An inequality connecting entropy distance, Fisher Information and large deviations

Available online at www.sciencedirect.com

ScienceDirect

Stochastic Processes and their Applications 130 (2020) 2596–2638

stochastic processes and their applications

www.elsevier.com/locate/spa

An inequality connecting entropy distance, Fisher Information and large deviations

Bastian Hilder[a,*], Mark A. Peletier[b], Upanshu Sharma[c], Oliver Tse[d]

[a] *Institut für Analysis, Dynamik und Modellierung, Universität Stuttgart, Pfaffenwaldring 57, 70569 Stuttgart, Germany*

[b] *Department of Mathematics and Computer Science and Institute for Complex Molecular Systems, Eindhoven University of Technology, 5600 MB Eindhoven, The Netherlands*

[c] *CERMICS, École des Ponts ParisTech, 6-8 Avenue Blaise Pascal, Cité Descartes, Marne-la-Vallée, 77455, France*

[d] *Department of Mathematics and Computer Science, Eindhoven University of Technology, 5600 MB Eindhoven, The Netherlands*

Received 30 January 2019; received in revised form 1 July 2019; accepted 29 July 2019
Available online 2 August 2019

Abstract

In this paper we introduce a new generalisation of the relative Fisher Information for Markov jump processes on a finite or countable state space, and prove an inequality which connects this object with the relative entropy and a large deviation rate functional. In addition to possessing various favourable properties, we show that this *generalised Fisher Information* converges to the classical Fisher Information in an appropriate limit. We then use this generalised Fisher Information and the aforementioned inequality to qualitatively study coarse-graining problems for jump processes on discrete spaces.

MSC: 28A33; 34C29; 34D15; 46N20; 49J40; 60B10; 60F10; 60J27; 60J28

Keywords: Markov jump process; Relative entropy; Fisher Information; Large deviations

* Corresponding author.
E-mail addresses: bastian.hilder@mathematik.uni-stuttgart.de (B. Hilder), M.A.Peletier@tue.nl (M.A. Peletier), upanshu.sharma@enpc.fr (U. Sharma), o.t.c.tse@tue.nl (O. Tse).

https://doi.org/10.1016/j.spa.2019.07.012

1. Introduction

Lyapunov functions are important tools in the study of evolution equations. The relative entropy, which for two probability measures $\mu, \rho \in \mathcal{P}(\mathcal{X})$ is given by

$$\mathscr{H}(\mu|\rho) = \begin{cases} \int_{\mathcal{X}} f \log f \, d\rho, & \text{if } f = \dfrac{d\mu}{d\rho} \text{ exists,} \\ +\infty, & \text{otherwise,} \end{cases} \tag{1}$$

is one such Lyapunov function that plays a crucial role in the study of forward Kolmogorov equations. These equations describe the evolution of the distribution of a Markov process. In recent years, extensive research has been devoted to the study of the relative entropy and the Fisher Information (entropy production) which, amongst other things, are used to study the trend to equilibrium for both continuous [1,26] and discrete state-space Markov processes [3,9]. Typically this involves studying the time evolution of the relative entropy (1) where ρ is the stationary solution and μ_t is the time-dependent solution of the forward Kolmogorov equation under consideration. Although it is not a metric on the space of probability measures, relative entropy has been used as a notion of distance to equilibrium due to its favourable properties and natural connections to statistical physics.

As opposed to what was described above, in certain cases the relative entropy is also used to compare the time-dependent distributions of two different Markov processes. In the context of hydrodynamic limits, Yau [41] uses the relative entropy to compare the evolution of finite particle evolution with certain local-Gibbs states. Legoll and Leliévre [23] use relative entropy to compare an approximate solution with the true solution of a Fokker–Planck equation arising in molecular dynamics, and Bogachev et al. [4] compare solutions of two different Fokker–Planck equations in the context of mean-field games.

It has recently been shown [11] that the relative entropy comparing an arbitrary time-dependent probability measure to the solution of a Fokker–Planck equation is directly linked to the Fisher Information and the large-deviation rate functional via an inequality. We refer to [37, Chapter 2] for a detailed overview. In [12] the authors present a new variational approach that uses this inequality to qualitatively study coarse-graining problems in (nonlocal) Fokker–Planck equations. In [11] this inequality has been used to quantitatively estimate coarse-graining errors.

While all the aforementioned references deal with diffusion processes, not much is known about the relative entropy of two time-dependent distributions for jump processes. In recent years, for processes on discrete spaces, new Wasserstein-like gradient-flow structures with relative entropy as the driving functional have been discovered [7,14,24,27,28]. In this paper we ask if the ideas described above for the continuous case can be generalised to the discrete case, specifically for Markov jump processes:

> Starting with Markov jump processes, can the relative entropy of two time-dependent curves be connected to the large-deviation rate functional? Furthermore, can this connection be exploited to study coarse-graining problems?

In this paper we provide an answer to these questions by generalising the notion of Fisher Information for Markov processes. In addition to studying its properties, we will show that this generalised Fisher Information is naturally related to the relative entropy and the large-deviation rate functional. Finally we apply this inequality to study a coarse-graining problem on a discrete state space.

1.1. Relative Fisher Information and large-deviation rate functional

Before we present our contributions to answering the questions mentioned above (see Section 1.2), we introduce the classical relative Fisher Information and the large-deviation rate functional. Unlike the relative entropy, these two objects explicitly depend on the evolution equation under consideration.

In this paper we are interested in jump processes on a *finite or countable state space* $\mathcal{X}$. Using $\mathcal{P}(\mathcal{X})$ to denote the space of probability measures on $\mathcal{X}$, the law of the jump process $\rho : [0, T] \to \mathcal{P}(\mathcal{X})$ satisfies the evolution equation

$$\begin{cases} \partial_t \rho = L^T \rho, \\ \rho_{t=0} = \rho_0. \end{cases} \tag{2}$$

In Eq. (2), L^T is the adjoint of $L : c_0(\mathcal{X}) \to c_0(\mathcal{X})$, the generator of the process. Note that in the context of stochastic processes Eq. (2) is usually referred to as the forward Kolmogorov equation. Since $\mathcal{X}$ is discrete, we use matrix notation and write the operator L as a (potentially infinite) matrix $L \in \mathbb{R}^{\mathcal{X}\times\mathcal{X}}$. The generator L satisfies

$$(L1) \qquad L(x, y) \geq 0 \text{ for all } x \neq y \text{ and } \sum_{y\in\mathcal{X}} L(x, y) = 0 \text{ for all } x \in \mathcal{X}, \tag{3a}$$

$$(L2) \qquad \sup_{x\in\mathcal{X}} |L(x, x)| < \infty, \tag{3b}$$

$$(L3) \qquad L \text{ is irreducible.} \tag{3c}$$

These conditions are sufficient for L to be a bounded Markov operator $L : c_0(\mathcal{X}) \to c_0(\mathcal{X})$, where $c_0(\mathcal{X})$ is the Banach space of functions on $\mathcal{X}$ that converge to zero outside of large finite subsets of $\mathcal{X}$, equipped with the supremum norm. Since L^T generates a uniformly continuous semigroup in $\ell^1(\mathcal{X})$ [13, Proposition 2.11], Eq. (2) admits a unique solution $\rho \in \mathcal{C}^1([0, T]; \ell^1(\mathcal{X}))$ [13, Theorem 6.6]; since Eq. (2) preserves non-negativity and total mass, we have $\rho \in \mathcal{C}([0, T], \mathcal{P}(\mathcal{X}))$ whenever $\rho_0 \in \mathcal{P}(\mathcal{X})$.

Remark 1.1. The space $\mathcal{P}(\mathcal{X})$ is a subset of $\ell^1(\mathcal{X})$, and the weak measure topology on $\mathcal{P}(\mathcal{X})$ coincides with the $\sigma(\ell^1, \ell^\infty)$-topology on $\ell^1(\mathcal{X})$. Recall that by Schur's theorem, weak and strong *convergence* on $\ell^1(\mathcal{X})$ are the same, even though the weak and strong topologies may be different; therefore functions $f : [0, T] \to \ell^1(\mathcal{X})$ are strongly continuous if and only they are weakly continuous. Since 'weak measure convergence' in $\mathcal{P}(\mathcal{X})$ is the same as the $\sigma(\ell^1, \ell^\infty)$-convergence in $\ell^1(\mathcal{X})$, we will omit the term 'weak' in our discussion and notation, and simply talk about 'continuous' functions from $[0, T]$ to $\mathcal{P}(\mathcal{X})$ or to $\ell^1(\mathcal{X})$. □

The classical definition of 'relative Fisher Information' arises from the time derivative of the relative entropy along two solutions of (2). Indeed, for two positive solutions μ, ρ of (2), we have

$$\frac{d}{dt}\mathscr{H}(\mu_t|\rho_t) = -\mathscr{R}_L(\mu_t|\rho_t), \tag{4}$$

where μ_t, ρ_t denote the time slice at time t, and (4) is used to define the right-hand side as follows.

Definition 1.2. For $\mu, \rho \in \mathcal{P}_+(\mathcal{X})$, where $\mathcal{P}_+(\mathcal{X})$ is the set of strictly positive probability measures on $\mathcal{X}$, the (classical) *relative Fisher Information* is defined as

$$\mathscr{R}_L(\mu|\rho) := \sum_{x,y\in\mathcal{X}} \rho(x)L(x,y)\left[v(y) - v(x) - v(x)\log\left(\frac{v(y)}{v(x)}\right)\right], \qquad v = \mu/\rho \tag{5}$$

This sum is well-defined in $[0,\infty]$, since $L(x,y) \geq 0$ for $x \neq y$, and the term between brackets is non-negative and vanishes if $x = y$. Especially, the relative Fisher Information is non-negative. This corresponds to the well-known fact that the relative entropy decays in time along two solutions of the same forward Kolmogorov equation (see [40, Theorem 1.1]). It should be noted that the definition (5) of the Fisher Information coincides with the classical notion of Fisher Information with respect to the stationary measure, i.e. when $L^T\rho = 0$ (see [3, Equation 1.4]). Alternatively, the relative Fisher Information (5) can also be seen as the Bregman divergence of the Fisher Information with respect to the stationary measure (see [16, Section 5.1] for details).

Remark 1.3. The notation of 'classical' relative Fisher information is used in Definition 1.2 to highlight that $\mathscr{R}_L$ is connected to the well-known relative Fisher information in the continuous setting, see e.g. [37, Equation 1.2.8]. However, it is necessary to state the precise definition here, since the different ways of writing the Fisher information in the continuous setting are no longer equivalent in the discrete setting (see [3, Section 3]). This arises due to the lack of a chain rule. □

Apart from the classical connection between (linear) Markov processes and forward Kolmogorov equations described above, the forward Kolmogorov equations can also be viewed as the many-particle limit of some underlying system of Markov processes. To make this precise, consider a sequence $(X^n)_{n\in\mathbb{N}}$ of independent and identical Markov processes on state space $\mathcal{X}$ and generated by L. Under fairly general conditions (see for instance [10, Theorem 11.4.1]), the sequence of empirical measures

$$\rho^N := \frac{1}{N}\sum_{i=1}^{N} \delta_{X^i}, \tag{6}$$

converges almost surely to the solution of (2).

This convergence is the starting point for a large-deviation result. In particular it has been shown (see Theorem 1.4) that the sequence ρ^N has a *large-deviation property* which characterises the probability of finding the empirical measure far from the limit ρ, written informally as

$$\mathrm{Prob}(\rho^N \approx \rho) \sim e^{-N(\mathscr{I}_0(\rho_0)+\mathscr{I}_L(\rho))} \quad \text{as } N\to\infty,$$

in terms of *rate functionals* $\mathscr{I}_0$ and $\mathscr{I}_L$ of the initial data $(\rho_0^N)_{N\in\mathbb{N}}$ and the path $(t\mapsto\rho_t^N)_{N\in\mathbb{N}}$ respectively. Formally, the rate functionals satisfy the inequalities

$$\limsup_{N\to\infty}\frac{1}{N}\log(\mathrm{law}(\rho^N)(M_c)) \leq -\inf_{\rho\in M_c}(\mathscr{I}_0(\rho_0)+\mathscr{I}_L(\rho)),$$
$$\liminf_{N\to\infty}\frac{1}{N}\log(\mathrm{law}(\rho^N)(M_o)) \geq -\inf_{\rho\in M_o}(\mathscr{I}_0(\rho_0)+\mathscr{I}_L(\rho))$$

for any measurable, closed set $M_c \subset \mathcal{C}([0,T];\mathcal{P}(\mathcal{X}))$ and open set $M_o \subset \mathcal{C}([0,T];\mathcal{P}(\mathcal{X}))$, see [39] for a general definition. Here law$(\cdot)$ denotes the law of a random variable.

In this paper we will focus on $\mathscr{I}_L : \mathcal{C}([0,T];\mathcal{P}(\mathcal{X})) \to [0,\infty]$ which has a characterisation of the form

$$\mathscr{I}_L(\rho) = \begin{cases} \displaystyle\int_0^T \mathcal{L}(\rho_t, \partial_t \rho_t)\,dt, & \text{if } \rho \in AC([0,T];\mathcal{P}(\mathcal{X})), \\ +\infty, & \text{otherwise.} \end{cases} \tag{7}$$

Here $AC([0,T];\mathcal{P}(\mathcal{X}))$ is the space of absolutely continuous trajectories in the space of probability measures (see Appendix A).

The *Lagrangian* $\mathcal{L} : \mathcal{P}(\mathcal{X}) \times \ell^1(\mathcal{X}) \to [0,\infty]$ in the definition above of $\mathscr{I}_L$ is non-negative and convex in its second argument, and satisfies $\mathcal{L}(\rho_t, \partial_t\rho_t) = 0$ if and only if ρ solves $\partial_t \rho_t = L^T \rho_t$. The rate functional $\mathscr{I}_L$ therefore has the crucial properties

$$\text{(a) } \mathscr{I}_L(\rho) \geq 0, \qquad \text{and} \qquad \text{(b) } \rho \text{ solves (2)} \iff \mathscr{I}_L(\rho) = 0, \tag{8}$$

and consequently the equation "$\mathscr{I}_L(\rho) = 0$" can be viewed as a variational characterisation of the forward Kolmogorov equation.

The Lagrangian $\mathcal{L}$ is defined as the Legendre dual of a *Hamiltonian* $\mathcal{H} : \mathcal{P}(\mathcal{X}) \times \ell^\infty(\mathcal{X}) \to [0,\infty]$,

$$\mathcal{L}(\mu, s) := \sup_{\xi \in \ell^\infty(\mathcal{X})} \left\{ \sum_{x\in\mathcal{X}} \xi(x)s(x) - \mathcal{H}(\mu,\xi) \right\}. \tag{9}$$

In our setting of a Markov process on a discrete state space with generator L, the Hamiltonian is explicitly given by

$$\mathcal{H}(\mu,\xi) := \sum_{x,y\in\mathcal{X}} \mu(x)L(x,y)\left[e^{\xi(y)-\xi(x)} - 1\right], \tag{10}$$

and by Legendre duality it has the alternative characterisation

$$\mathcal{H}(\mu,\xi) = \sup_{s\in\ell^1(\mathcal{X})} \left\{ \sum_{x\in\mathcal{X}} \xi(x)s(x) - \mathcal{L}(\mu,s) \right\}. \tag{11}$$

The following result places the preceding remarks in a rigorous context. We denote the space of right-continuous functions with left limits mapping $[0,T]$ into $\mathcal{P}(\mathcal{X})$ by $D_{\mathcal{P}(\mathcal{X})}[0,T]$, and the dual pairing between $\ell^\infty(\mathcal{X})$ and $\mathcal{P}(\mathcal{X})$ by $\langle f, \mu\rangle = \sum_{x\in\mathcal{X}} f(x)\mu(x)$ for any $f \in \ell^\infty(\mathcal{X})$ and $\mu \in \mathcal{P}(\mathcal{X})$, then the following result holds.

Theorem 1.4. *Let $\rho^N \in \mathcal{P}(\mathcal{X})$ be the empirical process (6) generated by $N \in \mathbb{N}$ independent Markov processes $(X^i)_{i=1,\ldots N}$ on the state space $\mathcal{X}$ with generator L. Furthermore, assume that the initial values $(\rho_0^N)_{N\in\mathbb{N}}$ are deterministic and converge in $\mathcal{P}(\mathcal{X})$ to some ρ_0. Then, $(\rho^N)_{N\in\mathbb{N}}$ satisfies a large deviations principle in $D_{\mathcal{P}(\mathcal{X})}[0,T]$ with rate functional $\mathscr{I}_L : \mathcal{C}([0,T];\mathcal{P}(\mathcal{X})) \to \mathbb{R}$ given by (7), and which has the alternative representation*

$$\mathscr{I}_L(\mu) = \sup_{f\in\mathcal{C}^1([0,T];\ell^\infty(\mathcal{X}))} \left\{ \langle f_T, \mu_T\rangle - \langle f_0, \mu_0\rangle - \int_0^T \Big(\langle \partial_t f_t, \mu_t\rangle + \mathcal{H}(\mu_t, f_t) \Big)\,dt \right\} \tag{12}$$

where $\mu \in \mathcal{C}([0,T];\mathcal{P}(\mathcal{X}))$ with $\mu|_{t=0} = \rho_0$ and the Hamiltonian $\mathcal{H}$ is defined in (10). Additionally, if for some $\mu \in \mathcal{C}([0,T];\mathcal{P}(\mathcal{X}))$ we have $\mathscr{I}_L(\mu) < \infty$, then $t \mapsto \mu_t \in \mathcal{P}(\mathcal{X})$ is *absolutely continuous, and the rate functional can be reformulated as*

$$\mathscr{I}_L(\mu) = \sup_{f \in L^\infty(0,T;\ell^\infty(\mathcal{X}))} \int_0^T \Big(\langle f_t, \partial_t \mu_t \rangle - \mathcal{H}(\mu_t, f_t) \Big)\, dt. \tag{13}$$

The existence of the large-deviation principle is a reformulation of [19, Proposition 5.10], while the main statement of the theorem is the alternative characterisation (12); we give the proof in Appendix B. Appendix A collects some results on absolutely-continuous curves and integration.

1.2. Main results

As mentioned earlier, the main goal of this work is to connect relative entropy, Fisher Information and large-deviation rate functional in the context of Markov processes on a discrete state space. While the connection between the relative entropy and the rate functional is fairly classical, it does not connect to the Fisher Information. As pointed out earlier, these objects have been connected recently in the case when $\mathcal{X} = \mathbb{R}^n$ and L is a diffusion operator via the inequality (see [37, Chapter 2] and [11, Section 2.5] for details)

$$\mathscr{H}(\mu_T|\rho_T) + \int_0^T \mathscr{R}_L(\mu_s|\rho_s)\, ds \le \mathscr{H}(\mu_0|\rho_0) + \mathscr{I}_L(\mu), \tag{14}$$

where μ is a measure-valued curve (such that the right-hand side of the estimate is well defined) and ρ solves $\partial_t \rho = L^T \rho$. In [37] this relation is called the free-energy-relative-Fisher-Information-rate-functional (FIR) inequality, a terminology that we will use throughout this paper.

We shall demonstrate in Section 2.1 that such an inequality already fails in fairly simple situations for a Markov jump process. To get around this issue, we generalise the notion of the relative Fisher Information.

Definition 1.5. Let $\lambda \in (0,1)$. We define the *generalised relative Fisher Information* $\mathscr{R}_L^\lambda : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to [0,\infty]$ (corresponding to a generator L) as follows.

1. If $\rho, \mu \in \mathcal{P}_+(\mathcal{X})$ and $\sup_{x\in\mathcal{X}} \max\{\mu(x)/\rho(x), \rho(x)/\mu(x)\} < \infty$, then

$$\mathscr{R}_L^\lambda(\mu|\rho) := \sum_{x,y\in\mathcal{X}} L(x,y) \frac{\mu(y)}{\rho(y)} \rho(x) - \frac{1}{\lambda} \mathcal{H}\left(\mu, \lambda \log\left(\frac{\mu}{\rho} \right) \right) \tag{15a}$$

$$= \sum_{x,y\in\mathcal{X}} L(x,y) \left[\frac{\mu(y)}{\rho(y)} \rho(x) - \mu(x) - \frac{1}{\lambda} \left(\mu(x)^{1-\lambda} \rho(x)^\lambda \left(\frac{\mu(y)}{\rho(y)} \right)^\lambda - \mu(x) \right) \right]. \tag{15b}$$

Here $\mathcal{H}$ is the Hamiltonian (10) that arises in the context of large deviations.

2. If $\rho, \mu \in \mathcal{P}(\mathcal{X})$, then

$$\mathscr{R}_L^\lambda(\mu|\rho) := \sum_{x,y\in\mathcal{X}} L(x,y) \psi_\lambda(x,y), \tag{15c}$$

where ψ_λ is defined as

$$\psi_\lambda(x,y) := \begin{cases} \dfrac{\mu(y)}{\rho(y)}\rho(x) - \mu(x) & \\ \quad -\dfrac{1}{\lambda}\left(\mu(x)^{1-\lambda}\rho(x)^{\lambda}\left(\dfrac{\mu(y)}{\rho(y)}\right)^{\lambda} - \mu(x)\right), & \text{if } \rho(y) > 0, \text{ and } \rho(x) > 0, \\ +\infty & \text{if } \rho(y) = 0,\ \rho(x) > 0, \\ & \qquad \text{and } \mu(y) > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Both these definitions of the generalised relative Fisher Information are consistent, i.e. whenever both definitions apply, they give the same value (see Lemma 2.4). To motivate these definitions, we use the characterisation (12) of the rate functional and reason formally as follows. Let $\mu : [0,T] \to \mathcal{P}(\mathcal{X})$ be a smooth curve with $\mathscr{I}_L(\mu) < \infty$ and $\rho : [0,T] \to \mathcal{P}(\mathcal{X})$ be a smooth solution of the forward Kolmogorov equation (2) such that $\log(\mu/\rho)$ is sufficiently regular. Using $f = \lambda \log(\mu/\rho)$ with $\lambda \in (0,1)$ in (12), we obtain

$$\begin{aligned} \frac{1}{\lambda}\mathscr{I}_L(\mu) \geq & \sum_{x\in\mathcal{X}} \log\left(\frac{\mu_T(x)}{\rho_T(x)}\right)\mu_T(x) - \textstyle\sum_{x\in\mathcal{X}} \log\left(\frac{\mu_0(x)}{\rho_0(x)}\right)\rho_0(x) \\ & \qquad - \int_0^T \left(\sum_{x\in\mathcal{X}} \partial_t \log\left(\frac{\mu_t(x)}{\rho_t(x)}\right)\mu_t(x) + \frac{1}{\lambda}\mathcal{H}\left(\mu_t, \lambda\log\left(\frac{\mu_t}{\rho_t}\right)\right)\right) dt \\ = & \mathscr{H}(\mu_T|\rho_T) - \mathscr{H}(\mu_0|\rho_0) \\ & \qquad + \int_0^T \left(\sum_{x,y\in\mathcal{X}} L(x,y)\mu_t(y)\frac{\rho_t(x)}{\rho_t(y)} - \frac{1}{\lambda}\mathcal{H}\left(\mu_t, \lambda\log\left(\frac{\mu_t}{\rho_t}\right)\right) dt\right), \end{aligned}$$

where the equality follows since

$$\begin{aligned} \sum_{x\in\mathcal{X}} \partial_t \log\left(\frac{\mu_t(x)}{\rho_t(x)}\right)\mu_t(x) &= \sum_{x\in\mathcal{X}} \partial_t\mu_t(x) - \sum_{x\in\mathcal{X}} \frac{\mu_t(x)}{\rho_t(x)}(L^T\rho)(x) \\ &= 0 - \sum_{x,y\in\mathcal{X}} L(x,y)\mu_t(y)\frac{\rho_t(x)}{\rho_t(y)}. \end{aligned}$$

The formal inequality above resembles (14), where the integrand in the time integral is precisely the generalised Fisher Information given in (15a). These formal calculations can and will be made rigorous, resulting in the first main result of this article which we now state.

Theorem 1.6. *Let $\rho \in AC([0,T];\mathcal{P}(\mathcal{X}))$ be a solution of (2) and $\mu \in \mathcal{C}([0,T];\mathcal{P}(\mathcal{X}))$ satisfy*

$$\mathscr{I}_L(\mu) + \mathscr{H}(\mu_0|\rho_0) < \infty,$$

with $\mu|_{t=0} = \mu_0$. Then for any $\lambda \in (0,1)$ we have

$$\mathscr{H}(\mu_T|\rho_T) + \int_0^T \mathscr{R}_L^\lambda(\mu_t|\rho_t)\,dt \leq \mathscr{H}(\mu_0|\rho_0) + \frac{1}{\lambda}\mathscr{I}_L(\mu). \tag{FIR$_\lambda$}$$

It is important to note that the roles of μ and ρ in the FIR inequality (FIR$_\lambda$) cannot be interchanged, i.e. μ is a solution to the forward Kolmogorov equation and ρ is arbitrary, since the relative entropy is not symmetric. As evident from the formal calculations above, the generalised relative Fisher Information (15) is constructed such that the proof of the FIR inequality goes through. In addition to satisfying (FIR$_\lambda$), the generalised Fisher Information has several favourable properties which we now summarise (see Sections 2.2–2.3 for details).

In order to state the precise result, we introduce two concepts. First, we note that in the case of jump processes we can identify the state space $\mathcal{X}$ naturally with a graph by connecting two points of $\mathcal{X}$ by an edge if there is a non-zero probability to jump directly between these states. In particular this enables us to define connected components in $\mathcal{X}$ (see Section 2.3 for additional details). Second, to study the behaviour of $\mathscr{R}_L^\lambda$ for $\lambda \to 0$, we use the notion of Gamma convergence (see e.g. [5] for an introduction), which is the natural type of convergence in a variational setting.

Theorem 1.7. *For $\lambda \in (0, 1)$, the generalised Fisher Information satisfies:*

(i) $\mathscr{R}_L^\lambda$ is non-negative and lower-semicontinuous on $\mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X})$.
(ii) If $\mu, \rho \in \mathcal{P}(\mathcal{X})$ with $\mathscr{R}_L^\lambda(\mu|\rho) = 0$, then μ is a constant multiple of ρ on each connected component of the support of ρ. In particular, if $\rho \in \mathcal{P}_+(\mathcal{X})$, then $\mu = \rho$ on $\mathcal{X}$.
(iii) $\mathscr{R}_L^\lambda \to \mathscr{R}_L$ as $\lambda \to 0$ on $\mathcal{P}_+(\mathcal{X}) \times \mathcal{P}_+(\mathcal{X})$ in the sense of Gamma convergence.

Whenever two measures ρ and μ satisfy $\mathscr{R}_L^\lambda(\mu|\rho) = 0$, Theorem 1.7(ii) provides information on how they are related, similar to that of a logarithmic version of a Dirichlet form in continuous state spaces. The name ‘generalised’ Fisher Information is motivated by the fact that we can recover the relative Fisher Information (5) as a limit for $\lambda \to 0$ (cf. Theorem 1.7(iii)). In addition to this asymptotic relation, the generalised and the classical relative Fisher Information can also be compared directly by an inequality in a fairly restrictive setting, thereby allowing us to prove a FIR inequality with the classical Fisher Information (see Section 2.4 for details).

We point out that the FIR inequality bears similarity to the entropy-dissipation identity that arises in the context of reversible Markov processes and more generally gradient flows (see [30] for details). However in Theorem 1.6 (and throughout this article) we do not assume the generator L to be reversible and therefore our results go beyond the existing results on gradient flows. Additionally, the FIR inequality compares two curves, which is not the case for the entropy-dissipation identity.

1.3. Application to coarse-graining

Coarse-graining is an umbrella term used for techniques which approximate a complex or high-dimensional system by a simpler or lower-dimensional one. While there are many formal techniques for achieving this (see [15] and references therein), rigorous mathematical analysis is typically restricted to situations that exhibit explicit separation of temporal and/or spatial scales, i.e. the presence of fast and slow variables. In these situations, as the ratio of ‘fast’ to ‘slow’ increases, some form of averaging or homogenisation allows one to remove the fast scales, and obtain a limiting system that focuses on the slow ones. Recently, a new variational technique based on studying the large-deviation rate functional has been introduced in [12,37] to study coarse-graining limits arising in the context of diffusion processes (see Section 3.1 for details). As mentioned earlier, in this paper we apply this variational technique to study a coarse-graining problem arising in the discrete setting (described below). The generalised Fisher Information (15) and the FIR inequality (FIR$_\lambda$) described in the last section play a crucial role in this study.

The coarse-graining problem we study here is inspired by kinetic Monte-Carlo methods in molecular dynamics (see [21, Chapter 5] for details). Consider a particle moving in a potential-energy landscape, which consists of small and large barriers as described in Fig. 1. The large

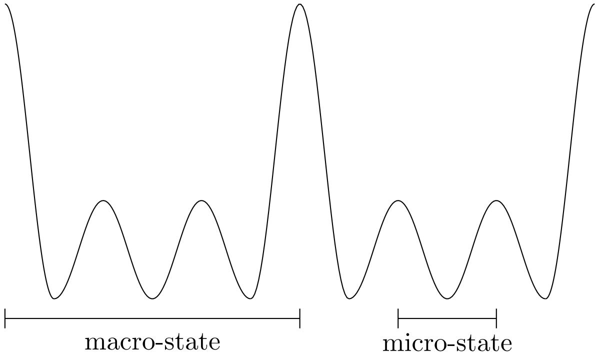

Fig. 1. Energy landscape with two macro-states.

energy barriers introduce a natural scale-separation since it is harder for the particle to jump across them compared to the smaller barriers. More precisely we can model the behaviour of such a particle as a Markov jump process on $\mathcal{X} = \mathcal{Y} \times \mathcal{Z}$ where $\mathcal{Y}$ corresponds to the states separated by the large energy barriers while $\mathcal{Z}$ is the part of the state space separated by small energy barriers. For simplicity, we assume that there is only one large barrier, i.e. $\mathcal{Y} = \{0, 1\}$ and finitely many small barriers corresponding to each of these large barriers, i.e. $\mathcal{Z} = \{1, \dots, n\}$. This intuitively means that the state space is divided up into two *macro-states*, each of which contain $n \in \mathbb{N}$ easily accessible *micro-states*.

We consider the Markov process which evolves according to the generator

$$\tilde{L}^\varepsilon = Q + \varepsilon C := \begin{pmatrix} Q_0 & 0 \\ 0 & Q_1 \end{pmatrix} + \varepsilon \begin{pmatrix} D_0 & C_{0,1} \\ C_{1,0} & D_1 \end{pmatrix},$$

where Q and C are ε-independent matrices with

$$\forall x_1 \in \mathcal{X} : \sum_{x_2 \in \mathcal{X}} Q(x_1, x_2) = 0 = \sum_{x_2 \in \mathcal{X}} C(x_1, x_2). \tag{16}$$

The diagonal matrix D_y, $y \in \mathcal{Y}$, is constructed so that C satisfies the aforementioned property, i.e.

$$\forall z_1 \in \mathcal{Z} : \; D_y(z_1, z_1) := - \sum_{z_2 \in \mathcal{Z}} C_{y,1-y}(z_1, z_2).$$

We assume that Q_y is irreducible for every $y \in \mathcal{Y}$ and $\tilde{L}^\varepsilon$ is irreducible. The irreducibility of $\tilde{L}^\varepsilon$ is equivalent to assuming that $C_{1,0}$ and $C_{0,1}$ have at least one positive entry.

Now let us take a closer look at each of these components. The small parameter $\varepsilon > 0$ models the scale-separation arising due to the difference in the heights of the barriers. The matrix $Q_y \in \mathbb{R}^{n \times n}$ encodes the jumps between micro-states within the yth macro-state. The matrix $C_{y,1-y} \in \mathbb{R}^{n \times n}$ encodes the transition from the yth macro-state to $(1 - y)$th macro-state. The summability condition (16) ensures that $\tilde{L}^\varepsilon$ is a generator, i.e. an operator satisfying (3a).

When ε is small, the dynamics of the particle evolving according to $\tilde{L}^\varepsilon$ splits into slow and fast components. The fast component moves the particle within a macro-state, and the slow component is visible as a rare jump to a different macro-state. Following [22], in order to focus on the slow component we rescale time by ε^{-1} and arrive at

$$L^\varepsilon = \frac{1}{\varepsilon}Q + C := \frac{1}{\varepsilon}\begin{pmatrix} Q_0 & 0 \\ 0 & Q_1 \end{pmatrix} + \begin{pmatrix} D_0 & C_{0,1} \\ C_{1,0} & D_1 \end{pmatrix}. \tag{17}$$

The main goal of the second part of this work is to study the behaviour of the Markov jump process described by the forward Kolmogorov equation

$$\begin{cases} \partial_t \mu^\varepsilon = (L^\varepsilon)^T \mu^\varepsilon, \\ \mu^\varepsilon_{t=0} = \mu_0, \end{cases} \tag{18}$$

in the limit $\varepsilon \to 0$. In this limit it is natural to expect that the solution μ^ε equilibrates in each macro-state and the limit can be described by a jump process on $\mathcal{Y}$, i.e. a two-point Markov jump process. In the second part of this article we make this intuition precise (see Section 3 for details).

To state the precise result we need to introduce two objects: (1) the stationary measure of (18), denoted by $\pi^\varepsilon \in \mathcal{P}(\mathcal{X})$, which exists since L^ε is irreducible, and (2) the *coarse-graining map* $\xi : \mathcal{X} \to \mathcal{Y}$ as $\xi(x) = y$ for every $x = (y, z) \in \mathcal{X}$.

For more details on this coarse-graining map see Section 3.

Theorem 1.8. *Consider a sequence $\mu^\varepsilon \in \mathcal{C}([0,T];\mathcal{P}(\mathcal{X}))$ of solutions to (18). Assume that the initial data satisfies*

$$\sup_{\varepsilon>0} \mathscr{H}(\mu_0^\varepsilon|\pi^\varepsilon) < \infty.$$

We then find for a subsequence (not relabelled) such that the following holds.

1. *(Compactness) The sequence $\mu^\varepsilon \to \mu$ in $\mathcal{M}([0,T]\times\mathcal{X})$, the space of non-negative, finite measures on $[0,T]\times\mathcal{X}$, with respect to the narrow topology, and $\xi_\#\mu^\varepsilon \to \xi_\#\mu$ in $\mathcal{C}([0,T];\mathcal{P}(\mathcal{Y}))$ uniformly in time.*
2. *(Local equilibrium) There exists $\hat\mu \in \mathcal{C}([0,T];\mathcal{P}(\mathcal{Y}))$ such that for almost all $t \in [0,T]$*

$$\forall y \in \mathcal{Y},\ \ A \subset \mathcal{Z},\ \ \mu_t(\{y\}\times A) = \hat\mu_t(y)\pi_y(A),$$

   where for each $y \in \mathcal{Y}$, $\pi_y \in \mathcal{P}(\mathcal{Z})$ is the stationary measure corresponding to Q_y. Furthermore $\xi_\#\mu^\varepsilon \to \hat\mu$ in $\mathcal{C}([0,T];\mathcal{P}(\mathcal{Y}))$ uniformly in time.
3. *(Limit dynamics) The limit $\hat\mu \in \mathcal{C}([0,T];\mathcal{P}(\mathcal{Y}))$ solves*

$$\partial_t \hat\mu = L^T \hat\mu$$

   with the (limiting) generator

$$L := \begin{pmatrix} -\lambda_0 & \lambda_0 \\ \lambda_1 & -\lambda_1 \end{pmatrix}, \quad \lambda_y := \sum_{z,z'\in\mathcal{Z}} \pi_y(z) C_{y,1-y}(z,z').$$

The narrow topology used in Theorem 1.8(i) is defined via weak convergence in duality with test functions in C_b. Furthermore, note that we do not specify the topology on $\mathcal{P}(\mathcal{X})$ in this result, since $\mathcal{X}$ is finite and thus $\mathcal{P}(\mathcal{X})$ is a subset of a finite-dimensional space (also see Remark 3.1). Finally, we point out that this result is a special case of our analysis in Section 3, which also applies to the case of *approximate solutions* (see Remark 3.6 for details).

1.4. Comparison with other work

We now comment on the novelties developed in this paper compared with other work.

1. *In comparison with other works on the FIR inequality.* As mentioned earlier, the idea of an FIR inequality connecting the free energy (which, in our case, is the relative entropy), the relative Fisher Information and the large deviation rate functional was discussed in the context of diffusion processes [4,11,12,37], although most of these works do not explicitly refer to this inequality as the FIR inequality. Our contribution lies in the extension of the FIR inequality to the discrete settings which is substantially different from the diffusion case treated in the references above. The main difference is that the Hamiltonian in the discrete case has a different scaling behaviour which ensures that the classical FIR inequality fails in the discrete setting (see Section 2.1 for details). For a more detailed review of these connections see Section 2.5.
2. *In comparison with other work on the example treated in this paper.* The coarse-graining example introduced in Section 1.3 is an averaging problem for Markov chains [22,33]. In these references, martingale techniques are used to prove a pathwise convergence result while our proof relies on the variational framework given by the large deviations result. Although the convergence result in this work is weaker, we obtain an explicit local-equilibrium statement and our result also applies to approximate solutions, i.e. curves with finite rate functional, rather than zero. This allows us to work with a larger class of measures (see Remark 3.6). This latter property also distinguishes our approach from other classical strategies such as geometric singular perturbation theory, see for instance [20].
3. *Comparison with variational evolutionary methods.* In recent years, variational-evolutionary structures akin to gradient flows have been developed for forward Kolmogorov equations on finite state spaces [7,24,27,28]. This structure can also be used to investigate singular limits [29,34,35]. However these structures are limited to reversible Markov chains, while the approach discussed in this paper does not require reversibility since we only use the variational structure provided by the large-deviations principle.
4. *Quantitative coarse-graining.* As in the diffusion case [11,37], a natural next step is to derive explicit error estimates for 'finite' scale separation. However, the strategy to obtain those estimates does not use the full FIR inequality but only a related result inspired by [41] and is thus omitted in this paper. For details we refer to [16, Chapter 8].

1.5. *Outline of the article*

In the rest of the paper we present the details of the ideas introduced above. In Section 2 we construct the generalised Fisher Information and prove the FIR inequality. In Section 3 we study the coarse-graining problem using the variational technique developed in [12]. Section 4 provides further discussions and generalisations and certain details on the rate functional are discussed in Appendix B. In Appendix A we collect some results on integration in infinite-dimensional spaces and in Appendix C we provide a result on positivity of solutions for irreducible generators.

2. Generalised relative Fisher Information and FIR inequality

In Section 2.1 we discuss a simple example where the FIR inequality fails when working with the classical relative Fisher Information (5), following which we prove the FIR inequality with the generalised relative Fisher Information (15) in Section 2.2. We then prove the main properties of the generalised Fisher Information in Section 2.3. Finally in Section 2.5 we connect these ideas to diffusions and compare to existing results in the literature.

Remark 2.1 (*Extension to Finite Measures*)**.** We restrict the treatment in what follows to probability measures to keep the notation simple. However, the definition as well as the properties of the generalised Fisher Information can be generalised to non-negative, finite measures with no additional difficulties. □

2.1. Failure of FIR inequality with relative Fisher Information

Before we present the proof of the FIR inequality with the generalised Fisher Information (described in Theorem 1.6), we first show a simple example where such an inequality (14) fails when working with the 'classical' relative Fisher Information (5). Note that this is distinctly different from the case of diffusions on continuous state space where the FIR inequality holds for the relative Fisher Information (for a detailed discussion see Section 2.5).

The idea is to construct a sequence of curves for which the rate functional stays bounded while the classical relative Fisher Information is unbounded in the limit, which would prove that the FIR inequality does not hold in this setting. We consider a two-point space $\mathcal{X} = \{0, 1\}$ and a generator given by

$$L = \begin{pmatrix} -a & a \\ b & -b \end{pmatrix},$$

for $a, b > 0$. Furthermore we consider a constant-in-time curve $\mu \in \mathcal{P}(\mathcal{X})$. For any $f \in \ell^\infty(\mathcal{X})$ and $s := f(0) - f(1)$, the Hamiltonian (10) can be written as

$$\mathcal{H}(\mu, f) = a\mu(0)\left(e^{-s} - 1\right) + b(1 - \mu(0))\left(e^{s} - 1\right).$$

There exists a constant $c > 0$ such that for any f and μ we have $\mathcal{H}(\mu, f) > -c$. Therefore using the definition of the rate functional (13) we find

$$\forall \mu \in \mathcal{P}(\mathcal{X}): \ \mathscr{I}_L(\mu) = \sup_{f \in L^\infty([0,T];\ell^\infty(\mathcal{X}))} \int_0^T -\mathcal{H}(\mu, f_t)\, dt \leq cT.$$

Next let us look at the classical relative Fisher Information (5) with $\rho = (a+b)^{-1}(b, a) \in \mathcal{P}_+(\mathcal{X})$ which satisfies $L^T \rho = 0$. Writing $\mu = (\mu_0, 1 - \mu_0)$ and $\rho = (\rho_0, 1 - \rho_0)$ we find

$$\mathscr{R}_L(\mu|\rho) = a\left[\frac{(1-\mu_0)\rho_0}{1-\rho_0} - \mu_0 - \mu_0 \log\left(\frac{(1-\mu_0)\rho_0}{\mu_0(1-\rho_0)}\right)\right] + b\left[\frac{\mu_0(1-\rho_0)}{\rho_0} - (1-\mu_0) - (1-\mu_0)\log\left(\frac{\mu_0(1-\rho_0)}{(1-\mu_0)\rho_0}\right)\right].$$

Choosing a sequence (μ^n) with $\mu_0^n \to 0$, we have $\mathscr{R}_L(\mu^n|\rho) \to \infty$, and therefore for any $C > 0$, $\mathscr{R}_L(\mu^n|\rho) \geq C\mathscr{I}(\mu)$ for a large enough n. As a result, the FIR inequality with the classical relative Fisher Information (14) does not hold in the discrete setting in general.

Remark 2.2. Note that this example did not exploit any pathological behaviour of the generator and works for all irreducible generators L on this two-point state space. Therefore we do not expect that there is a simple restriction on the class of admissible generators such that the FIR inequality (14) holds. A careful look at the example reveals that the FIR inequality fails since

$\log(\mu_0/\rho_0) \to -\infty$ as $\mu_0 \to 0$, and if such choices of μ are excluded than an FIR inequality with the relative Fisher Information might hold. This is indeed the case, as will be discussed in Lemma 2.10.

On the other hand, the generalised relative Fisher Information (15) does not suffer from the issue above since in this setting for any fixed $\lambda \in (0,1)$ we find

$$\mathscr{R}_L^\lambda(\mu|\rho) = 0 - \frac{1}{\lambda}\mathcal{H}\left(\mu, \lambda \log\left(\frac{\mu}{\rho}\right)\right) < \frac{c}{\lambda}. \tag{19}$$

It is not a coincidence that the FIR inequality holds for the generalised Fisher Information, as we prove below. □

2.2. FIR inequality with generalised relative Fisher Information

In what follows we first prove an auxiliary lemma on the structure of the generalised relative Fisher Information, which we use in Lemma 2.4 to study the consistency of its definition and discuss some simple properties. We conclude this section by giving the proof of Theorem 1.6.

For any $\rho(y), \rho(x) > 0$, the function ψ_λ in (15c) may be rewritten as

$$\psi_\lambda(x,y) = \frac{r_\lambda(v(x), v(y))}{\lambda}\rho(x), \qquad v = \frac{\mu}{\rho}, \tag{20}$$

where $(\xi,\eta) \mapsto r_\lambda(\xi,\eta) := (1-\lambda)\xi - \xi^{1-\lambda}\eta^\lambda + \lambda\eta$.

Lemma 2.3. *For any $\lambda \in (0,1)$, the function $r_\lambda : [0,\infty) \times [0,\infty) \to \mathbb{R}$ defined by*

$$r_\lambda(\xi,\eta) = (1-\lambda)\xi - \xi^{1-\lambda}\eta^\lambda + \lambda\eta,$$

satisfies the following properties:

(i) $r_\lambda \ge 0$ on $[0,\infty) \times [0,\infty)$;
(ii) $r_\lambda(\xi,\eta) = 0$ if and only if $\xi = \eta$;
(iii) For any $\xi,\eta \ge 0$, the function $\lambda \mapsto \lambda^{-1} r_\lambda(\xi,\eta)$ is monotonically decreasing on $(0,1)$;
(iv) For any $\xi,\eta > 0$, $\lim_{\lambda\to 0} \lambda^{-1} r_\lambda(\xi,\eta) = \eta - \xi + \xi\log(\frac{\xi}{\eta})$ monotonically increasing.

Proof.

(i) For any $\lambda \in (0,1)$ and $\xi,\eta \ge 0$, the Young's inequality yields

$$\xi^{1-\lambda}\eta^\lambda \le (1-\lambda)\xi + \lambda\eta,$$

and the non-negativity of r_λ follows by simply rearranging the terms.

(ii) The reverse implication follows trivially by inserting $\xi = \eta$. Now assume that $r_\lambda(\xi,\eta) = 0$. If $\xi = 0$, it follows that $\eta = 0$ and vice versa. Therefore without the loss of generality we assume that $\xi > 0$, which implies that $\eta > 0$. By rewriting

$$r_\lambda(\xi,\eta) = \xi\big((1-\lambda) - s^\lambda + \lambda s\big), \qquad s = \eta/\xi,$$

and noting that the function $s \mapsto s^\lambda$ is strictly concave on $(0,\infty)$, we deduce that the expression within the bracket vanishes if and only if $s = 1$, i.e. $\eta = \xi$.

(iii) If $\xi = 0 = \eta$, there is nothing to show. Suppose $\xi = 0$, then $\lambda^{-1} r_\lambda(\xi,\eta) = \eta$, i.e. $\lambda^{-1} r_\lambda(\xi,\eta)$ is constant in λ and therefore monotonically decreasing. If $\eta = 0$ and $\xi > 0$, then $\lambda^{-1} r_\lambda(\xi,\eta) = (1/\lambda - 1)\xi$, which is monotonically decreasing in λ since

$\lambda \mapsto 1/\lambda$ is monotonically decreasing. For $\xi, \eta > 0$, we begin by observing that $\lambda \mapsto \lambda^{-1} r_\lambda(\xi, \eta) \in \mathcal{C}^1((0,1))$, with

$$\frac{d}{d\lambda}\frac{r_\lambda(\xi,\eta)}{\lambda} = \frac{\xi}{\lambda^2}\big(s^\lambda - 1 - s^\lambda \log s^\lambda\big), \qquad s = \eta/\xi.$$

Since $\alpha \mapsto \alpha \log \alpha$ is convex on $(0, \infty)$, it follows that $s^\lambda \log s^\lambda \geq s^\lambda - 1$, and therefore $\lambda^{-1} r_\lambda(\xi, \eta)$ is monotonically decreasing in λ.

(iv) Let $\xi, \eta > 0$ and set $s = \eta/\xi$. Using l'Hospital's formula it follows that

$$\lim_{\lambda\to 0}\frac{r_\lambda(\xi,\eta)}{\lambda} = \eta - \xi - \xi \lim_{\lambda\to 0}\left(\frac{s^\lambda - 1}{\lambda}\right) = \eta - \xi - \xi \lim_{\lambda\to 0}\left(\frac{e^{\lambda \log(s)} - 1}{\lambda}\right) = \eta - \xi - \xi \log(s), \tag{21}$$

The monotonically increasing convergence holds due to (iii). □

Lemma 2.4. *The two definitions in Definition 1.5 are consistent; that is, whenever both definitions apply, they give the same value. Additionally,*

(i) $\mathscr{R}^\lambda_L(\mu|\rho) \geq 0$ *for all* $\mu, \rho \in \mathcal{P}(\mathcal{X})$*;*
(ii) $\mathscr{R}^\lambda_L$ *is lower-semicontinuous on* $\mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X})$*.*

Proof. Using the Hamiltonian (10), it is easy to check that the definitions (15a) and (15b) agree for any $\rho, \mu \in \mathcal{P}_+(\mathcal{X})$ with $\sup_{x\in\mathcal{X}} \max\{\mu(x)/\rho(x), \rho(x)/\mu(x)\} < \infty$, which proves the consistency of Definition 1.5.

(i) Since $\psi_\lambda(x, x) = 0$ for all x, the diagonal in the double sum in (15c) vanishes. So we consider $x \neq y$, for which $L(x, y) \geq 0$. If $\mu(x) = 0$ or $\rho(y) = 0$, then $\psi_\lambda(x, y) \geq 0$; if $\rho(y) > 0$, $\psi_\lambda(x, y) = 0$ if $\rho(x) = 0$ and $\psi_\lambda(x, y) \geq 0$ (due to (20) and the non-negativity of r_λ in Lemma 2.3) if $\rho(x) > 0$. Therefore $L(x, y)\psi_\lambda(x, y) \geq 0$ for all x, y, and $\mathscr{R}^\lambda_L(\mu|\rho) \geq 0$.

(ii) Let $((\mu^n, \rho^n))_{n\in\mathbb{N}} \subset \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X})$ be a sequence that converges to (μ, ρ). In particular, $\mu^n(x) \to \mu(x)$ and $\rho^n(x) \to \rho(x)$ for every $x \in \mathcal{X}$ (cf. Remark 1.1).
Now let $x \in \mathcal{X}$ be arbitrary and consider $y \in \mathcal{X}$ with $L(x, y) > 0$. For simplicity, we denote

$$\psi^n(x,y) = \frac{\mu^n(y)}{\rho^n(y)}\rho^n(x) - \mu^n(x) - \frac{1}{\lambda}\left(\mu^n(x)^{1-\lambda}\rho^n(x)^\lambda\left(\frac{\mu^n(y)}{\rho^n(y)}\right)^\lambda - \mu^n(x)\right).$$

Case 1: $(\rho(y) = \alpha > 0)$ Due to the pointwise convergence, there exists an $\alpha' > 0$ such that $\rho^n(y) > \alpha'$ for sufficiently large n. In this case, we easily conclude that $\psi^n(x, y) \to \psi(x, y)$ as $n \to \infty$.

Case 2: $(\rho(y) = 0,\ \rho(x), \mu(y) \geq \beta > 0)$ As before, there exists a $\beta' > 0$ such that $\rho^n(x), \mu^n(y) > \beta'$ for sufficiently large n. Further, we have $\mu(x), \rho(x) \in [0, M]$ for all $x \in \mathcal{X}$, with some $M \geq 1$. Therefore,

$$\begin{aligned}\psi^n(x,y) &\geq (\beta')^2\frac{1}{\rho^n(y)} - M - \frac{1}{\lambda}M^{1+\lambda}\left(\frac{1}{\rho^n(y)}\right)^\lambda \\ &= \frac{1}{\rho^n(y)}\underbrace{\left[(\beta')^2 - \frac{1}{\lambda}M^{1+\lambda}(\rho^n(y))^{1-\lambda}\right]}_{(*)} - M.\end{aligned}$$

Since $(\rho^n(y))^{1-\lambda} \to 0$ as $n \to \infty$, it follows that $(\beta')^2 \geq (*) \geq \delta$ for some $\delta > 0$ and sufficiently large n. Consequently, $\psi^n(x, y) \to \infty$ as $n \to \infty$.

The other cases are trivial since $\psi^n(x, y) \geq 0$. An application of Fatou's lemma yields

$$\liminf_{n\to\infty} \mathscr{R}_L^\lambda(\mu^n|\rho^n) \geq \sum_{x,y\in\mathcal{X}} L(x, y) \liminf_{n\to\infty} \psi^n(x, y) \geq \sum_{x,y\in\mathcal{X}} L(x, y)\psi(x, y) = \mathscr{R}_L^\lambda(\mu|\rho),$$

thereby concluding the proof. □

We are now in a position to prove the first main result of this paper.

Proof of Theorem 1.6. We proceed by approximation. Let $\rho \in AC([0, T]; \mathcal{P}(\mathcal{X}))$ be a solution of (2). Since we assume the generator L to be bounded (3c) and irreducible (3b), it follows that $\rho_t(x) > 0$ for any $t > 0$ and $x \in \mathcal{X}$ (see Lemma C.1 for a proof). Without loss of generality we can assume that $\mu \in AC([0, T]; \mathcal{P}(\mathcal{X}))$, since by Theorem 1.4 this is implied by $\mathscr{I}_L(\mu) < \infty$. Using Lemma A.1 we find $\rho, \mu \in W^{1,1}(0, T; \ell^1(\mathcal{X}))$ and therefore $\partial_t\rho, \partial_t\mu \in L^1(0, T; \ell^1(\mathcal{X}))$.

For $\varepsilon > 0$, define the function $\rho_t^\varepsilon(x) := \rho_t(x) + \varepsilon\mu_t(x)$. Since $\mu \ll \rho^\varepsilon$, we can define the density

$$v_t^\varepsilon(x) := \frac{\mu_t(x)}{\rho_t^\varepsilon(x)} \in \left[0, \frac{1}{\varepsilon}\right].$$

Note that $v_t^\varepsilon(x) \to \mu_t(x)/\rho_t(x)$ as $\varepsilon \to 0$ for all $x \in \mathcal{X}$ and $t > 0$.

Since $\log(v^\varepsilon + \delta) \in L^\infty(0, T; \ell^\infty(\mathcal{X}))$ for any $\delta \in (0, 1)$, using the representation (13) we find

$$\frac{1}{\lambda}\mathscr{I}_L(\mu) \geq \int_0^T \langle\log(v_t^\varepsilon + \delta), \partial_t\mu_t\rangle - \frac{1}{\lambda}\mathcal{H}(\mu_t, \lambda\log(v_t^\varepsilon + \delta))\, dt.$$

We split the proof into two steps, where the first step deals with passing $\delta \to 0$ and the second step with passing $\varepsilon \to 0$.

Step 1: Taking the liminf ($\delta \to 0$) in the previous inequality yields

$$\begin{aligned}\frac{1}{\lambda}\mathscr{I}_L(\mu) &\geq \liminf_{\delta\to 0}\left\{\int_0^T \langle\log(v_t^\varepsilon + \delta), \partial_t\mu_t\rangle\, dt\right\}\\ &\quad - \frac{1}{\lambda}\limsup_{\delta\to 0}\left\{\int_0^T \mathcal{H}(\mu_t, \lambda\log(v_t^\varepsilon + \delta))\, dt\right\}\\ &= (I) - \frac{1}{\lambda}(II).\end{aligned}$$

We now study both these terms.

Part (I): Define the function $g_{\varepsilon,\delta} : [0, \infty) \times (0, \infty) \to \mathbb{R}$ by

$$g_{\varepsilon,\delta}(\eta, \xi) := \eta\log\left(\frac{\eta}{\varepsilon\eta + \xi} + \delta\right).$$

For fixed ε, δ, the function $g_{\varepsilon,\delta}$ is globally Lipschitz on $A := [0, \infty)\times(0, \infty)$, and differentiable at each $(\eta, \xi) \in A$. Since $\rho_t(x) > 0$ for all $t > 0$ and $x \in \mathcal{X}$, by Lemma A.3 the function $t \mapsto g_{\varepsilon,\delta}(\mu_t, \rho_t) = \mu_t(x)\log(v_t^\varepsilon(x) + \delta)$ is an element of $AC([0, T]; \ell^1(\mathcal{X}))$, and the following chain rule holds for almost every $t \in [0, T]$:

$$\begin{aligned}\frac{d}{dt}\sum_{x\in\mathcal{X}}\mu_t(x)\log(v_t^\varepsilon(x)+\delta) = \sum_{x\in\mathcal{X}}&\left(\frac{v_t^\varepsilon(x)}{v_t^\varepsilon(x)+\delta}\frac{\rho_t(x)}{\varepsilon\mu_t(x)+\rho_t(x)}+\log(v_t^\varepsilon(x)+\delta)\right)\partial_t\mu_t(x)\\ &-\sum_{x\in\mathcal{X}}v_t^\varepsilon(x)\frac{v_t^\varepsilon(x)}{v_t^\varepsilon(x)+\delta}\partial_t\rho_t^\varepsilon(x).\end{aligned}$$

From this chain rule we easily deduce

$$\begin{aligned}\int_0^T\langle\log(v_t^\varepsilon+\delta),\partial_t\mu_t\rangle\,dt &= \sum_{x\in\mathcal{X}}\mu_T(x)\log(v_T^\varepsilon(x)+\delta)-\sum_{x\in\mathcal{X}}\mu_0(x)\log(v_0^\varepsilon(x)+\delta)\\ &\quad-\int_0^T\sum_{x\in\mathcal{X}}\frac{v_t^\varepsilon(x)}{v_t^\varepsilon(x)+\delta}\frac{\rho_t(x)}{\varepsilon\mu_t(x)+\rho_t(x)}\partial_t\mu_t(x)\,dt\\ &\quad+\int_0^T\sum_{x\in\mathcal{X}}v_t^\varepsilon(x)\frac{v_t^\varepsilon(x)}{v_t^\varepsilon(x)+\delta}\partial_t\rho_t^\varepsilon(x)\,dt.\end{aligned}\tag{22}$$

We now pass to the limit $\delta \to 0$ in each of the terms on the right-hand side.

Since $\partial_t\mu, \partial_t\rho^\varepsilon \in L^1(0, T; \ell^1(\mathcal{X}))$ and $v^\varepsilon \in L^\infty(0, T; \ell^\infty(\mathcal{X}))$ we may pass to the limit $\delta \to 0$ using the dominated convergence theorem to obtain

$$\begin{aligned}&\int_0^T\sum_{x\in\mathcal{X}}v_t^\varepsilon(x)\frac{v_t^\varepsilon(x)}{v_t^\varepsilon(x)+\delta}\partial_t\rho_t^\varepsilon(x)\,dt\\ &\xrightarrow{\delta\to 0}\quad\int_0^T\sum_{x\in\mathcal{X}}v_t^\varepsilon(x)\partial_t\rho_t^\varepsilon(x)\,dt=\int_0^T\sum_{x\in\mathcal{X}}v_t^\varepsilon(x)\left[(L^T\rho_t)(x)+\varepsilon\partial_t\mu_t\right]\,dt.\end{aligned}$$

A similar argument gives

$$\begin{aligned}&\int_0^T\sum_{x\in\mathcal{X}}\frac{v_t^\varepsilon(x)}{v_t^\varepsilon(x)+\delta}\frac{\rho_t(x)}{\varepsilon\mu_t(x)+\rho_t(x)}\partial_t\mu_t(x)\,dt\\ &\xrightarrow{\delta\to 0}\quad\int_0^T\sum_{x\in\mathcal{X}}\mathbb{1}\{\mu_t(x)>0\}\,\frac{\rho_t(x)}{\varepsilon\mu_t(x)+\rho_t(x)}\partial_t\mu_t(x)\,dt.\end{aligned}$$

Turning to the first term in (22), using $\mu_t(x)\log(v_t^\varepsilon(x)+\delta) \geq \mu_t(x)\log(v_t^\varepsilon(x))$ for any $(t, x) \in [0, T] \times \mathcal{X}$, we find

$$\sum_{x\in\mathcal{X}}\mu_t(x)\log(v_t^\varepsilon(x)+\delta)\geq\sum_{x\in\mathcal{X}}\mu_t(x)\log(v_t^\varepsilon(x))=\mathscr{H}(\mu_t|\rho_t^\varepsilon).$$

At time zero, the finiteness of $\mathscr{H}(\mu_0|\rho_0)$ implies that whenever $\mu_0(x) > 0$ we have $\rho_0(x) > 0$, and therefore the density $v_0(x) := \mu_0(x)/\rho_0(x)$ is well-defined μ_0-almost-everywhere. Using the concavity and monotonicity of the natural logarithm, for the second term in (22) we find

$$\begin{aligned}\sum_{x\in\mathcal{X}}\mu_0(x)\log(v_0^\varepsilon(x)+\delta)&=\sum_{x\in\mathcal{X}}v_0^\varepsilon(x)\log(v_0^\varepsilon(x)+\delta)\rho_0^\varepsilon(x)\\ &\leq\sum_{x\in\mathcal{X}}\mu_0(x)\log(v_0^\varepsilon(x))+\delta(1+\varepsilon)\\ &\leq\sum_{x\in\mathcal{X}}\mu_0(x)\log(v_0(x))+\delta(1+\varepsilon)=\mathscr{H}(\mu_0|\rho_0)+\delta(1+\varepsilon),\end{aligned}$$

where we have used $\rho_0^\varepsilon \geq \rho_0$ to arrive at the second inequality. Altogether, we obtain

$$\begin{aligned}\liminf_{\delta\to 0}\int_0^T \langle \log(v_t^\varepsilon+\delta), \partial_t\mu_t\rangle\, dt \geq \mathscr{H}(\mu_T|\rho_T^\varepsilon) - \mathscr{H}(\mu_0|\rho_0)&\\ +\int_0^T \sum_{x\in\mathcal{X}} v_t^\varepsilon(x)\left[(L^T\rho_t)(x)+\varepsilon\partial_t\mu_t(x)\right] dt&\\ +\int_0^T \sum_{x\in\mathcal{X}} \mathbb{1}\{\mu_t>0\}\,\frac{\rho_t(x)}{\varepsilon\mu_t(x)+\rho_t(x)}\partial_t\mu_t(x)\, dt&,\end{aligned}$$

which concludes part (I).

Part (II)*:* Using the definition (10) of the Hamiltonian, and $\sum_{y\in\mathcal{X}} L(x,y)=0$ we find

$$\begin{aligned}\mathcal{H}(\mu_t,\lambda\log(v_t^\varepsilon+\delta)) &= \sum_{x,y\in\mathcal{X}} \mu_t(x)L(x,y)\left[e^{\lambda\log(v_t^\varepsilon+\delta)(y)-\lambda\log(v_t^\varepsilon+\delta)(x)}-1\right]\\ &= \sum_{x,y\in\mathcal{X}} \mu_t(x)L(x,y)\left(\frac{v_t^\varepsilon(y)+\delta}{v_t^\varepsilon(x)+\delta}\right)^\lambda\\ &= \sum_{x,y\in\mathcal{X}} \rho_t^\varepsilon(x)(v_t^\varepsilon(x))^{1-\lambda}L(x,y)\left(\frac{v_t^\varepsilon(x)}{v_t^\varepsilon(x)+\delta}\right)^\lambda (v_t^\varepsilon(y)+\delta)^\lambda.\end{aligned}$$

We have the upper bound

$$\left|\rho_t^\varepsilon(x)(v_t^\varepsilon(x))^{1-\lambda}L(x,y)\left(\frac{v_t^\varepsilon(x)}{v_t^\varepsilon(x)+\delta}\right)^\lambda (v_t^\varepsilon(y)+\delta)^\lambda\right| \leq \varepsilon^{\lambda-1}(\varepsilon^{-1}+1)^\lambda\,\rho^\varepsilon(x)|L(x,y)|,$$

where we have used $|v_t^\varepsilon|\leq\varepsilon^{-1}$ and $\delta\in(0,1)$. Note that the right-hand side is an element of $\ell^1(\mathcal{X}\times\mathcal{X})$ since $\rho^\varepsilon\in\ell^1(\mathcal{X})$ and L satisfies (3b). Using the dominated convergence theorem we find

$$\limsup_{\delta\to 0}\int_0^T \mathcal{H}(\mu_t,\lambda\log(v_t^\varepsilon+\delta))\, dt = \int_0^T \sum_{x,y\in\mathcal{X}} \rho_t^\varepsilon(x)(v_t^\varepsilon(x))^{1-\lambda}L(x,y)(v_t^\varepsilon(y))^\lambda\, dt.$$

This concludes part (II).

Putting both the parts together, we obtain

$$\begin{aligned}\frac{1}{\lambda}\mathscr{I}_L(\mu) &\geq (I) - \frac{1}{\lambda}(II)\\ &\geq \mathscr{H}(\mu_T|\rho_T^\varepsilon) - \mathscr{H}(\mu_0|\rho_0)\\ &\qquad + \int_0^T \sum_{x,y\in\mathcal{X}} L(x,y)\rho_t^\varepsilon(x)\left[v_t^\varepsilon(y)-\frac{1}{\lambda}(v_t^\varepsilon(x))^{1-\lambda}(v_t^\varepsilon(y))^\lambda\right] dt\\ &\qquad + \varepsilon\int_0^T \sum_{x\in\mathcal{X}} v_t^\varepsilon(x)\partial_t\mu_t(x)\, dt\\ &\qquad + \int_0^T \sum_{x\in\mathcal{X}} \mathbb{1}\{\mu_t>0\}\,\frac{\rho_t(x)}{\varepsilon\mu_t(x)+\rho_t(x)}\partial_t\mu_t(x)\, dt\\ &= \mathscr{H}(\mu_T|\rho_T^\varepsilon) - \mathscr{H}(\mu_0|\rho_0) + \int_0^T \mathscr{R}_L^\lambda(\mu_t|\rho_t^\varepsilon)\, dt + \varepsilon\int_0^T \sum_{x\in\mathcal{X}} v_t^\varepsilon(x)\partial_t\mu_t(x)\, dt\\ &\qquad ++\int_0^T \sum_{x\in\mathcal{X}} \mathbb{1}\{\mu_t>0\}\,\frac{\rho_t(x)}{\varepsilon\mu_t(x)+\rho_t(x)}\partial_t\mu_t(x)\, dt,\end{aligned} \tag{23}$$

where in the final identity we used the property $\sum_{y\in\mathcal{X}} L(x,y)=0$ and (20). This inequality clearly resembles the FIR inequality.

Step 2: We now take the limit $\varepsilon\to 0$. For any $t\in(0,T]$ we have

$$\begin{aligned}\mathscr{H}(\mu_t|\rho_t^\varepsilon) &= \sum_{x\in\mathcal{X}}\mu_t(x)\log(v_t^\varepsilon(x)) = \sum_{x\in\mathcal{X}} v_t^\varepsilon(x)\log(v_t^\varepsilon(x))\rho_t^\varepsilon(x)\\ &= \sum_{x\in\mathcal{X}}\Big[v_t^\varepsilon(x)(\log(v_t^\varepsilon(x))-1)+1\Big]\rho_t^\varepsilon(x) + \sum_{x\in\mathcal{X}}[\mu_t(x)-\rho_t^\varepsilon(x)]\\ &= \sum_{x\in\mathcal{X}}\Big[v_t^\varepsilon(x)(\log(v_t^\varepsilon(x))-1)+1\Big]\rho_t^\varepsilon(x)-\varepsilon.\end{aligned}$$

The final inequality follows since $\sum_{x\in\mathcal{X}}\rho_t^\varepsilon(x)=1+\varepsilon$. The summand in the final right-hand side is non-negative, and for each x and t such that $\rho_t(x)>0$ we have $v_t^\varepsilon(x)\to v_t(x)=\mu_t(x)/\rho_t(x)$ for $\varepsilon\to 0$. We therefore apply Fatou's lemma to obtain

$$\begin{aligned}\liminf_{\varepsilon\to 0}\mathscr{H}(\mu_t|\rho_t^\varepsilon) &\ge \liminf_{\varepsilon\to 0}\sum_{x\in\mathcal{X}}\Big[v_t^\varepsilon(x)\log(v_t^\varepsilon(x))-v_t^\varepsilon(x)+1\Big]\rho_t^\varepsilon(x)\\ &\ge \liminf_{\varepsilon\to 0}\sum_{x\in\mathcal{X}}\Big[v_t^\varepsilon(x)\log(v_t^\varepsilon(x))-v_t^\varepsilon(x)+1\Big]\rho_t(x)\\ &= \sum_{x\in\mathcal{X}}\Big[v_t(x)(\log(v_t(x))-1)+1\Big]\rho_t(x) = \mathscr{H}(\mu_t|\rho_t).\end{aligned}$$

As for the other expression, we use the non-negativity and lower-semicontinuity of $\mathscr{R}_L^\lambda$ (recall Lemma 2.4 and Remark 2.1) to obtain

$$\liminf_{\varepsilon\to 0}\int_0^T \mathscr{R}_L^\lambda(\mu_t|\rho_t^\varepsilon)\,dt \ge \int_0^T \liminf_{\varepsilon\to 0}\mathscr{R}_L^\lambda(\mu_t|\rho_t^\varepsilon)\,dt = \int_0^T \mathscr{R}_L^\lambda(\mu_t|\rho_t)\,dt. \tag{24}$$

Since $\varepsilon v_t^\varepsilon(x)$ is uniformly bounded for every $t\in(0,T]$ and $x\in\mathcal{X}$, we can pass $\varepsilon\to 0$ in the final term of (23) using the dominated convergence theorem, which gives

$$\lim_{\varepsilon\to 0}\varepsilon\int_0^T\sum_{x\in\mathcal{X}} v_t^\varepsilon(x)\partial_t\mu_t(x)dt = 0.$$

Finally, since $\frac{\rho_t(x)}{\varepsilon\mu_t(x)+\rho_t(x)}\le 1$ and $\rho_t(x)>0$ for any $t>0$ by irreducibility of L, we can again apply dominated convergence theorem and obtain

$$\lim_{\varepsilon\to 0}\int_0^T\sum_{x\in\mathcal{X}}\mathbb{1}\{\mu_t>0\}\,\frac{\rho_t(x)}{\varepsilon\mu_t(x)+\rho_t(x)}\partial_t\mu_t(x)\,dt = \int_0^T\sum_{x\in\mathcal{X}}\mathbb{1}\{\mu_t(x)>0\}\,\partial_t\mu_t(x)\,dt$$

This limit is equal to zero, as we now show using another application of the dominated convergence theorem. Let $H_m:\mathbb{R}\to[0,1]$ be a smooth approximation of the Heaviside function H with $H_m(s)=0$ for $s\le 0$ and $H_m(s)\uparrow 1$ for $s>0$ as $m\to\infty$; set $f_m(s)=\int_0^s H_m(\sigma)\,d\sigma$. Since f_m is Lipschitz, $t\mapsto f_m(\mu_t(\cdot))$ is again absolutely continuous by Lemma A.3, and we have the chain rule

$$\sum_{x\in\mathcal{X}}\Big[f_m(\mu_T(x))-f_m(\mu_0(x))\Big] = \int_0^T\sum_{x\in\mathcal{X}} H_m(\mu_t(x))\partial_t\mu_t(x)\,dt.$$

Using the dominated convergence theorem on both sides, we pass to the limit $m\to\infty$ to find

$$0=\sum_{x\in\mathcal{X}}\Big[\mu_T(x)-\mu_0(x)\Big] = \int_0^T\sum_{x\in\mathcal{X}}\mathbb{1}\{\mu_t(x)>0\}\,\partial_t\mu_t(x)\,dt.$$

Putting the results of the two steps together, we obtain

$$\frac{1}{\lambda}\mathscr{I}_L(\mu) \geq \mathscr{H}(\mu_T|\rho_T) - \mathscr{H}(\mu_0|\rho_0) + \int_0^T \mathscr{R}_L^\lambda(\mu_t|\rho_t)\,dt,$$

which concludes the proof of the FIR inequality. □

2.3. Properties of the generalised relative Fisher Information

Given the set $\mathcal{X}$ and the operator L, we define a graph with vertices $\mathcal{X}$ and un-oriented edges $\mathcal{E} \subset \mathcal{X} \times \mathcal{X}$ as follows:

$$(x, y) \in \mathcal{E} \quad \Longleftrightarrow \quad L(x, y) > 0 \quad \text{or} \quad L(y, x) > 0.$$

The interpretation of this graph is that two vertices are connected if they are a single jump of the Markov process apart, in either direction. In this graph, the support $\operatorname{supp}(\rho) := \{x \in \mathcal{X} : \rho(x) > 0\}$ is a subset of the vertices, and defines a subgraph by deleting all edges that do not connect two vertices in $\operatorname{supp}(\rho)$. Furthermore, we can decompose $\operatorname{supp}(\rho)$ into connected components Ω_i, i.e. $\operatorname{supp}(\rho) = \cup_{i\in I}\Omega_i$ and for every pair $x, y \in \Omega_i$ there exists a finite sequence $(x_n)_{n=1,\dots,N}$ in Ω_i with $x_1 = x$, $x_N = y$ and the vertices x_n and x_{n+1} are connected for all $n = 1, \dots, N-1$.

Lemma 2.5. *Let $\mu, \rho \in \mathcal{P}(\mathcal{X})$, and let* $\operatorname{supp}(\rho)$ *be decomposed into connected components Ω_i. If $\mu = \rho$ then $\mathscr{R}_L^\lambda(\mu|\rho) = 0$. Further, if $\mathscr{R}_L^\lambda(\mu|\rho) = 0$, then there exist numbers $a_i \geq 0$, $i \in I$, such that $\mu(x) = a_i\rho(x)$ for all $x \in \Omega_i$. In particular, if $\rho(x) > 0$ for all $x \in \mathcal{X}$ and L is irreducible, then $\mu = \rho$.*

Proof. The fact that $\mu = \rho$ implies $\mathscr{R}_L^\lambda(\mu|\rho) = 0$ follows from the definition of $\mathscr{R}_L^\lambda$. Assume now that $\mathscr{R}_L^\lambda(\mu|\rho) = 0$ for $\mu, \rho \in \mathcal{P}(\mathcal{X})$. Let Ω_i be a connected component of the support of ρ, where we exclude the trivial cases that μ vanishes identically on Ω_i or that Ω_i only contains one vertex. We now show that if μ does not vanish identically it is strictly positive on Ω_i. Assume that $\mu|_{\Omega_i} \not> 0$; since Ω_i is a connected subgraph there exists $x, y \in \Omega_i$ such that $L(x, y) > 0$ and either $\mu(x) > 0$ and $\mu(y) = 0$ or $\mu(x) = 0$ and $\mu(y) > 0$. In the first case, we estimate using (15c) (recall that $\rho(x) > 0$ and $\rho(y) > 0$) that

$$\mathscr{R}_L^\lambda(\mu|\rho) \geq L(x, y)\left(-\mu(x) + \frac{1}{\lambda}\mu(x)\right) > 0,$$

since $\lambda \in (0, 1)$. In the second case, we obtain

$$\mathscr{R}_L^\lambda(\mu|\rho) \geq L(x, y)\left(\frac{\mu(y)}{\rho(y)}\rho(x)\right) > 0.$$

Therefore, in both cases we obtain a contradiction to $\mathscr{R}_L^\lambda(\mu|\rho) = 0$ and thus, $\mu|_{\Omega_i} > 0$.

Now, let $x, y \in \Omega_i$ be arbitrary. Since Ω_i is a connected, there exists a finite sequence $(x_n)_{n=1,\dots,N}$ with $x_1 = x, x_N = y$ and either $L(x_n, x_{n+1}) > 0$ or $L(x_{n+1}, x_n) > 0$ for all $n = 1, \dots, N-1$. Furthermore, $\rho > 0$ on Ω_i and thus (cf. (20)),

$$0 = \mathscr{R}_L^\lambda(\mu|\rho) \geq L(x, y)\rho(x)\frac{r_\lambda(v(x), v(y))}{\lambda} \geq 0, \qquad v = \mu/\rho$$

for all $x, y \in \Omega_i$ and hence, especially

$$r_\lambda(v(x_n), v(x_{n+1})) = 0 \text{ or } r_\lambda(v(x_{n+1}), v(x_n)) = 0.$$

Using Lemma 2.3, this is true if and only if $v(x_n) = v(x_{n+1})$ and thus,

$$\frac{\mu(x_{n-1})}{\rho(x_{n-1})} = \frac{\mu(x_n)}{\rho(x_n)} = \frac{\mu(x_{n+1})}{\rho(x_{n+1})} \qquad \text{for all } n = 2, \dots, N-1.$$

Since the pair x, y was arbitrarily chosen, it follows that there exists a constant $a > 0$ such that $\mu(x) = a\rho(x)$ for all $x \in \Omega_i$.

Finally, if $\rho(x) > 0$ for every $x \in \mathcal{X}$ and L is irreducible, then $\mathcal{X}$ itself is a connected component and we can apply the previous result. Furthermore, since μ, ρ have the same mass, i.e. $\mu(\mathcal{X}) = \rho(\mathcal{X})$, we have $a = 1$ in this case. □

Remark 2.6. Note that no claim is made about $\mu(x)$ for $x \notin \operatorname{supp}(\rho)$; see Example 2.7 in which $\mathscr{R}^\lambda_L(\mu|\rho) = 0$, but there exist $x \in \mathcal{X}$ with $\rho(x) = 0$ and $\mu(x) > 0$. However, if one assumes additionally that $\mathscr{H}(\mu|\rho) < \infty$, then necessarily $\mu(x) = 0$ for all $x \notin \operatorname{supp}(\rho)$. □

Example 2.7. We now give an example of ρ, μ, such that $\mathscr{R}^\lambda_L(\mu|\rho) = 0$ and $\rho(x) = 0$ but $\mu(x) > 0$ for some $x \in \mathcal{X}$. Let $w, z \in \mathcal{X}$ and L such that $L(x,z) = 0$ as well as $L(z,x) = 0$ for all $x \neq w$. We consider $\mu = \delta_z$ and ρ with $\operatorname{supp}(\rho) = \mathcal{X} \setminus \{w,z\}$. The corresponding generalised relative Fisher information (15c) is

$$\begin{aligned}\mathscr{R}^\lambda_L(\mu|\rho) = &\sum_{x,y\in\mathcal{X}\setminus\{w,z\}} L(x,y)\psi_\lambda(x,y) \\ &+ \sum_{x\in\mathcal{X}\setminus\{w,z\}} [L(x,z)\psi_\lambda(x,z) + L(z,x)\psi_\lambda(z,x) \\ &+ L(x,w)\psi_\lambda(x,w) + L(w,x)\psi_\lambda(w,x)] \\ &+ L(w,z)\psi_\lambda(w,z) + L(z,w)\psi_\lambda(z,w).\end{aligned}$$

By the definition of ψ_λ, the first summation vanishes since $\mu(x) = \mu(y) = 0$ for $x, y \in \mathcal{X}\setminus\{w,z\}$. Regarding the second summation, note that $L(x,z) = L(z,x) = 0$ by assumption and thus the first two terms vanish. Furthermore, $\psi_\lambda(x,w) = 0$ since $\mu(w) = 0$ and $\psi_\lambda(w,x) = 0$ since $\rho(w) = 0$, and thus the remaining two terms vanish. The last two terms in the equality above also vanish since $\rho(w) = \rho(z) = 0$. This show that $\mathscr{R}^\lambda_L(\mu|\rho) = 0$ but $\mu(z) = 1 > 0$ while $\rho(z) = 0$, i.e. there does not exist any $a > 0$ such that $\mu(x) \neq a\rho(x)$ for $x \notin \operatorname{supp}(\rho)$. Additionally, this gives an example for which $\mu = a\rho$ holds on a subgraph $\Omega = \mathcal{X} \setminus \{w,z\}$ with $a = 0$. □

Next we turn to the asymptotic behaviour of $\mathscr{R}^\lambda_L$ in the limit $\lambda \to 0$, described by Lemma 2.8. Before presenting the result, we first formally derive the limit which in this case is the relative Fisher Information (5). Using (11), for any $\lambda \in (0,1)$ and $f \in \ell^\infty(\mathcal{X})$ we find

$$\frac{1}{\lambda}\mathcal{H}(\mu, \lambda f) = \sup_{s\in\ell^1(\mathcal{X})} \left\{ \sum_{x\in\mathcal{X}} f(x)s(x) - \frac{1}{\lambda}\mathcal{L}(\mu, s) \right\} \geq \sum_{x\in\mathcal{X}} f(x)(L^T\mu)(x),$$

where we have chosen $s = L^T\mu$ and used $\mathcal{L}(\mu, L^T\mu) = 0$ (cf. (8)) to arrive at the inequality. Substituting this into (15a) we arrive at

$$\mathscr{R}^\lambda_L(\mu|\rho) \leq \sum_{x,y\in\mathcal{X}} L(x,y)\frac{\mu(y)}{\rho(y)}\rho(x) - \sum_{x\in\mathcal{X}} L\log\left(\frac{\mu}{\rho}\right)(x)\mu(x) = \mathscr{R}_L(\mu|\rho),$$

where $\mathscr{R}_L(\cdot|\cdot)$ is defined in (5). Since $\mathcal{L}$ is the Lagrangian corresponding to the operator L, it follows that $\mathcal{L}(\mu, s) > 0$ if $s \neq L^T\mu$ (recall the properties below (7)). Hence for small λ, the deviations from $s = L^T\mu$ are penalised in the definition of the Hamiltonian (11) and therefore for $\lambda \to 0$ we expect that the supremum is attained at $s = L^T\mu$, i.e.

$$\lim_{\lambda \searrow 0} \frac{1}{\lambda}\mathcal{H}(\mu, \lambda f) = \sum_{x\in\mathcal{X}} f(x)(L^T\mu)(x) = \sum_{x,y\in\mathcal{X}} \mu(x)L(x,y)(f(y) - f(x)).$$

Substituting this in (15a) we expect that $\mathscr{R}_L^\lambda \xrightarrow{\lambda\to 0} \mathscr{R}_L$. We make this intuition rigorous in the next result.

Lemma 2.8. (i) *For all* $\mu, \rho \in \mathcal{P}_+(\mathcal{X})$, $\lim_{\lambda\searrow 0}\mathscr{R}_L^\lambda(\mu|\rho) = \mathscr{R}_L(\mu|\rho)$ *monotonically increasing.*
(ii) Γ-$\lim_{\lambda\searrow 0}\mathscr{R}_L^\lambda = \mathscr{R}_L$ *on* $\mathcal{P}_+(\mathcal{X}) \times \mathcal{P}_+(\mathcal{X})$.

Proof. (i) Set $v = \mu/\rho$. Using (15c), (20) we find

$$\mathscr{R}_L^\lambda(\mu|\rho) = \sum_{x,y\in\mathcal{X}} L(x,y)\rho(x)\frac{r_\lambda(v(x), v(y))}{\lambda}.$$

Using Lemma 2.3 and applying the monotone convergence theorem we find

$$\begin{aligned}\lim_{\lambda\to 0}\mathscr{R}_L^\lambda(\mu_\lambda|\rho_\lambda) &= \sum_{x,y\in\mathcal{X}} L(x,y)\rho(x)\left(\lim_{\lambda\to\infty}\frac{r_\lambda(v(x), v(y))}{\lambda}\right)\\ &= \sum_{x,y\in\mathcal{X}} L(x,y)\rho(x)\left[v(y) - v(x) + v(x)\log\left(\frac{v(x)}{v(y)}\right)\right] = \mathscr{R}_L(\mu|\rho).\end{aligned}$$

The monotonicity of the convergence follows from the monotonicity of $\lambda \mapsto \lambda^{-1}r_\lambda$ in Lemma 2.3.

(ii) The proof of the Γ-limit consists of a liminf and a limsup inequality (see [5, Section 1.2] for details).

The liminf inequality states that for any sequences $(\mu_\lambda)_{\lambda\geq 0}, (\rho_\lambda)_{\lambda\geq 0} \subset \mathcal{P}_+(\mathcal{X})$ which converge in $\ell^1(\mathcal{X})$ (and therefore pointwisely) to $\mu, \rho \in \mathcal{P}_+(\mathcal{X})$ as $\lambda \to 0$, we have

$$\liminf_{\lambda\to 0}\mathscr{R}_L^\lambda(\mu_\lambda|\rho_\lambda) \geq \mathscr{R}_L(\mu|\rho). \tag{25}$$

Using the definition (15c) of $\mathscr{R}_L^\lambda$, (20) and Lemma 2.3, we find with Fatou's lemma that

$$\begin{aligned}\liminf_{\lambda\to 0}\mathscr{R}_L^\lambda(\mu_\lambda|\rho_\lambda) &= \liminf_{\lambda\to 0}\sum_{x,y\in\mathcal{X}} \rho_\lambda(x)L(x,y)\frac{r_\lambda(v_\lambda(x), v_\lambda(y))}{\lambda}\\ &\geq \sum_{x,y\in\mathcal{X}} \rho(x)L(x,y)\liminf_{\lambda\to 0}\frac{r_\lambda(v_\lambda(x), v_\lambda(y))}{\lambda},\end{aligned}$$

where $v_\lambda := \mu_\lambda/\rho_\lambda$. To complete the proof of the liminf inequality (25) we need to bound the right hand side of the inequality above by the relative Fisher Information. Setting $s_\lambda(x,y) = v_\lambda(y)/v_\lambda(x)$, we find

$$\liminf_{\lambda\to 0}\frac{r_\lambda(v_\lambda(x), v_\lambda(y))}{\lambda} = v(y) - v(x) - \limsup_{\lambda\to 0}\left\{v_\lambda(x)\left(\frac{s_\lambda(x,y)^\lambda - 1}{\lambda}\right)\right\}.$$

Due to the pointwise convergence $v_\lambda \to v$, we have that $s_\lambda(x,y) \to s(x,y) = v(y)/v(x)$. In particular, for any $\varepsilon > 0$, we find a $\lambda_\varepsilon > 0$ such that $|s_\lambda(x,y) - s(x,y)| < \varepsilon$ for all $\lambda \in (0, \lambda_\varepsilon)$. Consequently, $0 < s_\lambda(x,y) < s(x,y) + \varepsilon$ for $\lambda \in (0, \lambda_\varepsilon)$, which yields

$$\frac{s_\lambda(x,y)^\lambda - 1}{\lambda} < \frac{(s(x,y)+\varepsilon)^\lambda - 1}{\lambda} \qquad \text{for all } \lambda \in (0, \lambda_\varepsilon).$$

Multiplication with $v_\lambda(x)$ and passing to the limit $\lambda \to 0$, we then obtain (cf. (21))

$$\limsup_{\lambda\to 0} \left\{ v_\lambda(x) \left(\frac{s_\lambda(x,y)^\lambda - 1}{\lambda} \right) \right\} \le v(x)\log(s(x,y)+\varepsilon).$$

Since $\varepsilon > 0$ may be chosen arbitrarily small, we obtain

$$\liminf_{\lambda\to 0} \mathscr{R}^\lambda_L(\mu_\lambda|\rho_\lambda) \ge \sum_{x,y\in\mathcal{X}} \rho(x)L(x,y)\left[v(y) - v(x) + v(x)\log\left(\frac{v(x)}{v(y)}\right)\right] = \mathscr{R}_L(\mu|\rho),$$

as required.

Next we prove the limsup inequality, wherein for fixed $\mu, \rho \in \mathcal{P}_+(\mathcal{X})$ we need to prove the existence of a sequence $(\mu_\lambda)_{\lambda\ge 0}, (\rho_\lambda)_{\lambda\ge 0}$ in $\mathcal{P}_+(\mathcal{X})$ which satisfies

$$\limsup_{\lambda\to 0} \mathscr{R}^\lambda_L(\mu_\lambda|\rho_\lambda) \le \mathscr{R}_L(\mu|\rho).$$

Due to (i) we immediately see that the constant sequence for $(\mu_\lambda)_{\lambda\ge 0}, (\rho_\lambda)_{\lambda\ge 0}$, i.e. $\mu_\lambda = \mu$ and $\rho_\lambda = \rho$ for all $\lambda > 0$ does the job, which completes the proof. □

Remark 2.9 (*Role of Irreducibility*)**.** While from the very outset we have assumed that the generator L is irreducible (cf. (3c)), it is worth noting that the definition of the generalised Fisher Information (15c) is well defined even when this does not hold. Furthermore the various properties of the generalised Fisher Information outlined in this and the previous section do not require irreducibility as well. However, irreducibility of the generator is required to prove the FIR inequality in Theorem 1.6. □

2.4. Modified FIR for classical relative Fisher Information

In what follows, we use the convergence result in Lemma 2.8 to prove a FIR-inequality with the classical relative Fisher Information (5) by restricting the class of admissible curves μ. In the next result we provide sufficient conditions under which

$$(1-\gamma)\mathscr{R}_L(\mu|\rho) \le \mathscr{R}^\lambda_L(\mu|\rho)$$

for some $\gamma \in (0,1)$. Recall from our discussion in Section 2.1 that this is not true in general since we can construct a sequence for which the relative Fisher Information is unbounded while the rate functional is bounded (and therefore the generalised Fisher Information is bounded by Theorem 1.6). In fact, from Lemma 2.8 we know that the generalised Fisher Information $\mathscr{R}^\lambda_L$ is always bounded from above by the Fisher Information $\mathscr{R}_L$, and in the following result we show that the inequality can be reversed under certain conditions.

Lemma 2.10. *Fix $K < \infty$, $\lambda \in (0,1)$ and let $\mu, \rho \in \mathcal{P}_+(\mathcal{X})$ satisfy*

$$\sup_{x\in\mathcal{X}} \left| \log\left(\frac{\mu(x)}{\rho(x)}\right)\right| \le K.$$

Then there exists a $\gamma = \gamma(K,\lambda) > 0$ such that

$$(1-\gamma)\mathscr{R}_L(\mu|\rho) \le \mathscr{R}^\lambda_L(\mu|\rho). \tag{26}$$

Furthermore for every $K < \infty$ there exists a $\lambda_0 \in (0,1)$ such that $\gamma(K,\lambda) \in (0,1)$ for all $\lambda \in (0,\lambda_0)$.

Proof. The uniform bound on the logarithm implies that $\mathscr{R}_L(\mu|\rho)$ is well-defined. Using the definitions of these objects we can rewrite (26) as

$$\begin{aligned}\frac{1}{\lambda}\mathcal{H}\left(\mu,\lambda\log\left(\frac{\mu}{\rho}\right)\right)-\sum_{x,y\in\mathcal{X}}\mu(x)L(x,y)\log\left(\frac{\mu(y)\rho(x)}{\rho(y)\mu(x)}\right)&=\mathscr{R}_L(\mu|\rho)-\mathscr{R}_L^{\lambda}(\mu|\rho)\\ \le\gamma\mathscr{R}_L(\mu|\rho)=\gamma\left[\mathcal{H}\left(\mu,\log\left(\frac{\mu}{\rho}\right)\right)-\sum_{x,y\in\mathcal{X}}\mu(x)L(x,y)\log\left(\frac{\mu(y)\rho(x)}{\rho(y)\mu(x)}\right)\right].\end{aligned}$$

To simplify the notation, we define

$$\begin{aligned}\mathcal{D}(\mu,f)&:=\mathcal{H}(\mu,f)-\sum_{x,y\in\mathcal{X}}\mu(x)L(x,y)(f(y)-f(x))\\&=\sum_{x,y\in\mathcal{X}}\mu(x)L(x,y)\left[e^{\nabla f(y,x)}-(1+\nabla f(y,x))\right],\end{aligned}$$

where $\nabla f(y,x)=f(y)-f(x)$. Using the Taylor expansion of the exponential, we estimate

$$\begin{aligned}\mathcal{D}(\mu,\lambda f)&\le\sum_{x,y\in\mathcal{X}}\mu(x)L(x,y)\sum_{n\ge2}\lambda^n\frac{|\nabla f(y,x)|^n}{n!}\\&=\lambda^2\sum_{x,y\in\mathcal{X}}\mu(x)L(x,y)\sum_{n\ge2}\lambda^{n-2}\frac{|\nabla f(y,x)|^n}{n!}\\&\le\lambda^2\sum_{x,y\in\mathcal{X}}\mu(x)L(x,y)\sum_{n\ge2}\frac{|\nabla f(y,x)|^n}{n!}\\&=\lambda^2\sum_{x,y\in\mathcal{X}}\mu(x)L(x,y)\left[e^{|\nabla f(y,x)|}-(1+|\nabla f(y,x)|)\right]=:\lambda^2\tilde{\mathcal{D}}(\mu,f),\end{aligned}$$

where the second inequality follows since $\lambda\in(0,1)$. Next, we show that there exists a $c_K>0$ only depending on K such that $\mathcal{D}(\mu,f)\ge c_K\tilde{\mathcal{D}}(\mu,f)$ uniformly for all f with $\|f\|_\infty\le K$. This is equivalent to proving that

$$\varphi(\alpha):=\frac{e^\alpha-(1+\alpha)}{e^{|\alpha|}-(1+|\alpha|)}\ge c_K$$

for $\alpha\in[-2K,2K]$. If $\alpha>0$, then $\varphi(\alpha)=1$ and hence, it is sufficient to consider $\alpha\le0$. By using l'Hospital, we can continuously extend φ to $\alpha=0$ by defining $\varphi(0)=1$. Furthermore, φ is positive and monotonically decreasing for $\alpha<0$. Since $[-2K,2K]$ is compact, the existence of $c_K>0$ follows from the continuity and positivity of φ.

We thus established that for every $K<\infty$, there exists a $c_K>0$ only depending on K such that

$$\frac{1}{\lambda}\mathcal{D}\left(\mu,\lambda\log\left(\frac{\mu}{\rho}\right)\right)\le\frac{\lambda}{c_K}\mathcal{D}\left(\mu,\log\left(\frac{\mu}{\rho}\right)\right).$$

Choosing $\gamma=\lambda/c_K>0$ then yields (26) and for all $\lambda<c_K$, we obtain $\gamma\in(0,1)$. □

Using this result along with Theorem 1.6 we arrive at a modified FIR inequality for the classical relative Fisher Information.

Proposition 2.11. *Let $\rho \in AC([0,T];\mathcal{P}(\mathcal{X}))$ be a solution of (2) and $\mu \in C([0,T];\mathcal{P}_+(\mathcal{X}))$ satisfy $\mathscr{I}_L(\mu)+\mathscr{H}(\mu_0|\rho_0)<\infty$. Furthermore assume that there exists a $K<\infty$ such that*

$$\sup_{t\in[0,T]}\sup_{x\in\mathcal{X}}\left|\log\left(\frac{\mu_t(x)}{\rho_t(x)}\right)\right| \le K.$$

Then there exists a sufficiently small λ (see Lemma 2.10) such that

$$\mathscr{H}(\mu_T|\rho_T)+(1-\gamma)\int_0^T \mathscr{R}_L(\mu_t|\rho_t)\,dt \le \mathscr{H}(\mu_0|\rho_0)+\frac{1}{\lambda}\mathscr{I}_L(\mu),$$

with $\gamma\in(0,1)$.

Remark 2.12 (*Convexity of Generalised Fisher Information*)**.** Let $\mu,\rho\in\mathcal{P}_+(\mathcal{X})$. Using the explicit representation for the Hamiltonian (10) we find

$$\begin{aligned}\mathscr{R}^\lambda_L(\mu|\rho) &= \sum_{x,y\in\mathcal{X}} L(x,y)\left[\mu(y)\frac{\rho(x)}{\rho(y)}-\mu(x)\right]-\frac{1}{\lambda}\sum_{x,y\in\mathcal{X}}\mu(x)L(x,y)\\ &\quad\times\left[\left(\frac{\mu(y)\rho(x)}{\mu(x)\rho(y)}\right)^\lambda-1\right]\\ &= \sum_{x,y\in\mathcal{X}} L(x,y)\left[\mu(y)\frac{\rho(x)}{\rho(y)}-\mu(x)\right]-\frac{1}{\lambda}\sum_{x,y\in\mathcal{X}}L(x,y)\\ &\quad\times\left[\left(\frac{\rho(x)}{\rho(y)}\right)^\lambda\mu(y)^\lambda\mu(x)^{1-\lambda}-\mu(x)\right]\end{aligned}$$

Since $\alpha^\lambda\beta^{1-\lambda}$ is concave for $\alpha,\beta>0$ and $\lambda\in(0,1)$ it follows that the third term on the right hand side is concave in μ. Since the rest of the terms on the right hand side are linear in μ it follows that the generalised Fisher Information is convex in the first entry. □

2.5. Comparison with diffusion processes

So far we have limited our discussion to Markov jump processes. In this section we will apply the connections between the relative entropy, the generalised Fisher Information and the rate functional described earlier to the case of diffusions. In what comes next, we first define each of these objects for diffusions and then connect to the existing literature. Since our focus in this paper is on the discrete setting, we will keep the treatment in this section formal.

Consider a stochastic differential equation on $\mathbb{R}^d$,

$$dX_t = b(X_t)dt+\sqrt{2}\sigma(X_t)dB_t, \tag{27}$$

where $b:\mathbb{R}^d\to\mathbb{R}^d$, $\sigma:\mathbb{R}^d\to\mathbb{R}^{d\times d}$, B_t is a standard Brownian motion in $\mathbb{R}^d$ and $X_0\in\mathbb{R}^d$ is the initial data. The corresponding forward Kolmogorov equation (also called the Fokker–Planck equation in this case) evolves according to

$$\begin{cases}\partial_t\rho = L^T\rho := \mathrm{div}(b\rho)+\nabla^2 : A\rho\\ \rho_{t=0}=\rho_0,\end{cases} \tag{28}$$

where $A:=\sigma\sigma^T\in\mathbb{R}^{d\times d}$, $\rho_0\in\mathcal{P}(\mathbb{R}^d)$ is the initial data, ∇^2 is the Hessian and for two matrices $B,\tilde B\in\mathbb{R}^{d\times d}$, $B:\tilde B=\mathrm{tr}(B^T\tilde B)$. Here L^T is the adjoint corresponding to the generator

$$Lf(x) := -b(x)\cdot\nabla f(x)+A(x):\nabla^2 f(x). \tag{29}$$

Throughout this section we assume that the coefficients and the solution to (28) are sufficiently smooth (for a more general setup see [11]). For any probability measures $\mu, \rho \in \mathcal{P}(\mathbb{R}^d)$ and a Markov generator L, we define the relative Fisher Information as

$$\mathscr{R}_L(\mu|\rho) := \int_{\mathbb{R}^d} \left[-L \log\left(\frac{\mu}{\rho}\right)\mu + L\left(\frac{\mu}{\rho}\right)\rho \right] = \int_{\mathbb{R}^d} \left| \nabla \log\left(\frac{\mu}{\rho}\right) \right|_A^2 \mu,$$

where $|x|_A^2 := x^T Ax$. This is continuous version of the classical relative Fisher Information (5). Here we have inherently assumed that μ, ρ have sufficiently smooth densities (not renamed) such that this object is well defined. Note that, since we are working with 'linear' diffusion processes, the Fisher Information depends on the generator L only via the matrix A. As in the discrete case (recall (4)), when μ, ρ are solutions to (28), the relative Fisher Information satisfies the relation

$$\mathscr{R}_L(\mu_t|\rho_t) = -\frac{d}{dt}\mathscr{H}(\mu_t|\rho_t).$$

The corresponding large-deviation rate functional $\mathscr{I}_L : \mathcal{C}([0,T];\mathcal{P}(\mathcal{X})) \to \mathbb{R}$ is (see e.g. [8,32])

$$\mathscr{I}_L(\mu) = \sup_{f \in C^1([0,T];C_b^2(\mathbb{R}^d))} \int_{\mathbb{R}^d} f_T \, d\mu_T - \int_{\mathbb{R}^d} f_0 \, d\mu_0 - \int_0^T \left(\int_{\mathbb{R}^d} \partial_t f \, d\mu_t + \mathcal{H}(\mu_t, f_t) \right) dt, \tag{30}$$

with the Hamiltonian

$$\mathcal{H}(\mu, f) := \int_{\mathbb{R}^d} e^{-f} L e^f \, d\mu = \int_{\mathbb{R}^d} Lf + \Gamma(f,f) \, d\mu. \tag{31}$$

Here Γ is the carré-du-champ operator corresponding to the Markov generator L (see [2, Section 1.4.2])

$$\Gamma(f,g) := \frac{1}{2}\left[L(fg) - fLg - gLf\right] = \nabla f \cdot A\nabla g.$$

The (A-weighted) quadratic structure on the right hand side is particular to the diffusion processes.

For any $\lambda \in (0,1)$, and probability measures $\mu, \rho \in \mathcal{P}_+(\mathbb{R}^d)$, the continuous state-space counterpart of the generalised Fisher Information (15) is

$$\mathscr{R}_L^\lambda(\mu|\rho) := \int_{\mathbb{R}^d} \frac{\mu}{\rho} L^* \rho - \frac{1}{\lambda}\mathcal{H}\left(\mu, \lambda \log\left(\frac{\mu}{\rho}\right)\right) = (1-\lambda)\,\mathscr{R}_L(\mu|\rho),$$

where L^* denotes the $L^2(\mathbb{R}^d, \rho)$-adjoint of L. The equality here follows by using (31). Note that this is different from the discrete case where the generalised Fisher Information is bounded from above by the relative Fisher Information (recall Lemma 2.8) and the reversed inequality only holds in a fairly restrictive setting (see Lemma 2.10). This is due to the simpler structure of the Hamiltonian (31) which can be written as a combination of a linear and a quadratic term, as opposed to a genuine exponential structure in the discrete case.

Following the formal approach used for deriving the FIR inequality (cf. Section 1.2), we arrive at

$$\mathscr{H}(\mu_T|\rho_T) + (1-\lambda)\int_0^T \mathscr{R}_L(\mu_t|\rho_t)dt \le \mathscr{H}(\mu_0|\rho_0) + \frac{1}{\lambda}\mathscr{I}_L(\mu),$$

which has been derived recently in [11], and without the connection to large deviations in [4]. In [4] such an inequality is proven rigorously by directly studying the time derivative of the relative entropy and using appropriate regularity results for a very wide class of Fokker–Planck equations, while here we derive this inequality by studying the dual formulation of the rate functional. Similar ideas have also been developed for the (nonlinear) Vlasov–Fokker–Planck equation in [12, Theorem 2.3].

3. Coarse-graining

In this section we study the coarse-graining problem introduced in Section 1.3, which we now recall. Consider a family of forward Kolmogorov equations

$$\begin{cases} \partial_t \mu^\varepsilon = (L^\varepsilon)^T \mu^\varepsilon, \\ \mu^\varepsilon_{t=0} = \mu_0, \end{cases} \tag{32}$$

on $\mathcal{X} = \mathcal{Y} \times \mathcal{Z}$ with $\mathcal{Y} = \{0, 1\}$ and $\mathcal{Z} = \{1, \dots, n\}$, generated by the family of operators

$$L^\varepsilon = \frac{1}{\varepsilon} Q + C := \frac{1}{\varepsilon} \begin{pmatrix} Q_0 & 0 \\ 0 & Q_1 \end{pmatrix} + \begin{pmatrix} D_0 & C_{0,1} \\ C_{1,0} & D_1 \end{pmatrix}, \tag{33}$$

i.e. with

$$Q((y,z),(y',z')) = \begin{cases} Q_y(z,z') & \text{if } y' = y \\ 0 & \text{otherwise,} \end{cases}$$

$$C((y,z),(y',z')) = \begin{cases} C_{y,y'}(z,z') & \text{if } y' \neq y \\ D_y(z) & \text{if } y' = y \text{ and } z' = z \\ 0 & \text{otherwise} \end{cases}$$

for $x = (y,z)$, $x' = (y',z') \in \mathcal{X}$ satisfying

$$\forall x \in \mathcal{X} : \sum_{x' \in \mathcal{X}} Q(x,x') = 0 = \sum_{x \in \mathcal{X}} C(x,x'),$$

and diagonal matrix D_y, $y \in \mathcal{Y}$, which satisfies

$$\forall z \in \mathcal{Z} : \; D_y(z) := - \sum_{z' \in \mathcal{Z}} C_{y,1-y}(z,z'). \tag{34}$$

Here L^ε is irreducible, and therefore (32) admits a stationary solution $\pi^\varepsilon \in \mathcal{P}(\mathcal{X})$. Additionally we assume that Q_0 and Q_1 are irreducible as well. In what follows we will use $\nabla f(y,x) := f(y) - f(x)$.

Remark 3.1 (*Topologies on $\mathcal{P}(\mathcal{X})$*)**.** Since $\mathcal{X}$ is a finite set, $\mathcal{P}(\mathcal{X})$ can be identified with a closed, bounded (and thus compact) subset of the finite-dimensional vector space $\mathbb{R}^{\mathcal{X}}$. Therefore, there is no necessity to distinguish between different notions of convergence on $\mathcal{P}(\mathcal{X})$, since there is a unique topology which makes $\mathbb{R}^{\mathcal{X}}$ a (Hausdorff) topological vector space. In particular, the notion of uniform convergence (generated by the total variation distance) and narrow convergence (weak convergence with test functions in C_b) are equivalent and coincide with the standard convergence on $\mathbb{R}^{\mathcal{X}}$. □

The rest of this section is devoted to studying the behaviour of (32) in the limit of $\varepsilon \to 0$. We now outline an abstract variational framework, developed in [12], that will be used to study this problem.

3.1. A variational framework for coarse-graining

Let $\rho^\varepsilon : [0,T] \to \mathcal{P}(\mathcal{X})$ be a family of solutions to the forward Kolmogorov equations (32), and let $\mathscr{I}_{L^\varepsilon}$ be the corresponding family of large-deviation rate functionals associated to the underlying stochastic process (recall Theorem 1.4). Since the solutions ρ^ε are characterised by $\mathscr{I}_{L^\varepsilon}$ via $\mathscr{I}_{L^\varepsilon}(\rho^\varepsilon) = 0$, establishing the limit behaviour as $\varepsilon \to 0$ consists of answering two questions:

(1) *Compactness:* Do solutions of $\mathscr{I}_{L^\varepsilon}(\rho^\varepsilon) = 0$ have useful compactness properties, allowing one to extract a subsequence that converges in a suitable topology, say τ?
(2) *Liminf inequality:* Is there a limit functional $\mathscr{I} \geq 0$ such that

$$\rho^\varepsilon \xrightarrow{\tau} \rho \implies \liminf_{\varepsilon \searrow 0} \mathscr{I}_{L^\varepsilon}(\rho^\varepsilon) \geq \mathscr{I}(\rho)? \tag{35}$$

And if so, does one have

$$\mathscr{I}(\rho) = 0 \iff \partial_t \rho = L^T \rho,$$

for some limiting operator L?

As we shall see in the coming sections, the method we use answers both these questions for *approximate solutions*. By this we mean that we work with a sequence of time-dependent probability measures which satisfy $\sup_{\varepsilon>0} \mathscr{I}_{L^\varepsilon}(\mu^\varepsilon) < \infty$. The exact solutions are special cases when $\mathscr{I}_{L^\varepsilon}(\mu^\varepsilon) = 0$. Consequently, all our results follow from this uniform bound and assumptions on well-prepared initial data (which is exactly the right hand side of the FIR inequality (FIR$_\lambda$)).

The question of compactness will be answered by the uniform bound on the rate functional. Since our state space is finite, this bound along with the Arzelà–Ascoli theorem will provide us with suitable compactness properties (see Section 3.2 for details).

In answering the second question, we will make use of two crucial ingredients. First, that the rate functional has a duality relation of the type (recall Theorem 1.4),

$$\mathscr{I}_L(\mu) = \sup_f \mathcal{J}_L(\mu, f), \tag{36}$$

where the supremum is taken over an appropriate class of functions. Second, that the problem is of coarse-graining type as we expect that in the limit of $\varepsilon \to 0$, the dynamics in each macro-state equilibrates and the limiting object is a jump process across the macro-states (recall discussion in Section 1.3). We characterise this behaviour by means of a coarse-graining map which identifies the relevant degrees of freedom. In our setting we choose this to be a mapping onto the macro-states, i.e. $\xi : \mathcal{X} \to \mathcal{Y}$ with $\xi(x) = y$ for every $x = (y,z) \in \mathcal{X}$. The coarse-grained equivalent of $\rho^\varepsilon : [0,T] \to \mathcal{P}(\mathcal{X})$ is the push-forward $\hat{\rho}^\varepsilon := \xi_\# \rho^\varepsilon : [0,T] \to \mathcal{P}(\mathcal{Y})$. For a discussion on coarse-graining mappings in other contexts see [37, Section 1.4].

The core of the argument for the liminf inequality (35) is summarised in the following formal calculation:

$$\begin{aligned} \mathscr{I}_{L^\varepsilon}(\rho^\varepsilon) \quad &= \quad \sup_f \; \mathcal{J}_{L^\varepsilon}(\rho^\varepsilon, f) \\ &\overset{f = g\circ\xi}{\geq} \quad \sup_g \; \mathcal{J}_{L^\varepsilon}(\rho^\varepsilon, g \circ \xi) \\ &\qquad\qquad \Big\downarrow \; \varepsilon \to 0 \end{aligned} \tag{37}$$

$$\begin{aligned} &\sup_g \mathcal{J}(\rho, g\circ\xi) \\ \overset{(*)}{=:}\ &\sup_g \hat{\mathcal{J}}(\hat\rho, g) \overset{(**)}{=:} \hat{\mathscr{I}}(\hat\rho) \end{aligned}$$

Let us now go through each of these lines. The first line is the dual characterisation of the rate functional (36). The inequality on the second line follows by restricting the class of admissible functions f to functions of the type $f = g\circ\xi$. Here we have made a choice to restrict ourselves to functions of the form $f = g \circ \xi$. Following this inequality we pass to the limit using the compactness results derived earlier. The choice of coarse-graining map is crucial here since we cannot expect convergence for functions f which still have access to the full information.

In the next step $(*)$, we pass from the full limit measure ρ to the coarse-grained measure $\hat\rho$. To do that rigorously we need a *local-equilibrium* result, which describes how we can reconstruct the full information in ρ which is lost by considering only $\hat\rho$. As we shall see in Section 3.3, this result crucially depends on the generalised Fisher Information and the FIR inequality.

Finally, we define in $(**)$ a new functional $\hat{\mathscr{I}}$. In a successful application of coarse-graining, this functional is connected to an evolution equation similar to (8). In our example it turns out that $\hat{\mathscr{I}}$ is again a large deviations rate functional and connected to a lower dimensional effective equation.

In what follows we go through each of the steps described above to derive the behaviour of (32) as $\varepsilon \to 0$. In Section 3.2 we prove compactness results, Section 3.3 contains the local-equilibrium result and in Section 3.4 we prove the liminf inequality.

3.2. Compactness

In the following result we discuss the compactness properties. We prove a two-level compactness result, a weaker result on the original space $\mathcal{X}$ and a stronger result on the coarse-grained space $\mathcal{Y}$.

Lemma 3.2. *Let a sequence $\mu^\varepsilon \in C([0,T];\mathcal{P}(\mathcal{X}))$ satisfy*

$$\sup_{\varepsilon>0} \mathscr{I}_{L^\varepsilon}(\mu^\varepsilon) < \infty.$$

Then there exist $\mu \in \mathcal{M}([0,T]\times\mathcal{X})$ and a subsequence (not relabelled) such that

(i) $\mu^\varepsilon \to \mu$ in $\mathcal{M}([0,T]\times\mathcal{X})$ narrowly with $\mu = \int_0^T \mu_t$ for a Borel family $\{\mu_t\}_{t\in(0,T)}$.
(ii) $\xi_\#\mu^\varepsilon \to \xi_\#\mu$ in $C([0,T];\mathcal{P}(\mathcal{Y}))$ with respect to the uniform topology in time.

Proof. Since $[0,T]\times\mathcal{X}$ is compact, every subset of $\mathcal{M}([0,T]\times\mathcal{X})$ is tight. Furthermore, since $\mu^\varepsilon = \int_0^T \mu_t^\varepsilon$ with $\mu_t^\varepsilon \in \mathcal{P}(\mathcal{X})$ the set $\{\mu^\varepsilon, \varepsilon > 0\}$ is uniformly bounded in $\mathcal{M}([0,T]\times\mathcal{X})$, and so by Prokhorov's theorem and the equi-integrability of the map $t \mapsto \mu_t^\varepsilon(\mathcal{X})$, we have that $\mu^\varepsilon \to \mu$ narrowly in $\mathcal{M}([0,T]\times\mathcal{X})$ for some $\mu \in \mathcal{M}([0,T]\times\mathcal{X})$. Furthermore where μ has the representation $\mu = \int_0^T \mu_t$ for a Borel family $\{\mu_t\}_{t\in(0,T)}$ due to the disintegration theorem.

To prove the second statement we use the Arzelà–Ascoli theorem [31, Theorem 45.4]. Using the characterisation (13) of the rate functionals $\mathscr{I}_{L^\varepsilon}$, we obtain

$$M \geq \mathscr{I}_{L^\varepsilon}(\mu^\varepsilon) \geq \int_0^T \left[\left\langle \mathbb{1}_{[s_1,s_2]}(t)\frac{g\circ\xi}{\lambda}, \partial_t\mu_t^\varepsilon\right\rangle - \mathcal{H}^\varepsilon\left(\mu_t^\varepsilon, \mathbb{1}_{[s_1,s_2]}(t)\frac{g\circ\xi}{\lambda}\right)\right] dt, \tag{38}$$

for any $s_1, s_2 \in [0,T]$, $g \in \ell^\infty(\mathcal{Y})$ and $\lambda > 0$, where $\mathcal{H}^\varepsilon$ is the Hamiltonian corresponding to the generator L^ε (see (10)). We then calculate

$$\begin{aligned}\mathcal{H}^\varepsilon\left(\mu_t^\varepsilon, \mathbb{1}_{[s_1,s_2]}(t)\frac{g\circ\xi}{\lambda}\right) &= \sum_{x_1\in\mathcal{X}}\mu_t^\varepsilon(x_1)\sum_{z_2\in\mathcal{Z}}\varepsilon^{-1}Q_{y_1}(z_1,z_2)\left(e^{-\frac{1}{\lambda}\mathbb{1}_{[s_1,s_2]}(t)\nabla g(y_1,y_1)}-1\right)\\ &\quad+\sum_{x_1\in\mathcal{X}}\mu_t^\varepsilon(x_1)\sum_{z_2\in\mathcal{Z}}C_{y_1,1-y_1}(z_1,z_2)\\ &\quad\times\left(e^{-\frac{1}{\lambda}\mathbb{1}_{[s_1,s_2]}(t)\nabla g(y_1,1-y_1)}-1\right)\\ &\le 0+\bar{C}\left(e^{\frac{1}{\lambda}2\|g\|_\infty}-1\right)\mathbb{1}_{[s_1,s_2]}(t),\end{aligned}$$

where $\bar{C} := \sup_{y\in\mathcal{Y}}\|C_{y,1-y}\|$ is independent of $\varepsilon > 0$ and $s \in [0,T]$ and the zero in the final inequality follows since $\nabla g(y_1,y_1) = 0$. Note that D_0 and D_1 do not contribute to the equality above. Substituting this bound into (38) with $\lambda = -\|g\|_\infty/\log\sqrt{|s_2-s_1|}$ and using absolute continuity on $t \mapsto \mu_t^\varepsilon$ we find

$$\begin{aligned}\langle g, \xi_\#\mu_{s_2}^\varepsilon - \xi_\#\mu_{s_1}^\varepsilon\rangle &= \int_{s_1}^{s_2}\langle g\circ\xi, \partial_t\mu_t^\varepsilon\rangle\,dt\\ &\le \lambda M + \lambda\bar{C}|s_2-s_1|\left(e^{\frac{1}{\lambda}2\|g\|_\infty}-1\right)\\ &= \frac{\|g\|_\infty M}{-\log\sqrt{|s_2-s_1|}} + \frac{\|g\|_\infty\bar{C}|s_2-s_1|}{-\log\sqrt{|s_2-s_1|}}\left(\frac{1}{|s_2-s_1|}-1\right)\\ &\le 2\|g\|_\infty\frac{M+\bar{C}|1-|s_2-s_1||}{|\log|s_2-s_1||}.\end{aligned}$$

Since the narrow topology coincides with the uniform topology and the upper bound does not depend on ε this gives equicontinuity of $(\xi_\#\mu^\varepsilon)$. Furthermore, $\xi_\#\mu^e$ is naturally bounded from above in $\mathcal{C}([0,T];\mathcal{P}(\mathcal{Y}))$ and thus, we can apply the Arzelà–Ascoli theorem which gives the statement. □

3.3. Local-equilibrium

As stated earlier, our interest is in studying the slow behaviour of the dynamics and we do this by focussing on a coarse-grained description of the model (via ξ). However information is lost in the coarse-graining procedure, and in this section we reconstruct this lost information by proving a 'local-equilibrium' result, which crucially depends on the FIR inequality.

The central idea is to pass $\varepsilon \to 0$ in the FIR inequality, obtain a vanishing bound on the generalised Fisher Information and then study the properties of the limiting object. More precisely, we combine the lower-semicontinuity property of $\mathscr{R}_L^\lambda$ with the FIR inequality (FIR$_\lambda$) to show that in the limit of $\varepsilon \to 0$, the time-dependent sequence μ^ε becomes stationary in the micro-state variable and the time dependence completely shifts onto the macro-state variable. We first prove an auxiliary lemma which discusses the limit of the stationary measure π^ε and then prove the local-equilibrium result.

Lemma 3.3. *Let $(\pi^\varepsilon)_{\varepsilon>0} \subset \mathcal{P}(\mathcal{X})$ be a sequence of stationary measures corresponding to L^ε, i.e. $(L^\varepsilon)^T\pi^\varepsilon = 0$ for every $\varepsilon > 0$. Then there exists a positive probability measure $\pi \in \mathcal{P}_+(\mathcal{X})$ satisfying $Q^T\pi = 0$, with $\pi^\varepsilon \to \pi$ in $\mathcal{P}_+(\mathcal{X})$.*

Proof. Due to the compactness of $\mathcal{P}(\mathcal{X})$, we find some $\pi \in \mathcal{P}(\mathcal{X})$ such that $\pi^\varepsilon \to \pi$ as $\varepsilon \to 0$. Passing $\varepsilon \to 0$ in $\varepsilon(L^\varepsilon)^T\pi^\varepsilon = 0$ yields

$$Q^T\pi = 0 \iff \exists\, \alpha \in [0,1] \text{ such that } \pi = \begin{pmatrix} \alpha\pi_0 \\ (1-\alpha)\pi_1 \end{pmatrix}, \tag{39}$$

where $\pi_y \in \mathcal{P}(\mathcal{Z})$ is the stationary measure of Q_y, $y \in \mathcal{Y}$.

We now show that $\pi \in \mathcal{P}_+(\mathcal{X})$, which follows if $\alpha \in (0,1)$ since $\pi_y \in \mathcal{P}_+(\mathcal{Z})$ due to the irreducibility of Q_y. Using $(L^\varepsilon)^T\pi^\varepsilon = 0$ and $\sum_{z' \in \mathcal{Z}} Q_y(z,z') = 0$, for every $y \in \mathcal{Y}$ we find

$$\begin{aligned} 0 = \sum_{z\in\mathcal{Z}} ((L^\varepsilon)^T\pi^\varepsilon)(y,z) &= \sum_{z,z'\in\mathcal{Z}} \left[\frac{1}{\varepsilon} Q_y(z',z)\pi^\varepsilon(y,z') + C_{1-y,y}(z',z)\pi^\varepsilon(1-y,z')\right] \\ &\quad + \sum_{z\in\mathcal{Z}} D_y(z)\pi^\varepsilon(y,z) \\ &= \sum_{z,z'\in\mathcal{Z}} C_{1-y,y}(z',z)\pi^\varepsilon(1-y,z') + \sum_{z\in\mathcal{Z}} D_y(z)\pi^\varepsilon(y,z), \end{aligned}$$

Furthermore passing $\varepsilon \to 0$ and using (34) we obtain

$$0 = -\sum_{z\in\mathcal{Z}} D_{1-y}(z)\pi(1-y,z) + \sum_{z\in\mathcal{Z}} D_y(z)\pi(y,z).$$

Finally, using (39) and $\lambda_y := -\sum_{z\in\mathcal{Z}} D_y(z)\pi_y(z)$ we have

$$-\alpha\lambda_0 + (1-\alpha)\lambda_1 = 0 \implies \alpha = \frac{\lambda_1}{\lambda_0 + \lambda_1}.$$

Since $\lambda_y > 0$ (recall that L^ε is irreducible if and only if $C_{y,1-y}$ has at least one positive entry for all $y \in \{0,1\}$) we have $\alpha \in (0,1)$ and therefore $\pi \in \mathcal{P}_+(\mathcal{X})$. □

Lemma 3.4. *Let a sequence $\mu^\varepsilon \in \mathcal{C}([0,T];\mathcal{P}(\mathcal{X}))$ satisfy*

$$\sup_{\varepsilon>0}\left\{\mathscr{I}_{L^\varepsilon}(\mu^\varepsilon) + \mathscr{H}(\mu_0^\varepsilon|\pi^\varepsilon)\right\} < \infty, \tag{40}$$

where $(\pi^\varepsilon)_{\varepsilon>0} \subset \mathcal{P}(\mathcal{X})$ is a sequence of stationary measures of L^ε converging to $\pi \in \mathcal{P}_+(\mathcal{X})$ as $\varepsilon \to 0$. Then there is a $\hat\mu \in \mathcal{C}([0,T];\mathcal{P}(\mathcal{Y}))$ such that for almost every $t \in [0,T]$,

$$\forall y \in \mathcal{Y},\ A_\mathcal{Z} \subset \mathcal{Z},\ \mu_t(\{y\}\times A_\mathcal{Z}) = \hat\mu_t(y)\pi_y(A_\mathcal{Z}). \tag{41}$$

Here μ is the limit of $(\mu^\varepsilon)_{\varepsilon>0}$ (see Lemma 3.2) and for each $y \in \mathcal{Y}$, $\pi_y \in \mathcal{P}(\mathcal{Z})$ is the stationary measure corresponding to Q_y. Furthermore $\xi_\#\mu^\varepsilon \to \hat\mu$ in $\mathcal{C}([0,T];\mathcal{P}(\mathcal{Y}))$ uniformly in time.

Proof. Using (40) and the FIR inequality in Theorem 1.6, we find

$$\begin{aligned} &\mathscr{H}(\mu_T^\varepsilon|\pi^\varepsilon) + \int_0^T \mathscr{R}^\lambda_{L^\varepsilon}(\mu_t^\varepsilon|\pi^\varepsilon)\,dt \le \frac{1}{\lambda}\mathscr{I}_{L^\varepsilon}(\mu^\varepsilon) + \mathscr{H}(\mu_0^\varepsilon|\pi^\varepsilon) \le M \\ &\quad\implies \int_0^T \mathscr{R}^\lambda_{L^\varepsilon}(\mu_t^\varepsilon|\pi^\varepsilon)\,dt \le M, \end{aligned}$$

for some constant $M < \infty$ independent of ε. Recall that $L^\varepsilon = \varepsilon^{-1}Q + C$. Due to the linearity of $\mathscr{R}^\lambda_L$ with respect to L, we find that

$$\varepsilon^{-1}\int_0^T \mathscr{R}^\lambda_Q(\mu_t^\varepsilon|\pi^\varepsilon)\,dt + \int_0^T \mathscr{R}^\lambda_C(\mu_t^\varepsilon|\pi^\varepsilon)\,dt \le M.$$

Multiplying with ε and letting $\varepsilon \to 0$ we find

$$\liminf_{\varepsilon\to 0} \int_0^T \mathscr{R}^\lambda_Q(\mu^\varepsilon_t|\pi^\varepsilon)\, dt \le 0.$$

Using the non-negativity and lower-semicontinuity property of the generalised relative Fisher Information (cf. Lemma 2.4), together with the Borel-measurability of the non-negative functions $t \mapsto \mathscr{R}_Q(\mu^\varepsilon_t|\pi^\varepsilon)$, we obtain from Fatou's lemma that

$$\mathscr{R}^\lambda_Q(\mu_t|\pi) = 0 \qquad \text{for almost every } t \in (0,T). \tag{42}$$

In what follows, for $y \in \mathcal{Y}$ we use $\mu_t(\cdot|y) \in \mathcal{P}(\mathcal{Z})$ for the family of conditional measures corresponding to μ_t, i.e. we write $\mu_t(y,z) = \mu_t(z|y)(\xi_\#\mu_t)(y)$. We show that $\mathscr{R}^\lambda_Q(\mu_t|\pi) = 0$ if and only if $\mu_t(z|y) = \pi_y(z)$ for any $x = (y,z) \in \mathcal{X}$ with $(\xi_\#\mu_t)(y) > 0$. Using the representation (15b) and by disintegration we find

$$\begin{aligned}
\mathscr{R}^\lambda_Q(\mu_t|\pi) &= \sum_{x,x'\in\mathcal{X}} Q(x,x') \left[\mu_t(x')\frac{\pi(x)}{\pi(x')} - \frac{1}{\lambda}\mu_t(x)^{1-\lambda}\mu_t(x')^\lambda \left(\frac{\pi(x)}{\pi(x')}\right)^\lambda \right] \\
&= \sum_{y\in\mathcal{Y}} \sum_{z,z'\in\mathcal{Z}} (\xi_\#\mu_t)(y) Q_y(z,z') \\
&\quad \times \left[\mu_t(z'|y)\frac{\pi_y(z)}{\pi_y(z')} - \frac{1}{\lambda}\mu_t(z|y)^{1-\lambda}\mu_t(z'|y)^\lambda \left(\frac{\pi_y(z)}{\pi_y(z')}\right)^\lambda \right] \\
&= \sum_{y\in\mathcal{Y}} (\xi_\#\mu_t)(y) \mathscr{R}^\lambda_{Q_y}(\mu_t(\cdot|y)|\pi_y).
\end{aligned}$$

Here, we used that the conditional measure $\pi(\cdot|y) \in \mathcal{P}(\mathcal{Z})$ is the stationary measure π_y of Q_y since $(\xi_\#\pi)(y) > 0$. Using (42) along with the irreducibility of Q_y, the fact that $\pi_y \in \mathcal{P}_+(\mathcal{Z})$ and Lemma 2.5 we find $\mu_t(z|y) = \pi_y(z)$ for any $(y,z) \in \mathcal{X}$ with $(\xi_\#\mu_t)(y) > 0$, and therefore (41) follows since it holds trivially whenever $(\xi_\#\mu_t)(y) = 0$. By the convergence properties of $\xi_\#\mu^\varepsilon$ given in Lemma 3.2, we find $\hat\mu := \xi_\#\mu \in \mathcal{C}([0,T];\mathcal{P}(\mathcal{Y}))$ such that $\xi_\#\mu^\varepsilon \to \hat\mu$ uniformly in time. □

3.4. Liminf inequality

As discussed in Section 3.1, the final step is to prove a liminf inequality which will also provide us with the limit dynamics. We prove this result in the next theorem.

We define the (limiting) functional $\mathscr{I}_L : \mathcal{C}([0,T];\mathcal{P}(\mathcal{Y})) \to \mathbb{R}$ by

$$\begin{aligned}
\mathscr{I}_L(\hat\mu) := \sup_{g\in\mathcal{C}^1([0,T];\ell^\infty(\mathcal{Y}))} \Bigg\{ &\sum_{y\in\mathcal{Y}} g_T(y)\hat\mu_T(y) - \sum_{y\in\mathcal{Y}} g_0(y)\hat\mu_0(y) \\
&- \int_0^T \Bigg[\sum_{y\in\mathcal{Y}} \partial_t g_t(y)\hat\mu_t(y) + \sum_{y,y'\in\mathcal{Y}} \hat\mu_t(y)L(y,y') \\
&\times \left(e^{\nabla g_t(y',y)} - 1\right) \Bigg] dt \Bigg\},
\end{aligned} \tag{43}$$

with the (limiting) generator L defined as

$$L := \begin{pmatrix} -\lambda_0 & \lambda_0 \\ \lambda_1 & -\lambda_1 \end{pmatrix}, \quad \lambda_y := \sum_{z,z' \in \mathcal{Z}} \pi_y(z) C_{y,1-y}(z,z'). \tag{44}$$

Here $\pi_y \in \mathcal{P}_+(\mathcal{Z})$ is the stationary measure of Q_y (recall Lemma 3.4). Since $g = 0$ is admissible, $\mathscr{I}_L \geq 0$. Furthermore we have the equivalence

$$\mathscr{I}_L(\hat{\mu}) = 0 \iff \partial_t \hat{\mu} = L^T \hat{\mu}. \tag{45}$$

Lemma 3.5. *Under the same assumptions of Lemma 3.4 we assume that $\mu^\varepsilon \to \mu$ narrowly in $\mathcal{M}([0,T] \times \mathcal{X})$ and $\xi_\# \mu^\varepsilon \to \hat{\mu}$ in $\mathcal{C}([0,T]; \mathcal{P}(\mathcal{Y}))$ (recall Lemma 3.2). Then*

$$\liminf_{\varepsilon \to 0} \mathscr{I}_{L^\varepsilon}(\mu^\varepsilon) \geq \mathscr{I}_L(\hat{\mu}).$$

Proof. We write the rate functional $\mathscr{I}_{L^\varepsilon} : \mathcal{C}([0,T]; \mathcal{P}(\mathcal{X})) \to \mathbb{R}$ (defined in (12)) as

$$\mathscr{I}_{L^\varepsilon}(\mu^\varepsilon) = \sup_{f \in \mathcal{C}^1([0,T]; \ell^\infty(\mathcal{X}))} \mathcal{J}^\varepsilon(\mu^\varepsilon, f),$$

with

$$\mathcal{J}^\varepsilon(\mu^\varepsilon, f) := \langle f_t, \mu^\varepsilon_T \rangle - \langle f_0, \mu^\varepsilon_0 \rangle - \int_0^T \sum_{x,x' \in \mathcal{X}} \mu_t(x) \\ \times \left(\partial_t f_t(x) + L^\varepsilon(x,x') \left[e^{\nabla f(x',x)} - 1 \right] \right) dt.$$

Using $\mathcal{A} := \{ f = g \circ \xi \, : \, g \in \mathcal{C}^1([0,T]; \ell^\infty(\mathcal{Y})) \}$ we have

$$\mathscr{I}_{L^\varepsilon}(\mu^\varepsilon) \geq \sup_{f \in \mathcal{A}} \mathcal{J}^\varepsilon(\mu^\varepsilon, f),$$

where

$$\mathcal{J}^\varepsilon(\mu^\varepsilon, g \circ \xi) = \langle g_T \circ \xi, \mu^\varepsilon_T \rangle - \langle g_0 \circ \xi, \mu^\varepsilon_0 \rangle - \int_0^T \langle \partial_t (g_t \circ \xi), \mu^\varepsilon_t \rangle \, dt \\ - \int_0^T \sum_{(y,z) \in \mathcal{Y} \times \mathcal{Z}} \mu^\varepsilon_t((y,z)) \sum_{z' \in \mathcal{Z}} C_{y,1-y}(z,z') \left(e^{-\nabla g_t(y,1-y)} - 1 \right). \tag{46}$$

We now show that (46) converges to (43) term by term. Since $\xi_\# \mu^\varepsilon_t \to \hat{\mu}_t$ uniformly in $t \in [0,T]$, for the first three terms in the right hand side of (46) we find

$$\langle g_T, \xi_\# \mu^\varepsilon_T \rangle - \langle g_0, \xi_\# \mu^\varepsilon_0 \rangle - \int_0^T \langle \partial_t g_t, \xi_\# \mu^\varepsilon_t \rangle \, dt \\ \xrightarrow{\varepsilon \to 0} \quad \langle g_T, \xi_\# \mu_T \rangle - \langle g_0, \xi_\# \mu_0 \rangle - \int_0^T \langle \partial_t g_t, \xi_\# \mu_t \rangle \, dt.$$

Using Lemma 3.4 for the final term in (46) yields

$$\begin{aligned}
&\int_0^T \sum_{(y,z)\in\mathcal{Y}\times\mathcal{Z}} \mu_t^\varepsilon((y,z))\left(e^{-\nabla g_t(y,1-y)}-1\right)\sum_{z'\in\mathcal{Z}} C_{y,1-y}(z,z')\,dt \\
&\qquad \xrightarrow{\varepsilon\to 0} \int_0^T \sum_{(y,z)\in\mathcal{Y}\times\mathcal{Z}} \pi_y(z)\hat{\mu}_t(y)\left(e^{-\nabla g(y,1-y)}-1\right)\sum_{z'\in\mathcal{Z}} C_{y,1-y}(z,z')\,dt \\
&\qquad = \int_0^T \sum_{y\in\mathcal{Y}} \hat{\mu}_t(y)\left(e^{-\nabla g(y,1-y)}-1\right)\lambda_y\,dt.
\end{aligned}$$

where λ_y is defined in (44). Altogether, we obtain

$$\begin{aligned}
\liminf_{\varepsilon\to 0} \mathscr{I}_{L^\varepsilon}(\mu^\varepsilon) &\geq \liminf_{\varepsilon\to 0} \mathcal{J}^\varepsilon(\mu^\varepsilon, g\circ\xi) \\
&= \langle g_T, \hat{\mu}_T\rangle - \langle g_0, \hat{\mu}_0\rangle - \int_0^T \langle \partial_t g_t, \hat{\mu}_t\rangle \\
&\quad + \sum_{y\in\mathcal{Y}} \hat{\mu}_t(y)\left(e^{-\nabla g(y,1-y)}-1\right)\lambda_y\,dt
\end{aligned}$$

for every $g \in C^1([0,T];\ell^\infty(\mathcal{Y}))$. Taking the supremum over such functions concludes the proof. □

Remark 3.6 (*Limiting Behaviour of Solutions*)**.** So far, in all the steps we have assumed that the sequence μ^ε are *approximate solutions* in the sense that they satisfy $\sup_{\varepsilon>0}\mathscr{I}_{L^\varepsilon}(\mu^\varepsilon) < \infty$. The case when μ^ε is a sequence of solutions to the forward Kolmogorov equation (32) is a special case of our analysis, which corresponds to the choice $\mathscr{I}_{L^\varepsilon}(\mu^\varepsilon) = 0$. Lemma 3.5 implies that the limiting evolution for a sequence of solutions is given by (45). Theorem 1.8 summarises the results for a sequence of solutions. □

4. Conclusion and discussion

In this paper we construct a *generalised relative Fisher Information* in the context of Markov jump processes on possibly countable discrete state space. This generalised Fisher Information has various favourable properties, and connects naturally to the relative entropy and the large deviation rate functional. We then use these connections to solve a coarse-graining problem in the context of Markov jump processes.

We now discuss some open questions and connected problems.

Coarse-graining in more general setting. As mentioned in the introduction, our coarse-graining example was already discussed using martingale techniques in [22]. Related ideas have also been discussed in [33, Chapter 16]. We now discuss whether more general settings can also be treated by our method. For that we distinguish two cases, finite state spaces and countable state spaces. In the case of finite state spaces, we expect that our proofs straightforwardly generalise to the case there are more than two macro-states which each have a different (finite) number of macro-states, i.e. $\mathcal{Y}$ is an arbitrary finite set and $\mathcal{X} = \cup_{y\in\mathcal{Y}}\{y\}\times\mathcal{Z}_y$.

In contrast the case of infinite state spaces provides more difficulties. A particular one is that the compactness argument in Lemma 3.2 via Prokhorov's theorem relies on the fact that the state space is finite and thus compact. In [12] this is solved by using the FIR inequality to obtain bounds on the free energy which are in turn used to obtain compactness results. However, it is an open question, whether such a strategy is applicable in the discrete case.

Other stochastic processes. The approach to the FIR inequality presented in this work is rather general, which we now formally outline. Let X be a smooth manifold with tangent bundle TX and $\mathscr{L} : X \times TX \to \mathbb{R}$ a Lagrangian, or more generally an L-function [30], i.e. $\mathscr{L}$ is nonnegative, convex in its second argument and induces an evolution equation via

$$\mathscr{L}(x, s) = 0 \Longleftrightarrow s = \mathcal{A}(x).$$

Note that we do not assume that $\mathscr{L}$ originates from a large deviations principle. Furthermore, suppose that there is a smooth Lyapunov function $\mathscr{F} : X \to \mathbb{R}$ connected to the evolution equation $\partial_t x = \mathcal{A}(x)$.

We now construct a relative entropy-type functional comparing two elements from X by using the Bregman divergence of $\mathscr{F}$,

$$\mathscr{F}(x|y) := \mathscr{F}(x) - \mathscr{F}(y) - \langle d\mathscr{F}(y), x - y\rangle ,$$

where $d\mathscr{F}$ is the Fréchet derivative of $\mathscr{F}$. Then, we can formally define the generalised relative Fisher Information in this case as

$$\mathscr{R}^\lambda_{\mathcal{A}}(x|y) := \left\langle d^2\mathscr{F}(y)(\mathcal{A}(y)), x - y\right\rangle - \frac{1}{\lambda}\mathcal{H}(x, \lambda(d\mathscr{F}(x) - d\mathscr{F}(y))),$$

where $\mathcal{H}(x, \cdot)$ is the Legendre transform of $\mathcal{L}(x, \cdot)$ for fixed $x \in X$. By construction, these functionals satisfy the FIR-type inequality

$$\mathscr{F}(x_T|y_T) - \mathscr{F}(x_0|y_0) + \int_0^T \mathscr{R}^\lambda_{\mathcal{A}}(x_t|y_t)\, dt \le \frac{1}{\lambda}\int_0^T \mathcal{L}(x_t, \partial_t x_t)\, dt,$$

with $y : [0, T] \to X$ satisfying $\partial_t y = \mathcal{A}(y)$. We still expect that $\mathscr{R}^\lambda_{\mathcal{A}}$ converges for $\lambda \to 0$ to the classical relative Fisher Information $\mathscr{R}_{\mathcal{A}}$, similar to the motivation of Lemma 2.8. However, whether $\mathscr{R}^\lambda_{\mathcal{A}}$ is also a non-negative functional is an open question. We suspect that the Lagrangian and the Lyapunov function have to be connected in some appropriate sense for this to hold. One example of such a connection would be when both originate from a large deviations principle.

This is also related to the important question, ‘How to construct Lyapunov functions?’. There are, in principle, multiple approaches to do this. For example, a specific choice can be motivated via a gradient flow result or via a large deviations principle. In the case discussed in this work, both methods are valid. While the fact that the relative entropy can be obtained via a large deviations principle is well known, gradient flow results for discrete state spaces are relatively new, see e.g. [24]. Further results for both these approaches also exist for certain nonlinear systems, see e.g. [14,18]. However it is not clear if and how these are connected and whether they can be used in the construction of a generalised relative Fisher Information as described above.

Quantification of coarse-graining error. The FIR inequality has been successfully used to quantify error in relative entropy between two different forward Kolmogorov equations in the context of diffusion equations. Similar questions can be asked in the Markov jump process context, for instance to prove rates of convergence — note that in this paper we only prove qualitative convergence. However the role of the generalised Fisher Information and the FIR inequality in proving such quantitative estimates is an open problem. To do this, we expect that the right object to consider is not the FIR inequality but a related result inspired by [41] (see [16, Chapter 8] for preliminary results).

Declaration of competing interest

No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.spa.2019.07.012.

Acknowledgements

BH was supported by the German Research Foundation (DFG) within the Cluster of Excellence in Simulation Technology (EXC 310/2) at the University of Stuttgart. MAP and US kindly acknowledge support from the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) VICI grant 639.033.008. OT was funded by the NWO Vidi grant number 016.Vidi.189.102.

Appendix A. Banach-space-valued functions

In this appendix we briefly summarise some properties of functions from an interval $[0, T]$ into the Banach space $\ell^1(\mathcal{X})$; we follow the treatment in [17] and use their terminology. While in this paper the set $\mathcal{X}$ is assumed to be either finite or countable, in this appendix we assume that $\mathcal{X}$ is countable, and to simplify notation we assume that $\mathcal{X} = \mathbb{N}$; the results for the finite case are all classical.

First we define the space $\mathcal{AC}([0, T]; \mathcal{P}(\mathcal{X}))$ of absolutely continuous trajectories in the space of probability measures. This is the space of curves $\mu : [0, T] \to \mathcal{P}(\mathcal{X})$ that satisfy

> For every $\varepsilon > 0$, there exists $\delta > 0$ such that for any finite set of disjoint intervals $([a_k, b_k])_{k\in I} \subset [0, T]$ with $\sum_{k\in I} |b_k - a_k| < \delta$ we have $\sum_{k\in I} \|\mu(b_k) - \mu(a_k)\|_{\ell^1(\mathcal{X})} < \varepsilon$.

Note that the metric used in the definition above is the ℓ^1-norm, which is consistent because strong and weak continuity coincide.

Next we turn to Bochner spaces. We refer to [17] for the concepts of measurability and Bochner integrability of a function $u : [0, T] \to \ell^1(\mathbb{N})$. The Bochner space $L^1(0, T; \ell^1(\mathbb{N}))$ is defined as the space of equivalence classes of strongly Lebesgue-measurable functions with finite norm

$$\|u\|_{L^1(0,T;\ell^1(\mathbb{N}))} := \int_0^T \|u(t)\|_{\ell^1(\mathbb{N})}\, dt.$$

The space $W^{1,1}(0, T; \ell^1(\mathbb{N}))$ is defined as the subset of $L^1(0, T; \ell^1(\mathbb{N}))$ of functions with weak derivatives in $L^1(0, T; \ell^1(\mathbb{N}))$.

Lemma A.1. *Let $u : [0, T] \to \ell^1(\mathbb{N})$; then $u \in \mathcal{AC}([0, T]; \ell^1(\mathbb{N}))$ iff $u \in W^{1,1}(0, T; \ell^1(\mathbb{N}))$. In this case the derivative $\partial_t u(t)$ exists in the classical sense at almost all t, it is a.e. equal to the weak derivative of u, and we have*

$$u(\tau) - u(\sigma) = \int_\sigma^\tau \partial_t u(t)\, dt, \qquad \textit{for all } 0 \le \sigma \le \tau \le T,$$

where the integral is in the sense of Bochner.

Proof. The space $\ell^1(\mathbb{N})$ is separable and is the dual of the space

$$c_0(\mathbb{N}) = \Big\{ (u_n)_{n\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}} : \lim_{n\to\infty} u_n = 0 \Big\},$$

equipped with the supremum norm. This implies that $\ell^1(\mathbb{N})$ has the *Radon–Nikodym property* [17, Th. 1.3.21]. The assertion then follows from [17, Th. 2.5.12 and Prop. 2.5.9]. □

For the proof of Theorem 1.6 we need a generalisation of the chain rule to absolutely continuous functions with values in $\ell^1(\mathbb{N})$. When $u \in AC([0,T];\mathbb{R})$ and $f \in C^1(\mathbb{R})$, the chain rule is standard and can be found e.g. in [6, Cor. 8.11]; the extension to functions f that are only Lipschitz was first proved by De La Vallée Poussin [38, p. 467] (see also [36, Remark A.3] and a more in-depth treatment in [25]). The following lemma generalises this extension to compositions of the form $f(u(t), v(t))$ under special conditions on f:

Lemma A.2. *Let $A \subset E \subset \mathbb{R}^2$, and let $f : E \to \mathbb{R}$ be globally Lipschitz continuous and differentiable at each point of A. Let $u, v \in AC([0,T];\mathbb{R})$ satisfy $(u(t), v(t)) \in A$ for all t. Define $w(t) := f(u(t), v(t))$. Then w is absolutely continuous, and the chain rule holds in the following sense. There exists a null set $N \subset [0,T]$ such that w, u, and v are differentiable at each $t \in [0,T] \setminus N$, and such that*

$$w'(t) = \partial_1 f(u(t), v(t))u'(t) + \partial_2 f(u(t), v(t))v'(t) \qquad \textit{for all } t \in [0,T] \setminus N. \tag{47}$$

Proof. First note that by the Lipschitz continuity of f, w is absolutely continuous. To prove the chain rule (47), we restrict ourselves to the set of t for which u, v, and w each are differentiable; the remainder N of $[0,T]$ is a null set. Consider such a $t \in [0,T] \setminus N$; since $(u(t), v(t)) \in A$, f is differentiable at $(u(t), v(t))$, and therefore (47) follows from the classical chain rule. □

We then use the previous lemma to prove the chain rule for two nonnegative ℓ^1-valued functions.

Lemma A.3. *As in Lemma A.2, let $f : A \to \mathbb{R}$ be globally Lipschitz continuous and differentiable at each point of A. Let $u, v \in AC([0,T]; \ell^1(\mathbb{N}))$ satisfy $(u(t,x), v(t,x)) \in A$ for all t and x. Define the function*

$$w(t,x) = f(u(t,x), v(t,x)) \qquad \textit{for each } x \in \mathbb{N} \textit{ and } t \in [0,T].$$

Then $w \in AC([0,T]; \ell^1(\mathbb{N}))$ and

$$\partial_t w(t,x) = \partial_1 f(u(t,x), v(t,x))\partial_t u(t,x) + \partial_2 f(u(t,x), v(t,x))\partial_t v(t,x) \\ \textit{for a.e. } t \in [0,T] \textit{ and all } x \in \mathbb{N}. \tag{48}$$

Note that pointwise evaluation is a continuous operation on $\ell^1(\mathbb{N})$, and therefore commutes with time differentiation; this shows that there is no ambiguity in the notation $\partial_t w(t,x)$, since $[w'(t)](x) = d/dt\,[w(t,x)]$ for almost all t and all x.

Proof. The absolute continuity of w follows directly from the Lipschitz continuity of f. To prove the chain rule (48), fix $x \in \mathbb{N}$ and observe that $t \mapsto u(t,x)$ and $t \mapsto v(t,x)$ are elements of $AC([0,T]; [0,\infty))$; therefore

$$\begin{aligned}\partial_t[w(t,x)] &= \partial_t[f(u(t,x), v(t,x))] \\ &\overset{\text{Lemma A.2}}{=} \partial_1 f(u(t,x), v(t,x))\partial_t u(t,x) + \partial_2 f(u(t,x), v(t,x))\partial_t v(t,x),\end{aligned}$$

for all x and all $t \in [0,T] \setminus N_x$ for some null set N_x. Defining the null set $N := \cup_{x \in \mathbb{N}} N_x$ we find that this expression holds for all x and all $t \in [0,T] \setminus N$, which proves the lemma. □

Appendix B. Proof of Theorem 1.4

The large deviation result and the corresponding rate functional (see (7)) for Markov chains on a finite or countable state space have been discussed in [19, Proposition 5.10]. The main objective of Theorem 1.4 is to give a different characterisation of the rate functional which is more useful in the context of coarse-graining (discussed in Section 3.1). The proof is inspired by techniques developed in [8, Section 4], where the authors study large deviation principles in the context of weakly-interacting diffusions.

We define

$$\tilde{\mathcal{J}}_{s,t}(\mu, f) := \sum_{x\in\mathcal{X}} f_t(x)\mu_t(x) - \sum_{x\in\mathcal{X}} f_s(x)\mu_s(x) - \int_s^t \sum_{x\in\mathcal{X}} \partial_u f_u(x)\mu_u(x) + \mathcal{H}(\mu_u, f_u)\,du. \tag{49}$$

Corollary B.1. *Let $\mu \in \mathcal{C}([0,T];\mathcal{P}(\mathcal{X}))$, $\mathcal{I} \subset \mathbb{N}$ a finite index set and $[s_k, t_k] \subset [0,T]$, $k \in \mathcal{I}$ be a finite family of pairwise disjoint intervals. Then for any function $g = \sum_{k\in\mathcal{I}} \varphi_k \chi_{[s_k,t_k]} \in L^\infty(0,T;\ell^\infty(\mathcal{X}))$, with $\varphi_k \in \ell^\infty(\mathcal{X})$ and indicator function χ_I (on interval I), there exists a monotonically decreasing sequence $g^n \in \mathcal{C}^1([0,T];\ell^\infty(\mathcal{X}))$ such that $\|g^n - g\|_{\ell^\infty(\mathcal{X})} \to 0$ pointwise almost everywhere in $(0,T)$ as $n \to \infty$ and*

$$\tilde{\mathcal{J}}_{0,T}(\mu, g^n) \quad \xrightarrow{n\to\infty} \quad \sum_{k\in\mathcal{I}} \tilde{\mathcal{J}}_{s_k,t_k}(\mu, g),$$

where $\tilde{\mathcal{J}}_{s,t}$ is defined by (49).

Proof. For every $k \in \mathcal{I}$ there exists a decreasing sequence $(h_{k,n})_{n\in\mathbb{N}} \subset \mathcal{C}^1([0,T];\mathbb{R})$ such that $h_{k,n}(t) \in [0,1]$ for every $t \in [0,T]$ and $h_{n,k} \to \chi_{[s_k,t_k]}$ pointwise almost everywhere for $n \to \infty$. Furthermore, since there are only finitely many k we can choose the $h_{n,k}$ such that they have pairwise disjoint support for n large enough. Finally, we assume that there exists a $C < \infty$ not depending on n such that

$$\sum_{k\in\mathcal{I}} \int_0^T |\partial_t h_{k,n}|\,dt \le C.$$

We define $g_t^n(x) := \sum_{k\in\mathcal{I}} \varphi_k(x) h_{k,n}(t) \in \mathcal{C}^1([0,T];\ell^\infty(\mathcal{X}))$. This sequence is monotonically decreasing and satisfies $g^n \to g$ pointwise almost everywhere for $n \to \infty$.

Now, we recall that

$$\tilde{\mathcal{J}}_{0,T}(\mu, g^n) = \sum_{x\in\mathcal{X}} g_T^n(x)\mu_T(x) - \sum_{x\in\mathcal{X}} g_0^n(x)\mu_0(x) - \int_0^T \sum_{x\in\mathcal{X}} \partial_t g_t^n(x)\mu_t(x) + \mathcal{H}(\mu_t, g_t^n)\,dt. \tag{50}$$

We first consider the asymptotic behaviour of $\int_0^T \mathcal{H}(\mu_t, g_t^n)\,dt$. Since $h_{k,n}$ have pairwise-disjoint support for large n, we find by the monotone convergence theorem

$$\int_0^T \mathcal{H}(\mu_t, g_t^n)\,dt = \int_0^T \sum_{x,y\in\mathcal{X}} \mu_t(x)L(x,y)\left[e^{\nabla g_t^n(y,x)} - 1\right] dt$$

$$= \sum_{k\in\mathcal{I}} \int_0^T \chi_{\mathrm{supp}(h_{k,n})} \sum_{x,y\in\mathcal{X}} \mu_t(x)L(x,y)\left[e^{h_{k,n}(t)\nabla\varphi_k(y,x)} - 1\right] dt$$

$$\xrightarrow{n\to\infty} \sum_{k\in\mathcal{I}} \int_{s_k}^{t_k} \sum_{x,y\in\mathcal{X}} \mu_t(x) L(x,y) \left[e^{\nabla\varphi_k(y,x)} - 1\right] dt = \sum_{k\in\mathcal{I}} \int_{s_k}^{t_k} \mathcal{H}(\mu_t, g_t)\, dt.$$

To study the first three terms on the right side of (50), for any $\phi \in \mathcal{C}^1([0,T];\mathbb{R})$ we define

$$\mathcal{F}_{k,n}(\phi) := h_{k,n}(T)\phi(T) - h_{k,n}(0)\phi(0) \qquad - \int_0^T \partial_t h_{k,n}(t)\phi(t)\, dt = \int_0^T h_{k,n}(t)\partial_t\phi(t)\, dt,$$

$$\mathcal{F}_k(\phi) := \phi(t_k) - \phi(s_k) = \int_{s_k}^{t_k} \partial_t\phi(t)\, dt,$$

where the second equality follows from the integration by parts formula. Note that both $\mathcal{F}_{k,n}$ and $\mathcal{F}_k$ are linear in ϕ and

$$|\mathcal{F}_{k,n}(\phi)| \le |h_{k,n}(T)\phi(T)| + |h_{k,n}(0)\phi(0)| + \int_0^T |\partial_t h_{k,n}(t)||\phi(t)|\, dt \le (2+C)\,\|\phi\|_\infty,$$

$$|\mathcal{F}_k(\phi)| \le 2\,\|\phi\|_\infty,$$

where the bounds are uniform in n, and that $\lim_{n\to\infty}\mathcal{F}_{k,n}(\phi) = \mathcal{F}_k(\phi)$ for all $\phi\in\mathcal{C}^1([0,T];\mathbb{R})$ and k. Now consider an arbitrary $\phi \in \mathcal{C}([0,T];\mathbb{R})$ and sequence $\phi_l \in \mathcal{C}^1([0,T];\mathbb{R})$ which uniformly converges to ϕ for $l\to\infty$. Then for every k we find

$$\lim_{n\to\infty}\mathcal{F}_{k,n}(\phi) = \lim_{n\to\infty}\lim_{l\to\infty}\mathcal{F}_{k,n}(\phi_l) = \lim_{l\to\infty}\lim_{n\to\infty}\mathcal{F}_{k,n}(\phi_l) = \lim_{l\to\infty}\mathcal{F}_k(\phi_l) = \mathcal{F}_k(\phi).$$

Using this, for any $\mu\in\mathcal{C}([0,T];\mathcal{P}(\mathcal{X}))$ we find

$$\begin{aligned}
&\sum_{x\in\mathcal{X}} g_T^n(x)\mu_T(x) - \sum_{x\in\mathcal{X}} g_0^n(x)\mu_0(x) - \int_0^T \sum_{x\in\mathcal{X}} \partial_t g_t^n(x)\mu_t(x)\, dt \\
&= \sum_{k\in\mathcal{I}}\sum_{x\in\mathcal{X}} \varphi_k(x)\left[h_{k,n}(T)\mu_T(x) - h_{k,n}(0)\mu_0(x) - \int_0^T \partial_t h_{k,n}(t)\mu_t(x)\, dt\right] \\
&= \sum_{k\in\mathcal{I}}\sum_{x\in\mathcal{X}} \varphi_k(x)\mathcal{F}_{k,n}(\mu(x)) \xrightarrow{n\to\infty} \sum_{k\in\mathcal{I}}\sum_{x\in\mathcal{X}} \varphi_k(x)\mathcal{F}_k(\mu(x)) \\
&= \sum_k \left[\sum_{x\in\mathcal{X}} \varphi_k(x)\mu_{t_k}(x) - \sum_{x\in\mathcal{X}} \varphi_k(x)\mu_{s_k}(x)\right],
\end{aligned}$$

where we have used Fubini's theorem to arrive at the first equality and the dominated convergence theorem to pass to the limit. Together with the convergence of the Hamiltonian proved earlier, we have the result. □

Proof of Theorem 1.4. We first prove the large-deviation principle itself. Applying [19] to the generator L, we take for its core D the space $c_0(\mathcal{X})$, equipped with the supremum norm, so that the dual D' is isomorphic to $\ell^1(\mathcal{X})$. Then [19, Proposition 5.10] implies that ρ^N satisfies a large-deviation principle in $D_{\mathcal{P}(\mathcal{X})}[0,T]$ with rate function

$$\widehat{\mathscr{I}}_L(\mu) = \begin{cases} \displaystyle\int_0^T \widehat{\mathcal{L}}(\mu_t, \partial_t\mu_t)\, dt, & \text{if } \mu \in D\text{-}AC([0,T];\mathcal{P}(\mathcal{X})), \\ +\infty, & \text{otherwise.} \end{cases} \tag{51}$$

Here the Lagrangian $\widehat{\mathcal{L}} : \mathcal{P}(\mathcal{X}) \times \ell^1(\mathcal{X}) \to [0,\infty]$ given in terms of $\mathcal{H}$ in (10) by

$$\widehat{\mathcal{L}}(\mu, s) := \sup_{f\in c_0(\mathcal{X})} \langle f, s\rangle - \mathcal{H}(\mu, f),$$

and the space $D\text{-}AC([0,T];\mathcal{P}(\mathcal{X}))$ is the space of curves $\nu : [0,\infty) \to \mathcal{P}(\mathcal{X})$ such that $t \mapsto \langle f, \nu(t)\rangle$ is absolutely continuous for all $f \in D = c_0(\mathcal{X})$, with a unique weak-star measurable derivative $u : [0,\infty) \to D' = \ell^1(\mathcal{X})$ in the sense that $(d/dt)\langle \nu(t), f\rangle = \langle f, u(t)\rangle$ for all $f \in c_0(\mathcal{X})$ and $t \geq 0$.

The rate function $\widehat{\mathscr{I}_L}$ in (51) differs from $\mathscr{I}_L$ in (7) in two ways. First, the explicit domain of definition in (7) is $AC([0,T];\mathcal{P}(\mathcal{X}))$, the space of curves that are absolutely continuous in $\ell^1(\mathcal{X})$; this is a subspace of $D\text{-}AC([0,T];\mathcal{P}(\mathcal{X}))$. Secondly, $\mathcal{L}(\mu, s)$ is defined as a supremum over $\ell^\infty(\mathcal{X})$, while $\widehat{\mathcal{L}}(\mu, s)$ is defined as the same supremum but over the smaller space $c_0(\mathcal{X})$, implying that $\widehat{\mathcal{L}} \leq \mathcal{L}$.

Nonetheless, we have $\widehat{\mathscr{I}_L} = \mathscr{I}_L$. To show this, we first note that for $s \in \ell^1(\mathcal{X})$ and $\mu \in \mathcal{P}(\mathcal{X})$, we have

$$\sup_{f\in\ell^\infty(\mathcal{X})} \langle f, s\rangle - \mathcal{H}(\mu, f) = \sup_{f\in c_0(\mathcal{X})} \langle f, s\rangle - \mathcal{H}(\mu, f), \tag{52}$$

and therefore $\widehat{\mathcal{L}}(\mu, s) = \mathcal{L}(\mu, s)$ for all $s \in \ell^1(\mathcal{X})$. Indeed, fix $s \in \ell^1(\mathcal{X})$ and $f \in \ell^\infty(\mathcal{X})$, and let $f_n \in c_0(\mathcal{X})$ be the truncation of f to the first n elements of $\mathcal{X}$. Then

$$\sum_{x\in\mathcal{X}} f_n(x)s(x) \to \sum_{x\in\mathcal{X}} f(x)s(x) \qquad \text{and}$$
$$\sum_{x,y\in\mathcal{X}} \mu(x)L(x,y)\left[e^{f_n(y)-f_n(x)} - 1\right] \to \sum_{x,y\in\mathcal{X}} \mu(x)L(x,y)\left[e^{f(y)-f(x)} - 1\right],$$

both by the dominated convergence theorem, since $s \in \ell^1(\mathcal{X})$ and $(x,y) \mapsto \mu(x)L(x,y) \in \ell^1(\mathcal{X}\times\mathcal{X})$. This proves (52), and shows that for $s \in \ell^1(\mathcal{X})$, $\widehat{\mathcal{L}}(\mu, s) = \mathcal{L}(\mu, s)$.

Next, by [19, Proposition 2.12], curves μ with $\widehat{\mathscr{I}_L}(\mu) < \infty$ satisfy $\mu \in AC([0,T];\mathcal{P}(\mathcal{X}))$. Since curves in $AC([0,T];\mathcal{P}(\mathcal{X}))$ have derivatives in ℓ^1, any curve with $\widehat{\mathscr{I}_L}(\mu) < \infty$ satisfies

$$\widehat{\mathscr{I}_L}(\mu) = \int_0^T \widehat{\mathcal{L}}(\mu_t, \partial_t\mu_t)\,dt = \int_0^T \mathcal{L}(\mu_t, \partial_t\mu_t)\,dt = \mathscr{I}_L(\mu).$$

This proves that $\widehat{\mathscr{I}_L} = \mathscr{I}_L$ whenever $\widehat{\mathscr{I}_L} < \infty$. For the remaining case $\widehat{\mathscr{I}_L}(\mu) = \infty$ there are three possibilities:

1. $\mu \notin D\text{-}AC([0,T];\mathcal{P}(\mathcal{X}))$, therefore $\mu \notin AC([0,T];\mathcal{P}(\mathcal{X}))$ and $\mathscr{I}_L(\mu) = \infty$ also;
2. $\mu \in D\text{-}AC([0,T];\mathcal{P}(\mathcal{X}))$ but $\mu \notin AC([0,T];\mathcal{P}(\mathcal{X}))$ and again $\mathscr{I}_L(\mu) = \infty$;
3. $\mu \in AC([0,T];\mathcal{P}(\mathcal{X}))$ but
$$\infty = \int_0^T \widehat{\mathcal{L}}(\mu_t, \partial_t\mu_t)\,dt \leq \int_0^T \mathcal{L}(\mu_t, \partial_t\mu_t)\,dt,$$
so that again $\mathscr{I}_L(\mu) = \infty$.

This proves that $\mathscr{I}_L = \widehat{\mathscr{I}_L}$ and concludes the proof of the large-deviation principle.

We now continue with the characterisation (12). We define

$$\tilde{\mathscr{I}_L}(\mu) := \sup_{f\in C^1([0,T];\ell^\infty(\mathcal{X}))} \tilde{\mathcal{J}}_{0,T}(\mu, f),$$

where $\tilde{\mathcal{J}}_{0,T}$ is given by (49).

The plan of the proof is now as follows. We first show that $\tilde{\mathscr{I}_L}(\mu) < \infty$ for $\mu \in C([0,T];\mathcal{P}(\mathcal{X}))$ implies that $\mu \in AC([0,T];\mathcal{P}(\mathcal{X}))$. We then show that $\mathscr{I}_L(\mu) \geq \tilde{\mathscr{I}_L}(\mu)$ and vice versa which yields the equality. In particular, applying integration by parts in (49) since $\mu \in \mathcal{AC}([0,T];\mathcal{P}(\mathcal{X}))$, yields

$$\mathscr{I}_L(\mu) = \sup_{f\in L^\infty(0,T;\ell^\infty(\mathcal{X}))} \int_0^T \langle f_t, \partial_t\mu_t\rangle - \mathcal{H}(\mu_t, f_t)\,dt,$$

which is the last part of the statement.

We now show by contradiction that $\mu \in \mathcal{C}([0,T];\mathcal{P}(\mathcal{X}))$ and $\tilde{\mathscr{I}}_L(\mu) < \infty$ implies $\mu \in \mathcal{AC}([0,T];\mathcal{P}(\mathcal{X}))$. Suppose $\tilde{\mathscr{I}}_L(\mu) < \infty$, but $\mu \notin \mathcal{AC}([0,T];\mathcal{P}(\mathcal{X}))$, i.e. there exists an $\varepsilon > 0$ such that for any $\delta > 0$, there exists a finite family of pairwise-disjoint intervals $[s_k, t_k] \subset [0,T]$, $k \in \mathcal{I}$ with

$$\sum_{k\in\mathcal{I}} |t_k - s_k| < \delta \quad \text{and} \quad \sum_{k\in\mathcal{I}}\sum_{x\in\mathcal{X}} |\mu_{t_k}(x) - \mu_{s_k}(x)| \geq \varepsilon.$$

Next, for an arbitrary $A > 0$, we define $g \in L^\infty(0,T;\ell^\infty(\mathcal{X}))$ as

$$g_t(x) := A\sum_{k\in\mathcal{I}} \operatorname{sign}(\mu_{t_k}(x) - \mu_{s_k}(x))\chi_{[s_k,t_k]}(t).$$

Using Corollary B.1, there exists a sequence $g^n \in \mathcal{C}^1([0,T];\ell^\infty(\mathcal{X}))$ such that

$$\tilde{\mathcal{J}}_{0,T}(\mu, g^n) \quad \xrightarrow{n\to\infty} \quad \sum_{k\in\mathcal{I}} \tilde{\mathcal{J}}_{s_k,t_k}(\mu, g). \tag{53}$$

Note that the latter expression is well defined since $g|_{[s_k,t_k]} \in \mathcal{C}^1([s_k,t_k];\ell^\infty(\mathcal{X}))$ for all $k \in \mathcal{I}$. Moreover, there exists a $C < \infty$ which only depends on μ and L such that

$$\sum_{k\in\mathcal{I}} \int_{s_k}^{t_k} \mathcal{H}(\mu_t, g_t)\,dt \leq Ce^A \sum_{k\in\mathcal{I}} |t_k - s_k| < Ce^A\delta,$$

since $\operatorname{sign}(\mu_{t_k} - \mu_{s_k})$ is uniformly bounded in $\mathcal{X}$. Furthermore, we find

$$\begin{aligned}
&\sum_{k\in\mathcal{I}}\left[\sum_{x\in\mathcal{X}} g_{t_k}(x)\mu_{t_k}(x) - \sum_{x\in\mathcal{X}} g_{s_k}(x)\mu_{s_k}(x)\right] \\
&\quad = A\sum_{k\in\mathcal{I}}\sum_{x\in\mathcal{X}} \operatorname{sign}(\mu_{t_k}(x) - \mu_{s_k}(x))(\mu_{t_k}(x) - \mu_{s_k}(x)) \\
&\quad = A\sum_{k\in\mathcal{I}}\sum_{x\in\mathcal{X}} |\mu_{t_k}(x) - \mu_{s_k}(x)| \geq A\varepsilon.
\end{aligned}$$

Thus, using (53) we find

$$\tilde{\mathcal{J}}_{0,T}(\mu, g^n) \geq \frac{1}{2}\sum_{k\in\mathcal{I}} \tilde{\mathcal{J}}_{s_k,t_k}(\mu, g) \geq \frac{1}{2}\left(A\varepsilon - Ce^A\delta\right),$$

for sufficiently large n. Since $\delta > 0$ and $A > 0$ were arbitrary, the right-hand side can be arbitrarily large. More specifically, for a given A, we choose $\delta = \varepsilon Ae^{-A}/(2C)$, thereby yielding

$$\tilde{\mathscr{I}}_L(\mu) \geq \tilde{\mathcal{J}}_{0,T}(\mu, g^n) \geq \frac{1}{4}\varepsilon A.$$

Since A can be made arbitrarily large, this contradicts $\tilde{\mathscr{I}}_L(\mu) < \infty$. Hence, $\mu \in \mathcal{C}([0,T];\mathcal{P}(\mathcal{X}))$ and $\tilde{\mathscr{I}}_L(\mu) < \infty$ imply that $\mu \in \mathcal{AC}(0,T;\ell^1(\mathcal{X}))$.

Next, we show that $\mathscr{I}_L(\mu) \ge \tilde{\mathscr{I}_L}(\mu)$. For $\mu \notin \mathcal{AC}([0,T];\mathcal{P}(\mathcal{X}))$ we have $\mathscr{I}_L(\mu) = \infty$ and therefore $\mathscr{I}_L(\mu) \ge \tilde{\mathscr{I}_L}(\mu)$. For $\mu \in \mathcal{AC}([0,T];\mathcal{P}(\mathcal{X}))$, on the other hand, we have

$$\mathscr{I}_L(\mu) = \int_0^T \mathcal{L}(\mu_t, \partial_t\mu_t) \ge \int_0^T \langle f_t, \partial_t\mu_t\rangle - \mathcal{H}(\mu_t, f_t)\,dt = \tilde{\mathcal{J}}_{0,T}(\mu, f),$$

for any curve $f \in \mathcal{C}^1([0,T];\ell^\infty(\mathcal{X}))$, where we used integration by parts to arrive at the final equality. This yields $\mathscr{I}_L(\mu) \ge \tilde{\mathscr{I}_L}(\mu)$.

We complete the proof by showing that $\mathscr{I}_L(\mu) \le \tilde{\mathscr{I}_L}(\mu)$ for $\mu \in \mathcal{AC}([0,T];\mathcal{P}(\mathcal{X}))$. Note that since $\mu \in \mathcal{AC}([0,T];\ell^1(\mathcal{X})) \equiv W^{1,1}(0,T;\ell^1(\mathcal{X}))$ and

$$\tilde{\mathcal{J}}_{0,T}(\mu, f) = \int_0^T \langle f_t, \partial_t\mu_t\rangle - \mathcal{H}(\mu_t, f_t)\,dt \qquad \text{for all } f \in \mathcal{C}^1([0,T];\ell^\infty(\mathcal{X})),$$

we have that $\tilde{\mathcal{J}}_{0,T}(\mu,\cdot) : L^\infty(0,T;\ell^\infty(\mathcal{X})) \to \mathbb{R}$ is a continuous (nonlinear) functional. Since every element in $L^\infty(0,T;\ell^\infty(\mathcal{X}))$ can be approximated pointwise by a sequence in $\mathcal{C}^1([0,T];\ell^\infty(\mathcal{X}))$, we can extend the above representation to all $f \in L^\infty((0,T);\ell^\infty(\mathcal{X}))$ by using the dominated convergence theorem. In particular, this yields that

$$\sup_{f\in\mathcal{C}^1([0,T];\ell^\infty(\mathcal{X}))} \tilde{\mathcal{J}}_{0,T}(\mu, f) = \sup_{f\in L^\infty(0,T;\ell^\infty(\mathcal{X}))} \tilde{\mathcal{J}}_{0,T}(\mu, f).$$

Now, for any fixed $\varepsilon > 0$ and for almost every $t \in (0,T)$ exists a $g_t \in \ell^\infty(\mathcal{X})$ such that

$$\sum_{x\in\mathcal{X}} g_t(x)\partial_t\mu_t(x) - \mathcal{H}(\mu_t, g_t) \ge \max\{\mathcal{L}(\mu_t, \partial_t\mu_t) - \varepsilon, 0\},$$

where we used the definition of the Lagrangian. Note that $t \mapsto g_t$ might not be an element of $L^\infty(0,T;\ell^\infty(\mathcal{X}))$. Therefore, we define the sequence

$$f_t^k(x) := \begin{cases} g_t(x) & \text{if } \|g_t\|_{\ell^\infty(\mathcal{X})} \le k, \\ 0 & \text{otherwise,} \end{cases}$$

with $k \in \mathbb{N}$. Then, by construction we have that $0 \le \sum_{x\in\mathcal{X}} f_t^k(x)\partial_t\mu_t(x) - \mathcal{H}(\mu_t, f_t^k) \le \mathcal{L}(\mu_t, \partial_t\mu_t)$ for all $k \in \mathbb{N}$, where $t \mapsto \mathcal{L}(\mu_t, \partial_t\mu_t) \in L^1(0,T;[0,\infty))$ since $\mu \in \mathcal{AC}([0,T];\mathcal{P}(\mathcal{X}))$. Furthermore, using $f_t^k(x) \le g_t(x)$ for all $x \in \mathcal{X}$ and almost all $t \in (0,T)$ and the dominated convergence theorem, we find

$$\sum_{x\in\mathcal{X}} f_t^k(x)\partial_t\mu_t(x) - \mathcal{H}(\mu_t, f_t^k) \quad \xrightarrow{k\to\infty} \quad \sum_{x\in\mathcal{X}} g_t(x)\partial_t\mu_t(x) - \mathcal{H}(\mu_t, g_t),$$

for almost every $t \in (0,T)$. Hence, we can apply the dominated convergence theorem to obtain

$$\lim_{k\to\infty} \tilde{\mathcal{J}}_{0,T}(\mu, f^k) = \tilde{\mathcal{J}}_{0,T}(\mu, g) \ge \int_0^T \mathcal{L}(\mu_t, \partial_t\mu_t)\,dt - \varepsilon T.$$

Finally, since $f^k \in L^\infty(0,T;\ell^\infty(\mathcal{X}))$ for all $k \in \mathbb{N}$ we obtain that the left-hand side is bounded from above by $\tilde{\mathscr{I}_L}(\mu)$. Therefore, since $\varepsilon > 0$ was arbitrary, we obtain $\tilde{\mathscr{I}_L}(\mu) \ge \mathscr{I}_L(\mu)$ which proves the statement. □

Appendix C. Positivity of solution to the forward Kolmogorov equation

In this appendix we show that the solution to the forward Kolmogorov equation with a bounded and irreducible generator is strictly positive. While we expect this result to be known, we could not find a reference for it, and therefore provide the result here for completeness.

Lemma C.1. *Let $\rho \in AC([0,T];\mathcal{P}(\mathcal{X}))$ be a solution to (2), where the generator L satisfies (3a)–(3c). Then $\rho_t \in \mathcal{P}_+(\mathcal{X})$ for every $t > 0$.*

Proof. Since L is a bounded Markov generator with $\nu := \sup_{x\in\mathcal{X}} |L(x,x)| < \infty$, we can write $L = P - \nu I$ for a matrix P with non-negative entries and identity matrix I. Note that L^T generates a uniformly continuous semigroup on $\ell^1(\mathcal{X})$ which conserves mass, i.e. if $\mu_t = e^{tL^T}\mu_0$, then $\sum_{x\in\mathcal{X}} \mu_t(x) = \sum_{x\in\mathcal{X}} \mu_0(x)$. Therefore, we can write $e^{tL^T} = \sum_{n\geq 0} \frac{t^n (L^T)^n}{n!}(x,y) = e^{t(P^T-\nu I)} = e^{-\nu t} e^{tP^T}$. We will show that $e^{tL^T}(x,y) > 0$, by proving that $e^{tP^T}(x,y) > 0$.

Since L is irreducible, for every $x, y \in \mathcal{X}$ with $x \neq y$, there exists a finite sequence $x_0, x_1, \dots, x_N \in \mathcal{X}$ containing no doubled points with $x_0 = x,\ x_N = y$ and $L(x_n, x_{n+1}) > 0$. Using $L = P - \nu I$, $P(x_n, x_{n+1}) = L(x_n, x_{n+1}) > 0$ we find

$$(P^T)^N(x,y) \geq \sum_{i=1}^{N-1} P^T(x_{i+1}, x_i) P^T(x_i, x_{i-1}) = \sum_{i=1}^{N-1} P(x_i, x_{i+1}) P(x_{i-1}, x_i) > 0.$$

Therefore

$$e^{tP^T}(x,y) \geq \frac{t^N (P^T)^N}{N!}(x,y) > 0.$$

Since $x, y \in \mathcal{X}$ are arbitrary, it follows that e^{tL^T} is a positive semigroup and therefore $e^{tL^T} : \mathcal{P}(\mathcal{X}) \to \mathcal{P}_+(\mathcal{X})$ for all $t > 0$. □

References

[1] A. Arnold, J.A. Carrillo, L. Desvillettes, J. Dolbeault, A. Jüngel, C. Lederman, P.A. Markowich, G. Toscani, C. Villani, Entropies and equilibria of many-particle systems: An essay on recent research, Monatsh. Math. 142 (1) (2004) 35–43.

[2] D. Bakry, I. Gentil, M. Ledoux, Analysis and Geometry of Markov Diffusion Operators, in: Grundlehren der mathematischen Wissenschaften, vol. 348, Springer International Publishing, 2014.

[3] S.G. Bobkov, P. Tetali, Modified logarithmic Sobolev inequalities in discrete settings, J. Theoret. Probab. 19 (2) (2006) 289–336.

[4] V. Bogachev, M. Röckner, S. Shaposhnikov, Distances between transition probabilities of diffusions and applications to nonlinear Fokker–Planck–Kolmogorov equations, J. Funct. Anal. 271 (5) (2016) 1262–1300.

[5] A. Braides, Gamma-convergence for Beginners, in: Oxford Lecture Series in Mathematics and its Applications, vol. 22, Oxford University Press, 2002.

[6] H. Brezis, Functional Analysis, Sobolev Spaces and Partial Differential Equations, in: Universitext, Springer-Verlag New York, 2011.

[7] S.-N. Chow, W. Huang, Y. Li, H. Zhou, Fokker–planck equations for a free energy functional or Markov process on a graph, Arch. Ration. Mech. Anal. 203 (3) (2012) 969–1008.

[8] D.D. Dawson, J. Gärtner, Large deviations from the McKean-Vlasov limit for weakly interacting diffusions, Stochastics 20 (4) (1987) 247–308.

[9] P. Diaconis, L. Saloff-Coste, Logarithmic Sobolev inequalities for finite markov chains, Ann. Appl. Probab. 6 (3) (1996) 695–750.

[10] R. Dudley, Real Analysis and Probablity, Wadsworth & Brooks/Cole, 1989.

[11] M.H. Duong, A. Lamacz, M.A. Peletier, A. Schlichting, U. Sharma, Quantification of coarse-graining error in langevin and overdamped langevin dynamics, Nonlinearity 31 (10) (2018) 4517–4566.

[12] M.H. Duong, A. Lamacz, M.A. Peletier, U. Sharma, Variational approach to coarse-graining of generalized variational approach to coarse-graining of generalized gradient flows, Calc. Var. Partial Differential Equations 56 (4) (2017).

[13] K.-J. Engel, R. Nagel, in: S. Axler, K. Ribet (Eds.), A Short Course on Operator Semigroups, in: Universitext, Springer-Verlag New York, 2006.

[14] M. Erbar, M. Fathi, V. Laschos, A. Schlichting, Gradient flow structure for McKean-Vlasov equations on discrete spaces, Discrete Contin. Dyn. Syst. Ser. A 36 (12) (2016) 6799–6833.

[15] D. Givon, R. Kupferman, A. Stuart, Extracting macroscopic dynamics: model problems and algorithms, Nonlinearity 17 (6) (2004) R55.

[16] B. Hilder, An FIR Inequality for Markov Jump Processes on Discrete State Spaces (Master Thesis), Eindhoven University of Technology / University of Stuttgart, 2017, Master Thesis, Eindhoven University of Technology/University of Stuttgart (https://goo.gl/6jn8AE).

[17] T. Hytönen, J. Van Neerven, M. Veraar, L. Weis, Analysis in Banach spaces, Volume I: Martingales and Littlewood-Paley Theory, in: A Series of Modern Surveys in Mathematics, vol. 63, Springer, 2016.

[18] R. Kraaij, Large deviations for Markov jump processes with mean-field interaction via the comparison principle for an associated Hamilton-Jacobi, J. Stat. Phys. 164 (2) (2016) 321–345.

[19] R. Kraaij, Large deviations of the trajectory of empirical distributions of Feller processes on locally compact spaces, Ann. Probab. 46 (2) (2018) 775–828.

[20] C. Kuehn, Multiple Time Scale Dynamics, Springer International Publishing, 2015.

[21] S. Lahbabi, Étude Mathématique de Modèles Quantiques et Classigue Pour les Matériaux Aléatoires à l'échelle Atomique (Ph.D. thesis), Université de Cergy-Pontoise, 2013.

[22] S. Lahbabi, F. Legoll, Effective dynamics for a kinetic Monte–Carlo model with slow and fast time scales, J. Stat. Phys. 153 (6) (2013) 931–966.

[23] F. Legoll, T. Lelièvre, Effective dynamics using conditional expectations, Nonlinearity 23 (9) (2010) 2131–2163.

[24] J. Maas, Gradient flows of the entropy for finite Markov chains, J. Funct. Anal. 261 (8) (2011) 2250–2292.

[25] M. Marcus, V.J. Mizel, Absolute continuity on tracks and mappings of sobolev spaces, Arch. Ration. Mech. Anal. 45 (4) (1972) 294–320.

[26] P. Michel, S. Mischler, B. Perthame, General relative entropy inequality: an illustration on growth models, J. Math. Pures Appl. 84 (9) (2005) 1235–1260.

[27] A. Mielke, A gradient structure for reaction–diffusion systems and for energy-drift-diffusion systems, Nonlinearity 24 (4) (2011) 1329–1346.

[28] A. Mielke, Geodesic convexity of the relative entropy in reversible Markov chains, Calc. Var. Partial Differential Equations 48 (1) (2013) 1–31.

[29] A. Mielke, On evolutionary Γ-convergence for gradient systems, in: A. Muntean, J. Rademacher, A. Zagaris (Eds.), Macroscopic and Large Scale Phenomena: Coarse Graining, Mean Field Limits and Ergodicity, Springer International Publishing, 2016.

[30] A. Mielke, M.A. Peletier, M. Renger, On the relation between gradient flows and the large-deviation principle, with applications to Markov chains and diffusion, Potential Anal. 41 (4) (2014) 1293–1327.

[31] J.R. Munkres, Topology, second ed., Prentice Hall, 2000.

[32] K. Oelschlager, A martingale approach to the law of large numbers for weakly interacting stochastic processes, Ann. Probab. (1984) 458–479.

[33] G.A. Pavliotis, A. Stuart, Multiscale Methods: Averaging and Homogenization, Springer Science & Business Media, 2008.

[34] E. Sandier, S. Serfaty, Gamma-convergence of gradient flows with applications to Ginzburg-Landau, Comm. Pure Appl. Math. 57 (12) (2004) 1627–1672.

[35] S. Serfaty, Gamma-convergence of gradient flows on Hilbert spaces and metric spaces and appliations, Discrete Contin. Dyn. Syst. Ser. A 31 (4) (2011) 1427–1451.

[36] E. Shargorodsky, J.F. Toland, Bernoulli Free-Boundary Problems, Vol. 912–918, American Mathematical Soc., 2008.

[37] U. Sharma, Coarse-graining of Fokker-Planck Equations (Ph.D. thesis), Eindhoven University of Technology, 2017.

[38] C.d.l. Vallée Poussin, Sur l'intégrale de Lebesgue, Trans. Amer. Math. Soc. (1915) 435–501.

[39] S.R.S. Varadhan, Asymptotic probabilities and differential equations, Comm. Pure Appl. Math. 19 (3) (1966) 261–286.

[40] J. Voigt, Stochastic operators, information, and entropy, Comm. Math. Phys. 81 (1) (1981) 31–38.

[41] H.-T. Yau, Relative entropy and hydrodynamics of Ginzburg-Landau models, Lett. Math. Phys. 22 (1) (1991) 63–80.