

ABSTRACT. This paper analyzes the convergence properties of an iterative Monte Carlo procedure proposed in the Physics literature for estimating the quasi-stationary distribution on a transient set of a Markov chain (see, [8,7,10]). In contrast to existing linear algebra methods, this approach eliminates the need for explicit transition matrix manipulations in order to compute the principal eigenvector. Our paper analyzes the procedure proposed in the physics literature by casting it as a stochastic approximation algorithm ([24, 17]). Using this connection we are able to not only verify the consistency of the estimator but also provide a rate of convergence in the form of a Central Limit Theorem. We provide a simple example showing that convergence might occur very slowly and indicate how this issue can be alleviated by using averaging.

1. INTRODUCTION

The motivation for the algorithm that we study here comes from the physics literature. The papers [7,8,10] propose and study an iterative Monte Carlo procedure to estimate the quasi-stationary distribution of an interacting particle system (IPS), [18]. As we shall review, a quasi-stationary distribution can be computed via the left principal eigenvector of the substochastic matrix corresponding to non-absorbing states. Consequently, it is natural to use numerical linear algebra methods for computing such principal eigenvector. Nevertheless, the application of these methods become difficult when the underlying matrix is large. In particular, the application of numerical linear algebra methods is specially prohibitive in IPS. In contrast, the approach described in [7,8,10] only uses a small portion of the underlying matrix in each iteration and, thus, the method can be executed in the setting of huge state spaces. We also shall briefly discuss additional Monte Carlo methods in Section 2.1. The idea is to highlight the features that make the algorithm that we analyze desirable in some settings and thus provide motivation for our convergence analysis and further improvements.

The method suggested in [7,8,10], it turns out, is equivalent to a class of algorithms studied in the context of urn processes, [1,2,22]. So, the convergence of the sampling procedure has been rigorously established in the urn process literature. Moreover, in [2], some results on rates of convergence have been obtained. These results involve a Central Limit Theorem (CLT) for the inner product of the following two quantities: any non-principal eigenvector (or linear combinations thereof) and the estimated quasi-stationary vector (i.e. the estimated principal eigenvector). One of our contributions in this paper is the development of a multidimensional CLT for the estimated quasi-stationary vector. Therefore, we can obtain a CLT for the inner product between the estimated quasi-stationary vector and any other vector (not necessarily one which must be represented as the linear combination of non-principal eigenvectors). More generally, our main contributions are as follows:

- Our paper recognizes the algorithm in [7,8,10] as a stochastic approximation algorithm (Section 3.2).
- Using the stochastic approximation connection we prove the convergence of the underlying estimator and provide sufficient conditions for a CLT (Theorem 1) which is not restricted to inner products involving linear combinations of non-principal eigenvectors.
- We recognize common scenarios based on the spectral gap of the underlying substochastic matrix (Section 4) where the CLT of the estimator fails and convergence is slow.

Date: October 7, 2014.

Key words and phrases. Stochastic approximations, quasi-stationary distribution, ODE method.
• More importantly, using Polyak-Ruppert averaging, [23], we suggest an improved algorithm (Section 4.2.1) which exhibits a valid CLT under no additional restriction on the spectral gap of the underlying substochastic matrix.

• We provide an estimator which allows us to compute the variance in the CLT for quasi-stationary expectations, which is particularly well suited for sparse functions, see Section 4.2.2.

We concentrate on discrete-time Markov chains. The adaptation of our results to continuous-time Markov chains is relatively straightforward and it is given in [25].

The rest of the paper is organized as follows. In Section 2 we review some background material on quasi-stationary distributions. We also provide a quick review of other methods which can be used to estimate quasi-stationary distributions, based on numerical linear algebra or Monte Carlo. In Section 3 we provide a quick overview of stochastic approximation methods and we also sketch the proof of convergence (the full proof is given in Sections 6.1 and 6.2). In Section 4.2.1 discusses an improved version of the algorithm using projection along with averaging. In Section 5 provides several numerical experiments designed to show the improvement of projection and averaging suggested by the stochastic approximations connection. All the proofs and technical results are elaborated in Section 6.

2. Background, Notation, and Related Literature

2.1. Quasi-stationary Distribution: Definition and Methods of Computation. We first introduce some notation and basic definitions concerning quasi-stationary distributions. We then explain the main alternative methods for computing quasi-stationary distributions and highlight features that are improved in the algorithm that we analyze.

2.1.1. Basic Definitions. An early reference for quasi-stationary distributions is the work in [5]. Let \( \{X_n : n \geq 0\} \) be a discrete-time, finite-state-space, Markov chain \( \{X_n : n \geq 0\} \) with transition matrix \( P \). Let us assume that 0 is an absorbing state and that \( 1, ..., d \) are non-absorbing. We can partition \( P \) as

\[
P = \begin{bmatrix} 1 & 0 \\ \alpha & Q \end{bmatrix},
\]

where the upper left corner represents the fact that 1 = \( P(X_1 = 0 \mid X_0 = 0) \). We then have that \( \alpha (i) = P(X_1 = 0 \mid X_0 = i) \) and \( Q \) is a substochastic transition matrix of size \( d \) by \( d \).

Now, given a probability distribution \( \pi \) supported on \( \{1, ..., d\} \) we can define for all \( n \geq 1 \)

\[
\mu^n_j (n) = P(X_n = j \mid X_0 \sim \pi, X_1, ..., X_{n-1} \notin \{0\}) = \frac{\pi'Q^n e_j}{\pi'Q^n e},
\]

where \( \{e_j\}_{j=1}^d \) is the standard basis for \( \mathbb{R}^d \) and \( e \) is a vector with all coordinates equal to one. We use the notation \( X_0 \sim \pi \) inside the conditioning to indicate that \( X_0 \) follows the distribution \( \pi \). Throughout the rest of the paper we use “‘” to denote transposition. Then, we can provide the definition of a quasi-stationary distribution.

If there is a distribution \( \pi \) over the transient states \( \{1, ..., d\} \) such that \( \tilde{\mu^n} (n) := (\tilde{\mu^n}_j (n) : 1 \leq j \leq n) \) is independent of \( n \), then we call \( \tilde{\mu^n} := \mu^n (n) \) a quasi-stationary distribution.

Under the assumption that the substochastic matrix \( Q \) is irreducible (although not necessarily aperiodic) it is straightforward to see (from the Perron-Frobenius Theorem, [14]) that there exists a unique quasi-stationary distribution which can be computed via the solution to the principal eigenvector problem

\[
\mu^n Q = \lambda^n \mu^n,
\]

where \( \lambda^n \geq 0 \). The paper [20] discusses the existence of quasi-stationary distributions for infinite spaces.

Assumption: Throughout the rest of the paper we shall assume that \( Q \) is irreducible and that \( Q^n \to 0 \) as \( n \to \infty \).

2.1.2. Computation via Linear Algebra Methods. Classical linear algebra methods such as the power method, [13], might suffer from the curse of dimensionality when the state-space is huge because these methods require manipulation of the whole matrix in each iteration (multiplying by a vector). Monte Carlo power methods, for example [11], can be adapted to produce eigenvectors but they also require working with a significant portion of the underlying matrix. Similar considerations apply to methods discussed in [15, 16].
in which matrix multiplications are required. This property might render these methods difficult to apply in some model settings, such as those given by typical IPS. An important example of such a class of models, motivated by the physics literature, is known as the contact process which we shall discuss in Section [5] see also [15,16,21]. These types of models render existing linear algebra methods difficult to apply. In contrast, the method that we discuss here is an iterative procedure that requires only a few entries of the transition matrix in each iteration. In particular, it only requires the entries corresponding to transitions visited by the chain in what we call “a tour”, which is simply the excursion from a state from \{1,...,d\} suitably selected, to the absorbing state 0.

2.1.3. Computation via the Fleming-Viot Interacting Particle System. A powerful Monte Carlo based alternative for computing the quasi-stationary distribution is given by the Fleming-Viot method, [4,12,20]. This method is based on an IPS which can be applied to countable continuous time Markov chains and even to diffusion process. It consists of \(N\) particles evolving independently, each according to the dynamics of the underlying Markov process. If one particle gets absorbed (i.e. jumps to state 0), it is immediately restarted. As both time, \(t\), and the number of particles, \(N\), tend to infinity the empirical measure of the particle’s position at time \(t\) converges almost surely to the underlying quasi-stationary distribution. A significant advantage of the method is that it can be run in parallel. A disadvantage, however, is that it is asymptotically biased. That is, for a fixed value \(N\), if only \(N\) is send to infinity, the method will not converge to the underlying quasi-stationary distribution.

In contrast to the Fleming-Viot method, the method that we analyze here is asymptotically unbiased as the number of iterations, \(N\), tends to infinity. Moreover, as we shall show, a small modification can be made to the method to ensure convergence at rate \(N^{-1/2}\). Note that such rate of convergence is not possible to achieve in the Fleming-Viot method because of the presence of the bias appearing by truncating the evolution at time \(t\).

2.1.4. Connections to Urn Processes. The algorithm that we study here has been previously analyzed as a generalized Polya’s urn process, [1,2]. The paper [22] provides a comprehensive survey of urn processes. We extend the rate of convergence results obtained in the urn process literature. In particular, the result offered in [22] provides a CLT of the inner product of the estimated quasi-stationary distribution with linear combinations of non-principal right eigenvectors. In contrast, we obtain a multidimensional CLT for the estimated quasi-stationary distribution. Note that the set of non-principal right eigenvectors can never span the whole \(n\) dimensional Euclidian space because the principal left eigenvector is easily seen to be orthogonal to all of the non-principal right eigenvectors (a feature that facilitates the proofs in [2] because their processes are centered). In addition, and perhaps more importantly, the methods that we use here are completely different from those used in the urn process literature (we use stochastic approximation techniques, whereas the urn process literature relies on branching processes). The connection to stochastic approximations allows to obtain significant algorithmic improvements (due to significant gains in the convergence rate via averaging techniques) as we shall explain in Section 4.2.1 and explore empirically in Section 5.

3. Stochastic Approximations Analysis of the Algorithm

3.1. Basic Notions of Stochastic Approximations. We focus on one of the simplest martingale-difference-type noise driving the stochastic approximation sequence which takes the form

\[
\begin{equation}
\theta_{n+1} = \theta_n + \varepsilon_n W_n, \tag{1}
\end{equation}
\]

for \(n \geq 0\), where \(\{\varepsilon_n\}_{n \geq 0}\) is a step-size sequence (user defined) of non-negative numbers and the \(n\)-th noise observation, \(W_n\), depends on the whole history \(\mathcal{F}_n = \sigma\{(\theta_k, W_{k-1}) : 1 \leq k \leq n\}\). The initial condition \(\theta_0\) is given. We require the existence of a function \(g(\cdot)\) such that

\[
E(W_n | \mathcal{F}_n) = g(\theta_n).
\]

The step-size sequence is such that \(\sum \varepsilon_n = \infty\) but \(\sum \varepsilon_n^2 < \infty\) and under mild regularity conditions, to be reviewed momentarily in our setting, we have that \(\theta_n\) converges almost surely to the stable attractors of the ODE

\[
\dot{\theta}(t) = g(\theta(t)).
\]
3.2. The Precise Algorithm in Stochastic Approximation Form. The papers in [7,8,10] provide a derivation of a non-linear equilibrium equation for the quasi-stationary distribution. Such derivation is based on the Kolmogorov’s forward equation and the conceptualization of a quasi-stationary distribution as the long term limiting distribution of the chain conditioned on not absorption. Although the derivation is clever and instructive, it is somewhat heuristic and it does not lend itself to a rate of convergence analysis. A streamlined version of this derivation is given in [25].

In simple words the algorithm proceeds as follows. Suppose that we have \( d \) bins (one for each element in the underlying transient set). At the beginning of the \( n \)-th iteration we have a certain distribution of balls across the bins and we select an initial position according to such distribution. For example, if \( d = 2 \) and there are 3 balls in the first bin and 5 balls in the second bin, then state 1 is selected with probability 3/8 and state 2 is selected with probability 5/8. The first iteration then proceeds by running a Markov chain starting from the selected state \( i \in \{1,\ldots,d\} \) according to the underlying dynamics, until absorption (i.e. until hitting state 0) and we call such a trajectory a tour. We count the number of times state \( j \) is visited during such tour (for \( j \in \{1,\ldots,d\} \), note for example that the \( i \)-th state is visited at least once) and update the distribution of balls across bins by adding these counts accordingly. So, for example if \( d = 2 \) and during the tour state 1 was visiting 2 times, while state 2 was visited 4 times, then the distribution of balls at the beginning of the \((n+1)\)-th iteration will be \((5=3+2, 9=5+4)\). The output of the algorithm is the normalized distribution of balls (in order to obtain a probability vector) after many iterations as an estimate of the quasi-stationary distribution.

We now explain how this procedure can be described in terms of a stochastic approximation recursion.

**Notation:**

- \( \mu_n \) is the sequence of probability vectors of the transient set \( \{1,\ldots,d\} \) obtained at the \( n \)-th iteration of the algorithm. This vector will store the cumulative empirical measure up to, and including, the \( n \)-th iteration of the algorithm. We use \( \mu_n (x) \) to denote the particular value at the transient state \( x \).

- \( \{X_k^{(n)}\}_{k \geq 0} \) is the Markov chain ran during the \( n \)-th iteration of the algorithm. These Markov chains are conditionally independent (given the values \( X_0^{(n)} \)). The \( n \)-th Markov chain has an initial position drawn from the vector \( \mu_n \).

- We define \( \tau(n) = \inf\{k \geq 0 : X_k^{(n)} = 0\} \). (Recall that 0 is the underlying absorbing state.)

We are interested in analyzing the recursion

\[
\mu_{n+1} (x) = \left( \frac{n \sum_{k=0}^{\tau(k)} \mu_n (x)}{n + 1} \right) + \left( \frac{\sum_{k=0}^{\tau(n+1)-1} I \left( X_k^{(n+1)} = x \mid X_0^{(n)} \sim \mu_n \right)}{n + 1} \right),
\]

for all \( x \in \{1,\ldots,d\} \), where the notation

\[
I \left( X_k^{(n+1)} = x \mid X_0^{(n)} \sim \mu_n \right)
\]

described the indicator of the event \( \{X_k^{(n+1)} = x\} \) and we emphasize that \( X_0^{(n)} \) is sampled using the distribution \( \mu_n \). We may select the initial probability distribution \( \mu_0 \) supported on \( \{1,\ldots,d\} \) in an arbitrary way.

We transform \( \mu_n \) into a more familiar stochastic approximation form by writing

\[
\mu_{n+1} (x) = \mu_n (x) + \frac{1}{n + 1} \left( \frac{\sum_{j=0}^{\tau(n+1)-1} I \left( X_j^{(n+1)} = x \mid X_0^{(n)} \sim \mu_n \right) - \mu_n (x)}{\sum_{j=0}^{\tau(n+1)} \tau(j) / (n + 1)} \right).
\]

Compared to the standard form in [1] we recognize that \( \varepsilon_n = 1/(n + 1) \), however, if we attempt to make a direct translation into [1] we see that the denominator is a bit problematic because its conditional expectation (given the whole history of the algorithm up to the end of the \( n \)-th iteration) is not only a function of \( \mu_n \). To
address this issue, we add another variable, $T_n$, leading to the recursions (assuming $T_0 = 0$)

$$T_{n+1} = T_n + \frac{1}{n+2} \left( \tau^{(n+1)} - T_n \right) = \frac{1}{n+1} \sum_{j=0}^{n} \tau^{(j)},$$

$$\mu_{n+1} (x) = \mu_n (x) + \frac{1}{n+1} \left( \sum_{k=0}^{\tau^{(n+1)}-1} \left( \frac{I \left( X_k^{(n+1)} = x \mid X_0^{(n)} \sim \mu_n \right) - \mu_n (x)}{T_n + \tau^{(n+1)}/(n+1)} \right) \right).$$

In order to provide a more succinct notation let us define

$$Y_n \left( \mu', T \right) (x) := \sum_{k=0}^{\tau^{(n+1)}-1} \left( \frac{I \left( X_k = x \mid X_0 \sim \mu \right) - \mu (x)}{T + \tau/(n+1)} \right),$$

$$Z \left( \mu', T \right) := (\tau - T),$$

where $\{X_l : l \geq 0\}$ denotes a generic Markov chain with transition matrix $P$, $X_0$ is distributed according to $\mu$ (supported on $\{1, \ldots, n\}$), and $\tau$ corresponds to the first hitting time to 0 of the chain $\{X_l : l \geq 0\}$. We also write

$$Y \left( \mu', T \right) (x) := \sum_{k=0}^{\tau^{(n+1)}-1} \left( \frac{I \left( X_k = x \mid X_0 \sim \mu \right) - \mu (x)}{T} \right).$$

Note that $Y$ is time homogeneous where as $Y_n$ is not. Then, we can rewrite the stochastic approximation recursion in distribution via

$$\mu_{n+1} (x) = \mu_n (x) + \frac{1}{n+1} Y_n \left( \mu, T_n \right) (x),$$

$$T_{n+1} = T_n + \frac{1}{n+2} Z \left( \mu_n, T_n \right).$$

If we let $\theta_n = (\mu_n', T_n)$ we now have a setting very close to that described in [1], except for the fact that $g (\cdot)$ is time homogeneous (i.e. of the form $g_n (\cdot)$).

We have the following remarks:

- As noted earlier, the term $Y_n$ is not time homogeneous because of the presence of the term $\tau/(n+1)$. It is not difficult to argue (as we shall do in Lemma [1]) that such term is asymptotically negligible because $\tau^{(n)} = O \left( \log (n) \right)$ almost surely.
- Note that each $\mu_n$ during the course of the algorithm is a probability vector, that is, $\mu_n \in H := \{ x \in \mathbb{R}^d_+ : e' x = 1 \}$. So, the boundedness requirement in Theorem 5.2.1 of [17] holds automatically at least for $\mu_n$ although we shall need to argue boundedness for the coordinate $T_n$.
- A similar algorithm can be defined for continuous-time Markov chains by keeping track of the amount of time spent in each transient state. The details are omitted for the purpose of saving space, but the description is given in [25].

3.2.1. Convergence Result: Consistency and CLT. We now state the main result of this section.

**Theorem 1.** Suppose that $\mu_0$ is any probability vector supported on $\{1, \ldots, n\}$ and pick $T_0 \geq 1$. Then $(\mu_n', T_n) \rightarrow \theta^* := (\mu^*, 1/(1 - \lambda_*))$ with probability one, where the left principal eigenvector $\mu^*$ of $Q$ is normalized so that $\mu^* e = 1$. Finally, if $\lambda$ is any non-principal eigenvalue of $Q$ (i.e. $\lambda \neq \lambda_*$) and

$$\text{Re} \left( \frac{1}{1 - \lambda} \right) < \frac{1}{2} \left( \frac{1}{1 - \lambda_*} \right),$$

then

$$n^{1/2} \left( \mu_n - \mu_* \right) \Rightarrow N \left( 0, V_0 \right),$$

for some $V_0$, explicitly characterized by equation [20].

**Sketch of Proof of Theorem** [1]. The full proof is given in Sections 6.1 and 6.2 but we outline the main idea here. The technique uses the ODE method (Theorem 5.2.1 in [17]), which requires us to examine the asymptotic
behavior of the following coupled dynamic system:

\[
\dot{\mu}(t) = \frac{1}{T(t)} E \left( \sum_{k=0}^{\tau-1} (I(X_k = \cdot | X_0 \sim \mu(t)) - \tau \mu(t)) \right) = \frac{1}{T(t)} (\mu(t) \cdot R - (\mu(t) \cdot Re) \mu(t))',
\]

\[
\dot{T}(t) = E(\tau | X_0 \sim \mu(t)) - T(t) = \mu(t) \cdot Re - T(t),
\]

where \( R = (I - Q)^{-1} \). In Section 6.1 we are able to show using Duhamel’s principle that for a given initial position in the probability simplex \( H \), the solution to a suitably reduced dynamical system (obtained by ignoring \( \dot{T}(t) \) and assuming that \( T(t) = 1 \) in the evolution of \( \mu(t) \)) exists and converges as \( t \to \infty \) to its stationary point. This stationary point is the unique solution to the eigenvalue problem \( \mu_\ast R = \rho_\ast \mu_\ast \), \( \mu_\ast e = 1 \), and \( \rho_\ast \geq 0 \), where \( \rho_\ast = 1/(1 - \lambda_\ast) \). The uniqueness of the solution of this eigenvalue problem follows from Perron-Frobenius’ theorem. The complete dynamical system (given above for \( \mu(t) \) and \( T(t) \)) is a time change of the reduced one, so we can connect them via a simple transformation.

Thus, applying Theorem 5.2.1 from [17] we can conclude that \( \mu_n \) converges to the quasi-stationary distribution for all initial configurations \( (\mu_0, T_0) \in H \times [1, \infty) \).

For the CLT we invoke Theorem 10.2.1 in [17]. Because the recursion in (3) uses step size \( \varepsilon_n = 1/(n + 1) \), we need to verify that the Jacobian matrix of the ODE vector field, evaluated at the stability point, has spectral radius less than \(-1/2\). As we show in Section 6.2 this is equivalent to requiring (5). The expression for \( V_0 \) is extracted from the variance of an associated Ornstein-Uhlenbeck process as in p. 332 of [17].

\[ \square \]

4. Variations on the Algorithm with Improved Rate of Convergence

In this section we study what occurs to the rate of convergence when the sufficient conditions for the CLT are not met. We shall study a simple example consisting of two states.

4.1. Counterexample to Square Root Convergence. Consider the Markov chain with states \{0, 1, 2\}, the state 0 is absorbing and the matrix \( Q \) satisfies

\[
Q = \begin{pmatrix}
\frac{(1 - \varepsilon)}{2} & \frac{(1 - \varepsilon)}{2} \\
\frac{(1 - \varepsilon)}{2} & \frac{(1 - \varepsilon)}{2}
\end{pmatrix}.
\]

By symmetry the recursion that we analyze, namely (2), can be tracked by a simple process, \{\( \bar{X}_m : m \geq 0 \)\}, which we describe now. Assume that the distribution of \( X_0 \) is given. At step \( m \), the value of \( \bar{X}_m \) is decided according to a Bernoulli trial which call the type. The type is Bernoulli with success parameter equal to \( 1 - \varepsilon \).

If the type is a success, we sample a second Bernoulli trial with probability 1/2 of success, if the second trial is successful we let \( \bar{X}_m = 1 \), if it is a failure we let \( \bar{X}_m = 2 \).

If the type is a failure (which occurs with probability \( \varepsilon \)), the we sample state 1 or 2 according to the empirical measure of \{\( \bar{X}_k : 0 \leq k \leq m - 1 \)\}.

Let \( T_n \) be time at which the \( n \)-th failure type occurs. Then we can have that recursion (2) is equivalent to studying

\[
\mu_n(x) = \frac{1}{T_n} \sum_{k=0}^{T_n} I(\bar{X}_k = x).
\]

The process \{\( \bar{X}_m : m \geq 0 \)\} is known as a self-interacting Markov chain, see [9]. Inequality (6) in Theorem 1 applied to this case corresponds to requiring \( \varepsilon < 1/2 \). Of course, \( \mu = (1/2, 1/2) \). Reference [9] is applicable to this example and shows that if \( f \neq 0 \) there exists \( \delta > 0 \) such that for \( n \geq 1 \)

\[
\frac{1}{n^{2(1-\varepsilon)}} \leq E \left( (\mu_n f - \mu f)^2 \right) \leq \delta^{-1} \frac{1}{n^{2(1-\varepsilon)}}.
\]

This results indicates that the rate of convergence is not \( O(n^{-1/2}) \) but rather \( O(n^{-1(1-\varepsilon)}) \). In the numerical examples in Section 5 we simulate this simple Markov chain to demonstrate the slow rate of convergence that is present when \( \varepsilon < 1/2 \) and how the approach that we discuss next can be used to speed up convergence.
4.2. Doeblinization. The previous example shows that the rate of convergence can deteriorate substantially if (5) does not hold. We now argue that in fact that there are natural algorithmic “tricks” that one might attempt to use and which are likely to induce a violation of (5). Note that the expected time to absorption starting from the quasi-stationary distribution satisfies $E(\tau \mid X_0 \sim \mu_\ast) = (1 - \lambda_\ast)^{-1}$. If this expected time is large the iterations of the algorithm will tend to be long. In order to shorten the length of the iterations one might “Doeblinize” the chain by multiplying $Q$ by a constant $\alpha \in (0, 1)$. The term Doeblinization is adopted from the name of the French-German probabilist Wolfgang Doeblin, who studied Markov chains satisfying certain minorization condition which is satisfied when multiplying $Q$ by $\alpha$ as indicated earlier.) This operation, of course, does not change the principal eigenvector of $Q$, but it does change all the eigenvalues by the same proportion. Because of the non-linearity of the $1/(1 - \lambda_\ast)$ as a function of $\lambda_\ast$ and its presence in (5), we conclude choosing $\alpha > 0$ too small might result in a significant deterioration in rate of convergence (despite the gain in speed at each iteration). This observation further motivates the need for a technique that allows to obtain a CLT with a square-root convergence which can be guaranteed regardless of the eigenvalues of $Q$.

Doeblinization can also be done for continuous time Markov chain by subtracting $\alpha I$ from the substochastic rate matrix associated with the transient states $\{1, \ldots, n\}$.

4.2.1. Projection and Averaging. Now that the method is under the stochastic approximation umbrella, we can modify the algorithm in order to change the step-size by using its projection variant, given by the recursion

$$
\bar{\mu}_{n+1} = \Pi_H \left( \bar{\mu}_n + \varepsilon_n \left( \sum_{k=0}^{\tau(n+1)-1} \left( I \left( X_k^{(n+1)} = \cdot \mid X_0^{(n)} \sim \mu_n \right) - \bar{\mu}_n \left( \cdot \right) \right) \right) \right),
$$

where $\Pi_H$ denotes the $L_2$-projection into the probability simplex $H$. We still require $\sum \varepsilon_n = \infty$ and $\sum \varepsilon_n^2 < \infty$. In practice we only need to perform a small number of projections. The vector inside the projection operator always has components which add up to one. So, projection is only needed when any of the components of the vector inside $\Pi_H (\cdot)$ is negative. By conveniently splitting the recursion defining the iterates in (7) we can gain insight into when one or more components of the updated vector $\bar{\mu}_n$ can become negative. In particular, we can write

$$
\bar{\mu}_n + \varepsilon_n \left( \sum_{k=0}^{\tau(n+1)-1} \left( I \left( X_k^{(n+1)} = \cdot \mid X_0^{(n)} \sim \mu_n \right) - \bar{\mu}_n \left( \cdot \right) \right) \right) = \bar{\mu}_n \left( 1 - \varepsilon_n \tau^{(n+1)} \right) + \varepsilon_n \sum_{k=0}^{\tau(n+1)-1} I \left( X_k^{(n+1)} = \cdot \mid X_0^{(n)} \sim \bar{\mu}_n \right).
$$

So, for a component to become negative, it is necessary that $\tau^{(n+1)} > \varepsilon_n^{-1}$. It is not difficult to argue, as we shall do in Lemma 3 (part III), that there exists $\delta > 0$ such that $\tau^{(n+1)} > \delta \log (n)$ for only finitely many values of $n \geq 1$ with probability one, thus $\tau^{(n+1)} > \varepsilon_n^{-1}$ occurs only finitely many times if $\varepsilon_n = O \left( n^{-\alpha} \right)$ for $\alpha > 0$. Moreover, it is quite easy to perform the $L_2$ projection into a probability simplex. In particular,

$$
\bar{\mu}_{n+1} = \left( \bar{\mu}_n \left( 1 - \varepsilon_n \tau^{(n+1)} \right) + \varepsilon_n \sum_{k=0}^{\tau(n+1)-1} I \left( X_k^{(n+1)} = \cdot \mid X_0^{(n)} \sim \bar{\mu}_n \right) - u_{n+1} e \right),
$$

where $u_{n+1} > 0$ is the unique constant such that $\bar{\mu}_{n+1} e = 1$ (see (3)). The advantage of the projection version is that we are free to choose slower step sizes so that we can weaken the condition for the required CLT to hold. In particular, when $\varepsilon_n = n^{-\alpha}$ and $\alpha \in (1/2, 1)$ we always obtain a $\varepsilon_n^{1/2}$-CLT. We summarize this observation in the following result proved in Section 7.

**Proposition 1.** If $\varepsilon_n = n^{-\alpha}$ for $\alpha \in (1/2, 1)$ we have that

$$
\varepsilon_n^{-1/2} \left( \bar{\mu}_n - \mu_\ast \right) \Rightarrow N \left( 0, V_1 \right),
$$

where $V_1$ can be characterized via (21).
The Polyak-Ruppert averaging technique, [23], can be applied jointly with the projection algorithm to ensure “square root convergence”, regardless of whether (5) holds or not as the next theorem shows. Its proof takes advantage of the analysis behind Proposition 1 and Theorem 1 and it is given in [7] based on [23].

**Theorem 2.** Suppose that \( \mu_0 \) is any probability vector supported on \( \{1, \ldots, n\} \) and pick \( T_0 \geq 1 \). Selecting \( \varepsilon_n = n^{-\alpha} \) for \( \alpha \in (1/2, 1) \), let

\[
v_n = \frac{1}{n} \sum_{k=1}^{n} \bar{\mu}_k.
\]

Then,

\[
n^{1/2} (v_n - \mu_\ast) \Rightarrow N (0, \tilde{V}_1),
\]

where \( \tilde{V}_1 \) is given in equation (25).

We can apply Theorem 2 in the estimation of quasi-stationary expectations of the form

\[
E (s (X) | X \sim \mu_\ast) = \mu_\ast' s,
\]

using the outcomes of the improved algorithm, which we ultimately advocate using.

**Proposition 2.** Let \( v_n \) be defined as in Theorem 2. For \( \varepsilon_n = n^{-\alpha} \) with \( \alpha \in (1/2, 1) \), set \( n_k = \lfloor k^{\beta/\alpha} \rfloor \) and \( N_n = \lfloor n^{\alpha/\beta} \rfloor \), where \( \beta \in (\alpha, 1) \). Then,

\[
\frac{1}{N_n} \sum_{k=0}^{N_n} \varepsilon_{n_k}^{-1} (\bar{\mu}_{n_k} - v_{n_k}) (\bar{\mu}_{n_k} - v_{n_k})' \to \tilde{V}_1,
\]

as \( n \to \infty \) in probability.

In the context of Corollary 1 we can use Proposition 2 noting that

\[
\frac{1}{N_n} \sum_{k=0}^{N_n} \varepsilon_{n_k}^{-1} (\bar{\mu}_{n_k} - v_{n_k}) (\bar{\mu}_{n_k} - v_{n_k})' s = \frac{1}{N_n} \sum_{k=0}^{N_n} \varepsilon_{n_k}^{-1} (\bar{\mu}_{n_k} - v_{n_k})' s^2.
\]

In turn, we can recursively compute \( \bar{\mu}_{n_k} s - v_{n_k} s \) the previous estimator can be cheaply computed because the \( n \)-th update of the associated recursions takes \( O (E (\tau^{(n+1)} | X_{n+1}^{n+1} \sim \mu_n)) \) function evaluations, in expectation and we saw that the “Doeblinization” procedure allows us to impose a uniform (in \( n \)) upper bound on \( E (\tau^{(n+1)} | X_{n+1}^{n+1} \sim \mu_n) \).

5. **Numerical Experiments**

5.1. **Loopy Markov chain.** We considered the matrix (6). As mentioned before, condition (5) is equivalent to \( \varepsilon < 1/2 \). We tested the original estimator (underlying Theorem 1), and the Polyak-Ruppert estimator (underlying Theorem 2) for the case \( \varepsilon = .98 \), which is well outside the sufficient condition (5). The result can be seen in Figure 1 where the Polyak-Ruppert estimator improves convergence by a significant margin.

5.2. **Birth-Death Chain with Finite Capacity.** We simulated a birth and death process on the set \( \{0, 1, \ldots, 100\} \) with an absorbing boundary at 0 and a reflecting boundary at 100. The probability of going up is 5/9. The expected time to absorption is quite large because the process tends to drift upwards. The Doeblinized Markov chain no longer satisfies the CLT. One can see in Figure 2 that the Polyak-Ruppert estimator significantly outperforms the plain vanilla estimator from Theorem 1.
5.3. **Contact Process.** Let us briefly introduce the contact process, for a motivation from the physics literature see for example [7,8,10,19]. For a given graph \((V,E)\), \((V\) is for vertices and \(E\) is for edges) a Markov jump process \(((X_1(t),...,X_m(t)) : t \geq 0)\) is a contact process if \(X_i(t) \in \{0,1\}\) and the following holds,

- \(|V| = m\) and \(X_i(t)\) is the state of the \(i\)-th vertex at time \(t\).
- The \(i\)-th vertex transitions from state 1 to state 0 at an exponential rate equal to 1.
- The \(i\)-th vertex transitions from state 0 to state 1 at rate \(\lambda r_i(t)\), where \(r_i(t)\) is the fraction of the neighbors of \(i\) which are in state 1 and \(\lambda > 0\) is a parameter called the infection rate.

Note that this continuous time Markov chain has \(2^m\) states and that \((0,...,0)\) is an absorbing state. All of the remaining states are transient.

The contact process will eventually reach absorption, but the motivation in [7,8,10,19] is to investigate the behavior of the system before absorption occurs after a long period of time.

We simulated the contact process on a complete graph (i.e. each vertex is connected with any other vertex). If the infection rate is changed to 1.5 and \(m = 100\) then each iteration of the algorithm can take an extremely long time. We applied Doeblinization by subtracting \(1/2\) the identity matrix. The details of the continuous time adaptation of the algorithm are given in [25]. The eigenvalue condition \(\epsilon > 0\) fails, resulting in a slow speed of convergence of the estimator underlying Theorem 1. The Polyak-Ruppert averaging estimator from Theorem 2 significantly outperforms the plain vanilla estimator, as can be seen from the results in Figure 3.

6. **Proofs of Main Results**

6.1. **Proof of Theorem 1: Convergence.** We first restate a series of assumptions and notations that are used in Theorem 5.2.1 from [17]. We adopt the abstract form of the recursion \(\theta_{n+1} = \theta_n + \epsilon_n W_n\). In our setting
Figure 2. This is the simulation of a M/M/1 queue with 100 queue capacity and ρ = 1.25. We are considering the embedded discrete-time chain at the jump times of the system. We had to Doeblinize the process (multiply transition matrix by 0.95) in order to deal with the large $E[τ]$ due to the system being in heavy-traffic regime. As you can see, the Polyak-Ruppert averaging (red) is significantly better than the original algorithm (green) on the log-log plot. The eigenvalue condition for the CLT is not satisfied after Doeblinization.

$θ_n = (µ_n, T_n)$ and $W_n = (Y_n (θ_n), Z (θ_n))$ as defined in (4). Recall that $F_n$ is the $σ$-field generated by the iterates of the algorithm, namely, $F_n = σ(θ_0, θ_i, W_{i−1} : 1 ≤ i ≤ n)$.

For the almost sure convergence of $µ_n$ to $µ$, we must verify the following conditions.

1. $ε_n → 0$, $∑ ε_n = ∞$, $∑ ε^2_n < ∞$. This is immediately satisfied with the choice $ε_n = 1/(n + 1)$, as in our case. Moreover, define $t_n = ∑_{j=1}^n ε_j$ (with $t_0 = 0$) and let $m(s) = max{n : t_n ≤ s}$.

2. Uniformly bounded variance: $sup_n E ∥W_n∥^2 < ∞$. This is shown in Lemma 4 below. (We can use any norm, but we choose the norm $∥x∥_∞ = max_i |x_i|$.)

3. Local averaging condition: Define $g_n (θ_n) := E (W_n | F_n)$. The family of functions $\{g_n (·)\}_{n≥0}$ must be uniformly equicontinuous, and there must exist a continuous function $g (·)$ such that for each $θ$, and each $t > 0$,

$$\sum_{k=n}^{m(t_n+t)} ε_k (g_k (θ) − g (θ)) → 0,$$

almost surely. This local averaging condition is proved in Lemma 3 below.

In our setting, we write $θ = (µ', T)$ and define $g (θ) = (f' (θ), h (θ))$, where $f$ and $h$ are given via

$$f (µ', T) = \frac{1}{T} E \left( ∑_{k=0}^{T−1} (I (X_k = ·) − µ) \mid X_0 ∼ µ \right) = \frac{1}{T} (µ' R − (µ' Re)µ)',$$

$$h (µ', T) = E (τ | X_0 ∼ µ) − T = µ' Re − T,$$
FIGURE 3. This is a simulation of the contact process on a complete graph where $\lambda = 1.5$ with 100 nodes. The plot is the log-log plot of the number of steps vs. MSE. The sufficient condition for CLT cannot be met in this case after subtracting $0.5I$ from the rate matrix. The Polyak’s averaging algorithm (red) significantly outperforms the vanilla algorithm (green).
6.1.1. Auxiliary Results. Define $\bar{\tau}(x)$ to be a random variable with the distribution of the first passage time to the absorbing state, 0, given that the initial condition of the chain is the transient state $x \in \{1, \ldots, d\}$. Suppose that the random variables $\{\bar{\tau}(j) : 1 \leq j \leq d\}$ are all independent. Then, let $\bar{\tau} = \max\{\bar{\tau}(j) : 1 \leq j \leq d\}$. We have the following simple but useful result.

**Lemma 1.** The following claims hold: i) $\tau^{(n+1)}$ is stochastically bounded by $\bar{\tau}$, ii) there exists $\delta > 0$ such that $E \exp(\delta \bar{\tau}) < \infty$, iii) $P(\tau^{(n+1)} > \log(n) \ i.o.) = 0$, iv) almost surely we have that $1 \leq \lim \sum_{k=1}^{n} \tau^{(k)}/n \leq E(\bar{\tau}) < \infty$.

Proof of Lemma 1: The proof is almost immediate part i) follows regardless of any assumption, for parts ii) to iv) we need $Q^n \rightarrow 0$ as $n \rightarrow \infty$ because this ensures that $\bar{\tau}(x)$ has a finite moment generating function in a neighborhood of the origin.

We also have a useful expression for $f_n(\mu', T)$. Define $v(x,s) := E(\exp(-s\bar{\tau}) | X_0 = x)$.

**Lemma 2.** For each $(\mu', T) \in H \times (0, \infty)$, the $x$-th component of $f_n(\mu', T)$, namely $f_n(\mu', T)(x)$ satisfies

$$f_n(\mu', T)(x) = \int_{0}^{\infty} e^{-Tu} \left[ v\left( x, \frac{u}{n+1} \right) \mu' \left( I - e^{-\frac{u}{n+1} Q} \right)^{-1} e_x - \left( \mu' \left( I - e^{-\frac{u}{n+1} Q} \right)^{-1} v \left( \cdot, \frac{u}{n+1} \right) \right) \mu(x) \right] du,$$

where $e_x$ denotes the vector which has 1 in the $x$-th coordinate and zeroes elsewhere.

Proof of Lemma 2: First note that

$$\frac{1}{T + \tau \left/ (n + 1) \right.} = \int_{0}^{\infty} \exp \left( -(T + \tau \left/ (n + 1) \right.) u \right) du.$$

Then we have (applying Fubini's theorem since $\int_{0}^{\infty} E \left( e^{-Tu} \tau | X_0 \sim \mu \right) du < \infty$),

$$f_n(\mu', T)(x) = \int_{0}^{\infty} E \left( e^{-(T + \frac{u}{n+1})} \sum_{k=0}^{\tau-1} (I(X_k = x) - \mu(x)) \right) | X_0 \sim \mu) du.$$

Again, another application of Fubini's theorem (also valid because $E(\tau|X_0 \sim \mu) < \infty$) yields that the previous expression equals

$$\int_{0}^{\infty} E \left( e^{-(T + \frac{u}{n+1})} \sum_{k=0}^{\infty} I(\tau > k) (I(X_k = x) - \mu(x)) \right) | X_0 \sim \mu) du$$

$$= \int_{0}^{\infty} e^{-Tu} \sum_{k=0}^{\infty} e^{-\frac{k}{n+1} \frac{u}{n+1}} E \left( I(\tau > k) (I(X_k = x) - \mu(x)) \right) | X_0 \sim \mu) du$$

$$= \int_{0}^{\infty} e^{-Tu} \sum_{k=0}^{\infty} e^{-\frac{k}{n+1} \frac{u}{n+1}} \left[ v \left( X_k, \frac{u}{n+1} \right) \mu(X_k - \mu(x)) \right] \mu(\mu | X_0 \sim \mu) du$$

$$= \int_{0}^{\infty} e^{-Tu} \sum_{k=0}^{\infty} e^{-\frac{k}{n+1} \frac{u}{n+1}} \left[ v \left( x, \frac{u}{n+1} \right) \mu'(I - e^{-\frac{u}{n+1} Q}^{-1}) e_x - \left( \mu' \left( I - e^{-\frac{u}{n+1} Q} \right)^{-1} v \left( \cdot, \frac{u}{n+1} \right) \right) \mu(x) \right] du.$$

\[\square\]

6.1.2. Local Averaging and Uniformly Bounded Variance. We first verify the uniformly bounded variance condition

**Lemma 3.** We have that $\{g_n(\cdot)\}$ is a sequence of uniformly equicontinuous functions on $H \times [1, \infty)$ and we have that for each $t > 0$,

$$\lim_{n \rightarrow \infty} \left| \sum_{k=n}^{m(t_n+t)} \varepsilon_k g_k(\mu', T) - g(\mu', T) \right| = 0.$$
Proof of Lemma 4. Clearly, we have that
\[
|f_k (\mu', T) (x) - f (\mu', T) (x)|
\leq E \left( \sum_{k=0}^{\tau - 1} \frac{|I (X_k = x) - \mu (x)|}{T (k + 1 + \tau)} \right) \leq E \left( \sum_{k=0}^{\tau - 1} \frac{|I (X_k = x) - \mu (x)|}{T (k + 1)} \right) \leq E \left( \frac{\varphi^2}{k + 1} \right).
\]
Therefore,
\[
\sum_{k=0}^{\tau - 1} \frac{|I (X_k = x) - \mu (x)|}{T (k + 1 + \tau)} \leq E \left( \frac{\varphi^2}{k + 1} \right) \leq E \left( \frac{\varphi^2}{n + 1} \right) \rightarrow 0
\]
as \( n \rightarrow \infty \). Finally, we need to argue that \( g_n (\cdot) \) is uniformly equicontinuous on compact sets in \( H \times [1, \infty) \). This follows easily by noting from the expression obtained in Lemma 2 that the Jacobian \( (Df_n) (\mu', T) \) is uniformly bounded for a neighborhood around any point \((\mu', T) \in H \times [1, \infty) \).
The coordinate of \( g_n \) corresponding to \( h \) does not depend on \( n \) and thus the result follows immediately in this case.

Now we turn our attention to condition 2., namely, uniformly bounded variance.

\[
\text{Lemma 4. We have that sup}_n E \| W_n \|_\infty^2 < \infty.
\]

Proof of Lemma 4. Clearly, we have that \( \| Y_n (\mu_n, T_n) \|_\infty \leq \tau (n + 1) \) and therefore Lemma 1 parts i) and ii) yield that
\[
E \| Y_n (\mu_n, T_n) \|_\infty^2 \leq E \tau^2 < \infty.
\]
We also have that \( |Z (\mu_n, T_n)| \leq \tau (n + 1) \) and therefore a similar bound to (10) applies, thus concluding the proof.

6.1.3. Stability of the Dynamical System and Final Convergence Argument. The dynamical system of interest, namely \( \dot{\theta} (t) = g (\theta (t)) \) takes the form
\[
\dot{\mu} (t)' = f (\mu (t), T (t)) = \frac{1}{T (t)} (\mu (t)' R - (\mu (t)' Re) \mu' (t))
\]
\[
\dot{T} (t) = h (\mu (t), T (t)) = \mu (t)' Re - T (t).
\]
In the proof of Theorem 5.2.1 in [17] it is shown that any converging subsequence of the suitably normalized iterates converges to a function \( \theta (\cdot, \omega) \) which is a solution to the ODE \( \dot{\theta} (t) = g (\theta (t)) \). We only need to show that these solutions converge as \( t \rightarrow \infty \) to \((\mu_*, 1/(1 - \lambda_*)) \).

We will actually show that all solutions of a suitably reduced ODE (starting from points in \( H \)) converge to the quasi-stationary distribution and then we will show how to map the solutions \( \theta (\cdot, \omega) \) for the full system (11) into solutions to the reduced ODE.

We can invoke then Theorem 5.2.1 in [17] after arguing that \( \{(\mu_n, T_n)\} \) visits a compact set of \( H \times [1, \infty) \) infinitely often almost sure; actually we will show that \( \{(\mu_n, T_n)\} \) will stay inside a compact set eventually. We now define the reduced ODE as follows
\[
\dot{v} (t) = (v (t)' R - (v (t)' Re) v' (t))'
\]
\[
v (0) = \mu_0 \in H.
\]
Note that the gradient of the vector field in (12) is continuously differentiable in \( H \), therefore \( v (\cdot) \) has a unique solution given \( v (0) \in H \).

Suppose that \( T_0 \geq 1 \) and let \((\mu (\cdot), T (\cdot))\) be a solution to (11) obtained by the subsequence procedure in Theorem 5.2.1 from [17], then define \( \Gamma (t) = \int_0^t (1/T (s)) ds \). It follows by formal differentiation that \( v (t) = \mu (\Gamma^{-1} (t)) \) solves (12). The following result ensures regularity properties of \( \Gamma (\cdot) \).

\[
\text{Lemma 5.} \quad \Gamma (t) > 0, \Gamma (\cdot) \text{ is strictly increasing, and} \quad \Gamma (t) \rightarrow \infty \text{ as} \quad t \rightarrow \infty.
\]
Proof of Lemma 5. Clearly, we have that $T(s) \geq 1$ so $\Gamma(\cdot)$ is strictly increasing and non-negative. Now, suppose that $\Gamma(\infty) < \infty$, this implies that

$$T(t) = T_0 \exp \left( \int_0^t E\left( \tau \mid X_0 \sim \mu(s) \right) ds - t \right) \leq T_0 \exp \left( E(\tau) \int_0^t \frac{ds}{T(s)} - t \right) \leq T_0 \exp \left( E(\tau) \Gamma(\infty) \right) \in (0, \infty).$$

This bound would imply that there exists $\delta > 0$ so that $1/T(t) > \delta$ and obtaining a contradiction to the assumption that $\Gamma(\infty) < \infty$. Thus we must have that $\Gamma(\infty) = \infty$. \hfill \Box

Lemma 6. Any solution to the reduced ODE in (12) (regardless of $v(0) \in H$) converges to the quasi-stationary distribution $\mu$ as $t \to \infty$.

Proof of Lemma 6. By applying inner product with the vector $e$ we can see that the entire trajectory stays in $H$. We let $R = (I - Q)^{-1} = \sum_{k=0}^{\infty} Q^k$ which has only non-negative entries (actually, the entries are strictly positive if $Q$ is irreducible). By Duhamel’s principle all the solutions $v(\cdot)$ can be represented by

$$(13) \quad v(t) = v(0) \exp \left( Rt - \int_0^t (v(s)' Re) ds \right).$$

Because $R$ has non-negative entries we have that $\exp(Rt)$ has non-negative entries and thus $v(t) \geq 0$ (in fact the entries are strictly positive if $Q$ is irreducible). Rearranging equation (13) we arrive at

$$(14) \quad v(t) \exp \left( \int_0^t (v(s)' Re) ds - \lambda_R t \right) = v(0) \exp (Rt - \lambda_R t),$$

where $\lambda_R$ denotes the principal eigenvalue of $R$ (which is positive). The matrix $\exp(R)/\exp(\lambda_R)$ is a matrix with strictly positive entries and it has spectral radius equal to one. By the Perron-Frobenius theorem, see [14], we have that there exists a strictly positive vector $w$ such that

$$v(0) \exp(Rn - \lambda_R n) \to w'$$

as $n \to \infty$, where $w$ is a principal eigenvector of $R - \lambda_R I$. The convergence holds also along real numbers $t$ (not only natural numbers $n$) by virtue of a continuity argument noting that $(R - \lambda_R I)/m$ has the same eigenvectors regardless of the value of $m > 0$. Now, take the inner product with $e$ in both sides of equation (14) to obtain (because $v(t) \in H$) that

$$\exp \left( \int_0^t (v(s)' Re) ds - \lambda_R t \right) \to \gamma := w' e \in (0, \infty).$$

Finally, rewrite (13) as

$$v(t) = v(0) \exp (Rt - \lambda_R t) \exp \left( - \left( \int_0^t (v(s)' Re) ds - \lambda_R t \right) \right) \to \frac{w'}{\gamma}.$$

The fact that finishes the proof of the lemma is that the Perron-Frobenius eigenvectors of $R$ and $Q$ are identical, so we see that $v(t)$ converges to the quasi-stationary distribution as $t \to \infty$. \hfill \Box

Now we are ready to conclude the consistency portion of Theorem 1 by invoking Theorem 5.2.1 in [17] together with the following proposition.

Proposition 3. Any subsequence solution (obtained as in Theorem 5.2.1 in [17]) of the system (11) satisfies that $\mu_0 \in H$ and $T_0 \geq 1$ and it is such that $\mu(t) \to \mu_*$ and

$$T(t) \to 1/(1 - \lambda) = E(\tau \mid X_0 \sim \mu_*),$$

as $t \to \infty$. The sequence $\{\mu_n, T_n\}$ stays in a compact set of the attractor domain $H \times [1, \infty)$ eventually. Therefore, $\mu_n \to \mu_*$ and $T_n \to 1/(1 - \lambda)$ with probability one.

Proof of Proposition 3. We have that $\mu_n \in H$ and $T_n \geq 1$ because $T_n$ is the average of the $\tau^{(n)}$’s which are greater than one, so the subsequence procedure in Theorem 5.2.1 in [17] produces trajectories that lie in $H \times [1, \infty)$ and which are solutions to (11). Now, we have noted that $\Gamma(\cdot)$ is non-negative and strictly
increasing, according to Lemma 5 and thus Lemma 6 implies that \( v(t) = \mu(\Gamma^{-1}(t)) \to \mu_* \). Moreover, Lemma 5 indicates that \( \Gamma^{-1}(t) \to \infty \), therefore we have that \( \mu(t) \to \mu_* \) as \( t \to \infty \). Now, observe that

\[
T(t) = \int_t^\infty \frac{E(\tau | X_0 \sim \mu(s)) \exp(s) \, ds + T_0}{\exp(t)}.
\]

Because \( E(\tau | X_0 \sim \mu(t)) \to E(\tau | X_0 \sim \mu_*) \), we can use L'Hopital's rule and conclude that

\[
\lim_{t \to \infty} T(t) = \lim_{t \to \infty} \frac{E(\tau | X_0 \sim \mu(t)) \exp(t)}{\exp(t)} = E(\tau | X_0 \sim \mu_*) = 1/(1 - \lambda_*).
\]

The fact that \( \{(\mu'_n, T_n)\} \) stays in a compact set of \( H \times [1, \infty) \) follows from Lemma 1 part iv).

**6.2. Proof of Theorem 1: CLT.** In order to prove the CLT portion of Theorem 1 we shall invoke Theorem 10.2.1 in [17]; this requires verifying the following conditions.

1. The sequence \( \{W_n I(\|\theta_n - \theta_*\| \leq \delta)\} \) is uniformly integrable. (This follows immediately because due to Lemma 4 we have that \( W_n \) is \( L_2 \) bounded.)

2. \( \theta_* \) is an isolated stable point of the ODE. (This follows from Perron-Frobenius theorem and the analysis in Proposition 5.)

3. The following expansion is valid

\[
g_n(\theta) = g_n(\theta_*) + (Dg_n)(\theta_*) (\theta - \theta_*) + o(\|\theta - \theta_*\|),
\]

where the error term is uniform in \( n \). (This estimate will be elaborated in Lemma 7.)

4. We must have that

\[
\lim_{n,m \to \infty} \frac{1}{m^{1/2}} \sum_{k=n}^{n+m-1} (Dg_k)(\theta_*) = 0
\]

uniformly over \( t \) in compact sets. (See Lemma 9)

5. There exists a matrix \( A \) such that

\[
\lim_{n,m \to \infty} \sum_{k=n}^{n+m-1} ((Dg_k)(\theta_*) - A) = 0.
\]

(Let \( A = (Dg)(\theta_*) \), then this condition will hold true by Lemma 10 which shows that \( (Dg_n)(\theta_*) \to (Dg)(\theta_*) \).)

6. The matrix \( A \) must also be such that \( A + I/2 \) is Hurwitz (i.e. all its eigenvalues have negative real part). (This corresponds precisely to condition 5 and it will be established in Proposition 4)

7. The sequence \( \{\theta_n - \theta_*\}/\varepsilon_n^{1/2} \) is tight. (See Lemma 12)

8. Define \( \delta M_n = W_n - g_n(\theta_n) \), then there exists \( p > 0 \) such that

\[
\sup_n E\|\delta M_n\|^{2+p} < \infty,
\]

and a non-negative matrix \( \Sigma \) such that

\[
E(\delta M_n(\delta M_n)'|\mathcal{F}_n) \to \Sigma
\]

in probability as \( n \to \infty \). (This is established in Lemma 13)

**6.2.1. Analysis of the Jacobian: Conditions 3, 4, and 5.**

**Lemma 7.** We have that

\[
\lim_{\|\theta - \theta_*\|_\infty \to 0, n \geq 1} \sup_{n \geq 1} \left| g_n(\theta) - g_n(\theta_*) - (Dg_n)(\theta_*) (\theta - \theta_*) \right| = 0.
\]

**Proof of Lemma 7.** We consider the analysis only for \( f_n(\mu', T) \) because \( h \) is a simpler quantity and does not depend on \( n \). The analysis follows as an application of the representation derived in Lemma 2. It is easy to justify the interchange of differentiation and integration in the representation given in Lemma 2 because the integrand consists of products of a second degree polynomial in \( \mu \), the exponential factor \( \exp(-uT) \) on the region of interest which is \( T \geq 1 \), and the term including \( v(x, s) \in (0, 1] \). Thus the second derivatives of \( f_n \) will be bounded uniformly in \( n \) around a neighborhood of the stationary point \( \theta_* \). \( \square \)
Next we turn to condition 4., but first we have an auxiliary result.

Lemma 8. It follows that \( g_n (\theta_*) = O \left( \frac{1}{n} \right) \).

Proof of Lemma 8. Note that \( h (\theta_*) = 0 \), so we only focus on \( f_n (\theta_*) \). On the other hand, we observed in the proof of Lemma 8 that \( |f_n (\mu', T) (x) - f (\mu', T) (x)| \leq E (\bar{v}^2) / (n + 1) \), but (with \( R = (I - Q)^{-1} \)),

\[
 f (\mu'_*, T) = \left( \frac{1}{T} (\mu'_* R - (\mu'_* R e) \mu'_*) \right) = \frac{1}{T} \left( \frac{1}{1 - x} \mu'_* - \frac{1}{1 - x} \mu'_* \right) = 0.
\]

Hence the Lemma 8 follows.\( \square \)

Lemma 9. We have that

\[
 \lim_{n,m \to \infty} \frac{1}{m^{1/2}} \sum_{k=n}^{n+mt-1} \| (Dg_k) (\theta_*) \|_\infty \leq \frac{c}{m^{1/2}} \log \left( 1 + \frac{mt}{n} \right).
\]

Changing the variables via the transformation \( u = mt/n \) we have that

\[
 \frac{1}{m^{1/2}} \sum_{k=n}^{n+mt-1} \| (Dg_k) (\theta_*) \|_\infty \leq \frac{cm^{1/2}}{n} \times \sup_{u \geq 0} \frac{\log (1 + u)}{u^{1/2}}.
\]

We have that \( \sup_{u \geq 0} \log (1 + u) / u^{1/2} < \infty \) and therefore, we can send \( n \to \infty \) in the right hand side to conclude the statement of Lemma 9.\( \square \)

Lemma 10. \( (Dg_n) (\theta_*) \to (Dg) (\theta_*) \)

as \( n \to \infty \).

Proof of Lemma 10. Once again, it suffices to concentrate on \( f_n \). Because of Lemma 7 we know that

\[
 f_n (\theta) = f_n (\theta_*) + (Df_n) (\theta_*) (\theta - \theta_*) + o (\theta - \theta_*).
\]

Taking the limit as \( n \to \infty \), we arrive at

\[
 f (\theta) = \lim_{n \to \infty} (Df_n) (\theta_*) (\theta - \theta_*) + o (\theta - \theta_*).
\]

Expanding the left hand side, we have that

\[
 f (\theta) = (Df) (\theta_*) (\theta - \theta_*) + o (\theta - \theta_*).
\]

Matching these terms and noting that \( \theta - \theta_* \) can have any direction, we conclude the result of the lemma.\( \square \)

6.2.2. The Hurwitz Property: Condition 6.

Proposition 4. Let \( A = (Dg) (\theta_*) \), then \( A + I/2 \) is Hurwitz assuming that the eigenvalues of \( Q \) satisfy condition 5.

Proof of Proposition 4. Recall that \( g (\mu', T) = (f' (\mu', T), h (\mu', T)) \) and expressions (8) and (9). Letting \( B = R' = (I - Q')^{-1} \) we have that the Jacobians are given by (using \( D_\mu \) and \( D_T \) to denote the derivatives with respect to \( \mu \) and \( T \) respectively),

\[
 D_\mu f (\mu', T) = \frac{1}{T} [B - (\mu' B e) I - \mu e' B],
\]

\[
 D_T f (\mu', T) = -\frac{1}{T^2} [B \mu - (\mu' B e) \mu],
\]

\[
 D_\mu h (\mu', T) = e' B, \quad D_T h (\mu', T) = -1.
\]
We consider the stationary point and note that $\mu'_t B e = T_*$. Then, define

$$ J := D_{\mu} f (\mu'_*, T_*) = \frac{1}{T_*} (B - T_* I - \mu_* e' B) . $$

Also, note that $D_T f (\mu'_*, T_*) = 0$. We now establish a one-to-one correspondence between the eigenvectors of $J$ and the eigenvectors of $B$. The overall Jacobian in block form would take the form

$$ A = \begin{bmatrix} J & 0 \\ e' B & -1 \end{bmatrix} . $$

This matrix has the same eigenvalues as $J$, with the addition of the eigenvalue $-1$ which has no effect on the Hurwitz property. Hence we only need to ensure that $J + I/2$ is Hurwitz.

Let $y$ be any vector such that $J y = \lambda_J y$ and such that $y$ is linearly independent of $\mu_*$. Note that $J y = \lambda_J y$ is equivalent to the equation

$$ B y = T_* \lambda_J y + T_* y + (e' B y) \mu_* . $$

and therefore if we let $x = y + r \mu_*$, for some $r$ to be characterized momentarily, we have that

$$ B x = B y + r T_* \mu_* = T_*(\lambda_J + 1)y + (e' B y + r T_*) \mu_* . $$

So, the value of $r$ that would make $x$ an eigenvector of $B$ is such that $r T_* \lambda_J = e' B y$. Since $T_* > 0$ the existence of $r$ is guaranteed if $\lambda_J \neq 0$ and the corresponding eigenvalue for $B$ would be $\lambda_B = T_* (1 + \lambda_J)$. On the other hand, if $y$ is a multiple of $\mu_*$, its eigenvalue for the matrix $J$ equals $-1$ (the eigenvalue for $B$ is, of course, $T_*$). In Lemma 11 below we will argue that $\lambda_J$ cannot be zero, so the argument just given shows that every eigenvector of $J$ is an eigenvector of $B$.

Conversely, given any vector $z$, such that $B z = \lambda_B z$, we can define $u = z + r \mu_*$. If choose $r = (T_* + \lambda_B e' z) / (T_* - \lambda_B)$ then we conclude that

$$ J u = (\lambda_B / T_* - 1) u . $$

This selection of $r$ is valid if $z$ is not the principal right eigenvector of $B$ (i.e. in case $z$ is not $\mu_*$). In case we select $z = \mu_*$, then trivially $J z = -z$.

Consequently, we conclude that there is a one-to-one correspondence between the eigenvectors (and eigenvalues) of $J$ and $B$, and the relationship between the eigenvalues is as follows

$$ \begin{align*} 
\lambda_J &= \frac{\lambda_B}{T_*} - 1, \text{ for } \lambda_B \neq T_* \text{ or } \lambda_J \neq 0, \\
\lambda_J &= -1 \text{ if } \lambda_B = T_* . 
\end{align*} $$

Therefore, in order to ensure that $J + I/2$ is Hurwitz, we must have that

$$ \text{Re} (\lambda_B) < T_*/2 $$

for all $\lambda_B \neq T_*$, which is precisely condition 5. □

We finish the analysis of the Hurwitz condition, with the following result invoked in the previous proof.

**Lemma 11.** We have that $\lambda_J \neq 0$.

**Proof of Lemma 11.** Assume that $y$ is such that $J y = 0$. This implies that

$$ B y = T_* y + \mu_* (e' B y) = T_* y + (\mu_* e') B y . $$

Therefore

$$ (I - (\mu_* e')) B y = T_* y . $$

We recognize that $\tilde{P} = (I - (\mu_* e'))$ is a (non-orthogonal) projection in the sense that $\tilde{P}^2 = \tilde{P}$. Also we have that $\tilde{P} \mu_* = 0$ and $e' \tilde{P} = 0$. This means that $T_*$ is an eigenvalue of $\tilde{P} B$, which in turn implies that there would exist a left eigenvector $x$ such that

$$ x' \tilde{P} B = T_* x' , $$

or

$$ x' \tilde{P} = T_* x' B^{-1} . $$

Now let $x' \tilde{P} = z'$ and consider all possible solutions $w$ such that $w' \tilde{P} = z'$ which must be written as the sum of an element of the null space and a particular solution. Observe, because $\tilde{P}^2 = \tilde{P}$, that $z' \tilde{P} = z'$ is
a particular solution and therefore any solution \( x \) (i.e., any eigenvector corresponding to \( T_s \) for \( \bar{P}B \)) must take the form \( x = ce + z \) for some constant \( c \). Observe from (17), multiplying by \( \mu_* \) from the right, that
\[
0 = T_s x' B^{-1} \mu_* = T_s x' (I - Q') \mu_* = x' \mu_* .
\]
Therefore, we have that \( c = 0 \) because \( x' \mu_* = 0 \) and
\[
0 = x' \mu_* = c + z' \mu_* = c + 0.
\]
Consequently, \( x' = z' = T_s x B^{-1} \), which implies that \( x' B = T_s x' \), therefore concluding that \( x \) is the principal left-eigenvector of \( B \). Consequently, \( x \) must have strictly positive entries and, in turn, we must have that \( x' \mu_* > 0 \) thus arriving at a contradiction. So, there is no eigenvalue \( T_s \) for the matrix \( PB \) and thus \( \lambda_s = 0 \) is not possible. \( \square \)

6.2.3. Tightness: Condition 7.

**Lemma 12.** The sequence \( \{ (\theta_n - \theta_s) / \varepsilon_n^{1/2} \} \) is tight.

**Proof of Lemma** \[ \square \] We use the local techniques discussed in Section 10.5.2 of \[ 17 \] and apply them as in the proof of Theorem 10.4.1 in \[ 17 \], albeit with some modifications. We shall use the Lyapunov function \( V(\theta) = ||\theta - \theta_s||_2^2 \). However, we now have to deal with the gradient of \( g_n \) as opposed to the gradient of \( g \) as in the proof of Theorem 10.4.1 in \[ 17 \]. We shall control the changes in \( g_n \) by expanding around the stationary point. We have that
\[
E(V(\theta_{n+1})|\mathcal{F}_n) - V(\theta_n)
\]
\[= 2\varepsilon_n (\theta_n - \theta_s)' g_n(\theta_n) + O(\varepsilon_n^2)\]
\[= 2\varepsilon_n (\theta_n - \theta_s)' g_n(\theta_s) + 2\varepsilon_n (\theta_n - \theta_s)' A_n(\theta_n - \theta_s)' + \varepsilon_n o(||\theta_n - \theta_s||_2^2) + O(\varepsilon_n^2),\]
where the first equality uses an idea similar to that of the proof of Lemma 4 to arrive at the error term \( O(\varepsilon_n^2) \), and the second inequality is just an expansion of \( g_n \) around \( \theta_s \) followed by an application of Lemma 7. Since \( 2A_n \) has eigenvalues with negative real part less than \(-1\) (i.e., \( A_n + I/2 \) is Hurwitz) for \( n \) large enough, we conclude that there exists \( \delta > 0 \) such that for all \( n \) sufficiently large
\[
(\theta_n - \theta_s)' A_n(\theta_n - \theta_s)' < -(1 + 2\delta) ||\theta_n - \theta_s||_2^2 .
\]
Moreover, because \( g_n(\theta_s) = O(1/n) \) due to Lemma 8, we have that
\[
2\varepsilon_n (\theta_n - \theta_s)' g_n(\theta_s) \leq O \left( \varepsilon_n^2 (\theta_n - \theta_s)' \right)
\]
and
\[
o(||\theta_n - \theta_s||_2^2) \leq \delta V(\theta_n).
\]
We then conclude that
\[
E(V(\theta_{n+1})|\mathcal{F}_n) - V(\theta_n) \leq -\varepsilon_n (1 + \delta) V(\theta_n) + O \left( \varepsilon_n^2 \right).
\]
The rest of the proof now can be concluded as in Theorem 10.4.1 from \[ 17 \]. \( \square \)

6.2.4. Quadratic Variation of the Martingales: Condition 8.

**Lemma 13.** Let \( \delta M_n = W_n - g_n(\theta_n), \) then
\[(18) \quad \sup_{n \geq 0} E ||\delta M_n||_2^4 < \infty .\]
Moreover,
\[(19) \quad E(\delta M_n \delta M_n' | \mathcal{F}_n) \to \Sigma \]
for some matrix \( \Sigma \) in probability.

**Proof of Lemma** \[ \square \] We use the notation \( E_n(\cdot) \) for \( E(\cdot | \mathcal{F}_n) \),
\[
||\delta M_n||_2^4 \leq 2(||Y_n(\theta_n) - E_n Y_n(\theta_n)||_2^4 + ||Z_n(\theta_n) - E_n Z_n(\theta_n)||_2^4)
\]
\[ \leq 16 \left( ||Y_n(\theta_n)||_2^4 + ||E Y_n(\theta_n)||_2^4 + ||Z_n(\theta_n)||_2^4 + ||E_n Z_n(\theta_n)||_2^4 \right) .
\]
An argument similar to Lemma 3 yields that \( \|Y_n(\theta_n)\|^4 \leq \bar{\tau}^4 \) and \( |Z_n(\theta_n)|^4 \leq \bar{\tau}^4 \) (stochastically) and therefore, by Lemma 1, we conclude bound (18). To establish (19) let us write

\[
\begin{align*}
    \delta M_n \delta M_n' &= \left[ (Y_n(\theta_n) - f_n(\theta_n)) (Y_n(\theta_n) - f_n(\theta_n))' \right] \left[ (Z_n(\theta_n) - h(\theta_n)) (Z_n(\theta_n) - h(\theta_n))' \right].
\end{align*}
\]

By Lemma 8 we have that \( f_n(\theta_n) \to 0 \) and \( h(\theta_n) \to 0 \). Note that the distribution of \( Z(\theta) \) and \( Y(\theta) \) can be written in a way that is continuous in \( \theta \) (as a mixture of the initial distribution), therefore \( Z(\theta_n) \Rightarrow Z(\theta_\ast) \) and \( Y(\theta_n) \Rightarrow Y(\theta_\ast) \); consequently we also have that \( Y_n(\theta_n) \Rightarrow Y(\theta_\ast) \). We observe that each entry of the matrix is stochastically dominated by \( 2\bar{\tau} \) and thus we can apply Lemma 1 to conclude uniform integrability (since \( E(\bar{\tau}^2) < \infty \)), thereby obtaining that

\[
E_n(\delta M_n \delta M_n') \rightarrow \Sigma := E \left[ \begin{bmatrix} Y(\theta_\ast) \\ Z(\theta_\ast) \end{bmatrix} \begin{bmatrix} Y(\theta_\ast)' & Z(\theta_\ast) \end{bmatrix} \right].
\]

In turn, following the expression in p. 332 from [17], we have that the asymptotic covariance equals

\[
V_0 = \int_0^\infty \exp((J + I/2) t) \Sigma \exp((J' + I/2) t) \, dt.
\]

\( \boxdot \)

7. PROOF OF PROPOSITION 4

Proof of Proposition 4: The proof is almost identical to that of Theorem 1. In fact, the analysis is somewhat simpler because there is no denominator and so we just need to analyze the reduced ODE in (12). The proof of tightness also follows similar steps as the argument given in Theorem 10.4.1 in [17], which distinguishes the cases \( \varepsilon_n = 1/n \) and the case \( \varepsilon_n = n^{-\alpha} \) for \( \alpha \in (1/2, 1) \) as we do here.

Now, recall that \( J \) be the Jacobian of the vector field obtained in (15), evaluated at the unique stability point \( \mu_\ast \). We need to ensure that \( J \) is Hurwitz (i.e. all the eigenvalues have strictly negative real part). This is in contrast to requiring that \( J + I/2 \) is Hurwitz - which is a stronger condition because then one needs that all the eigenvalues have real part less than \(-1/2\), which leads to (5). Instead, requiring that \( J \) be Hurwitz is equivalent to the condition that \( Re(\bar{\lambda}) < \lambda \), for all non-principal eigenvalue \( \bar{\lambda} \) of the matrix \( Q \), which is automatic by Perron-Frobenius Theorem, [14]. Hence we can conclude the result by invoking Theorem 10.2.1 from [17].

The asymptotic covariance matrix in this case takes the form

\[
V_1 = \int_0^\infty \exp(Jt) \Sigma_0 \exp(J't) \, dt,
\]

where

\[
\Sigma_0 = E \left( \bar{Y}(\mu_\ast) \bar{Y}(\mu_\ast)' \right),
\]

with \( \bar{Y}(\mu) = \sum_{k=0}^{\lfloor \frac{t}{\tau} \rfloor - 1} (I(X_k = \cdot | X_0 \sim \mu) - \mu(\cdot)) \).

We now are ready to discuss the proof of Theorem 2.

Proof of Theorem 2: We shall verify the conditions in Theorem 2 of [23]. First, define \( \tilde{f}(\mu) = f(\mu', 1) \), recall that \( f(\cdot) \) was introduced in (8) and it coincides with \( f(\mu', 1) = E\bar{Y}(\mu') \). We must first verify that,

- There exists a function \( L \) a (globally) Lipschitz continuous function \( V(\cdot) \) such that \( L(\mu_\ast) = 0 \), and for some positive definite matrix \( G' \),

\[
DL(\mu)G\tilde{f}(\mu') < 0,
\]

for \( \mu \neq \mu_\ast \), there exists \( \varepsilon, \delta > 0 \) such that

\[
DL(\mu)G\tilde{f}(\mu') \leq -\delta V(\mu)
\]

if \( ||\mu - \mu_\ast|| \leq \varepsilon \), and \( L(\mu - \mu_\ast) \geq \delta ||\mu - \mu_\ast||^2 \) for some \( \delta > 0 \).

This condition is satisfied if we construct \( L(\cdot) \) by noting that \( \mu_\ast \) is the unique root of \( \tilde{f}(\mu_\ast) = 0 \) and we have established in Lemma 4 that \( \tilde{J} := (D\tilde{f})(\mu_\ast) = JT_\ast \) is Hurwitz (in fact \( J + I/2 = J/T_\ast + I/2 \) is Hurwitz), therefore

\[
\tilde{f}(\mu) = \tilde{J}(\mu - \mu_\ast) + O(||\mu - \mu_\ast||^2).
\]
In this section we provide the proof of consistency for the estimator of and its quadratic variation, the verification is completely parallel to that of Proposition 4. ε satisfied for are two more conditions to check for the application of Theorem 2 of [23], the fourth condition is triviallyhood of µ is Hurwitz. There are two more conditions to check for the application of Theorem 2 of [23], the fourth condition is trivially satisfied for εn = n−α with α ∈ (1/2, 1) and the third condition involves the martingale difference process and its quadratic variation, the verification is completely parallel to that of Proposition 4.

8. Proof of Proposition 2

In this section we provide the proof of consistency for the estimator of

\[ \hat{V}_1 = J^{-1} \Sigma_0 J^{-1}. \]

Expression (25) follows from Theorem 11.1.1 in [17].

Proof of Proposition 2 We write

\[ \varepsilon_n^{-1} (\bar{x}_n - v_n) (\bar{x}_n - v_n)^\prime = \varepsilon_n^{-1} (\bar{x}_n - \mu_*) (\bar{x}_n - \mu_*)^\prime + \varepsilon_n^{-1} (v_n - \mu_*) (v_n - \mu_*)^\prime - \varepsilon_n^{-1} (\bar{x}_n - \mu_*) (v_n - \mu_*)^\prime - \varepsilon_n^{-1} (\bar{x}_n - \mu_*) (v_n - \mu_*)^\prime. \]

Note that

\[ \varepsilon_n^{-1/2} (\bar{x}_n - \mu_*) (\varepsilon_n n)^{-1/2} n^{1/2} (v_n - \mu_*)^\prime \to 0 \]

in probability as \( n \to \infty \) because of Proposition 1 and Theorem 2 since \( \varepsilon_n n \to \infty \) as \( n \to \infty \). A similar argument applies to all the terms in (26) involving \( (v_n - \mu_*) \); so it suffices to study the limit of

\[ \frac{1}{N_n} \sum_{n_k=1}^{N_n} \varepsilon_n^{-1} (\bar{x}_{n_k} - \mu_*) (\bar{x}_{n_k} - \mu_*)^\prime \]

as \( n \to \infty \). The rest of the calculation is similar to analysis of the asymptotic covariance in Theorem 11.3.1 of [17]. The idea is that the sequence \( \{ (\bar{x}_{n_k} - \mu_*) \varepsilon_n^{-1/2} \} \) is weakly dependent and each term is asymptotically normal (as \( k \to \infty \)) with variance \( \hat{V}_1 \).

9.

Acknowledgement: We acknowledge support from the NSF through the grants CMMI-1069064 and DMS-1320550.

References


