Functional inequalities and convergence of stochastic processes

by

Francesco Pedrotti

July, 2024

A thesis submitted to the
Graduate School
of the
Institute of Science and Technology Austria
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

Committee in charge:

Matthew Robinson, Chair
Jan Maas
Matthias Erbar
Matthew Kwan

Institute of Science and Technology Austria

The thesis of Francesco Pedrotti, titled *Functional inequalities and convergence of stochastic processes*, is approved by:

Supervisor: Prof. Jan Maas, ISTA, Klosterneuburg, Austria

Signature: ______________________________

Committee Member: Prof. Matthias Erbar, Universität Bielefeld, Bielefeld, Germany

Signature: ______________________________

Committee Member: Prof. Matthew Kwan, ISTA, Klosterneuburg, Austria

Signature: ______________________________

Defense Chair: Prof. Matthew Robinson, ISTA, Klosterneuburg, Austria

Signature: ______________________________

Signed page is on file

ISTA Thesis, ISSN: 2663-337X

I hereby declare that this thesis is my own work and that it does not contain other people's work without this being so stated; this thesis does not contain my previous work without this being stated, and the bibliography contains all the literature that I used in writing the dissertation.

I accept full responsibility for the content and factual accuracy of this work, including the data and their analysis and presentation, and the text and citation of other work.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee, and that this thesis has not been submitted for a higher degree to any other university or institution.

I certify that any republication of materials presented in this thesis has been approved by the relevant publishers and co-authors.

Signature: ____________________________

Francesco Pedrotti
July, 2024

Signed page is on file

Abstract

This thesis deals with the study of stochastic processes and their ergodicity properties. The variety of problems encountered calls for a set of different approaches, ranging from classical to modern ones: a special place is held by probabilistic methods based on couplings, by functional inequalities, and by the theory of gradient flows in the space of measures.

The material is organized as follows. Chapter 1 contains the introduction to this thesis, starting with a general presentation of some of the relevant topics. Section 1.1 is dedicated to the theory of gradient flows in metric spaces, and introduces the first contribution of this thesis [DSMP24], which is presented in detail in Chapter 2. Section 1.2 moves to the topic of curvature of Markov chains, concluding with a brief description of our second contribution [Ped23], which is included in Chapter 3. Section 1.3 discusses applications of stochastic processes to the theory of sampling, in particular the recent framework of score-based diffusion models, and our contribution [PMM24], which is contained in Chapter 4. Section 1.4 discusses some related problems, concerning the regularization properties of the heat flow. It serves as a motivation for the work [BP24], which we report in Chapter 5. Finally, Section 1.5 discusses the last contribution of this thesis, which can be found in Chapter 6. It deals with the convergence to equilibrium of a particular stochastic model from quantitative genetics: this is established via some functional inequalities, which we prove with probabilistic arguments based on couplings.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor, Jan Maas. When I started the PhD program, I had really broad interests and was unsure about which research directions I would have liked to work on. It still amazes me how Jan managed to guide me through different problems, apparently unrelated and meeting many of my interests, but that turned out to share some intriguing connections.

I am also grateful to the members of my thesis committee, Matthias Erbar and Matthew Kwan, for their time, interest and valuable suggestions, and to Matthew Robinson for acting as a defense chair for the thesis defense.

Further thanks are due to my colleagues and friends at ISTA, responsible for an always great atmosphere and for sharing many interesting conversations. I would like to thank especially Marco Mondelli for our work together, and all the members of Jan's group in the past years: Giovanni Brigati, Lorenzo Dello Schiavo, Ksenia Khuadiakova, Hila Lifchitz, Lorenzo Portinale, Filippo Quattrocchi, Melchior Wirth, and Haonan Zhang.

Next, I would like to extend my heartfelt thanks to my roommates, Marko Kojic and Simone Bombari: I could always count on you, and I will remember fondly our time together.

Moving away from Vienna, I am especially indebted to my family for their constant support and help on so many occasions. You have made me miss home and cherish every visit.

I am also grateful to my friends in Trento for staying close despite many years abroad.

Last but not least, I cannot thank enough Ofelia for being at my side since the beginning and always believing in me.

Funding. The research contained in this thesis has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 716117) and from the Austrian Science Fund (FWF) project 10.55776/F65.

About the Author

Francesco Pedrotti completed his undergraduate studies in mathematics at the University of Trento, in July 2019. In July 2020, he obtained a Master's degree in mathematics at the University of Cambridge (*Part III of the Mathematical Tripos*), where he specialized in probability theory. Subsequently, he started his PhD studies at the Institute of Science and Technology Austria in September 2020, under the supervision of Prof. Jan Maas. His main research interests are probability theory and optimal transport, with a special emphasis on functional inequalities that are related to the concentration of measure phenomenon and to fast convergence of stochastic processes to equilibrium.

List of Collaborators and Publications

During my studies at the Institute of Science and Technology Austria (ISTA) in the years 2020-2024, I was lucky to collaborate with different colleagues, including my supervisor Jan Maas, Giovanni Brigati, Lorenzo Dello Schiavo, Ksenia Khudiakova and Marco Mondelli.

This thesis is based in particular on the following articles:

- Lorenzo Dello Schiavo, Jan Maas, and Francesco Pedrotti. Local conditions for global convergence of gradient flows and proximal point sequences in metric spaces. *Trans. Am. Math. Soc.*, 377(6):3779–3804, 2024.

  This work, referred to as [DSMP24] in this thesis, is presented in Chapter 2.

- Francesco Pedrotti, Jan Maas, and Marco Mondelli. Improved convergence of score-based diffusion models via prediction-correction. *Transactions on Machine Learning Research*, 2024.

  This work, referred to as [PMM24] in this thesis, is presented in Chapter 4.

- Francesco Pedrotti. Contractive coupling rates and curvature lower bounds for Markov chains. *arXiv preprint arXiv:2308.00516*, 2023.

  This work, referred to as [Ped23] in this thesis, is presented in Chapter 3.

- Ksenia A Khudiakova, Jan Maas, and Francesco Pedrotti. L^∞-optimal transport of anisotropic log-concave measures and exponential convergence in Fisher's infinitesimal model. *arXiv preprint arXiv:2402.04151*, 2024.

  This work, referred to as [KMP24] in this thesis, is presented in Chapter 6.

- Giovanni Brigati and Francesco Pedrotti. Heat flow, log-concavity, and Lipschitz transport maps. *arXiv preprint arXiv:2404.15205*, 2024.

  This work, referred to as [BP24] in this thesis, is presented in Chapter 5.

Table of Contents

CHAPTER 1

Introduction

A general motivation for this thesis is the study of convergence properties of stochastic processes. Needless to say, this is a vast topic with applications in several fields: a few notable examples, relevant to this thesis, include Markov chains, diffusion processes on $\mathbb{R}^d$, algorithms for sampling, and evolutionary models from quantitative genetics. This variety of settings is reflected in a broad range of approaches to tackle the problem, which often draw inspiration from different areas of mathematics, creating deep and fruitful connections.

A first example of this, and a recurrent theme in this thesis, is given by the topic of functional inequalities. It turns out that some functional inequalities play a fundamental role in the study of convergence to equilibrium of Markov processes, by governing the rate at which an appropriate distance to equilibrium decays to 0 with time. On the other hand, such inequalities often describe also interesting properties of the equilibrium measure of the stochastic process, typically related to the *concentration of measure phenomenon*. Roughly speaking, this expression indicates the remarkable observation that, for many probability distributions naturally arising in high-dimensional spaces, most of the mass lives close to each other. In other words, there are relatively small regions that capture almost all the mass of these probability distributions. Perhaps the most fundamental example is given by the standard Gaussian distribution, which satisfies many functional inequalities of this type [BGL14].

Not surprisingly, by involving notions of closeness/distance, the concentration of measure phenomenon is also intimately connected to geometric properties of the space; in particular, a key role is played by the concept of (positive) curvature. Seminal results in this direction are due to Lévy [L5́1], who proved the concentration of measure phenomenon for the unit sphere in $\mathbb{R}^d$, and to Gromov [MS86], who extended this observation to Riemannian manifolds having positive Ricci curvature, in such a way that the more the space is curved, the stronger the implied concentration phenomenon is. In these examples, the reference measure and distance are naturally taken to be the uniform one and the geodesic one, respectively. The celebrated work by Bakry and Émery further generalized these results, as we now recall. In the setting of Riemannian manifolds, denote by Ric the Ricci curvature, and consider a diffusion process $(X_t)_t$ with generator $L = \Delta - \nabla V \cdot \nabla$, for a smooth potential V. The corresponding semigroup P_t admits a reversible measure $\pi \propto e^{-V} d\lambda$, where λ denotes the uniform measure. It was proved in [BE85] that, if $\nabla^2 V + \mathrm{Ric} > 0$, then π satisfies a log-Sobolev inequality. This fundamental inequality is a prominent member of the aforementioned family of functional inequalities, as it implies both strong concentration properties for the measure π, and fast convergence in law of the diffusion process to π itself. Notice that by choosing $V = 0$ one recovers the Gromov–Lévy theorem, and by choosing $V(x) = |x|^2/2$ in the Euclidean space one recovers the Gaussian log-Sobolev inequality. This result, however, extends the concentration properties to a much larger class of measures defined on Riemannian manifolds, elucidating the favourable effect of positive curvature. At the same time, this suggests also a change of perspective, that is, to treat the quantity $\nabla^2 V + \mathrm{Ric}$ as an "effective" curvature of the random process $(X_t)_t$. The development by Bakry–Émery of this idea gave rise to a powerful theory of curvature for Markov processes, based on an abstract set of rules known as Γ-calculus and the corresponding *curvature-dimension condition* $\mathrm{CD}(K, d)$ [BGL14].

The established connections between curvature and powerful functional inequalities motivated the effort to extend these definitions and implications to even more general settings. It is not surprising that the theory of optimal transport, lying at the intersection of probability theory and geometry, turned out to be a key tool in this respect. It is building on this theory that Sturm [Stu06] and Lott and Villani [LV09] independently developed a notion of curvature lower bound for a large class of geodesic metric measure spaces (X, d, m) (here X is a set, d a distance and m a measure). The curvature of the space is characterized in terms of geodesic convexity properties of the relative entropy functional (with respect to the reference measure m), in the 2-Wasserstein space, i.e. the space of probability measures with finite second moment equipped with the 2-Wasserstein distance from optimal transport. Remarkably, this definition is consistent with the classical notion of Ricci curvature for Riemannian manifolds, and even in this very abstract setting positive curvature is shown to imply many functional inequalities.

Optimal transport gives also new insights into the role played by functional inequalities both in the concentration of measure phenomenon and in the speed of convergence of some stochastic processes to equilibrium, and into the connections with the geometry of the space (the curvature in particular). The key observation, starting with the seminal work [JKO98], is that the evolution of some random processes (interpreted as curves in the space of probability measures), can be interpreted as a *deterministic gradient flow* for an appropriate functional in the 2-Wasserstein space. An important example comes from considering the relative entropy as the driving functional, for which the associated stochastic process is the Langevin dynamics. Recalling that in the Lott–Sturm–Villani theory positive curvature is defined in terms of convexity of this functional, it appears now rather intuitive that this implies fast convergence of the Langevin dynamics to equilibrium, based on the common knowledge that gradient flows for convex functionals exhibit favourable properties. This variational perspective turns out to be very useful for the convergence analysis of these random processes: indeed, by transforming the original problem into the study of the convergence of a gradient flow, one can draw inspiration from the field of optimization to come up with new results or new proofs and interpretations of existing ones. Because of the great influence of this point of view on this thesis, the next subsection is devoted to the theory of gradient flows in metric spaces and their convergence properties.

1.1 Gradient flows in metric spaces

For a smooth function $F \colon \mathbb{R}^d \to \mathbb{R}$, a gradient flow (started at a point $x_0 \in \mathbb{R}^d$) is a solution $(y_t)_t = (y_t)_{t \geq 0}$ to the Cauchy problem

$$\begin{cases} y_0 &= x_0, \\ y_t' &= -\nabla F(y_t). \end{cases} \tag{1.1.1}$$

There are many different possible ways of extending this definition to the setting of abstract metric spaces (lacking, in general, a differential structure), which differ, e.g., in the level of generality, or in the conditions for existence and uniqueness, and that are inspired by corresponding properties holding in the smooth Euclidean case. The reader is referred to [AGS08] for a comprehensive study, and to [San17] for a concise overview. In this introduction, for the sake of exposition, we restrict our attention to complete metric spaces (X, d) and to non-negative proper lower semicontinuous functionals $F\colon X \to \mathbb{R}_{\geq 0} \cup \{+\infty\}$. These assumptions are not everywhere necessary in what follows, but are anyway satisfied for the main examples considered. We briefly describe below two possible ways of defining gradient flows for the functional F.

Curve of maximal slope. To motivate the first definition, consider the following simple observation in Euclidean spaces. Let $F\colon \mathbb{R}^d \to \mathbb{R}$ and $(y_t)_t$ be respectively a smooth function and a smooth curve in $\mathbb{R}^d$. Then, by the Cauchy–Schwarz and Young's inequality,

$$-\frac{\mathrm{d}}{\mathrm{d}t}F(y_t) = -\langle \nabla F(y_t), y_t' \rangle \leq \frac{1}{2}|\nabla F(y_t)|^2 + \frac{1}{2}|y_t'|^2.$$

Moreover, equality holds if and only if $y_t' = -\nabla F(y_t)$ for all t; therefore, one could take the converse inequality

$$-\frac{\mathrm{d}}{\mathrm{d}t}F(y_t) \geq \frac{1}{2}|\nabla F(y_t)|^2 + \frac{1}{2}|y_t'|^2$$

as a definition of gradient flow. To make sense of this in the abstract setting of lower semicontinuous functionals on (X, d), we need to define the analogue of the quantities $|\nabla F|$ and $|y_t'|$. For the first one, we consider the notion of descending slope

$$|\mathrm{D}^- F|(x) \coloneqq \limsup_{y \to x} \frac{[F(y) - F(x)]_-}{\mathrm{d}(y, x)}. \tag{1.1.2}$$

For the second, suppose that $(y_t)_t \in \mathrm{AC}^1_{\mathrm{loc}}(\mathbb{R}_{>0}; X)$, i.e. the curve $(y_t)_t$ is locally absolutely continuous. Then, we can naturally consider its metric speed, defined by

$$|\dot{y}_t| \coloneqq \lim_{s \to t} \frac{\mathrm{d}(y_s, y_t)}{|s - t|}.$$

With these choices, we arrive at the following first definition of gradient flow for F in (X, d), which is known in the literature as the definition of *curve of maximal slope* (cf. [MS20, Dfn. 4.1] and Chapter 2 for more details). Recall that for a function $F\colon X \to \mathbb{R} \cup \{+\infty\}$ we define $\mathrm{dom}(F) = \{x \in X \mid F(x) \neq +\infty\}$ and we say that F is proper if $\mathrm{dom}(F) \neq \emptyset$.

Definition 1.1.1. *Let $F\colon X \to \mathbb{R}_{\geq 0} \cup \{+\infty\}$ be proper and lower semicontinuous. We say that $(y_t)_t$ is a* curve of maximal slope for F *started at $y_0 \in \mathrm{dom}(F)$ if*

- $(y_t)_t \in \mathrm{AC}^1_{\mathrm{loc}}(\mathbb{R}_{>0}; X)$ *and* $\lim_{t \to 0} y_t = y_0$*;*
- $(F(y_t))_t \in \mathrm{AC}^1_{\mathrm{loc}}(\mathbb{R}_{\geq 0}; \mathbb{R})$*;*
- *the following* Energy Dissipation Inequality *holds:*

$$-\frac{\mathrm{d}}{\mathrm{d}t}F(y_t) \geq \frac{1}{2}|\dot{y}_t|^2 + \frac{1}{2}|\mathrm{D}^- F|(y_t)^2 \quad \textit{for a.e. } t.$$

Evolution variational inequality To motivate the second definition, suppose again to start with that $F\colon \mathbb{R}^d \to \mathbb{R}$ is smooth, and assume in addition that F is λ-convex for some $\lambda \in \mathbb{R}$. This means that we have the uniform lower bound $\nabla^2 F \succcurlyeq \lambda I_d$ for the Hessian of F, or equivalently, that for all $x, y \in \mathbb{R}^d$ and $0 \le t \le 1$

$$F(tx + (1-t)y) \le tF(x) + (1-t)F(y) - \lambda \frac{t(1-t)}{2}|x-y|^2. \tag{1.1.3}$$

Then, the vector $\nabla F(x)$ can be characterized as the unique $v \in \mathbb{R}^d$ such that

$$F(y) \ge F(x) + \langle v, y - x\rangle + \frac{\lambda}{2}|x-y|^2 \qquad \text{for all } y \in \mathbb{R}^d.$$

For any point $x \in \mathbb{R}^d$, an easy computation using the above shows that a classical gradient flow (1.1.1) satisfies the inequality

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}|y_t - x|^2 \le F(x) - F(y_t) - \frac{\lambda}{2}|y_t - x|^2.$$

The idea is then to require the above to hold for every x as an alternative definition of gradient flow. Before stating the corresponding definition for our functional $F\colon X \to \mathbb{R} \cup \{+\infty\}$, let us discuss how to extend the notion of λ-convexity. Notice that (1.1.3) considers a linear interpolation $t \to tx + (1-t)y$ between the points x and y in the left-hand side, i.e. it connects the points x and y with a straight line. This does not make sense in a general metric space (X, d), but a natural way to mimic this construction is to make the additional assumptions that the space (X, d) is geodesic. Recall that a curve $(\gamma_t)_{t\in[0,1]}$ in X is a constant speed geodesic if and only if

$$\mathrm{d}(\gamma_s, \gamma_t) = |t-s|\mathrm{d}(\gamma_0, \gamma_1) \qquad \text{for all } s, t \in [0,1].$$

The space (X, d) is said to be geodesic if for every points $x, y \in X$ there exists a constant speed geodesic $(\gamma_t)_{t\in[0,1]}$ connecting x and y, i.e. such that $\gamma_0 = x, \gamma_1 = y$. The following definition is then natural.

Definition 1.1.2. *Let (X, d) be a geodesic space. A functional $F\colon X \to \mathbb{R}_{\ge 0} \cup \{+\infty\}$ is λ-geodesically convex if and only if for all $x, y \in X$ there exists a constant speed geodesic $(\gamma_t)_{t\in[0,1]}$ between x and y such that for all $t \in [0,1]$*

$$F(\gamma_t) \le tF(x) + (1-t)F(y) - \frac{\lambda}{2}t(1-t)\mathrm{d}(x,y)^2.$$

We can now move to the definition of gradient flow for a λ-geodesically convex functional. Actually, in the next definition, it is not strictly necessary to assume that F is λ-geodesically convex, but doing so is helpful to prove existence of a solution (in fact, it is also essentially necessary to have existence from every starting point $y_0 \in \mathrm{dom}(F)$, as shown in [DS08]).

Definition 1.1.3. *Let $F\colon X \to \mathbb{R}_{\ge 0} \cup \{+\infty\}$ be proper and lower semicontinuous. We say that a curve $(y_t)_t$ is an EVI_λ-gradient flow started at $y_0 \in \mathrm{dom}(F)$ if*

- $(y_t)_t \in \mathrm{AC}^2_{\mathrm{loc}}(\mathbb{R}_{>0}; X)$ *and* $\lim_{t\to 0} y_t = y_0$*;*
- *the following* evolution variational inequality *holds for all $x \in \mathrm{dom}(F)$:*

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{d}^2(y_t, x)^2 \le F(x) - F(y_t) - \frac{\lambda}{2}\mathrm{d}^2(y_t, t) \quad \textit{for a.e. } t.$$

It was proved by Savaré that, under our assumptions, a solution in the above sense is also a curve of maximal slope (see also [San17, AG13]),while the converse is false [San17]. However, EVI_λ-gradient flows enjoy many useful properties, e.g. in terms of uniqueness and stability of the solutions (see [AGS08] for a detailed study).

Gradient flows in the space of probability measures

As anticipated, the relevance to this thesis of the topic of gradient flows in metric spaces mainly concerns the case where X is (a subset of) the space of probability measures $\mathcal{P}(\Omega)$ on a set Ω. Indeed, given a stochastic process $(X_t)_t$ on Ω and denoting by $\mu_t := \mathrm{law}(X_t)$ its marginal law, it turns out that in some cases we can equip $\mathcal{P}(\Omega)$ with an appropriate distance d and consider a functional $F\colon \mathcal{P}(\Omega) \to \mathbb{R}_{\geq 0}$ such that $(\mu_t)_t$ becomes a gradient flow for F in the metric space $(\mathcal{P}(\Omega), \mathrm{d})$. This point of view provides useful insights for the study of the properties of the stochastic process $(X_t)_t$: in particular, the convergence to equilibrium in law can be interpreted as the convergence of a gradient flow to the unique minimizer of its driving functional F.

Langevin dynamics

The first remarkable example that we consider is given by the *Langevin dynamics*. Consider a potential $V\colon \mathbb{R}^d \to \mathbb{R}$, and let us assume that it is smooth and such that ∇V is globally Lipschitz and $\int \exp(-V(x))\,\mathrm{d}x < \infty$. We can consider a probability density $\pi \propto \exp(-V)$ associated with this potential, where the proportionality symbol $\propto$ means that $\pi = \frac{1}{Z}\exp(-V) \in \mathcal{P}(\mathbb{R}^d)$ for a normalizing constant $Z > 0$ (with abuse of notation, we often identify (probability) densities with the corresponding (probability) measures). The Langevin dynamics is the stochastic process obtained as a solution of the following stochastic differential equation

$$X_0 \sim \mu_0, \qquad dX_t = -\nabla V(X_t)dt + \sqrt{2}dB_t, \tag{1.1.4}$$

where $(B_t)_t$ denotes the standard Brownian motion and μ_0 is a regular enough probability measure. Existence and uniqueness of the solution hold under our assumptions by the standard theory of stochastic differential equations [KS91]. Correspondingly, we have that $\mu_t = \mathrm{law}(X_t)$ solves the *Fokker–Planck equation*

$$\partial_t \mu_t - \nabla \cdot [\mu_t \nabla V] - \Delta \mu_t = 0. \tag{1.1.5}$$

From substituting $\mu_0 = \pi$, it can be checked that π is stationary, and in fact X_t converges weakly to π under mild assumptions. A remarkable variational structure for the Langevin dynamics was observed in [JKO98]. Denote by $\mathcal{P}_{2,\mathrm{ac}}\left(\mathbb{R}^d\right)$ the space of probability densities in $\mathbb{R}^d$ with finite second moment, and by W_2 the 2-Wasserstein distance, which is defined for $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ by

$$W_2(\mu, \nu) = \inf_{X,Y} \mathbb{E}\left[|X - Y|^2\right],$$

where the infimum runs over all $\mathbb{R}^d$-valued random vectors X and Y defined on the same probability space $(\Omega, \mathcal{F}, \mathcal{P})$ with $\mathrm{law}(X) = \mu$ and $\mathrm{law}(Y) = \nu$. Finally, we consider the relative entropy functional (with respect to π), which is given by

$$\mathcal{D}_{\mathrm{KL}}(\nu \,\|\, \pi) = \begin{cases} \int_{\mathbb{R}^d} \rho \log \rho \,\mathrm{d}\pi & \text{if } \nu \ll \pi \text{ with } \rho := \frac{\mathrm{d}\nu}{\mathrm{d}\pi}, \\ +\infty & \text{otherwise.} \end{cases} \tag{1.1.6}$$

Notice that this functional is non-negative and it is equal to 0 if and only if $\nu = \pi$, so that it can be interpreted as a measure of distance of ν from π. A fundamental contribution of the work [JKO98] was observing that the evolution $(\mu_t)_t$ corresponds to the gradient flow of the relative entropy functional $\mathcal{D}_{\mathrm{KL}}(\cdot \,\|\, \pi)$ in the Wasserstein space $\left(\mathcal{P}_{2,\mathsf{ac}}\left(\mathbb{R}^d\right), W_2\right)$ (in a sense to be made precise, which depends on the properties of the potential V). Additionally, Otto [Ott01] derived also a formalism known as *Otto's calculus*, which allows to *formally* interpret the 2-Wasserstein space as a Riemannian manifold, correspondingly yielding yet another description of the gradient flows in this space and of the quantities involved. Let us also mention that, besides the Fokker–Planck equation (1.1.5), many other evolution PDEs fall into the framework of Wasserstein gradient flows of appropriate functionals [AGS08].

Reversible Markov chains

Inspired by the theory of Wasserstein gradient flows, one is motivated to find analogous description of other stochastic processes. The work of Maas [Maa11] constitutes an important contribution in this direction, in a discrete setting. More precisely, let Ω be a finite space, P an irreducible stochastic matrix on Ω and $\pi \in \mathcal{P}(\Omega)$ a probability measure satisfying the detailed balanced conditions $\pi(x)P(x,y) = \pi(y)P(y,x)$ for all $x, y \in \Omega$. We can consider the associated continuous time Markov chain with generator $L = P - I$ and semigroup $P_t = \exp(tL)$, which is reversible (and ergodic) with respect to π. A major contribution of [Maa11] was the construction of a metric $\mathscr{W}$ on $\mathcal{P}(\Omega)$ which plays the role of the Wasserstein distance W_2. More precisely, this metric is geodesic on $\mathcal{P}(\Omega)$, Riemannian on the subspace $\mathcal{P}_*(\Omega)$ of probability measures with full support, and, most importantly, such that the evolution $(\mu_t)_t \coloneqq (\mu P_t)_t$ becomes the gradient flow for the relative entropy functional $\mathcal{D}_{\mathrm{KL}}(\cdot \,\|\, \pi)$ in the space $(\mathcal{P}(\Omega), \mathscr{W})$.

Convergence of gradient flows

As discussed in the previous subsection, both the Langevin dynamics and the evolution of a reversible Markov chain can be interpreted as gradient flows in the space of probability measures, equipped with an appropriate distance. In both cases, the driving functional is the relative entropy $\mathcal{D}_{\mathrm{KL}}(\cdot \,\|\, \pi)$ with respect to a reference measure π, which constitutes the unique minimizer of this functional. Therefore, the ergodicity of these processes to π is reinterpreted as the convergence of a gradient flow to the unique minimizer of its functional. Motivated by this observation, one is naturally led to the study of sufficient conditions for convergence of gradient flows in metric spaces, ideally also with quantitative estimates on the speed of convergence. This strategy turns out to be particularly effective: studying the convergence of a deterministic curve can indeed be easier, and many existing arguments from the theory of optimization can be often adapted to this very abstract setting. In what follows, we briefly and informally discuss three conditions of this type.

Geodesic convexity

As usual, let us start with the smooth Eucliden setting. It is then well known that one of the best assumptions to prove convergence of a solution of (1.1.1) to a minimizer of F is the (strong) convexity of F. Indeed, suppose for example that $\nabla^2 F \succcurlyeq \lambda I_d$ for some $\lambda > 0$. Then, this automatically guarantees that F admits a unique minimizer x^*. Moreover, for two gradient flows $(y_t)_t$, $(\tilde{y}_t)$ with different initialization, an easy computation shows that $|y_t - \tilde{y}_t|$

decays exponentially fast with time, which in turns implies fast convergence of any gradient flow to x^* in a rather strong sense.

Remarkably, similar conclusions hold in abstract geodesics spaces, when the functional F is λ-geodesically convex for positive λ and the gradient flows are intended in the EVI_λ sense of Definition 1.1.3. Because of these strong consequences, it is then natural to investigate when geodesic convexity of the relative entropy holds for the examples discussed before.

For the Langevin dynamics, there is a simple characterization due to McCann [McC97]: the relative entropy functional $\mathcal{D}_{\mathrm{KL}}(\cdot \,\|\, \pi)$ with respect to a probability density $\pi \propto \exp(-V)$ is λ-geodesically convex in $\mathcal{P}_2\Big(\mathbb{R}^d, W_2\Big)$ if and only if the potential V is λ-convex (in this case, we also say that π is λ-log-concave). This result combined with the current line of reasoning allows to elegantly establish several important consequences of the strong convexity of V, such as the exponential decay of the relative entropy $\mathcal{D}_{\mathrm{KL}}(\mu_t \,\|\, \pi)$ along the Langevin dynamics, or functional inequalities for the invariant measure π.

For Markov chains, the situation is more complicated. The geodesic convexity of the relative entropy in $(\mathcal{P}(\Omega), \mathscr{W})$ was studied in [EM12] (see also [Mie13]): inspired by the Lott–Sturm–Villani theory, Erbar and Maas defined the (entropic) Ricci curvature of a reversible Markov chain in terms of the geodesic convexity of this functional. Moreover, they showed that under λ-geodesic convexity the evolution $(\mu_t)_t$ of the Markov chain is an EVI_λ-gradient flow for the relative entropy, and that, when λ is positive, many powerful inequalities hold in analogy with the Langevin diffusion. Equivalent characterizations of the λ-geodesic convexity are also provided, but a simple criterion in the spirit of the log-concavity of π is missing, and in practice computing the entropic Ricci curvature of a Markov chain can be quite challenging.

Polyak–Łojasiewicz condition

Next, we consider a simple but important criterion from optimization, known as the *Polyak–Łojasiewicz condition* (PL) [Pol63]. For smooth $F\colon \mathbb{R}^d \to \mathbb{R}_{\geq 0}$, it takes the form of a *gradient domination* inequality

$$|\nabla F|^2 \geq 2KF \tag{1.1.7}$$

for some $K > 0$. It can be checked that this is weaker than the convexity lower bound $\nabla^2 F \succcurlyeq K I_d$, but still yields powerful consequences. For example, since along a solution of (1.1.1) we have $\frac{\mathrm{d}}{\mathrm{d}t} F(y_t) = -|\nabla F(y_t)|^2 \leq -2KF(y_t)$, by Grönwall's lemma it follows that $F(y_t) \leq \exp(-2KT) F(y_0)$; hence, $F(y_t)$ converges to the minimum value of F exponentially fast (recall that F is non-negative). In addition, the minimum of F is attained at some point x^* such that $|y_t - x^*| \leq \sqrt{\frac{2F(y_t)}{K}}$.

The PL condition makes perfect sense also for curves of maximal slope as in Definition 1.1.1, simply replacing the norm of the gradient with the descending slope (1.1.2), and it still implies exponential convergence to a minimizer. For the Langevin dynamics and the case of reversible Markov chains, the condition (1.1.7) corresponds to an important functional inequality, which yields the exponential decay of the relative entropy along the corresponding evolution $(\mu_t)_t$. In the continuous setting, this is the celebrate log-Sobolev inequality [BGL14]

$$2K\mathcal{D}_{\mathrm{KL}}(\cdot \,\|\, \pi) \leq \mathcal{I}_2(\cdot \,\|\, \pi), \tag{1.1.8}$$

where the right hand side is the relative Fisher information, defined for an absolutely continuous (with respect to π) probability measure $d\mu = \rho d\pi$ by

$$\mathcal{I}_2(\mu \,\|\, \pi) = \int_{\mathbb{R}^d} |\nabla \log \rho|^2 d\mu.$$

For reversible Markov chains, the inequality is called modified log-Sobolev inequality [BT06] (to distinguish it from similar versions that are not equivalent in discrete spaces): it is defined analogously to (1.1.8), with the difference that the Fisher information is now given by

$$\mathcal{I}_2(\mu \,\|\, \pi) = \frac{1}{2} \sum_x \pi(x) P(x,y)(\rho(x) - \rho(y))(\log \rho(x) - \log \rho(y)).$$

Coherently with the smooth Euclidean case, the (modified) log Sobolev inequality is implied by the strong geodesic convexity of the relative entropy. This fact specialized to the Langevin dynamics (where the geodesic convexity is equivalent to the log-concavity of π) recovers in particular the celebrated result by Bakry and Émery [BE85], which also gains a new interpretation.

Local conditions

Finally, we consider a third condition. This constitutes the first contribution of this thesis, which is based on joint work with Lorenzo Dello Schiavo and Jan Maas [DSMP24], and is presented in full detail in Chapter 2.

The general motivation was to find weaker conditions that imply convergence of gradient flows to minimizers with quantitative rates. In fact, while the (PL) condition is weaker than strong convexity, it is still quite demanding, by requiring an inequality to hold globally, i.e. for all points in the space. A remarkable *local condition* was recently introduced by Chatterjee [Cha22] and Oymak and Soltanolkotabi [OS19], for a solution to the Cauchy problem (1.1.1) and a smooth non-negative functional F. The main assumptions of these works is still a gradient domination condition as in (1.1.7), but which is required to hold only on a big enough ball around the starting point x_0:

$$|\nabla F(x)|^2 \geq 2KF(x) \qquad \text{for all } x \in B_r(x_0). \tag{1.1.9}$$

Remarkably, these authors showed that provided that $r^2 > \frac{2F(x_0)}{K}$, we automatically get the main conclusions of the global (PL) inequality: there exists a minimizer $x^* \in B_r(x_0)$ such that $F(x^*) = 0$ and $\frac{K}{2}|y_t - x^*|^2 \leq F(y_t) \leq \exp(-2Kt)F(x_0)$.

Inspired by this result, the work [DSMP24] deals with generalization of the above criterion. Our main theorem in continuous time does so by (i) dealing with curves of maximal slope in complete metric spaces and (ii) providing more general local *Kurdyka–Łojasiewicz inequalities*, which include the (PL) inequality as a particular case. We state below one instance of the results of [DSMP24].

Theorem 1.1.4. *Let (X, d) be a complete metric space and $F\colon \mathbb{R}^d \to \mathbb{R}_{\geq 0} \cup \{+\infty\}$ be lower semicontinuous. Let also $\theta\colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be continuous, continuously differentiable on $\mathbb{R}_{>0}$ and such that $\theta(0) = 0$, $\theta'(t) > 0$ for $t > 0$. Finally, suppose that for some $x_0 \in \mathrm{dom}(F)$ the following holds:*

$$(\theta' \circ f)(x) \cdot |\mathrm{D}^- f|(x) \geq 1 \quad \text{for all} \quad x \in B_{(\theta \circ F)(x_0)}(x_0). \tag{1.1.10}$$

Then, there exists at least one minimizer $x^* \in \overline{B_{(\theta \circ F)(x_0)}(x_0)}$ such that $F(x^*) = 0$. Moreover, every curve of maximal slope $(y_t)_t$ started at x_0 converges to a minimizer of F in $\overline{B_{(\theta \circ F)(x_0)}(x_0)}$ with explicit rates.

For the proof and more general and precise results we refer the reader to Chapter 2. Let us notice that the above Theorem includes the local (PL) condition as a particular case by choosing $\theta(t) = \sqrt{\frac{2t}{K}}$.

1.2 Curvature of Markov chains

The second contibution [Ped23] of this thesis deals with curvature of Markov chains, and is treated in detail in Chapter 3. More precisely, we consider reversible Markov chains on a finite space Ω with generator L, transition semigroup P_t and invariant measure π. For a measure $\mu \in \mathcal{P}(\Omega)$, we denote by $\mu_t = \mu P_t$ the law of the Markov chain at time t which is initially distributed accorded to μ. We often think of $L = \lambda(P - I)$ for a stochastic matrix P and some $\lambda > 0$, so that there is also a naturally associated Markov chain in discrete time. We also consider the graph distance d on Ω obtained by drawing an edge between two points $x \neq y$ if and only if $L(x, y) > 0$, in which case we write $x \sim y$.

Because of the success of the Bakry–Émery and Lott–Sturm–Villani theories, considerable effort has been put into developing a notion of Ricci curvature for discrete spaces. In spite of its generality, the definition of synthetic Ricci curvature by Lott, Sturm and Villani does not apply in this setting: in fact, as observed in [Maa11, EM12], for the 2-Wasserstein distance the space $\mathcal{P}(\Omega)$ is not geodesic, and moreover the curve $(\mu_t)_t$ does not correspond to a Wasserstein gradient flow. As anticipated, a good candidate notion is the *entropic Ricci curvature* of [EM12], obtained by replacing the 2-Wasserstein distance with the modified metric $\mathscr{W}$ of [Maa11] in the geodesic convexity requirement. This definition is a meaningful one: many non-trivial Markov chains are positively curved, and at the same time positive curvature is related to fast mixing of the Markov chain and good concentration properties of the invariant measure π. Unfortunately, in many situations it is not easy to provide good estimates for the entropic curvature of a Markov chains. The geodesic convexity of the relative entropy in $(\mathcal{P}(\Omega), \mathscr{W})$, in fact, corresponds to a "Riemannian Hessian lower bound", which requires to establish an involved inequality uniformly over all pair (ρ, ψ) of densities and functions on Ω [EM12, Thm. 4.5].

Besides the entropic curvature, several other notions of discrete curvature have been introduced, by adapting different equivalent characterizations valid for Riemannian manifolds. For Markov chains, these definitions are typically not easily comparable, one notable obstacle being the lack of a chain rule in discrete spaces. One of the most important notions is the *coarse Ricci curvature* Ric_c introduced by Ollivier [Oll09], which involves the graph distance d on Ω and the corresponding 1-Wasserstein metric on $\mathcal{P}(\Omega)$. The coarse Ricci curvature was initially defined for discrete time Markov chains; for a stochastic matrix P, we say that the curvature lower bound $\mathrm{Ric}_\mathrm{c} \geq K$ holds if and only if for every $\mu, \nu \in \mathcal{P}\left(\mathbb{R}^d\right)$ and integers $n > 0$ we have the contraction

$$W_1(\mu P^n, \nu P^n) \leq (1 - K)^n W_1(\mu, \nu). \tag{1.2.1}$$

In continuous time, we similarly require that for all times $t \geq 0$

$$W_1(\mu P_t, \nu P_t) \leq \exp(-Kt) W_1(\mu, \nu).$$

It is not difficult to see that, if $K > 0$, the above inequalities imply upper bounds on the mixing time of the Markov chain; moreover, some functional inequalities for the invariance measure also follow, e.g. Poincaré inequality. Unfortunately, this does not include the more powerful modified log-Sobolev inequality, as Münch recently showed [Mün23], disproving a conjecture of Peres and Tetali. One reason for the popularity of this notion of curvature, on the other hand, lies in its simplicity, which makes it easy to estimate it accurately. For example, it suffices to check the condition (1.2.1) only for $n = 1$ and $x \sim y$.

The requirement (1.2.1) also amounts to exhibiting a coupling between the laws of the Markov chains with different starting points, which contracts the distance on average. The construction of couplings that are "contractive" in some sense has frequently appeared before and in different forms in the study of mixing times of Markov chains; these methods are often referred to as *probabilistic* approaches. Ollivier's positive curvature corresponds in particular to the *Dobrushin criterion* [Dob70, DS85]. Very recently, Conforti [Con22] introduced another variation of these methods, based on *contractive coupling rates*. Remarkably, Conforti showed on several examples that these coupling rates can be used to establish some functional inequalities for the Markov chain (including the modified log-Sobolev inequality), thus providing an intriguing connection between the probabilistic approach and more analytic methods based on functional inequalities.

Since the modified log-Sobolev inequality is implied in particular by positive entropic curvature, it is natural to wonder whether contractive coupling rates can also be used to establish such curvature lower bounds. Moreover, because of the strict relationship between contractive couplings and Ollivier's work, this would also extend the probabilistic-analytic connection at the level of discrete curvature, suggesting possible links between the coarse and the entropic notions.

Our contribution [Ped23] is aimed at this. First, we develop a general strategy to estimate the entropic curvature: coupling rates are used to provide an alternative expression and a simpler lower bound for the Riemannian Hessian of the relative entropy functional; subsequently, it is explained heuristically how contractions in the couplings further simplify this expression. Then, we illustrate these ideas in most of the examples discussed in [Con22]: in particular, we provide new estimates for the entropic curvature of the Ising, Curie–Weiss and the hardcore models, and for an interactive random walk on the discrete grid $\mathbb{N}^d$.

1.3 Score-based diffusion models

Stochastic processes are useful tools in applications to data science. For example, it is often desirable to *sample* efficiently from a target distribution π, i.e. to generate random points $(X_i)_i$ that are (approximately) independently and identically distributed according to π. The third contribution of this thesis [PMM24] provides a theoretical analysis of an algorithm for sampling which belongs to the recent framework of score-based generative modeling. This work is in collaboration with Jan Maas and Marco Mondelli and it is included in Chapter 4

A popular way to sample from a distribution π is to design a stochastic process $(X_t)_t$ that can be simulated efficiently and that converges in law to π, so that by running it for a long enough time T the output X_T is approximately distributed according to the desired distribution. Of course, to understand the performance of the algorithm, one is led to the study of the speed of convergence to equilibrium of this process.

A fundamental choice in the theory of sampling is given by the Langevin diffusion (1.1.4). One reason for its popularity lies in the fact that the dynamics (1.1.4) only requires knowledge of the *score function* $-\nabla V = \nabla \log \pi$ of the data distribution. Modeling the score function is typically easier than modeling a probability density $\pi \propto \exp(-V)$, because the latter typically requires to estimate the normalizing constant $\int_{\mathbb{R}^d} \exp(-V)$, a notoriously difficult task in high-dimensional spaces. Unfortunately, for complex distributions π, the Langevin dynamics suffers from a few drawbacks. One of this is related to the speed of convergence to equilibrium, which can be very slow in the absence of appropriate functional inequalities for π (which are often not satisfied by the distributions encountered in practice).

These observations about the Langevin dynamics have motivated the search for other algorithms for sampling, which retain its advantages (i.e. being based on score functions) while solving some of its problems, e.g. the reliance on strong functional inequalities for fast convergence. In the last years, a remarkable success has been achieved by *score-based diffusion models* [SE19, SDWMG15, HJA20, SGSE20], which we now introduce in a simplified particular form.

We are interested in the setting where the target distribution π is unknown, but we are given a collection of samples $(x_i)_i \overset{\text{iid}}{\sim} p$ independent and distributed according to π. Let us start by looking at the *Ornstein–Uhlenbeck (OU) flow*, which is just the Langevin dynamics targeting the standard Gaussian distribution γ, initially distributed according to π:

$$X_0 \sim \pi, \qquad dX_t = -X_t dt + \sqrt{2} dB_t. \tag{1.3.1}$$

Since the standard Gaussian density γ is 1-log-concave, this flow converges to γ exponentially fast in various metrics (e.g. relative entropy and Wasserstein distance). An interesting fact, dating back to Anderson [And82], is that this flow can be reversed. Fix a time horizon $T_1 > 0$ and a number $M \geq 1$, denote $\pi_t = \operatorname{law}(X_t)$ and consider the process

$$U_0 \sim \pi_{T_1}, \qquad dU_t = U_t dt + M \nabla \log \pi_{T_1 - t}(U_t)\, dt + \sqrt{2(M-1)}\, dB_t. \tag{1.3.2}$$

Classical choices are $M = 2$ or $M = 1$, and for the latter the process $(U_t)_t$ is deterministic (except for its initialization). A remarkable feature of this process is that $U_{T_1} \sim \pi$: therefore, we could simulate it until time T_1 to generate samples from π. Two obstacles arise: first, to initialize the process we need to sample from π_{T_1}, which is unknown; as a solution, the framework of score-based diffusion models suggests to take T_1 big enough and sample from γ instead, the error being small because of the ergodicity of the Ornstein–Uhlenbeck flow. Secondly, the reverse process (1.3.2) involves the vector field $\nabla \log \pi_{T_1 - t}$, which is also unknown: however, since these are score functions, they can be efficiently estimated thanks to techniques known as *score-matching* [SE19, SE20], producing a good estimator $s_\theta(t, x) \approx \nabla \log \pi_t(x)$ to be plugged in (1.3.2). More precisely, common and realistic assumptions are a control on the weighted L^2-error of the score approximation, for example that the loss

$$\mathbb{E}_{\pi_t}\Big[|\nabla \log \pi_t - s_\theta(t, \cdot)|^2 \Big]$$

is small for all $t \in [0, T_1]$.

Because of the impressive empirical performance of this algorithm, several works were aimed at providing theoretical guarantees for its success [Bor22, CCL$^+$23b, CLL22, CDS23, BDBDD23]. Denoting by π_θ the output distribution of the algorithm, the goal is to provide an upper bound on the error in the approximation $\pi \approx \pi_\theta$, which can be measured with different metrics e.g. Wasserstein distance, relative entropy, total variation. Three sources of errors are given by: (i) starting the reverse process (1.3.2) at the Gaussian distribution γ instead of π_{T_1}; (ii) using an estimator $s_\theta(T_1 - t, \cdot)$ for the score functions $\nabla \log \pi_{T_1 - t}$; (iii) (possibly) considering a numerical scheme to simulate the reverse process. Our paper [PMM24] belongs to this line of work: a novelty in our approach lies in the study of a slightly modified algorithm, aimed at addressing differently the error in the approximation $\pi_{T_1} \approx \gamma$, and which can be seen as an instance of the popular predictor-corrector methods [SSDK$^+$21]. In fact, reducing this source of error normally requires to take T_1 larger and larger: this is undesirable, because it brings the need to approximate the score function on a larger time interval $[0, T_1]$ and potentially increases the error propagation with the simulation of the reverse process. Instead, we consider a two-stage algorithm: we first fix T_1 large enough; then, we sample from π_{T_1} by running an approximate Langevin dynamics (1.3.3) for some time $T_2 > 0$. Finally, we start the usual approximate reverse process (1.3.4) from the output of the Langevin dynamics (instead than from a Gaussian sample): the random point Y_{T_1} constitutes the output of the algorithm.

$$Z_0 \sim \gamma, \quad dZ_t = s_\theta(T_1, Z_t)\,dt + \sqrt{2}\,dB_t, \qquad 0 \le t \le T_2; \quad (1.3.3)$$

$$Y_0 = Z_{T_2}, \quad dY_t = Y_t\,dt + M s_\theta(T_1 - t, Y_t)\,dt, + \sqrt{2(M-1)}dB_t \qquad 0 \le t \le T_1. \quad (1.3.4)$$

Figure 1.1 provides a schematic representation of the method.

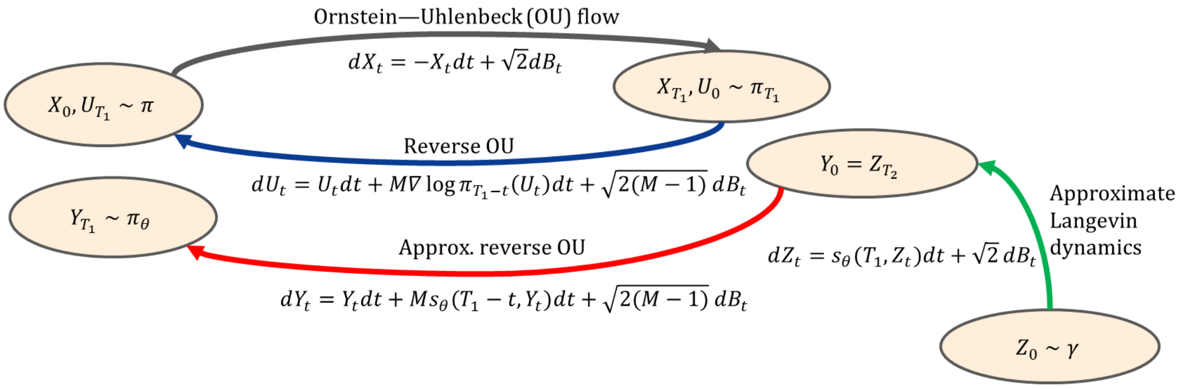


Figure 1.1: A two-stage score-based method for sampling

The work [PMM24] contains some convergence results for this algorithm. These are in the form of upper bounds for the error in the approximation $\pi \approx \pi_\theta$ which become arbitrarily small provided that T_1 is fixed but big enough, $T_2 \to \infty$ and the error in the score approximation and the step size in the discrete scheme vanish. We refer to Chapter 4 for details. Let us just briefly mention two key ideas in deriving these results, regarding in particular the initial step (1.3.3) of the algorithm. In fact, this involves a Langevin dynamics, which might seem counterintuitive, given that one motivation for the design of the diffusion models was to overcome its slow convergence. The crucial point here is that the Langevin dynamics is targeting not the complex distribution π, but its perturbed version π_{T_1}: the smoothing and ergodicity properties of the Ornstein–Uhlenbeck flow make π_{T_1} "more similar" to γ, so that it inherits some of its properties, including the log-Sobolev inequality, which in turns imply the desired exponential convergence of (1.3.3). A second obstacle arises from the fact that (1.3.3) is using an approximation of the score function $\nabla \log \pi_{T_1}$, which we can assume to be accurate only in L^2-sense. Unfortunately, this is in general not enough to ensure that in the limit where $T_2 \to \infty$ the random variable Z_{T_2} is approximately distributed according to π_{T_1}. To overcome this obstacle, we show how to convert control of the L^2-error into control of a stronger norm. This exploits again the ergodicity of the Ornstein–Uhlenbeck process, to provide a priori estimate for the true score function $\nabla \log \pi_{T_1}$ in comparison to the limiting score $\nabla \log \gamma(x) = -x$. These estimates are crucially used to correct wrong predictions from the estimator $s_\theta(T_1, \cdot)$ that are very far from the correct value, allowing to improve its accuracy.

1.4 Log-concavity along the heat flow and Lipschitz transport maps

In the previous section, the Ornstein–Uhlenbeck flow played a key role, both for its regularization properties and for providing a natural interpolation between a measure π and the Gaussian distribution γ. Together with Giovanni Brigati, we investigated related questions in the work [BP24], which can be found in Chapter 5.

Log-concavity along the heat flow. Consider again the Ornstein–Uhlenbeck process (1.3.1), and recall that we denote $\pi_t = \mathrm{law}(X_t)$. Crucial to the analysis of [PMM24] were the observations that, after some big enough time T_1, (i) the distribution π_{T_1} satisfies a log-Sobolev inequality [CCNW21], and that (ii) the score function $\nabla \log \pi_{T_1}$ is reasonably close to the one of the Gaussian distribution γ. It is natural to wonder, and useful for applications, whether stronger conclusions holds, for example whether π_{T_1} eventually becomes strongly log-concave, corresponding to a control on the log-Hessian $\nabla^2 \log \pi_t$. The first result of [BP24] shows that, in general, this is not the case. More precisely, we construct a subgaussian distribution π (or with even thinner tails) such that the Hessian $\nabla^2 \log \pi_t$ cannot be bounded from above uniformly in space, for any time $t > 0$.

This raises the question of identifying a class of distributions for which a positive answer holds. In our second result, we consider log-Lipschitz perturbations of strongly log-concave measures, i.e. probability measures of the form $\pi = \exp(-V - H)$ where V is strongly convex and H is Lipschitz. For these measures, we provide bound on $\nabla^2 \log \pi_t$ for all times $t > 0$, which show in particular that for t big enough the measure μ_t becomes strongly log-concave.

Lipschitz transport maps. The key idea of score-based diffusion models is to revert the Ornstein–Uhlenbeck flow, so as to transform a sample from the Gaussian distribution into a sample from a target distribution π. A similar construction has appeared before also in [OV00], and was then exploited in [KM12] with a completely different motivation. More precisely, the aim of [KM12] was to identify a class of distributions π for which they can construct a map $T\colon \mathbb{R}^d \to \mathbb{R}^d$ which is Lipschitz and such that $T\#\gamma = \pi$ (this last condition means that if $X \sim \gamma$ then $T(X) \sim \pi$). This line of research started with the seminal work of Caffarelli [Caf00]: one important motivation is that if T is Lipschitz and $\pi = T\#\gamma$, then one can show that many functional inequalities satisfied by the Gaussian distribution are satisfied also by π. Further works along this line include [MS23, FMS24, CF21, Nee22, KP21].

Let us give a rough idea of the construction of [OV00, KM12] and its connections with the framework of diffusion models. Consider the process (1.3.2) with $M = 1$, and recall that $U_{T_1} \sim \pi$; moreover, by the choice of M, this evolution is deterministic, i.e. U_t is given by solving an ordinary differential equation. Denote then by $T_{T_1}(x)$ the solution U_{T_1} to the differential equation (1.3.2) with $M = 1$ and $U_0 = x$. It follows that $T_{T_1}\#\pi_{T_1} = \pi$; but then, by letting $T_1 \to \infty$ and since $\pi_{T_1} \to \gamma$, one can construct a transport map T such $T\#\gamma \to \pi$. Of course, these heuristic arguments need to be justified rigorously as in [KM12, MS23], and we refer the reader to Chapter 5 and the references therein for more details. Interestingly, the log-Hessian bounds of [BP24] allow to estimate the Lipschitz norm of this transport map when π is a log-Lipschitz perturbation of a strongly log-concave measure. More precisely, suppose that $\pi = \exp(-V - H)$ where V is α-convex and H is L-Lipschitz. Then, we prove in [BP24] that there exists a map T such that $T\#\gamma = \pi$ and T is $\frac{1}{\sqrt{\alpha}} \exp\Big(\frac{L^2}{2\alpha} + \frac{2L}{\sqrt{\alpha}}\Big)$-Lipschitz.

1.5 L^∞-Transport-Information inequalities and applications to Fisher's infinitesimal model

The last included contribution [KMP24], based on joint work with Ksenia Khudiakova and Jan Maas, is presented in Chapter 6, and is quite representative of the spirit of this thesis. In fact, on the one hand it contains an approach based on analysing stochastic processes (again, Langevin dynamics) to prove abstract functional inequalities; on the other hand, it crucially exploits these new inequalities to prove convergence results for another stochastic process, Fisher's infinitesimal model from quantitative genetics.

To motivate the first point, recall that if the probability measure π is κ-log-concave, it satisfies the 2-Transport-Information inequality $W_2(\mu, \pi) \le \frac{1}{\kappa}\sqrt{\mathcal{I}_2(\mu \,\|\, \pi)}$ for all other $\mu \in \mathcal{P}\big(\mathbb{R}^d\big)$. A first natural question is whether the above inequality also holds for $p \neq 2$, and in particular $p = \infty$ is relevant for the applications to the Fisher's model. Here, for smooth densities μ, π, the L^∞-Fisher information is defined by $\mathcal{I}_\infty(\mu \,\|\, \pi) = \left\|\nabla \log \frac{d\mu}{d\pi}\right\|_\infty$. A first observation we make (see also [GTU23, CPS23]) is that the analogous inequality

$$W_\infty(\mu, \pi) \le \frac{1}{\kappa}\mathcal{I}_\infty(\mu \,\|\, \pi) \tag{1.5.1}$$

holds if π is κ-log-concave. One way to prove this is based on synchronous coupling of two Langevin diffusions, targeting π and μ respectively:

$$\begin{aligned} X_0 &\sim \mu, & dX_t &= \nabla \log \mu(X_t)dt + \sqrt{2}dB_t,\\ Y_0 &= X_0, & dY_t &= \nabla \log \pi(Y_t)dt + \sqrt{2}dB_t. \end{aligned}$$

An simple arguement shows that $|X_t - Y_t|$ is a.s. uniformly bounded, so that the desired conclusion follows by letting $t \to \infty$. One of the main contributions of [KMP24], however, is that the constant κ in (1.5.1) can be significantly improved in the presence of anisotropy, both in the potential $-\nabla \log \pi$ and in the log-relative density $\log \frac{d\mu}{d\pi}$. The proof is based on the same synchronous coupling, but requires a more careful analysis to provide tighter uniform bounds on the distance $|X_t - Y_t|$.

As anticipated, the L^∞-Transport-Information inequality, and specifically the refined anisotropic version, finds application in the study of Fisher's infinitesimal model. We refer the reader to Chapter 6 and references therein for a more detailed discussion; here, we just focus on the connection with the Transport-Information inequality. In short, the setting is as follows. We are given a strongly log-concave equilibrium density $\mathbf{F} \in \mathcal{P}\big(\mathbb{R}^d\big)$; for an initial distribution $F_0 = u_0 F$ with $\|\nabla u_0\|_\infty < \infty$ (i.e. $\mathcal{I}_\infty(F_0 \,\|\, \mathbf{F}) < \infty$), we consider the discrete-time dynamics $F_n = u_n \mathbf{F}$ defined recursively by

$$u_n(x) \propto \int_{\mathbb{R}^d \times \mathbb{R}^d} P(x_1, x_2; x) u_{n-1}(x_1) u_{n-1}(x_2) \mathrm{d}x_1 \mathrm{d}x_2,$$

for an appropriate transition kernel $P(\cdot\,; x)$ on $\mathbb{R}^d \times \mathbb{R}^d$. The non-linearity of this dynamics, which involves a backward in time integral on a higher-dimensional space, makes it difficult to apply standard techniques from the theory of Markov processes. Instead, Calvez, Poyato and Santambrogio discovered a remarkable duality relationship with the L^∞-Fisher information, which can be expressed as

$$\mathcal{I}_\infty(F_n \,\|\, \mathbf{F}) \le \sqrt{2}\mathcal{I}_\infty(F_{n-1} \,\|\, \mathbf{F}) \cdot \sup_{x \neq y} \frac{W_\infty(P(\cdot\,; x), P(\cdot\,; y))}{|x - y|}. \tag{1.5.2}$$

From the above, it is clear that a uniform bound $W_\infty(P(\cdot\,;x),P(\cdot\,;y)) < \frac{\sqrt{2}}{2}K|x-y|$ for some $0<K<1$ implies by induction that $\mathcal{I}_\infty(F_n \,\|\, \mathbf{F}) \leq K^n \mathcal{I}_\infty(F_0 \,\|\, \mathbf{F})$, and so the dynamics converges exponentially fast to equilibrium in L^∞-Fisher information. While the naive bound (1.5.1) falls short of this goal, the refined anisotropic inequalities in [KMP24] serve the purpose, providing sharp estimates in this setting.

CHAPTER 2

Local Conditions for Global Convergence of Gradient Flows and Proximal Point Sequences in Metric Spaces

This chapter corresponds to the publication [DSMP24].

This paper[1] deals with local criteria for the convergence to a global minimiser for gradient flow trajectories and their discretisations. To obtain quantitative estimates on the speed of convergence, we consider variations on the classical Kurdyka–Łojasiewicz inequality for a large class of parameter functions. Our assumptions are given in terms of the initial data, without any reference to an equilibrium point. The main results are convergence statements for gradient flow curves and proximal point sequences to a global minimiser, together with sharp quantitative estimates on the speed of convergence. These convergence results apply in the general setting of lower semicontinuous functionals on complete metric spaces, generalising recent results for smooth functionals on $\mathbb{R}^n$. While the non-smooth setting covers very general spaces, it is also useful for (non)-smooth functionals on $\mathbb{R}^n$.

2.1 Introduction

For given $x_0 \in \mathbb{R}^n$ and $f \in \mathrm{C}^2(\mathbb{R}^n)$ we consider the *gradient flow* equation

$$\frac{\mathrm{d}}{\mathrm{d}t} y_t = -\nabla f(y_t)\,, \qquad y_0 = x_0\,. \tag{2.1.1}$$

It is of great interest in many applications to find conditions which guarantee convergence of gradient-flow trajectories $(y_t)_{t\geq 0}$ to a global minimizer of f as $t \to \infty$, and to quantify the speed of convergence. This also applies to the associated discrete-time schemes, such as *gradient descent* (or *forward Euler*), the discrete-time scheme with step-size $\tau > 0$ given by

$$y_{k+1} = y_k - \tau \nabla f(y_k)\,, \qquad y_0 = x_0\,, \tag{2.1.2}$$

and the *backward Euler* scheme

$$y_{k+1} = y_k - \tau \nabla f(y_{k+1})\,, \qquad y_0 = x_0\,. \tag{2.1.3}$$

[1]First published in Transactions of the American Mathematical Society 377(6) in 2024, published by the American Mathematical Society.

The Polyak–Łojasiewicz condition

A very simple celebrated criterion for convergence to a global minimum is the *Polyak–Łojasiewicz condition* [Pol63], which requires neither the uniqueness of a minimizer nor the convexity of the function f. The condition holds if, for some $\beta > 0$,

$$|\nabla f(x)|^2 \geq \beta(f(x) - f^\star)\,, \qquad x \in \mathbb{R}^n\,, \tag{PŁ}$$

where $f^\star$ is the global minimum of f, which is assumed to be attained. Since $\frac{\mathrm{d}}{\mathrm{d}t}f(y_t) = -|\nabla f(y_t)|^2$ along any solution y_t to the gradient-flow equation $\frac{\mathrm{d}}{\mathrm{d}t}y_t = -\nabla f(y_t)$, an application of Gronwall's inequality yields the exponential bound

$$f(y_t) - f^\star \leq e^{-\beta t}\Big(f(y_0) - f^\star\Big)\,, \qquad t \geq 0\,.$$

Moreover, a short argument shows that y_t converges to a global minimizer $x^\star$, with the bound

$$|y_t - x^\star|^2 \leq \frac{4}{\beta}\Big(f(y_t) - f^\star\Big)\,, \qquad t \geq 0\,.$$

These inequalities together yield exponentially fast convergence to $x^\star$. Analogous results hold for the associated gradient-descent scheme (2.1.2) and for certain proximal-gradient methods [KNS16]. Interestingly, in spite of its simplicity, it has been argued in [KNS16] that the PŁ condition «*is actually weaker than the main conditions that have been explored to show linear convergence rates without strong convexity over the last 25 years.*»

The Kurdyka–Łojasiewicz condition

An important generalization of the PŁ condition is the *Kurdyka–Łojasiewicz inequality* (KŁ), that was introduced by Łojasiewicz [Łoj63, Łoj93] and later generalized by Kurdyka [Kur98].

Definition 2.1.1. *Let $\theta \in C([0,\infty)) \cap C^1((0,\infty))$ satisfy $\theta(0) = 0$ and $\theta'(u) > 0$ for $u > 0$. We say that the KŁ inequality is satisfied in a neighbourhood U of an equilibrium point $x^\star \in \mathbb{R}^n$ if*

$$\theta'(f(x) - f(x^\star)) \cdot |\nabla f(x)| \geq 1 \qquad \textit{for all } x \in U \cap \{f > f(x^\star)\}\,. \tag{2.1.4}$$

In applications, θ is often of the form $\theta(u) := \frac{c}{\gamma}u^\gamma$ with $\gamma \in (0,1]$ and $c > 0$. In this case, (2.1.4) reads as

$$c\,|\nabla f(x)| \geq \Big(f(x) - f(x^\star)\Big)^{1-\gamma}\,, \qquad x \in U \cap \{f > f(x^\star)\}\,.$$

In particular, if $\gamma = \frac{1}{2}$ and $c = 1/\sqrt{\beta}$ one recovers the PŁ inequality.

The KŁ condition is a powerful tool to obtain convergence properties for gradient-flow solutions and discrete schemes. An important feature of the KŁ condition is that the inequality is only required to hold *locally*, on a suitable neighbourhood U of an equilibrium point.

To obtain convergence results for gradient-flow trajectories to an equilibrium point, the KŁ condition is often combined with additional information, typically an upper bound on the length of the trajectory, to ensure that the solution is eventually contained in U; cf. [Sim83, HM19] for results of this type for gradient flows and [AB09, ABRS10, ABS13, BDLM10] for discrete schemes.

Let us also remark that the KŁ condition does not in general yield convergence to a global minimizer of f, but merely to a stationary point. To deduce convergence to a global minimizer, it is often required to know *a priori* that the starting point is close enough to this minimizer (whose existence is often part of the assumption); cf. [ABRS10, Thm. 10] and [ABS13, Thm. 2.12] for such results for discrete schemes.

A PŁ condition around the starting point

A remarkable variant of the PŁ condition was discussed by Oymak and Soltanolkotabi [OS19] and by Chatterjee [Cha22] for *non-negative* functions $f \in \mathrm{C}^2(\mathbb{R}^n)$. For fixed $x_0 \in \mathbb{R}^n$, these authors consider the local quantity

$$\alpha = \alpha(x_0, r) := \inf_{\substack{x \in B_r(x_0) \\ f(x) \neq 0}} \frac{|\nabla f(x)|^2}{f(x)},$$

where $B_r(x_0)$ denotes the open ball of radius $r > 0$ around x_0.[2] The criterion in [Cha22] requires that

$$\alpha(x_0, r) > \frac{4f(x_0)}{r^2} \tag{2.1.5}$$

for some $x_0 \in \mathbb{R}^n$ and some $r > 0$. In other words, the inequality $|\nabla f(x)|^2 \geq \beta f(x)$ is imposed to hold for all $x \in B_r(x_0)$, with a sufficiently large constant β, namely, $\beta > 4f(x_0)/r^2$.

Under (2.1.5), it is shown in [Cha22] that the unique gradient flow curve $(y_t)_{t\geq 0}$ starting at $y_0 = x_0$ stays within $B_r(x_0)$ for all times $t \geq 0$, converges to a global minimizer $x^\star \in \overline{B_r(x_0)}$, and satisfies the exponential bounds

$$f(y_t) \leq e^{-\alpha t} f(x_0) \quad \text{and} \quad |y_t - x^\star|^2 \leq r^2 e^{-\alpha t} \tag{2.1.6}$$

for $t \geq 0$, where we write $\alpha = \alpha(x_0, r)$ for brevity.

Like the KŁ condition, (2.1.5) is a local version of the PŁ condition. However, while the KŁ condition involves information of f in a neighbourhood of an equilibrium point (whose location is often unknown in applications), (2.1.5) is formulated in terms of the starting point x_0 of the gradient-flow trajectory. The existence of a global minimizer and the boundedness of the gradient-flow trajectory are not assumed; these statements are part of the conclusion. The specific constant $\frac{4f(x_0)}{r^2}$ in (2.1.5) is important, as it ensures that the gradient-flow curve does not leave the ball $B_r(x_0)$, so that local information suffices to draw conclusions on the long-term behaviour.

Chatterjee also proves analogous bounds for the gradient descent (2.1.2) starting at $y_0 = x_0$, namely

$$f(y_k) \leq (1-\delta)^k f(x_0) \quad \text{and} \quad |y_k - x^\star| \leq r^2 (1-\delta)^k$$

for all $k \in \mathbb{N}$ and any $\delta < \alpha\tau$, provided that the step-size $\tau > 0$ is sufficiently small, depending also on the size of the derivatives of f in $B_r(x_0)$. Similar results for gradient descent were obtained previously in [OS19]. Applications of (2.1.5) to neural networks can be found in [OS19, BPVF22, Cha22, BAM22, BMR21].

[2] In a general metric space, the closure $\overline{B_r(x_0)}$ is a subset of the closed ball $\{x \in X : \mathsf{d}(x, x_0) \leq r\}$ and the inclusion may be strict.

2.1.1 Main results

In this work we generalise some of the results of [Cha22] in the following ways. First, we replace C^2 functions on $\mathbb{R}^n$ with lower semicontinuous functionals on complete metric spaces. Secondly, we replace the local PŁ-like condition with a KŁ-like assumption with a more general parameter function θ. Thirdly, we prove convergence results for the proximal point method, which corresponds to the backward Euler scheme in smooth settings.

Let (X, d) be a complete metric space. In this generality, the ODE (2.1.1) does not have a direct interpretation, as the velocity of a curve and the gradient of a function are not defined. However (2.1.1) admits an equivalent variational characterisation, as a *curve of maximal slope*, and this notion naturally extends to metric spaces. We refer to Section 2.2 for the definition of the metric slope $|\mathrm{D}^- f|(x)$ and other concepts from analysis in metric spaces relevant to our work. Gradient flows in metric spaces are ubiquitous in applications; notable examples are dissipative PDEs in the Wasserstein space [JKO98, AGS14] and related gradient flows on spaces of (probability) measures; see, e.g., [DNS09, Maa11, Mie11, KV18]. A systematic treatment can be found in the monograph [AGS08]. The metric point of view can also be useful to deal with non-differentiable functionals on $\mathbb{R}^n$; see §2.4 for some toy examples.

We first define the functions appearing in the assumption and the main results.

Definition 2.1.2 (Parameter function). *We say that* $\theta \in C^1((0,\infty)) \cap C([0,\infty))$ *is a* parameter function *if* $\theta'(u) > 0$ *for* $u > 0$*, and* $\theta(0) = 0$*. Furthermore, we consider the auxiliary functions* $\eta : [0,\infty) \to [-\infty,\infty)$ *and* $\Gamma : [0,\theta(\infty)) \to [-\infty,\infty)$ *defined by*

$$\eta(u) := \int_1^u \big(\theta'(s)\big)^2 \, \mathrm{d}s \quad \text{and} \quad \Gamma(u) := (\eta \circ \theta^{-1})(u).$$

The next definition contains a generalisation of (2.1.5) to the metric setting for a general class of parameter functions; see Remark 2.4.1 below for a precise comparison.

Definition 2.1.3 (Conditions (A) and (A')). *For* $x_0 \in \mathrm{dom}(f)$ *and* $r > 0$*, we say that condition* (A) *is satisfied with parameter function* θ *(as in Definition 2.1.2) if*

$$(\theta \circ f)(x_0) \le r \quad \text{and} \quad (\theta' \circ f)(x) \cdot |\mathrm{D}^- f|(x) \ge 1\,, \qquad x \in B_r(x_0) \cap \{0 < f \le f(x_0)\}\,. \tag{2.1.7}$$

Similarly, we say that condition (A') *is satisfied if the first inequality in* (2.1.7) *is replaced by the strict inequality* $(\theta \circ f)(x_0) < r$*.*

Under Condition (A), our first main result asserts that gradient-flow trajectories stay within in a bounded set and converge to a global minimum, with quantitative bounds on the rate of convergence.

Theorem 2.1.4 (Convergence of gradient flows). *Let* $f : X \to [0,\infty]$ *be proper and lower semicontinuous, and suppose that* $x_0 \in \mathrm{dom}(f)$ *and* $r > 0$ *satisfy Condition* (A) *for some parameter function* θ*. For some* $T \in (0,\infty]$*, let* $(y_t)_{t\in[0,T)}$ *be a curve of maximal slope for* f *starting at* x_0*. Then:*

(*i*) (confinement) $y_t \in \overline{B_r(x_0)}$ *for all* $0 \le t < T$*. Moreover,* $y_t \in B_r(x_0)$ *for all* $0 \le t < T$ *with* $f(y_t) > 0$*.*

(ii) (convergence) $y_T := \lim_{t\to T} y_t$ *exists and belongs to* $\overline{B_r(x_0)}$. *Moreover,*

$$(\theta \circ f)(y_s) - (\theta \circ f)(y_t) \geq \mathsf{d}(y_t, y_s) \tag{2.1.8}$$

for all $0 \leq s \leq t \leq T$. *In particular,* $y_T \in B_r(x_0)$ *if* $f(y_T) > 0$.

(iii) (convergence rates) *Set* $t_* := \inf\{t \in [0,T) : f(y_t) = 0\} \wedge T$. *The following bounds hold for* $0 \leq t \leq t_*$*:*

$$\Gamma\Big(\mathsf{d}(y_t, y_T)\Big) \leq \Gamma(r) - t, \tag{2.1.9}$$

$$(\eta \circ f)(y_t) \leq (\eta \circ f)(x_0) - t. \tag{2.1.10}$$

Moreover, if $T = \infty$ *then* $f(y_\infty) = 0$.

In the special case of Remark 2.4.1, the previous theorem yields the following generalisation of [Cha22, Thm. 2.1] to the setting of metric spaces; see also Cor. 2.3.8 below for a version with more general parameter functions. For $x_0 \in \operatorname{dom}(f)$ and $r > 0$ we define

$$\alpha = \alpha(x_0, r) := \inf_{\substack{x \in B_r(x_0) \\ 0 < f(x) \leq f(x_0)}} \frac{|D^- f|(x)^2}{f(x)}.$$

Corollary 2.1.5. *Let* $f : X \to [0,\infty]$ *be proper and lower semicontinuous, and suppose that* $\alpha(x_0, r) \geq 4f(x_0)/r^2$ *for some* $x_0 \in \operatorname{dom}(f)$ *and* $r > 0$. *For some* $T \in (0,\infty]$, *let* $(y_t)_{t\in[0,T)}$ *be a curve of maximal slope for* f *starting at* x_0. *Then* $y_T := \lim_{t\to T} y_t$ *exists,* y_t *belongs to* $\overline{B_r(x_0)}$ *for all* $t \in [0,T]$, *and*

$$\mathsf{d}(y_t, y_T) \leq r\, e^{-\alpha t/2} \qquad \text{and} \qquad f(y_t) \leq e^{-\alpha t} f(x_0)$$

for all $t \in [0,T]$, *where, conventionally,* $e^{-\infty} := 0$.

Various works deal with convergence of gradient-flow trajectories under a KŁ condition in the setting of metric spaces [BB18, HM19]; see also [BDLM10, AB09, ABRS10] for related work on proximal point sequences. Applications have been found to convergence of mean-field birth-death processes [LMS22] and to swarm gradient dynamics [BMV22].

The main estimates in our paper are obtained by adapting known arguments from, e.g., [BB18, HM19]. However, as in [OS19, Cha22], our point of view differs from these works, as we work under a local condition in terms of the starting point without referring to an equilibrium point in the assumption.

Discrete schemes

In the general setting of metric spaces, we are not aware of any way to formulate a forward Euler scheme (2.1.2). However, the backward Euler scheme admits an equivalent metric formulation as a minimising movement scheme (or proximal point method). This scheme was originally introduced by Martinet [Mar70] and Rockafellar [Roc76] as a natural regularisation method in optimisation problems:

$$y_{k+1} \in \operatorname*{arg\,min}_{x\in X} \left\{ f(x) + \frac{1}{2\tau}\mathsf{d}(y_k, x)^2 \right\}, \qquad y_0 = x_0\,.$$

Any (finite or infinite) sequence $(y_k)_k$ arising in this way is called a *proximal point sequence* (or τ*-minimising movement sequence*).

Our second main result is an analogue of Theorem 2.1.4 for the proximal point method, under the slightly stronger assumption (A').

Theorem 2.1.6. *Let $f : X \to [0, \infty]$ be proper and lower semicontinuous, and suppose that $x_0 \in \operatorname{dom}(f)$ and $r > 0$ satisfy Condition (A') for some parameter function θ. Suppose further that there exists $\bar{\tau} > 0$ such that, for all $x \in B_r(x_0) \cap \{f \le f(x_0)\}$ and $\tau \in (0, \bar{\tau})$, the functional*

$$X \ni y \longmapsto f(x) + \frac{1}{2\tau}\mathsf{d}(x,y)^2$$

has at least one global minimizer. Then there exists an infinite proximal point sequence starting from x_0, for any step-size $\tau < \bar{\tau}$. Moreover, for any such sequence $(y_k)_{k=0}^{\infty}$, the following statements hold:

(i) (confinement) $y_k \in B_r(x_0)$ *for all* $k \ge 0$;

(ii) (convergence) $y_\infty := \lim_{k\to\infty} y_k$ *exists and belongs to* $B_r(x_0)$. *Moreover,* $f(y_\infty) = 0$;

(iii) (distance bound) *For all* $0 \le i \le j < \infty$ *we have*

$$\mathsf{d}(y_i, y_j) \le (\theta \circ f)(y_i) - (\theta \circ f)(y_j) \quad \text{and} \quad \mathsf{d}(y_i, y_\infty) \le (\theta \circ f)(y_i)\,. \tag{2.1.11}$$

In the particular case where θ takes the form $\theta(u) := 2c\sqrt{u}$, we obtain the following result. In this case we also obtain an estimate for the speed of convergence of $f(y_k)$ to 0. Other special cases of Theorem 2.1.6 are presented in Corollary 2.6.7 below.

Corollary 2.1.7 (see Cor. 2.6.7)**.** *Let $f : X \to [0, \infty]$ be proper and lower semicontinuous, and suppose that $\alpha(x_0, r) > 4f(x_0)/r^2$ for some $x_0 \in \operatorname{dom}(f)$ and $r > 0$. Suppose further that there exists $\bar{\tau} > 0$ such that, for all $x \in B_r(x_0) \cap \{f \le f(x_0)\}$ and $\tau \in (0, \bar{\tau})$, the functional*

$$X \ni y \longmapsto f(x) + \frac{1}{2\tau}\mathsf{d}(x,y)^2$$

has at least one global minimizer. Then there exists an infinite proximal point sequence starting from x_0, for any step-size $\tau < \bar{\tau}$. Moreover, for any such sequence, the following statements hold:

(i) (confinement) $y_k \in B_r(x_0)$ *for all* $k \ge 0$;

(ii) (convergence) $y_\infty := \lim_{k\to\infty} y_k$ *exists and belongs to* $B_r(x_0)$. *Moreover,* $f(y_\infty) = 0$;

(iii) (convergence rates) *The following bounds hold for all* $k \ge 0$:

$$f(y_k) \le \Big(1 + \alpha\tau\Big)^{-k} f(x_0) \quad \text{and} \quad \mathsf{d}(y_k, y_\infty) \le \Big(1 + \alpha\tau\Big)^{-k/2} r\,.$$

Plan of the work

Preliminaries on gradient flows in metric spaces are collected in §2.2. Our main results in the continuum case are proved in §2.3 and extended to piecewise gradient-flow curves in §2.5. In §2.4 we discuss Condition (A) and its variants together with some examples. Our main results in the discrete case are proved in §2.6. Auxiliary results are presented in the subsequent sections.

Acknowledgement

The authors gratefully acknowledges support by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 716117). LDS gratefully acknowledges funding of his current position by the Austrian Science Fund (FWF), ESPRIT Fellowship Project 208. JM also acknowledges support by the Austrian Science Fund (FWF), Project SFB F65.

2.2 Gradient flows in metric spaces

In this section we collect some known facts about gradient flows in metric spaces. We assume throughout that (X, d) is a complete metric space and $J \subset \mathbb{R}$ is a (not necessarily open, nor closed) interval.

Let $1 \leq p < \infty$. A measurable function $m : J \to \mathbb{R}$ belongs to $L^p_{\mathrm{loc}}(J)$ if $\mathbf{1}_K m \in L^p(J)$ for every *compact* set $K \subset J$. A curve $(y_t)_{t \in J}$ is said to be locally p-absolutely continuous on J —in short: it belongs to $\mathrm{AC}^p_{\mathrm{loc}}(J; X)$— if there exists $m \in L^p_{\mathrm{loc}}(J)$ so that

$$\mathsf{d}(y_t, y_s) \leq \int_s^t m(r)\,\mathrm{d}r \tag{2.2.1}$$

for all $s, t \in J$ with $s < t$. Similarly, we write $(y_t)_{t \in J} \in \mathrm{AC}^p(J; X)$ if $m \in L^p(J)$.

Whenever $(y_t)_{t \in J}$ is in $\mathrm{AC}^1_{\mathrm{loc}}(J; X)$, the *metric speed*

$$|\dot{y}_t| := \lim_{s \to t} \frac{\mathsf{d}(y_s, y_t)}{|s - t|} \tag{2.2.2}$$

exists for a.e. $t \in J$. Furthermore, the metric speed coincides a.e. with the smallest function m satisfying (2.2.1); see, e.g., [AGS08, Thm. 1.1.2].

Remark 2.2.1. *Every curve in $\mathrm{AC}^1_{\mathrm{loc}}(J; X)$ is continuous on J. Note however that if $(y_t)_{t \in J} \in \mathrm{AC}^1_{\mathrm{loc}}(J; X)$ with $J = (a, b]$ for some $a < b$, then the existence of $\lim_{t \downarrow a} y_t$ does not imply that $(y_t)_{t \in J} \in \mathrm{AC}^1(J; X)$.*

The *domain* of a function $f : X \to (-\infty, \infty]$ is the set

$$\mathrm{dom}(f) := \{x \in X : f(x) < \infty\} .$$

In order to rule out trivial statements, we always assume that f is *proper*, i.e., $\mathrm{dom}(f) \neq \varnothing$. The (*descending*) *slope* of f at $x \in X$ is the quantity

$$|\mathrm{D}^- f|(x) := \limsup_{y \to x} \frac{\big[f(y) - f(x)\big]_-}{\mathsf{d}(y, x)} ,$$

where $a_- := \max\{-a, 0\}$ denotes the negative part of $a \in \mathbb{R}$. Conventionally, $|\mathrm{D}^- f|(x) := 0$ when $x \in \mathrm{dom}(f)$ is isolated, and $|\mathrm{D}^- f|(x) = +\infty$ if $x \notin \mathrm{dom}(f)$.

2.2.1 Gradient flows in metric spaces: curves of maximal slope

The next definition provides a natural notion of gradient flow in a metric space; cf. [AGS08] for an extensive treatment. The motivation for this definition comes from the following simple argument in Euclidean space. Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a smooth function. For any smooth curve $(u_t)_{t\in[0,T)}$ in $\mathbb{R}^n$ and $t \in (0,T)$, we have

$$-\frac{\mathrm{d}}{\mathrm{d}t} f(u_t) = -\nabla f(u_t)\cdot \dot{u}_t \leq \tfrac{1}{2}|\nabla f(u_t)|^2 + \tfrac{1}{2}|\dot{u}_t|^2 \,.$$

Since equality holds if and only if $\dot{u}_t = -\nabla f(u_t)$, the reverse inequality $-\frac{\mathrm{d}}{\mathrm{d}t} f(u_t) \geq \frac{1}{2}|\nabla f(u_t)|^2 + \frac{1}{2}|\dot{u}_t|^2$ is an equivalent formulation of the gradient-flow equation, which admits a natural generalisation to metric spaces.

Definition 2.2.2 (curve of maximal slope, gradient flow, cf. [MS20, Dfn. 4.1])**.** *Let $J \subset \mathbb{R}$ be an interval and let $f : X \to (-\infty, \infty]$ be proper. We say that $(y_t)_{t\in J}$ is a* curve of maximal slope for f *if*

- (a) $(y_t)_{t\in J} \in \mathrm{AC}^1_{\mathrm{loc}}(J;X)$*;*
- (b) $(f(y_t))_{t\in J} \in \mathrm{AC}^1_{\mathrm{loc}}(J;\mathbb{R})$*;*
- (c) *the following* Energy Dissipation Inequality *holds:*

$$-\frac{\mathrm{d}}{\mathrm{d}t} f(y_t) \geq \tfrac{1}{2}|\dot{y}_t|^2 + \tfrac{1}{2}|\mathrm{D}^- f|(y_t)^2 \quad \textit{for } \mathrm{d}t\textit{-a.e. } t \in J\,. \qquad (\mathrm{EDI})$$

If additionally $y_{\inf J} := \lim_{t\downarrow \inf J} y_t$ exists, we say that $(y_t)_{t\in J}$ is a curve of maximal slope for f starting at $y_{\inf J}$.

Remark 2.2.3. *From (EDI) and the absolute continuity of f we conclude that $t \mapsto f(y_t)$ is non-increasing along curves of maximal slope $(y_t)_{t\in[0,T)}$.*

There are several slightly different notions of curve of maximal slope in the literature, and the distinction matters for our purposes. In particular, it is important here to include the absolute continuity of the function along gradient-flow trajectories in our definition, as this allows one to deduce the following well-known fact, asserting the equality of the speed of the gradient flow and the slope of the driving functional.

Lemma 2.2.4. *Let $f : X \to (-\infty,\infty]$ be proper and let $(y_t)_{t\in J}$ be a curve of maximal slope. For a.e. $t \in J$ we have*

$$-\frac{\mathrm{d}}{\mathrm{d}t} f(y_t) = |\dot{y}_t|^2 = |D^- f|(y_t)^2. \qquad (2.2.3)$$

In particular, equality holds for a.e. $t \in J$ in (EDI).

Proof. Let $t \in J$ be such that the metric speed $|\dot{y}_t|$ and the derivative $\frac{\mathrm{d}}{\mathrm{d}t} f(y_t)$ exist. By local absolute continuity of $(y_t)_t$ and $(f(y_t))_t$, this property holds almost everywhere. Using the definitions we obtain

$$\begin{aligned}
-\frac{\mathrm{d}}{\mathrm{d}t} f(y_t) &= \limsup_{s\downarrow t} \frac{f(y_t) - f(y_s)}{|t-s|} \\
&\leq \limsup_{s\downarrow t} \frac{f(y_t) - f(y_s)}{\mathsf{d}(y_t,y_s)} \cdot \limsup_{s\downarrow t} \frac{\mathsf{d}(y_t,y_s)}{|t-s|} \\
&\leq |D^- f|(y_t)\cdot|\dot{y}_t|.
\end{aligned}$$

Combining this inequality with (EDI), we find that

$$\tfrac{1}{2}|\dot{y}_t|^2 + \tfrac{1}{2}|D^- f|(y_t)^2 \le -\frac{\mathrm{d}}{\mathrm{d}t} f(y_t) \le |D^- f|(y_t) \cdot |\dot{y}_t|,$$

which, again by Young's inequality, implies the desired identities. ■

In light of Lemma 2.2.4, every curve of maximal slope satisifies the *Energy Dissipation Equality* for a.e. $t \in J$:

$$-\frac{\mathrm{d}}{\mathrm{d}t} f(y_t) = \tfrac{1}{2}|\dot{y}_t|^2 + \tfrac{1}{2}|D^- f|(y_t)^2. \qquad \text{(EDE)}$$

Remark 2.2.5. *Let $J = [a, b)$ with $-\infty < a < b \le \infty$ and suppose that $y_a \in \mathrm{dom}(f)$. If f is bounded from below by some constant $M \in \mathbb{R}$, then $t \mapsto |\dot{y}_t|$ belongs to $L^2(a, b)$ for every curve of maximal slope $(y_t)_{t\in J}$. Indeed, for $t \in (a, b)$, integration of (EDI) yields*

$$\frac{1}{2}\int_a^t |\dot{y}_r|^2 \,\mathrm{d}r + \frac{1}{2}\int_a^t |\mathrm{D}^- f|(y_r)^2 \,\mathrm{d}r \le f(y_a) - f(y_t) \le f(y_a) - M. \qquad (2.2.4)$$

The conclusion follows by passing to the limit $t \uparrow b$.

Remark 2.2.6 (Comparison with [HM19, Dfn.s 2.12, 2.13])**.** *Our Definition 2.2.2 is more restrictive than [HM19, Dfn.s 2.12, 2.13] as we additionally require the condition in (b). This condition guarantees that $|\mathrm{D}^- f|$ is a strong upper gradient of f along $(y_t)_{t\in J}$; see e.g. [AGS14, Rmk. 2.8]. Furthermore, by (b) we may integrate (EDI) to conclude that $t \mapsto f(y_t)$ is non-increasing, which is rather an assumption in [HM19, Dfn. 2.12]. Everywhere below, following [HM19], we could drop the assumption of (b) and replace $|\mathrm{D}^- f|$ by any given strong upper gradient g. For the sake of simplicity however, we confine our exposition to the case $g := |\mathrm{D}^- f|$ for which the assumptions in [HM19] are verified in light of (b) as discussed above.*

2.3 Convergence of gradient flows

This section is devoted to the proof of Theorem 2.1.4, which deals with the converence of gradient flows under Assumption (A).

Definition 2.3.1 (Equilibrium point)**.** *We say that $x^\star \in X$ is an* equilibrium point *for f if $x^\star \in \mathrm{dom}(|\mathrm{D}^- f|)$ and $|\mathrm{D}^- f|(x^\star) = 0$.*

We refer to cf. [HM19, Dfn. 2.35] for a more general definition for strong upper gradients.

Clearly, every local minimizer $x^\star \in \mathrm{dom}(f)$ is an equilibrium point for f.

It will be useful to first investigate gradient flow curves starting from an equilibrium point.

Lemma 2.3.2 (Trivial flows)**.** *Let $f \colon X \to (-\infty, \infty]$ be proper and $T \in (0, \infty]$.*

(i) If $x^\star \in \mathrm{dom}(f)$ is an equilibrium point for f, then the constant curve $(y_t)_{t\in[0,T)}$ defined by $y_t \equiv x^\star$ is a curve of maximal slope for f starting at $x^\star$.

(ii) If $x^\star \in \mathrm{dom}(f)$ is a local minimizer for f, then the constant curve $(y_t)_{t\in[0,T)}$ defined by $y_t \equiv x^\star$ is the only curve of maximal slope for f starting at $x^\star$.

Proof. (i): This follows immediately from the definitions.

(ii): Let $x^\star \in \mathrm{dom}(f)$ be a local minimizer and let U be a neighbourhood of $x^\star$ such that $f \geq f(x^\star)$ on U. Furthermore, let $(y_t)_{t\in[0,T)}$ be a curve of maximal slope for f starting at $x^\star$, and set

$$t_0 := \inf\{t > 0 : y_t \notin U\} \wedge T\,.$$

Note that $t_0 > 0$, since $t \mapsto y_t$ is continuous. Since $y_t \in U$ for $t \in [0, t_0)$, we have $f(y_t) \geq f(x^\star)$ for $t \in [0, t_0)$. As $t \mapsto f(y_t)$ is non-increasing by (EDI), we thus infer that $f(y_t) = f(x^\star)$ for $t \in [0, t_0)$. Therefore, $t \mapsto \frac{\mathrm{d}}{\mathrm{d}t} f(y_t)$ is identically 0, hence $|\dot{y}_t| = 0$ for $t \in [0, t_0)$ again by (EDI). Applying (2.2.1) to the metric speed, we infer that $\mathsf{d}(y_t, y_0) \leq \int_0^t |\dot{y}_r|\,\mathrm{d}r = 0$, hence $y_t = y_0 := x^\star$ for all $t \in [0, t_0)$. By continuity of $t \mapsto y_t$ we conclude that $t_0 = T$, which proves the assertion. ∎

For convenience of the reader we recall the following definition from the introduction.

Definition 2.3.3 (Auxiliary function)**.** *Given a parameter function $\theta : [0,\infty) \to [0,\infty)$ we consider the auxiliary function*

$$\eta : [0,\infty) \to [-\infty,\infty), \qquad \eta(u) := \int_1^u \big(\theta'(s)\big)^2\,\mathrm{d}s \qquad \textit{for } u \in [0,\infty),$$
$$\Gamma : [0,\theta(\infty)) \to [-\infty,\infty), \qquad \Gamma(u) := (\eta \circ \theta^{-1})(u).$$

Here we use the convention that $\theta(\infty) := \lim_{u\to\infty} \theta(u)$. Note that θ is indeed invertible and nonnegative, so that Γ is well-defined. The following lemma collects some elementary properties of θ. We leave the proof to the reader.

Lemma 2.3.4 (Properties of the auxiliary function)**.** *The function η is strictly increasing, $\eta(1) = 0$, and $\eta(0)$ is possibly $-\infty$. Moreover, η is continuously differentiable on $(0,\infty)$ and $\eta'(u) = (\theta'(u))^2$ for all $u > 0$.*

Remark 2.3.5. *In the special case where $\theta(u) = \frac{c}{\gamma} u^\gamma$ we have the explicit formulas*

$$\eta(u) = \frac{c^2}{2\gamma - 1}(u^{2\gamma-1} - 1) \quad \textit{if } \gamma > 0,\ \gamma \neq \frac{1}{2}, \qquad \textit{and} \qquad \eta(u) = c^2 \log u \quad \textit{if } \gamma = \frac{1}{2}\,.$$

The following lemma contains the crucial quantitative bounds on the distance and the driving functional that can be derived from Condition (A), for suitable gradient-flow trajectories that stay within the ball $B_r(x_0)$.

Lemma 2.3.6 (Distance bound and energy bound)**.** *Let $f\colon X \to [0,\infty]$ be lower semicontinuous, and suppose that $x_0 \in \mathrm{dom}(f)$ and $r > 0$ satisfy Condition (A) for some parameter function θ. Let $(y_t)_{t\in[0,T)}$, with $T \in (0,\infty]$, be a curve of maximal slope starting at x_0. Let $0 \leq s \leq t < T$ and assume that $y_u \in B_r(x_0)$ and $f(y_u) > 0$ for all $u \in [s,t]$. Then:*

$$(\theta \circ f)(y_s) - (\theta \circ f)(y_t) \geq \mathsf{d}(y_t, y_s), \tag{2.3.1}$$
$$(\eta \circ f)(y_s) - (\eta \circ f)(y_t) \geq t - s. \tag{2.3.2}$$

Proof. As θ and η are continuously differentiable on $(0,\infty)$, and $t \mapsto f(y_t)$ is locally absolute continuous, we conclude that also $\mathcal{H}\colon t \mapsto (\theta \circ f)(y_t)$ and $t \mapsto (\eta \circ f)(y_t)$ are locally absolutely continuous on $(0,T)$. For almost every $u \in [s,t]$, we obtain by absolute continuity of $\mathcal{H}$, by (2.2.3), and by (A),

$$-\mathcal{H}'(u) = -(\theta' \circ f)(y_u) \cdot \tfrac{\mathrm{d}}{\mathrm{d}u} f(y_u) = (\theta' \circ f)(y_u) \cdot |\mathrm{D}^- f|(y_u)|\dot{y}_u| \geq |\dot{y}_u|. \tag{2.3.3}$$

Since $t \mapsto y_t$ is locally absolute continuous, we obtain

$$\mathsf{d}(y_t, y_s) \leq \int_s^t |\dot{y}_u| \,\mathrm{d}u \leq \int_s^t -\mathcal{H}'(u) \,\mathrm{d}u = \mathcal{H}(s) - \mathcal{H}(t), \tag{2.3.4}$$

which proves (2.3.1).

Moreover, using again that $f(y_u) > 0$ for all $u \in [s,t]$, we obtain for a.e. $u \in (s,t)$, by Lemma 2.3.4, by Lemma 2.2.4, and by (A),

$$-\tfrac{\mathrm{d}}{\mathrm{d}u}(\eta \circ f)(y_u) = -(\eta' \circ f)(y_u) \cdot \tfrac{\mathrm{d}}{\mathrm{d}u} f(y_u) = \Big((\theta' \circ f)(y_u) \cdot |\mathrm{D}^- f|(y_u)\Big)^2 \geq 1.$$

Integration of this inequality yields (2.3.2). ■

Example 2.3.7. *An explicit computation shows that in the special case where* $\theta(u) = \frac{c}{\gamma} u^\gamma$*, the energy estimate* (2.3.2) *becomes*

$$f(y_t) \leq \begin{cases} \left(f(y_s)^{2\gamma-1} - \dfrac{2\gamma-1}{c^2}(t-s) \right)^{1/(2\gamma-1)} & \text{if } \gamma > 0\,,\ \gamma \neq \frac{1}{2}\,, \\ e^{-(t-s)/c^2} f(y_s) & \text{if } \gamma = \frac{1}{2}\,. \end{cases} \tag{2.3.5}$$

We are now ready to prove our first main result.

Proof of Theorem 2.1.4. We assume that $f(x_0) > 0$, as the result would otherwise follow immediately from Lemma 2.3.2.

(i) We define

$$t_0 \coloneqq \inf \Big\{ t \in [0,T) : y_t \in \partial B_r(x_0) \Big\} \wedge T$$

and note that $t_0 > 0$, since $(y_t)_{t \in [0,T)}$ is continuous. If $t_0 = T$ the conclusion follows, hence it suffices to treat the case where $t_0 < T$.

If $f(y_{t_0}) = 0$, the conclusion follows from Lemma 2.3.2 and the definition of t_0. It thus remains to treat the case where $t_0 < T$ and $f(y_{t_0}) > 0$. We will show that these conditions yield a contradiction, which completes the proof.

Indeed, (2.3.1) and Assumption (A) yield, for $0 < t < t_0$,

$$\mathsf{d}(y_t, x_0) \leq (\theta \circ f)(x_0) - (\theta \circ f)(y_t) \leq r - (\theta \circ f)(y_{t_0})\,.$$

Since $(\theta \circ f)(y_{t_0}) > 0$ and $t \mapsto y_t$ is continuous, it follows by passing to the limit $t \uparrow t_0$ that $\mathsf{d}(y_{t_0}, x_0) < r$. This is the desired contradiction, since $\mathsf{d}(y_{t_0}, x_0) = r$ by construction.

(ii) Since $t \mapsto f(y_t)$ is continuous, it follows that

$$t_* \coloneqq \inf \{ t \in [0,T) : f(y_t) = 0 \} \wedge T > 0\,.$$

We first claim that $y_T := \lim_{t\to T} y_t$ exists and belongs to $\overline{B_r(x_0)}$.

If $t_* < T$, then $y_t = y_{t_*}$ for every $t \in [t_*, T)$ by Lemma 2.3.2, and the claim follows.

If otherwise $t_* = T$, then (2.3.1) holds for all $0 \le s \le t < T$. Write $\mathcal{H}(t) := (\theta \circ f)(y_t)$. Then $\mathcal{H}\colon [0,T) \to [0,\infty)$ is continuous, non-increasing and bounded from below, so it admits a continuous non-increasing extension on $[0,T]$. Thus, the bound (for $0 \le s < t < T$)

$$\mathsf{d}(y_s, y_t) \le \mathcal{H}(s) - \mathcal{H}(t) \le \mathcal{H}(s) - \mathcal{H}(T)$$

combined with $\mathcal{H}(s) \downarrow \mathcal{H}(T) \ge 0$ as $s \to T$ implies the Cauchy property of $(y_t)_t$, hence the existence of the limit, which proves the claim.

By lower semicontinuity of f and Lemma 2.3.2 and in view of (i), we infer that (2.3.1) holds for all $0 \le s \le t \le T$ (even if $t_* < T$). Choosing $s = 0$ and $t = T$, the last part of the statement follows using 2.1.7.

(iii) Let $0 \le t < t_*$. In view of (i), (2.1.10) follows from (2.3.2). Next, by (2.1.8) we have

$$\mathsf{d}(y_t, y_T) \le \theta(f(y_t)) - \theta(f(y_T)) \le \theta(f(y_t)).$$

Using this bound and (2.1.10), we obtain

$$(\eta \circ \theta^{-1})(\mathsf{d}(y_t, y_T)) \le (\eta \circ f)(y_t) \le (\eta \circ f)(x_0) - t \le (\eta \circ \theta^{-1})(r) - t,$$

which shows (2.1.9). By continuity of $(y_t)_t$ and lower semicontinuity of f, (2.1.9) and (2.1.10) extend to $t = t_*$.

Finally, suppose that $T = \infty$. If $t_* < \infty$, then clearly $f(y_\infty) = 0$. If on the other hand $t_* = \infty$, it follows from (2.1.10) that $(\eta \circ f)(y_t) \to -\infty$ as $t \to \infty$, hence $f(y_t) \to 0$. By lower semicontinuity of f the result follows. ■

In the special case where the parameter function θ takes the form $\theta(u) = \frac{c}{\gamma} u^\gamma$, we obtain the following more explicit result. The notation t^* was introduced in Theorem 2.1.4.

Corollary 2.3.8. *Let $f\colon X \to [0,\infty]$ be lower semicontinuous and suppose that $x_0 \in \mathrm{dom}(f)$ and $r > 0$ satisfy Condition (A) with parameter function $\theta(u) = \frac{c}{\gamma} u^\gamma$ for some $c > 0$ and $\gamma \in (0,1]$. Let $(y_t)_{t\in[0,T)}$ be a curve of maximal slope for f starting at x_0, for some $T \in (0,\infty]$. Then $y_T := \lim_{t\to T} y_t$ exists, y_t belongs to $\overline{B_r(x_0)}$ for all $t \in [0,T]$, and for all $0 \le t \le t^*$ we have*

$$\begin{cases} \mathsf{d}(y_t, y_T) & \le \frac{c}{\gamma}\left(\left(\frac{\gamma r}{c}\right)^{\frac{2\gamma-1}{\gamma}} - \frac{2\gamma-1}{c^2} t\right)^{\frac{\gamma}{2\gamma-1}}, \\ f(y_t) & \le \left(f(x_0)^{2\gamma-1} - \frac{2\gamma-1}{c^2} t\right)^{\frac{1}{2\gamma-1}}, \end{cases} \qquad \text{if } \gamma \ne \tfrac{1}{2}, \tag{2.3.6}$$

$$\begin{cases} \mathsf{d}(y_t, y_T) & \le r e^{-\frac{t}{2c^2}}, \\ f(y_t) & \le f(x_0)\, e^{-\frac{t}{c^2}}, \end{cases} \qquad \text{if } \gamma = \tfrac{1}{2}. \tag{2.3.7}$$

Moreover, if $\frac{1}{2} < \gamma \le 1$, we have $t_* \le \frac{c^2}{2\gamma-1} f(x_0)^{2\gamma-1}$.

Proof. The estimates for $\mathsf{d}(y_t, y_T)$ and $f(y_t)$ are obtained from (2.1.9) and (2.1.10) by rearranging terms. The final assertion follows from the second bound in (2.3.6). ■

Remark 2.3.9 (The case $\theta(u) = \frac{c}{2}\sqrt{u}$). *Under the above assumptions, the distance estimate in Corollary 2.3.8 can be improved if $\gamma = \frac{1}{2}$, using the following the ideas of [Cha22]. Let $0 \le s \le t \le T$ and assume that $y_u \in B_r(x_0)$ and $f(y_u) > 0$ for all $u \in [s,t]$. Then:*

$$\mathsf{d}(y_t, y_s)^2 \le 4c^2\Big(e^{-\frac{s}{2c^2}} - e^{-\frac{t}{2c^2}}\Big)\sqrt{f(x_0)}\Big(\sqrt{f(y_s)} - \sqrt{f(y_t)}\Big) \tag{2.3.8}$$

$$\le 4c^2 e^{-\frac{s}{2c^2}}\Big(e^{-\frac{s}{2c^2}} - e^{-\frac{t}{2c^2}}\Big)f(x_0)\,. \tag{2.3.9}$$

Proof. We can assume $t < T$ and then extend the result to $t = T$ by taking limits. Using the local 2-absolute continuity of $(y_t)_{t\in[0,T)}$, the Cauchy–Schwarz inequality, and the assumption that $f(y_u) > 0$ for all $u \in [s,t]$, we find

$$\mathsf{d}(y_t, y_s) \le \int_s^t |\dot y_u|\,\mathrm{d}u \le \Big(\int_s^t \sqrt{f(y_u)}\,\mathrm{d}u\Big)^{1/2}\left(\int_s^t \frac{|\dot y_u|^2}{\sqrt{f(y_u)}}\,\mathrm{d}u\right)^{1/2}. \tag{2.3.10}$$

By local absolute continuity of $u \mapsto f(y_u)$ we see that $u \mapsto \sqrt{f(y_u)}$ too is locally absolutely continuous, and $\frac{\mathrm{d}}{\mathrm{d}u}\sqrt{f(y_u)}(t) = \Big(2\sqrt{f(y_u)}\Big)^{-1}\frac{\mathrm{d}}{\mathrm{d}u}f(y_u)$ holds a.e. on $(0,T)$. Since $\frac{\mathrm{d}}{\mathrm{d}u}f(y_u) = -|\dot y_u|^2$ by (2.2.3), we obtain

$$\int_s^t \frac{|\dot y_u|^2}{\sqrt{f(y_u)}}\,\mathrm{d}u = -\int_s^t \frac{\frac{\mathrm{d}}{\mathrm{d}u}f(y_u)}{\sqrt{f(y_u)}}\,\mathrm{d}u = -2\int_s^t \frac{\mathrm{d}}{\mathrm{d}u}\sqrt{f(y_u)}\,\mathrm{d}u = 2\Big(\sqrt{f(y_s)} - \sqrt{f(y_t)}\Big)\,. \tag{2.3.11}$$

Since $(y_u)_{u\in[s,t]} \subseteq B_r(x_0)$, we can take the square root of the second bound in (2.3.7) to see that

$$\int_s^t \sqrt{f(y_u)}\,\mathrm{d}u \le \sqrt{f(x_0)}\int_s^t e^{-\frac{u}{2c^2}}\,\mathrm{d}u = 2c^2\Big(e^{-\frac{s}{2c^2}} - e^{-\frac{t}{2c^2}}\Big)\sqrt{f(x_0)}\,. \tag{2.3.12}$$

Inserting (2.3.11) and (2.3.12) into (2.3.10), we arrive at (2.3.8).

Finally, another application of the second bound in (2.3.7) yields

$$\sqrt{f(y_s)} - \sqrt{f(y_t)} \le \sqrt{f(y_s)} \le \sqrt{f(x_0)}\,e^{-\frac{s}{2c^2}}\,.$$

Inserting this inequality into the right-hand side of (2.3.8) we obtain (2.3.9). ∎

2.4 Comments on the assumption

In this section we collect some comments on the main assumption of this paper, Conditions (A) and (A') introduced in Definition 2.1.3.

Remark 2.4.1 (Comparison with [Cha22]). *Let $f\colon X \to [0,\infty]$ be proper. For $r > 0$ and $x_0 \in \mathrm{dom}(f)$ with $f(x_0) > 0$ we define*

$$\alpha = \alpha(x_0, r) := \inf_{\substack{x\in B_r(x_0)\\ 0<f(x)\le f(x_0)}} \frac{|D^- f|(x)^2}{f(x)}\,. \tag{2.4.1}$$

If $0 < \alpha < \infty$, it follows immediately from the definitions that the following statements are equivalent:

(i) Condition (A) holds for the parameter function $\theta(u) := 2\sqrt{u/\alpha(x_0, r)}$;

(ii) The following inequality holds:

$$\alpha(x_0, r) \geq \frac{4f(x_0)}{r^2}\,. \tag{C}$$

Similarly, the slightly stronger Condition (A') from Definition 2.1.3 is equivalent to Condition (C'), the strict inequality $\alpha(x_0, r) > \frac{4f(x_0)}{r^2}$. The latter condition is essentially identical to the main standing assumption in [Cha22], in the setting of C^2 functions on $\mathbb{R}^n$. The difference is that we restrict in (2.4.1) *to a sub-level set of f and work with an open ball instead of a closed ball of radius r around x_0.*

The following example illustrates that it is occasionally useful to consider the weaker Condition (C) instead of Condition (C').

Example 2.4.2. *Fix $x_0 > 0$ and consider the function $f : \mathbb{R} \to \mathbb{R}$ defined by (see Fig. 2.1)*

$$f(x) = \begin{cases} x^2 & \text{if } x \geq 0 \\ \frac{x_0^2}{2} & \text{if } x < 0 \end{cases}. \tag{2.4.2}$$

Then $\alpha(x_0, r) = 4$ for $0 < r \leq x_0$ and $\alpha(x_0, r) = 0$ for $r > x_0$. Therefore, Condition (C')

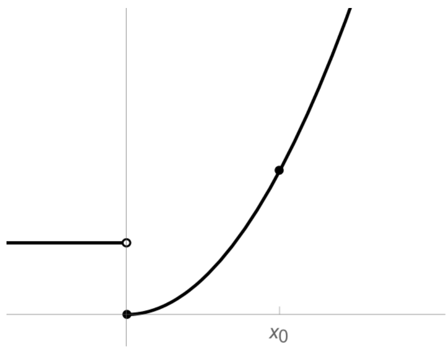


Figure 2.1: The function in (2.4.2).

fails to hold regardless of the choice of $r > 0$, but Condition (C) is satisfied for $r = x_0$.

Remark 2.4.3 (Attainment of the minimum)**.** *Assumption (A') implies the existence of a global minimizer $x^\star$ of f satisfying $\mathsf{d}(x^\star, x_0) \leq (\theta \circ f)(x_0)$ and $f(x^\star) = 0$. This follows from a result by Ioffe [Iof77], which we recall in Lemma 2.6.1 below. To derive the conclusion, Ioffe's result should be to be applied to the function $\theta \circ f$, and a metric version of the chain rule is required to relate the slope of f to the slope of $\theta \circ f$. For completeness, we give a proof of this chain rule in Lemma 2.7.1.*

In light of this observation, it is possible to derive results similar to Theorem 2.1.4 by applying existing results for convergence to a global minimum under the KŁ condition that assume the existence of a global minimum close to the starting point x_0; see, e.g., [ABRS10, Thm. 10] and [ABS13, Thm. 2.12] for such results for discrete schemes. However, a combination of these results with Ioffe's result yields a non-optimal criterion, as the KŁ inequality is required to hold on a bigger set than necessary. Moreover, some additional assumptions are made in the aforementioned results.

Remark 2.4.4 (Sharpness of Condition (A)). *To guarantee the existence and the proximity of a global minimizer of f under Condition (A), the constant r in the inequality $(\theta \circ f)(x_0) \le r$ cannot be replaced by any larger constant.*

To see this, fix $M < \infty$ (large) and consider for (small) $\varepsilon \ge 0$ the function $f_\varepsilon : [0, \infty) \to [0, \infty)$ defined by $f_\varepsilon(x) = \theta^{-1}(x + \varepsilon)$ for $0 \le x < M$ and $f_\varepsilon(x) = 0$ for $x \ge M$. Fix $x_0 \in (0, M/2)$. Differentiating the identity $\theta(f_\varepsilon(x)) = x + \varepsilon$, we find that $(\theta' \circ f_\varepsilon)(x) f_\varepsilon'(x) = 1$ for $0 < x < M$. In particular, the second inequality in Condition (A) is satisfied in an open ball of radius x_0 around x_0.

If $\varepsilon = 0$, the identity $\theta(f_0(x_0)) = x_0$ implies that Condition (A) holds with $r = x_0$, and indeed, the distance of x_0 to the nearest global minimizer of f (which is 0) equals x_0.

If $\varepsilon > 0$, Condition (A) fails to hold just barely (since $\theta(f_\varepsilon(x_0)) = x_0 + \varepsilon$), but the distance of x_0 to the nearest global minimizer (which is M) is enormous (namely, $M - x_0$) and the gradient flow curve starting from x_0 will converge to 0, which is not a global minimizer.

The following non-smooth example in $\mathbb{R}$ shows that Condition (A) can be applied in a setting where there is no uniqueness of gradient flow curves with a given starting point.

Example 2.4.5 (Non-uniqueness). *Let $\lambda > 0$ and $a > 0$, and consider the function $f : \mathbb{R} \to \mathbb{R}$ (see Fig. 2.2a)*

$$f(x) = \min\left\{\frac{\lambda}{2}(x-a)^2,\ \frac{\lambda}{2}(x+a)^2\right\}. \tag{2.4.3}$$

This function is everywhere smooth except at the origin. For each $x_0 \neq 0$, there exists a unique gradient-flow trajectory starting at x_0, given by $y_t := e^{-\lambda t} x_0 \pm \left(1 - e^{-\lambda t}\right) a$ for $x_0 \gtrless 0$. However, there are two distinct gradient-flow trajectories y^+ and y^- starting at the origin, given $y_t^\pm := \pm\left(1 - e^{-\lambda t}\right) a$ for $t \ge 0$.

In spite of this non-uniqueness, we shall verify that this example satisfies our assumptions. Note that $|D^- f|(x) = \lambda \Big| |x| - a \Big|$ for all $x \in \mathbb{R}$. In particular, f has finite slope at 0, although it is not differentiable. Consequently, $\frac{|D^- f|(x)^2}{f(x)} = 2\lambda$ for all $x \in \mathbb{R}$. It follows that Condition (C) holds for all $x_0 \in \mathbb{R}$ with $\alpha(x_0, r) = 2\lambda$ (hence Condition (A) holds with $\theta(u) = \sqrt{2u/\lambda}$), provided $r \ge |x_0 - a| \wedge |x_0 + a|$. Thus, at every point $x_0 \in \mathbb{R}$, the criterion provides the optimal result, in the sense that it yields the smallest possible ball centered at x_0 containing each gradient-flow trajectory starting at x_0.

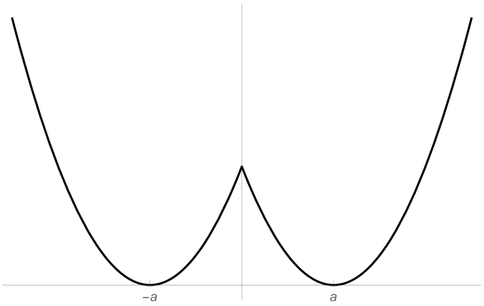
Remark 2.4.6 (Restriction to path connected component). *The second inequality in Condition (A) is required to hold for all $y \in B_r(x_0) \cap \{0 < f \le f(x_0)\}$. However, in the proof of Theorem 2.1.4, this bound is needed only on the set $G(x_0, r)$ consisting of all points inside the ball that are reachable by the considered curve of maximal slope starting at x_0. Therefore, Theorem 2.1.4 would still hold if one replaces the set $B_r(x_0) \cap \{0 < f \le f(x_0)\}$ by $G(x_0, r)$ in the definition in (A). Of course, in practice $G(x_0, r)$ is often not explicitly known, so this condition might be not easy to check. Instead of $G(x_0, r)$, one could also consider the path connected component $P(x_0, r)$ of x_0 in $B_r(x_0) \cap \{0 < f \le f(x_0)\}$ and modify the definition of (A) accordingly.*

The following modification of Example 2.4.5 provides an example where it is useful to employ the modified assumption. Let $\lambda > 0$ and $a > 0$, and consider the function $f : \mathbb{R} \to \mathbb{R}$ (see

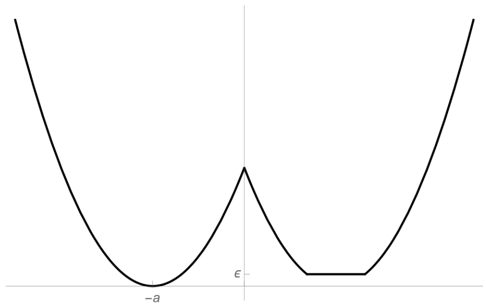
Fig. 2.2b) given by

$$f(x) = \min\left\{\max\left\{\frac{\lambda}{2}(x-a)^2, \varepsilon\right\}\ \frac{\lambda}{2}(x+a)^2\right\}. \qquad (2.4.4)$$

For this function, Assumption (C) is satisfied for every $x_0 < 0$ and suitable $r > 0$ when $\alpha(x_0, r)$ is defined with $P(x_0, r)$ in place of $B_r(x_0) \cap \{f \leq f(x_0)\}$. However, it is not satisfied for any $x_0 < 0$ yet sufficiently close to 0 when $\alpha(x_0, r)$ is defined as in (2.4.1).



(a) A non-smooth double-well potential (2.4.3).



(b) An asymmetric double-well potential (2.4.4).

Figure 2.2: The objective functions in Example 2.4.5 and Remark 2.4.6

2.5 Extension of gradient-flow trajectories

It is possible, even under Condition (A), that a curve of maximal slope defined on a finite interval $[0, T)$ does not extend to a curve of maximal slope on $[0, \infty)$. The following simple example illustrates this phenomenon.

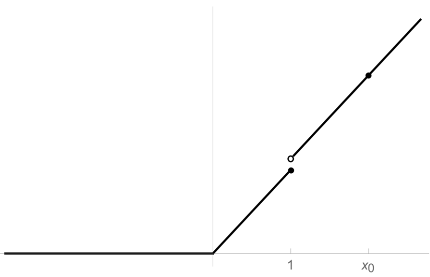


Figure 2.3: There is no curve of maximal slope with $T = \infty$ starting at x_0.

Example 2.5.1. *For fixed $m, \varepsilon > 0$, consider the lower-semicontinuous function $f\colon \mathbb{R} \to [0, \infty)$ defined by*

$$f(x) = m\, x\, \mathbf{1}_{[0,1]}(x) + (mx + \varepsilon)\mathbf{1}_{(1,\infty)}(x)\,.$$

See Figure 2.3. Let $x_0 > 1$ and fix $r > 0$. Then $f(x_0) = mx_0 + \varepsilon$ and Condition (A) is satisfied with

$$\theta(u) = \frac{2}{m}\sqrt{u(mx_0 + \varepsilon)}$$

and $r \geq 2(mx_0 + \varepsilon)/m$. On the interval $[0, T)$ with $T = \frac{x_0 - 1}{m}$, there exists a unique curve of maximal slope $(y_t)_{t\in[0,T)}$ starting from x_0. This is the curve which travels at constant speed m towards the discontinuity of f, namely $y_t = x_0 - mt$. However, there is no extension of $(y_t)_{t\in[0,T)}$ to a curve of maximal slope defined on $[0, T')$ for any $T' > T$, since $t \mapsto f(y_t)$ cannot be (absolutely) continuous on $[0, T')$.

Of course, the curve in this example can be naturally extended to $[0, \infty)$ by defining $y_T = 1$, and concatenating a new curve of maximal slope starting from there. The resulting curve, given by $y_t = (x_0 - mt)_+$ for $t \geq 0$, satisfies the exponential convergence rates of Corollary 2.1.5, even though it is not a curve of maximal slope in the sense of Definition 2.2.2.

Theorem 2.5.3 below shows that, under Condition (A), concatenated curves of maximal slope always satisfy the convergence rates of Theorem 2.1.4. The key ingredient is the following simple observation, which shows that Condition (A) is preserved under curves of maximal slope $(y_t)_{t\geq 0}$ in a suitable sense: if the condition holds at time 0 for x_0 and some $r > 0$, then it holds at any time $t \geq 0$ for the point y_t and the radius $r - \mathsf{d}(y_t, y_0)$ and with the same parameter function θ.

Remark 2.5.2 (Assumption preserved along the flow)**.** *Suppose that Condition (A) holds for $x_0 \in \mathrm{dom}(f)$ and $r > 0$. Let $(y_t)_{t\in[0,T)}$ be a curve of maximal slope, and extend it to $(y_t)_{t\in[0,T]}$ using Theorem 2.1.4. Then we have by* (2.3.1) *that for $t \in [0, T)$,*

$$(\theta \circ f)(y_t) \leq (\theta \circ f)(x_0) - \mathsf{d}(x_0, y_t) \leq r - \mathsf{d}(x_0, y_t),$$

which implies that Condition (A) holds for y_t (in place of x_0) and $r - \mathsf{d}(x_0, y_t)$ (in place of r); note that if $f(y_t) = 0$ it is possible that $r - \mathsf{d}(x_0, y_t) = 0$; otherwise this quantity is strictly positive. If f is lower semicontinuous, then Condition (A) holds also for y_T and $r - \mathsf{d}(x_0, y_T)$, as can be seen by taking limits.

Theorem 2.5.3. *Let $f : X \to [0, \infty]$ be a lower semicontinuous function on a complete metric space (X, d) and suppose that $x_0 \in X$ and $r > 0$ satisfy Condition (A) for some parameter function θ. Let $K \geq 1$, $0 = T_0 < T_1 < \ldots < T_K = T \leq \infty$ and $(y_t)_{t\in[T_i,T_{i+1})}$ be curves of maximal slope starting from x_i, where $x_i = \lim_{t\uparrow T_i} y_t \in B_r(x_0)$ for $1 \leq i \leq K - 1$. Then, setting $t_* := \inf\{t \in [0, T) : f(y_t) = 0\} \wedge T$, the following assertions hold:*

- (i) $y_t \in \overline{B_r(x_0)}$ *for all* $0 \leq t < T$;
- (ii) $y_T := \lim_{t\to T} y_t$ *exists and belongs to* $\overline{B_r(x_0)}$;
- (iii) $\mathsf{d}(y_s, y_t) \leq (\theta \circ f)(y_s) - (\theta \circ f)(y_t)$ *for all* $0 \leq t \leq T$.
- (iv) *for all* $0 \leq t \leq t_*$

$$\Gamma\Big(\mathsf{d}(y_t, y_T)\Big) \leq \Gamma(r) - t, \tag{2.5.1}$$

$$(\eta \circ f)(y_t) \leq (\eta \circ f)(x_0) - t. \tag{2.5.2}$$

Proof. (i) and (ii) follow from a repeated application of Theorem 2.1.4 and Remark 2.5.2.

(iii): Recall first that, for $K = 0, \ldots, K - 1$ and $T_k \leq s \leq t \leq T_{k+1}$, by (2.1.8)

$$\mathsf{d}(y_s, y_t) \leq (\theta \circ f)(y_s) - (\theta \circ f)(y_t)\,.$$

Therefore by a telescoping sum argument, the same inequality holds for all $0 \leq s \leq t \leq T$.

(iv): If $K = 1$, the claim follows from Theorem 2.1.4. Proceeding by induction, we assume that the claim holds for all $K \leq \bar{K}$. We shall show that it also holds for $K = \bar{K} + 1$. For this purpose, suppose that $t_* \geq \bar{K}$ and let $T_{\bar{K}} \leq t \leq t_*$, otherwise the conclusion is trivial.

Then notice that the induction hypothesis yields $(\eta \circ f)(x_{\bar{K}}) \leq (\eta \circ f)(x_0) - T_{\bar{K}}$. Moreover, applying Theorem 2.1.4 for $(y_t)_{t \in [T_{\bar{K}}, T_{\bar{K}+1}]}$, we find

$$(\eta \circ f)(y_t) \leq (\eta \circ f)(x_{\bar{K}}) - (t - T_{\bar{K}}).$$

Combining these bounds, (2.5.2) follows. Finally, (2.5.1) follows from (iii) and (2.5.2) in the same way as in the proof of Theorem 2.1.4. ■

Remark 2.5.4. *Theorem 2.5.3 still holds true if we replace the sub-level set of f in condition (A) with the path-connected component $P(x_0, r)$ as in Remark 2.4.6; however the proof does not work if we instead use $G(x_0, r)$, since, in this case, the inequality $(\theta' \circ f) \cdot |\mathrm{D}^- f| \geq 1$ may not hold on $G\big(y_{T_i}, r - \mathsf{d}(x_0, y_{T_i})\big)$.*

2.6 Convergence of the discrete scheme

This section contains the proof of Theorem 2.1.6, which deals with the convergence of proximal point sequences to a global minimizer. Our proof is based on adaptation of the arguments in [BDLM10, Thm. 24]). A key tool is the following result by Ioffe [Iof00]; see also [DIL15, Lemma 2.5].

Lemma 2.6.1 (Ioffe's Lemma)**.** *Let $g \colon X \to [-\infty, \infty]$ be a lower semicontinuous functional on a complete metric space (X, d). Let $x \in \mathrm{dom}(g)$ and suppose that there are constants $\delta \leq g(x)$ and $R > 0$ such that* [3]

$$|D^- g|(u) \geq v \quad \text{for all } u \in B_R(x) \cap \{\delta < g \leq g(x)\}$$

for some $v > (g(x) - \delta)/R$. Then:

$$\mathsf{d}\Big(x, \{g \leq \delta\}\Big) \leq \frac{g(x) - \delta}{v}. \tag{2.6.1}$$

Throughout the remainder of this section we impose the following standing assumptions that are in force without further mentioning:

- $f \colon X \to [0, \infty]$ is a proper and lower semicontinuous functional on a complete metric space (X, d);
- $x_0 \in \mathrm{dom}(f)$ and $r > 0$ satisfy Condition (A') for some parameter function θ;
- there exists a time-step $\bar{\tau} > 0$ (that will be fixed from now on) such that, for all $x \in B_r(x_0) \cap \{f \leq f(x_0)\}$ and $\tau \in (0, \bar{\tau})$, the functional

$$X \ni y \longmapsto f(y) + \frac{1}{2\tau} \mathsf{d}(x, y)^2 \tag{2.6.2}$$

  has at least one global minimizer. The non-empty set of minimizers will be denoted by $J_\tau(x)$.

[3] Note that the corresponding result in [DIL15] involves the closed ball $\{y \in X : \mathsf{d}(x, y) \leq R\}$ instead of the open ball B_R. It is easy to see that the statements are equivalent, possibly after taking a slightly smaller radius.

The latter condition is satisfied with $\bar{\tau} = \infty$ if (X, d) is proper, i.e., if all closed d-bounded sets in X are compact.

The following result contains some fundamental properties of J_τ, which can be found in [AGS08] under slightly different assumptions. The same proofs apply to our setting.

Lemma 2.6.2. *For $x \in B_r(x_0) \cap \{f \leq f(x_0)\}$ the following assertions hold:*

$$f(z_0) \geq f(z_1) \qquad \forall\, 0 < \tau_0 \leq \tau_1 < \bar{\tau}, z_0 \in J_{\tau_0}(x), z_1 \in J_{\tau_1}(x)\,; \tag{2.6.3}$$

$$\mathsf{d}(x, z) \geq \tau |D^- f|(z) \qquad \forall\, 0 < \tau < \bar{\tau}, z \in J_\tau(x)\,; \tag{2.6.4}$$

$$f(z) + \frac{\mathsf{d}(x,z)^2}{2\tau} + \int_0^\tau \frac{\mathsf{d}(x, z_s)^2}{2s^2}\,\mathrm{d}s = f(x) \quad \forall\, 0 < \tau < \bar{\tau}, z \in J_\tau(x), z_s \in J_s(x)\,. \tag{2.6.5}$$

Proof. Inequality (2.6.3) can be found in [AGS08, Lem. 3.1.2]; (2.6.4) can be found in [AGS08, Lem. 3.1.3]; (2.6.5) can be found in [AGS08, Thm. 3.1.4, Eqn. (3.1.12)]. ■

In the following result we consider a slightly more general notion of proximal point sequences, as we allow the step-size $\tau = \tau_k$ to depend on the step k. This will be useful in Lemma 2.6.6 below.

Lemma 2.6.3 (Confinement and distance bound)**.** *Let $(y_k)_{k=0}^N$ with $N \in \mathbb{N}$ be a proximal point sequence starting at x_0 with step-sizes $\tau_k \in (0, \bar{\tau})$ for $0 \leq k < N$. Then for all $0 \leq k \leq \ell \leq N$ we have the distance bound*

$$\mathsf{d}(y_k, y_\ell) \leq (\theta \circ f)(y_k) - (\theta \circ f)(y_\ell). \tag{2.6.6}$$

In particular, $y_k \in B_r(x_0)$ for all $0 \leq k \leq N$.

Proof. We will prove that (2.6.6) holds for all $0 \leq i \leq j \leq N$ by induction on N, noting that the case $N = 0$ is trivial.

We thus suppose that the claim is true for some $N \geq 0$, and let $(y_k)_{k=0}^{N+1}$ be a proximal point sequence starting at x_0. By the induction hypothesis and the triangle inequality it suffices to prove that $\mathsf{d}(y_N, y_{N+1}) \leq (\theta \circ f)(y_N) - (\theta \circ f)(y_{N+1})$. If $f(y_N) = 0$, we have $y_N = y_{N+1}$ and the claim follows. We thus assume that $f(y_N) > 0$.

By Condition (A') there exists $\varepsilon > 0$ such that $(1+\varepsilon)(\theta \circ f)(x_0) < r$. We will apply Lemma 2.6.1 to

$$g = \theta \circ f, \quad x = y_N, \quad \delta = (\theta \circ f)(y_{N+1}), \quad R = (1+\varepsilon)(\theta \circ f)(y_N), \quad v = 1.$$

We will show that the assumptions of Lemma 2.6.1 are satisfied.

- Firstly, we claim that $|D^- g|(u) \geq 1$ for $u \in B_R(y_N) \cap \{\delta < g \leq g(y_N)\}$. Indeed, by the triangle inequality and the induction hypothesis,

$$\begin{aligned} \mathsf{d}(x_0, u) &\leq \mathsf{d}(x_0, y_N) + \mathsf{d}(y_N, u) \\ &\leq \Big((\theta \circ f)(x_0) - (\theta \circ f)(y_N) \Big) + (1+\varepsilon)(\theta \circ f)(y_N) \\ &= (\theta \circ f)(x_0) + \varepsilon (\theta \circ f)(y_N) \leq (1+\varepsilon)(\theta \circ f)(x_0) < r. \end{aligned}$$

This shows that $u \in B_r(x_0)$. Moreover, since θ is strictly increasing, $f(u) \le f(y_N) \le f(x_0)$. Since furthermore $f(u) > 0$, Condition (A') implies that $\theta'(f(u))|D^- f|(u) \ge 1$. In particular, $|D^- f|(u) > 0$, hence u is not an isolated point. Therefore, Lemma 2.7.1 yields the desired inequality

$$|D^- g|(u) = \theta'(f(u))|D^- f|(u) \ge 1.$$

- Secondly, we claim that $g(y_N) - \delta < R$. Indeed,

$$g(y_N) - \delta = (\theta \circ f)(y_N) - (\theta \circ f)(y_{N+1}) < (1+\varepsilon)(\theta \circ f)(y_N) = R.$$

Using that θ is strictly increasing, we deduce from Lemma 2.6.1 that

$$\mathsf{d}\Big(y_N, \big\{f \le f(y_{N+1})\big\}\Big) \le (\theta \circ f)(y_N) - (\theta \circ f)(y_{N+1}).$$

This means that for any $\kappa > 0$ there exists $\bar{x} \in X$ such that

$$f(\bar{x}) \le f(y_{N+1}) \quad \text{and} \quad \mathsf{d}(\bar{x}, y_N) < (\theta \circ f)(y_N) - (\theta \circ f)(y_{N+1}) + \kappa.$$

On the other hand, since $y_{N+1} \in J_{\tau_N}(y_N)$, we have

$$\mathsf{d}^2(y_{N+1}, y_N) \le 2\tau_N\Big(f(\bar{x}) - f(y_{N+1})\Big) + \mathsf{d}^2(\bar{x}, y_N).$$

As $\kappa > 0$ can be chosen arbitrarily small, a combination of these bounds yields

$$\mathsf{d}\Big(y_N, y_{N+1}\Big) \le (\theta \circ f)(y_N) - (\theta \circ f)(y_{N+1}),$$

which completes the induction step and the proof of (2.6.6).

The final assertion follows from (2.6.6) since $\theta(f(x_0)) < r$ by Condition (A') . ■

Lemma 2.6.4. *For any $\tau \in (0, \bar{\tau})$ there exists an* infinite *proximal point sequence $(y_k)_{k \ge 0}$ with $y_0 = x_0$.*

Proof. It suffices to iteratively construct the sequence $(y_k)_k$ by letting $y_0 := x_0$ and y_{k+1} be a minimizer in (2.6.2) with y_k in place of x, noting that $y_k \in B_r(x_0)$ by Lemma 2.6.3 and $f(y_k) \le f(x_0)$ by (2.6.5). ■

Lemma 2.6.5. *Let $(y_k)_{k \ge 0}$ be a proximal point sequence and suppose that $y_\infty := \lim_{k\to\infty} y_k$ exists. Then*

$$f(y_\infty) = \lim_{k\to\infty} f(y_k) \quad \textit{and} \quad \lim_{k\to\infty} |D^- f|(y_k) = 0.$$

Proof. The lower semicontinuity of f yields $f(y_\infty) \le \liminf_{k\to\infty} f(y_k)$. On the other hand, since $y_k \in J_\tau(y_{k-1})$ we have

$$f(y_k) \le f(y_\infty) + \frac{1}{2\tau}\mathsf{d}(y_{k-1}, y_\infty)^2,$$

hence $\limsup_{k\to\infty} f(y_k) \le f(y_\infty)$. Combining these inequalities, we obtain the first identity.

As for the second identity, note that

$$0 \le |D^- f|(y_k) \le \frac{\mathsf{d}(y_{k-1}, y_k)}{\tau}$$

for $k \ge 1$ by (2.6.4). The conclusion follows by letting $k \to \infty$. ■

Proof of Theorem 2.1.6. Fix $\tau \in (0, \bar{\tau})$. The existence of an infinite proximal point sequence $(y_k)_{k\geq 0}$ with $y_0 = x_0$ and step-size τ was proved in Lemma 2.6.4.

Statement (i) was proved in Lemma 2.6.3.

The distance bound in (iii) was proved for $0 \leq i \leq j < \infty$ in Lemma 2.6.3 as well.

To prove (ii), note that $(\theta(f(y_k)))_k$ is a Cauchy sequence, as it is non-negative and non-increasing. Therefore, (iii) implies the Cauchy property of $(y_k)_k$ and hence the existence of $y_\infty := \lim_{k\to\infty} y_k$. Using (2.6.6) and Condition (A') we infer for $0 \leq i < \infty$ that

$$\mathsf{d}(y_i, y_\infty) \leq \liminf_{j\to\infty} \mathsf{d}(y_i, y_j) \leq (\theta \circ f)(y_i) < r.$$

This show that $y_\infty \in B_r(x_0)$ and the distance bound in (iii) for $j = \infty$ follows as well. To show that $f(y_\infty) = 0$, we may assume that $f(y_k) > 0$ for all k, since otherwise there is nothing to prove. Since $|D^- f|(y_k) \to 0$ as $k \to \infty$ by Lemma 2.6.5 and

$$\theta'(f(y_k)) \cdot |D^- f|(y_k) \geq 1$$

by Condition (A'), we infer that $\theta'(f(y_k)) \to \infty$. As θ' is continuous on $(0, \infty)$ and the sequence $(f(y_k))_k$ is non-increasing, it follows that $f(y_k) \to 0$, hence $f(y_\infty) = 0$ by lower semicontinuity of f. ■

For specific choices of the parameter function θ it is possible to obtain more explicit estimates on the decay of $f(y_k)$ and $\mathsf{d}(y_k, y_\infty)$ as $k \to \infty$. We adapt and refine some arguments from [AB09], where similar results are proved.

Lemma 2.6.6. *Let $(y_k)_{k=0}^{\infty}$ be a proximal point sequence with step-size $\tau \in (0, \bar{\tau})$ starting at x_0. If the parameter function θ is concave, we have, for all $k \geq 0$,*

$$f(y_k) - f(y_{k+1}) \geq \frac{\tau}{(\theta' \circ f)(y_{k+1})^2}\,. \tag{2.6.7}$$

Proof. Fix $k \geq 0$ and take $z_s \in J_s(y_k)$ for $s \in (0, \tau)$. Using De Giorgi's formula (2.6.5), the inequality (2.6.4), and Condition (A'), we obtain

$$\begin{aligned} f(y_k) - f(y_{k+1}) &= \frac{\mathsf{d}(y_k, y_{k+1})^2}{2\tau} + \int_0^\tau \frac{\mathsf{d}(y_k, z_s)^2}{2s^2}\,\mathrm{d}s \geq \frac{\tau}{2}|D^- f|(y_{k+1})^2 + \frac{1}{2}\int_0^\tau |D^- f|(z_s)^2\,\mathrm{d}s \\ &\geq \frac{\tau}{2(\theta' \circ f)(y_{k+1})^2} + \int_0^\tau \frac{1}{2(\theta' \circ f)(z_s)^2}\,\mathrm{d}s\,. \end{aligned}$$

Note that Condition (A') can be applied to z_s since $f(z_s) \leq f(y_k) \leq f(x_0)$ and Lemma 2.6.3 implies that $z_s \in B_r(x_0)$. Since θ is concave and $f(z_s) \geq f(y_{k+1})$ by (2.6.3), the result follows. ■

Note that a weaker decay estimate with an additional factor of $1/2$ in the right-hand side of (2.6.7) can be obtained without using De Giorgi's identity (2.6.5).

Corollary 2.6.7. *Suppose that the parameter function θ is given by $\theta(u) = \frac{c}{\gamma}u^\gamma$ for some $c > 0$ and $\gamma \in (0, 1]$. Let $(y_k)_{k\geq 0}$ be a proximal point sequence starting at x_0 with step-size $\tau \in (0, \bar{\tau})$, and set $y_\infty := \lim_{k\to\infty} y_k$. The following assertions hold:*

(i) *If $\gamma = 1$, then $y_k = y_\infty$ and $f(y_k) = 0$ for all $k \geq \lceil cr/\tau \rceil$.*

(*ii*) *If* $\frac{1}{2} < \gamma < 1$, *then, for* $k \geq 0$,

$$\begin{aligned} f(y_k) &\leq \Big(1 + \frac{\tau}{c^2} f(x_0)^{1-2\gamma}\Big)^{-k} f(x_0)\,, \\ \mathsf{d}(y_k, y_\infty) &\leq \frac{c}{\gamma} f(y_k)^\gamma = \mathcal{O}\bigg(\Big(1 + \frac{\tau}{c^2} f(x_0)^{1-2\gamma}\Big)^{-k\gamma}\bigg)\,, \end{aligned}$$

and, for $k \geq k_0 := \log_{1+\tilde\alpha}\big(2\tilde\alpha^{-1/(2\gamma-1)}\big)$ *with* $\tilde\alpha := \frac{\tau}{c^2} f(x_0)^{1-2\gamma}$,

$$\begin{aligned} f(y_k) &\leq \Big(\frac{\tau}{c^2}\Big)^{\frac{1}{2\gamma-1}} 2^{-(2-2\gamma)^{-(k-k_0)}}\,, \\ \mathsf{d}(y_k, y_\infty) &\leq \frac{c}{\gamma} f(y_k)^\gamma = \mathcal{O}\Big(2^{-\gamma(2-2\gamma)^{-(k-k_0)}}\Big)\,. \end{aligned}$$

(*iii*) *If* $\gamma = \frac{1}{2}$ *then, for* $k \geq 0$,

$$\begin{aligned} f(y_k) &\leq \Big(1 + \frac{\tau}{c^2}\Big)^{-k} f(x_0)\,, \\ \mathsf{d}(y_k, y_\infty) &\leq 2c\sqrt{f(y_k)} \leq 2c\Big(1 + \frac{\tau}{c^2}\Big)^{-k/2} \sqrt{f(x_0)}\,. \end{aligned}$$

(*iv*) *If* $0 < \gamma < \frac{1}{2}$ *then, for* $k \geq 0$,

$$\begin{aligned} f(y_k) &\leq \Big(f(x_0)^{-(1-2\gamma)} + C_1 k\Big)^{-\frac{1}{1-2\gamma}} = \mathcal{O}\Big(k^{-\frac{1}{1-2\gamma}}\Big), \\ \mathsf{d}(y_k, y_\infty) &\leq \frac{c}{\gamma} f(y_k)^\gamma = \mathcal{O}\Big(k^{-\frac{\gamma}{1-2\gamma}}\Big), \end{aligned}$$

where $C_1 = \sup_{R>1} \min\Big\{\frac{\tau(1-2\gamma)}{c^2 R}, \Big(R^{\frac{1-2\gamma}{2-2\gamma}} - 1\Big) f(x_0)^{2\gamma-1}\Big\}$.

Proof. (*i*): Suppose that $f(y_K) > 0$ for some $K \geq 0$. Condition (A') yields $|D^- f|(y_k) \geq \frac{1}{c}$ for all $0 \leq k \leq K$, hence $\mathsf{d}(y_k, y_{k+1}) \geq \frac{\tau}{c}$ for all $0 \leq k \leq K-1$ by (2.6.4). Using Condition (A') and Lemma 2.6.3 we infer that

$$r > c(f(x_0) - f(y_K)) = c\sum_{k=0}^{K-1}\Big(f(y_k) - f(y_{k+1})\Big) \geq \sum_{k=0}^{K-1} d(y_k, y_{k+1}) \geq \frac{K\tau}{c},$$

hence $K < cr/\tau$. It follows that $f(y_k) = 0$ and therefore $y_k = y_\infty$ for all $k \geq \lceil cr/\tau \rceil$.

$(ii) - (iv)$: Write $f_k := f(y_k)$. From (2.6.7) we deduce the recursive inequality

$$f_{k-1} - f_k \geq \frac{\tau}{c^2} f_k^{2-2\gamma}\,.$$

This inequality yields the decay of f_k using Lemma 2.8.1 with $\alpha = \tau/c^2$ and $\delta = 2 - 2\gamma$. The decay estimates for $d_k := \mathsf{d}(y_k, y_\infty)$ follow from the bounds for f_k combined with Theorem 2.1.6(*iii*). ■

Remark 2.6.8. *In general — unlike in the continuous setting of Corollary 2.3.8 — the discrete scheme does not converge in a finite number of steps when* $\frac{1}{2} < \gamma < 1$. *For suitable* $f \in C^1(\mathbb{R}^n)$, *this is easy to deduce from the equivalent formulation in* (2.1.3). *However, the corollary above shows that the rate of convergence is (asymptotically) faster than exponential.*

Example 2.6.9 (Non-uniqueness revisited)**.** *Let us consider again the function f defined in* (2.4.3) *(see Fig. 2.2a) and let $\tau > 0$. For any $x_0 \neq 0$, the resolvent $J_\tau(x_0)$ is single-valued. However, if $x_0 = 0$, the resolvent $J_\tau(x_0)$ contains two elements, say x_τ and $-x_\tau$. Consequently, there are two distinct proximal point sequences starting at $x_0 = 0$; once $x_\tau^1 \in J_\tau(x_0)$ is selected, the rest of the sequence is determined. Corollary 2.1.7 implies the exponential convergence for both of these sequences.*

2.7 A nonsmooth chain rule

In practice, it can be difficult to compute the slope of a non differentiable function, because tools such as the chain rule are missing. We have however the following basic substitute Lemma.

Lemma 2.7.1. *Let $f\colon X \to (-\infty, \infty]$ be proper and lower semicontinuous and let $g\colon \mathbb{R} \to \mathbb{R}$ be lower semicontinuous and non-decreasing. Then $g \circ f\colon X \to (-\infty, \infty]$ is proper and lower semicontinuous. Furthermore, if $x \in \operatorname{dom}(f)$ is not isolated and is such that there exists the left derivative $\partial_- g(f(x)) \geq 0$ of g at $f(x)$, then*

$$|\mathrm{D}^-(g \circ f)|(x) = \partial_- g(f(x)) \cdot |\mathrm{D}^- f|(x)\,. \tag{2.7.1}$$

Proof. The lower semicontinuity of $g \circ f$ is well known, so we only prove (2.7.1). Set

$$S \coloneqq \{(y_n)_n \subset X : y_n \to x\,,\ y_n \neq x\}\,.$$

Since x is not isolated, $S \neq \varnothing$. For a function $h\colon X \to (-\infty, \infty]$ we have

$$\limsup_{y \to x} h(y) = \sup\left\{\limsup_{n\to\infty}\ h(y_n) : (y_n)_n \in S\right\} = \max\left\{\limsup_{n\to\infty}\ h(y_n) : (y_n)_n \in S\right\}. \tag{2.7.2}$$

We first prove that

$$|\mathrm{D}^-(g \circ f)|(x) \leq \partial_- g(f(x)) \cdot |\mathrm{D}^- f|(x)\,. \tag{2.7.3}$$

Let $(y_n)_n \in S$. We need to show that

$$\limsup_{n\to\infty} \frac{[(g\circ f)(y_n) - (g \circ f)(x)]_-}{\mathsf{d}(y_n, x)} \leq \partial_- g(f(x)) \cdot |\mathrm{D}^- f|(x)\,.$$

Observe that if $f(y_n) \geq f(x)$ then, since g is non decreasing,

$$\frac{[(g\circ f)(y_n) - (g \circ f)(x)]_-}{\mathsf{d}(y_n, x)} = 0\,.$$

Therefore without loss of generality (by changing sequence and/or restricting to a subsequence) we can assume that $f(y_n) < f(x)$. Then, passing to the limit superior as $n \to \infty$ we see that $\limsup_n f(y_n) \leq f(x)$. By lower semicontinuity of f we have as well that $\liminf_n f(y_n) \geq f(x)$, thus there exists $\lim_n f(y_n) = f(x)$. Noting also that

$$\begin{aligned} \frac{[(g\circ f)(y_n) - (g \circ f)(x)]_-}{\mathsf{d}(y_n, x)} &= \frac{\big[(g\circ f)(y_n) - (g \circ f)(x)\big]_-}{\big[f(y_n) - f(x)\big]_-} \cdot \frac{\big[f(y_n) - f(x)\big]_-}{\mathsf{d}(y_n, x)} \\ &= \frac{g(f(x)) - g(f(y_n))}{f(x) - f(y_n)} \cdot \frac{[f(y_n) - f(x)]_-}{\mathsf{d}(y_n, x)} \end{aligned} \tag{2.7.4}$$

we conclude that

$$\limsup_{n\to\infty} \frac{\Big[(g\circ f)(y_n) - (g\circ f)(x)\Big]_-}{\mathsf{d}(y_n, x)} \le \partial_- g(f(x)) \cdot |\mathrm{D}^- f|(x)$$

as desired.

We now prove the converse inequality:

$$|\mathrm{D}^-(g\circ f)|(x) \ge \partial_- g(f(x)) \cdot |\mathrm{D}^- f|(x)\,. \tag{2.7.5}$$

If $|\mathrm{D}^- f|(x) = 0$ the claim is trivial, so we assume now that $|\mathrm{D}^- f|(x) > 0$. By (2.7.2) there exists $(y_n)_n \in S$ such that

$$\limsup_{n\to\infty} \frac{\Big[f(y_n) - f(x)\Big]_-}{\mathsf{d}(y_n, x)} = |\mathrm{D}^- f|(x) > 0\,.$$

By restricting to a subsequence we can assume that $f(y_n) < f(x)$. Arguing as before we have the existence of $\lim_n f(y_n) = f(x)$ and that (2.7.4) holds. Taking the limit superior as $n \to \infty$ in (2.7.4) gives

$$|\mathrm{D}^-(g\circ f)|(x) \ge \limsup_{n\to\infty} \frac{[(g\circ f)(y_n) - (g\circ f)(x)]_-}{\mathsf{d}(y_n, x)} \ge \partial_- g(f(x)) \cdot |\mathrm{D}^- f|(x)$$

as desired. Combing (2.7.5) with the opposite inequality (2.7.3) yields the assertion. ∎

2.8 Estimating recursive inequalities

The following lemma contains some estimates that are used in the proof of Corollary 2.6.7.

Lemma 2.8.1. *Let $(f_k)_{k\in\mathbb{N}}$ be a sequence of non-negative real numbers with $f_0 > 0$ and suppose that for some $\alpha, \delta > 0$ the recursive relation*

$$f_{k-1} - f_k \ge \alpha f_k^\delta \tag{2.8.1}$$

holds for all $k \ge 1$. Then, for all $k \ge 0$:

$$f_k \le \begin{cases} \Big(f_0^{-(\delta-1)} + Ck\Big)^{-1/(\delta-1)} & \text{if } \delta > 1\,, & (2.8.2a)\\ (1+\alpha)^{-k} f_0 & \text{if } \delta = 1\,, & (2.8.2b)\\ (1+\tilde\alpha)^{-k} f_0 & \text{if } \delta < 1\,, & (2.8.2c)\end{cases}$$

where $\tilde\alpha = \alpha / f_0^{1-\delta}$ and

$$C \coloneqq \sup_{R>1} \min\left\{\frac{\alpha(\delta-1)}{R}, \Big(R^{(\delta-1)/\delta} - 1\Big) f_0^{1-\delta}\right\}.$$

Furthermore, if $\delta < 1$, we also have

$$f_k \le \alpha^{1/(1-\delta)} \cdot 2^{-\delta^{-(k-k_0)}} \quad \text{for } \; k \ge k_0 \coloneqq \log_{1+\tilde\alpha}\Big(2\tilde\alpha^{-1/(1-\delta)}\Big)\,. \tag{2.8.3}$$

Proof. Note first that in all cases the sequence $(f_k)_k$ is non-increasing.

(2.8.2a): We follow the arguments of [AB09]. Fix $R \in (1,\infty)$ and consider the concave function $H(s) = \frac{1}{1-\delta}s^{1-\delta}$ and its derivative $h(s) = H'(s) = s^{-\delta}$. Observe that (2.8.1) can be equivalently written as

$$\alpha \le (f_{k-1} - f_k)h(f_k)\,.$$

Suppose first that $h(f_k) \le Rh(f_{k-1})$. Using the concavity of H we obtain

$$\begin{aligned}\alpha &\le (f_{k-1} - f_k)h(f_k) \le R(f_{k-1} - f_k)h(f_{k-1}) \le R(H(f_{k-1}) - H(f_k))\\ &= \frac{R}{\delta - 1}\Big(f_k^{1-\delta} - f_{k-1}^{1-\delta}\Big).\end{aligned}$$

Writing $C_1(R) \coloneqq \alpha(\delta-1)/R > 0$, this shows that

$$f_k^{1-\delta} - f_{k-1}^{1-\delta} \ge C_1(R)\,.$$

Suppose next that instead $h(f_k) > Rh(f_{k-1})$. Raising this inequality to the power $\frac{\delta-1}{\delta}$ we obtain

$$f_k^{1-\delta} > R^{\frac{\delta-1}{\delta}} f_{k-1}^{1-\delta}$$

and hence, since $(f_k)_k$ is non-increasing,

$$f_k^{1-\delta} - f_{k-1}^{1-\delta} \ge \Big(R^{\frac{\delta-1}{\delta}} - 1\Big) f_{k-1}^{1-\delta} \ge \Big(R^{\frac{\delta-1}{\delta}} - 1\Big) f_0^{1-\delta} \eqqcolon C_2(R) > 0.$$

Finally, defining $C \coloneqq \sup_{R\in(1,\infty)} \min\{C_1(R), C_2(R)\} > 0$, the above inequalities combined yield

$$f_k^{1-\delta} - f_{k-1}^{1-\delta} \ge C.$$

Evaluating a telescopic sum, we obtain $f_k^{1-\delta} - f_0^{1-\delta} \ge Ck$. Rearranging terms we obtain the desired inequality (2.8.2a).

(2.8.2b): This is straightforward.

(2.8.2c): Suppose without loss of generality that $f_k > 0$. From (2.8.1) it follows that

$$\frac{f_{k-1}}{f_k} \ge 1 + \frac{\alpha}{f_k^{1-\delta}} \ge 1 + \frac{\alpha}{f_0^{1-\delta}} = 1 + \tilde{\alpha},$$

from which we deduce (2.8.2c).

(2.8.3): Writing $\tilde{f}_k \coloneqq \alpha^{-1/(1-\delta)} f_k$ we note that (2.8.1) implies $\tilde{f}_k^{\delta} \le \tilde{f}_{k-1}$ and therefore

$$\tilde{f}_k \le \Big(\tilde{f}_{k_0}\Big)^{\delta^{-(k-k_0)}}.$$

Moreover, (2.8.2c) and the definition of k_0 yield $\tilde{f}_{k_0} \le (1+\tilde{\alpha})^{-k_0}\tilde{f}_0 = \frac{1}{2}$. Combining these estimates, we obtain the desired estimate $\tilde{f}_k \le 2^{-\delta^{-(k-k_0)}}$. ■

CHAPTER 3

Contractive coupling rates and curvature lower bounds for Markov chains

This chapter corresponds to the preprint [Ped23].

Contractive coupling rates have been recently introduced by Conforti as a tool to establish convex Sobolev inequalities (including modified log-Sobolev and Poincaré inequality) for some classes of Markov chains. In this work, for most of the examples discussed by Conforti, we use contractive coupling rates to prove stronger inequalities, in the form of curvature lower bounds (in entropic and discrete Bakry–Émery sense) and geodesic convexity of some entropic functionals. In addition, we recall and give straightforward generalizations of some notions of coarse Ricci curvature, and we discuss some of their properties and relations with the concepts of couplings and coupling rates: as an application, we show exponential contraction of the p-Wasserstein distance along the heat flow in the aforementioned examples.

3.1 Introduction

In this work, we are mostly concerned with finite state space continuous time Markov chains and we assume that they are irreducible and reversible: we use the letter Ω for the state space, L for the generator and m for the invariant measure. A fundamental problem in the theory of Markov chains consists in estimating the speed of convergence to the stationary distribution and giving upper bound for its mixing time. Convex Sobolev inequalities are a particularly useful tool to address this task. Given a convex function $\phi\colon \mathbb{R}_{\geq 0} \to \mathbb{R}$ such that $\phi \in C^1(\mathbb{R}_{>0})$, we define the ϕ-entropy of a function $\rho\colon \Omega \to \mathbb{R}_{>0}$ by

$$\mathcal{H}^\phi(\rho) := \mathbb{E}_m[\phi \circ \rho] - \phi(\mathbb{E}_m[\rho]). \tag{3.1.1}$$

Notice that by Jensen's inequality $\mathcal{H}^\phi(\rho) \geq 0$ and $\mathcal{H}^\phi(C) = 0$ for any constant $C > 0$. Therefore, if we denote by $\rho = \frac{d\mu}{dm}$ the density of a probability measure μ with respect to m, we can think of $\mathcal{H}^\phi(\rho)$ as a (non-symmetric) measure of distance of μ from m. We also recall the definition of the Dirichlet form $\mathcal{E}\colon \mathbb{R}^\Omega \times \mathbb{R}^\Omega \to \mathbb{R}$ via

$$\mathcal{E}(f, g) := -\mathbb{E}_m[f(Lg)],$$

and of the $\phi-$Fisher information

$$\mathcal{I}^\phi(\rho) := \mathcal{E}(\rho, \phi' \circ \rho). \tag{3.1.2}$$

We then say that a ϕ-convex Sobolev inequality holds with constant $K > 0$ (notation: $\mathrm{CSI}_\phi(K)$) if for all positive functions $\rho\colon \Omega \to \mathbb{R}_{>0}$ we have that

$$K\,\mathcal{H}^\phi(\rho) \le \mathcal{I}^\phi(\rho). \tag{3.1.3}$$

The interest behind this inequality lies in the fact that, denoting by $P_t = e^{tL}$ the semigroup associated to the generator L, we have the well-known identity

$$\frac{d}{dt}\mathcal{H}^\phi(P_t\rho) = -\mathcal{I}^\phi(P_t\rho)$$

for any $\rho\colon \Omega \to \mathbb{R}_{\ge 0}$. Thus, by Grönwall Lemma, (3.1.3) is equivalent to the exponential decay of the entropy along the heat flow

$$\mathcal{H}^\phi(P_t\rho) \le e^{-Kt}\mathcal{H}^\phi(\rho),$$

and therefore quantifies the speed of convergence to equilibrium of the Markov chain. Classical choices of the function ϕ include the function ϕ_α for $\alpha \in [1,2]$ defined by

$$\phi_\alpha(t) = \begin{cases} t\log t - t + 1 & \text{if } \alpha = 1, \\ \frac{t^\alpha - t}{\alpha - 1} - t + 1 & \text{if } \alpha \in (1,2]. \end{cases} \tag{3.1.4}$$

When $\phi = \phi_1$, we get the relative entropy and inequality (3.1.3) is the celebrated modified log-Sobolev inequality [BT06] (notation: $\mathrm{MLSI}(K)$); when $\phi = \phi_2$, we find the variance and (3.1.3) is the Poincaré inequality (notation: $\mathrm{PO}(K)$). For $\alpha \in (1,2)$, inequalities (3.1.3) are known as Beckner inequalities, which interpolate between modified log-Sobolev and Poincaré [BT06, JY17].

Curvature of Markov chains In the setting of Riemannian manifolds, positive lower bounds for the Ricci curvature have been linked to many functional inequalities: this has motivated the seminal independent works of Sturm [Stu06] and Lott and Villani [LV09], who extended the notion of curvature lower bound and many of its consequences (including some logarithmic Sobolev inequalities) to a large class of geodesic metric measure spaces. In spite of its generality, this theory does not apply to Markov chains on discrete spaces; for this reason, several adapted notions of curvature have been proposed, based on different equivalent characterisations of Ricci curvature for Riemannian manifolds. Among these, we recall in particular the entropic curvature by Erbar and Maas [EM12], which is based on displacement convexity of the relative entropy with respect to an adapted Wasserstein-like metric $\mathscr{W}$ introduced in [Maa11] (see also the work of Mielke [Mie13]). This theory shares many similarities with the classical Lott–Sturm–Villani one, and among its merits it is such that many of the desired functional inequalities follow from positive lower bounds on the Ricci curvature, including in particular the modified log-Sobolev inequality. Moreover, as shown in [EM14], the role of the classical relative entropy (with respect to m) can be taken over by other ϕ-entropy functionals as defined in (3.1.1), provided that one changes accordingly a parameter function in the definition of $\mathscr{W}$: once again, from positive geodesic convexity one can derive many consequences, including the convex Sobolev inequality (3.1.3). Unfortunately, establishing positive lower bounds for the entropic Ricci curvature (or more generally K-geodesic convexity of an entropic functional $\mathcal{H}^\phi$) of a Markov chain can be challenging, and in many interesting examples good estimates are not available.

Coupling rates and Conforti's results While studying the entropic curvature of a Markov chain is a difficult task, in general even finding good estimates on the best constant for the modified log-Sobolev inequality (or other convex Sobolev inequalities) can be difficult. For this reason, in the recent paper [Con22], Conforti introduced a method based on the new notion of *coupling rates* to study general convex Sobolev inequalities, and applied it to some interesting classes of Markov chains. Coupling rates are a modification of the familiar notion of coupling: roughly speaking, they are used to "couple" the action of the generator L from two different states. While couplings have been extensively used to establish fast convergence of Markov chains ("probabilistic" approach to fast mixing), their use to establish convex Sobolev inequalities (which belong to the "analytic" approach to fast mixing) is less common, and the results of [Con22] give an interesting connection between these two families of methods.

Our contribution and organization of the paper In this work, we show that the coupling rates introduced by Conforti are a powerful tool to establish entropic curvature lower bounds and other related inequalities for some classes of Markov chains. As an illustration of the applications of these methods, we state below one particular instance of our results. We refer to Sections 3.2–3.4 below for precise definitions.

Theorem 3.1.1 (Cf. Sections 3.4.1, 3.4.3)**.** *Denote by* $\mathrm{Ric_e}$ *the entropic curvature of a continuous time reversible Markov chain [EM12, FM16, EHMT17].*

- *For the Curie–Weiss model with size N and parameter $\beta > 0$, in the limit $N \to \infty$ we have*
$$\mathrm{Ric_e} \geq (1-\beta) + (1-2\beta)e^{-\beta}$$
for $\beta \leq \frac{1}{2}$.
- *For the Ising model in dimension d with parameter $\beta > 0$, we have*
$$\mathrm{Ric_e} \geq 1 + e^{-2\beta d} - 3d(1-e^{-2\beta})e^{2\beta d}$$
if $2d\Big(1-e^{-2\beta}\Big)e^{4d\beta} \leq 1$.
- *For the hardcore model on a graph with maximum degree Δ and parameter $\beta > 0$, set $\kappa_* = 1 - \beta(\Delta - 1)$ and $\overline{\kappa_*} = \min\{\beta, 1-\beta\Delta\}$. Then*
$$\mathrm{Ric_e} \geq \frac{\kappa_*}{2} + \overline{\kappa_*}$$
provided that $\beta\Delta \leq 1$.

We remark that in all the examples above we find new estimates for the entropic curvature of those Markov chains: these estimates imply in particular MLSI with the same constant obtained in [Con22], but also other interesting functional inequalities, such as exponential contractivity along the heat flow of the popular Wasserstein-like metric $\mathscr{W}$ of [Maa11] (see Section 3.2.1 and references therein for more details). Therefore, in this sense, we provide a strengthening of the results of [Con22].

To conclude this section, we briefly present below the organization of this paper while giving a more complete overview of our contributions.

- In Section 3.2 we give some preliminary definitions and define the general abstract inequality that we consider: it reads $\mathscr{B}(\rho, \psi) \geq K\mathscr{A}(\rho, \psi)$ for a constant K and all $\rho\colon \Omega \to \mathbb{R}_{>0}$ and $\psi\colon \Omega \to \mathbb{R}$, and it depends on an additional weight function $\theta\colon \mathbb{R}_{>0} \times \mathbb{R}_{>0} \to \mathbb{R}_{>0}$. This inequality was first introduced by Erbar and Maas in their works on the entropic curvature of a Markov chain [EM12] and, more generally, on the geodesic convexity of some entropic functionals [EM14]. To motivate the interest in studying this inequality and for the sake of completeness, in the following subsections we recall some results by Erbar and Maas in connection with specific choices of the weight function θ. In particular:
  - In Section 3.2.1 we recall that when θ is the logarithmic mean then the inequality corresponds to an entropic curvature lower bound for the Markov chain, as proved in [EM12]. For the convenience of the reader, we also recall from [EM12] some necessary definitions and consequences of such curvature lower bound, including $\mathrm{MLSI}(2K)$.
  - In Section 3.2.2, by the results of [EM14], we extend the considerations of the previous section to the case where the relative entropy is replaced by some other ϕ-entropy functionals, and in particular we explain how we recover a family of convex Sobolev inequalities as in (3.1.3) (consistently with [Con22]), together with other functional inequalities.
  - In Section 3.2.3 we recall that if θ is the arithmetic mean then the inequality of interest corresponds to a lower bound for another notion of discrete curvature, namely the discrete Bakry–Émery one (see [Sch99]).
- In Section 3.3 we recall the definition of coupling rates. Moreover, under some basic assumptions on the weight function θ, we use coupling rates to provide a lower bound for the quantity $\mathscr{B}(\rho, \psi)$: this is the content of Lemma 3.3.1, which will be crucial for our method and the applications to the particular Markov chains considered later in this paper. This section also includes some heuristic considerations, explaining the favorable role of "contractivity" of the couplings in proving the abstract inequality $\mathscr{B}(\rho, \psi) \geq K\mathscr{A}(\rho, \psi)$, which can also guide possible future applications of this tool.
- In Section 3.4 we illustrate the considerations of the previous section by considering most of the examples discussed in [Con22] and correspondingly establishing the general abstract inequality $\mathscr{B}(\rho, \psi) \geq K\mathscr{A}(\rho, \psi)$ introduced in Section 3.2. More specifically:
  - In Section 3.4.1 we consider Glauber dynamics, which includes in particular the Ising model and the Curie–Weiss model.
  - In Section 3.4.2 we consider a simplified version of the Bernoulli–Laplace model.
  - In Section 3.4.3 we consider the classical hardcore model.
  - In Section 3.4.4 we consider the case of interacting random walks on the discrete grid $\mathbb{N}^d$.

  In particular, by choosing θ to be the logarithmic mean, we find new estimates for the entropic curvature of the Ising model, the Curie–Weiss model and the hardcore model, and we recover the best known lower bound for the entropic curvature of the Bernoulli–Laplace model.
- Finally, in Section 3.5 we explain how the coupling rates constructed by Conforti are naturally connected with the notion of coarse Ricci curvature by Ollivier [Oll09]. We

recall some well-known definitions and properties of the coarse Ricci curvature, and we also provide some natural generalizations. In particular, inspired by the properties of the coupling rates constructed by Conforti, we consider a stronger notion of coarse Ricci curvature which, roughly speaking, is based on simultaneous contraction of 1-Wasserstein distance W_1 and non-expansion of the ∞-Wasserstein distance W_∞ along the Markov chain dynamics (where both Wasserstein distances are defined with respect to the natural graph distance d on Ω). Correspondingly, we raise the question of connecting positive lower bounds for this notion of curvature to a modified log-Sobolev inequality, formulating a weaker version of a conjecture by Peres and Tetali.

As a further application of the discussion in Section 3.5, and as already done by Conforti for the specific case of the interacting random walks of Section 3.4.4, we show that in all the other examples discussed in Section 3.4 Conforti's coupling rates imply an exponential contraction of the Wasserstein distance of the form

$$W_p(P_t\rho, P_t\sigma) \leq e^{-\frac{Kt}{p}} W_p(\rho, \sigma)$$

for all starting densities ρ, σ and $p \geq 1$.

As a final remark, we emphasize how, in some sense, this section shows that the connections between probabilistic and analytic methods emerging in [Con22] carry over at the level of curvature. In fact, while contractive couplings are naturally linked to the coarse Ricci curvature, we use them to establish (for some classes of Markov chains) lower bounds on the entropic curvature, which is a rather analytic notion of curvature.

3.2 Preliminaries and main inequality

Following [Con22], we work with a so-called "mapping representation" of the Markov chain, which we briefly recall. We are given a finite set of moves G (where a move $\sigma \in G$ is a function $\sigma\colon \Omega \to \Omega$) together with a transition rate function $c\colon \Omega \times G \to \mathbb{R}_{\geq 0}$, so that $c(\eta, \sigma)$ represents the rate of using the move σ starting from the state η. Such a mapping representation has already proved useful before in establishing functional inequalities and curvature lower bounds for Markov chains [PP13, CPP09, EM12, FM16, EHMT17]. Typically (and if not otherwise specified) we use the letter η for a state and $\sigma, \gamma, \bar{\gamma}$ for moves, and to lighten the notation we write for example $\sigma\eta$ instead of $\sigma(\eta)$ for the state reached after jumping with the move σ from the state η. We make the assumption that for each move σ there exists a unique inverse $\sigma^{-1} \in G$ such that $\sigma^{-1}\sigma\eta = \eta$ whenever $m(\eta)c(\eta, \sigma) > 0$ (recall that we denote by m the unique invariant measure). We also denote by $e\colon \Omega \to \Omega$ the 'null move" (i.e. the identity map); without loss of generality, we assume that $e \notin G$ and we denote by $G^* := G \cup \{e\}$ the enlarged set of moves, as in [Con22]. With this notation, we can write explicitly the action of the generator L of the continuous time Markov chain in the form

$$L\psi(\eta) = \sum_{\sigma \in G} c(\eta, \sigma)(\psi(\sigma\eta) - \psi(\eta)) \tag{3.2.1}$$

for any bounded $\psi\colon \Omega \to \mathbb{R}$. Notice that in (3.2.1) we could also take the sum for $\sigma \in G^*$, and that in this context the rates $c(\eta, e)$ can be arbitrarily defined. The state space Ω is at most countable, and we assume that

$$\sum_{\eta \in \Omega, \sigma \in G} m(\eta)c(\eta, \sigma) < \infty.$$

We also use the notation

$$\nabla_\sigma \psi(\eta) := \psi(\sigma\eta) - \psi(\eta)$$

for the discrete gradient and

$$S := \{(\eta, \sigma) \in \Omega \times G \mid c(\eta, \sigma) > 0\}.$$

We assume that for all bounded functions $F\colon \Omega \times G \to \mathbb{R}$ we have

$$\sum_{\eta\in\Omega,\sigma\in G} m(\eta)c(\eta,\sigma)F(\eta,\sigma) = \sum_{\eta\in\Omega,\sigma\in G} m(\eta)c(\eta,\sigma)F(\sigma\eta,\sigma^{-1}). \tag{3.2.2}$$

As observed in e.g. [FM16, Def 3.1] (cf. also [PP13, EM12, EHMT17, Con22]), the condition (3.2.2) expresses the reversibility of the Markov chain, and every irreducible and reversible Markov chain admits a mapping representation satisfying the above requirements. In practice, it is typically useful to consider such a description where the set of moves G is small. In all the concrete examples of Markov chains considered in Section 3.4, following [Con22], we will work with mapping representations satisfying the above conditions.

We now proceed to introduce in an abstract way the main inequality of interest in this paper. For this, we first need an additional ingredient, i.e. a weight function $\theta\colon \mathbb{R}_{>0} \times \mathbb{R}_{>0} \to \mathbb{R}_{\geq 0}$. In this paper, we always work under the following basic

Assumption 1. *The weight function θ is such that:*

1. *θ is not identically 0;*
2. *$\theta(s,t) = \theta(t,s)$;*
3. *θ is differentiable;*
4. *θ is concave.*

Given a weight function θ satisfying the above, the inequality reads

$$\mathscr{B}(\rho,\psi) \geq K\mathscr{A}(\rho,\psi) \tag{3.2.3}$$

for all positive functions $\rho\colon \Omega \to \mathbb{R}_{>0}$, functions $\psi\colon \Omega \to \mathbb{R}$ and for a constant $K \in \mathbb{R}$ independent of ρ, ψ, where we have

$$\mathscr{A}(\rho,\psi) = \frac{1}{2}\sum_{(\eta,\sigma)\in S} m(\eta)c(\eta,\sigma)\theta(\rho(\eta),\rho(\sigma\eta))[\psi(\eta)-\psi(\sigma\eta)]^2,$$
$$\mathscr{B}(\rho,\psi) = \mathscr{C}(\rho,\psi) - \mathscr{D}(\rho,\psi),$$

with

$$\mathscr{C}(\rho,\psi) = \frac{1}{4}\sum_{(\eta,\sigma)\in S} m(\eta)c(\eta,\sigma)\left\{\nabla\theta(\rho(\eta),\rho(\sigma\eta))\cdot\begin{pmatrix} L\rho(\eta) \\ L\rho(\sigma\eta)\end{pmatrix}\right\}[\psi(\eta)-\psi(\sigma\eta)]^2,$$
$$\mathscr{D}(\rho,\psi) = \frac{1}{2}\sum_{(\eta,\sigma)\in S} m(\eta)c(\eta,\sigma)\theta(\rho(\eta),\rho(\sigma\eta))(\psi(\eta)-\psi(\sigma\eta))(L\psi(\eta)-L\psi(\sigma\eta)).$$

This inequality was introduced in the work of Erbar and Maas [EM12, EM14]: depending on the choice of θ, it has different interpretations and consequences, which we discuss in the next subsections.

3.2.1 Logarithmic mean and entropic curvature

The main reason for studying inequality (3.2.3) comes from the work [EM12] and for choosing as θ the logarithmic mean

$$\theta_1(s,t) := \int_0^1 s^{1-p}t^p\,dp = \begin{cases} \frac{s-t}{\log s - \log t} & \text{if } s \neq t, \\ s & \text{if } s = t. \end{cases} \tag{3.2.4}$$

Let us now denote by $\mathcal{P}(\Omega)$ the set of probability densities on Ω with respect to m, i.e. functions $\rho\colon \Omega \to \mathbb{R}_{\geq 0}$ such that $\mathbb{E}_m(\rho) = 1$, and by $\mathcal{P}_*(\Omega)$ the set of strictly positive densities. In [Maa11], Maas introduced a Wasserstein-like metric $\mathscr{W}$ on $\mathcal{P}(\Omega)$ via a discrete variant of the Benamou–Brenier formula, and showed that, as in the classical setting, for any $\rho \in \mathcal{P}(\Omega)$ the heat flow $t \to P_t\rho = e^{tL}\rho$ is the gradient flow of the relative entropy functional in $(\mathcal{P}(\Omega), \mathscr{W})$ started at ρ, where the relative entropy is the restriction of the functional $\mathcal{H}^{\phi_1}$ (as in (3.1.1)) to $\mathcal{P}(\Omega)$. Writing it explicitly, for $\rho \in \mathcal{P}(\Omega)$ we have that

$$\mathcal{H}^{\phi_1}(\rho) = \sum_{\eta\in\Omega} m(\eta)\rho(\eta)\log(\rho(\eta)),$$

with the convention that $0\log 0 = 0$ in the above sum. In this setting, inequality (3.2.3) can be interpreted as a lower bound for the Hessian of $\mathcal{H}^{\phi_1}$ with respect to $\mathscr{W}$ (see [EM12, Thm. 4.5]), or equivalently as a statement of K-geodesic convexity (see also the independent work of Mielke [Mie13]). As gradient flows of geodesically K-convex functionals enjoy many useful properties, functional inequalities can be subsequently derived for the Markov chains. In the next Proposition we collect in particular some results proved in [EM12] (cf. Proposition 4.7 and Theorems $7.3, 7.4$ therein). To be precise, [EM12] considers actually the case where the generator corresponds to a Markov Kernel K (whose rows sum to 1), but these definitions and properties easily extend to non-normalised transition rates, as considered in the subsequent literature [EM14, FM16, EHMT17].

Proposition 3.2.1. *Assume that θ is the logarithmic mean and that inequality* (3.2.3) *holds for some constant $K \in \mathbb{R}$ and for all $\psi\colon \Omega \to \mathbb{R}$, $\rho \in \mathcal{P}_*(\Omega)$. Then:*

- *the* $\mathrm{HWI}(K)$ *inequality*

$$H^{\phi_1}(\rho) \leq \mathscr{W}(\rho,\mathbf{1})\sqrt{I^{\phi_1}(\rho)} - \frac{K}{2}\mathscr{W}(\rho,\mathbf{1})^2$$

  holds for all $\rho \in \mathcal{P}(\Omega)$;

- *for any $\rho, \sigma \in \mathcal{P}(\Omega)$*

$$\mathscr{W}(P_t\rho, P_t\sigma) \leq e^{-Kt}\mathscr{W}(\rho,\sigma);$$

- *if $K > 0$ then the modified log-Sobolev inequality* $\mathrm{MLSI}(2K)$

$$2K\mathcal{H}^{\phi_1}(\rho) \leq \mathcal{I}^{\phi_1}(\rho)$$

  holds for all $\rho \in \mathcal{P}_*(\Omega)$.

Following [EM12], when (3.2.3) holds for the logarithmic mean, we say that the *entropic curvature* of the Markov chain is bounded from below by K, and we use the notation

$$\mathrm{Ric}_{\mathrm{e}} \geq K.$$

For more consequences of entropic curvature lower bounds we refer the reader to [EM12, EF18].

3.2.2 Weight functions and convex Sobolev inequalities

In some cases, it is possible to let other ϕ-entropies take over the role of the relative entropy in the previous section, following [EM14] and by choosing an appropriate weight function θ. First, we consider a function $\phi\colon \mathbb{R}_{\geq 0} \to \mathbb{R}$ and correspondingly we make the following

Assumption 2. *The function $\phi\colon \mathbb{R}_{\geq 0} \to \mathbb{R}$ is such that*

1. *ϕ is continuous and $\phi \in C^2(\mathbb{R}_{>0})$;*
2. *ϕ is strictly convex.*

Moreover, the weight function $\theta = \theta_\phi$ defined by

$$\theta(s,t) := \begin{cases} \frac{s-t}{\phi'(s)-\phi'(t)} & \text{if } s \neq t, \\ \frac{1}{\phi''(s)} & \text{if } s = t \end{cases} \tag{3.2.5}$$

satisfies Assumption 1.

A first motivation for defining θ as in (3.2.5) comes from the following result due to [EM14] (see Theorem 4.8 therein).

Proposition 3.2.2. *Suppose that Assumption 2 is satisfied and that inequality* (3.2.3) *holds for some $K > 0$. Then the convex Sobolev inequality* (3.1.3) *holds with constant $2K$ (notation: $\mathrm{CSI}_\phi(2K)$).*

For completeness, we provide the proof of this proposition in Section 3.6, since it was proved in [EM14] under slightly more restrictive assumptions, as a consequence of stronger geodesic convexity results, as explained later in this section. The idea for the proof of Proposition 3.2.2 is that, when restricting to the specific choice $\psi := \phi' \circ \rho$, inequality (3.2.3) is equivalent to the second order differential inequality

$$\left.\frac{\mathrm{d}^2}{\mathrm{d}t^2}\right|_{t=0} \mathcal{H}^\phi(P_t\rho) \geq 2K\mathcal{I}^\phi(\rho). \tag{3.2.6}$$

From this, it is standard to deduce the convex Sobolev inequality (3.1.3) with constant $2K$, essentially by integration, following what is known as the "Bakry–Émery argument". Actually, this is exactly the approach used by Conforti in [Con22], that is, he uses couplings rates to establish (3.2.6) and subsequently deduces the convex Sobolev inequality $\mathrm{CSI}_\phi(2K)$. Since (3.2.6) is a particular case of (3.2.3) for a specific choice of ψ, it is clear that proving (3.2.3) under Assumption 2 gives a stronger result as compared to (3.2.6) and is in general more challenging to achieve as one has to deal with two unknown functions (ρ and ψ) as opposed to just one (i.e. ρ). Another difference with the work of Conforti lies in the assumptions on the convex function ϕ: indeed, our Assumption 2 requires that θ is concave. This assumption was present also in [JY17] and (as already observed there) implies in particular that $\frac{1}{\phi''}$ is concave, which is a classical assumption for the continuous setting. On the other hand, Conforti does not assume concavity of θ, but requires instead that the function

$$(s,t) \to (s-t)\cdot(\phi'(s) - \phi'(t)) \tag{3.2.7}$$

is convex.

While both assumptions are enough to deduce convex Sobolev inequalities, it is possible to make another more demanding one and deduce stronger consequences from inequality (3.2.3).

Assumption 3. *Assumption 2 is satisfied. Moreover, with θ as in* (3.2.5)*, we have that:*

- $\phi \in C^\infty(\mathbb{R}_{>0})$;
- $\theta \in C^\infty(\mathbb{R}_{>0} \times \mathbb{R}_{>0})$;
- *θ extends to a continuous function defined on $\mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0}$;*
- *$\theta(r,s) \leq \theta(r,t)$ for all $0 \leq s \leq t$ and $0 \leq r$.*

If the above is satisfied, then in [EM14] the authors showed it is possible to adapt some of the results of Section 3.2.1. Replacing the $\mathscr{W}$ metric with a suitable modified metric $\mathscr{W}_\phi$ (where the new weight function θ replaces the logarithmic mean), it holds that for any starting density $\rho \in \mathcal{P}(\Omega)$ the heat flow $t \to P_t\rho$ is the gradient flow of the ϕ-entropy $\mathcal{H}^\phi$ in $(\mathcal{P}(\Omega), \mathscr{W}_\phi)$. Moreover, as in the previous section, inequality (3.2.3) si equivalent to K-geodesic convexity of $\mathcal{H}^\phi$ in $(\mathcal{P}(\Omega), \mathscr{W}_\phi)$ and the following result holds (cf. Propositions 4.2, 4.6 and Theorem 4.8 in [EM14]).

Proposition 3.2.3. *Under Assumption 3, suppose that inequality* (3.2.3) *holds for some constant $K \in \mathbb{R}$ and for all $\psi \colon \Omega \to \mathbb{R}$, $\rho \in \mathcal{P}_*(\Omega)$. Then:*

- *the inequality*
$$H^\phi(\rho) \leq \mathscr{W}_\phi(\rho, \mathbf{1})\sqrt{I^\phi(\rho)} - \frac{K}{2}\mathscr{W}_\phi(\rho, \mathbf{1})^2$$
holds for all $\rho \in \mathcal{P}(\Omega)$;
- *for any $\rho, \sigma \in \mathcal{P}(\Omega)$*
$$\mathscr{W}_\phi(P_t\rho, P_t\sigma) \leq e^{-Kt}\mathscr{W}_\phi(\rho, \sigma);$$
- *if $K > 0$ then the ϕ-convex Sobolev inequality $\mathrm{CSI}_\phi(2K)$*
$$2K\mathcal{H}^\phi(\rho) \leq \mathcal{I}^\phi(\rho)$$
holds for all $\rho \in \mathcal{P}_*(\Omega)$.

All the functions ϕ_α defined in (3.1.4) satisfy Assumption 3 (see also [JY17, Lemma 16]): for the corresponding weight function, we use the notation θ_α. Notice in particular that for $\alpha = 1$ we recover the logarithmic mean of the previous subsection, while for $1 < \alpha < 2$ we have

$$\theta_\alpha(s,t) = \begin{cases} \frac{\alpha-1}{\alpha}\frac{s-t}{s^{\alpha-1}-t^{\alpha-1}} & \text{if } s \neq t, \\ \frac{1}{\alpha}s^{2-\alpha} & \text{if } s = t. \end{cases} \tag{3.2.8}$$

Remark 3.2.4 (Case $\alpha = 2$)**.** *The case $\alpha = 2$ is particular and should be studied separately. In this case, indeed, the weight function satisfies $\theta_2 \equiv \frac{1}{2}$. Therefore, the quantities $\mathscr{B}(\rho, \psi)$ and $\mathscr{A}(\rho, \psi)$ become independent of ρ, which makes establishing inequality* (3.2.3) *significantly simpler. Actually, for $\theta = \theta_2$ establishing* (3.2.3) *for all ρ, ψ is equivalent to establishing* (3.2.6) *for all ρ, as done by Conforti. Therefore, in this case it is usually possible to establish inequality* (3.2.3) *with a better constant than what would happen just under Assumption 1. In all the examples of Section 3.4, this can be achieved by a simple modification of the arguments after substituting $\theta \equiv \frac{1}{2}$, or alternatively, given the equivalence of* (3.2.3) *and* (3.2.6)*, by just applying the results of [Con22]. For this reasons, for the results of Section 3.4 applied to $\theta = \theta_\alpha$ we focus on $\alpha \in [1, 2)$ when comparing to [Con22].*

3.2.3 Arithmetic mean and discrete Bakry–Émery curvature

If θ is the arithmetic mean, it is well known that inequality (3.2.3) is equivalent to a lower bound for the discrete Bakry–Émery curvature (for example, it was already observed in [Maa17]). For completeness, and since we did not find a detailed proof in the literature, we recall the definitions and include a proof of this fact. In analogy with the classical setting discussed in great detail in [BGL14], for $f, g \colon \Omega \to \mathbb{R}$ define

$$\Gamma(f,g)(\eta) := \frac{1}{2}\sum_{\sigma\in G} c(\eta,\sigma)(f(\sigma\eta) - f(\eta))(g(\sigma\eta) - g(\eta)),$$

$\Gamma(f) := \Gamma(f,f)$ and

$$\Gamma_2(f) := \frac{1}{2}L\Gamma(f) - \Gamma(f, Lf).$$

Definition 3.2.5 ([Sch99]). *We say that the curvature condition* $\mathrm{CD}(K,\infty)$ *is satisfied if for all* $f \colon \Omega \to \mathbb{R}$

$$\Gamma_2(f) \geq K\Gamma(f). \tag{3.2.9}$$

Proposition 3.2.6. *Suppose that θ is the arithmetic mean. Then for any $K \in \mathbb{R}$ inequality* (3.2.3) *holds if and only if* $\mathrm{CD}(K,\infty)$ *holds.*

Proof. Notice that, using reversibility,

$$\begin{aligned}
\mathscr{A}(\rho,\psi) &= \frac{1}{4}\sum_{(\eta,\sigma)\in S} m(\eta)c(\eta,\sigma)(\rho(\eta)+\rho(\sigma\eta))[\psi(\eta)-\psi(\sigma\eta)]^2 \\
&= \frac{1}{2}\sum_{\eta\in\Omega} m(\eta)\rho(\eta)\left[\sum_{\sigma\in G} c(\eta,\sigma)[\psi(\eta)-\psi(\sigma\eta)]^2\right] \\
&= \sum_{\eta\in\Omega} m(\eta)\rho(\eta)\Gamma(\psi)(\eta).
\end{aligned}$$

Moreover, using reversibility multiple times (cf. also (3.2.2)),

$$\begin{aligned}
\mathscr{C}(\rho,\psi) &= \frac{1}{8}\sum_{(\eta,\sigma)\in S} m(\eta)c(\eta,\sigma)(L\rho(\eta)+L\rho(\sigma\eta))[\psi(\eta)-\psi(\sigma\eta)]^2\\
&= \frac{1}{4}\sum_{(\eta,\sigma)\in S} m(\eta)c(\eta,\sigma)L\rho(\eta)[\psi(\eta)-\psi(\sigma\eta)]^2\\
&= \frac{1}{4}\sum_{\substack{\eta\in\Omega\\ \sigma,\gamma\in G}} m(\eta)c(\eta,\sigma)c(\eta,\gamma)(\rho(\gamma\eta)-\rho(\eta))[\psi(\eta)-\psi(\sigma\eta)]^2\\
&= \frac{1}{4}\sum_{\substack{\eta\in\Omega\\ \sigma,\gamma\in G}} m(\eta)c(\gamma\eta,\sigma)c(\eta,\gamma)\rho(\eta)[\psi(\gamma\eta)-\psi(\sigma\gamma\eta)]^2\\
&- \frac{1}{4}\sum_{\substack{\eta\in\Omega\\ \sigma,\gamma\in G}} m(\eta)c(\eta,\sigma)c(\eta,\gamma)\rho(\eta)[\psi(\eta)-\psi(\sigma\eta)]^2\\
&= \frac{1}{4}\sum_{\eta\in\Omega} m(\eta)\rho(\eta)\sum_{\gamma\in G} c(\eta,\gamma)\sum_{\sigma\in G}\Big\{c(\gamma\eta,\sigma)[\psi(\gamma\eta)-\psi(\sigma\gamma\eta)]^2\\
&- c(\eta,\sigma)[\psi(\eta)-\psi(\sigma\eta)]^2\Big\}\\
&= \sum_{\eta\in\Omega} m(\eta)\rho(\eta)\frac{1}{2}L\Gamma\psi(\eta),
\end{aligned}$$

and

$$\begin{aligned}
\mathscr{D}(\rho,\psi) &= \frac{1}{4}\sum_{(\eta,\sigma)\in S} m(\eta)c(\eta,\sigma)(\rho(\eta)+\rho(\sigma\eta))(\psi(\eta)-\psi(\sigma\eta))(L\psi(\eta)-L\psi(\sigma\eta))\\
&= \frac{1}{2}\sum_{\eta\in\Omega} m(\eta)\rho(\eta)\sum_{\sigma\in G} c(\eta,\sigma)(\psi(\eta)-\psi(\sigma\eta))(L\psi(\eta)-L\psi(\sigma\eta))\\
&= \sum_{\eta\in\Omega} m(\eta)\rho(\eta)\Gamma(\psi,L\psi)(\eta).
\end{aligned}$$

Therefore, (3.2.3) is equivalent to

$$\sum_{\eta\in\Omega} m(\eta)\rho(\eta)\Gamma_2\psi(\eta) \geq K\sum_{\eta\in\Omega} m(\eta)\rho(\eta)\Gamma\psi(\eta). \tag{3.2.10}$$

From this, it is clear that $\mathrm{CD}(K,\infty)$ implies (3.2.3) by choosing $f=\psi$. Conversely, choosing $\rho=\frac{d\delta_\eta}{dm}$ to be the density of a Dirac and $\psi=f$ in (3.2.10) gives the converse implication. ■

For more details about and consequences of the discrete Bakry–Émery curvature we refer the reader to [FS18] and the references therein.

3.3 Coupling rates and curvature lower bound

Coupling rates were introduced by Conforti in [Con22] as a tool to establish convex Sobolev inequalities. They are a modification of the usual notion of coupling, and they apply to continuous time Markov chains. Roughly speaking, they are a way of letting the generator L act at the same time at two different states, in a way that is consistent with equation (3.2.1) when one looks separately at the two states. More precisely, for any pair of different states $\eta, \bar{\eta} \in \Omega$, we consider coupling rates between them in the form of a function $c^{\mathsf{cpl}}(\eta, \bar{\eta}, \cdot, \cdot) \colon G^* \times G^* \to \mathbb{R}_{\geq 0}$ such that

$$\forall \gamma \in G, \quad \sum_{\bar{\gamma} \in G^*} c^{\mathsf{cpl}}(\eta, \bar{\eta}, \gamma, \bar{\gamma}) = c(\eta, \gamma),$$
$$\forall \bar{\gamma} \in G, \quad \sum_{\gamma \in G^*} c^{\mathsf{cpl}}(\eta, \bar{\eta}, \gamma, \bar{\gamma}) = c(\bar{\eta}, \bar{\gamma}).$$

It can be seen easily seen that, for any fixed states $\eta \neq \bar{\eta}$, coupling rates between them always exist; for example, one can consider the "product coupling rates", constructed as follows. Suppose without loss of generality (otherwise, exchange the role of $\eta, \bar{\eta}$) that

$$\sum_{\sigma \in G} c(\eta, \sigma) \leq \sum_{\sigma \in G} c(\bar{\eta}, \sigma) =: M,$$

and notice $0 < M < \infty$ (since G is finite and the chain is irreducible). Then, set $c(\eta, e) = M - \sum_{\sigma \in G} c(\eta, \sigma)$ and $c(\bar{\eta}, e) = 0$, where e is the null move. Finally, for $\gamma, \bar{\gamma} \in G^*$, define

$$c^{\mathsf{cpl}}(\eta, \bar{\eta}, \gamma, \bar{\gamma}) = \frac{1}{M} c(\eta, \gamma) c(\bar{\eta}, \bar{\gamma}),$$

which is easily seen to define appropriate coupling rates between η and $\bar{\eta}$.

From the definition of coupling rates, it follows immediately that one can jointly express the action of the generator (3.2.1) on a function ψ at the states η and $\bar{\eta}$ as follows:

$$\begin{aligned} L\psi(\eta) &= \sum_{\gamma, \bar{\gamma} \in G^*} c^{\mathsf{cpl}}(\eta, \bar{\eta}, \gamma, \bar{\gamma})(\psi(\gamma\eta) - \psi(\eta)), \\ L\psi(\bar{\eta}) &= \sum_{\gamma, \bar{\gamma} \in G^*} c^{\mathsf{cpl}}(\eta, \bar{\eta}, \gamma, \bar{\gamma})(\psi(\bar{\gamma}\bar{\eta}) - \psi(\bar{\eta})). \end{aligned} \tag{3.3.1}$$

In [Con22], Conforti showed that coupling rates are useful for organizing the terms appearing in the inequality (3.2.6), and thus (if one manages to establish it), in proving convex Sobolev inequalities via the Bakry–Émery argument. Heuristically, it turns out it is convenient to consider not arbitrary coupling rates, but rather "contractive" ones. Informally, this means that, for neigbouring states $\eta \neq \sigma\eta$ with $\eta \in \Omega, \sigma \in G$, the coupling rates $c^{\mathsf{cpl}}(\eta, \sigma\eta, \gamma, \bar{\gamma})$ between them are such that "as often as possible" $\gamma\eta = \bar{\gamma}\sigma\eta$ (and, in particular, a fruitful choice is $(\gamma, \bar{\gamma}) = (\sigma, e)$ or (e, σ^{-1})). Indeed, when this is achieved, some terms cancellations going into the right direction occur when studying inequality (3.2.6).

In the rest of this section, we show that similar considerations also hold when studying the stronger inequality (3.2.3). In particular, we will derive a lower bound for $\mathscr{B}(\rho, \psi)$ using coupling rates under only Assumption 1 on θ: this gives a sufficient condition for establishing the inequality $\mathscr{B}(\rho, \psi) \geq K\mathscr{A}(\rho, \psi)$. In general, this is a more challenging situation compared to (3.2.6), since now we are dealing with two unknowns (ρ, ψ) as opposed to just ρ, and we also have an additional non linear weight function θ to deal with; moreover, as discussed in Section 3.2.2, inequality (3.2.6) corresponds to a particular case of (3.2.3) with the choice $\psi = \phi' \circ \rho$ and θ as in Assumption 2. Next, we will conclude the section by discussing heuristically how contractions in the coupling rates can help proving the inequality (3.2.3) too, similarly to what happened in [Con22].

We now proceed to show how arbitrary coupling rates can help rewrite the main inequality (3.2.3) in a convenient way, by organizing the involved terms. Notice first that we can write, using coupling rates as in (3.3.1),

$$\begin{aligned}
&\nabla\theta(\rho(\eta),\rho(\sigma\eta))\cdot\begin{pmatrix}L\rho(\eta)\\ L\rho(\sigma\eta)\end{pmatrix}\\
=&\nabla\theta(\rho(\eta),\rho(\sigma\eta))\cdot\left[\sum_{\gamma,\bar\gamma\in G^*}c^{\mathsf{cpl}}(\eta,\sigma\eta,\gamma,\bar\gamma)\begin{pmatrix}\rho(\gamma\eta)-\rho(\eta)\\ \rho(\bar\gamma\sigma\eta)-\rho(\sigma\eta)\end{pmatrix}\right]\\
=&\sum_{\gamma,\bar\gamma\in G^*}c^{\mathsf{cpl}}(\eta,\sigma\eta,\gamma,\bar\gamma)\nabla\theta(\rho(\eta),\rho(\sigma\eta))\cdot\left[\begin{pmatrix}\rho(\gamma\eta)\\ \rho(\bar\gamma\sigma\eta)\end{pmatrix}-\begin{pmatrix}\rho(\eta)\\ \rho(\sigma\eta)\end{pmatrix}\right]\\
\geq&\sum_{\gamma,\bar\gamma\in G^*}c^{\mathsf{cpl}}(\eta,\sigma\eta,\gamma,\bar\gamma)[\theta(\rho(\gamma\eta),\rho(\bar\gamma\sigma\eta))-\theta(\rho(\eta),\rho(\sigma\eta))],
\end{aligned}\tag{3.3.2}$$

where we used concavity of θ in the last line. Therefore, for $\mathscr{C}(\rho,\psi)$ we have the lower bound

$$\begin{aligned}
\mathscr{C}(\rho,\psi)\geq&\frac{1}{4}\sum_{(\eta,\sigma)\in S}\sum_{\gamma,\bar\gamma\in G^*}m(\eta)c(\eta,\sigma)\,c^{\mathsf{cpl}}(\eta,\sigma\eta,\gamma,\bar\gamma)\theta(\rho(\gamma\eta),\rho(\bar\gamma\sigma\eta))[\psi(\eta)-\psi(\sigma\eta)]^2\\
&-\frac{1}{4}\sum_{(\eta,\sigma)\in S}\sum_{\gamma,\bar\gamma\in G^*}m(\eta)c(\eta,\sigma)c^{\mathsf{cpl}}(\eta,\sigma\eta,\gamma,\bar\gamma)\theta(\rho(\eta),\rho(\sigma\eta))[\psi(\eta)-\psi(\sigma\eta)]^2.
\end{aligned}$$

As for the term $\mathscr{D}(\rho,\psi)$, we can write

$$\begin{aligned}
\mathscr{D}(\rho,\psi)=&\frac{1}{2}\sum_{(\eta,\sigma)\in S}\sum_{\gamma,\bar\gamma\in G^*}m(\eta)c(\eta,\sigma)c^{\mathsf{cpl}}(\eta,\sigma\eta,\gamma,\bar\gamma)\theta(\rho(\eta),\rho(\sigma\eta))\\
&\cdot\Big\{(\psi(\eta)-\psi(\sigma\eta))(\psi(\gamma\eta)-\psi(\bar\gamma\sigma\eta))-(\psi(\eta)-\psi(\sigma\eta))^2\Big\}.
\end{aligned}$$

Combining the bound for $\mathscr{C}$ and the expression for $\mathscr{D}$ we derive the following:

Lemma 3.3.1. *Let θ be a weight function satisfying Assumption 1. We have*

$$\mathscr{B}(\rho,\psi)\geq\frac{1}{4}\sum_{(\eta,\sigma)\in S}\sum_{\gamma,\bar\gamma\in G^*}m(\eta)c(\eta,\sigma)\,c^{\mathsf{cpl}}(\eta,\sigma\eta,\gamma,\bar\gamma)J(\eta,\sigma,\gamma,\bar\gamma)\tag{3.3.3}$$

for all $\rho\colon\Omega\to\mathbb{R}_{>0}$ and $\psi\colon\Omega\to\mathbb{R}$, where we define the function $J\colon\Omega\times G^*\times G^*\times G^*\to\mathbb{R}$ by

$$\begin{aligned}
J(\eta,\sigma,\gamma,\bar\gamma):=&\{\theta(\rho(\gamma\eta),\rho(\bar\gamma\sigma\eta))+\theta(\rho(\eta),\rho(\sigma\eta))\}[\psi(\eta)-\psi(\sigma\eta)]^2\\
&-2\theta(\rho(\eta),\rho(\sigma\eta))(\psi(\eta)-\psi(\sigma\eta))(\psi(\gamma\eta)-\psi(\bar\gamma\sigma\eta)).
\end{aligned}$$

It is also convenient to define the function $I\colon\Omega\times G^*\times G^*\times G^*\to\mathbb{R}$ by

$$\begin{aligned}
I(\eta,\sigma,\gamma,\bar\gamma)&=I_1(\eta,\sigma,\gamma,\bar\gamma)-I_2(\eta,\sigma,\gamma,\bar\gamma),\\
I_1(\eta,\sigma,\gamma,\bar\gamma)&=\theta(\rho(\gamma\eta),\rho(\bar\gamma\sigma\eta))[\psi(\eta)-\psi(\sigma\eta)]^2\geq 0,\\
I_2(\eta,\sigma,\gamma,\bar\gamma)&=\theta(\rho(\eta),\rho(\sigma\eta))[\psi(\gamma\eta)-\psi(\bar\gamma\sigma\eta)]^2\geq 0.
\end{aligned}$$

Notice that we have

$$\begin{aligned}
J(\eta,\sigma,\gamma,\bar\gamma)&=I(\eta,\sigma,\gamma,\bar\gamma)+\theta(\rho(\eta),\rho(\sigma\eta))[\psi(\eta)-\psi(\sigma\eta)-\psi(\gamma\eta)+\psi(\bar\gamma\sigma\eta)]^2\\
&\geq I(\eta,\sigma,\gamma,\bar\gamma).
\end{aligned}\tag{3.3.4}$$

At this point, we can explain at least heuristically why it is useful to consider especially *contractive* coupling rates. In view of Lemma 3.3.1, to establish inequality (3.2.3) it suffices to prove that for some coupling rates

$$\begin{aligned}&\frac{1}{2}\sum_{(\eta,\sigma)\in S}\sum_{\gamma,\bar\gamma\in G^*} m(\eta)c(\eta,\sigma)\, c^{\mathsf{cpl}}(\eta,\sigma\eta,\gamma,\bar\gamma)J(\eta,\sigma,\gamma,\bar\gamma)\\ &\geq K\sum_{(\eta,\sigma)\in S} m(\eta)c(\eta,\sigma)\theta(\rho(\eta),\rho(\sigma\eta))[\psi(\sigma\eta)-\psi(\eta)]^2 = 2K\,\mathscr{A}(\rho,\psi).\end{aligned} \tag{3.3.5}$$

Notice first of all that whenever $\gamma\eta=\bar\gamma\sigma\eta$ the second term in the definition of $J(\eta,\sigma\eta,\gamma,\bar\gamma)$ is 0, so J is non-negative, suggesting a first lower bound for $\mathscr{B}(\rho,\psi)$. More precisely, when $\gamma\eta=\bar\gamma\sigma\eta$, looking at the corresponding terms in the left-hand-side of the inequality (3.3.5), we see that

$$\begin{aligned}&c^{\mathsf{cpl}}(\eta,\sigma\eta,\gamma,\bar\gamma)J(\eta,\sigma,\gamma,\bar\gamma)\\ &= c^{\mathsf{cpl}}(\eta,\sigma\eta,\gamma,\bar\gamma)\{\theta(\rho(\gamma\eta),\rho(\bar\gamma\sigma\eta))+\theta(\rho(\eta),\rho(\sigma\eta))\}[\psi(\eta)-\psi(\sigma\eta)]^2\\ &\geq c^{\mathsf{cpl}}(\eta,\sigma\eta,\gamma,\bar\gamma)\theta(\rho(\eta),\rho(\sigma\eta))[\psi(\eta)-\psi(\sigma\eta)]^2.\end{aligned} \tag{3.3.6}$$

Hence, we recognise some terms appearing in the sum in the right-and-side of (3.3.5) defining $\mathscr{A}(\rho,\psi)$, multiplied by the factor $c^{\mathsf{cpl}}(\eta,\sigma\eta,\gamma,\bar\gamma)$: therefore, if we have a uniform positive lower bound for

$$\inf_{(\eta,\sigma)\in S}\sum_{\substack{\gamma,\bar\gamma\in G^*\\ \gamma\eta=\bar\gamma\sigma\eta}} c^{\mathsf{cpl}}(\eta,\sigma\eta,\gamma,\bar\gamma) > 0, \tag{3.3.7}$$

we are in a good position to prove the inequality (3.3.5), provided that we can also accomplish the non-trivial task of dealing with the other terms appearing in the left-hand-side of (3.3.5) (corresponding to the pairs of moves $(\gamma,\bar\gamma)$ not realising a contraction). This is indeed the general strategy that we will use in Section 3.4, where we analyse specific classes of Markov chains.

A second point we wish to make is that sometimes, depending also on the weight function θ, we can improve the bounds obtained with the strategy described before. The first observation is that in (3.3.6) we have thrown away some non-negative terms, corresponding to

$$c^{\mathsf{cpl}}(\eta,\sigma\eta,\gamma,\bar\gamma)\theta(\rho(\gamma\eta),\rho(\bar\gamma\sigma\eta))[\psi(\eta)-\psi(\sigma\eta)]^2. \tag{3.3.8}$$

We now restrict our attention to two particular pairs of moves that are "contractive", given respectively by (e,σ^{-1}) and (σ,e). In this case, the terms in (3.3.8) sum up to

$$\Big\{c^{\mathsf{cpl}}(\eta,\sigma\eta,e,\sigma^{-1})\theta(\rho(\eta),\rho(\eta))+c^{\mathsf{cpl}}(\eta,\sigma\eta,\sigma,e)\theta(\rho(\sigma\eta),\rho(\sigma\eta))\Big\}[\psi(\eta)-\psi(\sigma\eta)]^2.$$

These terms could also be related to the ones appearing in the definition of $\mathscr{A}(\rho,\psi)$, if we knew that for some constant M_θ

$$\theta(\rho(\eta),\rho(\eta))+\theta(\rho(\sigma\eta),\rho(\sigma\eta))\geq 2M_\theta\theta(\rho(\eta),\rho(\sigma\eta)),$$

and if we had a uniform positive lower bound

$$\inf_{(\eta,\sigma)\in S}\min\Big\{c^{\mathsf{cpl}}(\eta,\sigma\eta,\sigma,e),c^{\mathsf{cpl}}(\eta,\sigma\eta,e,\sigma^{-1})\Big\}>0, \tag{3.3.9}$$

similarly to (3.3.7). For this reason, for a given weight function θ satisfying Assumption 1, it is natural to define the quantity

$$M_\theta := \inf_{\substack{s,t>0:\\ \theta(s,t)>0}} \frac{\theta(s,s)+\theta(t,t)}{2\theta(s,t)} \in [0,1], \tag{3.3.10}$$

so that for all $s, t \geq 0$

$$2M_\theta \theta(s,t) \leq \theta(s,s) + \theta(t,t).$$

By choosing $s = t$ we can see that $M_\theta \leq 1$. The next proposition, whose proof is given in Section 3.7, provides the value of M_θ for the explicit examples of θ considered in Section 3.2.

Proposition 3.3.2.
- *For $\alpha \in [1,2]$ and θ_α as in equations* (3.2.4), (3.2.8), *we have*

$$M_{\theta_\alpha} = \begin{cases} 1 & \text{if } \alpha \in \left[1, \frac{3}{2}\right] \text{ or } \alpha = 2; \\ \frac{1}{2(\alpha-1)} & \text{if } \alpha \in \left(\frac{3}{2}, 2\right). \end{cases}$$

- *For the arithmetic mean we have $M_\theta = 1$.*

As a concluding remark for this section, we emphasize that, while the method described in this section potentially applies to a wide variety of settings, in general it seems that some extra assumptions on the Markov chains are helpful to get the desired conclusions. In particular, reversibility of the model and an underlying symmetry of the structure of the Markov chain can help obtain useful terms cancellations to deal with the "non-contractive" pairs of moves $(\gamma, \bar{\gamma})$ in the left-hand-side of (3.3.5).

3.4 Applications

In this section, we apply Lemma 3.3.1 to establish the general inequality of interest (3.2.3) for most of the examples considered in [Con22], under just Assumption 1. In particular, Section 3.4.1, 3.4.2, 3.4.3 and 3.4.4 corresponds to Section 4, 5.1, 5.2 and 3 of [Con22] respectively. Not surprisingly, the proofs are similar to the ones of Conforti, and in all these examples the considered contractive coupling rates are the ones constructed in [Con22]. The case of the interacting random walks of Section 3.4.4 is the only one where an additional assumption is present compared to [Con22]: moreover, as done by Conforti, in that section a localization procedure is used to deal with the infinite cardinality of the state space.

3.4.1 Glauber dynamics

We work in the setting of Section 4 of [Con22] (i.e. Glauber dynamics) and we use the same notation, which we briefly recall. The state space is a finite set Ω. We assume that $\sigma = \sigma^{-1}$ and $\sigma\gamma\eta = \gamma\sigma\eta$ for all moves $\sigma, \gamma \in G$ and states $\eta \in \Omega$. Given an inverse temperature parameter $\beta > 0$ and an Hamiltonian function $H\colon \Omega \to \mathbb{R}$, the rates are defined by

$$c(\eta, \sigma) = \exp\left(-\frac{\beta}{2}\nabla_\sigma H(\eta)\right),$$

where we recall the notation $\nabla_\sigma H(\eta) = H(\sigma\eta) - H(\eta)$ for the discrete gradient. The reversible measure is then the Gibbs measure

$$m(\eta) = \frac{1}{Z_\beta}\exp(-\beta H(\eta))$$

where $Z_\beta > 0$ is the appropriate normalization constant. Finally we make the key assumption that $\kappa(\eta,\sigma) \geq 0$ for all states $\eta \in \Omega$ and moves $\sigma \in G$, where we define

$$\kappa(\eta,\sigma) := c(\sigma\eta,\sigma) - \sum_{\gamma:\gamma\neq\sigma} \max\{-\nabla_\sigma c(\eta,\gamma), 0\}.$$

This assumption is crucial for the construction of appropriate contractive coupling rates, for which we will apply Lemma 3.3.1. We also define the quantities

$$\kappa_* := \inf_{\eta,\sigma} \kappa(\eta,\sigma) + \kappa(\sigma\eta,\sigma), \qquad \overline{\kappa_*} := \inf_{\eta,\sigma} \kappa(\eta,\sigma),$$

which correspond respectively to the infimums in (3.3.7) and (3.3.9). Notice that $2\overline{\kappa_*} \leq \kappa_*$.

Theorem 3.4.1. *With the previous notation, suppose that for all $\eta \in \Omega$ and $\sigma, \gamma \in G$ we have $\sigma\gamma = \gamma\sigma$, $\sigma = \sigma^{-1}$ and $\kappa(\eta,\sigma) \geq 0$. Let θ be a weight function satisfying Assumption 1. Then the inequality* (3.2.3) *holds with constant*

$$K = M_\theta \overline{\kappa_*} + \frac{\kappa_*}{2}.$$

Remark 3.4.2 (Comparison with [Con22])**.** *In [Con22, Thm. 4.1], under the same assumptions on the model, Conforti establishes inequality* (3.2.6) *and thus $\mathrm{CSI}_\phi(2K)$ with constant K equal to*

- $\frac{\kappa_*}{2}$ *for general convex ϕ satisfying convexity of* (3.2.7)*;*
- $\frac{\kappa_*}{2} + \overline{\kappa_*}$ *for $\phi = \phi_1$ (thus $\mathrm{MLSI}(\kappa_* + 2\overline{\kappa_*})$);*
- $\frac{\alpha}{2}\kappa_*$ *for $\phi = \phi_\alpha$ with $\alpha \in (1,2]$.*

Thus, by Proposition 3.3.2 and by the discussion in Section 3.2 (i.e. recalling for example Proposition 3.2.1 and that (3.2.6) *is particular case of* (3.2.3)*), we obtain a stronger result for the case $\theta = \theta_1$ and complementary results for other choices of θ.*

Proof of Theorem 3.4.1

As done in [Con22], we define

$$\begin{aligned}
\Upsilon^<(\eta) &= \{(\sigma,\gamma) \in G \times G \mid \sigma \neq \gamma, \nabla_\sigma c(\eta,\gamma) < 0\},\\
\Upsilon^>(\eta) &= \{(\sigma,\gamma) \in G \times G \mid \sigma \neq \gamma, \nabla_\sigma c(\eta,\gamma) > 0\},\\
\Upsilon^=(\eta) &= \{(\sigma,\gamma) \in G \times G \mid \sigma \neq \gamma, \nabla_\sigma c(\eta,\gamma) = 0\},
\end{aligned}$$

where we recall the notation

$$\begin{aligned}
\nabla_\sigma c(\eta,\gamma) &= c(\sigma\eta,\gamma) - c(\eta,\gamma)\\
&= \exp\left(-\frac{\beta}{2}[H(\gamma\sigma\eta) - H(\sigma\eta)]\right) - \exp\left(-\frac{\beta}{2}[H(\gamma\eta) - H(\eta)]\right).
\end{aligned}$$

We then define the same coupling rates: for $\eta \in \Omega, \sigma \in G$ set

$$c^{\mathsf{cpl}}(\eta,\sigma\eta,\gamma,\bar\gamma) = \begin{cases}
\min\{c(\sigma\eta,\gamma), c(\eta,\gamma)\} & \text{if } \gamma = \bar\gamma \text{ and } \sigma \neq \gamma, \gamma \in G,\\
-\nabla_\sigma c(\eta,\gamma) & \text{if } \bar\gamma = \sigma \text{ and } (\sigma,\gamma) \in \Upsilon^<(\eta),\\
\nabla_\sigma c(\eta,\bar\gamma) & \text{if } \gamma = \sigma \text{ and } (\sigma,\bar\gamma) \in \Upsilon^>(\eta),\\
\kappa(\sigma\eta,\sigma) & \text{if } \gamma = \sigma, \bar\gamma = e,\\
\kappa(\eta,\sigma) & \text{if } \gamma = e, \bar\gamma = \sigma,\\
0 & \text{otherwise.}
\end{cases}$$

Notice that these are indeed admissible coupling rates between η and $\sigma\eta$, since by assumption $\kappa(\sigma\eta,\sigma),\kappa(\eta,\sigma)\geq 0$. With these coupling rates and using Lemma 3.3.1 and (3.3.4) we can write $\mathscr{B}(\rho,\psi)\geq\frac{1}{4}(A+B+C+D)$ with

$$A=\sum_{\substack{\eta\in\Omega,\sigma,\gamma\in G,\\ \sigma\neq\gamma}} m(\eta)c(\eta,\sigma)\min\{c(\eta,\gamma),c(\sigma\eta,\gamma)\}I(\eta,\sigma,\gamma,\gamma),$$

$$B=-\sum_{\substack{\eta\in\Omega,\\ (\sigma,\gamma)\in\Upsilon^{<}(\eta)}} m(\eta)c(\eta,\sigma)\nabla_\sigma c(\eta,\gamma)I(\eta,\sigma,\gamma,\sigma),$$

$$C=\sum_{\substack{\eta\in\Omega,\\ (\sigma,\bar\gamma)\in\Upsilon^{>}(\eta)}} m(\eta)c(\eta,\sigma)\nabla_\sigma c(\eta,\bar\gamma)I(\eta,\sigma,\sigma,\bar\gamma),$$

$$D=\sum_{\eta\in\Omega,\sigma\in G} m(\eta)c(\eta,\sigma)(\kappa(\sigma\eta,\sigma)J(\eta,\sigma,\sigma,e)+\kappa(\eta,\sigma)J(\eta,\sigma,e,\sigma)).$$

We show below that $A=B=C=0$ and that $D\geq(4M_\theta\overline{\kappa_*}+2\kappa_*)\mathscr{A}(\rho,\psi)$, which concludes the proof of the theorem. It is useful to have an auxiliary lemma:

Lemma 3.4.3. *For all $\eta\in\Omega$ and $\sigma,\gamma\in G$ with $\sigma\neq\gamma$ the following hold:*

1. $c(\eta,\sigma)\nabla_\sigma c(\eta,\gamma)=c(\eta,\gamma)\nabla_\gamma c(\eta,\sigma)$.
2. $c(\eta,\sigma)c(\sigma\eta,\gamma)=c(\eta,\gamma)c(\gamma\eta,\sigma)$.
3. $\nabla_\sigma c(\sigma\eta,\gamma)=-\nabla_\sigma c(\eta,\gamma)$.
4. $(\sigma,\gamma)\in\Upsilon^{<}(\eta)\iff(\gamma,\sigma)\in\Upsilon^{<}(\eta)$.
5. $(\sigma,\gamma)\in\Upsilon^{>}(\gamma\eta)\iff(\sigma,\gamma)\in\Upsilon^{<}(\eta)$.
6. $(\sigma,\gamma)\in\Upsilon^{>}(\sigma\eta)\iff(\sigma,\gamma)\in\Upsilon^{<}(\eta)$.
7. $(\sigma,\gamma)\in\Upsilon^{=}(\gamma\eta)\iff(\sigma,\gamma)\in\Upsilon^{=}(\eta)$.
8. $I(\eta,\sigma,\gamma,\sigma)=-I(\eta,\gamma,\sigma,\gamma)$.
9. $I(\sigma\eta,\sigma,\sigma,\gamma)=-I(\gamma\eta,\gamma,\gamma,\sigma)$.
10. $I(\eta,\sigma,\gamma,\gamma)=-I(\gamma\eta,\sigma,\gamma,\gamma)$.

Proof of Lemma. Statements 1–7 were already observed in the proof of [Con22, Thm. 4.1] and are easy to check, while statements 8–10 are immediate from the definitions. ■

Term D We have

$$J(\eta,\sigma,\sigma,e)=\{\theta(\rho(\sigma\eta),\rho(\sigma\eta))+\theta(\rho(\eta),\rho(\sigma\eta))\}[\psi(\eta)-\psi(\sigma\eta)]^2$$
$$J(\eta,\sigma,e,\sigma)=\{\theta(\rho(\eta),\rho(\eta))+\theta(\rho(\eta),\rho(\sigma\eta))\}[\psi(\eta)-\psi(\sigma\eta)]^2$$

and so

$$\begin{aligned}&\kappa(\sigma\eta,\sigma)J(\eta,\sigma,\sigma,e)+\kappa(\eta,\sigma)J(\eta,\sigma,e,\sigma)\\ \geq&\{\overline{\kappa_*}[\theta(\rho(\sigma\eta),\rho(\sigma\eta))+\theta(\rho(\eta),\rho(\eta))]+\kappa_*\theta(\rho(\eta),\rho(\sigma\eta))\}[\psi(\eta)-\psi(\sigma\eta)]^2\\ \geq&(2M_\theta\overline{\kappa_*}+\kappa_*)\theta(\rho(\eta),\rho(\sigma\eta))[\psi(\eta)-\psi(\sigma\eta)]^2.\end{aligned}$$

Therefore we get

$$D\geq(4M_\theta\overline{\kappa_*}+2\kappa_*)\mathscr{A}(\rho,\psi).$$

Term B We have that

$$\begin{aligned}
B &= -\sum_{\substack{\eta\in\Omega,\\(\sigma,\gamma)\in\Upsilon^{<}(\eta)}} m(\eta)c(\eta,\sigma)\nabla_\sigma c(\eta,\gamma)I(\eta,\sigma,\gamma,\sigma)\\
&= -\sum_{\substack{\eta\in\Omega,\\(\gamma,\sigma)\in\Upsilon^{<}(\eta)}} m(\eta)c(\eta,\sigma)\nabla_\sigma c(\eta,\gamma)I(\eta,\sigma,\gamma,\sigma)\\
&= -\sum_{\substack{\eta\in\Omega,\\(\sigma,\gamma)\in\Upsilon^{<}(\eta)}} m(\eta)c(\eta,\gamma)\nabla_\gamma c(\eta,\sigma)I(\eta,\gamma,\sigma,\gamma)\\
&= -\sum_{\substack{\eta\in\Omega,\\(\sigma,\gamma)\in\Upsilon^{<}(\eta)}} m(\eta)c(\eta,\sigma)\nabla_\sigma c(\eta,\gamma)I(\eta,\gamma,\sigma,\gamma)\\
&= \sum_{\substack{\eta\in\Omega,\\(\sigma,\gamma)\in\Upsilon^{<}(\eta)}} m(\eta)c(\eta,\sigma)\nabla_\sigma c(\eta,\gamma)I(\eta,\sigma,\gamma,\sigma)\\
&= -B
\end{aligned}$$

which implies that $B=0$. In the above, the second equality is by 4. of Lemma 3.4.3, the third by exchanging the role of σ and γ, the fourth by 1. of Lemma 3.4.3 and the fifth by 8. of Lemma 3.4.3.

Term C This is similar to term B using reversibility. We have

$$C = \sum_{\substack{\eta\in\Omega,\\(\sigma,\bar\gamma)\in\Upsilon^{>}(\eta)}} m(\eta)c(\eta,\sigma)\nabla_\sigma c(\eta,\bar\gamma)I(\eta,\sigma,\sigma,\bar\gamma).$$

Using the reversibility property (3.2.2) with $F(\eta,\sigma)=\sum_{\bar\gamma:(\sigma,\bar\gamma)\in\Upsilon^{>}(\eta)}\nabla_\sigma c(\eta,\bar\gamma)I(\eta,\sigma,\sigma,\bar\gamma)$, the assumption $\sigma=\sigma^{-1}$ and properties 3. and 6. of Lemma 3.4.3, we get

$$C = -\sum_{\substack{\eta\in\Omega,\\(\sigma,\bar\gamma)\in\Upsilon^{<}(\eta)}} m(\eta)c(\eta,\sigma)\nabla_\sigma c(\eta,\bar\gamma)I(\sigma\eta,\sigma,\sigma,\bar\gamma).$$

We want to show that this expression is 0, analogously to what was done for B. Notice that

$$\begin{aligned}
C &= -\sum_{\substack{\eta\in\Omega,\\(\sigma,\bar\gamma)\in\Upsilon^{<}(\eta)}} m(\eta)c(\eta,\sigma)\nabla_\sigma c(\eta,\bar\gamma)I(\sigma\eta,\sigma,\sigma,\bar\gamma)\\
&= -\sum_{\substack{\eta\in\Omega,\\(\bar\gamma,\sigma)\in\Upsilon^{<}(\eta)}} m(\eta)c(\eta,\sigma)\nabla_\sigma c(\eta,\bar\gamma)I(\sigma\eta,\sigma,\sigma,\bar\gamma)\\
&= -\sum_{\substack{\eta\in\Omega,\\(\sigma,\bar\gamma)\in\Upsilon^{<}(\eta)}} m(\eta)c(\eta,\bar\gamma)\nabla_{\bar\gamma} c(\eta,\sigma)I(\bar\gamma\eta,\bar\gamma,\bar\gamma,\sigma)\\
&= -\sum_{\substack{\eta\in\Omega,\\(\sigma,\bar\gamma)\in\Upsilon^{<}(\eta)}} m(\eta)c(\eta,\sigma)\nabla_\sigma c(\eta,\bar\gamma)I(\bar\gamma\eta,\bar\gamma,\bar\gamma,\sigma)\\
&= \sum_{\substack{\eta\in\Omega,\\(\sigma,\bar\gamma)\in\Upsilon^{<}(\eta)}} m(\eta)c(\eta,\sigma)\nabla_\sigma c(\eta,\bar\gamma)I(\sigma\eta,\sigma,\sigma,\bar\gamma)\\
&= -C,
\end{aligned}$$

which implies that $C=0$. In the above, the second equality is by 4. of Lemma 3.4.3, the third by exchanging the role of σ and $\bar\gamma$, the fourth by 1. of Lemma 3.4.3 and the fifth by 9. of Lemma 3.4.3.

Term A We split A into three different terms: $A = A_1 + A_2 + A_3$, where

$$A_1 = \sum_{\substack{\eta\in\Omega,\\(\sigma,\gamma)\in\Upsilon^<(\eta)}} m(\eta)c(\eta,\sigma)c(\sigma\eta,\gamma)I(\eta,\sigma,\gamma,\gamma),$$
$$A_2 = \sum_{\substack{\eta\in\Omega,\\(\sigma,\gamma)\in\Upsilon^>(\eta)}} m(\eta)c(\eta,\sigma)c(\eta,\gamma)I(\eta,\sigma,\gamma,\gamma),$$
$$A_3 = \frac{1}{2}\sum_{\substack{\eta\in\Omega,\\(\sigma,\gamma)\in\Upsilon^=(\eta)}} m(\eta)c(\eta,\sigma)(c(\sigma\eta,\gamma)+c(\eta,\gamma))I(\eta,\sigma,\gamma,\gamma).$$

We want to show that $A_1 + A_2 = 0$ and $A_3 = 0$. We have that

$$\begin{aligned}A_2 &= \sum_{\substack{\eta\in\Omega,\\(\sigma,\gamma)\in\Upsilon^>(\eta)}} m(\eta)c(\eta,\sigma)c(\eta,\gamma)I(\eta,\sigma,\gamma,\gamma)\\ &= \sum_{\substack{\eta\in\Omega,\\(\sigma,\gamma)\in\Upsilon^>(\gamma\eta)}} m(\eta)c(\eta,\gamma)c(\gamma\eta,\sigma)I(\gamma\eta,\sigma,\gamma,\gamma)\\ &= -\sum_{\substack{\eta\in\Omega,\\(\sigma,\gamma)\in\Upsilon^<(\eta)}} m(\eta)c(\eta,\sigma)c(\sigma\eta,\gamma)I(\eta,\sigma,\gamma,\gamma)\\ &= -A_1.\end{aligned}$$

In the above, the second equality is by the reversibility property (3.2.2) with $F(\eta,\gamma) = \sum_{\sigma:(\sigma,\gamma)\in\Upsilon^>(\eta)} c(\eta,\sigma)I(\eta,\sigma,\gamma,\gamma)$ and the assumption $\gamma = \gamma^{-1}$, while the second equality is by the properties $2.$, $5.$ and $10.$ of Lemma 3.4.3. It follows that the contribution of the first two terms is 0.
It remains to show that $A_3 = 0$, which is done in a similar way: notice that

$$\begin{aligned}&\sum_{\substack{\eta\in\Omega,\\(\sigma,\gamma)\in\Upsilon^=(\eta)}} m(\eta)c(\eta,\sigma)c(\eta,\gamma)I(\eta,\sigma,\gamma,\gamma)\\ =&\sum_{\substack{\eta\in\Omega,\\(\sigma,\gamma)\in\Upsilon^=(\gamma\eta)}} m(\eta)c(\eta,\gamma)c(\gamma\eta,\sigma)I(\gamma\eta,\sigma,\gamma,\gamma)\\ =&-\sum_{\substack{\eta\in\Omega,\\(\sigma,\gamma)\in\Upsilon^=(\eta)}} m(\eta)c(\eta,\sigma)c(\sigma\eta,\gamma)I(\eta,\sigma,\gamma,\gamma).\end{aligned}$$

In the above, the first equality is by the reversibility property (3.2.2) with the function $F(\eta,\gamma) = \sum_{\sigma:(\sigma,\gamma)\in\Upsilon^=(\eta)} c(\eta,\sigma)I(\eta,\sigma,\gamma,\gamma)$ and by the assumption $\gamma = \gamma^{-1}$, while the second equality holds by properties $2.$, $7.$ and $10.$ of Lemma 3.4.3. It follows that $A_3 = 0$, thus concluding the proof of the theorem.

Examples of Glauber dynamics

Below, following [Con22], we present two examples of Glauber dynamics models satisfying the assumptions of Theorem 3.4.1.

Curie Weiss model For the Curie Weiss model, the state space is the discrete hypercube $\Omega = \{-1,1\}^N$ for some integer $N > 0$. The set of moves G is given by $G = \{\sigma_1,\ldots,\sigma_N\}$

where $\sigma_i \colon \Omega \to \Omega$ corresponds to flipping the i-th bit. Finally, the Hamiltonian function is

$$H(\eta) \coloneqq -\frac{1}{2N}\sum_{i,j}\eta_i\eta_j.$$

In this setting, the assumptions of Theorem 3.4.1 and the explicit values of $\kappa_*, \overline{\kappa_*}$ were checked by Conforti, who proves the following

Theorem 3.4.4 (Thm. 4.2 of [Con22])**.** *Assume that*

$$(N-1)\Big(e^{\frac{2\beta}{N}}-1\Big) \leq 1.$$

Then the assumptions of Theorem 3.4.1 are satisfied with

$$\begin{aligned}\kappa_* &= f_{\mathsf{CW},\beta,N}\Big(\Big\lfloor\frac{N-1}{2}\Big\rfloor\Big),\\ \overline{\kappa_*} &= e^{-\frac{\beta}{N}(N-1)}\Big[1-(N-1)\Big(1-e^{\frac{2\beta}{N}}\Big)\Big],\end{aligned}$$

where $f_{\mathsf{CW},\beta,N} \colon \mathbb{N} \to \mathbb{R}$ is defined by

$$\begin{aligned}f_{\mathsf{CW},\beta,N}(m) \coloneqq{}& e^{-\frac{\beta}{N}(N-1-2m)}\Big[1-(N-1-m)\Big(e^{\frac{2\beta}{N}}-1\Big)\Big]\\ &+e^{\frac{\beta}{N}(N-1-2m)}\Big[1-m\Big(e^{\frac{2\beta}{N}}-1\Big)\Big].\end{aligned}$$

Remark 3.4.5 (Comparison with Cor. 4.5 of [EHMT17])**.** *In particular, in the limit $N \to \infty$, the condition above reads $\beta \leq \frac{1}{2}$. Thus by choosing $\theta = \theta_1$ and combining Theorems 3.4.1 and 3.4.4 we have that as $N \to \infty$ the entropic curvature of the Curie–Weiss model satisfies*

$$\mathrm{Ric}_{\mathrm{e}} \geq (1-\beta) + (1-2\beta)e^{-\beta}$$

for $\beta \leq \frac{1}{2}$. This improves both in the estimate and in the range of admissible β over [EHMT17, Cor. 4.5], where it was proved that, as $N \to \infty$, $\mathrm{Ric}_{\mathrm{e}} \geq 2(1-2\beta e^{2\beta})e^{-\beta}$ for $\beta \lessapprox 0.284$.

Ising model The second example of Glauber dynamics that we consider is the Ising model. For the Ising model, we let $\Lambda \subset \mathbb{Z}^d$ be a connected subset of $\mathbb{Z}^d$, endowed with the inherited graph structure $\sim$ of the discrete grid, and consider the state space $\Omega = \{-1,1\}^{\Lambda}$. The set of moves is $G = \{\sigma_x\}_{x\in\Lambda}$ where $\sigma_x \colon \Omega \to \Omega$ acts on a state η by flipping the spin η_x at site x. Finally, the Hamiltonian is defined by

$$H(\eta) \coloneqq -\frac{1}{2}\sum_{x\sim y}\eta_x\eta_y.$$

Again, the assumptions and values of $\kappa_*, \overline{\kappa_*}$ in Theorem 3.4.1 were checked in [Con22], where the following result is proved.

Theorem 3.4.6 (Thm. 4.3 of [Con22])**.** *Assume that*

$$2d\Big(1-e^{-2\beta}\Big)e^{4d\beta} \leq 1. \tag{3.4.1}$$

Then the assumptions of Theorem 3.4.1 are satisfied and we have

$$\begin{aligned}\kappa_* &= 2-2d(1-e^{-2\beta})e^{2\beta d}\\ \overline{\kappa_*} &= e^{-2\beta d}-2d(1-e^{-2\beta})e^{2\beta d}.\end{aligned}$$

Remark 3.4.7 (Comparison with Cor. 4.4 of [EHMT17]). *By combining Theorems 3.4.1 and 3.4.4 and by choosing the logarithmic mean $\theta = \theta_1$, it follows that*

$$\mathrm{Ric_e} \geq 1 + e^{-2\beta d} - 3d(1 - e^{-2\beta})e^{2\beta d} \qquad \text{if} \qquad 2d\Big(1 - e^{-2\beta}\Big)e^{4d\beta} \leq 1. \tag{3.4.2}$$

On the other hand, in [EHMT17, Cor. 4.4], it was proved for the Ising model that

$$\mathrm{Ric_e} \geq 2\Big[1 - (2d-1)\Big(1 - e^{-2\beta}\Big)e^{4\beta d}\Big]e^{-2\beta d} \qquad \text{if} \qquad (2d-1)\Big(1 - e^{-2\beta}\Big)e^{4\beta d} \leq 1. \tag{3.4.3}$$

As observed by Conforti, the condition in (3.4.2) *is a bit more demanding then the one in* (3.4.3)*, but when it is satisfied then the corresponding lower bound for the entropic Ricci curvature is better (for $d \geq 2$).*

3.4.2 Bernoulli–Laplace model

In this subsection we analyze a simplified version of the Bernoulli–Laplace model, following Section 5.1 of [Con22]. Given integers $L > N \in \mathbb{N}$, where L represents the number of sites and N the number of particles, the state space is

$$\Omega = \Big\{\eta \in \{0,1\}^{[L]} \mid \sum_{i=1}^{L} \eta_i = N\Big\},$$

where $[L] = \{1, \dots, L\}$. Let $\delta_i \in \{0,1\}^{[L]}$ be defined by $\delta_i(k) = \begin{cases} 1 & \text{if } i = k, \\ 0 & \text{otherwise} \end{cases}$. Then the set of moves is $G = \{\sigma_{ik} \mid i, k \in [L]\}$ where σ_{ik} moves a particle from site i to site k if possible, i.e.

$$\sigma_{ij}(\eta) = \begin{cases} \eta - \delta_i + \delta_j & \text{if } \eta_i(1-\eta_j) > 0, \\ \eta & \text{otherwise.} \end{cases}$$

The transition rates are given by

$$c(\eta, \sigma_{ij}) = \eta_i(1 - \eta_j)$$

and the reversible measure m is the uniform one on Ω. Finally, notice that we have $\sigma_{ij}^{-1} = \sigma_{ji}$.

Theorem 3.4.8. *For the Bernoulli–Laplace model and for all weight functions θ satisfying Assumption 1, the inequality* (3.2.3) *holds with constant*

$$K = M_\theta + \frac{L}{2}.$$

Remark 3.4.9 (Comparison with [Con22]). *In [Con22, Thm. 5.1], under the same assumptions on the model, Conforti establishes inequality* (3.2.6) *and thus $\mathrm{CSI}_\phi(2K)$ with constant K equal to*

- $\frac{L}{2}$ *for general convex ϕ satisfying convexity of* (3.2.7)*;*
- $\frac{L}{2} + 1$ *for $\phi = \phi_1$, corresponding to $\mathrm{MLSI}(L+2)$;*
- $\frac{\alpha L}{2}$ *for $\phi = \phi_\alpha$ with $\alpha \in (1, 2]$.*

Thus, by the discussion in Section 3.2, we obtain a stronger result for the case $\theta = \theta_1$ and complementary results for other choices of θ.

Remark 3.4.10. *The entropic curvature of the Bernoulli–Laplace model has been studied before [EMT15], also in the more general case of non-homogeneous rates [FM16]. In the homogeneous setting, our result for $\theta = \theta_1$ recovers the same (best known) lower bound of [EMT15].*

Proof of Theorem 3.4.8

Again, the proof is based on Lemma 3.3.1 and adapts the arguments of [Con22], from which we use the same coupling rates: for $(\eta, \sigma_{ij}) \in S$ (i.e. $\eta_i = 1, \eta_j = 0$) set

$$c^{\mathsf{cpl}}(\eta, \sigma_{ij}\eta, \gamma, \bar{\gamma}) = \begin{cases} \min\{c(\eta,\gamma), c(\sigma_{ij}\eta,\gamma)\} & \text{if } \gamma = \bar{\gamma}, \\ 1 & \text{if } \gamma = \sigma_{ij}, \bar{\gamma} = e \text{ or if } \gamma = e, \bar{\gamma} = \sigma_{ji}, \\ 1 - \eta_l & \text{if } \gamma = \sigma_{il}, \bar{\gamma} = \sigma_{jl} \text{ and } l \notin \{i,j\}, \\ \eta_k & \text{if } \gamma = \sigma_{kj}, \bar{\gamma} = \sigma_{ki} \text{ and } k \notin \{i,j\}, \\ 0 & \text{otherwise.} \end{cases}$$

With these coupling rates and by (3.3.4), the right hand side of equation (3.3.3) from Lemma 3.3.1 is bounded from below by $\frac{1}{4}(A + B + C + D)$, where we define

$$\begin{aligned} A &= \sum_{\eta,i,j,k,l} m(\eta)c(\eta,\sigma_{ij}) \min\{c(\eta,\sigma_{kl}), c(\sigma_{ij}\eta,\sigma_{kl})\} I(\eta,\sigma_{ij},\sigma_{kl},\sigma_{kl}), \\ B &= \sum_{\eta,i,j} m(\eta)c(\eta,\sigma_{ij})[J(\eta,\sigma_{ij},\sigma_{ij},e) + J(\eta,\sigma_{ij},e,\sigma_{ji})], \\ C &= \sum_{\eta,i,j} m(\eta)c(\eta,\sigma_{ij}) \left[\sum_{l \neq i,j} (1-\eta_l) J(\eta,\sigma_{ij},\sigma_{il},\sigma_{jl}) \right], \\ D &= \sum_{\eta,i,j} m(\eta)c(\eta,\sigma_{ij}) \left[\sum_{k \neq i,j} \eta_k J(\eta,\sigma_{ij},\sigma_{kj},\sigma_{ki}) \right]. \end{aligned}$$

We show that

- $A = 0$,
- $B \geq (4 + 4M_\theta)\mathscr{A}(\rho,\psi)$,
- $C \geq 2(L - N - 1)\mathscr{A}(\rho,\psi)$,
- $D \geq 2(N-1)\mathscr{A}(\rho,\psi)$,

from which the theorem follows by Lemma 3.3.1. It is convenient to prove first the following

Lemma 3.4.11. *For all $\eta \in \Omega$ and $i,j,k,l \in [L]$ the following hold:*

1. $c(\eta,\sigma_{ij}) \min\{c(\eta,\sigma_{kl}), c(\sigma_{ij}\eta,\sigma_{kl})\} = \begin{cases} 1 & \textit{if } i \neq k, j \neq l, \eta_i = \eta_k = 1, \eta_j = \eta_l = 0, \\ 0 & \textit{otherwise}. \end{cases}$
2. $c(\eta,\sigma_{ij}) \min\{c(\eta,\sigma_{kl}), c(\sigma_{ij}\eta,\sigma_{kl})\} = c(\eta,\sigma_{kl}) \min\{c(\eta,\sigma_{ij}), c(\sigma_{kl}\eta,\sigma_{ij})\}$.
3. $I(\eta,\sigma_{ij},\sigma_{kl},\sigma_{kl}) = -I(\sigma_{kl}\eta,\sigma_{ij},\sigma_{lk},\sigma_{lk})$ *if* $c(\eta,\sigma_{ij}) \min\{c(\eta,\sigma_{kl}), c(\sigma_{ij}\eta,\sigma_{kl})\} > 0$.

Proof of Lemma. Statements 1–2 were already observed by Conforti in the proof of [Con22, Thm 5.1] and are easy to check, while statement 3 is immediate from the definitions. ∎

Term A We have, using 2. of Lemma 3.4.11,

$$\begin{aligned}A &= \sum_{\eta,i,j,k,l} m(\eta)c(\eta,\sigma_{ij})\min\{c(\eta,\sigma_{kl}),c(\sigma_{ij}\eta,\sigma_{kl})\}I(\eta,\sigma_{ij},\sigma_{kl},\sigma_{kl})\\ &= \sum_{\eta,i,j,k,l} m(\eta)c(\eta,\sigma_{kl})\min\{c(\eta,\sigma_{ij}),c(\sigma_{kl}\eta,\sigma_{ij})\}I(\eta,\sigma_{ij},\sigma_{kl},\sigma_{kl})\\ &= \sum_{\eta,k,l} m(\eta)c(\eta,\sigma_{kl})F(\eta,\sigma_{kl})\end{aligned}$$

with

$$F(\eta,\sigma_{kl}) = \sum_{i,j}\min\{c(\eta,\sigma_{ij}),c(\sigma_{kl}\eta,\sigma_{ij})\}I(\eta,\sigma_{ij},\sigma_{kl},\sigma_{kl}).$$

Next, using the reversibility property (3.2.2), the fact that $\sigma_{lk} = \sigma_{kl}^{-1}$ for the first and second equality, and properties 2 and 3 of Lemma 3.4.11 respectively for the third and fourth equality, we deduce that

$$\begin{aligned}A &= \sum_{\eta,k,l} m(\eta)c(\eta,\sigma_{kl})F(\sigma_{kl}\eta,\sigma_{lk})\\ &= \sum_{\eta,i,j,k,l} m(\eta)c(\eta,\sigma_{kl})\min\{c(\eta,\sigma_{ij}),c(\sigma_{kl}\eta,\sigma_{ij})\}I(\sigma_{kl}\eta,\sigma_{ij},\sigma_{lk},\sigma_{lk})\\ &= \sum_{\eta,i,j,k,l} m(\eta)c(\eta,\sigma_{ij})\min\{c(\eta,\sigma_{kl}),c(\sigma_{ij}\eta,\sigma_{kl})\}I(\sigma_{kl}\eta,\sigma_{ij},\sigma_{lk},\sigma_{lk})\\ &= -\sum_{\eta,i,j,k,l} m(\eta)c(\eta,\sigma_{ij})\min\{c(\eta,\sigma_{kl}),c(\sigma_{ij}\eta,\sigma_{kl})\}I(\eta,\sigma_{ij},\sigma_{kl},\sigma_{kl})\\ &= -A.\end{aligned}$$

This implies that $A = 0$, as desired.

Term B Notice that for $(\eta,\sigma_{ij}) \in S$

$$\begin{aligned}J(\eta,\sigma_{ij},\sigma_{ij},e) &= \{\theta(\rho(\sigma_{ij}\eta),\rho(\sigma_{ij}\eta)) + \theta(\rho(\eta),\rho(\sigma_{ij}\eta))\}[\psi(\eta)-\psi(\sigma_{ij}\eta)]^2\\ J(\eta,\sigma_{ij},e,\sigma_{ji}) &= \{\theta(\rho(\eta),\rho(\eta)) + \theta(\rho(\eta),\rho(\sigma_{ij}\eta))\}[\psi(\eta)-\psi(\sigma_{ij}\eta)]^2\end{aligned}$$

and so

$$J(\eta,\sigma_{ij},\sigma_{ij},e) + J(\eta,\sigma_{ij},e,\sigma_{ji}) \geq \{2M_\theta+2\}\theta(\rho(\eta),\rho(\sigma_{ij}\eta))[\psi(\eta)-\psi(\sigma_{ij}\eta)]^2.$$

Therefore we get

$$B \geq 4(M_\theta+1)\mathscr{A}(\rho,\psi).$$

Term C Notice that for $(\eta,\sigma_{ij}) \in S$ we have $\eta_i = 1, \eta_j = 0$ and there are $L-N-1$ empty sites left. Moreover when $l \neq i,j$ and $\eta_l = 0$ we have that $\sigma_{jl}\sigma ij\eta = \sigma_{il}\eta$ and so

$$J(\eta,\sigma_{ij},\sigma_{il},\sigma_{jl}) \geq \theta(\rho(\eta),\rho(\sigma_{ij}\eta))[\psi(\eta)-\psi(\sigma_{ij}\eta)]^2.$$

Therefore we get

$$C \geq 2(L-N-1)\mathscr{A}(\rho,\psi).$$

Term D Notice that for $(\eta, \sigma_{ij}) \in S$ we have $\eta_i = 1, \eta_j = 0$ and there are $N-1$ other occupied sites. Moreover when $k \neq i, j$ and $\eta_k = 1$ we have $\sigma_{ki}\sigma ij\eta = \sigma_{kj}\eta$ and so

$$J(\eta, \sigma_{ij}, \sigma_{kj}, \sigma_{ki}) \geq \theta(\rho(\eta), \rho(\sigma_{ij}\eta))[\psi(\eta) - \psi(\sigma_{ij}\eta)]^2.$$

Therefore we get

$$D \geq 2(N-1)\mathscr{A}(\rho, \psi).$$

Combining these estimates for A, B, C, D, an application of Lemma 3.3.1 gives

$$\mathscr{B}(\rho, \psi) \geq \left(1 + M_\theta + \frac{L}{2} - 1\right)\mathscr{A}(\rho, \psi) = \left(M_\theta + \frac{L}{2}\right)\mathscr{A}(\rho, \psi).$$

This concludes the proof of the theorem.

3.4.3 Hardcore model

Following Section 5.2 of [Con22], we consider the classical hardcore model. Let (V, E) be a simple, finite, connected graph and write $x \sim y$ if the vertices x, y are connected (in the rest of this subsection, x, y always denote general elements of V). The state space is

$$\Omega = \left\{\eta \in \{0,1\}^V \mid \eta_x\eta_y = 0 \text{ if } x \sim y\right\}.$$

In other words, each vertex can either be empty or be occupied by a particle, with the rule that if a site is occupied then its neighbors are all free. Let $N_x = \{y \in V \mid x \sim y\}$ be the set of neighbors of vertex x, $\bar{N}_x = N_x \cup \{x\}$ and as before let $\delta_x \in \{0,1\}^V$ be defined by $\delta_x(y) = \begin{cases} 1 & \text{if } x = y, \\ 0 & \text{otherwise }. \end{cases}$ Then, the set of moves is given by $G = \{\gamma_x^+, \gamma_x^- \mid x \in V\}$ where γ_x^+ adds a particle to site x if possible and γ_x^- removes it if possible, i.e.

$$\gamma_x^+(\eta) = \begin{cases} \eta + \delta_x & \text{if } \eta + \delta_x \in \Omega, \\ \eta & \text{otherwise }, \end{cases}$$

$$\gamma_x^-(\eta) = \begin{cases} \eta - \delta_x & \text{if } \eta - \delta_x \in \Omega, \\ \eta & \text{otherwise.} \end{cases}$$

We also denote $G^+ = \{\gamma_x^+ \mid x \in V\}$ and $G^- = \{\gamma_x^- \mid x \in V\}$. For a given parameter $\beta \in (0,1)$, the transition rates are defined by

$$c(\eta, \gamma_x^+) = \beta \prod_{y \in \bar{N}_x} (1 - \eta_y),$$
$$c(\eta, \gamma_x^-) = \eta_x$$

for all $\eta \in \Omega, x \in V$. With these choices, we have $(\gamma_x^+)^{-1} = \gamma_x^-$ and the reversible measure is given by

$$m(\eta) = \frac{1}{Z}\mathbb{1}_{\eta \in \Omega} \prod_{x \in V} \beta^{\eta_x},$$

where $Z > 0$ is the normalization constant.

Theorem 3.4.12. *Let Δ be the maximum degree of (V, E) and assume that*

$$\beta\Delta \leq 1. \tag{3.4.4}$$

Set

$$\kappa_* = 1 - \beta(\Delta - 1), \qquad \overline{\kappa_*} = \min\{\beta, 1 - \beta\Delta\}.$$

Then, for all weight functions θ satisfying Assumption 1, inequality (3.2.3) *holds with constant*

$$K = \frac{\kappa_*}{2} + M_\theta \overline{\kappa_*}.$$

Remark 3.4.13 (Comparison with [Con22])**.** *In [Con22, Thm. 5.2], under the same assumptions on the model, Conforti establishes inequality* (3.2.6) *and thus $\mathrm{CSI}_\phi(2K)$ with constant K equal to*

- $\frac{\kappa_*}{2}$ *for general convex ϕ satisfying convexity of* (3.2.7)*;*
- $\frac{\kappa_*}{2} + \overline{\kappa_*}$ *for $\phi = \phi_1$, corresponding to $\mathrm{MLSI}(\kappa_* + 2\overline{\kappa_*})$;*
- $\frac{\alpha\kappa_*}{2}$ *for $\phi = \phi_\alpha$ with $\alpha \in (1, 2]$.*

Thus, by the discussion in Section 3.2, we obtain a stronger result for the case $\theta = \theta_1$ and complementary results for other choices of θ.

Remark 3.4.14. *The entropic curvature of the hardcore model has been studied before in [EHMT17], where a more general version of the model is considered. When restricting to the classical version discussed in this section, it was proved in [EHMT17, Cor. 4.8] that*

$$\mathrm{Ric}_\mathrm{e} \geq \frac{\kappa_*}{2}$$

under condition (3.4.4)*. Therefore, in this setting, by choosing $\theta = \theta_1$ in Theorem 3.4.12 we find a better lower bound for the entropic curvature.*

Proof of Theorem 3.4.12

The proof is based on Lemma 3.3.1 and on the arguments in [Con22], from which we use the same coupling rates. For $(\eta, \gamma_x^+) \in S$ (i.e. $\eta\,|_{\bar{N}_x} = 0$) we set

$$c^{\mathsf{cpl}}(\eta, \gamma_x^+\eta, \gamma, \bar{\gamma}) = \begin{cases} \min\{c(\eta, \gamma), c(\gamma_x^+\eta, \gamma)\} & \text{if } \gamma = \bar{\gamma}, \\ \beta & \text{if } \gamma = \gamma_y^+, \bar{\gamma} = \gamma_x^- \text{ with } y \sim x, \eta|_{\bar{N}_y} = 0, \\ \beta & \text{if } \gamma = \gamma_x^+, \bar{\gamma} = e, \\ 1 - \beta\Big|\Big\{y : y \sim x, \eta|_{\bar{N}_y} = 0\Big\}\Big| & \text{if } \gamma = e, \bar{\gamma} = \gamma_x^-, \\ 0 & \text{otherwise.} \end{cases}$$

If $(\eta, \gamma_x^-) \in S$ then also $(\gamma_x^-\eta, \gamma_x^+) \in S$, and so we can set

$$c^{\mathsf{cpl}}(\eta, \gamma_x^-\eta, \gamma, \bar{\gamma}) = c^{\mathsf{cpl}}(\gamma_x^-\eta, \gamma_x^+\gamma_x^-\eta, \bar{\gamma}, \gamma) = c^{\mathsf{cpl}}(\gamma_x^-\eta, \eta, \bar{\gamma}, \gamma).$$

With these coupling rates the right hand side of equation (3.3.3) from Lemma 3.3.1 reads

$$\frac{1}{4} \sum_{(\eta, \gamma_x^+) \in S} \sum_{\gamma, \bar{\gamma} \in G^*} m(\eta) c(\eta, \gamma_x^+) c^{\mathsf{cpl}}(\eta, \gamma_x^+\eta, \gamma, \bar{\gamma}) J(\eta, \gamma_x^+, \gamma, \bar{\gamma})$$
$$+ \frac{1}{4} \sum_{(\eta, \gamma_x^-) \in S} \sum_{\gamma, \bar{\gamma} \in G^*} m(\eta) c(\eta, \gamma_x^-) c^{\mathsf{cpl}}(\gamma_x^-\eta, \eta, \bar{\gamma}, \gamma) J(\eta, \gamma_x^-, \gamma, \bar{\gamma}).$$

Using the reversibility property (3.2.2) with

$$F(\eta,\sigma) = \mathbb{1}_{G^-}(\sigma) \sum_{\gamma,\bar{\gamma}\in G^*} c^{\mathsf{cpl}}(\sigma\eta,\eta,\bar{\gamma},\gamma) J(\eta,\sigma,\gamma,\bar{\gamma})$$

and that $J(\eta,\sigma\eta,\gamma,\bar{\gamma}) = J(\sigma\eta,\sigma^{-1},\bar{\gamma},\gamma)$ (when $\sigma^{-1}\sigma\eta=\eta$) the second term is equal to the first, so we can rewrite the previous quantity as

$$\frac{1}{2} \sum_{(\eta,\gamma_x^+)\in S} \sum_{\gamma,\bar{\gamma}\in G^*} m(\eta) c(\eta,\gamma_x^+) c^{\mathsf{cpl}}(\eta,\gamma_x^+\eta,\gamma,\bar{\gamma}) J(\eta,\gamma_x^+,\gamma,\bar{\gamma}). \tag{3.4.5}$$

Similarly we have

$$\begin{aligned}
\mathscr{A}(\rho,\psi) =& \frac{1}{2} \sum_{(\eta,\gamma_x^+)\in S} m(\eta)c(\eta,\gamma_x^+)\theta\Big(\rho(\eta),\rho(\gamma_x^+\eta)\Big)\Big[\psi(\eta)-\psi(\gamma_x^+\eta)\Big]^2 \\
&+\frac{1}{2} \sum_{(\eta,\gamma_x^-)\in S} m(\eta)c(\eta,\gamma_x^-)\theta\Big(\rho(\eta),\rho(\gamma_x^-\eta)\Big)\Big[\psi(\eta)-\psi(\gamma_x^-\eta)\Big]^2
\end{aligned}$$

and using again reversibility (3.2.2) the second term is equal to the first, so that we can write

$$\mathscr{A}(\rho,\psi) = \sum_{(\eta,\gamma_x^+)\in S} m(\eta)c(\eta,\gamma_x^+)\theta\Big(\rho(\eta),\rho(\gamma_x^+\eta)\Big)\Big[\psi(\eta)-\psi(\gamma_x^+\eta)\Big]^2.$$

We then have that (using that $\forall \eta,x,y\ c(\eta,\gamma_y^-)\le c(\gamma_x^+\eta,\gamma_y^-)$ and $c(\eta,\gamma_y^+)\ge c(\gamma_x^+\eta,\gamma_y^+)$) the quantity (3.4.5) (and in particular $\mathscr{B}(\rho,\psi)$ too) is lower bounded by $\frac{1}{2}(A+B+C)$ with

$$\begin{aligned}
A =& \sum_{\eta,x,y} m(\eta)c(\eta,\gamma_x^+)c(\eta,\gamma_y^-)I(\eta,\gamma_x^+,\gamma_y^-,\gamma_y^-) \\
&+ \sum_{\eta,x,y} m(\eta)c(\eta,\gamma_x^+)c(\gamma_x^+\eta,\gamma_y^+)I(\eta,\gamma_x^+,\gamma_y^+,\gamma_y^+), \\
B =& \beta \sum_{\substack{\eta,x,y:\\ x\sim y,\eta|_{\bar{N}_y}=0}} m(\eta)c(\eta,\gamma_x^+)I(\eta,\gamma_x^+,\gamma_y^+,\gamma_x^-), \\
C =& \sum_{\eta,x} m(\eta)c(\eta,\gamma_x^+)\Big[\Big(1-\beta\Big|\Big\{y : y\sim x, \eta|_{\bar{N}_y}=0\Big\}\Big|\Big)J(\eta,\gamma_x^+,e,\gamma_x^-)+\beta J(\eta,\gamma_x^+,\gamma_x^+,e)\Big].
\end{aligned}$$

We will show that

- $A=0$,
- $B=0$,
- $C\ge(\kappa_*+2M_\theta\overline{\kappa_*})\mathscr{A}(\rho,\psi)$.

An application of Lemma 3.3.1 then concludes the proof of the theorem. To do so, we will use the following:

Lemma 3.4.15. *For all $\eta\in\Omega$ and $x,y\in V$ the following hold:*

1. $c(\eta,\gamma_x^+)c(\gamma_x^+\eta,\gamma_y^+) = c(\eta,\gamma_x^+)c(\gamma_x^+\eta,\gamma_y^+) = \begin{cases} \beta^2 & \text{if } x \nsim y, \eta|_{\bar N_x \cup \bar N_y} = 0, \\ 0 & \text{otherwise.} \end{cases}$
2. $x \nsim y, \eta|_{\bar N_x \cup \bar N_y} = 0 \implies \gamma_x^+\gamma_y^+\eta = \gamma_y^+\gamma_x^+\eta$.
3. *If* $\eta|_{\bar N_x \cup \bar N_y} = 0$ *then* $c(\eta,\gamma_x^+) = c(\eta,\gamma_y^+) = \beta$.
4. $I(\gamma_y^+\eta,\gamma_x^+,\gamma_y^-,\gamma_y^-) = -I(\eta,\gamma_x^+,\gamma_y^+,\gamma_y^+)$ *if* $x \nsim y, \eta|_{\bar N_x \cup \bar N_y} = 0$.
5. $I(\eta,\gamma_x^+,\gamma_y^+,\gamma_x^-) = -I(\eta,\gamma_y^+,\gamma_x^+,\gamma_y^-)$ *if* $\eta|_{\bar N_x \cup \bar N_y} = 0$.

Proof of Lemma. Statements 1–3 were already in the proof of [Con22, Thm. 5.2] and are easy to check, while statement 4–5 are immediate from the definitions. ■

Term A We look at the first term in the sum defining A: we can write it as

$$\sum_{\eta,\gamma} m(\eta)c(\eta,\gamma)F(\eta,\gamma)$$

with

$$F(\eta,\gamma) = \mathbb{1}_{G^-}(\gamma)\sum_x c(\eta,\gamma_x^+)I(\eta,\gamma_x^+,\gamma,\gamma).$$

Using reversibility (3.2.2) we can rewrite it as

$$\sum_{\eta,x,y} m(\eta)c(\eta,\gamma_y^+)c(\gamma_y^+\eta,\gamma_x^+)I(\gamma_y^+\eta,\gamma_x^+,\gamma_y^-,\gamma_y^-).$$

Then we have

$$\begin{aligned} &\sum_{\eta,x,y} m(\eta)c(\eta,\gamma_y^+)c(\gamma_y^+\eta,\gamma_x^+)I(\gamma_y^+\eta,\gamma_x^+,\gamma_y^-,\gamma_y^-) \\ &= \sum_{\eta,x,y} m(\eta)c(\eta,\gamma_x^+)c(\gamma_x^+\eta,\gamma_y^+)I(\gamma_y^+\eta,\gamma_x^+,\gamma_y^-,\gamma_y^-) \\ &= -\sum_{\eta,x,y} m(\eta)c(\eta,\gamma_x^+)c(\gamma_x^+\eta,\gamma_y^+)I(\eta,\gamma_x^+,\gamma_y^+,\gamma_y^+) \end{aligned}$$

using Lemma 3.4.15. Therefore, the first term in the sum of A is the opposite of the second, which implies $A = 0$.

Term B We have

$$B = \beta \sum_{\substack{\eta,x,y:\\ x\sim y, \eta|_{\bar N_y}=0}} m(\eta)c(\eta,\gamma_x^+)I(\eta,\gamma_x^+,\gamma_y^+,\gamma_x^-).$$

Noticing as in [Con22] that $(\eta,\gamma_x^+) \in S$ if and only if $\eta|_{\bar N_x = 0}$ we can write

$$B = \beta \sum_{\substack{\eta,x,y:\\ x\sim y, \eta|_{\bar N_y}=0, \eta|_{\bar N_x}=0}} m(\eta)c(\eta,\gamma_x^+)I(\eta,\gamma_x^+,\gamma_y^+,\gamma_x^-). \tag{3.4.6}$$

By exchanging x, y we therefore also have

$$B = \beta \sum_{\substack{\eta,x,y:\\ x\sim y, \eta|_{\bar N_y}=0, \eta|_{\bar N_x}=0}} m(\eta)c(\eta,\gamma_y^+)I(\eta,\gamma_y^+,\gamma_x^+,\gamma_y^-). \tag{3.4.7}$$

Summing these two expressions we get

$$\begin{aligned}
2B &= \beta \sum_{\substack{\eta,x,y:\\ x\sim y, \eta|_{\bar N_y}=0, \eta|_{\bar N_x}=0}} m(\eta)\Big[c(\eta,\gamma_x^+)I(\eta,\gamma_x^+,\gamma_y^+,\gamma_x^-)+c(\eta,\gamma_y^+)I(\eta,\gamma_y^+,\gamma_x^+,\gamma_y^-)\Big]\\
&= \beta \sum_{\substack{\eta,x,y:\\ x\sim y, \eta|_{\bar N_y}=0, \eta|_{\bar N_x}=0}} m(\eta)\Big[c(\eta,\gamma_x^+)-c(\eta,\gamma_y^+)\Big]I(\eta,\gamma_x^+,\gamma_y^+,\gamma_x^-)\\
&= 0,
\end{aligned}$$

where we used Lemma 3.4.15.

Term C We have

$$\begin{aligned}
C &= \sum_{\eta,x} m(\eta)c(\eta,\gamma_x^+)\Big[\Big(1-\beta\Big|\Big\{y: y\sim x, \eta|_{\bar N_y}=0\Big\}\Big|\Big)J(\eta,\gamma_x^+,e,\gamma_x^-)+\beta J(\eta,\gamma_x^+,\gamma_x^+,e)\Big]\\
&= \sum_{\eta,x} m(\eta)c(\eta,\gamma_x^+)\Big[\psi(\eta)-\psi(\gamma_x^+\eta)\Big]^2\\
&\cdot \Big[\Big(1-\beta\Big|\Big\{y: y\sim x, \eta|_{\bar N_y}=0\Big\}\Big|\Big)\Big(\theta(\rho(\eta),\rho(\eta))+\theta\Big(\rho(\eta),\rho(\gamma_x^+\eta)\Big)\Big)\\
&+\beta\Big(\theta\Big(\rho(\gamma_x^+\eta),\rho(\gamma_x^+\eta)\Big)+\theta\Big(\rho(\eta),\rho(\gamma_x^+\eta)\Big)\Big)\Big]\\
&\geq \sum_{\eta,x} m(\eta)c(\eta,\gamma_x^+)\theta\Big(\rho(\eta),\rho(\gamma_x^+\eta)\Big)\Big[\psi(\eta)-\psi(\gamma_x^+\eta)\Big]^2(\kappa_*+2M_\theta\overline{\kappa_*})\\
&= (\kappa_*+2M_\theta\overline{\kappa_*})\mathscr{A}(\rho,\psi).
\end{aligned}$$

This concludes the proof of the theorem.

3.4.4 Interacting random walks

Following Section 3 of [Con22], we now consider the case of interacting random walks. One motivation in this subsection is to find a discrete analogue of the following classical result.

Proposition 3.4.16. *Consider $\mathbb{R}^d$ equipped with the standard Euclidean distance d. Let $V\colon \mathbb{R}^d \to \mathbb{R}$ be **convex**, γ_d be the law of a standard Gaussian in $\mathbb{R}^d$ and $Z>0$ be a normalizing constant so that $\frac{1}{Z}e^{-V}d\gamma_d$ is a probability measure. Then the metric measure space $\left(\mathbb{R}^d,\mathrm{d},\frac{1}{Z}e^{-V}d\gamma_d\right)$ has Ricci curvature $\mathrm{Ric}\geq 1$ in the sense of the Lott–Sturm–Villani theory.*

To find a discrete analogue, here the role of $\mathbb{R}^d$ is taken over by the discrete state space $\Omega=\mathbb{N}^d$, while the Gaussian measure γ_d is replaced by the multivariate Poisson distribution μ_λ given by the product measure of d one-dimensional Poisson distribution of intensity $\frac{1}{\lambda}$, i.e. for $\eta\in\Omega$ we have

$$\mu_\lambda(\eta)=\prod_{i=1}^d e^{-\frac{1}{\lambda}}\frac{\lambda^{-\eta_i}}{\eta_i!}.$$

It remains to define a Markov chain on this state space. We consider the set of moves G containing γ_i^+,γ_i^- for $i\in[d]$, where

$$\begin{aligned}
\gamma_i^+\eta &= \eta+\mathbf{e_i},\\
\gamma_i^-\eta &= \eta-\mathbf{e_i}\,\mathbb{1}_{\eta_i>0},
\end{aligned}$$

and we denote as usual by e the null move. Consider two potentials $V^+, V^-\colon \mathbb{N}^d \to \mathbb{R}$ and correspondingly define transition rates

$$
\begin{aligned}
c(\eta, \gamma_i^+) &= \exp\Big(-\nabla_i^+ V^+(\eta)\Big),\\
c(\eta, \gamma_i^-) &= \begin{cases} \exp\Big(-\nabla_i^- V^-(\eta)\Big) & \text{if } \eta_i > 0,\\ 0 & \text{if } \eta_i = 0,\end{cases}
\end{aligned}
$$

where we define $\nabla_i^\pm = \nabla_{\gamma_i^\pm}$. Then the reversible measure takes the form

$$
m = \frac{1}{Z}\exp(-V^+ - V^-),
$$

where Z is athe normalizing constant. An interesting choice is given by

$$
V^-(\eta) = \sum_{i=1}^{d} \log(\lambda)\eta_i + \log(\eta_i!), \tag{3.4.8}
$$

which corresponds to $c(\eta, \gamma_i^-) = \lambda\eta_i \mathbb{1}_{\eta_i>0} = \lambda\eta_i$. In this case, we write $V = V^+$ and the reversible measure becomes

$$
\frac{1}{Z} e^{-V} d\mu_\lambda,
$$

which is reminiscent of the setting of Proposition 3.4.16. Therefore, to find an analogous discrete result, we are left to look for conditions on the potential V that resemble convexity and yield positive entropic curvature of the corresponding Markov chain; we will do this in Section 3.4.4, as an application of the main theorem of this section below.

The first assumption that we make on the model is that for all $\eta \in \mathbb{N}^d$ and $i, j \in [d]$ we have

$$
\nabla_i^+ c(\eta, \gamma_j^+) \leq 0; \tag{3.4.9}
$$

$$
\nabla_i^+ c(\eta, \gamma_j^-) \geq 0. \tag{3.4.10}
$$

Notice that for V^- as in (3.4.8) the condition (3.4.10) is always satisfied. We remark that this assumption was not needed in [Con22], but it will be useful later in the proof of the main theorem of this section to obtain some terms cancellations.

Next, following [Con22], we make the crucial assumptions that for all $\eta \in \mathbb{N}^d, i \in [d]$ the following quantities are non-negative:

$$
\kappa^+(\eta, i) := -\nabla_i^+ c(\eta, \gamma_i^+) \quad - \sum_{j\in[d], j\neq i} \nabla_i^+ c(\eta, \gamma_j^-) \geq 0, \tag{3.4.11}
$$

$$
\kappa^-(\eta, i) := \nabla_i^+ c(\eta, \gamma_i^-) \quad + \sum_{j\in[d], j\neq i} \nabla_i^+ c(\eta, \gamma_j^+) \quad \geq 0. \tag{3.4.12}
$$

Correspondingly we set

$$
\kappa_* = \inf_{\eta\in\mathbb{N}^d, i\in[d]} \kappa^+(\eta, i) + \kappa^-(\eta, i).
$$

It is natural to introduce the additional quantity

$$
\overline{\kappa_*} = \min\left\{ \inf_{\eta\in\mathbb{N}^d, i\in[d]} \kappa^+(\eta, i),\ \inf_{\eta\in\mathbb{N}^d, i\in[d]} \kappa^-(\eta, i) \right\}.
$$

As in [Con22] and in analogy with the previous examples, the assumptions in (3.4.11), (3.4.12) are needed for the construction of appropriate contractive coupling rates; compared to [Con22], the expressions are slightly simplified thanks to the additional assumptions (3.4.9), (3.4.10). With regard to the heuristic discussion in Section 3.3, the quantities κ_* and $\overline{\kappa_*}$ correspond respectively to (3.3.7) and (3.3.9).

It is also important to notice that this is the only example that we discuss where the cardinality of the state space is not finite and that, because of this, there are some additional technical difficulties and not all the considerations of Section 3.2 can be directly applied here. In this paper, to deal with the infinite cardinality of the state space, we proceed as in [Con22] and make use of a localization argument, which we now briefly describe; more precisely, with this procedure and with ϕ and θ satisfying Assumption 2, we explain how to derive inequality (3.1.3) from establishing (3.2.3) for a localizing sequence of finite state space Markov chains. Given an integer $N \geq 2$, let $\Omega_N = \left\{\eta \in \mathbb{N}^d \mid \eta_i \leq N \, \forall i \in [d]\right\}$. On Ω_N, consider the Markov chain with generator L_N described by the set of moves $G_N = \left\{\gamma_i^{+,N}, \gamma_i^{-,N} \mid i \in [d]\right\}$, where $\gamma_i^{+,N}(\eta) = \eta + \mathbf{e_i}\mathbb{1}_{\eta_i<N}$ and $\gamma_i^{-,N}(\eta) = \eta - \mathbf{e_i}\mathbb{1}_{\eta_i>0} = \gamma_i^-|_{\Omega_N}\eta$ (and $\gamma_i^-|_{\Omega_N}$ denotes the restriction of γ_i^- to Ω_N), and by the transition rates $c_N(\eta, \gamma_i^{+,N}) = c(\eta, \gamma_i^+)\mathbb{1}_{\eta_i<N}$ and $c_N(\eta, \gamma_i^{-,N}) = c(\eta, \gamma_i^-)$. Clearly $\left(\gamma_i^{\pm,N}\right)^{-1} = \gamma_i^{\mp,N}$; moreover, as observed by Conforti, it is easy to check that this Markov chain is reversible with respect to the probability measure $m_N = \frac{m}{m(\Omega_N)}$ on Ω_N and (3.2.2) holds. Finally, denote by $\mathscr{B}_N$ and $\mathscr{A}_N$ the corresponding quantities $\mathscr{B}$ and $\mathscr{A}$ for this Markov chain, set $G_N^+ = \left\{\gamma_i^{+,N} \mid i \in [d]\right\}$, $G_N^- = \left\{\gamma_i^{-,N} \mid i \in [d]\right\}$, and, as usual, consider the enlarged set of moves $G_N^* = G_N \cup \{e\}$. The main theorem of this section reads as follows.

Theorem 3.4.17. *With the previous notation, suppose that for all $\eta \in \mathbb{N}^d$ and $i, j \in [d]$ the assumptions* (3.4.9), (3.4.10), (3.4.11) *and* (3.4.12) *are satisfied. Let also θ be a weight function satisfying Assumption 1. Then we have that*

$$\mathscr{B}_N(\rho, \psi) \geq \left(\frac{\kappa_*}{2} + M_\theta \overline{\kappa_*}\right)\mathscr{A}_N(\rho, \psi)$$

for all integers $N \geq 2$ and for all functions $\rho\colon \Omega_N \to \mathbb{R}_{>0}$ and $\psi\colon \Omega_N \to \mathbb{R}$.

Remark 3.4.18. *With θ as in Assumption 2, by the discussion of Section 3.2 the previous theorem allows to deduce the convex Sobolev inequality* (3.1.3) *for all Markov chains (Ω_N, L_N, m_N) with uniform constant. As observed in [Con22], this allows us to deduce the same convex Sobolev inequality for the original Markov chain by taking limits, when ϕ is lower bounded (as it is the case for $\phi = \phi_\alpha$ with $\alpha \in [1, 2]$ in particular), cf. Corollary 2.1 of [Con22].*

Remark 3.4.19 (Comparison with [Con22])**.** *In [Con22, Thm. 3.1], Conforti establishes inequality* (3.2.6) *for the localizing sequence of Markov chain on Ω_N with a uniform constant K, and thus also $\mathrm{CSI}_\phi(2K)$ for the original Markov chain, with constant K equal to*

- $\frac{\kappa_*}{2}$ *for general convex ϕ satisfying convexity of* (3.2.7)*;*
- $\frac{\alpha\kappa_*}{2}$ *for $\phi = \phi_\alpha$ with $\alpha \in (1, 2]$.*

Compared to our assumptions on the model, he does not assume non-negativity in (3.4.9), (3.4.10)*; however, these additional assumptions are satisfied in the examples of Section 3.4.4. By the discussion in Section 3.2, we therefore obtain a complementary result to [Con22, Thm. 3.1].*

Proof of Theorem 3.4.17

The proof adapts the one in [Con22], with slightly different notation and choices. Fix an integer $N \geq 2$ and consider the Markov chain described by the triple (Ω_N, L_N, m_N). Notice that we have

$$\begin{aligned} S_N &:= \{(\eta,\sigma) \in \Omega_N \times G_N \mid c_N(\eta,\sigma) > 0\} \\ &= \Big\{(\eta,\gamma_i^{+,N}) \mid \eta \in \Omega_N, i \in [d], \eta_i < N\Big\} \cup \Big\{(\eta,\gamma_i^{-,N}) \mid \eta \in \Omega_N, i \in [d], \eta_i > 0\Big\}. \end{aligned}$$

Notice that from our definitions it follows that if $\left(\eta,\gamma_i^{+,N}\right) \in S_N$ then $\gamma_i^{+,N}\eta = \gamma_i^{+}|_{\Omega_N}\eta$. To lighten the notation, with a slight abuse of notation, we will take advantage of this and drop the superscript N in $\gamma_i^{\pm,N}$ (i.e. we just write $\gamma_i^{\pm}$), and we will also write $c_N(\eta,\gamma_i^{\pm}) = c_N(\eta,\gamma_i^{\pm,N})$. Similarly, for a function $\psi\colon \Omega_N \to \mathbb{R}$, a state $\eta \in \Omega_N$ and $i \in [d]$, we will write $\nabla_i^{\pm}\psi(\eta)$ instead of $\nabla_{\gamma_i^{\pm,N}}\psi(\eta)$. Again, this minor abuse of notation is justified by the fact that whenever $\nabla_i^{+}\psi(\eta)$ appears in the computations below, it will be multiplied by a jump rate equal to 0 if $\eta_i = N$. In analogy with (3.4.11), for the localized Markov chain and for $(\eta,\gamma_i^{+}) \in S_N$, we consider the quantity

$$\kappa^{+,N}(\eta,i) := -\nabla_i^{+}c_N(\eta,\gamma_i^{+}) - \sum_{j\in[d], j\neq i} \nabla_i^{+}c_N(\eta,\gamma_j^{-}),$$

and we observe that

$$\kappa^{+,N}(\eta,i) = \kappa^{+}(\eta,i) + c(\gamma_i^{+}\eta,\gamma_i^{+})\mathbb{1}_{\eta_i=N-1} \geq \kappa^{+}(\eta,i) \geq 0, \tag{3.4.13}$$

where the first equality is due to the fact that we have set $c_N(\tilde\eta,\gamma_i^{+}) = 0$ if $\tilde\eta_i = N$, as opposed to $c(\tilde\eta,\gamma_i^{+})$. Similarly, we define

$$\kappa^{-,N}(\eta,i) := \nabla_i^{+}c_N(\eta,\gamma_i^{-}) + \sum_{j\in[d], j\neq i} \nabla_i^{+}c_N(\eta,\gamma_j^{+})$$

and we notice that

$$\kappa^{-,N}(\eta,i) = \kappa^{-}(\eta,i) - \sum_{j\in[d], j\neq i} \mathbb{1}_{\eta_j=N}\cdot\nabla_i^{+}c(\eta,\gamma_j^{+}) \geq \kappa^{-}(\eta,i) \geq 0, \tag{3.4.14}$$

where the first equality follows from the fact that we have set $c_N(\tilde\eta,\gamma_j^{+}) = 0$ if $\tilde\eta_j = N$, as opposed to $c(\tilde\eta,\gamma_j^{+})$, and the first inequality is due to (3.4.9).

With these definitions, we are now ready to construct appropriate coupling rates, analogously to [Con22]. For $(\eta,\gamma_i^{+}) \in S_N$ and $\gamma,\bar\gamma \in G_N^*$ set

$$c^{\mathsf{cpl}}(\eta,\gamma_i^{+}\eta,\gamma,\bar\gamma) = \begin{cases} \min\{c_N(\eta,\gamma), c_N(\gamma_i^{+}\eta,\gamma)\} & \text{if } \gamma = \bar\gamma \neq e, \\ \max\Big\{\nabla_i^{+}c_N(\eta,\bar\gamma), 0\Big\} & \text{if } \gamma = \gamma_i^{+} \text{ and } \bar\gamma \neq \gamma_i^{+},\gamma_i^{-},e, \\ \max\Big\{-\nabla_i^{+}c_N(\eta,\gamma), 0\Big\} & \text{if } \gamma \neq \gamma_i^{+},\gamma_i^{-},e \text{ and } \bar\gamma = \gamma_i^{-}, \\ \kappa^{+,N}(\eta,i) & \text{if } \gamma = \gamma_i^{+}, \bar\gamma = e, \\ \kappa^{-,N}(\eta,i) & \text{if } \gamma = e, \bar\gamma = \gamma_i^{-}, \\ 0 & \text{otherwise.} \end{cases}$$

Next, notice that if $(\eta,\gamma_i^{-}) \in S_N$ then $(\gamma_i^{-}\eta,\gamma_i^{+}) \in S_N$ and so we can set $c^{\mathsf{cpl}}(\eta,\gamma_i^{-}\eta,\gamma,\bar\gamma) = c^{\mathsf{cpl}}(\gamma_i^{-}\eta,\gamma_i^{+}\gamma_i^{-}\eta,\bar\gamma,\gamma) = c^{\mathsf{cpl}}(\gamma_i^{-}\eta,\eta,\bar\gamma,\gamma)$.

By Lemma 3.3.1, to prove the theorem it suffices to show that

$$\frac{1}{4}\sum_{(\eta,\sigma)\in S_N}\sum_{\gamma,\bar{\gamma}\in G_N^*} m(\eta)c_N(\eta,\sigma)c^{\mathsf{cpl}}(\eta,\sigma\eta,\gamma,\bar{\gamma})J(\eta,\sigma,\gamma,\bar{\gamma}) \geq \left(\frac{\kappa_*}{2}+M_\theta\overline{\kappa_*}\right)\mathscr{A}_N(\rho,\psi) \tag{3.4.15}$$

for all $\rho\colon \Omega_N\to\mathbb{R}_{>0}$ and $\psi\colon\Omega_N\to\mathbb{R}$.

The left hand side of equation (3.4.15) reads

$$\begin{aligned}&\frac{1}{4}\sum_{(\eta,\gamma_i^+)\in S_N}\sum_{\gamma,\bar{\gamma}\in G_N^*} m(\eta)c_N(\eta,\gamma_i^+)c^{\mathsf{cpl}}(\eta,\gamma_i^+\eta,\gamma,\bar{\gamma})J(\eta,\gamma_i^+,\gamma,\bar{\gamma})\\ +&\frac{1}{4}\sum_{(\eta,\gamma_i^-)\in S_N}\sum_{\gamma,\bar{\gamma}\in G_N^*} m(\eta)c_N(\eta,\gamma_i^-)c^{\mathsf{cpl}}(\gamma_i^-\eta,\eta,\bar{\gamma},\gamma)J(\eta,\gamma_i^-,\gamma,\bar{\gamma}).\end{aligned}$$

Using reversibility (3.2.2) and that $J(\eta,\sigma,\gamma,\bar{\gamma})=J(\sigma\eta,\sigma^{-1},\bar{\gamma},\gamma)$ when $\sigma^{-1}\sigma\eta=\eta$, we get that the second summand is equal to the first and so we can rewrite our quantity as

$$\frac{1}{2}\sum_{(\eta,\gamma_i^+)\in S_N}\sum_{\gamma,\bar{\gamma}\in G_N^*} m(\eta)c_N(\eta,\gamma_i^+)c^{\mathsf{cpl}}(\eta,\gamma_i^+\eta,\gamma,\bar{\gamma})J(\eta,\gamma_i^+,\gamma,\bar{\gamma}).$$

With our explicit choice of coupling rates it follows that we can write the left-hand side of (3.4.15) as $\frac{1}{2}(\tilde{A}+\tilde{B}+\tilde{C}+D)$, where

$$\begin{aligned}\tilde{A} &= \sum_{(\eta,\gamma_i^+)\in S_N}\sum_{\gamma\in G_N} m(\eta)c_N(\eta,\gamma_i^+)\min\Big\{c_N(\eta,\gamma),c_N(\gamma_i^+\eta,\gamma)\Big\}J(\eta,\gamma_i^+,\gamma,\gamma),\\ \tilde{B} &= \sum_{(\eta,\gamma_i^+)\in S_N}\sum_{\substack{\gamma\in G_N\\ \gamma\neq\gamma_i^\pm}} m(\eta)c_N(\eta,\gamma_i^+)\max\{\nabla_i^+c_N(\eta,\gamma),0\}J(\eta,\gamma_i^+,\gamma_i^+,\gamma),\\ \tilde{C} &= \sum_{(\eta,\gamma_i^+)\in S_N}\sum_{\substack{\gamma\in G_N\\ \gamma\neq\gamma_i^\pm}} m(\eta)c_N(\eta,\gamma_i^+)\max\{-\nabla_i^+c_N(\eta,\gamma),0\}J(\eta,\gamma_i^+,\gamma,\gamma_i^-),\\ D &= \sum_{(\eta,\gamma_i^+)\in S_N} m(\eta)c_N(\eta,\gamma_i^+)\Big[\kappa^{+,N}(\eta,i)J(\eta,\gamma_i^+,\gamma_i^+,e)+\kappa^{-,N}(\eta,i)J(\eta,\gamma_i^+,e,\gamma_i^-)\Big].\end{aligned}$$

This is then lower bounded (since $J\geq I$ by (3.3.4)) by $\frac{1}{2}(A+B+C+D)$ where

$$\begin{aligned}A &= \sum_{(\eta,\gamma_i^+)\in S_N}\sum_{\gamma\in G_N} m(\eta)c_N(\eta,\gamma_i^+)\min\Big\{c_N(\eta,\gamma),c_N(\gamma_i^+\eta,\gamma)\Big\}I(\eta,\gamma_i^+,\gamma,\gamma),\\ B &= \sum_{(\eta,\gamma_i^+)\in S_N}\sum_{\substack{\gamma\in G_N\\ \gamma\neq\gamma_i^\pm}} m(\eta)c_N(\eta,\gamma_i^+)\max\{\nabla_i^+c_N(\eta,\gamma),0\}I(\eta,\gamma_i^+,\gamma_i^+,\gamma),\\ C &= \sum_{(\eta,\gamma_i^+)\in S_N}\sum_{\substack{\gamma\in G_N\\ \gamma\neq\gamma_i^\pm}} m(\eta)c_N(\eta,\gamma_i^+)\max\{-\nabla_i^+c_N(\eta,\gamma),0\}I(\eta,\gamma_i^+,\gamma,\gamma_i^-).\end{aligned}$$

Next, using the assumptions (3.4.9) and (3.4.10) we can rewrite the expressions of A, B, C as

$$
\begin{aligned}
A = & \sum_{(\eta,\gamma_i^+)\in S_N} \sum_{j\in[d]} m(\eta) c_N(\eta,\gamma_i^+) c_N(\gamma_i^+\eta,\gamma_j^+) I(\eta,\gamma_i^+,\gamma_j^+,\gamma_j^+) \\
& + \sum_{(\eta,\gamma_i^+)\in S_N} \sum_{j\in[d]} m(\eta) c_N(\eta,\gamma_i^+) c_N(\eta,\gamma_j^-) I(\eta,\gamma_i^+,\gamma_j^-,\gamma_j^-), \\
B = & \sum_{(\eta,\gamma_i^+)\in S_N} \sum_{\substack{j\in[d] \\ j\neq i}} m(\eta) c_N(\eta,\gamma_i^+) \nabla_i^+ c_N(\eta,\gamma_j^-) I(\eta,\gamma_i^+,\gamma_i^+,\gamma_j^-), \\
C = & \sum_{(\eta,\gamma_i^+)\in S_N} \sum_{\substack{j\in[d] \\ j\neq i}} m(\eta) c_N(\eta,\gamma_i^+) \{-\nabla_i^+ c_N(\eta,\gamma_j^+)\} I(\eta,\gamma_i^+,\gamma_j^+,\gamma_i^-).
\end{aligned}
$$

Finally, we notice that we can write

$$
\begin{aligned}
\mathscr{A}_N(\rho,\psi) = & \frac{1}{2} \sum_{(\eta,\gamma_i^+)\in S_N} m(\eta) c_N(\eta,\gamma_i^+) \theta\Big(\rho(\eta),\rho(\gamma_i^+\eta)\Big) \Big[\psi(\eta)-\psi(\gamma_i^+\eta)\Big]^2 \\
& + \frac{1}{2} \sum_{(\eta,\gamma_i^-)\in S_N} m(\eta) c_N(\eta,\gamma_i^-) \theta\Big(\rho(\eta),\rho(\gamma_i^-\eta)\Big) \Big[\psi(\eta)-\psi(\gamma_i^-\eta)\Big]^2 \\
= & \sum_{(\eta,\gamma_i^+)\in S_N} m(\eta) c_N(\eta,\gamma_i^+) \theta\Big(\rho(\eta),\rho(\gamma_i^+\eta)\Big) \Big[\psi(\eta)-\psi(\gamma_i^+\eta)\Big]^2
\end{aligned}
$$

using reversibility (3.2.2) again for the last equality.

In what follows, we will show that $D \geq (2M_\theta \overline{\kappa_*} + \kappa_*)\mathscr{A}_N(\rho,\psi)$ and that $A = B = C = 0$, thus verifying (3.4.15) and concluding the proof of the theorem.

Term D We have

$$
J(\eta,\gamma_i^+,\gamma_i^+,e) = \Big(\theta\Big(\rho(\gamma_i^+\eta),\rho(\gamma_i^+\eta)\Big) + \theta(\rho(\eta),\rho(\gamma_i^+\eta))\Big) \Big[\psi(\eta)-\psi(\gamma_i^+\eta)\Big]^2
$$

and

$$
J(\eta,\gamma_i^+,e,\gamma_i^-) = \Big(\theta(\rho(\eta),\rho(\eta)) + \theta(\rho(\eta),\rho(\gamma_i^+\eta))\Big) \Big[\psi(\eta)-\psi(\gamma_i^+\eta)\Big]^2
$$

and so, remembering also (3.4.13) and (3.4.14),

$$
\begin{aligned}
& \kappa^{+,N}(\eta,i) J(\eta,\gamma_i^+,\gamma_i^+,e) + \kappa^{-,N}(\eta,i) J(\eta,\gamma_i^+,e,\gamma_i^-) \\
\geq & \Big\{\overline{\kappa_*}\Big[\theta\Big(\rho(\gamma_i^+\eta),\rho(\gamma_i^+\eta)\Big) + \theta(\rho(\eta),\rho(\eta))\Big] + \kappa_*\theta\Big(\rho(\eta),\rho(\gamma_i^+\eta)\Big)\Big\} \Big[\psi(\eta)-\psi(\gamma_i^+\eta)\Big]^2 \\
\geq & (2M_\theta\overline{\kappa_*} + \kappa_*)\theta\Big(\rho(\eta),\rho(\gamma_i^+\eta)\Big) \Big[\psi(\eta)-\psi(\gamma_i^+\eta)\Big]^2.
\end{aligned}
$$

Therefore it follows that

$$
D \geq (2M_\theta\overline{\kappa_*} + \kappa_*)\mathscr{A}_N(\rho,\psi).
$$

Other terms We now show that each one of the other terms is 0, concluding the proof of the theorem. To show that $A = B = C = 0$ we proceed similarly to [Con22]. It is useful to have an auxiliary lemma.

Lemma 3.4.20. *For all* $\eta \in \mathbb{N}^d$, $\tilde{\eta} \in \Omega_N$ *and* $i, j \in [d]$

1. $c(\eta,\gamma_j^+)c(\gamma_j^+\eta,\gamma_i^+) = c(\eta,\gamma_i^+)c(\gamma_i^+\eta,\gamma_j^+)$ *and similarly* $c_N(\tilde\eta,\gamma_j^+)c_N(\gamma_j^+\tilde\eta,\gamma_i^+) = c_N(\tilde\eta,\gamma_i^+)c_N(\gamma_i^+\tilde\eta,\gamma_j^+)$.

2. $c(\eta,\gamma_i^-)\nabla_i^- c(\eta,\gamma_j^-) = c(\eta,\gamma_j^-)\nabla_j^- c(\eta,\gamma_i^-)$ *and similarly* $c_N(\tilde\eta,\gamma_i^-)\nabla_i^- c_N(\tilde\eta,\gamma_j^-) = c_N(\tilde\eta,\gamma_j^-)\nabla_j^- c_N(\tilde\eta,\gamma_i^-)$.

3. $c(\eta,\gamma_i^+)\nabla_i^+ c(\eta,\gamma_j^+) = c(\eta,\gamma_j^+)\nabla_j^+ c(\eta,\gamma_i^+)$ *and similarly when* $\eta_i,\eta_j < N$ $c_N(\tilde\eta,\gamma_i^+)\nabla_i^+ c_N(\tilde\eta,\gamma_j^+) = c_N(\tilde\eta,\gamma_j^+)\nabla_j^+ c_N(\tilde\eta,\gamma_i^+)$.

4. $I(\gamma_j^+\eta,\gamma_i^+,\gamma_j^-,\gamma_j^-) = -I(\eta,\gamma_i^+,\gamma_j^+,\gamma_j^+)$.

5. $I(\gamma_i^-\eta,\gamma_i^+,\gamma_i^+,\gamma_j^-) = -I(\gamma_j^-\eta,\gamma_j^+,\gamma_j^+,\gamma_i^-)$ *if* $\eta_i,\eta_j > 0$.

6. $I(\eta,\gamma_i^+,\gamma_j^+,\gamma_i^-) = -I(\eta,\gamma_j^+,\gamma_i^+,\gamma_j^-)$.

Proof of Lemma. Statements 1–3 were already observed in the proof of [Con22, Thm. 3.1] and are easy to check, while statements 4–6 are immediate from the definitions and the assumptions on the model. ■

Term A Using the reversibility property (3.2.2) in the second summand defining A with

$$F(\eta,\sigma) = 1_{G_N^-}(\sigma)\sum_{i\in[d]} c_N(\eta,\gamma_i^+)I(\eta,\gamma_i^+,\sigma,\sigma).$$

we find that

$$\begin{aligned}
&\sum_{\substack{\eta\in\Omega_N\\ i,j\in[d]}} m(\eta)c_N(\eta,\gamma_i^+)c_N(\eta,\gamma_j^-)I(\eta,\gamma_i^+,\gamma_j^-,\gamma_j^-)\\
&= \sum_{\substack{\eta\in\Omega_N\\ i,j\in[d]}} m(\eta)c_N(\eta,\gamma_j^+)c_N(\gamma_j^+\eta,\gamma_i^+)I(\gamma_j^+\eta,\gamma_i^+,\gamma_j^-,\gamma_j^-)\\
&= -\sum_{\substack{\eta\in\Omega_N\\ i,j\in[d]}} m(\eta)c_N(\eta,\gamma_i^+)c_N(\gamma_i^+\eta,\gamma_j^+)I(\eta,\gamma_i^+,\gamma_j^+,\gamma_j^+),
\end{aligned}$$

where in the last equality we have used properties $1.$ and $4.$ of Lemma 3.4.20. This implies that $A = 0$, as desired.

Term B First, using the reversibility property (3.2.2) with

$$F(\eta,\sigma) = 1_{G_N^+}(\sigma)\sum_{\gamma\in G_N^-,\gamma\neq\sigma,\sigma^{-1}} \nabla_\sigma c_N(\eta,\gamma)I(\eta,\sigma,\sigma,\gamma),$$

and the fact that $\gamma_i^+\gamma_i^-\eta = \eta$ if $\eta_i > 0$ we find

$$\begin{aligned}
B &= \sum_{\substack{\eta\in\Omega_N,\\ i\neq j\in[d]}} m(\eta)c_N(\eta,\gamma_i^-)\nabla_i^+ c_N(\gamma_i^-\eta,\gamma_j^-)I(\gamma_i^-\eta,\gamma_i^+,\gamma_i^+,\gamma_j^-)\\
&= \sum_{\substack{\eta\in\Omega_N,\\ i\neq j\in[d]}} m(\eta)c_N(\eta,\gamma_i^-)\{-\nabla_i^- c_N(\eta,\gamma_j^-)\}I(\gamma_i^-\eta,\gamma_i^+,\gamma_i^+,\gamma_j^-).
\end{aligned} \tag{3.4.16}$$

By exchanging i, j first and then using properties $5.$ and $2.$ of Lemma 3.4.20 we deduce that

$$\begin{aligned}
B &= \sum_{\substack{\eta\in\Omega_N,\\ i\neq j\in[d]}} m(\eta)c_N(\eta,\gamma_j^-)\{-\nabla_j^- c_N(\eta,\gamma_i^-)\}I(\gamma_j^-\eta,\gamma_j^+,\gamma_j^+,\gamma_i^-)\\
&= -\sum_{\substack{\eta\in\Omega_N,\\ i\neq j\in[d]}} m(\eta)c_N(\eta,\gamma_j^-)\{-\nabla_j^- c_N(\eta,\gamma_i^-)\}I(\gamma_i^-\eta,\gamma_i^+,\gamma_i^+,\gamma_j^-)\\
&= -\sum_{\substack{\eta\in\Omega_N,\\ i\neq j\in[d]}} m(\eta)c_N(\eta,\gamma_i^-)\{-\nabla_i^- c_N(\eta,\gamma_j^-)\}I(\gamma_i^-\eta,\gamma_i^+,\gamma_i^+,\gamma_j^-).
\end{aligned}$$

Comparing this with the expression of B in (3.4.16) we deduce that $B=-B$, hence $B=0$.

Term C We have

$$\begin{aligned}
C &= \sum_{\substack{\eta\in\Omega_N,\\ i\neq j\in[d]}} m(\eta)c_N(\eta,\gamma_i^+)\{-\nabla_i^+ c_N(\eta,\gamma_j^+)\}I(\eta,\gamma_i^+,\gamma_j^+,\gamma_i^-)\\
&= \sum_{\substack{\eta\in\Omega_N,\\ i\neq j\in[d]}} m(\eta)c_N(\eta,\gamma_j^+)\{-\nabla_j^+ c_N(\eta,\gamma_i^+)\}I(\eta,\gamma_j^+,\gamma_i^+,\gamma_j^-)\\
&= -\sum_{\substack{\eta\in\Omega_N,\\ i\neq j\in[d]}} m(\eta)c_N(\eta,\gamma_j^+)\{-\nabla_j^+ c_N(\eta,\gamma_i^+)\}I(\eta,\gamma_i^+,\gamma_j^+,\gamma_i^-)\\
&= -\sum_{\substack{\eta\in\Omega_N,\\ i\neq j\in[d]}} m(\eta)c_N(\eta,\gamma_i^+)\{-\nabla_i^+ c_N(\eta,\gamma_j^+)\}I(\eta,\gamma_i^+,\gamma_j^+,\gamma_i^-)
\end{aligned}$$

where we have exchanged i, j in the first equality and used properties $6.$ and $3.$ of Lemma 3.4.20 in the last two. This shows that $C=-C$, hence $C=0$.

This concludes the proof of the theorem.

Examples of interacting random walks

As anticipated, as an application of Theorem 3.4.17 and looking for a discrete analogue of Proposition 3.4.16, we now revisit some particular examples of interacting random walks considered in [Con22] . In this subsection, we stick to the particular choice of V^- given in (3.4.8) (for which (3.4.10) is satisfied) and we simply write $V=V^+$. Notice first of all that our assumption (3.4.9) can be written equivalently as

$$\nabla_i^+\nabla_j^+V(\eta)\geq 0 \quad \text{for all } \eta\in\mathbb{N}^d, i,j\in[d]. \tag{3.4.17}$$

Interestingly, while in Proposition 3.4.16 the key assumption was the convexity of V (i.e. the positive semi-definiteness of the Hessian $\nabla^2 V$), here we see that the non-negativity of the entries of a "discrete Hessian" of the potential V comes into play.

Under this assumption (3.4.17), the conditions (3.4.11) and (3.4.12) were checked by Conforti [Con22], and in particular we have the following

Corollary 3.4.21 (Cor 3.2 of [Con22])**.** *With the notation of this section, suppose that* (3.4.17) *holds and that for all* $\eta\in\mathbb{N}^d, i\in[d]$

$$\lambda-\sum_{\substack{j=1,\\ j\neq i}}^{d}\left[e^{-\nabla_j^+V(\eta)}-e^{-\nabla_j^+V(\gamma_i^+\eta)}\right]\geq 0.$$

Then, the assumptions of Theorem 3.4.17 are satisfied and

$$\kappa_* = \inf_{\substack{\eta \in \mathbb{N}^d \\ i \in [d]}} \lambda + \Big[e^{-\nabla_i^+ V(\eta)} - e^{-\nabla_i^+ V(\gamma_i^+ \eta)}\Big] - \sum_{\substack{j=1, \\ j \neq i}}^{d} \Big[e^{-\nabla_j^+ V(\eta)} - e^{-\nabla_j^+ V(\gamma_i^+ \eta)}\Big].$$

If in addition

$$\min_{i \in [d]} \lambda - \sum_{\substack{j=1 \\ j \neq i}} e^{-\nabla_j^+ V(\mathbf{0})} \geq 0 \tag{3.4.18}$$

then κ_* is bounded from below by the expression in (3.4.18).

We consider now a particular example of potential V satisfying (3.4.17). Given a function $h \colon \mathbb{R}_{\geq 0} \to \mathbb{R}$ and some $\beta > 0$, set

$$V(\eta) \coloneqq \beta h(|\eta|),$$

where $|\eta| = \sum_i \eta_i$ for $\eta \in \mathbb{N}^d$. Using the notation $\nabla^+ h(m) \coloneqq h(m+1) - h(m)$, Conforti observed the following

Corollary 3.4.22 (Cor 3.1 of [Con22])**.** *With the notation of this section, suppose that h is* ***convex*** *and that*

$$\inf_{m \in \mathbb{N}} \lambda - (d-1)\Big[e^{-\beta \nabla^+ h(m)} - e^{-\beta \nabla^+ h(m+1)}\Big] \geq 0.$$

Then the assumptions of Theorem 3.4.17 are satisfied and

$$\kappa_* = \inf_{m \in \mathbb{N}} \lambda - (d-2)\Big[e^{-\beta \nabla^+ h(m)} - e^{-\beta \nabla^+ h(m+1)}\Big].$$

In particular, if $h(1) > h(0)$ and

$$\beta \geq \frac{\log(d-1) - \log(\lambda)}{h(1) - h(0)}$$

then the assumptions of Theorem 3.4.17 are satisfied with

$$\kappa_* \geq \lambda - (d-2)e^{-\beta \nabla^+ h(0)}.$$

Interestingly, as before, a notion of convexity of the potential is naturally involved. Note also that the necessary condition of a lower bound on β means this is a non perturbative criterion, since the resulting reversible measure is far from being a product measure. It is known that product measures behave well with the entropic curvature, i.e. they tensorize (see Theorem 6.2 of [EM12]). Therefore, it is particularly interesting to have conditions implying positive entropic curvature for a Markov chain whose stationary measure is not close to a product measure.

3.5 Couplings and coarse Ricci curvature

In this section, we recall some well-known related definitions of curvature for discrete and continuous time Markov chains (generally referred to as "coarse Ricci curvature"), give some natural generalizations and discuss the relations among them and with the concept of coupling. As an application, we show that in all the examples of Section 3.4 (except for Section 3.4.4 -

which was however covered in [Con22, Sec. 3.3] - since we restrict to the case of finite state space Markov chains), when the assumptions of the respective main theorems are satisfied then the coarse curvature is positive, and for all starting probability densities we have exponential contraction of the p-Wasserstein distances along the heat flow P_t (see the precise details later). The coarse Ricci curvature was first introduced by Ollivier for discrete time Markov chains (see [Oll09, Oll10]) and later modified to apply to continuous time models (see [LLY11, Vey12]). Compared to the entropic Ricci curvature, it is often easier to establish positive curvature for this notion. However, it is not known whether positive coarse Ricci curvature implies some functional inequalities, and in particular the modified log-Sobolev inequality (see Section 3.5.4 for a discussion).

Throughout this section, we switch to a more standard notation, and we don't employ the description of the Markov chain in terms of its allowed moves. As anticipated, we always assume in this section that we are working with irreducible and reversible Markov chains on a finite state space. We use the letter Ω for the state space, x, y, v, w, z for elements of Ω, P for a stochastic transition matrix and L, Q for a generator L with transition rates Q, so that the action of L on a function $\psi\colon \Omega \to \mathbb{R}$ is given by

$$L\psi(x) = \sum_{y\in\Omega} Q(x,y)(\psi(y) - \psi(x)).$$

Here we have $Q(x, y) \geq 0$ and we do not assume $Q(x, x) = 0$, and in fact we often change the value of $Q(x, x)$ (without loss of generality) depending on convenience; on the other hand, we sometimes identify L, as a linear operator, with a matrix, in which case by construction $L(x, x) = -\sum_{y\neq x} Q(x, y)$. We typically identify measures with row vectors: in particular, let δ_x be the row vector with entry 1 corresponding to x and 0 everywhere else, which is identified with the Dirac measure at x; on the other hand, densities with respect to π and other functions on the state space are identified with column vectors. We also introduce a simple graph structure, where $x \sim y$ if and only if $x \neq y$ and $P(x, y) > 0$ or $Q(x, y) > 0$ respectively, and correspondingly consider the unweighted graph distance d. With respect to this graph distance d, we will consider the p-Wasserstein distances W_p. Couplings for the transition measures/rates from starting points $x \neq y \in \Omega$ will be described by non-negative functions $\Pi(x, y, \cdot, \cdot), C(x, y, \cdot, \cdot)\colon \Omega \times \Omega \to \mathbb{R}_{\geq 0}$ respectively in discrete and continuous time, so that

$$\begin{cases} \sum_{w\in\Omega} \Pi(x,y,w,z) &= P(y,z) \text{ for all } z\in\Omega, \\ \sum_{z\in\Omega} \Pi(x,y,w,z) &= P(x,w) \text{ for all } w\in\Omega, \end{cases} \tag{3.5.1}$$

$$\begin{cases} \sum_{w\in\Omega} C(x,y,w,z) &= Q(y,z) \text{ for all } z\in\Omega, \\ \sum_{z\in\Omega} C(x,y,w,z) &= Q(x,w) \text{ for all } w\in\Omega. \end{cases} \tag{3.5.2}$$

Similarly to what we observed in Section 3.3, the set of such admissible couplings is non-empty, provided that in continuous time one redefines $Q(x, x)$ appropriately without loss of generality: notice indeed that the existence of the coupling rates implies (by summing over $w, z \in \Omega$) that $\sum_w Q(x, w) = \sum_z Q(y, z) =: Z$; if this holds, the "product" coupling rates $C(x, y, w, z) = \frac{1}{Z} Q(x, w)Q(y, z)$ are admissible. Notice also that, to compare with the notation of the previous sections of this work, we could write, for states $x, y, w, z \in \Omega$ and enlarged set of moves G^*,

$$C(x,y,w,z) = \sum_{\substack{\gamma,\bar\gamma\in G^* \\ \gamma x = w, \bar\gamma y = z}} c^{\mathsf{cpl}}(x,y,\gamma,\bar\gamma).$$

Preliminaries definitions

Before turning to a detailed discussion of the coarse Ricci curvature, we recall a few definitions from optimal transport, which will be needed in the sequel. Throughout this section, we denote by $\mathcal{P}(X)$ the set of probability measures on a space X.

In what follows, we let X, Y be two finite sets. Given two probability measures $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$, we denote by $\Gamma(\mu,\nu)$ the family of couplings between μ and ν, i.e. the family of probability measures $\gamma \in \mathcal{P}(X \times Y)$ having marginals μ and ν respectively.

Given a cost function $c\colon X \times Y \to \mathbb{R}_{\geq 0}$, the optimal transport cost $\mathcal{T}_c(\mu,\nu)$ between two probability measures $\mu \in \mathcal{P}(X), \nu \in \mathcal{P}(Y)$ is defined by

$$\mathcal{T}_c(\mu,\nu) = \inf_{\gamma \in \Gamma(\mu,\nu)} \sum_{x \in X, y \in Y} c(x,y)\gamma(x,y). \tag{3.5.3}$$

When $X = Y$ and we are given a distance d on X, we can define the Wasserstein distance of any order $p \in (1,\infty)$ as follows: for $\mu, \nu \in \mathcal{P}(X)$

$$W_p^p(\mu,\nu) = \mathcal{T}_{d^p}(\mu,\nu) = \inf_{\gamma \in \Gamma(\mu,\nu)} \sum_{x \in X, y \in Y} d(x,y)^p \gamma(x,y).$$

When discussing the coarse Ricci curvature of a Markov chain with finite state space Ω, we will consider the Wasserstein distances with respect to the natural graph distance d mentioned before.

Finally, we will consider also the total variation distance between two probability measures $\mu, \nu \in \mathcal{P}(X)$, which is defined by

$$\|\mu - \nu\|_{\mathsf{TV}} = \frac{1}{2} \sum_{x \in X} |\mu(x) - \nu(x)| \in [0,1].$$

3.5.1 Discrete time

An important notion of curvature for Markov chains is the *coarse Ricci curvature* introduced by Ollivier (see [Oll09] and [Oll10]).

Definition 3.5.1. *Given $p \geq 1$ and $x \neq y$, we say that the Markov chain has (discrete time) p-coarse Ricci curvature $K_{dc,p}(x,y)$ in direction (x,y) if*

$$W_p(\delta_x P, \delta_y P) = (1 - K_{dc,p}(x,y))d(x,y).$$

Remark 3.5.2. *Ollivier focused in particular on the case $p = 1$; in this paper, however, it will be useful to consider also other values of p.*

We also give the following definition, inspired by the properties of the couplings constructed in the previous sections.

Definition 3.5.3. *For $x \neq y$, we define the (discrete time) ∞-coarse Ricci curvature $K_{dc,\infty}(x,y)$ in direction (x,y) to be the supremum of all $K \in \mathbb{R}$ such that there exists a coupling $\Pi(x,y,\cdot,\cdot)$ of $(\delta_x P, \delta_y P)$ satisfying:*

- $\sum_{w,z \in \Omega} \Pi(x,y,w,z)\mathbb{1}_{d(w,z) > d(x,y)} = 0$;

- $\sum_{w,z\in\Omega}\Pi(x,y,w,z)d(w,z)\leq(1-K)d(x,y)$.

We use the convention that $\sup\emptyset=-\infty$.

Remark 3.5.4. *We have that $K_{dc,\infty}(x,y)\in\mathbb{R}_{\geq 0}\cup\{-\infty\}$; if $K_{dc,\infty}(x,y)\geq 0$ then the supremum in the definition is attained. Notice that, equivalently, $K_{dc,\infty}(x,y)\geq 0$ means that there exists a coupling (X,Y) of the one-step probability distributions $(\delta_x P,\delta_y P)$ which proves $K_{dc,1}(x,y)\geq K_{dc,\infty}(x,y)$ (i.e. $\mathbb{E}[d(X,Y)]\leq(1-K_{dc,\infty}(x,y))d(x,y)$) and at the same time satisfies $d(X,Y)\leq d(x,y)$ almost surely.*

For $p\in[1,\infty]$ we write

$$\mathrm{Ric_{dc,p}}\geq K$$

if $K_{dc,p}(x,y)\geq K$ for all $x\neq y\in\Omega$. The next proposition collects some useful results.

Proposition 3.5.5. *The following hold:*

1. *For $p\in[1,\infty]$, if $K_{dc,p}(x,y)\geq K$ for all $x\sim y$ then $\mathrm{Ric_{dc,p}}\geq K$.*
2. *For $1\leq p\leq q<\infty$ we have $K_{dc,p}(x,y)\geq K_{dc,q}(x,y)$. Moreover if $x\sim y$ we have*
$$K_{dc,p}(x,y)\geq 1-(1-K_{dc,\infty}(x,y))^{\frac{1}{p}}\geq\frac{K_{dc,\infty}(x,y)}{p}.$$
3. *If $x\sim y$ and $\lim_{p\to\infty}K_{dc,p}(x,y)\geq 0$ we have*
$$K_{dc,\infty}(x,y)\geq 1-e^{-\limsup_{p\to\infty}p\cdot K_{dc,p}(x,y)}.$$
4. *For $p\in[1,\infty)$, if $\mathrm{Ric_{dc,p}}\geq K$ then for any starting probability measures μ,ν and $n\geq 0$ we have*
$$W_p(\mu P^n,\nu P^n)\leq(1-K)^n W_p(\mu,\nu).$$

Proof. 1. If $p<\infty$, this is done as in [Oll09, Prop. 19]: suppose $d(x,y)=n$ and let $x=z_0\sim z_1\sim\ldots\sim z_n=y$. Then

$$W_p(\delta_x P,\delta_y P)\leq\sum_{i=0}^{n-1}W_p\Big(\delta_{z_{i-1}}P,\delta_{z_i}P\Big)\leq(1-K)n=(1-K)d(x,y).$$

Suppose now that $p=\infty$: if $K=-\infty$ the conclusion is trivial, hence assume that $K\geq 0$. Let again $n=d(x,y)$ and $x=z_0\sim z_1\sim\ldots\sim z_n=y$: we prove the claim by induction over n. The base case $n=1$ follows directly by the assumption. Now suppose $n>1$ and that the inductive hypothesis holds. Let $\Pi(x,z_{n-1},\cdot,\cdot)$ and $\Pi(z_{n-1},y,\cdot,\cdot)$ be such that

$$\begin{aligned}
&\sum_{v,w\in\Omega}\Pi(x,z_{n-1},v,w)\mathbb{1}_{d(v,w)>d(x,z_{n-1})}=0,\\
&\sum_{v,w\in\Omega}\Pi(x,z_{n-1},v,w)d(v,w)\leq(1-K)d(x,z_{n-1}),\\
&\sum_{v,w\in\Omega}\Pi(z_{n-1},y,v,w)\mathbb{1}_{d(v,w)>1}=0,\\
&\sum_{v,w\in\Omega}\Pi(z_{n-1},y,v,w)d(v,w)\leq(1-K).
\end{aligned}$$

By the Gluing lemma [Vil09], there exists $\hat{\Pi}(\cdot,\cdot,\cdot) = \hat{\Pi}(x, z_{n-1}, y, \cdot,\cdot,\cdot) \in \mathcal{P}(\Omega\times\Omega\times\Omega)$ such that $p_{1,2}\#\hat{\Pi} = \Pi(x, z_{n-1}, \cdot,\cdot)$ and $p_{2,3}\#\hat{\Pi} = \Pi(z_{n-1}, y, \cdot,\cdot)$, where $p_{i,j}$ is the projection on coordinates i, j and $\#$ denotes the pushforward of a measure via a map. The measure $\Pi(x, y, \cdot,\cdot) := p_{1,3}\#\hat{\Pi}$ then realizes a coupling with the desired properties, since, given that $d(v, w) \le d(v, s) + d(s, w)$, we have

$$\begin{aligned}
&\sum_{v,w\in\Omega} \Pi(x, y, v, w) d(v, w)\\
=&\sum_{v,w,s\in\Omega} \hat{\Pi}(x, z_{n-1}, y, v, s, w) d(v, w)\\
\le&\sum_{v,w,s\in\Omega} \hat{\Pi}(x, z_{n-1}, y, v, s, w)(d(v, s) + d(s, w))\\
=&\sum_{v,s\in\Omega} \Pi(x, z_{n-1}, v, s) d(v, s) + \sum_{s,w\in\Omega} \Pi(z_{n-1}, y, s, w) d(s, w)\\
\le&(1 - K)(d(x, z_{n-1}) + d(z_{n-1}, y))\\
=&(1 - K) d(x, y),
\end{aligned}$$

and similarly

$$\begin{aligned}
&\sum_{v,w\in\Omega} \Pi(x, y, v, w) \mathbb{1}_{d(v,w)>d(x,y)}\\
=&\sum_{v,w,s\in\Omega} \hat{\Pi}(x, z_{n-1}, y, v, s, w) \mathbb{1}_{d(v,w)>d(x,y)}\\
\le&\sum_{v,w,s\in\Omega} \hat{\Pi}(x, z_{n-1}, y, v, s, w) \mathbb{1}_{d(v,s)+d(s,w)>d(x,z_{n-1})+d(z_{n-1},y)}\\
\le&\sum_{v,w,s\in\Omega} \hat{\Pi}(x, z_{n-1}, y, v, s, w)\Big(\mathbb{1}_{d(v,s)>d(x,z_{n-1})} + \mathbb{1}_{d(s,w)>d(z_{n-1},y)}\Big)\\
=&\sum_{v,s\in\Omega} \Pi(x, z_{n-1}, v, s)\mathbb{1}_{d(v,s)>d(x,z_{n-1})} + \sum_{s,w\in\Omega} \Pi(z_{n-1}, y, s, w)\mathbb{1}_{d(s,w)>d(z_{n-1},y)}\\
=&0.
\end{aligned}$$

2. The first statement follows by the inequality $W_p(\mu, \nu) \le W_q(\mu, \nu)$ for $1 \le p \le q < \infty$. For the second statement, suppose that $x \sim y$ and $K := K_{dc,\infty}(x, y) \ge 0$. Let Π be the optimal coupling in the definition of $K_{dc,\infty}(x, y)$. Then notice that

$$W_p(\delta_x P, \delta_y P) \le [1 - K]^{\frac{1}{p}} \le 1 - \frac{K}{p},$$

from which the conclusion follows.

3. Let $K = \limsup_{p\to\infty} p \cdot K_{dc,p}(x, y) \ge 0$ and consider a sequence $(p_n)_n \subset [1, \infty)$ such that $p_n \to +\infty$ as $n \to \infty$, and denote by $\Pi_n(x, y, \cdot,\cdot)$ associated couplings of $(\delta_x P, \delta_y P)$ that show $W_{p_n}(\delta_x P, \delta_y P) \le (1 - K_n)$, i.e.

$$\left(\sum_{w,z} \Pi_n(x, y, w, z) d(w, z)^{p_n}\right)^{\frac{1}{p_n}} \le 1 - K_n.$$

By viewing $\Pi_n(x, y, \cdot,\cdot)$ as elements of $[0, 1]^{\Omega\times\Omega}$ and by a compactness argument (recall that Ω is finite), we can pass to a subsequence (which we denote in the same way) and assume that there exists $\Pi \in [0,1]^{\Omega\times\Omega}$ such that $\Pi_n(x,y,\cdot,\cdot) \to \Pi(x,y,\cdot,\cdot)$ entrywise as $n\to\infty$. It is straightforward to check that Π is still a coupling of $(\delta_x P, \delta_y P)$, and we claim that Π has the desired properties. Indeed, by construction for $n \geq 1$ we have

$$\begin{aligned}
&2^{p_n}\sum_{w,z}\Pi_n(x,y,w,z)\mathbb{1}_{d(w,x)\geq 2} + \sum_{w,z}\Pi_n(x,y,w,z)\mathbb{1}_{d(w,x)=1}\\
\leq & W_{p_n}^{p_n}(\delta_x P, \delta_y P)\\
\leq & (1-K_n)^{p_n}\\
\leq & e^{-K_n p_n}.
\end{aligned}$$

Letting $n\to\infty$ in the previous we deduce

$$\begin{cases}
\sum_{w,z}\Pi(x,y,w,z)\mathbb{1}_{d(w,x)\geq 2} = 0,\\
\sum_{w,z}\Pi(x,y,w,z)\mathbb{1}_{d(w,x)=1} \leq e^{-K},
\end{cases}$$

which yields the desired conclusion.

4. Let $\Pi_{\mu,\nu}$ and $\Pi_{x,y}$ be the optimal couplings in the definitions of $W_p(\mu,\nu)$ and $K_{dc,p}(x,y)$ respectively. Then

$$\begin{aligned}
W_p^p(\mu P, \nu P) &\leq \sum_{x,y} d(x,y)^p \sum_{w,z}\Pi_{\mu,\nu}(w,z)\Pi_{w,z}(x,y)\\
&\leq (1-K)^p\sum_{w,z}\Pi_{\mu,\nu}(w,z)d(w,z)^p\\
&\leq (1-K)^p W_p^p(\mu,\nu).
\end{aligned}$$

The conclusion follows by induction over n.

■

3.5.2 Continuous time

In this subsection, we describe the analogous notions of coarse Ricci curvature for continuous time Markov chains with generator L on a finite state space (see [Vey12, LLY11, MW19]). Recalling that we identify L with a matrix with zero row sums, we use the notation

$$\tilde{P}_t = I + tL,$$

so that for $t>0$ small enough $\tilde{P}_t$ is a stochastic matrix too and we expect it to approximate the transition matrix P_t (given it corresponds to the first order Taylor expansion of $P_t = e^{tL}$). The next definition is motivated by the study of the idleness function in [BCL+18].

Definition 3.5.6. *Let* $T = (\max_{x\in\Omega} -L(x,x))^{-1} = \left(\max_{x\in\Omega}\sum_{y\neq x} Q(x,y)\right)^{-1}$. *Fix* $x \neq y \in \Omega$ *and let* $p\in[1,\infty)$. *Then we define the function* $I_{p,x,y}\colon [0,T]\to\mathbb{R}$ *by*

$$I_{p,x,y}(t) \coloneqq W_p^p(\delta_x\tilde{P}_t, \delta_y\tilde{P}_t).$$

For $0\leq t\leq T$ we have that $\tilde{P}_t$ is a stochastic matrix, so $\delta_x\tilde{P}_t$ and $\delta_y\tilde{P}_t$ are probability measures on Ω and $I_{p,x,y}(t)$ is well-defined. For simplicity, we drop the subscripts x,y when there is no confusion. Notice that $I_p(0) = d(x,y)^p$. The next propositions and lemma are straightforward adaptations of the results in [BCL+18], where the case $p=1$ was considered.

Our motivation here is to establish linearity of $t \to I_p(t)$ for small values of t, hence the differentiability at $t = 0$ (cf. Proposition 3.5.9): this allows us to consider a first definition of coarse Ricci curvature in continuous time, cf. Definition 3.5.11. This definition was first given for $p = 1$ in [LLY11] for combinatorial graphs (i.e. simple random walks on graphs), and later studied in a more general setting in [MW19].

Proposition 3.5.7. *The function $t \to I_p(t)$ is convex.*

Proof. We use Kantorovich duality: for a function $\psi\colon \Omega \to \mathbb{R}$, we denote by $\psi^{c,p}$ its d^p-transform, i.e.

$$\psi^{c,p}(x) := \inf_y \{d(x,y)^p - \psi(y)\}.$$

Then we have for $t \in [0,T]$

$$\begin{aligned}
&W_p^p(\delta_x \tilde{P}_t, \delta_y \tilde{P}_t) \\
=&\sup_\psi \left\{ \sum_{z\in\Omega} \psi(z) \cdot \Big(\delta_x \tilde{P}_t\Big)(z) + \psi^{c,p}(z) \cdot \Big(\delta_y \tilde{P}_t\Big)(z) \right\} \\
=&\sup_\psi \left\{ \psi(x)\left(1 - t\sum_z Q(x,z)\right) + \psi^{c,p}(y)\left(1 - t\sum_z Q(y,z)\right) \right. \\
&\left. +t\left[\sum_z Q(x,z)\psi(z) + Q(y,z)\psi^{c,p}(z)\right] \right\}.
\end{aligned}$$

This is the supremum of affine functions of t, hence $t \to W_p^p(\delta_x \tilde{P}_t, \delta_y \tilde{P}_t) = I_{p,x,y}(t)$ is convex. ∎

Lemma 3.5.8. *Let $0 \le t_1 < t_2 \le T$ and suppose that $(\psi, \psi^{c,p})$ is a pair of optimal Kantorovich potentials in the definition of both $W_p(\delta_x \tilde{P}_{t_1}, \delta_y \tilde{P}_{t_1})$ and $W_p(\delta_x \tilde{P}_{t_2}, \delta_y \tilde{P}_{t_2})$. Then $t \to I_p(t)$ is linear over $[t_1, t_2]$.*

Proof. We already know the function $t \to W_p^p\Big(\delta_x \tilde{P}_t, \delta_y \tilde{P}_t\Big)$ is convex, hence is suffices to show it is also concave over $[t_1, t_2]$. This follows by the assumption using Kantorovich duality, since for $\alpha \in [0,1]$

$$\begin{aligned}
&W_p^p\Big(\delta_x \tilde{P}_{\alpha t_1 + (1-\alpha)t_2}, \delta_y \tilde{P}_{\alpha t_1 + (1-\alpha)t_2}\Big) \\
\ge&\sum_{z\in\Omega} \psi(z) \cdot \Big(\delta_x \tilde{P}_{\alpha t_1 + (1-\alpha)t_2}\Big)(z) + \psi^{c,p}(z) \cdot \Big(\delta_y \tilde{P}_{\alpha t_1 + (1-\alpha)t_2}\Big)(z) \\
=&\psi(x)\left(1 - (\alpha t_1 + (1-\alpha)t_2)\sum_z Q(x,z)\right) + \psi^{c,p}(y)\left(1 - (\alpha t_1 + (1-\alpha)t_2)\sum_z Q(y,z)\right) \\
&+(\alpha t_1 + (1-\alpha)t_2)\left[\sum_z Q(x,z)\psi(z) + Q(y,z)\psi^{c,p}(z)\right] \\
=&\alpha W_p^p\Big(\delta_x \tilde{P}_{t_1}, \delta_y \tilde{P}_{t_1}\Big) + (1-\alpha) W_p^p\Big(\delta_x \tilde{P}_{t_2}, \delta_y \tilde{P}_{t_2}\Big).
\end{aligned}$$

∎

Proposition 3.5.9. *The function I_p is linear over $[0, \frac{T}{2}]$.*

Proof. Fix $t < \frac{T}{2}$ and let ψ and γ be a couple of optimal Kantorovich potential and optimal transport plan for $W_p\big(\delta_x\tilde{P}_t, \delta_y\tilde{P}_t\big)$. Then, it is easy to see that $\gamma(x,y) > 0$ (since the laziness of $\tilde{P}_t$ is at least $1 - t/T > \frac{1}{2}$, and so $\delta_x\tilde{P}_t(x), \delta_y\tilde{P}_t(y) > \frac{1}{2}$). It follows that $\psi(x) + \psi^c(y) = d(x,y)^p$ (cf. [Vil09, Thm. 5.10] for example). This implies that (ψ, ψ^c) is optimal also for $W_p(\delta_x\tilde{P}_0, \delta_y\tilde{P}_0) = W_p(\delta_x, \delta_y) = d(x,y)$. The conclusion follows by Lemma 3.5.8 and by letting $t \to \frac{T}{2}$ (by continuity at $\frac{T}{2}$, which follows from the convexity of I_p). ■

Since linear functions are easily differentiated, the previous proposition immediately implies the following corollary, which gives an expression for the derivative of $I_p(t)$ at $t = 0$.

Corollary 3.5.10. *For $x \neq y$ there exists*

$$\frac{d}{dt}\bigg|_{t=0} W_p^p\Big(\delta_x\tilde{P}_t, \delta_y\tilde{P}_t\Big) = -\frac{1}{s}\Big(d(x,y)^p - W_p^p\Big(\delta_x\tilde{P}_s, \delta_y\tilde{P}_s\Big)\Big) \tag{3.5.4}$$

for all $0 < s \leq \frac{T}{2}$.

Definition 3.5.11. *The continuous time p-coarse Ricci curvature in direction (x,y) is defined via*

$$\begin{aligned} K_{cc,p}(x,y) &:= -\frac{1}{d(x,y)}\frac{d}{dt}\bigg|_{t=0} W_p\Big(\delta_x\tilde{P}_t, \delta_y\tilde{P}_t\Big) \\ &= -\frac{1}{p\cdot d(x,y)^p}\frac{d}{dt}\bigg|_{t=0} W_p^p\Big(\delta_x\tilde{P}_t, \delta_y\tilde{P}_t\Big) \\ &= \frac{1}{sp}\left(1 - \frac{W_p^p\Big(\delta_x\tilde{P}_s, \delta_y\tilde{P}_s\Big)}{d(x,y)^p}\right) \end{aligned} \tag{3.5.5}$$

for all $0 < s \leq \frac{T}{2}$.

Remark 3.5.12. *The original definition in [LLY11] (see also [MW19]) focused on the case $p = 1$, but in this paper it will be useful to consider also other values of p.*

Next, we prove a few preliminary results needed for the proof of Corollary 3.5.16, which will show that we can define $K_{cc,p}$ also by involving P_t instead of $\tilde{P}_t$. For $p = 1$, this shows the equivalence with the notion of continuous time coarse Ricci curvature defined in [Vey12], which was first proved in [MW19, Thm. 5.8].

Lemma 3.5.13. *Let X, Y be finite sets and $\mu, \tilde{\mu} \in \mathcal{P}(X), \nu \in \mathcal{P}(Y)$ be probability measures. Suppose that $\gamma \in \Gamma(\mu,\nu)$ is a coupling between μ and ν. Then, there exists a coupling $\tilde{\gamma} \in \Gamma(\tilde{\mu},\nu)$ between $\tilde{\mu}$ and ν that satisfies*

$$\|\tilde{\gamma} - \gamma\|_{TV} = \|\tilde{\mu} - \mu\|_{TV}.$$

Proof. We will construct a coupling $\tilde{\gamma} \in \Gamma(\tilde{\mu},\nu)$ such that

$$\tilde{\gamma}(x,y) \geq \gamma(x,y) \text{ if and only if } \tilde{\mu}(x) \geq \mu(x). \tag{3.5.6}$$

Notice that this implies the thesis since we would have

$$\begin{aligned}
\|\tilde{\gamma}-\gamma\|_{\mathsf{TV}} &= \frac{1}{2}\sum_{x\in X, y\in Y}|\tilde{\gamma}(x,y)-\gamma(x,y)| \\
&= \frac{1}{2}\sum_{x,y:\tilde{\mu}(x)\geq\mu(x)}[\tilde{\gamma}(x,y)-\gamma(x,y)] + \frac{1}{2}\sum_{x,y:\tilde{\mu}(x)<\mu(x)}[\gamma(x,y)-\tilde{\gamma}(x,y)] \\
&= \frac{1}{2}\sum_{x:\tilde{\mu}(x)\geq\mu(x)}[\tilde{\mu}(x)-\mu(x)] + \frac{1}{2}\sum_{x:\tilde{\mu}(x)<\mu(x)}[\mu(x)-\tilde{\mu}(x)] \\
&= \frac{1}{2}\sum_{x\in X}|\tilde{\mu}(x)-\mu(x)| \\
&= \|\tilde{\mu}-\mu\|_{\mathsf{TV}}.
\end{aligned}$$

Let us thus construct $\tilde{\gamma}$ that satisfies (3.5.6). To do that, for every $x \in X, y \in Y$ set

$$\begin{aligned}
\alpha(x) &= \tilde{\mu}(x) - (\tilde{\mu}\wedge\mu)(x), \\
\beta(y) &= \nu(y) - \sum_{x\in X:\tilde{\mu}(x)<\mu(x)}\gamma(x,y)\frac{\tilde{\mu}(x)}{\mu(x)} - \sum_{x\in X:\tilde{\mu}(x)\geq\mu(x)}\gamma(x,y).
\end{aligned}$$

It is easy to see that $\sum_{x\in X}\alpha(x) = \sum_{y\in Y}\beta(y) = \|\mu-\tilde{\mu}\|_{\mathsf{TV}}$. Then we define

$$\tilde{\gamma}(x,y) \coloneqq \mathbb{1}_{\tilde{\mu}(x)<\mu(x)}\frac{\tilde{\mu}(x)}{\mu(x)}\gamma(x,y) + \mathbb{1}_{\tilde{\mu}(x)\geq\mu(x)}\gamma(x,y) + \frac{1}{\|\mu-\tilde{\mu}\|_{\mathsf{TV}}}\alpha(x)\beta(y).$$

It is easy to check that $\tilde{\gamma}\in\Gamma(\tilde{\mu},\nu)$ and that $\tilde{\gamma}$ satisfies (3.5.6). ■

Corollary 3.5.14. *Let X, Y be finite sets, $\mu_1,\mu_2\in\mathcal{P}(X)$, $\nu_1,\nu_2\in\mathcal{P}(Y)$ and $c\colon X\times Y\to\mathbb{R}_{\geq 0}$ be a cost function. For any probability measures $\mu\in\mathcal{P}(X),\nu\in\mathcal{P}(Y)$, let $\mathcal{T}(\mu,\nu)=\mathcal{T}_c(\mu,\nu)$ be the optimal transport cost associated to the cost c. Then*

$$|\mathcal{T}(\mu_1,\nu_1)-\mathcal{T}(\mu_2,\nu_2)| \leq 2\max_{x\in X,y\in Y}c(x,y)\cdot(\|\mu_1-\mu_2\|_{TV}+\|\nu_1-\nu_2\|_{TV}).$$

Proof. Let $M \coloneqq \max_{x\in X,y\in Y}c(x,y)\geq 0$. Notice that

$$|\mathcal{T}(\mu_1,\nu_1)-\mathcal{T}(\mu_2,\nu_2)| \leq |\mathcal{T}(\mu_1,\nu_1)-\mathcal{T}(\mu_2,\nu_1)| + |\mathcal{T}(\mu_2,\nu_1)-\mathcal{T}(\mu_2,\nu_2)|,$$

hence it suffices to show that

$$\begin{aligned}
|\mathcal{T}(\mu_1,\nu_1)-\mathcal{T}(\mu_2,\nu_1)| &\leq 2M\cdot\|\mu_1-\mu_2\|_{\mathsf{TV}}, \\
|\mathcal{T}(\mu_2,\nu_1)-\mathcal{T}(\mu_2,\nu_2)| &\leq 2M\cdot\|\nu_1-\nu_2\|_{\mathsf{TV}}.
\end{aligned}$$

We prove only the first inequality, since the proof of the second one is similar. Let γ be an optimal coupling in the definition of $\mathcal{T}(\mu_1,\nu_1)$ and $\tilde{\gamma}$ be the coupling for (μ_2,ν_1) given by Lemma 3.5.13. Then it follows that

$$\begin{aligned}
\mathcal{T}(\mu_2,\nu_1) &\leq \sum_{x,y}c(x,y)\tilde{\gamma}(x,y) \\
&\leq \sum_{x,y}c(x,y)(\gamma(x,y)+|\tilde{\gamma}-\gamma|(x,y)) \\
&\leq \mathcal{T}(\mu_1,\nu_1) + 2\,M\|\tilde{\gamma}-\gamma\|_{\mathsf{TV}} \\
&\leq \mathcal{T}(\mu_1,\nu_1) + 2\,M\|\mu_1-\mu_2\|_{\mathsf{TV}}.
\end{aligned}$$

Similarly one shows that

$$\mathcal{T}(\mu_1,\nu_1) \le \mathcal{T}(\mu_2,\nu_1) + 2\,M\|\mu_1-\mu_2\|_{\mathsf{TV}},$$

which concludes the proof of the corollary. ■

Corollary 3.5.15. *For any probability measures $\mu,\nu \in \mathcal{P}(\Omega)$ and $t>0$ small enough we have*

$$\left|W_p^p(\mu P_t,\nu P_t) - W_p^p\Big(\mu\tilde{P}_t,\nu\tilde{P}_t\Big)\right| = O(t^2)$$

as $t \to 0$.

Proof. Fix $\mu,\nu \in \mathcal{P}(X)$ and $t>0$ small and let $D := \operatorname{diam}(\Omega)$. Since Ω is finite, we have that $D<\infty$ and $\|\mu\tilde{P}_t - \mu P_t\|_{\mathsf{TV}}, \|\nu\tilde{P}_t - \nu P_t\|_{\mathsf{TV}} = O(t^2)$. By applying Corollary 3.5.14 for $t>0$ small enough with $\mu_1 = \mu P_t, \mu_2 = \mu\tilde{P}_t, \nu_1 = \nu P_t, \nu_2 = \nu\tilde{P}_t$, we find

$$\left|W_p^p(\mu P_t,\nu P_t) - W_p^p\Big(\mu\tilde{P}_t,\nu\tilde{P}_t\Big)\right| \le 2D^p \cdot \Big(\|\mu\tilde{P}_t - \mu P_t\|_{\mathsf{TV}} + \|\nu\tilde{P}_t - \nu P_t\|_{\mathsf{TV}}\Big) = O(t^2),$$

as desired.

■

We can finally give an equivalent definition of $K_{cc,p}$ using P_t instead of $\tilde{P}_t$.

Corollary 3.5.16. *For $x \ne y \in \Omega$ we have*

$$K_{cc,p}(x,y) = -\frac{1}{d(x,y)}\frac{d}{dt}\bigg|_{t=0} W_p(\delta_x P_t,\delta_y P_t). \tag{3.5.7}$$

Proof. Recalling Definition 3.5.11, it suffices to show that

$$\frac{d}{dt}\bigg|_{t=0} W_p\Big(\delta_x\tilde{P}_t,\delta_y\tilde{P}_t\Big) = \frac{d}{dt}\bigg|_{t=0} W_p(\delta_x P_t,\delta_y P_t).$$

Equivalently, since $W_p(\delta_x P_0,\delta_y P_0) = W_p\Big(\delta_x\tilde{P}_0,\delta_y\tilde{P}_0\Big) = d(x,y) > 0$, it is enough to show

$$\frac{d}{dt}\bigg|_{t=0} W_p^p\Big(\delta_x\tilde{P}_t,\delta_y\tilde{P}_t\Big) = \frac{d}{dt}\bigg|_{t=0} W_p^p(\delta_x P_t,\delta_y P_t),$$

but this follows from Corollary 3.5.15. ■

Remark 3.5.17. *As anticipated, for $p=1$ the right hand side in* (3.5.7) *corresponds to the definition of curvature for continuous time Markov chain introduced by Veysseire (see [Vey12]). For $p=1$, the identity in the corollary was established in [MW19] (in a more general setting).*

As in discrete time, we define also the curvature for $p=\infty$.

Definition 3.5.18. *For $x \ne y$, we define the continuous time ∞-coarse Ricci curvature $K_{cc,\infty}(x,y)$ in direction (x,y) to be the supremum of all $K \in \mathbb{R}$ such that there exist coupling rates $C(x,y,\cdot,\cdot)$ satisfying:*

- $\sum_{v,z\in\Omega} C(x,y,v,z)\mathbb{1}_{d(v,z)>d(x,y)} = 0$*;*

- $\sum_{v,z\in\Omega} C(x,y,v,z)d(v,z) \leq \Big(\sum_{v,z} C(x,y,v,z) - K\Big)d(x,y)$.

Remark 3.5.19. *We have that $K_{cc,\infty}(x,y) \in \mathbb{R}_{\geq 0} \cup \{-\infty\}$; if $K_{cc,\infty}(x,y) \geq 0$, the supremum in the definition is attained. Notice that if $C(x,y,\cdot,\cdot)$ gives optimal coupling rates, then we can change the value $C(x,y,x,y)$ arbitrarily obtaining other optimal coupling rates.*

As in discrete time, for $p \in [1,\infty]$ we write

$$\mathrm{Ric_{cc,p}} \geq K$$

if $K_{cc,p}(x,y) \geq K$ for all $x \neq y \in \Omega$.

The next result shows that, given coupling rates from $x \neq y$, we can construct a coupling for the probability measures $\left(\delta_x \tilde{P}_t, \delta_t \tilde{P}_t\right)$ for any $t > 0$ small enough. This will be useful to connect the notions of coarse Ricci curvature for discrete and continuous time Markov chains, when we consider a natural correspondence between stochastic transition matrices P and generators L, see Section 3.5.3.

Lemma 3.5.20. *Let $x \neq y \in \Omega$ and $C(x,y,\cdot,\cdot)$ denote coupling rates from x, y, and set*

$$M := \sum_{v,w} C(x,y,v,w) = \sum_v Q(x,v) = \sum_w Q(y,w).$$

Then, for any $0 < t \leq \frac{1}{M - C(x,y,x,y)}$ we have that γ_t is a coupling for $\left(\delta_x \tilde{P}_t, \delta_t \tilde{P}_t\right)$, where we define

$$\gamma_t(v,w) = \begin{cases} C(x,y,v,w)\cdot t & \text{if } (v,w) \neq (x,y), \\ 1 - M\cdot t + C(x,y,x,y)\cdot t & \text{if } (v,w) = (x,y). \end{cases}$$

Proof. Clearly for $0 < t \leq \frac{1}{M - C(x,y,x,y)}$ and $v, w \in \Omega$ we have that $0 \leq \gamma_t(v,w) \leq 1$. Now, if $v \neq x$ we have that

$$\sum_w \gamma_t(v,w) = t \sum_w C(x,y,v,w) = tQ(x,v) = (\delta_x \tilde{P}_t)(v).$$

If $v = x$ instead then

$$\begin{aligned} \sum_w \gamma_t(x,w) &= t \sum_{w \neq y} C(x,y,x,w) + 1 - M\cdot t + C(x,y,x,y)\cdot t \\ &= 1 + Q(x,x)\cdot t - M\cdot t \\ &= (\delta_x \tilde{P}_t)(x). \end{aligned}$$

Hence, this shows that the first marginal of γ_t is $\delta_x \tilde{P}_t$. Similarly, one checks that the second marginal of γ_t is $\delta_y \tilde{P}_t$, as desired. ■

The next proposition collects some useful results, and it is a continuous time analogue of Proposition 3.5.5.

Proposition 3.5.21. *The following hold:*

1. *For $p \in [1,\infty]$, if $K_{cc,p}(x,y) \geq K$ for all $x \sim y$ then $\mathrm{Ric_{cc,p}} \geq K$.*

2. *For $1 \leq p \leq q < \infty$ we have $K_{cc,p}(x,y) \geq K_{cc,q}(x,y)$. Moreover if $x \sim y$ we have that*

$$K_{cc,p}(x,y) \geq \frac{K_{cc,\infty}(x,y)}{p}.$$

3. *For $p \in [1,\infty)$, if $\mathrm{Ric_{cc,p}} \geq K$ then for any starting probability measures $\mu, \nu \in \mathcal{P}(\Omega)$ and any $t \geq 0$ we have that*

$$W_p(\mu P_t, \nu Pt) \leq e^{-Kt} W_p(\mu,\nu).$$

Proof. 1. Suppose first that $p < \infty$ and let $n = d(x,y)$ with $x = z_0 \sim z_1 \sim \ldots \sim z_n = y$. Then for $t > 0$ small enough we have that

$$-W_p(\delta_x \tilde{P}_t, \delta_y \tilde{P}_t) \geq \sum_{i=0}^{n-1} -W_p\Big(\delta_{z_i}\tilde{P}_t, \delta_{z_{i+1}}\tilde{P}_t\Big)$$

by the triangle inequality. Therefore adding $n = d(x,y)$ and dividing by t

$$\frac{d(x,y) - W_p(\delta_x \tilde{P}_t, \delta_y \tilde{P}_t)}{t} \geq \sum_{i=0}^{n-1} \frac{1 - W_p\Big(\delta_{z_i}\tilde{P}_t, \delta_{z_{i+1}}\tilde{P}_t\Big)}{t}.$$

Letting $t \to 0$ and using the assumption gives the conclusion.

Suppose now that $p = \infty$: if $K = -\infty$ the conclusion is trivial, hence assume that $K \geq 0$. Let again $n = d(x,y)$ and $x = z_0 \sim z_1 \sim \ldots \sim z_n = y$: we prove the claim by induction over n. The base case $n = 1$ follows directly by the assumption. Now suppose $n > 1$ and that the inductive hypothesis holds. Let $C(x, z_{n-1}, \cdot, \cdot)$ and $C(z_{n-1}, y, \cdot, \cdot)$ be such that

$$\sum_{v,w\in\Omega} C(x, z_{n-1}, v, w)\mathbb{1}_{d(v,w) > d(x,z_{n-1})} = 0,$$

$$\sum_{v,w\in\Omega} C(x, z_{n-1}, v, w) d(v,w) \leq \left(\sum_{v,w\in\Omega} C(x, z_{n-1}, v, w) - K\right) d(x, z_{n-1}),$$

$$\sum_{v,w\in\Omega} C(z_{n-1}, y, v, w)\mathbb{1}_{d(v,w)>1} = 0,$$

$$\sum_{v,w\in\Omega} C(z_{n-1}, y, v, w) d(v,w) \leq \left(\sum_{v,w\in\Omega} C(z_{n-1}, y, v, w) - K\right).$$

Without loss of generality, by changing if needed the values of $C(x, z_{n-1}, x, z_{n-1})$, $C(z_{n-1}, y, z_{n-1}, y)$, $Q(x,x)$, $Q(z_{n-1}, z_{n-1})$, $Q(y,y)$, we can assume that

$$\begin{aligned}
&\sum_{s\in\Omega} Q(x,s) = \sum_{s\in\Omega} Q(z_{n-1}, s) = \sum_{s\in\Omega} Q(y,s) \\
&= \sum_{v,w\in\Omega} C(x, z_{n-1}, v, w) = \sum_{v,w\in\Omega} C(z_{n-1}, y, v, w) \\
&=: M > 0.
\end{aligned}$$

Therefore, we can apply the Gluing lemma (which easily extends to measures having the same total mass) to conclude that there exists $\hat{C}(\cdot,\cdot,\cdot) = \hat{C}(x, z_{n-1}, y, \cdot, \cdot, \cdot) \colon \Omega \times \Omega \times$

$\Omega \to \mathbb{R}_{\geq 0}$ such that $p_{1,2}\#\hat{C} = C(x, z_{n-1}, \cdot, \cdot)$ and $p_{2,3}\#\hat{C} = C(z_{n-1}, y, \cdot, \cdot)$, where $p_{i,j}$ is the projection on coordinates i, j and $\#$ is the pushforward, so that

$$\sum_{w\in\Omega} \hat{C}(x, z_{n-1}, y, v, s, w) = C(x, z_{n-1}, v, s),$$
$$\sum_{v\in\Omega} \hat{C}(x, z_{n-1}, y, v, s, w) = C(z_{n-1}, y, s, w).$$

Defining $C(x, y, \cdot, \cdot) := p_{1,3}\#\hat{C}$ gives coupling rates with the desired properties, analogously to the proof of Proposition 3.5.5. Indeed, we have that

$$\begin{aligned}
\sum_{v,w} C(x, y, v, w)d(v, w) &= \sum_{v,w,s} \hat{C}(x, z_{n-1}, y, v, s, w)d(v, w) \\
&\leq \sum_{v,w,s} \hat{C}(x, z_{n-1}, y, v, s, w)(d(v, s) + d(s, w)) \\
&= \sum_{v,s} C(x, z_{n-1}, v, s)d(v, s) + \sum_{s,w} C(z_{n-1}, y, s, w)d(s, w) \\
&\leq (M - K)(d(x, z_{n-1}) + d(z_{n-1}, y)) \\
&= \left(\sum_{v,w} C(x, y, v, w) - K\right)d(x, y),
\end{aligned}$$

and similarly

$$\begin{aligned}
&\sum_{v,w} C(x, y, v, w)\mathbb{1}_{d(v,w)>d(x,y)} \\
&= \sum_{v,w,s} \hat{C}(x, z_{n-1}, y, v, s, w)\mathbb{1}_{d(v,w)>d(x,y)} \\
&\leq \sum_{v,w,s} \hat{C}(x, z_{n-1}, y, v, s, w)\mathbb{1}_{d(v,s)+d(s,w)>d(x,z_{n-1})+d(z_{n-1},y)} \\
&\leq \sum_{v,w,s} \hat{C}(x, z_{n-1}, y, v, s, w)\Big(\mathbb{1}_{d(v,s)>d(x,z_{n-1})} + \mathbb{1}_{d(s,w)>d(z_{n-1},y)}\Big) \\
&= \sum_{v,s} C(x, z_{n-1}, v, s)\mathbb{1}_{d(v,s)>d(x,z_{n-1})} + \sum_{s,w} C(z_{n-1}, y, s, w)\mathbb{1}_{d(s,w)>d(z_{n-1},y)} \\
&= 0.
\end{aligned}$$

2. The first statement follows by the inequality $W_p(\mu, \nu) \leq W_q(\mu, \nu)$: indeed it implies

$$\frac{d(x, y) - W_p(\delta_x\tilde{P}_t, \delta_y\tilde{P}_t)}{t\, d(x, y)} \geq \frac{d(x, y) - W_q(\delta_x\tilde{P}_t, \delta_y\tilde{P}_t)}{t\, d(x, y)},$$

from which the conclusion follows by letting $t \to 0$. For the second statement, suppose $K_{cc,\infty}(x, y) \geq 0$ for $x \sim y$. Let $C(x, y, \cdot, \cdot)$ be optimal coupling rates in the definition of $K_{cc,\infty}(x, y)$. Then for $t > 0$ small enough consider the coupling γ_t given by Lemma 3.5.20. It is easy to see that this coupling is such that

$$\sum_{v,w} \gamma_t(v, w)\mathbb{1}_{d(v,w)>1} = 0;$$
$$\sum_{v,w} \gamma_t(v, w)\mathbb{1}_{d(v,w)=0} \geq K_{cc,\infty}(x, y) \cdot t.$$

Therefore, it shows that

$$W_p\Big(\delta_x\tilde{P}_t, \delta_y\tilde{P}_t\Big) \leq (1 - K_{cc,\infty}(x, y) \cdot t)^{\frac{1}{p}} \leq 1 - \frac{K_{cc,\infty}(x, y) \cdot t}{p}.$$

Hence we have

$$\frac{1 - W_p(\delta_x \tilde{P}_t, \delta_y \tilde{P}_t)}{t} \geq \frac{K_{cc,\infty}(x,y)}{p},$$

from which the conclusion follows by letting $t \to 0$.

3. Let $\gamma_{\mu,\nu}$ and $\gamma_{x,y,t}$ be optimal couplings in the definitions of $W_p(\mu,\nu)$ and $W_p(\delta_x P_t, \delta_y P_t)$ respectively. Then notice that

$$\begin{aligned} W_p^p(\mu P_t, \nu P_t) &\leq \sum_{x,y} d(x,y)^p \sum_{w,z} \gamma_{\mu,\nu}(w,z)\gamma_{w,z,t}(x,y) \\ &= \sum_{w,z} \gamma_{\mu,\nu}(w,z) W_p^p(\delta_w P_t, \delta_z P_t), \end{aligned}$$

and so

$$\frac{W_p^p(\mu P_t, \nu P_t) - W_p^p(\mu,\nu)}{t} \leq \sum_{w,z} \gamma_{\mu,\nu}(w,z) \frac{W_p^p(\delta_w P_t, \delta_z P_t) - d(w,z)^p}{t}.$$

Taking the $\limsup_{t\to 0^+}$ and denoting by $\frac{d^+}{dt}$ the upper Dini derivative (cf. Section 3.8) we find that

$$\begin{aligned} &\left.\frac{d^+}{dt}\right|_{t=0} W_p^p(\mu P_t, \nu P_t) \\ \leq &\sum_{w,z} \gamma_{\mu,\nu}(w,z) \left.\frac{d}{dt}\right|_{t=0} W_p^p(\delta_w P_t, \delta_z P_t) \\ = &\sum_{w,z} \gamma_{\mu,\nu}(w,z)\, p\, d(w,z)^{p-1} \left.\frac{d}{dt}\right|_{t=0} W_p(\delta_w P_t, \delta_z P_t) \\ \leq &- K\, p \sum_{w,z} \gamma_{\mu,\nu}(w,z) d(w,z)^p \\ = &- K\, p\, W_p^p(\mu,\nu), \end{aligned}$$

where we also used Corollary 3.5.16. Therefore,

$$\left.\frac{d^+}{dt}\right|_{t=0} W_p^p(\mu P_t, \nu P_t) \leq -Kp W_p^p(\mu,\nu).$$

and by Markovianity this extends to every $\bar{t} > 0$, i.e.

$$\left.\frac{d^+}{dt}\right|_{t=\bar{t}} W_p^p(\mu P_t, \nu P_t) \leq -Kp W_p^p(\mu P_{\bar{t}}, \nu P_{\bar{t}}).$$

Noticing also that $t \to W_p^p(\mu P_t, \nu P_t)$ is continuous by Corollary 3.5.14, we can apply Lemma 3.8.1 to conclude that

$$W_p^p(\mu P_t, \nu P_t) \leq e^{-Kpt} W_p^p(\mu,\nu),$$

for any $t \geq 0$, as desired.

■

3.5.3 Comparison discrete and continuous time

There is a natural way to construct a continuous time Markov chain from a discrete time one: namely, for $\lambda > 0$ and a stochastic matrix P, the generator is defined by $L = \lambda(P - I)$. On the other hand, it is readily seen that, given a generator L, for any $\lambda > 0$ big enough there exists a corresponding stochastic matrix $P = I + \frac{1}{\lambda}L$ (recall we are assuming finiteness of the state space, so the entries of L are bounded).

Remark 3.5.22. *Let P be a stochastic matrix and $\lambda > 0$. In view of Definition 3.5.11, we see that for $p = 1$ the $\mathrm{Ric}_{\mathrm{dc},1}$ and $\mathrm{Ric}_{\mathrm{cc},1}$ notions of curvature are essentially equivalent for the continuous time Markov chain with generator $L = \lambda(P - I)$ (and transition semigroup P_t) and the **lazy** discrete time Markov chain with transition matrix $\tilde{P} := \frac{I+P}{2}$. Indeed, it follows from the identity in* (3.5.5) *with $s = \frac{1}{2\lambda}$ that*

$$K^{(\tilde{P})}_{dc,1}(x, y) = \frac{1}{2\lambda} K_{cc,1}(x, y), \tag{3.5.8}$$

where above $K^{(\tilde{P})}_{dc,1}(x, y)$ is the curvature $K_{dc,1}(x, y)$ for the Markov chain with transition matrix $\tilde{P}$, not P. Notice that here it is fundamental to consider the lazy version with transition matrix $\tilde{P}$. To see why, consider the following simple example: the state space is the two-point space $\Omega = \{x, y\}$, the stochastic matrix is

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

and $\lambda = 1$. Then we have $K^{(P)}_{dc,1}(x, y) = 0$ for P, $K^{(\tilde{P})}_{dc,1}(x, y) = 1$ for $\tilde{P}$ and $K_{cc,1}(x, y) = 2$ for $L = P - I$. Hence (3.5.8) *is satisfied only when $K_{dc,1}(x, y)$ is defined for $\tilde{P}$ and not for P.*

The next proposition shows that an analogous relation as the one described in the above remark holds true also for $\mathrm{Ric}_{\mathrm{dc},\infty}$ and $\mathrm{Ric}_{\mathrm{cc},\infty}$.

Proposition 3.5.23. *Suppose $L = \lambda(P - I)$ for a stochastic matrix P and $\lambda > 0$. Then the following hold for any $x \neq y$:*

1. *$K_{cc,\infty}(x, y) \geq \lambda K^{(P)}_{dc,\infty}(x, y)$.*
2. *For the lazy Markov chain with transition matrix $\tilde{P} = \frac{P+I}{2}$, we have $K^{(\tilde{P})}_{dc,\infty}(x, y) \geq \frac{1}{2\lambda} K_{cc,\infty}(x, y)$.*

Proof. 1. Assume $K^{(P)}_{dc,\infty}(x, y) \geq 0$, otherwise the claim is trivial. Let $\Pi(x, y, \cdot, \cdot)$ be an optimal coupling in the definition of $K^{(P)}_{dc,\infty}(x, y)$ and define then the coupling rates $C(x, y, w, z) = \lambda \cdot \Pi(x, y, w, z)$. It is easy to check that these coupling rates are admissible and yield the first conclusion.

2. Again, assume $K_{cc,\infty}(x, y) \geq 0$, otherwise the conclusion is trivial. Let $C(x, y, \cdot, \cdot)$ be optimal coupling rates in the definition of $K_{cc,\infty}(x, y)$. Notice that for $L = \lambda(P - I)$ and $t = \frac{1}{2\lambda}$ we have

$$\tilde{P}_t = \tilde{P}_{\frac{1}{2\lambda}} = I + \frac{1}{2\lambda} \cdot \lambda(P - I) = \tilde{P}.$$

Therefore, since

$$\begin{aligned} M \coloneqq \sum_{v,w} C(x,y,v,w) &= C(x,y,x,y) + \sum_{v,w:(v,w)\neq(x,y)} C(x,y,v,w) \\ &\leq C(x,y,x,y) + \sum_{v,w:v\neq x} C(x,y,v,w) + \sum_{v,w:w\neq y} C(x,y,v,w) \\ &= C(x,y,x,y) + \sum_{v\neq x} Q(x,v) + \sum_{w\neq y} Q(y,w) \\ &\leq C(x,y,x,y) + 2\lambda, \end{aligned}$$

we can apply Lemma 3.5.20 to obtain a coupling $\Pi(x,y,\cdot,\cdot)$ for $\left(\delta_x\tilde{P},\delta_y\tilde{P}\right)$, i.e.

$$\Pi(x,y,v,w) = \begin{cases} C(x,y,v,w)\cdot\frac{1}{2\lambda} & \text{if } (v,w)\neq(x,y), \\ 1 - M\cdot\frac{1}{2\lambda} + C(x,y,x,y)\cdot\frac{1}{2\lambda} & \text{if } (v,w)=(x,y). \end{cases}$$

Clearly, by the assumptions on $C(x,y,\cdot,\cdot)$ we have that

$$\sum_{v,w\in\Omega} \Pi(x,y,v,w)\mathbb{1}_{d(v,w)>d(x,y)} = 0.$$

Moreover

$$\begin{aligned} \sum_{v,w\in\Omega} \Pi(x,y,v,w)d(v,w) =& \left(1-\frac{M}{2\lambda}\right)d(x,y) + \frac{1}{2\lambda}\sum_{v,w} C(x,y,v,w)d(v,w) \\ \leq& \left(1-\frac{M}{2\lambda}\right)d(x,y) + \frac{1}{2\lambda}(M - K_{cc,\infty}(x,y))d(x,y) \\ =& \left(1-\frac{K_{cc,\infty}(x,y)}{2\lambda}\right)d(x,y). \end{aligned}$$

This shows that $K^{(\tilde{P})}_{dc,\infty}(x,y) \geq \frac{1}{2\lambda}K_{cc,\infty}(x,y)$ for the lazy Markov chain $\tilde{P}$, as desired.

■

3.5.4 Applications and related problems

Recalling Definition 3.5.18 and Proposition 3.5.21, we see that for a continuous time Markov chain on a finite state space Ω we have that $\mathrm{Ric}_{\mathrm{cc},\infty} \geq K > 0$ if and only if for all *neighbouring* states $x\sim y\in\Omega$ there exist coupling rates $C(x,y,\cdot,\cdot)$ satisfying

- $\sum_{v,z\in\Omega} C(x,y,v,z)\mathbb{1}_{d(v,z)>1} = 0$;
- $\sum_{v,z\in\Omega} C(x,y,v,z)\mathbb{1}_{d(v,z)=0} \geq K$.

Bearing this mind, we can see that in the examples of Section 3.4 the constructed coupling rates immediately yield positive $\mathrm{Ric}_{\mathrm{cc},\infty}$ curvature: more precisely, we have the following result.

Theorem 3.5.24. *The following hold:*

- *Under the assumptions of Theorem 3.4.1, we have* $\mathrm{Ric}_{\mathrm{cc},\infty} \geq \kappa_*$ *for Glauber dynamics.*

- *Under the assumptions of Theorem 3.4.8, we have* $\mathrm{Ric}_{\mathrm{cc},\infty} \geq L$ *for the Bernoulli–Laplace model.*
- *Under the assumptions of Theorem 3.4.12, we have* $\mathrm{Ric}_{\mathrm{cc},\infty} \geq \kappa_*$ *for the hardcore model.*

In particular, by Proposition 3.5.21, under the respective theorems' assumptions for all probability measures $\mu, \nu \in \mathcal{P}(\Omega)$, $p \geq 1$ *and* $t \geq 0$ *we have*

$$W_p(\mu P_t, \nu P_t) \leq e^{-\frac{K}{p}t} W_p(\mu, \nu), \tag{3.5.9}$$

with $K = \begin{cases} \kappa_* & \text{for Glauber dynamics,} \\ L & \text{for the Bernoulli–Laplace model,} \\ \kappa_* & \text{for the hardcore model.} \end{cases}$

We remark here that the estimate (3.5.9) was already established in [Con22] only for the specific case of interacting random walks on the grid of Section 3.4.4 (see Theorem 3.2 of [Con22]), which doesn't follow directly from the arguments of this section since we restricted our discussion to finite state space Markov chains.
We also remark that Theorem 3.5.24 shows that, for some Markov chains, assumptions that are strictly connected to positive $\mathrm{Ric}_{\mathrm{cc},\infty}$ curvature are useful for establishing the modified log-Sobolev inequality, positive entropic curvature, positive discrete Bakry–Èmery curvature and other related inequalities (cf. the discussion in Section 3.2). This suggests interesting connections with some other open problems in the theory of functional inequalities and discrete curvature for Markov chains, as we discuss next.

Peres–Tetali Conjecture One notable example is the following important unpublished conjecture by Peres and Tetali, which links coarse Ricci curvature to the modified log-Sobolev inequality in the setting of lazy simple random walks on finite graphs (see also [ELL17, Con. 3.1], [Fat19, Con. 4], [BCC+22, Rmk. 1.1]).

Conjecture 3.5.25. *There exists a universal constant* $\alpha > 0$ *such that the following holds. Let* Ω *be a finite unweighted graph and consider the stochastic matrix* P *associated with the simple random walk on this graph,* $\tilde{P} = \frac{P+I}{2}$ *associated to the lazy simple random walk and the generator* $L = P - I$*. If* $\mathrm{Ric}_{\mathrm{dc},1} \geq K > 0$ *for the lazy stochastic matrix* $\tilde{P}$ *(or, equivalently, if* $\mathrm{Ric}_{\mathrm{cc},1} \geq 2K$ *for* L*), then* $\mathrm{MLSI}(\alpha K)$ *holds.*

In all the examples of Theorem 3.5.24 we have positive $\mathrm{Ric}_{\mathrm{cc},\infty}$ curvature, which implies in particular positive $\mathrm{Ric}_{\mathrm{cc},1}$ curvature by Proposition 3.5.21. Therefore, it is natural to study the following problem related to the above conjecture: assuming a strictly positive lower bound of $\mathrm{Ric}_{\mathrm{cc},\infty}$ (and under some additional assumptions), is it possible to deduce a lower bound of the same order for the MLSI constant? In particular, if the additional assumptions are that we are in the setting of simple random walks on finite graphs, this problem constitutes a weaker form of the Peres–Tetali Conjecture.

Coarse and entropic curvature Another important open problem consists in comparing the different notions of discrete curvature. For example, it is not known when a positive lower bound for the coarse Ricci curvature implies a positive lower bound of the same order for the entropic curvature, or vice versa (and similarly for the discrete Bakry–Émery curvature). In light of the results of this paper, the following is also a natural question: assuming a strictly positive lower bound of $\mathrm{Ric}_{\mathrm{cc},\infty}$ (and under some additional assumption), is it possible to deduce a lower bound of the same order for the entropic curvature? Interestingly, we remark that positive lower bounds for the entropic curvature are linked to exponential contraction with respect to the metric $\mathscr{W}$. Precisely, if $\mathrm{Ric}_{\mathrm{e}} \geq K$ then

$$\mathscr{W}(\mu P_t, \nu P_t) \leq e^{-Kt}\mathscr{W}(\mu, \nu)$$

for all $t \geq 0$ and starting probability measures μ, ν (see [EM12, Prop 4.7] and cf. Proposition 3.2.1). This is of course reminiscent of the exponential decay of (3.5.9), which follows from $\mathrm{Ric}_{\mathrm{cc},\infty} \geq K$. However, not much is known about the relationship between $\mathscr{W}$ and W_p (see [EM12, Prop. 2.12, 2.14] for a lower(/upper) bound of $\mathscr{W}$ in terms of $W_1(/W_2)$), so it is not clear how these exponential decay estimates are connected.

3.6 Proof of Proposition 3.2.2

Here, we work in the setting of Section 3.5 and with that notation; recall in particular that the state space is finite and the Markov chain is irreducible and reversible.

Proof. Suppose Ω is finite and Assumption 2 is satisfied and that (3.2.3) holds for all $\rho\colon \Omega \to \mathbb{R}_{>0}$ and $\psi\colon \Omega \to \mathbb{R}$. Fix now $\rho\colon \Omega \to \mathbb{R}_{>0}$ and choose $\psi = \phi' \circ \rho$. Notice then that

$$\begin{aligned}&\mathscr{A}(\rho, \phi' \circ \rho)\\ =&\frac{1}{2}\sum_{x,y:\psi(x)\neq\psi(y)} \pi(x)Q(x,y)\frac{\rho(x)-\rho(y)}{(\phi'\circ\rho)(x)-(\phi'\circ\rho)(y)}[(\phi'\circ\rho)(x)-(\phi'\circ\rho)(y)]^2\\ =&\mathcal{E}(\rho, \phi'\circ\rho).\end{aligned}$$

Moreover if $\psi(x) \neq \psi(y)$ we have that

$$\nabla\theta(\rho(x),\rho(y)) = \frac{1}{\psi(x)-\psi(y)}[1-\phi''(\rho(x))\,\theta(\rho(x),\rho(y)), -1+\phi''(\rho(y))\,\theta(\rho(x),\rho(y))].$$

It follows that

$$\begin{aligned}\mathscr{C}(\rho,\phi'\circ\rho) &= \frac{1}{4}\sum_{x,y}\pi(x)Q(x,y)[(\phi'\circ\rho)(x)-(\phi'\circ\rho)(y)]\\ &\quad\cdot\left\{[1-\phi''(\rho(x))\,\theta(\rho(x),\rho(y)), -1+\phi''(\rho(y))\,\theta(\rho(x),\rho(y))]\cdot\begin{pmatrix}L\rho(x)\\ L\rho(y)\end{pmatrix}\right\}\\ &= \frac{1}{2}\mathcal{E}(L\rho,\phi'\circ\rho) - \frac{1}{2}\mathcal{E}(\rho,(\phi''\circ\rho)\cdot L\rho)\\ &= \frac{1}{2}\mathcal{E}(\rho, L(\phi'\circ\rho)) - \frac{1}{2}\mathcal{E}(\rho,(\phi''\circ\rho)\cdot L\rho),\\ \mathscr{D}(\rho,\phi'\circ\rho) &= \frac{1}{2}\sum_{x,y}\pi(x)Q(x,y)(\rho(x)-\rho(y))(L(\phi'\circ\rho)(x)-L(\phi'\circ\rho)(y))\\ &= \mathcal{E}(\rho, L(\phi'\circ\rho)).\end{aligned}$$

Therefore

$$\mathscr{B}(\rho,(\phi'\circ\rho)) = -\frac{1}{2}\mathcal{E}[\rho,(\phi''\circ\rho)\cdot L\rho + L(\phi'\circ\rho)].$$

From this, letting $\rho_t = P_t\rho$, we see that

$$\frac{d}{dt}\bigg|_{t=0} \mathcal{E}(\rho_t, \phi' \circ \rho_t) = -2\mathscr{B}(\rho, \phi' \circ \rho).$$

Therefore the inequality $\mathscr{B}(\rho, \phi' \circ \rho) \geq K\mathscr{A}(\rho, \phi' \circ \rho)$ is equivalent to inequality (3.2.6), from which $\mathrm{CSI}_\phi(2K)$ follows. ■

Remark 3.6.1. *In the particular case when θ is the logarithmic mean, for any constant $Z > 0$ we have $\mathscr{B}(Z \cdot \rho, \psi) = Z\mathscr{B}(\rho, \psi)$ and $\mathscr{A}(Z \cdot \rho, \psi) = Z\mathscr{A}(\rho, \psi)$, so it suffices to consider the case where ρ is a density with respect to π when proving inequality* (3.2.3).

3.7 Proof of Proposition 3.3.2

In this section we prove Proposition 3.3.2, which gives the value of M_θ from (3.3.10) for some of the weight functions considered in Section 3.2.

Proof. Recall the definition

$$M_\theta \coloneqq \inf_{\substack{s,t\geq 0:\\ \theta(s,t)>0}} \frac{\theta(s,s)+\theta(t,t)}{2\theta(s,t)} \in [0,1].$$

The statement about the arithmetic mean is trivial, while the one about the logarithmic mean θ_1 follows from equation (2.1) in [EM12]. Let us therefore consider the case of the weight function θ_α, which was defined in (3.2.5) with $\phi = \phi_\alpha$ as in (3.1.4), corresponding to Beckner functionals. In other words, we have

$$\theta(s,t) = \frac{\alpha-1}{\alpha}\frac{s-t}{s^{\alpha-1}-t^{\alpha-1}}$$

for $s \neq t > 0$ and $\theta(s,s) = \frac{1}{\alpha}s^{2-\alpha}$. Again, for $\alpha = 2$ the result is trivial, so we assume henceforth $1 < \alpha < 2$. Without loss of generality we can minimize over $s > t > 0$, so that substituting the expression of θ_α from (3.2.8) the problem reduces to computing

$$\begin{aligned} M_{\theta_\alpha} &= \inf_{s>t>0} \frac{1}{2(\alpha-1)} \frac{(s^{2-\alpha}+t^{2-\alpha})\cdot(s^{\alpha-1}-t^{\alpha-1})}{s-t} \\ &= \frac{1}{2(\alpha-1)}\left\{1 + \inf_{s>t>0}\frac{s^{\alpha-1}t^{2-\alpha}-s^{2-\alpha}t^{\alpha-1}}{s-t}\right\} \\ &= \frac{1}{2(\alpha-1)}\left\{1+\inf_{\lambda>1}\frac{\lambda^{\alpha-1}-\lambda^{2-\alpha}}{\lambda-1}\right\} \end{aligned} \tag{3.7.1}$$

where we set $\lambda \coloneqq \frac{s}{t} > 1$. If $\alpha \geq \frac{3}{2}$, the conclusion follows by noticing that $\frac{\lambda^{\alpha-1}-\lambda^{2-\alpha}}{\lambda-1} \geq 0$ for $\lambda > 1$ and by letting $\lambda \to \infty$. Suppose hence now that $\alpha \in \left(1, \frac{3}{2}\right)$: to conclude that in this case $M_{\theta_\alpha} = 1$ it is enough to show that for $\lambda > 1$

$$\frac{\lambda^{\alpha-1}-\lambda^{2-\alpha}}{\lambda-1} \geq 2\alpha-3. \tag{3.7.2}$$

Notice that equality holds as $\lambda \to 1^+$. By density of $\mathbb{Q}$ in $\mathbb{R}$ and by a continuity argument, it suffices to show that for all $\lambda > 1$ and all *even* integers $p, s \in \mathbb{N}$ with $p > s$ we have that

$$\frac{\lambda^{\frac{s}{p+s}}-\lambda^{\frac{p}{p+s}}}{\lambda-1} \geq -\frac{p-s}{p+s},$$

where we used the substitution $\alpha = 1 + \frac{s}{p+s}$. Rearranging this and renaming $\lambda \leftarrow \lambda^{\frac{1}{p+s}}$ we need to prove equivalently that for all $\lambda > 1$

$$\frac{\lambda^{p+s} - 1}{p+s} \geq \frac{\lambda^p - \lambda^s}{p-s}.$$

Equivalently, dividing both sides by $\lambda - 1$ and denoting by AM the arithmetic mean, we need to prove that

$$\mathrm{AM}\Big(1, \dots, \lambda^{p+s-1}\Big) \geq \mathrm{AM}\Big(\lambda^s, \dots, \lambda^{p-1}\Big).$$

This last inequality holds true, since for $\lambda > 1$ and integers $0 < i < j$ we have that

$$\lambda^{i-1} + \lambda^{j+1} \geq \lambda^i + \lambda^j$$

by the classical rearrangement inequality. ∎

3.8 Grönwall's lemma with Dini derivative

For a continuous function $f\colon I \to \mathbb{R}$, where $I \subset \mathbb{R}$ is an interval, we consider the *upper/lower Dini derivative*, which are defined respectively by

$$\begin{aligned}\frac{d^+}{dt}\bigg|_{t_0} f(t) &= \limsup_{h \to 0^+} \frac{f(t+h) - f(t)}{h},\\ \frac{d^-}{dt}\bigg|_{t_0} f(t) &= \liminf_{h \to 0^+} \frac{f(t+h) - f(t)}{h}.\end{aligned}$$

Clearly, $\frac{d^-}{dt} f(t) \leq \frac{d^+}{dt} f(t)$. It may be useful to apply Grönwall's lemma in absence of differentiability, by considering instead the Dini derivatives. In particular, the following variant holds.

Lemma 3.8.1 (Grönwall's lemma)**.** *Let $I = [a, b) \subset \mathbb{R}$ be an interval,with $a < b \leq \infty$, and consider real valued continuous functions $u, \beta\colon I \to \mathbb{R}$. Suppose that for all $t \in I$*

$$\frac{d^-}{dt} u(t) \leq \beta(t) u(t). \tag{3.8.1}$$

Then for all $t \in I$

$$u(t) \leq \exp\Big(\int_0^t \beta(s) ds\Big) u(a).$$

Proof. Set $v(t) = \exp\Big(-\int_0^t \beta(s) ds\Big)$, which satisfies $v(t) > 0$, $v'(t) = -\beta(t)v(t)$. Notice that for all $t \in I$

$$\begin{aligned}\frac{d^-}{dt}[u(t)v(t)] =& \liminf_{h \to 0^+} \frac{[u(t+h) - u(t)]v(t+h)}{h} + \frac{u(t)[v(t+h) - v(t)]}{h}\\ =& \left[\frac{d^-}{dt} u(t)\right] v(t) + u(t)v'(t)\\ \leq& \beta(t)u(t)v(t) - u(t)\beta(t)v(t)\\ =& 0.\end{aligned}$$

Next, we notice as in [DS08, Eqn. (3.8)] that this implies that the continuous function $u(t)v(t)$ is non-increasing on I. In particular, we have $u(t)v(t) \leq u(a)$, which implies the thesis by substituting the expression for $v(t)$ and rearranging. ∎

CHAPTER 4

Improved Convergence of Score-Based Diffusion Models via Prediction-Correction

This chapter corresponds to the publication [PMM24].

Score-based generative models (SGMs) are powerful tools to sample from complex data distributions. Their underlying idea is to *(i)* run a forward process for time T_1 by adding noise to the data, *(ii)* estimate its score function, and *(iii)* use such estimate to run a reverse process. As the reverse process is initialized with the stationary distribution of the forward one, the existing analysis paradigm requires $T_1 \to \infty$. This is however problematic: from a theoretical viewpoint, for a given precision of the score approximation, the convergence guarantee fails as T_1 diverges; from a practical viewpoint, a large T_1 increases computational costs and leads to error propagation. This paper addresses the issue by considering a version of the popular *predictor-corrector* scheme: after running the forward process, we first estimate the final distribution via an inexact Langevin dynamics and then revert the process. Our key technical contribution is to provide convergence guarantees which require to run the forward process *only for a fixed finite time* T_1. Our bounds exhibit a mild logarithmic dependence on the input dimension and the subgaussian norm of the target distribution, have minimal assumptions on the data, and require only to control the L^2 loss on the score approximation, which is the quantity minimized in practice.

4.1 Introduction

Score matching models [SE19, SE20] and diffusion probabilistic models [SDWMG15, HJA20] – recently unified into the single framework of score-based generative models (SGMs) [SGSE20] – have shown remarkable performance in sampling from unknown complex data distributions, achieving the state of the art in image [SGSE20, DN21] and audio [PVG+21, KPH+21, CZZ+21] generation; see also the recent surveys [YZS+23, CHIS23]. The idea is to gradually perturb the data by adding noise, and then to learn to revert the process. Both the forward process that adds noise and the reverse process can be described by a stochastic differential equation and, specifically, the reverse process is defined in terms of the *score function* (*i.e.*, the gradient of the logarithm of the perturbed density at all noise scales, see Section 4.3 for details). This time-dependent score function can be learned with a neural network, using efficient techniques such as sliced score matching [SGSE20] or denoising score matching [Vin11]. Then, to start the reverse process, one would ideally need to sample from the perturbed distribution, which is in principle unknown: instead, one runs the forward process for a long enough time T_1 so that the perturbed distribution p_{T_1} is well approximated by the stationary distribution π, which is known and can be readily sampled from.

A central theoretical question is to understand the quality of the sampling, *i.e.*, measure a distance between the output distribution of the reverse process and the true one. Three sources of error are given by *(i)* starting the reverse process from the stationary distribution π, rather than from the perturbed distribution p_{T_1}, *(ii)* approximating the score function (*e.g.*, with a neural network), and *(iii)* discretizing the reverse stochastic differential equation. The quantitative characterization of such errors has been carried out in a number of recent papers, see [SDME21, KFL22, LLT22, CCL+23b, CLL22, Bor22] and Section 4.2. However, to achieve convergence of the output distribution to the ground truth, this line of work requires to run the forward process for $T_1 \to \infty$. This is due to the first source of error mentioned above, *i.e.*, the approximation of p_{T_1} with π. At the same time, large values of T_1 amplify the other two sources of error and are also responsible for an increased computational cost in the training procedure, because of the need to approximate the score function on a large time interval $[0, T_1]$. Thus, there appears to be a subtle trade-off between the precision in the score approximation and the running time T_1: on the one hand, one needs to take $T_1 \to \infty$ so that p_{T_1} approaches π; on the other hand, for a given precision in the score, the convergence guarantees fail as $T_1 \to \infty$,[1] which highlights the instability of existing results. For these reasons, it is of great interest to characterize an appropriate time T_1 where to stop the forward process [YZS+23, Sec. 8].

Main contribution. In this work, we address the trade-off in the choice of the running time T_1 of the forward process by considering a variant of the predictor-corrector methods of [SSDK+21]. More precisely, after obtaining p_{T_1} via the forward process, we sample from p_{T_1} via an inexact Langevin dynamics that leverages the approximation of the score at time T_1. Then, we use the resulting sample to start a standard reverse process. Our main convergence result (Theorem 4.4.1) focuses on the deterministic reverse process, which has the form of an ordinary differential equation (ODE), and analyzes the proposed algorithm: in particular, it provides convergence guarantees in Wasserstein distance for a vast class of data distributions and under realistic assumptions on the score estimation, which are compatible with the training loss used in practice. We highlight that our bounds require a perturbation time T_1 that only depends logarithmically on the dimension of the space and on the subgaussian norm of the target distribution, and *not* on the desired sampling precision. The mild logarithmic dependence suffices to ensure the regularity – in the form of a log-Sobolev inequality – of the perturbed measure p_{T_1}, which in turn allows the inexact Langevin dynamics to converge exponentially fast.

Our analysis improves upon earlier bounds in Wasserstein distance [KFL22] by removing both the need for $T_1 \to \infty$ and an assumption on the score estimator (more precisely, on its one-sided Lipschitz constant, see the discussion after Theorem 4.4.1 for details). This comes at the cost of requiring a control on a loss function for the score estimate at time T_1, which is stronger than the usual L^2 loss. To address this issue, we exploit the fast convergence

[1] See, *e.g.*, [CCL+23b, Thm. 2] and note that the third term in the bound diverges when ε_{score} is fixed and $T \to \infty$.

of the forward process to its stationary distribution in order to correct the score estimator, which in turn allows us to translate upper bounds for the L^2 loss into upper bounds for the stronger loss (Theorem 4.4.2). Finally, when considering instead a stochastic reverse process (SDE), we show how the algorithm is compatible with the existing analyses for the convergence of the discretized reverse SDE in information-divergence metrics. This allows us to deduce convergence guarantees in total variation distance for a discretization of the algorithm. Consequently, we highlight some advantages over previous results that arise from the choice of a fixed perturbation time T_1, related again to the decreased computational cost in the training procedure and to the stability of the error bounds.

Paper organization. Section 4.2 discusses related works. Section 4.3 sets up the technical framework by recalling the formal description of SGMs. Section 4.4 presents our main contribution: after describing the algorithm, we state the convergence result in Wasserstein distance. Section 4.4.1 contains a sketch of the proofs, with the full arguments deferred to Section 4.8. Our main convergence results in Section 4.4 are stated in continuous time and in Wasserstein distance. Then, Section 4.5 contains a discussion about discretizations of the algorithm, as well as a convergence result in total variation distance. Section 4.6 concludes the paper with some final remarks.

Acknowledgement

Francesco Pedrotti and Jan Maas acknowledge support by the Austrian Science Fund (FWF) project 10.55776/F65. Marco Mondelli acknowledges support by the 2019 Lopez-Loreta prize.

4.2 Related work

The empirical success of SGMs has led to extensive research aimed at providing theoretical guarantees on their performance. Specifically, the goal is to give upper bounds on the distance between the true data distribution p and the output distribution p_θ of the sampling method. Adopting the description of forward and reverse process in terms of stochastic differential equations [SSDK+21], an upper bound on the KL-divergence $\mathcal{D}_{\mathrm{KL}}(p \,\|\, p_\theta)$ is provided in [SDME21]: under some regularity assumptions, this KL-divergence goes to 0, as $T_1 \to \infty$ and the score approximation error vanishes. In a similar vein, using the theory of optimal transport instead of stochastic tools, an upper bound on the Wasserstein distance[2] $W_2(p, p_\theta)$ is provided in [KFL22]. In many important situations, results in Wasserstein distance are more meaningful than in KL-divergence or total variation: for example, under the manifold hypothesis, it is not possible to obtain non-trivial convergence guarantees in those metrics, as one has $\mathcal{D}_{\mathrm{KL}}(p \,\|\, p_\theta) = \infty$ and $\|p - p_\theta\|_{\mathsf{TV}} = 1$ (cf. [Bor22, CCL+23b]). However, to obtain convergence, [KFL22] imposes a strong assumption on the score estimator. A line of work has focused on the discretization of the reverse stochastic differential equation. Specifically, convergence in Wasserstein distance of order 1 is provided in [Bor22] under the manifold hypothesis, but the results depend poorly on important parameters, such as the sampling precision, the input dimension and the diameter of the support of p. The works

[2] For general data distributions μ, ν, there is no relation between $\mathcal{D}_{\mathrm{KL}}(\nu \,\|\, \mu)$ and $W_2(\mu, \nu)$, therefore the results in KL-divergence cannot be translated trivially to results in W_2, and vice versa. In particular, the analysis in W_2 seems to be more challenging, because of an expansive term in the reverse process (cf. [CCL+23b, Sec. 4]).

[LLT23, CCL+23b, CLL22] provide convergence guarantees for general distributions in KL-divergence and total variation (which is weaker by Pinsker's inequality). From these results, bounds in Wasserstein distance are also deduced for some classes of distributions, including bounded support ones. This improves upon [Bor22], but the improvement requires an extra projection step and comes at the cost of a worse dependence on the problem's parameters in comparison with the bounds in KL-divergence and total variation. Other works have provided convergence results for SGMs, but they suffer from at least one of the following drawbacks [CCL+23b]: *(i)* non-quantitative bounds [Pid22, BTHD21] or poor dependence on important problem parameters [BMR22], *(ii)* strong assumptions on the data distribution, typically in the form of a functional inequality [LLT22, YYW23], and *(iii)* strong assumptions on the score estimation error, such as a uniform pointwise control [BTHD21], which is not observed in practice [ZC23].

In most of these works, to guarantee convergence of p_θ to p, it is necessary to have $T_1 \to \infty$. In contrast, [BTHD21] introduces a different approach based on solving the Schrödinger bridge problem (see also [CLT22, SBDD22, Son22]), which allows for a finite time perturbation of the target measure; however, this strategy does not come with quantitative convergence results under realistic assumptions. The work [FRY+23] adopts an approach closer to ours: an auxiliary model is used to bridge the gap between the limiting distribution of the forward process and the true perturbed distribution, which is then followed by a standard reverse process. However, the design of the auxiliary model appears to be *ad hoc* for the data distribution, and a theoretical convergence result for general data distributions is missing. Our algorithm can also be considered as an instance of the predictor-corrector approach of [SSDK+21]: there, the authors suggest to alternate one step of the reverse process with a few steps of a corrector method based on the score function, such as Langevin dynamics, and provide extensive empirical evidence showing an improved performance. In this work, we consider instead the case where all the corrector steps (Langevin dynamics) are performed at the beginning, and they are then followed by a standard (non-corrected) reverse process. We remark that error bounds for (a different variant of) the predictor-corrector schemes were first provided in [LLT22]. The results therein, however, impose strong assumptions on the data distribution, in the form of a log-Sobolev inequality; the log-Sobolev constant enters crucially in the derived bounds, which depend polynomially on it, and not just logarithmically.

Concurrent work. The concurrent work [CCL+23a] also studies the performance of a predictor corrector scheme. Instead of our two-stage algorithm, [CCL+23a] considers an implementation closer to [SSDK+21], that alternates deterministic predictor steps with corrector steps based on Langevin dynamics. For this method, the authors provide convergence guarantees in total variation distance: interestingly, they show that when the corrector steps are based on the *underdamped Langevin dynamics* (rather then the classical overdamped version), it is possible to obtain a better dependence on the dimension d. Contrary to our results, however, the error bounds in [CCL+23a] still require $T_1 \to \infty$ to obtain convergence, with analogous disadvantages as we discussed for pre-existing results.

After the first version of our paper appeared online, further progress on the theoretical study of SGMs has been made. The work [LWCC23] obtains convergence guarantees for both the standard reverse SDE and the deterministic reverse ODE (without corrector), using an approach based on studying directly discrete time methods rather than controlling the errors in approximating the continuous time dynamics. The paper [BDBDD23] improves the dimension dependence of convergence guarantees for the reverse SDE, when one does not assume Lipschitzness of $\nabla \log p_t$, by exploiting a connection with stochastic localization

[Eld13, Mon23]. The work [CDS23] proves convergence in KL-divergence without early stopping when replacing the classical Lipschitzness assumption on $\nabla \log p_t$ with finiteness of the relative Fisher information.

In general, these works can be considered complementary to the present contribution, in that their analysis techniques can be combined with our methods to provide convergence results for the two-stage algorithm under a set of different assumptions, gaining the advantage of a fixed perturbation time T_1. We refer the reader to Section 4.5 for an example of this.

4.3 Preliminaries

To sample from an unknown data distribution p on $\mathbb{R}^d$, the framework of SGMs consists in perturbing p through a forward process that converges to a known prior distribution and then approximately reverse this process. The forward process is described by a stochastic differential equation (SDE)

$$\begin{cases} X_0 \sim p, \\ dX_t = f(t, X_t)\, dt + g(t)\, dB_t, \end{cases} \tag{4.3.1}$$

where B_t is a standard Brownian motion, f and g are sufficiently smooth, and the SDE is run until some time $T_1 > 0$. The law of X_t, denoted by p_t, correspondingly solves the Fokker–Planck equation

$$\begin{cases} p_0 = p, \\ \partial_t p_t + \nabla \cdot \Big[p_t\Big(f(t,x) - \frac{g(t)^2}{2}\nabla \log p_t(x)\Big)\Big] = 0. \end{cases} \tag{4.3.2}$$

Remarkably, under some regularity conditions, this SDE admits a reverse process, in the sense that, for any smooth function $M\colon [0, T_1] \to \mathbb{R}_{\geq 1}$, the process $(U_t)_t$ defined by

$$\begin{cases} U_0 \sim p_{T_1}, \\ dU_t = -f(T_1 - t, U_t)\, dt + \frac{M(t)}{2} g(T_1 - t)^2 \nabla \log p_{T_1 - t}(U_t)\, dt + \sqrt{M(t) - 1}\, g(T_1 - t)\, dB_t, \end{cases} \tag{4.3.3}$$

is such that $U_{T_1} \sim p$. Usual choices are $M(t) \equiv 2$ or $M(t) \equiv 1$ and, considering the latter, the reverse process is deterministic, except for its initialization, see [SSDK+21, Sec. 4.3]. Below, we will first focus on $M \equiv 1$ for simplicity, but similar results can be readily deduced for general $M(t)$.

To simulate the reverse SDE (4.3.3) and sample from p, one needs to *(i)* approximately sample from p_{T_1} to initialize the backward process, and *(ii)* estimate the score function with $s_\theta(t, \cdot) \approx \nabla \log p_t$ for $t \in [0, T_1]$. For the first point, one chooses T_1 big enough so that p_{T_1} is close to a known distribution π and samples from $Y_0 \sim \pi$. For the second point, one learns a function $s_\theta(t, x)$ that approximates $\nabla \log p_t(x)$ *e.g.* with a neural network: specifically, the training loss considered in practice is

$$J_{\mathsf{SM}}(\theta, \lambda) = \int_0^{T_1} \lambda(t) \mathbb{E}_{p_t}\Big[\|\nabla \log p_t - s_\theta(t, \cdot)\|^2\Big]\, dt, \tag{4.3.4}$$

for some strictly positive weight function $t \to \lambda(t)$. Notably, although the score $\nabla \log p_t$ is unknown, this loss can be estimated with standard score-matching techniques [Vin11, SGSE20, SSDK+21]. When $\lambda(t) = g(t)^2$, which corresponds to the likelihood weighting of [SDME21], we simply write $J_{\mathsf{SM}}(\theta)$. The corresponding sampling algorithm simulates the process

$$\begin{cases} Y_0 \sim \pi, \\ dY_t = -f(T_1 - t, Y_t)\,dt + \frac{M(t)}{2} g(T_1 - t)^2 s_\theta(T_1 - t, Y_t)\,dt + \sqrt{M(t) - 1} g(T_1 - t)\,dB_t, \end{cases} \tag{4.3.5}$$

until time T_1, and it takes Y_{T_1} as an approximate sample of p. This reverse SDE can be approximated via standard general-purpose numerical solvers, or by taking advantage of the additional knowledge of the approximated score function $s_\theta(t,\cdot) \approx \nabla \log p_t$: for example, the *predictor-corrector* methods of [SSDK+21] alternate one discretized step for the reverse process (4.3.5) with several steps of a score-based corrector algorithm, such as Langevin dynamics or Hamiltonian Monte Carlo.

For ease of exposition, we will focus on the popular *Ornstein–Uhlenbeck* (OU) forward process

$$\begin{cases} X_0 \sim p, \\ dX_t = -X_t\,dt + \sqrt{2}\,dB_t, \end{cases} \tag{4.3.6}$$

which corresponds to a method known as *Denoising Diffusion Probabilistic Models (DDPMs)* [HJA20], and is also referred to as Variance Preserving SDE in [SSDK+21]. Its Fokker–Planck equation reads

$$\begin{cases} p_0 = p, \\ \partial_t p_t + \nabla \cdot [p_t(-x - \nabla \log p_t(x))] = 0. \end{cases} \tag{4.3.7}$$

The standard Gaussian γ is the limiting distribution of the OU process: more precisely, if $Z \sim \gamma$ is independent of X_0, then $X_t \sim e^{-t}X_0 + \sqrt{1 - e^{-2t}}Z$, and p_t converges to γ *e.g.* in Wasserstein distance W_2 and in relative entropy $\mathcal{D}_{\mathrm{KL}}(\cdot \,\|\, \gamma)$ [Vil03, Chap. 9]. Restricting to the OU process (or its time reparametrization) is commonly done in the theoretical literature, see [LLT22, Bor22, CCL+23b, CLL22]; as for these works, our techniques can be extended to other choices of the forward process, such as those considered in [SGSE20, Sec. 3.4].

Notation. We denote by $\gamma_{y,t}$ the density of a normal random variable in $\mathbb{R}^d$ with mean y and variance tI_d, and for compactness we write $\gamma_t = \gamma_{0,t}$ and $\gamma = \gamma_1$. With abuse of notation, we identify the law of a random variable with the corresponding probability density. Given two probability measures μ, ν, the KL-divergence is defined by $\mathcal{D}_{\mathrm{KL}}(\mu \,\|\, \nu) = \int \log\left(\frac{d\mu}{d\nu}\right) d\mu$ if μ is absolutely continuous with respect to ν, and $\mathcal{D}_{\mathrm{KL}}(\mu \,\|\, \nu) = +\infty$ otherwise; if μ, ν have finite second moment, the 2-Wasserstein distance is defined by $W_2^2(\mu,\nu) = \inf_{X \sim \mu, Y \sim \nu} \mathbb{E}[\|X - Y\|^2]$. We denote by $\mathcal{P}(\mathbb{R}^d)$ the space of probability measures on $\mathbb{R}^d$. Throughout this paper, we denote by p the target probability measure and by p_t its law following the OU process; correspondingly, we consider random variables $X \sim p$, $X_t \sim p_t$. We use the symbol $\lesssim$ to denote an inequality up to an absolute positive multiplicative constant.

4.4 Improved Wasserstein-convergence in continuous time via prediction-correction

Description of the algorithm. We consider the following predictor-corrector algorithm. First, we run the OU forward process (4.3.6) until time T_1 and, for $0 \le t \le T_1$, we approximate

$\nabla \log p_t(x)$ with $s_\theta(t,x)$. Next, we approximate p_{T_1} by following an *inexact* Langevin dynamics started at γ until time T_2:

$$\begin{cases} Z_0 \sim \gamma, \\ dZ_t = s_\theta(T_1, Z_t)\,dt + \sqrt{2}\,dB_t, \qquad 0 \le t \le T_2. \end{cases} \tag{4.4.1}$$

We remark that the Langevin dynamics (4.4.1) is *inexact* as it uses $s_\theta(T_1,x)$ in place of $\nabla \log p_{T_1}(x)$, since we have access to the former but not to the latter. The idea is that, as these two quantities are close, the random variable Z_{T_2} provides an approximate sample of p_{T_1}, for sufficiently large T_2. Then, we approximate the original distribution p by following a deterministic reverse process that starts from Z_{T_2}:

$$\begin{cases} Y_0 = Z_{T_2}, \\ dY_t = Y_t\,dt + s_\theta(T_1 - t, Y_t)\,dt, \qquad 0 \le t \le T_1. \end{cases} \tag{4.4.2}$$

We let $q_t = \mathrm{law}(Y_t)$ for $t \in [0,T_1]$ and $\sigma_t = \mathrm{law}(Z_t)$ for $t \in [0,T_2]$. In particular, we have $q_0 = \sigma_{T_2}$. Here, the *prediction-correction* consists in starting (4.4.2) from Z_{T_2}, instead of γ. We also note that this reverse process is deterministic except for its initialization $Y_0 = Z_{T_2}$. This allows to use standard numerical methods for solving ordinary differential equations, and in particularly exponential integrator schemes have shown remarkable performances [LZB+22]. Finally, we take Y_{T_1} to be an approximate sample from p. For stability reasons, we can also choose a small time $0 < \tau \ll T_1$ and stop the reverse process (4.4.2) at time $T_1 - \tau$, taking $Y_{T_1-\tau}$ as an approximate sample from p. This is commonly done in practice, see *e.g.* [SSDK+21, Sec. C].

Assumptions. Throughout this section, we consider the following assumptions.

(A1) The estimator $s_\theta \colon [0,T_1] \times \mathbb{R}^d \to \mathbb{R}^d$ is Lipschitz continuous. Moreover, for $t \in [0,T_1]$ we denote by $L_s(t) \in \mathbb{R}$ the *one-sided Lipschitz constant* for $s_\theta(t,\cdot)$, such that for all $x,y \in \mathbb{R}^d$,
$$(s_\theta(t,x) - s_\theta(t,y)) \cdot (x-y) \le L_s(t)\|x-y\|^2.$$

(A2) $X \sim p$ is norm-subgaussian.

Condition (A1) is mild: in fact, s_θ is typically given by a neural network, which corresponds to a Lipschitz function for most practical activations. We emphasize that the requirement on the Lipschitz constant of s_θ is purely qualitative, in the sense that it does not enter our bounds (as opposed to the one-sided Lipschitz constant, which instead plays a quantitative role in the bounds).

As for condition (A2), we recall that an $\mathbb{R}^d$-valued random variable X is norm-subgaussian if its euclidean norm $\|X\|$ is subgaussian (for details and a formal definition, see Section 4.7). We denote by $\|X\|_{\mathrm{SG}}$ the corresponding norm. Bounded random variables and subgaussian ones (in the sense of [Ver18, Def. 3.4.1]) are norm-subgaussian, which covers most practical cases (e.g., in image generation pixels are usually rescaled in $[0,1]$). Other properties of norm-subgaussian random vectors are established in [JNG+19].

Performance of the proposed algorithm. Our main result is stated below.

Theorem 4.4.1. *Let Assumptions (A1)-(A2) hold, and let p_t be obtained via the forward OU process in* (4.3.6). *Pick $0 < \delta < 1$, $T_2 > 0$, $T_1 \geq \frac{1}{2}\log\Big(2 + 172\frac{\|X\|_{\mathrm{SG}}^2}{\delta} + \frac{d}{2\delta}\Big)$ and a small early stopping time $0 < \tau \leq \min\Big\{T_1, \frac{d}{2\|X\|_{\mathrm{SG}}^2}\Big\}$. Let $p_\theta = q_{T_1-\tau} = \mathrm{law}(Y_{T_1-\tau})$, where $Y_{T_1-\tau}$ is obtained from the reverse process in* (4.4.2). *Consider the loss functions given by*

$$b(t) = \mathbb{E}_{p_t}\Big[\|\nabla \log p_t - s_\theta(t,\cdot)\|^2\Big], \quad \varepsilon_{\mathrm{MGF}} = \log \mathbb{E}_{p_{T_1}}\Big[\exp\Big(\frac{1}{1-\delta}\|\nabla \log p_{T_1} - s_\theta(T_1,\cdot)\|^2\Big)\Big].$$

Then, the distance between the output p_θ and the target distribution p can be bounded as follows:

$$W_2(p,p_\theta) \leq \sqrt{3\tau d} + \int_\tau^{T_1}\sqrt{b(t)}I_\tau(t)\,dt + I_\tau(T_1)\sqrt{\frac{2}{1-\delta}\Big(\delta e^{-\frac{(1-\delta)T_2}{2}} + 2\varepsilon_{\mathrm{MGF}}\Big)} \tag{4.4.3}$$

$$\leq \sqrt{3\tau d} + \sqrt{\frac{J_{\mathsf{SM}}(\theta)}{2}\int_\tau^{T_1} I_\tau(t)^2\,dt} + I_\tau(T_1)\sqrt{\frac{2}{1-\delta}\Big(\delta e^{-\frac{(1-\delta)T_2}{2}} + 2\varepsilon_{\mathrm{MGF}}\Big)}, \tag{4.4.4}$$

where $I_\tau(t) = \exp\Big(t - \tau + \int_\tau^t L_s(r)\,dr\Big)$.

The right-hand side of (4.4.3)–(4.4.4) consists of three error terms, due to *(i)* the early stopping of the reverse process (4.4.2), *(ii)* the approximation of the score function $s_\theta(t,\cdot) \approx \nabla \log p_t$ in (4.4.2), and *(iii)* the approximation $q_0 = \sigma_{T_2} \approx p_{T_1}$ from the output of the Langevin dynamics (4.4.1). A key feature of Theorem 4.4.1 is that, to have a vanishing sampling error, one does *not* need $T_1 \to \infty$. In fact, consider a sequence of estimators satisfying the additional technical assumption[3] $\limsup_{J_{\mathsf{SM}}\to 0}\int_\tau^{T_1}\exp\Big(2\int_\tau^t L_s(r)\,dr\Big)\,dt < \infty$. Then, letting $\tau \to 0$, $J_{\mathsf{SM}} \to 0$, $\varepsilon_{\mathrm{MGF}} \to 0$ and $T_2 \to \infty$, with T_1 fixed, the convergence of p_θ to p in W_2 distance follows from Theorem 4.4.1. We now discuss the roles of T_1, T_2, τ and of the free parameter δ.

The role of T_1 is to ensure the regularity of p_{T_1}. Specifically, we show that p_{T_1} satisfies a log-Sobolev inequality with constant at least $1-\delta$, which leads to the mild logarithmic dependence of T_1 on the subgaussian norm $\|X\|_{\mathrm{SG}}$. In fact, the dependence on d can be even dropped (although the subgaussian norm $\|X\|_{\mathrm{SG}}$ may still depend on d): if $T_1 \geq \frac{1}{2}\log\Big(2 + 172\frac{\|X\|_{\mathrm{SG}}^2}{\delta}\Big)$, then (4.4.3) and (4.4.4) hold with $\frac{d}{3}\exp\Big(-\frac{(1-\delta)T_2}{2}\Big)$ in place of $\delta\exp\Big(-\frac{(1-\delta)T_2}{2}\Big)$ (see the last term of the expression). Choosing $T_1 > 0$ is necessary, as performing Langevin dynamics directly for p works poorly. This was already observed in [SE19] and served precisely as a motivation for SGMs.

The role of T_2 is to improve the accuracy of sampling from p_{T_1} and, due to the regularity of this distribution, the convergence is exponential in T_2. Taking T_2 large, instead of T_1, is beneficial, as the neural network needs to approximate the score of the forward process only until time T_1 and, correspondingly, J_{SM} increases with T_1. In contrast, [KFL22, Thm. 1, Cor. 2] requires $T_1 \to \infty$ to achieve convergence, and a full quantitative analysis of the dependence of the bounds on T_1 is missing.

[3]This condition was implicitly needed in [KFL22] for the same reasons. To see that it is reasonable, notice that L_s is upper bounded by the Lipschitz constant $\mathrm{lip}(s_\theta)$, which is expected to be similar to $\mathrm{lip}(\nabla \log p_t)$ as $J_{\mathsf{SM}} \to 0$. The latter is well behaved by the regularization properties of the OU flow: if p has bounded support, then $\int_\tau^{T_1}\mathrm{lip}(\nabla \log p_t)\,dt < \infty$ for all $0 < \tau < T_1$ [CCL+23b, Lem. 20]. Note also that the one-sided Lipschitz constant can be negative (e.g., for γ_t it is equal to $-\frac{1}{t}$), which helps with the convergence of the integral.

The role of τ is to allow for early stopping in the reverse process (4.4.2). If that is not needed (since *e.g.* the distribution p is sufficiently regular), then one can simply take the limit $\tau \to 0$ in (4.4.3)-(4.4.4).[4]

The role of δ is to provide a trade-off between T_1 and T_2. We remark that the result of Theorem 4.4.1 holds for any $\delta \in (0,1)$, and smaller values of δ give a faster decay of the error term coming from the Langevin dynamics, due to the larger log-Sobolev constant of p_{T_1}. This improvement comes at the price of a tighter lower bound for T_1.

We highlight that our convergence result does not need the condition $\lim_{t\to\infty} L_s(t) = -1$, which was introduced in [KFL22]. This requirement was heuristically justified by the one-sided Lipschitz constant of the stationary distribution γ of the forward process (4.3.6) being -1, but obtaining a rigorous control on $\lim_{t\to\infty} L_s(t)$ only from the L^2 loss of the score estimation remained challenging. To circumvent this issue, we instead introduce the additional loss $\varepsilon_{\mathrm{MGF}}$ in Theorem 4.4.1, which concerns the score estimation *only* at time T_1. By Jensen's inequality, this is a stronger loss than the usual L^2 one ($\varepsilon_{\mathrm{MGF}} \geq \frac{1}{1-\delta} b(T_1)$). However, a simple truncation of the estimator $s_\theta(T_1, \cdot)$ allows us to control $\varepsilon_{\mathrm{MGF}}$ in terms of $b(T_1)$. This is formalized in the result below.

Theorem 4.4.2. *Let Assumptions (A1)-(A2) hold, and let p_t be obtained via the forward OU process in* (4.3.6) *starting from $X \sim p$. Pick $0 < \delta < 1$ and $T_1 \geq \log\Big(\frac{16}{\delta} d(\|X\|_{\mathrm{SG}} + 1)\Big)$. Define the estimator $\widetilde{s_\theta} \colon \mathbb{R}^d \to \mathbb{R}^d$ by*

$$[\widetilde{s}_\theta(x)]_i = \begin{cases} [s_\theta(T_1, x)]_i, & \textit{if } |-x_i - [s_\theta(T_1,x)]_i| \leq \ell, \\ -x_i - \ell, & \textit{if } [s_\theta(T_1,x)]_i < -x_i - \ell, \\ -x_i + \ell, & \textit{if } [s_\theta(T_1,x)]_i > -x_i + \ell, \end{cases} \tag{4.4.5}$$

where $\ell = \ell(x) = \frac{\delta}{2d}(1 + \|x\|)$. Fix $0 < \varepsilon < 1$. Then, for all $0 < \beta \leq \frac{1}{36\delta^2}$, we have

$$\log \mathbb{E}_{p_{T_1}}\Big[\exp\Big(\beta \|\nabla \log p_{T_1} - \widetilde{s}_\theta\|^2\Big)\Big] \leq \varepsilon, \tag{4.4.6}$$

provided that

$$b(T_1) \leq \frac{1}{34\beta} \varepsilon^{1+36\beta\delta^2}. \tag{4.4.7}$$

A combination of Theorems 4.4.1 and 4.4.2 gives an end-to-end convergence result in W_2 distance requiring only a control on the L^2 loss $\{b(t)\}_{t\in[0,T_1]}$, which is the object minimized in practice. We regard the truncation of the estimator $s_\theta(T_1,\cdot)$ carried out in Theorem 4.4.2 as purely technical. In fact, even if the estimator $s_\theta(T_1,\cdot)$ explicitly minimizes the training loss $b(T_1)$, one also expects the stronger loss $\varepsilon_{\mathrm{MGF}}$ to be small, when δ is sufficiently small. This is confirmed by the numerical results of Figure 4.1, for which we do not replace the estimator $s_\theta(T_1,\cdot)$ with $\widetilde{s}_\theta$. Specifically, the plots show that, having fixed T_1, both $W_2(p, p_\theta)$ (in orange) and $W_2(p_{T_1}, q_0)$ (in blue) decrease as a function of the running time T_2 of the inexact Langevin dynamics (4.4.1) for different standard datasets.

4.4.1 Proof ideas

Analysis of the reverse process (4.4.2)**.** To obtain Theorem 4.4.1, we start with the analysis of the reverse process (4.4.2). By adapting the argument in [KFL22], we derive the following bound on the Wasserstein distance between p_τ and $q_{T_1-\tau}$.

[4]This passage can be justified by an application of monotone convergence, after estimating the one-sided Lipschitz constant $L_s(t)$ with the Lipschitz constant of $s_\theta(t,\cdot)$.

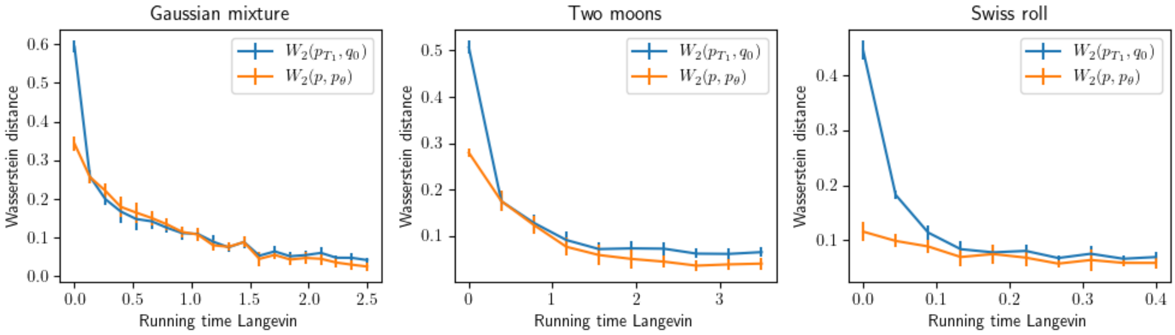


Figure 4.1: Simulation results for an asymmetric mixture of two gaussians (left), the two moons dataset (center) and the rescaled swiss roll (right). For fixed T_1 and variable T_2, we plot in blue the W_2 distance between the perturbed measure p_{T_1} and the output of (4.4.1), while we plot in orange the W_2 distance between the true distribution p and the output of the algorithm p_θ. As expected, both quickly decrease as T_2 increases.

Proposition 4.4.3. *Under Assumptions (A1)-(A2), for $0 \le \tau < T_1$, we have*

$$W_2(p_\tau, q_{T_1-\tau}) \le I_\tau(T_1) W_2(p_{T_1}, q_0) + \int_\tau^{T_1} I_\tau(t)\sqrt{b(t)}\, dt. \tag{4.4.8}$$

For completeness we include a proof in Section 4.8, since compared to [KFL22] we consider $M(t) \equiv 1$ (and not $M(t) \equiv 2$), a different starting distribution for Y_0, and early stopping. In addition, we need the following short-time estimate on $W_2(p, p_\tau)$, which is proved in Section 4.8.

Lemma 4.4.4. *Suppose that $M := \int_{\mathbb{R}^d} \|x\|^2 p(x) dx < \infty$. Then, for $0 < \tau < \frac{d}{M^2}$, we have*

$$W_2(p, p_\tau) \le \sqrt{3\tau d}.$$

Using the triangle inequality for W_2, the combination of Proposition 4.4.3 and Lemma 4.4.4 readily gives

$$W_2(p, q_{T_1-\tau}) \le \sqrt{3\tau d} + I_\tau(T_1) W_2(p_{T_1}, q_0) + \int_\tau^{T_1} \sqrt{b(t)} I_\tau(t)\, dt. \tag{4.4.9}$$

This bound shows that to achieve convergence, we need *(i)* small τ, *(ii)* $b(t) \to 0$ (which is reasonable, since s_θ is obtained by minimizing the L^2 loss), and *(iii)* $W_2(q_0, p_{T_1}) \to 0$. The latter condition corresponds to choosing a good starting distribution for the reverse process (instead of γ), and it is ensured by the inexact Langevin dynamics (4.4.1), which will be analyzed next.

Analysis of inexact Langevin dynamics (4.4.1)**.** Recall the definition of the log-Sobolev inequality.

Definition 4.4.5. *A probability measure μ satisfies a log-Sobolev inequality with constant $\kappa > 0$ (notation: $\mathrm{LSI}(\kappa)$) if, for all probability measures ν,*

$$\mathcal{D}_{\mathrm{KL}}(\nu \,\|\, \mu) \le \frac{1}{2\kappa} \mathcal{I}_\mu(\nu), \tag{4.4.10}$$

where $\mathcal{I}_\mu(\nu)$ is the relative Fisher Information, defined by

$$\mathcal{I}_\mu(\nu) = \begin{cases} \int_{\mathbb{R}^d} \left\| \nabla \log \frac{d\nu}{d\mu} \right\|^2 d\nu = 4 \int_{\mathbb{R}^d} \left\| \nabla \sqrt{\frac{d\nu}{d\mu}} \right\|^2 d\mu, & \text{if } \nu \ll \mu \text{ and } \sqrt{\frac{d\nu}{d\mu}} \in H^1(\mu), \\ +\infty, & \text{otherwise.} \end{cases} \tag{4.4.11}$$

Here, $H^1(\mu)$ denotes the weighted Sobolev space. If μ satisfies a log-Sobolev inequality, we use the notation $C_{\mathrm{LS}}(\mu)$ for its optimal constant.

Remark 4.4.6. *If μ satisfies $\mathrm{LSI}(\kappa)$ for some $\kappa > 0$, then it also satisfies the transport-entropy inequality [OV00] (see also [GL10, Thm. 8.12] for an overview of related results)*

$$W_2(\nu, \mu) \leq \sqrt{\frac{2}{\kappa} \mathcal{D}_{\mathrm{KL}}(\nu \,\|\, \mu)}. \tag{4.4.12}$$

It is well known that Langevin dynamics for a measure μ converges exponentially fast in KL-divergence if μ satisfies a log-Sobolev inequality, see *e.g.* [Vil03, VW19]. A similar convergence result has recently been proved in [YYW23] for the inexact Langevin dynamics

$$\begin{cases} Z_0 \sim \nu, \\ dZ_t = s_\theta(Z_t)\, dt + \sqrt{2}\, dB_t, \end{cases} \tag{4.4.13}$$

where s_θ approximates the score $\nabla \log \mu$. For convenience, we state Theorem 1 of [YYW23] below.

Theorem 4.4.7. *Let μ, ν be probability measures with full support that admit densities with respect to the Lebesgue measure. Suppose that μ satisfies a log-Sobolev inequality and let $\kappa = C_{\mathrm{LS}}(\mu)$. Then, the time-marginal law ν_t of the inexact Langevin dynamics* (4.4.13) *satisfies, for $t \geq 0$,*

$$\mathcal{D}_{\mathrm{KL}}(\nu_t \,\|\, \mu) \leq e^{-\frac{1}{2}\kappa t} \mathcal{D}_{\mathrm{KL}}(\nu \,\|\, \mu) + 2 \log \mathbb{E}_\mu \Big[\exp\Big(\frac{1}{\kappa} \|\nabla \log \mu - s_\theta\|^2 \Big) \Big].$$

The application of Theorem 4.4.7 requires p_{T_1} to satisfy a log-Sobolev inequality, as well as the estimation of $C_{\mathrm{LS}}(p_{T_1})$. To ensure this, we notice that $p_{T_1} = \mathrm{law}\Big(e^{-T_1} X + \sqrt{1 - e^{-2T_1}} Z\Big)$ where $Z \sim \gamma$ is an independent Gaussian, and apply recent results from [CCNW21] that quantify the log-Sobolev constant of the convolution of a subgaussian probability measure with a Gaussian having sufficiently high variance (cf. Lemma 4.8.3 and Theorem 4.8.4 in Section 4.8). In this way, we deduce that $C_{\mathrm{LS}}(p_{T_1}) \geq 1 - \delta$. Next, in order to apply Theorem 4.4.7 with $\mu = p_{T_1}$ and $\nu = \gamma$, we need an estimate of $\mathcal{D}_{\mathrm{KL}}(\gamma \,\|\, p_{T_1})$. This is provided by the result below, which is proved in Section 4.8.

Lemma 4.4.8. *Let $p \in \mathcal{P}(\mathbb{R}^d)$ with $M := \int_{\mathbb{R}^d} \|x\|^2 \, dp(x) < \infty$. Then, for $t > 0$ we have*

$$\mathcal{D}_{\mathrm{KL}}(\gamma \,\|\, p_t) \leq \frac{d}{2\sigma_t} \Big(\frac{M}{d} e^{-2t} + \sigma_t \log \sigma_t - \sigma_t + 1 \Big),$$

with $\sigma_t = 1 - e^{-2t}$. Thus, for $t \geq \max\Big(\log(\sqrt{2}), \log\Big(\sqrt{(M + d/2)/\delta}\Big)\Big)$, we have the upper bound $\mathcal{D}_{\mathrm{KL}}(\gamma \,\|\, p_t) \leq \delta$, while for $t \geq \max\Big(\log(\sqrt{2}), \log\Big(\sqrt{3M}\Big)\Big)$ we have $\mathcal{D}_{\mathrm{KL}}(\gamma \,\|\, p_t) \leq \frac{d}{3}$.

By combining the estimate on $\mathcal{D}_{\mathrm{KL}}(\sigma_{T_2} \,\|\, p_{T_1})$ given by Theorem 4.4.7 with the transport-entropy inequality (4.4.12) (which also uses $C_{\mathrm{LS}}(p_{T_1}) \geq 1 - \delta > 0$) and the above lemma, we obtain an upper bound on $W_2(\sigma_{T_2}, p_{T_1}) = W_2(q_0, p_{T_1})$. Combining this upper bound with (4.4.9) gives the desired inequality (4.4.3). The inequality (4.4.4) then follows from the Cauchy-Schwarz inequality, which concludes the proof of Theorem 4.4.1. The complete argument is contained in Section 4.8.

Controlling a stronger loss (Theorem 4.4.2). It is not difficult to construct an estimator $s_\theta(T_1, \cdot)$ such that $b(T_1)$ is arbitrarily small and $\varepsilon_{\mathrm{MGF}}$ diverges. In fact, consider estimating the score of the standard Gaussian γ, given by $\nabla \log \gamma(x) = -x$, via a sequence of estimators of the form $s_M(x) = -x\mathbb{1}_{\|x\| \leq M}$. Then, we have $\lim_{M \to \infty} b(T_1) = 0$, but $\varepsilon_{\mathrm{MGF}}$ is infinite for all $M \geq 0$. This might seem discouraging, but we can prevent such pathological problems by leveraging our knowledge about the target score function. This allows us to fix predictions of the estimator $s_\theta(T_1, x)$ that happen to be very far from the target value $\nabla \log p_{T_1}(x)$. More precisely, by choosing T_1 according to the prescription of Theorem 4.4.2, we can ensure that $\nabla \log p_{T_1}(x)$ lies in a region around the score of the standard Gaussian $\nabla \log \gamma = -x$, i.e.,

$$|-x_i - \partial_i \log p_{T_1}| \leq \frac{\delta}{2d}(1 + \|x\|), \quad \text{for all } i \in \{1, \dots, d\},$$

see Lemma 4.9.2. This is illustrated in Figure 4.2, in one dimension: the green dashed line represents $\nabla \log \gamma = -x$, and our choice of T_1 guarantees that $\nabla \log p_{T_1}$ lies in the region delimited by the blue and purple lines. Whenever s_θ returns a value outside this region, we correct it by choosing the closest value on the boundary. This leads to the definition of the estimator $\widetilde{s}_\theta$ given by (4.4.5). For this new estimator, we can now convert an L^2 error into an upper bound for $\varepsilon_{\mathrm{MGF}}$, which gives Theorem 4.4.2. The argument crucially relies on the fast decay of the tails of p_{T_1}, thanks to the similarity with the standard Gaussian γ, and exploits the confinement knowledge to deduce a converse bound to Jensen's inequality, cf. Lemma 4.7.4. The complete proof is contained in Section 4.9.

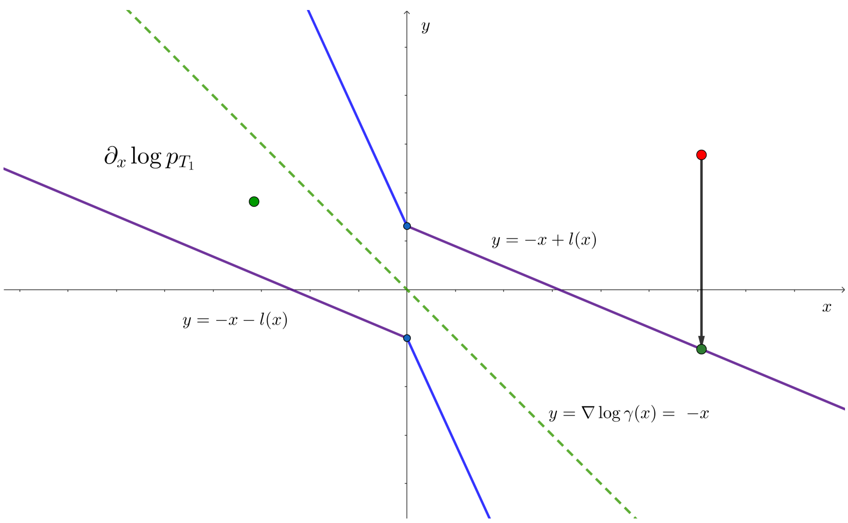


Figure 4.2: The confinement region for $\nabla \log p_{T_1}$.

4.5 Convergence of discretized schemes in total variation distance

In practical implementations, the continuous time dynamics needs to be approximated by a discrete time scheme. The existing literature for discretizations of the reverse process [CCL+23b, CLL22, LLT22] focuses on the reverse SDE instead of the ODE (*i.e.*, $M(t) = 2$ instead of $M(t) = 1$ in (4.3.3)), and on information-divergence metrics instead of the Wasserstein distance, as for the latter the analysis seems significantly more complicated [CCL+23b, Sec. 4]. The two-stage algorithm (with a stochastic reverse process) considered in this paper is compatible with such recent analyses: we illustrate this in the present section by providing convergence guarantees for a natural discretization of the algorithm. The argument proceeds in a similar way to Section 4.4. In particular, a key role is played again by Lemmas 4.8.3 and 4.4.8: the first guarantees that p_{T_1} satisfies a log-Sobolev inequality with a good constant, so that a discretization of the inexact Langevin dynamics performs well, while the second ensures that the Gaussian distribution γ is a good initialization for the algorithm.

For the discretization of (4.4.1), we consider following the inexact Langevin algorithm with variable step sizes $h_k > 0$:

$$\begin{cases} Z_0 \sim \gamma, \\ Z_{k+1} = Z_k + h_k s_\theta(T_1, Z_k) + \sqrt{2h_k} B_k, \end{cases} \tag{4.5.1}$$

where $(B_k)_k \overset{\text{iid}}{\sim} \mathcal{N}(0, I_d)$. This is run for a variable number of steps N_2, and we now denote by σ_k the law of Z_k.

Convergence guarantees for Z_k as $k \to \infty$ can be obtained from the analysis in [YYW23, Thm. 2], which we slightly modify to take into account a decaying step size: this will give a logarithmic improvement on the computational complexity, in the same spirit as [DK19]. As in Section 4.4 (and as in [YYW23]), the analysis of the Langevin algorithm introduces a modified loss $\tilde{\varepsilon}_{\text{MGF}}$, which is stronger than the standard L^2 error on the accuracy of the score estimate. However, at time T_1 this loss can be controlled again thanks to Theorem 4.4.2 (cf. also Remark 4.5.2).

As for the reverse process, to take advantage of the existing literature, we consider a discretization of the reverse SDE instead of the reverse ODE (corresponding to $M(t) = 2$ instead of $M(t) = 1$ in (4.3.5)). The chosen numerical method is the popular exponential integrator scheme [ZC23], so that the second stage of the algorithm is given by

$$\begin{cases} Y_0 = Z_{N_2}, \\ Y_{k+1} = Y_k e^{\tilde{h}_k} + 2 s_\theta(t_k, Y_k)\left(e^{\tilde{h}_k} - 1\right) + \sqrt{e^{2\tilde{h}_k} - 1}\, \tilde{B}_k, \\ t_k = T_1 - \sum_{i=0}^{k-1} \tilde{h}_i, \end{cases} \tag{4.5.2}$$

for a sequence of step sizes $(\tilde{h}_k)_{k=0}^{N_1-1}$ such that $\sum_{k=0}^{N_1-1} \tilde{h}_k \le T_1$ and for $(\tilde{B}_k)_k \overset{\text{iid}}{\sim} \mathcal{N}(0, I_d)$. The output Y_{N_1} is finally taken as an approximate sample from p, and we denote now by p_θ its law; under appropriate assumptions, we derive convergence guarantees of p_θ to p in total variation distance. In place of the integrated L^2 loss in (4.3.4), for discrete time schemes it is natural to introduce the analogous loss

$$\widehat{J_{\mathsf{SM}}} = \sum_{k=0}^{N_1-1} \tilde{h}_k \mathbb{E}_{p_{t_k}}\left[\|\nabla \log p_{t_k} - s_\theta(t_k, \cdot)\|^2\right].$$

We can think of $\widehat{J_{\mathsf{SM}}}$ as an approximation of J_{SM}: it is a standard and realistic assumption that $\widehat{J_{\mathsf{SM}}}$ can be made arbitrarily small with sufficient data and model capacity. The errors arising in the reverse process (4.5.2) (due to the inaccuracy of the score estimate and the use of the discrete scheme) can be bounded thanks to the recent theoretical literature on the performance of diffusion models, see e.g. [CCL+23b, CLL22]; together with the analysis of (4.5.1), this allows us to deduce end-to-end convergence guarantees for the two-stage algorithm.

To illustrate this, we present below one such result which exploits the analysis of [CLL22, Thm. 1], and we refer the reader to Section 4.10 for the proof. Additional results under different assumptions and choices of the step size can be deduced by adapting other arguments (e.g. [CLL22, Thm. 2]). Specifically, we consider a constant step size for the reverse process and a standard smoothness condition [CCL+23b, CLL22].

Theorem 4.5.1. *Under Assumption (A1), pick $0 < \delta \leq \frac{1}{2}$ and suppose that $T_1 \geq \frac{1}{2}\log\Big(2 + 172\frac{\|X\|_{\mathrm{SG}}^2}{\delta} + \frac{d}{2\delta}\Big)$. Assume in addition that $\nabla \log p_t$ is L_1-Lipschitz for $t \in [0, T_1]$ and that $s_\theta(T_1, \cdot)$ is L_2-Lipschitz with $L_1, L_2 \geq 1$, and consider the modified loss at time T_1*

$$\tilde{\varepsilon}_{\mathrm{MGF}} := \log \mathbb{E}_{p_{T_1}}\Big[\exp\Big(\frac{9}{1-\delta}\|\nabla \log p_{T_1} - s_\theta(T_1, \cdot)\|^2\Big)\Big]. \tag{4.5.3}$$

Then, for the algorithm described above with step sizes $h_k = \frac{1}{24L_1L_2 + \frac{k+1}{16}}$ and $\tilde{h}_k = \tilde{h} = \frac{T_1}{N_1} \leq 1$, we have that

$$\begin{aligned} \|p - p_\theta\|_{\mathsf{TV}} &\lesssim \sqrt{\widehat{J_{\mathsf{SM}}} + \frac{dL_1^2T_1^2}{N_1}} + \sqrt{\Big(\frac{L_1L_2}{N_2+1}\Big)^2 + \frac{dL_2^2}{N_2+1} + \tilde{\varepsilon}_{\mathrm{MGF}}} \\ &= \sqrt{\widehat{J_{\mathsf{SM}}} + dL_1^2T_1\tilde{h}} + \sqrt{\Big(\frac{L_1L_2}{N_2+1}\Big)^2 + \frac{dL_2^2}{N_2+1} + \tilde{\varepsilon}_{\mathrm{MGF}}}. \end{aligned} \tag{4.5.4}$$

Remark 4.5.2. *To deduce convergence from the above result assuming only L^2 accuracy of the score, we need to control $\tilde{\varepsilon}_{\mathrm{MGF}}$. This can be done again using Theorem 4.4.2, where we now choose $\beta = \frac{9}{1-\delta}$ and take $0 < \delta < 0.054$ to fulfill $\beta \leq \frac{1}{36\delta^2}$.*

Remark 4.5.3 (Complexity of sampling)**.** *Suppose that the goal is to achieve $\|p - p_\theta\|_{\mathsf{TV}} \leq \varepsilon$ for some $0 < \varepsilon < 1$. A typical assumption is to be able to control the L^2 error of the score approximation; hence, by the remark above, we can assume that $\widehat{J_{\mathsf{SM}}}, \tilde{\varepsilon}_{\mathrm{MGF}} \lesssim \varepsilon^2$, as needed in the bound* (4.5.4)*. Consequently,* (4.5.4) *shows that the algorithm needs at most N steps to ensure $\|p - p_\theta\|_{\mathsf{TV}} \leq \varepsilon$, with*

$$N = N_1 + N_2 \lesssim \frac{d}{\varepsilon^2} \cdot \Big(L_1^2 \log^2(d + \|X\|_{\mathrm{SG}}) + L_2^2\Big).$$

Remark 4.5.4. *If we choose also for the inexact Langevin algorithm* (4.5.1) *a fixed step size $h_k = h \leq \frac{1}{24L_1L_2}$, we obtain instead the bound*

$$\|p - p_\theta\|_{\mathsf{TV}} \lesssim \sqrt{\widehat{J_{\mathsf{SM}}} + \frac{dL_1^2T_1^2}{N_1}} + \sqrt{e^{-\frac{1}{8}hN_2} + L_2(L_1 + L_2)dh + \tilde{\varepsilon}_{\mathrm{MGF}}}.$$

In this case, for the number of steps N to achieve $\|p - p_\theta\|_{\mathsf{TV}} \leq \varepsilon$, we have that

$$N = N_1 + N_2 \lesssim \frac{d}{\varepsilon^2} \cdot \Big(L_1^2 \log^2(d + \|X\|_{\mathrm{SG}}) + L_2(L_1 + L_2)\log\frac{1}{\varepsilon}\Big).$$

4.5.1 Comparison with previous results

To highlight a few favorable properties of the predictor-corrector algorithm, we compare the result above with [CLL22, Thm. 1], translated into total variation distance via Pinsker's inequality. The authors of [CLL22] consider, under similar assumptions, the standard sampling scheme based on a discretization of the reverse SDE without the prior Langevin algorithm or any corrector; in other words, they consider (4.5.2) with constant step size initialized at $Z_0 \sim \gamma$. For the corresponding output distribution $\widehat{p_\theta}$ they prove the bound

$$\begin{aligned}\|p - \widehat{p_\theta}\|_{\mathsf{TV}} &\lesssim \sqrt{(M+d)e^{-2T_1} + \widehat{J_{\mathsf{SM}}} + \frac{dL^2T_1^2}{N_1}} \\ &= \sqrt{(M+d)e^{-2T_1} + \widehat{J_{\mathsf{SM}}} + dL^2T_1\tilde{h}},\end{aligned} \tag{4.5.5}$$

where M is the second moment of p. Therefore, to achieve $\|p - \widehat{p_\theta}\|_{\mathsf{TV}} \leq \varepsilon$ the reverse SDE needs at most N steps with

$$N \lesssim \frac{dL_1^2}{\varepsilon^2} \log^2\left(\frac{M+d}{\varepsilon^2}\right).$$

Comparing the bounds (4.5.4) and (4.5.5) shows some advantages of the predictor-corrector schemes, arising from a fixed choice of T_1 independent of the desired sampling accuracy ε.

1. The convergence result in (4.5.4), unlike (4.5.5), is stable with T_1. In other words, for a fixed choice of step size $\tilde{h}$ in the reverse process, the bound in (4.5.5) explodes as $T_1 \to \infty$, which is however necessary to minimize the error $(M+d)e^{-2T_1}$ arising from the approximation $p_{T_1} \approx \gamma$. Thus, there is a trade-off between the choice of T_1 and the step size in the reverse process. In contrast, this problem does not occur for the bound in (4.5.4), since T_1 is now fixed. For the specified choice of step sizes h_k (which is independent of the desired accuracy), the error goes to 0 as $\tilde{h} \to 0$ and $N_2 \to \infty$. At the same time, the bound is now stable for any choice of the variable quantities $\tilde{h} \leq 1$ and $N_2 \geq 1$. [YYW23] is also aimed at obtaining stable convergence. However, the results therein apply only to distributions satisfying a log-Sobolev inequality and under stronger assumptions on the accuracy of the score.

2. To achieve convergence, the bound (4.5.5) requires learning the score $\nabla \log p_t$ on a time interval which increases as $T_1 \to \infty$: correspondingly, the error term $\widehat{J_{\mathsf{SM}}}$ increases too with T_1. For example, assuming that we have L^2-accuracy of ε^2 at every time (i.e., $b(t_k) \leq \varepsilon^2$ for every t_k, cf. Assumption 3 of [CCL+23b]) we have that $\widehat{J_{\mathsf{SM}}} = \varepsilon^2 T_1$, which diverges as $T_1 \to \infty$ with fixed $\varepsilon > 0$ (see also [CCL+23b, Thm. 2]). These problems do not occur with the bound in (4.5.4), since T_1 is fixed; this simplifies the training procedure of the neural network learning the score and contributes to the stability of the convergence result. At the same time, this further pushes the observation by [CCL+23b] that "sampling is as easy as learning the score", in the sense that knowledge of the score $\nabla \log p_t$ for all times t suffices for efficient sampling. Indeed, Theorem 4.5.1 shows the stronger result that, for norm-subgaussian distributions, knowledge of the score *on a fixed finite time interval* is actually enough.

3. Finally, when looking at the dependence on the desired accuracy ε, the bound of (4.5.4) removes a factor of $\log^2\left(\frac{1}{\varepsilon}\right)$ in the number of steps required by the algorithm.

4.6 Concluding remarks

In this work, we give convergence guarantees for a variant of the popular predictor-corrector approach in the context of score-based generative modeling. Our analysis provides bounds that *(i)* require running the forward process only for a fixed time T_1, which does not depend on the final sampling accuracy, *(ii)* make minimal assumptions on the data (subgaussianity of the norm), *(iii)* exhibit a mild logarithmic dependence on the input dimension and on the tails of the data distribution, and *(iv)* allow for realistic assumptions on the score estimation, in the form of a control on the standard L^2 loss integrated over the finite time T_1.

4.7 Additional and auxiliary lemmas

Notation

Recall that a scalar random variable X is *subgaussian* if there exists a constant $K > 0$ such that $\mathbb{E}\left[e^{\frac{X^2}{K^2}}\right] \leq 2$. Its subgaussian norm is defined by $\|X\|_{\psi_2} := \inf\left\{t > 0 : \mathbb{E}\left[e^{\frac{X^2}{t^2}}\right] \leq 2\right\}$. Furthermore, an $\mathbb{R}^d$-valued random variable X is said to be norm-subgaussian if its euclidean norm $\|X\|$ is subgaussian. We define $\|X\|_{\mathrm{SG}} := \|\|X\|\|_{\psi_2}$ Note that both $\|\cdot\|_{\psi_2}$ and $\|\cdot\|_{\mathrm{SG}}$ are norms and, if $\operatorname{Supp} X \subset B(0, R)$ for some radius $R > 0$, then $\|X\|_{\mathrm{SG}} \leq \frac{R}{\sqrt{\log(2)}}$.

For the convenience of the reader, in the table below we recall the relevant notation used in the paper.

Notation	**Meaning**
$\|X\|_{\psi_2}$	Subgaussian norm of random variable
$\|X\|_{\mathsf{SG}} = \|\|X\|\|_{\psi_2}$	Subgaussian norm of random vector
p	Target distribution
p_t	Perturbed distribution at time t
p_θ	Output distribution of the algorithm
$s_\theta(t, x) \approx \nabla \log p_t(x)$	Estimator for the score function
$L_s(t)$	One-sided Lipschitz constant of $s_\theta(t, \cdot)$
$\tau \geq 0$	Early stopping time
$T_1 > 0$	Running time of the forward process
$N_1 > 0$	Number of steps for reverse process
$T_2 \geq 0$	Running time of Langevin dynamics
$N_2 \geq 0$	Number of steps for Langevin algorithm
$b(t) = \mathbb{E}_{p_t}\left[\|\nabla \log p_t - s_\theta(t, \cdot)\|^2\right]$	L^2-error for score approximation
$\varepsilon_{\mathrm{MGF}} = \log \mathbb{E}_{p_{T_1}}\left[e^{\frac{1}{1-\delta}\|\nabla \log p_{T_1} - s_\theta(T_1, \cdot)\|^2}\right]$ $\tilde{\varepsilon}_{\mathrm{MGF}} = \log \mathbb{E}_{p_{T_1}}\left[e^{\frac{9}{1-\delta}\|\nabla \log p_{T_1} - s_\theta(T_1, \cdot)\|^2}\right]$	Stronger losses for the score at time T_1

Auxiliary lemmas

First of all, we recall the following classical properties of the Gaussian distribution.

Lemma 4.7.1. *Let $Z \sim \gamma_t$. Then,*

$$\|Z\|_{\mathrm{SG}} \leq 2\sqrt{dt}.$$

Moreover, $C_{\mathrm{LS}}(\gamma_{x,t}) \geq \frac{1}{t}$ for all $x \in \mathbb{R}^d$ and $t > 0$.

Proof. Without loss of generality, suppose that $Z \sim \gamma$ (i.e. $t = 1$). Recalling the moment generating function of the χ^2-distribution, we have that, for $0 \leq c \leq \frac{1}{4d} < \frac{1}{2}$,

$$\mathbb{E}\Big[e^{c\|Z\|^2}\Big] = (1-2c)^{-\frac{d}{2}} \leq \frac{1}{1-2cd} \leq 2,$$

which proves the first claim.

The statement about the log-Sobolev constant is well known (it follows for example from the Bakry–Émery criterion, cf. [Vil03, Thm. 9.9],[BGL14]). ∎

Lemma 4.7.2. *For all $t, c > 0$, we have*

$$\int_t^\infty e^{-cx^2}\,dx \leq \frac{1}{2ct}e^{-ct^2}.$$

Proof. As in [Ver18, Prop 2.1.2], we have

$$\int_t^\infty e^{-cx^2}\,dx = \int_0^\infty e^{-c(x^2+2xt+t^2)}\,dx \leq e^{-ct^2}\int_0^\infty e^{-2ctx}\,dx = \frac{1}{2ct}e^{-ct^2}.$$

∎

The next lemma provides some useful estimates for norm-subgaussian random vectors (cf. [Ver18, JNG+19]).

Lemma 4.7.3. *Let $X \sim p$ be a norm-subgaussian random vector. The following hold:*

(i) *For all $s \geq 0$,*

$$\mathbb{P}(\|X\| \geq s) \leq 2e^{-\frac{s^2}{\|X\|_{\mathrm{SG}}^2}}. \tag{4.7.1}$$

Moreover,

$$\mathbb{E}\Big[\|X\|^2\Big] \leq 2\|X\|_{\mathrm{SG}}^2. \tag{4.7.2}$$

(ii) *For any $L > 0$ and $0 < c < \frac{1}{\|X\|_{\mathrm{SG}}^2}$,*

$$\int_{B(0,L)^c} e^{c\|X\|^2}\,dp(x) \leq 2\left(1 + \frac{c\|X\|_{\mathrm{SG}}^2}{1-c\|X\|_{\mathrm{SG}}^2}\right)e^{-L^2\cdot\left(\frac{1}{\|X\|_{\mathrm{SG}}^2}-c\right)}.$$

(iii) *For any $L > 0$,*

$$\int_{B(0,L)^c}\|x\|\,dp(x) \leq \left(2L + \frac{\|X\|_{\mathrm{SG}}^2}{L}\right)e^{-\frac{L^2}{\|X\|_{\mathrm{SG}}^2}}.$$

Proof. (i) As in [Ver18, Prop. 2.5.2], we have

$$\mathbb{P}(\|X\| \geq s) = \mathbb{P}\left(e^{\frac{\|X\|^2}{\|X\|_{\mathrm{SG}}^2}} \geq e^{\frac{s^2}{\|X\|_{\mathrm{SG}}^2}}\right) \leq e^{-\frac{s^2}{\|X\|_{\mathrm{SG}}^2}} \mathbb{E}\left[e^{\frac{\|X\|^2}{\|X\|_{\mathrm{SG}}^2}}\right] \leq 2e^{-\frac{s^2}{\|X\|_{\mathrm{SG}}^2}},$$

where the first inequality follows from Markov inequality and the second uses the definition of norm-subgaussianity. Using (4.7.1), we obtain

$$\mathbb{E}\Big[\|X\|^2\Big] = \int_0^\infty \mathbb{P}\Big(\|X\|^2 \geq k\Big) dk \leq 2\int_0^\infty e^{-\frac{k}{\|X\|_{\mathrm{SG}}^2}} dk = 2\|X\|_{\mathrm{SG}}^2,$$

which proves (4.7.2).

(ii) Let $Y = e^{c\|X\|^2} \mathbb{1}_{\{\|X\| \geq L\}}$. Then, $\int_{B(0,L)^c} e^{c\|X\|^2} p(x) dx = \mathbb{E}[Y]$. Moreover, the following chain of inequalities holds:

$$\begin{aligned}
\mathbb{E}[Y] &= \int_0^\infty \mathbb{P}(Y \geq k)\, dk \\
&\leq e^{cL^2} \cdot \mathbb{P}(\|X\| \geq L) + \int_{e^{cL^2}}^\infty \mathbb{P}\left(\|X\|^2 \geq \frac{\log k}{c}\right) dk \\
&\leq 2e^{cL^2} e^{-\frac{L^2}{\|X\|_{\mathrm{SG}}^2}} + 2\int_{e^{cL^2}}^\infty k^{-\frac{1}{c\|X\|_{\mathrm{SG}}^2}}\, dk \\
&= 2\left(1 + \frac{c\|X\|_{\mathrm{SG}}^2}{1 - c\|X\|_{\mathrm{SG}}^2}\right) e^{-L^2\left(\frac{1}{\|X\|_{\mathrm{SG}}^2} - c\right)},
\end{aligned}$$

where the third line follows from point (i).

(iii) Similarly to the proof of the previous point, let $Y = \|X\| \mathbb{1}_{\|X\| \geq L}$. Then,

$$\int_{B(0,L)^c} \|X\| p(x) dx = \mathbb{E}[Y].$$

Moreover, the following chain of inequalities holds:

$$\begin{aligned}
\mathbb{E}[Y] &= \int_0^\infty \mathbb{P}(Y \geq k)\, dk \\
&\leq L \cdot \mathbb{P}(\|X\| \geq L) + \int_L^\infty \mathbb{P}(\|X\| \geq k)\, dk \\
&\leq 2Le^{-\frac{L^2}{\|X\|_{\mathrm{SG}}^2}} + 2\int_L^\infty e^{-\frac{k^2}{\|X\|_{\mathrm{SG}}^2}}\, dk \\
&\leq 2Le^{-\frac{L^2}{\|X\|_{\mathrm{SG}}^2}} + \frac{\|X\|_{\mathrm{SG}}^2}{L} e^{-\frac{L^2}{\|X\|_{\mathrm{SG}}^2}} \\
&= \left(2L + \frac{\|X\|_{\mathrm{SG}}^2}{L}\right) e^{-\frac{L^2}{\|X\|_{\mathrm{SG}}^2}},
\end{aligned}$$

where the second inequality follows from point (i) and the third one follows from Lemma 4.7.2.

■

The next lemma will be useful to find an upper bound for $\varepsilon_{\mathrm{MGF}}$ in terms of $b(T_1)$; the corresponding lower bound is easy and follows immediately from Jensen's inequality.

Lemma 4.7.4. *Consider a random variable F such that $0 \leq F \leq M$ for some $M > 0$. Then,*

$$\mathbb{E}\Big[e^F\Big] \leq 1 + e^M \mathbb{E}[F].$$

Proof. By the mean value theorem, for $x \geq 0$ we have the bound $e^x \leq 1 + xe^x$. Hence, $e^F \leq 1 + e^M F$, from which the conclusion follows by taking the expectation on both sides. ■

4.8 Proof of Theorem 4.4.1

We begin with the proof of Proposition 4.4.3, which gives W_2-estimates for the reverse process Y_t satisfying (4.4.2). Its time-marginals $q_t = \mathrm{law}(Y_t)$ satisfy the continuity equation

$$\begin{cases} q_0 = \mathrm{law}(Z_{T_2}), \\ \partial_t q_t(x) + \nabla \cdot [q_t(x)(x + \widehat{s_\theta}(t,x))] = 0. \end{cases} \tag{4.8.1}$$

Here and below we use the symbol $\widehat{(\cdot)}$ to denote the time-reversal of a function on $[0, T_1]$, e.g., $\hat{p}_t = p_{T_1 - t}$, $\widehat{s_\theta}(t, \cdot) = s_\theta(T_1 - t, \cdot)$, $\hat{b}(t) = b(T_1 - t)$, $\widehat{L_s}(t) = L_s(T_1 - t)$.

Our goal is to obtain an upper bound for $W_2(p_\tau, q_{T_1-\tau})$, where $(\hat{p}_t)_t$ satisfies the Fokker–Planck equation

$$\begin{cases} \hat{p}_0 = p_{T_1}, \\ \partial_t \hat{p}_t(x) + \nabla \cdot [\hat{p}_t(x)(x + \nabla \log \hat{p}_t(x))] = 0. \end{cases} \tag{4.8.2}$$

Following [KFL22], we apply the following well-known formula for the derivative of the Wasserstein distance between two curves of probability measures, cf. [AGS08, Thm. 8.4.7, Rmk. 8.4.8], [Vil09, Thm. 23.9].

Theorem 4.8.1. *Let $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ be the space of probability measures on $\mathbb{R}^d$ with finite second moment equipped with the Wasserstein distance W_2. Consider two weakly continuous curves $(\mu_t)_t, (\nu_t)_t$ in $\mathcal{P}_2(\mathbb{R}^d)$ that solve the continuity equations*

$$\partial_t \mu_t + \nabla \cdot (\xi_t \mu_t) = 0, \qquad \partial_t \nu_t + \nabla \cdot \left(\tilde{\xi}_t \nu_t\right) = 0.$$

Suppose moreover that, for some $0 \leq t_1 < t_2 < \infty$, we have

$$\int_{t_1}^{t_2} \left(\mathbb{E}_{\mu_t}\Big[\|\xi_t\|^2\Big] + \mathbb{E}_{\nu_t}\Big[\big\|\tilde{\xi}_t\big\|^2\Big]\right) dt < \infty.$$

Then, denoting by π_t an optimal coupling for $W_2(\mu_t, \nu_t)$, we have

$$\frac{d}{dt} \frac{W_2^2(\mu_t, \nu_t)}{2} = \mathbb{E}_{\pi_t}\Big[(x - y) \cdot (\xi_t(x) - \tilde{\xi}_t(y))\Big],$$

for a.e. $t \in (t_1, t_2)$.

To apply the theorem above and deduce a differential inequality, we first need to prove the following result.

Lemma 4.8.2. *For $0 < \tau < T_1 < \infty$, we have*

$$\int_0^{T_1-\tau} \mathbb{E}_{q_t}\Big[\|x + \widehat{s_\theta}(t,x)\|^2\Big]dt < \infty, \tag{4.8.3}$$

$$\int_0^{T_1-\tau} \mathbb{E}_{\hat{p}_t}\Big[\|x + \nabla \log \hat{p}_t(x)\|^2\Big]dt < \infty. \tag{4.8.4}$$

Hence, the curves $(q_t)_{t\in[0,T_1-\tau]}$ and $(\hat{p}_t)_{t\in[0,T_1-\tau]}$ are absolutely continuous in $\Big(\mathcal{P}_2(\mathbb{R}^d), W_2\Big)$.

Proof. We start with (4.8.4). Recall first that for $t > 0$,

$$-\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{D}_{\mathrm{KL}}(p_t \,\|\, \gamma) = \mathcal{I}_\gamma(p_t) = \mathbb{E}_{p_t}\left[\left\|\nabla \log \frac{dp_t}{d\gamma}\right\|^2\right] = \mathbb{E}_{p_t}\Big[\|x + \nabla \log p_t(x)\|^2\Big],$$

where, with abuse of notation, we have identified the probability measures p_t, γ with their densities with respect to the Lebesgue measure. Integrating this inequality between τ and T_1 we find

$$\int_\tau^{T_1} \mathbb{E}_{p_t}\Big[\|x + \nabla \log p_t(x)\|^2\Big]dt \le \mathcal{D}_{\mathrm{KL}}(p_\tau \,\|\, \gamma) < \infty,$$

where we used non-negativity of the KL-divergence and that $\mathcal{D}_{\mathrm{KL}}(p_\tau \,\|\, \gamma) < \infty$, cf. [Vil03, Rmk. 9.4]. The conclusion follows by a change of variable in the integral.

As for (4.8.3), we argue as in the proof of [Vil09, Thm. 23.9]. Note first that $q_0 \in \mathcal{P}_2(\mathbb{R}^d)$, since $p_{T_1} \in \mathcal{P}_2(\mathbb{R}^d)$ and $W_2(q_0, p_{T_1}) < \infty$. Let $v_t(x)$ denote the velocity field $v_t(x) = x + \widehat{s_\theta}(t,x)$. Since T_1 is finite, it follows from our assumption (A1) in Section 4.4 that there exists a constant $C > 0$ such that $\|v_t(x)\| \le C(1 + \|x\|)$ for all $x \in \mathbb{R}^s$ and $0 \le t \le T_1$. The Lipschitz assumption on s_θ in our assumption (A1) in Section 4.4 also implies that v is Lipschitz. Therefore, there exists a unique trajectory map $T_t \colon \mathbb{R}^d \to \mathbb{R}^d$ associated to the continuity equation (4.8.1), i.e.,

$$\begin{cases} T_0(x) &= x, \\ \frac{d}{dt}T_t(x) &= v_t(T_t(x)). \end{cases}$$

Then, by the conservation of mass formula [Vil09], we have $q_t = (T_t)_\# q$, where $\#$ denotes the pushforward of a measure by a map. Notice now that, for all $0 \le t \le T_1$,

$$\|T_t(x)\| = \left\|x + \int_0^t v_t(T_t)(x)dt\right\| \le \|x\| + CT_1 + C\int_0^t \|T_t(x)\|dt.$$

Therefore, by the integral version of Gronwall's lemma applied to the continuous function $t \to \|T_t(x)\|$, we deduce that

$$\|T_t(x)\| \le (\|x\| + CT_1)e^{CT_1}.$$

It follows that, for $0 \le t \le T_1$,

$$\int_{\mathbb{R}^d} \|x\|^2 q_t(dx) = \int_{\mathbb{R}^d} \|T_t(x)\|^2 q_0(dx) \le e^{2CT_1}\int_{\mathbb{R}^d} (\|x\| + CT_1)^2 q_0(dx) =: \tilde{C} < \infty,$$

thus the second moment of q_t is uniformly bounded by $\tilde{C}$ for $0 \le t \le T_1$. Replacing $\tilde{C}$ with $\tilde{C} + 1$, we note that also the first moment is uniformly bounded by $\tilde{C}$. Therefore, recalling that $\|x + s_\theta(t,x)\| = \|v_{T_1-t}(x)\| \le C(1 + \|x\|)$, we obtain the following bound, for all $0 \le t \le T_1$,

$$\mathbb{E}_{q_t}\Big[\|x + s_\theta(t,x)\|^2\Big] \le C(1 + \mathbb{E}_{q_t}\Big[\|x\|^2\Big]) \le C(1 + \tilde{C}).$$

This implies the desired estimate (4.8.3).

Finally, the absolute continuity of the curves $(q_t)_{t\in[0,T_1-\tau]}$ and $(\hat{p}_t)_{t\in[0,T_1-\tau]}$ is an immediate consequence of the bounds (4.8.3) and (4.8.4) in view of [AGS08, Thm. 8.3.1]. ∎

Proof of Proposition 4.4.3. Thanks to Lemma 4.8.2, we can apply Theorem 4.8.1. Let π_t be an optimal coupling in W_2 for $\hat{p}_t$ and q_t, so that by definition we have $\mathbb{E}_{\pi_t}[\|x-y\|^2] = W_2^2(\hat{p}_t, q_t)$. Then, we deduce that, for a.e. $t \in [0, T_1 - \tau]$,

$$
\begin{aligned}
&\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t} W_2^2(\hat{p}_t, q_t) \\
&= \mathbb{E}_{\pi_t}[(x-y)\cdot(x-y)] + \mathbb{E}_{\pi_t}[(x-y)\cdot(\nabla \log \hat{p}_t(x) - \widehat{s_\theta}(t,y))] \\
&= W_2^2(\hat{p}_t, q_t) + \mathbb{E}_{\pi_t}[(x-y)\cdot(\widehat{s_\theta}(t,x) - \widehat{s_\theta}(t,y))] + \mathbb{E}_{\pi_t}[(x-y)\cdot(\nabla \log \hat{p}_t(x) - \widehat{s_\theta}(t,x))] \\
&\le W_2^2(\hat{p}_t, q_t) + \widehat{L_s}(t)\mathbb{E}_{\pi_t}\left[\|x-y\|^2\right] + \sqrt{\mathbb{E}_{\pi_t}[\|x-y\|^2]}\cdot\sqrt{\mathbb{E}_{\pi_t}[\|\nabla \log \hat{p}_t(x) - \widehat{s_\theta}(t,x)\|^2]} \\
&= \left(1 + \widehat{L_s}(t)\right) W_2^2(\hat{p}_t, q_t) + \sqrt{\hat{b}(t)}\, W_2(\hat{p}_t, q_t).
\end{aligned}
$$

From this we deduce the differential inequality

$$
\frac{\mathrm{d}}{\mathrm{d}t} W_2(\hat{p}_t, q_t) \le \left(1 + \widehat{L_s}(t)\right) W_2(\hat{p}_t, q_t) + \sqrt{\hat{b}(t)}.
$$

We can solve this differential inequality by introducing the auxiliary function $I(\tau, t) := \exp\left(t - \tau + \int_\tau^t L_s(r)dr\right)$, which satisfies

$$
I(\tau, r)I(r, t) = I(\tau, t), \quad I(t,t) = 1, \quad \text{and} \quad \frac{d}{dt} I(\tau, t) = (1 + L_s(t))I(\tau, t). \tag{4.8.5}
$$

Combining the latter identity with the differential inequality above, we find

$$
\frac{d}{dt}\Big(I(T_1, T_1 - t) W_2(\hat{p}_t, q_t)\Big) \le I(T_1, T_1 - t)\sqrt{\hat{b}(t)},
$$

for a.e. t. Since the curve $t \to I(\tau, t)$ is Lipschitz by (A1) in Section 4.4, it is also absolutely continuous. Moreover, the triangle inequality for W_2 yields

$$
\begin{aligned}
|W_2(\hat{p}_t, q_t) - W_2(\hat{p}_s, q_s)| &\le |W_2(\hat{p}_t, q_t) - W_2(\hat{p}_t, q_s)| + |W_2(\hat{p}_t, q_s) - W_2(\hat{p}_s, q_s)| \\
&\le W_2(q_t, q_s) + W_2(\hat{p}_t, \hat{p}_s).
\end{aligned}
$$

Using the absolute continuity of $\hat{p}_t, q_t$ in $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ (cf. Lemma 4.8.2) we deduce that $t \to W_2(\hat{p}_t, q_t)$ is absolutely continuous too on $[0, T_1 - \tau]$. Therefore, also the function

$$
t \to I(T_1, T_1 - t) W_2(\hat{p}_t, q_t)
$$

is absolutely continuous on $[0, T_1 - \tau]$. Hence, we can apply the second fundamental theorem of calculus for the Lebesgue integral and integrate the differential inequality between 0 and $T_1 - \tau$. Doing this gives

$$
I(T_1, \tau) W_2(p_\tau, q_{T_1-\tau}) \le W_2(p_{T_1}, q_0) + \int_0^{T_1-\tau} I(T_1, T_1 - t)\sqrt{\hat{b}(t)}dt.
$$

Using the properties of I from (4.8.5) we find

$$
W_2(p_\tau, q_{T_1-\tau}) \le I(\tau, T_1) W_2(p_{T_1}, q_0) + \int_0^{T_1-\tau} I(\tau, T_1 - t)\sqrt{\hat{b}(t)}dt,
$$

which yields the desired expression after a change of variables $t' := T_1 - t$. ∎

We now prove Lemma 4.4.4, which gives a well-known Hölder continuity bound in Wasserstein distance for the Ornstein-Uhlenbeck flow.

Proof of Lemma 4.4.4. Let X be a random variable with law p, and let be Z be a standard Gaussian random variable that is independent of X. Then $X_\tau := e^{-\tau}X + \sqrt{1-e^{-2\tau}}Z$ has law p_τ. Using independence, we obtain

$$\begin{aligned} W_2^2(p, p_\tau) \le \mathbb{E}\Big[|X - X_\tau|^2\Big] &= \mathbb{E}\Big[|(1-e^{-\tau})X - \sqrt{1-e^{-2\tau}}Z|^2\Big] \\ &= (1-e^{-\tau})^2\mathbb{E}\Big[|X|^2\Big] + (1-e^{-2\tau})\mathbb{E}\Big[|Z|^2\Big] \le \tau^2 M^2 + 2\tau d, \end{aligned}$$

which implies the result. ∎

As discussed in Section 4.4.1, to establish fast convergence of the approximate Langevin dynamics in (4.4.1) we need a quantitative estimate for the log-Sobolev constant of p_{T_1}, which is provided in the next result.

Lemma 4.8.3. *Let $(p_t)_{t\ge 0}$ be the law of the Ornstein–Uhlenbeck flow starting from a norm-subgaussian random vector X. Then, p_t satisfies a log-Sobolev inequality with constant*

$$C_{\mathrm{LS}}(p_t) \ge \frac{1}{1 + 172\|X\|_{\mathrm{SG}}^2 e^{-2t}},$$

for all $t > t_0 := \frac{1}{2}\log(1 + 4\|X\|_{\mathrm{SG}}^2)$.

Consequently, for any $\delta \in (0,1)$, the log-Sobolev constant of p_t satisfies $C_{\mathrm{LS}}(p_t) \ge 1-\delta$ whenever $t \ge \max\Big(t_0, \frac{1}{2}\log(172\|X\|_{\mathrm{SG}}^2/\delta)\Big)$.

The proof is based on the following recent result from [CCNW21, Thm. 2], which gives an estimate for the log-Sobolev constant of Gaussian convolutions of sub-Gaussian distributions.

Theorem 4.8.4. *Let μ be a probability measure and $\sigma, C_{\mathrm{SG}} > 0$ be such that*

$$\int\int e^{\frac{\|x-x'\|^2}{\sigma^2}}\mu(dx)\mu(dx') \le C_{\mathrm{SG}}. \tag{4.8.6}$$

For all $t > \sigma^2$, the measure $\mu * \gamma_t$ satisfies a log-Sobolev inequality with the constant

$$C_{\mathrm{LS}}(\mu * \gamma_t) \ge \left(3t\left[\frac{t}{t-\sigma^2} + C_{\mathrm{SG}}^{\frac{\sigma^2}{t-\sigma^2}}\right]\left[1 + \frac{\sigma^2}{t-\sigma^2}\log C_{\mathrm{SG}}\right]\right)^{-1}. \tag{4.8.7}$$

Proof of Lemma 4.8.3. Note that p_t is the law of $e^{-t}X + \sqrt{1-e^{-2t}}Z$, with $X \sim p$ and $Z \sim \gamma$ independent. Consequently, $p_t = \mu_t * \gamma_{1-e^{-2t}}$, where μ_t denotes the law of $e^{-t}X$. Suppose now that $1 + 4\|X\|_{\mathrm{SG}}^2 < e^{2t}$ and define $\sigma := \sqrt{2}e^{-t}\|X\|_{\mathrm{SG}}$. Since $1 - e^{-2t} - 2\sigma^2 > 0$, we may proceed as in [CCNW21, Rmk. 3] and write

$$p_t = \mu_t * \gamma_{2\sigma^2} * \gamma_{1-e^{-2t}-2\sigma^2}. \tag{4.8.8}$$

We claim that μ_t satisfies the assumption of Theorem 4.8.4 with $C_{\mathrm{SG}} = 4$ and σ as defined above. Indeed,

$$\begin{aligned} \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} e^{\frac{\|x-x'\|^2}{\sigma^2}}\mu_t(dx)\mu_t(dx') &\le \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} e^{2\frac{\|x\|^2+\|x'\|^2}{\sigma^2}}\mu_t(dx)\mu_t(dx') \\ &= \Big(\int_{\mathbb{R}^d} e^{\frac{2\|x\|^2}{\sigma^2}}\mu_t(dx)\Big)^2 \le 4, \end{aligned} \tag{4.8.9}$$

where the last step uses our definition of σ. Therefore, an application of Theorem 4.8.4 yields

$$C_{\mathrm{LS}}(\mu_t * \gamma_{2\sigma^2}) \geq \Big[6\sigma^2(2+C)(1+\log C)\Big]^{-1} \geq \Big[86\sigma^2\Big]^{-1}.$$

Using the subadditivity of C_{LS}^{-1} under convolution (cf. [WW16, Prop. 1.1]) and the estimate $C_{\mathrm{LS}}(\gamma_{1-e^{-2t}-2\sigma^2}) \geq C_{\mathrm{LS}}(\gamma) = 1$ from Lemma 4.7.1, we obtain using (4.8.8),

$$C_{\mathrm{LS}}(p_t) \geq \left[\frac{1}{C_{\mathrm{LS}}(\mu_t * \gamma_{2\sigma^2})} + \frac{1}{C_{\mathrm{LS}}(\gamma_{1-e^{-2t}-2\sigma^2})}\right]^{-1} \geq \Big[86\sigma^2+1\Big]^{-1},$$

which proves the first part of the statement. The second part follows immediately. ∎

Proof of Lemma 4.4.8. We proceed in two steps. Suppose first that $p = \delta_x$ for some $x \in \mathbb{R}^d$. Then $p_t = \gamma_{e^{-t}x,\sigma_t}$ is Gaussian with $\sigma_t := 1 - e^{-2t}$. An explicit calculation gives

$$\mathcal{D}_{\mathrm{KL}}(\gamma \,\|\, \gamma_{e^{-t}x,\sigma_t}) = \frac{d}{2\sigma_t}\left(e^{-2t}\frac{\|x\|^2}{d} + \sigma_t \log \sigma_t - \sigma_t + 1\right). \tag{4.8.10}$$

In the general case where $p \in \mathbb{R}^d$ has finite second moment, we condition on the initial value using the disintegration formula

$$p_t(dy) = \int_{\mathbb{R}^d} \gamma_{e^{-t}x,\sigma_t}(dy)\, p(dx).$$

Using this formula, we employ the joint convexity of the KL-divergence and (4.8.10) to obtain

$$\begin{aligned}\mathcal{D}_{\mathrm{KL}}(\gamma \,\|\, p_{T_1}) &= \mathcal{D}_{\mathrm{KL}}\Big(\gamma \,\Big\|\, \int_{\mathbb{R}^d} \gamma_{e^{-t}x,\sigma_t}\, dp(x)\Big) = \mathcal{D}_{\mathrm{KL}}\Big(\int_{\mathbb{R}^d} \gamma\, dp(x) \,\Big\|\, \int_{\mathbb{R}^d} \gamma_{e^{-t}x,\sigma_t}\, dp(x)\Big) \\ &\leq \int_{\mathbb{R}^d} \mathcal{D}_{\mathrm{KL}}(\gamma \,\|\, \gamma_{e^{-t}x,\sigma_t})\, dp(x) = \frac{d}{2\sigma_t}\left(\frac{M_2(p)}{d}e^{-2t} + \sigma_t \log \sigma_t - \sigma_t + 1\right),\end{aligned}$$

where $M_2(p) := \int \|x\|^2\, dp(x)$.

For the second claim, we use the scalar inequalities $r \log r - r + 1 \leq (r-1)^2$ for $r \geq 0$. Thus, whenever $e^{-2t} \leq \frac{1}{2}$, we have

$$\mathcal{D}_{\mathrm{KL}}(\gamma \,\|\, p_{T_1}) \leq \frac{d}{2(1-e^{-2t})}\left(\frac{M_2(p)}{d}e^{-2t} + e^{-4t}\right) \leq \left(M_2(p) + \frac{d}{2}\right)e^{-2t}.$$

This implies the desired result. ∎

We are now ready to prove Theorem 4.4.1.

Proof of Theorem 4.4.1. Note first that, by the triangle inequality for W_2, we have

$$W_2(p, p_\theta) = W_2(p, q_{T_1-\tau}) \leq W_2(p, p_\tau) + W_2(p_\tau, q_{T_1-\tau}).$$

The first term can be estimated with Lemma 4.4.4, the second with Proposition 4.4.3. Plugging in these estimates gives

$$W_2(p, p_\theta) \leq \sqrt{3d\tau} + I_\tau(T_1) W_2(p_{T_1}, q_0) + \int_\tau^{T_1} I_\tau(t)\sqrt{b(t)}dt.$$

Therefore, to prove (4.4.3), it suffices to show that

$$W_2(p_{T_1}, q_0) \le \sqrt{\frac{2}{1-\delta}\Big(\delta e^{-\frac{(1-\delta)T_2}{2}} + 2\varepsilon_{\mathrm{MGF}}\Big)}.$$

To see this, observe that $C_{\mathrm{LS}}(p_{T_1}) \ge 1-\delta$ by Lemma 4.8.3: therefore, we can combine (4.4.12) with Theorem 4.4.7 to deduce that

$$W_2(p_{T_1}, q_0) \le \sqrt{\frac{2}{1-\delta}\Big(\mathcal{D}_{\mathrm{KL}}(\gamma \,\|\, p_{T_1}) e^{-\frac{(1-\delta)T_2}{2}} + 2\varepsilon_{\mathrm{MGF}}\Big)}.$$

Recalling that $\mathbb{E}[\|X\|^2] \le 2\|X\|_{\mathrm{SG}}^2$ by (i) of Lemma 4.7.3, we can use the estimate $\mathcal{D}_{\mathrm{KL}}(\gamma \,\|\, p_{T_1})$ from Lemma 4.4.8 to prove (4.4.3); an application of Cauchy–Schwarz inequality then gives (4.4.4). Finally, if we only know $T_1 \ge \frac{1}{2}\log\Big(2 + 172\frac{\|X\|_{\mathrm{SG}}^2}{\delta}\Big)$, then we can instead estimate $\mathcal{D}_{\mathrm{KL}}(p_{T_1} \,\|\, \gamma) \le \frac{d}{3}$, again by Lemma 4.4.8, which proves our claim in the discussion on the role of T_1 after Theorem 4.4.1. ∎

4.9 Proof of Theorem 4.4.2

When starting an Ornstein–Uhlenbeck flow from a norm-subgaussian distribution, the distribution at time T_1 will be norm-subgaussian too, and we can estimate its norm.

Lemma 4.9.1. *Let $(X_t)_{t\ge 0}$ be an Ornstein–Uhlenbeck process* (4.3.6) *starting from a norm-subgaussian random vector X. Then, if $T_1 \ge \log\frac{\|X\|_{\mathrm{SG}}}{\sqrt{d}}$, we have*

$$\|X_{T_1}\|_{\mathrm{SG}} \le 3\sqrt{d}. \tag{4.9.1}$$

Proof. Let $Z \sim \gamma$ be independent of X. Then, X_{T_1} is equal in law to $e^{-T_1}X + \sqrt{1-e^{-2T_1}}Z$. Consequently,

$$\|X_{T_1}\|_{\mathrm{SG}} = \Big\|e^{-T_1}X + \sqrt{1-e^{-2T_1}}Z\Big\|_{\mathrm{SG}} \le e^{-T_1}\|X\|_{\mathrm{SG}} + \|Z\|_{\mathrm{SG}} \le 3\sqrt{d},$$

where in the last inequality we use Lemma 4.7.1 and the choice of T_1. ∎

The following lemma gives an a priori estimate for $\nabla \log p_{T_1}$ which can be used to correct predictions of $s_\theta(T_1, \cdot)$ that are far from the ground-truth.

Lemma 4.9.2. *Let $(p_t)_{t\ge 0}$ be the law of the Ornstein–Uhlenbeck flow starting from a norm-subgaussian random vector X with law p. Fix $0<\delta<1$ and take $T_1 \ge \log\Big(\frac{16}{\delta}d(\|X\|_{\mathrm{SG}}+1)\Big)$. Then, for all $x \in \mathbb{R}^d$ and $i \in \{1,\dots,d\}$, we have*

$$|-x_i - \partial_i \log p_{T_1}(x)| \le \frac{\delta}{2d}(1+\|x\|).$$

Proof. Let μ be the law of $e^{-T_1}X$ and set $\sigma^2 = 1-e^{-2T_1}$. By our choice of T_1, we have

$$\Big\|e^{-T_1}X\Big\|_{\mathrm{SG}} \le \frac{\delta}{16d} \quad \text{and} \quad e^{-2T_1} = 1-\sigma^2 \le \left(\frac{\delta}{16d}\right)^2. \tag{4.9.2}$$

Notice that $p_{T_1} = \mu * \gamma_{\sigma^2}$. Therefore, for all $x \in \mathbb{R}^d$, we can write as in [BGMZ18]

$$p_{T_1}(x) = \int_{\mathbb{R}^d} (2\pi\sigma^2)^{-\frac{d}{2}} \exp\left(-\frac{\|x-z\|^2}{2\sigma^2}\right)\mu(dz) = (2\pi\sigma^2)^{-\frac{d}{2}} \exp\left(-\left(\frac{\|x\|^2}{2\sigma^2} + W_\sigma(x)\right)\right),$$

where we set

$$W_\sigma(x) = -\log \int_{\mathbb{R}^d} \exp\left(\frac{x\cdot z}{\sigma^2} - \frac{\|z\|^2}{2\sigma^2}\right)\mu(dz).$$

Taking the logarithm and differentiating, we find that

$$\partial_i \log p_{T_1}(x) = -\frac{1}{\sigma^2}x_i - \partial_i W_\sigma(x). \tag{4.9.3}$$

Observe now that

$$\left|\sigma^2 \partial_i W_\sigma(x)\right| \le \frac{\int_{\mathbb{R}^d} |z_i| \exp\left(\frac{x\cdot z}{\sigma^2} - \frac{\|z\|^2}{2\sigma^2}\right)\mu(dz)}{\int_{\mathbb{R}^d} \exp\left(\frac{x\cdot z}{\sigma^2} - \frac{\|z\|^2}{2\sigma^2}\right)\mu(dz)} = \frac{\int_{\mathbb{R}^d} |z_i| \gamma_{x,\sigma^2}(z)\mu(dz)}{\int_{\mathbb{R}^d} \gamma_{x,\sigma^2}(z)\mu(dz)}, \tag{4.9.4}$$

where, with some abuse of notation, $\gamma_{x,t}$ denotes the density of a gaussian $\mathcal{N}(x, tI_d)$. We claim that

$$\int_{\mathbb{R}^d} |z_i| \gamma_{x,\sigma^2}(z)\mu(dz) \le \frac{\delta(1+\|x\|)}{4d} \int_{\mathbb{R}^d} \gamma_{x,\sigma^2}(z)\mu(dz), \tag{4.9.5}$$

which we prove later. Using this bound in (4.9.4) we deduce that

$$|\partial_i W_\sigma(x)| \le \frac{\delta(1+\|x\|)}{4d\sigma^2}.$$

Inserting this estimate in (4.9.3), it follows using (4.9.2) that

$$\begin{aligned}|-\partial_i \log p_{T_1}(x) - x_i| &\le \frac{1-\sigma^2}{\sigma^2}|x_i| + \frac{\delta(1+\|x\|)}{4d\sigma^2} \le \frac{1+\|x\|}{\sigma^2}\left[\left(\frac{\delta}{16d}\right)^2 + \frac{\delta}{4d}\right] \\ &\le \frac{\delta}{2d}(1+\|x\|),\end{aligned}$$

where in the last inequality we used that

$$\frac{1}{\sigma^2}\left[\left(\frac{\delta}{16d}\right)^2 + \frac{\delta}{4d}\right] \le \frac{1}{1-e^{-2T_1}}\left[\frac{1}{256} + \frac{1}{4}\right]\frac{\delta}{d} \le \frac{1}{1-1/256}\frac{\delta}{3d} \le \frac{\delta}{2d},$$

since $0 < \delta < 1$ and $T_1 \ge \log(16)$. This the desired estimate.

It remains to prove (4.9.5). To do so, we start by writing

$$\begin{aligned}\int_{\mathbb{R}^d} \gamma_{x,\sigma^2}(z)\mu(dz) &\ge \int_{B(0,\delta)} \gamma_{x,\sigma^2}(z)\mu(dz) \\ &\ge (2\pi\sigma^2)^{-\frac{d}{2}} \exp\left(-\frac{(\|x\|+\delta)^2}{2\sigma^2}\right)\mu(B(0,\delta)) \\ &\ge (2\pi\sigma^2)^{-\frac{d}{2}} \exp\left(-\frac{(\|x\|+\delta)^2}{2\sigma^2}\right)[1-2\exp(-256)],\end{aligned} \tag{4.9.6}$$

where in the last step we use (i) of Lemma 4.7.3 and the estimate $\left\|e^{-T_1}X\right\|_{\mathrm{SG}} \le \frac{\delta}{16d} \le \frac{\delta}{16}$, which holds in view of (4.9.2). We now split the integral in the left-hand side of (4.9.5) into two terms that we will estimate separately. Set $r = \frac{\delta(1+\|x\|)}{8d}$. Then,

$$\int_{B(0,r)} |z_i| \gamma_{x,\sigma^2}(z)\mu(dz) \le r \int_{B(0,r)} \gamma_{x,\sigma^2}(z)\mu(dz) \le \frac{\delta(1+\|x\|)}{8d} \int_{\mathbb{R}^d} \gamma_{x,\sigma^2}(z)\mu(dz).$$

Therefore, recalling (4.9.6), to conclude the proof of (4.9.5) it is enough to show that

$$\int_{B(0,r)^c} |z_i| \gamma_{x,\sigma^2}(z)\mu(dz) \le \frac{\delta(1+\|x\|)}{8d}[1 - 2\exp(-256)](2\pi\sigma^2)^{-\frac{d}{2}} \exp\left(-\frac{(\|x\|+\delta)^2}{2\sigma^2}\right).$$

To do this, we write

$$\begin{aligned}
(2\pi\sigma^2)^{\frac{d}{2}} \int_{B(0,r)^c} |z_i| \gamma_{x,\sigma^2}(z)\mu(dz) &\le \int_{B(0,r)^c} \|z\| \mu(dz) \\
&\le \left(2r + \frac{\delta^2}{256 d^2 r}\right) \exp\left(-\frac{256 d^2 r^2}{\delta^2}\right) \\
&= \left(\frac{2\delta}{8d}(1+\|x\|) + \frac{\delta}{32d}\frac{1}{1+\|x\|}\right) \exp\left(-4(1+\|x\|)^2\right) \\
&\le \frac{3\delta}{8d}(1+\|x\|) \exp\left(-4(1+\|x\|)^2\right),
\end{aligned}$$

where in the second line we have used (iii) of Lemma 4.7.3. As $\sigma^2 \ge \frac{1}{2}$ by (4.9.2), $\delta < 1$, and $3\exp(-4) \le [1 - 2\exp(-256)]$, we have

$$3\exp\left(-4(1+\|x\|)^2\right) \le [1 - 2\exp(-256)] \exp\left(-\frac{(\|x\|+\delta)^2}{2\sigma^2}\right)$$

for all x, which concludes the proof. ∎

We are now ready to move to the proof of Theorem 4.4.2.

Proof of Theorem 4.4.2. Notice that, thanks to Lemma 4.9.2 and to our definition of $\widetilde{s_\theta}$, for all $x \in \mathbb{R}^d$ we have

$$\|\nabla \log p_{T_1}(x) - \widetilde{s_\theta}(x)\|^2 \le \frac{\delta^2}{d}(1+\|x\|)^2, \tag{4.9.7}$$

$$\|\nabla \log p_{T_1}(x) - \widetilde{s_\theta}(x)\| \le \|\nabla \log p_{T_1}(x) - s_\theta(T_1, x)\|. \tag{4.9.8}$$

As $\log(1+\varepsilon) \le \varepsilon$, it suffices to show that

$$\int_{\mathbb{R}^d} \exp\left(\beta\|\nabla \log p_{T_1}(x) - \widetilde{s_\theta}(x)\|^2\right) dp_{T_1}(x) \le 1 + \varepsilon.$$

Now let us fix a radius $R > 0$, whose value we specify later. We will show that, for an appropriate choice of $R > 0$,

$$\int_{B(0,R)^c} \exp\left(\beta\|\nabla \log p_{T_1}(x) - \widetilde{s_\theta}(x)\|^2\right) dp_{T_1}(x) \le \frac{\varepsilon}{2}, \tag{4.9.9}$$

and

$$\int_{B(0,R)} \exp\left(\beta\|\nabla \log p_{T_1}(x) - \widetilde{s_\theta}(x)\|^2\right) dp_{T_1}(x) \le 1 + \frac{\varepsilon}{2}, \tag{4.9.10}$$

thus concluding the proof.

First, we consider (4.9.9). Notice that

$$\begin{aligned}\exp\Big(\beta\|\nabla\log p_{T_1}(x)-\widetilde{s_\theta}(x)\|^2\Big) &\leq \exp\bigg(\beta\frac{\delta^2}{d}(1+\|x\|)^2\bigg)\\ &\leq \exp\bigg(\frac{2\beta\delta^2}{d}\bigg)\exp\bigg(\frac{2\beta\delta^2}{d}\|x\|^2\bigg)\\ &\leq \bigg(1+\frac{4\beta\delta^2}{d}\bigg)\exp\bigg(\frac{2\beta\delta^2}{d}\|x\|^2\bigg)\\ &\leq 2\exp\Big(\frac{1}{18d}\|x\|^2\Big).\end{aligned}$$

Here, for the first inequality we use (4.9.7); for the third inequality, we use that $e^s \leq 1+2s$ for $0\leq s\leq 1$ and the condition on β; the condition on β is used again for the last inequality. Therefore, we deduce that

$$\begin{aligned}\int_{B(0,R)^c}\exp\Big(\beta\|\nabla\log p_{T_1}(x)-\widetilde{s_\theta}(x)\|^2\Big)\,dp_{T_1}(x) &\leq 2\int_{B(0,R)^c}\exp\Big(\frac{1}{18d}\|x\|^2\Big)\,dp_{T_1}(x)\\ &\leq 8e^{-\frac{R^2}{18d}},\end{aligned}$$

where for the last inequality we use (ii) of Lemma 4.7.3 and Lemma 4.9.1. Therefore, picking

$$R=\sqrt{18d\log\frac{16}{\varepsilon}}$$

readily gives (4.9.9). With this choice of R, we now consider (4.9.10). Let us define the random variable

$$F=\beta\|\nabla\log p_{T_1}(X_{T_1})-\widetilde{s_\theta}(X_{T_1})\|^2\mathbb{1}_{\big\{\|X_{T_1}\|\leq R\big\}},$$

where $X_{T_1}\sim p_{T_1}$ as usual. We note that

$$\int_{B(0,R)}\exp\Big(\beta\|\nabla\log p_{T_1}-\widetilde{s_\theta}\|^2\Big)p_{T_1}(x)dx\leq\mathbb{E}\Big[e^F\Big].$$

It remains to estimate $\mathbb{E}\Big[e^F\Big]$, which we will do using Lemma 4.7.4. To show that F satisfies the conditions of Lemma 4.7.4, we check its boundedness. Using (4.9.7) and the constraint on β we obtain

$$0\leq F\leq\beta\frac{\delta^2}{d}(1+R)^2\leq\frac{1}{18d}+36\beta\delta^2\log\frac{16}{\varepsilon}=:M.$$

Furthermore, using (4.9.8) and (4.4.7) we can estimate $\mathbb{E}[F]$ by

$$\mathbb{E}[F]\leq\beta\mathbb{E}_{p_{T_1}}\Big[\|\nabla\log p_{T_1}-s_\theta(T_1,\cdot)\|^2\Big]\leq\frac{1}{34}\varepsilon^{1+36\beta\delta^2}.$$

Notice also that

$$e^M=e^{\frac{1}{18d}}16^{36\beta\delta^2}\varepsilon^{-36\beta\delta^2}\leq 16e^{\frac{1}{18}}\varepsilon^{-36\beta\delta^2},$$

where we used once more the constraint on β. We can now apply Lemma 4.7.4 to deduce that

$$\mathbb{E}\Big[e^F\Big]-1\leq e^M\mathbb{E}[F]\leq\frac{16e^{\frac{1}{18}}}{34}\varepsilon\leq\frac{\varepsilon}{2},$$

which concludes the proof. ∎

4.10 Proof of Theorem 4.5.1

The first step of the argument in the proof of Theorem 4.5.1 consists in giving an upper bound for $\|p - p_\theta\|_{\mathsf{TV}}$ which allows to control separately the errors originating from *(i)* taking Z_{N_2} as an approximate sample from p_{T_1}, and *(ii)* approximating the reverse process with a discretized scheme and with an L^2 accurate score. To do so, we follow the strategy of [CCL+23b, CLL22]. Let us denote by S the Markov kernel which associates to a probability measure μ the law of the random variable U_{T_1}, where $(U_t)_t$ satisfies the true backward SDE initialised at μ, *i.e.*,

$$\begin{cases} U_0 \sim \mu, \\ dU_t = U_t dt + 2\nabla \log p_{T_1 - t}(U_t)dt + \sqrt{2}dB_t. \end{cases}$$

In particular, we have $p = p_{T_1} S$. Similarly, we denote by $\hat{S}$ the Markov kernel which corresponds to following the approximate reverse process in (4.5.2) initialised at μ. In particular, we have $p_\theta = \sigma_{N_2} \hat{S}$ (recall that $\sigma_k = \mathrm{law}(Z_k)$). The following chain of inequalities holds:

$$\begin{aligned} \|p - p_\theta\|_{\mathsf{TV}} &= \|p_{T_1} S - \sigma_{N_2} \hat{S}\|_{\mathsf{TV}} \\ &\leq \|p_{T_1} S - p_{T_1} \hat{S}\|_{\mathsf{TV}} + \|p_{T_1} \hat{S} - \sigma_{N_2} \hat{S}\|_{\mathsf{TV}} \\ &\lesssim \sqrt{\mathcal{D}_{\mathrm{KL}}\Big(p_{T_1} S \,\Big\|\, p_{T_1} \hat{S}\Big)} + \|p_{T_1}, \sigma_{N_2}\|_{\mathsf{TV}} \\ &\lesssim \sqrt{\mathcal{D}_{\mathrm{KL}}\Big(p_{T_1} S \,\Big\|\, p_{T_1} \hat{S}\Big)} + \sqrt{\mathcal{D}_{\mathrm{KL}}(\sigma_{N_2} \,\|\, p_{T_1})}. \end{aligned} \tag{4.10.1}$$

In the above, we have used the triangle inequality for the total variation distance, Pinsker's inequality and the data-processing inequality. This achieves the desired decomposition, so that we can study the two processes (4.5.1), (4.5.2) separately.

Convergence of inexact Langevin algorithm.

To control the error term $\mathcal{D}_{\mathrm{KL}}(\sigma_{N_2} \,\|\, p_{T_1})$, we need to study convergence of the process (4.5.1). This is done by adapting the results of [YYW23] to the case of a decaying step size. In particular, we prove the following

Proposition 4.10.1. *Suppose that Assumption (A2) holds and pick* $0 < \delta < \frac{1}{2}$ *and* $T_1 \geq \frac{1}{2} \log\Big(2 + 172 \frac{\|X\|_{\mathrm{SG}}^2}{\delta} + \frac{d}{2\delta}\Big)$. *Assume in addition that* $\nabla \log p_{T_1}$ *is* L_1*-Lipschitz and that* $s_\theta(T_1, \cdot)$ *is* L_2*-Lipschitz, with* $L_1, L_2 \geq 1$. *Then, for the inexact Langevin algorithm* (4.5.1) *with step sizes* $h_k = \frac{1}{24 L_1 L_2 + \frac{k+1}{16}}$, *we have that*

$$\mathcal{D}_{\mathrm{KL}}(\sigma_{N_2} \,\|\, p_{T_1}) \lesssim \left(\frac{L_1 L_2}{N_2 + 1}\right)^2 + \frac{d L_2^2}{N_2 + 1} + \tilde{\varepsilon}_{\mathrm{MGF}}, \tag{4.10.2}$$

where $\tilde{\varepsilon}_{\mathrm{MGF}}$ *is defined in* (4.5.3).

The proof is postponed to the end of this section. Using Proposition 4.10.1 gives the desired upper bound for the second error term in (4.10.1), when using a decaying step size. For the analogous result with a constant step size, see [YYW23, Thm. 2].

Analysis of the reverse process.

It remains to give an upper bound for the error due to the discretization and approximation of the score in the reverse process, corresponding to the first term in the right hand side of (4.10.1). This has been analysed in a number of recent works, under different assumptions. We recall in particular the following results, proved in [CLL22] building on the Girsanov framework developed in [CCL+23b].

Lemma 4.10.2. *Suppose that p has finite second moment and that $\nabla \log p_t$ is L_1-Lipschitz for $t \in [0, T_1]$ with $T_1, L_1 \geq 1$. Then, choosing a constant step size $\tilde{h}_k = \frac{T_1}{N_1} \leq 1$ gives*

$$\mathcal{D}_{\mathrm{KL}}\Big(p \,\Big\|\, p_{T_1}\hat{S}\Big) = \mathcal{D}_{\mathrm{KL}}\Big(p_{T_1}S \,\Big\|\, p_{T_1}\hat{S}\Big) \lesssim \widehat{J_{SM}} + \frac{dL^2T_1^2}{N_1}.$$

Inserting this bound in (4.10.1) concludes the proof of Theorem 4.5.1.

Proof of Proposition 4.10.1

The proof of Proposition 4.10.1 is based on the results of [YYW23], and in particular on Lemma 6 therein, which upper bounds the relative entropy after one step of the inexact Langevin algorithm and which we recall below.

Lemma 4.10.3 (Lemma 6 of [YYW23])**.** *Let μ, ν_0 be probability measures with full support that admit densities with respect to the Lebesgue measure. Suppose that μ satisfies a log-Sobolev inequality and let $0 < \kappa \leq C_{\mathrm{LS}}(\mu)$. In addition, suppose that s_θ is an approximation of $\nabla \log \mu$, that s_θ is L_2-Lipschitz and that $\nabla \log \mu$ is L_1-Lipschitz with $L_1, L_2 \geq 1$. Set $Z_0 \sim \nu_0$ and*

$$Z_1 = Z_0 + hs_\theta(Z_0) + \sqrt{2h}B$$

where $B \sim \mathcal{N}(0, I_d)$ is independent of Z_0 and $0 < h < \min\Big\{\frac{\kappa}{12L_1L_2}, \frac{1}{2\kappa}\Big\}$. Then, letting $\nu_1 = \mathrm{law}(Z_1)$, we have

$$\mathcal{D}_{\mathrm{KL}}(\nu_1 \,\|\, \mu) \leq e^{-\frac{1}{4}\kappa h}\mathcal{D}_{\mathrm{KL}}(\nu_0 \,\|\, \mu) + 144L_2^2L_1dh^3 + 24dL_2^2h^2 + h\frac{\kappa}{2}\log \mathbb{E}_\mu\Big[e^{\frac{9}{\kappa}\|s_\theta - \nabla \log \mu\|^2}\Big].$$

With this lemma at hand, we can prove Proposition 4.10.1.

Proof of Proposition 4.10.1. By the choice of T_1, we have that $C_{\mathrm{LS}}(p_{T_1}) \geq 1 - \delta \geq \frac{1}{2}$. In particular, the step sizes $(h_k)_{k=0}^{N_2-1}$ defined by

$$h_k = \frac{1}{24L_1L_2 + \frac{k+1}{16}}$$

satisfy the constraint in Lemma 4.10.3 with $\mu = p_{T_1}$ and $s_\theta = s_\theta(T_1, \cdot)$. We can therefore apply the lemma to deduce that, after each step of the inexact Langevin algorithm,

$$\mathcal{D}_{\mathrm{KL}}(\sigma_{k+1} \,\|\, p_{T_1}) \leq e^{-\frac{1}{8}h_k}\mathcal{D}_{\mathrm{KL}}(\sigma_k \,\|\, \mu) + 30dL_2^2h_k^2 + h_k\tilde{\varepsilon}_{\mathrm{MGF}},$$

for $0 \leq k \leq N_2 - 1$. By iterating the above result we find that

$$\mathcal{D}_{\mathrm{KL}}(\sigma_{N_2} \,\|\, p_{T_1}) \leq e^{-\frac{1}{8}\sum_{i=0}^{N_2-1} h_i} + \sum_{i=0}^{N_2-1}\Big\{\Big(30dL_2^2h_i^2 + h_i\tilde{\varepsilon}_{\mathrm{MGF}}\Big) \cdot e^{-\frac{1}{8}\sum_{j=i+1}^{N_2-1} h_j}\Big\}, \quad (4.10.3)$$

where we have also used that $\mathcal{D}_{\mathrm{KL}}(\sigma_0 \,\|\, p_{T_1}) \le \delta \le 1$ by Lemma 4.4.8.

Now, notice that, for $0 \le j \le k$, we have the bound

$$\sum_{i=j}^{k} h_i \ge \int_j^{k+1} \frac{1}{24L_1L_2 + \frac{x+1}{16}} dx = 16 \log \frac{24L_1L_2 + \frac{k+2}{16}}{24L_1L_2 + \frac{j+1}{16}} \tag{4.10.4}$$

and

$$1 > \frac{h_{k+1}}{h_k} \ge \frac{h_1}{h_0} = \frac{24L_1L_2 + \frac{1}{16}}{24L_1L_2 + \frac{1}{8}} \ge \frac{1}{2}.$$

Using this in (4.10.3), we find that

$$\begin{aligned}
&\mathcal{D}_{\mathrm{KL}}(\sigma_{N_2} \,\|\, p_{T_1}) \\
&\le \left(\frac{24L_1L_2 + \frac{1}{16}}{24L_1L_2 + \frac{N_2+1}{16}}\right)^2 + \sum_{i=0}^{N_2-1} \left\{ \left(30dL_2^2 h_i^2 + h_i \tilde{\varepsilon}_{\mathrm{MGF}}\right) \cdot \left(\frac{24L_1L_2 + \frac{i+2}{16}}{24L_1L_2 + \frac{N_2+1}{16}}\right)^2 \right\} \\
&\le \left(\frac{2}{1 + \frac{N_2}{384L_1L_2}}\right)^2 + 30dL_2^2 \sum_{i=0}^{N_2-1} h_i^2 \frac{h_{N_2}^2}{h_{i+1}^2} + \tilde{\varepsilon}_{\mathrm{MGF}} \sum_{i=0}^{N_2-1} h_i \cdot \frac{h_{N_2}}{h_{i+1}} \\
&\le \left(\frac{2}{1 + \frac{N_2}{384L_1L_2}}\right)^2 + 120dL_2^2 N_2 \cdot \left(\frac{1}{24L_1L_2 + \frac{N_2}{16}}\right)^2 + 2\tilde{\varepsilon}_{\mathrm{MGF}} \cdot \frac{N_2}{24L_1L_2 + \frac{N_2}{16}} \\
&\le \left(\frac{2}{1 + \frac{N_2}{384L_1L_2}}\right)^2 + 30720 \frac{dL_2^2}{N_2 + 1} + 32\tilde{\varepsilon}_{\mathrm{MGF}} \\
&\lesssim \left[\left(\frac{L_1L_2}{N_2+1}\right)^2 + \frac{dL_2^2}{N_2+1} + \tilde{\varepsilon}_{\mathrm{MGF}}\right].
\end{aligned} \tag{4.10.5}$$

This concludes the proof of the proposition. ■

CHAPTER 5

Heat flow, log-concavity, and Lipschitz transport maps

This chapter corresponds to the preprint [BP24].

In this paper we derive estimates for the Hessian of the logarithm (log-Hessian) for solutions to the heat equation. For initial data in the form of log-Lipschitz perturbation of strongly log-concave measures, the log-Hessian admits an explicit, uniform (in space) lower bound. This yields a new estimate for the Lipschitz constant of a transport map pushing forward the standard Gaussian to a measure in this class. Further connections are discussed with score-based diffusion models and improved Gaussian logarithmic Sobolev inequalities. Finally, we show that assuming only fast decay of the tails of the initial datum does not suffice to guarantee uniform log-Hessian upper bounds.

5.1 Introduction

Let $d \geq 1$. We say that a function $V\colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is α-convex, and that a probability density $\mu \in L^1_+(\mathbb{R}^d)$ is α-log-concave, if, respectively, $x \to V(x) - \frac{\alpha}{2}\|x\|^2$ is convex, and $\mu(x) = e^{-V(x)}$ for some α-convex function such that $\int_{\mathbb{R}^d} e^{-V(x)} dx = 1$. In case $\alpha = 0$, μ is a log-concave probability density; if $\alpha > 0$, μ is strongly log-concave. We also consider the *heat flow* over $\mathbb{R}^d$:

$$\begin{cases} \partial_t f = \frac{1}{2}\Delta f, \\ \lim_{t\to 0} f(t,\cdot) = \mu. \end{cases} \tag{5.1.1}$$

Taking $\mu = \delta_0$, the Dirac delta centered in zero, then the *fundamental solution* to (5.1.1) is

$$f(t,x) = \gamma_t(x) := (2\pi t)^{-d/2}\, \mathrm{e}^{-\|x\|^2/2t},$$

where γ_t is the isotropic Gaussian density with zero mean and covariance matrix equal to tI_d. Any other solution to (5.1.1) is then given by $\mu * \gamma_t$, where $*$ is the symbol of convolution: $(g_1 * g_2)(x) = \int_{\mathbb{R}^d} g_1(x-y)\, g_2(y)\, dy$. Denote by $(P_t)_t$ the corresponding heat semigroup, i.e.

$$P_t\mu := \mu * \gamma_t, \quad t > 0, \tag{5.1.2}$$

which is induced by the flow of (5.1.1). As solutions to (5.1.1) are Gaussian convolutions of the initial datum μ, it is expected that those would inherit some features from the Gaussian.

There is a vast literature on the subject, which can be roughly classified into three types of results.

(1). Properties holding as soon as $t > 0$. For example, for all $t > 0$, $f(t, \cdot)$ is smooth [Eva22].
(2). Asymptotic behaviour, in the limit $t \to \infty$, for which we refer to [BE85, DT15, Váz17].
(3). Properties which are satisfied by $f(t, \cdot)$ for $t \geq T$, after a finite time $T > 0$.

5.1.1 Log-concavity in finite time

Observing that the fundamental solution to (5.1.1) is log-concave for all $t > 0$, we pose the following, in the spirit of *(3).*

Question. *Given a probability measure μ on $\mathbb{R}^d$, does there exist a time $T > 0$, such that the solution $f(t, x)$ to* (5.1.1) *is log-concave for $t \geq T$?*

In general, we cannot expect *instantaneous* creation of log-concavity, as suggested by the example $\mu = \frac{1}{2}(\delta(1) + \delta(-1)) \in \mathcal{P}(\mathbb{R})$, see [Bri23]. In addition, some hypotheses on the behaviour at infinity of μ shall be required, as suggested by [Her99]. On the other hand, our question has a positive answer in two known cases.

- If μ is already log-concave, the solution to (5.1.1) is log-concave at all times, see [SW14, Pre73, Lei72, BL76]. Then, by the semigroup property, if a solution to (5.1.1) is log-concave at a time $T > 0$, this property will be propagated to all $t \geq T$.
- If μ is supported in $B(0, R)$, then $f(t, \cdot)$ is log-concave for all $t \geq R^2$, as pointed out first in [LV03]. More precisely, in [BGMZ18] it is shown that for all $t > 0$

$$-\nabla^2 \log(\mu * \gamma_t) \succcurlyeq \frac{1}{t}\left(1 - \frac{R^2}{t}\right) I_d. \tag{5.1.3}$$

One aim of ours is to extend the class of measures for which creation of log-concavity in finite time holds, beyond the compactly supported case, motivated also by the series of papers [IST21, IST22, Ish23, IST24], concerning various concavity property of solutions for the heat flow.

An analogous question can be posed in the context of functional inequalities satisfied by the Gaussian distribution. Starting from the case of compactly supported measures, previously analysed in [Zim13, WW16, BGMZ18], Chen, Chewi, and Niles-Weed prove in [CCNW21] that if μ is subgaussian, i.e. for some $\varepsilon, \mathcal{K} > 0$

$$\int_{\mathbb{R}^d} e^{\varepsilon \|x\|^2} \mu(dx) \leq \mathcal{K}, \tag{5.1.4}$$

then the solution $\mu_t := f(t, \cdot)\, dx$ to (5.1.1) satisfies a log-Sobolev inequality, for $t \geq T(\varepsilon, \mathcal{K})$. Moreover, the subgaussianity assumption is also necessary. Indeed, if μ_T satisfies a log-Sobolev inequality for some $T > 0$, then μ_T is also subgaussian [BGL14, Prop. 5.4.1], which implies that μ is subgaussian in the first place. On the other hand, strongly log-concave measures do also satisfy a logarithmic Sobolev inequality, see [BE85]. Then, one might wonder if (5.1.4) would be sufficient for a measure to become log-concave along the heat flow. The following theorem implies that this is not the case.

Theorem 5.1.1. *For all non-decreasing function $\Psi\colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, there exists an explicit probability measure on $\mathbb{R}$ such that*

- $\int_{\mathbb{R}} \mathrm{e}^{\Psi(x)} \mu(dx) < \infty$;
- *for all* $t > 0$, $\inf_{x\in\mathbb{R}}\Big\{-\frac{d^2}{dx^2}\log \mu * \gamma_t\Big\} = -\infty$.

Remark 5.1.2. *Similar conclusions hold in arbitrary dimension, as it can be seen by considering the product probability measure $\mu \times \delta_0 \times \ldots \times \delta_0$, with μ given by Theorem 5.1.1.*

Our result shows that the creation of log-concavity cannot be guaranteed by assuming only some control on the tails of the distributions μ. Therefore, we restrict our analysis to a perturbation regime, i.e. we take measures μ which are close to being strongly log-concave, and we show that they become log-concave after a finite time along (5.1.1). More precisely, we prove the following

Theorem 5.1.3. *Suppose that $\mu = \mathrm{e}^{-(V+H)} \in L^1_+(\mathbb{R}^d)$, where $V\colon \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ is α-convex and $H\colon \mathbb{R}^d \to \mathbb{R}$ is L-Lipschitz for some $\alpha \in \mathbb{R}$, $L \geq 0$. Then for every $t > 0$ such that $\alpha t + 1 > 0$ we have*

$$\frac{1}{t}\left[1 - \frac{1}{t}\left(\frac{L}{\alpha + \frac{1}{t}} + \sqrt{\frac{1}{\alpha+\frac{1}{t}}}\right)^2\right] I_d \preccurlyeq -\nabla^2 \mathrm{log}(\mu * \gamma_t) \preccurlyeq \frac{1}{t} I_d. \tag{5.1.5}$$

In particular, for $\alpha > 0$ and $t \geq \left(\frac{L}{\alpha} + \sqrt{\frac{1}{\alpha}}\right)^2$, we have that $\mu * \gamma_t$ is strongly log-concave.

Equation (5.1.5) goes beyond the problem of log-concavity, yielding interesting consequences, as explained in the next subsections.

5.1.2 Application to Lipschitz transport maps

In a seminal paper [Caf00], Caffarelli showed that the Brenier map [Bre91] from optimal transport between the standard Gaussian γ and an α-log-concave probability measure μ is $(1/\sqrt{\alpha})$-Lipschitz. This result is useful because Lipschitz transport maps transfer functional inequalities (including isoperimetric, log-Sobolev and Poincaré inequalities) from a probability measure to another one, and it is typically much easier to prove these inequalities for the Gaussian measure in the first place. For example, suppose that a probability measure μ satisfies the log-Sobolev inequality $\mathrm{LSI}(C)$ for some $C > 0$, i.e. for all regular enough probability measures $\rho \ll \mu$

$$\int \frac{d\rho}{d\mu} \log \frac{d\rho}{d\mu}\, d\mu \leq 2C \int \left|\nabla\sqrt{d\rho/d\mu}\right|^2 d\mu, \tag{LSI(C)}$$

where the two sides of the inequalities go under the name of *relative entropy* and *relative Fisher information*, respectively. Suppose, furthermore, that $T\colon \mathbb{R}^d \to \mathbb{R}^d$ is L-Lipschitz and consider the pushforward probability measure $\nu := T\#\mu$. Then, ν satisfies $\mathrm{LSI}(L^2\cdot C)$. Therefore, Caffarelli's result (together with the Gaussian LSI [Gro75]) immediately implies that strongly α-log-concave probability densities satisfy $\mathrm{LSI}(1/\alpha)$, recovering the celebrated result by Bakry and Émery [BE85]. Further details and many more applications of Lipschitz transport maps are discussed in [MS23, CE02] and the references therein.

More recently, Kim and Milman [KM12] generalized Caffarelli's result by constructing another transport map, which is obtained by reverting an appropriate heat flow, and is referred to as the *heat-flow map* (notation: T^{flow}). Other Lipschitz estimates for this transport map were then provided in [MS23], where the authors considered different types of assumptions on the target measure ν (namely, measures that satisfy a combination of boundedness and (semi-)log-concavity and some Gaussian convolutions).

Several works dealt with the study of Lipschitz transport maps [KP21, CF21, DGH$^+$23, MS21, She24, CFJ17, CFS24, CP23]; the recent paper [FMS24] in particular considers an analogous class of target measure as in the present contribution. For comparison, we recall below its main result in the Euclidean setting.

Theorem 5.1.4 ([FMS24], Theorem 1). *Let $\mu = e^{-(V+H)}$, $\nu = e^{-V}$ be probability densities on $\mathbb{R}^d$ such that for all $x \in \mathbb{R}^d$ we have*

$$\|\nabla H\| \leq L, \qquad \nabla^2 V(x) \geq \alpha I_d, \qquad \left|\nabla^3 V(x)(w,w)\right| \leq K \qquad \text{for all } w \in \mathbb{S}^{d-1},$$

for some $\alpha > 0$, L, $K \geq 0$. Then, there exists a transport map $T \colon \mathbb{R}^d \to \mathbb{R}^d$ such that $T\#\nu = \mu$ and T is $\exp\Big(\frac{5L^2}{\alpha} + \frac{5\sqrt{\pi}L}{\sqrt{\alpha}} + \frac{LK}{2\alpha^2}\Big)$-Lipschitz.

Since Lipschitz transport maps can be composed, this result (combined with Caffarelli's theorem [Caf00]) implies in particular the existence of transport map $\tilde{T}$ such that $\tilde{T}\#\gamma = \mu$ and $\tilde{T}$ is Lipschitz with constant

$$\frac{1}{\sqrt{\alpha}} \exp\bigg(\frac{5L^2}{\alpha} + \frac{5\sqrt{\pi}L}{\sqrt{\alpha}} + \frac{LK}{2\alpha^2}\bigg). \tag{5.1.6}$$

On the other hand, we will prove in Section 5.3 that our Theorem 5.1.3 implies new upper bounds on the Lipschitz norm for the heat-flow map from γ to μ.

Theorem 5.1.5. *Let $\mu = e^{-(V+H)} \in L^1_+(\mathbb{R}^d)$ be a probability density on $\mathbb{R}^d$ such that V is α-convex for $\alpha > 0$ and H is L-Lipschitz for $L \geq 0$. Then, there exists a map $T^{\text{flow}} \colon \mathbb{R}^d \to \mathbb{R}^d$ such that $T^{\text{flow}}\#\gamma = \mu$ and T^{flow} is $\frac{1}{\sqrt{\alpha}} \exp\Big(\frac{L^2}{2\alpha} + 2\frac{L}{\sqrt{\alpha}}\Big)$-Lipschitz.*

Remark 5.1.6. *Consider the case where $d = 1$, $V(x) = \frac{1}{2}x^2$ and $H(x) = L|x| + \log(Z)$ for a normalizing constant Z, so that the assumptions of Theorem 5.1.5 are satisfied with $\alpha = 1$. Then, it was observed in [FMS24] that the Lipschitz norm of any map T such that $T\#\gamma = \mu$ is at least $\mathrm{e}^{\frac{L^2}{2}}$. Hence, the dependence on L^2 in Theorem 5.1.5 is sharp.*

The estimate for the Lipschitz constant of T^{flow} in Theorem 5.1.5 improves in particular on the value in (5.1.6), yielding the best available bound in this setting. Moreover, Theorem 5.1.5 does not need any assumption on $\nabla^3 V$.

On the technical side, in Theorem 5.1.5 we transport directly γ to μ via the heat-flow map, and our proof only exploits elementary log-Hessian estimates for the heat semigroup, as in Theorem 5.1.3. On the other hand, [FMS24] employs a construction based on reverting the overdamped Langevin dynamics targeting the measure $\nu = \mathrm{e}^{-V}$: this requires estimates for the corresponding semigroup (cf. [FMS24, Proposition 2]), which is less explicit and needs more sophisticated arguments. We remark that the results of [FMS24] are of independent interest, due to the construction of a Lipschitz map transporting ν to μ therein, and the extension to some non-Euclidean spaces.

5.1.3 Score-based diffusions models and the Gaussian LSI

To further motivate our results, we briefly describe here two more applications of Theorem 5.1.1 and 5.1.3.

Score-based diffusion models A similar construction as in Section 5.1.2, based on reverting an ergodic diffusion process, has also recently found application in the machine learning community, within the framework of score-based diffusion models [SSDK+21, HJA20]. Let μ be a probability measure, from which we want to generate random samples. Consider the Ornstein–Uhlenbeck process (initialized at μ)

$$X_0 \sim \mu, \qquad dX_t = -X_t dt + \sqrt{2} dB_t,$$

and denote by Q_t the associate semigroup, i.e.

$$Q_t f(x) = \int f\Big(e^{-t}x + \sqrt{1-e^{-2t}}\Big)\gamma(x)\,dx. \tag{5.1.7}$$

The key observation is that this process can be reverted, i.e. for $T_1 > 0$ the *reverse SDE*

$$Y_0 \sim \mathrm{law}(X_{T_1}), \qquad dY_t = -Y_t\,dt + 2\nabla \log Q_t\left(\frac{d\mu}{d\gamma}\right)(Y_t)dt + \sqrt{2}dB_t \tag{5.1.8}$$

is such that $Y_{T_1} \sim \mu$, see [And82, CCGL23, SSDK+21]. Therefore, one can simulate the process $(Y_t)_t$ until time T_1 to sample from μ. A common assumption in theoretical works aimed at analysing this method is some control on the Lipschitz constant of $\nabla \log Q_t\big(\frac{d\mu}{d\gamma}\big)$ [CCL+23a, CCL+23b, CLL22] or on the one-sided one [KFL22, PMM24]. These assumptions are indeed useful to control the discretization errors when employing a numerical scheme to simulate the process or some sort of "contractivity" along the reverse dynamics. On the one hand, Theorem 5.1.3 enlarges the class of distributions μ for which these assumptions can be justified, by implying bounds on the Hessian $\nabla^2 \log Q_t\big(\frac{d\mu}{d\gamma}\big)$ (cf. Corollary 5.3.2), beyond the setting where the initial distribution μ has bounded support.

On the other hand, Theorem 5.1.1 shows that, for some distributions μ, such assumptions can be too restrictive. Thus, complementary analysis is needed, as done in [CDS23, BDBDD23, CLL22].

Improvements in the Gaussian LSI The standard Gaussian measure γ satisfies $\mathrm{LSI}(1)$. Henceforth, let $\nu \ll \gamma$ be a probability measure, and set $u^2 := \frac{d\nu}{d\gamma} \in L^1(\gamma)$: then

$$\int_{\mathbb{R}^d} |\nabla u|^2\,d\gamma - \frac{1}{2}\int_{\mathbb{R}^d} u^2\,\log u^2\,d\gamma \geq 0, \quad \text{if } u \in \mathrm{H}^1(d\gamma). \tag{γ-LSI}$$

The Gaussian logarithmic Sobolev inequality was written first in [Gro75], although it can be deduced from [Sha48]. The related literature is wide: see [BDS23b] for a recent review, and [RV08, Car91] for accurate historical comments. The constant $C = 1$ is optimal, with extremal probability measures belonging to $\mathcal{M} := \{\nu_{a,b} := \mathrm{e}^{a+\langle b,x\rangle}\,\gamma, \;\; a \in \mathbb{R},\, b \in \mathbb{R}^d\}$, according to [Car91]. Then, one may investigate whether the constant in the Gaussian LSI can be improved on a subclass of measures ν, under orthogonality constraints. Contributions in this direction appear in [FIL16, BDS23b], and they are closely related to *stability inequalities*, for which the reader may refer to [DEF+22, DEF+24, BDS23a, BDS24, IK21], and references quoted therein.

In [BDS23b, Theorem 1], an improved Gaussian LSI is shown: for all $\varepsilon, \mathcal{K} > 0$, there exists a constant $\eta(\varepsilon, \mathcal{K}) > 0$, such that for all probability measures $\nu = u^2 \gamma$ satisfying $\int_{\mathbb{R}^d} x \, \nu = 0$, and (5.1.4), we have

$$\int_{\mathbb{R}^d} |\nabla u|^2 \, d\gamma - \frac{1}{2} \int_{\mathbb{R}^d} u^2 \log u^2 \, d\gamma \geq \eta \int_{\mathbb{R}^d} |\nabla u|^2 \, d\gamma, \qquad \text{if } u \in \mathrm{H}^1(d\gamma). \tag{5.1.9}$$

The core of the proof for (5.1.9) is showing that – after a finite time $T > 0$ – the solution $f(t, \cdot) = \nu * \gamma_t$ to (5.1.1), starting at ν, satisfies a Poincaré inequality:

$$\int_{\mathbb{R}^d} \varphi(x)^2 \, f(t,x) \, dx - \left(\int_{\mathbb{R}^d} \varphi(x) \, f(t,x) \, dx \right)^2 \leq C_P \int_{\mathbb{R}^d} |\nabla \varphi(x)|^2 \, f(t,x) \, dx,$$

for all functions $\varphi \in \mathrm{H}^1(f(t, \cdot) \, dx)$, and some constant $C_P > 0$. Condition (5.1.4) guarantees such a Poincaré inequality in finite time, see [CCNW21]. Then, [FIL16, Theorem 1] applies, and an improved inequality like (5.1.9) holds for $u(T, \cdot) = \sqrt{f(T, \cdot)/\gamma}$, after a finite time. The proof is completed by *integrating backwards in time* via [BDS23b, Lemma 2].

- We notice first that, if $\nu = \mathrm{e}^{-(V+H)}$ is a log-Lipschitz perturbation of a strongly log-concave measure, then the ideas of [BDS23b, Theorem 1] apply, since (5.1.4) holds true. Alternatively, one could estimate the Poincaré constant of $f(t, \cdot) = \nu * \gamma_t$, for any $t \geq 0$, either via the Lipschitz transport map of Section 5.3, or by a perturbation argument [CG22], and apply [FIL16, Theorem 1]. Finally, one can optimise the resulting constant in (5.1.9) over the parameter $t \geq 0$.

- The scheme of proof for [BDS23b, Theorem 1] can be adapted to measures $\nu = u^2 \, d\gamma$ which become α-log-concave in finite time along (5.1.1), for $\alpha > 0$. In this case, the Poincaré inequality is given by the Bakry-Émery method [BE85].

- The same can be done for measures ν which become just log-concave ($\alpha = 0$) in finite time along (5.1.1), provided an a priori bound on the second-order moment $\int_{\mathbb{R}^d} |x|^2 \, \nu$, see [Bob99] and the discussion of [BDS23b, Section 2].

5.1.4 Structure of the paper

The proof of Theorem 5.1.3 is given in Section 5.2, followed by Subsection 5.2.1, where sufficient conditions in order to apply Theorem 5.1.3 are discussed. In Section 5.3, we detail our main application to the existence of Lipschitz transport maps, with the proof of Theorem 5.1.5. Finally, in Section 5.4, we prove the negative result for the creation of log-concavity in finite time, namely Theorem 5.1.1.

Acknowledgement

The authors thank Professors Jean Dolbeault, Jan Maas, and Nikita Simonov for many useful comments, and Professors Kazuhiro Ishige, Asuka Takatsu, and Yair Shenfeld for inspiring interactions.
This research was funded in part by the Austrian Science Fund (FWF) project 10.55776/F65 and by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 101034413.

5.2 Log-Lipschitz perturbations of log-concave measures: proof of Theorem 5.1.3

Let μ be a probability measure on $\mathbb{R}^d$. For $t > 0$ and $z \in \mathbb{R}^d$, define the probability measure $\mu_{z,t}$ by

$$\mu_{z,t} \propto \exp\left(\frac{z \cdot x}{t} - \frac{\|x\|^2}{2t}\right)\mu(x) \propto \gamma_{z,t}(x)\mu(x), \tag{5.2.1}$$

where $\gamma_{z,t}$ is the Gaussian density with mean z and covariance matrix tI_d. We will make frequent use of the following well-known probabilistic characterization of the Hessian of $\log(\mu * \gamma_t)$, cf. [BGMZ18, KP21]:

$$-\nabla^2 \log(\mu * \gamma_t)(z) = \frac{1}{t}\left(I_d - \frac{\mathrm{Cov}_{\mu_{z,t}}}{t}\right). \tag{5.2.2}$$

Consequently, bounds on $\nabla^2 \log(\mu * \gamma_t)$ are given by bounds on covariance matrices. For this purpose, we provide the following lemma, which gives an upper bound for the covariance matrix of a probability measure μ in terms of the covariance of another probability measure ν and of the Wasserstein distance between the two.

Lemma 5.2.1. *Let μ, ν be probability measures on $\mathbb{R}^d$. For any unit vector $w \in \mathbb{S}^{d-1}$*

$$\langle w, \mathrm{Cov}_\mu w\rangle \le \left(W_2(\mu,\nu) + \sqrt{\langle w, \mathrm{Cov}_\nu w\rangle}\right)^2. \tag{5.2.3}$$

Proof. Let (X, Y) be an optimal coupling for $W_2(\mu, \nu)$. Fix a unit vector $w \in \mathbb{R}^d$ and let $X_w \coloneqq \langle w, X\rangle$ and $Y_w \coloneqq \langle w, Y\rangle$. We have that

$$\begin{aligned}
\langle w, \mathrm{Cov}_\mu w\rangle &= \mathbb{E}\Big[(X_w - \mathbb{E}[X_w])^2\Big] \\
&\le \mathbb{E}\Big[(X_w - \mathbb{E}[Y_w])^2\Big] = \mathbb{E}\Big[(X_w - Y_w + Y_w - \mathbb{E}[Y_w])^2\Big] \\
&\le \left(\sqrt{\mathbb{E}\Big[(X_w - Y_w)^2\Big]} + \sqrt{\mathbb{E}\Big[(Y_w - \mathbb{E}[Y_w])^2\Big]}\right)^2 \qquad \text{(by Cauchy–Schwarz)} \\
&\le \left(W_2(\mu,\nu) + \sqrt{\mathbb{E}\Big[(Y_w - \mathbb{E}[Y_w])^2\Big]}\right)^2 = \left(W_2(\mu,\nu) + \sqrt{\langle w, \mathrm{Cov}_\mu w\rangle}\right)^2.
\end{aligned}$$

∎

Proof of Theorem 5.1.3. The upper bound in (5.1.5) is well known, and holds for arbitrary probability measures μ (cf., for example, [EL18, Lemma 1.3]); alternatively, it follows from (5.2.2) and the fact that covariance matrices are positive semidefinite. Let us then turn to the first inequality. Fix $t > 0$ and $z \in \mathbb{R}^d$. Define the probability density $\nu_{z,t} \in L^1_+(\mathbb{R}^d)$ by $\nu_{z,t} \propto e^{-V}\gamma_{z,t}$. Notice that $\nu_{z,t}$ is $(\alpha + \frac{1}{t})$-log-concave: therefore, $\mathrm{Cov}_{\nu_{z,t}} \preccurlyeq \frac{1}{\alpha+\frac{1}{t}} I_d$ by the Brascamp–Lieb inequality [BL76] (cf. also [EL14, Lemma 5]). Moreover we have $\mu_{z,t} \propto e^{-H}\nu_{z,t}$: it follows from [KMP24, Corollary 2.4] that

$$W_2(\mu_{z,t}, \nu_{z,t}) \le W_\infty(\mu_{z,t}, \nu_{z,t}) \le \frac{L}{\alpha + \frac{1}{t}}.$$

We are now in position to apply Lemma 5.2.1: for any unit vector $v \in \mathbb{R}^d$ we have

$$\begin{aligned}\langle v, \mathrm{Cov}_{\mu_{z,t}} v\rangle &\leq \Big(W_2(\mu_{z,t}, \nu_{z,t}) + \sqrt{\langle v, \mathrm{Cov}_{\nu_{z,t}} v\rangle}\Big)^2 \\ &\leq \left(\frac{L}{\alpha + \frac{1}{t}} + \sqrt{\frac{1}{\alpha + \frac{1}{t}}}\right)^2.\end{aligned}$$

This shows that $\mathrm{Cov}_{\mu_{z,t}} \preccurlyeq \left(\frac{L}{\alpha+\frac{1}{t}} + \sqrt{\frac{1}{\alpha+\frac{1}{t}}}\right)^2 I_d$, and the conclusion follows from (5.2.2). ■

Remark 5.2.2. *In the proof of Theorem 5.1.3, we estimated from above $W_2(\mu_{z,t}, \nu_{z,t})$ with the L^∞-Wasserstein distance $W_\infty(\mu_{z,t}, \nu_{z,t})$. Alternatively, we could have achieved the same conclusion as follows, using that $\nu_{z,t}$ satisfies $\mathrm{LSI}\big(\frac{t}{\alpha t+1}\big)$. First, a transport-entropy inequality [OV00] allows to estimate $W_2(\mu_{z,t}, \nu_{z,t})$ in terms of the relative entropy of $\mu_{z,t}$ with respect to $\nu_{z,t}$; then, the relative entropy is bounded from above by the relative Fisher information using the logarithmic Sobolev inequality of $\nu_{z,t}$; finally, the relative Fisher information is easily estimated using that $\mu_{z,t} \propto e^{-H}\nu_{z,t}$ and H is L-Lipschitz.*

5.2.1 Sufficient conditions

By Theorem 5.1.3, log-Lipschitz perturbations of strongly log-concave measures become log-concave in finite time along (5.1.1); by Theorem 5.1.5, they are the pushforward of the Gaussian measure via a Lipschtzt transport map. The purpose of this subsection is to give sufficient conditions for a measure μ to be a log-Lipschitz perturbation of a strongly log-concave measure.

Example 5.2.3. *Suppose that μ is a probability measure supported on the Euclidean ball $B(0, R)$ for some radius $R > 0$. Then, proceeding as in [BGMZ18], for any $s > 0$ we can write*

$$\mu * \gamma_s = e^{-H}\gamma_s$$

where $H\colon \mathbb{R}^d \to \mathbb{R}^d$ is $\frac{R}{s}$-Lipschitz. By $\frac{1}{s}$-log-concavity of γ_s, Theorem 5.1.3 applied to $\mu * \gamma_s$ and $t > 0$ yields that

$$\frac{1}{t}\left[1 - \frac{1}{t}\left(\frac{R/s}{1/s + \frac{1}{t}} + \sqrt{\frac{1}{1/s + \frac{1}{t}}}\right)^2\right] I_d \preccurlyeq -\nabla^2 \mathrm{log}(\mu * \gamma_{t+s}) \preccurlyeq \frac{1}{t+s} I_d.$$

By letting $s \to 0$, we recover the classical estimate

$$\frac{1}{t}\left(1 - \frac{R^2}{t}\right) I_d \preccurlyeq -\nabla^2 \log(\mu * \gamma_t) \preccurlyeq \frac{1}{t} I_d,$$

cf. [BGMZ18, Sec. 2.1]. In this sense, we can say that the class of densities considered in Theorem 5.1.3 contains both log-concave ones (taking $L = 0$) and the ones with bounded support.

Consider now a probability density $\mu = \mathrm{e}^{-U} \in L^1_+(\mathbb{R}^d)$ for some $U \in C^2(\mathbb{R}^d)$. The following result asserts that, if we have a uniform positive lower bound for the Hessian of U outside some Euclidean ball, then we can rewrite μ as a log-Lipschitz perturbation of a strongly log-concave measure.

Lemma 5.2.4. *Let $U \in C^2(\mathbb{R}^d)$ be such that for some $\alpha, \beta, R \geq 0$ it holds that*

$$\begin{cases} \nabla^2 U(x) \succcurlyeq \alpha I_d & \text{if } \|x\| \geq R, \\ \nabla^2 U(x) \succcurlyeq -\beta I_d & \text{if } \|x\| < R. \end{cases}$$

Then there exists $V, H \in C^1(\mathbb{R}^d)$ such that $U = V + H$, V is α-convex and H is $2(\alpha+\beta)R$-Lipschitz.

Proof. Let $H \colon \mathbb{R}^d \to \mathbb{R}$ be defined by

$$-H(x) = \begin{cases} (\alpha+\beta)\|x\|^2 & \text{if } \|x\| \leq R, \\ 2(\alpha+\beta)R\|x\| - 2(\alpha+\beta)R^2 & \text{if } \|x\| \geq R, \end{cases}$$

and set $V(x) = U(x) - H(x)$. Then we have that $U = V + H$, $V \in C^1(\mathbb{R}^d)$ is α-convex and $\|\nabla H\| \leq 2(\alpha+\beta)R$, as desired. ∎

The above lemma can be useful to study linear combinations of strongly log-concave densities, via the following

Proposition 5.2.5. *Consider a measure $\mu = \sum_{i=1}^N \alpha_i \, \mathrm{e}^{-U_i}$ for some $N > 0$, weights $\alpha_i > 0$ and potentials $U_i \in C^2(\mathbb{R}^d)$ such that $\mathrm{e}^{-U_i} \in L^1_+(\mathbb{R}^d)$. Assume $\nabla^2 U_i \succcurlyeq K I_d$ for all i and some $K > 0$. Then*

$$-\nabla^2 \log \mu \succcurlyeq K I_d - \frac{\sum_{i>j} \alpha_i \alpha_j \mathrm{e}^{-U_i - U_j} (\nabla U_i - \nabla U_j)^{\otimes 2}}{\mu^2} \tag{5.2.4}$$

$$\succcurlyeq K I_d - \sum_{i>j} \frac{(\nabla U_i - \nabla U_j)^{\otimes 2}}{\left(2 + \frac{\alpha_i}{\alpha_j} \mathrm{e}^{U_j - U_i} + \frac{\alpha_j}{\alpha_i} \mathrm{e}^{U_i - U_j}\right)}. \tag{5.2.5}$$

Proof. Notice that

$$-\nabla^2 \log \mu = \mu^{-2}(\nabla\mu \otimes \nabla\mu - \mu\nabla^2\mu).$$

Set $\mu_i := \alpha_i \mathrm{e}^{-U_i}$ so that $\mu = \sum_{i=1}^N \mu_i$. By construction

$$\nabla\mu_i = -\nabla U_i \, \mu_i, \qquad \nabla^2 \mu_i = (-\nabla^2 U_i + \nabla U_i \otimes \nabla U_i)\mu_i, \quad \forall i = 1, \ldots, N.$$

Then,

$$\begin{aligned} -\nabla^2 \log \mu &= \frac{\left(\sum_{i=1}^N \nabla U_i \, \mu_i\right)^{\otimes 2} - \left(\sum_{i=1}^N \mu_i\right) \left(\sum_{i=1}^N (-\nabla^2 U_i + \nabla U_i \otimes \nabla U_i)\mu_i\right)}{\mu^2} \\ &= \frac{\mu \sum_{i=1}^N \nabla^2 U_i \, \mu_i - \sum_{i,j=1}^N \mu_i \, \mu_j (\nabla U_i \otimes \nabla U_j - \nabla U_j \otimes \nabla U_j)}{\mu^2} \\ &\succcurlyeq K \, I_d - \frac{\sum_{i>j} \mu_i \mu_j (\nabla U_i - \nabla U_j)^{\otimes 2}}{\mu^2}, \end{aligned}$$

which shows (5.2.4). The crude estimate

$$\mu^2 = \sum_{l,m=1}^N \mu_l \, \mu_m \geq 2\mu_i \, \mu_j + \mu_i^2 + \mu_j^2 \quad \text{for } i \neq j$$

then gives (5.2.5). ∎

From the above proposition, it is clear that when the right-hand-side of (5.2.4) is uniformly positive definite outside a Euclidean ball, then by Lemma 5.2.4 we can recast μ as a log-Lipschitz perturbation of a strongly log-concave measure. Therefore, the assumptions of Theorem 5.1.3 are satisfied, and $\mu * \gamma_t$ becomes strongly log-concave in finite time along the heat flow (5.1.1). We illustrate this in the following example, where μ is a finite mixture of Gaussians in dimension 1.

Example 5.2.6. *Let μ be a linear combination of one-dimensional Gaussians, i.e. $\mu = \sum_{i=1}^N \alpha_i \,\mathrm{e}^{-U_i}$ for some $N \geq 2$, weights $\alpha_i > 0$ and potentials U_i of the form*

$$U_i(x) = \frac{(x - m_i)^2}{\sigma_i^2}$$

for some $m_i \in \mathbb{R}, \sigma_i^2 > 0$. Without loss of generality we can assume that $U_i \neq U_j$ for $i \neq j$. By Proposition 5.2.5, we have that

$$-\frac{d^2}{dx^2} \log \mu \succcurlyeq \frac{1}{\max_i \sigma_i^2} - \sum_{i>j} \frac{(U_i' - U_j')^2}{\left(2 + \frac{\alpha_i}{\alpha_j} \mathrm{e}^{U_j - U_i} + \frac{\alpha_j}{\alpha_i} \mathrm{e}^{U_i - U_j}\right)}.$$

It is then not difficult to see that the argument of the sum in the right-hand-side converges to 0 as $|x| \to \infty$. By the previous discussion, it follows that the assumptions of Theorem 5.1.3 are satisfied for some $L, \alpha > 0$: hence, a finite linear combination of Gaussian densities on $\mathbb{R}$ becomes strongly log-concave in finite time along the heat flow.

5.3 Lipschitz transport maps: proof of Theorem 5.1.5

Construction of the heat-flow map. Let $\mu \in L^1_+(\mathbb{R}^d)$ be a probability density on $\mathbb{R}^d$. Assume, furthermore, that μ has finite second-order moment. We begin by sketching the construction of the heat-flow map, and refer the reader to [KM12, MS23] for details. The idea is to interpolate between μ and γ along the Ornstein–Uhlenbeck flow

$$X_0 \sim \mu, \qquad dX_t = -X_t dt + \sqrt{2} dB_t. \tag{5.3.1}$$

Let us denote by Q_t the associated transition semigroup (5.1.7) and by μ_t the law of X_t. Then, μ_t satisfies the Fokker–Planck equation

$$\partial \mu_t - \nabla \cdot \left[\mu_t \nabla \log Q_t\left(\frac{d\mu}{d\gamma}\right)\right] = 0.$$

Correspondingly, we can consider the flow maps $(S_t)_{t\geq 0}$ obtained by solving

$$S_0(x) = x, \quad \frac{\mathrm{d}}{\mathrm{d}t} S_t(x) = -\nabla \log Q_t\left(\frac{d\mu}{d\gamma}\right)$$

for all $x \in \mathbb{R}^d$. Under some regularity assumptions (cf. [KM12, MS23, OV00, Vil03]), this defines a flow of diffeomorphisms such that $S_t \# \mu = \mu_t$; conversely, $T_t := S_t^{-1}$ is such that $T_t \# \mu_t = \mu$. The heat-flow map is then heuristically defined by $T^{\mathsf{flow}} = \lim_{t\to\infty} T_t$ and is such that $T^{\mathsf{flow}} \# \gamma = \mu$. To make things rigorous, we recall/adapt the following result from [MS23].

Lemma 5.3.1. *Suppose that $\mu \in L^1_+(\mathbb{R}^d)$ is a probability density with finite second-order moment. Suppose, furthermore, that for all $t > 0$ there exist $\theta_t^{max}, \theta_t^{min} \in \mathbb{R}$ such that*

$$\theta_t^{min} I_d \preccurlyeq \nabla^2 \log Q_t\left(\frac{d\mu}{d\gamma}\right) \preccurlyeq \theta_t^{max} I_d \tag{5.3.2}$$

and for all $s > 1$

$$\sup_{\frac{1}{s}<t<s} \max\left\{|\theta_t^{min}|, |\theta_t^{max}|\right\} < \infty.$$

Then, provided that $L := \limsup_{t\to\infty} \int_{\frac{1}{t}}^{t} \theta_t^{max} dt < \infty$, there exists a map $T\colon \mathbb{R}^d \to \mathbb{R}^d$ such that $T\#\gamma = \mu$ and T is e^L-Lipschitz.

Proof. Notice first of all that μ_t is a smooth density for every $t > 0$. Fix $s > 0$: by the assumptions in the Lemma and by [MS23, Lemma 2 and 3] there exists a map T_s which is $\exp\left(\int_{\frac{1}{s}}^{s} \theta_t^{max}\, dt\right)$-Lipschitz and such that $T_s\#\mu_s = \mu_{\frac{1}{s}}$. Since $\mu_s \to \gamma$ and $\mu_{\frac{1}{s}} \to \mu$ in W_2-distance (hence weakly) as $s \to \infty$, the conclusion follows from [MS23, Lemma 1]. ■

New estimates. In view of Lemma 5.3.1, the goal is to provide estimates on $\nabla^2 \log Q_t\left(\frac{d\mu}{d\gamma}\right)$, for some classes of probability measures μ on $\mathbb{R}^d$. The Ornstein–Uhlenbeck semigroup Q_t is related to the heat semigroup P_t in (5.1.2) by the identity $Q_t f(x) = P_{1-e^{-2t}} f(e^{-t}x)$ for $f \in L^1(\gamma)$. Combining this with Theorem 5.1.3 yields the following

Corollary 5.3.2 (Corollary of Thm. 5.1.3)**.** *Let $\mu = e^{-V-H} \in L^1_+(\mathbb{R}^d)$ be a probability density on $\mathbb{R}^d$ such that V is α-convex and H is L-Lipschitz, for some $\alpha, \in \mathbb{R}, L \geq 0$. Then for every $0 < t$ such that $\alpha t + 1 > 0$ we have*

$$\begin{aligned} -\frac{1}{e^{2t}-1} I_d \preccurlyeq & \nabla^2 \log Q_t\left(\frac{d\mu}{d\gamma}\right) \\ & \preccurlyeq \left(\frac{1-\alpha}{\alpha(e^{2t}-1)+1} + \frac{e^{2t}L^2}{(\alpha(e^{2t}-1)+1)^2} + \frac{2Le^{2t}}{\sqrt{(e^{2t}-1)}\,(\alpha(e^{2t}-1)+1)^{3/2}}\right) I_d \end{aligned} \tag{5.3.3}$$

Proof of Theorem 5.1.5. We integrate the upper bound in (5.3.3). An elementary computation using the change of variable $\tau = e^{2t} - 1$ shows that

$$\begin{aligned} &\int_0^\infty \left(\frac{1-\alpha}{\alpha(e^{2t}-1)+1} + \frac{e^{2t}L^2}{(\alpha(e^{2t}-1)+1)^2} + \frac{2Le^{2t}}{\sqrt{(e^{2t}-1)}\,(\alpha(e^{2t}-1)+1)^{3/2}}\right) dt \\ &= \int_0^\infty \left(\frac{1-\alpha}{\tau\alpha+1} + L^2 \frac{\tau+1}{(\tau\alpha+1)^2} + 2L\frac{\tau+1}{\sqrt{\tau}\,(\tau\alpha+1)^{3/2}}\right) \frac{1}{2(\tau+1)} d\tau \\ &= -\frac{1}{2}\log(\alpha) + \frac{L^2}{2\alpha} + 2\frac{L}{\sqrt{\alpha}}. \end{aligned}$$

The desired conclusion then follows from Lemma 5.3.1. ■

5.4 The negative result: proof of Theorem 5.1.1

Before proving the actual theorem, we give some heuristics behind the proof. The leading idea is the following. If one considers (5.1.1) with $\mu = \delta_0$, then the solution is immediately log-concave for $t > 0$. However, this behaviour is not stable.

Proposition 5.4.1. *Fix $x_0 \in \mathbb{R}$. Let $\mu = \frac{\alpha}{\alpha+\beta}\delta_0 + \frac{\beta}{\alpha+\beta}\delta_{x_0}$, for some $\alpha, \beta > 0$. Then, $\mu * \gamma_t$ is log-concave (if and) only if $t \geq \frac{1}{4}x_0^2$.*

Proof. We prove only the *only if* part, since the other implication follows directly from (5.1.3). It is not difficult to see that with $x_0, t, \alpha, \beta > 0$ fixed, there exists $\bar{z} \in \mathbb{R}$ for which

$$\alpha e^{-\frac{\bar{z}^2}{2t}} = \beta e^{-\frac{(\bar{z}-x_0)^2}{2t}}.$$

Then, using (5.2.2), we have that

$$\frac{d^2}{dx^2}(-\log \mu * \gamma_t)(\bar{z}) = \frac{1}{t}\left(1 - \frac{x_0^2}{4t}\right),$$

which is negative if $t < x_0^2/4$. ∎

From equation (5.1.3) we see that a compactly-supported distribution becomes log-concave along (5.1.1) after a time $T = O(R^2)$. Proposition 5.4.1 gives a simple account of this time scale being correct. In addition, we see that the time needed for the measure μ of Proposition 5.4.1 to become log-concave along (5.1.1) does not depend on the mass of the perturbation δ_{x_0}. Exploiting these observations allows us to create mixtures of Dirac deltas with arbitrarily thin tails, which never become log-concave along (5.1.1).

Proof of Theorem 5.1.1. For $i \geq 0$, set $x_i = \frac{i(i+1)}{2} \geq 0$. Define the probability measure μ on $\mathbb{R}$ by

$$\mu \propto \sum_{i=0}^{\infty} \frac{1}{(i+1)^2} e^{-\Psi(x_i)} \delta_{x_i}$$

and let $X \sim \mu$. It is immediate to check that $\mathbb{E}\left[e^{\Psi(X)}\right] < \infty$. Let us now fix $t \geq 0$. Recall from (5.2.2) that

$$-\frac{d^2}{dx^2} \log \mu * \gamma_t(z) = \frac{1}{t}\left(1 - \frac{\operatorname{Var} \mu_{z,t}}{t}\right),$$

where

$$\mu_{z,t}(x) \propto e^{\frac{zx}{t} - \frac{x^2}{2t}} \mu(x) \propto \sum_{i=0}^{\infty} \frac{1}{(i+1)^2} e^{-\Psi(x_i) + \frac{zx_i}{t} - \frac{x_i^2}{2t}} \delta_{x_i}.$$

Therefore, it suffices to prove that, for every $M > 0$, there exists z such that $\operatorname{Var} \mu_{z,t} \geq M^2$. To this end, fix M and choose $j \geq \sqrt{2}M$ so that

$$|x_j - x_{j-1}|^2 = j^2 \geq 2M^2.$$

To conclude, it suffices to show that there exists $z \in \mathbb{R}$ such that

$$\mu_{z,t}([0, x_{j-1}]) = \frac{1}{2} = \mu_{z,t}([x_j, +\infty]). \tag{5.4.1}$$

Indeed, the above implies that $\operatorname{Var}\mu_{z,t} \geq M^2$. Notice now that (5.4.1) is equivalent to finding a solution to the equation $F(z) = 0$, where

$$F(z) = \sum_{i=0}^{j-1} \frac{1}{(i+1)^2} e^{-\Psi(x_i) + \frac{zx_i}{t} - \frac{x_i^2}{2t}} - \sum_{i=j}^{\infty} \frac{1}{(i+1)^2} e^{-\Psi(x_i) + \frac{zx_i}{t} - \frac{x_i^2}{2t}}. \tag{5.4.2}$$

It is straightforward to check that $F(0) \geq 0$, e.g. using that $1 > \sum_{i=1}^{\infty} \frac{1}{(i+1)^2}$ and that Ψ is non-decreasing. Moreover, F is continuous, since for any compact interval $[a, b] \subset \mathbb{R}$, the series in (5.4.2) converges uniformly in $C([a, b])$. To conclude, we show now that $\lim_{z\to\infty} F(z) = -\infty$. To this end, notice that

$$\begin{aligned} F(z) &\leq j e^{-\Psi(0) + \frac{zx_{j-1}}{t}} - \frac{1}{(j+1)^2} e^{-\Psi(x_j) - \frac{x_j^2}{2t} + \frac{zx_j}{t}} \\ &= e^{\frac{zx_{j-1}}{t}} \left(j e^{-\Psi(0)} - \frac{1}{(j+1)^2} e^{-\Psi(x_j) - \frac{x_j^2}{2t}} e^{\frac{zj}{t}} \right), \end{aligned}$$

which yields the desired conclusion since $j > 0$. ■

CHAPTER 6

L^∞-optimal transport of anisotropic log-concave measures and exponential convergence in Fisher's infinitesimal model

This chapter corresponds to the preprint [KMP24].

We prove upper bounds on the L^∞-Wasserstein distance from optimal transport between strongly log-concave probability densities and log-Lipschitz perturbations. In the simplest setting, such a bound amounts to a transport-information inequality involving the L^∞-Wasserstein metric and the relative L^∞-Fisher information. We show that this inequality can be sharpened significantly in situations where the involved densities are anisotropic. Our proof is based on probabilistic techniques using Langevin dynamics. As an application of these results, we obtain sharp exponential rates of convergence in Fisher's infinitesimal model from quantitative genetics, generalising recent results by Calvez, Poyato, and Santambrogio in dimension 1 to arbitrary dimensions.

6.1 Introduction

Upper bounds on transport distances to log-concave probability densities play a central role in the theory of optimal transport and in applications in high-dimensional geometry and probability.

One fundamental example is *Talagrand's inequality* [Tal96], which provides a remarkable upper bound for the 2-Wasserstein distance to the standard Gaussian measure γ. For all probability measures ν having finite relative entropy $\mathcal{D}_{\mathrm{KL}}(\nu \,\|\, \gamma) = \int \log \frac{\mathrm{d}\nu}{\mathrm{d}\gamma}(x)\mathrm{d}\nu(x)$, Talagrand's inequality asserts that $W_2(\nu,\gamma) \le \sqrt{2\mathcal{D}_{\mathrm{KL}}(\nu \,\|\, \gamma)}$. More generally, Otto and Villani [OV00] showed that

$$W_2(\nu,\mu) \le \sqrt{\frac{2}{\kappa}\mathcal{D}_{\mathrm{KL}}(\nu \,\|\, \mu)} \tag{6.1.1}$$

for all ν, whenever μ satisfies a logarithmic Sobolev inequality with constant $\kappa > 0$. This includes in particular the class of all κ-log-concave densities. (A probability density μ is said to be κ-*log-concave* for some $\kappa \in \mathbb{R}$, if $\mu = e^{-U}$ where $U : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is κ-*convex*; i.e., $x \mapsto U(x) - \frac{\kappa}{2}|x|^2$ is convex.) The main reason for the great interest of this inequality is that it implies dimension-free Gaussian concentration for μ.

Another seminal result of a similar flavour is *Caffarelli's contraction theorem* [Caf00], which asserts that any 1-log-concave probability density μ can be obtained as the image (or push-forward) of the standard Gaussian measure γ under a 1-Lipschitz map $T : \mathbb{R}^d \to \mathbb{R}^d$. In fact, the optimal transport map for the W_2-distance (the so-called Brenier map) does the job. This theorem is a powerful tool to transfer functional inequalities from the Gaussian measure to the large class of 1-log-concave measures.

6.1.1 L^∞-optimal transport of log-concave densities

This paper deals with yet another class of bounds on the transport distance to a log-concave reference density, involving the transport distance W_∞ instead of the more common distance W_2. For probability measures μ, ν on $\mathbb{R}^d$, $W_\infty(\mu,\nu)$ can be defined in probabilistic terms by

$$W_\infty(\mu,\nu) = \inf_{X,Y} \left\{ \operatorname*{esssup}_{\omega\in\Omega} |X(\omega) - Y(\omega)| \right\},$$

where the infimum runs over all $\mathbb{R}^d$-valued random vectors X and Y defined on the same probability space $(\Omega, \mathcal{F}, \mathcal{P})$ with $\mathrm{law}(X) = \mu$ and $\mathrm{law}(Y) = \nu$.

Our goal is to obtain quantitative bounds on the transport distance $W_\infty(\mu,\nu)$ to a log-concave reference density μ for a large class of measures. The following result is a prototypical example, which we obtain as a consequence of our main result; see Corollary 6.2.4 below.

Proposition 6.1.1. *Let μ and ν be probability densities on $\mathbb{R}^d$. Suppose that μ is κ-log-concave for some $\kappa > 0$, and that $\nu = e^{-H}\mu$, where $H \in C(\mathbb{R}^d)$ is L-Lipschitz for some $L < \infty$. Then:*

$$W_\infty(\mu,\nu) \le \frac{L}{\kappa}\,. \tag{6.1.2}$$

This bound is sharp, as can be seen by considering two shifted isotropic Gaussian measures. Under the more restrictive assumptions that both densities μ and ν are κ-log-concave, supported on a Euclidean ball and bounded away from 0 on it, such a bound was recently proved in [CPS23, Prop. 3.1] by completely different methods.

Proposition 6.1.1 can also be formulated as a functional inequality involving the L^∞ *relative Fisher information* $\mathcal{I}_\infty(\nu \,\|\, \mu)$ defined by

$$\mathcal{I}_\infty(\nu \,\|\, \mu) = \left\| \nabla \log\Big(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\Big) \right\|_{L^\infty(\mathbb{R}^d,\mu)}$$

for sufficiently regular densities $\nu \ll \mu$. Indeed, Proposition 6.1.1 asserts that any probability density $\mu \in L^1_+(\mathbb{R}^d)$ that is κ-log-concave for some $\kappa > 0$ satisfies the L^∞ *transport-information inequality*

$$W_\infty(\mu,\nu) \le \frac{1}{\kappa}\mathcal{I}_\infty(\nu \,\|\, \mu)$$

for all sufficiently regular probability densities ν. This inequality can be viewed as an L^∞-analogue of well known L^2-based transport-information inequalities; see Section 6.2 for more details.

One of the main contributions of this paper is the insight that the estimate (6.1.2) can be improved significantly when the involved probability densities are anisotropic. Anisotropic densities are ubiquitous in applications, e.g., when densities are concentrated near a lower-dimensional manifold. To formulate the improved estimate, it will be convenient to introduce some more notation.

Let $K \in \mathbb{R}^{d\times d}$ be a symmetric matrix. A function $U : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is *K-convex* if $x \mapsto U(x) - \frac{1}{2}\langle x, Kx\rangle$ is convex. If $U \in C^2(\mathbb{R}^d)$, then U is K-convex if and only if $\nabla^2 U(x) \succcurlyeq K$ for all $x \in \mathbb{R}^d$. A function $\mu \in L^1_+(\mathbb{R}^d) \setminus \{0\}$ is said to be *K-log-concave* if $\mu = e^{-U}$ for some K-convex function U. The special case $K = \kappa I_d$ corresponds to the notions of κ-convexity and κ-log-concavity introduced above. If $K = 0$, we recover the usual notions of convexity and log-concavity.

Let A and B be orthogonal subspaces satisfying $A \oplus B = \mathbb{R}^n$, and let P_A and P_B denote the corresponding orthogonal projections. The following result (see Corollary 6.2.7 below) is a generalisation of Proposition 6.1.1, capturing different behaviour of the involved measures on the subspaces A and B. In the special case where $A = \mathbb{R}^n$ and $B = \emptyset$ we recover (6.1.2).

Theorem 6.1.2. *Let μ and ν be probability densities on $\mathbb{R}^d$. Suppose that μ is K-log-concave where $K = \kappa_A \mathsf{P}_A + \kappa_B \mathsf{P}_B$ for some $\kappa_A, \kappa_B > 0$, and that $\nu = e^{-H}\mu$, with $H \in C(\mathbb{R}^d)$ satisfying, for some $L_A < \infty$,*

$$\|H(x) - H(y)\| \leq L_A |\mathsf{P}_A(x-y)| \quad \textit{for all } x, y \in \mathbb{R}^d\,.$$

Then:

$$W_\infty(\mu,\nu) \leq \begin{cases} \frac{L_A}{\kappa_A} & \textit{if } \kappa_A \leq 2\kappa_B\,, \\ \frac{L_A}{2\sqrt{\kappa_B(\kappa_A - \kappa_B)}} & \textit{if } \kappa_A \geq 2\kappa_B\,. \end{cases}$$

In the regime $1 \leq \frac{\kappa_A}{\kappa_B} \leq 2$, observe that the constants in the denominator depends only on the directional log-concavity constant κ_A, and not on the uniform log-concavity constant κ_B.

Proposition 6.1.1 and Theorem 6.1.2 will be proved as corollaries to a general criterion (Theorem 6.2.1). The proof is based on a probabilistic argument using careful estimates for Langevin dynamics for μ and ν.

While our main results are general, our investigation is partly motivated by applications to the long-term behaviour of Fisher's infinitesimal model from quantitative genetics, as will be discussed in Section 6.1.2. The improvement of Theorem 6.1.2 over Proposition 6.1.1 is crucial to obtain sharp rates of convergence in this model, as we will discuss below.

6.1.2 Application to Fisher's infinitesimal model

Fisher's infinitesimal model from quantitative genetics describes the distribution $F_n \in L^1_+(\mathbb{R}^d)$ of a d-dimensional trait $x \in \mathbb{R}^d$ in an evolving population at discrete times $n \in \mathbb{N}_0$. The trait distribution evolves according to the rule $F_{n+1} = \mathcal{T}[F_n]$, where $\mathcal{T} = \mathcal{S} \circ \mathcal{R}$ consists of a reproduction operator $\mathcal{R}$ and a selection operator $\mathcal{S}$ acting on $L^1_+(\mathbb{R}^d)$. The reproduction operator $\mathcal{R}$ is Fisher's infinitesimal operator given by

$$\mathcal{R}[F](x) = \int_{\mathbb{R}^d\times\mathbb{R}^d} G\Big(x - \frac{x_1+x_2}{2}\Big)\frac{F(x_1)F(x_2)}{\|F\|_{L^1}}\mathrm{d}x_1\mathrm{d}x_2$$

for $F \in L^1_+(\mathbb{R}^d)$ and $x \in \mathbb{R}^d$, where $G(x) = (2\pi)^{-d/2} \exp(-|x|^2/2)$ is the standard Gaussian kernel on $\mathbb{R}^d$. We use the natural convention that $\mathcal{R}[0] = 0$. This operator describes sexual reproduction in a mean-field model where individuals mate independently and produce offspring whose traits are (isotropic) Gaussian centred at the average traits of their two parents. The operator $\mathcal{R}$ preserves the size of the population: $\|\mathcal{R}[F]\|_{L^1} = \|F\|_{L^1}$ for all $F \in L^1_+(\mathbb{R}^d)$. Selection effects are modelled using the multiplication operator $\mathcal{S}$, which is given by

$$\mathcal{S}[F](x) = e^{-m(x)} F(x)$$

for a fixed mortality function $m : \mathbb{R}^d \to [0, \infty)$. This operator reflects the idea that individuals with certain traits have a higher survival probability than others. In this paper, m will be strictly convex, which means that individuals with intermediate trait values have a higher survival probability. This is the regime of *stabilising selection*.

Fisher's infinitesimal model was introduced in [Fis19] and explicitly formulated in [Bul85]. Though the model has been influential in quantitative genetics since it was proposed, it was proved only recently that the model emerges as a limit of models subject to the laws of Mendelian inheritance when the number of discrete loci tends to infinity [BEV17]. We refer to [WL18, Ch. 24] for the biological background of various different infinitesimal models.

Long-term behaviour

Significant recent progress has been obtained in understanding the long-term behaviour of the model as $n \to \infty$ under suitable assumptions on the mortality function m. In particular, it is natural to ask whether there exists a (unique) probability distribution $\mathbf{F}$ that is *quasi-invariant* in the sense that $\mathcal{T}[\mathbf{F}] = \boldsymbol{\lambda}\mathbf{F}$ for some $\boldsymbol{\lambda} > 0$. Then one may ask whether the renormalised densities $F_n/\boldsymbol{\lambda}^n$ converge to $\mathbf{F}$ for a general class of initial probability distributions F_0, and to quantify the speed of convergence using suitable metrics or functionals.

A comprehensive investigation has been carried out in the special case of quadratic selection, namely $m(x) = \frac{\alpha}{2}|x|^2$ for some $\alpha > 0$ [CLP24]. In this situation, the model preserves the class of Gaussian distributions and it is shown that there exists a unique quasi-equilibrium $\mathbf{F}$, which is an explicit Gaussian distribution. Moreover, the authors prove exponential convergence to $\mathbf{F}$ (in the sense of relative entropy) for general initial data.

The remarkable recent paper [CPS23] treats more general uniformly convex selection in dimension 1. Namely, under the assumption that $m : \mathbb{R} \to [0, \infty]$ satisfies $m'' \geq \alpha$ for some $\alpha > 0$, the authors show the existence of a (non-explicit) β-log-concave quasi-equilibrium $\mathbf{F}$, without establishing its uniqueness. The parameter $\beta > \max\{\frac{1}{2}, \alpha\}$ depends on α in an explicit way. Moreover, [CPS23] uncovers a remarkable central role played by the L^∞ relative Fisher information. The authors show that the one-step contractivity estimate

$$\mathcal{I}_\infty(\mathcal{T}[F] \,\|\, \mathbf{F}) \leq \left(\tfrac{1}{2} + \beta\right)^{-1} \mathcal{I}_\infty(F \,\|\, \mathbf{F}) \tag{6.1.3}$$

holds for all $F \in L^1_+(\mathbb{R}^d)$. This inequality immediately yields the exponential convergence bound $\mathcal{I}_\infty(F_n \,\|\, \mathbf{F}) \leq (\frac{1}{2}+\beta)^{-n} \mathcal{I}_\infty(F_0 \,\|\, \mathbf{F})$ for all initial distributions F_0 with $\mathcal{I}_\infty(F_0 \,\|\, \mathbf{F}) < \infty$. Observe that the latter condition is a strong assumption on the initial datum F_0; e.g., if G and G' are 1-dimensional Gaussian distributions with different variances, then $\mathcal{I}_\infty(G \,\|\, G') = \infty$.

Proof of the one-step contractivity

Let us briefly discuss the strategy of the proof of (6.1.3) from [CPS23]. After proving the existence of a β-log-concave quasi-equilibrium $\mathbf{F}$, the authors consider the renormalised densities $u_n := F_n/\boldsymbol{\lambda}^n\mathbf{F}$, which satisfy the recursive equation

$$u_{n+1}(x) = \int_{\mathbb{R}^d\times\mathbb{R}^d} \frac{u_n(x_1)u_n(x_2)}{\|u_n\mathbf{F}\|_{L^1(\mathbb{R}^d)}} P(x_1,x_2;x)\mathrm{d}x_1\mathrm{d}x_2\,,$$

where $P(x_1,x_2;x)$ denotes the weighted transition rates from parental traits (x_1,x_2) to a child with trait x. These rates are given by

$$P(x_1,x_2;x) = \frac{1}{Z(x)}\mathbf{F}(x_1)\mathbf{F}(x_2)G\Big(x - \frac{x_1+x_2}{2}\Big)\,, \tag{6.1.4}$$

where $Z(x) = \int_{\mathbb{R}^d\times\mathbb{R}^d}\mathbf{F}(x_1)\mathbf{F}(x_2)G\Big(x-\frac{x_1+x_2}{2}\Big)\mathrm{d}x_1\mathrm{d}x_2$ denotes the normalising constant which ensures that $P(\cdot;x)$ is a probability distribution on $\mathbb{R}^d\times\mathbb{R}^d$ for all $x\in\mathbb{R}^d$.

The proof of the one-step contractivity estimate (6.1.3) relies on two key inequalities. Firstly, for all strictly positive initial data $u_0\in C^1(\mathbb{R}^d)$ and all $x,\tilde{x}\in\mathbb{R}^d$, it is shown in [CPS23, Lem. 2.4] that

$$|\log u_1(x) - \log u_1(\tilde{x})| \le \Big\|\nabla\log u_0\Big\|_{L^\infty(\mathbb{R}^d)} W_{\infty,1}\Big(P(\cdot;x),P(\cdot;\tilde{x})\Big)\,. \tag{6.1.5}$$

Here, $W_{\infty,1}$ denotes the ∞-Wasserstein metric over the base space $\mathbb{R}^{2d}$ endowed with the norm $|(x_1,x_2)|_1 := |x_1|+|x_2|$, with $|x_i|$ denoting the euclidean norm of $x_i\in\mathbb{R}^d$ for $i=1,2$. While (6.1.5) is stated in [CPS23] for $d=1$, the proof extends verbatim to arbitrary dimensions.

The second key inequality from [CPS23] is a sharp bound on the $W_{\infty,1}$-distance appearing in the above inequality. Namely, in the special case $d=1$, it is shown that, for all $x,\tilde{x}\in\mathbb{R}$,

$$W_{\infty,1}\Big(P(\cdot;x),P(\cdot;\tilde{x})\Big) \le \Big(\tfrac{1}{2}+\beta\Big)^{-1}|x-\tilde{x}|\,. \tag{6.1.6}$$

The inequalities (6.1.5) and (6.1.6) combined yield the crucial one-step contractivity inequality (6.1.3) for the L^∞ relative Fisher information.

However, as pointed out in [CPS23, Rem. 1.6], there are non-trivial obstacles that prevent an extension of the proof of (6.1.6) to higher dimensions. The reason is that this proof employs the Brenier map (the optimal transport map for the W_2-distance), which satisfies the Monge-Ampère equation. The required L^∞-bound on the Brenier map between $P(\cdot;x)$ and $P(\cdot;\tilde{x})\in L^1_+(\mathbb{R}^2)$ is then obtained by using a maximum principle for the Monge-Ampère equation in convex but not uniformly convex domains, exploiting recent progress on the regularity theory for the Monge-Ampère equation in two-dimensional domains with special symmetries [Jha19].

Results

In this paper we obtain a sharp multi-dimensional version of (6.1.6) by a completely different (probabilistic) method, as a consequence of Theorem 6.1.2. Using the notation from above, we first establish the existence of a quasi-invariant distribution in the multi-dimensional setting.

Theorem 6.1.3 (Existence of a quasi-equilibrium)**.** *Let $m\in C^1(\mathbb{R}^d)$ be α-convex for some $\alpha>0$. Then there exist $\boldsymbol{\lambda}\in(0,1)$ and a probability density $\mathbf{F}\in L^1_+(\mathbb{R}^d)$ such that $\mathcal{T}[\mathbf{F}]=\boldsymbol{\lambda}\mathbf{F}$. Moreover, $\mathbf{F}$ is β-log-concave, where $\beta>\max\big\{\frac{1}{2},\alpha\big\}$ satisfies $\beta=\alpha+\frac{\beta}{\frac{1}{2}+\beta}$.*

The proof of this result adapts the arguments from [CPS23], where the corresponding result was obtained for $d = 1$. The key technical tool is the L^∞-transport bound from Theorem 6.1.2, which yields a Cauchy property for a sequence of iterates, and hence a candidate quasi-equilibrium. The properties of the L^∞ relative Fisher information require us to work first with a localised problem on a bounded domain, and subsequently identify a quasi-equilibrium for the original operator $\mathcal{T}$ by an approximation procedure. The extension of this argument to higher dimensions brings additional technicalities to deal with the boundedness of the domains and to show tightness of a sequence of quasi-equilibria.

As Theorem 6.1.3 yields the existence of a quasi-equilibrium $\mathbf{F}$, we can define the weighted transition kernels $P(\cdot;x)$ by $P(x_1,x_2;x) = \frac{1}{Z(x)}\mathbf{F}(x_1)\mathbf{F}(x_2)G\Big(x - \frac{x_1+x_2}{2}\Big)$ as in (6.1.4), where $Z(x)$ denotes a normalising constant. Using Theorem 6.1.2 we obtain the following d-dimensional generalisation of (6.1.6).

Theorem 6.1.4 (W_∞-contractivity)**.** *Let $m \in C^1(\mathbb{R}^d)$ be α-convex for some $\alpha > 0$. Then:*

$$W_\infty\Big(P(\cdot;x), P(\cdot;\tilde{x})\Big) \leq 2^{-1/2}\Big(\tfrac{1}{2}+\beta\Big)^{-1}|x-\tilde{x}|$$

for all $x, \tilde{x} \in \mathbb{R}^d$, where $\beta > \max\{\frac{1}{2}, \alpha\}$ satisfies $\beta = \alpha + \frac{\beta}{\frac{1}{2}+\beta}$.

Since $W_{\infty,1} \leq \sqrt{2}W_\infty$ in view of the trivial inequality $|(x_1,x_2)|_1 \leq \sqrt{2}\,\|(x_1,x_2)\|$, this result implies the desired bound (6.1.6). Consequently, the main conclusions of [CPS23] carry over to multi-dimensional traits. The following result summarises these conclusions.

Corollary 6.1.5. *Let $m \in C^1(\mathbb{R}^d)$ be α-convex for some $\alpha > 0$, and let $(\boldsymbol{\lambda}, \mathbf{F})$ be as in Theorem 6.1.3. Take $0 \neq F_0 \in L^1_+(\mathbb{R}^d)$ with $\mathcal{I}_\infty(F_0 \,\|\, \mathbf{F}) < \infty$, and set $F_n = \mathcal{T}^n[F_0]$ for $k \geq 0$. Then:*

(i) (Convergence of the relative L^∞-Fisher information) For all $n \in \mathbb{N}$ we have

$$\mathcal{I}_\infty(F_n \,\|\, \mathbf{F}) \leq \Big(\tfrac{1}{2}+\beta\Big)^{-n}\mathcal{I}_\infty(F_0 \,\|\, \mathbf{F})\,.$$

(ii) (Convergence of the relative entropy) There exists a constant $C > 0$ depending on F_0 such that for all $n \in \mathbb{N}$ we have

$$\mathcal{D}_{\mathrm{KL}}\left(\frac{F_n}{\|F_n\|_{L^1}} \,\middle\|\, \mathbf{F}\right) \leq C\Big(\tfrac{1}{2}+\beta\Big)^{-2n} \quad \textit{and} \quad \left|\frac{\|F_n\|_{L^1}}{\|F_{n-1}\|_{L^1}} - \boldsymbol{\lambda}\right| \leq C\Big(\tfrac{1}{2}+\beta\Big)^{-n}.$$

One may wonder whether analogues of the contraction property in ((i)) hold with the same rate for functionals other than $\mathcal{I}_\infty(\cdot \,\|\, \mathbf{F})$, such as the relative entropy and the relative L^2-Fisher information. In Section 6.4 we show that this is not the case, not even in the setting of quadratic selection ($m(x) = \frac{\alpha}{2}|x|^2$) and Gaussian initial data. We refer the reader to Section 6.4 for the details.

6.1.3 Structure of the paper

Section 6.2 deals with L^∞-optimal transport bounds for perturbations of log-concave densities, containing a general criterion (Theorem 6.2.1) and the proofs of Proposition 6.1.1 and Theorem 6.1.2. The applications to Fisher's infinitesimal model, and in particular the proof of Theorems

6.1.3 and 6.1.4 and Corollary 6.1.5, can be found in Section 6.3. The discussion after Corollary 6.1.5 is expanded in Section 6.4, which deals with the relative L^2-Fisher information and the relative entropy instead of the relative L^∞-Fisher information. Finally, Section 6.5 contains two lemmas on log-concave distributions that are used in the proof of Theorem 6.1.3 in Section 6.3.

6.1.4 Notation and preliminaries

Let $L^1_+(\mathbb{R}^d)$ denote the cone of non-negative functions in $L^1(\mathbb{R}^d)$. Throughout the paper, we identify (probability) densities in $L^1_+(\mathbb{R}^d)$ with the corresponding (probability) measures.

Weak convergence of densities (or measures) denotes convergence in duality with bounded continuous functions. We will frequently use that $(\mu,\nu)\mapsto W_\infty(\mu,\nu)$ is jointly continuous with respect to weak convergence of probability measures. This follows from the corresponding result for W_p, since $W_p \to W_\infty$ pointwise as $p\to\infty$; see [GS84].

Definition 6.1.6. *Suppose that $\mu \in L^1_+(\mathbb{R}^d)$ is a κ-log-concave density for some $\kappa\in\mathbb{R}$, not necessarily normalised, so that $\operatorname{Supp}\mu$ is closed and convex. If $\nu\in L^1_+(\mathbb{R}^d)$ satisfies $\nu \ll \mu$ and $\log\Big(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\Big) = f$ μ-a.e. for some Lipschitz function $f\colon \operatorname{Supp}\mu\to\mathbb{R}$, then*

$$\mathcal{I}_\infty(\nu\,\|\,\mu) := \sup\left\{\frac{|f(x)-f(y)|}{|x-y|} \ :\ x,y\in\operatorname{Supp}\mu\,,\ x\neq y\right\}. \tag{6.1.7}$$

Otherwise, $\mathcal{I}_\infty(\nu\,\|\,\mu) := +\infty$.

Remark 6.1.7. *In particular, if $\nu\ll\mu$ and $\log\Big(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\Big) = f$ μ-a.e. for some $f\in C^1(\operatorname{Supp}\mu)$, then $\mathcal{I}_\infty(\nu\,\|\,\mu) = \|\nabla f\|_{L^\infty(\mathbb{R}^d,\mu)}$.*

The relative entropy (or Kullback-Leibler divergence) of a probability density ν with respect to a probability density μ is defined by

$$\mathcal{D}_{\mathrm{KL}}(\nu\,\|\,\mu) = \begin{cases} \displaystyle\int_{\mathbb{R}^d}\rho\log\rho\,\mathrm{d}\mu & \text{if } \nu\ll\mu \text{ with } \rho := \frac{\mathrm{d}\nu}{\mathrm{d}\mu}\,,\\ +\infty & \text{otherwise}\,. \end{cases} \tag{6.1.8}$$

$B_r(x)$ denotes the open ball of radius $r>0$ around $x\in\mathbb{R}^d$. Its closure will be denoted $\overline{B}_r(x)$

$\gamma_{\mu,C}$ denotes the centred Gaussian density with mean $\mu\in\mathbb{R}^d$ and covariance matrix $C\in\mathbb{R}^{d\times d}$. If $\mu=0$ we simply write γ_C.

The following well-known property of log-concave densities will be useful in the sequel; see, e.g., [SW14, Thm. 3.7.2].

Lemma 6.1.8 (Preservation of log-concavity)**.** *For $i=1,2$, let $\mu_i\in L^1_+(\mathbb{R}^d)$ be K_i-log-concave for some matrix $K_i\in\mathbb{R}^{d\times d}$ with $K_i\succ 0$. Then $\mu_1 * \mu_2$ is K-log-concave with*

$$K^{-1} = K_1^{-1} + K_2^{-1}\,.$$

We also use the following well-known result in the reverse direction; see [EL18, Lem. 1.3].

Lemma 6.1.9 (Log-convexity along the heat flow)**.** *Let μ be a probability measure on $\mathbb{R}^d$. For any $t>0$ the probability density $\mu_t := \mu * \gamma_{tI_d}$ is $(-\frac{1}{t})$-log-convex, in the sense that, for all $x\in\mathbb{R}^d$,*

$$\nabla^2\Big(-\log\mu_t(x)\Big) \preccurlyeq \frac{1}{t}I_d\,. \tag{6.1.9}$$

Acknowledgement

This research was funded in part by the Austrian Science Fund (FWF) project 10.55776/F65 and the Austrian Academy of Science, DOC fellowship nr. 26293.

6.2 L^∞-optimal transport of log-concave measures

In this section we present several bounds for the ∞-Wasserstein distance $W_\infty(\mu, \nu)$ between a log-concave measure μ and a log-Lipschitz perturbation ν. Unless specified otherwise, the Wasserstein distance is taken with respect to the Euclidean distance on the underlying space. Our bounds will be derived from the following general criterion.

Theorem 6.2.1. *Let μ and ν be probability densities on $\mathbb{R}^d$ satisfying the following assumptions:*

(i) μ is K-log-concave for some matrix $K \in \mathbb{R}^{d\times d}$ with $K \succ 0$.

(ii) $\nu = e^{-H}\mu$ with $H \in C(\mathbb{R}^d)$ satisfying

$$|H(y) - H(x)| \leq \ell(x - y) \quad \textit{for all } x, y \in \mathbb{R}^d\,,$$

for some positively 1-homogeneous function $\ell \in C(\mathbb{R}^d)$.

Then we have

$$W_\infty(\mu, \nu) \leq M\,,$$

where

$$M := \sup_{z\in\mathbb{R}^d}\Big\{|z| \,:\, \langle z, Kz\rangle \leq \ell(z)\Big\}\,.$$

Remark 6.2.2. *Note that the assumptions imply that H is Lipschitz continuous with Lipschitz constant $L := \sup_{|z|=1} \ell(z)$. A possible choice of ℓ is given by $\ell(z) = L\|z\|$. However, it is important to allow for other choices of ℓ which take anisotropy into account. This will indeed be crucial to get optimal bounds in our application to the Fisher model. When $H \in C^1(\mathbb{R}^d)$, the assumed bound on H can be written equivalently as*

$$\langle \nabla H(x), z\rangle \leq \ell(z) \quad \textit{for all } x, z \in \mathbb{R}^d\,.$$

Proof. The proof consists of three steps.

Step 1. Suppose first that $\mu = e^{-U}$ for some $U \in C^2(\mathbb{R}^d)$ such that ∇U is Lipschitz, and that $H \in C^1(\mathbb{R}^d)$. It then follows from the standard theory of stochastic differential equations [KS91, Thm. 5.2.9] that there exists a unique strong solution to the following system of SDEs, driven by *the same Brownian motion B_t*, for all times $t \geq 0$:

$$\mathrm{d}X_t = -\nabla U(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t\,, \qquad X_0 \sim \nu\,, \tag{6.2.1}$$
$$\mathrm{d}Y_t = -\nabla U(Y_t)\mathrm{d}t - \nabla H(Y_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t\,, \qquad Y_0 = X_0\,. \tag{6.2.2}$$

Subtracting these equations in their integral form we note that the Brownian term vanishes, and since X and Y have a.s. continuous sample paths, we infer that the sample paths of

$Z := X - Y$ are continuously differentiable a.s. Using the chain rule and our assumptions, we find

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|Z_t|^2 &= -\Big\langle X_t - Y_t, \nabla U(X_t) - \nabla U(Y_t)\Big\rangle + \Big\langle X_t - Y_t, \nabla H(Y_t)\Big\rangle \\ &\le -\langle Z_t, K\, Z_t\rangle + \ell(Z_t)\,.\end{aligned}$$

Observe now that, for any differentiable function $h\colon \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ with $h(0) = 0$ we have $\sup h = \sup_{x\ge 0}\{h(x) : h'(x) \ge 0\}$. Applying this identity to $h(t) = \frac{1}{2}|Z_t|^2$, we obtain

$$|X_t - Y_t| \le M \quad \text{for all } t \ge 0\,. \tag{6.2.3}$$

Since μ is strongly log-concave and X_0 has finite second moment, $\mathrm{law}(X_t)$ converges to μ in W_2-distance as $t \to \infty$, hence weakly. Using the joint lower semicontinuity of W_∞ with respect to weak convergence [GS84] we deduce that $W_\infty(\mu, \nu) \le M$.

Step 2. We now remove the extra assumptions on μ. To this end, set $\mu_n = \mu * \gamma_{\frac{1}{n}I_d}$ and define the probability density $\nu_n \propto e^{-H}\mu_n$. Note that $U_n = -\log \mu_n$ is smooth with

$$K_n := \Big(K^{-1} + \frac{1}{n}I_d\Big)^{-1} \preccurlyeq \nabla^2 U_n \preccurlyeq \frac{1}{n}I_d$$

by Lemma 6.1.8. Therefore, we are in a position to apply Step 1 and we obtain the bound $W_\infty(\mu_n, \nu_n) \le M_n$, where

$$M_n := \sup_{z\in\mathbb{R}^d}\Big\{|z| \,:\, \langle z, K_n z\rangle \le \ell(z)\Big\}\,.$$

Note that $\mu_n \to \mu$ weakly. Moreover, Lemma 6.2.3 below implies that $\nu_n \to \nu$ weakly too. Hence, using again the joint lower semicontinuity of W_∞ with respect to weak convergence we find

$$W_\infty(\mu, \nu) \le \liminf_{n\to\infty} W_\infty(\mu_n, \nu_n) \le \liminf_{n\to\infty} M_n\,.$$

It thus remains to show that $M_n \to M$.

For this purpose, we define the sets

$$C_n = \Big\{z \in \mathbb{R}^d \,:\, \langle z, K_n z\rangle \le \ell(z)\Big\} \quad \text{and} \quad C = \Big\{z \in \mathbb{R}^d \,:\, \langle z, K z\rangle \le \ell(z)\Big\}\,.$$

Since $t \mapsto t^{-1}$ is operator monotone (see, e.g., [Car10, Lemma 2.7]), we have $\langle z, K_n z\rangle \ge \langle z, K_{n-1} z\rangle$ for all z, hence $C_n \subseteq C_{n-1}$ and $M_n \le M_{n-1}$. Moreover, since $\langle z, K_n z\rangle \to \langle z, K z\rangle$ monotonically for all z, we have $C = \bigcap_n C_n$.

Using the continuity and the positive 1-homogeneity of ℓ, we infer that the sets C_n are non-empty and compact. Consequently, there exists $z_n \in C_n \subseteq C_1$ with $|z_n| = M_n$. Since C_1 is compact, we may extract a subsequence $\{z_{n_k}\}_k$ converging to some $\hat z \in C_1$. Since each C_m is closed, and since $z_{n_k} \in C_m$ whenever $n_k \ge m$, it follows that $\hat z \in C_m$, hence $\hat z \in \bigcap_m C_m = C$. Therefore, $M \ge |\hat z| = \lim_{k\to\infty} |z_{n_k}| = \lim_{k\to\infty} M_{n_k}$. Since $M \le M_n \le M_{n-1}$ for all n, it follows that $\lim_{n\to\infty} M_n = M$.

Step 3. We remove the differentiability assumptions on H. Write

$$L := \sup_{\|x\|=1} \ell(x) < \infty\,,$$

so that H is L-Lipschitz. Let $j\colon \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ be a smooth mollifier supported in the unit ball of $\mathbb{R}^d$. We write $j_n(x) := n^d j(nx)$ and $H_n := j_n * H$, so that

$$H_n(x) = n^d \int_{\mathbb{R}^d} H(x-y) j(ny)\mathrm{d}y = \int_{\mathbb{R}^d} H\Big(x - \frac{y}{n}\Big) j(y)\mathrm{d}y\,.$$

Since $\operatorname{Supp} j \subseteq B_1(0)$, we have for all $x, y \in \mathbb{R}^d$,

$$|H_n(x) - H(x)| \leq \frac{L}{n}\,, \tag{6.2.4}$$

$$|H_n(x) - H_n(y)| \leq \ell(x-y) \leq L\|x-y\|\,. \tag{6.2.5}$$

Define the probability measures $\nu_n \propto e^{-H_n}\mu$. Since H_n is a smooth function satisfying (6.2.5), an application of Step 2 yields

$$W_\infty(\mu, \nu_n) \leq M\,.$$

Hence, since W_∞ is jointly weakly lower semicontinuous, it suffices to show that $\nu_n \to \nu$ weakly. For this purpose, it is in turn sufficient to prove that e^{-H_n} converges to e^{-H} in $L^1(\mu)$, which we will do next.

Fix $\varepsilon > 0$. Since μ is κ-log-concave with $\kappa > 0$, we have $-\log \mu(x) \geq \frac{\kappa}{2}|x - \bar{x}|^2$ for some $\bar{x} \in \mathbb{R}^d$. Furthermore, since $|H(x)| \leq |H(0)| + L|x|$, (6.2.4) implies that $|H_n(x)| \leq C + L\|x\|$ with $C := |H(0)| + L$. Therefore, there exists $R > 0$ such that, for all $n \geq 1$,

$$\int_{B_R(0)^c} e^{-H_n}\mathrm{d}\mu + \int_{B_R(0)^c} e^{-H}\mathrm{d}\mu \leq \frac{\varepsilon}{2}\,.$$

Furthermore, since the function $x \mapsto e^{-x}$ is uniformly continuous on bounded intervals, (6.2.4) implies that there exists $\bar{n} \geq 1$ such that for all $n \geq \bar{n}$,

$$\sup_{x \in \overline{B}_R(0)} \Big|e^{-H_n(x)} - e^{-H(x)}\Big| \leq \frac{\varepsilon}{2}\,.$$

Consequently, for $n \geq \bar{n}$,

$$\int_{\mathbb{R}^d} \Big|e^{-H_n} - e^{-H}\Big|\mathrm{d}\mu \leq \int_{B_R(0)^c} e^{-H_n}\mathrm{d}\mu + \int_{B_R(0)^c} e^{-H}\mathrm{d}\mu + \sup_{\overline{B}_R(0)} \Big\|e^{-H_n} - e^{-H}\Big\| \leq \varepsilon\,,$$

which implies that $e^{-H_n} \to e^{-H}$ in $L^1(\mu)$ as $n \to \infty$. ∎

Lemma 6.2.3. *Let $\mu \in L^1_+(\mathbb{R}^d)$ be a K-log-concave probability density for some matrix $K \in \mathbb{R}^{d\times d}$ with $K \succ 0$, and define $\mu_n = \mu * \gamma_{\frac{1}{n}I_d}$ for $n \geq 1$. Then*

$$\int f\mathrm{d}\mu_n \to \int f\mathrm{d}\mu$$

for all continuous functions $f\colon \mathbb{R}^d \to \mathbb{R}$ satisfying $|f(x)| \leq C\exp(C\|x\|)$ for some $C > 0$.

Proof. We show first that the integrals above are finite. Let f be as in the statement, let $X \sim \mu$ and $Z \sim \gamma_{I_d}$ be independent, and set $X_n = X + \frac{Z}{\sqrt{n}}$.

Let $\kappa > 0$ be the smallest eigenvalue of K, and fix $\kappa' \in (0, \kappa)$. It follows from Lemma 6.1.8 that μ_n is κ'-log-concave for all n sufficiently large. Therefore, the Bakry-Émery criterion implies that the measures μ and μ_n satisfy a logarithmic Sobolev inequality with the same constant. Using this, the growth assumption on f, and the fact that $\mathbb{E}[X_n] = \mathbb{E}[X]$, the so-called Herbst argument [BGL14, Prop. 5.4.1] implies that $f(X_n) \in L^2$ and that the sequence $\{f(X_n)\}_n$ is bounded in L^2. In particular, $f(X), f(X_n) \in L^1$, hence the integrals above are finite. It remains to show that

$$\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)].$$

For this purpose, note first that $X_n \to X$ in probability. Since f is continuous, $f(X_n) \to f(X)$ in probability as well; see, e.g. [Kal21, Lem. 5.3]. Therefore, to conclude that $f(X_n) \to f(X)$ in L^1 it suffices to show that $\{f(X_n)\}_n$ is uniformly integrable; see, e.g. [Kal21, Thm. 5.12]. But this follows from the fact that the sequence $\{f(X_n)\}_n$ is bounded in L^2, which we proved above. ■

6.2.1 Isotropic case

The simplest non-trivial case of Theorem 6.2.1 is the following estimate, which we stated as Proposition 6.1.1 above.

Corollary 6.2.4. *Let μ and ν be probability densities on $\mathbb{R}^d$. Suppose that μ is κ-log-concave for some $\kappa > 0$, and that $\nu = e^{-H}\mu$, where $H \in C(\mathbb{R}^d)$ is L-Lipschitz for some $L < \infty$. Then:*

$$W_\infty(\mu, \nu) \le \frac{L}{\kappa}\,. \tag{6.2.6}$$

Proof. This is an application of Theorem 6.2.1 with $K = \kappa I_d$ and $\ell(z) = L|z|$. ■

The following result is a reformulation of Corollary 6.2.4 as a functional inequality.

Theorem 6.2.5 (∞-Transport-Information Inequality)**.** *Let $\mu \in L^1_+(\mathbb{R}^d)$ be a κ-log-concave probability density for some $\kappa > 0$. Then the* transport-information inequality

$$W_\infty(\mu, \nu) \le \frac{1}{\kappa}\mathcal{I}_\infty(\nu \,\|\, \mu) \tag{6.2.7}$$

holds for all probability densities $\nu \in L^1_+(\mathbb{R}^d)$.

Proof. Suppose that $\mathcal{I}_\infty(\nu \,\|\, \mu) < +\infty$; otherwise there is nothing to prove. In view of Definition 6.1.6 there exists a Lipschitz function $h : \operatorname{Supp}\mu \to \mathbb{R}$ with Lipschitz constant $L := \mathcal{I}_\infty(\nu \,\|\, \mu)$, that agrees with $\log\!\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right)$ μ-a.e.. By the Kirszbraun theorem, h can be extended to a Lipschitz function H on $\mathbb{R}^d$ with the same Lipschitz constant L. Since $\nu = e^{-H}\mu$, the result follows from Corollary 6.2.4. ■

Remark 6.2.6. *The inequality* (6.2.7) *is an L^∞-analogue of the well-known L^2-based transport-information inequality*

$$W_2(\mu,\nu) \le \frac{1}{\kappa}\sqrt{\mathcal{I}_2(\nu \,\|\, \mu)}\ , \tag{6.2.8}$$

where $\mathcal{I}_2(\nu \,\|\, \mu)$ denotes the L^2-relative Fisher Information, i.e. $\mathcal{I}_2(\nu \,\|\, \mu) := \left\|\nabla \log\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right)\right\|^2_{L^2(\nu)}$ for sufficiently regular densities ν.

The latter inequality holds under the assumption that μ satisfies a logarithmic Sobolev inequality $\mathcal{D}_{\mathrm{KL}}(\nu \,\|\, \mu) \le \frac{1}{2\kappa}\mathcal{I}_2(\nu \,\|\, \mu)$, and thus for every κ-log concave measure μ by the Bakry–Émery criterion. To prove (6.2.8), *note that the logarithmic Sobolev inequality implies the transport-entropy inequality $W_2(\mu,\nu) \le \sqrt{\frac{2}{\kappa}\mathcal{D}_{\mathrm{KL}}(\nu \,\|\, \mu)}$ by the work of Otto and Villani [OV00]. Combining these two inequalities immediately yields* (6.2.8)*. For a systematic study of transport-information inequalities we refer to [GLWY09].*

6.2.2 Anisotropic case

We will now develop a more refined criterion, that yields improved bounds in situations where the measures behave differently in different directions. Let A and B be non-empty subspaces of $\mathbb{R}^d$ that are orthogonal and satisfy $A \oplus B = \mathbb{R}^d$. Let P_A and P_B be the corresponding orthogonal projections.

Corollary 6.2.7. *Let μ and ν be probability densities on $\mathbb{R}^d$ satisfying the following assumptions:*

(i) μ is K-log-concave, with $K = \kappa_A \mathsf{P}_A + \kappa_B \mathsf{P}_B$ for some $\kappa_A, \kappa_B > 0$.

(ii) $\nu = e^{-H}\mu$ with $H \in C(\mathbb{R}^d)$ satisfying, for some $L_A < \infty$,

$$\|H(x) - H(y)\| \le L_A |\mathsf{P}_A(x-y)| \quad \textit{for all } x, y \in \mathbb{R}^d\,.$$

Then:

$$W_\infty(\mu,\nu) \le \begin{cases} \frac{L_A}{\kappa_A} & \textit{if } \kappa_A \le 2\kappa_B\,,\\ \frac{L_A}{2\sqrt{\kappa_B(\kappa_A-\kappa_B)}} & \textit{if } \kappa_A \ge 2\kappa_B\,. \end{cases} \tag{6.2.9}$$

Proof. Applying Theorem 6.2.1 with $K = \kappa_A \mathsf{P}_A + \kappa_B \mathsf{P}_B$ and $\ell(z) = L_A|\mathsf{P}_A z|$, we infer that $W_\infty(\mu,\nu) \le M$, where

$$M := \sup_{z_A, z_B \ge 0}\left\{\sqrt{z_A^2 + z_B^2} \ :\ \kappa_A z_A^2 + \kappa_B z_B^2 \le L_A z_A\right\}. \tag{6.2.10}$$

Performing the maximisation over z_B first, we observe that

$$\begin{aligned} M^2 &= \sup\left\{z_A^2 + \frac{1}{\kappa_B}\Big(L_A z_A - \kappa_A z_A^2\Big) \ :\ 0 \le z_A \le \frac{L_A}{\kappa_A}\right\}\\ &= \frac{1}{\kappa_B}\sup\left\{p(z_A) \ :\ 0 \le z_A \le \frac{L_A}{\kappa_A}\right\}, \end{aligned}$$

where $p(z) = L_A z - (\kappa_A - \kappa_B)z^2$. We now distinguish two cases.

If $\kappa_A \le 2\kappa_B$, then p is non-decreasing on the interval $[0, L_A/\kappa_A]$. Therefore, the supremum of p on $[0, L_A/\kappa_A]$ is attained at the right endpoint of this interval, hence

$$M^2 = \frac{1}{\kappa_B} p\Big(\frac{L_A}{\kappa_A}\Big) = \frac{L_A^2}{\kappa_A^2}.$$

If $\kappa_A > 2\kappa_B$, then p attains its global maximum in the open interval $(0, L_A/\kappa_A)$, at $z := \frac{L_A}{2(\kappa_A - \kappa_B)}$. Therefore,

$$M^2 = \frac{1}{\kappa_B} p\Big(\frac{L_A}{2(\kappa_A - \kappa_B)}\Big) = \frac{L_A^2}{4\kappa_B(\kappa_A - \kappa_B)},$$

as desired. ∎

Remark 6.2.8. *Note that the right-hand side of* (6.2.9) *involves the the ratio of a "directional Lipschitz constant" and an "effective convexity parameter". In this sense, the bound has the same form as* (6.2.6)*. The bound* (6.2.9) *is sharp for $\kappa_A \le 2\kappa_B$, as we will see in the application to the Fisher model below.*

We finally state a corollary that will be used in the application to Fisher's infinitesimal model. Let $F = e^{-V}$ be a κ-log-concave probability density on $\mathbb{R}^{2d}$ for some $\kappa > 0$. For $x \in \mathbb{R}^d$ we consider the probability density $P(\cdot; x)$ on $\mathbb{R}^{2d}$ defined by

$$P(x_1, x_2; x) = \frac{1}{Z_x} \exp\Big(-V(x_1, x_2) - \frac{1}{2}\Big|x - \frac{x_1 + x_2}{2}\Big|^2\Big), \tag{6.2.11}$$

where $Z_x > 0$ is the normalising constant which ensures that P_x is a probability density. The transition rates appearing in the Fisher model are precisely of this form; see Theorem 6.1.4.

Corollary 6.2.9. *Let F be a κ-log-concave probability density on $\mathbb{R}^{2d}$ for some $\kappa > \frac{1}{2}$. Then, for any $x, \tilde{x} \in \mathbb{R}^d$,*

$$W_\infty\Big(P(\cdot; x), P(\cdot; \tilde{x})\Big) \le \frac{1}{\frac{1}{2} + \kappa} \frac{|x - \tilde{x}|}{\sqrt{2}}. \tag{6.2.12}$$

Before proving this result, we first show that an application of the isotropic criterion from Corollary 6.2.4 yields a suboptimal result. For ease of notation, suppose that $V \in C^2(\mathbb{R}^{2d})$. Fix $x, \tilde{x} \in \mathbb{R}^d$ and let us write $\mu_x = e^{-U} := P(\cdot; x)$ and $\mu_{\tilde{x}} = e^{-H}\mu := P(\cdot; \tilde{x})$. Then:

$$\nabla^2 U(x_1, x_2) = \nabla^2 V(x_1, x_2) + \frac{1}{4}\begin{pmatrix} I_d & I_d \\ I_d & I_d \end{pmatrix} \quad \text{and} \quad \nabla H(x_1, x_2) = \frac{1}{2}\begin{pmatrix} x - \tilde{x} \\ x - \tilde{x} \end{pmatrix}.$$

Taking into account that $\nabla^2 V \succcurlyeq \kappa I_{2d}$ by assumption, we have the bounds

$$\nabla^2 U(x) \succcurlyeq \kappa I_{2d} \quad \text{and} \quad |\nabla H(x)| \le \frac{|x - \tilde{x}|}{\sqrt{2}}. \tag{6.2.13}$$

An application of Corollary 6.2.4 then yields the estimate $W_\infty(\mu_x, \mu_{\tilde{x}}) \le \frac{|x - \tilde{x}|}{\kappa\sqrt{2}}$, which is weaker than the desired inequality (6.2.12). (In particular, in the application to the Fisher model, where $\kappa = \beta$, the comparison of norms $|x|_1 \le \sqrt{2}|x|_2$ implies that $W_{\infty,1}(\mu_x, \mu_{\tilde{x}}) \le \frac{|x - \tilde{x}|}{\beta}$, which is weaker than the desired inequality (6.1.6).)

The following proof crucially exploits anisotropy to obtain the sharp constant.

Proof of Corollary 6.2.9. Consider the orthogonal decomposition of $\mathbb{R}^{2d}$ into symmetric and anti-symmetric vectors: $\mathbb{R}^{2d} = \mathbb{R}^{2d}_{\mathrm{s}} \oplus \mathbb{R}^{2d}_{\mathrm{a}}$, where

$$\mathbb{R}^{2d}_{\mathrm{s}} := \left\{ \begin{pmatrix} x \\ x \end{pmatrix} \in \mathbb{R}^{2d} \ : \ x \in \mathbb{R}^d \right\} \quad \text{and} \quad \mathbb{R}^{2d}_{\mathrm{a}} := \left\{ \begin{pmatrix} x \\ -x \end{pmatrix} \in \mathbb{R}^{2d} \ : \ x \in \mathbb{R}^d \right\}.$$

The corresponding orthogonal projections $\mathsf{P}_{\mathrm{s}}, \mathsf{P}_{\mathrm{a}} : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ have the form

$$\mathsf{P}_{\mathrm{s}} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} x_1 + x_2 \\ x_1 + x_2 \end{pmatrix} \quad \text{and} \quad \mathsf{P}_{\mathrm{a}} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} x_1 - x_2 \\ x_2 - x_2 \end{pmatrix}.$$

The crucial observation is now that the isotropic bounds (6.2.13) can be replaced by more refined estimates that take into account how U and H behave in symmetric and anti-symmetric directions. Namely, since $\left(\begin{smallmatrix} I_d & I_d \\ I_d & I_d \end{smallmatrix} \right) = 2\mathsf{P}_{\mathrm{s}}$, we have the following improvement over (6.2.13):

$$\nabla^2 U(x) \succcurlyeq \Big(\frac{1}{2} + \kappa\Big)\mathsf{P}_{\mathrm{s}} + \kappa \mathsf{P}_{\mathrm{a}} \quad \text{and} \quad |\langle \nabla H(x), z\rangle| \leq \frac{|x - \tilde{x}|}{\sqrt{2}} |\mathsf{P}_{\mathrm{s}} z|.$$

(The first inequality holds when $V \in C^2(\mathbb{R}^{2d})$. In the general case, the corresponding nonsmooth statement holds, which asserts that U is K-convex with $K = \big(\frac{1}{2} + \kappa\big)\mathsf{P}_{\mathrm{s}} + \kappa \mathsf{P}_{\mathrm{a}}$.) Therefore, an application of Corollary 6.2.7 to $A = \mathbb{R}^{2d}_{\mathrm{s}}$ and $B = \mathbb{R}^{2d}_{\mathrm{a}}$ with parameters

$$\kappa_A = \frac{1}{2} + \kappa, \quad \kappa_B = \kappa, \quad L_A = \frac{|x - \tilde{x}|}{\sqrt{2}},$$

yields, if $\kappa \geq \frac{1}{2}$,

$$W_\infty(\mu_x, \mu_{\tilde{x}}) \leq \frac{1}{\frac{1}{2} + \kappa} \frac{|x - \tilde{x}|}{\sqrt{2}},$$

which is the desired inequality. ■

Remark 6.2.10 (Optimality)**.** *The constants in* (6.2.12) *are sharp. In fact, it was observed in [CPS23, Remark 2.7] that equality holds in the context of Fisher's infinitesimal model with quadratic selection in dimension* 1*, which means that* $m(x) = \frac{\alpha}{2} x^2$ *with* $\alpha > 0$*. In this case, we have* $V(x_1, x_2) = \frac{\beta}{2}(x_1^2 + x_2^2)$*, with* $\beta > \frac{1}{2}$ *as in Theorem 6.1.3. The measures* $P(\cdot; x)$ *are then Gaussian with mean* $(\frac{1}{2} + \beta)^{-1}(\frac{x}{2}, \frac{x}{2})$ *and the same covariance matrix. The* W_∞*-distance between two such measures is simply the euclidean distance between the respective means, which corresponds to the right-hand side in* (6.2.12)*.*

To show that this bound can not be improved, take arbitrary densities μ *and* ν *with finite first moment, and random variables* X *and* Y *with marginals* μ *and* ν *respectively. Then:*

$$\left| \int x \mathrm{d}\mu(x) - \int x \mathrm{d}\nu(x) \right| = \left| \mathbb{E}[X - Y] \right| \leq \mathbb{E}[|X - Y|],$$

which implies that, $|\int x \mathrm{d}\mu(x) - \int x \mathrm{d}\nu(x)| \leq W_\infty(\mu, \nu)$*.*

6.2.3 Boundedness of the forward-flow transport map

In this subsection we sketch an alternative argument to prove the transport bound of Corollary 6.2.4. Instead of constructing a suitable coupling, we provide an upper bound on the displacement of the *forward-flow map*, whose inverse is the so-called *Langevin transport map*. The Langevin transport map and the forward-flow map were introduced by Kim and Milman [KM12] in their work on generalisations of Cafferelli's contraction theorem [Caf00]. Subsequently, there has been a lot of interest in Lipschitz bounds for the forward-flow map [MS23, FMS24, Nee22, KP21], as such bounds allow one to transfer functional inequalities from log-concave measures to their image under the forward-flow map. Here we show that L^∞-bounds can be obtained as well.

As we already provided a rigorous proof of Corollary 6.2.4 by a different method, we keep the arguments in this section formal, so as not to obscure the main ideas. In particular, we do not discuss the delicate issues of existence of flow maps. For more details on the construction and rigorous justifications we refer the reader to [OV00, KM12, MS23, FMS24].

Construction of the forward-flow map Consider probability densities μ and ν. Here we assume that $\mu = e^{-U}$ and $\nu = e^{-H}\mu$ with smooth $U, H \colon \mathbb{R}^d \to \mathbb{R}$. Moreover, μ is assumed to be κ-log-concave (i.e., $\nabla^2 U \geq \kappa I_d$) for some $\kappa > 0$ and ν is a log-Lipschitz perturbation (i.e. $|\nabla H| \leq L$ for some $L < \infty$).

We shall briefly and informally describe the construction of the forward-flow map $S \colon \mathbb{R}^d \to \mathbb{R}^d$, which pushes-forward ν onto μ (i.e., $S_\# \nu = \mu$), referring the reader to the aforementioned references for details.

The key idea is to interpolate between ν and μ using the Langevin dynamics

$$X_0 \sim \nu, \qquad \mathrm{d}X_t = -\nabla U(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t\,.$$

Denoting $\rho_t := \mathrm{law}(X_t)$, we have $\rho_0 = \nu$ and $\rho_t \to \mu$ weakly as $t \to \infty$. Moreover, ρ_t satisfies the Fokker-Planck equation, which we formulate here as a continuity equation

$$\partial_t \rho_t - \nabla \cdot \Big(\rho_t \nabla \log f_t\Big) = 0\,, \tag{6.2.14}$$

where $f_t := \frac{\mathrm{d}\rho_t}{\mathrm{d}\mu}$. Since $f_0 = \frac{\mathrm{d}\nu}{\mathrm{d}\mu} = e^{-H}$, our assumptions imply the pointwise bound $|\nabla \log f_0| = |\nabla H| \leq L$. We will show that

$$|\nabla \log f_t| \leq L e^{-\kappa t} \tag{6.2.15}$$

for all $t \geq 0$.

For this purpose, let $(P_t)_{t\geq 0}$ be the transition semigroup associated to the Langevin dynamics, and note that $f_t = P_t f_0$ by reversibility. Since μ is κ-log-concave, the Bakry-Émery theory [BGL14, Thm. 3.3.18] implies the pointwise gradient estimate

$$|\nabla P_t f| \leq e^{-\kappa t} P_t |\nabla f| \tag{6.2.16}$$

for all sufficiently regular $f : \mathbb{R}^d \to \mathbb{R}$. Using this inequality, the inequality $|\nabla f_0| \leq L f_0$, and the positivity of P_t, we obtain

$$|\nabla f_t| \leq e^{-\kappa t} P_t |\nabla f_0| \leq L e^{-\kappa t} P_t f_0 = L e^{-\kappa t} f_t\,,$$

which yields the claimed bound (6.2.15).

For $t \geq 0$, consider the flow map $S_t : \mathbb{R}^d \to \mathbb{R}^d$ associated to the vector field $(t, x) \mapsto -\nabla \log f_t(x)$, which satisfies

$$S_0(x) = x\,, \qquad \frac{\mathrm{d}}{\mathrm{d}t} S_t(x) = -\nabla \log f_t\Big(S_t(x)\Big)\,.$$

Then, by construction, $(S_t)_\# \nu = \rho_t$, and for $0 \le s \le t$, (6.2.15) yields

$$\|S_s - S_t\|_{L^\infty} \le \int_s^t \|\nabla \log(f_r \circ S_r)\|_{L^\infty} \mathrm{d}r \le \frac{L}{\kappa}\Big(e^{-\kappa s} - e^{-\kappa t}\Big). \tag{6.2.17}$$

Passing to the limit, it is now simple to deduce that the forward-flow map $S = \lim_{t\to\infty} S_t$ is well-defined, that $S_\# \nu = \mu$, and that

$$\|S - I\|_{L^\infty(\mathbb{R}^d)} \le \frac{L}{\kappa}. \tag{6.2.18}$$

This is the desired bound, which immediately implies the bound $W_\infty(\mu,\nu) \le \frac{L}{\kappa}$ from (6.2.6).

The inverse of the forward-flow map S is known as the Langevin transport map. In general, these maps do not coincide with the Brenier map [Tan21, LS22], except in dimension 1. An analogous bound to (6.2.18) was proved in [CPS23, Prop. 3.1] for the Brenier map, under the stronger conditions that μ and ν are κ-log-concave, supported on a euclidean ball, and bounded away from 0 on it.

6.3 Applications to Fisher's infinitesimal model

Throughout this section, we fix $\alpha > 0$ and an α-convex mortality function $m \in C^1(\mathbb{R}^d)$. We assume that $m \ge 0$ and $m(0) = 0$. These assumptions are without loss of generality, except for the claim in Theorem 6.3.2 below that $\boldsymbol{\lambda} \in (0,1)$. We also fix $\beta > \frac{1}{2}$ through the identity $\beta = \alpha + \frac{\beta}{\frac{1}{2}+\beta}$, as in Theorem 6.1.3.

The following result is taken from [CPS23, Lemma 2.4]. For the convenience of the reader we include their proof. Recall that the metric $W_{\infty,1}$ was defined after (6.1.5).

Lemma 6.3.1. *Let $c > 0$, and let $P(\cdot\,;x)$ be a probability density on $\mathbb{R}^{2d}$ for each $x \in \mathbb{R}^d$. Suppose that $u_0, u_1 \in C(\mathbb{R}^d)$ are strictly positive functions, that $\log u_0$ is L-Lipschitz, and that*

$$u_1(x) = c \int_{\mathbb{R}^{2d}} P(x_1, x_2; x) u_0(x_1) u_0(x_2) \mathrm{d}x_1 \mathrm{d}x_2 \tag{6.3.1}$$

for all $x \in \mathbb{R}^d$. Then we have

$$|\log u_1(x) - \log u_1(\tilde{x})| \le L\, W_{\infty,1}\Big(P(\cdot; x), P(\cdot; \tilde{x})\Big) \tag{6.3.2}$$

for all $x, \tilde{x} \in \mathbb{R}^d$.

Proof. Fix $x, \tilde{x} \in \mathbb{R}^d$, and let $\gamma \in \mathcal{P}\Big(\mathbb{R}^{2d} \times \mathbb{R}^{2d}\Big)$ be an optimal coupling in the definition of $W_{\infty,1}\Big(P(\cdot; x), P(\cdot; \tilde{x})\Big)$. For $(x_1, x_2), (\tilde{x}_1, \tilde{x}_2) \in \mathbb{R}^{2d}$ we have

$$\begin{aligned}
&\log\Big(u_0(x_1)u_0(x_2)\Big) - \log\Big(u_0(\tilde{x}_1)u_0(\tilde{x}_2)\Big) \\
=&\log u_0(x_1) - \log u_0(\tilde{x}_1) + \log u_0(x_2) - \log u_0(\tilde{x}_2) \\
\le& L\Big(|x_1 - \tilde{x}_1| + |x_2 - \tilde{x}_2|\Big).
\end{aligned}$$

Writing $W := W_{\infty,1}\Big(P(\cdot;x), P(\cdot;\tilde{x})\Big)$, it follows using the bound above that

$$\begin{aligned}
u_1(x) &= c\int_{\mathbb{R}^{4d}} u_0(x_1)u_0(x_2)\,\gamma(\mathrm{d}x_1,\mathrm{d}x_2,\mathrm{d}\tilde{x}_1,\mathrm{d}\tilde{x}_2)\\
&\le c\int_{\mathbb{R}^{4d}} \exp\Big(L\Big(|x_1-\tilde{x}_1|+|x_2-\tilde{x}_2|\Big)\Big)u_0(\tilde{x}_1)u_0(\tilde{x}_2)\,\gamma(\mathrm{d}x_1,\mathrm{d}x_2,\mathrm{d}\tilde{x}_1,\mathrm{d}\tilde{x}_2)\\
&\le ce^{LW}\int_{\mathbb{R}^{4d}} u_0(\tilde{x}_1)u_0(\tilde{x}_2)\,\gamma(\mathrm{d}x_1,\mathrm{d}x_2,\mathrm{d}\tilde{x}_1,\mathrm{d}\tilde{x}_2) = e^{LW}u_1(\tilde{x})\,.
\end{aligned}$$

The desired conclusion follows after exchanging the roles of x and $\tilde{x}$. ∎

6.3.1 Analysis of a localised problem

As in [CPS23, Sec. 4], we study an auxiliary localised problem. Specifically, for $R > 0$, we consider the localised selection function

$$m_R(x) := m(x) + \chi_{\overline{B_R}}\,,$$

where χ denotes the convex indicator function, i.e.,

$$\chi_A(x) = \begin{cases} 0 & \text{if } x \in A\,,\\ +\infty & \text{otherwise}\,. \end{cases}$$

The corresponding localised operator is given by

$$\mathcal{T}_R[F](x) = e^{-m_R(x)}\int_{\mathbb{R}^{2d}} G\Big(x - \frac{x_1+x_2}{2}\Big)\frac{F(x_1)F(x_2)}{\|F\|_{L^1}}\mathrm{d}x_1\mathrm{d}x_2\,.$$

In this section we establish the existence of a quasi-stationary distribution for the localised problem, adapting the proof of [CPS23, Thm. 4.1(i)].

Theorem 6.3.2. *Let $R > 0$. There exists $\boldsymbol{\lambda}_R \in (0,1)$ and a β-log-concave probability density $\mathbf{F}_R$ on $\mathbb{R}^d$ that is bounded away from 0 on its support $\overline{B_R}$, and satisfies*

$$\mathcal{T}_R[\mathbf{F}_R] = \boldsymbol{\lambda}_R\mathbf{F}_R\,.$$

The following result, proved in [CPS23, Lemma 2.2, 2.3], is an immediate consequence of the fact that log-concavity is preserved by convolution (Lemma 6.1.8) and pointwise multiplication with log-concave functions.

Lemma 6.3.3 (Preservation of log-concavity)**.** *Let $R > 0$. If F is κ-log-concave for some $\kappa > 0$, then $\mathcal{T}[F]$ and $\mathcal{T}_R[F]$ are κ'-log-concave with $\kappa' := \alpha + \frac{2\kappa}{1+2\kappa}$. In particular, if F is β-log-concave, then $\mathcal{T}[F]$ and $\mathcal{T}_R[F]$ are β-log-concave as well.*

The key ingredient in the proof of Theorem 6.3.2 is the following contractivity estimate.

Lemma 6.3.4. *Define $F_0 \in L^1_+(\mathbb{R}^d)$ by $F_0(x) = \exp\Big(-\frac{\beta}{2}\|x\|^2 - \chi_{\overline{B}_R(0)}(x)\Big)$ and set $F_{n+1} := T_R[F_n]$ for $n \ge 0$. Then, for all $n \ge 1$:*

$$\mathcal{I}_\infty(F_{n+1}\,\|\,F_n) \le \Big(\tfrac{1}{2}+\beta\Big)^{-1}\mathcal{I}_\infty(F_n\,\|\,F_{n-1})\,.$$

Proof. Set $B_R := B_R(0)$ for brevity, and define $u_n := \frac{F_n}{F_{n-1}}$ for $n \geq 1$. Note that u_n is strictly positive and of class C^1 on $\overline{B_R}$. Using the identities

$$F_n(x) = \frac{e^{-m_R(x)}}{\|F_{n-1}\|_{L^1}} \int_{\mathbb{R}^{2d}} F_{n-1}(x_1)F_{n-1}(x_2)G\Big(x - \frac{x_1+x_2}{2}\Big)\mathrm{d}x_1\mathrm{d}x_2\,,$$
$$F_{n+1}(x) = \frac{e^{-m_R(x)}}{\|F_n\|_{L^1}} \int_{\mathbb{R}^{2d}} F_{n-1}(x_1)u_n(x_1)F_{n-1}(x_2)u_n(x_2)G\Big(x - \frac{x_1+x_2}{2}\Big)\mathrm{d}x_1\mathrm{d}x_2\,,$$

we obtain the recursion relation

$$u_{n+1}(x) = \frac{\|F_{n-1}\|_{L^1}}{\|F_n\|_{L^1}} \int_{\mathbb{R}^{2d}} P_n(x_1,x_2;x)u_n(x_1)u_n(x_2)\mathrm{d}x_1\mathrm{d}x_2 \tag{6.3.3}$$

for $x \in \overline{B_R}$ and $n \geq 1$, with n-dependent transition rates

$$P_n(x_1,x_2;x) = \frac{1}{Z_n(x)}F_{n-1}(x_1)F_{n-1}(x_2)G\Big(x - \frac{x_1+x_2}{2}\Big)\,,$$

where $Z_n(x) > 0$ is the normalising constant ensuring that $P_n(\cdot\,;x)$ is a probability density for all $x \in \mathbb{R}^d$. Arguing as in the proof of Lemma 6.3.1, we infer that

$$|\log u_{n+1}(x) - \log u_{n+1}(\tilde{x})| \leq \|\nabla \log u_n\|_{L^\infty(\overline{B}_R)}\, W_{\infty,1}\Big(P_n(\cdot;x),P_n(\cdot;\tilde{x})\Big)$$

for all $x,\tilde{x} \in \overline{B_R}$. Since F_n is β-log-concave by Lemma 6.3.3, Corollary 6.2.9 yields, in view of the elementary comparison of norms $|(x_1,x_2)|_1 \leq \sqrt{2}|(x_1,x_2)|$ for $x_1,x_2 \in \mathbb{R}^d$,

$$W_{\infty,1}\Big(P_n(\cdot;x),P_n(\cdot;\tilde{x})\Big) \leq \sqrt{2}W_\infty\Big(P_n(\cdot;x),P_n(\cdot;\tilde{x})\Big) \leq \frac{|x-\tilde{x}|}{\frac{1}{2}+\beta}\,.$$

Combining these inequalities, we find

$$\|\nabla \log u_{n+1}\|_{L^\infty(\overline{B}_R)} \leq \Big(\tfrac{1}{2}+\beta\Big)^{-1}\|\nabla \log u_n\|_{L^\infty(\overline{B}_R)}\,,$$

which is the desired inequality. ■

Proof of Theorem 6.3.2. Set $F_0 = \frac{1}{Z}\exp\Big(-\frac{\beta}{2}\|x\|^2 - \chi_{\overline{B_R}}\Big)$ as in Lemma 6.3.4 and define $F_{n+1} = \mathcal{T}_R[F_n]$ for $n \geq 0$, and write $V_n := -\log F_n$. Clearly, the restriction of F_n to $\overline{B_R}$ (which will simply be denoted by F_n as well) is bounded away from 0 and it belongs to $C^1(\overline{B_R})$ for all $n \geq 0$. Adapting arguments from [CPS23], we will show that $\log\Big(F_n/\|F_n\|_{L^1}\Big)$ converges in $C(\overline{B_R})$ as $n \to \infty$. This statement will follow from two claims.

Firstly, we claim that $\nabla\log\Big(F_n/\|F_n\|_{L^1}\Big) = -\nabla V_n$ converges in $C(\overline{B_R})$ as $n \to \infty$. To prove this, we observe that Lemma 6.3.4 yields

$$\mathcal{I}_\infty(F_{n+1}\,\|\,F_n) \leq \Big(\tfrac{1}{2}+\beta\Big)^{-n}\mathcal{I}_\infty(F_1\,\|\,F_0)\,.$$

Since $\mathcal{I}_\infty(F_{n+1}\,\|\,F_n) = \|\nabla V_n - \nabla V_{n+1}\|_{C(\overline{B_R})}$, the sequence ∇V_n is Cauchy in $C(\overline{B_R})$, hence convergent.

Secondly, we claim that $\frac{F_n(0)}{\|F_n\|_{L^1}}$ converges in $\mathbb{R}$ as $n \to \infty$. To show this, we use the identity

$$\frac{F_n(x)}{\|F_n\|_{L^1}} = \frac{\int_{\overline{B_R}\times\overline{B_R}} G\Big(x - \frac{x_1+x_2}{2}\Big)\exp\Big(-m(x)-v_n(x_1)-v_n(x_2)\Big)\mathrm{d}x_1\mathrm{d}x_2}{\iint_{\overline{B_R}\times\overline{B_R}\times\overline{B_R}} G\Big(x' - \frac{x_1+x_2}{2}\Big)\exp\Big(-m(x')-v_n(x_1)-v_n(x_2)\Big)\mathrm{d}x_1\mathrm{d}x_2\mathrm{d}x'}\,,$$

where we write $v_n(x) = V_{n-1}(x) - V_{n-1}(0)$ for brevity. Note that the artifically introduced factors $e^{V_{n-1}(0)}$ cancel out. Writing $v_n(x) = x \cdot \int_0^1 \nabla V_{n-1}(\theta x)\mathrm{d}\theta$ we infer from the first claim that v_n converges uniformly. Therefore, the second claim follows using dominated convergence.

The two claims combined imply that $\log\Big(F_n/\|F_n\|_{L^1}\Big)$ converges in $C(\overline{B_R})$ as $n \to \infty$. Let $-\mathbf{V}_R$ be its limit, and define $\mathbf{F}_R := \exp\Big(-\mathbf{V}_R - \chi_{\overline{B_R}}\Big)$. It remains to verify that $\mathbf{F}_R$ has the desired properties.

Since $\mathbf{V}_R$ is bounded, it follows that $\mathbf{F}_R$ is bounded away from 0 on its support $\overline{B_R}$.

To prove the identity $\mathcal{T}_R[\mathbf{F}] = \boldsymbol{\lambda}_R \mathbf{F}_R$ we write

$$\frac{\|F_{n+1}\|_{L^1}}{\|F_n\|_{L^1}} = \int_{\overline{B_R}\times\overline{B_R}} H_R(x_1,x_2)\frac{F_n(x_1)}{\|F_n\|_{L^1}}\frac{F_n(x_2)}{\|F_n\|_{L^1}}\mathrm{d}x_1\mathrm{d}x_2\,,$$

where $H(x_1,x_2) = \int_{\overline{B_R}} e^{-m(x)}G\Big(x - \frac{x_1+x_2}{2}\Big)\mathrm{d}x$ is bounded. Since H_R is bounded and $F_n/\|F_n\|_{L^1}$ converges uniformly by the first part of the proof, we infer that $\frac{\|F_{n+1}\|_{L^1}}{\|F_n\|_{L^1}} \to \boldsymbol{\lambda}_R$ for some $\boldsymbol{\lambda}_R > 0$. Since $\frac{F_n}{\|F_n\|_{L^1}} \to \mathbf{F}$ in $C(\overline{B_R})$, it follows that $\mathcal{T}_R\Big[\frac{F_n}{\|F_n\|_{L^1}}\Big] \to \mathcal{T}_R[\mathbf{F}]$. On the other hand,

$$\mathcal{T}_R\left[\frac{F_n}{\|F_n\|_{L^1}}\right] = \frac{\|F_{n+1}\|_{L^1}}{\|F_n\|_{L^1}}\frac{F_{n+1}}{\|F_{n+1}\|_{L^1}} \to \boldsymbol{\lambda}_R\mathbf{F}_R$$

as $n \to \infty$. This yields the desired identity $\mathcal{T}_R[\mathbf{F}] = \boldsymbol{\lambda}_R\mathbf{F}_R$.

Since F_0 is a β-log-concave density, so are all F_n by Lemma 6.3.3. Therefore, the functions $-\log\Big(F_n/\|F_n\|_{L^1}\Big)$ are β-convex, and so is their uniform limit $\mathbf{V}_R$. It follows that $\mathbf{F}_R$ is β-log-concave.

Finally, we will show that $\boldsymbol{\lambda}_R \in (0,1)$. Indeed, since $\mathbf{F}_R$ is quasi-stationary, we have

$$\boldsymbol{\lambda}_R\mathbf{F}_R(x)e^{m_R(x)} = \int_{\mathbb{R}^{2d}} G\Big(x - \frac{x_1+x_2}{2}\Big)\mathbf{F}_R(x_1)\mathbf{F}_R(x_2)\mathrm{d}x_1\mathrm{d}x_2\,.$$

From this, it is immediate to see that $\boldsymbol{\lambda}_R > 0$, by choosing $x = 0$. To see that $\boldsymbol{\lambda}_R < 1$, it suffices to integrate over $x \in \mathbb{R}^d$ on both sides. Indeed, there exists a small $\delta \in (0,R)$ such that $c_\delta := \int_{B_\delta(0)} \mathbf{F}_R(x)\mathrm{d}x < 1$. But then, using the assumptions on m,

$$\int_{\mathbb{R}^d} e^{m_R(x)}\mathbf{F}_R(x)\mathrm{d}x \geq e^{\alpha\delta^2/2}\int_{B_\delta(0)}\mathbf{F}_R(x)\mathrm{d}x + \int_{B_\delta(0)^c}\mathbf{F}_R(x)\mathrm{d}x = c_\delta e^{\alpha\delta^2/2} + (1-c_\delta) > 1\,,$$

while

$$\int_{\mathbb{R}^d}\int_{\mathbb{R}^{2d}} G\Big(x - \frac{x_1+x_2}{2}\Big)\mathbf{F}_R(x_1)\mathbf{F}_R(x_2)\mathrm{d}x_1\mathrm{d}x_2\,\mathrm{d}x = 1\,.$$

Consequently, $\boldsymbol{\lambda}_R \in (0,1)$. ∎

6.3.2 Existence of a β-log-concave quasi-equilibrium

The following second-moment bound is an analogue of [CPS23, Prop. 5.1], but the proof is based on different arguments that seem more convenient in the multi-dimensional setting. In particular, we use various properties of maxima of strongly log-concave densities, that are proved in Section 6.5.

Proposition 6.3.5. *For $R > 0$, let F_R be a solution to the localised problem given in Theorem 6.3.2. Then:*

$$\sup_{R>0} \int_{\mathbb{R}^d} |x|^2 F_R(x)\mathrm{d}x < \infty\,.$$

Proof. Let $\mu_R = \int_{\mathbb{R}^d} x\mathbf{F}_R(x)\mathrm{d}x$ be the barycenter of $\mathbf{F}_R = e^{-\mathbf{V}_R}$. Since the measures $\mathbf{F}_R$ are β-log-concave, it follows from the Poincaré inequality [BGL14, Prop. 4.8.1] that

$$\sup_{R>0} \int_{\mathbb{R}^d} |x-\mu_R|^2 F_R(x)\mathrm{d}x \le \frac{d}{\beta} < \infty\,.$$

Therefore, it suffices to show that $\sup_{R>0} |\mu_R| < \infty$. Let $v_R \in \mathbb{R}^d$ be the unique minimizer of $\mathbf{V}_R$. Since $\mathbf{F}_R$ is β-log-concave, Lemma 6.5.1 implies that

$$|v_R - \mu_R| \le \sqrt{\frac{d}{\beta}}\,.$$

Define $G_R := \mathcal{R}[\mathbf{F}_R]$ and write $G_R = e^{-U_R}$. The barycenter μ_R of $\mathbf{F}_R$ is also the barycenter of G_R, since G_R can be written in probabilistic terms as $G_R = \mathrm{law}\Big(\frac{X_R+\tilde{X}_R}{2} + Z\Big)$, where $X_R, \tilde{X}_R$ are independent random variables with law F_R and Z is standard Gaussian, and we have $\mathbb{E}\Big[\frac{X_R+\tilde{X}_R}{2} + Z\Big] = \mathbb{E}[X_R]$. Moreover, Lemma 6.3.3 implies that G_R is τ-log-concave with $\tau := \beta/(\frac{1}{2}+\beta)$. Therefore, another application of Lemma 6.5.1 yields

$$|u_R - \mu_R| \le \sqrt{\frac{d}{\tau}}\,, \tag{6.3.4}$$

where $u_R \in \mathbb{R}^d$ denotes the unique minimizer of U_R.

Since $\mathcal{T}[\mathbf{F}_R] = \boldsymbol{\lambda}\mathbf{F}_R$, it follows that $\mathbf{V}_R = m_R + U_R + \log \boldsymbol{\lambda}_R$. Recall that m_R has its unique minimizer at 0 and satisfies $\nabla^2 m_R \succcurlyeq \alpha I_d$. Observe that Lemma 6.1.9 implies that $\nabla^2 U_R \preccurlyeq I_d$. Therefore, Lemma 6.5.2 implies that

$$\alpha|u_R| \le (1+\alpha)|u_R - v_R|\,.$$

Combining the three inequalities above, we find

$$\frac{\alpha}{1+\alpha}|u_R| \le |u_R - v_R| \le |u_R - \mu_R| + |v_R - \mu_R| \le \sqrt{\frac{d}{\tau}} + \sqrt{\frac{d}{\beta}}\,.$$

Another application of (6.3.4) implies that $\sup_{R>0} |\mu_R| < \infty$, as desired. ■

To prove Theorem 6.1.3, we can now follow the argument from [CPS23, Thm. 5.2].

Proof of Theorem 6.1.3. It follows from Proposition 6.3.5 that the family of probability measures $\{\mathbf{F}_R\}_{R>0}$ is tight. Therefore, there exists a sequence of radii $(R_n)_n$ with $R_n \uparrow \infty$ and a limiting probability measure $\mathbf{F}$ such that $\mathbf{F}_{R_n} \to \mathbf{F}$ weakly. Then, proceeding as in the proof of [CPS23, Thm. 5.2], it follows that $\boldsymbol{\lambda}_{R_n}$ converges to some $\boldsymbol{\lambda} \in [0,1]$, that $\mathbf{F}$ is β-log-concave, and that the pair $(\boldsymbol{\lambda}, \mathbf{F})$ satisfies $\mathcal{T}[\mathbf{F}] = \boldsymbol{\lambda}\mathbf{F}$. Proceeding as in the proof of Theorem 6.3.2 we also find that $\boldsymbol{\lambda} \in (0,1)$. ■

6.3.3 Exponential convergence to quasi-equilibrium

Proof of Theorem 6.1.4. Recall from (6.1.4) that

$$P(x_1, x_2; x) = \frac{1}{Z_x}\mathbf{F}(x_1)\mathbf{F}(x_2)\exp\Big(-\frac{1}{2}\Big|x - \frac{x_1+x_2}{2}\Big|^2\Big), \tag{6.3.5}$$

where $\mathbf{F}$ is a β-log-concave quasi-equilibrium obtained in Theorem 6.1.3. Therefore, the result follows from Corollary 6.2.9. ■

Proof of Corollary 6.1.5. We follow the proof of [CPS23, Thm. 1.1]; for the convenience of the reader, we reproduce the argument here.

((i)): Take $0 \neq F_0 \in L^1_+(\mathbb{R}^d)$ with $\mathcal{I}_\infty(F_0 \,\|\, \mathbf{F}) < \infty$. Then we can write $F_0 = u_0\mathbf{F}$ for some strictly positive $u_0 \in C(\mathbb{R}^d)$ such that $\log u_0$ is L-Lipschitz with $L := \mathcal{I}_\infty(F_0 \,\|\, \mathbf{F})$. For $n \geq 1$, set $u_n = \frac{F_n}{\boldsymbol{\lambda}^n\mathbf{F}}$. We will show by induction that $\log u_n$ is L_n-Lipschitz with $L_n = \left(\frac{1}{2}+\beta\right)^{-n}\mathcal{I}_\infty(F_0 \,\|\, \mathbf{F})$, which implies ((i)) in Corollary 6.1.5. To this end, recall that we have the recursion

$$u_{n+1}(x) = \int_{\mathbb{R}^d\times\mathbb{R}^d} \frac{u_n(x_1)u_n(x_2)}{\|u_n\mathbf{F}\|_{L^1}} P(x_1, x_2; x)\mathrm{d}x_1\mathrm{d}x_2$$

for all $x \in \mathbb{R}^d$. Therefore, Lemma 6.3.1 implies

$$\|\log u_{n+1}(x_1) - \log u_{n+1}(x_2)\| \leq \mathcal{I}_\infty(F_n \,\|\, \mathbf{F})\, W_{\infty,1}\Big(P(\cdot;x), P(\cdot;\tilde{x})\Big)$$

for all $x_1, x_2 \in \mathbb{R}^d$ with $x_1 \neq x_2$. Using the elementary bound $W_{\infty,1} \leq \sqrt{2}W_\infty$ and Theorem 6.1.4, we obtain

$$W_{\infty,1}\Big(P(\cdot;x), P(\cdot;\tilde{x})\Big) \leq \sqrt{2}W_\infty\Big(P(\cdot;x), P(\cdot;\tilde{x})\Big) \leq \frac{|x-\tilde{x}|}{\frac{1}{2}+\beta}\,.$$

Combining these inequalities, we find $\mathcal{I}_\infty(F_{n+1} \,\|\, \mathbf{F}) \leq \left(\frac{1}{2}+\beta\right)^{-1}\mathcal{I}_\infty(F_n \,\|\, \mathbf{F})$, which implies the desired conclusion.

((ii)): For brevity, write $\widehat{F}_n := F_n/\|F_n\|_{L^1}$. Since $\mathbf{F}$ is β-log-concave, it satisfies a logarithmic Sobolev inequality by the Bakry–Émery theory [BGL14, Cor. 5.7.2]). Using this and the trivial bound $\mathcal{I}_2(\cdot \,\|\, \mathbf{F}) \leq \mathcal{I}_\infty(\cdot \,\|\, \mathbf{F})^2$, we deduce that

$$\mathcal{D}_{\mathrm{KL}}\Big(\widehat{F}_n \,\Big\|\, \mathbf{F}\Big) \leq \frac{1}{2\beta}\mathcal{I}_2\Big(\widehat{F}_n \,\Big\|\, \mathbf{F}\Big) \leq \frac{1}{2\beta}\mathcal{I}_\infty(F_n \,\|\, \mathbf{F})^2 \leq \frac{\mathcal{I}_\infty(F_0 \,\|\, \mathbf{F})^2}{2\beta\left(\frac{1}{2}+\beta\right)^{2n}}\,.$$

As for the last conclusion, set $\phi := e^{-m} * G \in C_b(\mathbb{R}^d)$, and note that

$$\frac{\|F_{n+1}\|_{L^1}}{\|F_n\|_{L^1}} = \int_{\mathbb{R}^{2d}} \phi\Big(\frac{x_1+x_2}{2}\Big)\widehat{F}_n(x_1)\widehat{F}_n(x_2)\mathrm{d}x_1\mathrm{d}x_2\,,$$

$$\boldsymbol{\lambda} = \int_{\mathbb{R}^{2d}} \phi\Big(\frac{x_1+x_2}{2}\Big)\mathbf{F}(x_1)\mathbf{F}(x_2)\mathrm{d}x_1\mathrm{d}x_2\,,$$

hence, by Hölder's inequality,

$$\left|\frac{\|F_{n+1}\|_{L^1}}{\|F_n\|_{L^1}} - \boldsymbol{\lambda}\right| \leq \|\phi\|_{L^\infty}\Big\|\widehat{F}_n \otimes \widehat{F}_n - \mathbf{F}\otimes\mathbf{F}\Big\|_{L^1}\,.$$

Using Pinsker's inequality, the tensorization of the relative entropy, and the previous step we deduce that

$$\left\|\widehat{F}_n\otimes\widehat{F}_n-\mathbf{F}\otimes\mathbf{F}\right\|_{L^1}\le\sqrt{\frac{1}{2}\mathcal{D}_{\mathrm{KL}}\Big(\widehat{F}_n\otimes\widehat{F}_n\,\Big\|\,\mathbf{F}\otimes\mathbf{F}\Big)}\le\sqrt{\mathcal{D}_{\mathrm{KL}}\Big(\widehat{F}_n\,\Big\|\,\mathbf{F}\Big)}\le\frac{\mathcal{I}_\infty(F_0\,\|\,\mathbf{F})}{\sqrt{2\beta}\left(\frac{1}{2}+\beta\right)^n},$$

which gives the desired conclusion. ■

6.4 Other information metrics

In view of the crucial contraction estimate $\mathcal{I}_\infty(\mathcal{T}[F]\,\|\,\mathbf{F})\le\left(\frac{1}{2}+\beta\right)^{-1}\mathcal{I}_\infty(F\,\|\,\mathbf{F})$, it is natural to ask whether analogous inequalities hold for other functionals $\mathcal{F}$, such as the relative entropy $\mathcal{D}_{\mathrm{KL}}(\cdot\,\|\,\mathbf{F})$ and the L^2 relative Fisher information $\mathcal{I}_2(\cdot\,\|\,\mathbf{F})$, which play a central role in the Bakry–Émery theory for diffusion equations.

Here we consider the case of quadratic selection $m(x)=\frac{\alpha}{2}|x|^2$ for some $\alpha>0$ in dimension $d=1$, which has been analysed in detail in [CLP24]. In this case, the operator $\mathcal{T}$ maps Gaussian densities to multiples of Gaussian densities. Indeed, for $\mu\in\mathbb{R}$ and $\sigma>0$, we have

$$\mathcal{T}[\gamma_{\mu,\sigma^2}]\propto\gamma_{\tilde\mu,\tilde\sigma^2}\,,\qquad\text{where }\ \tilde\mu=\frac{\mu}{1+\alpha\left(1+\frac{\sigma^2}{2}\right)}\quad\text{and}\quad\tilde\sigma^2=\frac{1+\frac{\sigma^2}{2}}{1+\alpha\left(1+\frac{\sigma^2}{2}\right)}\,,\tag{6.4.1}$$

see (3.5) in [CLP24]. Moreover, the unique quasi-stationary probability distribution $\mathbf{F}$ is the centered Gaussian density with variance $\frac{1}{\beta}$, where $\beta>\frac{1}{2}$ is the log-concavity parameter in Theorem 6.1.3; see (1.12) and (1.13) in [CLP24].

To analyse the behaviour of the three functionals under $\mathcal{T}$, we consider the renormalised operator $\widehat{\mathcal{T}}$ given by $\widehat{\mathcal{T}}[F]:=\mathcal{T}[F]/\|\mathcal{T}[F]\|_{L^1}$ that preserves probability densities. Let us first consider the case where $G_\mu:=\gamma_{\mu,\frac{1}{\beta}}$ is a Gaussian density having the variance $\frac{1}{\beta}$ of the quasi-equilibrium $\mathbf{F}$ with arbitrary nonzero mean $\mu\in\mathbb{R}$. Then $\widehat{\mathcal{T}}[G_\mu]$ is Gaussian with variance $\frac{1}{\beta}$ as well, and the three functionals contract with the same rate:

$$\frac{\mathcal{D}_{\mathrm{KL}}\Big(\widehat{\mathcal{T}}[G_\mu]\,\Big\|\,\mathbf{F}\Big)}{\mathcal{D}_{\mathrm{KL}}(G_\mu\,\|\,\mathbf{F})}=\frac{\mathcal{I}_2\Big(\widehat{\mathcal{T}}[G_\mu]\,\Big\|\,\mathbf{F}\Big)}{\mathcal{I}_2(G_\mu\,\|\,\mathbf{F})}=\left(\frac{\mathcal{I}_\infty\Big(\widehat{\mathcal{T}}[G_\mu]\,\Big\|\,\mathbf{F}\Big)}{\mathcal{I}_\infty(G_\mu\,\|\,\mathbf{F})}\right)^2=\left(\tfrac{1}{2}+\beta\right)^{-2}<1\,.\tag{6.4.2}$$

These equalities readily follow from the following Gaussian identities, which hold for $\mu,\bar\mu\in\mathbb{R}$ and $\sigma^2,\bar\sigma^2>0$:

$$\mathcal{D}_{\mathrm{KL}}(\gamma_{\mu,\sigma^2}\,\|\,\gamma_{\bar\mu,\bar\sigma^2})=\frac{1}{2}\left(\frac{(\mu-\bar\mu)^2}{\bar\sigma^2}+\log\left(\frac{\bar\sigma^2}{\sigma^2}\right)-1+\frac{\sigma^2}{\bar\sigma^2}\right),\tag{6.4.3}$$

$$\mathcal{I}_2(\gamma_{\mu,\sigma^2}\,\|\,\gamma_{\bar\mu,\bar\sigma^2})=\frac{(\mu-\bar\mu)^2}{\bar\sigma^4}+\frac{(\sigma^2-\bar\sigma^2)^2}{\sigma^2\bar\sigma^4}\,,\tag{6.4.4}$$

$$\mathcal{I}_\infty(\gamma_{\mu,\sigma^2}\,\|\,\gamma_{\bar\mu,\bar\sigma^2})=\frac{|\mu-\bar\mu|}{\sigma^2}\quad\text{if }\sigma=\bar\sigma;\quad\text{otherwise, }\ \mathcal{I}_\infty(\gamma_{\mu,\sigma^2}\,\|\,\gamma_{\bar\mu,\bar\sigma^2})=+\infty\,.\tag{6.4.5}$$

Next, let us suppose that $G=\gamma_{\mu,\sigma^2}$ is a Gaussian density with arbitrary mean $\mu\in\mathbb{R}$ and variance $\sigma^2\neq\frac{1}{\beta}$. In this case, $\mathcal{I}_\infty(G\,\|\,\mathbf{F})=\mathcal{I}_\infty(T[G]\,\|\,\mathbf{F})=+\infty$. However, the relative entropy and the L^2 relative Fisher information are finite, so one might wonder whether these functionals contract under $\widehat{\mathcal{T}}$ with the rate suggested by (6.4.2). The following result shows that this is not the case.

Proposition 6.4.1. *Let* $m(x) = \frac{\alpha}{2}|x|^2$ *for some* $\alpha > 0$, *and define* $\beta > \max\{\frac{1}{2}, \alpha\}$ *by* $\beta = \alpha + \frac{\beta}{\frac{1}{2}+\beta}$, *as in Theorem 6.1.3. Then there exist Gaussian probability densities* $G \in L^1_+(\mathbb{R})$ *such that*

$$\frac{\mathcal{I}_2\big(\widehat{\mathcal{T}}[G] \,\big\|\, \mathbf{F}\big)}{\mathcal{I}_2(G \,\|\, \mathbf{F})} > \left(\tfrac{1}{2}+\beta\right)^{-2} \quad \textit{and} \quad \frac{\mathcal{D}_{\mathrm{KL}}\big(\widehat{\mathcal{T}}[G] \,\big\|\, \mathbf{F}\big)}{\mathcal{D}_{\mathrm{KL}}(G \,\|\, \mathbf{F})} > \left(\tfrac{1}{2}+\beta\right)^{-2}.$$

Proof. Let $\mathcal{F}$ be either $\mathcal{D}_{\mathrm{KL}}(\cdot \,\|\, \mathbf{F})$ or $\mathcal{I}_2(\cdot \,\|\, \mathbf{F})$. Using (6.4.1) we observe that

$$\lim_{\mu\to\infty} \frac{\mathcal{F}\big(\widehat{\mathcal{T}}[\gamma_{\mu,\sigma^2}]\big)}{\mathcal{F}(\gamma_{\mu,\sigma^2})} = \left(1+\alpha\left(1+\tfrac{\sigma^2}{2}\right)\right)^{-2}.$$

Consequently,

$$\lim_{\sigma^2\to 0}\lim_{\mu\to\infty} \frac{\mathcal{F}\big(\widehat{\mathcal{T}}[\gamma_{\mu,\sigma^2}]\big)}{\mathcal{F}(\gamma_{\mu,\sigma^2})} = (1+\alpha)^{-2}. \tag{6.4.6}$$

Since $(1+\alpha)^{-2} > (\frac{1}{2}+\beta)^{-2}$, the claim follows. ∎

We illustrate the behaviour of the contraction factor $C_{\mathcal{F}}(\mu,\sigma^2) := \dfrac{\mathcal{F}\big(\widehat{\mathcal{T}}[\gamma_{\mu,\sigma^2}]\big)}{\mathcal{F}(\gamma_{\mu,\sigma^2})}$ in Fig. 6.1.

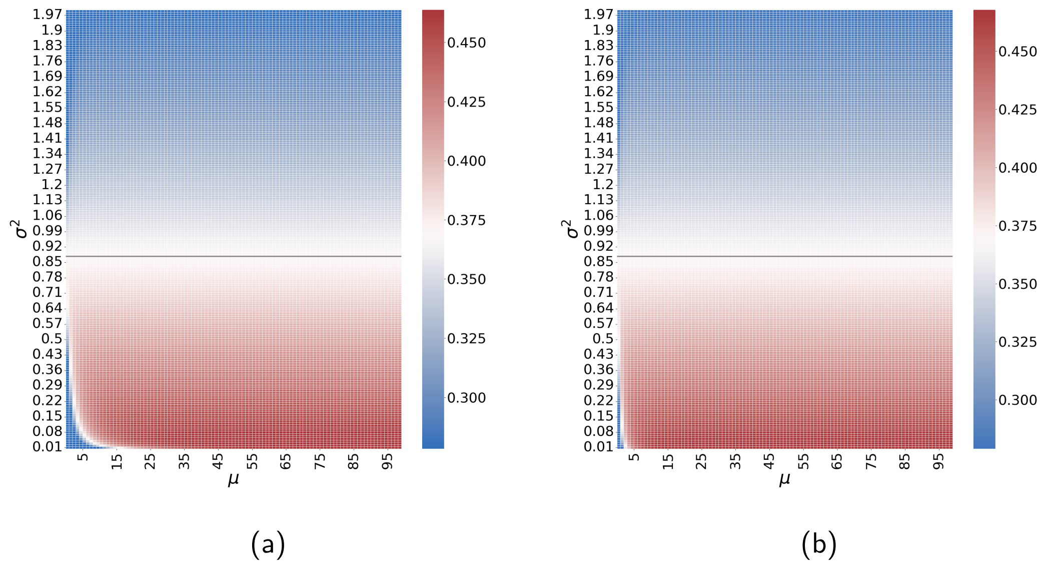


Figure 6.1: Heatmaps of $C_{\mathcal{F}}(\mu,\sigma^2)$ for (a) $\mathcal{F} = \mathcal{I}_2(\cdot \,\|\, \mathbf{F})$ and (b) $\mathcal{F} = \mathcal{D}_{\mathrm{KL}}(\cdot \,\|\, \mathbf{F})$. The chosen parameter value is $\alpha = 0.45$, and the variance of the Gaussian quasi-equilibrium is $1/\beta \approx 0.87$. This value is indicated by the grey line. The corresponding contraction factor is $\left(\frac{1}{2}+\beta\right)^{-2} \approx 0.37$ as computed in (6.4.2). As $\sigma^2 \to 0$ after $\mu \to \infty$, the contraction factor $C_{\mathcal{F}}(\mu,\sigma^2)$ approaches $(1+\alpha)^{-2} \approx 0.48$, as computed in (6.4.6).

6.5 Peaks of strongly of log-concave densities

The following standard result asserts that strongly log-concave distributions concentrate around the minimizer of their potential. Since we apply the result for general log-concave densities (not necessarily having full support on $\mathbb{R}^d$), we provide a detailed proof.

Lemma 6.5.1. *Let $\mu = e^{-V}$ be a κ-log-concave probability density on $\mathbb{R}^d$ for some $\kappa > 0$. Assume that V is lower semicontinuous, and set $\hat{x} := \arg\min V$. Then we have*

$$\int_{\mathbb{R}^d} |x - \hat{x}|^2 \mu(x) \mathrm{d}x \le \frac{d}{\kappa} . \tag{6.5.1}$$

Proof. Note first that since V is lower semicontinuous and κ-convex, it indeed admits a minimizer. The proof then consists of two steps.

Step 1. Assume that $V : \mathbb{R}^d \to \mathbb{R}$ is of class C^2 and such that ∇V is Lipschitz. In this case, $\nabla^2 V \succcurlyeq \kappa I_d$ and there exists a solution to the Langevin equation

$$dX_t = -\nabla V(X_t)dt + \sqrt{2}dB_t , \qquad X_0 = \hat{x} .$$

Using Itô's formula, the κ-convexity of V, and the fact that $\nabla V(\hat{x}) = 0$, we find

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\Big[|X_t - \hat{x}|^2\Big] = -\mathbb{E}[\nabla V(X_t) \cdot (X_t - \hat{x})] + d \le -\kappa \mathbb{E}\Big[|X_t - \hat{x}|^2\Big] + d .$$

Hence,

$$\mathbb{E}\Big[|X_t - \hat{x}|^2\Big] \le \frac{d}{\kappa}$$

for all $t \ge 0$. As $\mathbb{E}\Big[\|X_t - \hat{x}\|^2\Big] = W_2(\mathrm{law}(X_t), \delta_{\hat{x}})^2$, the conclusion follows by passing to the limit $t \to \infty$, since $W_2(\mathrm{law}(X_t), \mu) \to 0$; see e.g., [AGS08, Thm. 11.2.1].

Step 2. We remove the additional assumptions on μ. To this end, define $\mu_n := \mu * \gamma_{\frac{1}{n}}$, set $V_n := -\log \mu_n$, and $\hat{x}_n := \arg\min V_n$. Then μ_n is $\frac{n\kappa}{n+\kappa}$-log-concave by Lemma 6.1.8. Using the triangle inequality in $L^2(\mu_n)$ and an application of Step 1 to μ_n we find

$$\begin{aligned}\left(\int |x - \hat{x}|^2 \mu_n(x) \mathrm{d}x\right)^{1/2} &\le \left(\int |x - \hat{x}_n|^2 \mu_n(x) \mathrm{d}x\right)^{1/2} + |\hat{x}_n - \hat{x}| \\ &\le \left(d\frac{n+\kappa}{n\kappa}\right)^{1/2} + |\hat{x}_n - \hat{x}| .\end{aligned}$$

Since μ_n converges weakly to μ, and $x \mapsto |x - \hat{x}|^2$ is continuous and bounded from below, we have

$$\left(\int |x - \hat{x}|^2 \mu(x) \mathrm{d}x\right)^{1/2} \le \liminf_{n\to\infty} \left(\int |x - \hat{x}|^2 \mu_n(x) \mathrm{d}x\right)^{1/2} .$$

Thus, to obtain the desired result, it remains to show that $\|\hat{x}_n - \hat{x}\| \to 0$.

For this purpose, fix $\varepsilon \in (0, 1)$. It remains to show that there exists $\hat{n} \ge 1$ such that μ_n attains its maximum in a ball of radius ε around $\hat{x}$ whenever $n \ge \hat{n}$.

Let $\delta > 0$ be a small parameter, only depending on ε, that will be specified later.

First we will argue that μ_n attains a large value near $\hat{x}$. For this purpose, observe that $\mathrm{dom}(V)$ has non-empty interior, since μ is a log-concave density. Take $z \in \mathrm{dom}(V)^\circ$. Since V is continuous on its domain, V is bounded on an open ball around z. Therefore, by convexity of V, we can find $y \in B_{\frac{\varepsilon}{2}}(\hat{x}) \cap \mathrm{dom}(V)^\circ$ and a radius $h > 0$ such that $\mu(x) \geq \mu(\hat{x}) - \delta$ for all $x \in B_h(y)$. Without loss of generality, we choose $h \leq \min\{\delta, \frac{\varepsilon}{2}\}$. Observe now that there exists a constant $\hat{n} \geq 1$ depending only on h and the dimension d, such that

$$\int_{B_h(0)} \gamma_{\frac{1}{n}}(x)\mathrm{d}x \geq 1 - h \tag{6.5.2}$$

for all $n \geq \hat{n}$. Hence, for $n \geq \hat{n}$, (6.5.2) yields

$$\mu_n(y) \geq \Big(\mu(\hat{x}) - \delta\Big)(1-h) \geq \Big(\mu(\hat{x}) - \delta\Big)(1-\delta)\,. \tag{6.5.3}$$

Next we will quantify the fact that μ_n decreases fast if $|x - \hat{x}|$ increases. Indeed, since V is κ-convex and $\hat{x} = \arg\min V$, we have $V(x) \geq V(\hat{x}) + \frac{\kappa}{2}|x - \hat{x}|^2$ for all $x \in \mathbb{R}^d$, hence

$$\mu(x) \leq e^{-\frac{\kappa}{2}|x-\hat{x}|^2}\mu(\hat{x})\,.$$

Therefore, if $|x - \hat{x}| > \varepsilon$, another application of (6.5.2) yields, taking into account that $h \leq \frac{\varepsilon}{2}$,

$$\mu_n(x) \leq \sup_{|y-x|\leq h} \mu(y) + h \sup_{y\in\mathbb{R}^d} \mu(y) \leq \sup_{|y-\hat{x}|\geq\frac{\varepsilon}{2}} \mu(y) + h\mu(\hat{x}) \leq \Big(e^{-\frac{\kappa}{8}\varepsilon^2} + \delta\Big)\mu(\hat{x})\,. \tag{6.5.4}$$

Choosing $\delta > 0$ small enough (depending on ε), it follows by combining (6.5.3) and (6.5.4) that

$$\mu_n(y) > \sup_{x:|x-\hat{x}|>\varepsilon} \mu_n(x)\,,$$

hence $\hat{x}_n \in \overline{B}_\varepsilon(\tilde{x})$ whenever $n \geq \hat{n}$, which completes the poof. ■

Lemma 6.5.2. *Let $V, U\colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be strictly convex functions such that*

(i) V is lower semicontinuous and α-convex for some $\alpha > 0$;

(ii) U belongs to $C^1(\mathbb{R}^d)$, it admits a minimizer, and ∇U is β-Lipschitz for some $\beta > 0$.

Define $x = \arg\min V$, $y = \arg\min U$, and $z = \arg\min(V+U)$. Then:

$$\frac{1}{\alpha}\|z - y\| \geq \max\left\{\frac{1}{\beta}\|z - x\|\,, \frac{1}{\alpha+\beta}\|y - x\|\right\}.$$

Proof. Note first that since $V + U$ is lower semicontinuous and α-convex, it indeed admits a minimizer.

Step 1. Assume additionally that $V \in C^1(\mathbb{R}^d)$. Then one of the two desired inequalities follows from

$$\alpha\|z - x\| \leq \|\nabla V(z)\| = \|\nabla U(z)\| \leq \beta\|z - y\|\,.$$

The other one follows by combining this inequality with the triangle inequality $\|z - x\| \geq \|y - x\| - \|z - y\|$.

Step 2. We now remove the additional assumption that $V \in C^1(\mathbb{R}^d)$. For $\lambda > 0$, we consider the Moreau-Yosida approximation V_λ of V defined by

$$V_\lambda(x) = \inf_{y \in \mathbb{R}^d} \Big\{ V(y) + \frac{1}{2\lambda} \|x - y\|^2 \Big\}.$$

It is classical that V_λ is of class C^1 and α_λ-convex with $\alpha_\lambda \downarrow \alpha > 0$ as $\lambda \to 0$ (cf. [Clé09, Prop. 3.1]). Clearly, $x = \arg\min V_\lambda$.

Write $z_\lambda := \arg\min(V_\lambda + U)$. An application of Step 1 yields

$$\frac{1}{\alpha_\lambda} \|z_\lambda - y\| \geq \max \Big\{ \frac{1}{\beta} \|z_\lambda - x\| \, , \frac{1}{\alpha_\lambda + \beta} \|y - x\| \Big\}.$$

Therefore, to derive the desired conclusion, it remains to prove that $z_\lambda \to z$ as $\lambda \to 0$.

To show this, define $\tilde{z}_\lambda := \arg\min_y \Big\{ V(y) + \frac{1}{2\lambda} \|y - z_\lambda\|^2 \Big\}$, so that

$$V_\lambda(z_\lambda) = V(\tilde{z}_\lambda) + \frac{1}{2\lambda} \|z_\lambda - \tilde{z}_\lambda\|^2 . \tag{6.5.5}$$

We claim that there exists a compact set $\mathcal{C}$ such that $z_\lambda, \tilde{z}_\lambda \in \mathcal{C}$ for all $\lambda \in (0, 1]$. Let us show this. Since $V_\lambda \leq V$, $z_\lambda = \arg\min(V_\lambda + U)$, and (6.5.5), we obtain

$$V(z) + U(z) \geq V_\lambda(z) + U(z) \geq V_\lambda(z_\lambda) + U(z_\lambda) \geq V(\tilde{z}_\lambda) + \frac{1}{2\lambda} \|z_\lambda - \tilde{z}_\lambda\|^2 + U(z_\lambda) . \tag{6.5.6}$$

Using this inequality and the fact that $x = \arg\min V$ and $y = \arg\min U$, we find

$$V(z) + U(z) \geq V(x) + \frac{1}{2\lambda} \|z_\lambda - \tilde{z}_\lambda\|^2 + U(y) .$$

Consequently,

$$\|z_\lambda - \tilde{z}_\lambda\|^2 \leq 2\lambda M \, , \qquad \text{where } M := V(z) + U(z) - V(x) - U(y) . \tag{6.5.7}$$

Since $x = \arg\min V$ and V is α-convex, $y = \arg\min U$, and (6.5.6), we deduce

$$\begin{aligned}
\frac{1}{2} |\tilde{z}_\lambda - z|^2 &\leq |\tilde{z}_\lambda - x|^2 + |x - z|^2 \leq \frac{2}{\alpha} V(\tilde{z}_\lambda) + |x - z|^2 \\
&\leq \frac{2}{\alpha} \Big(V(\tilde{z}_\lambda) + \frac{1}{2\lambda} \|z_\lambda - \tilde{z}_\lambda\|^2 + U(z_\lambda) - U(y) \Big) + |x - z|^2 \\
&\leq \frac{2}{\alpha} \Big(V(z) + U(z) - U(y) \Big) + |x - z|^2 .
\end{aligned}$$

Together with (6.5.7), this estimate yields the claim.

Fix $\varepsilon > 0$. Since U is uniformly continuous on $\mathcal{C}$, there exists $\delta \in (0, \frac{\varepsilon}{2})$ such that

$$|U(x_1) - U(x_2)| \leq \frac{\alpha \varepsilon^2}{8} \tag{6.5.8}$$

for all $x_1, x_2 \in \mathcal{C}$ with $\|x_1 - x_2\| \leq \delta$. Define $\hat{\lambda} := \min\Big\{1, \frac{\delta^2}{2M}\Big\}$. To complete the proof, we shall show that $\|z - z_\lambda\| \leq \varepsilon$ whenever $\lambda \leq \hat{\lambda}$.

Note first that $\|z_\lambda - \tilde{z}_\lambda\| \le \delta$ for all $0 < \lambda \le \hat{\lambda}$ by (6.5.7) and the definition of $\hat{\lambda}$. Using (6.5.6), (6.5.8), the α-convexity of $V + U$ and the fact that $z = \arg\min(V + U)$, we further deduce that

$$\begin{aligned} V(z) + U(z) \ge V(\tilde{z}_\lambda) + U(z_\lambda) &\ge V(\tilde{z}_\lambda) + U(\tilde{z}_\lambda) - \frac{\alpha\varepsilon^2}{8} \\ &\ge V(z) + U(z) + \frac{\alpha}{2}\|z - \tilde{z}_\lambda\|^2 - \frac{\alpha\varepsilon^2}{8}\,. \end{aligned}$$

This implies $\|z - \tilde{z}_\lambda\| \le \frac{\varepsilon}{2}$. Since $\|z_\lambda - \tilde{z}_\lambda\| \le \delta < \frac{\varepsilon}{2}$, we obtain the desired result. ■

Bibliography

[AB09] Hedy Attouch and Jérôme Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Math. Program.*, 116(1-2 (B)):5–16, 2009.

[ABRS10] Hédy Attouch, Jérôme Bolte, Patrick Redont, and Antoine Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the Kurdyka-Łojasiewicz inequality. *Math. Oper. Res.*, 35(2):438–457, 2010.

[ABS13] Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. *Math. Program.*, 137(1-2 (A)):91–129, 2013.

[AG13] Luigi Ambrosio and Nicola Gigli. *A User's Guide to Optimal Transport*, pages 1–155. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[AGS08] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, second edition, 2008.

[AGS14] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. Calculus and heat flow in metric measure spaces and applications to spaces with Ricci bounds from below. *Invent. Math.*, 195(2):289–391, 2014.

[And82] Brian D. O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes Appl.*, 12:313–326, 1982.

[BAM22] Simone Bombari, Mohammad Hossein Amani, and Marco Mondelli. Memorization and optimization in deep neural networks with minimum over-parameterization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[BB18] Adrien Blanchet and Jérôme Bolte. A family of functional inequalities: Łojasiewicz inequalities and displacement convex functions. *J. Funct. Anal.*, 275(7):1650–1673, 2018.

[BCC$^+$22] Antonio Blanca, Pietro Caputo, Zongchen Chen, Daniel Parisi, Daniel Štefankovič, and Eric Vigoda. On mixing of Markov chains: coupling, spectral independence, and entropy factorization. *Electron. J. Probab.*, 27:42, 2022. Id/No 142.

[BCL+18] D. P. Bourne, D. Cushing, S. Liu, F. Münch, and N. Peyerimhoff. Ollivier–ricci idleness functions of graphs. *SIAM Journal on Discrete Mathematics*, 32(2):1408–1424, 2018.

[BDBDD23] Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Linear convergence bounds for diffusion models via stochastic localization. *arXiv preprint arXiv:2308.03686*, 2023.

[BDLM10] Jérôme Bolte, Aris Daniilidis, Olivier Ley, and Laurent Mazet. Characterizations of Łojasiewicz inequalities: subgradient flows, talweg, convexity. *Trans. Am. Math. Soc.*, 362(6):3319–3363, 2010.

[BDS23a] Giovanni Brigati, Jean Dolbeault, and Nikita Simonov. Logarithmic Sobolev and interpolation inequalities on the sphere: constructive stability results. *Annales de l'Institut Henri Poincaré C*, 2023.

[BDS23b] Giovanni Brigati, Jean Dolbeault, and Nikita Simonov. Stability for the logarithmic sobolev inequality. *preprint arXiv:2303.12926*, 2023.

[BDS24] Giovanni Brigati, Jean Dolbeault, and Nikita Simonov. On Gaussian interpolation inequalities. *C. R. Math. Acad. Sci. Paris*, 362:21–44, 2024.

[BE85] D. Bakry and Michel Émery. Diffusions hypercontractives. In *Séminaire de probabilités, XIX, 1983/84*, volume 1123 of *Lecture Notes in Math.*, pages 177–206. Springer, Berlin, 1985.

[BEV17] N.H. Barton, A.M. Etheridge, and A. Véber. The infinitesimal model: Definition, derivation, and implications. *Theoretical Population Biology*, 118:50 – 73, 2017.

[BGL14] Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and geometry of Markov diffusion operators*, volume 348 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer, Cham, 2014.

[BGMZ18] Jean-Baptiste Bardet, Nathaël Gozlan, Florent Malrieu, and Pierre-André Zitt. Functional inequalities for Gaussian convolutions of compactly supported measures: Explicit bounds and dimension dependence. *Bernoulli*, 24(1):333–353, 2018.

[BL76] Herm Jan Brascamp and Elliott H. Lieb. On extensions of the Brunn-Minkowski and Prekopa-Leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation. *J. Funct. Anal.*, 22:366–389, 1976.

[BMR21] Peter L. Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *Acta Numer.*, 30:87–201, 2021.

[BMR22] Adam Block, Youssef Mroueh, and Alexander Rakhlin. Generative modeling with denoising auto-encoders and Langevin sampling. *arXiv preprint arXiv:2211.01916*, 2022.

[BMV22] Jérôme Bolte, Laurent Miclo, and Stéphane Villeneuve. Swarm gradient dynamics for global optimization: the density case, 2022.

[Bob99] Sergey Bobkov. Isoperimetric and analytic inequalities for log-concave probability measures. *Ann. Probab.*, 27(4):1903–1921, 1999.

[Bor22] Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*, 2022.

[BP24] Giovanni Brigati and Francesco Pedrotti. Heat flow, log-concavity, and Lipschitz transport maps. *arXiv preprint arXiv:2404.15205*, 2024.

[BPVF22] Etienne Boursier, Loucas Pillaud-Vivien, and Nicolas Flammarion. Gradient flow dynamics of shallow ReLU networks for square loss and orthogonal inputs, 2022.

[Bre91] Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Comm. Pure Appl. Math.*, 44(4):375–417, 1991.

[Bri23] Giovanni Brigati. *Kinetic equations, functional inequalities, and distances in the space of probability measures*. PhD thesis, Université Paris sciences et lettres; Università degli studi di Pavia, 2023.

[BT06] Sergey G. Bobkov and Prasad Tetali. Modified logarithmic Sobolev inequalities in discrete settings. *J. Theor. Probab.*, 19(2):289–336, 2006.

[BTHD21] Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion Schrödinger bridge with applications to score-based generative modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[Bul85] M.G. Bulmer. *The Mathematical Theory of Quantitative Genetics*. Oxford Science Publications. Clarendon Press, 1985.

[Caf00] Luis A. Caffarelli. Monotonicity properties of optimal transportation and the FKG and related inequalities. *Comm. Math. Phys.*, 214(3):547–563, 2000.

[Car91] Eric A. Carlen. Superadditivity of Fisher's information and logarithmic Sobolev inequalities. *J. Funct. Anal.*, 101(1):194–211, 1991.

[Car10] Eric Carlen. Trace inequalities and quantum entropy: an introductory course. In *Entropy and the quantum*, volume 529 of *Contemp. Math.*, pages 73–140. Amer. Math. Soc., Providence, RI, 2010.

[CCGL23] Patrick Cattiaux, Giovanni Conforti, Ivan Gentil, and Christian Léonard. Time Reversal of diffusion processes under a finite entropy condition. *Ann. Inst. Henri Poincaré, Probab. Stat.*, 59(4):1844–1881, 2023.

[CCL+23a] Sitan Chen, Sinho Chewi, Holden Lee, Yuanzhi Li, Jianfeng Lu, and Adil Salim. The probability flow ode is provably fast. *arXiv preprint arXiv:2305.11798*, 2023.

[CCL+23b] Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *International Conference on Learning Representations (ICLR)*, 2023.

[CCNW21] Hong-Bin Chen, Sinho Chewi, and Jonathan Niles-Weed. Dimension-free log-Sobolev inequalities for mixture distributions. *Journal of Functional Analysis*, 281(11):109236, 2021.

[CDS23] Giovanni Conforti, Alain Durmus, and Marta Gentiloni Silveri. Score diffusion models without early stopping: finite fisher information is all you need. *arXiv preprint arXiv:2308.12240*, 2023.

[CE02] Dario Cordero-Erausquin. Some applications of mass transport to Gaussian-type inequalities. *Arch. Ration. Mech. Anal.*, 161(3):257–269, 2002.

[CF21] Maria Colombo and Max Fathi. Bounds on optimal transport maps onto log-concave measures. *J. Differential Equations*, 271:1007–1022, 2021.

[CFJ17] Maria Colombo, Alessio Figalli, and Yash Jhaveri. Lipschitz changes of variables between perturbations of log-concave measures. *Ann. Sc. Norm. Super. Pisa Cl. Sci. (5)*, 17(4):1491–1519, 2017.

[CFS24] Guillaume Carlier, Alessio Figalli, and Filippo Santambrogio. On optimal transport maps between 1/d-concave densities, 2024.

[CG22] Patrick Cattiaux and Arnaud Guillin. Functional inequalities for perturbed measures with applications to log-concave measures and to some Bayesian problems. *Bernoulli*, 28(4):2294–2321, 2022.

[Cha22] Sourav Chatterjee. Convergence of gradient descent for deep neural networks, 2022.

[CHIS23] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[Clé09] Philippe Clément. Introduction to gradient flows in metric spaces. https://igk.math.uni-bielefeld.de/sites/default/files/study_materials/notes2.pdf, 2009. Lecture Notes, University of Bielefeld.

[CLL22] Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. *arXiv preprint arXiv:2211.01916*, 2022.

[CLP24] Vincent Calvez, Thomas Lepoutre, and David Poyato. Ergodicity of the Fisher infinitesimal model with quadratic selection. *Nonlinear Anal.*, 238:Paper No. 113392, 59, 2024.

[CLT22] Tianrong Chen, Guan-Horng Liu, and Evangelos Theodorou. Likelihood training of Schrödinger bridge using forward-backward SDEs theory. In *International Conference on Learning Representations (ICLR)*, 2022.

[Con22] Giovanni Conforti. A probabilistic approach to convex (ϕ)-entropy decay for Markov chains. *Ann. Appl. Probab.*, 32(2):932–973, 2022.

[CP23] Sinho Chewi and Aram-Alexandre Pooladian. An entropic generalization of Caffarelli's contraction theorem via covariance inequalities. *C. R. Math. Acad. Sci. Paris*, 361:1471–1482, 2023.

[CPP09] Pietro Caputo, Paolo Dai Pra, and Gustavo Posta. Convex entropy decay via the Bochner-Bakry-Emery approach. *Ann. Inst. Henri Poincaré, Probab. Stat.*, 45(3):734–753, 2009.

[CPS23] Vincent Calvez, David Poyato, and Filippo Santambrogio. Uniform contractivity of the Fisher infinitesimal model with strongly convex selection. *arXiv:2302.12063*, 2023.

[CZZ+21] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J. Weiss, Mohammad Norouzi, and William Chan. Wavegrad: Estimating gradients for waveform generation. In *International Conference on Learning Representations (ICLR)*, 2021.

[DEF+22] Jean Dolbeault, Maria J. Esteban, Alessio Figalli, Rupert L. Frank, and Michael Loss. Sharp stability for Sobolev and log-Sobolev inequalities, with optimal dimensional dependence. *preprint arXiv:2209.08651*, 2022.

[DEF+24] Jean Dolbeault, Maria J. Esteban, Alessio Figalli, Rupert L. Frank, and Michael Loss. A short review on improvements and stability for some interpolation inequalities. *preprint arXiv:2402.08527*, 2024.

[DGH+23] Yin Dai, Yuan Gao, Jian Huang, Yuling Jiao, Lican Kang, and Jin Liu. Lipschitz transport maps via the follmer flow. *preprint arXiv:2309.03490*, 2023.

[DIL15] D. Drusvyatskiy, A. D. Ioffe, and A. S. Lewis. Curves of descent. *SIAM J. Control Optim.*, 53(1):114–138, 2015.

[DK19] Arnak S. Dalalyan and Avetik Karagulyan. User-friendly guarantees for the langevin monte carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 129(12):5278–5311, 2019.

[DN21] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[DNS09] Jean Dolbeault, Bruno Nazaret, and Giuseppe Savaré. A new class of transport distances between measures. *Calc. Var. Partial Differential Equations*, 34(2):193–231, 2009.

[Dob70] R. L. Dobrušin. Definition of a system of random variables by means of conditional distributions. *Teor. Verojatnost. i Primenen.*, 15:469–497, 1970.

[DS85] R. L. Dobrushin and S. B. Shlosman. Constructive criterion for the uniqueness of Gibbs field. In *Statistical physics and dynamical systems (Köszeg, 1984)*, volume 10 of *Progr. Phys.*, pages 347–370. Birkhäuser Boston, Boston, MA, 1985.

[DS08] Sara Daneri and Giuseppe Savaré. Eulerian calculus for the displacement convexity in the wasserstein distance. *SIAM Journal on Mathematical Analysis*, 40(3):1104–1122, 2008.

[DSMP24] Lorenzo Dello Schiavo, Jan Maas, and Francesco Pedrotti. Local conditions for global convergence of gradient flows and proximal point sequences in metric spaces. *Trans. Am. Math. Soc.*, 377(6):3779–3804, 2024.

[DT15] Jean Dolbeault and Giuseppe Toscani. Best matching Barenblatt profiles are delayed. *J. Phys. A*, 48(6):065206, 14, 2015.

[EF18] Matthias Erbar and Max Fathi. Poincaré, modified logarithmic sobolev and isoperimetric inequalities for markov chains with non-negative ricci curvature. *Journal of Functional Analysis*, 274(11):3056–3089, 2018.

[EHMT17] Matthias Erbar, Christopher Henderson, Georg Menz, and Prasad Tetali. Ricci curvature bounds for weakly interacting Markov chains. *Electronic Journal of Probability*, 22(none):1 – 23, 2017.

[EL14] Ronen Eldan and Joseph Lehec. Bounding the norm of a log-concave vector via thin-shell estimates. In *Geometric aspects of functional analysis. Proceedings of the Israel seminar (GAFA) 2011–2013*, pages 107–122. Cham: Springer, 2014.

[EL18] Ronen Eldan and James R. Lee. Regularization under diffusion and anticoncentration of the information content. *Duke Math. J.*, 167(5):969–993, 2018.

[Eld13] Ronen Eldan. Thin shell implies spectral gap up to polylog via a stochastic localization scheme. *Geometric and Functional Analysis*, 23(2):532–569, 2013.

[ELL17] Ronen Eldan, James R. Lee, and Joseph Lehec. Transport-entropy inequalities and curvature in discrete-space Markov chains. In *A journey through discrete mathematics*, pages 391–406. Springer, 2017.

[EM12] Matthias Erbar and Jan Maas. Ricci curvature of finite Markov chains via convexity of the entropy. *Arch. Ration. Mech. Anal.*, 206(3):997–1038, 2012.

[EM14] Matthias Erbar and Jan Maas. Gradient flow structures for discrete porous medium equations. *Discrete and Continuous Dynamical Systems*, 34(4):1355–1374, 2014.

[EMT15] Matthias Erbar, Jan Maas, and Prasad Tetali. Discrete Ricci curvature bounds for Bernoulli-Laplace and random transposition models. *Ann. Fac. Sci. Toulouse, Math. (6)*, 24(4):781–800, 2015.

[Eva22] Lawrence C. Evans. *Partial differential equations*, volume 19. American Mathematical Society, 2022.

[Fat19] Max Fathi. *Quelques applications du transport optimal en analyse et en probabilités*. Habilitation à diriger des recherches, Université Paul Sabatier (Toulouse 3), April 2019.

[FIL16] Max Fathi, Emanuel Indrei, and Michel Ledoux. Quantitative logarithmic Sobolev inequalities and stability estimates. *Discrete Contin. Dyn. Syst.*, 36(12):6835–6853, 2016.

[Fis19] R. A. Fisher. XV.—The correlation between relatives on the supposition of Mendelian inheritance. *Earth and Environmental Science Transactions of The Royal Society of Edinburgh*, 52(2):399–433, 1919.

[FM16] Max Fathi and Jan Maas. Entropic Ricci curvature bounds for discrete interacting systems. *Ann. Appl. Probab.*, 26(3):1774–1806, 2016.

[FMS24] Max Fathi, Dan Mikulincer, and Yair Shenfeld. Transportation onto log-Lipschitz perturbations. *Calc. Var. Partial Differential Equations*, 63(3):Paper no. 61, 2024.

[FRY+23] Giulio Franzese, Simone Rossi, Lixuan Yang, Alessandro Finamore, Dario Rossi, Maurizio Filippone, and Pietro Michiardi. How much is enough? A study on diffusion times in score-based generative models. *Entropy*, 25(4), 2023.

[FS18] Max Fathi and Yan Shu. Curvature and transport inequalities for Markov chains in discrete spaces. *Bernoulli*, 24(1):672 – 698, 2018.

[GL10] N. Gozlan and C. Léonard. Transport inequalities. A survey. *Markov Process. Relat. Fields*, 16(4):635–736, 2010.

[GLWY09] Arnaud Guillin, Christian Léonard, Liming Wu, and Nian Yao. Transportation-information inequalities for Markov processes. *Probab. Theory Related Fields*, 144(3-4):669–695, 2009.

[Gro75] Leonard Gross. Logarithmic Sobolev inequalities. *Amer. J. Math.*, 97(4):1061–1083, 1975.

[GS84] Clark R. Givens and Rae Michael Shortt. A class of Wasserstein metrics for probability distributions. *Michigan Math. J.*, 31(2):231–240, 1984.

[GTU23] Arun Ganesh, Abhradeep Thakurta, and Jalaj Upadhyay. Universality of langevin diffusion for private optimization, with applications to sampling from rashomon sets. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 1730–1773. PMLR, 12–15 Jul 2023.

[Her99] Luis Herraiz. Asymptotic behaviour of solutions of some semilinear parabolic problems. volume 16, pages 49–105, 1999.

[HJA20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[HM19] Daniel Hauer and José M. Mazón. Kurdyka-Łojasiewicz-Simon inequality for gradient flows in metric spaces. *Trans. Am. Math. Soc.*, 372(7):4917–4976, 2019.

[IK21] Emanuel Indrei and Daesung Kim. Deficit estimates for the logarithmic Sobolev inequality. *Differential Integral Equations*, 34(7-8):437–466, 2021.

[Iof77] A. D. Ioffe. On lower semicontinuity of integral functionals. I. *SIAM J. Control Optim.*, 15:521–538, 1977.

[Iof00] A. D. Ioffe. Metric regularity and subdifferential calculus. *Russ. Math. Surv.*, 55(3):501–558, 2000.

[Ish23] Kazuhiro Ishige. Eventual concavity properties of the heat flow. *preprint arXiv:2310.14475*, 2023.

[IST21] Kazuhiro Ishige, Paolo Salani, and Asuka Takatsu. Power concavity and Dirichlet heat flow. *preprint arXiv:2105.02574*, 2021.

[IST22] Kazuhiro Ishige, Paolo Salani, and Asuka Takatsu. New characterizations of log-concavity via Dirichlet heat flow. *Ann. Mat. Pura Appl. (4)*, 201(4):1531–1552, 2022.

[IST24] Kazuhiro Ishige, Paolo Salani, and Asuka Takatsu. Hierarchy of deformations in concavity. *Inf. Geom.*, 7:S251–S269, 2024.

[Jha19] Yash Jhaveri. On the (in)stability of the identity map in optimal transportation. *Calc. Var. Partial Differential Equations*, 58(3):Paper No. 96, 25, 2019.

[JKO98] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker-Planck equation. *SIAM J. Math. Anal.*, 29(1):1–17, 1998.

[JNG+19] Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M Kakade, and Michael I Jordan. A short note on concentration inequalities for random vectors with subgaussian norm. *arXiv preprint arXiv:1902.03736*, 2019.

[JY17] Ansgar Jüngel and Wen Yue. Discrete beckner inequalities via the bochner–bakry–emery approach for markov chains. *The Annals of Applied Probability*, 27(4):2238–2269, 2017.

[Kal21] Olav Kallenberg. *Foundations of modern probability*, volume 99 of *Probability Theory and Stochastic Modelling*. Springer, Cham, third edition, 2021.

[KFL22] Dohyun Kwon, Ying Fan, and Kangwook Lee. Score-based generative modeling secretly minimizes the Wasserstein distance. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[KM12] Young-Heon Kim and Emanuel Milman. A generalization of Caffarelli's contraction theorem via (reverse) heat flow. *Math. Ann.*, 354(3):827–862, 2012.

[KMP24] Ksenia A Khudiakova, Jan Maas, and Francesco Pedrotti. L^∞-optimal transport of anisotropic log-concave measures and exponential convergence in Fisher's infinitesimal model. *arXiv preprint arXiv:2402.04151*, 2024.

[KNS16] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases*, pages 795–811, Cham, 2016. Springer International Publishing.

[KP21] Bo'az Klartag and Eli Putterman. Spectral monotonicity under gaussian convolution. *preprint arXiv:2107.09496*, 2021.

[KPH+21] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations (ICLR)*, 2021.

[KS91] Ioannis Karatzas and Steven E. Shreve. *Brownian motion and stochastic calculus*, volume 113 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, second edition, 1991.

[Kur98] Krzysztof Kurdyka. On gradients of functions definable in o-minimal structures. *Ann. Inst. Fourier*, 48(3):769–783, 1998.

[KV18] Stanislav Kondratyev and Dmitry Vorotnikov. Spherical hellinger-kantorovich gradient flows, 2018.

[L5́1] Paul Lévy. *Problèmes concrets d'analyse fonctionnelle. Avec un complément sur les fonctionnelles analytiques par F. Pellegrino*. Gauthier-Villars, Paris, 1951. 2d ed.

[Lei72] László Leindler. On a certain converse of Hölder's inequality. II. *Acta Sci. Math. (Szeged)*, 33(3-4):217–223, 1972.

[LLT22] Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence for score-based generative modeling with polynomial complexity. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[LLT23] Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence of score-based generative modeling for general data distributions. In Shipra Agrawal and Francesco Orabona, editors, *Proceedings of The 34th International Conference on Algorithmic Learning Theory*, volume 201 of *Proceedings of Machine Learning Research*, pages 946–985. PMLR, 2023.

[LLY11] Yong Lin, Linyuan Lu, and Shing-Tung Yau. Ricci curvature of graphs. *Tohoku Mathematical Journal*, 63(4):605 – 627, 2011.

[LMS22] Linshan Liu, Mateusz B. Majka, and Łukasz Szpruch. Polyak-Łojasiewicz inequality on the space of measures and convergence of mean-field birth-death processes, 2022.

[Łoj63] Stanisław Łojasiewicz. A topological property of real analytic subsets. In *Equ. Derivees partielles, Paris 1962, Colloques internat. Centre nat. Rech. sci.*, volume 117, pages 87–89, 1963.

[Łoj93] Stanislas Łojasiewicz. Sur la géométrie semi- et sous- analytique. *Ann. I. Fourier*, 43(5):1575–1595, 1993.

[LS22] Hugo Lavenant and Filippo Santambrogio. The flow map of the Fokker-Planck equation does not provide optimal transport. *Appl. Math. Lett.*, 133:Paper No. 108225, 7, 2022.

[LV03] Ki-Ahm Lee and Juan Luis Vázquez. Geometrical properties of solutions of the porous medium equation for large times. *Indiana Univ. Math. J.*, 52(4):991–1016, 2003.

[LV09] John Lott and Cedric Villani. Ricci curvature for metric-measure spaces via optimal transport. *Ann. Math. (2)*, 169(3):903–991, 2009.

[LWCC23] Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Towards faster non-asymptotic convergence for diffusion-based generative models. *arXiv preprint arXiv:2306.09251*, 2023.

[LZB+22] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[Maa11] Jan Maas. Gradient flows of the entropy for finite Markov chains. *J. Funct. Anal.*, 261(8):2250–2292, 2011.

[Maa17] Jan Maas. Entropic Ricci curvature for discrete spaces. In *Modern approaches to discrete curvature*, volume 2184 of *Lecture Notes in Math.*, pages 159–174. Springer, Cham, 2017.

[Mar70] B. Martinet. Régularisation d'inéquations variationnelles par approximations successives. *Rev. Franç. Inform. Rech. Opér.*, 4(R-3):154–158, 1970.

[McC97] Robert J. McCann. A convexity principle for interacting gases. *Adv. Math.*, 128(1):153–179, 1997.

[Mie11] Alexander Mielke. A gradient structure for reaction-diffusion systems and for energy-drift-diffusion systems. *Nonlinearity*, 24(4):1329–1346, 2011.

[Mie13] Alexander Mielke. Geodesic convexity of the relative entropy in reversible Markov chains. *Calc. Var. Partial Differ. Equ.*, 48(1-2):1–31, 2013.

[Mon23] Andrea Montanari. Sampling, diffusions, and stochastic localization. *arXiv preprint arXiv:2305.10690*, 2023.

[MS86] Vitali D. Milman and Gideon Schechtman. *Asymptotic theory of finite-dimensional normed spaces*, volume 1200 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1986. With an appendix by M. Gromov.

[MS20] Matteo Muratori and Giuseppe Savaré. Gradient flows and evolution variational inequalities in metric spaces. I: Structural properties. *J. Funct. Anal.*, 278(4):108347, 67, 2020.

[MS21] Dan Mikulincer and Yair Shenfeld. The brownian transport map. *preprint arXiv:2111.11521*, 2021.

[MS23] Dan Mikulincer and Yair Shenfeld. On the Lipschitz properties of transportation along heat flows. In *Geometric aspects of functional analysis*, volume 2327 of *Lecture Notes in Math.*, pages 269–290. Springer, Cham, [2023] ©2023.

[Mün23] Florentin Münch. Ollivier curvature, isoperimetry, concentration, and log-sobolev inequalitiy. *arXiv preprint arXiv:2309.06493*, 2023.

[MW19] Florentin Münch and Radosław K. Wojciechowski. Ollivier Ricci curvature for general graph Laplacians: heat equation, Laplacian comparison, non-explosion and diameter bounds. *Adv. Math.*, 356:45, 2019. Id/No 106759.

[Nee22] Joe Neeman. Lipschitz changes of variables via heat flow. *arXiv:2201.03403*, 2022.

[Oll09] Yann Ollivier. Ricci curvature of Markov chains on metric spaces. *J. Funct. Anal.*, 256(3):810–864, 2009.

[Oll10] Yann Ollivier. A survey of Ricci curvature for metric spaces and Markov chains. In *Probabilistic approach to geometry*, volume 57 of *Adv. Stud. Pure Math.*, pages 343–381. Math. Soc. Japan, Tokyo, 2010.

[OS19] Samet Oymak and Mahdi Soltanolkotabi. Overparameterized nonlinear learning: Gradient descent takes the shortest path? In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4951–4960. PMLR, 09–15 Jun 2019.

[Ott01] Felix Otto. The geometry of dissipative evolution equations: The porous medium equation. *Commun. Partial Differ. Equations*, 26(1-2):101–174, 2001.

[OV00] F. Otto and C. Villani. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *J. Funct. Anal.*, 173(2):361–400, 2000.

[Ped23] Francesco Pedrotti. Contractive coupling rates and curvature lower bounds for Markov chains. *arXiv preprint arXiv:2308.00516*, 2023.

[Pid22] Jakiw Pidstrigach. Score-based generative models detect manifolds. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[PMM24] Francesco Pedrotti, Jan Maas, and Marco Mondelli. Improved convergence of score-based diffusion models via prediction-correction. *Transactions on Machine Learning Research*, 2024.

[Pol63] B.T. Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

[PP13] Paolo Dai Pra and Gustavo Posta. Entropy decay for interacting systems via the Bochner-Bakry-émery approach. *Electron. J. Probab.*, 18:21, 2013. Id/No 52.

[Pre73] Andras Prekopa. On logarithmic concave measures and functions. *Acta Sci. Math.*, 34:335–343, 1973.

[PVG+21] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail A. Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning (ICML)*, 2021.

[Roc76] R. Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.

[RV08] Fraydoun Rezakhanlou and Cédric Villani. Entropy production and convergence to equilibrium. *Entropy Methods for the Boltzmann Equation: Lectures from a Special Semester at the Centre Émile Borel, Institut H. Poincaré, Paris, 2001*, pages 1–70, 2008.

[San17] Filippo Santambrogio. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bull. Math. Sci.*, 7(1):87–154, 2017.

[SBDD22] Yuyang Shi, Valentin De Bortoli, George Deligiannidis, and Arnaud Doucet. Conditional simulation using diffusion Schrödinger bridges. In *Uncertainty in Artificial Intelligence (UAI)*, 2022.

[Sch99] Michael Schmuckenschläger. Curvature of nonlocal Markov generators. In *Convex geometric analysis*, pages 189–197. Cambridge: Cambridge University Press, 1999.

[SDME21] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[SDWMG15] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML)*, 2015.

[SE19] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[SE20] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[SGSE20] Yang Song, Sahaj Garg, Jiaxin Shi, and Stefano Ermon. Sliced score matching: A scalable approach to density and score estimation. In *Uncertainty in Artificial Intelligence Conference (UAI)*, 2020.

[Sha48] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[She24] Yair Shenfeld. Exact renormalization groups and transportation of measures. *Ann. Henri Poincaré*, 25(3):1897–1910, 2024.

[Sim83] Leon Simon. Asymptotics for a class of non-linear evolution equations, with applications to geometric problems. *Ann. Math. (2)*, 118:525–571, 1983.

[Son22] Ki-Ung Song. Applying regularized Schrödinger-bridge-based stochastic process in generative modeling. *arXiv preprint arXiv:2208.07131*, 2022.

[SSDK+21] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.

[Stu06] Karl-Theodor Sturm. On the geometry of metric measure spaces. I. *Acta Math.*, 196(1):65–131, 2006.

[SW14] Adrien Saumard and Jon A. Wellner. Log-concavity and strong log-concavity: a review. *Stat. Surv.*, 8:45–114, 2014.

[Tal96] M. Talagrand. Transportation cost for Gaussian and other product measures. *Geom. Funct. Anal.*, 6(3):587–600, 1996.

[Tan21] Anastasiya Tanana. Comparison of transport map generated by heat flow interpolation and the optimal transport Brenier map. *Commun. Contemp. Math.*, 23(6):Paper No. 2050025, 7, 2021.

[Váz17] Juan Luis Vázquez. Asymptotic behaviour methods for the Heat Equation. Convergence to the Gaussian. *preprint arXiv:1706.10034*, 2017.

[Ver18] Roman Vershynin. *High Dimensional Probability*. Cambridge University Press, Cambridge, 2018.

[Vey12] Laurent Veysseire. Coarse ricci curvature for continuous-time markov processes, 2012.

[Vil03] Cédric Villani. *Topics in optimal transportation*, volume 58 of *Grad. Stud. Math.* Providence, RI: American Mathematical Society (AMS), 2003.

[Vil09] Cédric Villani. *Optimal transport*, volume 338 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2009. Old and new.

[Vin11] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.

[VW19] Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[WL18] Bruce Walsh and Michael Lynch. *Evolution and Selection of Quantitative Traits*. Oxford University Press, July 2018.

[WW16] Feng-Yu Wang and Jian Wang. Functional inequalities for convolution probability measures. *Ann. Inst. Henri Poincaré, Probab. Stat.*, 52(2):898–914, 2016.

[YYW23] Kaylee Yingxi Yang and Andre Wibisono. Convergence of the inexact Langevin algorithm and score-based generative models in KL divergence. *arxiv preprint arXiv:2211.01512*, 2023.

[YZS+23] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *arXiv preprint arXiv:2209.00796*, 2023.

[ZC23] Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator. In *International Conference on Learning Representations (ICLR)*, 2023.

[Zim13] David Zimmermann. Logarithmic Sobolev inequalities for mollified compactly supported measures. *J. Funct. Anal.*, 265(6):1064–1083, 2013.