W. Li, L. Ying *Res Math Sci* (2019) 6:34
https://doi.org/10.1007/s40687-019-0198-9

Research in the Mathematical Sciences

RESEARCH

Hessian transport gradient flows

Wuchen Li[1*] and Lexing Ying[2]

*Correspondence:
wcli@math.ucla.edu
[1]Department of Mathematics, University of California, Los Angeles, USA
Full list of author information is available at the end of the article
Wuchen Li is supported by AFOSR MURI FA9550-18-1-0502

Abstract

We derive new gradient flows of divergence functions in the probability space embedded with a class of Riemannian metrics. The Riemannian metric tensor is built from the transported Hessian operator of an entropy function. The new gradient flow is a generalized Fokker–Planck equation and is associated with a stochastic differential equation that depends on the reference measure. Several examples of Hessian transport gradient flows and the associated stochastic differential equations are presented, including the ones for the reverse Kullback–Leibler divergence, α-divergence, Hellinger distance, Pearson divergence, and Jenson–Shannon divergence.

Keywords: Optimal transport, Information/Hessian geometry, Hessian transport, Hessian transport stochastic differential equations, Generalized de Bruijn identity

1 Introduction

The de Bruijn identity plays crucial roles in information theory, probability, statistics, geometry, and machine learning [11–13,30,34,38]. It states that the dissipation of the relative entropy, also known as the Kullback–Leibler (KL) divergence function, along the heat flow is equal to the relative Fisher information functional. This identity is important for many applications in Bayesian statistics and Markov chain Monte Carlo methods.

It turns out that there are two geometric structures in the probability space related to the de Bruijn identity. One is Wasserstein geometry (WG) [16,36], which refers to the heat flows or Gaussian kernels. In [15,32], it shows that the gradient flow of the negative Boltzmann Shannon entropy in WG is the heat equation. The de Bruijn identity can be understood as the rate of entropy dissipation within WG. The other one is information geometry (IG) [1,5], which relates to the differential structures of the entropy. IG studies various families of Hessian geometry of entropy and divergence functions. Here, the Boltzmann–Shannon entropy, the Fisher–Rao metric, and the further induced KL divergence function are of particular importance. Besides these classical cases, one also studies generalized entropy and divergence functions, such as Tsallis entropy and Tsallis divergence [2,35].

A natural question arises: *What are natural families of geometries in the probability space that connect entropy/divergence functions, heat flows, and the de Bruijn identity*?

In this paper, we positively answer this question by introducing a family of Riemannian metrics in the probability space. Consider a compact space Ω and a positive smooth probability space $\mathcal{P}(\Omega)$. For a strictly convex entropy function $\mathcal{H}\colon \mathcal{P}(\Omega) \to \mathbb{R}$, we introduce a

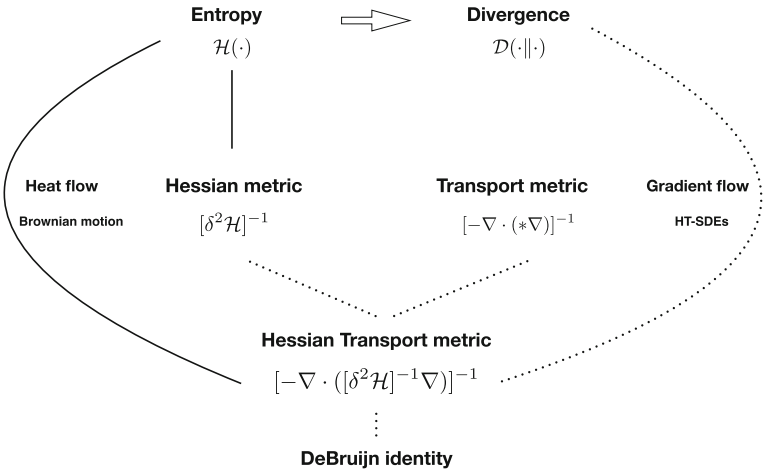

Fig. 1 Derivation diagram of the Hessian transport metric. The real lines are results of classical WG or IG communities. The dot lines are results derived in this paper

new Riemannian metric tensor $G_{\mathcal{H}}$ in the probability space

$$G_{\mathcal{H}}(\rho)^{-1} = [-\nabla\cdot([\delta^2\mathcal{H}(\rho)]^{-1}\nabla)],$$

where $\rho \in \mathcal{P}(\Omega)$ is a probability density function, $\delta^2\mathcal{H}(\rho)$ is the L^2 Hessian operator of the entropy function, and finally, ∇ and $\nabla\cdot$ are gradient and divergence operators on Ω, respectively. We refer to Definition 1 for the formal definition. Notice that the proposed metric involves both the Hessian geometry of $\mathcal{H}$ and the transport metric (gradient and divergence operator on sample space). For this reason, it is called the *Hessian transport metric* (HT-metric) (see Fig. 1 for a schematic diagram).

As a simple but motivating example, the heat equation is the gradient flow of the entropy function $\mathcal{H}$ under the HT-metric $G_{\mathcal{H}}$ induced by $\mathcal{H}$ itself:

$$\begin{aligned}\partial_t\rho &= -G_{\mathcal{H}}(\rho)^{-1}\delta\mathcal{H}(\rho)\\ &= -\Big(-\nabla\cdot([\delta^2\mathcal{H}(\rho)]^{-1}\nabla\delta\mathcal{H}(\rho))\Big)\\ &= \nabla\cdot([\delta^2\mathcal{H}(\rho)]^{-1}[\delta^2\mathcal{H}(\rho)]\nabla\rho)\\ &= \nabla\cdot(\nabla\rho) = \Delta\rho.\end{aligned}$$

In a more general setting, we consider a family of entropy functions of form $\mathcal{H}_f(\rho) = \int_\Omega f(\rho)\mathrm{d}x$, where $f(\cdot)$ is convex, $f(1) = 0$ and f'' is homogeneous of degree $(-\gamma)$. For a fixed reference measure μ, there is an associated divergence function $\mathcal{D}_f(\rho\|\mu) = \int_\Omega f(\frac{\rho}{\mu})\mu\mathrm{d}x$ for each $\mathcal{H}_f$. By considering the gradient flow of $\mathcal{D}_f(\rho\|\mu)$ in $(\mathcal{P}(\Omega), G_{\mathcal{H}_f})$, we derive a generalized Fokker–Planck equation

$$\partial_t\rho(t,x) = \nabla\cdot\left(\mu(x)^\gamma\nabla\left(\frac{\rho(t,x)}{\mu(x)}\right)\right),$$

along with a stochastic differential equation for independent particle dynamics

$$\mathrm{d}X_t = \gamma\mu(X_t)^{\gamma-2}\nabla\mu(X_t)\mathrm{d}t + \sqrt{2\mu(X_t)^{\gamma-1}}\mathrm{d}B_t,$$

where B_t is the standard Brownian motion. Such a SDE is called a *Hessian transport stochastic differential equation* (HT-SDE).

It is worth mentioning that two special cases of HT-metrics and their induced HT-SDEs are particularly relevant. When $\gamma = 1$, $\mathcal{H}(\rho)$ is the Boltzmann–Shannon entropy, the HT-metric is the usual Wasserstein-2 metric and the HT-SDE is the classical Langevin dynamics. When $\gamma = 0$, $\mathcal{H}(\rho)$ is the Pearson divergence, the HT-metric is the H^{-1} metric, and the associated HT-SDE is a diffusion process with zero drift. We refer to Table 1 for a summary of the results.

Currently, there are several efforts in combining both Wasserstein metric and information/Hessian metric [3,6–8,10,21,28,37] from various perspectives. There has also been recent work on using novel diffusion/mobility tensors for numerical and modeling purposes [23,27]. Within the Gaussian families, several extensions are studied in [29]. In [9,14], a generalization of optimal transport metrics have been studied, in which $\delta^2\mathcal{H}$ is only chosen as a scale function of ρ. Another example from the machine learning community is the Stein variational gradient descent method [24–26]. Here we introduce a new geometry structure, which keeps heat flows as gradient flows of general entropy functions. The emphasis of this work is the interaction between the Hessian of the entropy and the transport metric. By deriving the gradient flow of the entropy-generated divergence functional, we introduce a new class of stochastic differential equations. In this angle, our approach is a natural extension to both IG and WG. It can also be viewed as a generalization for the field of Wasserstein information geometry [17,19,20].

We summarize the main contributions of this paper as follows:

(i) We propose a framework of Riemannian metrics in probability space, which combines both transport operator and Hessian operator of entropy functional.
(ii) The new metrics allow us to derive gradient flows of various divergence functionals. These flows are probability transition equations, which further introduce new stochastic differential equations.

The rest of this paper is organized as follows. In Sect. 2, we define the Hessian transport metric and show that the heat flow can be interpreted as the gradient flow of several energy functions under appropriate HT-metrics. We then move on to derive, for general divergence functions, the HT-metric gradient flows and the associated HT-SDEs. In Sect. 3, we introduce the Hessian transport distance (HT-distance) and derive the corresponding HT-geodesic equation. Several numerical examples are given in Sect. 4.

2 Hessian transport gradient flows

In this section, we introduce the Hessian transport metrics and derive the gradient flows under these metrics.

2.1 Motivations

Consider a compact space $\Omega \subset \mathbb{R}^d$. Following the usual convention, we denote by ∇ and $\nabla\cdot$ the gradient and divergence operators in Ω, by $\|\cdot\|$ the Euclidean norm in $\mathbb{R}^d$ and by δ, δ^2 the first and the second L^2 variations. From now on, the boundary conditions on Ω are given by either Neumann or periodic boundary conditions.

Table 1 Hessian transport for KL divergence, reverse KL divergence, α-divergence, Hellinger distance, Pearson divergence, and Jensen–Shannon divergence

Divergence	Inverse HT-metric	Relative Fisher divergence	HT-gradient flow	HT-SDE	HT-geodesics
$\int_\Omega \rho \log \frac{\rho}{\mu} \mathrm{d}x$	$-\nabla \cdot (\rho \nabla)$	$\int_\Omega \left\| \nabla \log \frac{\rho}{\mu} \right\|^2 \rho \mathrm{d}x$	$\partial_t \rho = \nabla \cdot (\rho \nabla \log \mu) + \Delta \rho$	$\mathrm{d}X_t = \nabla \log \mu \mathrm{d}t + \sqrt{2} B_t$	$\partial_t \rho + \nabla \cdot (\rho \nabla \Phi) = 0$ $\partial_t \Phi + \frac{1}{2}(\nabla \Phi, \nabla \Phi) = 0$
$-\int_\Omega \mu \log \frac{\mu}{\rho} \mathrm{d}x$	$-\nabla \cdot (\rho^2 \nabla)$	$\int_\Omega \left\| \nabla \left(\frac{\rho}{\mu} \right)^{-1} \right\|^2 \rho^2 \mathrm{d}x$	$\partial_t \rho = \nabla \cdot \left(\mu^2 \nabla \left(\frac{\rho}{\mu} \right) \right)$	$\mathrm{d}X_t = 2 \nabla \mu \mathrm{d}t + \sqrt{2\mu} \mathrm{d}B_t$	$\partial_t \rho + \nabla \cdot (\rho^2 \nabla \Phi) = 0$ $\partial_t \Phi + (\nabla \Phi, \nabla \Phi) \rho = 0$
$\int_\Omega \frac{4}{1-\alpha^2} (1 - (\frac{\rho}{\mu})^{\frac{1+\alpha}{2}}) \mu \mathrm{d}x$	$-\nabla \cdot (\rho^{\frac{3-\alpha}{2}} \nabla)$	$(\frac{2}{\alpha-1})^2 \int_\Omega \left\| \nabla \left(\frac{\rho}{\mu} \right)^{\frac{\alpha-1}{2}} \right\|^2 \rho^{\frac{3-\alpha}{2}} \mathrm{d}x$	$\partial_t \rho = \nabla \cdot \left(\mu^{(3-\alpha)/2} \nabla \left(\frac{\rho}{\mu} \right) \right)$	$\mathrm{d}X_t = \frac{3-\alpha}{2} \mu^{\frac{-1-\alpha}{2}} \nabla \mu \mathrm{d}t + \sqrt{2 \mu^{\frac{1-\alpha}{2}}} \mathrm{d}B_t$	$\partial_t \rho + \nabla \cdot (\rho^{\frac{3-\alpha}{2}} \nabla \Phi) = 0$ $\partial_t \Phi + \frac{3-\alpha}{2} (\nabla \Phi, \nabla \Phi) \rho^{\frac{1-\alpha}{2}} = 0$
$\int_\Omega (\sqrt{\rho} - \sqrt{\mu})^2 \mathrm{d}x$	$-\nabla \cdot (\rho^{\frac{3}{2}} \nabla)$	$4 \int_\Omega \left\| \nabla \left(\frac{\rho}{\mu} \right)^{-\frac{1}{2}} \right\|^2 \rho^{\frac{3}{2}} \mathrm{d}x$	$\partial_t \rho = \nabla \cdot \left(\mu^{3/2} \nabla \left(\frac{\rho}{\mu} \right) \right)$	$\mathrm{d}X_t = \frac{3}{2} \mu^{-\frac{1}{2}} \nabla \mu \mathrm{d}t + \sqrt{2 \mu^{\frac{1}{2}}} \mathrm{d}B_t$	$\partial_t \rho + \nabla \cdot (\rho^{\frac{3}{2}} \nabla \Phi) = 0$ $\partial_t \Phi + \frac{3}{2} (\nabla \Phi, \nabla \Phi) \rho^{\frac{1}{2}} = 0$
$\frac{1}{2} \int_\Omega (\frac{\rho}{\mu} - 1)^2 \mu \mathrm{d}x$	$-\nabla \cdot (\nabla)$	$\int_\Omega \left\| \nabla \left(\frac{\rho}{\mu} \right) \right\|^2 \mathrm{d}x$	$\partial_t \rho = \nabla \cdot (\nabla (\frac{\rho}{\mu}))$	$\mathrm{d}X_t = \sqrt{2 \mu^{-1}} \mathrm{d}B_t$	$\partial_t \rho + \Delta \Phi = 0$ $\partial_t \Phi = 0$
$\int_\Omega \rho \log \frac{\rho}{\frac{1}{2}(\rho+\mu)} + \mu \log \frac{\mu}{\frac{1}{2}(\rho+\mu)} \mathrm{d}x$	$-\nabla \cdot (\rho(1+\rho) \nabla)$	$\int_\Omega \left\| \nabla \log (\frac{2\rho}{\rho+\mu}) \right\|^2 \rho(1+\rho) \mathrm{d}x$	$\partial_t \rho = \nabla \cdot \left(\frac{(1+\rho)\mu^2}{\rho+\mu} \nabla (\frac{\rho}{\mu}) \right)$	-	$\partial_t \rho + \nabla \cdot (\rho(1+\rho) \nabla \Phi) = 0$ $\partial_t \Phi + \frac{1}{2} (\nabla \Phi, \nabla \Phi)(2\rho + 1) = 0$

In the case $\alpha = 1$, the HT-metric recovers the Wasserstein-2 metric from the classical optimal transport theory. In the case of $\alpha = 3$, the HT-metric recovers the H^{-1} metric

The heat equation

$$\partial_t \rho(t,x) = \Delta \rho(t,x), \tag{1}$$

can be written in several equivalent ways as follows:

$$\begin{aligned}
\partial_t \rho(t,x) =& \nabla \cdot \Big(\nabla \rho(t,x) \Big), \\
\partial_t \rho(t,x) =& \nabla \cdot \Big(\rho(t,x) \nabla \log \rho(t,x) \Big), \\
\partial_t \rho(t,x) =& \nabla \cdot \left(\rho(t,x)^2 \nabla \left(-\frac{1}{\rho(t,x)} \right) \right),
\end{aligned}$$

where the following relation is used:

$$\nabla \rho = \rho \nabla \log \rho = \rho^2 \nabla \left(-\frac{1}{\rho} \right).$$

These formulas show that the heat flow has multiple gradient descent flow interpretations. Recall that a general gradient flow takes the form

$$\partial_t \rho = -G(\rho)^{-1} \delta \mathcal{E}(\rho),$$

where $\mathcal{E}(\cdot)$ is an energy function and the operator $G(\rho)\colon C^\infty(\Omega) \to C^\infty(\Omega)$ represents the metric tensor. Under this framework, the heat equation can be interpreted in several ways:

(i) Dirichlet energy formulation:

$$\mathcal{E}(\rho) = \int_\Omega \|\nabla \rho(x)\|^2 \mathrm{d}x, \quad \delta \mathcal{E}(\rho) = -\Delta \rho, \quad G(\rho)^{-1} = \mathbb{I},$$

where $\mathbb{I}\colon C^\infty(\Omega) \to C^\infty(\Omega)$ is the identity operator. Then

$$\partial_t \rho = -G(\rho)^{-1} \delta \mathcal{E}(\rho) = -(-\Delta \rho) = \Delta \rho.$$

(ii) Boltzmann–Shannon entropy formulation:

$$\begin{aligned}
\mathcal{E}(\rho) &= \int_\Omega \rho(x) \log \rho(x) \mathrm{d}x, \quad \delta \mathcal{E}(\rho) = \log \rho + 1, \quad G(\rho)^{-1} = -\nabla \cdot (\rho \nabla), \\
\partial_t \rho &= -G(\rho)^{-1} \delta \mathcal{E}(\rho) = -(-\nabla \cdot (\rho \nabla \log \rho)) = \Delta \rho.
\end{aligned}$$

(iii) Cross-entropy formulation:[1]

$$\begin{aligned}
\mathcal{E}(\rho) &= -\int_\Omega \log \rho(x) \mathrm{d}x, \quad \delta \mathcal{E}(\rho) = -\frac{1}{\rho}, \quad G(\rho)^{-1} = -\nabla \cdot (\rho^2 \nabla), \\
\partial_t \rho &= -G(\rho)^{-1} \delta \mathcal{E}(\rho) = -\left(-\nabla \cdot \left(\rho^2 \nabla \left(-\frac{1}{\rho} \right) \right) \right) = \Delta \rho.
\end{aligned}$$

[1] Given $\mu \in C^\infty(\Omega)$, the cross-entropy is defined as follows

$$\mathcal{H}(\rho, \mu) = -\int_\Omega \mu(x) \log \rho(x) \mathrm{d}x$$

Here we let $\mu(x) = 1$, for all $x \in \Omega$.

It is clear that the metric G and the energy $\mathcal{E}$ need to be compatible in order to give rise to the heat equation. In fact, given a strictly convex energy function $\mathcal{E}$ in the probability space, it induces a compatible metric operator

$$G(\rho)^{-1} := \Big(-\nabla \cdot ([\delta^2 \mathcal{E}(\rho)]^{-1} \nabla) \Big),$$

which combines both the transport operator (gradient, divergence operator in Ω) and the L^2 Hessian operator of $\mathcal{E}$. Following this relation, the heat equation can be viewed as the gradient flow of the energy $\mathcal{E}$ under the $\mathcal{E}$-induced metric operator. Below we include the calculations of the above three cases for the sake of completeness.

(i) Dirichlet energy formulation:

$$\begin{aligned} \mathcal{E}(\rho) &= \int_\Omega \|\nabla \rho(x)\|^2 \mathrm{d}x, \quad \delta^2 \mathcal{E}(\rho) = -\Delta, \\ G(\rho)^{-1} &= \Big(-\nabla \cdot ([\delta^2 \mathcal{E}(\rho)]^{-1} \nabla) \Big) = \Big(-\nabla \cdot ([-\Delta]^{-1} \nabla) \Big). \\ &= \mathbb{I}. \end{aligned}$$

(ii) Boltzmann–Shannon entropy formulation:

$$\begin{aligned} \mathcal{E}(\rho) &= \int_\Omega \rho(x) \log \rho(x) \mathrm{d}x, \quad \delta^2 \mathcal{E}(\rho) = \frac{1}{\rho}, \\ G(\rho)^{-1} &= -\nabla \cdot ([\delta^2 \mathcal{E}(\rho)]^{-1} \nabla) = -\nabla \cdot (\rho \nabla). \end{aligned}$$

(iii) Cross-entropy formulation:

$$\begin{aligned} \mathcal{E}(\rho) &= -\int_\Omega \log \rho(x) \mathrm{d}x, \quad \delta^2 \mathcal{E}(\rho) = \frac{1}{\rho^2} \\ G(\rho)^{-1} &= -\nabla \cdot ([\delta^2 \mathcal{E}(\rho)]^{-1} \nabla) \\ &= -\nabla \cdot (\rho^2 \nabla). \end{aligned}$$

2.2 Hessian transport gradient flows

In this subsection, we will make the discussion in Sect. 2.1 precise. Consider the set of smooth and strictly positive densities

$$\mathcal{P}(\Omega) = \Big\{ \rho \in C^\infty(\Omega) \colon \rho(x) > 0, \ \int_\Omega \rho(x) \mathrm{d}x = 1 \Big\}.$$

The tangent space of $\mathcal{P}(\Omega)$ at $\rho \in \mathcal{P}(\Omega)$ is given by

$$T_\rho \mathcal{P}(\Omega) = \Big\{ \sigma \in C^\infty(\Omega) \colon \int_\Omega \sigma(x) \mathrm{d}x = 0 \Big\}.$$

For a strictly convex entropy function $\mathcal{H} \colon \mathcal{P}(\Omega) \to \mathbb{R}$, we first define the following $\mathcal{H}$-induced metric tensor in the probability space.

Definition 1 (*Hessian transport metric tensor*) The inner product $G_{\mathcal{H}}(\rho) \colon T_\rho \mathcal{P}(\Omega) \times T_\rho \mathcal{P}(\Omega) \to \mathbb{R}$ is defined as for any σ_1 and $\sigma_2 \in T_\rho \mathcal{P}(\Omega)$:

$$G_{\mathcal{H}}(\rho)(\sigma_1, \sigma_2) = \int_\Omega \int_\Omega \Big(\sigma_1(x), \Big(-\nabla \cdot ([\delta^2 \mathcal{H}(\rho)]^{-1} \nabla) \Big)^{-1} (x, y) \sigma_2(y) \Big) \mathrm{d}x \, \mathrm{d}y,$$

where $[\delta^2 \mathcal{H}(\rho)]^{-1}$ is the inverse of L^2 Hessian operator of $\mathcal{H}$, and

$$\Big(-\nabla \cdot ([\delta^2 \mathcal{H}(\rho)]^{-1} \nabla) \Big)^{-1} \colon T_\rho \mathcal{P}(\Omega) \to T_\rho \mathcal{P}(\Omega)$$

is the inverse of weighted elliptic operator $-\nabla\cdot([\delta^2\mathcal{H}(\rho)]^{-1}\nabla)$.

The proposed metric tensor is an extension of the Wasserstein metric. To see it, we represent the metric tensor into a cotangent bundle [17,32]. Denote the space of potential functions on Ω by $\mathcal{F}(\Omega)$, and consider the quotient space $\mathcal{F}(\Omega)/\mathbb{R}$. Here each $\Phi\in\mathcal{F}(\Omega)/\mathbb{R}$ is a function defined up to an additive constant.

We first show that $\mathcal{F}(\Omega)/\mathbb{R}$ is the cotangent bundle $T^*_\rho\mathcal{P}(\Omega)$. Consider the identification map $\mathbf{V}\colon \mathcal{F}(\Omega)/\mathbb{R}\to T_\rho\mathcal{P}(\Omega)$ defined by

$$\mathbf{V}(\Phi)=-\nabla\cdot([\delta^2\mathcal{H}(\rho)]^{-1}\nabla\Phi).$$

At any ρ, define the elliptic operator

$$L_{\mathcal{H},\rho}=-\nabla\cdot([\delta^2\mathcal{H}(\rho)]^{-1}\nabla). \tag{2}$$

The uniform elliptic property of $L_{\mathcal{H},\rho}$ guarantees that $\mathbf{V}\colon \mathcal{F}(\Omega)/\mathbb{R}\to T_\rho\mathcal{P}(\Omega)$ is well defined, linear, and one to one. In other words, $\mathcal{F}(\Omega)/\mathbb{R}=T^*_\rho\mathcal{P}(\Omega)$. This identification further induces the following inner product on $T_\rho\mathcal{P}(\Omega)$.

Definition 2 (*Hessian transport metric on the cotangent bundle*) The inner product $G_{\mathcal{H}}(\rho):T_\rho\mathcal{P}(\Omega)\times T_\rho\mathcal{P}(\Omega)\to\mathbb{R}$ is defined as for any two tangent vectors $\sigma_1=\mathbf{V}(\Phi_1)$ and $\sigma_2=\mathbf{V}(\Phi_2)\in T_\rho\mathcal{P}(\Omega)$

$$\begin{aligned}G_{\mathcal{H}}(\rho)(\sigma_1,\sigma_2)&=\int_\Omega\sigma_1\Phi_2\mathrm{d}x=\int_\Omega\sigma_2\Phi_1\mathrm{d}x\\&=\int_\Omega\int_\Omega(\nabla\Phi_1(x),[\delta^2\mathcal{H}(\rho)]^{-1}(x,y)\nabla\Phi_2(y))\mathrm{d}x\,\mathrm{d}y.\end{aligned}$$

Here the equivalence of Definition 1 and 2 is shown as follows. By denoting $\sigma_i(x)=\mathbf{V}(\Phi_i)=L_{\mathcal{H},\rho}\Phi_i$ for $i=1,2$, i.e.,

$$\sigma_i(x)=-\nabla\cdot\left(\int_\Omega[\delta^2\mathcal{H}(\rho)]^{-1}(x,y)\nabla\Phi_i(y)\mathrm{d}y\right)(x),$$

one has

$$\begin{aligned}&\int_\Omega\int_\Omega(\nabla\Phi_1,[\delta^2\mathcal{H}(\rho)]^{-1}\nabla\Phi_2)\mathrm{d}x\,\mathrm{d}y=\int_\Omega\int_\Omega(\Phi_1,L_{\mathcal{H},\rho}\Phi_2)\mathrm{d}x\,\mathrm{d}y\\&=\int_\Omega\int_\Omega\mathbf{V}(\Phi_1)L^{-1}_{\mathcal{H},\rho}L_{\mathcal{H},\rho}L^{-1}_{\mathcal{H},\rho}\mathbf{V}(\Phi_2)\mathrm{d}x\,\mathrm{d}y=\int_\Omega\int_\Omega\sigma_1L^{-1}_{\mathcal{H},\rho}\sigma_2\mathrm{d}x\,\mathrm{d}y,\end{aligned}$$

where in the first equality we apply the integration by parts with respect to Ω using the boundary condition.

Remark 1 In particular, if $\mathcal{H}(\rho)=\int_\Omega\rho(x)\log\rho(x)\mathrm{d}x$, then $[\delta^2\mathcal{H}(\rho)]^{-1}=\rho$ and the Hessian transport metric takes the form

$$G_{\mathcal{H}}(\rho)(\sigma_1,\sigma_2)=\int_\Omega(\nabla\Phi_1,\nabla\Phi_2)\rho\mathrm{d}x,$$

with $\sigma_i=-\nabla\cdot(\rho\nabla\Phi_i)$, $i=1,2$. In this case, the Hessian transport metric is the Wasserstein-2 metric [32,36].

We are now ready to introduce the gradient flows in $(\mathcal{P}(\Omega), G_{\mathcal{H}})$.

Lemma 3 (Hessian transport Gradient flow) *Given an energy functional* $\mathcal{E}\colon \mathcal{P}(\Omega) \to \mathbb{R}$, *the gradient flow of* $\mathcal{E}$ *in* $(\mathcal{P}(\Omega), G_{\mathcal{H}})$ *is*

$$\partial_t \rho(t,x) = \nabla \cdot \left(\int_\Omega [\delta^2 \mathcal{H}(\rho)]^{-1}(x,y) \nabla \delta \mathcal{E}(\rho)(y) \mathrm{d}y \right).$$

Proof The proof follows the definition. The Riemannian gradient in $(\mathcal{P}(\Omega), G_{\mathcal{H}})$ is defined as

$$G_{\mathcal{H}}(\rho)(\sigma, \mathrm{grad}_{\mathcal{H}} \mathcal{E}(\rho)) = \int_\Omega \delta \mathcal{E}(\rho)(x) \sigma(x) \mathrm{d}x, \quad \text{for any } \sigma(x) \in T_\rho \mathcal{P}(\Omega). \tag{3}$$

Denote

$$\sigma(x) = -\nabla \cdot \left(\int_\Omega [\delta^2 \mathcal{H}(\rho)]^{-1}(x,y) \nabla \Phi(y) \mathrm{d}y \right). \tag{4}$$

Thus,

$$\Phi(x) = \int_\Omega \Big(-\nabla \cdot ([\delta^2 \mathcal{H}(\rho)]^{-1} \nabla \Big)^{-1}(x,y) \sigma(y) \mathrm{d}y.$$

Notice that

$$\begin{aligned}
\text{LHS of (3)} &= G_{\mathcal{H}}(\rho)(\sigma, \mathrm{grad}_{\mathcal{H}} \mathcal{E}(\rho)) \\
&= \int_\Omega \Big(\int_\Omega \Big(-\nabla \cdot ([\delta^2 \mathcal{H}(\rho)]^{-1} \nabla \Big)^{-1}(x,y) \sigma(y) \mathrm{d}y \Big) \mathrm{grad}_{\mathcal{H}} \mathcal{E}(\rho)(x) \mathrm{d}x \\
&= \int_\Omega \Phi(x) \mathrm{grad}_{\mathcal{H}} \mathcal{E}(\rho)(x) \mathrm{d}x,
\end{aligned}$$

where we applies the definitions of the metric tensor and σ in (4). On the other hand,

$$\begin{aligned}
\text{RHS of (3)} &= \int_\Omega \delta \mathcal{E}(\rho)(x) \sigma(x) \mathrm{d}x \\
&= \int_\Omega \delta \mathcal{E}(\rho)(x) \Big(-\nabla \cdot \left(\int_\Omega [\delta^2 \mathcal{H}(\rho)]^{-1}(x,y) \nabla \Phi(y) \mathrm{d}y \right) \Big) \mathrm{d}x \\
&= \int_\Omega \int_\Omega \Big(\nabla \delta \mathcal{E}(\rho)(x), [\delta^2 \mathcal{H}(\rho)]^{-1}(x,y) \nabla \Phi(y) \Big) \mathrm{d}x \, \mathrm{d}y \\
&= \int_\Omega \Phi(y) \left(-\nabla \cdot \left(\int_\Omega [\delta^2 \mathcal{H}(\rho)]^{-1}(x,y) \nabla \delta \mathcal{E}(\rho)(x) \mathrm{d}x \right) \right) \mathrm{d}y,
\end{aligned}$$

where the second equality is obtained by integration by parts with respect to x and the third equality holds by integration by parts with respect to y. Interchanging x and y in the RHS and comparing the LHS and RHS of (3) for any Φ, we obtain the gradient operator

$$\mathrm{grad}_{\mathcal{H}} \mathcal{E}(\rho)(x) = -\nabla \cdot \left(\int_\Omega [\delta^2 \mathcal{H}(\rho)]^{-1}(x,y) \nabla \Phi(y) \mathrm{d}y \right).$$

Thus, the Riemannian gradient flow in $(\mathcal{P}(\Omega), G_{\mathcal{H}})$ satisfies

$$\partial_t \rho(t,x) = -\mathrm{grad}_{\mathcal{H}} \mathcal{E}(\rho)(t,x) = \nabla \cdot \left(\int_\Omega [\delta^2 \mathcal{H}(\rho)]^{-1}(x,y) \nabla \delta \mathcal{E}(\rho)(y) \mathrm{d}y \right).$$

□

In particular, when $\mathcal{E}(\rho) = \mathcal{H}(\rho) = \int_\Omega f(\rho(x))\mathrm{d}x$, the gradient flow of $\mathcal{H}$ in $(\mathcal{P}(\Omega), G_{\mathcal{H}})$ satisfies the heat equation (1) because

$$\delta\mathcal{H}(\rho)(x) = f'(\rho)(x), \quad \delta^2\mathcal{H}(\rho)(x,y) = f''(\rho)(x)\delta_{x=y}$$

and

$$\begin{aligned}\mathrm{grad}_{\mathcal{H}}\mathcal{H}(\rho)(x) &= -\nabla\cdot\left(\frac{1}{f''(\rho)(x)}\nabla(f'(\rho)(x))\right)\\ &= -\nabla\cdot\left(\frac{1}{f''(\rho)(x)}f''(\rho)(x)\nabla\rho(x)\right) = -\Delta\rho(x).\end{aligned}$$

The gradient flow of $\mathcal{H}(\rho)$ in $(\mathcal{P}(\Omega), G_{\mathcal{H}}(\rho))$ is then given by

$$\partial_t\rho(t,x) = -\mathrm{grad}_{\mathcal{H}}\mathcal{H}(\rho)(t,x) = -(-\Delta\rho(t,x)) = \Delta\rho(t,x),$$

which is the heat equation as demonstrated in Sect. 2.1.

2.3 Divergence and Hessian transport SDE

By taking $\mathcal{E}$ to be the divergence function associated with the entropy $\mathcal{H}$, we derive here a class of generalized Fokker–Planck equations as the gradient flows under the Hessian transport metrics. In addition, we also give the associated Hessian transport stochastic differential equations (HT-SDEs).

To the entropy function $\mathcal{H}_f(\rho) = \int_\Omega f(\rho(x))\mathrm{d}x$, we can associate a corresponding divergence function:

$$\mathcal{D}_f(\rho\|\mu) = \int_\Omega f\left(\frac{\rho(x)}{\mu(x)}\right)\mu(x)\mathrm{d}x.$$

Here $\rho, \mu \in \mathcal{P}(\Omega)$ and $f:\mathbb{R}\to\mathbb{R}$ is a convex function such that $f(1) = 0$. In the literature, $\mathcal{D}_f(\cdot\|\cdot)$ is called the f-divergence function.

Theorem 4 (Hessian transport stochastic differential equations) *Given a reference measure $\mu \in \mathcal{P}(\Omega)$, the gradient flow of $\mathcal{D}_f(\rho\|\mu)$ in $(\mathcal{P}(\Omega), G_{\mathcal{H}_f})$ satisfies*

$$\partial_t\rho(t,x) = \nabla\cdot\left(f''(\rho)(t,x)^{-1}\nabla f'\left(\frac{\rho}{\mu}\right)(t,x)\right). \tag{5}$$

In addition, when $f''(\cdot)$ is homogeneous of degree $-\gamma$, i.e.,

$$f''(t) = f''(1)t^{-\gamma}.$$

Equation (5) can be simplified

$$\partial_t\rho(t,x) = \nabla\cdot\left(\mu(x)^\gamma\nabla\left(\frac{\rho(t,x)}{\mu(x)}\right)\right), \tag{6}$$

and it is the Kolmogorov forward equation of the stochastic differential equation

$$\mathrm{d}X_t = \gamma\mu(X_t)^{\gamma-2}\nabla\mu(X_t)\mathrm{d}t + \sqrt{2\mu(X_t)^{\gamma-1}}\mathrm{d}B_t, \tag{7}$$

where B_t is the standard Brownian motion in Ω.

Proof We first derive the gradient flow in $(\mathcal{P}, G_{\mathcal{H}_f})$. Notice that

$$\delta \mathcal{D}_f(\rho\|\mu)(x) := \frac{\delta}{\delta\rho(x)}\mathcal{D}_f(\rho\|\mu) = f'\left(\frac{\rho}{\mu}\right)(x), \tag{8}$$

and that the transport metric is

$$G_{\mathcal{H}_f}(\rho) = \Big(-\nabla\cdot(f''(\rho)^{-1}\nabla)\Big)^{-1}$$

from Definition 1 and $\mathcal{H}_f(\rho) = \int_\Omega f(\rho(x))\mathrm{d}x$. Thus, the gradient flow of $\mathcal{D}_f(\rho\|\mu)$ in $(\mathcal{P}(\Omega), G_{\mathcal{H}_f})$ satisfies

$$\begin{aligned}\partial_t\rho(t,x) =& -G_{\mathcal{H}_f}(\rho)^{-1}\delta\mathcal{H}_f(\rho\|\mu) = -\Big([-\nabla\cdot(f''(\rho)^{-1}\nabla)]^{-1}\Big)^{-1}f'\left(\frac{\rho}{\mu}\right)\\ =&\nabla\cdot\left(f''(\rho)^{-1}\nabla f'\left(\frac{\rho}{\mu}\right)\right) = \nabla\cdot\left(f''(\rho)^{-1}f''\left(\frac{\rho}{\mu}\right)\nabla\left(\frac{\rho}{\mu}\right)\right).\end{aligned}$$

Notice $f''\left(\frac{\rho}{\mu}\right) = \mu^\gamma f''(\rho)$ with $\gamma\in\mathbb{R}$ due to the homogeneity assumption. Then the gradient flow (5) can be simplified to

$$\partial_t\rho(t,x) = \nabla\cdot\left(\mu(x)^\gamma\nabla\left(\frac{\rho(t,x)}{\mu(x)}\right)\right).$$

This equation is a Kolmogorov forward equation (see for example [31,33]) $\partial_t\rho = L^*\rho$ with the forward operator given by $L^* = \nabla\cdot(\mu^\gamma\cdot\nabla(\frac{1}{\mu}\cdot))$. In order to obtain the corresponding stochastic differential equation, we first write down its adjoint equation, the Kolmogorov backward equation $\partial_t u = Lu$ for functions on Ω with $L = \left(\frac{1}{\mu}\cdot\right)\nabla\cdot(\mu^\gamma\cdot\nabla)$. Expanding L explicitly gives

$$\partial_t u = \left(\frac{1}{\mu}\cdot\right)\nabla\cdot(\mu^\gamma\cdot\nabla u) = \gamma(\mu(x)^{\gamma-2}\nabla\mu(x))\cdot\nabla u(x) + \mu(x)^{\gamma-1}\Delta u(x).$$

By identifying the coefficient $\gamma(\mu(x)^{\gamma-2}\nabla\mu(x))$ before the drift term $\nabla u(x)$ and the coefficient $\mu(x)^{\gamma-1}$ before the diffusion term $\Delta u(x)$, one notices that L is the generator of the stochastic differential equation

$$\mathrm{d}X_t = \gamma\mu(X_t)^{\gamma-2}\nabla\mu(X_t)\mathrm{d}t + \sqrt{2\mu(X_t)^{\gamma-1}}\mathrm{d}B_t,$$

which finishes the proof. □

Following the gradient flow relation (6), the reference measure μ is the invariant measure for the HT-SDE (7). We next derive a generalized de Bruijn identity that characterizes the dissipation of the divergence function along the gradient flow.

Corollary 5 (Hessian transport de Bruijn identity) *Suppose $\rho(t,x)$ satisfies (5), then*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{D}_f(\rho(t,\cdot)\|\mu) = -I_f(\rho(t,\cdot)\|\mu),$$

where the f-relative Fisher information functional $I_f(\rho\|\mu)$ is given by

$$I_f(\rho\|\mu) = \int_\Omega \left\|\nabla f'\left(\frac{\rho}{\mu}\right)\right\|^2 f''(\rho)^{-1}\mathrm{d}x. \tag{9}$$

Proof The proof follows the dissipation of energy along gradient flows in the probability space. Notice that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{D}_f(\rho(t,\cdot)\|\mu) &= -\int_\Omega \delta\mathcal{D}_f(\rho(t,\cdot)\|\mu)\partial_t\rho\mathrm{d}x\\ &= \int_\Omega \delta\mathcal{D}_f(\rho(t,\cdot)\|\mu)\nabla\cdot(f''(\rho)^{-1}\nabla\delta\mathcal{D}_f(\rho(t,\cdot)\|\mu))\mathrm{d}x\\ &= -\int_\Omega (\nabla\delta\mathcal{D}_f(\rho(t,\cdot)\|\mu),\nabla\delta\mathcal{D}_f(\rho(t,\cdot)\|\mu))f''(\rho)^{-1}\mathrm{d}x\\ &= -\int_\Omega \left\|\nabla f'(\frac{\rho}{\mu})\right\|^2 f''(\rho)^{-1}\mathrm{d}x,\end{aligned}$$

where the last equality holds by formula (11). □

Consider the case $f(\rho) = \rho\log\rho$. The f-entropy is the negative Boltzmann–Shannon entropy $\int_\Omega \rho\log\rho\mathrm{d}x$, and the f-divergence is the usual relative entropy

$$\mathcal{D}_f(\rho\|\mu) = \int_\Omega \frac{\rho(x)}{\mu(x)}\log\frac{\rho(x)}{\mu(x)}\mu(x)\mathrm{d}x = \int_\Omega \rho(x)\log\frac{\rho(x)}{\mu(x)}\mathrm{d}x.$$

In this case, $\delta\mathcal{D}_f(\rho\|\mu) = \log\frac{\rho}{\mu}+1$, and thus,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{D}_f(\rho(t,\cdot)\|\mu) = -\int_\Omega \left\|\nabla\log\frac{\rho(t,x)}{\mu(x)}\right\|^2\rho(t,x)\mathrm{d}x.$$

Here we recover the classical result that the dissipation of the relative entropy is equal to the negative relative Fisher information functional. Our result extends this relation to any f-divergence functions. For this reason, I_f in (9) is called the *f-relative Fisher information functional*.

Remark 2 Here we demonstrate the relations between our approaches and the ones in literature [4,8,38]. The generalized de Bruijn identity and f-relative Fisher information functional (9) recovers exactly the ones in [38] when μ is a uniform measure. They differ from [38] when μ is a non-uniform reference measure. Our approach always generalizes the entropy dissipation as the geometric dissipation as gradient flows of the probability manifold $(\mathcal{P}(\Omega), G_{\mathcal{H}})$, while [38] studies the dissipation of relative entropy among two heat flows for two variables in the divergence function. Our approach is also different from the one in [4]. We derive a class of Fokker–Planck equation (6) with parameter γ, while [4] studies the Fokker–Planck equation(6) with $\gamma = 1$. Lastly, our approach differs from [8]. While [8] proposes a *reference-measure-dependent* metric under which the Fokker–Planck equation (6) with $\gamma = 1$ is the gradient flow of the Renyi entropy, our approach introduces a class of *reference-measure-independent* metrics. They only depend on the L^2 Hessian operator of the convex entropy function and allow us to derive a new class of Fokker–Planck equations (6).

2.4 Examples

Below we consider a few special but important cases of f-divergences and present the f-divergence induced HT-SDE in Theorem 4.

Example 1 (KL divergence HT-SDE)

$$f(\rho) = \rho \log \rho, \quad f'(\rho) = \log \rho + 1, \quad f''(\rho) = 1/\rho, \quad \gamma = 1.$$

The gradient flow, the HT-SDE, and the relative Fisher information functional are, respectively,

$$\begin{aligned}
\partial_t \rho(t,x) &= \nabla \cdot \left(\mu(x) \nabla \left(\frac{\rho(t,x)}{\mu(x)} \right) \right), \\
\mathrm{d}X_t &= \mu(X_t)^{-1} \nabla \mu(X_t) \mathrm{d}t + \sqrt{2} \mathrm{d}B_t, \\
\mathcal{I}_f(\rho \| \mu) &= \int_\Omega \left\| \nabla \log \frac{\rho(x)}{\mu(x)} \right\|^2 \rho(x) \mathrm{d}x.
\end{aligned}$$

Example 2 (Reverse KL divergence HT-SDE)

$$f(\rho) = -\log \rho, \quad f'(\rho) = -1/\rho, \quad f''(\rho) = 1/\rho^2, \quad \gamma = 2.$$

The gradient flow, the HT-SDE, and the relative Fisher information functional are, respectively,

$$\begin{aligned}
\partial_t \rho(t,x) &= \nabla \cdot \left(\mu(x)^2 \nabla \left(\frac{\rho(t,x)}{\mu(x)} \right) \right), \\
\mathrm{d}X_t &= 2 \nabla \mu(X_t) \mathrm{d}t + \sqrt{2\mu(X_t)} \mathrm{d}B_t, \\
\mathcal{I}_f(\rho \| \mu) &= \int_\Omega \left\| \nabla \Big(\frac{\rho(x)}{\mu(x)} \Big)^{-1} \right\|^2 \rho(x)^2 \mathrm{d}x.
\end{aligned}$$

Example 3 (α-divergence HT-SDE)

$$f(\rho) = \frac{4}{1-\alpha^2}(1 - \rho^{\frac{1+\alpha}{2}}), \quad f'(\rho) = \frac{2}{\alpha - 1} \rho^{\frac{\alpha-1}{2}}, \quad f''(\rho) = \rho^{\frac{\alpha-3}{2}}, \quad \gamma = \frac{3-\alpha}{2}.$$

The gradient flow, the HT-SDE, and the relative Fisher information functional are, respectively,

$$\begin{aligned}
\partial_t \rho(t,x) &= \nabla \cdot \left(\mu(x)^{(3-\alpha)/2} \nabla \left(\frac{\rho(t,x)}{\mu(x)} \right) \right), \\
\mathrm{d}X_t &= \frac{3-\alpha}{2} \mu(X_t)^{\frac{-1-\alpha}{2}} \nabla \mu(X_t) \mathrm{d}t + \sqrt{2\mu(X_t)^{\frac{1-\alpha}{2}}} \mathrm{d}B_t, \\
\mathcal{I}_f(\rho \| \mu) &= \left(\frac{2}{\alpha - 1} \right)^2 \int_\Omega \left\| \nabla \left(\frac{\rho(x)}{\mu(x)} \right)^{\frac{\alpha-1}{2}} \right\|^2 \rho(x)^{\frac{3-\alpha}{2}} \mathrm{d}x.
\end{aligned}$$

Example 4 (Hellinger distance HT-SDE)

$$f(\rho) = (\sqrt{\rho} - 1)^2,$$

and it is a special case of α-divergence with $\alpha = 0$ and hence $\gamma = 3/2$. The gradient flow, the HT-SDE, and the relative Fisher information functional are, respectively,

$$\begin{aligned}
\partial_t \rho(t,x) &= \nabla \cdot \left(\mu(x)^{3/2} \nabla \left(\frac{\rho(t,x)}{\mu(x)} \right) \right), \\
\mathrm{d}X_t &= \frac{3}{2} \mu(X_t)^{-\frac{1}{2}} \nabla \mu(X_t) \mathrm{d}t + \sqrt{2\mu(X_t)^{\frac{1}{2}}} \mathrm{d}B_t, \\
\mathcal{I}_f(\rho \| \mu) &= 4 \int_\Omega \left\| \nabla \left(\frac{\rho(x)}{\mu(x)} \right)^{-1/2} \right\|^2 \rho(x)^{\frac{3}{2}} \mathrm{d}x.
\end{aligned}$$

Example 5 (Pearson divergence HT-SDE)

$$f(\rho) = (\rho - 1)^2,$$

and it is a special case of α-divergence with $\alpha = 3$ and hence $\gamma = 0$. The gradient flow, the HT-SDE, and the relative Fisher information functional are, respectively,

$$\begin{aligned}
\partial_t \rho(t,x) &= \nabla \cdot \left(\nabla \left(\frac{\rho(t,x)}{\mu(x)} \right) \right), \\
\mathrm{d}X_t &= \sqrt{2\mu(X_t)^{-1}}\mathrm{d}B_t, \\
\mathcal{I}_f(\rho\|\mu) &= \int_\Omega \left\| \nabla \Big(\frac{\rho(x)}{\mu(x)} \Big) \right\|^2 \mathrm{d}x.
\end{aligned}$$

Example 6 (Jensen–Shannon divergence HT-SDE)

$$f(\rho) = -(\rho+1)\log\frac{1+\rho}{2} + \rho\log\rho, \quad f'(\rho) = -\log\frac{1+\rho}{2} + \log\rho, \quad f''(\rho) = \frac{1}{\rho(1+\rho)}$$

The above discussion does not apply since $f''(\rho)$ is not homogeneous in ρ. However, one can still obtain a gradient flow PDE

$$\partial_t \rho(t,x) = \nabla \cdot \left(\frac{(1+\rho(t,x))\mu(x)^2}{(\rho(t,x)+\mu(x))} \nabla \left(\frac{\rho(t,x)}{\mu(x)} \right) \right),$$

and the f−relative Fisher information functional

$$\mathcal{I}_f(\rho\|\mu) = \int_\Omega \left\| \nabla \log(\frac{2\rho}{\rho+\mu}) \right\|^2 \rho(1+\rho)\mathrm{d}x.$$

Remark 3 We note that the proposed metrics and the gradient flows could be useful in machine learning. Matching a target distribution given some reference distribution is a routine task. Here we derive a class of stochastic differential equations depending on the reference measure μ. Many numerical methods in MCMC focus on preconditioning the drift gradient operator of Langevin dynamics, while our method modifies both the drift and diffusion coefficient terms by the reference measure. For example, the Pearson transport SDE has no drift term. We notice that this approach brings more challenge in numerical computation with new convergence rate to be analyzed. The related numerical issues will be studied in future work.

3 Hessian transport distance

In this section, we introduce the Hessian transport distance following the metric tensor proposed in Definition 1 and derive the geodesic equations.

Definition 6 (*Hessian transport distance*) Given a convex entropy function $\mathcal{H}$, the distance function $W_{\mathcal{H}} \colon \mathcal{P}(\Omega) \times \mathcal{P}(\Omega) \to \mathbb{R}$ between two densities ρ^0 and ρ^1 is

$$W_{\mathcal{H}}(\rho^0, \rho^1) = \left(\inf_{v,\rho} \int_0^1 \int_\Omega \int_\Omega \big(v(t,x), [\delta^2 \mathcal{H}(\rho)]^{-1}(x,y) v(t,y) \big) \, \mathrm{d}x\mathrm{d}y\mathrm{d}t \right)^{1/2}, \tag{10}$$

such that the infimum is taken among all density path $\rho \colon [0,1] \times \Omega \to \mathbb{R}$ and vector field $v \colon [0,1] \times \Omega \to T\Omega$, satisfying

$$\partial_t \rho(t,x) + \nabla \cdot \left(\int_\Omega [\delta^2 \mathcal{H}(\rho)]^{-1}(x,y) v(t,y) \mathrm{d}y \right) = 0,$$

with $\rho(0,x) = \rho^0(x)$ and $\rho(1,x) = \rho^1(x)$.

We first illustrate that $W_{\mathcal{H}}$ is a Riemannian distance. For a fixed density $\rho(x)$, the *Hodge decomposition* for a vector function $v(x)$ is

$$v(x) = \nabla \Phi(x) + u(x),$$

where $\Phi \colon \Omega \to \mathbb{R}$ and $u \colon \Omega \to T_x\Omega$ is the divergence free vector for ρ in the following sense

$$\nabla \cdot \left(\int_\Omega [\delta^2 \mathcal{H}(\rho)]^{-1}(x,y) u(y) \mathrm{d}y \right)(x) = 0. \tag{11}$$

Thus,

$$\begin{aligned}
&\int_\Omega \int_\Omega (v(x), [\delta^2 \mathcal{H}(\rho)]^{-1}(x,y) v(y)) \mathrm{d}x\, \mathrm{d}y \\
&= \int_\Omega \int_\Omega (\nabla \Phi(x), [\delta^2 \mathcal{H}(\rho)]^{-1}(x,y) \nabla \Phi(y)) \mathrm{d}x\, \mathrm{d}y + \int_\Omega \int_\Omega (u(x), [\delta^2 \mathcal{H}(\rho)]^{-1}(x,y) u(y)) \mathrm{d}x\, \mathrm{d}y \\
&\quad + 2 \int_\Omega \int_\Omega (u(x), [\delta^2 \mathcal{H}(\rho)]^{-1}(x,y) \nabla \Phi(y)) \mathrm{d}x\, \mathrm{d}y \\
&= \int_\Omega \int_\Omega (\nabla \Phi(x), [\delta^2 \mathcal{H}(\rho)]^{-1}(x,y) \nabla \Phi(y)) \mathrm{d}x\, \mathrm{d}y + \int_\Omega \int_\Omega (u(x), [\delta^2 \mathcal{H}(\rho)]^{-1}(x,y) u(y)) \mathrm{d}x\, \mathrm{d}y \\
&\geq \int_\Omega \int_\Omega (\nabla \Phi(x), [\delta^2 \mathcal{H}(\rho)]^{-1}(x,y) \nabla \Phi(y)) \mathrm{d}x\, \mathrm{d}y,
\end{aligned}$$

where the second equality uses the divergence free relation (11). Thus, the minimization problem (10) is same as the one over variable $(\rho(t,x), \Phi(t,x))$, where $\Phi(t,x)$ is the first part of the Hodge decomposition of $v(t,x)$. By denoting $\partial_t \rho(t,x) = L_{\mathcal{H},\rho} \Phi(t,x)$ with $L_{\mathcal{H},\rho}$ defined in (2), we arrive at

$$\begin{aligned}
G_{\mathcal{H}}(\rho)(\partial_t \rho, \partial_t \rho) &= \int_\Omega \int_\Omega \left(\partial_t \rho(t,x), L^{-1}_{\mathcal{H},\rho} \partial_t \rho(t,y) \right) \mathrm{d}x\, \mathrm{d}y \\
&= \int_\Omega \int_\Omega \left(L_{\mathcal{H},\rho} \Phi(t,x), L^{-1}_{\mathcal{H},\rho} L_{\mathcal{H},\rho} \Phi(t,y) \right) \mathrm{d}x\, \mathrm{d}y \\
&= \int_\Omega \int_\Omega (\Phi(t,x), L_{\mathcal{H},\rho} \Phi(t,y)) \mathrm{d}x\, \mathrm{d}y.
\end{aligned}$$

Thus, the distance function defined in (10) can be formulated as

$$\left(W_{\mathcal{H}}(\rho^0, \rho^1) \right)^2 = \inf_{\rho \colon [0,1] \to \mathcal{P}(\Omega)} \left\{ \int_0^1 G_{\mathcal{H}}(\rho)(\partial_t \rho, \partial_t \rho) \mathrm{d}t \colon \rho^0,\ \rho^1 \text{ fixed} \right\}.$$

This is exactly the geometric action functional in $(\mathcal{P}(\Omega), G_{\mathcal{H}})$, and therefore, $W_{\mathcal{H}}$ is a Riemannian distance on $\mathcal{P}(\Omega)$.

Next, we prove that the distance is well defined and derive the formulations of geodesics equations.

Theorem 7 (Hessian transport geodesic (HT-geodesic)) *The Hessian transport distance is well defined in* $\mathcal{P}(\Omega)$*, i.e.,* $W_{\mathcal{H}}(\rho^0, \rho^1) < +\infty$*. The geodesic equation is*

$$\begin{cases} \partial_t \rho(t,x) + \nabla \cdot \left(\int_\Omega [\delta^2 \mathcal{H}(\rho)]^{-1}(x,y) \nabla \Phi(t,y) \mathrm{d}y \right) = 0 \\ \partial_t \Phi(t,x) + \frac{1}{2} \delta_\rho \int_\Omega \int_\Omega \nabla \Phi(t,x), [\delta^2 \mathcal{H}(\rho)]^{-1}(x,y) \nabla \Phi(t,y) \mathrm{d}x \, \mathrm{d}y = 0. \end{cases} \tag{12}$$

If $\mathcal{H}(\rho) = \mathcal{H}_f(\rho) = \int_\Omega f(\rho)(x) \mathrm{d}x$*, the geodesic equation simplifies to*

$$\begin{cases} \partial_t \rho + \nabla \cdot (f''(\rho)^{-1} \nabla \Phi) = 0 \\ \partial_t \Phi - \frac{1}{2} (\nabla \Phi, \nabla \Phi) \frac{f'''(\rho)}{f''(\rho)^2} = 0. \end{cases}$$

Proof We first prove that the distance function is well defined. First, by denoting $m(t,x) = [\delta^2 \mathcal{H}(\rho)]^{-1} v(t,x)$, we can rewrite the minimization problem (10) as

$$W_{\mathcal{H}}(\rho^0, \rho^1)^2 = \inf_{\rho, m} \int_0^1 \int_\Omega \int_\Omega (m(t,x), \delta^2 \mathcal{H}(\rho)(x,y) m(t,y)) \mathrm{d}x \, \mathrm{d}y \mathrm{d}t \tag{13}$$

along with the constraints

$$\partial_t \rho(t,x) + \nabla \cdot m(t,x) = 0,$$

with fixed initial and terminal densities ρ^0, ρ^1. We show that there exists a feasible path for any $\rho^0, \rho^1 \in \mathcal{P}(\Omega)$. Notice that $\min\{\min_{x\in\Omega} \rho^0, \min_{x\in\Omega} \rho^1\} > 0$. We construct a path $\tilde{\rho}(t,x) = (1-t)\rho^0 + t\rho^1$, where $t \in [0,1]$. Thus, $\tilde{\rho}(t,x) \in \mathcal{P}(\Omega)$ and $\min_{t,x\in\Omega} \tilde{\rho}(t,x) > 0$. Construct a feasible flux function $\tilde{m}(t,x) = \nabla \tilde{\Phi}(t,x)$, with $\tilde{\Phi}(t,x) = -\Delta^{-1} \partial_t \tilde{\rho}(t,x) \in C^\infty(\Omega)$. Thus,

$$\int_0^1 \int_\Omega \int_\Omega (\nabla \tilde{\Phi}(t,x), [\delta^2 \mathcal{H}(\rho)]^{-1}(x,y) \nabla \tilde{\Phi}(t,y)) \mathrm{d}x \, \mathrm{d}y \mathrm{d}t < \infty,$$

Then $(\tilde{\rho}(t,x), \tilde{m} = \nabla \tilde{\Phi}(t,x))$ is a feasible path for minimization problem (13) for any ρ^0, $\rho^1 \in \mathcal{P}(\Omega)$.

We next derive the geodesic equation within $\mathcal{P}(\Omega)$. The first step is to write down the Lagrangian multiplier $\Phi(t,x)$ of the continuity equation $\partial_t \rho + \nabla \cdot m = 0$

$$\begin{aligned} \mathcal{L}(m, \rho, \Phi) = & \frac{1}{2} \int_0^1 \int_\Omega \int_\Omega (m(t,x), \delta^2 \mathcal{H}(\rho)(x,y) m(t,y)) \mathrm{d}x \, \mathrm{d}y \mathrm{d}t \\ & + \int_0^1 \int_\Omega \Phi(x) (\partial_t \rho(t,x) + \nabla \cdot m(t,x)) \mathrm{d}x. \end{aligned}$$

At $\rho \in \mathcal{P}(\Omega)$, $\delta_\rho \mathcal{L} = 0$, $\delta_m \mathcal{L} = 0$, and $\delta_\Phi \mathcal{L} = 0$, we know that the minimizer satisfies

$$\begin{cases} \int_\Omega \delta^2 \mathcal{H}(\rho)(x,y) m(t,y) \mathrm{d}y = \nabla \Phi(t,x), \\ \frac{1}{2} \delta_\rho \int_\Omega \int_\Omega (m(t,x), \delta^2 \mathcal{H}(\rho)(x,y) m(t,y)) \mathrm{d}x \, \mathrm{d}y = \partial_t \Phi(t,x), \\ \partial_t \rho(t,x) + \nabla \cdot m(t,x) = 0. \end{cases}$$

Finally, by denoting $m(t,x) = \int_\Omega \delta^2\mathcal{H}(\rho)^{-1}(x,y)\nabla\Phi(t,y)\mathrm{d}y$ and using the fact

$$\delta_\rho[\delta^2\mathcal{H}(\rho)] = -[\delta^2\mathcal{H}(\rho)]^{-1}\delta_\rho[\delta^2\mathcal{H}(\rho)^{-1}][\delta^2\mathcal{H}(\rho)]^{-1},$$

we derive the geodesic equation (12). □

Remark 4 Consider the special case that $\mathcal{H}(\rho)$ is the f-entropy with $f''(\cdot)$ homogeneous of degree $-\gamma$. The objective function

$$\int_0^1\int_\Omega\int_\Omega (m(t,x), \delta^2\mathcal{H}(\rho)(x,y)m(t,y))\mathrm{d}x\,\mathrm{d}y\mathrm{d}t = \int_0^1\int_\Omega \frac{\|m(t,x)\|^2}{\rho(t,x)^\gamma}\mathrm{d}x\mathrm{d}t$$

is convex jointly in (m,ρ) if and only if $\gamma \in [0,1]$. As two special cases, the proposed minimal flux minimization (13) for the optimal KL ($\gamma = 1$) and the Pearson ($\gamma = 0$) Hessian transport is convex.

3.1 Examples

Here we list the geodesic equation for the f-divergence functions.

Example 7 (KL divergence HT-geodesic)

$$f(\rho) = \rho\log\rho, \quad f'(\rho) = \log\rho + 1, \quad f''(\rho) = \frac{1}{\rho}, \quad f'''(\rho) = -\frac{1}{\rho^2}.$$

Thus, $f'''(\rho)/f''(\rho)^2 = -1$ and the geodesic equation is

$$\begin{cases} \partial_t\rho + \nabla\cdot(\rho\nabla\Phi) = 0, \\ \partial_t\Phi + \dfrac{1}{2}(\nabla\Phi,\nabla\Phi) = 0. \end{cases}$$

This is the classical geodesic equation in Wasserstein geometry, including both the continuity equation and the Hamilton–Jacobi equation.

Example 8 (Reverse KL divergence HT-geodesic)

$$f(\rho) = -\log\rho, \quad f'(\rho) = -\frac{1}{\rho}, \quad f''(\rho) = \frac{1}{\rho^2}, \quad f'''(\rho) = -\frac{2}{\rho^3}.$$

Thus, $f'''(\rho)/f''(\rho)^2 = -2\rho$, then the geodesic equation is

$$\begin{cases} \partial_t\rho + \nabla\cdot(\rho^2\nabla\Phi) = 0 \\ \partial_t\Phi + (\nabla\Phi,\nabla\Phi)\rho = 0. \end{cases}$$

Example 9 (α-divergence HT-geodesic)

$$f(\rho) = \frac{4}{1-\alpha^2}(1-\rho^{\frac{1+\alpha}{2}}), \quad f'(\rho) = \frac{2}{\alpha-1}\rho^{\frac{\alpha-1}{2}}, \quad f''(\rho) = \rho^{\frac{\alpha-3}{2}}, \quad f'''(\rho) = \frac{\alpha-3}{2}\rho^{\frac{\alpha-5}{2}}.$$

Thus, $f'''(\rho)/f''(\rho)^2 = \frac{\alpha-3}{2}\rho^{\frac{1-\alpha}{2}}$, then the geodesic equation is

$$\begin{cases} \partial_t\rho + \nabla\cdot(\rho^{\frac{3-\alpha}{2}}\nabla\Phi) = 0, \\ \partial_t\Phi + \dfrac{3-\alpha}{2}(\nabla\Phi,\nabla\Phi)\rho^{\frac{1-\alpha}{2}} = 0. \end{cases}$$

Example 10 (Hellinger distance HT-geodesic)

$$f(\rho) = (\sqrt{\rho} - 1)^2,$$

and it is a special case of α-divergence with $\alpha = 0$. Hence, the geodesic equation takes the form

$$\begin{cases} \partial_t \rho + \nabla \cdot (\rho^{\frac{3}{2}} \nabla \Phi) = 0, \\ \partial_t \Phi + \dfrac{3}{2} (\nabla \Phi, \nabla \Phi) \rho^{\frac{1}{2}} = 0. \end{cases}$$

Example 11 (Pearson divergence HT-geodesic)

$$f(\rho) = (\rho - 1)^2,$$

and it is a special case of α-divergence with $\alpha = 3$. Hence, the geodesic equation is

$$\begin{cases} \partial_t \rho + \Delta \Phi = 0, \\ \partial_t \Phi = 0. \end{cases}$$

This geodesic equation satisfies $\frac{\partial^2}{\partial t^2} \rho(t, x) = 0$, which implies that

$$\rho(t, x) = (1 - t)\rho^0(x) + t\rho^1(x).$$

It states that the geodesic equation in optimal Hellinger distance transport metric is a straight line in the probability space.

Example 12 (Jensen–Shannon divergence HT-geodesic)

$$f(\rho) = -(\rho + 1) \log \frac{1 + \rho}{2} + \rho \log \rho, \quad f''(\rho) = \frac{1}{\rho(1 + \rho)}, \quad f'''(\rho) = -\frac{2\rho + 1}{(\rho + 1)^2 \rho^2}$$

Thus, $f'''(\rho)/f''(\rho)^2 = -(2\rho + 1)$. Hence, the geodesic equation is

$$\begin{cases} \partial_t \rho + \nabla \cdot (\rho(1 + \rho) \nabla \Phi) = 0, \\ \partial_t \Phi + \dfrac{1}{2} (\nabla \Phi, \nabla \Phi)(2\rho + 1) = 0. \end{cases}$$

Remark 5 The derivation of geodesics equation works also for the Stein metric [24]. In particular, when the kernel is equal to the delta measure, the reverse KL divergence HT-geodesic satisfies exactly the geodesic equation for the Stein metric. In addition, there are interesting directions in machine learning where geometric structures are applied in statistical manifold to accelerate the optimization method in MCMC. Our results provide a general class of metrics and gradient operators in probability space, which could be helpful in this direction.

4 Numerical examples

In this section, we demonstrate the properties of the newly derived equations with several examples. Since the gradient flow equations are linear, the flow dynamics are governed mostly by the spectrum of the Kolmogorov forward and backward operators. Here, we consider the simple setting of Ω equal to the unit interval $[0, 1]$ with periodic boundary condition. The PDEs are numerically discretized with a finite element method with a uniform discretization.

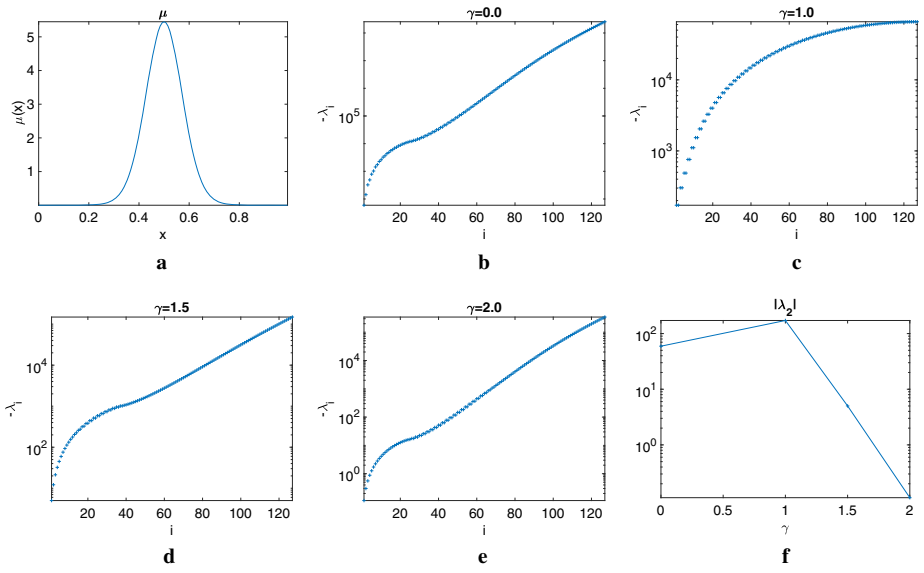

Fig. 2 Unimodal case. **a** Reference measure $\mu(x)$. **b**–**e** The spectrum (part close to the zero) in the log-scale for the gradient flow PDEs with $\gamma = 0, 1, 1.5, 2$. **f** The smallest nonzero eigenvalue λ_2 for the gradient flow PDEs with different choices of γ

We consider two simple examples in this setting. In the first example, the reference measure $\mu(x)$ is a unimodal distribution (shown in Fig. 2a). Figure 2b–e plots the bottom part of the spectrum of the gradient flow PDEs for $\gamma = 0, 1, 1.5, 2$. These γ values correspond to the Pearson, KL, Hellinger, and reverse KL divergence. We also summarize the magnitude of the smallest nonzero eigenvalue λ_2 for these choices of γ in Fig. 2f. For these linear gradient flow PDEs, $|\lambda_2|$ controls the convergence rate to the reference measure $\mu(x)$ for a generic initial condition $\rho(t = 0)$. The plot suggests that among various choices of γ, the standard Fokker–Planck equation ($\gamma = 1$) has the largest $|\lambda_2|$ and hence the fastest convergence rate.

In the second example, the reference measure $\mu(x)$ is a multimodal distribution (shown in Fig. 3a). Figure 3b–e plots the bottom part of the spectrum of the gradient flow PDEs for $\gamma = 0, 1, 1.5, 2$. We again summarize the magnitude of the smallest nonzero eigenvalue λ_2 for these choices of γ in Fig. 3f. It is a well-known fact that, for the multimodal distribution, there exists a gap between the first few lowest eigenvalues (the number of which is equal to the number of modes) and the rest of the spectrum, due to the metastable states. For the standard Fokker–Planck equation ($\gamma = 1$), this gap is shown clearly in Fig. 3d. From the plots in Fig. 3, one can make two observations concerning the gradient flow PDEs introduced in Sect. 2. The first is that, although the gap seems to persist for γ greater than 1, it decreases when γ increases from 1. For example in Fig. 3, the gap is significantly smaller at $\gamma = 0$. The second observation is that, in contrast to the unimodal case, $|\lambda_2|$ for the multimodal case is no longer obtained at $\gamma = 1$. In fact, $|\lambda_2|$ increases quite rapidly as γ decreases from 1, thus implying that the gradient flow PDE of the Pearson ($\gamma = 0$) divergence converges at a faster rate compared to the one of the standard Fokker–Planck equations ($\gamma = 1$).

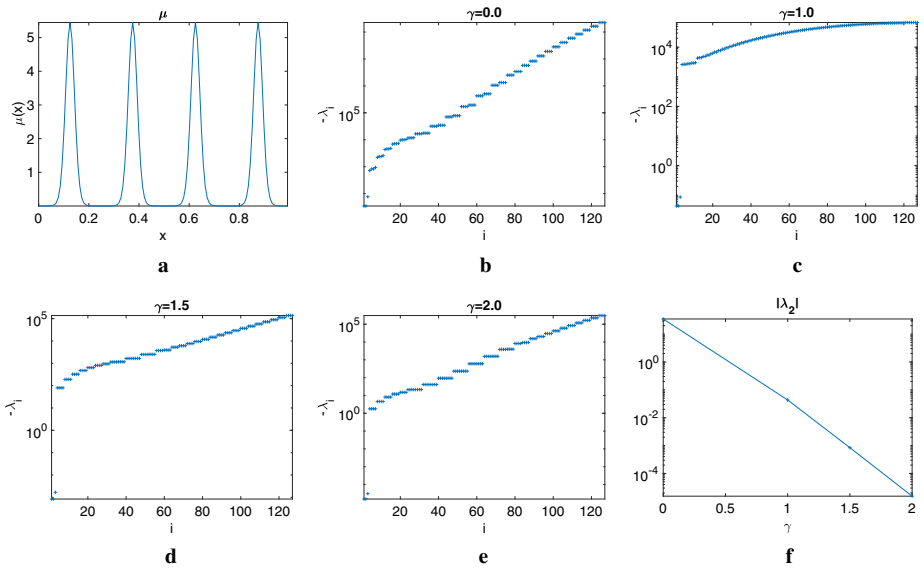

Fig. 3 Multimodal case. **a** Reference measure $\mu(x)$. **b**–**e** The spectrum (part close to the zero) in the log-scale for the gradient flow PDEs with $\gamma = 0, 1, 1.5, 2$. **f** The smallest nonzero eigenvalue λ_2 for the gradient flow PDEs with different choices of γ

5 Discussions

In this paper, we propose a family of Riemannian metrics in the probability space, named Hessian transport metric. We demonstrate that the heat flow is the gradient flow of several energy functions under the HT-metrics. Following this, we further introduce the gradient flows of divergence functions in the HT-metrics, which can be interpreted as Kolmogorov forward equations of the associated HT-SDEs.

Our study is the first step to bridge Hessian geometry, Wasserstein geometry, and divergence functions. Several fundamental questions arise. Firstly, there are many entropies and divergence functions in information theory [1]. Besides the α divergences and α entropy, which type of entropy's HT-gradient flows of divergence functions is the probability transition equation of HT-SDEs? Secondly, in machine learning applications, especially the parametric statistics, our new geometry structure leads to a new class of metrics in parameter spaces/statistical manifold. We expect some of these metrics will help the training process [18,22]. Lastly and most importantly, we introduce a new class of stochastic differential equations, named HT-SDEs. In the future, we shall study the convergence rate of these HT-SDEs and apply them for related machine learning applications.

Author details
[1]Department of Mathematics, University of California, Los Angeles, USA, [2]Department of Mathematics, Stanford University and Facebook AI Research, Stanford, USA.

Received: 20 May 2019 Accepted: 12 October 2019 Published online: 28 October 2019

References

1. Amari, S.: Information Geometry and Its Applications, 1st edn. Springer, New York (2016)
2. Amari, S., Cichocki, A.: Information geometry of divergence functions. Bull. Polish Acad. Sci. Tech. Sci. **58**(1), 183–195 (2010)

3. Amari, S., Karakida, R., Oizumi, M.: Information Geometry Connecting Wasserstein Distance and Kullback–Leibler Divergence via the Entropy-Relaxed Transportation Problem. arXiv:1709.10219 [cs, math] (2017)
4. Arnold, A., Markowich, P., Toscani, G., Unterreiter, A.: On convex Sobolev inequalities and the rate of convergence to equilibrium for Fokker–Planck type equations. Commun. Part. Differ. Equ. **26**(1–2), 43–100 (2001)
5. Ay, N., Jost, J., Lê, H.V., Schwachhöfer, L.: Information Geometry, vol. 64. Springer, Cham (2017)
6. Bauer, M., Joshi, S., Modin, K.: Diffeomorphic density matching by optimal information transport. SIAM J. Imaging Sci. **8**(3), 1718–1751 (2015)
7. Bauer, M., Modin, K.: Semi-invariant Riemannian Metrics in Hydrodynamics. arXiv:1810.03424 [math] (2018)
8. Cao, Y., Lu, J., Lu, Y.: Exponential Decay of Renyi Divergence Under Fokker–Planck Equations. arXiv:1805.06554 [math] (2018)
9. Carrillo, J.A., Lisini, S., Savare, G., Slepcev, D.: Nonlinear mobility continuity equations and generalized displacement convexity. J. Funct. Anal. **258**(4), 1273–1309 (2010)
10. Chizat, L., Peyré, G., Schmitzer, B., Vialard, F.-X.: An interpolating distance between optimal transport and Fisher–Rao metrics. Found. Comput. Math. **18**(1), 1–44 (2018)
11. Chow, S.-N., Li, W., Zhou, H.: Entropy dissipation of Fokker-Planck equations on graphs. Discrete Contin. Dyn. Syst. **38**(10), 4929–4950 (2018)
12. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley Series in Telecommunications. Wiley, New York (1991)
13. Csiszár, I., Shields, P.C.: Information theory and statistics: a tutorial. Found. Trends Commun. Inf. Theory **1**(4), 417–528 (2004)
14. Dolbeault, J., Nazaret, B., Savaré, G.: A new class of transport distances between measures. Calc. Var. Part. Differ. Equ. **34**(2), 193–231 (2009)
15. Jordan, R., Kinderlehrer, D., Otto, F.: The variational formulation of the Fokker–Planck equation. SIAM J. Math. Anal. **29**(1), 1–17 (1998)
16. Lafferty, J.D.: The density manifold and configuration space quantization. Trans. Am. Math. Soc. **305**(2), 699–741 (1988)
17. Li, W.: Geometry of Probability Simplex via Optimal Transport. arXiv:1803.06360 [math] (2018)
18. Li, W., Lin, A.T., Montufar, G.: Affine natural proximal learning. In: Geometric Science of Information, pp. 705–714 (2019)
19. Li, W., Montúfar, G.: Natural gradient via optimal transport. Inf. Geom. **1**(2), 181–214 (2018)
20. Li, W., Montufar, G.: Ricci curvature for parametric statistics via optimal transport. CAM report 18-52 (2018)
21. Liero, M., Mielke, A., Savaré, G.: Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. Invent. Math. **211**(3), 969–1117 (2018)
22. Lin, A.T., Li, W., Osher, S., Montufar, G.: Wasserstein proximal of GANs. CAM report 18-53 (2019)
23. Liu, J.-G., Lu, J., Margetis, D., Marzuola, J.L.: Asymmetry in crystal facet dynamics of homoepitaxy by a continuum model. Phys. D Nonlinear Phenom. **393**, 54–67 (2019)
24. Liu, Q.: Stein variational gradient descent as gradient flow. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 30, pp. 3115–3123. Curran Associates Inc., New York (2017)
25. Liu, Q., Wang, D.: Stein variational gradient descent: a general purpose bayesian inference algorithm. In: Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16, pp. 2378–2386. Curran Associates Inc., USA (2016)
26. Lu, J., Lu, Y., Nolen, J.: Scaling limit of the stein variational gradient descent: the mean field regime. SIAM J. Math. Anal. **51**(2), 648–671 (2019)
27. Lu, J., Vanden-Eijnden, E.: Exact dynamical coarse-graining without time-scale separation. J. Chem. Phys. **141**(4), 044109 (2014)
28. Malagò, L., Montrucchio, L., Pistone, G.: Wasserstein Riemannian Geometry of Positive Definite Matrices. arXiv:1801.09269 [math, stat] (2018)
29. Minh, H.Q.: A unified formulation for the Bures–Wasserstein and Log-Euclidean/Log-Hilbert–Schmidt distances between positive definite operators. In: Geometric Science of Information, pp. 475–483 (2019)
30. Nelson, E.: Quantum Fluctuations. Princeton Series in Physics. Princeton University Press, Princeton (1985)
31. Oksendal, B.K.: Stochastic Differential Equations: An Introduction with Applications, 2nd edn. Springer, Berlin (2013)
32. Otto, F.: The geometry of dissipative evolution equations the porous medium equation. Commun. Part. Differ. Equ. **26**(1–2), 101–174 (2001)
33. Pavliotis, G.A.: Stochastic processes and applications, volume 60 of Texts in Applied Mathematics. In: Diffusion Processes, the Fokker–Planck and Langevin Equations. Springer, New York (2014)
34. Shlyakhtenko, D.: Free Fisher Information for Non-tracial States. arXiv:math/0101137 (2001)
35. Tsallis, C.: Possible generalization of Boltzmann–Gibbs statistics. J. Stat. Phys. **52**(1), 479–487 (1988)
36. Villani, C.: Optimal Transport: Old and New. Number 338 in Grundlehren Der Mathematischen Wissenschaften. Springer, Berlin (2009)
37. Wong, T.K.L.: Logarithmic Divergences from Optimal Transport and Renyi Geometry. arXiv:1712.03610 [cs, math, stat] (2017)
38. Zozor, S., Brossier, J.-M.: deBruijn identities: from Shannon, Kullback–Leibler and Fisher to generalized ϕ-entropies, ϕ-divergences and ϕ-Fisher informations. AIP Conf. Proc. **1641**(1), 522–529 (2015)

Publisher's Note