Journal of Machine Learning Research 25 (2024) 1-45 Submitted 2/24; Revised 9/24; Published 12/24

Fisher information dissipation for time-inhomogeneous stochastic differential equations

Qi Feng QFENG2@FSU.EDU
Department of Mathematics
Florida State University
Tallahassee, FL 32306, USA

Xinzhe Zuo ZXZ@MATH.UCLA.EDU
Department of Mathematics
University of California, Los Angeles
Los Angeles, CA 90095, USA

Wuchen Li WUCHEN@MAILBOX.SC.EDU
Department of Mathematics
University of South Carolina
Columbia, SC 29208, USA

Editor: Qiang Liu

Abstract

We provide a Lyapunov convergence analysis for time-inhomogeneous variable coefficient stochastic differential equations (SDEs). Three typical examples include overdamped, irreversible drift, and underdamped Langevin dynamics. We first formulate the probability transition equation of Langevin dynamics as a modified gradient flow of the Kullback-Leibler divergence in the probability space with respect to time-dependent optimal transport metrics. This formulation contains both gradient and non-gradient directions depending on a class of time-dependent target distribution. We then select a time-dependent relative Fisher information functional as a Lyapunov functional. We develop a time-dependent Hessian matrix condition, which guarantees the convergence of the probability density function of the SDE. We verify the proposed conditions for several time-inhomogeneous Langevin dynamics. For the overdamped Langevin dynamics, we prove the $O(t^{-1/2})$ convergence in L^1 distance for the simulated annealing dynamics with a strongly convex potential function. For the irreversible drift Langevin dynamics, we prove an improved convergence towards the target distribution in an asymptotic regime. We also verify the convergence condition for the underdamped Langevin dynamics. Numerical examples demonstrate the convergence results for the time-dependent Langevin dynamics.

Keywords: Time-dependent Fisher information dissipation; Time-dependent Langevin dynamics.

1. Introduction

Time-inhomogeneous (time-dependent) stochastic dynamics are an essential class of equations, which are widely used in modeling engineering problems, designing Bayesian sampling algorithms of a target distribution, and approximating global optimization problems with applications in machine learning (Chiang et al., 1987; Geman and Hwang, 1986; Ma et al., 2021; Tang and Zhou, 2021). An important example is the stochastic dynamics from the simulated annealing method (Cerny, 1985; Kirkpatrick et al., 1983). It finds a global minimizer of a function with a time-dependent diffusion constant. The diffusion constant converges to zero when time approaches infinity. Eventually, the solution of stochastic dynamics will be a global minimizer of such a function. In recent years, general time-dependent stochastic dynamics have also been designed to maintain desired invariant distributions, such as the nonreversible Langevin sampler (Duncan et al., 2016, 2017; Zhang et al., 2022). The discretized stochastic dynamics are useful stochastic algorithms in practice. In these studies, a key consideration is the rate at which these stochastic dynamics converge to their stationary distributions. The convergence analysis can be leveraged to design and refine sampling algorithms that exhibit faster convergence.

This paper presents the convergence analysis for time-inhomogeneous stochastic dynamics, including three equations: overdamped, nonreversible drift, and underdamped Langevin dynamics. We use the time-dependent Fisher information as a Lyapunov functional to study convergence behaviors of the probability density functions of stochastic dynamics. Applying some convex analysis tools in generalized Gamma calculus (Feng and Li., 2023, 2021), we derive a time-dependent Hessian matrix condition to characterize convergence behaviors of time-dependent stochastic dynamics in Theorem 6. Lastly, we present three examples for the proposed convergence analysis. We first study the Lyapunov analysis of time-dependent overdamped Langevin dynamics based on the continuous limit of simulated annealing algorithms. When the potential function is strongly convex, we show that the Fisher information converges at a rate of $O(\frac{1}{t})$ when the diffusion coefficient is $O(\frac{1}{\log t})$, where $t > 0$ is a time variable. We then analyze the time-dependent Langevin dynamics with nonreversible drift and a nondegenerate diffusion matrix. We prove the speed-up of the convergence near the global minimizer of the potential function. Lastly, we study the convergence analysis for the inhomogeneous underdamped Langevin dynamics. Several numerical experiments are provided to justify our theoretical results.

In literature, the convergence study of time-dependent stochastic dynamics is an emerging area for stochastic algorithms in machine learning (Chizat, 2022). In this direction, the continuous-time simulated annealing based on time-dependent overdamped Langevin dynamics was first studied in Geman and Hwang (1986). It was shown in Chiang et al. (1987); Geman and Hwang (1986) that the correct order of diffusion constant for the time-dependent Lanvegin dynamics to converge to the global minimum of the objective function V is of order $(\log t)^{-1}$. Recent works (Chizat, 2022; Monmarché, 2018; Menz et al., 2018; Tang and Zhou, 2021) have shown polynomial convergence in both L^1 distance and tail probability. The state-dependent overdamped Langevin dynamics version of simulated annealing was studied in Fang et al. (1997); Gao et al. (2020).

Compared to previous results, we focus on the convergence analysis using time-dependent Fisher information functional for general time-inhomogeneous Langevin dynamics. This allows us to derive a Hessian matrix condition in establishing the convergence rates. As a special example, in time-dependent overdamped Langevin dynamics, we obtain a $O(t^{-\frac{1}{2}})$ convergence in L^1 distance under the strongly convex assumption of the potential function. On the other hand, analysis on the time-dependent Fisher information dissipation in nonreversible and underdamped Langevin dynamics is still a work in progress. This paper initializes the convergence analysis of these stochastic dynamics.

The paper is organized as follows. We formulate the main results in sections 2 and 3. Using the decay of a time-dependent Fisher information functional, we state the condition for the convergence of general stochastic differential equations. We then present several examples of convergence analysis. Section 4 provides the detailed convergence analysis for simulating annealing dynamics with a strongly convex potential function. Section 5 presents the convergence analysis for the Langevin dynamics with an irreversible drift and nondegenerate diffusion matrices. Section 6 shows the convergence analysis of underdamped Langevin dynamics. Several numerical examples are provided to verify the convergence analysis.

2. Setting

In this section, we provide the main setting of this paper. We consider the general time-dependent stochastic differential equation. We also formulate its Fokker-Planck equation, for which we develop a time-dependent decomposition of gradient and non-gradient directions in the probability density space. We then introduce the time-dependent relative Fisher information functional, which will be used in the convergence analysis of the solution of the Fokker-Planck equation.

2.1 General setting

Consider degenerate Itô type stochastic differential equations (SDEs) in $\mathbb{R}^{n+m}$ as follows:

$$dX_t = b(t, X_t)dt + \sqrt{2}a(t, X_t)dB_t. \tag{1}$$

For $m, n \in \mathbb{Z}_+$, we assume that $a \in \mathbb{C}^\infty(\mathbb{R}_+ \times \mathbb{R}^{n+m}; \mathbb{R}^{(n+m)\times n})$ is a degenerate (i.e. rectangular) time-dependent diffusion matrix, $b \in \mathbb{C}^\infty(\mathbb{R}_+ \times \mathbb{R}^{n+m}; \mathbb{R}^{n+m})$ is a time-dependent vector field, and B_t is a standard $\mathbb{R}^n$-valued Brownian motion. We denote n as the number of the columns and $n + m$ as the number of the rows for the diffusion matrix a. In what follows, we shall assume the rank of diffusion matrix a to be n, and its codimension to be m. We denote $a(t, x)^\mathsf{T}$ as the transpose of matrix $a(t, x)$, and $a(t, x)a(t, x)^\mathsf{T}$ as the standard matrix multiplication. For $i = 1, \cdots, n$, we denote $a_i^\mathsf{T} = (a(t, x)^\mathsf{T})_i$ as the row vectors of $a(t, x)^\mathsf{T}$, and $a_{\cdot i} = a(t, x)_{\cdot i}$ as the column vectors of $a(t, x)$, i.e. $a_{i\hat{i}}^\mathsf{T} = a_{\hat{i}i}$, for $\hat{i} = 1, \cdots, n + m$. For each row vector $a_i^\mathsf{T} \in \mathbb{R}^{n+m}$ with $i = 1, \cdots, n$, we denote $\mathbf{A}_i(t, x) := \sum_{\hat{i}=1}^{n+m} a_{i\hat{i}}^\mathsf{T} \frac{\partial}{\partial x_{\hat{i}}}$ as the corresponding vector fields for each row vector a_i^T. Similarly, we denote $\mathbf{A}_0(t, x) := \sum_{\hat{i}=1}^{n+m} b_{\hat{i}}(t, x) \frac{\partial}{\partial x_{\hat{i}}}$ as the vector field associated to the drift

term b. In this paper, we assume Hörmander like conditions (Hörmander, 1967) for the vector fields such that the probability density function $p(t,x)$ for the diffusion process X_t exists and is smooth. In the current time inhomogeneous setting, such conditions may include the Hörmander condition (Cattiaux and Mesnager, 2002), weak Hörmander condition (Höpfner et al., 2017), the UFG (uniformly finitely generated) condition (Cass et al., 2021), and the restricted Hörmander's hypothesis (Chaleyat-Maurel and Michel, 1984). Denote $[\mathbf{A}_i(t,x), \mathbf{A}_j(t,x)]$, for $i,j \in \{0,\cdots,n\}$, as the Lie bracket of two vector fields. The Hörmander type condition means that the Lie algebra generated by $\mathbf{A}_i(t,x)$, $1 \le i \le n$, and $\mathbf{A}_0(t,x) + \frac{\partial}{\partial t}$ has full rank. For all (t,x), we assume

$$\text{Span Lie}\Big\{\mathbf{A}_0(t,x) + \frac{\partial}{\partial t}, \mathbf{A}_1(t,x), \cdots, \mathbf{A}_n(t,x)\Big\} = \mathbb{R}^{n+m}.$$

Under the above assumptions, $p(t,x)$, which is the probability density function for X_t, satisfies the following Fokker-Planck equation of the SDE (1),

$$\partial_t p(t,x) = -\nabla \cdot (p(t,x)b(t,x)) + \sum_{i=1}^{n+m}\sum_{j=1}^{n+m} \frac{\partial^2}{\partial x_i \partial x_j}\Big(\Big(a(t,x)a(t,x)^{\mathsf{T}}\Big)_{ij} p(t,x)\Big), \qquad (2)$$

with the following initial condition

$$p_0(x) = p(0,x), \quad p_0 \in \mathcal{P}.$$

Here we denote $\mathcal{P}$ as a probability density space supported on $\mathbb{R}^{n+m}$, defined as

$$\mathcal{P} = \Big\{p \in L^1(\mathbb{R}^{n+m}) \colon \int_{\mathbb{R}^{n+m}} p(x)dx = 1, \quad p \ge 0\Big\}.$$

2.2 Time-dependent Gradient and Non-gradient decompositions

To study the convergence of the probability density function $p(t,x)$ towards the invariant distribution or the reference distribution $\pi(t,x)$. We make the following decomposition of Fokker-Planck equation (2). We assume that $\pi(t,x) \in \mathbb{C}^{2,2}(\mathbb{R}_+ \times \mathbb{R}^{n+m}; \mathbb{R})$ has an explicit analytical formula. If $\pi(t,x)$ indeed solves the equation,

$$-\nabla_x \cdot (\pi(t,x)b(t,x)) + \sum_{i=1}^{n+m}\sum_{j=1}^{n+m} \frac{\partial^2}{\partial x_i \partial x_j}\Big(\Big(a(t,x)a(t,x)^{\mathsf{T}}\Big)_{ij} \pi(t,x)\Big) = \partial_t \pi(t,x) = 0,$$

then $\pi(t,x) = \pi(x)$ is the invariant distribution. Otherwise, we use $\pi(t,x)$ as a reference distribution for the probability density function $p(t,x)$ at each time t.

The Fokker-Planck equation (2) can be decomposed into a gradient and a non-gradient part by introducing a non-gradient vector field $\gamma(t,x) : \mathbb{R}_+ \times \mathbb{R}^{n+m} \to \mathbb{R}$. The same decomposition has been used in Feng and Li. (2021), where the non-gradient vector field $\gamma(x)$ does not depend on the time variable. For self-consistency, we show the decomposition

below for the time-dependent vector fields. We first introduce the following notation, for $t \geq 0$,

$$\nabla \cdot \Big(a(t,x)a(t,x)^{\mathsf{T}}\Big) = \Big(\sum_{j=1}^{n+m} \frac{\partial}{\partial x_j} \Big(a(t,x)a(t,x)^{\mathsf{T}}\Big)_{ij}\Big)_{i=1}^{n+m} \in \mathbb{R}^{n+m}. \tag{3}$$

We then have the following decomposition.

Proposition 1 (Decomposition) *For the Fokker-Planck equation* (2) *and a reference distribution density function $\pi(t,x)$, we define a non-gradient vector field $\gamma(t,x) : \mathbb{R}_+ \times \mathbb{R}^{n+m} \to \mathbb{R}^{n+m}$ as*

$$\gamma(t,x) := \Big(a(t,x)a(t,x)^{\mathsf{T}}\Big)\nabla \log \pi(t,x) - b(x) + \nabla \cdot \Big(a(t,x)a(t,x)^{\mathsf{T}}\Big).$$

Then the Fokker-Planck equation (2) *is equivalent to the following equation:*

$$\partial_t p(t,x) = \nabla \cdot \Big(p(t,x)\Big(a(t,x)a(t,x)^{\mathsf{T}}\Big)\nabla \log \frac{p(t,x)}{\pi(t,x)}\Big) + \nabla \cdot (p(t,x)\gamma(t,x)). \tag{4}$$

Proof The proof is based on a direct calculation. For simplicity of notations, we skip the variables (t,x) below. We note

$$\begin{aligned}
&\sum_{i=1}^{n+m}\sum_{j=1}^{n+m} \frac{\partial^2}{\partial x_i \partial x_j}\Big(\Big(aa^{\mathsf{T}}\Big)_{ij} p\Big) = \sum_{i=1}^{n+m} \frac{\partial}{\partial x_i} \sum_{j=1}^{n+m} \frac{\partial}{\partial x_j}\Big(\Big(aa^{\mathsf{T}}\Big)_{ij} p(t,x)\Big) \\
=&\sum_{i=1}^{n+m} \frac{\partial}{\partial x_i} \sum_{j=1}^{n+m} \Big(\frac{\partial}{\partial x_j}\Big(aa^{\mathsf{T}}\Big)_{ij} p + \Big(aa^{\mathsf{T}}\Big)_{ij} \frac{\partial}{\partial x_j} p\Big) \\
=&\sum_{i=1}^{n+m} \frac{\partial}{\partial x_i}\Big(p \frac{\partial}{\partial x_j} \sum_{j=1}^{n+m} \Big(aa^{\mathsf{T}}\Big)_{ij}\Big) + \sum_{i,j=1}^{n+m} \frac{\partial}{\partial x_i}\Big(\Big(aa^{\mathsf{T}}\Big)_{ij} \frac{\partial}{\partial x_j} p\Big) \\
=&\sum_{i=1}^{n+m} \frac{\partial}{\partial x_i}\Big(p \frac{\partial}{\partial x_j} \sum_{j=1}^{n+m} \Big(a(x)a(x)^{\mathsf{T}}\Big)_{ij}\Big) + \sum_{i,j=1}^{n+m} \frac{\partial}{\partial x_i}\Big(\Big(aa^{\mathsf{T}}\Big)_{ij} p \frac{\partial}{\partial x_j} \log p\Big),
\end{aligned}$$

where we used the fact $\frac{\partial}{\partial x_j} p = p \frac{\partial}{\partial x_j} \log p$. From the definition of γ and the above observation, we show that the R.H.S. of the Fokker-Planck equation (2) can be written as

$$\begin{aligned}
&-\nabla \cdot (pb) + \sum_{i=1}^{n+m}\sum_{j=1}^{n+m} \frac{\partial^2}{\partial x_i \partial x_j}\Big((aa^{\mathsf{T}})_{ij} p\Big) \\
=&-\nabla \cdot (pb) + \sum_{i,j=1}^{n+m} \frac{\partial}{\partial x_i}\Big(p \frac{\partial}{\partial x_j}(aa^{\mathsf{T}})_{ij}\Big) + \nabla \cdot (p(aa^{\mathsf{T}})\nabla \log p) \\
=&-\nabla \cdot (pb) + \sum_{i,j=1}^{n+m} \frac{\partial}{\partial x_i} \cdot \Big(p \frac{\partial}{\partial x_j}(aa^{\mathsf{T}})_{ij}\Big) + \nabla \cdot (p(aa^{\mathsf{T}})\nabla \log \pi) \\
&-\nabla \cdot (p(aa^{\mathsf{T}})\nabla \log \pi) + \nabla \cdot (p(aa^{\mathsf{T}})\nabla \log p) \\
=&\nabla \cdot \Big(p(-b + \nabla \cdot (aa^{\mathsf{T}}) + aa^{\mathsf{T}} \nabla \log \pi)\Big) + \nabla \cdot (paa^{\mathsf{T}} \nabla \log \frac{p}{\pi}) \\
=&\nabla \cdot (p\gamma) + \nabla \cdot \Big(p(aa^{\mathsf{T}})\nabla \log \frac{p}{\pi}\Big),
\end{aligned}$$

where we used the definition of γ and the fact that $\nabla \log \frac{p}{\pi} = \nabla \log p - \nabla \log \pi$. ■

Remark 2 *The time-dependent hypoelliptic operator $-\nabla \cdot (paa^{\mathsf{T}}\nabla)$ is a "modified gradient operator in the Wasserstein-2 type metric space" (Villani et al., 2009). And the vector field γ is not a gradient direction (Villani, 2009). Interested readers may look for relevant discussions in the time-homogenous case (Feng and Li., 2021).*

2.3 Lyapunov functionals

To measure the distance between $p(t,x)$ and $\pi(t,x)$, as well as the corresponding convergence rate towards $\pi(t,x)$, we define the Kullback–Leibler (KL) divergence

$$\mathrm{D}_{\mathrm{KL}}(p(t,\cdot)\|\pi(t,\cdot)) := \int_{\mathbb{R}^{n+m}} p(t,x) \log \frac{p(t,x)}{\pi(t,x)} dx. \tag{5}$$

For $t \geq 0$, and a diffusion matrix $a(t,x) \in \mathbb{C}^\infty(\mathbb{R}_+ \times \mathbb{R}^{n+m}; \mathbb{R}^{(n+m)\times n})$ associated with SDE (1) with rank n, we introduce a complementary matrix, defined as,

$$z(t,x) \in \mathbb{C}^\infty(\mathbb{R}_+ \times \mathbb{R}^{n+m}; \mathbb{R}^{(n+m)\times m}), \tag{6}$$

such that, for all $t \geq 0$,

$$\mathrm{Rank}\Big(a(t,x)a(t,x)^{\mathsf{T}} + z(t,x)z(t,x)^{\mathsf{T}}\Big) = n+m, \quad \text{for all} \;\; x \in \mathbb{R}^{n+m}. \tag{7}$$

Adapted from the previous notation, we denote a^{T} and z^{T} as the transpose of matrices $a(t,x)$ and $z(t,x)$. We denote $\{a_i^{\mathsf{T}}\}_{i=1}^n$ and $\{z_j^{\mathsf{T}}\}_{j=1}^m$ as the row vectors of a^{T} and z^{T}. Here the matrix z is selected in a way such that the full rank of aa^{T} (with rank n) and zz^{T} (with rank m) equals to $n+m$. Thus, the matrix $aa^{\mathsf{T}} + zz^{\mathsf{T}}$ can be used as a non-degenerate metric for the entire space $\mathbb{R}^{n+m}$. The condition (7) means that the linear span of the row vectors $\{a_i^{\mathsf{T}}\}_{i=1}^n$ and $\{z_j^{\mathsf{T}}\}_{j=1}^m$ generate the entire space $\mathbb{R}^{n+m}$ for all $t \geq 0$. Furthermore, to ensure that the Bochner's formula (Feng and Li., 2021, Theorem 1) holds, we assume that, for $0 \leq k \leq m$, $0 \leq i \leq n$,

$$\mathbf{Z}_k(t,x)\mathbf{A}_i(t,x) \in \mathrm{Span}\{\mathbf{A}_j(t,x), 0 \leq j \leq n\}, \quad \text{for all} \quad t \geq 0, \quad \text{and} \quad x \in \mathbb{R}^{n+m}, \tag{8}$$

where we denote $\mathbf{Z}_k(t,x)$ as the corresponding vector field for each row vector z_k^{T}. For a smooth function $f \in \mathbb{C}^\infty(\mathbb{R}^{n+m})$, we denote the gradient vector field ∇f as a column vector,

$$\nabla f(x) = \Big(\frac{\partial f}{\partial x_1,} \cdots, \frac{\partial f}{\partial x_{n+m}}\Big)^{\mathsf{T}}. \tag{9}$$

We keep the following notation throughout the paper. A standard multiplication of a row vector and a column vector has the following form,

$$a_k^{\mathsf{T}} \nabla f = \sum_{k'=1}^{n+m} a_{kk'}^{\mathsf{T}} \frac{\partial f}{\partial x_{k'}}. \tag{10}$$

Similarly, we denote

$$a_k^{\mathsf{T}}\nabla a_i^{\mathsf{T}}\nabla f = a_k^{\mathsf{T}}(\nabla a_i^{\mathsf{T}})\nabla f = \sum_{k',i'=1}^{n+m} a_{kk'}^{\mathsf{T}}\frac{\partial a_{ii'}^{\mathsf{T}}}{\partial x_{k'}}\frac{\partial f}{\partial x_{i'}}, \tag{11}$$

where the gradient is always applied to the function next to it. Given matrices $a(t,x)$, $z(t,x)$, and the reference measure $\pi(t,x)$ as above, we introduce the following relative Fisher information functionals as our Lyapunov functionals. Denote $\langle u, v\rangle = \sum_{i=1}^{n+m} u_i v_i$, for any vectors $u, v \in \mathbb{R}^{n+m}$.

Definition 3 (Fisher information functionals) *Define a functional $\mathcal{I}_a\colon \mathcal{P} \to \mathbb{R}_+$ as*

$$\mathrm{I}_a(p(t,\cdot)\|\pi(t,\cdot)) := \int_{\mathbb{R}^{n+m}} \Big\langle \nabla \log\frac{p(t,x)}{\pi(t,x)}, a(t,x)a(t,x)^{\mathsf{T}}\nabla\log\frac{p(t,x)}{\pi(t,x)}\Big\rangle p(t,x)dx, \tag{12}$$

Define an auxiliary functional $\mathcal{I}_z\colon \mathcal{P} \to \mathbb{R}_+$ as

$$\mathrm{I}_z(p(t,\cdot)\|\pi(t,\cdot)) := \int_{\mathbb{R}^{n+m}} \Big\langle \nabla \log\frac{p(t,x)}{\pi(t,x)}, z(t,x)z(t,x)^{\mathsf{T}}\nabla\log\frac{p(t,x)}{\pi(t,x)}\Big\rangle p(t,x)dx. \tag{13}$$

3. Time-dependent Fisher information decay

In this section, we present the main theoretical analysis. We use the time-dependent original and auxiliary Fisher information functionals as Lyapunov functionals for the convergence of the Fokker-Planck equation in Theorem 6.

We shall derive the dissipation of KL divergence and Fisher information along time-inhomogenous equations. We first show the relation between the KL divergence and the Fisher information functional in this time-dependent setting.

Proposition 4 *For $t \geq 0$, we have*

$$\begin{aligned}\partial_t \mathrm{D}_{\mathrm{KL}}(p\|\pi) = &-\int_{\mathbb{R}^{n+m}} \langle \nabla\log\frac{p(t,x)}{\pi(t,x)}, aa^{\mathsf{T}}\nabla\log\frac{p(t,x)}{\pi(t,x)}\rangle p(t,x)dx \\ &-\int_{\mathbb{R}^{n+m}} \mathcal{R}(t,x,\pi)p(t,x)dx,\end{aligned} \tag{14}$$

where we define the correction term $\mathcal{R}(t,x,\pi) : \mathbb{R}_+ \times \mathbb{R}^{n+m} \times \mathcal{P} \to \mathbb{R}$ as below,

$$\mathcal{R}(t,x,\pi) := \frac{\partial_t\pi(t,x) - \nabla\cdot(\pi(t,x)\gamma(t,x))}{\pi(t,x)}. \tag{15}$$

Remark 5 *Note that if $\pi(t,x) = \pi(x)$ is the invariant measure, we have $\nabla\cdot(\pi\gamma) = 0$, hence $\mathcal{R}(t,x,\pi) = 0$. However, in the more general setting, $\nabla\cdot(\pi(t,x)\gamma(t,x)) \neq 0$ for a general reference measure $\pi(t,x)$. Thus, we introduce the correction term $\mathcal{R}(t,x,\pi)$ in (15), which is equivalent to $\partial_t\log\pi(t,x) - \langle\nabla\log\pi(t,x),\gamma\rangle - \nabla\cdot(\gamma)$. The first and second terms correspond to the time and spacial derivatives of $\log\pi(t,x)$, while the third term coppresponds to the non-reversible vector field γ.*

For simplicity of notation in all proofs, we shall denote $\int_{\mathbb{R}^{n+m}}$ as $\int$. We also skip the variables (t, x) to simplify the notation.

Proof We derive the entropy dissipation as below,

$$\begin{aligned}\partial_t \mathrm{D}_{\mathrm{KL}}(p\|\pi) &= \int \partial_t p \log \frac{p}{\pi} dx + \int p \partial_t \log p dx - \int p \partial_t \log \pi dx \\ &= \int \big[\nabla \cdot (p\gamma) + \nabla \cdot (paa^{\mathsf{T}} \nabla \log \frac{p}{\pi})\big] \log \frac{p}{\pi} dx + \int \partial_t p dx - \int p \partial_t \log \pi dx \\ &= -\int \langle \nabla \log \frac{p}{\pi}, aa^{\mathsf{T}} \nabla \log \frac{p}{\pi} \rangle p dx + \int \nabla \cdot (p\gamma) \log \frac{p}{\pi} dx - \int p \partial_t \log \pi dx.\end{aligned}$$

Furthermore, we have

$$\begin{aligned}\int \nabla \cdot (p\gamma) \log \frac{p}{\pi} dx &= -\int \langle \nabla \log \frac{p}{\pi}, \gamma \rangle p dx \\ &= -\int \langle \nabla p, \gamma \rangle dx + \int \langle \nabla \log \pi, \gamma \rangle p dx \\ &= \int (\nabla \cdot \gamma + \langle \nabla \log \pi, \gamma \rangle) p dx \\ &= \int \frac{p}{\pi} [(\nabla \cdot \gamma)\pi + \langle \nabla \pi, \gamma \rangle] dx \\ &= \int \frac{\nabla \cdot (\pi\gamma)}{\pi} p dx.\end{aligned}$$

Combining the above terms, we complete the proof. ■

3.1 Fisher information decay

In this subsection, we first present the Fisher information functional dissipation result. The proof will be postponed to Section 3.3, and Section 3.4. To simplify our notation, we define

$$\mathrm{I}_{a,z}(t) = \mathrm{I}_a(p\|\pi) + \mathrm{I}_z(p\|\pi). \tag{16}$$

Theorem 6 (Fisher decay) *We define $\mathfrak{R}(t, x) : \mathbb{R}_+ \times \mathbb{R}^{n+m} \to \mathbb{R}^{(n+m)\times(n+m)}$ as the corresponding time-dependent Hessian matrix function, which is defined in the Appendix A. Assume that*

$$\mathfrak{R}(t, x) - \frac{1}{2}\partial_t(aa^{\mathsf{T}} + zz^{\mathsf{T}})(t, x) \succeq \lambda(t)[aa^{\mathsf{T}} + zz^{\mathsf{T}}](t, x), \tag{17}$$

for all $x \in \mathbb{R}^{n+m}$, and for all $t \geq 0$. We have

$$\mathrm{I}_{a,z}(t) \leq e^{-2\int_{t_0}^t \lambda(r)dr} \Big(\int_{t_0}^t 2[\mathrm{A}(r) + \mathrm{Z}(r)] e^{2\int_{t_0}^r \lambda(\tau)d\tau} dr + \mathrm{I}_{a,z}(t_0) \Big),$$

where the correction term $\mathcal{R}(\cdot,\cdot,\cdot)$ *is introduced in* (15). *And*

$$\begin{aligned}\mathrm{A}(r) &= \int_{\mathbb{R}^{n+m}} \big[\nabla\cdot(aa^{\mathsf{T}}\nabla\mathcal{R}) + \langle\nabla\mathcal{R}, aa^{\mathsf{T}}\nabla\log\pi\rangle\big]p(r,x)dx,\\ \mathrm{Z}(r) &= \int_{\mathbb{R}^{n+m}} \big[\nabla\cdot(zz^{\mathsf{T}}\nabla\mathcal{R}) + \langle\nabla\mathcal{R}, zz^{\mathsf{T}}\nabla\log\pi\rangle\big]p(r,x)dx.\end{aligned}$$

Proof From the definition, we have $\mathrm{I}_{a,z}(t) = \mathrm{I}_{a,z}(p\|\pi) = \mathrm{I}_a(p\|\pi) + \mathrm{I}_z(p\|\pi)$. According to Proposition 10 in the next section, and Assumption (17), we have

$$\partial_t \mathrm{I}_{a,z}(p\|\pi) \le -2\lambda(t)\mathrm{I}_{a,z}(p\|\pi) + 2[\mathrm{A}(t)+\mathrm{Z}(t)].$$

We next construct a function $Q(t)$, such that

$$\partial_t Q(t) + 2\lambda(t)Q(t) = 2[\mathrm{A}(t)+\mathrm{Z}(t)].$$

Let $F(t) = -2\int_{t_0}^t \lambda(r)dr$. We obtain $Q(t) = Q(t_0)e^{F(t)} + e^{F(t)}\int_{t_0}^t 2[\mathrm{A}(s)+\mathrm{Z}(s)]e^{-F(s)}ds$, which implies

$$\partial_t(\mathrm{I}_{a,z}(t) - Q(t)) \le -2\lambda(t)(\mathrm{I}_{a,z}(t) - Q(t)).$$

From Gronwall's inequality, we have

$$\begin{aligned}\mathrm{I}_{a,z}(t) &\le Q(t) + (\mathrm{I}_{a,z}(t_0) - Q(t_0))e^{-2\int_{t_0}^t \lambda(r)dr}\\ &= Q(t_0)e^{-2\int_{t_0}^t \lambda(r)dr} + e^{\int_{t_0}^{\lambda}(r)dr}\int_{t_0}^t 2[\mathrm{A}(r)+\mathrm{Z}(r)]e^{\int_{t_0}^r 2\lambda(\tau)d\tau}dr\\ &\quad + (\mathrm{I}_a(t_0) - Q(t_0))e^{-2\int_{t_0}^t \lambda(r)dr}\\ &= e^{-2\int_{t_0}^t \lambda(r)dr}\Big(\int_{t_0}^t 2[\mathrm{A}(r)+\mathrm{Z}(r)]e^{2\int_{t_0}^r \lambda(\tau)d\tau}dr + \mathrm{I}_{a,z}(t_0)\Big).\end{aligned}$$

This finishes the proof. ■

3.2 Information Gamma calculus

To derive the dissipation of the Fisher information functional, we first introduce the information Gamma calculus in the current setting. These information Gamma operators are generalized from the Carrè du champ operators, see Bakry and Émery (2006) for elliptic operator setting, and Baudoin and Garofalo (2016) for hypoelliptic operator setting. We refer to Feng and Li. (2023, 2021); Bayraktar et al. (2024) for more motivations and detailed discussions on these operators. We follow closely the notations as in Feng and Li. (2021, Definition 2) below. Following the decomposition in Proposition 1, the diffusion operator L associated with SDE (1) is defined in the following form, for smooth function $f:\mathbb{R}^{n+m}\to\mathbb{R}$,

$$Lf = \widetilde{L}f - \langle\gamma(t,x),\nabla f\rangle, \tag{18}$$

where we define the reversible component of the diffusion operator L as below,

$$\widetilde{L}f \quad = \quad \nabla \cdot (a(t,x)a(t,x)^{\mathsf{T}}\nabla f) + \langle a(t,x)a(t,x)^{\mathsf{T}}\nabla \log \pi(t,x), \nabla f\rangle. \tag{19}$$

For the diffusion matrix function $a(t,x)$, we construct a matrix $z(t,x) \in \mathbb{C}^{\infty}(\mathbb{R}_+ \times \mathbb{R}^{n+m}; \mathbb{R}^{(n+m)\times m})$ such that conditions (7), (8), and the Hörmander condition hold true. We then introduce the following z-direction differential operator as

$$\widetilde{L}_z f = \nabla \cdot (z(t,x)z(t,x)^{\mathsf{T}}\nabla f) + \langle z(t,x)z(t,x)^{\mathsf{T}}\nabla \log \pi(t,x), \nabla f\rangle.$$

The Gamma one bilinear forms (also known as Carrè du champ operators) for the matrices $a(t,x)$ and $z(t,x)$ are defined as below, $\Gamma_1, \Gamma_1^z \colon \mathbb{C}^{\infty}(\mathbb{R}^{n+m}) \times \mathbb{C}^{\infty}(\mathbb{R}^{n+m}) \to \mathbb{C}^{\infty}(\mathbb{R}^{n+m})$ as

$$\Gamma_1(f,f) = \langle a(t,x)^{\mathsf{T}}\nabla f, a(t,x)^{\mathsf{T}}\nabla f\rangle_{\mathbb{R}^n}, \quad \Gamma_1^z(f,f) = \langle z(t,x)^{\mathsf{T}}\nabla f, z(t,x)^{\mathsf{T}}\nabla f\rangle_{\mathbb{R}^m}. \tag{20}$$

Definition 7 (Time-dependent Information Gamma operators) *For operators $\widetilde{L}$ and $\widetilde{L}_z$, we define the following Information Gamma operators.*

(i) Gamma two operator:

$$\widetilde{\Gamma}_2(f,f) := \frac{1}{2}\widetilde{L}\Gamma_1(f,f) - \Gamma_1(\widetilde{L}f, f).$$

(ii) Generalized Gamma z operator:

$$\begin{aligned}\widetilde{\Gamma}_2^{z,\pi}(f,f) \quad := \quad & \frac{1}{2}\widetilde{L}\Gamma_1^z(f,f) - \Gamma_1^z(\widetilde{L}f, f) \\ & + \mathbf{div}_z^{\pi}\Big(\Gamma_{1,\nabla(aa^{\mathsf{T}})}(f,f)\Big) - \mathbf{div}_a^{\pi}\Big(\Gamma_{1,\nabla(zz^{\mathsf{T}})}(f,f)\Big).\end{aligned}$$

Here $\mathbf{div}_a^{\pi}$, $\mathbf{div}_z^{\pi}$ are divergence operators defined by

$$\mathbf{div}_a^{\pi}(F) := \frac{1}{\pi}\nabla \cdot (\pi a a^{\mathsf{T}} F), \quad \mathbf{div}_z^{\pi}(F) := \frac{1}{\pi}\nabla \cdot (\pi z z^{\mathsf{T}} F),$$

for any smooth vector field $F \in \mathbb{R}^{n+m}$, and $\Gamma_{1,\nabla(aa^{\mathsf{T}})}$, $\Gamma_{1,\nabla(zz^{\mathsf{T}})}$ are vector Gamma one bilinear forms defined by

$$\begin{aligned}\Gamma_{1,\nabla(aa^{\mathsf{T}})}(f,f) \quad &:= \quad \langle \nabla f, \nabla(aa^{\mathsf{T}})\nabla f\rangle = (\langle \nabla f, \frac{\partial}{\partial x_{\hat{k}}}(aa^{\mathsf{T}})\nabla f\rangle)_{\hat{k}=1}^{n+m}, \\ \Gamma_{1,\nabla(zz^{\mathsf{T}})}(f,f) \quad &:= \quad \langle \nabla f, \nabla(aa^{\mathsf{T}})\nabla f\rangle = (\langle \nabla f, \frac{\partial}{\partial x_{\hat{k}}}(zz^{\mathsf{T}})\nabla f\rangle)_{\hat{k}=1}^{n+m}.\end{aligned}$$

Definition 8 (Irreversible Gamma operator)

$$\Gamma_{\mathcal{I}_a}(f,f) + \Gamma_{\mathcal{I}_z}(f,f) \quad := \quad (\widetilde{L}f + \widetilde{L}_z f)\langle \nabla f, \gamma\rangle - \frac{1}{2}\langle \nabla\big(\Gamma_1(f,f) + \Gamma_1^z(f,f)\big), \gamma\rangle.$$

Remark 9 *One key difference in the current setting compared to (Feng and Li., 2021) is the fact* $\nabla \cdot (\pi(t,x)\gamma(t,x))$ *can be non-zero. Due to the decomposition of operator* L *in* (18)*, the time-dependent vector field* $\gamma(t,x)$ *does not make a difference for the second order operators* $\widetilde{\Gamma}_2$ *and* $\widetilde{\Gamma}_2^{z,\pi}$*. Thus the Bochner's formula for* $\widetilde{\Gamma}_2$ *and* $\widetilde{\Gamma}_2^{z,\pi}$ *remains the same as in (Feng and Li., 2021), we postpone the Bochner's formula to the Appendix. The expressions for* $\Gamma_{\mathcal{I}_a}(f,f)$ *and* $\Gamma_{\mathcal{I}_z}(f,f)$ *are different, see Lemma 14 and Lemma 17 below.*

By using the time-dependent Information Gamma operator defined above, we have the following estimates for the first order dissipation of the Fisher Information functional.

Proposition 10

$$\begin{aligned}
&\partial_t[\mathrm{I}_a(p\|\pi)+\mathrm{I}_z(p\|\pi)]\\
\leq& -2\int \mathfrak{R}(\nabla\log\frac{p}{\pi},\nabla\log\frac{p}{\pi})pdx+\int\langle\nabla\log\frac{p}{\pi},\partial_t(aa^{\mathsf{T}}+zz^{\mathsf{T}})\nabla\log\frac{p}{\pi}\rangle pdx\\
&+\int\big[2\nabla\cdot((aa^{\mathsf{T}}+zz^{\mathsf{T}})\nabla\mathcal{R})+2\langle\nabla\mathcal{R},(aa^{\mathsf{T}}+zz^{\mathsf{T}})\nabla\log\pi\rangle\big]pdx,
\end{aligned}\tag{21}$$

where the correction term $\mathcal{R}(t,x,\pi)$ *is defined in* (15)*. And the Hessian matrix* $\mathfrak{R}(t,x)$ *is defined in the Appendix A.*

Proof Combining Proposition 11 and Proposition 15 below, we have

$$\begin{aligned}
&\partial_t[\mathrm{I}_a(p\|\pi)+\mathrm{I}_z(p\|\pi)]\\
=&-2\int\big[\widetilde{\Gamma}_2(\log\frac{p}{\pi},\log\frac{p}{\pi})\big]pdx+\int\langle\nabla\log\frac{p}{\pi},\partial_t(aa^{\mathsf{T}})\nabla\log\frac{p}{\pi}\rangle pdx\\
&-\int\langle\gamma,\langle\nabla\log\frac{p}{\pi},\nabla(aa^{\mathsf{T}})\nabla\log\frac{p}{\pi}\rangle\rangle pdx+2\int\langle aa^{\mathsf{T}}\nabla\log\frac{p}{\pi},\nabla\gamma\nabla\log\frac{p}{\pi}\rangle pdx\\
&-2\int\widetilde{\Gamma}_2^{z,\pi}(\log\frac{p}{\pi},\log\frac{p}{\pi})pdx+\int\langle\nabla\log\frac{p}{\pi},\partial_t(zz^{\mathsf{T}})\nabla\log\frac{p}{\pi}\rangle pdx\\
&-\int\langle\gamma,\langle\nabla\log\frac{p}{\pi},\nabla(zz^{\mathsf{T}})\nabla\log\frac{p}{\pi}\rangle\rangle pdx+2\int\langle zz^{\mathsf{T}}\nabla\log\frac{p}{\pi},\nabla\gamma\nabla\log\frac{p}{\pi}\rangle pdx\\
&+\int\big[2\nabla\cdot((aa^{\mathsf{T}}+zz^{\mathsf{T}})\nabla\mathcal{R})+2\langle\nabla\mathcal{R},(aa^{\mathsf{T}}+zz^{\mathsf{T}})\nabla\log\pi\rangle\big]pdx\\
=&-2\int\big[\widetilde{\Gamma}_2(\log\frac{p}{\pi},\log\frac{p}{\pi})\big]pdx-2\int\widetilde{\Gamma}_2^{z,\pi}(\log\frac{p}{\pi},\log\frac{p}{\pi})pdx\\
&+\int\langle\nabla\log\frac{p}{\pi},\partial_t(aa^{\mathsf{T}})\nabla\log\frac{p}{\pi}\rangle pdx+\int\langle\nabla\log\frac{p}{\pi},\partial_t(zz^{\mathsf{T}})\nabla\log\frac{p}{\pi}\rangle pdx\\
&-\int\langle\gamma,\langle\nabla\log\frac{p}{\pi},\nabla(aa^{\mathsf{T}})\nabla\log\frac{p}{\pi}\rangle\rangle pdx+2\int\langle aa^{\mathsf{T}}\nabla\log\frac{p}{\pi},\nabla\gamma\nabla\log\frac{p}{\pi}\rangle pdx\\
&-\int\langle\gamma,\langle\nabla\log\frac{p}{\pi},\nabla(zz^{\mathsf{T}})\nabla\log\frac{p}{\pi}\rangle\rangle pdx+2\int\langle zz^{\mathsf{T}}\nabla\log\frac{p}{\pi},\nabla\gamma\nabla\log\frac{p}{\pi}\rangle pdx\\
&+\int\big[2\nabla\cdot((aa^{\mathsf{T}}+zz^{\mathsf{T}})\nabla\mathcal{R})+2\langle\nabla\mathcal{R},(aa^{\mathsf{T}}+zz^{\mathsf{T}})\nabla\log\pi\rangle\big]pdx.
\end{aligned}\tag{22}$$

Applying the Information Bochner's formula from Proposition 32 in the Appendix, for the first term in (22), we have

$$\begin{aligned} &-2\int \big[\widetilde{\Gamma}_2(\log\frac{p}{\pi},\log\frac{p}{\pi})\big]pdx - 2\int \widetilde{\Gamma}_2^{z,\pi}(\log\frac{p}{\pi},\log\frac{p}{\pi})pdx \\ =&-2\int \Big(\|\mathfrak{Hess}_\beta f\|_{\mathrm{F}}^2 + (\mathfrak{R}_a+\mathfrak{R}_z+\mathfrak{R}_\pi)(\nabla\log\frac{p}{\pi},\nabla\log\frac{p}{\pi})\Big)pdx, \end{aligned} \tag{23}$$

Note that

$$\begin{aligned} &-\int\langle\gamma,\langle\nabla\log\frac{p}{\pi},\nabla(aa^{\mathsf{T}}+zz^{\mathsf{T}})\nabla\log\frac{p}{\pi}\rangle\rangle pdx + 2\int\langle(aa^{\mathsf{T}}+zz^{\mathsf{T}})\nabla\log\frac{p}{\pi},\nabla\gamma\nabla\log\frac{p}{\pi}\rangle pdx \\ &=-2\int(\mathfrak{R}_{\gamma_a}+\mathfrak{R}_{\gamma_z})(\nabla\log\frac{p}{\pi},\nabla\log\frac{p}{\pi})pdx. \end{aligned} \tag{24}$$

Plugging (23) and (24) into (22), we finish the proof, since $\|\mathfrak{Hess}_\beta f\|_{\mathrm{F}}^2 \geq 0$. ■

3.3 Dissipation of $\mathrm{I}_a(p\|\pi)$

Now we are ready to present the following technical lemmas for first order dissipation of $\mathrm{I}_a(p\|\pi)$.

Proposition 11

$$\begin{aligned} \partial_t \mathrm{I}_a(p\|\pi) &= -2\int\big[\widetilde{\Gamma}_2(\log\frac{p}{\pi},\log\frac{p}{\pi})\big]pdx + \int\langle\nabla\log\frac{p}{\pi},\partial_t(aa^{\mathsf{T}})\nabla\log\frac{p}{\pi}\rangle pdx \\ &\quad -\int\langle\gamma,\langle\nabla\log\frac{p}{\pi},\nabla(aa^{\mathsf{T}})\nabla\log\frac{p}{\pi}\rangle\rangle pdx + 2\int\langle aa^{\mathsf{T}}\nabla\log\frac{p}{\pi},\nabla\gamma\nabla\log\frac{p}{\pi}\rangle pdx \\ &\quad +\int\big[2\nabla\cdot(aa^{\mathsf{T}}\nabla\mathcal{R}) + 2\langle\nabla\mathcal{R},aa^{\mathsf{T}}\nabla\log\pi\rangle\big]pdx. \end{aligned}$$

Proof The proof of Proposition 11 follows from Lemma 13 and Lemma 14. According to Lemma 14, we have

$$\begin{aligned} \int\widetilde{\Gamma}_{\mathcal{I}_a}(f,f)pdx &= \frac{1}{2}\int\langle\gamma,\langle\nabla f,\nabla(aa^{\mathsf{T}})\nabla f\rangle\rangle pdx - \int\langle aa^{\mathsf{T}}\nabla f,\nabla\gamma\nabla f\rangle pdx \\ &\quad +\int\frac{\nabla\cdot(\pi\gamma)}{\pi}\Gamma_1(f,f)pdx. \end{aligned}$$

Plugging into Lemma 13 with $f=\log\frac{p}{\pi}$, we prove the results. ■

Remark 12 *The extra term in Lemma 14 involving $\nabla\cdot(\pi\gamma)$ is canceled by the extra term in Lemma 13. This makes it possible to define the tensor $\mathfrak{R}_{\gamma_a}$ the same as in the time independent setting (Feng and Li., 2021).*

Lemma 13

$$\begin{aligned}\partial_t \mathrm{I}_a(p\|\pi) &= -2\int \big[\widetilde{\Gamma}_2(\log\frac{p}{\pi},\log\frac{p}{\pi})+\widetilde{\Gamma}_{\mathcal{I}_a}(\log\frac{p}{\pi},\log\frac{p}{\pi})\big]pdx \\ &+\int\langle\nabla\log\frac{p}{\pi},[\partial_t(aa^{\mathsf{T}})+2aa^{\mathsf{T}}\frac{\nabla\cdot(\pi\gamma)}{\pi}]\nabla\log\frac{p}{\pi}\rangle pdx \\ &+\int\big[2\nabla\cdot(aa^{\mathsf{T}}\nabla\mathcal{R})+2\langle\nabla\mathcal{R},aa^{\mathsf{T}}\nabla\log\pi\rangle\big]pdx.\end{aligned} \tag{25}$$

Proof

$$\begin{aligned}\partial_t \mathrm{I}_a(p\|\pi) &= 2\int\langle\nabla\partial_t\log\frac{p}{\pi},aa^{\mathsf{T}}\nabla\log\frac{p}{\pi}\rangle pdx+\int\langle\nabla\log\frac{p}{\pi},\partial_t(aa^{\mathsf{T}})\nabla\log\frac{p}{\pi}\rangle pdx \\ &+\int\langle\nabla\log\frac{p}{\pi},aa^{\mathsf{T}}\nabla\log\frac{p}{\pi}\rangle\partial_t pdx \\ &= 2\int\langle\nabla\partial_t\log p,aa^{\mathsf{T}}\nabla\log\frac{p}{\pi}\rangle pdx+\int\langle\nabla\log\frac{p}{\pi},aa^{\mathsf{T}}\nabla\log\frac{p}{\pi}\rangle\partial_t pdx \\ &+\int\langle\nabla\log\frac{p}{\pi},\partial_t(aa^{\mathsf{T}})\nabla\log\frac{p}{\pi}\rangle pdx-2\int\langle\nabla\partial_t\log\pi,aa^{\mathsf{T}}\nabla\log\frac{p}{\pi}\rangle pdx.\end{aligned}$$

We first observe that,

$$\begin{aligned}&2\int\langle\nabla\partial_t\log p,aa^{\mathsf{T}}\nabla\log\frac{p}{\pi}\rangle pdx+\int\langle\nabla\log\frac{p}{\pi},aa^{\mathsf{T}}\nabla\log\frac{p}{\pi}\rangle\partial_t pdx \\ =&\int\Gamma_1(\log\frac{p}{\pi},\log\frac{p}{\pi})\partial_t p-2\frac{\nabla\cdot(paa^{\mathsf{T}}\nabla\log\frac{p}{\pi})}{p}\partial_t pdx \\ =&\int\Gamma_1(\log\frac{p}{\pi},\log\frac{p}{\pi})\partial_t p-2\Big(\langle\nabla\log p,aa^{\mathsf{T}}\nabla\log\frac{p}{\pi}\rangle+\nabla\cdot(aa^{\mathsf{T}}\nabla\log\frac{p}{\pi})\Big)\partial_t pdx \\ =&\int\Gamma_1(\log\frac{p}{\pi},\log\frac{p}{\pi})\partial_t p-2\Big(\langle\nabla\log\frac{p}{\pi},aa^{\mathsf{T}}\nabla\log\frac{p}{\pi}\rangle+\widetilde{L}\log\frac{p}{\pi}\Big)\partial_t pdx \\ =&-2\int\Big\{\frac{1}{2}\Gamma_1(\log\frac{p}{\pi},\log\frac{p}{\pi})\partial_t p+\widetilde{L}\log\frac{p}{\pi}\partial_t p\Big\}dx \\ =&-2\int\Big\{\quad\frac{1}{2}\Gamma_1(\log\frac{p}{\pi},\log\frac{p}{\pi})\nabla\cdot(p\gamma)+\widetilde{L}\log\frac{p}{\pi}\nabla\cdot(p\gamma) \\ &\qquad+\frac{1}{2}\Gamma_1(\log\frac{p}{\pi},\log\frac{p}{\pi})\widetilde{L}^*p+\widetilde{L}\log\frac{p}{\pi}\widetilde{L}^*p\Big\}dx \\ =&-2\int\Big\{-\frac{1}{2}\langle\nabla\Gamma_1(\log\frac{p}{\pi},\log\frac{p}{\pi}),\gamma\rangle+\widetilde{L}\log\frac{p}{\pi}\langle\nabla\log\frac{p}{\pi},\gamma\rangle+\widetilde{L}\log\frac{p}{\pi}\frac{\nabla\cdot(\pi\gamma)}{\pi} \\ &\qquad+\frac{1}{2}\widetilde{L}\Gamma_1(\log\frac{p}{\pi},\log\frac{p}{\pi})-\Gamma_1(\widetilde{L}\log\frac{p}{\pi},\log\frac{p}{\pi})\Big\}pdx \\ =&-2\int[\widetilde{\Gamma}_2(\log\frac{p}{\pi},\log\frac{p}{\pi})+\widetilde{\Gamma}_{\mathcal{I}_a}(\log\frac{p}{\pi},\log\frac{p}{\pi})]pdx-2\int\tilde{L}\log\frac{p}{\pi}\frac{\nabla\cdot(\pi\gamma)}{\pi}pdx.\end{aligned}$$

In the second last equality, we apply the following fact $p\nabla \log p = \nabla p$, $\pi \nabla \log \pi = \nabla \pi$, such that

$$\begin{aligned}\nabla \cdot (p\gamma) &= p\Big(\langle \nabla \log p, \gamma\rangle + \nabla \cdot \gamma\Big) \\ &= p\Big(\langle \nabla \log p, \gamma\rangle + \frac{\nabla \cdot (\pi\gamma)}{\pi} - \langle \nabla \log \pi, \gamma\rangle\Big) \\ &= p\langle \nabla \log \frac{p}{\pi}, \gamma\rangle + p\frac{\nabla \cdot (\pi\gamma)}{\pi}. \end{aligned} \tag{26}$$

We then have,

$$\begin{aligned}\partial_t \mathrm{I}_a(p\|\pi) &= -2\int \Big[\widetilde{\Gamma}_2(\log \frac{p}{\pi}, \log \frac{p}{\pi}) + \widetilde{\Gamma}_{\mathcal{I}_a}(\log \frac{p}{\pi}, \log \frac{p}{\pi})\Big] p dx - 2\int \tilde{L} \log \frac{p}{\pi} \frac{\nabla \cdot (\pi\gamma)}{\pi} p dx \\ &\quad + \int \langle \nabla \log \frac{p}{\pi}, \partial_t(aa^{\mathsf{T}}) \nabla \log \frac{p}{\pi}\rangle p dx - 2\int \langle \nabla \partial_t \log \pi, aa^{\mathsf{T}} \nabla \log \frac{p}{\pi}\rangle p dx.\end{aligned}$$

Observing the following equality, we have

$$\begin{aligned}& -2\int \langle \nabla \partial_t \log \pi, aa^{\mathsf{T}} \nabla \log \frac{p}{\pi}\rangle p dx \\ =\ & 2\int \partial_t \log \pi \frac{\nabla \cdot (paa^{\mathsf{T}} \nabla \log \frac{p}{\pi})}{p} p dx \\ =\ & 2\int \Big[\nabla \cdot (aa^{\mathsf{T}} \nabla \log \frac{p}{\pi}) + \langle \nabla \log \frac{p}{\pi}, aa^{\mathsf{T}} \nabla \log \frac{p}{\pi}\rangle + \langle \nabla \log \pi, aa^{\mathsf{T}} \nabla \log \frac{p}{\pi}\rangle\Big] \partial_t \log \pi p dx \\ =\ & 2\int \Big[\tilde{L} \log \frac{p}{\pi} + \langle \nabla \log \frac{p}{\pi}, aa^{\mathsf{T}} \nabla \log \frac{p}{\pi}\rangle\Big] \partial_t \log \pi p dx, \end{aligned} \tag{27}$$

which implies

$$\begin{aligned}\partial_t \mathrm{I}_a(p\|\pi) &= -2\int \Big[\widetilde{\Gamma}_2(\log \frac{p}{\pi}, \log \frac{p}{\pi}) + \widetilde{\Gamma}_{\mathcal{I}_a}(\log \frac{p}{\pi}, \log \frac{p}{\pi})\Big] p dx + \int \langle \nabla \log \frac{p}{\pi}, \partial_t(aa^{\mathsf{T}}) \nabla \log \frac{p}{\pi}\rangle p dx \\ &\quad + 2\int \tilde{L} \log \frac{p}{\pi} \frac{\partial_t \pi - \nabla \cdot (\pi\gamma)}{\pi} p dx + 2\int \langle \nabla \log \frac{p}{\pi}, aa^{\mathsf{T}} \nabla \log \frac{p}{\pi}\rangle \partial_t \log \pi p dx.\end{aligned}$$

We also have

$$\begin{aligned}& 2\int \tilde{L} \log \frac{p}{\pi} \mathcal{R} p dx \\ =\ & 2\int \nabla \cdot (aa^{\mathsf{T}} \nabla \log \frac{p}{\pi}) \mathcal{R} p dx + 2\int \langle \nabla \log \pi, aa^{\mathsf{T}} \nabla \log \frac{p}{\pi}\rangle \mathcal{R} p dx \\ =\ & -2\int \langle \nabla(\mathcal{R}p), aa^{\mathsf{T}} \nabla \log \frac{p}{\pi}\rangle dx + 2\int \langle \nabla \log \pi, aa^{\mathsf{T}} \nabla \log \frac{p}{\pi}\rangle \mathcal{R} p dx \\ =\ & -2\int \langle \nabla \mathcal{R}, aa^{\mathsf{T}} \nabla \log \frac{p}{\pi}\rangle p dx - 2\int \langle \nabla p, aa^{\mathsf{T}} \nabla \log \frac{p}{\pi}\rangle \mathcal{R} dx + 2\int \langle \nabla \log \pi, aa^{\mathsf{T}} \nabla \log \frac{p}{\pi}\rangle \mathcal{R} p dx \\ =\ & -2\int \langle \nabla \mathcal{R}, aa^{\mathsf{T}} \nabla \log \frac{p}{\pi}\rangle p dx - 2\int \langle \nabla \log \frac{p}{\pi}, aa^{\mathsf{T}} \nabla \log \frac{p}{\pi}\rangle \mathcal{R} p dx.\end{aligned}$$

Combining the above terms, we have

$$\begin{aligned}\partial_t \mathrm{I}_a(p\|\pi) &= -2\int \big[\widetilde{\Gamma}_2(\log\frac{p}{\pi},\log\frac{p}{\pi})+\widetilde{\Gamma}_{\mathcal{I}_a}(\log\frac{p}{\pi},\log\frac{p}{\pi})\big]pdx\\ &\quad+\int\langle\nabla\log\frac{p}{\pi},[\partial_t(aa^{\mathsf{T}})+2aa^{\mathsf{T}}\partial_t\log\pi]\nabla\log\frac{p}{\pi}\rangle pdx\\ &\quad-2\int\langle\nabla\mathcal{R},aa^{\mathsf{T}}\nabla\log\frac{p}{\pi}\rangle pdx-2\int\langle\nabla\log\frac{p}{\pi},aa^{\mathsf{T}}\nabla\log\frac{p}{\pi}\rangle\mathcal{R}pdx.\end{aligned}$$

Note that,

$$\begin{aligned}-\int\langle\nabla\mathcal{R},aa^{\mathsf{T}}\nabla\log\frac{p}{\pi}\rangle pdx &= -\int\langle\nabla\mathcal{R},aa^{\mathsf{T}}\nabla p\rangle dx+\int\langle\nabla\mathcal{R},aa^{\mathsf{T}}\nabla\log\pi\rangle pdx\\ &= \int\nabla\cdot(aa^{\mathsf{T}}\nabla\mathcal{R})pdx+\int\langle\nabla\mathcal{R},aa^{\mathsf{T}}\nabla\log\pi\rangle pdx.\end{aligned}$$

We conclude with

$$\begin{aligned}\partial_t \mathrm{I}_a(p\|\pi) &= -2\int \big[\widetilde{\Gamma}_2(\log\frac{p}{\pi},\log\frac{p}{\pi})+\widetilde{\Gamma}_{\mathcal{I}_a}(\log\frac{p}{\pi},\log\frac{p}{\pi})\big]pdx\\ &\quad+\int\langle\nabla\log\frac{p}{\pi},[\partial_t(aa^{\mathsf{T}})+2aa^{\mathsf{T}}\partial_t\log\pi-2aa^{\mathsf{T}}\mathcal{R}]\nabla\log\frac{p}{\pi}\rangle pdx\\ &\quad+\int\big[2\nabla\cdot(aa^{\mathsf{T}}\nabla\mathcal{R})+2\langle\nabla\mathcal{R},aa^{\mathsf{T}}\nabla\log\pi\rangle\big]pdx.\end{aligned}$$

And the results follow the fact $\partial_t\log\pi-\mathcal{R}=\frac{\nabla\cdot(\pi\gamma)}{\pi}$. ■

Recall that the irreversible Gamma operator associated with a is defined as

$$\widetilde{\Gamma}_{\mathcal{I}_a}(f,f)=\widetilde{L}f\langle\nabla f,\gamma\rangle-\frac{1}{2}\langle\nabla\Gamma_1(f,f),\gamma\rangle.$$

We next show the following equivalence identity in a weak form for the irreversible Gamma operator.

Lemma 14 *Denote $f=\log\frac{p}{\pi}$, we have*

$$\begin{aligned}\int\widetilde{\Gamma}_{\mathcal{I}_a}(f,f)pdx &= \frac{1}{2}\int\langle\gamma,\langle\nabla f,\nabla(aa^{\mathsf{T}})\nabla f\rangle\rangle pdx-\int\langle aa^{\mathsf{T}}\nabla f,\nabla\gamma\nabla f\rangle pdx\\ &\quad+\int\frac{\nabla\cdot(\pi\gamma)}{\pi}\Gamma_1(f,f)pdx.\end{aligned}$$

Proof We first observe that,

$$\frac{\nabla\cdot(p\gamma)}{p}=\langle\nabla\log p,\gamma\rangle+\nabla\cdot\gamma=\langle\nabla\log\frac{p}{\pi},\gamma\rangle+\langle\nabla\log\pi,\gamma\rangle+\nabla\cdot\gamma=\langle\nabla f,\gamma\rangle+\frac{\nabla\cdot(\pi\gamma)}{\pi}.$$

According to the definition of the information Gamma operator, we have

$$\begin{aligned}
&\int \widetilde{\Gamma}_{\mathcal{I}_a}(f,f)pdx \\
=\ & \int [\widetilde{L}f\langle\nabla f,\gamma\rangle - \frac{1}{2}\langle\nabla\Gamma_1(f,f),\gamma\rangle]pdx \\
=\ & \int \Big[\nabla\cdot(aa^{\mathsf{T}}\nabla f)\langle\nabla f,\gamma\rangle + \langle aa^{\mathsf{T}}\nabla\log\pi,\nabla f\rangle\langle\nabla f,\gamma\rangle\Big]pdx + \frac{1}{2}\int\nabla\cdot(p\gamma)\Gamma_1(f,f)dx \\
=\ & \int \Big[\nabla\cdot(aa^{\mathsf{T}}\nabla f)\langle\nabla f,\gamma\rangle + \langle aa^{\mathsf{T}}\nabla\log\pi,\nabla f\rangle\langle\nabla f,\gamma\rangle\Big]pdx \\
& +\frac{1}{2}\int[\langle\nabla f,\gamma\rangle + \frac{\nabla\cdot(\pi\gamma)}{\pi}]\Gamma_1(f,f)pdx
\end{aligned}$$

$$\begin{aligned}
=\ & \int \Big[-\langle aa^{\mathsf{T}}\nabla f,\nabla\log p\rangle\langle\nabla f,\gamma\rangle + \langle aa^{\mathsf{T}}\nabla\log\pi,\nabla f\rangle\langle\nabla f,\gamma\rangle\Big]pdx + \frac{1}{2}\int\langle\nabla f,\gamma\rangle\Gamma_1(f,f)pdx \\
& -\int[\langle aa^{\mathsf{T}}\nabla f,\nabla^2 f\gamma\rangle - \langle aa^{\mathsf{T}}\nabla f,\nabla\gamma\nabla f\rangle]pdx + \frac{1}{2}\int\frac{\nabla\cdot(\pi\gamma)}{\pi}\Gamma_1(f,f)pdx \\
=\ & -\frac{1}{2}\int\langle\nabla f,\gamma\rangle\Gamma_1(f,f)pdx - \int[\langle aa^{\mathsf{T}}\nabla f,\nabla^2 f\gamma\rangle - \langle aa^{\mathsf{T}}\nabla f,\nabla\gamma\nabla f\rangle]pdx + \frac{1}{2}\int\frac{\nabla\cdot(\pi\gamma)}{\pi}\Gamma_1(f,f)pdx \\
=\ & -\frac{1}{2}\int\langle\nabla p,\gamma\rangle\Gamma_1(f,f)dx + \frac{1}{2}\int\langle\nabla\log\pi,\gamma\rangle\Gamma_1(f,f)pdx \\
& -\int[\langle aa^{\mathsf{T}}\nabla f,\nabla^2 f\gamma\rangle - \langle aa^{\mathsf{T}}\nabla f,\nabla\gamma\nabla f\rangle]pdx + \frac{1}{2}\int\frac{\nabla\cdot(\pi\gamma)}{\pi}\Gamma_1(f,f)pdx \\
=\ & \frac{1}{2}\int\langle\gamma,\langle\nabla f,\nabla(aa^{\mathsf{T}})\nabla f\rangle\rangle pdx - \int\langle aa^{\mathsf{T}}\nabla f,\nabla\gamma\nabla f\rangle pdx + \int\frac{\nabla\cdot(\pi\gamma)}{\pi}\Gamma_1(f,f)pdx.
\end{aligned}$$

The last equality follows from the fact that

$$\begin{aligned}
&-\frac{1}{2}\int\langle\nabla p,\gamma\rangle\Gamma_1(f,f)dx = \frac{1}{2}\int\nabla\cdot(\gamma\Gamma_1(f,f))pdx \\
=\ & \frac{1}{2}\int\nabla\cdot\gamma\Gamma_1(f,f)pdx + \int\langle aa^{\mathsf{T}}\nabla f,\nabla^2 f\gamma\rangle pdx + \frac{1}{2}\int\langle\gamma,\langle\nabla f,\nabla(aa^{\mathsf{T}})\nabla f\rangle\rangle pdx,
\end{aligned}$$

and

$$\frac{\nabla\cdot(\pi\gamma)}{\pi} = \langle\nabla\log\pi,\gamma\rangle + \nabla\cdot\gamma.$$

■

3.4 Dissipation of auxillary Fisher information

Similar to the first-order dissipation of the Fisher information functional, we have the following decay for the auxiliary Fisher information functional.

Proposition 15

$$\begin{aligned}\partial_t \mathrm{I}_z(p\|\pi) &= -2\int \widetilde{\Gamma}_2^{z,\pi}(\log\frac{p}{\pi},\log\frac{p}{\pi})pdx + \int\langle\nabla\log\frac{p}{\pi},\partial_t(zz^{\mathsf{T}})\nabla\log\frac{p}{\pi}\rangle pdx\\ &\quad -\int\langle\gamma,\langle\nabla\log\frac{p}{\pi},\nabla(zz^{\mathsf{T}})\nabla\log\frac{p}{\pi}\rangle\rangle pdx + 2\int\langle zz^{\mathsf{T}}\nabla\log\frac{p}{\pi},\nabla\gamma\nabla\log\frac{p}{\pi}\rangle pdx\\ &\quad +\int\big[2\nabla\cdot(zz^{\mathsf{T}}\nabla\mathcal{R}) + 2\langle\nabla\mathcal{R},zz^{\mathsf{T}}\nabla\log\pi\rangle\big]pdx. \end{aligned}\tag{28}$$

The proof follows from the following Lemma 16 and Lemma 17.

Lemma 16

$$\begin{aligned}\partial_t \mathrm{I}_z(p\|\pi) &= -2\int\big[\widetilde{\Gamma}_2^{z,\pi}(\log\frac{p}{\pi},\log\frac{p}{\pi}) + \Gamma_{\mathcal{I}_z}(\log\frac{p}{\pi},\log\frac{p}{\pi})\big]pdx\\ &\quad +\int\langle\nabla\log\frac{p}{\pi},[\partial_t(zz^{\mathsf{T}}) + 2zz^{\mathsf{T}}\frac{\nabla\cdot(\pi\gamma)}{\pi}]\nabla\log\frac{p}{\pi}\rangle pdx\\ &\quad +\int\big[2\nabla\cdot(zz^{\mathsf{T}}\nabla\mathcal{R}) + 2\langle\nabla\mathcal{R},zz^{\mathsf{T}}\nabla\log\pi\rangle\big]pdx. \end{aligned}\tag{29}$$

Proof

$$\begin{aligned}\partial_t \mathrm{I}_z(p\|\pi) &= 2\int\langle\nabla\partial_t\log\frac{p}{\pi},zz^{\mathsf{T}}\nabla\log\frac{p}{\pi}\rangle pdx + \int\langle\nabla\log\frac{p}{\pi},\partial_t(zz^{\mathsf{T}})\nabla\log\frac{p}{\pi}\rangle pdx\\ &\quad +\int\langle\nabla\log\frac{p}{\pi},zz^{\mathsf{T}}\nabla\log\frac{p}{\pi}\rangle\partial_t pdx\\ &= 2\int\langle\nabla\partial_t\log p,zz^{\mathsf{T}}\nabla\log\frac{p}{\pi}\rangle pdx + \int\langle\nabla\log\frac{p}{\pi},zz^{\mathsf{T}}\nabla\log\frac{p}{\pi}\rangle\partial_t pdx\\ &\quad +\int\langle\nabla\log\frac{p}{\pi},\partial_t(zz^{\mathsf{T}})\nabla\log\frac{p}{\pi}\rangle pdx - 2\int\langle\nabla\partial_t\log\pi,zz^{\mathsf{T}}\nabla\log\frac{p}{\pi}\rangle pdx.\end{aligned}$$

Similar to the derivation for $\mathrm{I}_a(p\|\pi)$, we first observe the following the fact,

$$\begin{aligned}&\int\Gamma_1^z(\log\frac{p}{\pi},\log\frac{p}{\pi})\partial_t p + 2\Gamma_1^z(\frac{\partial_t p}{p},\log\frac{p}{\pi})pdx\\ =\ &\int\Gamma_1^z(\log\frac{p}{\pi},\log\frac{p}{\pi})\partial_t p - 2\frac{\nabla\cdot(pzz^{\mathsf{T}}\nabla\log\frac{p}{\pi})}{p}\partial_t pdx\\ =\ &\int\Gamma_1^z(\log\frac{p}{\pi},\log\frac{p}{\pi})\partial_t p - 2\Big(\langle\nabla\log p,zz^{\mathsf{T}}\nabla\log\frac{p}{\pi}\rangle + \nabla\cdot(zz^{\mathsf{T}}\nabla\log\frac{p}{\pi})\Big)\partial_t pdx\\ =\ &\int\Gamma_1^z(\log\frac{p}{\pi},\log\frac{p}{\pi})\partial_t p - 2\Big(\langle\nabla\log\frac{p}{\pi},zz^{\mathsf{T}}\nabla\log\frac{p}{\pi}\rangle + \widetilde{L}_z\log\frac{p}{\pi}\Big)\partial_t pdx\\ =\ &-2\int\Big\{\frac{1}{2}\Gamma_1^z(\log\frac{p}{\pi},\log\frac{p}{\pi})\partial_t p + \widetilde{L}_z\log\frac{p}{\pi}\partial_t p\Big\}dx\\ =\ &-2\int\Big\{\quad\frac{1}{2}\Gamma_1^z(\log\frac{p}{\pi},\log\frac{p}{\pi})\nabla\cdot(p\gamma) + \widetilde{L}_z\log\frac{p}{\pi}\nabla\cdot(p\gamma)\\ &\qquad +\frac{1}{2}\Gamma_1^z(\log\frac{p}{\pi},\log\frac{p}{\pi})\widetilde{L}^*p + \widetilde{L}_z\log\frac{p}{\pi}\widetilde{L}^*p\Big\}dx\end{aligned}$$

$$
\begin{aligned}
&= -2\int \Big\{\widetilde{\Gamma}_{\mathcal{I}_z}(\log\frac{p}{\pi},\log\frac{p}{\pi}) + \frac{1}{2}\widetilde{L}_z\Gamma_1(\log\frac{p}{\pi},\log\frac{p}{\pi}) - \Gamma_1(\widetilde{L}_z\log\frac{p}{\pi},\log\frac{p}{\pi})\Big\}pdx \\
&\quad -2\int \tilde{L}_z\log\frac{p}{\pi}\frac{\nabla\cdot(\pi\gamma)}{\pi}pdx,
\end{aligned}
$$

where we apply the equality in (26) for $\nabla\cdot(p\gamma)$. Now applying [Proposition 5.11](Feng and Li., 2023) (see also [Proposition 8](Feng and Li., 2021)), we have the following equality

$$
\int \Big\{\frac{1}{2}\widetilde{L}_z\Gamma_1(\log\frac{p}{\pi},\log\frac{p}{\pi}) - \Gamma_1(\widetilde{L}_z\log\frac{p}{\pi},\log\frac{p}{\pi})\Big\}pdx = \int \widetilde{\Gamma}_2^{z,\pi}(\log\frac{p}{\pi},\log\frac{p}{\pi})pdx. \tag{30}
$$

We then have,

$$
\begin{aligned}
\partial_t \mathrm{I}_z(p\|\pi) &= -2\int \Big[\widetilde{\Gamma}_2^{z,\pi}(\log\frac{p}{\pi},\log\frac{p}{\pi}) + \widetilde{\Gamma}_{\mathcal{I}_z}(\log\frac{p}{\pi},\log\frac{p}{\pi})\Big]pdx - 2\int \tilde{L}_z\log\frac{p}{\pi}\frac{\nabla\cdot(\pi\gamma)}{\pi}pdx \\
&\quad + \int \langle\nabla\log\frac{p}{\pi}, \partial_t(zz^{\mathsf{T}})\nabla\log\frac{p}{\pi}\rangle pdx - 2\int\langle\nabla\partial_t\log\pi, zz^{\mathsf{T}}\nabla\log\frac{p}{\pi}\rangle pdx.
\end{aligned}
$$

Observing the following equality, we have

$$
\begin{aligned}
&-2\int\langle\nabla\partial_t\log\pi, zz^{\mathsf{T}}\nabla\log\frac{p}{\pi}\rangle pdx \\
&= 2\int \partial_t\log\pi\frac{\nabla\cdot(pzz^{\mathsf{T}}\nabla\log\frac{p}{\pi})}{p}pdx \\
&= 2\int\Big[\nabla\cdot(zz^{\mathsf{T}}\nabla\log\frac{p}{\pi}) + \langle\nabla\log\frac{p}{\pi}, zz^{\mathsf{T}}\nabla\log\frac{p}{\pi}\rangle + \langle\nabla\log\pi, zz^{\mathsf{T}}\nabla\log\frac{p}{\pi}\rangle\Big]\partial_t\log\pi pdx \\
&= 2\int\Big[\tilde{L}_z\log\frac{p}{\pi} + \langle\nabla\log\frac{p}{\pi}, zz^{\mathsf{T}}\nabla\log\frac{p}{\pi}\rangle\Big]\partial_t\log\pi pdx,
\end{aligned} \tag{31}
$$

which implies

$$
\begin{aligned}
\partial_t \mathrm{I}_z(p\|\pi) &= -2\int \Big[\widetilde{\Gamma}_2^{z,\pi}(\log\frac{p}{\pi},\log\frac{p}{\pi}) + \widetilde{\Gamma}_{\mathcal{I}_z}(\log\frac{p}{\pi},\log\frac{p}{\pi})\Big]pdx + \int\langle\nabla\log\frac{p}{\pi}, \partial_t(zz^{\mathsf{T}})\nabla\log\frac{p}{\pi}\rangle pdx \\
&\quad +2\int \tilde{L}_z\log\frac{p}{\pi}\frac{\partial_t\pi - \nabla\cdot(\pi\gamma)}{\pi}pdx + 2\int\langle\nabla\log\frac{p}{\pi}, zz^{\mathsf{T}}\nabla\log\frac{p}{\pi}\rangle\partial_t\log\pi pdx.
\end{aligned}
$$

We also have

$$
\begin{aligned}
&2\int \tilde{L}_z\log\frac{p}{\pi}\mathcal{R}pdx \\
&= 2\int\nabla\cdot(zz^{\mathsf{T}}\nabla\log\frac{p}{\pi})\mathcal{R}pdx + 2\int\langle\nabla\log\pi, zz^{\mathsf{T}}\nabla\log\frac{p}{\pi}\rangle\mathcal{R}pdx \\
&= -2\int\langle\nabla\mathcal{R}, zz^{\mathsf{T}}\nabla\log\frac{p}{\pi}\rangle pdx - 2\int\langle\nabla\log\frac{p}{\pi}, zz^{\mathsf{T}}\nabla\log\frac{p}{\pi}\rangle\mathcal{R}pdx.
\end{aligned}
$$

Combining the above terms, we have

$$
\begin{aligned}
\partial_t \mathrm{I}_z(p\|\pi) &= -2\int \big[\widetilde{\Gamma}_2^{z,\pi}(\log\frac{p}{\pi},\log\frac{p}{\pi}) + \widetilde{\Gamma}_{\mathcal{I}_z}(\log\frac{p}{\pi},\log\frac{p}{\pi})\big]pdx \\
&\quad + \int\langle\nabla\log\frac{p}{\pi}, [\partial_t(zz^{\mathsf{T}}) + 2zz^{\mathsf{T}}\partial_t\log\pi]\nabla\log\frac{p}{\pi}\rangle pdx \\
&\quad -2\int\langle\nabla\mathcal{R}, zz^{\mathsf{T}}\nabla\log\frac{p}{\pi}\rangle pdx - 2\int\langle\nabla\log\frac{p}{\pi}, zz^{\mathsf{T}}\nabla\log\frac{p}{\pi}\rangle\mathcal{R}pdx.
\end{aligned}
$$

Note that,

$$\begin{aligned}-2\int\langle\nabla\mathcal{R},zz^{\mathsf{T}}\nabla\log\frac{p}{\pi}\rangle pdx &= -2\int\langle\nabla\mathcal{R},zz^{\mathsf{T}}\nabla p\rangle dx+2\int\langle\nabla\mathcal{R},zz^{\mathsf{T}}\nabla\log\pi\rangle pdx\\ &= 2\int\nabla\cdot(zz^{\mathsf{T}}\nabla\mathcal{R})pdx+2\int\langle\nabla\mathcal{R},zz^{\mathsf{T}}\nabla\log\pi\rangle pdx.\end{aligned}$$

We conclude with

$$\begin{aligned}\partial_t\mathrm{I}_z(p\|\pi) &= -2\int\big[\widetilde{\Gamma}_2^{z,\pi}(\log\frac{p}{\pi},\log\frac{p}{\pi})+\widetilde{\Gamma}_{\mathcal{I}_z}(\log\frac{p}{\pi},\log\frac{p}{\pi})\big]pdx\\ &+\int\langle\nabla\log\frac{p}{\pi},[\partial_t(zz^{\mathsf{T}})+2zz^{\mathsf{T}}\partial_t\log\pi-2zz^{\mathsf{T}}\mathcal{R}]\nabla\log\frac{p}{\pi}\rangle pdx\\ &+\int\big[2\nabla\cdot(zz^{\mathsf{T}}\nabla\mathcal{R})+2\langle\nabla\mathcal{R},zz^{\mathsf{T}}\nabla\log\pi\rangle\big]pdx,\end{aligned}$$

and the result follows the fact $\partial_t\log\pi-\mathcal{R}=\frac{\nabla\cdot(\pi\gamma)}{\pi}$. ■

The irreversible Gamma operator associated with matrix z has the following equivalent form.

Lemma 17 *Denote $f=\log\frac{p}{\pi}$. We have*

$$\begin{aligned}\int\widetilde{\Gamma}_{\mathcal{I}_z}(f,f)pdx &= \frac{1}{2}\int\langle\gamma,\langle\nabla f,\nabla(zz^{\mathsf{T}})\nabla f\rangle\rangle pdx-\int\langle zz^{\mathsf{T}}\nabla f,\nabla\gamma\nabla f\rangle pdx\\ &+\int\frac{\nabla\cdot(\pi\gamma)}{\pi}\Gamma_1^z(f,f)pdx.\end{aligned}$$

Proof We will use the following fact again

$$\frac{\nabla\cdot(p\gamma)}{p}=\langle\nabla\log p,\gamma\rangle+\nabla\cdot\gamma=\langle\nabla\log\frac{p}{\pi},\gamma\rangle+\langle\nabla\log\pi,\gamma\rangle+\nabla\cdot\gamma=\langle\nabla f,\gamma\rangle+\frac{\nabla\cdot(\pi\gamma)}{\pi}.$$

We have

$$
\begin{aligned}
&\int \widetilde{\Gamma}_{\mathcal{I}_z}(f,f)pdx \\
=\ & \int [\widetilde{L}_z f\langle \nabla f,\gamma\rangle - \frac{1}{2}\langle \nabla\Gamma_1^z(f,f),\gamma\rangle]pdx \\
=\ & \int \Big[\nabla\cdot(zz^{\mathsf{T}}\nabla f)\langle\nabla f,\gamma\rangle + \langle zz^{\mathsf{T}}\nabla\log\pi,\nabla f\rangle\langle\nabla f,\gamma\rangle\Big]pdx + \frac{1}{2}\int\nabla\cdot(p\gamma)\Gamma_1^z(f,f)dx \\
=\ & \int \Big[\nabla\cdot(zz^{\mathsf{T}}\nabla f)\langle\nabla f,\gamma\rangle + \langle zz^{\mathsf{T}}\nabla\log\pi,\nabla f\rangle\langle\nabla f,\gamma\rangle\Big]pdx \\
& +\frac{1}{2}\int[\langle\nabla f,\gamma\rangle + \frac{\nabla\cdot(\pi\gamma)}{\pi}]\Gamma_1^z(f,f)pdx \\
=\ & \int \Big[-\langle zz^{\mathsf{T}}\nabla f,\nabla\log p\rangle\langle\nabla f,\gamma\rangle + \langle zz^{\mathsf{T}}\nabla\log\pi,\nabla f\rangle\langle\nabla f,\gamma\rangle\Big]pdx + \frac{1}{2}\int\langle\nabla f,\gamma\rangle\Gamma_1^z(f,f)pdx \\
& -\int[\langle zz^{\mathsf{T}}\nabla f,\nabla^2 f\gamma\rangle - \langle zz^{\mathsf{T}}\nabla f,\nabla\gamma\nabla f\rangle]pdx + \frac{1}{2}\int\frac{\nabla\cdot(\pi\gamma)}{\pi}\Gamma_1^z(f,f)pdx \\
=\ & -\frac{1}{2}\int\langle\nabla f,\gamma\rangle\Gamma_1^z(f,f)pdx - \int[\langle zz^{\mathsf{T}}\nabla f,\nabla^2 f\gamma\rangle - \langle zz^{\mathsf{T}}\nabla f,\nabla\gamma\nabla f\rangle]pdx + \frac{1}{2}\int\frac{\nabla\cdot(\pi\gamma)}{\pi}\Gamma_1^z(f,f)pdx \\
=\ & -\frac{1}{2}\int\langle\nabla p,\gamma\rangle\Gamma_1^z(f,f)dx + \frac{1}{2}\int\langle\nabla\log\pi,\gamma\rangle\Gamma_1^z(f,f)pdx \\
& -\int[\langle zz^{\mathsf{T}}\nabla f,\nabla^2 f\gamma\rangle - \langle zz^{\mathsf{T}}\nabla f,\nabla\gamma\nabla f\rangle]pdx + \frac{1}{2}\int\frac{\nabla\cdot(\pi\gamma)}{\pi}\Gamma_1^z(f,f)pdx \\
=\ & \frac{1}{2}\int\langle\gamma,\langle\nabla f,\nabla(zz^{\mathsf{T}})\nabla f\rangle\rangle pdx - \int\langle zz^{\mathsf{T}}\nabla f,\nabla\gamma\nabla f\rangle pdx + \int\frac{\nabla\cdot(\pi\gamma)}{\pi}\Gamma_1^z(f,f)pdx.
\end{aligned}
$$

The last equality follows from the fact that

$$
\begin{aligned}
& -\frac{1}{2}\int\langle\nabla p,\gamma\rangle\Gamma_1^z(f,f)dx = \frac{1}{2}\int\nabla\cdot(\gamma\Gamma_1^z(f,f))pdx \\
=\ & \frac{1}{2}\int\nabla\cdot\gamma\Gamma_1^z(f,f)pdx + \int\langle zz^{\mathsf{T}}\nabla f,\nabla^2 f\gamma\rangle pdx + \frac{1}{2}\int\langle\gamma,\langle\nabla f,\nabla(zz^{\mathsf{T}})\nabla f\rangle\rangle pdx,
\end{aligned}
$$

and $\frac{\nabla\cdot(\pi\gamma)}{\pi} = \langle\nabla\log\pi,\gamma\rangle + \nabla\cdot\gamma$.

■

4. Example I: reversible SDE

This example considers an inhomogeneous stochastic differential equation (SDE).

$$
\begin{aligned}
dX_t =& \Big(-\alpha(t,X_t)\alpha(t,X_t)^{\mathsf{T}}\nabla V(X_t) + \beta(t)\nabla\cdot\Big(\alpha(t,X_t)\alpha(t,X_t)^{\mathsf{T}}\Big)\Big)dt \\
&+\sqrt{2\beta(t)}\alpha(t,X_t)dB_t,
\end{aligned}
\tag{32}
$$

where $n=d$, $m=0$, $X_t\in\mathbb{R}^d$, B_t is a standard d-dimensional Brownian motion, $\beta(t)\in\mathbb{R}_+^1$ is a positive, twice continuously differentiable, decreasing function, $V\in\mathbb{C}^2(\mathbb{R}^d;\mathbb{R})$ and

$\alpha(t,x) \in \mathbb{R}^{d\times d}$ is a positive definite matrix function with at least twice differentiable in x and differentiable in t. We denote $a(t,x) = \sqrt{\beta(t)}\alpha(t,x)$. And we assume that a satisfies the uniform non-degenerate condition (see, e.g., (Kusuoka and Stroock, 1984)). Hence there exists a smooth density function for the solution X_t, denoted as $p(t,x)$. Furthermore, we denote $\pi(t,x) \in \mathbb{R}_+$ as a time-dependent probability density function with

$$\pi(t,x) := \frac{1}{Z(t)} e^{-\frac{V(x)}{\beta(t)}}, \tag{33}$$

where we assume that the normalization constant is finite, i.e., $Z(t) = \int_{\mathbb{R}^d} e^{-\frac{V(y)}{\beta(t)}} dy < \infty$. We note that $\pi(t,x)$ is not the stationary distribution of the SDE (32).

Remark 18 *In non-convex optimization,* (32) *is of great importance. Generally speaking, finding the global minimum of a non-convex convex function is much more difficult than if the function is convex/strongly convex. One popular method for non-convex optimization is simulated annealing (SA) Pincus (1970); Khachaturyan et al. (1979, 1981); van Laarhoven and Aarts (1987). The SA process generally consists of a proposal step, an accept/reject step and a cooling scheme. At each iteration, a new move is proposed. This new move will be accepted with some probability that depends the temperature. The lower the temperature, the lower the acceptance rate. At the end of this iteration, temperature is cooled further according to the cooling scheme. The hope is that with appropriate cooling speed, the system will be able to explore enough landscape and escape local minima before settling down near the global minimum. Geman and Hwang (1986) proposed to view* (32) *as the continuous-time version of simulated annealing. Indeed, if $\beta(t) \to 0$ as $t \to \infty$, then $\pi(t,x)$* (33) *converges in distribution to the delta measure supported on the global minimum of V. It was shown by Geman and Hwang (1986); Chiang et al. (1987) that the correct cooling scheme $\beta(t)$ for the process* (32) *to converge to the global minimum is $\log(t)^{-1}$.*

4.1 Convergence analysis

As a special case of Proposition 11, with $\gamma \equiv 0$ and $z(t,x) \equiv 0$, we have the following lemma.

Lemma 19 *For any $t \geq 0$, consider $p(t,x)$ as the probability density function of SDE* (32)*. Denote*

$$\mathrm{I}_a(p\|\pi) = \int (\nabla \log \frac{p}{\pi}, aa^{\mathsf{T}} \nabla \log \frac{p}{\pi}) p dx.$$

We have

$$\begin{aligned} \partial_t \mathrm{I}_a(p\|\pi) &= -2\int \widetilde{\Gamma}_2(\log\frac{p}{\pi}, \log\frac{p}{\pi}) p dx + \int \langle \nabla \log \frac{p}{\pi}, \partial_t(aa^{\mathsf{T}}) \nabla \log \frac{p}{\pi} \rangle p dx \\ &\quad +2\int \big[\nabla \cdot (aa^{\mathsf{T}} \nabla \partial_t \log \pi) + \langle \nabla \partial_t \log \pi, aa^{\mathsf{T}} \nabla \log \pi \rangle\big] p dx. \end{aligned} \tag{34}$$

As a special case of Proposition 10, following Lemma 19, we have

$$\widetilde{\Gamma}_2(\log\frac{p}{\pi}, \log\frac{p}{\pi}) \geq \mathfrak{R}(\nabla \log \frac{p}{\pi}, \nabla \log \frac{p}{\pi}),$$

where $\mathfrak{R}$ (as defined in Appendix A) denotes the Ricci curvature tensor in this example with $\gamma(t,x) \equiv 0$, $z(t,x) \equiv 0$, and $a(t,x) = \sqrt{\beta(t)}\alpha(t,x)$. We then have the following Fisher information functional decay for $\mathrm{I}_a(t) := \mathrm{I}_a(p(t,\cdot)\|\pi(t,\cdot))$.

Theorem 20 *Consider $p(t,x)$ as the probability density function of SDE (32). Suppose $\mathfrak{R}$ is defined in Appendix A. Assume that there exists a positive function $\lambda(t) > 0$, such that*

$$\mathfrak{R} - \frac{1}{2}\partial_t(aa^{\mathsf{T}}) \succeq \lambda(t)aa^{\mathsf{T}}, \tag{35}$$

for $t \geq t_0$ with some constant $t_0 > 0$. Then we have

$$\mathrm{I}_a(t) \leq e^{-2\int_{t_0}^t \lambda(r)dr}\Big(\int_{t_0}^t 2\mathrm{A}(r)e^{2\int_{t_0}^r \lambda(\tau)d\tau}dr + \mathrm{I}_a(t_0)\Big),$$

where $A(t)$ is a function depending on the time variable, such that

$$\mathrm{A}(t) := \int \big[\nabla\cdot(aa^{\mathsf{T}}\nabla\partial_t\log\pi) + \langle\nabla\partial_t\log\pi, aa^{\mathsf{T}}\nabla\log\pi\rangle\big]pdx. \tag{36}$$

Proof The proof follows from Theorem 6 with $\mathcal{R} = \partial_t\log\pi$, and our choice of parameters for SDE (32). ■

4.2 Time-dependent overdamped Langevin dynamics

In this section, we present an explicit example of the convergence result in Theorem 20. Consider the overdamped Langevin dynamics

$$dX_t = -\nabla V(X_t) + \sqrt{2\beta(t)}dB_t. \tag{37}$$

And the diffusion matrix $a(t,x) \in \mathbb{R}^{d\times d}$ has the following form,

$$a(t,x) = \sqrt{\beta(t)}\mathbb{I}, \tag{38}$$

where $\mathbb{I} \in \mathbb{R}^{d\times d}$ is an identity matrix.

Corollary 21 *Let $\beta(t) = \frac{C}{\log t}$ for some constant $C > 0$, and $t \geq t_0 > e$ for some constant $t_0 > 0$. Assume $\nabla^2_{xx}V \succeq \lambda_0\mathbb{I}$, for some constant $\lambda_0 > 0$, and $\int_{\mathbb{R}^d}(\|\nabla_x V\|^2 + |\Delta_x V|)p(t,x)dx \leq \bar{C}$, for some constant $\bar{C} > 0$. Denote*

$$\mathrm{I}_a(p(t,\cdot)\|\pi(t,\cdot)) := \beta(t)\int\|\nabla\log\frac{p(t,x)}{\pi(t,x)}\|^2 p(t,x)dx.$$

Then there exists a constant $C_0 > 0$, such that

$$\mathrm{I}_a(p(t,\cdot)\|\pi(t,\cdot)) \leq \frac{C_0}{t}.$$

Proof The matrix function $\mathfrak{R}$ defined in Appendix (A) is simply $\mathfrak{R} = \beta(t)\nabla^2_{xx}V(x)$ for equation (37). Applying Theorem 20, the Assumption (35) in Theorem 20 is then reduced to the following condition,

$$\beta(t)\nabla^2_{xx}V(x) - \frac{1}{2}\partial_t\beta(t)\mathbb{I} \succeq \lambda(t)\beta(t)\mathbb{I}. \tag{39}$$

For $\beta(t) = \frac{C}{\log t}$, the above condition is equivalent to

$$\frac{C}{\log t}\nabla^2_{xx}V + \frac{1}{2}\frac{C/t}{(\log t)^2}\mathbb{I} \succeq \lambda(t)\frac{C}{\log t}\mathbb{I}.$$

Based on assumption $\nabla^2_{xx}V \succeq \lambda_0\mathbb{I}$ with $\lambda_0 > 0$, for $t \geq t_0$, and we let $\lambda(t) \equiv \lambda_0$, then

$$\nabla^2_{xx}V + \frac{1}{2t\log t}\mathbb{I} \succeq (\lambda_0 + \frac{1}{2t\log t})\mathbb{I} \succeq \lambda_0\mathbb{I} = \lambda(t)\mathbb{I}. \tag{40}$$

Now we turn to the estimate for $\mathrm{A}(t)$ in Theorem 20. Plugging in $\beta(t) = \frac{C}{\log t}$, we obtain

$$\begin{aligned}\nabla\cdot(aa^{\mathsf{T}}\nabla\partial_t\log\pi) + \langle\nabla\partial_t\log\pi, aa^{\mathsf{T}}\nabla\log\pi\rangle &= \beta\Delta(\partial_t\log\pi) + \beta\langle\nabla\partial_t\log\pi, \nabla\log e^{-V/\beta}\rangle \\ &= \frac{\partial_t\beta}{\beta}\Delta_x V - \frac{\partial_t\beta}{\beta^2}\|\nabla_x V\|^2.\end{aligned}$$

Applying the assumption that $\int_{\mathbb{R}^d}(\|\nabla_x V\|^2 + |\Delta_x V|)p(t,x)dx \leq \overline{C}$, we get

$$\begin{aligned}\mathrm{A}(t) = 2\int[\frac{\partial_t\beta}{\beta}\Delta_x V - \frac{\partial_t\beta}{\beta^2}\|\nabla_x V\|^2]pdx &\leq 2\overline{C}(|\frac{\partial_t\beta}{\beta}| + |\frac{\partial_t\beta}{\beta^2}|) \\ &\leq 2\overline{C}(|\frac{1}{t\log t}| + |\frac{1}{Ct}|) \leq \frac{C_A}{t},\end{aligned}$$

where we denote C_A as the upper bound of $\mathrm{A}(t)$ for $t > t_0 > e$. Following the proof of Theorem 20, we have

$$\frac{d}{dt}\mathrm{I}_a(t) \leq -2\lambda_0 I_a(t) + \frac{C_A}{t}.$$

Hence

$$\begin{aligned}\mathrm{I}_a(t) \leq& e^{-2\lambda_0(t-t_0)}\Big(\int_{t_0}^t 2\frac{C_A}{r}e^{2\lambda_0(r-t_0)}dr + \mathrm{I}_a(t_0)\Big) \\ =& e^{-2\lambda_0(t-t_0)}\mathrm{I}_a(t_0) + 2C_A e^{-2\lambda_0 t}\int_{t_0}^t\frac{e^{2\lambda_0 r}}{r}dr.\end{aligned}$$

We notice that

$$\lim_{t\to+\infty}\frac{e^{-2\lambda_0 t}\int_{t_0}^t\frac{e^{2\lambda_0 r}}{r}dr}{\frac{1}{t}} = \lim_{t\to+\infty}\frac{t\int_{t_0}^t\frac{e^{2\lambda_0 r}}{r}dr}{e^{2\lambda_0 t}} = \lim_{t\to+\infty}\frac{\int_{t_0}^t\frac{e^{2\lambda_0 r}}{r}dr + e^{2\lambda_0 t}}{2\lambda_0 e^{2\lambda_0 t}} = \frac{1}{2\lambda_0}.$$

For a sufficient small $\epsilon > 0$, there exists a constant $T > 0$, such that when $t > T$,

$$e^{-2\lambda_0 t}\int_{t_0}^t\frac{e^{2\lambda_0 r}}{r}dr \leq (\frac{1}{2\lambda_0} + \epsilon)\frac{1}{t}.$$

Denote $M = \sup_{t\in[0,T]} e^{-2\lambda_0 t}\int_{t_0}^{t} \frac{e^{2\lambda_0 r}}{r}dr$. Thus, when $t_0 \le t \le T$, we have

$$e^{-2\lambda_0 t}\int_{t_0}^{t}\frac{e^{2\lambda_0 r}}{r}dr \le M = \frac{M}{T}T \le \frac{MT}{t}.$$

Thus, there exists a constant $C_0 > 0$, such that

$$\mathrm{I}_a(t) \le e^{-2\lambda_0(t-t_0)}\mathrm{I}_a(t_0) + 2\frac{C_A}{t}\max\{\frac{1}{2\lambda_0}+\epsilon, MT\} \le \frac{C_0}{t}.$$

This finishes the proof.

■

Following the Fisher information decay in Corollary 21, we get the decay of the KL divergence of the density for the dynamics (37) as below.

Corollary 22 *Under the assumptions in Theorem 20, for any $\tau \ge t_0$, we have*

$$\mathrm{D}_{\mathrm{KL}}(p(\tau)\|\pi(\tau)) \le \frac{1}{2\lambda_0}\mathrm{I}_a(p(\tau)\|\pi(\tau)) \le \frac{C_0}{2\lambda_0\tau},$$

and

$$\int_{\mathbb{R}^d}|p(\tau,x)-\pi(\tau,x)|dx \le \sqrt{\frac{C_0}{\lambda_0\tau}}.$$

Proof For any fixed $\tau \ge t_0$, we consider the standard overdamped Langevin dynamics:

$$dX_t^\tau = -\nabla V(X_t^\tau)dt + \sqrt{2\beta(\tau)}dB_t,$$

which is equipped with the invariant measure $\pi(x;\tau) = \frac{1}{Z}e^{-\frac{V(x)}{\beta(\tau)}}$. Denote $p(t;\tau)$ as the density for X_t^τ. Since V is strongly convex, i.e. $\nabla^2_{xx}V \succeq \lambda_0\mathbb{I}$, we have the classical log-Sobolev inequality, such that

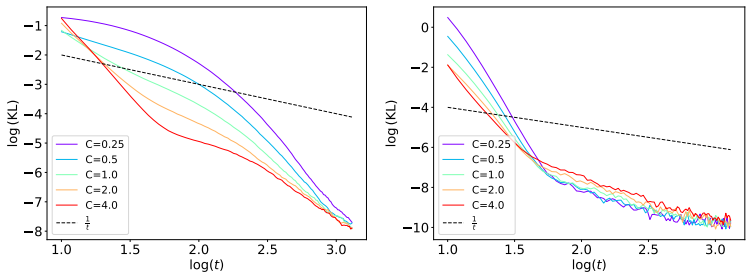
$$\mathrm{D}_{\mathrm{KL}}(p(\tau)\|\pi(\tau)) \le \frac{1}{2\lambda_0}\mathrm{I}_a(p(\tau)\|\pi(\tau)) \le \frac{C_0}{2\lambda_0\tau},$$

where the last inequality follows from Corollary 21. From Pinsker's inequality, we have

$$\int_{\mathbb{R}^d}|p(\tau,x)-\pi(\tau,x)|dx \le \sqrt{2\mathrm{D}_{\mathrm{KL}}(p(\tau)\|\pi(\tau))} \le \sqrt{\frac{C_0}{\lambda_0\tau}}.$$

■

Remark 23 *When we use the upper bound of $\lambda(t)$ in (40) as $\frac{1}{2t\log t}$, $A(t)$ can be infinity. Thus, the current convergence analysis does not work. In other words, the choice of $\beta = \frac{C}{\log t}$ is essential for the current convergence proof, as discussed in Tang and Zhou (2021).*



(a) $V(x) = \frac{(x-1)^2}{8}$ (b) $V(x) = \frac{(x+1)^2}{2} - \frac{\cos(x)}{2}$

Figure 1: Convergence rate of two strongly convex functions in one-dimension. Here y-axis represents the logarithm of the KL divergence between the empirical distribution and the invariant measure $\pi(t,x)$ given by (33). And the x-axis is $\log(t)$. We have also added a dotted line representing t^{-1} on a logarithmic scale for comparison.

4.3 Numerics

In this section, we perform numerical experiments to demonstrate the convergence rate in Corollary 22.

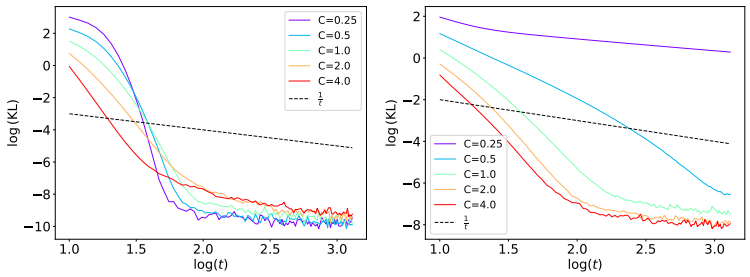
We consider $V : \mathbb{R}^d \to \mathbb{R}$, $d = 1, 2$, and $\beta(t) = \frac{C}{\log(t)}$ for some choices of constant C. We would like to compare the KL divergence between the invariant measure $\pi(t,x)$ given by (33) and the sample distribution of X_t that follows (37) for different choices of $V(x)$ and $\beta(t)$. We first sample $M = 10^6$ particles from $\mathcal{N}(0,1)$. Then we evolve (37) using the Euler-Maruyama scheme shown below for $N = 10000$ steps with a step size of $h = 0.002$:

$$X_{n+1} = X_n - h\nabla V(X_n) + \sqrt{\frac{2C}{\log(nh+t_0)}} B_n \,, \tag{41}$$

where $B_n \sim \mathcal{N}(0, \sqrt{h})$, and $t_0 = e$. During each iteration, we compute the discrete KL divergence between the empirical distribution of the M particles and the invariant measure $\pi(t,x)$ given by (33). The KL divergence between two discrete distributions is given by $\mathrm{D}_{\mathrm{KL}}(p\|q) = \sum p_i \log(p_i/q_i)$. At each iteration, we can use the histogram of the empirical distribution to get p_i for $i = 1, ..., K$. Here K is the number of bins of the histogram and we choose $K = 50$ in our numerical experiment. Let x_i denote the location (midpoint between the left and right bin edge) of each of the bins. Then at the n-th iteration, we can compute $q_i = \frac{1}{Z}\exp(-\frac{V(x_i)}{\beta(nh+t_0)})$, where $Z = \int_{\mathbb{R}} \exp(-\frac{V(x)}{\beta(nh+t_0)})dx$ is the normalization constant and can be estimated numerically. The results are plotted (on a logarithmic scale) in Fig. 1 for strongly convex $V(x)$ and Fig. 2 for non-convex function $V(x)$ with different constant C in the expression of $\beta(t)$.

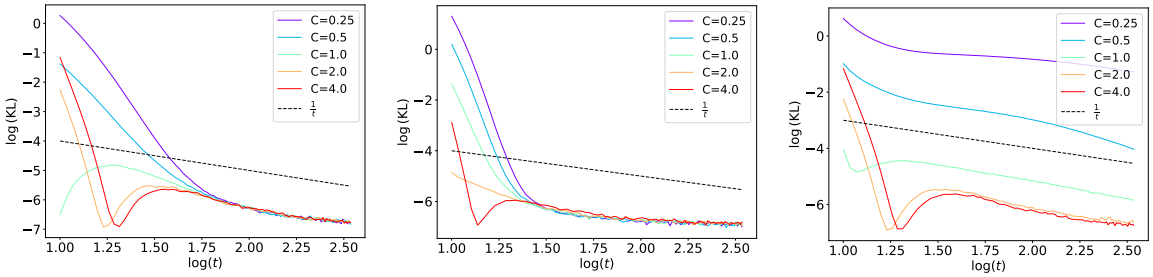
In the strongly-convex setting (Fig. 1), we see that the KL divergence between empirical distribution and π decreases at a rate faster than $O(1/t)$ for all choices of C. In the non-convex setting (Fig. 2), we observe a convergence rate faster than $O(1/t)$ at the beginning which then drops to $O(1/t)$ as t becomes larger. In two-dimension (Fig. 3), we observe the

$O(1/t)$ convergence in both the strongly convex examples (3a and 3b) and the non-convex example (3c). In our two-dimensional examples, we used $M = 10^6$ particles, $N = 10000$ steps with a stepsize of $h = 0.001$. We have 50 bins in both x and y direction which gives a total of 2500 bins.



(a) $V(x) = \frac{(x+2)^4}{4} - \frac{x^2}{2} + \frac{x}{8}$ (b) $V(x) = \frac{(x-2)^2}{2} - \frac{\sin(5x)}{2}$

Figure 2: Convergence rate of two non-convex functions in one-dimension.



(a) $\frac{x^2+y^2}{2}$ (b) $x^2 + y^2 - \frac{\cos(x+y)}{2}$ (c) $x^2 + y^2 - \sin(2x + 2y)$

Figure 3: Convergence rate of two strongly convex functions (A) and (B), and a non-convex function (C) in two dimensions.

5. Example II: non-degenerate, non-reversible SDEs

In this section, we apply Theorem 6 to study the following non-degenerate and non-reversible SDE,

$$\begin{aligned} dX_t =& \Big(-\alpha(t, X_t)\alpha(t, X_t)^{\mathsf{T}} \nabla V(X_t) + \beta(t) \nabla \cdot \Big(\alpha(t, X_t)\alpha(t, X_t)^{\mathsf{T}} \Big) - \gamma(t, X_t) \Big) dt \\ &+ \sqrt{2\beta(t)}\alpha(t, X_t) dB_t. \end{aligned} \tag{42}$$

Again, we have $d = n$, $m = 0$. The above SDE is a variant of SDE (32) by adding a smooth irreversible vector field $\gamma(t, x) \in \mathbb{R}^d$, which is assumed to satisfy

$$\nabla \cdot (e^{-V(x)} \gamma(t, x)) = 0.$$

Remark 24 *Besides overdamped and underdamped Langevin dynamics, there is one other class of SDEs known as non-reversible Langevin dynamics that can also be used for sampling (we refer our readers to the beginning of 6 for a short literature review on sampling a posterior distribution using overdamped and underdamped Langevin dynamics). It has been noted in several papers (Duncan et al., 2017, 2016; Hwang et al., 2015; Lelievre et al., 2013; Rey-Bellet and Spiliopoulos, 2015; Wu et al., 2014) that adding an appropriate non-reversible component to the SDE* (32) *could be beneficial. In particular, the non-reversible component could help speed up the convergence to the target distribution and reduce the asymptotic variance.*

In the current setting, we focus on a special case with $a(t,x) = \sqrt{\beta(t)}\alpha(t,x)$. In particular, we consider the diffusion matrix a in a special form, which satisfies $a_{ii}(t,x) = a_{ii}(t,x_i) = \sqrt{\beta(t)}\alpha_{ii}(x_i) > 0$, for all $x_i \in \mathbb{R}$, $i = 1, \cdots, n$, with

$$\alpha(x) = \begin{pmatrix} \alpha_{11}(x_1) & 0 & \cdots & 0 \\ 0 & \alpha_{22}(x_2) & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \alpha_{nn}(x_n) \end{pmatrix}. \tag{43}$$

Proposition 25 *The Hessian matrix $\mathfrak{R}$ for the above time-dependent non-reversible SDE* (42) *has the following form,*

$$\mathfrak{R} - \frac{1}{2}\partial_t(aa^\mathsf{T}) = \mathfrak{R}_a + \mathfrak{R}_{\gamma_a} - \frac{1}{2}\partial_t(aa^\mathsf{T}),$$

where

$$\begin{cases} \mathfrak{R}_{a,ii} &= \beta(t)\alpha_{ii}^3\partial_{x_i}\alpha_{ii}\partial_{x_i}V(x) + \beta(t)\alpha_{ii}^4\partial^2_{x_ix_i}V(x) - \beta^2(t)\alpha_{ii}^3\partial^2_{x_ix_i}\alpha_{ii}, \quad i = 1, \cdots, n; \\ \mathfrak{R}_{a,ij} &= \beta(t)\alpha_{ii}^2\alpha_{jj}^2\partial^2_{x_ix_j}V(x), \qquad\qquad i,j = 1, \cdots, n, i \neq j; \\ \mathfrak{R}_{\gamma_a,ii} &= \beta(t)\gamma_i\alpha_{ii}\partial_{x_i}\alpha_{ii} - \beta(t)\partial_{x_i}\gamma_i(\alpha_{ii})^2, \qquad i = 1, \cdots, n; \\ \mathfrak{R}_{\gamma_a,ij} &= -\frac{1}{2}\beta(t)[\partial_{x_i}\gamma_j(\alpha_{jj})^2 + \partial_{x_j}\gamma_i(\alpha_{ii})^2], \qquad i,j = 1, \cdots, n, i \neq j. \end{cases} \tag{44}$$

Proof Following Feng and Li. (2021, Proposition 2), we have

$$\begin{cases} \mathfrak{R}_{a,ii} &= -a_{ii}^3\partial_{x_i}a_{ii}\partial_{x_i}\log\pi - a_{ii}^4\partial^2_{x_ix_i}\log\pi - a_{ii}^3\partial^2_{x_ix_i}a_{ii}, i = 1, \cdots, n; \\ \mathfrak{R}_{a,ij} &= -a_{ii}^2a_{jj}^2\partial^2_{x_ix_j}\log\pi, \qquad\qquad i,j = 1, \cdots, n, i \neq j; \\ \mathfrak{R}_{\gamma_a,ii} &= \gamma_i a_{ii}\partial_{x_i}a_{ii} - \partial_{x_i}\gamma_i(a_{ii})^2, \qquad i = 1, \cdots, n; \\ \mathfrak{R}_{\gamma_a,ij} &= -\frac{1}{2}[\partial_{x_i}\gamma_j(a_{jj})^2 + \partial_{x_j}\gamma_i(a_{ii})^2], \qquad i,j = 1, \cdots, n, i \neq j. \end{cases}$$

Plugging in the matrix $a(t,x) = \sqrt{\beta(t)}\alpha(x)$, we derive the desired matrix $\mathfrak{R}$. ■

As in the previous section, if there exists a constant $\lambda > 0$, such that

$$\mathfrak{R} - \frac{1}{2}\partial_t(aa^\mathsf{T}) \succeq \lambda aa^\mathsf{T},$$

then the Fisher information decay in Theorem 6 holds.

5.1 Time-dependent non-reversible Langevin dynamics

In this section, we consider a special case with $n = 2$, $\alpha \equiv \mathbb{I}$, and $\gamma = \frac{1}{\beta(t)}\mathsf{J}\nabla V$, where the matrix J has the following form, for some smooth function $c(t) : \mathbb{R}^+ \to \mathbb{R}$

$$\mathsf{J} = \begin{pmatrix} 0 & \beta(t)c(t) \\ -\beta(t)c(t) & 0 \end{pmatrix}, \quad i.e. \quad \gamma(t,x) = \begin{pmatrix} c(t)\partial_{x_2}V(x) \\ -c(t)\partial_{x_1}V(x) \end{pmatrix}.$$

It is easy to check that $\nabla \cdot (\pi(t,x)\gamma(t,x)) = 0$ (e.g.: see (47) below). Applying Proposition 25, we have

$$\begin{aligned} \mathfrak{R} &= \beta(t) \begin{pmatrix} \partial_{x_1x_1}V - c(t)\partial_{x_1x_2}V & \partial_{x_1x_2}V - c(t)\frac{1}{2}(-\partial_{x_1x_1}V + \partial_{x_2x_2}V) \\ \partial_{x_2x_1}V - c(t)\frac{1}{2}(-\partial_{x_1x_1}V + \partial_{x_2x_2}V) & \partial_{x_2x_2}V + c(t)\partial_{x_2x_1}V \end{pmatrix} \\ &=: \beta(t)\mathsf{B}(t,x). \end{aligned} \tag{45}$$

Comparing with the Corollary 21 and Corollary 22, the irreversible vector field $\gamma(t,x)$ only changes the matrix $\mathfrak{R}$, but does not change the estimate of $\mathrm{A}(t)$. If the smallest eigenvalue of $\mathsf{B}(t,x)$ is bigger than the smallest eigenvalue of $\nabla^2_{xx}V$ for a proper choice of the function $c(t)$, the convergence of stochastic dynamics (42) can be faster than the underdamped Langevin dynamics (32).

Variable matrices J. We also study a case with the variable coefficient anti-symmetric vector field. Consider a two-dimensional stochastic differential equation:

$$dX_t = (-\nabla V(X_t) - \mathsf{J}(t,X_t)\nabla V(X_t) - \beta(t)\nabla \cdot \mathsf{J}(t,X_t))dt + \sqrt{2\beta(t)}dB_t, \tag{46}$$

where we define

$$\mathsf{J} = \begin{pmatrix} 0 & c(t,x) \\ -c(t,x) & 0 \end{pmatrix},$$

and

$$\begin{aligned} \gamma(t,x) &= aa^{\mathsf{T}}\nabla \log \pi - b + \beta(t)(\frac{\partial}{\partial x_j}(\alpha\alpha^{\mathsf{T}})_{ij})_{i=1}^n \\ &= \begin{pmatrix} c(t,x)\partial_{x_2}V(x) \\ -c(t,x)\partial_{x_1}V(x) \end{pmatrix} + \beta(t) \begin{pmatrix} -\partial_{x_2}c(t,x) \\ \partial_{x_1}c(t,x) \end{pmatrix}. \end{aligned}$$

Here $\pi(t,x) = \frac{1}{Z(t)} e^{-\frac{V(x)}{\beta(t)}}$. We also have the fact that $\nabla \cdot (\pi\gamma) = \frac{1}{Z}\nabla \cdot (e^{-V}\gamma) = 0$, since

$$\begin{aligned}
\nabla \cdot (\pi\gamma) &= \nabla \cdot \Big(\pi \begin{pmatrix} c(t,x)\partial_{x_2} V(x) \\ -c(t,x)\partial_{x_1} V(x) \end{pmatrix} + \pi\beta \begin{pmatrix} -\partial_{x_2} c(t,x) \\ \partial_{x_1} c(t,x) \end{pmatrix} \Big) \\
&= \langle \nabla\pi, \begin{pmatrix} c(t,x)\partial_{x_2} V(x) \\ -c(t,x)\partial_{x_1} V(x) \end{pmatrix} + \beta \begin{pmatrix} -\partial_{x_2} c(t,x) \\ \partial_{x_1} c(t,x) \end{pmatrix} \rangle \\
&\quad + \pi\nabla \cdot \Big(\begin{pmatrix} c(t,x)\partial_{x_2} V(x) \\ -c(t,x)\partial_{x_1} V(x) \end{pmatrix} + \beta \begin{pmatrix} -\partial_{x_2} c(t,x) \\ \partial_{x_1} c(t,x) \end{pmatrix} \Big) \\
&= -\frac{\pi}{\beta} \Big\langle \begin{pmatrix} \partial_{x_1} V \\ \partial_{x_2} V \end{pmatrix}, \begin{pmatrix} c(t,x)\partial_{x_2} V(x) \\ -c(t,x)\partial_{x_1} V(x) \end{pmatrix} + \beta \begin{pmatrix} -\partial_{x_2} c(t,x) \\ \partial_{x_1} c(t,x) \end{pmatrix} \Big\rangle \\
&\quad + \pi\partial_{x_1}[c(t,x)\partial_{x_2} V - \beta\partial_{x_2} c(t,x)] + \pi\partial_{x_2}[-c(t,x)\partial_{x_1} V + \beta\partial_{x_1} c(t,x)] \\
&= -\frac{\pi}{\beta}(c\partial_{x_1} V\partial_{x_2} V - c\partial_{x_1} V\partial_{x_2} V) + \pi(\partial_{x_1} V\partial_{x_2} c - \partial_{x_2} V\partial_{x_1} c) \\
&\quad + \pi[\partial_{x_1} c\partial_{x_2} V + c\partial_{x_1 x_2} V - \beta\partial_{x_1 x_2} c - \partial_{x_2} c\partial_{x_1} V - c\partial_{x_1 x_2} V + \beta\partial_{x_1 x_2} c] \\
&= 0. \qquad (47)
\end{aligned}$$

For the matrix $\mathfrak{R}$, with $\gamma_1 = c(t,x)\partial_{x_2} V - \beta(t)\partial_{x_2} c(t,x)$, and $\gamma_2 = -c(t,x)\partial_{x_1} V + \beta(t)\partial_{x_1} c(t,x)$, we have

$$\begin{aligned}
\mathfrak{R} = &\quad \beta \begin{pmatrix} \partial_{x_1 x_1} V & \partial_{x_1 x_2} V \\ \partial_{x_2 x_1} V & \partial_{x_2 x_2} V \end{pmatrix} - \beta \begin{pmatrix} \partial_{x_1}\gamma_1 & \frac{1}{2}(\partial_{x_1}\gamma_2 + \partial_{x_2}\gamma_1) \\ \frac{1}{2}(\partial_{x_1}\gamma_2 + \partial_{x_2}\gamma_1) & \partial_{x_2}\gamma_2 \end{pmatrix} \\
= &\quad \beta \begin{pmatrix} \partial_{x_1 x_1} V & \partial_{x_1 x_2} V \\ \partial_{x_2 x_1} V & \partial_{x_2 x_2} V \end{pmatrix} \\
&\quad -\beta \begin{pmatrix} \partial_{x_1} c\partial_{x_2} V + c\partial_{x_1 x_2} V - \beta\partial_{x_1 x_2} c & \mathfrak{R}_{\gamma,12} \\ \mathfrak{R}_{\gamma,12} & -\partial_{x_2} c\partial_{x_1} V - c\partial_{x_2 x_1} V + \beta\partial_{x_2 x_1} c \end{pmatrix},
\end{aligned}$$

where $\mathfrak{R}_{\gamma,12} = \frac{1}{2}[c(\partial_{x_2 x_2} V - \partial_{x_1 x_1} V) + \beta(-\partial_{x_2 x_2} c + \partial_{x_1 x_1} c) + \partial_{x_2} c\partial_{x_2} V - \partial_{x_1} c\partial_{x_1} V]$.

Example 1 *Let us consider an example where V is a two dimensional quadratic form with*

$$\nabla^2_{xx} V = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix},$$

such that $a > b > 0$. This implies that $\nabla^2_{xx} V$ is positive definite. We assume that the minimum of V is at $(0,0)$. Let $c(t,x) = c(x)$ be another quadratic form such that it has the same global minimum as V. Denote by $c''_{ij} = \partial_{x_i x_j} c$. We now consider the neighbourhood near the global minimum, so that all first-order partial derivatives of V can be neglected, and $c \approx 0$. Then the matrix $\mathfrak{R}$ is approximated by

$$\mathfrak{R} \approx \beta \begin{pmatrix} a + \beta c''_{12} & -\frac{1}{2}\beta(c''_{11} - c''_{22}) \\ -\frac{1}{2}\beta(c''_{11} - c''_{22}) & b - \beta c''_{12} \end{pmatrix}.$$

There are many choices of c to make the smallest eigenvalues of $\mathfrak{R}$ larger than that of V. For instance, take $c''_{11} = c''_{22} = 0$, $c''_{12} = (b-a)/2\beta(t)$. In this case, the smallest eigenvalue is $(a+b)/2$ whereas the smallest eigenvalue of V is b. A visualization is shown Fig. 4

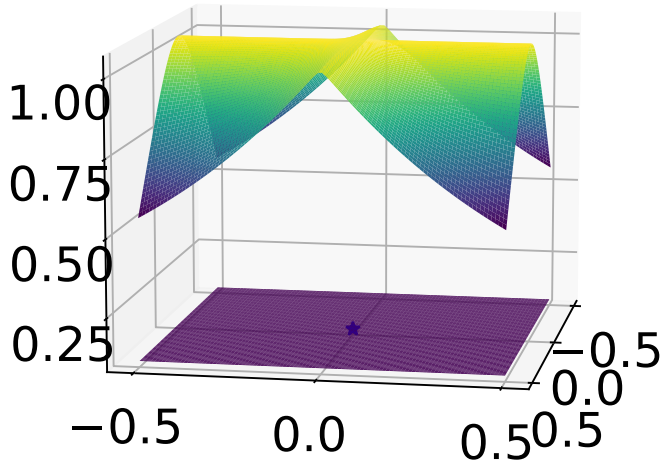


Figure 4: Eigenvalue comparison between $V(x)$ and $\mathfrak{R}(x)$ for $x \in [-0.5, 0.5] \times [-0.5, 0.5]$. The parameters are chosen as in Example 1. The yellow surface represents the smallest eigenvalue of $\mathfrak{R}(x)$ and the purple surface represents the smallest eigenvalue of $V(x)$. The global minimum of V is marked with a blue asterisk. As shown in the figure, the smallest eigenvalue of $\mathfrak{R}$ is larger than that of V near the global minimum.

when $a = 2$, $b = 0.1$, $\beta = 1$. *Now let us consider* $\beta(t) = \frac{1}{t_0+t}$ *for* $t_0 = 1$. *We use the Euler-Maruyama scheme to run* (37) *and* (46) *with a step size of* $dt = 5 \times 10^{-5}$ *for* $3 * 10^5$ *iterations. We use* 10^4 *particles initially sampled from a standard Gaussian distribution for our comparison. The result is demonstrated in Fig. 5. We observe that equation* (46) *yields a faster convergence towards the global minimum than equation* (37).

6. Example III: underdamped Langevin dynamics

In this section, we consider an underdamped Langevin dynamics with variable diffusion coefficients:

$$\begin{cases} dx_t = v_t dt \\ dv_t = (-r(t, x_t)v_t - \nabla_x V(x_t))dt + \sqrt{2r(t, x_t)}dB_t, \end{cases} \tag{48}$$

where $n = m = 1$, $d = n+m = 2$, $X_t = (x_t, v_t) \in \mathbb{R}^2$ is a two dimensional stochastic process, $V \in \mathbb{C}^2(\mathbb{R}^1)$ is a Lipschitz potential function with assumption $\int_{\mathbb{R}^1} e^{-V(x)}dx < +\infty$, B_t is a standard Brownian motion in $\mathbb{R}$, and $r : \mathbb{R}_+ \times \mathbb{R}^2 \to \mathbb{R}_+$ is a positive smooth Lipschitz

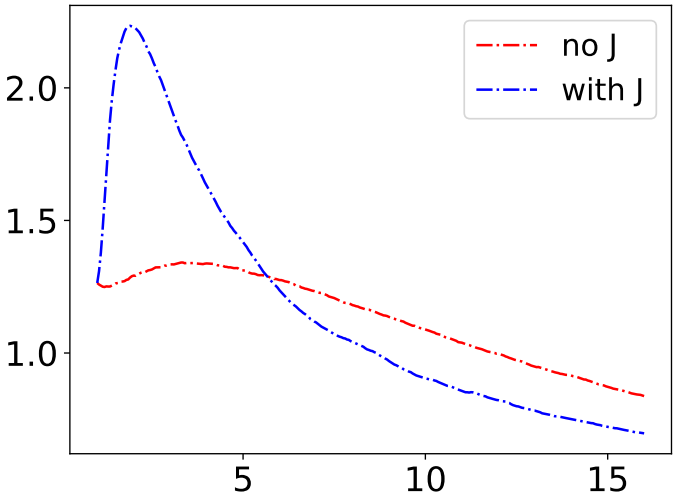


Figure 5: Convergence comparison between (37) and (46) in Example 1. x-axis represents time. y-axis represents the average distance of the particles to the global minimum using the two SDEs.

function. Indeed, the reference measure $\pi(t,x,v) = \pi(x,v)$ is the invariant measure, defined as

$$\pi(x,v) = \frac{1}{Z} e^{-H(x,v)}, \quad H(x,v) = \frac{v^2}{2} + V(x),$$

where $Z = \int_{\mathbb{R}^2} e^{-H(x,v)} dxdv < +\infty$ is a normalization constant. Following the definition of diffusion matrix a, the vector field γ and the correction term $\mathcal{R}$, we have

$$a = \begin{pmatrix} 0 \\ \sqrt{r(t,x)} \end{pmatrix}, \quad \gamma = \begin{pmatrix} -v \\ \nabla V(x) \end{pmatrix}, \quad \text{and} \quad \mathcal{R}(t,x,\pi) = 0, \tag{49}$$

since $\partial_t \pi(x,v) = 0$, and $\nabla \cdot (\pi\gamma) = 0$.

Remark 26 *Sampling from a posterior distribution has numerous applications in scientific computing, including Bayesian inference (Gelman et al., 1995; Newman and Barkema, 1999), inverse problems (Stuart, 2010; Dashti and Stuart, 2013), as well as Bayesian machine learning and Bayesian neural network (Welling and Teh, 2011; Andrieu et al., 2003; Izmailov et al., 2021). To sample a distribution of the form* $\exp(-V)/Z$*, one popular choice is to use the overdamped Langevin dynamics* (37) *with* $\beta(t) = \beta$*. In a seminal work, Jordan et al. (1998) showed that the Kolmogorov forward equation of the overdamped Langevin dynamics corresponds to the gradient flow of the relative entropy functional in the space of measures with the Wasserstein metric. On the other hand, the Kolmogorov forward equation of the underdamped Langevin dynamics corresponds to the accelerated gradient flow (Su et al., 2016) of the relative entropy functional. Following this idea, researchers have been*

trying to prove the accelerated convergence speed and designing better numerical implementations of underdamped Langevin dynamics (Ma et al., 2021; Cao et al., 2023; Cheng et al., 2018; Zhang et al., 2023; Shen and Lee, 2019) .

Note that the diffusion matrix $a(t,x)$ has rank $n=1$. Following from the Condition in (7), we construct matrix $z(t,x)$ such that $z(t,x)$ also has rank 1, and Condition (7) holds true. Under this consideration, we can select a time-dependent vector field $z=\begin{pmatrix} z_1(t,x) \\ z_2(t,x)\end{pmatrix}$ in the most general form satisfying the above assumptions. We have the following proposition. The derivation follows similar studies in the time-independent case as shown in (Feng and Li., 2021). We skip the details here.

Proposition 27 *For the time-dependent underdamped Langevin dynamics* (48)*, the time-dependent Hessian matrix function $\mathfrak{R}(t,x):\mathbb{R}_+\times\mathbb{R}^2\to\mathbb{R}^{2\times 2}$ has the following form,*

$$\mathfrak{R} \;=\; \mathfrak{R}_a+\mathfrak{R}_z+\mathfrak{R}_\pi-\mathfrak{M}_\Lambda+\mathfrak{R}_{\gamma_a}+\mathfrak{R}_{\gamma_z},$$

where $a_{21}=\sqrt{r}$, and

$$\begin{aligned}
\mathfrak{R}_a &= \begin{pmatrix} 0 & 0 \\ 0 & -\frac{\partial^2\log\pi}{\partial v^2}|a_{21}|^4\end{pmatrix}, \quad \mathfrak{R}_\pi=\begin{pmatrix}0&0\\0&C_\pi\end{pmatrix},\\
\mathfrak{R}_z &= \frac{1}{2}\Big[\begin{pmatrix}0\\ -z_1^\mathsf{T}\nabla((a_{21})^2\frac{\partial\log\pi}{\partial v})\end{pmatrix}z_1^\mathsf{T}+z_1\begin{pmatrix}0 & -z_1^\mathsf{T}\nabla((a_{21})^2\frac{\partial\log\pi}{\partial v})\end{pmatrix}\Big],\\
\mathfrak{R}_{\gamma_a} &= \frac{1}{2}\gamma_1\nabla_1(aa^\mathsf{T})-\frac{1}{2}\Big[(\nabla\gamma)^\mathsf{T}aa^\mathsf{T}+aa^\mathsf{T}\nabla\gamma\Big],\\
\mathfrak{R}_{\gamma_z} &= \frac{1}{2}\gamma_1\nabla_1(zz^\mathsf{T})-\frac{1}{2}\Big[(\nabla\gamma)^\mathsf{T}zz^\mathsf{T}+zz^\mathsf{T}\nabla\gamma\Big], \quad \mathfrak{M}_\Lambda=\frac{1}{(a_{21})^2}\mathsf{K}^\mathsf{T}(aa^\mathsf{T}+zz^\mathsf{T})^{-1}\mathsf{K},
\end{aligned}$$

with

$$\begin{aligned}
C_\pi &= 2\Big[z_1^\mathsf{T}z_1^\mathsf{T}\nabla^2a_{21}a_{21}+(z_1^\mathsf{T}\nabla a_{21})^2+(z_1^\mathsf{T}\nabla\log\pi)[z_1^\mathsf{T}\nabla a_{21}a_{21}]\Big],\\
\mathsf{K} &= \begin{pmatrix}0 & 2z_1^2\partial_x[a_{21}]a_{21}-\frac{1}{2}\beta\gamma_1(a_{21})^2\\ -z_1^2\partial_x[a_{21}]a_{21}+\frac{1}{2}\beta\gamma_1(a_{21})^2 & z_1z_2\partial_x[a_{21}]a_{21}\end{pmatrix}.
\end{aligned}$$

If there exists a constant $\lambda>0$, such that

$$\mathfrak{R}-\frac{1}{2}\partial_t(aa^\mathsf{T}+zz^\mathsf{T})\succeq\lambda(aa^\mathsf{T}+zz^\mathsf{T}),$$

then the Fisher information decay in Theorem 6 holds.

In the following, we consider a special case where we choose $r(t,x)=r(t)$, and $z=\begin{pmatrix}z_1\\z_2\end{pmatrix}$, for some constants $z_1,z_2\in\mathbb{R}$. For such a constant matrix z, it is easily verified that conditions (7) and (8) hold true for positive constants z_1 and z_2.

In this case, the matrix $\mathfrak{R}(t,x)$ is simplified into the following form,

$$\mathfrak{R}=\mathfrak{R}_a+\mathfrak{R}_z+\mathfrak{R}_{\gamma_a}+\mathfrak{R}_{\gamma_z},$$

where we have

$$\begin{aligned}\mathfrak{R}_a &= \begin{pmatrix}0 & 0\\ 0 & (r(t))^2\end{pmatrix}, \quad \mathfrak{R}_z = r(t)\begin{pmatrix}0 & \frac{z_1z_2}{2}\\ \frac{z_1z_2}{2} & z_2^2\end{pmatrix},\\ \mathfrak{R}_{\gamma_a} &= r(t)\begin{pmatrix}0 & \frac{1}{2}\\ \frac{1}{2} & 0\end{pmatrix}, \quad \mathfrak{R}_{\gamma_z} = \begin{pmatrix}z_1z_2 & \frac{1}{2}(z_2^2 - z_1^2\nabla_{xx}^2V)\\ \frac{1}{2}(z_2^2 - z_1^2\nabla_{xx}^2V) & -z_1z_2\nabla_{xx}^2V\end{pmatrix}.\end{aligned}$$

Proposition 28 (Sufficient conditions) *In the above example,*

$$\begin{aligned}&\mathfrak{R} - \frac{1}{2}\partial_t(aa^{\mathsf{T}})\\ =\ &\begin{pmatrix}z_1z_2 & \frac{1}{2}[r(t) + r(t)z_1z_2 + z_2^2 - z_1^2\nabla_{xx}^2V(x)]\\ \frac{1}{2}[r(t) + r(t)z_1z_2 + z_2^2 - z_1^2\nabla_{xx}^2V(x)] & (r(t))^2 + r(t)z_2^2 - z_1z_2\nabla_{xx}^2V(x) - \frac{1}{2}\partial_t r(t)\end{pmatrix}.\end{aligned}$$

Assume that $0 < \underline{\lambda} \le \partial_{xx}^2V \le \overline{\lambda}$, *and there exist constants* $z_2 \in (0, \frac{r(t)+\sqrt{r(t)^2+4r(t)}}{2})$, *for all* $t \ge t_0$, *such that* $\underline{\lambda}$, $\overline{\lambda}$ *satisfy the following conditions:*

$$-\overline{\lambda}^2 + [2(r(1+z_2) - z_2^2)]\underline{\lambda} - [(1-z_2)r + z_2^2]^2 - 2z_2\partial_t r > 0, \quad r^2 + rz_2^2 - \overline{\lambda}z_2 - \frac{1}{2}\partial_t r > 0. \tag{50}$$

Then there exists a function $\lambda(t) > 0$, *for* $t > t_0$, *such that*

$$\mathfrak{R} - \frac{1}{2}\partial_t(aa^{\mathsf{T}}) \succeq \lambda(aa^{\mathsf{T}} + zz^{\mathsf{T}}). \tag{51}$$

Proof For notation convenience, we take $r = r(t)$. It is sufficient to prove $\mathsf{det}(\mathfrak{R}) > 0$ for $z_1z_2 > 0$ and $r^2 + rz_2^2 - \partial_{xx}^2Vz_1z_2 - \frac{1}{2}\partial_t r > 0$, which is equivalent to

$$z_1z_2(r^2 + rz_2^2 - \partial_{xx}^2Vz_1z_2 - \frac{1}{2}\partial_t r) - \frac{1}{4}(rz_1z_2 + z_2^2 - \partial_{xx}^2Vz_1^2 + r)^2 > 0.$$

It is equivalent to the following inequality:

$$-z_1^4(\partial_{xx}^2V)^2 + [2(r(1+z_1z_2) - z_2^2)z_1^2]\partial_{xx}^2V - [(1-z_1z_2)r + z_2^2]^2 - 2z_1z_2\partial_t r > 0. \tag{52}$$

According to the assumption of ∂_{xx}^2V, it is sufficient to prove the following conditions:

$$\begin{cases} z_1z_2 > 0, \quad r^2 + rz_2^2 - \overline{\lambda}z_1z_2 - \frac{1}{2}\partial_t r > 0, \quad (r(1+z_1z_2) - z_2^2) > 0;\\ -z_1^4\overline{\lambda}^2 + [2(r(1+z_1z_2) - z_2^2)z_1^2]\underline{\lambda} - [(1-z_1z_2)r + z_2^2]^2 - 2z_1z_2\partial_t r > 0. \end{cases} \tag{53}$$

Let $z_1 = 1$, then (50) is equivalent to (53). We complete the proof. ■

The next corollary estimates λ in (51) under some specific choices of parameters.

Corollary 29 *If* $z_2 = z_1 = 1$, $\overline{\lambda} > \frac{1}{2\underline{\lambda}} + \frac{\underline{\lambda}}{2} + 1$, $\beta \ge \overline{\lambda}/2$, *we have* $\mathfrak{R} - \frac{1}{2}\partial_t(aa^{\mathsf{T}}) \succeq 0$ *as* $t \to \infty$. *Suppose further that* $\beta = \overline{\lambda}/2$, *and* $\overline{\lambda} \ge \underline{\lambda} + 2$, *then we have* $\lambda = \mathcal{O}(\frac{2\underline{\lambda}}{\overline{\lambda}} - \frac{1}{\overline{\lambda}^2})$.

Proof Since $r(t) = \beta + C/\log t$, we have that $r(t) \to \beta$ and $\partial_t r \to 0$ as $t \to \infty$. Denote by $u(x) = \partial^2_{xx} V(x)$ and let $\beta = \overline{\lambda}/2$. We directly compute

$$\begin{aligned}\mathsf{det}(\mathfrak{R} - \frac{1}{2}\partial_t(aa^\mathsf{T})) &= \frac{\overline{\lambda}^2}{4} + \frac{\overline{\lambda}}{2} - u(x) - \frac{1}{4}(\overline{\lambda} + 1 - u(x))^2 \\ &= \frac{\overline{\lambda}u(x)}{2} - \frac{u(x)}{2} - \frac{u(x)^2}{4} - \frac{1}{4}\,, \end{aligned} \tag{54}$$

which is a quadratic function in $u(x)$. One can check that since $0 < \underline{\lambda} \le u(x) \le \overline{\lambda}$ for all x, we have $\mathsf{det}(\mathfrak{R} - \frac{1}{2}\partial_t(aa^\mathsf{T})) > 0$ as long as

$$\overline{\lambda} > \max\{\frac{1}{2\underline{\lambda}} + \frac{\underline{\lambda}}{2} + 1, 1 + \sqrt{2}\} = \frac{1}{2\underline{\lambda}} + \frac{\underline{\lambda}}{2} + 1\,.$$

Now let $\beta = \frac{\overline{\lambda}}{2}$. We want to find the largest λ, such that

$$\mathfrak{R} - \frac{1}{2}\partial_t(aa^\mathsf{T}) - \lambda(aa^\mathsf{T} + zz^\mathsf{T}) = \begin{pmatrix} 1-\lambda & \frac{1}{2}(\overline{\lambda} + 1 - u(x)) - \lambda \\ \frac{1}{2}(\overline{\lambda} + 1 - u(x)) - \lambda & \frac{\overline{\lambda}^2}{4} + \frac{\overline{\lambda}}{2} - u(x) - (1 + \frac{\overline{\lambda}}{2})\lambda \end{pmatrix} \succeq 0,$$

as $t \to \infty$. This translates to

$$1 - \lambda \ge 0, \tag{55}$$

$$\frac{\overline{\lambda}^2}{4} + \frac{\overline{\lambda}}{2} - u(x) - (1 + \frac{\overline{\lambda}}{2})\lambda \ge 0, \tag{56}$$

$$-1 + 2(\overline{\lambda} - 1)u(x) - u(x)^2 - \overline{\lambda}(\overline{\lambda} - 2\lambda)\lambda \ge 0. \tag{57}$$

Define $f(u, \lambda) = -1 + 2(\overline{\lambda} - 1)u(x) - u(x)^2 - \overline{\lambda}(\overline{\lambda} - 2\lambda)\lambda$. It is clear that when λ is fixed, f is quadratic in u and peaks at $u = \overline{\lambda} - 1$. We also have that by definition of $u(x)$, we have $\underline{\lambda} \le u(x) \le \overline{\lambda}$ for all x. Therefore,

$$\min_u f(u, \lambda) = \min\{f(\overline{\lambda}, \lambda), f(\underline{\lambda}, \lambda), f(\overline{\lambda} - 1, \lambda)\}\,.$$

When $\overline{\lambda} \ge \underline{\lambda} + 2$, the above implies $\min_u f(u, \lambda) = f(\underline{\lambda}, \lambda)$. Thus (57) is satisfied as long as $f(\underline{\lambda}, \lambda) \ge 0$. We would like to maximize λ subject to the constraints $f(\underline{\lambda}, \lambda) \ge 0$ together with (55) and (56). From (56) we have that

$$\lambda \le \frac{\overline{\lambda}}{2} - \frac{\underline{\lambda}}{1 + \overline{\lambda}/2}\,.$$

Using our assumption $\overline{\lambda} \ge \underline{\lambda} + 2$ we get that

$$\frac{\overline{\lambda}}{2} - \frac{\underline{\lambda}}{1 + \overline{\lambda}/2} > 1\,.$$

Therefore, (55) and (56) together imply $\lambda \le 1$. Observe that $f(\underline{\lambda}, \lambda)$ is a quadratic function of λ, which produces two roots. It is straightforward to check that the larger root of f is

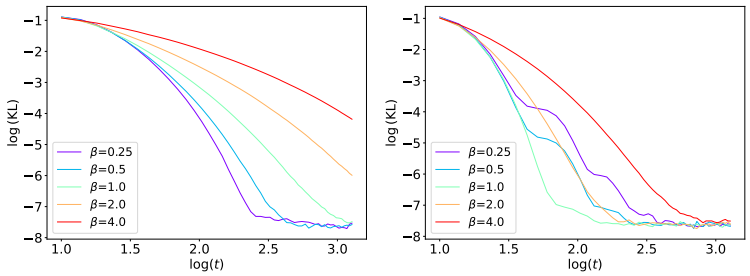
greater than 1. Hence, we conclude that λ cannot be larger than the smaller root of f. We have

$$\begin{aligned}\lambda_{\max} &= \frac{\overline{\lambda}}{4} - \frac{1}{4}\sqrt{\frac{8}{\overline{\lambda}} + \overline{\lambda}^2 + 16\frac{\underline{\lambda}}{\overline{\lambda}} - 16\underline{\lambda} + 8\frac{\underline{\lambda}^2}{\overline{\lambda}}} \\ &\approx \frac{\overline{\lambda}}{4} - \frac{1}{4}\sqrt{\frac{8}{\overline{\lambda}} + \overline{\lambda}^2 - 16\underline{\lambda}} \\ &\approx \frac{2\underline{\lambda}}{\overline{\lambda}} - \frac{1}{\overline{\lambda}^2}\,,\end{aligned} \tag{58}$$

where our approximation holds when $\underline{\lambda}/\overline{\lambda} \ll 1$. ■

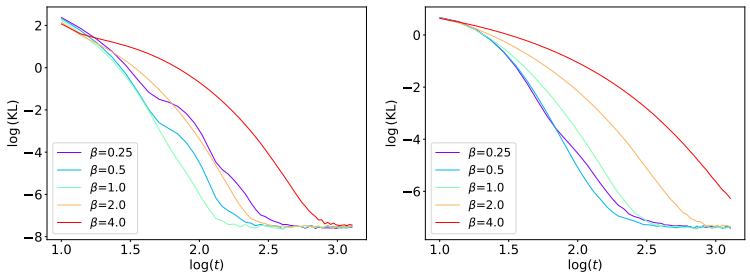
6.1 Numerics

We plot the convergence of (48) in Fig. 6 for strongly convex functions and in Fig. 7 for non-convex functions. We have also plotted the KL divergence for x variable only in Fig. 8 and Fig. 9. We used the same experiment setting as described in Section 4.3. And we use the Euler-Maruyama discretization for underdamped Langevin dynamics. In all of our numerical experiments, we observe that the KL divergence converges to 0. Comparing Fig. 1 with Fig. 8, we observe that the convergence speed of the underdamped Langevin dynamics (48) has a greater dependence on the constant than overdamped Langevin dynamics (37) does (recall that there is a constant C in $\beta(t)$ in (37) and a constant β in $r(t)$ in (48)). If the constant is chosen appropriately, the underdamped Langevin dynamics could converge much faster to the invariant measure than the overdamped Langevin dynamics. In both Fig. 8 and Fig. 9, we observe oscillations of the error, which is a typical phenomenon in accelerated convex optimization methods (Attouch et al., 2020, 2021; Zuo et al., 2023). Designing the optimal constant β in $r(t)$ with fast convergence speed is a delicate issue that is left for future studies.



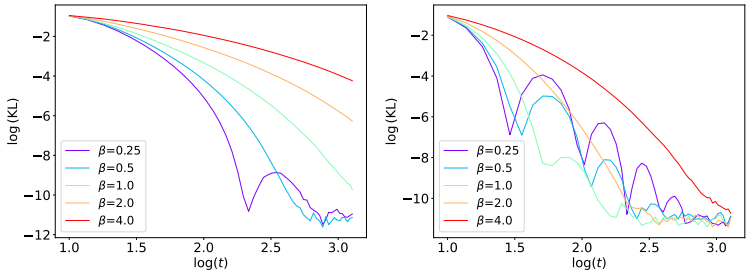
(a) $V(x) = \frac{(x-1)^2}{8}$ (b) $V(x) = \frac{(x+1)^2}{2} - \frac{\cos(x)}{2}$

Figure 6: Convergence rate of two strongly convex functions in one-dimension for (48) with $r(t) = \beta + 1/\log(t)$, where we measure the KL divergence in both x and v.



(a) $V(x) = \frac{(x+2)^4}{4} - \frac{x^2}{2} + \frac{x}{8}$ (b) $V(x) = \frac{(x-2)^2}{2} - \frac{\sin(5x)}{2}$

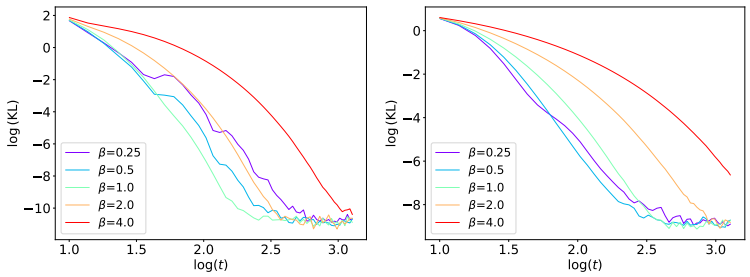
Figure 7: Convergence rate of two non-convex functions in one-dimension for (48) with $r(t) = \beta + 1/\log(t)$, where we measure the KL divergence in both x and v variables.



(a) $V(x) = \frac{(x-1)^2}{8}$ (b) $V(x) = \frac{(x+1)^2}{2} - \frac{\cos(x)}{2}$

Figure 8: Convergence rate of two strongly convex functions in one-dimension for (48), where we only measure the KL divergence in the x variable.

Remark 30 *We briefly review some theoretical efforts on proving the algorithmic dimensional dependence of Langevin dyanmics for sampling. Note that dimensional dependence stems from discretization schemes of the continuous SDEs. The first non-asymptotic analysis of unadjusted Langevin algorithm (ULA) is performed by Dalalyan (2017) in which the author proved that in order to achieve a total variation error less than ε, one needs $\mathcal{O}(d/\varepsilon^2)$ iterations, where d is the dimension of the problem. Then Durmus and Moulines (2016) obtained the same complexity for Wasserstein-2 distance. Cheng et al. (2018) proposed a discretization scheme for underdamped Langevin dynamics based on Hamiltonian Monte Carlo (HMC) (Simon et al., 1987; Neal, 2010), that is able to achieve an ε error in Wasserstein-2 distance in $\mathcal{O}(d^{1/2}/\varepsilon)$ steps. Later Shen and Lee (2019) proposed a randomized midpoint method for discretizing underdamped Langevin dynamics based on Cheng et al. (2018) that is able to improve the dimensional dependence further to $\mathcal{O}(d^{1/3})$. As discretization is not the focus of this paper, we only demonstrate an Gaussian example in higher dimensions using the Euler-Maruyama discretization of underdamped Langevin dynamics in Fig. 10. The y-axis is* $\log(\mathrm{KL}/d)$. *Error curves representing different dimensions seem to coincide with each other. This implies that for our quadratic potential, the KL di-*



(a) $V(x) = \frac{(x+2)^4}{4} - \frac{x^2}{2} + \frac{x}{8}$ (b) $V(x) = \frac{(x-2)^2}{2} - \frac{\sin(5x)}{2}$

Figure 9: Convergence rate of two non-convex functions in one-dimension for (48), where we only measure the KL divergence in the x variable.

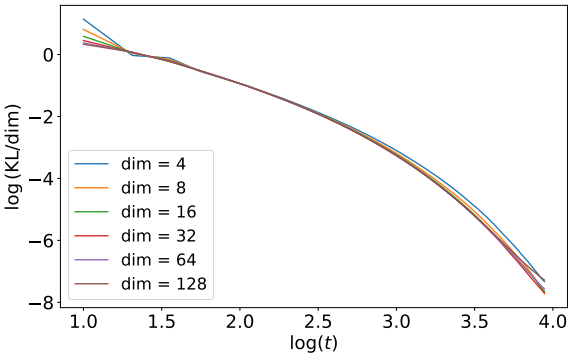


Figure 10: Quadratic potential in higher dimensions. For each example, we fix the smallest and largest eigenvalues of the covariance to be 0.05 and 100. The rest of the eigenvalues are uniformly spaced between 0.05 and 100. We choose $\beta = 0.1$. We use $M = 10^5$ particles, and step size $h = 0.02$. The KL divergence is measaured only in the x variable.

vergence depends linearly on the dimension when we fix the smallest and largest eigenvalues of the target covariance matrix and use an Euler-Maruyama discretization. We leave more detailed study of discretization schemes to future works.

7. Discussion

This paper studies the convergence analysis of time-dependent stochastic dynamics. We obtain a time-dependent Hessian matrix condition, which characterizes the convergence behavior of stochastic dynamics in terms of generalized Fisher information functionals. Examples of convergence speeds are shown, including over-damped, irreversible drift and degenerate diffusion, and underdamped Langevin dynamics. We also present several numerical experiments to verify the current convergence analysis of general stochastic dynamics.

In future work, we shall investigate the "optimal" choice of time-dependent matrix function a and vector field γ to find the global minimizer of a non-convex function V. Here, the "optimal" is in the sense of fast convergence speed towards the global minimizer. However, as we see in this paper, the convergence analysis for stochastic algorithms is more delicate than their deterministic counterparts. This requires us to estimate the general Hessian matrix, a.k.a. Ricci curvature lower bound, from both diffusion matrices a and non-gradient vector from γ. They depend on the second derivatives of coefficients in stochastic dynamics. The other practical issue is the estimation of step sizes in the Euler-Maruyama scheme (41). The related discrete-time convergence analysis of stochastic algorithms is left in future studies.

Acknowledgments

Q. Feng is partially supported by the National Science Foundation under grant DMS-2306769, DMS-2420029. X. Zuo and W. Li are supported by AFOSR MURI FA9550-18-1-0502, AFOSR YIP award No. FA9550-23-1-0087, and NSF RTG: 2038080.

References

C. Andrieu, F. N. De, A. Doucet, and M. I. Jordan. An introduction to MCMC for machine learning. *Machine Learning*, 50:5–43, 2003.

H. Attouch, Z. Chbani, J. Fadili, and H. Riahi. First-order optimization algorithms via inertial systems with Hessian driven damping. *Mathematical Programming*, pages 1–43, 2020.

H. Attouch, Z. Chbani, J. Fadili, and H. Riahi. Convergence of iterates for first-order optimization algorithms with inertia and Hessian driven damping. *Optimization*, pages 1–40, 2021.

D. Bakry and M. Émery. Diffusions hypercontractives. In *Séminaire de Probabilités XIX 1983/84: Proceedings*, pages 177–206. Springer, 2006.

F. Baudoin and N. Garofalo. Curvature-dimension inequalities and Ricci lower bounds for sub-riemannian manifolds with transverse symmetries. *Journal of the European Mathematical Society*, 19(1):151–219, 2016.

E. Bayraktar, Q. Feng, and W. Li. Exponential Entropy Dissipation for Weakly Self-Consistent Vlasov–Fokker–Planck Equations. *Journal of Nonlinear science*, 34(1):7, 2024.

Y. Cao, J. Lu, and L. Wang. On explicit L_2-convergence rate estimate for underdamped Langevin dynamics. *Archive for Rational Mechanics and Analysis*, 247(5):90, 2023.

T. Cass, D. Crisan, P. Dobson, and M. Ottobre. Long-time behaviour of degenerate diffusions: UFG-type SDEs and time-inhomogeneous hypoelliptic. *Electronic Journal of Probability*, 26:1–72, 2021.

P. Cattiaux and L. Mesnager. Hypoelliptic non-homogenous diffusions. *Probability Theory and Related Fields*, 123:453–483, 2002.

V. Cerny. Thermodynamical approach to the traveling salesman problem: an efficient simulation algorithm. *Journal of Optimization Theory and Applications*, 45:41–51, 1985.

M. Chaleyat-Maurel and D. Michel. Hypoellipticity Theorems and Conditional Laws. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 65(4):573–597, 1984.

X. Cheng, N. S. Chatterji, P. L. Bartlett, and M. I. Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. In *Conference on Learning Theory*, pages 300–323. PMLR, 2018.

T.-S. Chiang, C.-R. Hwang, and S. J. Sheu. Diffusion for global optimization in $\mathbf{R}^n$. *SIAM Journal on Control and Optimization*, 25:737–753, 1987.

L. Chizat. Mean-Field Langevin Dynamics: Exponential Convergence and Annealing. *Transactions on Machine Learning Research*, pages 2835–8856, 2022.

A. S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(3):651–676, 2017.

M. Dashti and A. M. Stuart. The Bayesian approach to inverse problems. *arXiv preprint arXiv:1302.6989*, 2013.

A. B. Duncan, T. Lelievre, and G. A. Pavliotis. Variance Reduction Using Nonreversible Langevin Samplers. *Journal of Statistical Physics*, 163:457–491, 2016.

A. B. Duncan, G. A. Pavliotis, and K. Zygalakis. Nonreversible Langevin samplers: Splitting schemes, analysis and implementation. *arXiv preprint arXiv:1701.04247*, 2017.

A. Durmus and É. Moulines. Sampling from a strongly log-concave distribution with the Unadjusted Langevin Algorithm. *arXiv preprint arXiv:1605.01559*, 2016.

H. Fang, M. Qian, and G. Gong. An improved annealing method and its large-time behavior. *Stochastic Processes and their Applications*, 71:55–74, 1997.

Q. Feng and W. Li. Hypoelliptic Entropy dissipation for stochastic differential equations. *arXiv preprint arXiv:2102.00544*, 2021.

Q. Feng and W. Li. Entropy Dissipation for Degenerate Stochastic Differential Equations via Sub-Riemannian Density Manifold. *Entropy*, 25:786, 2023.

X. Gao, Z. Q. Xu, and X. Y. Zhou. State-dependent temperature control for Langevin diffusions. *arXiv preprint arXiv2005.04507*, 2020.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

S. Geman and C. R. Hwang. Diffusions for global optimization. *SIAM Journal on Control and Optimization*, 24:1031–1043, 1986.

R. Höpfner, E. Löcherbach, and M. Thieullen. Strongly degenerate time inhomogeneous SDEs: Densities and support properties. *Bernoulli*, 23(4):2587–2616, 2017.

L. Hörmander. Hypoelliptic second order differential equations. *Acta Mathematica*, 119: 147–171, 1967.

C.-R. Hwang, W. Raoul Normand, and S.-J. Wu. Variance reduction for diffusions. *Stochastic Processes and their Applications*, 125(9):3522–3540, 2015.

P. Izmailov, S. Vikram, M. D. Hoffman, and A. G. Wilson. What are Bayesian neural network posteriors really like? In *International Conference on Machine Learning*, pages 4629–4640. PMLR, 2021.

R. Jordan, D. Kinderlehrer, and F. Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

A. Khachaturyan, S. Semenovsovskaya, and B. Vainshtein. Statistical-thermodynamic approach to determination of structure amplitude phases. *Sov. Phys. Crystallography*, 24 (5):519–524, 1979.

A. Khachaturyan, S. Semenovsovskaya, and B. Vainshtein. The thermodynamic approach to the structure analysis of crystals. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 37(5):742–754, 1981.

S. Kirkpatrick, J. Gelatt, and M. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.

S. Kusuoka and D. Stroock. Applications of the Malliavin calculus, part i. In *North-Holland Mathematical Library*, volume 32, pages 271–306. Elsevier, 1984.

T. Lelievre, F. Nier, and G. A. Pavliotis. Optimal non-reversible linear drift for the convergence to equilibrium of a diffusion. *Journal of Statistical Physics*, 152(2):237–274, 2013.

Y.-A. Ma, N. S. Chatterji, X. Cheng, N. Flammarion, P. L. Bartlett, and M. I. Jordan. Is there an analog of Nesterov acceleration for gradient-based MCMC? *Bernoulli*, 27(3): 1942–1992, 2021.

G. Menz, A. Schlichting, W. Tang, and T. Wu. Ergodicity of the infinite swapping algorithm at low temperature. *arXiv preprint arXiv:1811.10174*, 2018.

P. Monmarché. Hypocoercivity in metastable settings and kinetic simulated annealing. *Probability Theory and Related Fields*, pages 1–34, 2018.

R. M. Neal. Mcmc using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, pages 113–162, 2010.

M. E. J. Newman and G. T. Barkema. *Monte Carlo methods in statistical physics*. Clarendon Press, 1999.

M. Pincus. A Monte Carlo method for the approximate solution of certain types of constrained optimization problems. *Operations research*, 18(6):1225–1228, 1970.

L. Rey-Bellet and K. Spiliopoulos. Irreversible Langevin samplers and variance reduction: a large deviations approach. *Nonlinearity*, 28(7):2081, 2015.

R. Shen and Y. T. Lee. The randomized midpoint method for log-concave sampling. *Advances in Neural Information Processing Systems*, 32, 2019.

S. Duane Simon, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Physics letters B*, 195(2):216–222, 1987.

A. M. Stuart. Inverse problems: a Bayesian perspective. *Acta numerica*, 19:451–559, 2010.

W. Su, S. Boyd, and E. J. Candes. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17 (153):1–43, 2016.

W. Tang and X. Zhou. Simulated annealing from continuum to discretization: a convergence analysis via the Eyring-Kramers law. *arXiv preprint arXiv:2102.02339*, 2021.

P. J. M. van Laarhoven and E. H. L. Aarts. *Simulated annealing*. Springer, 1987.

C. Villani. *Hypocoercivity*. Memoirs of the American Mathematical Society, 2009.

C. Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 681–688. Citeseer, 2011.

S.-J. Wu, C.-R. Hwang, and M. T. Chu. Attaining the optimal Gaussian diffusion acceleration. *Journal of Statistical Physics*, 155:571–590, 2014.

B. J. Zhang, Y. M. Marzouk, and K. Spiliopoulos. Geometry-informed irreversible perturbations for accelerated convergence of Langevin dynamics. *Statistics and Computing*, 32 (5):78, 2022.

S. Zhang, S. Chewi, M. Li, K. Balasubramanian, and M. A. Erdogdu. Improved Discretization Analysis for Underdamped Langevin Monte Carlo. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 36–71. PMLR, 12–15 Jul 2023. URL `https://proceedings.mlr.press/v195/zhang23a.html`.

X. Zuo, S. Osher, and W. Li. Primal-Dual damping algorithms for optimization. *arXiv preprint arXiv:2304.14574*, 2023.

Appendix A.

The time-independent version of the Hessian matrix is first introduced in Feng and Li. (2021, Definition 1). For completeness of this paper, we introduce the time-dependent version of it for matrices $a(t,x)$ and $z(t,x)$, and we take the interpolation parameter $\beta = 0$ for Feng and Li. (2021, Definition 1), since we do not always have $\nabla \cdot (\pi(t,x)\gamma(t,x)) = 0$, which can be seen in Lemma 14 and Lemma 17. This is a major difference compared to Feng and Li. (2021, Proposition 9). Note that, both $\widetilde{\Gamma}_2$ and $\widetilde{\Gamma}_2^{z,\pi}$ only involve spacial derivatives, thus the following Bochner's formula (Feng and Li., 2021, Theorem 3) holds true. We first introduce the information Hessian matrix.

Definition 31 (Hessian matrix) *Let matrices $a(t,x)$ and $z(t,x)$ satisfy the Hörmander like condtion, and conditions* (7), (8)*. We define a bilinear form associated with SDE* (1)*, and matrices a, z as below, for a smooth vector field $\mathsf{U} \in C^\infty(\mathbb{R}^{n+m};\mathbb{R}^{n+m})$,*

$$\mathfrak{R}(\mathsf{U},\mathsf{U}) \;=\; (\mathfrak{R}_a + \mathfrak{R}_z + \mathfrak{R}_\pi + \mathfrak{R}_{\gamma_a} + \mathfrak{R}_{\gamma_z})(\mathsf{U},\mathsf{U}) - \Lambda_1^\mathsf{T}\Lambda_1 - \Lambda_2^\mathsf{T}\Lambda_2 + \mathsf{D}^\mathsf{T}\mathsf{D} + \mathsf{E}^\mathsf{T}\mathsf{E}. \tag{59}$$

We define $\mathfrak{R}(t,x) : \mathbb{R}_+ \times \mathbb{R}^{n+m} \to \mathbb{R}^{(n+m)\times(n+m)}$ as the corresponding time-dependent matrix function such that

$$\mathsf{U}^\mathsf{T}\mathfrak{R}(x)\mathsf{U} = \mathfrak{R}(\mathsf{U},\mathsf{U}), \tag{60}$$

for all vector fields U. The bilinear forms in (59) *are defined as below.*

$$\begin{aligned}
\mathfrak{R}_a(\mathsf{U},\mathsf{U}) \;=\;& \sum_{i,k=1}^n a_i^\mathsf{T}\nabla a_i^\mathsf{T}\nabla a_k^\mathsf{T}\mathsf{U}(a_k^\mathsf{T}\mathsf{U}) + \sum_{i,k=1}^n a_i^\mathsf{T} a_i^\mathsf{T}\nabla^2 a_k^\mathsf{T}\mathsf{U}(a_k^\mathsf{T}\mathsf{U}) \\
&- \sum_{i,k=1}^n a_k^\mathsf{T}\nabla a_i^\mathsf{T}\nabla a_i^\mathsf{T}\mathsf{U}(a_k^\mathsf{T}\mathsf{U}) - \sum_{i,k=1}^n a_k^\mathsf{T} a_i^\mathsf{T}\nabla^2 a_i^\mathsf{T}\mathsf{U}(a_k^\mathsf{T}\mathsf{U}) \\
&+ \sum_{i=1}^n \sum_{\hat{k}=1}^{n+m} \Big[(aa^\mathsf{T}\nabla\log\pi)_{\hat{k}}\nabla_{\hat{k}} a_i^\mathsf{T}\mathsf{U} - a_i^\mathsf{T}\nabla(aa^\mathsf{T}\nabla\log\pi)_{\hat{k}}\mathsf{U}_{\hat{k}}\Big] a_i^\mathsf{T}\mathsf{U} \\
&+ \nabla a \circ \Big(\sum_{k=1}^n \Big[a^\mathsf{T}\nabla a_k^\mathsf{T}\mathsf{U} - a_k^\mathsf{T}\nabla a^\mathsf{T}\mathsf{U})\Big] a_k^\mathsf{T}\mathsf{U}\Big) - \langle\Big(a^\mathsf{T}\nabla^2 a \circ (a^\mathsf{T}\mathsf{U})\Big), a^\mathsf{T}\mathsf{U}\rangle_{\mathbb{R}^n}, \\
\mathfrak{R}_z(\mathsf{U},\mathsf{U}) \;=\;& \sum_{i=1}^n\sum_{k=1}^m a_i^\mathsf{T}\nabla a_i^\mathsf{T}\nabla z_k^\mathsf{T}\mathsf{U}(z_k^\mathsf{T}\mathsf{U}) + \sum_{i,k=1}^n a_i^\mathsf{T} a_i^\mathsf{T}\nabla^2 z_k^\mathsf{T}\mathsf{U}(z_k^\mathsf{T}\mathsf{U}) \\
&- \sum_{i=1}^n\sum_{k=1}^m z_k^\mathsf{T}\nabla a_i^\mathsf{T}\nabla a_i^\mathsf{T}\mathsf{U}(z_k^\mathsf{T}\mathsf{U}) - \sum_{i,k=1}^n z_k^\mathsf{T} a_i^\mathsf{T}\nabla^2 a_i^\mathsf{T}\mathsf{U}(z_k^\mathsf{T}\mathsf{U}) \\
&+ \sum_{k=1}^m \sum_{\hat{k}=1}^{n+m} \Big[(aa^\mathsf{T}\nabla\log\pi)_{\hat{k}}\nabla_{\hat{k}} z_k^\mathsf{T}\mathsf{U} - z_k^\mathsf{T}\nabla(aa^\mathsf{T}\nabla\log\pi)_{\hat{k}}\mathsf{U}_{\hat{k}}\Big] z_k^\mathsf{T}\mathsf{U} \\
&+ \nabla a \circ \Big(\sum_{k=1}^m \Big[a^\mathsf{T}\nabla z_k^\mathsf{T}\mathsf{U} - z_k^\mathsf{T}\nabla a^\mathsf{T}\mathsf{U})\Big] z_k^\mathsf{T}\mathsf{U}\Big) - \langle\Big(z^\mathsf{T}\nabla^2 a \circ (a^\mathsf{T}\mathsf{U})\Big), z^\mathsf{T}\mathsf{U}\rangle_{\mathbb{R}^m},
\end{aligned}$$

$$
\begin{aligned}
\mathfrak{R}_{\pi}(\mathsf{U},\mathsf{U}) &= 2\sum_{k=1}^{m}\sum_{i=1}^{n}\Big[\nabla z_k^{\mathsf{T}} z_k^{\mathsf{T}}\nabla a_i^{\mathsf{T}}\mathsf{U}a_i^{\mathsf{T}}\mathsf{U} + z_k^{\mathsf{T}}\nabla z_k^{\mathsf{T}}\nabla a_i^{\mathsf{T}}\mathsf{U}a_i^{\mathsf{T}}\mathsf{U} + z_k^{\mathsf{T}} z_k^{\mathsf{T}}\nabla^2 a_i^{\mathsf{T}}\mathsf{U}a_i^{\mathsf{T}}\mathsf{U}\Big]\\
&\quad +2\sum_{k=1}^{m}\sum_{i=1}^{n}\Big[(z_k^{\mathsf{T}}\nabla a_i^{\mathsf{T}}\mathsf{U})^2 + (z^{\mathsf{T}}\nabla\log\pi)_k\Big[z_k^{\mathsf{T}}\nabla a_i^{\mathsf{T}}\mathsf{U}a_i^{\mathsf{T}}\mathsf{U}\Big]\Big]\\
&\quad -2\sum_{j=1}^{m}\sum_{l=1}^{n}\Big[\nabla a_l^{\mathsf{T}} a_l^{\mathsf{T}}\nabla z_j^{\mathsf{T}}\mathsf{U}z_j^{\mathsf{T}}\mathsf{U} + a_l^{\mathsf{T}}\nabla a_l^{\mathsf{T}}\nabla z_j^{\mathsf{T}}\mathsf{U}z_j^{\mathsf{T}}\mathsf{U} + a_l^{\mathsf{T}} a_l^{\mathsf{T}}\nabla^2 z_j^{\mathsf{T}}\mathsf{U}z_j^{\mathsf{T}}\mathsf{U}\Big]\\
&\quad -2\sum_{j=1}^{m}\sum_{l=1}^{n}\Big[(a_l^{\mathsf{T}}\nabla z_j^{\mathsf{T}}\mathsf{U})^2 + (a^{\mathsf{T}}\nabla\log\pi)_l\Big[a_l^{\mathsf{T}}\nabla z_j^{\mathsf{T}}\mathsf{U}z_j^{\mathsf{T}}\mathsf{U}\Big]\Big],\\
\mathfrak{R}_{\gamma_a}(\mathsf{U},\mathsf{U}) &= \frac{1}{2}\sum_{\hat{k}=1}^{n+m}\gamma_{\hat{k}}\langle\mathsf{U},\nabla_{\hat{k}}(aa^{\mathsf{T}})\mathsf{U}\rangle - \langle\nabla\gamma\mathsf{U}, aa^{\mathsf{T}}\mathsf{U}\rangle_{\mathbb{R}^{n+m}},\\
\mathfrak{R}_{\gamma_z}(\mathsf{U},\mathsf{U}) &= \frac{1}{2}\sum_{\hat{k}=1}^{n+m}\gamma_{\hat{k}}\langle\mathsf{U},\nabla_{\hat{k}}(zz^{\mathsf{T}})\mathsf{U}\rangle - \langle\nabla\gamma\mathsf{U}, zz^{\mathsf{T}}\mathsf{U}\rangle_{\mathbb{R}^{n+m}}.
\end{aligned}
$$

We define vector functions $\mathsf{D}:\mathbb{R}^{n+m}\to\mathbb{R}^{n^2\times 1}$, *and* $\mathsf{E}:\mathbb{R}^{n+m}\to\mathbb{R}^{(n\times m)\times 1}$ *as below,*

$$
\mathsf{D}_{ik} = \sum_{\hat{i},\hat{k}=1}^{n+m} a_{\hat{i}i}^{\mathsf{T}}\partial_{x_{\hat{i}}}a_{k\hat{k}}^{\mathsf{T}}\mathsf{U}_{\hat{k}}, \quad \mathsf{E}_{ik} = \sum_{\hat{i},\hat{k}=1}^{n+m} a_{\hat{i}i}^{\mathsf{T}}\partial_{x_{\hat{i}}}z_{k\hat{k}}^{\mathsf{T}}\mathsf{U}_{\hat{k}}. \tag{61}
$$

For $\beta\in\mathbb{R}$, *the vector functions* $\Lambda_1:\mathbb{R}^{n+m}\to\mathbb{R}^{n^2\times 1}$ *and* $\Lambda_2:\mathbb{R}^{n+m}\to\mathbb{R}^{(n\times m)\times 1}$ *are defined as, for* $i,l\in\{1,\cdots,n\}$,

$$
\begin{aligned}
(\Lambda_1)_{il} &= \sum_{k=1}^{n}[\sum_{i'=1}^{n+m}a_{ii'}^{\mathsf{T}}\lambda_l^{i'k} - \sum_{k'=1}^{n+m}a_{kk'}^{\mathsf{T}}\lambda_l^{k'i}]a_k^{\mathsf{T}}\mathsf{U} + \sum_{k=1}^{m}\Big(\sum_{i'=1}^{n+m}a_{ii'}^{\mathsf{T}}\omega_l^{i'k} - \sum_{k'=1}^{n+m}z_{kk'}^{\mathsf{T}}\lambda_l^{k'i}\Big)z_k^{\mathsf{T}}\mathsf{U}\\
&\quad -\sum_{k=1}^{m}\sum_{i'=1}^{n+m}a_{ii'}^{\mathsf{T}}\omega_l^{i'k}z_k^{\mathsf{T}}\mathsf{U} - \frac{\beta}{2}\alpha_l(a_i^{\mathsf{T}}\mathsf{U}) + \frac{\beta}{2}\langle\mathsf{U},\gamma\rangle\mathbf{1}_{\{i=l\}} + \mathsf{D}_{il},
\end{aligned}
$$

and for $i\in\{1,\cdots,n\}$, $l\in\{1,\cdots,m\}$,

$$
\begin{aligned}
(\Lambda_2)_{il} &= \sum_{k=1}^{n}[\sum_{i'=1}^{n+m}a_{ii'}^{\mathsf{T}}\lambda_{l+n}^{i'k} - \sum_{k'=1}^{n+m}a_{kk'}^{\mathsf{T}}\lambda_{l+n}^{k'i}]a_k^{\mathsf{T}}\mathsf{U} + \sum_{k=1}^{m}\Big(\sum_{i'=1}^{n+m}a_{ii'}^{\mathsf{T}}\omega_{l+n}^{i'k} - \sum_{k'=1}^{n+m}z_{kk'}^{\mathsf{T}}\lambda_{l+n}^{k'i}\Big)z_k^{\mathsf{T}}\mathsf{U}\\
&\quad +\sum_{k=1}^{n}\sum_{k'=1}^{n+m}z_{lk'}^{\mathsf{T}}\lambda_i^{k'k}a_k^{\mathsf{T}}\mathsf{U} + z_l^{\mathsf{T}}\nabla a_i^{\mathsf{T}}\mathsf{U} - \sum_{k=1}^{m}\sum_{i'=1}^{n+m}a_{ii'}^{\mathsf{T}}\omega_{l+n}^{i'k}z_k^{\mathsf{T}}\mathsf{U} - a_i^{\mathsf{T}}\nabla z_l^{\mathsf{T}}\mathsf{U} - \frac{\beta}{2}\alpha_{l+n}(a_i^{\mathsf{T}}\mathsf{U}) + \mathsf{E}_{il}.
\end{aligned}
$$

For each indices $i,k,\hat{k}$, *assume that there exist smooth functions* $\lambda_l^{i'k}$, $\omega_l^{i'k}$ *and* α_l *for* $l=1,\cdots,n+m$,

$$
\nabla_{i'}a_{k\hat{k}}^{\mathsf{T}} = \sum_{l=1}^{n}\lambda_l^{i'k}a_{l\hat{k}}^{\mathsf{T}} + \sum_{l=1}^{m}\lambda_{l+n}^{i'k}z_{l\hat{k}}^{\mathsf{T}}, \quad \nabla_{i'}z_{k\hat{k}}^{\mathsf{T}} = \sum_{l=1}^{n}\omega_l^{i'k}a_{l\hat{k}}^{\mathsf{T}} + \sum_{l=1}^{m}\omega_{l+n}^{i'k}z_{l\hat{k}}^{\mathsf{T}},
$$

and $\gamma_{\hat{k}} = \sum_{l=1}^{n}\alpha_l a_{l\hat{k}}^{\mathsf{T}} + \sum_{l=1}^{m}\alpha_{l+n}z_{l\hat{k}}^{\mathsf{T}}$. *For a vector function* $\gamma\in\mathbb{R}^{n+m}$, *we define* $\nabla\gamma\in\mathbb{R}^{(n+m)\times(n+m)}$ *with* $(\nabla\gamma)_{ij} = \nabla_i\gamma_j$.

Proposition 32 (Information Bochner's formula) *If the Assumption in* (8) *is satisfied, then the following decomposition holds. For any* $f = \log \frac{p}{\pi} \in C^\infty(\mathbb{R}^{n+m}, \mathbb{R})$ *and* $\beta = 0$,

$$\int \Big[\widetilde{\Gamma}_2(f,f) + \widetilde{\Gamma}_2^{z,\pi}(f,f)\Big] pdx = \int \Big[\|\mathfrak{Hess}_\beta f\|_{\mathrm{F}}^2 + (\mathfrak{R}_a + \mathfrak{R}_z + \mathfrak{R}_\pi)(\nabla f, \nabla f)\Big] pdx.$$

We denote

$$\|\mathfrak{Hess}_\beta f\|_{\mathrm{F}}^2 = [\mathsf{Q}\mathsf{X} + \Lambda_1]^{\mathsf{T}}[\mathsf{Q}\mathsf{X} + \Lambda_1] + [\mathsf{P}\mathsf{X} + \Lambda_2]^{\mathsf{T}}[\mathsf{P}\mathsf{X} + \Lambda_2],$$

where $\mathfrak{R}$, Λ_1, Λ_2 *are defined in Definition 31. And we define matrices* Q *and* P *by*

$$\mathsf{Q} = a^{\mathsf{T}} \otimes a^{\mathsf{T}} \in \mathbb{R}^{n^2 \times (n+m)^2}, \quad \mathsf{P} = a^{\mathsf{T}} \otimes z^{\mathsf{T}} \in \mathbb{R}^{(nm) \times (n+m)^2}, \tag{62}$$

with $\mathsf{Q}_{ik\hat{i}\hat{k}} = a_{i\hat{i}}^{\mathsf{T}} a_{k\hat{k}}^{\mathsf{T}}$ *and* $\mathsf{P}_{ik\hat{i}\hat{k}} = a_{i\hat{i}}^{\mathsf{T}} z_{k\hat{k}}^{\mathsf{T}}$. *More precisely, for each row (resp. column) of* Q, *the row (resp. column) indices of* $\mathsf{Q}_{ik\hat{i}\hat{k}}$ *follow the double summation* $\sum_{i=1}^n \sum_{k=1}^n$ *(resp.* $\sum_{\hat{i}=1}^{n+1} \sum_{\hat{k}=1}^{n+m}$*). For any smooth function* $f : \mathbb{R}^{n+m} \to \mathbb{R}$, *we define* $\mathsf{X} \in \mathbb{R}^{(n+m)^2 \times 1}$ *by the vectorization of the Hessian matrix for function* f *with*

$$\mathsf{X}_{\hat{i}\hat{k}} = \frac{\partial^2 f}{\partial x_{\hat{i}} \partial x_{\hat{k}}}, \quad \textit{for } \hat{i}, \hat{k} = 1, \cdots, n+m.$$

Proof For self-consistence, we present the key steps to prove the above Information Bochner's formulas. For any vector field $\mathsf{U} \in \mathcal{C}^\infty(\mathbb{R}^{n+m})$, we define vectors $\mathsf{C}, \mathsf{F}, \mathsf{G} \in \mathbb{R}^{(n+m)^2 \times 1}$ as below. For $\hat{i}, \hat{k} = 1, \cdots, n+m$,

$$\begin{aligned}
\mathsf{C}_{\hat{i}\hat{k}} &= \sum_{i,k=1}^{n} \Big(a_{i\hat{i}}^{\mathsf{T}} a_i^{\mathsf{T}} \nabla a_{k\hat{k}}^{\mathsf{T}} - a_{i\hat{k}}^{\mathsf{T}} a_k^{\mathsf{T}} \nabla a_{i\hat{i}}^{\mathsf{T}}\Big) a_k^{\mathsf{T}} \mathsf{U}, \quad \mathsf{F}_{\hat{i}\hat{k}} = \sum_{i=1}^{n} \sum_{k=1}^{m} \Big(a_{i\hat{i}}^{\mathsf{T}} a_i^{\mathsf{T}} \nabla z_{k\hat{k}}^{\mathsf{T}} - z_k^{\mathsf{T}} \nabla a_{i\hat{i}}^{\mathsf{T}} a_{i\hat{k}}^{\mathsf{T}}\Big) z_k^{\mathsf{T}} \mathsf{U}, \\
\mathsf{G}_{\hat{i}\hat{k}} &= \sum_{i=1}^{n} \sum_{k=1}^{m} \Big[\Big(z_{k\hat{k}}^{\mathsf{T}} z_k^{\mathsf{T}} \nabla a_{i\hat{i}}^{\mathsf{T}} a_i^{\mathsf{T}} \mathsf{U} + a_{i\hat{i}}^{\mathsf{T}} z_{k\hat{k}}^{\mathsf{T}} z_k^{\mathsf{T}} \nabla a_i^{\mathsf{T}} \mathsf{U}\Big) - \Big(a_{i\hat{i}}^{\mathsf{T}} a_i^{\mathsf{T}} \nabla z_{k\hat{k}}^{\mathsf{T}} z_k^{\mathsf{T}} \mathsf{U} + z_{k\hat{k}}^{\mathsf{T}} a_{i\hat{i}}^{\mathsf{T}} a_i^{\mathsf{T}} \nabla z_k^{\mathsf{T}} \mathsf{U}\Big)\Big].
\end{aligned}$$

For the Information Gamma operators defined in Definition 7, following from (Feng and Li., 2021, Proposition 11), we have

$$\begin{aligned}
\widetilde{\Gamma}_2(f,f) &= (\mathsf{Q}\mathsf{X} + \mathsf{D})^{\mathsf{T}}(\mathsf{Q}\mathsf{X} + \mathsf{D}) + 2\mathsf{C}^{\mathsf{T}}\mathsf{X} + \mathfrak{R}_a(\nabla f, \nabla f), \\
\widetilde{\Gamma}_2^z(f,f) &= (\mathsf{P}\mathsf{X} + \mathsf{E})^{\mathsf{T}}(\mathsf{P}\mathsf{X} + \mathsf{E}) + 2\mathsf{F}^{\mathsf{T}}\mathsf{X} + \mathfrak{R}_z(\nabla f, \nabla f), \\
\mathbf{div}_z^\pi(\Gamma_{\nabla(aa^{\mathsf{T}})} f, f) - \mathbf{div}_a^\pi(\Gamma_{\nabla(zz^{\mathsf{T}})} f, f) &= \mathfrak{R}_\pi(\nabla f, \nabla f) + 2\mathsf{G}^{\mathsf{T}}\mathsf{X}.
\end{aligned}$$

Here we denote $\widetilde{\Gamma}_2^z(f,f) = \frac{1}{2}\widetilde{L}\Gamma_1^z(f,f) - \Gamma_1^z(\widetilde{L}f, f)$. Thus, we end up with

$$\begin{aligned}
&\int \Big[\widetilde{\Gamma}_2(f,f) + \widetilde{\Gamma}_2^{z,\pi}(f,f)\Big] pdx \\
= &\int \Big[(\mathsf{Q}\mathsf{X} + \mathsf{D})^{\mathsf{T}}(\mathsf{Q}\mathsf{X} + \mathsf{D}) + 2\mathsf{C}^{\mathsf{T}}\mathsf{X} + \mathfrak{R}_a(\nabla f, \nabla f)\Big] pdx \\
&+ \int \Big[(\mathsf{P}\mathsf{X} + \mathsf{E})^{\mathsf{T}}(\mathsf{P}\mathsf{X} + \mathsf{E}) + 2\mathsf{F}^{\mathsf{T}}\mathsf{X} + \mathfrak{R}_z(\nabla f, \nabla f) + \mathfrak{R}_\pi(\nabla f, \nabla f) + 2\mathsf{G}^{\mathsf{T}}\mathsf{X}\Big] pdx.
\end{aligned}$$

By completing the squares for the above quadratic form, we have

$$\begin{aligned} & (\mathsf{QX}+\mathsf{D})^{\mathsf{T}}(\mathsf{QX}+\mathsf{D})+(\mathsf{PX}+\mathsf{E})^{\mathsf{T}}(\mathsf{PX}+\mathsf{E})+2\mathsf{C}^{\mathsf{T}}\mathsf{X}+2\mathsf{F}^{\mathsf{T}}\mathsf{X}+2\mathsf{G}^{\mathsf{T}}\mathsf{X} \\ = \; & [\mathsf{QX}+\Lambda_1]^{\mathsf{T}}[\mathsf{QX}+\Lambda_1]+[\mathsf{PX}+\Lambda_2]^{\mathsf{T}}[\mathsf{PX}+\Lambda_2]-\Lambda_1^{\mathsf{T}}\Lambda_1-\Lambda_2^{\mathsf{T}}\Lambda_2+\mathsf{D}^{\mathsf{T}}\mathsf{D}+\mathsf{E}^{\mathsf{T}}\mathsf{E}. \end{aligned}$$

This follows from the Assumption in (8). The above equality is a special case for (Feng and Li., 2021, Proof of Theorem 3) with $\beta = 0$. Combining the above terms, we complete the proof. Note here, we do not include the irreversible Gamma operators $\Gamma_{\mathcal{I}_a}$ and $\Gamma_{\mathcal{I}_z}$ above, since they are separately discussed in Lemma 14 and Lemma 17. ■