ON THE GEOMETRY AND DYNAMICAL FORMULATION OF THE SINKHORN ALGORITHM FOR OPTIMAL TRANSPORT

Journal of Computational Dynamics
Vol. 11, No. 4, October 2024, pp. 442-452
doi:10.3934/jcd.2024006

ON THE GEOMETRY AND DYNAMICAL FORMULATION OF THE SINKHORN ALGORITHM FOR OPTIMAL TRANSPORT

Klas Modin*

Chalmers and University of Gothenburg, Sweden

Abstract. The Sinkhorn algorithm is a numerical method for the solution of optimal transport problems. Here, I give a brief survey of this algorithm, with a strong emphasis on its geometric origin: it is natural to view it as a discretization, by standard methods, of a non-linear integral equation. In the appendix, I also provide a short summary of an early result of Beurling on product measures, directly related to the Sinkhorn algorithm.

1. **Introduction.** The Sinkhorn algorithm has quickly sailed up as a popular numerical method for optimal transport (OT) problems (see the book by Peyré and Cuturi [19] and references therein). Its standard derivation starts from the Kantorovich formulation of the discrete OT problem and then adds entropic regularization, which enables Sinkhorn's theorem [22] for the optimal coupling matrix (see the paper by Cuturi [5] for basic notions, and the work of Schmitzer [20] and the PhD thesis of Feydy [6] for more details, including optimized and stable implementation of the full algorithm). In this paper, I give a brief survey of the Sinkhorn algorithm from a geometric viewpoint, making explicit connections to well-known techniques in geometric hydrodynamics, such as the Madelung transform, and the dynamical formulation of OT as given by Benamou and Brenier [2]. This viewpoint on entropic regularization is known to experts since long before the Sinkhorn algorithm became popular for OT: it was first formulated by Schrödinger [21], then re-established by Zambrini [25], and is today well-known as the law governing the *Schrödinger bridge problem* in probability theory (cf. Leonard [13]). Here, I wish to complement this presentation by entirely avoiding probability theory and instead use the language of geometric analysis, particularly geometric hydrodynamics (*cf.* Arnold and Khesin [1]). The presentation of Léger [10], based on Wasserstein geometry, is close to mine and contains more details. A benefit of the hydrodynamic formulation is that it readily enables standard numerical tools for space and time discretization (and the accompanying numerical analysis).

The principal theme of my presentation is that the Sinkhorn algorithm on a smooth, connected, orientable Riemannian manifold M has a natural interpretation as a space and time discretization of the following non-linear evolution integral

2020 *Mathematics Subject Classification.* Primary: 49Q22, 35Q49, 37K65; Secondary: 65R20.
Key words and phrases. Optimal transport, Sinkhorn algorithm, Schrödinger bridge, product measures, Beurling, Madelung transform, groups of diffeomorphisms, geometric hydrodynamics.
Corresponding author: Klas Modin.

equation:

$$\begin{cases} \partial_s f = -f - \log\left(\displaystyle\int_M K_\varepsilon(\cdot, y)\, \mathrm{e}^{g(y)}\, \mathrm{d}y\right) + \log \rho_0, & f\colon M \times [0,\infty) \to \mathbb{R} \\ \partial_s g = -g - \log\left(\displaystyle\int_M K_\varepsilon(\cdot, y)\, \mathrm{e}^{f(y)}\, \mathrm{d}y\right) + \log \rho_1, & g\colon M \times [0,\infty) \to \mathbb{R}. \end{cases} \tag{1}$$

Here, ρ_0 and ρ_1 are two strictly positive and smooth densities with the same total mass relative to the Riemannian volume form, and $K_\varepsilon(x,y)$ is the heat kernel on M. In the special case $M = \mathbb{R}^n$ then

$$K_\varepsilon(x,y) = \frac{1}{(4\pi\varepsilon)^{n/2}} \exp\Big(-\frac{\|x-y\|^2}{4\varepsilon}\Big). \tag{2}$$

It is sometimes convenient to use the heat flow semi-group notation

$$\mathrm{e}^{\varepsilon\Delta} f \coloneqq \int_M K_\varepsilon(\cdot, y) f(y) \mathrm{d}y. \tag{3}$$

Remark 1.1. The equations (1) are almost linear for small ε. Indeed, they can be written

$$\begin{cases} \partial_s f = -f - g + \log \rho_0 - T_\varepsilon(g), \\ \partial_s g = -g - f + \log \rho_1 - T_\varepsilon(f), \end{cases} \qquad T_\varepsilon(f) \coloneqq \log\big(\mathrm{e}^{-f} \mathrm{e}^{\varepsilon\Delta} \mathrm{e}^{f}\big), \tag{4}$$

where $(f,\varepsilon) \mapsto T_\varepsilon(f)$ is a C^∞ mapping, for example in Sobolev topologies H^s. Local existence and uniqueness of solutions to the integral equations (1) are obtained by standard Picard iterations. The extension to global results probably follow by standard techniques, but the question of *convergence* as $s \to \infty$ (and possibly $\epsilon \to 0$ simultaneously) is subtle. To approach it, one could attempt backward error analysis on the continuous version of the system (11) below. I expect this would yield a connection to the non-linear parabolic equation studied by Berman [3], whose dynamics, he proved, approximates the dynamics of $\boldsymbol{f}^{(k)} \mapsto \boldsymbol{f}^{(k+1)}$ below.

Before discussing the geometric origin of the equations (1), let us see how the standard (discrete) Sinkhorn algorithm arises from the equations (1).

Let ρ_0 and ρ_1 be two atomic measures on M

$$\rho_0 = \sum_{i=1}^N p_i \delta_{x_i}, \qquad \rho_1 = \sum_{i=1}^N q_i \delta_{y_i}, \tag{5}$$

where $x_i, y_i \in \mathbb{R}^n$ and the weights $p_i > 0$ and $q_i > 0$ fulfill $\sum_i p_i = \sum_i q_i$. If we change variables $a(x,s) = \exp(f(x,s))$ and $b(x,s) = \exp(g(x,s))$ then the integral equations (1) become

$$\begin{cases} \partial_s a = -a \log\left(\dfrac{a \mathrm{e}^{\varepsilon\Delta} b}{\rho_0}\right), \\ \partial_s b = -b \log\left(\dfrac{b \mathrm{e}^{\varepsilon\Delta} a}{\rho_1}\right). \end{cases} \tag{6}$$

For this to make sense for (5) we need $a \ll \rho_0$ and $b \ll \rho_1$, i.e., $a = \sum_i a_i \delta_{x_i}$ and $b = \sum_i b_i \delta_{y_i}$ for some $a_i \geq 0$ and $b_i \geq 0$. Since the heat flow convolves a delta distribution to a smooth function, it follows if $\varepsilon > 0$ that

$$\left(\frac{a \mathrm{e}^{\varepsilon\Delta} b}{\rho_0}\right)(x_i) = \frac{a_i}{q_i} \sum_{j=1}^N K_\varepsilon(x_i, y_j) b_j = \frac{a_i}{q_i} \boldsymbol{K}_\varepsilon \boldsymbol{b}, \tag{7}$$

where $\boldsymbol{b} = (b_1, \dots, b_N)$ and $\boldsymbol{K}_\varepsilon$ is the $N \times N$ matrix with entries $(\boldsymbol{K}_\varepsilon)_{ij} = K_\varepsilon(x_i, y_i)$. Likewise,

$$\left(\frac{b\mathrm{e}^{\varepsilon\Delta}a}{\rho_1}\right)(y_i) = \frac{b_i}{p_i}\boldsymbol{K}_\varepsilon^\top \boldsymbol{a}, \tag{8}$$

where $\boldsymbol{a} = (a_1, \dots, a_N)$ and the heat kernel property $K_\varepsilon(y_i, x_j) = K_\varepsilon(x_j, y_i)$ is used. Expressed in $\boldsymbol{a}$ and $\boldsymbol{b}$, the equations (6) now become an ordinary differential equation (ODE) on M^{2N}

$$\begin{cases} \partial_s \boldsymbol{a} = -\boldsymbol{a} * \log\left(\dfrac{\boldsymbol{a}}{\boldsymbol{q}} * \boldsymbol{K}_\varepsilon \boldsymbol{b}\right), \\ \partial_s \boldsymbol{b} = -\boldsymbol{b} * \log\left(\dfrac{\boldsymbol{b}}{\boldsymbol{p}} * \boldsymbol{K}_\varepsilon^\top \boldsymbol{a}\right), \end{cases} \tag{9}$$

where $*$ denotes element-wise multiplication and log and divisions are also applied element-wise. Notice that this ODE loses its meaning (or rather its connection to (6)) if $\varepsilon = 0$. Moving back to the coordinates $\boldsymbol{f} = \log \boldsymbol{a}$ and $\boldsymbol{g} = \log \boldsymbol{b}$ yields the system

$$\begin{cases} \partial_s \boldsymbol{f} = -\boldsymbol{f} - \log\left(\boldsymbol{K}_\varepsilon \exp \boldsymbol{g}\right) + \log \boldsymbol{q}, \\ \partial_s \boldsymbol{g} = -\boldsymbol{g} - \log\left(\boldsymbol{K}_\varepsilon^\top \exp \boldsymbol{f}\right) + \log \boldsymbol{p}, \end{cases} \tag{10}$$

To proceed, we need a time discretization. For this, apply to (10) the Trotter splitting (*cf.* [14]) combined with the forward Euler method to obtain

$$\begin{cases} \boldsymbol{f}^{(k+1)} = (1-h)\boldsymbol{f}^{(k)} - h\log\left(\boldsymbol{K}_\varepsilon \exp \boldsymbol{g}^{(k)}\right) + h\log \boldsymbol{q}, \\ \boldsymbol{g}^{(k+1)} = (1-h)\boldsymbol{g}^{(k)} - h\log\left(\boldsymbol{K}_\varepsilon^\top \exp \boldsymbol{f}^{(k+1)}\right) + h\log \boldsymbol{p}, \end{cases} \tag{11}$$

where $h > 0$ is the time-step length.

Definition 1.2. The *discrete Sinkhorn algorithm on* M is given by the time discretization (11) with $h = 1$.

Remark 1.3. For $M = \mathbb{R}^n$ the heat kernel is given by (2). If we express (11) in the variables $\boldsymbol{a}$ and $\boldsymbol{b}$ and take $h = 1$ we recover the Euclidean Sinkhorn algorithm as presented in the literature

$$\boldsymbol{a}^{(k+1)} = \frac{\boldsymbol{q}}{\boldsymbol{K}_\varepsilon \boldsymbol{b}^{(k)}}, \qquad \boldsymbol{b}^{(k+1)} = \frac{\boldsymbol{p}}{\boldsymbol{K}_\varepsilon^\top \boldsymbol{a}^{(k+1)}}. \tag{12}$$

Remark 1.4. If we take $h \neq 1$ then (11) expressed in $\boldsymbol{a}$ and $\boldsymbol{b}$ become

$$\boldsymbol{a}^{(k+1)} = \left(\boldsymbol{a}^{(k)}\right)^{1-h}\left(\frac{\boldsymbol{q}}{\boldsymbol{K}_\varepsilon \boldsymbol{b}^{(k)}}\right)^h, \qquad \boldsymbol{b}^{(k+1)} = \left(\boldsymbol{b}^{(k)}\right)^{1-h}\left(\frac{\boldsymbol{p}}{\boldsymbol{K}_\varepsilon^\top \boldsymbol{a}^{(k+1)}}\right)^h. \tag{13}$$

For $h > 1$ this is the *over-relaxed Sinkhorn algorithm*, which convergences faster than (12) (see [23, 18, 12]). Indeed, the faster convergence is readily understood by applying linear stability theory to (11) in the vicinity $\varepsilon \to 0^+$. From a numerical ODE perspective, the accelerated convergence is expected: larger time steps typically yield faster marching toward asymptotics, provided that the time step is small enough for the method to remain stable. Indeed, Figure 1 shows an almost perfect match between the stability region of the Trotter-Euler method and the convergence of the time-step iterations (11). An interesting venue could be to look for other methods, with other numerical damping properties.

Remark 1.5. Notice that if $C = \int_M a\mathrm{e}^{\varepsilon\Delta}b$ then, along the flow (6),

$$\frac{d}{ds}C = -\int_M a\mathrm{e}^{\varepsilon\Delta}b\log\left(\frac{a\mathrm{e}^{\varepsilon\Delta}b}{\rho_0}\right) - \int_M b\mathrm{e}^{\varepsilon\Delta}a\log\left(\frac{b\mathrm{e}^{\varepsilon\Delta}a}{\rho_1}\right).$$

2. **Dynamical formulation of optimal transport.** This section reviews the dynamical (or fluid) formulation of smooth optimal transport problems as advocated by Benamou and Brenier [2]. See [8] for details on the notation and more information about infinite-dimensional manifold and Riemannian structures. For simplicity, I assume from here on that M is a compact Riemannian manifold without boundary. The non-compact or boundary cases can be handled by introducing suitable decay or boundary conditions.

Let $\mathrm{Dens}(M) = \{\rho \in C^\infty(M) \mid \rho(x) > 0, \int_M \rho = 1\}$ denote the space of smooth probability densities. It has the structure of a smooth Fréchet manifold [7]. Its tangent bundle is given by tuples $(\rho, \dot\rho)$ where $\dot\rho \in C^\infty_0(M) = \{f \in C^\infty(M) \mid \int_M f = 0\}$. Otto [17] suggested the following (weak) Riemannian metric on $\mathrm{Dens}(M)$

$$\langle \dot\rho, \dot\rho\rangle_\rho = \int_M |\nabla S|^2\rho, \qquad \dot\rho + \mathrm{div}(\rho\nabla S) = 0. \tag{14}$$

The beauty of this metric is that the distance it induces is exactly the L^2-Wasserstein distance, and the geodesic two-point boundary value problem corresponds to the optimal transport problem (in the smooth category). See [17, 9, 15] for details. In summary, L^2 optimal transport is a problem of Lagrangian (variational) mechanics on $\mathrm{Dens}(M)$: find a path $[0,1] \ni t \mapsto \rho_t \in \mathrm{Dens}(M)$ with fixed end-points ρ_0 and ρ_1 that extremizes (in this case minimizes) the action functional

$$A(\rho_t) = \int_0^1 L\Big(\rho_t, \frac{\partial\rho_t}{\partial t}\Big)\,\mathrm{d}t, \tag{15}$$

for the kinetic energy Lagrangian $L(\rho,\dot\rho) = \frac{1}{2}\langle\dot\rho,\dot\rho\rangle_\rho$. The optimal transport map is then recovered as the time-one flow map for the time dependent vector field on M given by $v(x,t) = \nabla S_t(x)$.

The dynamical formulation is now obtain by replacing the variational problem for the action (15) with an equivalent constrained variational problem on $\mathrm{Dens}(M) \times \Omega^1(M)$ (densities times one-forms) for the action

$$\bar A(\rho_t, m_t) = \frac{1}{2}\int_0^1 \langle\!\langle m_t/\rho_t, m_t\rangle\!\rangle_{L^2}\,\mathrm{d}t \tag{16}$$

subject to the constraint

$$\dot\rho_t + \mathrm{div}(m_t) = 0. \tag{17}$$

This is a convex optimization problem since $\bar A$ is convex and the constraint is linear (see [2] for details). Notice that the convexity of $\bar A$ is with respect to the linear structure of $C^\infty(M)\times\Omega^1(M)$, which is different from the non-linear convexity notion for the Levi-Civita connection associated with the Riemannian metric (14).

3. **Entropic regularization and the Madelung transform.** The aim of this section is to introduce entropic regularization of the dynamical formulation and show how it simplifies the problem via the imaginary Madelung (or Hopf–Cole) transform. This transformation is the analog, in the dynamical formulation of smooth OT, to Sinkhorn's theorem applied to the coupling matrix in discrete OT.

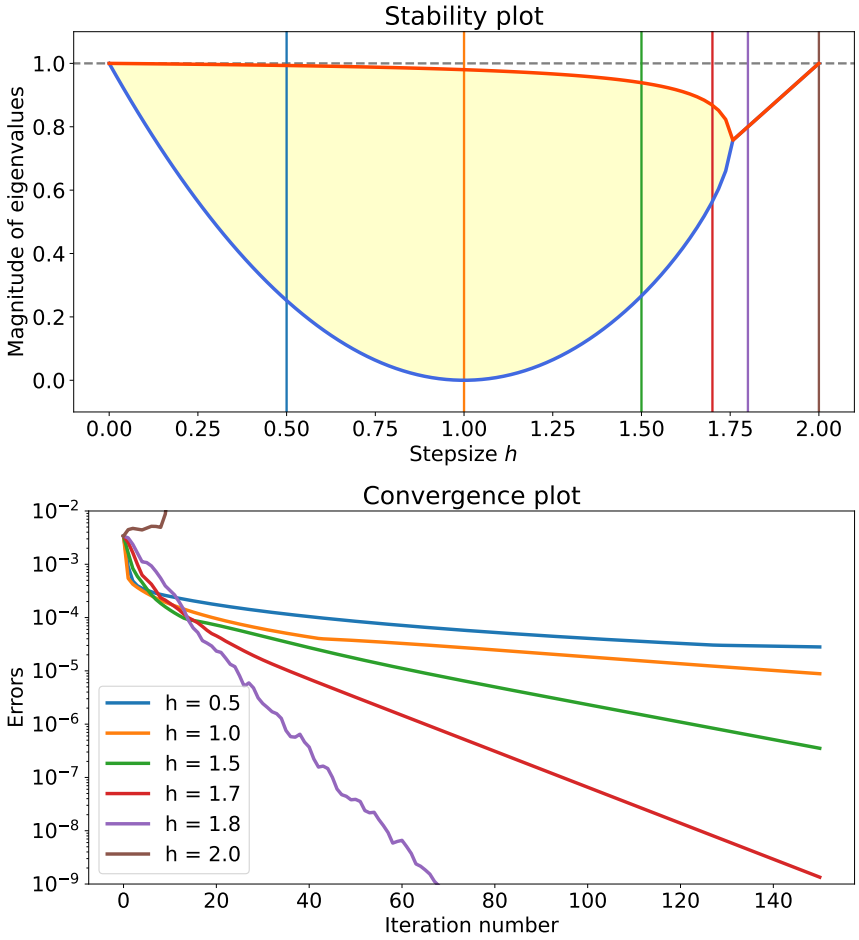

FIGURE 1. Stability analysis (upper figure) for the Trotter–Euler splitting method applied to the linear test equation $\dot{x} = -x - (1-\delta)y$, $\dot{y} = -(1-\delta)x - y$ for $\delta = 10^{-2}$, roughly corresponding to the system (11) with $\varepsilon \simeq \delta$. The method is stable whenever the magnitude of the corresponding two eigenvalues for the flow map are bounded by 1. The lower figure shows, for $\varepsilon = 10^{-2}$ and various step-sizes marked in the stability plot, how the L^2 norm of the right-hand side of (10) (the error) decreases with the number of time-steps taken with the Trotter–Euler method (11). The convergence plot matches the theoretical stability plot well. At $h = 1$ we observe an initially almost vertical decrease of the error due to the small eigenvalue near zero, thereafter the dynamics takes place in the eigenspace of the other eigenvalue close to one and here the convergence is slow. The optimal step-size is where the two eigenvalues meet at about $h \approx 1.75$. In the interval $h \in [1.75, 2)$ we expect an almost linear decrease of the error, as the two eigenvalues are the same. For $h \geq 2$ the method is not (linearly) stable.

Most of the results presented in this section are available in the papers by Leonhard [13] and by Léger [10]. See also Léger and Li [11] for a generalized Schrödinger bridge problem.

Let me first introduce two central functionals from information theory. The first is *entropy*, i.e., the functional on $\mathrm{Dens}(M)$ given by

$$E(\rho) = \int_M \rho \log \rho. \tag{18}$$

Its cousin, the *Fisher information functional*, is given by

$$I(\rho) = \frac{1}{2} \int_M \frac{|\nabla \rho|^2}{\rho}. \tag{19}$$

There are various ways to describe the relation between $E(\rho)$ and $I(\rho)$:

- $I(\rho)$ is the trace of the Hessian (with respect to (14)) of $E(\rho)$, or
- $I(\rho)$ is the rate of change of entropy along the heat flow $\dot\rho = \Delta\rho$.

In our context, $I(\rho)$ plays the role of negative potential energy in the Lagrangian

$$L(\rho, \dot\rho) = \frac{1}{2} \langle \dot\rho, \dot\rho \rangle_\rho + \varepsilon^2 I(\rho). \tag{20}$$

The corresponding action functional

$$A(\rho_t) = \int_0^1 \left(\frac{1}{2} \langle \partial_t \rho_t, \partial_t \rho_t \rangle_\rho + \varepsilon^2 I(\rho_t) \right) \mathrm{d}t \tag{21}$$

is called the *entropic regularization* of (15). Notice that the parameter ε has physical dimension as thermal diffusivity. This is not a coincidence. Indeed, as we shall soon see this regularization significantly simplifies the variational problem (imaginary ε also works!) by means of heat flows with thermal diffusivity ε. Before that, however, let me just point out that the dynamical formulation of Benamou and Brenier [2] is still applicable: if we change variables to (ρ, m) as before we obtain again a convex optimization problem (since the functional $I(\rho)$ is convex)

$$\min_{\rho_t, m_t} \int_0^1 \left(\frac{1}{2} \langle\!\langle m_t/\rho_t, m_t \rangle\!\rangle_{L^2} + \varepsilon^2 I(\rho_t) \right) \mathrm{d}t \quad \text{subject to} \quad \partial_t \rho_t + \mathrm{div}(m_t) = 0. \tag{22}$$

In a suitable setting one can apply convex analysis to obtain existence and uniqueness (*cf.* [2]).

So far I used Lagrangian mechanics to describe the dynamical formulation. Let me now switch to the Hamiltonian view-point. The Legendre transform of the Lagrangian (20) is

$$T\mathrm{Dens}(M) \ni (\rho, \dot\rho) \mapsto \left(\rho, \frac{\delta L}{\delta \dot\rho} \right) = (\rho, S) \in T^*\mathrm{Dens}(M) \tag{23}$$

where the co-tangent space $T^*_\rho \mathrm{Dens}(M)$ is given by co-sets of smooth functions defined up to addition of constants (if M is connected as I assume). Notice that the S in (23) is exactly the S in (14) (defined only up to a constant). The corresponding Hamiltonian on $T^*\mathrm{Dens}(M)$ is

$$H(\rho, S) = \frac{1}{2} \langle S, S \rangle_{*\rho} - \varepsilon^2 I(\rho), \tag{24}$$

where the *dual metric* $\langle \cdot, \cdot \rangle_*$ is given by

$$\langle S, S \rangle_{*\rho} = \int_M |\nabla S|^2 \rho. \tag{25}$$

Before I continue, consider a finite-dimensional analog of the Hamiltonian (24): on $T^*\mathbb{R}$ take $H(q,p) = p^2/2 - \varepsilon^2 q^2/2$. Of course, the analogy is $p^2/2 \leftrightarrow \langle S, S\rangle_{*\rho}$ and $q^2/2 \leftrightarrow I(\rho)$. The equations of motion are

$$\dot{q} = p, \quad \dot{p} = \varepsilon^2 q. \tag{26}$$

This system describes a harmonic oscillator when ε is *imaginary*. For real ε, the dynamics is not oscillatory. Indeed, if we change variables to $x = p - \varepsilon q$ and $y = p + \varepsilon q$ we obtain the Hamiltonian $H(x,y) = xy/2$ with dynamics

$$\dot{x} = x/2, \quad \dot{y} = -y/2. \tag{27}$$

These are two uncoupled equations where x is growing and y is decaying exponentially. Thus, we can expect this type of dynamics also for (24). Indeed, I shall now introduce a change of coordinates for (ρ, S), analogous to the change of coordinates $(q,p) \iff (x,y)$.

Definition 3.1. The *imaginary Madelung transform*[1] is given by

$$T^*\mathrm{Dens}(M) \ni (\rho, S) \longmapsto \left(\underbrace{\sqrt{\rho \mathrm{e}^{S/\epsilon}}}_{a}, \underbrace{\sqrt{\rho \mathrm{e}^{-S/\epsilon}}}_{\bar{b}} \right), \tag{28}$$

where $a, \bar{b} \in C^\infty(M)$ are defined up to $(\mathrm{e}^\sigma a, \mathrm{e}^{-\sigma}\bar{b})$ for $\sigma \in \mathbb{R}$ and should fulfill $\int_M a\bar{b} = 1$. The individual component a is known as the *Hopf-Cole transform*.

This transformation is a symplectomorphism (see [10] for $\varepsilon \in \mathbb{R}$ and [24, 8] for $\varepsilon \in \mathrm{i}\mathbb{R}$). The inverse transform is $\rho = a\bar{b}$ and $S = \varepsilon \log(a/\bar{b})$. The Hamiltonian (24) expressed in the new canonical coordinates $(a, \bar{b})$ thus become

$$\begin{aligned} H(a,\bar{b}) &= \int_M \frac{\epsilon^2}{2} \left(\left| \frac{\nabla a}{a} - \frac{\nabla \bar{b}}{\bar{b}} \right|^2 - \left| \frac{\nabla a}{a} + \frac{\nabla \bar{b}}{\bar{b}} \right|^2 \right) a\bar{b} \\ &= -\varepsilon^2 \int_M \nabla a \cdot \nabla \bar{b} = \varepsilon^2 \int_M a \Delta \bar{b}\,. \end{aligned} \tag{29}$$

Notice two things: (i) the Hamiltonian is quadratic, and (ii) it is of the form in the toy example $H(x,y)$ above. Hamilton's equations of motion for (29) are

$$\dot{a} = \varepsilon \Delta a, \qquad \dot{\bar{b}} = -\varepsilon \Delta \bar{b}\,. \tag{30}$$

Again, two decoupled equations as in the toy example, but now given by forward and backward heat flows with thermal diffusivity ε.

Remark 3.2. To be more precise, one should also take into account that $(a, \bar{b})$ is a co-set, so the general form of the equation should be

$$\dot{a} = \varepsilon \Delta a + \sigma a, \qquad \dot{\bar{b}} = -\varepsilon \Delta \bar{b} - \sigma \bar{b}\,, \tag{31}$$

where $t \mapsto \sigma(t) \in \mathbb{R}$ is arbitrary. However, since the scaling is arbitrary we can always represent the $(a, \bar{b})$ co-set by the element for which $\sigma = 0$. Notice also that the constraint $\int_M a\bar{b} = 1$ is preserved by the flow, as a short calculation shows.

It is, of course, not good to work with backward heat flows, but there is an easy fix. Let $b(x,t) \coloneqq \bar{b}(x, 1-t)$. Then b fulfills the forward heat equation (but

[1] In the standard Madelung transform $\varepsilon = \mathrm{i}\hbar$ which is why I say 'imaginary' here.

backwards in time). In the variables $a_t = a(\cdot, t)$ and $b_t = b(\cdot, t)$ the solution to the variational problem for the action (21) must therefore fulfill

$$\begin{cases} \partial_t a_t = \varepsilon \Delta a_t, & a_0 b_1 = \rho_0 \\ \partial_t b_t = \varepsilon \Delta b_t, & a_1 b_0 = \rho_1. \end{cases} \tag{32}$$

The two equations are coupled only through mixed boundary conditions at $t = 0$ and $t = 1$. With $a \coloneqq a_0$ and $b \coloneqq b_0$ these equations can be written in terms of the heat semigroup as

$$a \mathrm{e}^{\varepsilon \Delta} b = \rho_0, \quad b \mathrm{e}^{\varepsilon \Delta} a = \rho_1. \tag{33}$$

As you can see, a solution to (33) is a stationary point of the integral equations (6). Indeed, one should think of (6) as a gradient-type flow for solving the equations (33), as I shall now elaborate on.

If we take only the first part of the equations (33) we obtain the equation

$$\partial_s a = -a \log \left(\frac{a \mathrm{e}^{\varepsilon \Delta} b}{\rho_0} \right), \tag{34}$$

with b now as a fixed parameter. Let $\sigma = a \mathrm{e}^{\varepsilon \Delta} b$, so that $\partial_s \sigma = \mathrm{e}^{\varepsilon \Delta} b \partial_s a$ (since b is considered constant). The Fisher-Rao metric for $\partial_s \sigma$ is given by

$$\langle \partial_s \sigma, \partial_s \sigma \rangle_\sigma = \int_M \frac{(\partial_s \sigma)^2}{\sigma} = \int_M \frac{(\partial_s a)^2 \mathrm{e}^{\varepsilon \Delta} b}{a}. \tag{35}$$

Furthermore, the entropy of σ relative to ρ_0 is given by

$$H_{\rho_0}(\sigma) = \int_M \sigma \log \left(\frac{\sigma}{\rho_0} \right).$$

It is straightforward to check that the Riemannian gradient flow for the functional $F_1(a) = H_{\rho_0}(a \mathrm{e}^{\varepsilon \Delta} b)$ with respect to the metric (35) is given by equation (34). Likewise, the equation for b with fixed a is the Fisher-Rao gradient flow of $F_2(b) = H_{\rho_1}(b \mathrm{e}^{\varepsilon \Delta} a)$. Consequently, the Sinkhorn algorithm is the composition of steps for the first and second gradient flows. The question of assigning one Riemannian gradient structure to the entire flow is more intricate, since the functionals F_1 and F_2 depend on both a and b.

Appendix A. Beurling's "forgotten" result. Motivated by Einstein's work on Brownian motion governed in law by the heat flow, Schrödinger [21] arrived at the equations (33) by studying the most likely stochastic path for a system of particles from initial distribution ρ_0 to final distribution ρ_1. He gave physical arguments for why the problem should have a solution, but mathematically it was left open. S. Bernstein then addressed it at the 1932 International Congress of Mathematics in Zürich. A full resolution, however, did not come until 1960 through the work of Beurling [4]. The objective was, in Beurling's own words, "to derive general results concerning systems like (33) and, in particular, to disclose the inherent and rather simple nature of the problem." Beurling certainly succeeded in doing so. But to his astonishment (and slight annoyance) no-one took notice. In fact, Schrödinger's bridge problem was itself largely forgotten among physicists and mathematicians. Both Schrödinger's problem and the solution by Beurling were "rediscovered" and advocated by Zambrini [25] as he was working with an alternative version of Nelson's framework for *stochastic mechanics* (cf. [16]).

Beurling relaxed the problem (33) by replacing the functions ρ_0, ρ_1, a, b by measures $\mu_0, \mu_1, \alpha, \beta$ on M. By multiplying the right-hand sides, one obtains the product measure $\mu \equiv \mu_0 \wedge \mu_1$ on $M \times M$. For the left-hand side, Beurling went on as follows. Any measure ν on $M \times M$ gives rise to the generalized marginal measures ν_0 and ν_1 defined for all $h \in C_0(M)$ by

$$\int_M h \, d\nu_0 = \int_{M\times M} K_\epsilon(x,y) h(x) \, d\nu \quad \text{and} \quad \int_M h \, d\nu_1 = \int_{M\times M} K_\epsilon(x,y) h(y) \, d\nu. \tag{36}$$

Thus, we have a mapping from the space of measures to the space of product measures via the quadratic map

$$T_\epsilon \colon \nu \mapsto \nu_0 \wedge \nu_1. \tag{37}$$

Beurling noticed that the generalized version of Schrödinger's problem in equation (33) can be written

$$T_\epsilon(\alpha \wedge \beta) = \mu_0 \wedge \mu_1. \tag{38}$$

Let $\mathcal{M}$ denote the space of Radon measures on $M \times M$ (i.e., the continuous dual of compactly supported continuous functions on $M \times M$) and $\mathcal{P} \subset \mathcal{M}$ the sub-set of product measures. Further, let $\mathcal{M}^+ \subset \mathcal{M}$ and $\mathcal{P}^+ \subset \mathcal{P}$ denote the corresponding sub-sets of non-negative measures. Since $K_\epsilon > 0$, it follows that

$$T_\epsilon \colon \mathcal{M}^+ \to \mathcal{P}^+. \tag{39}$$

Let me now state Beurling's result adapted to the setting here.[2]

Theorem A.1 (Beurling [4], Thm. I)**.** *Let M be compact (possibly with boundary) and $\epsilon > 0$. Then the mapping (39) restricted to $\mathcal{P}^+$ is an automorphism (in the strong topology of $\mathcal{M}$).*

From this result, a solution to Schrödinger's problem (38) in the category of measures is obtained as $\alpha \wedge \beta = T_\epsilon^{-1}(\mu_0 \wedge \mu_1)$. Furthermore, the solution $\alpha \wedge \beta$ depends continuously (in operator norm) on the data $\mu_0 \wedge \mu_1$. Notice that, whereas $\alpha \wedge \beta$ is unique as a product measure, the components α, β themselves are only defined up to multiplication $e^f\alpha, e^{-f}\beta$ by an arbitrary function f on M. Thus, to work with product measures naturally captures the non-uniqueness pointed out in Remark 3.2 above.

The condition that M is compact is used to obtain a positive lower and upper bound on the kernel K_ϵ (these are, in fact, the only conditions that Beurling's proof imposes on K_ϵ). Such bounds are necessary for the map T_ϵ to be an automorphism (i.e., continuous with continuous inverse). Beurling also gave a second, weaker result, which can be applied to the case of non-compact M.

Theorem A.2 (Beurling [4], Thm. II)**.** *Let $\epsilon > 0$ and let $\mu_0 \wedge \mu_1 \in \mathcal{P}^+$ be such that*

$$\left| \int_M \int_M \log K_\epsilon \, d\mu_0 \, d\mu_1 \right| < \infty. \tag{40}$$

Then there exists a unique non-negative product measure ν on $M \times M$ that solves the equation

$$T_\epsilon(\nu) = \mu_0 \wedge \mu_1.$$

[2]The result proved by Beurling is much more general: it solves the problem for an n-fold product measure on the Cartesian product of n locally compact Hausdorff spaces.

Beurling's results can be viewed as a generalization from matrices to measures of Sinkhorn's theorem [22] on doubly stochastic matrices, only it came four years *before* Sinkhorn's result. I find it remarkable that Beurling came up with these results independently of Kantorovich's formulation of optimal transport in terms of measures on a product space (which came to general knowledge in the West in the late 1960's).

Acknowledgment. This work was supported by the Swedish Research Council (grant number 2022-03453) and the Knut and Alice Wallenberg Foundation (grant number WAF2019.0201). I would like to thank Ana Bela Cruzeiro, Christian Leonard, and Jean-Claude Zambrini, for helpful and intriguing discussions, and especially for pointing me to the "forgotten" work of Beurling.

REFERENCES

[1] V. I. Arnold and B. Khesin, *Topological Methods in Hydrodynamics*, Springer-Verlag, New York, 1998.
[2] J.-D. Benamou and Y. Brenier, A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem, *Numer. Math.*, **84** (2000), 375-393.
[3] R. J. Berman, The Sinkhorn algorithm, parabolic optimal transport and geometric Monge–Ampère equations, *Numer. Math.*, **145** (2020), 771-836.
[4] A. Beurling, An automorphism of product measures, *Ann. of Math.*, **72** (1960), 189-200.
[5] M. Cuturi, Sinkhorn distances: Lightspeed computation of optimal transport, *Advances in Neural Information Processing Systems 26, Curran Associates, Inc.*, (2013), 2292-2300.
[6] J. Feydy, *Analyse de Données Géométriques, au Delà des Convolutions*, PhD thesis, Université Paris-Saclay, 2020.
[7] R. S. Hamilton, The inverse function theorem of Nash and Moser, *Bull. Amer. Math. Soc. (N.S.)*, **7** (1982), 65-222.
[8] B. Khesin, G. Misiołek and K. Modin, Geometry of the Madelung transform, *Arch. Rational Mech. Anal.*, **234** (2019), 549-573.
[9] B. Khesin and R. Wendt, *The Geometry of Infinite-dimensional Groups*, Ergeb. Math. Grenzgeb., 51. Springer-Verlag, Berlin, 2009.
[10] F. Léger, A geometric perspective on regularized optimal transport, *Journal of Dynamics and Differential Equations*, **31** (2019), 1777-1791.
[11] F. Léger and W. Li, Hopf–Cole transformation via generalized Schrödinger bridge problem, *J. Differential Equations*, **274** (2021), 788-827.
[12] T. Lehmann, M.-K. von Renesse, A. Sambale and A. Uschmajew, A note on overrelaxation in the Sinkhorn algorithm, *Optimization Lett.*, **16** (2021), 2209-2220.
[13] C. Leonard, A survey of the Schrödinger problem and some of its connections with optimal transport, *Discrete Contin. Dyn. Syst.*, **34** (2014), 1533-1574.
[14] R. I. McLachlan and G. R. W. Quispel, Splitting methods, *Acta Numer.*, **11** (2002), 341-434.
[15] K. Modin, Geometry of matrix decompositions seen through optimal transport and information geometry, *J. Geom. Mech.*, **9** (2017), 335-390.
[16] E. Nelson, Review of stochastic mechanics, *Journal of Physics: Conference Series*, **361** (2012), 012011.
[17] F. Otto, The geometry of dissipative evolution equations: the porous medium equation, *Comm. Partial Differential Equations*, **26** (2001), 101-174.
[18] G. Peyré, L. Chizat, F.-X. Vialard and J. Solomon, Quantum entropic regularization of matrix-valued optimal transport, *Eur. J. Appl. Math.*, **30** (2019), 1079-1102.
[19] G. Peyré and M. Cuturi, Computational optimal transport, *Foundations and Trends in Machine Learning*, **11** (2020), 355-607.
[20] B. Schmitzer, Stabilized sparse scaling algorithms for entropy regularized transport problems, *SIAM J. Sci. Comput.*, **41** (2019), A1443-A1481.
[21] E. Schrödinger, *Über Die Umkehrung der Naturgesetze*, Verlag der Akademie der Wissenschaften in Kommission bei Walter De Gruyter, 1931.
[22] R. Sinkhorn, A relationship between arbitrary positive matrices and doubly stochastic matrices, *Ann. Math. Stat.*, **35** (1964), 876-879.

[23] A. Thibault, L. Chizat, C. Dossal and N. Papadakis, Overrelaxed Sinkhorn–Knopp algorithm for regularized optimal transport, *Algorithms*, **14** (2021), Paper No. 143, 16 pp.
[24] M.-K. von Renesse, An optimal transport view of Schrödinger's equation, *Canad. Math. Bull*, **55** (2012), 858-869.
[25] J. C. Zambrini, Variational processes and stochastic versions of mechanics, *J. Math. Phys.*, **27** (1986), 2307-2330.

Received August 2023; revised January 2024; early access January 2024.