Adv. Appl. Prob. **51,** 136–167 (2019)
*doi:*10.1017/apr.2019.6
© *Applied Probability Trust* 2019

EXTENDED LAPLACE PRINCIPLE FOR EMPIRICAL MEASURES OF A MARKOV CHAIN

STEPHAN ECKSTEIN[*] *University of Konstanz*

Abstract

We consider discrete-time Markov chains with Polish state space. The large deviations principle for empirical measures of a Markov chain can equivalently be stated in Laplace principle form, which builds on the convex dual pair of relative entropy (or Kullback–Leibler divergence) and cumulant generating functional $f \mapsto \ln \int \exp(f)$. Following the approach by Lacker (2016) in the independent and identically distributed case, we generalize the Laplace principle to a greater class of convex dual pairs. We present in depth one application arising from this extension, which includes large deviation results and a weak law of large numbers for certain robust Markov chains—similar to Markov set chains—where we model robustness via the first Wasserstein distance. The setting and proof of the extended Laplace principle are based on the weak convergence approach to large deviations by Dupuis and Ellis (2011).

Keywords: Large deviations; Markov chain; convex duality; distributional uncertainty

2010 Mathematics Subject Classification: Primary 60J05
Secondary 60F10

1. Introduction

Throughout the paper (E, d) denotes a Polish space, $\mathcal{P}(E)$ denotes the space of Borel probability measures on E endowed with the topology of weak convergence, and $C_b(E)$ denotes the space of continuous and bounded functions mapping E into $\mathbb{R}$. Let a Markov chain with state space E be given by its initial distribution $\pi_0 \in \mathcal{P}(E)$ and Borel measurable transition kernel $\pi : E \to \mathcal{P}(E)$, and denote by $\pi_n \in \mathcal{P}(E^n)$ the joint distribution of the first n steps of the Markov chain. Define the empirical measure map $L_n \colon E^n \to \mathcal{P}(E)$ by

$$L_n(x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i},$$

and recall the relative entropy $R \colon \mathcal{P}(E) \times \mathcal{P}(E) \to [0, \infty]$ given by

$$R(\nu, \mu) = \int_E \log\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right) \mathrm{d}\nu \quad \text{if } \nu \ll \mu, \qquad R(\nu, \mu) = \infty \quad \text{else.}$$

The main goal of this paper is to generalize the large deviations result for empirical measures of a Markov chain in its Laplace principle form. Under suitable assumptions on the Markov chain, the usual Laplace principle for empirical measures of a Markov chain states that, for all

Received 11 September 2017; revision received 23 August 2018.
The supplementary material for this article can be found at http://doi.org/10.1017/apr.2019.6.

[*] Postal address: Department of Mathematics, University of Konstanz, 78464 Konstanz, Germany.
Email address: stephan.eckstein@uni-konstanz.de

$F \in C_b(\mathcal{P}(E))$,

$$\lim_{n\to\infty} \frac{1}{n} \ln \int_{E^n} \exp(nF \circ L_n)\, \mathrm{d}\pi_n = \sup_{\nu\in\mathcal{P}(E)} (F(\nu) - I(\nu)). \tag{1.1}$$

Here, $I\colon \mathcal{P}(E) \to [0, \infty]$ is the rate function, given in the setting of [20, Chapter 8] by

$$I(\nu) = \inf_{q\colon \nu q=\nu} \int_E R(q(x), \pi(x))\nu(\mathrm{d}x),$$

where the infimum is over all stochastic kernels q on E that have ν as an invariant measure. In this paper a stochastic kernel q on E is a Borel measurable mapping $q\colon E \to \mathcal{P}(E)$, and $\nu q \in \mathcal{P}(E)$ is defined by $\nu q(A) := \int_E q(x, A)\nu(\mathrm{d}x)$ for $\nu \in \mathcal{P}(E)$, where we write $q(x, A) = q(x)(A)$ for $x \in E$ and Borel sets $A \subseteq E$. The Laplace principle (1.1)—in the mentioned setting of [20]—is equivalent to the more commonly used form of the large deviations result for empirical measures of a Markov chain, which states that, for all Borel sets $A \subseteq \mathcal{P}(E)$,

$$-\inf_{\nu\in\mathring{A}} I(\nu) \le \liminf \frac{1}{n} \ln \pi_n(L_n \in \mathring{A}) \le \limsup \frac{1}{n} \ln \pi_n(L_n \in \bar{A}) \le -\inf_{\nu\in\bar{A}} I(\nu),$$

where $\mathring{A}$ denotes the interior and $\bar{A}$ the closure of A. Large deviation probabilities of Markov chains have been studied in a variety of settings and under different assumptions; see e.g. [13], [16], [17], [18], [30], and [36].

The way we generalize the Laplace principle is by using the fact that both sides of the Laplace principle (1.1) can be stated solely in terms of relative entropy, its chain rule, and its convex dual pair. Equation (1.1) can therefore be formulated analogously for functionals resembling the relative entropy, in the sense that these functionals have to satisfy the same type of chain rule and duality. The kind of convex duality referred to is Fenchel Moreau duality, which is often studied in the context of convex risk measures; see, for example, [1], [4], [12], and [33].

The original idea for extensions of Laplace principles of this form is due to Lacker [34], who pursued this in the context of independent and identically distributed (i.i.d.) sequences of random variables instead of Markov chains. The initial goal was to provide a setting to study more than just exponential tail behavior of random variables, as is given by large deviations theory. The extension of Sanov's theorem he proved [34, Theorem 3.1] can be used to derive many interesting results, such as polynomial large deviation upper bounds, robust large deviation bounds, robust laws of large numbers, asymptotics of optimal transport problems, and more, while several possibilities remain unexplored.

In this paper the same type of extension for Markov chains is obtained. To this end, we work in a setting similar to that of [20, Chapter 8]. In particular, the results from [20, Chapter 8] are a special case of Theorem 1.1. To showcase the potential implications of Theorem 1.1, we focus on one broad application related to robust Markov chains, summarized in Theorems 1.2 and 1.3.

1.1. Main results

Let $\beta\colon \mathcal{P}(E) \times \mathcal{P}(E) \to (-\infty, \infty]$ be a Borel measurable function which is bounded from below and satisfies $\beta(\nu, \nu) = 0$ for all $\nu \in \mathcal{P}(E)$. One may think of $\beta(\cdot, \cdot) = R(\cdot, \cdot)$. To state the chain rule, we introduce the following notation for the decomposition of an n-dimensional measure $\nu \in \mathcal{P}(E^n)$ into kernels $\nu_{i,i+1}\colon E^i \to \mathcal{P}(E)$ for $i = 1, \dots, n-1$ and $\nu_{0,1} \in \mathcal{P}(E)$:

$$\nu(\mathrm{d}x_1, \dots, \mathrm{d}x_n) = \nu_{0,1}(\mathrm{d}x_1) \prod_{i=1}^{n-1} \nu_{i,i+1}(x_1, \dots, x_i,\ \mathrm{d}x_{i+1}).$$

https://doi.org/10.1017/apr.2019.6 Published online by Cambridge University Press

For $\theta \in \mathcal{P}(E)$, define $\beta_n^\theta \colon \mathcal{P}(E^n) \to (-\infty, \infty]$ by

$$\beta_n^\theta(\nu) = \beta(\nu_{0,1}, \theta) + \int_{E^n} \sum_{i=1}^{n-1} \beta(\nu_{i,i+1}(x_1, \dots, x_i), \pi(x_i))\nu(\mathrm{d}x_1, \dots, \mathrm{d}x_n),$$

where in the case in which $\beta(\cdot, \cdot) = R(\cdot, \cdot)$ we obtain $\beta_n^{\pi_0}(\nu) = R(\nu, \pi_n)$ for $\nu \in \mathcal{P}(E^n)$ by the chain rule for relative entropy. Note that $\beta_n^{\cdot}(\cdot)$ is well defined as the term inside the integral is Borel measurable, for example, by [6, Proposition 7.27]. Define ρ_n^θ as the convex dual of β_n^θ by

$$\rho_n^\theta(f) = \sup_{\mu \in \mathcal{P}(E^n)} \bigg(\int_{E^n} f \, \mathrm{d}\mu - \beta_n^\theta(\mu) \bigg)$$

for Borel measurable functions $f \colon E^n \to \mathbb{R}$, where we adopt the convention $\infty - \infty := -\infty$. For $\beta(\cdot, \cdot) = R(\cdot, \cdot)$, we obtain $\rho_n^{\pi_0}(f) = \ln \int_{E^n} \exp(f) \, \mathrm{d}\pi_n$ by the Donsker–Varadhan variational formula for the relative entropy. In the above definitions, θ is a placeholder for variable initial distributions, which is required as a tool in the proof. For the actual statement, only $\beta_n^{\pi_0}$ and $\rho_n := \rho_n^{\pi_0}$ are needed. We write $\rho := \rho_1$ and $\rho^\theta := \rho_1^\theta$.

The assumptions for the main theorem are stated below. Assumption M is [20, Condition 8.4.1.], and Assumption T is a direct generalization of [20, Condition 8.2.2].

Assumption M. *The following conditions hold on the Markov chain.*

(M1) *Define the k-step transition kernel $\pi^{(k)}$ of the Markov chain recursively by $\pi^{(k)}(x, A) := \int_E \pi(y, A)\pi^{(k-1)}(x, \mathrm{d}y)$ for $x \in E$ and Borel sets $A \subseteq E$.*
Assume that there exist $l_0, n_0 \in \mathbb{N}$ such that, for all $x, y \in E$,

$$\sum_{i=l_0}^{\infty} \frac{1}{2^i} \pi^{(i)}(x) \ll \sum_{j=n_0}^{\infty} \frac{1}{2^j} \pi^{(j)}(y).$$

(M2) *π has an invariant measure, i.e. there exists $\mu^* \in \mathcal{P}(E)$ such that $\mu^*\pi = \mu^*$.*

Assumption B. *The following assumptions hold on β.*

(B1) *The mapping $\mathcal{P}(E) \times \mathcal{P}(E^2) \ni (\theta, \mu) \mapsto \beta_2^\theta(\mu)$ is convex.*

(B2) *The mapping $\mathcal{P}(E) \times \mathcal{P}(E^2) \ni (\theta, \mu) \mapsto \beta_2^\theta(\mu)$ is lower semi-continuous.*

(B3) *If ν is not absolutely continuous with respect to μ then $\beta(\nu, \mu) = \infty$.*

Assumption T. *At least one of the following assumptions must hold in order to guarantee the tightness of certain families of random variables.*

(T1) *There exists a Borel measurable function $U \colon E \to [0, \infty)$ such that the following conditions hold:*

(a) $\inf_{x \in E} (U(x) - \rho^{\pi(x)}(U)) > -\infty$*;*

(b) *$\{x \in E \colon U(x) - \rho^{\pi(x)}(U) \le M\}$ is a relatively compact subset of E for all $M \in \mathbb{R}$;*

(c) $\rho(U) < \infty$.

(T1') *E is compact.*

In the case in which $\beta(\cdot, \cdot) = R(\cdot, \cdot)$, we usually impose another Condition on π in the form of the Feller property, i.e. continuity of $x \mapsto \pi(x)$; see, e.g. [20, Condition 8.3.1]. Here, this is

https://doi.org/10.1017/apr.2019.6 Published online by Cambridge University Press

implicitly included in Condition (B2). Indeed, we can quickly check that, for (B2) to hold in the case in which $\beta(\cdot,\cdot)=R(\cdot,\cdot)$, the following is sufficient: if $\theta_n \xrightarrow{w} \theta \in \mathcal{P}(E)$ then $\theta_n \otimes \pi \xrightarrow{w} \theta \otimes \pi \in \mathcal{P}(E^2)$ has to hold as well. The Feller property implies this; see [20, Lemma 8.3.2].

The following extension of the Laplace principle for empirical measures of a Markov chain is the main result.

Theorem 1.1. *Define the rate function* $I\colon \mathcal{P}(E)\to(-\infty,\infty]$ *by*

$$I(\nu) := \inf_{q:\nu q=\nu} \int_E \beta(q(x),\pi(x))\nu(\mathrm{d}x) = \inf_{q:\nu q=\nu} \beta_2^{\nu}(\nu\otimes q). \tag{1.2}$$

Under Condition (B1), (B2), and Assumption T, the upper bound

$$\limsup_{n\to\infty} \frac{1}{n}\rho_n(nF\circ L_n) \le \sup_{\nu\in\mathcal{P}(E)} (F(\nu)-I(\nu))$$

holds for all upper semicontinuous and bounded functions $F\colon\mathcal{P}(E)\to\mathbb{R}$.

Under Condition (M1), (M2), (B1), and (B3), the lower bound

$$\liminf_{n\to\infty} \frac{1}{n}\rho_n(nF\circ L_n) \ge \sup_{\nu\in\mathcal{P}(E)} (F(\nu)-I(\nu))$$

holds for all $F\in C_b(\mathcal{P}(E))$.

Intuition, applicability, and difficulties in dealing with the above result are very similar to the i.i.d. case and are described in detail in the introduction of [34]. The main differences for Markov chains are conditions (B1) and (B2). To verify these conditions, we would ideally like to have a better expression for $\beta_2^{\cdot}(\cdot)$ than is given by the definition, which is often not trivial. In the applications of this paper the choices of β are convenient in this regard. Some of the applications pursued in the i.i.d. case, e.g. [34, Chapter 4 and 6] appear more difficult to obtain for Markov chains. A thorough analysis of the range of applications of Theorem 1.1 remains incomplete for now, as the goal of this work is to give a detailed account of one application of Theorem 1.1 to robust Markov chains.

The following corollary complements Theorem 1.1.

Corollary 1.1. (a) *If* $(\theta,\nu)\mapsto\beta_2^{\theta}(\nu)$ *is lower semicontinuous then I is lower semicontinuous. If* $(\theta,\nu)\mapsto\beta_2^{\theta}(\nu)$ *is convex then I is convex.*

(b) *If the Theorem 1.1 upper bound holds and, additionally, I has compact sub-level sets, then the upper bound extends to all functions* $F\colon\mathcal{P}(E)\to[-\infty,\infty)$ *which are upper semicontinuous and bounded from above.*

1.2. Applications to robust Markov chains

In this paper robustness broadly refers to uncertainty about the correct model specification of the Markov chain. This type of uncertainty is often studied in terms of nonlinear expectations (see, e.g. [10], [35], [38], and [39]) and distributional robustness (see, e.g. [8], [23], [25], and [27]). Here, the main point is to take uncertainty with respect to the transition kernel π into consideration. Conceptually, a robust transition kernel is the following. If the Markov chain is at point $x\in E$, the next step of the Markov chain is not necessarily determined by a fixed measure $\pi(x)$, but rather can be determined by any measure $\hat{\pi}\in P(x)\subseteq\mathcal{P}(E)$. In our context, $P(x)$ will be defined as a neighborhood of $\pi(x)$ with respect to the first Wasserstein distance.

https://doi.org/10.1017/apr.2019.6 Published online by Cambridge University Press

The existing literature on robust Markov chains focuses on finite state spaces, where transition probabilities are uncertain in some convex and closed sets, usually expressed via matrix intervals. For example Škulj [42] gave a good overview of the field. Robust Markov chains are studied under the names of Markov set chains (see, e.g. [28], [29], and [32]), imprecise Markov chains (see, e.g. [14]), as well as Markov chains with interval probabilities (see, e.g. [41] and [42]). Recently, continuous-time versions have also received attention (see, e.g. [22] and [43]). While different types of laws of large numbers are studied frequently, large deviations theory seems to be absent in the current literature on robust Markov chains. Robust Markov chains find applications in several areas in machine learning and operations research, as well as in reinforcement learning [37], [46], [47], and [48], network control [22], [40], and server assignment [31].

In the following, the asymptotic behavior of such Markov chains is analyzed. The type of asymptotics studied is worst-case behaviors over all possible distributions, in the sense of large deviation probabilities (Theorem 1.2) and a law of large numbers (Theorem 1.3) of empirical measures of robust Markov chains. Worst-case behavior for large deviations means that the slowest possible rate of convergence to 0 of a tail event is identified. For laws of large numbers, we give upper bounds—or by changing signs, lower bounds—for law of large number type limits.

Define the first Wasserstein distance d_W on $\mathcal{P}(E)$ by

$$d_W(\mu, \nu) = \inf_{\tau \in \Pi(\mu,\nu)} \int_E d(x, y)\tau(\mathrm{d}x, \ \mathrm{d}y)$$

for $\mu, \nu \in \mathcal{P}(E)$, where $\Pi(\mu, \nu) \subseteq \mathcal{P}(E^2)$ denotes the set of measures with first marginal μ and second marginal ν. See, for example, [26] for an overview regarding the Wasserstein distance. In order to avoid complications with respect to compatibility of the weak convergence and Wasserstein distance, we assume that E is compact for the applications.

Fix $r \geq 0$. The set of possible joint distributions of the robust Markov chain up to step n is characterized by $M_n(\pi_0) \subseteq \mathcal{P}(E^n)$ defined by

$$M_n(\pi_0) := \{\nu \in \mathcal{P}(E^n) \colon d_W(\nu_{0,1}, \pi_0) \leq r \text{ and } d_W(\nu_{i,i+1}(x_1, \dots, x_i), \pi(x_i)) \leq r \nu\text{-a.s.} \\ \text{for } i = 1, \dots, n-1\}.$$

Note that elements of $M_n(\pi_0)$ do not have to be Markov chains. In this context the Markov property only applies to the evolution of the set of measures, but each individual measure can depend on the entire path.

For technical reasons related to Condition (B3), we also consider the following modification:

$$\underline{M}_n(\pi_0) :=\{\nu \in M_n(\pi_0) \colon \nu \ll \pi_0 \otimes \pi \otimes \cdots \otimes \pi\}.$$

Both definitions above can of course be stated for arbitrary $\theta \in \mathcal{P}(E)$ instead of π_0. We show that

$$\beta(\nu, \mu) := \inf_{\hat{\mu} \in M_1(\mu)} R(\nu, \hat{\mu})$$

satisfies the assumptions for the upper bound of Theorem 1.1, and that

$$\underline{\beta}(\nu, \mu) := \inf_{\hat{\mu} \in \underline{M}_1(\mu)} R(\nu, \hat{\mu})$$

satisfies the assumptions for the lower bound of Theorem 1.1. In Lemma 3.1 and Lemma 3.5 we will characterize β_n^θ and $\underline{\beta}_n^\theta$ in terms of $M_n(\theta)$ and $\underline{M}_n(\theta)$.

The rate functions I and $\underline{I}$ denote the rate functions for β and $\underline{\beta}$, respectively, as given by (1.2) and can be expressed as

$$I(\nu) = \inf_{q:\nu q=\nu} \inf_{\mu\in M_2(\nu)} R(\nu\otimes q, \mu), \qquad \underline{I}(\nu) = \inf_{q:\nu q=\nu} \inf_{\mu\in \underline{M}_2(\nu)} R(\nu\otimes q, \mu).$$

In the $r=0$ case, these rate functions simplify to those used in [20, Chapter 8], since, for $r=0$, it holds that $M_2(\nu)=\underline{M}_2(\nu)=\{\nu\otimes\pi\}$.

Theorem 1.1 yields the following result.

Theorem 1.2. *Assume that (E, d) is compact. Let β, $\underline{\beta}$ and $M_n(\theta)$, $\underline{M}_n(\theta)$ for $\theta\in\mathcal{P}(E)$ be given as above. Let I and $\underline{I}$ denote the rate functions for β and $\underline{\beta}$, respectively, as given by (1.2).*

(a) *If π satisfies the Feller property, it holds that, for Borel sets $A\subseteq\mathcal{P}(E)$,*

$$\limsup_{n\to\infty} \sup_{\mu\in M_n(\pi_0)} \frac{1}{n} \ln \mu(L_n\in\bar{A}) \le -\inf_{\nu\in\bar{A}} I(\nu).$$

(b) *If π satisfies (M), it holds that, for Borel sets $A\subseteq\mathcal{P}(E)$,*

$$\liminf_{n\to\infty} \sup_{\mu\in \underline{M}_n(\pi_0)} \frac{1}{n} \ln \mu(L_n\in\mathring{A}) \ge -\inf_{\nu\in\mathring{A}} \underline{I}(\nu).$$

For a (numerical) illustration of the above result, see Example 3.1. Among other things, the example showcases that one can identify conditions such that there is no difference between the upper and lower bounds, and, thus, the above identifies precise asymptotic rates. Note that in finite state spaces one can guarantee $M_n(\theta)=\underline{M}_n(\theta)$ by assuming that $\pi(x)(y)>0$ for all $x, y\in E$.

The following is the law of large numbers result for robust Markov chains, which is based on the choices

$$\beta(\mu,\nu) := \begin{cases} 0 & \text{if } d_W(\mu,\nu)\le r, \\ \infty & \text{otherwise,} \end{cases} \qquad \underline{\beta}(\mu,\nu) := \begin{cases} 0 & \text{if } d_W(\mu,\nu)\le r \text{ and } \mu\ll\nu, \\ \infty & \text{otherwise,} \end{cases}$$

again for fixed $r\ge 0$.

Theorem 1.3. *Assume that (E, d) is compact. Let $M_n(\theta)$, $\underline{M}_n(\theta)$ for $\theta\in\mathcal{P}(E)$ be given as above.*

(a) *If π satisfies the Feller property, it holds that, for all $F\colon\mathcal{P}(E)\to[-\infty,\infty)$, which are upper semicontinuous and bounded from above,*

$$\limsup_{n\to\infty} \sup_{\mu\in M_n(\pi_0)} \int_{E^n} F\circ L_n \,\mathrm{d}\mu \le \sup_{\nu\in\mathcal{P}(E)\colon \text{there exists } q,\nu q=\nu\colon\, \nu\otimes q\in M_2(\nu)} F(\nu).$$

(b) *If π satisfies (M), it holds that, for all $F\in C_b(\mathcal{P}(E))$,*

$$\liminf_{n\to\infty} \sup_{\mu\in \underline{M}_n(\pi_0)} \int_{E^n} F\circ L_n \,\mathrm{d}\mu \ge \sup_{\nu\in\mathcal{P}(E)\colon \text{there exists } q,\nu q=\nu\colon\, \nu\otimes q\in \underline{M}_2(\nu)} F(\nu).$$

This result is easiest interpreted by looking at the $r = 0$ case. If both the upper and lower bounds hold, the above states that

$$\pi_n \circ L_n^{-1} \overset{w}{\to} \delta_{\mu^*} \in \mathcal{P}(\mathcal{P}(E)),$$

where μ^* is the unique invariant measure under the Markov chain transition kernel π, which—under Assumption M—always exists (see [20, Lemma 8.6.2(a)]).

Specifically, the choices $F(\nu) := \int_E f \, \mathrm{d}\nu$ for $f \in C_b(E)$ in the above theorem can be interpreted as a robust Cesàro limit of a Markov chain. Indeed, for $r = 0$, this yields

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \pi_0 \pi^{(i-1)} \overset{w}{\to} \mu^*.$$

For $r > 0$, however, we obtain a result which strongly resembles, e.g. [28, Theorem 4.1], but in a more general state space.

1.3. Generalizations and relation to the literature

In this paper robustness is modeled via the first Wasserstein distance because it is both tractable and frequently used. Nevertheless, the question arises whether the presented approach can be applied more generally, specifically related to the existing literature in finite state spaces. In this section we roughly outline potential extensions.

In the existing literature regarding robust Markov chains in finite state spaces—where we mainly refer to [28] and [42] as references—the starting point is a robust transition kernel $P \colon E \to 2^{\mathcal{P}(E)}$ satisfying certain convexity and closedness conditions. For our approach however, we start with both a transition kernel $\pi : E \to \mathcal{P}(E)$ and a mapping $U \colon \mathcal{P}(E) \to 2^{\mathcal{P}(E)}$, with the relation of the approaches being $P = U \circ \pi$.

In Section 1.2 we used $U(\mu) = \{\hat{\mu} \in \mathcal{P}(E) \colon d_W(\mu, \hat{\mu}) \le r\}$. The setting of Section 1.2 translates to $\beta(\nu, \mu) = \inf_{\hat{\mu} \in U(\mu)} R(\nu, \hat{\mu})$ for large deviation results (Theorem 1.2) and $\beta(\nu, \mu) = \infty \cdot \mathbf{1}_{U(\mu)^C}(\nu)$ for law of large number results (Theorem 1.3). Furthermore, $M_n(\theta) = \{\mu \in \mathcal{P}(E^n) \colon \mu_{0,1} \in U(\theta), \mu_{i,i+1}(x_1, \ldots, x_i) \in U(\pi(x_i))\ \mu\text{-a.s. for } i = 1, \ldots, n-1\}$ for $\theta \in \mathcal{P}(E)$. In general, the following conditions on U would allow for a similar type of proof for analogs of Theorems 1.2 and 1.3, where the assumptions on E (compactness) and π (Feller property and/or Assumption M) stay the same.

(a) $\mu \in U(\mu)$ for all $\mu \in \mathcal{P}(E)$.

(b) The graph of U, i.e. $\{(\mu, \hat{\mu}) \in \mathcal{P}(E)^2 \colon \hat{\mu} \in U(\mu)\}$, is closed and convex.

Here, (a) implies that $\beta(\mu, \mu) = 0$ for all $\mu \in \mathcal{P}(E)$. That the graph of U is convex implies Condition (B1); see Lemma 3.2 and the subsequent paragraph, as well as Lemma 3.7. Closedness of the graph is used to verify Condition (B2); see Lemmas 3.3, 3.4, and 3.7. For the large deviations result, closedness of the graph also guarantees a representation of β_n^θ in terms of $M_n(\theta)$; see Lemmas 3.1 and 3.5.

The assumption that E has to be compact can likely be loosened by assuming that U is compact valued instead, even though an analog of Lemma 3.3 is then more difficult to obtain.

1.4. Further ideas and outlook

There are several possibilities for further directions that the main result, Theorem 1.1, can be used for. We refer again to the paper by Lacker [34] in which a range of applications are

https://doi.org/10.1017/apr.2019.6 Published online by Cambridge University Press

discussed in detail for the i.i.d. setting. For Markov chains, further natural choices for β that may lead to interesting applications are, for example, the following.

- φ-divergences, i.e. for a convex function $\varphi\colon \mathbb{R}_+ \to \mathbb{R}$,

$$\beta(\nu, \mu) := \int_E \varphi\Big(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\Big)\, \mathrm{d}\mu,$$

  which is understood to be ∞ if ν is not absolutely continuous with respect to μ.

- Transport costs, i.e. for a lower semicontinuous $c\colon E^2 \to [0, \infty]$,

$$\beta(\nu, \mu) := \inf_{\tau \in \Pi(\nu,\mu)} \int_E c\, \mathrm{d}\tau.$$

- The L^p norm of the Radon–Nikodym derivative

$$\beta(\nu, \mu) := \Big\| \frac{\mathrm{d}\nu}{\mathrm{d}\mu} \Big\|_{L^p(\mu)}.$$

- The sum of different choices for β, i.e. for $\beta^{(i)}$ $(i = 1, \ldots, K)$, define

$$\beta(\nu, \mu) := \sum_{i=1}^{K} \beta^{(i)}(\nu, \mu).$$

The first choice of φ-divergences is certainly interesting, and the key conditions (B1) and (B2) appear obtainable. Yet, the manifestations of this choice are difficult to interpret, since the resulting functionals β_n and ρ_n are very complex; see also [34, Chapter 1.5.]. The second choice of transport costs is analyzed in detail in [34, Chapter 6] in the i.i.d. case. For Markov chains, conditions (B1) and (B2) are satisfied in compact spaces. This is shown in detail in the supplementary material to this paper [21], as it nicely illustrates standard methods used to apply Theorem 1.1. Condition (B3) for the lower bound is, in general, not satisfied (consider $E = \mathbb{R}$ and the Euclidean distance c), but can be satisfied for specific c (e.g. $c(x, y) = 0$ if $x = y$ and $c(x, y) = \infty$ otherwise). The resulting limit theorem is loosely related to all kinds of optimization problems, including optimal transport balls; see, e.g. [2], [5], [7], [8], [23], and [25]. The third idea leads to polynomial large deviation bounds in the i.i.d. case (see [34, Chapter 4]). For Markov chains, while β is a very natural choice, β is in general not jointly convex, and, hence, Condition (B1) is not satisfied. Nevertheless, one should still keep this choice in mind, as there might be possible adaptations to make it applicable. The fourth point arises from the observation that all major assumptions on β, (B1), (B2), and (B3), are closed under finite summation, and, furthermore, the resulting dual ρ_n is tractable via the inf-convolution of the individual duals; see, e.g. [11, Chapter 1.6] for background.

Some choices of β have direct implications for applications, e.g. the large deviation bounds for robust Markov chains from Theorem 1.2 can be used to analyze and fine-tune systems using robust Markov chains; see, e.g. [22], [31], [40], [46], [47], and [48]. On the other hand, there may be some hidden applications that are more nuanced. One example is to use the results from Section 1.2 to analyze certain complex stochastic processes, which are themselves not Markovian (or even precisely determined at all), but can be shown to be *close* to a certain Markov chain (in the sense of Section 1.2). Processes like this may arise naturally, for example, in numerical implementations of certain Markovian algorithms. The stochastic

https://doi.org/10.1017/apr.2019.6 Published online by Cambridge University Press

process corresponding to the numerical implementation, including numerical inaccuracies and other potential sources of error, may be difficult to model precisely. However, it is natural to think of this process as *close* to the process corresponding to the theoretical algorithm, and, hence, the asymptotic behavior may be analyzed using the results from Section 1.2. The reason this requires the results from this paper as opposed to the results from the existing literature on robust Markov chains is that usually finite state spaces are too restrictive, while the assumption of compact state spaces (as in this paper) is often reasonable.

1.5. Structure of the paper

In Section 2 we prove Theorem 1.1 and Corollary 1.1. The method of proof is oriented at [20, Chapters 8 and 9], while also using tools from convex duality and measurable selection. In Section 2.1 we provide results relating to the lower bound and its proof, and in Section 2.2 we provide results relating to the upper bound and its proof. In Section 2.3 we provide the proof of Corollary 1.1.

In Section 3 we present in depth the applications to robust Markov chains. Aside from using Theorem 1.1 and Corollary 1.1, Section 3 is self-contained, so readers who prefer to read Section 3 before Section 2 can easily do so. A large part of Section 3 is devoted to verifying conditions (B1) and (B2) for the different choices of β. Furthermore, the large deviation results obtained are illustrated in Example 3.1.

Many of the smaller results not listed in the introduction are interesting in their own right, e.g. Lemmas 2.1, 2.2, and 3.3.

In the supplementary material [21] to this paper, applicability of the main theorem for the choice of β as a transport cost is shown.

2. Proofs of Theorem 1.1 and Corollary 1.1

2.1. The lower bound of Theorem 1.1

At some points in this section it is necessary to evaluate ρ_n^θ at universally measurable functions, which is still well defined. More precisely, upper semi-analytic functions are the object of interest, the reason made obvious in Lemma 2.1. In particular, upper semi-analytic functions are universally measurable; see, e.g. [6, Chapter 7] for background.

2.1.1. *Preliminary results*

Lemma 2.1. (See also [34, Proposition A.1].) *For $\theta \in \mathcal{P}(E)$, $f\colon E^n \to \mathbb{R}$ upper semi-analytic, and $0 < k < n$, it holds that*

$$\rho_n^\theta(f) = \rho_k^\theta(g),$$

where $g\colon E^k \to \mathbb{R}$ is defined by

$$g(x_1, \dots, x_k) = \rho_{n-k}^{\pi(x_k)}(f(x_1, \dots, x_k, \cdot)).$$

Furthermore, g is upper semi-analytic.

Proof. First, let $\nu \in \mathcal{P}(E^k)$ and let $K\colon E^k \to \mathcal{P}(E^{n-k})$ be a stochastic kernel. For notational purposes, we write $\bar{x} = (x_1, \dots, x_k)$ for $x_1, \dots, x_k \in E$ and

$$K(x_1, \dots, x_k) = K(\bar{x}) = K^{\bar{x}}.$$

Denote the decomposition of $K^{\bar{x}}$ in the usual way, i.e.

$$K^{\bar{x}} = K_{0,1}^{\bar{x}} \otimes K_{1,2}^{\bar{x}} \otimes \cdots \otimes K_{n-k-1,n-k}^{\bar{x}}.$$

https://doi.org/10.1017/apr.2019.6 Published online by Cambridge University Press

For the decompositions of ν and $\nu \otimes K$, the trivial $\nu \otimes K$-almost-sure equalities hold:

$$\nu_{i,i+1}(x_1, \ldots, x_i) = (\nu \otimes K)_{i,i+1}(x_1, \ldots, x_i) \qquad \text{for } i = 0, \ldots, k-1,$$
$$K^{\bar{x}}_{i,i+1}(x_{k+1}, \ldots, x_{k+i}) = (\nu \otimes K)_{k+i,k+i+1}(x_1, \ldots, x_{k+i}) \qquad \text{for } i = 0, \ldots, n-k-1.$$

Hence,

$$\begin{aligned}
&\beta^{\theta}_k(\nu) + \int_{E^k} \beta^{\pi(x_k)}_{n-k}(K^{\bar{x}})\nu(\mathrm{d}x_1, \ldots, \mathrm{d}x_k) \\
&\quad = \int_{E^n} \beta(\nu_{0,1}, \theta) + \Bigg(\sum_{i=1}^{k-1} \beta(\nu_{i,i+1}(x_1, \ldots, x_i), \pi(x_i))\Bigg) + \beta(K^{\bar{x}}_{0,1}, \pi(x_k)) \\
&\qquad + \Bigg(\sum_{i=1}^{n-k-1} \beta(K^{\bar{x}}_{i,i+1}(x_{k+1}, \ldots, x_{k+i}), \pi(x_{k+i}))\Bigg) \\
&\qquad \times K^{\bar{x}}(dx_{k+1}, \ldots, dx_n)\nu(\mathrm{d}x_1, \ldots, \mathrm{d}x_k) \\
&\quad = \beta^{\theta}_n(\nu \otimes K).
\end{aligned}$$

Using the above and a standard measurable selection argument [6, Proposition 7.50], we obtain

$$\begin{aligned}
\rho^{\theta}_k(g) &= \sup_{\nu \in \mathcal{P}(E^k)} \bigg(\int_{E^k} g \,\mathrm{d}\nu - \beta^{\theta}_k(\nu)\bigg) \\
&= \sup_{\nu \in \mathcal{P}(E^k)} \bigg(\int_{E^k} \sup_{\mu \in \mathcal{P}(E^{n-k})} \bigg(\int_{E^{n-k}} f(x_1, \ldots, x_n)\mu(\mathrm{d}x_{k+1}, \ldots, \mathrm{d}x_n) - \beta^{\pi(x_k)}_{n-k}(\mu)\bigg) \\
&\qquad \times \nu(\mathrm{d}x_1, \ldots, \mathrm{d}x_k) - \beta^{\theta}_k(\nu)\bigg) \\
&= \sup_{\nu \in \mathcal{P}(E^k)} \sup_{\substack{K \colon E^k \to \mathcal{P}(E^{n-k}), \\ K \text{ Borel}}} \bigg(\int_{E^n} f \,\mathrm{d}\nu \otimes K - \beta^{\nu}_n(\nu \otimes K)\bigg) \\
&= \rho^{\theta}_n(f).
\end{aligned}$$

That g is upper semi-analytic can be shown as follows. Both the mappings

$$(x_k, \nu) \mapsto -\beta^{\pi(x_k)}_{n-k}(\nu), \qquad (x_1, \ldots, x_k, \nu) \mapsto \int_{E^{n-k}} f(x_1, \ldots, x_k, \cdot) \,\mathrm{d}\nu$$

are upper semi-analytic by [6, Proposition 7.48], where, for the first mapping, we implicitly have to use [6, Proposition 7.27] as mentioned after the definition of $\beta^{\cdot}_n(\cdot)$. The sum of these mappings is therefore still upper semi-analytic (see, e.g. [6, Lemma 7.30(4)]) and, hence, by [6, Proposition 7.47], g is upper semi-analytic. □

Lemma 2.2. *Under Condition (B3), for all $\theta \in \mathcal{P}(E)$ and $f \colon E^n \to \mathbb{R}$ upper semi-analytic, it holds that*

$$\rho^{\theta}_n(f) \geq \int_E \rho^{\delta_x}_n(f)\theta(\mathrm{d}x).$$

Proof. By Condition (B3), it holds that, for all $x \in E$,

$$\rho_n^{\delta_x}(f) = \sup_{\nu \in \mathcal{P}(E^n)} \biggl(\int_{E^n} f \, \mathrm{d}\nu - \beta_n^{\delta_x}(\nu) \biggr)$$
$$= \sup_{\nu \in \mathcal{P}(E^n)\colon \nu_{0,1}=\delta_x} \biggl(\int_{E^n} f \, \mathrm{d}\nu - \beta_n^{\delta_x}(\nu) \biggr)$$
$$= \sup_{\nu \in \mathcal{P}(E^{n-1})} \biggl(\int_{E^n} f \, \mathrm{d}(\delta_x \otimes \nu) - \beta_n^{\delta_x}(\delta_x \otimes \nu) \biggr).$$

Hence, we obtain, for $\theta \in \mathcal{P}(E)$,

$$\int_E \rho_n^{\delta_{x_1}}(f)\theta(\mathrm{d}x_1)$$
$$= \int_E \sup_{\nu \in \mathcal{P}(E^{n-1})} \biggl(\int_{E^n} f \, \mathrm{d}(\delta_{x_1} \otimes \nu) - \beta_n^{\delta_{x_1}}(\delta_{x_1} \otimes \nu) \biggr) \theta(\mathrm{d}x_1)$$
$$= \int_E \sup_{\nu \in \mathcal{P}(E^{n-1})} \biggl(\int_{E^{n-1}} f(x_1, \cdot) \, \mathrm{d}\nu$$
$$- \int_{E^{n-1}} \sum_{k=2}^{n} \beta(\nu_{k-2,k-1}(x_2, \dots, x_{k-1}), \pi(x_{k-1}) \nu(\mathrm{d}x_2, \dots, \mathrm{d}x_n) \biggr) \theta(\mathrm{d}x_1)$$
$$\overset{(*)}{=} \sup_{\substack{K : E \to \mathcal{P}(E^{n-1}) \\ K \text{Borel}}} \biggl(\int_{E^n} f \, \mathrm{d}\theta \otimes K - \beta_n^{\theta}(\theta \otimes K) \biggr)$$
$$\leq \sup_{\nu \in \mathcal{P}(E^n)} \biggl(\int_{E^n} f \, \mathrm{d}\nu - \beta_n^{\theta}(\nu) \biggr)$$
$$= \rho_n^{\theta}(f).$$

Here, $(*)$ follows by a standard measurable selection argument; see, for example, [6, Proposition 7.50]. □

Lemma 2.3. *Let $(X_i)_{i \in \mathbb{N}}$ be an E-valued sequence of random variables such that $\lim_{n\to\infty} (1/n) \sum_{i=1}^{n} F(X_i) = \mathbb{E}[F(X_1)]$ holds almost surely for all $F \in C_b(E)$. Let $\nu^{(n)} = \mathbb{P} \circ (X_1, \dots, X_n)^{-1}$ be the distribution of $(X_1, \dots, X_n)$ for $n \in \mathbb{N}$. Then $\nu^{(n)} \circ L_n^{-1} \xrightarrow{w} \delta_{\nu^{(1)}}$.*

Proof. By a standard separability argument (cf. [44, Proof of Theorem 3.1]), it follows that $L_n(X_1, \dots, X_n) \xrightarrow{w} \nu^{(1)}$ holds $\mathbb{P}$-a.s. Hence, by dominated convergence,

$$\nu^{(n)} \circ L_n^{-1} \xrightarrow{w} \delta_{\nu^{(1)}}.$$ □

For the following results, note that, under Assumption M, π has a unique invariant measure, which we denote by μ^*; see [20, Lemma 8.6.2(a)]. Furthermore, recall that, for $k \in \mathbb{N}$, we denote by $\pi^{(k)}$ the k-step transition kernel of the Markov chain, as defined in Condition (M1).

Lemma 2.4. ([20, Lemma 8.6.2(b)].) *Let Assumption M be satisfied. Let $A \subseteq E$ be a Borel set such that $\pi^{(l_0)}(x_0, A) > 0$ for some $x_0 \in E$. Then $\mu^*(A) > 0$, where μ^* is the unique invariant measure under π.*

https://doi.org/10.1017/apr.2019.6 Published online by Cambridge University Press

Lemma 2.5. (Adapted version of [20, Lemma 8.6.2(c)].) *Let Assumption M and Condition (B3) be satisfied. Let $\nu \in \mathcal{P}(E)$ satisfy $\beta_2^\nu(\nu \otimes p) < \infty$ for some stochastic kernel p on E such that $\nu p = \nu$. Then it holds that $\nu \ll \mu^*$, where μ^* is the unique invariant measure under π.*

Proof. Let $\Omega_0 \subseteq E$ be a Borel set such that $\nu(\Omega_0) = 1$ and $p(x) \ll \pi(x)$ for all $x \in \Omega_0$, which we can choose by (B3) and since $\beta_2^\nu(\nu \otimes p) < \infty$. Define $\tilde{p}(x) := \mathbf{1}_{\Omega_0}(x)p(x) + \mathbf{1}_{\Omega_0^C}(x)\pi(x)$. Since $\tilde{p}(x) \ll \pi(x)$ for all $x \in E$, we have $\tilde{p}^{(l_0)}(x) \ll \pi^{(l_0)}(x)$ for all $x \in E$, where l_0 is the constant from Condition (M1).

Now choose a Borel set $A \subseteq E$ such that $\nu(A) > 0$. By iterating $\nu\tilde{p} = \nu$, we obtain a Borel set $B \subseteq E$ with $\nu(B) > 0$ and $\tilde{p}^{(l_0)}(x, A) > 0$ for all $x \in B$. Hence, $\pi^{(l_0)}(x, A) > 0$ for all $x \in B$ and, therefore, by Lemma 2.4, $\mu^*(A) > 0$. □

2.1.2. *Proof of Theorem 1.1: the lower bound.* Let $F \in C_b(\mathcal{P}(E))$ and $\varepsilon > 0$ be fixed. We have to show that

$$\liminf_{n\to\infty} \frac{1}{n} \rho_n(nF \circ L_n) \geq \sup_{\nu\in\mathcal{P}(E)} (F(\nu) - I(\nu)) - 4\varepsilon.$$

We do this by showing that every subsequence has a further subsequence which satisfies this inequality. So we fix a subsequence and relabel it $n \in \mathbb{N}$. Labeling subsequences by the same index as the original sequence will be a common practice throughout the remainder of this paper.

Outline of the proof. First, we show that there exists a Borel set $\Phi \subseteq E$ such that $\pi^{(l_0)}(y, \Phi) = 1$ for all $y \in E$, and, for all $x \in \Phi$,

$$\liminf_{n\to\infty} \frac{1}{n} \rho_{n-l_0}^{\delta_x}(nF \circ L_n(x_1, \ldots, x_{l_0}, \cdot)) \geq \sup_{\nu\in\mathcal{P}(E)} (F(\nu) - I(\nu)) - 3\varepsilon \tag{2.1}$$

holds for all $x_1, \ldots, x_{l_0} \in E$ and a further subsequence (the same subsequence for all $x_1, \ldots, x_{l_0}$). This subsequence then remains fixed for the rest of the proof and is again labeled by $n \in \mathbb{N}$.

The next step is to use Lemma 2.1, i.e. for all $f \in C_b(E^n)$,

$$\rho_n(f) = \rho_{l_0}((x_1, \ldots, x_{l_0}) \mapsto \rho_{n-l_0}^{\pi(x_{l_0})}(f(x_1, \ldots, x_{l_0}, \cdot)),$$

where l_0 is the constant from Condition (M1). This is used together with Lemma 2.2, i.e. for all $f \in C_b(E^n)$ and $\theta \in \mathcal{P}(E)$,

$$\rho_n^\theta(f) \geq \int_E \rho_n^{\delta_x}(f)\theta(\mathrm{d}x).$$

We then use these two results to show that

$$\rho_n(nF \circ L_n) \geq \rho_{l_0}(g_n), \tag{2.2}$$

where

$$g_n(x_1, \ldots, x_{l_0}) = \int_\Phi \rho_{n-l_0}^{\delta_x}(nF \circ L_n(x_1, \ldots, x_{l_0}, \cdot))\pi^{(l_0)}(x_{l_0}, \mathrm{d}x).$$

We conclude by combining the first limit result (2.1) and inequality (2.2), which works by Fatou's lemma, using monotonicity of ρ_n and the fact that $\rho_n(c) \geq c$ for all $c \in \mathbb{R}$.

Step 1. We show that (2.1) holds for all $x \in \Phi$ and $x_1, \ldots, x_{l_0} \in E$, where Φ and the required further subsequence is specified later.

https://doi.org/10.1017/apr.2019.6 Published online by Cambridge University Press

We can, without loss of generality, choose $\nu_0 \in \mathcal{P}(E)$ such that

$$-\infty < \sup_{\nu \in \mathcal{P}(E)} (F(\nu) - I(\nu)) \le F(\nu_0) - I(\nu_0) + \varepsilon < \infty,$$

since if the supremum equals $-\infty$, there is nothing to show. Then

$$\inf_{q : \nu_0 q = \nu_0} \int_E \beta(q(x), \pi(x)) \nu_0(dx) = I(\nu_0) < \infty.$$

Choose a stochastic kernel p on E with $\nu_0 p = \nu_0$ such that

$$\infty > I(\nu_0) + \varepsilon \ge \int_E \beta(p(x), \pi(x)) \nu_0(\mathrm{d}x) = \beta_2^{\nu_0}(\nu_0 \otimes p).$$

By (B3), we can choose a Borel set $N \subseteq E$ with $\nu_0(N) = 0$ such that $p(x) \ll \pi(x)$ for all $x \in N^C$. Define the stochastic kernel p_0 on E by $p_0(x) := \mathbf{1}_N(x)\pi(x) + \mathbf{1}_{N^C}(x)p(x)$ for $x \in E$ and show that

$$\infty > I(\nu_0) + \varepsilon \ge \beta_2^{\nu_0}(\nu_0 \otimes p) = \beta_2^{\nu_0}(\nu_0 \otimes p_0).$$

For all $x \in E$, $p_0(x) \ll \pi(x)$ holds. Next, we replace ν_0 and p_0 by ν_1 and p_1, such that $F(\nu_1) + \beta_2^{\nu_1}(\nu_1 \otimes p_1) \ge F(\nu_0) + \beta_2^{\nu_0}(\nu_0 \otimes p_0) - 2\varepsilon$ and, additionally, p_1 is pointwise equivalent to π.

By conditions (M1) and (M2), π has a unique invariant measure, denoted by μ^*; see [20, Lemma 8.6.2(a)]. By the lower boundedness of β, we can choose $\kappa_0 \in (0, 1)$ such that

$$(1 - \kappa_0)\beta_2^{\nu_0}(\nu_0 \otimes p_0) \le \beta_2^{\nu_0}(\nu_0 \otimes p_0) + \varepsilon.$$

By continuity of F, we can further choose $\kappa_1 > 0$ such that, for all $0 \le \hat{\kappa} \le \kappa_1$,

$$F((1 - \hat{\kappa})\nu_0 + \hat{\kappa}\mu^*) \ge F(\nu_0) - \varepsilon.$$

Choose $\kappa := \min\{\kappa_0, \kappa_1\}$ and define $\nu_1 := (1 - \kappa)\nu_0 + \kappa\mu^*$ and

$$p_1(x) = \frac{\mathrm{d}\nu_0}{\mathrm{d}\nu_1}(x)(1 - \kappa)p_0(x) + \frac{\mathrm{d}\mu^*}{\mathrm{d}\nu_1}(x)\kappa\pi(x).$$

Then one quickly checks that $\nu_1 \otimes p_1 = (1 - \kappa)(\nu_0 \otimes p_0) + \kappa(\mu^* \otimes \pi)$, in particular, therefore, $\nu_1 p_1 = \nu_1$. By the convexity of $\beta_2^{\cdot}(\cdot)$ and the assumption that $\beta(\nu, \nu) = 0$ for all $\nu \in \mathcal{P}(E)$, it holds that

$$\beta_2^{\nu_1}(\nu_1 \otimes p_1) \le (1 - \kappa)\beta_2^{\nu_0}(\nu_0 \otimes p_0) + \kappa\beta_2^{\mu^*}(\mu^* \otimes \pi) \le \beta_2^{\nu_0}(\nu_0 \otimes p_0) + \varepsilon,$$

and, thus,

$$F(\nu_1) - \beta_2^{\nu_1}(\nu_1 \otimes p_1) \ge F(\nu_0) - \beta_2^{\nu_0}(\nu_0 \otimes p_0) - 2\varepsilon.$$

Since $\beta_2^{\nu_1}(\nu_1 \otimes p_1) < \infty$, without loss of generality, $p_1(x) \ll \pi(x)$ for all $x \in E$. By Lemma 2.5 (which yields $\nu_1 \ll \mu^*$ and, hence, $\mathrm{d}\mu^*(x)/\,\mathrm{d}\nu_1 > 0$ for ν_1-almost all $x \in E$) and by the construction of p_1, it also holds that $\pi(x) \ll p_1(x)$, again without loss of generality for all $x \in E$.

So p_1 also satisfies (M1), as every kernel which is pointwise equivalent to π satisfies (M1), notably with the same constants l_0 and n_0.

It follows that the Markov chain with initial distribution ν_1 and transition kernel p_1 is ergodic; see [20, Lemma 8.6.2(a)]. It follows from the pointwise ergodic theorem (for both

https://doi.org/10.1017/apr.2019.6 Published online by Cambridge University Press

ergodic theorems used, see [20, Appendix A.4] or the references therein, i.e. [9, Corollaries 6.23 and 6.25]) that the sequence

$$(\mu^{(n)})_{n\in\mathbb{N}} := \Big(\nu_1 \otimes \Big(\bigotimes_{i=1}^{n-1} p_1\Big)\Big)_{n\in\mathbb{N}}$$

satisfies the conditions of Lemma 2.3 and, thus, $\mu^{(n)} \circ L_n^{-1} \xrightarrow{w} \delta_{\nu_1}$. This yields

$$\lim_{n\to\infty} \int_{E^n} |F \circ L_n - F(\nu_1)| \,\mathrm{d}\mu^{(n)} = \lim_{n\to\infty} \int_{\mathcal{P}(E)} |F - F(\nu_1)| \,\mathrm{d}\mu^{(n)} \circ L_n^{-1} = 0. \tag{2.3}$$

Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of E-valued random variables such that $(X_1, \dots, X_n) \sim \mu^{(n)}$ for all $n \in \mathbb{N}$. We see that

$$\begin{aligned} \mathbb{E}[\beta(p_1(X_1), \pi(X_1))] &= \beta_2^{\nu_1}(\nu_1 \otimes p_1), \\ \mathbb{E}[|\beta(p_1(X_1), \pi(X_1))|] &\le \Big|\min_{x\in\mathcal{P}(E)^2} \beta(x)\Big| + \beta_2^{\nu_1}(\nu_1 \otimes p_1) < \infty, \end{aligned}$$

and, thus, by the L_1-ergodic theorem,

$$\begin{aligned} &\lim_{n\to\infty} \mathbb{E}\bigg[\bigg|\frac{1}{n}\sum_{i=1}^{n-1} \beta(p_1(X_i), \pi(X_i)) - \beta_2^{\nu_1}(\nu_1 \otimes p_1)\bigg|\bigg] = 0 \\ \iff \quad &\lim_{n\to\infty} \int_{E^n} \bigg|\frac{1}{n}\sum_{i=1}^{n-1} \beta(p_1(x_i), \pi(x_i)) - \beta_2^{\nu_1}(\nu_1 \otimes p_1)\bigg| \mu^{(n)}(\mathrm{d}x_1, \dots, \mathrm{d}x_n) = 0. \end{aligned} \tag{2.4}$$

For $\theta \in \mathcal{P}(E)$ and a stochastic kernel $q \colon E \to \mathcal{P}(E)$, we define

$$(\mu^{(\theta,q,n)})_{n\in\mathbb{N}} := \Big(\theta \otimes \Big(\bigotimes_{i=1}^{n-1} q\Big)\Big)_{n\in\mathbb{N}}.$$

By the above limits (2.3) and (2.4), and the fact that L_1-convergence implies almost-sure convergence of a subsequence, we can choose a Borel set $\Phi \subseteq E$, $\nu_1(\Phi) = 1$ such that, for all $x \in \Phi$ and a subsequence (again labeled by $n \in \mathbb{N}$),

$$\lim_{n\to\infty} \int_{E^n} |F \circ L_n - F(\nu_1)| \mu^{(\delta_x, p_1, n)}(\mathrm{d}x_1, \dots, \mathrm{d}x_n) = 0 \tag{2.5}$$

and

$$\begin{aligned} &\lim_{n\to\infty} \int_{E^n} \bigg|\frac{1}{n}\sum_{i=1}^{n-1} \beta(p_1(x_i), \pi(x_i)) - \beta_2^{\nu_1}(\nu_1 \otimes p_1)\bigg| \mu^{(\delta_x, p_1, n)}(\mathrm{d}x_1, \dots, \mathrm{d}x_n) = 0 \\ &\qquad \implies \quad \lim_{n\to\infty} \beta_n^{\delta_x}(\mu^{(\delta_x, p_1, n)}) = \beta_2^{\nu_1}(\nu_1 \otimes p_1). \end{aligned}$$

Since ν_1 and μ^* are equivalent by Lemma 2.4, $\mu^*(\Phi) = 1$. Since $\mu^*(\Phi) = 1$, $\pi^{(l_0)}(\Phi) = 1$ holds, as otherwise Lemma 2.4 would imply that $\mu^*(\Phi^C) > 0$. So we have found the set Φ mentioned at the beginning of the proof and the required subsequence. It remains to show that (2.1) holds for all $x \in \Phi$ and $x_1, \dots, x_{l_0} \in E$.

https://doi.org/10.1017/apr.2019.6 Published online by Cambridge University Press

Let $x_1, \dots, x_{l_0} \in E$. By (2.5), dominated convergence, and the triangle inequality,

$$\begin{aligned}
&\int_{E^{n-l_0}} |F \circ L_n(x_1, \dots, x_n) - F(\nu_1)| \mu^{(\delta_x, p_1, n-l_0)}(\mathrm{d}x_{l_0+1}, \dots, \mathrm{d}x_n) \\
&\quad \le \int_{E^{n-l_0}} |F \circ L_n(x_1, \dots, x_n) - F \circ L_{n-l_0}(x_{l_0+1}, \dots, x_n)| \mu^{(\delta_x, p_1, n-l_0)}(\mathrm{d}x_{l_0+1}, \dots, \mathrm{d}x_n) \\
&\qquad + \int_{E^{n-l_0}} |F \circ L_{n-l_0}(x_{l_0+1}, \dots, x_n) - F(\nu_1)| \mu^{(\delta_x, p_1, n-l_0)}(\mathrm{d}x_{l_0+1}, \dots, \mathrm{d}x_n) \\
&\quad \to 0,
\end{aligned}$$

since F is continuous and $\|L_n(x_1, \dots, x_{l_0}, \cdot) - L_{n-l_0}\|_v \le 2l_0/n \to 0$, where $\| \cdot \|_v$ denotes the total variation norm. Thus,

$$\int_{E^{n-l_0}} F \circ L_n(x_1, \dots, x_n) \mu^{(\delta_x, p_1, n-l_0)}(\mathrm{d}x_{l_0+1}, \dots, \mathrm{d}x_n) \to F(\nu_1).$$

Finally, it follows that

$$\begin{aligned}
&\liminf_{n\to\infty} \frac{1}{n} \rho^{\delta_x}_{n-l_0}(nF \circ L_{n-l_0}(x_1, \dots x_{l_0}, \cdot)) \\
&\quad = \liminf_{n\to\infty} \sup_{\nu \in \mathcal{P}(E^{n-l_0})} \biggl(\int_{E^{n-l_0}} F \circ L_n(x_1, \dots, x_n) \nu(\mathrm{d}x_{l_0+1}, \dots, \mathrm{d}x_n) - \beta^{\delta_x}_{n-l_0}(\nu) \biggr) \\
&\quad \ge \liminf_{n\to\infty} \biggl(\int_{E^{n-l_0}} F \circ L_n(x_1, \dots, x_n) \mu^{(\delta_x, p_1, n-l_0)}(\mathrm{d}x_{l_0+1}, \dots, \mathrm{d}x_n) \\
&\qquad\qquad - \beta^{\delta_x}_{n-l_0}(\mu^{(\delta_x, p_1, n-l_0)}) \biggr) \\
&\quad = F(\nu_1) - \beta^{\nu_1}_2(\nu_1 \otimes p_1) \\
&\quad \ge \sup_{\nu \in \mathcal{P}(E)} (F(\nu) - I(\nu)) - 3\varepsilon.
\end{aligned}$$

Step 2. First, define $g_n \colon E^{l_0} \to \mathbb{R}$ for $n > l_0$ by

$$g_n(x_1, \dots, x_{l_0}) = \int_\Phi \rho^{\delta_x}_{n-l_0}(nF \circ L_n(x_1, \dots, x_{l_0}, \cdot)) \pi(x_{l_0}, \mathrm{d}x).$$

Then g_n is upper semi-analytic, since $(x, x_1, \dots, x_{l_0}) \mapsto \rho^{\delta_x}_{n-l_0}(nF \circ L_n(x_1, \dots, x_{l_0}, \cdot))$ is (by [6, Propositions 7.47 and 7.48]; see also Lemma 2.1) and, thus, g_n is as well (by [6, Prop. 7.48]).

By Fatou's lemma (applicable since $\rho^{\delta_x}_n(nF \circ L_n)/n \ge -\|F\|_\infty$), for all $x_1, \dots, x_{l_0} \in E$,

$$\begin{aligned}
\liminf_{n\to\infty} \frac{1}{n} g_n(x_1, \dots, x_n) &\ge \int_\Phi \liminf_{n\to\infty} \frac{1}{n} \rho^{\delta_x}_{n-l_0}(nF \circ L_n(x_1, \dots, x_{l_0}, \cdot)) \pi(x_{l_0}, \mathrm{d}x) \\
&\ge \sup_{\nu \in \mathcal{P}(E)} (F(\nu) - I(\nu)) - 3\varepsilon.
\end{aligned}$$

We define the sets

$$\Omega_n := \biggl\{ (x_1, \dots, x_{l_0}) \in E^{l_0} \colon \frac{1}{n} g_j(x_1, \dots, x_{l_0}) \ge \sup_{\nu \in \mathcal{P}(E)} (F(\nu) - I(\nu)) - 4\varepsilon \text{ for all } j \ge n \biggr\}$$

https://doi.org/10.1017/apr.2019.6 Published online by Cambridge University Press

for $n \in \mathbb{N}$, which are universally measurable and satisfy $\Omega_1 \subseteq \Omega_2 \subseteq \Omega_3 \cdots$ with $\bigcup_{i=1}^{\infty} \Omega_i = E^{l_0}$. For $n \in \mathbb{N}$, let $p_n := \mu^{\pi_0,\pi,l_0}(\Omega_n)$. Then, by continuity from below, $p_n \to 1$ for $n \to \infty$. We have, by Lemmas 2.1 and 2.2, and the monotonicity of ρ_{l_0},

$$\begin{aligned}\liminf_{n\to\infty} \frac{1}{n}\rho_n(nF \circ L_n) &\geq \liminf_{n\to\infty} \frac{1}{n}\rho_{l_0}(g_n)\\ &\geq \liminf_{n\to\infty} \frac{1}{n}\rho_{l_0}\Big(\mathbf{1}_{\Omega_n} n\Big(\sup_{\nu\in\mathcal{P}(E)} (F(\nu) - I(\nu)) - 4\varepsilon\Big) - \mathbf{1}_{\Omega_n^C} n\|F\|_\infty\Big)\\ &\geq \liminf_{n\to\infty} \Big(p_n\Big(\sup_{\nu\in\mathcal{P}(E)} (F(\nu) - I(\nu)) - 4\varepsilon\Big) - (1-p_n)\|F\|_\infty\Big)\\ &= \sup_{\nu\in\mathcal{P}(E)} (F(\nu) - I(\nu)) - 4\varepsilon,\end{aligned}$$

where the last inequality uses $\beta(\nu, \nu) = 0$ for all $\nu \in \mathcal{P}(E)$, which implies that $\beta_{l_0}^{\pi_0}(\mu^{\pi_0,\pi,l_0}) = 0$ and, hence, $\rho_{l_0}(f) \geq \int_{E^{l_0}} f \, \mathrm{d}\mu^{\pi_0,\pi,l_0}$ for all $f \in C_b(E^{l_0})$.

2.2. The upper bound of Theorem 1.1

2.2.1. *Preliminary results.* The following theorem is essential for the proof of the upper bound. It is based on Proposition 8.2.5 and Theorem 8.2.8 of [20].

Theorem 2.1. *Suppose that Assumption T holds, and let $(\mu^{(n)})_{n\in\mathbb{N}} \subseteq \mathcal{P}(E^n)$ be a sequence of measures such that*

$$\sup_{n\in\mathbb{N}} \frac{1}{n}\beta_n^{\pi_0}(\mu^{(n)}) < \infty.$$

For $n \in \mathbb{N}$, let $X_n = (X_{n,1}, \ldots, X_{n,n})$ be E^n-valued random variables with distribution $\mu^{(n)}$. Define the sequence of $\mathcal{P}(E \times E)$-valued random variables $(\gamma_n)_{n\in\mathbb{N}}$ by

$$\gamma_{n-1} := \frac{1}{n-1}\sum_{i=1}^{n-1} \delta_{X_{n,i}} \otimes \mu_{i,i+1}^{(n)}(X_{n,1}, \ldots, X_{n,i}).$$

It holds that

(i) *$(\gamma_n)_{n\in\mathbb{N}}$ is tight;*

(ii) *for every convergent (in distribution) subsequence of $(\gamma_n)_{n\in\mathbb{N}}$, there exists a probability space $(\bar{\Omega}, \bar{\mathcal{F}}, \bar{\mathbb{P}})$ such that, on this space, there exist random variables $\bar{\gamma}_n \sim \gamma_n$ and $\bar{\gamma} \sim \gamma$ with $\bar{\gamma}_n \xrightarrow{w} \bar{\gamma}$, $\bar{\mathbb{P}}$-a.s. Furthermore, $\bar{\gamma}^{(1)} = \bar{\gamma}^{(2)}$, $\bar{\mathbb{P}}$-a.s., where $\bar{\gamma}^{(1)}$ and $\bar{\gamma}^{(2)}$ are the first and second marginals of $\bar{\gamma}$, respectively.*

Proof. For the proof of (i), there is nothing to show if (T1') holds. So we only consider the case in which (T1) holds. Define the sequence of first marginals $(\tilde{L}_n)_{n\in\mathbb{N}} := (\gamma_n^{(1)})_{n\in\mathbb{N}}$. We first show that $(\tilde{L}_n)_{n\in\mathbb{N}}$ is tight. The idea is to use (T1) which yields a tightness function c on E defined by

$$c(x) := U(x) - \rho^{\pi(x)}(U),$$

and, thus, a tightness function G on $\mathcal{P}(E)$ defined by

$$G(\theta) := \int_E c \, \mathrm{d}\theta,$$

https://doi.org/10.1017/apr.2019.6 Published online by Cambridge University Press

where we refer to Appendix A.3.17 of [20] and the preceding definition, as well as Lemma 8.2.4 of [20] for properties of a tightness function. In the following, we show that $\mathbb{E}[\int_E c\,\mathrm{d}\tilde{L}_n] \le K \in \mathbb{R}$ uniformly in $n \in \mathbb{N}$, which is sufficient to yield the claim since

$$\mathbb{E}\bigg[\int_E c\,\mathrm{d}\tilde{L}_n\bigg] = \int_{\mathcal{P}(E)} \bigg(\int_E c\,\mathrm{d}\theta\bigg) \mathbb{P} \circ \tilde{L}_n^{-1}(\mathrm{d}\theta)$$

and the set $\{Q \in \mathcal{P}(\mathcal{P}(E)) \colon \int_{\mathcal{P}(E)} G(\theta) Q(\mathrm{d}\theta) \le M\}$ is tight for every $M \in \mathbb{R}$ by Lemma 8.2.4 of [20].

In a first step, we assume that U is bounded. Then, for all $x \in E$, by the definition of $\rho^{\pi(x)}$, for all $\nu \in \mathcal{P}(E)$,

$$\beta(\nu, \pi(x)) \ge \int_E U\,\mathrm{d}\nu - \rho^{\pi(x)}(U). \tag{2.6}$$

For $i \in \{1, 2, \dots, n-1\}$, $\mu^{(n)}_{i,i+1}(X_{n,1}, \dots, X_{n,i})$ is a regular conditional distribution of $X_{n,i+1}$ given $\sigma(X_{n,1}, \dots, X_{n,i})$ and, therefore (see, for example, [19, Theorem 10.2.5], where U was bounded),

$$\mathbb{E}[U(X_{n,i+1}) \mid X_{n,1}, \dots, X_{n,i}] = \int_E U\,\mathrm{d}\mu^{(n)}_{i,i+1}(X_{n,1}, \dots, X_{n,i}).$$

We calculate

$$\begin{aligned}
&\mathbb{E}[U(X_{n,i+1}) - U(X_{n,i})]\\
&\quad = \mathbb{E}[\mathbb{E}[U(X_{n,i+1}) | X_{n,1}, \dots, X_{n,i}] - U(X_{n,i})]\\
&\quad = \mathbb{E}\bigg[\int_E U d\mu^{(n)}_{i,i+1}(X_{n,1}, \dots, X_{n,i}) - U(X_{n,i})\bigg]\\
&\quad = \mathbb{E}\bigg[\int_E U d\mu^{(n)}_{i,i+1}(X_{n,1}, \dots, X_{n,i}) - \rho^{\pi(X_{n,i})}\bigg] + \mathbb{E}[\rho^{\pi(X_{n,i})} - U(X_{n,i})]\\
&\quad \overset{(2.6)}{\le} \mathbb{E}[\beta(\mu^{(n)}_{i,i+1}(X_{n,1}, \dots, X_{n,i}), \pi(X_{n,i}))] - \mathbb{E}[c(X_{n,i})].
\end{aligned}$$

Summing the above inequalities over $i \in \{1, 2, \dots, n-1\}$ yields

$$\begin{aligned}
\mathbb{E}[U(X_{n,n}) - U(X_{n,1})] &\le \sum_{i=1}^{n-1} (\mathbb{E}[\beta(\mu^{(n)}_{i,i+1}(X_{n,1}, \dots, X_{n,i}), \pi(X_{n,i}))] - \mathbb{E}[c(X_{n,i})])\\
&\Rightarrow \sum_{i=1}^{n-1} \mathbb{E}[c(X_{n,i})]\\
&\le \mathbb{E}[U(X_{n,1})] + \sum_{i=1}^{n-1} \mathbb{E}[\beta(\mu^{(n)}_{i,i+1}(X_{n,1}, \dots, X_{n,i}), \pi(X_{n,i}))],
\end{aligned}$$

where $\mathbb{E}[U(X_{n,n})] \ge 0$ is used. Dividing the above inequality by $(n-1)$, we obtain

$$\begin{aligned}
&\mathbb{E}\bigg[\int_E c d\tilde{L}_{n-1}\bigg]\\
&\quad = \frac{1}{n-1} \sum_{i=1}^{n-1} \mathbb{E}[c(X_{n,i})]
\end{aligned}$$

https://doi.org/10.1017/apr.2019.6 Published online by Cambridge University Press

$$
\begin{aligned}
&\leq \frac{1}{n-1}\bigg(\mathbb{E}[U(X_{n,1})] + \sum_{i=1}^{n-1}\mathbb{E}[\beta(\mu_{i,i+1}^{(n)}(X_{n,1},\dots,X_{n,i}),\pi(X_{n,i}))]\bigg)\\
&\leq \frac{1}{n-1}\bigg(\beta(\mu_{0,1}^{(n)},\pi_0) + \rho^{\pi_0}(U) + \sum_{i=1}^{n-1}\mathbb{E}[\beta(\mu_{i,i+1}^{(n)}(X_{n,1},\dots,X_{n,i}),\pi(X_{n,i}))]\bigg)\\
&= \frac{1}{n-1}\beta_n^{\pi_0}(\mu^{(n)}) + \frac{1}{n-1}\rho^{\pi_0}(U).
\end{aligned}
$$

The last term of the above inequality chain is uniformly bounded for all $n \geq 2$ by assumption and part (c) of (T1), and we denote this bound by $K \in \mathbb{R}$.

Now, let us show that the above holds for unbounded U. Let $U_k := U \wedge k$ (for $k \in \mathbb{N}$) and $c_k(x) := U_k(x) - \rho^{\pi(x)}(U_k)$. We have shown that

$$\mathbb{E}\bigg[\int_E c_k \,\mathrm{d}\tilde{L}_{n-1}\bigg] \leq \frac{1}{n-1}\beta_n^{\pi_0}(\mu^{(n)}) + \frac{1}{n-1}\rho^{\pi_0}(U_k) \leq \frac{1}{n-1}\beta_n^{\pi_0}(\mu^{(n)}) + \frac{1}{n-1}\rho^{\pi_0}(U).$$

One quickly verifies that $c_k \geq c \wedge (\inf_{\tau\in\mathcal{P}(E)^2}\beta(\tau))$. Indeed, $c_k(x) \geq \inf_{\tau\in\mathcal{P}(E)^2}\beta(\tau)$ if $U(x) \geq k$, and $c_k(x) \geq c(x)$ if $U(x) \leq k$. Hence, the c_k are uniformly bounded below by a constant owing to the lower boundedness of β and (T1). Furthermore, for all $x \in E$, $c(x) = \lim_{k\to\infty} c_k(x)$ by monotone convergence and, therefore, by Fatou's lemma

$$\mathbb{E}\bigg[\int_E c\,\mathrm{d}\tilde{L}_{n-1}\bigg] \leq \liminf_{k\to\infty}\mathbb{E}\bigg[\int_E c_k\,\mathrm{d}\tilde{L}_{n-1}\bigg] \leq \frac{1}{n-1}\beta_n^{\pi_0}(\mu^{(n)}) + \frac{1}{n-1}\rho^{\pi_0}(U) \leq K.$$

This shows that $(\tilde{L}_n)_{n\in\mathbb{N}}$ is tight.

Next, we show that the sequence of second marginals of $(\gamma_n)_{n\in\mathbb{N}}$ is tight, i.e. we prove tightness of the sequence $(\gamma_n^{(2)})_{n\in\mathbb{N}}$ given by $\gamma_{n-1}^{(2)} = (1/(n-1))\sum_{i=1}^{n-1}\mu_{i,i+1}^{(n)}(X_{n,1},\dots,X_{n,i})$. This follows from

$$
\begin{aligned}
\mathbb{E}\bigg[\int_E c\,\mathrm{d}\gamma_n^{(2)}\bigg] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\bigg[\int_E c\,\mathrm{d}\mu_{i,i+1}^{(n+1)}(X_{n+1,1},\dots,X_{n+1,i})\bigg]\\
&\overset{(*)}{=} \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\mathbb{E}[c(X_{n+1,i}) \mid X_{n+1,1},\dots,X_{n+1,i}]]\\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[c(X_{n+1,i})]\\
&= \mathbb{E}\bigg[\int_E c\,\mathrm{d}\tilde{L}_n\bigg]\\
&\leq K,
\end{aligned}
$$

where the last inequality is uniformly in $n \in \mathbb{N}$ as shown above. Note that while equality $(*)$ requires integrability, we can circumvent this requirement by using the same arguments as above, in that we first assume U is bounded and use Fatou's lemma for the transition to the general case.

https://doi.org/10.1017/apr.2019.6 Published online by Cambridge University Press

Tightness of $(\gamma_n)_{n\in\mathbb{N}}$ now follows from tightness of the marginals $(\gamma_n^{(2)})_{n\in\mathbb{N}}$ and $(\tilde{L}_n)_{n\in\mathbb{N}}$.

For part (ii), choose any subsequence still denoted by $(\gamma_n)_{n\in\mathbb{N}}$ that converges in distribution, which means there exists a $\mathcal{P}(E\times E)$-valued random variable γ such that

$$\mathbb{P}\circ\gamma_n^{-1}\xrightarrow{w}\mathbb{P}\circ\gamma^{-1}.$$

With Skorokhod's representation theorem (see, e.g. [24, page 102]), we can go over to a probability space $(\bar{\Omega}, \bar{\mathcal{F}}, \bar{\mathbb{P}})$ such that on this space there exist random variables $\bar{\gamma}_n\sim\gamma_n$ and $\bar{\gamma}\sim\gamma$ with $\bar{\gamma}_n\xrightarrow{w}\bar{\gamma}$, $\bar{\mathbb{P}}$-a.s..

It only remains to show that $\bar{\gamma}^{(1)}=\bar{\gamma}^{(2)}$ holds $\bar{\mathbb{P}}$-a.s. Since $\mu_{i,i+1}^{(n)}(X_{n,1},\dots,X_{n,i})$ is a regular conditional distribution of $X_{n,i+1}$ given $X_{n,1},\dots,X_{n,i}$,

$$\mathbb{E}\biggl[\biggl(f(X_{n,i+1})-\int_E f\,\mathrm{d}\mu_{i,i+1}^{(n)}(X_{n,1},\dots,X_{n,i})\biggr)\,\bigg|\,X_{n,1},\dots,X_{n,i}\biggr]=0$$

for $f\in C_b(E)$, $n\in\mathbb{N}$, and $i\in\{1,\dots,n-1\}$. This means that the terms inside the expectation form (for fixed n) a martingale difference sequence. For ease of notation, we write

$$a_{n,i}:=f(X_{n,i}),\quad\text{and}\quad b_{n,i}:=\int_E f\,\mathrm{d}\mu_{i-1,i}^{(n)}(X_{n,1},\dots,X_{n,i-1}),$$

and get for $n\geq 2$,

$$\begin{aligned}
&\bar{\mathbb{E}}\biggl[\biggl(\int_E f\,\mathrm{d}\bar{\gamma}_{n-1}^{(1)}-\int_E f\,\mathrm{d}\bar{\gamma}_{n-1}^{(2)}\biggr)^2\biggr]\\
&\quad=\mathbb{E}\biggl[\biggl(\int_E f\,\mathrm{d}\gamma_{n-1}^{(1)}-\int_E f\,\mathrm{d}\gamma_{n-1}^{(2)}\biggr)^2\biggr]\\
&\quad=\mathbb{E}\biggl[\biggl(\frac{1}{n-1}\sum_{i=1}^{n-1}a_{n,i}-b_{n,i+1}\biggr)^2\biggr]\\
&\quad=\frac{1}{(n-1)^2}\mathbb{E}\biggl[\biggl((b_{n,1}-b_{n,n})+\biggl(\sum_{i=1}^{n-1}a_{n,i}-b_{n,i}\biggr)\biggr)^2\biggr]\\
&\quad=\frac{1}{(n-1)^2}\mathbb{E}\biggl[(b_{n,1}-b_{n,n})^2+2(b_{n,1}-b_{n,n})\biggl(\sum_{i=1}^{n-1}a_{n,i}-b_{n,i}\biggr)\\
&\qquad\qquad+\biggl(\sum_{i=1}^{n-1}(a_{n,i}-b_{n,i})^2\biggr)\biggr]\\
&\quad\leq\frac{4+8(n-1)+4(n-1)}{(n-1)^2}\|f\|_\infty^2,
\end{aligned}$$

which converges to 0 for $n\to\infty$. By the triangle inequality,

$$\bar{\mathbb{E}}\biggl[\biggl(\int_E f\,\mathrm{d}\bar{\gamma}^{(1)}-\int_E f\,\mathrm{d}\bar{\gamma}^{(2)}\biggr)^2\biggr]=0,$$

which implies that $\int_E f\,\mathrm{d}\bar{\gamma}^{(1)}=\int_E f\,\mathrm{d}\bar{\gamma}^{(2)}$, $\bar{\mathbb{P}}$-a.s. for every $f\in C_b(E)$. By a standard separability argument (cf. [44, Proof of Theorem 3.1]), it follows that $\bar{\gamma}^{(1)}=\bar{\gamma}^{(2)}$, $\bar{\mathbb{P}}$-a.s. □

https://doi.org/10.1017/apr.2019.6 Published online by Cambridge University Press

2.2.2. *Proof of Theorem 1.1: the upper bound.* Let $F\colon \mathcal{P}(E)\to\mathbb{R}$ be bounded and upper semicontinuous. By definition,

$$\frac{1}{n}\rho_n(nF\circ L_n)=\sup_{\mu\in\mathcal{P}(E^n)}\bigg(\int_{E^n}F\circ L_n\,\mathrm{d}\mu-\frac{1}{n}\beta_n^{\pi_0}(\mu)\bigg).$$

Using the boundedness of F, the lower boundedness of β, and the fact that $\beta(\nu,\nu)=0$ for all $\nu\in\mathcal{P}(E)$, we can show that the right-hand side of the above equation is bounded below by $-\|F\|_\infty$ and bounded above by $\|F\|_\infty+\inf_{\tau\in\mathcal{P}(E)^2}|\beta(\tau)|$. Thus, for each $n\in\mathbb{N}$, we can choose $\mu^{(n)}\in\mathcal{P}(E^n)$ such that

$$\frac{1}{n}\rho_n(nF\circ L_n)-\frac{1}{n}\le\int_{E^n}F\circ L_n\,\mathrm{d}\mu^{(n)}-\frac{1}{n}\beta_n^{\pi_0}(\mu^{(n)}) \tag{2.7}$$

and

$$\sup_{n\in\mathbb{N}}\frac{1}{n}\beta_n^{\pi_0}(\mu^{(n)})<\infty.$$

The latter will be used to apply Theorem 2.1 in a few moments. First, we use $\beta(\nu,\nu)=0$ for all $\nu\in\mathcal{P}(E)$ and the convexity of $\beta_2^{\cdot}(\cdot)$ to calculate

$$\begin{aligned}\frac{1}{n}\beta_n^{\pi_0}(\mu^{(n)})&=\frac{1}{n}\beta(\mu_{0,1}^{(n)},\pi_0)+\frac{1}{n}\sum_{i=1}^{n-1}\int_{E^n}\beta(\mu_{i,i+1}^{(n)}(x_1,\dots,x_i),\pi(x_i))\mu^{(n)}(\mathrm{d}x_1,\dots,\mathrm{d}x_n)\\&=\frac{1}{n}\beta_2^{\pi_0}(\mu_{0,1}^{(n)}\otimes\pi)+\int_{E^n}\frac{1}{n}\sum_{i=1}^{n-1}\beta_2^{\delta_{x_i}}(\delta_{x_i}\otimes\mu_{i,i+1}^{(n)}(x_1,\dots,x_i))\mu^{(n)}(\mathrm{d}x_1,\dots,\mathrm{d}x_n)\\&\ge\int_{E^n}\beta_2^{(\pi_0+\sum_{i=1}^{n-1}\delta_{x_i})/n}\bigg(\frac{1}{n}\bigg(\mu_{0,1}^{(n)}\otimes\pi+\sum_{i=1}^{n-1}\delta_{x_i}\otimes\mu_{i,i+1}^{(n)}(x_1,\dots,x_i)\bigg)\bigg)\\&\qquad\times\mu^{(n)}(\mathrm{d}x_1,\dots,\mathrm{d}x_n),\end{aligned} \tag{2.8}$$

where '⊗' denotes the product measure if both arguments are measures.

For $n\in\mathbb{N}$, let $X_n=(X_{n,1},\dots,X_{n,n})$ be E^n-valued random variables with distribution $\mu^{(n)}$. Define the sequence of $\mathcal{P}(E\times E)$-valued random variables $(\gamma_n)_{n\in\mathbb{N}}$ by

$$\gamma_{n-1}:=\frac{1}{n-1}\sum_{i=1}^{n-1}\delta_{X_{n,i}}\otimes\mu_{i,i+1}^{(n)}(X_{n,1},\dots,X_{n,i}).$$

For any subsequence, Theorem 2.1(i) yields a further subsequence (again labeled by $n\in\mathbb{N}$ and fixed for the rest of the proof of the upper bound) such that $(\gamma_n)_{n\in\mathbb{N}}$ converges in distribution. By Theorem 2.1(ii), there exists a probability space $(\bar\Omega,\bar{\mathcal{F}},\bar{\mathbb{P}})$, such that on this space, there exist random variables $\bar\gamma_n\sim\gamma_n$ and $\bar\gamma\sim\gamma$ with $\bar\gamma_n\overset{w}{\to}\bar\gamma$, $\bar{\mathbb{P}}$-a.s. Furthermore, $\bar\gamma^{(1)}=\bar\gamma^{(2)}$, $\bar{\mathbb{P}}$-a.s., where $\bar\gamma^{(1)}$ and $\bar\gamma^{(2)}$ are the first and second marginals of $\bar\gamma$, respectively.

Define the sequence of first marginals of $(\bar\gamma_n)_{n\in\mathbb{N}}$ as $(\bar L_n)_{n\in\mathbb{N}}:=(\bar\gamma_n^{(1)})_{n\in\mathbb{N}}$ and $\bar L:=\bar\gamma^{(1)}$, and note that $\bar L_n\overset{w}{\to}\bar L$, $\bar{\mathbb{P}}$-a.s. With these definitions, (2.7), and (2.8), we obtain

$$\begin{aligned}&\frac{1}{n}\rho_n(nF\circ L_n)-\frac{1}{n}\\&\qquad\le\bar{\mathbb{E}}\bigg[F\bigg(\frac{n-1}{n}\bar L_{n-1}+\frac{1}{n}\delta_{\bar x_{n,n}}\bigg)-\beta_2^{\pi_0/n+(n-1)\bar L_{n-1}/n}\bigg(\frac{\mu_{0,1}^{(n)}\otimes\pi}{n}+\frac{n-1}{n}\bar\gamma_{n-1}\bigg)\bigg],\end{aligned}$$

https://doi.org/10.1017/apr.2019.6 Published online by Cambridge University Press

where the $\bar{X}_{n,n}$ are (redefined) random variables on $(\bar{\Omega}, \bar{\mathcal{F}}, \bar{\mathbb{P}})$ such that $(X_{n,n}, \gamma_{n-1}) \sim (\bar{x}_{n,n}, \bar{\gamma}_{n-1})$ for all $n \in \mathbb{N}$. For ease of notation, define

$$t_{n,0} := \frac{n-1}{n}\bar{L}_{n-1} + \frac{1}{n}\delta_{\bar{x}_{n,n}}, \qquad t_{n,1} := \frac{\pi_0}{n} + \frac{n-1}{n}\bar{L}_{n-1},$$
$$t_{n,2} := \frac{\mu_{0,1}^{(n)} \otimes \pi}{n} + \frac{n-1}{n}\bar{\gamma}_{n-1},$$

and note that $t_{n,0} \xrightarrow{w} \bar{L}$, $t_{n,1} \xrightarrow{w} \bar{L}$, and $t_{n,2} \xrightarrow{w} \bar{\gamma}$, all $\bar{\mathbb{P}}$-a.s. Therefore, by the upper semicontinuity of F and $-\beta_2^{\cdot}(\cdot)$,

$$\begin{aligned}\limsup_{n\to\infty} \frac{1}{n}\rho_n(nF \circ L_n) &\leq \limsup_{n\to\infty} \bar{\mathbb{E}}[F(t_{n,0}) - \beta_2^{t_{n,1}}(t_{n,2})] \\ &\leq \bar{\mathbb{E}}[F \circ \bar{L} - \beta_2^{\bar{L}}(\bar{\gamma})] \\ &= \bar{\mathbb{E}}\left[F \circ \bar{L} - \int_E \beta(\bar{\gamma}_{1,2}(x), \pi(x))\bar{L}(dx)\right] \\ &\leq \sup_{\nu \in \mathcal{P}(E)} \left(F(\nu) - \inf_{q:\nu q = \nu} \int_E \beta(q(x), \pi(x))\nu(\mathrm{d}x)\right),\end{aligned}$$

where in the last inequality we used the fact that $\bar{\gamma}^{(1)} = \bar{\gamma}^{(2)}$ holds $\bar{\mathbb{P}}$-a.s. We have shown that every subsequence has a further subsequence such that this inequality holds, which implies it also holds for the whole sequence.

2.3. Proof of Corollary 1.1

Claim 2.1. If $\beta_2^{\cdot}(\cdot)$ is lower semicontinuous then I is lower semicontinuous. If $\beta_2^{\cdot}(\cdot)$ is convex then I is convex.

Proof. To prove the lower semicontinuity, let $\nu_n \xrightarrow{w} \nu \in \mathcal{P}(E)$. We have to show that

$$\liminf_{n\to\infty} I(\nu_n) \geq I(\nu).$$

Note that I is bounded below. If the left-hand side of the above inequality equals ∞ then there is nothing to prove. So, for any subsequence, we can choose a further subsequence still denoted by $(\nu_n)_{n\in\mathbb{N}}$ such that $I(\nu_n) < \infty$ for all n. Thus, we can choose stochastic kernels q_n such that

$$\beta_2^{\nu_n}(\nu_n \otimes q_n) \leq I(\nu_n) + \frac{1}{n} \quad \text{and} \quad \nu_n q_n = \nu_n.$$

Since $\nu_n q_n = \nu_n$ and the sequence $(\nu_n)_{n\in\mathbb{N}}$ is tight by Prokhorov, the sequence $(\nu_n \otimes q_n)_{n\in\mathbb{N}}$ is tight as well. We go over to a further subsequence still denoted by $(\nu_n \otimes q_n)_{n\in\mathbb{N}}$ such that $\nu_n \otimes q_n \to \nu \otimes q$, where $\nu q = \nu$ follows by convergence of the marginals. By the lower semicontinuity of $\beta_2^{\cdot}(\cdot)$,

$$\liminf_{n\to\infty} I(\nu_n) \geq \liminf_{n\to\infty} \beta_2^{\nu_n}(\nu_n \otimes q_n) - \frac{1}{n} \geq \beta_2^{\nu}(\nu \otimes q) \geq I(\nu).$$

https://doi.org/10.1017/apr.2019.6 Published online by Cambridge University Press

To prove the convexity, note that $I(\nu)=\inf_{\tau\in\mathcal{P}(E^2):\,\tau_1=\tau_2=\nu}\beta_2^{\nu}(\tau)$. Let $\nu_1,\nu_2\in\mathcal{P}(E)$ and $\tau^{(1)},\tau^{(2)}\in\mathcal{P}(E^2)$ with $\tau_1^{(1)}=\tau_2^{(1)}=\nu_1$, and $\tau_1^{(2)}=\tau_2^{(2)}=\nu_2$. Then

$$\begin{aligned}\lambda\beta_2^{\nu_1}(\tau^{(1)})+(1-\lambda)\beta_2^{\nu_2}(\tau^{(2)})&\geq\beta_2^{\lambda\nu_1+(1-\lambda)\nu_2}(\lambda\tau^{(1)}+(1-\lambda)\tau^{(2)})\\&\geq\inf_{\tau\in\mathcal{P}(E^2):\,\tau_1=\tau_2=\lambda\nu_1+(1-\lambda)\nu_2}\beta_2^{\lambda\nu_1+(1-\lambda)\nu_2}(\tau)\\&=I(\lambda\nu_1+(1-\lambda)\nu_2).\end{aligned}$$

Taking the infimum on the left-hand side over all such $\tau^{(1)}$ and $\tau^{(2)}$ yields the claim. □

Claim 2.2. If the upper bound of Theorem 1.1 holds, and, additionally, I has compact sublevel sets, then the upper bound extends to all functions $F\colon\mathcal{P}(E)\to[-\infty,\infty)$ which are upper semicontinuous and bounded from above.

Proof. Let $F\colon\mathcal{P}(E)\to[-\infty,\infty)$ be upper semicontinuous and bounded from above. Define $F_m:=-m\vee F$ $(m\in\mathbb{N})$. By assumption, for all $m\in\mathbb{N}$,

$$\limsup_{n\to\infty}\frac{1}{n}\rho_n(nF\circ L_n)\leq\limsup_{n\to\infty}\frac{1}{n}\rho_n(nF_m\circ L_n)\leq\sup_{\nu\in\mathcal{P}(E)}(F_m(\nu)-I(\nu)),$$

so it remains to only show that

$$\limsup_{m\to\infty}S_m:=\limsup_{m\to\infty}\sup_{\nu\in\mathcal{P}(E)}(F_m(\nu)-I(\nu))\leq\sup_{\nu\in\mathcal{P}(E)}(F(\nu)-I(\nu))=:S.$$

The S_m are decreasing (for increasing m). If $S_m\to-\infty$ there is nothing to show. So assume that the S_m are bounded below by $C\in\mathbb{R}$. Choose $\nu_m\in\mathcal{P}(E)$ such that

$$F_m(\nu_m)-I(\nu_m)\geq S_m-\frac{1}{m}\geq C-1.$$

So the $I(\nu_m)$ are uniformly bounded. By compact sublevel sets of I, for any subsequence, we can choose a further subsequence still denoted by $(\nu_m)_{m\in\mathbb{N}}$ such that $\nu_m\overset{w}{\to}\nu_\infty$ for some $\nu_\infty\in\mathcal{P}(E)$. Then by the upper semicontinuity of F and $-I$,

$$\limsup_{m\to\infty}F_m(\nu_m)-I(\nu_m)\leq F(\nu_\infty)-I(\nu_\infty)\leq S.$$ □

3. Applications to robust Markov chains

3.1. Robust large deviations

In this section (E,d) is assumed to be compact. The main goal of this section is to show Theorem 1.2 and illustrate it in Example 3.1. To this end, we show the respective upper bound in Theorem 3.1 and the respective lower bound in Lemma 3.6. The intermediate results in this section are concerned with representation formulae for the functionals β_n (see Lemmas 3.1 and 3.5) and the verification of conditions (B1) and (B2) (see Lemmas 3.2, 3.3, and 3.4).

In the following part leading up the Theorem 3.1, we assume that π satisfies the Feller property. We work with

$$\beta(\nu,\mu):=\inf_{\hat{\mu}\colon d_W(\mu,\hat{\mu})\leq r}R(\nu,\hat{\mu})=\inf_{\hat{\mu}\in M_1(\mu)}R(\nu,\hat{\mu})$$

for some fixed $r\geq0$. Recall that

$$M_n(\theta):=\{\nu\in\mathcal{P}(E^n)\colon d_W(\nu_{0,1},\theta)\leq r\text{ and }d_W(\nu_{i,i+1}(x_1,\ldots,x_i),\pi(x_i))\leq r,\ \nu\text{-a.s.}\\ \text{for }i=1,\ldots,n-1\}.$$

To be precise, the above definition requires the Condition $d_W(\nu_{i,i+1}(x_1, \dots, x_i), \pi(x_i)) \le r$ to hold for ν-almost all $(x_1, \dots, x_n) \in E^n$ for every decomposition of ν, where the respective ν-null set may depend on the given decomposition. Equivalently, the definition could state that there has to exist one decomposition of ν such that this Condition holds pointwise. That this notion is equivalent follows from the fact that decompositions of ν are only unique up to ν-almost-sure equality.

Lemma 3.1. (See also [3, Lemma 4.4] and [34, Proposition 5.2].) *For all $n \in \mathbb{N}$, it holds that*

$$\beta_n^\theta(\nu) = \inf_{\hat{\mu} \in M_n(\theta)} R(\nu, \hat{\mu}).$$

Proof. Fix $\theta \in \mathcal{P}(E)$. Define the sets

$$Q_0 := \{\hat{\mu} \in \mathcal{P}(E) \colon d_W(\hat{\mu}, \theta) \le r\}$$

and, for $i = 1, \dots, n-1$ and $x_1, \dots, x_i \in E$,

$$Q_i(x_1, \dots, x_i) := \{\hat{\mu} \in \mathcal{P}(E) \colon d_W(\hat{\mu}, \pi(x_i)) \le r\}.$$

We note that $M_n(\theta) = Q_0 \otimes Q_1 \otimes \cdots \otimes Q_{n-1}$, where $Q_0 \otimes Q_1 \otimes \cdots \otimes Q_{n-1}$ is defined as the set of measures $\mu = K_0 \otimes K_1 \otimes K_2 \otimes \cdots \otimes K_{n-1} \in \mathcal{P}(E^n)$, where $K_0 \in Q_0$ and $K_i \colon E^i \to \mathcal{P}(E)$ are Borel measurable kernels such that $K_i(x_1, \dots, x_i) \in Q_i(x_1, \dots, x_i)$ for μ-almost all $x_1, \dots, x_i$. Since, for all $i = 1, \dots, n$, the set $\{(x_1, \dots, x_i, \hat{\mu}) \in E^i \times \mathcal{P}(E) \colon \hat{\mu} \in Q_i(x_1, \dots, x_i)\}$ is trivially Borel, a measurable selection argument (e.g. [6, Proposition 7.50]) yields, for $\nu \in \mathcal{P}(E^n)$,

$$\begin{aligned}
&\inf_{\hat{\mu} \in M_n(\theta)} R(\nu, \hat{\mu}) \\
&\quad = \inf_{K_0 \otimes \cdots \otimes K_{n-1} \in Q_0 \otimes \cdots \otimes Q_{n-1}} \sum_{i=0}^{n-1} \int_{E^n} R(\nu_{i,i+1}(x_1, \dots, x_i), K_i(x_1, \dots, x_i)) \nu(\mathrm{d}x_1, \dots, \mathrm{d}x_n) \\
&\quad \overset{(*)}{=} \sum_{i=0}^{n-1} \int_{E^n} \inf_{\hat{\mu} \in Q_i(x_1, \dots, x_i)} R(\nu_{i,i+1}(x_1, \dots, x_i), \hat{\mu}) \nu(\mathrm{d}x_1, \dots, \mathrm{d}x_n) \\
&\quad = \beta(\nu_{0,1}, \theta) + \sum_{i=1}^{n-1} \int_{E^n} \beta(\nu_{i,i+1}(x_1, \dots, x_i), \pi(x_i)) \nu(\mathrm{d}x_1, \dots, \mathrm{d}x_n) \\
&\quad = \beta_n^\theta(\nu),
\end{aligned}$$

where rigorously step $(*)$ works inductively; see the proofs of [3, Lemma 4.4] and [34, Proposition 5.2]. □

Lemma 3.2. *Let $\theta_1, \theta_2 \in \mathcal{P}(E)$, $\nu_1 \in M_2(\theta_1)$, $\nu_2 \in M_2(\theta_2)$, and $\lambda \in (0, 1)$. Then*

$$\lambda \nu_1 + (1 - \lambda)\nu_2 \in M_2(\lambda\theta_1 + (1 - \lambda)\theta_2).$$

Proof. Write $\nu_1 = \mu_1 \otimes K_1$ and $\nu_2 = \mu_2 \otimes K_2$ for some $\mu_1, \mu_2 \in \mathcal{P}(E)$ and K_1, K_2 stochastic kernels on E. Furthermore, K_1 and K_2 are chosen such that $d_W(K_i(x), \pi(x)) \le r$ for all $x \in E$ and $i \in \{1, 2\}$. We have the equality

$$\lambda \nu_1 + (1 - \lambda)\nu_2 = (\lambda\mu_1 + (1 - \lambda)\mu_2) \otimes K, \tag{3.1}$$

https://doi.org/10.1017/apr.2019.6 Published online by Cambridge University Press

where $K\colon E \to \mathcal{P}(E)$ is defined by

$$\begin{aligned} K(x) &= \frac{\mathrm{d}\mu_1}{\mathrm{d}(\lambda\mu_1 + (1-\lambda)\mu_2)}(x)\lambda K_1(x) + \frac{\mathrm{d}\mu_2}{\mathrm{d}(\lambda\mu_1 + (1-\lambda)\mu_2)}(x)(1-\lambda)K_2(x) \\ &=: \lambda_x K_1(x) + (1-\lambda_x)K_2(x). \end{aligned}$$

Equation (3.1) obviously holds for Borel sets of the form $A \times B \subseteq E^2$, which extends the equality to arbitrary Borel sets by Carathéodory. So K is a pointwise convex combination of K_1 and K_2. Since, for the first Wasserstein distance, the Kantorovich duality (see, e.g. [45, Chapter 5]) implies that

$$d_W(\lambda\mu_1 + (1-\lambda)\mu_2, \lambda\theta_1 + (1-\lambda)\theta_2) \le \lambda d_W(\mu_1, \theta_1) + (1-\lambda)d_W(\mu_2, \theta_2) \le r$$

and, for all $x \in E$,

$$\begin{aligned} &d_W(\lambda_x K_1(x) + (1-\lambda_x)K_2(x),\, \lambda_x\pi(x) + (1-\lambda_x)\pi(x)) \\ &\quad \le \lambda_x d_W(K_1(x), \pi(x)) + (1-\lambda_x)d_W(K_2(x), \pi(x)) \\ &\quad \le r, \end{aligned}$$

the claim follows. □

That $\beta_2^{\cdot}(\,\cdot\,)$ is convex follows by the previous lemma and the convexity of $R(\cdot, \cdot)$, since

$$\begin{aligned} &\beta_2^{\lambda\theta_1 + (1-\lambda)\theta_2}(\lambda\nu_1 + (1-\lambda)\ \nu_2) \\ &\quad = \inf_{\hat{\mu} \in M_2(\lambda\theta_1 + (1-\lambda)\theta_2)} R(\lambda\nu_1 + (1-\lambda)\nu_2, \hat{\mu}) \\ &\quad \overset{(3.2)}{\le} \inf_{\hat{\mu}_1 \in M_2(\theta_1),\, \hat{\mu}_2 \in M_2(\theta_2)} R(\lambda\nu_1 + (1-\lambda)\nu_2, \lambda\hat{\mu}_1 + (1-\lambda)\hat{\mu}_2) \\ &\quad \le \inf_{\hat{\mu}_1 \in M_2(\theta_1),\, \hat{\mu}_2 \in M_2(\theta_2)} \lambda R(\nu_1, \hat{\mu}_1) + (1-\lambda)R(\nu_2, \hat{\mu}_2) \\ &\quad = \lambda\beta_2^{\theta_1}(\nu_1) + (1-\lambda)\beta_2^{\theta_2}(\nu_2) \end{aligned}$$

It remains to show that $\beta_2^{\cdot}(\,\cdot\,)$ is lower semicontinuous. To this end, we first show the following.

Lemma 3.3. *If π satisfies the Feller property then $M_2(\theta)$ is closed.*

Proof. Recall that $\mu \otimes K \in M_2(\theta)$ if and only if both

$$d_W(\mu, \theta) \le r, \tag{3.2}$$

$$d_W(K(x), \pi(x)) \le r \quad \text{for } \mu\text{-a.a. } x \in E. \tag{3.3}$$

Condition (3.2) is closed (which is obvious once it is rewritten via the Kantorovich duality), so we focus on Condition (3.3). Since, by assumption, (E, d) is compact and thus totally bounded, the set of Lipschitz-1 functions mapping E onto $\mathbb{R}$ which are absolutely bounded by 1 (denoted by Lip_1) is separable with respect to the sup-norm (follows since the space of uniformly bounded and continuous functions is separable and every subset of a separable metric space is again separable). We denote by $\{f_1, f_2, \ldots\} \subseteq \mathrm{Lip}_1$ a countable dense subset. Furthermore, we are going to use the fact that, for bounded and measurable functions $h\colon E \to \mathbb{R}$ and $\nu \in \mathcal{P}(E)$,

$$(h \ge 0,\ \nu\text{-a.s.}) \quad \Longleftrightarrow \quad \left(\text{for all } g \in C_b(E),\, g \ge 0\colon \int_E g(x)h(x)\nu(\mathrm{d}x) \ge 0\right),$$

which holds because E is a Polish space and thus the function $\mathbf{1}_A$ for the Borel set $A := \{h < 0\}$ can be approximated in $L_1(\nu)$ by a sequence of nonnegative, continuous, and bounded functions.

We can rewrite Condition (3.3) as follows

$$\begin{aligned}
&d_W(K(x), \pi(x)) \le r \text{ for } \mu\text{-a.a. } x \in E\\
\iff\ &\left(\text{for all } f \in \mathrm{Lip}_1\colon \int_E f\,\mathrm{d}K(x) - \int_E f\,\mathrm{d}\pi(x) \le r\right) \quad \text{for } \mu\text{-a.a. } x \in E\\
\iff\ &\left(\text{for all } i \in \mathbb{N}\colon \int_E f_i\,\mathrm{d}K(x) - \int_E f_i\,\mathrm{d}\pi(x) \le r\right) \quad \text{for } \mu\text{-a.a. } x \in E\\
\iff\ &\Bigg(\text{for all } i \in \mathbb{N} \text{ and all } g \in C_b(E),\ g \ge 0\colon\\
&\quad \int_E g(x)\left(\int_E f_i(y)K(x,\ \mathrm{d}y) - \int_E f_i(y)\pi(x,\ \mathrm{d}y) - r\right)\mu(\mathrm{d}x) \le 0\Bigg)\\
\iff\ &\Bigg(\text{for all } i \in \mathbb{N} \text{ and all } g \in C_b(E),\ g \ge 0\colon \int_{E^2} g(x)f_i(y)\mu \otimes K(\mathrm{d}x,\ \mathrm{d}y)\\
&\quad - \int_E g(x)\left(\int_E f_i\,\mathrm{d}\pi(x) - r\right)\mu(\mathrm{d}x) \le 0\Bigg).
\end{aligned}$$

The last line expresses a closed Condition if π satisfies the Feller property, which guarantees that $x \mapsto \int_E f\,\mathrm{d}\pi(x)$ is continuous for all $f \in C_b(E)$. □

Lemma 3.4. *The function $\beta_2^{\cdot}(\cdot)$ is lower semicontinuous.*

Proof. Let $(\theta_n, \nu_n) \xrightarrow{w} (\theta, \nu) \in \mathcal{P}(E) \times \mathcal{P}(E^2)$ as $n \to \infty$. We have to show that

$$\liminf_{n\to\infty} \beta_2^{\theta_n}(\nu_n) \ge \beta_2^{\theta}(\nu),$$

which is done by choosing an arbitrary subsequence and showing that there exists a further subsequence such that the inequality holds. So we start with a subsequence still denoted by $(\theta_n, \nu_n)_{n\in\mathbb{N}}$. Let $\hat{\mu}_n \in M_2(\theta_n)$ such that

$$\beta_2^{\theta_n}(\nu_n) \ge R(\nu_n, \hat{\mu}_n) - \frac{1}{n},$$

and choose a further subsequence still denoted by $(\theta_n, \nu_n)_{n\in\mathbb{N}}$ such that $d_W(\theta_n, \theta) \le 1/n$ and $\hat{\mu}_n$ converges weakly to some $\hat{\mu} \in \mathcal{P}(E^2)$. We show that $\hat{\mu} \in M_2(\theta)$. To this end, define

$$M_2^{r,n}(\theta) := \left\{\mu \otimes K \in \mathcal{P}(E^2)\colon d_W(\mu, \theta) \le r + \frac{1}{n},\ d_W(K(x), \pi(x)) \le r \text{ for } \mu\text{-a.a. } x \in E\right\},$$

which is closed, as the proof of the previous lemma trivially carries over to this set. We see that $\hat{\mu}_m \in M_2^{r,n}(\theta)$ for all $m \ge n$, and, therefore, $\hat{\mu} \in M_2^{r,n}(\theta)$ for all $n \in \mathbb{N}$, which yields $\hat{\mu} \in M_2(\theta)$. Finally, by the lower semicontinuity of $R(\cdot, \cdot)$, we obtain

$$\liminf_{n\to\infty} \beta_2^{\theta_n}(\nu_n) \ge \liminf_{n\to\infty} R(\nu_n, \hat{\mu}_n) \ge R(\nu, \hat{\mu}) \ge \inf_{\mu\in M_2(\theta)} R(\nu, \mu) = \beta_2^{\theta}(\nu). \qquad \square$$

https://doi.org/10.1017/apr.2019.6 Published online by Cambridge University Press

The rate function I corresponding to the choice of β as defined at the beginning of the section is given by

$$I(\nu) := \inf_{q:\nu q=\nu} \int_E \inf_{K_x \in M(\pi(x))} R(q(x), K_x)\nu(\mathrm{d}x) \quad \text{for } \nu \in \mathcal{P}(E).$$

Using the above observations to apply the main theorem, we obtain the following result.

Theorem 3.1. *For all functions $F\colon \mathcal{P}(E) \to [-\infty, \infty)$ which are upper semicontinuous and bounded from above, it holds that*

$$\limsup_{n\to\infty} \sup_{\mu \in M_n(\pi_0)} \frac{1}{n} \ln \int_{E^n} \exp(nF \circ L_n)\,\mathrm{d}\mu \le \sup_{\nu \in \mathcal{P}(E)} (F(\nu) - I(\nu)).$$

Furthermore, for all closed sets $A \subseteq \mathcal{P}(E)$, it holds that

$$\limsup_{n\to\infty} \sup_{\mu \in M_n(\pi_0)} \frac{1}{n} \ln \mu(L_n \in A) \le -\inf_{\nu \in A} I(\nu).$$

Proof. For the first claim, apply Theorem 1.1, which by the compactness of E and, thus, by Corollary 1.1 extends to all functions $F\colon \mathcal{P}(E) \to [-\infty, \infty)$ which are upper semicontinuous and bounded from above. To arrive at the given form of ρ_n, we use Lemma 3.1 to obtain, for $f \in C_b(E^n)$,

$$\begin{aligned} \rho_n(f) &= \sup_{\nu \in \mathcal{P}(E^n)} \left(\int_{E^n} f\,\mathrm{d}\nu - \inf_{\mu \in M_n(\pi_0)} R(\nu, \mu) \right) \\ &= \sup_{\mu \in M_n(\pi_0)} \sup_{\nu \in \mathcal{P}(E^n)} \left(\int_{E^n} f\,\mathrm{d}\nu - R(\nu, \mu) \right) \\ &= \sup_{\mu \in M_n(\pi_0)} \ln \int_{E^n} \exp(f)\,\mathrm{d}\mu, \end{aligned}$$

where the last step follows by the Gibbs variational formula for the relative entropy.

With the first claim established, the second claim follows by choosing $F = -\infty\mathbf{1}_{A^C}$ for a closed set $A \subseteq \mathcal{P}(E)$. □

For the large deviations bound in Theorem 3.1 to be nonvacuous for a closed set, $A \subseteq \mathcal{P}(E)$ requires that

$$\inf_{\nu \in A} I(\nu) > 0. \tag{3.4}$$

Intuitively, (3.4) holds if and only if, for all pairs $\nu \in A$ and q with $\nu q = \nu$, there is some Borel set $S \subseteq E$ with $\nu(S) > 0$ such that $d_W(q(x), \pi(x)) > r$ for all $x \in S$.

To properly address the question of whether the attained bound is sharp, we need a lower bound in accordance with the upper bound. The choice of β that leads to Theorem 3.1 cannot yield a lower bound with our approach, since Condition (B3) is not satisfied for $r > 0$ and, hence, the lower bound of Theorem 1.1 cannot be applied.

In the following we therefore consider the functional $\underline{\beta}$, which is chosen such that it resembles β and satisfies (B3), albeit at the cost of not satisfying (B2). This will lead to the lower bound of Theorem 1.2 proven in Lemma 3.6. Define

$$\underline{\beta}(\nu, \mu) := \inf_{\hat{\mu}\colon d_W(\mu, \hat{\mu}) \le r,\ \hat{\mu} \ll \mu} R(\nu, \hat{\mu}), \qquad \underline{M}_n(\theta) := \{\nu \in M_n(\theta)\colon \nu \ll \theta \otimes \pi \otimes \cdots \otimes \pi\},$$

$$\underline{I}(\nu) := \inf_{q:\nu q=\nu} \int_E \inf_{K_x \in \underline{M}(\pi(x))} R(q(x), K(x))\nu(\mathrm{d}x)$$

https://doi.org/10.1017/apr.2019.6 Published online by Cambridge University Press

Furthermore, we assume that, for the analysis of the lower bound, π satisfies Assumption M, but no longer has to satisfy the Feller property.

Lemma 3.5. (See also [3, Lemma 4.4] and [34, Proposition 5.2].) *For all $n \in \mathbb{N}$, it holds that*

$$\underline{\beta}_n^{\theta}(\nu) = \inf_{\mu \in \underline{M}_n(\theta)} R(\nu, \mu).$$

Proof. The proof is the same as that of Lemma 3.1, except here we need measurability of the sets

$$S_i := \{(x_1, \ldots, x_i, \hat{\mu}) \in E^i \times \mathcal{P}(E) : d_W(\hat{\mu}, \pi(x_i)) \leq r \text{ and } \hat{\mu} \ll \pi(x_i)\}$$

for $i \in \{1, \ldots, n-1\}$. That these sets are indeed Borel measurable can be seen as follows: Define the function $g\colon \mathcal{P}(E) \times \mathcal{P}(E) \times E \to \mathbb{R}_+$ by

$$g(\mu, \nu, x) = \frac{\mathrm{d}\mu_{|\nu}}{\mathrm{d}\nu}(x).$$

Here $\mu_{|\nu}$ denotes the absolutely continuous part of μ with respect to ν as given by Lebesgue's decomposition theorem. Then g is Borel as shown in [15, V.58 and subsequent remark]. We have $\mu \ll \nu$ if and only if $\int_E g(\mu, \nu, \cdot)\,\mathrm{d}\nu = 1$, which shows that S_i is Borel (as the other conditions that define S_i are trivially Borel).

To arrive at the given form of $\underline{M}_n(\theta)$, we use the following equivalence for measures $\nu_1, \nu_2 \in \mathcal{P}(E)$ and stochastic kernels $K_1, K_2\colon E \to \mathcal{P}(E)$ (see, e.g. [3, Lemma A.2]):

$$\begin{aligned}&(\nu_1 \otimes K_1 \ll \nu_2 \otimes K_2 \in \mathcal{P}(E^2))\\ &\qquad\Longleftrightarrow \quad (\nu_1 \ll \nu_2 \text{ and } K_1(x) \ll K_2(x) \text{ for } \nu_1\text{-almost all } x \in E).\end{aligned}$$ □

In complete analogy to the choice of β leading to Theorem 3.1, we see that $\underline{\beta}$ satisfies (B1), which is a consequence of Lemma 3.5 in combination with Lemma 3.2, where one additionally uses the fact that

$$\begin{aligned}&(\mu_1 \ll \theta_1 \text{ and } \mu_2 \ll \theta_2)\\ &\quad\Longrightarrow \quad \lambda\mu_1 + (1-\lambda)\mu_2 \ll \lambda\theta_1 + (1-\lambda)\theta_2 \quad \text{for } \mu_1, \mu_2, \theta_1, \theta_2 \in \mathcal{P}(E),\ \lambda \in (0, 1).\end{aligned}$$

As (B3) and Assumption M are satisfied as well, Theorem 1.1 yields, for all $F \in C_b(\mathcal{P}(E))$,

$$\liminf_{n\to\infty} \sup_{\mu \in \underline{M}_n(\pi_0)} \frac{1}{n} \int_{E^n} \exp(F \circ L_n)\,\mathrm{d}\mu \geq \sup_{\nu \in \mathcal{P}(E)} (F(\nu) - \underline{I}(\nu)),$$

which leads to the following lemma.

Lemma 3.6. *Let Assumption M be satisfied. For $G \subseteq \mathcal{P}(E)$ open, it holds that*

$$\liminf_{n\to\infty} \sup_{\mu \in \underline{M}_n(\pi_0)} \frac{1}{n} \ln \mu(L_n \in G) \geq -\inf_{\nu \in G} \underline{I}(\nu).$$

Proof. The proof is an adapted version of [20, Theorem 1.2.3].

We work with the Laplace principle lower bound stated just before the lemma. Without loss of generality, assume that $\inf_{\nu \in G} \underline{I}(\nu) < \infty$. Let $\nu \in G$ such that $\underline{I}(\nu) < \infty$. Choose $M \in \mathbb{R}$

https://doi.org/10.1017/apr.2019.6 Published online by Cambridge University Press

such that $\underline{I}(\nu) < M$ and $k \in \mathbb{N}$ such that $B(\nu, 1/k) := \{\mu \in \mathcal{P}(E) \colon \hat{d}(\mu, \nu) \leq 1/k\} \subseteq G$, where $\hat{d}$ is some metric on $\mathcal{P}(E)$ compatible with weak convergence. Define

$$h(\theta) := -M((\hat{d}(\nu, \theta) \cdot k) \wedge 1).$$

We find that $-M \leq h \leq 0$, $h(\nu) = 0$, and $h(\theta) = -M$ for $\theta \in B(\nu, 1/k)^C$. Thus, for any $\mu \in \mathcal{P}(E^n)$,

$$\int_{E^n} \exp(nh \circ L_n)\, \mathrm{d}\mu \leq \exp(-nM) + \mu(L_n \in B(\nu, \delta)) \leq \max\{2\exp(-nM), 2\mu(L_n \in B(\nu, \delta))\}.$$

Therefore,

$$\begin{aligned}
&\max\left\{\liminf_{n\to\infty} \sup_{\mu \in \underline{M}_n(\pi_0)} \frac{1}{n} \ln \mu(L_n \in B(\nu, \delta)), -M\right\} \\
&\geq \liminf_{n\to\infty} \sup_{\mu \in \underline{M}_n(\pi_0)} \frac{1}{n} \ln \int_{E^n} \exp(nh \circ L_n)\, \mathrm{d}\mu \\
&\geq \sup_{\hat{\nu} \in \mathcal{P}(E)} (h(\hat{\nu}) - \underline{I}(\hat{\nu})) \\
&\geq h(\nu) - \underline{I}(\nu) \\
&= -\underline{I}(\nu).
\end{aligned}$$

Since $M > I(\nu)$,

$$\liminf_{n\to\infty} \sup_{\mu \in \underline{M}_n(\pi_0)} \frac{1}{n} \ln \mu(L_n \in B(\nu, \delta)) \geq -\underline{I}(\nu),$$

and using the facts that $B(\nu, \delta) \subseteq G$ and the above reasoning works for all $\nu \in G$ with $\underline{I}(\nu) < \infty$, we obtain the claim. □

The proof of Theorem 1.2 is now complete, as it follows from Theorem 3.1 and Lemma 3.6.

The following illustrates the obtained results. Note that to calculate the rates, as is usual in large deviations theory, the necessary minimization can be solved efficiently (at least in theory) over convex sets A, since I is convex.

Example 3.1. Consider the state space $\{1, 2, 3\}$ with discrete metric, i.e. $d(i, j) = 0$ if $i = j$ and $d(i, j) = 1$ otherwise. The Markov chain is given by its initial distribution $\pi_0 = \delta_3$ and transition kernel π with matrix representation

$$\begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.3 & 0.4 & 0.3 \\ 0 & 0.3 & 0.7 \end{bmatrix}.$$

Suppose that we are interested in the tail event that the empirical measure L_n under the Markov chain is close (in a certain sense) to the initial distribution π_0. We are uncertain of the precise model specification of the Markov chain and want to find the worst case (i.e. slowest possible) convergence rate to 0 of this tail event.

Formally, let $r = 0.05$ and take, for $\kappa = 0.2$, the set of measures $A = B_{d_W}(\delta_3, \kappa)$, i.e. the Wasserstein-1-ball around δ_3 with radius κ. The set $\{L_n \in A\}$ models the abovementioned tail event. What is the (exponential) asymptotic rate of convergence of

$$\sup_{\mu \in M_n(\delta_3)} \mu(L_n \in A) \to 0 \tag{3.5}$$

as $n \to \infty$? Note that r and the transition kernel are as always implicitly included in $M_n(\delta_3)$.

https://doi.org/10.1017/apr.2019.6 Published online by Cambridge University Press

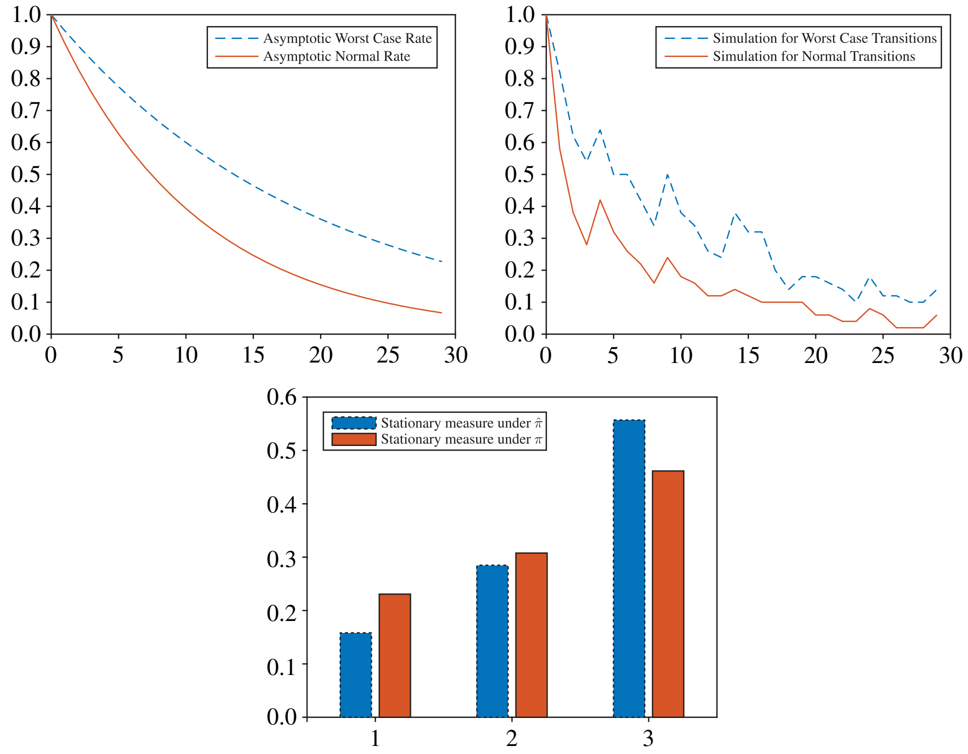

FIGURE 1: Illustration of convergence rates, simulated (100 paths) realized convergence, and the stationary distributions under the normal Markov chain and the robust worst-case Markov chain.

Calculating the upper bound of Theorem 1.2 yields a worst-case exponential rate:

$$r_{\text{worst case}} \approx 0.0511.$$

This is significantly lower than the normal rate for the Markov chain without the robustness (i.e. the $r = 0$ case), which is

$$r_{\text{normal}} \approx 0.0910.$$

In Figure 1 we present the difference in convergence speed. Notably, the optimizer of the optimization problem to obtain the worst-case rate also yields a kernel $\hat{\pi}$ such that $\pi_0 \otimes \hat{\pi} \otimes \cdots \otimes \hat{\pi} \in M_n(\pi_0)$ and the Markov chain with transition kernel $\hat{\pi}$ attains the worst-case rate, i.e.

$$\pi_0 \otimes \hat{\pi} \otimes \cdots \otimes \hat{\pi}(L_n \in A) \sim \exp\left(-n \cdot r_{\text{worst case}}\right).$$

In other words, the worst-case rate in (3.5) is obtained and one sequence of optimal measures is Markovian, with transition kernel $\hat{\pi}$ given by the matrix

$$\begin{bmatrix} 0.6 - r & 0.2 & 0.2 + r \\ 0.3 - r & 0.4 & 0.3 + r \\ 0 & 0.3 - r & 0.7 + r \end{bmatrix}.$$

In Figure 1 we also present a simulated convergence rate for both the initial Markov chain and the Markov chain with worst-case transition kernel $\hat{\pi}$ (100 paths simulated), and a comparison of the respective stationary distributions.

https://doi.org/10.1017/apr.2019.6 Published online by Cambridge University Press

Note that in the above example the rates are asymptotically sharp, as the worst-case kernel $\hat{\pi}$ for the rate function is already absolutely continuous with respect to π, so using $\underline{I}$ instead of I yields the same rate.

Using the above example, we can get an idea when the upper and lower bounds of Theorem 1.2 may not coincide. If we do not restrict ourselves to $\underline{I}$, it may happen that no optimal kernel $\hat{\pi}$ is absolutely continuous with respect to the initial kernel π. In this case, we can no longer guarantee that some near optimal kernel $\hat{\pi}$ satisfies Condition (M1), which is also needed in the nonrobust case to show the large deviations lower bound.

3.2. Robust weak law of large numbers

Let (E, d) be compact. In this section, Theorem 1.3 is proven. We first show the upper bound in Theorem 3.2 and then explain how to obtain the lower bound.

Up to Theorem 3.2, let π satisfy the Feller property. Define

$$\beta(\mu, \nu) := \begin{cases} 0 & \text{if } d_W(\mu, \nu) \leq r, \\ \infty & \text{otherwise,} \end{cases} .$$

for some $r \geq 0$ and obtain

$$\beta_n^{\theta}(\nu) = \infty \cdot \mathbf{1}_{(M_n(\theta))^C}(\nu).$$

Lemma 3.7. *The function $\beta_2^{\cdot}(\cdot)$ is convex and lower semicontinuous.*

Proof. We first show convexity. Let $\theta_1, \theta_2 \in \mathcal{P}(E)$, $\nu_1, \nu_2 \in \mathcal{P}(E^2)$, and $\lambda \in (0, 1)$. We have to show that

$$\beta_2^{\lambda\theta_1+(1-\lambda)\theta_2}(\lambda\nu_1 + (1-\lambda)\nu_2) \leq \lambda\beta_2^{\theta_1}(\nu_1) + (1-\lambda)\beta_2^{\theta_2}(\nu_2).$$

To this end, it suffices to show that if the right-hand side is 0, the left-hand side has to be 0 as well. If the right-hand side is 0 then both $\nu_1 \in M_2(\theta_1)$ and $\nu_2 \in M_2(\theta_2)$. It follows by Lemma 3.2 that $\lambda\nu_1 + (1-\lambda)\nu_2 \in M_2(\lambda\theta_1 + (1-\lambda)\theta_2)$ and, thus, the left-hand side is also 0.

We now show lower semicontinuity. Let $(\theta_n, \nu_n) \xrightarrow{w} (\theta, \nu) \in \mathcal{P}(E) \times \mathcal{P}(E^2)$. We have to show that

$$\liminf_{n\to\infty} \beta_2^{\theta_n}(\nu_n) \geq \beta_2^{\theta}(\nu).$$

Without loss of generality, the left-hand side is not equal to ∞. We have to show that the right-hand side is 0. We first choose an arbitrary subsequence and then a further subsequence still denoted by $(\theta_n, \nu_n)_{n\in\mathbb{N}}$ such that, for all $n \in \mathbb{N}$,

$$\beta_2^{\theta_n}(\nu_n) < \infty, \qquad d_W(\theta_n, \theta) \leq \frac{1}{n}.$$

It follows that $\nu_n \in M_2(\theta_n)$ for all $n \in \mathbb{N}$ and with the same notation and arguments as in the proof of Lemma 3.4, it follows that $\nu \in M_2^{r+1/n}(\theta)$ for all $n \in \mathbb{N}$ and, thus, $\nu \in M_2(\theta)$, i.e. $\beta_2^{\theta}(\nu) = 0$. □

By applying Theorem 1.1 and Corollary 1.1 we obtain the following result.

Theorem 3.2. *For all upper semicontinuous and bounded from above functions $F: \mathcal{P}(E) \to [-\infty, \infty)$, it holds that*

$$\limsup_{n\to\infty} \sup_{\mu\in M_n(\pi_0)} \int_{E^n} F \circ L_n \,\mathrm{d}\mu \leq \sup_{\nu\in\mathcal{P}(E):\, \textit{there exists } q, \nu q=\nu \textit{ such that } \nu\otimes q\in M_2(\nu)} F(\nu).$$

https://doi.org/10.1017/apr.2019.6 Published online by Cambridge University Press

We now focus on the lower bound in Theorem 1.3. Let π satisfy Assumption M (but no longer has to satisfy the Feller property). We define

$$\underline{\beta}(\mu, \nu) := \begin{cases} 0 & \text{if } d_W(\mu, \nu) \leq r \text{ and } \mu \ll \nu, \\ \infty & \text{otherwise,} \end{cases}$$

so that (B3) holds. We obtain

$$\underline{\beta}_n^{\theta}(\nu) = \infty \cdot \mathbf{1}_{(\underline{M}_n(\theta))^C}(\nu).$$

Proving (B1) for $\underline{\beta}$ works completely analogous to the case of β in Lemma 3.7 by replacing M_n by $\underline{M}_n$. Applying Theorem 1.1 yields

$$\liminf_{n \to \infty} \sup_{\mu \in \underline{M}_n(\pi_0)} \int_{E^n} F \circ L_n \, \mathrm{d}\mu \geq \sup_{\nu \in \mathcal{P}(E)\colon \text{there exists } q, \nu q = \nu \text{ such that } \nu \otimes q \in \underline{M}_2(\nu)} F(\nu)$$

for all $F \in C_b(\mathcal{P}(E))$. Theorem 1.3 is shown.

Acknowledgements

The author would like to thank Daniel Bartl, Michael Kupper, Daniel Lacker, and an anonymous referee for their valuable comments and suggestions. In particular, thanks to Daniel Lacker for providing the proof of Lemma 3.3.

References

[1] Acciaio, B. and Penner, I. (2011). Dynamic risk measures. In *Advanced Mathematical Methods for Finance*, Springer, Heidelberg, pp. 1–34.

[2] Agueh, M. and Carlier, G. (2011). Barycenters in the Wasserstein space. *SIAM J. Math. Anal.* **43,** 904–924.

[3] Bartl, D. (2016). Exponential utility maximization under model uncertainty for unbounded endowments. Preprint. Available at https://arxiv.org/abs/1610.00999.

[4] Bartl, D. (2016). Pointwise dual representation of dynamic convex expectations. Preprint. Available at https://arxiv.org/abs/1612.09103v1.

[5] Bartl, D., Drapeau, S. and Tangpi, L. (2017). Computational aspects of robust optimized certainty equivalents. Preprint. Available at https://arxiv.org/abs/1706.10186v1.

[6] Bertsekas, D. P. and Shreve, S. E. (1996). *Stochastic Optimal Control: The Discrete-Time Case*. Athena Scientific.

[7] Blanchet, A. and Carlier, G. (2016). Optimal transport and Cournot-Nash equilibria. *Math. Operat. Res.* **41,** 125–145.

[8] Blanchet, J. and Murthy, K. R. A. (2016). Quantifying distributional model risk via optimal transport. Preprint. Available at https://arxiv.org/abs/1604.01446.

[9] Breiman, L. (1992). *Probability* (Classics Appl. Math. **7**). Society for Industrial and Applied Mathematics, Philadelphia, PA.

[10] Cerreia-Vioglio, S., Maccheroni, F. and Marinacci, M. (2016). Ergodic theorems for lower probabilities. *Proceedings of the American Mathematical Society* **144,** 3381–3396.

[11] Cheridito, P. (2013). Convex analysis. Princeton University lecture notes.

[12] Cheridito, P. and Kupper, M. (2011). Composition of time-consistent dynamic monetary risk measures in discrete time. *Internat. J. Theoret. Appl. Finance* **14,** 137–162.

[13] De Acosta, A. (1990). Large deviations for empirical measures of Markov chains. *J. Theoret. Prob.* **3,** 395–431.

[14] De Cooman, G., Hermans, F. and Quaeghebeur, E. (2009). Imprecise Markov chains and their limit behavior. *Prob. Eng. Inf. Sci.* **23,** 597–635.

[15] Dellacherie, C. and Meyer, P.-A. (1982). *Probability and Potential B: Theory of Martingales*. Elsevier, Amsterdam.

[16] Dembo, A. and Zeitouni, O. (2010). *Large Deviations Techniques and Applications* (Stoch. Modelling Appl. Prob. **38**). Springer, Berlin.

[17] Donsker, M. D. and Varadhan, S. R. S. (1975). Asymptotic evaluation of certain Markov process expectations for large time, I. *Commun. Pure Appl. Math.* **28,** 1–47.

[18] Donsker, M. D. and Varadhan, S. R. S. (1976). Asymptotic evaluation of certain markov process expectations for large time. III. *Commun. Pure Appl. Math.* **29,** 389–461.

[19] Dudley, R. M. (2002). *Real Analysis and Probability* (Cambridge Studies in Advanced Mathematics **74**). Cambridge University Press.

[20] Dupuis, P. and Ellis, R. S. (2011). *A Weak Convergence Approach to the Theory of Large Deviations*. John Wiley, Hoboken, NJ.

[21] Eckstein, S. (2019). Extended Laplace principle for empirical measures of a Markov chain. Supplementary material. Available at http://doi.org/10.1017/apr.2019.6.

[22] Erreygers, A., Rottondi, C., Verticale, G. and De Bock, J. (2018). Imprecise Markov models for scalable and robust performance evaluation of flexi-grid spectrum allocation policies. Preprint. Available at https://arxiv.org/abs/1801.05700.

[23] Esfahani, P. M. and Kuhn, D. (2015). Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. Preprint. Available at https://arxiv.org/abs/1505.05116.

[24] Ethier, S. N. and Kurtz, T. G. (1986). *Markov Processes*. John Wiley, New York.

[25] Gao, R. and Kleywegt, A. J. (2016). Distributionally robust stochastic optimization with Wasserstein distance. Preprint. Available at https://arxiv.org/abs/1604.02199.

[26] Gibbs, A. L. and Su, F. E. (2002). On choosing and bounding probability metrics. *Internat. Statist. Rev.* **70,** 419–435.

[27] Hanasusanto, G. A., Roitch, V., Kuhn, D. and Wiesemann, W. (2015). A distributionally robust perspective on uncertainty quantification and chance constrained programming. *Math. Program.* **151,** 35–62.

[28] Hartfiel, D. and Seneta, E. (1994). On the theory of Markov set-chains. *Adv. Appl. Prob.* **26,** 947–964.

[29] Hartfiel, D. J. (2006). *Markov Set-Chains*. Springer, Heidelberg.

[30] Jain, N. C. (1990). Large deviation lower bounds for additive functionals of Markov processes. *Ann. Prob.* **18,** 1071–1098.

[31] Kirkizlar, E., Andradóttir, S. and Ayhan, H. (2010). Robustness of efficient server assignment policies to service time distributions in finite-buffered lines. *Naval Res. Logistics* **57,** 563–582.

[32] Kurano, M., Song, J., Hosaka, M. and Huang, Y. (1998). Controlled Markov set-chains with discounting. *J. Appl. Prob.* **35,** 293–302.

[33] Lacker, D. (2015). Law invariant risk measures and information divergences. Preprint. Available at https://arxiv.org/abs/1510.07030.

[34] Lacker, D. (2016). A non-exponential extension of sanov's theorem via convex duality. Preprint. Available at https://arxiv.org/abs/1609.04744.

[35] Lan, Y. and Zhang, N. (2017). Strong limit theorems for weighted sums of negatively associated random variables in nonlinear probability. Preprint. Available at https://arxiv.org/abs/1706.05788.

[36] Ney, P. and Nummelin, E. (1987). Markov additive processes II. large deviations. *Ann. Prob.* **15,** 593–609.

[37] Nilim, A. and El Ghaoui, L. (2005). Robust control of Markov decision processes with uncertain transition matrices. *Operat. Res.* **53,** 780–798.

[38] Peng, S. (2009). Survey on normal distributions, central limit theorem, brownian motion and the related stochastic calculus under sublinear expectations. *Sci. China Ser. A* **52,** 1391–1411.

[39] Peng, S. (2010). Nonlinear expectations and stochastic calculus under uncertainty. Preprint. Available at https://arxiv.org/abs/1002.4546.

[40] Rottondi, C., Erreygers, A., Verticale, G. and De Bock, J. (2017). Modelling spectrum assignment in a two-service flexi-grid optical link with imprecise continuous-time Markov chains. In *13th Internat. Conf. on Design of Reliable Communication Networks (DRCN 2017)*, Munich, pp. 1–8.

[41] Škulj, D. (2006). Finite discrete time Markov chains with interval probabilities. In *Soft Methods for Integrated Uncertainty Modelling* (Adv. Soft Comput. **37**), Springer, Heidelberg, pp. 299–306.

[42] Škulj, D. (2009). Discrete time Markov chains with interval probabilities. *Internat. J. Approx. Reason.* **50,** 1314–1329.

[43] Škulj, D. (2015). Efficient computation of the bounds of continuous time imprecise Markov chains. *Appl. Math. Comput.* **250,** 165–180.

[44] Varadarajan, V. S. (1958). Weak convergence of measures on separable metric spaces. *Sankhyā* **19,** 15–22.

[45] Villani, C. (2008). *Optimal Transport: Old and New*, Vol. 338. Springer, Heidelberg.

[46] Wiesemann, W., Kuhn, D. and Rustem, B. (2013). Robust Markov decision processes. *Math. Operat. Res.* **38,** 153–183.

[47] Yang, I. (2017). A convex optimization approach to distributionally robust Markov decision processes with Wasserstein distance. *IEEE Control Systems Lett.* **1,** 164–169.

[48] Yu, P. and Xu, H. (2016). Distributionally robust counterpart in Markov decision processes. *IEEE Trans. Automatic Control* **61,** 2538–2543.