Notes on Ricci curvature of finite markov chains

matilde dolfato

This is a collection of notes on the study behind the topic of *Ricci Curvature of Finite Markov Chains*, Erbar and Maas (2011), Erbar and Fathi (2016).

Geometry ↔ Optimal Transport ↔ Statistics and Statistical Mechanics.

Chapters 1 to 4 are on the mathematical areas behind the paper. Chapter 5 relates them in the rationale of the paper. Chapters 6 and 7 address the content of the studies.

Contents

High-level logical idea

In the paper, they try to study **Ricci curvature of discrete spaces** for the first time. Studying the Ricci curvature of continuous spaces, instead, is a whole established field of research. The main motivation is that lower bounding the Ricci curvature of a space gives nice asymptotic properties to functions on that space. The following example is the main one to have in mind.

Example (Heat equation asymptotic behavior). We study the development of a function f on a manifold M according to the heat equation ("apply" the heat equation to f). If we have an lower bound on the Ricci curvature of M, then we know that the heat semigroup will converge to a constant value (average temperature) after a large time t. I.e.,

$$\text{If}\, \mathrm{Ric} \geqslant k, \quad ||\nabla P_t f|| \leqslant e^{-tk} P_t |\nabla f|$$

where $P_t f$ indicates the development of f according to the heat equation, and its derivative for large t is bounded by 0, i.e. $P_t f$ converges to a constant value.

In the paper, the authors would like to find a notion of lower bound on the Ricci curvature for discrete spaces, like spaces of Markov Chains. However, the very definition of (Ricci) curvature depends on a geometric structure and differentiable properties that are undefinable for discrete spaces. Hence, what they decide is to use the equivalent statement proposed in Theorem (**??**), which is not a differentiable one. Indeed, the statement depends only on distances and integrals (which are easily extendable to a discrete sum). By doing so, we are able to gain precious insights on objects living on discrete spaces, like Markov Chains, as we did for functions like the heat equation. The treasure at the end of the rainbow is to study some "geometric" properties of the discrete space to discover asymptotic properties of Markov processes, such as if they converge to an equilibrium, a steady state.

CHAPTER 1

Notions from geometry

Given a smooth manifold M, we can define its tangent space.

Definition 0.1 (Tangent space, visual). *Consider all the smooth curves on M, $\gamma : [0, \varepsilon] \to M$ such that $\gamma(0) = p$ for some $p \in M$. We want to define an equivalence relation on these curves, namely that of having the same velocity: given $\gamma_1 : [0, \varepsilon_1] \to M$ and $\gamma_2 : [0, \varepsilon_2] \to M$, $\gamma_1 \sim \gamma_2$ if for any chart $\varphi : U \to V \subseteq \mathbb{R}^n$ it holds*

$$(\varphi \circ \gamma_1)'(0) = (\varphi \circ \gamma_2)'(0) \tag{1}$$

. The tangent space to M at p, T_pM, is the set of all equivalence classes according to the velocity equivalence relation (1).

Remark. We need to pass to the chart to define what the velocity of a curve is, through $(\varphi \circ \gamma)'$, otherwise this derivative doesn't make sense (with the technology we use here).

Remark. It is sufficient to verify condition (1) for one chart, when considering a smooth manifold.

An equivalent, more algebraic definition of tangent space follows, which looks at directional derivatives and first uses the concept of ***derivation***.

Definition 0.2 (Derivation). *Consider the space of all real-valued smooth function on M, denoted $C^\infty(M)$. A map $D : C^\infty(M) \to \mathbb{R}$ is called a derivation at p if it is linear and satisfies $D(fg) = D(f)g(p) + D(g)f(p)$ for all $f, g \in C^\infty(M)$.*

Definition 0.3 (Tangent space, algebraic). *Given a smooth manifold M, the tangent space of M at p, T_pM, is the set of all derivations at p, i.e. of all possible ways to take a directional derivative at p.*

Remark. Derivations are essentially directional derivatives, hence the tangent space is the set of all possible ways of taking a directional derivative of a function at p.

Remark. **The tangent space is a vector space.**

Citing Lee (2003), you should visualize tangent vectors to M as "arrows" that are tangent to M and whose base points are attached to M at the given point. Proofs of theorems about tangent vectors must, of course, be based on the abstract definition in terms of derivations, but your intuition should be guided as much as possible by the geometric picture.

When the manifold is immersed in some Euclidean space $\mathbb{R}^n$, we can view T_pM as a vector subspace. The affine subspace $p + T_pM$ corresponds to our intuition of a "Taylor approximation" of M. For a surface (2-dimensional manifold) in $\mathbb{R}^3$ that would be the tangent plane to the surface.

Now, let us define the tangent bundle to M.

DEFINITION 0.4 (Tangent bundle). *Given a smooth manifold M of dimension n, the tangent bundle to M is a $2n$-dimensional smooth manifold TM together with a smooth* projection map $\pi : TM \to M$ *such that*

$$TM = \bigsqcup_{p \in M} T_pM\,;$$

$$\pi(p, v) = p \quad \forall v \in T_pM,\ \forall p \in M$$

This means that TM is the disjoint union of the tangent spaces at all points in M, and π is the map sending each tangent vector back to the point it is tangent at. Each fiber $\pi^{-1}(p) \in TM$ has the structure of a vector space.

A tangent bundle glues together the tangent spaces of M.[1]

DEFINITION 0.5 (Vector field). *Given a smooth manifold M, a vector field is a continuous map $X : M \to TM$ such that*

$$\pi \circ X = Id_M$$

Equivalently, it is a map $p \mapsto X_p$ such that $X_p \in T_pM$.

REMARK. A vector field is a *section* of the map π.

Vector fields on M can be visualized just like we visualize vector fields in the Euclidean space: as an arrow attached to each point of M, chosen to be tangent to M and to vary continuously from point to point Lee (2003).[2]

[1] I do not go over the definition of a vector bundle in general, which glues together vector spaces, I did NOT understand it.

[2] I discard the notion of Jacobi field and a different interpretation of the relation between curvature and geodesics (pg 5 paper byBrue) for now (it was just a good excuse to look at the above notions).

Also, we can look at vector fields from a concrete perspective, as follows.

We first define a ***basis of vector fields***. Consider a manifold M and a smooth coordinate chart $\varphi : U \subseteq M \to \mathbb{R}^n$ that induces a coordinate system $\{x_1, \dots, x_n\}$ on all tangent spaces at U. [3]

Then, we can derive the basis vector fields as

$$\frac{\partial}{\partial x_i} =: \partial_i$$

that associate the basis vector at each point in U. They are constant vector fields.

Then, a vector field $X : M \to TM$ can be expressed in terms of coordinates as the sum of n real-valued functions $X^1, \dots, X^n$ applied to each basis vector field, i.e.

$$X = \sum_i X^i \partial_i$$

These functions are called *component functions of* X in the given chart Lee (2003).

Also, we write the value of the vector field X at any point p as X_p, and we have

$$X_p = X^i(p) \partial_i|_p$$

The next proposition says that given a tangent vector at some point p, we can always find a vector field that creates it.

PROPOSITION 0.6. *Let M be a smooth manifold with or without boundary. Given* $p \in M$ *and* $v \in T_pM$*, there is a smooth global vector field* X *on* M *such that* $X_p = v$.

Proof: the proof considers the vector field defined on the set $\{v\}$ *defined as* $v \mapsto p$ *and the applies the extension lemma (see A).*

We refer to $\mathfrak{X}(M)$ as the space of all smooth vector fields on M. It is a vector space (i.e. closed) under pointwise addition and scalar multiplication,

$$(aX + bY)_p = aX_p + bY_p$$

The zero element is the zero vector field that assigns $0 \in T_pM$ to each $p \in M$.

Also, smooth vector fields can be multiplied by smooth real-valued functions. Given $f \in C^\infty(M)$, we define $fX : M \to TM$ as

$$(fX)_p = f(p)X_p$$

and the result is smooth.

[3] In the Euclidean space we fix a coordinate system by choosing a basis, on a manifold we do so by choosing a chart that induces a basis on the tangent space to the manifold :)

Vector fields define operators on the space of smooth real-valued functions. Given $X \in \mathfrak{X}(M)$ and $f : U \to \mathbb{R}$, $U \subseteq M$, we define a new function $Xf : U \to \mathbb{R}$ as

$$(Xf)(p) = X_p f$$

Also, the function Xf is smooth.

To clarify:

$Xf \longrightarrow$ smooth function, the operator defined by X on smooth functions (derivation);

$fX \longrightarrow$ smooth vector field, obtained by multiplying f.

Smooth vector fields also satisfy the following product rule:

$$X(fg) = f \; Xg + g \; Xf \tag{2}$$

Where the blank space is there because we first apply X to g and then multiply by f, obtaining another smooth function.

Recalling definition 0.2, **derivations can be identified with smooth vector fields**, i.e.

$$X_p f = (Df)(p)$$

Hence, we sometimes use the same notation for these two objects (vector fields $X : M \to TM$ and derivations $D : C^\infty(M) \to C^\infty(M)$.

Now, if we consider a smooth function Xf, we can apply another vector field Y to it obtaining another smooth function $YXf = Y(Xf)$. The operator that brings $f \mapsto YXf$, however, does not satisfy the product rule and is not a vector field.

What we can do is we consider also the converse operation XYf and subtract the two, obtaining a new operator, the *Lie bracket*.

Definition 0.7 (Lie bracket). *Given two smooth vector fields $X, Y \in \mathfrak{X}(M)$ and a smooth function $f : C^\infty(M) \to C^\infty(M)$, the operator $[X, Y] : C^\infty(M) \to C^\infty(M)$ is called the Lie bracket operator and defined as*

$$[X, Y]f = X(Yf) - Y(Xf)$$

The key fact is that this operator *is* a (smooth) vector field.

In coordinates,

$$[X, Y] = (XY^i - YX^i)\partial_i$$

CHAPTER 2

Riemannian geometry

Going from smooth manifolds in general to Riemannian manifolds means adding a ***metric***. In general, Lee (2018) does a nice introduction on this. Everything we know about Euclidean geometry on $\mathbb{R}^n$ can be derived from its inner product. To extend this geometric ideas to abstract smooth manifolds, we define an object that amounts to a *smoothly varying choice of inner product* on the tangent spaces.

First, an auxiliary definition:

Definition 0.1 (Tensor). *A k-tensor is a multilinear function that maps k vector spaces $V \times V \times \ldots \times V$ to the reals Spivak (2018). Tensors are invariant to changes of coordinate system.*

Definition 0.2 (Riemannian metric). *Given a smooth manifold M, a Riemannian metric on M is a smooth covariant tensor field*[1] *g whose value g_p at each $p \in M$ is an inner product on T_pM.*

Hence, g is a symmetric 2-tensor field that is positive definite in the sense that $g_p(v, v) \geqslant 0$ for each $p \in M$ and each $v \in T_pM$, with equality if and only if $v = 0$.

Definition 0.3 (Riemannian manifold). *A Riemannian manifold is a pair (M, g) of a manifold M and a specific choice of Riemannian metric g.*

Proposition 0.4. *Every smooth manifold admits a Riemannian metric.*

Example. The n-dimensional Euclidean space $\mathbb{R}^n$ is a Riemannian manifold with the Euclidean metric g (the Euclidean norm).

Example. A second class of Riemannian manifolds is a family: given $R > 0$, let $\mathbb{S}^n(R)$ denote the sphere of radius R centered at the origin in $\mathbb{R}^{n+1}$, endowed with the metric $\mathring{g}_R$ (called the *round metric* of radius R) induced by the Euclidean metric on $\mathbb{R}^{n+1}$.

[1] I think of it as a collection of tensors that varies smoothly across tangent spaces, I disregard the precise definitions of tensor field for now.

1. Connections

To define what curvature is we need to generalize the concept of a straight line in $\mathbb{R}^n$ to a manifold M^n. This is captured by *geodesics*. Geodesics in $\mathbb{R}^n$, i.e. straight lines, have the property of being the shortest path between two points. This notion is quite difficult to work with. Hence, we consider an equivalent statement, that is that any straight line in $\mathbb{R}^n$ has 0 acceleration. To talk about acceleration on a manifold, we need to define a new concept of derivation, through the object of *connection*: a set of rules for taking directional derivatives of vector fields, **independently of the coordinate system**.

Recall some auxiliary definitions first:

Definition 1.1. *Let $\gamma : [0, t] \to \mathbb{R}^n$ be a smooth curve. The* position *(vector) at time t is $\gamma(t)$. The* velocity *(vector) is given by the derivatives of the position vector with respect to time, $\gamma'(t) = \frac{\partial}{\partial t}\gamma(t)$. The* speed *is the magnitude of the velocity vector, and it is a scalar quantity. So, the velocity includes both the speed and the direction of current motion. The* acceleration *is the derivative of the velocity and the second derivative of the position, $\gamma''(t) = \frac{\partial}{\partial t}\gamma'(t) = \frac{\partial^2}{\partial^2 t}\gamma(t)$.*

The velocity of a curve is a vector field along the curve. The definition of velocity can be extended to a generic manifold by considering the derivative of the composition of the curve with a smooth chart of the manifold as in Definition 0.1 (alternatively, with derivations, we don't go into that). For the acceleration, we use connections.

The following is an abstract from Lee (2018) that sheds light on the motivation behind connections.

The problem is this: to define $\gamma''(t)$ by differentiating $\gamma'(t)$ with respect to t, we have to take a limit of a difference quotient involving the vectors $\gamma'(t+h)$ and $\gamma'(t)$; but these live in different vector spaces ($T_{\gamma(t+h)}M$ and $T_{\gamma(t)}M$, respectively), so it does not make sense to subtract them. The definition of acceleration works in the special case of smooth curves in $\mathbb{R}^n$ expressed in standard coordinates (or more generally, curves in any finite-dimensional vector space expressed in linear coordinates) because each tangent space can be naturally identified with the vector space itself. On a general smooth manifold, there is no such natural identification. The velocity vector $\gamma'(t)$ is an example of a vector field along a curve, a concept for which we will give a rigorous definition presently. To interpret the acceleration of a curve in a manifold, what we need is some coordinate-independent way to differentiate vector fields *along curves. To do so, we need a way to compare values of the vector field at different points, or intuitively, to "connect" nearby tangent spaces. This is where a connection comes in: it will be an additional piece of data on a manifold, a rule for computing directional derivatives of vector fields.*

Definition 1.2 (Connections). *Given a smooth manifold M and two vector fields $X, Y \in \mathfrak{X}(M)$, a connection is a map $\nabla : \mathfrak{X}(M) \times \mathfrak{X}(M) \to \mathfrak{X}(M)$ that maps $(X, Y) \mapsto \nabla_X Y$ and satisfies the following properties:*

(i) Linearity over $C^\infty(M)$ in X*: for $f_1, f_2 \in C^\infty(M)$ and $X_1, X_2 \in \mathfrak{X}(M)$:*

$$\nabla_{f_1 X_1 + f_2 X_2} Y = f_1 \nabla_{X_1} Y + f_2 \nabla_{X_2} Y$$

(ii) Linearity over $\mathbb{R}$ in Y*: for $a_1, a_2 \in \mathbb{R}$ and $Y_1, Y_2 \in \mathfrak{X}(M)$:*

$$\nabla_X (a_1 Y_1 + a_2 Y_2) = a_1 \nabla_X Y_1 + a_2 \nabla_X Y_2$$

(iii) Product rule*: for $f \in C^\infty(M)$,*

$$\nabla_X (fY) = f \nabla_X Y + (Xf) Y$$

You can think of a connection as an object that takes the **derivative of a vector field Y along the direction of another field X**. Also, $\nabla_X Y$ is called the **covariant derivative of Y in the direction X**. A covariant derivative is just a derivative along a certain direction.

Remark (A connection is not a tensor.). $\nabla_X fY = X(f)Y + f\nabla_X Y$, which is different from what we would want by linearity for the second term. Now compute $\nabla_{fX} Y$:

$$\nabla_{fX} Y = f \nabla_X Y$$

so you get that **it is a tensor in the first component, and not in the second**. The reason why it is not in the second lies in the product rule, property (iii) of connections. This is the same rule as for the Euclidean space. Also, the derivative of the function is the term $X(f)$:

$$X(f) = \sum_i X^i \partial_i(f) = \sum_i X^i \frac{\partial f}{\partial x_i}$$

Hence, it is also clear that if $f(Y) = aY$ for $a \in \mathbb{R}$, the term $X(f)$ is 0 and we get linearity in the second component, property (i).

We can also write connections explicitly, given a smooth coordinate chart. Consider the basis vector fields $\{\partial_1, \ldots, \partial_n\}$ for a coordinate chart over $U \subseteq M$. For any i, j, we can express a connection as

$$\nabla_{\partial_i} \partial_j = \Gamma^k_{ij}\, \partial_k$$

This identifies n^3 functions $\Gamma^k_{ij} : U \to \mathbb{R}$ called the ***connection coefficients of*** ∇.

Proposition 1.3. *The connection coefficients entirely define a connection. In particular:*

$$\nabla_X Y = (X(Y^k) + X^i Y^j\, \Gamma^k_{ij})\, \partial_k \tag{3}$$

Notice that in this proposition: Y^k are functions, hence $XY^k =$ are functions; also $X^i Y^j$ are functions; when they get multiplied by the vector fields ∂_k we obtain vector fields (a connection spits out a vector field).

Remark. As equation 3 makes clear, the connection coefficients are a correction to properly take the derivative on the manifold, considering its *curvature*, with respect to taking a standard derivation as if we were in the Euclidean space ($XY^k \partial_k = XY$).

Let us look at some examples of connection..

Example (Euclidean connection). In the Euclidean space, Lee (2003) defines a connection directly as

$$\nabla^{\mathbb{R}^n}_X Y = \sum_i X(Y^i) \partial_i$$

Considering this definition, it is immediate to see that under the standard basis **a connection in the Euclidean space as trivial coefficients**.

Maybe Lee takes for granted that the metric is the Euclidean one: under a different metric I think we could have non-trivial coefficients.

Example (Tangential connection on a submanifold). When considering an embedded submanifold $M \subseteq \mathbb{R}^n$, the tangential connection ∇^t is

$$\nabla^t_X Y = \pi_{TM}(\nabla_{\tilde{X}} \tilde{Y}|_M)$$

where $\tilde{X}, \tilde{Y} \in \mathbb{R}^n$ are extensions of $X, Y \in M$ and π_{TM} is the orthogonal projection onto TM. Hence you extend the vector fields, take the derivative in $\mathbb{R}^n$ and project onto the tangent bundle of the manifold.

Now we turn to define a special connection, the ***Levi-Civita connection***. This definition is useful because to use geodesics and covariant derivatives as tools for studying Riemannian geometry, we need a way to single out a particular connection that reflects the property of the metric. The two properties below uniquely define a connection.

Definition 1.4 (Levi-Civita Connection). *Given a Riemannian manifold (M, g), the Levi-Civita connection is a connection that satisfies the following properties:*

- *it preserves the metric:*

$$X(g(Y, Z)) = g(\nabla_X Y, Z) + g(Y, \nabla_X Z)$$

- *it is* torsion-free *or* symmetric*:*

$$T(X, Y) = \nabla_X Y - \nabla_Y X - [X, Y] = 0;$$

$$\nabla_X Y - \nabla_Y X = [X, Y]$$

Theorem 1.5 (Fundamental theorem of Riemannian geometry). *Given a Riemannian manifold (M, g), there exists a unique connection ∇ on TM which is compatible with g and symmetric, which is the Levi-Civita connection of g.*[2]

Some remarks on the LC connection:

- The connection coefficients Γ^k_{ij} of the LC connection are called ***Christoffel symbols***.
- The Levi-Civita connection on a Euclidean space is equal to the Euclidean connection.
- The Levi-Civita connection on a submanifold embedded in a Euclidean space is equal to the tangential connection.

Now, we define what geodesics are.

Definition 1.6 (Geodesic). *Given a smooth manifold M and a connection ∇ on TM, a smooth curve γ is called a geodesic w.r.t. if it has acceleration 0, i.e. $\nabla_{\gamma'}\gamma' = 0$.*

Remark. As anticipated, we prefer the definition above to the characterization of geodesics as "curves that minimize distance". Indeed, this version breaks in many cases, like that of a sphere: in a sphere, geodesics go back to the starting point p. Hence, when we are the the antipodal point to p we have exactly 2 curves minimizing the distance,

[2] We just leave this there to remember that ∇_{LC} is a connection which is symmetric, and we disregard the proof for now.

and for points in the other emisphere our geodesic will not be the one which minimizes distance. Hence, this second characterization is correct only *locally*, in a neighborhood of p.

A nice remark is the following.

Example. In an embedded submanifold $M \in \mathbb{R}^n$, a smooth curve $\gamma(t) : I \to M$ is a geodesic wrt the tangential connection iff it has acceleration (ordinarily computed as $\gamma''(t)$) orthogonal to the tangent bundle TM. Think of the acceleration of a particle moving on a circle ! This makes sense because to get the tangential connection on the submanifold we project the Euclidean connection, hence we get acceleration 0.

2. Curvatures

Now we are ready to deal with curvatures.

First, we give an high-level introduction to the very basic notion of curvature, on the lines of Lee (2018) (refer to the textbook from page 1).

Lee says: *Riemannian geometry [...] is the branch of differential geometry in which "geometric" ideas, in the familiar sense of the word, come to the fore. It is the direct descendant of Euclid's plane and solid geometry, by way of Gauss's theory of curved surfaces in space, and it is a dynamic subject of contemporary research. The central unifying theme in current Riemannian geometry research is the notion of curvature and its relation to topology.* Lee (2018)

The curvature of a curve $\gamma(t)$ is defined as the **length of its acceleration vector**, $\kappa(t) = |\gamma''(t)|$.

The geometric interpretation of this is to consider the smallest circle inscripted on a curve (osculating circle), i.e. the smallest circle tangent to the curve with velocity and acceleration vectors at the tangent point equal to those of the curve. Then the curvature is $\kappa(t) = 1/R$ where R is the radius of that circle. For instance, a circle of radius r has constant curvature $1/r$, a straight line has curvature 0.

We can then extend this definition so that the curvature takes both positive and negative sign, by fixing a direction of positive sign. This direction is chosen by considering a normal vector field N on the curve, and setting the curvature to be negative when the curve is turning away from N. We result with a function κ_N that is the ***signed curvature***.

Then, to generalize Euclidean geometry, we start by thinking of a surface embedded in $\mathbb{R}^3$. The curvature of a surface is defined by two numbers, the ***principal curvatures***.

In general, the principal curvatures are the eigenvalues of a self-adjoint linear endomorphism $s : V \to V$ where V is our surface endowed with an inner product (chapter 8 pg 238 of Lee (2018)). We know that there are n real eigenvalues by the Spectral theorem!

However, principal curvatures are not *intrinsic* properties of surfaces. Intrinsic properties are those that do not depend on embedding the surface in a higher dimensional space, but are true also for a 2-dimensional being living on the surface. Formally, they are those properties that are preserved by isometries. In 1827 Gauss discovered that by combining principal curvatures in a specific way we get a property that *is* intrinsic, the ***Gaussian curvature***. The Gaussian curvature is the determinant of the operator s. In 2d, it is $K = \kappa_1\kappa_2$. Indeed, all isometric transformations of our surface V leave the determinant of s invariant.

Now, let us look at some definitions of curvature formally. The most important notion of curvature is the **Riemannian curvature tensor**.

Definition 2.1 (Riemannian curvature tensor). *Given a Riemannian manifold* (M, g), *the map* $R : \mathfrak{X}(M) \times \mathfrak{X}(M) \times \mathfrak{X}(M) \to \mathfrak{X}(M)$ *defined as*

$$\begin{aligned} R(X, Y)Z &:= \nabla_X\nabla_Y Z - \nabla_Y\nabla_X Z - \nabla_{[X,Y]}Z \\ &= (\nabla^2_{X,Y} - \nabla^2_{Y,X} - \nabla_{[X,Y]})Z \end{aligned}$$

is called the **Riemannian curvature tensor**. *It is a (1,3)-tensor as it is multilinear over* $C^\infty(M)$.

Remark. A nice insight is given by looking at the first two terms in R first. They are the "cross" derivative of the first two vector fields, but just subtracting these two quantities does not give a linear object, i.e. a tensor. It was Riemann who discovered that by subtracting the commutator between X and Y, $[X, Y]$, we get something linear.

Example. Two simple cases for computing the Riemannian curvature tensor are the following.

(1) The Euclidean space, where the Riemannian curvature is 0 as the Euclidean connection is flat.
(2) Relation with the Gaussian curvature: another simple case is that of a 2-dimensional manifold immersed in $\mathbb{R}^3$. This manifold takes the metric g induced by $\mathbb{R}^3$. In this case, the Riemannian curvature $R(X, Y, Z, W)$ is the product of a function

of the 4 vector fields and a real-valued function $K(X)$ which is the Gaussian curvature (defined above). It makes sense! because the Gaussian curvature is the determinant of a self-adjoint operator and it pops up when considering the LC connection, induced by the metric.

Remark (Tensor vs. real number). The Riemannian curvature as defined above is a tensor. We can equivalently take the inner product between $R(X,Y)Z$ and a *forth* vector field V to get the Riemannian curvature as a real $R(X,Y,Z,V) = (R(X,Y)Z, V) \in \mathbb{R}$. The two definitions hold the exact same amount of information. We use this second approach in the definitions below.

Next, we look at the sectional curvature of a space.

Definition 2.2 (Sectional curvature). *Given linearly independent tangent vectors $v, w \in T_pM$, the sectional curvature $\mathcal{K}(v,w) \in \mathbb{R}$ is defined as*

$$\mathcal{K}(v,w) = \frac{(R(v,w)w,v)}{(v,v)(w,w)-(v,w)^2} \quad \in \mathbb{R}$$

This is some kind of normalization of the Riemannian curvature. We can think of the **sectional curvature as the Gaussian curvature of the surface** on M tangent to T_pM. In fact, given two vector v, w, these define a region on the manifold M, tangent to the tangent space. This region is defined as follows. If you take a curve $\gamma(t)$ starting at p with initial velocity v, by changing the norm of v you get a different ending point $\gamma(1)$ along the curve. At the same time, if we parametrize the curve with two initial vectors v, w, we do not span only a curve but a whole region, with boundaries the curves parametrized by v and w alone.

Alternatively, the sectional curvature reflects the *second-order* asymptotic behavior of the distance function $d(\gamma(t), \eta(t))$ near $t = 0$ between geodesics $\gamma(t) = \exp_p(tv)$ and $\eta(t) = \exp_p(tw)$, where $\exp(\cdot)$ is the exponential map Ohta (2014).

Definition 2.3 (Exponential map). *Given a tangent vector $v \in T_pM$, there exists a unique geodesic $\gamma^v : [0,1] \to M$ such that $\gamma(0) = p$ and $(\gamma^v)'(0) = v$. The related exponential map is then defined as*

$$\exp_p(v) = \gamma^v(1)$$

Intuitively, you take a tangent vector v at p and walk in that direction for unit time. The point at which you arrive is the exponential map of v starting at p.

The exponential map in the Riemannian case is just the point at unit distance (for $||v|| = 1$) from p in direction v. Hence, $\exp(tv)$ can be thought of as an extension of convex combinations to Riemannian manifolds.

The 0-th order behavior of this distance is the distance between curves itself (which goes to 0); the first order is given by the velocity vectors v and w; the 2-nd order is given by the acceleration vectors, i.e. second derivatives of the curves along the direction of the first derivatives, the velocities. Indeed, the acceleration reflects curvature: in a 2-dimensional curve, the acceleration is the force you exercise in order to keep moving along the curve with a constant speed, with respect to the external space $\mathbb{R}^3$.

The notions of Riemannian curvature and sectional curvature hold exactly the same amount of information.

They induce the notion that of Ricci curvature, that is a weaker but crucial notion that in a way "summarizes" the information given by the other two. In particular, it keeps the information given by the *trace* of the other two tensors.[3]

Definition 2.4 (Ricci curvature tensor). *Given a Riemannian manifold (M, g) and vector fields $X, Y \in \mathfrak{X}(M)$, the Ricci curvature tensor of M is a map $\mathrm{Ric} : \mathfrak{X}(M) \times \mathfrak{X}(M) \to \mathbb{R}$ defined as*

$$\mathrm{Ric}(X, Y) = \sum_{i=1}^{n} (R(e_i, X)Y, e_i) \quad \in \mathbb{R}$$

I.e. the Ricci curvature is the Riemannian curvature tensor contracting on the first and third component. Hence, the Ricci curvature is an object eating two vectors and spitting out one number. We indeed look at the Ricci curvature as an object similar to an inner product (see the beginning of Section **??**).

Ohta (2014) defines the Ricci curvature along a direction v, starting from the sectional one. This is like feeding the Ricci tensor with the same vector v twice.

Definition 2.5 (Ricci curvature along v). *Given a unit vector $v \in T_pM$ we define the* Ricci curvature *of v as the trace of the sectional curvature $\mathcal{K}(v, \cdot)$, i.e.*

$$\mathrm{Ric}(v) := \sum_{i=1}^{n-1} \mathcal{K}(v, e_i) \quad \in \mathbb{R}$$

[3] In general, when we summarize information given by a matrix, we can do it with a spectrum of coordinate-independent objects. At the boundaries of this spectrum there are the determinant and the trace. All the objects between them are the other coefficients of the characteristic polynomial of the matrix.

where $\{e_i\}_{i=1}^{n-1} \cup v$ *is an orthonormal basis of* T_pM. *Hence,* $\mathrm{Ric}(v) = \mathit{Tr}(\mathcal{K}(v, \cdot))$.

This gives us a geometric interpretation of the Ricci curvature, as the sum of the sectional curvatures of the 2-planes spanned by $(v, e_1), \ldots, (v, e_{n-1})$.

In general, from both definitions and from how we introduced it, it is clear that the Ricci curvature is a sum/average of the sectional or Riemannian curvature. As such, it can not give information that has to do with surface as the sectional curvature did: it sums the information of that surface with all the other surfaces induced by that tangent space. Indeed, **the Ricci curvature has to do with *volumes***. This is formalized by the following result.

THEOREM 2.6 (Bishop-Gromov volume comparison for $K = 0$). *If we assume* $\mathrm{Ric} \geqslant 0$, *then for any* $p \in M$ *and* $0 < r < R$

$$\frac{\mathrm{vol}_g(B(x, R))}{\mathrm{vol}_g(B(x, r))} \leqslant \frac{\int_0^R t^{n-1}\,\mathrm{d}t}{\int_0^r t^{n-1}\,\mathrm{d}t} = \frac{R}{r} \tag{4}$$

We are not interested in the specific form of the bound in inequality 4; it is enough to consider that it is a function of K and it implies that a lower bound on the Ricci curvature controls the ratio of the volumes of two balls on the manifold, hence controls the volume vol_g.

When we refer to properties of the Ricci curvature of a whole manifold M, we mean that the property holds for all unit vectors $v \in TM$, e.g. $\mathrm{Ric} \geqslant K$ means $\mathrm{Ric}(v) \geqslant K \;\forall v \in TM$.

CHAPTER 3

Optimal transport

Optimal transport provides a robust approach to Ricci curvature lower bounds: the line of research linking ot to Riemannian geometry started in the 2000s. As we will see, optimal transport naturally inherits the geometric structure of the underlying space, and especially the Ricci curvature is crucial for describing optimal transport on Riemannian manifolds.

In general, optimal transport allows to define the metric structure on the space of measure, talking about how to optimally transport one probability measure onto another and thus about a *distance* between measures, which is indeed inherited by the distance on the underlying space. It has many applications like in statistics, physics and machine learning. Ambrosio et al. (2021) Villani et al. (2008) Ohta (2014)

1. Measure and probability theory

Here we collect some preliminary definitions useful for later.

Definition 1.1 (Hausdorff outer measure). *Given a metric space (X, g) and a set $A \subset X$, we define the Hausdorff outer measure $\mathscr{H}_*^d$ as*

$$\mathscr{H}_*^d(A) := \inf \left\{ \sum_{i=0}^{\infty} \big(\mathit{diam}(U_i)\big)^d : A \subseteq \bigcup_{i=0}^{\infty} U_i \right\}$$

where

$$\mathit{diam}\, U := \sup\{g(x,y) : x, y \in U\} \quad \mathit{diam}\, \varnothing := 0$$

The Hausdorff (outer) measure for $d = n$ in $\mathbb{R}^n$ is equivalent to the Lebesgue measure, they differ for a constant. The Hausdorff measure generalizes the Lebesgue measure as it allows to define lower dimensional volumes, for $d < n$. For instance, we can measure the lengths of a surface immersed in $\mathbb{R}^3$ with the Hausdorff measure of $\mathbb{R}^3$ (with the Euclidean distance) for $d = 2$.

Definition 1.2 (Borel σ-algebra). *Given a topological space $(X, \mathcal{T})$, the Borel σ-algebra is*

$$\mathscr{B}(X) := \sigma(\mathcal{T})$$

i.e., it is the σ-algebra generated by the collection of all open sets of X, $\mathcal{T}$. This means it is the smallest σ-algebra containing $\mathcal{T}$, equivalently, the intersection of all σ-algebras containing $\mathcal{T}$.

Definition 1.3 (Borel set). *A set $A \in \mathscr{B}(X)$ is a Borel set or Borel measurable set.*

Definition 1.4 (Borel function). *A function $f : X \to Y$ topological spaces is a Borel function if $f^{-1}(A)$ is a Borel set for any open set A.*

Then, the ***Hausdorff measure*** $\mathscr{H}^n$ is the Hausdorff outer measure restricted to Borel sets.

Definition 1.5 (Probability measure). *A probability measure on a σ-algebra $\mathcal{A}$ of a set X is a measure $\mu : \mathcal{A} \to [0, 1]$ such that $\mu(X) = 1$.*

We denote with $\mathscr{P}(X)$ the set of probability measures on the space X.

Given a probability measure μ we can see it as the *law* or *probability distribution* of a random variable X such that

$$\Pr(X \in A) = \mu(A)$$

Definition 1.6 (Push forward measure). *Given a Borel function $f : X \to Y$, we define the push forward operator $f_\# : \mathscr{M}(X) \to \mathscr{M}(Y)$, where $\mathscr{M}$ is the set of σ-additive functions (measures), by*

$$f_\#\mu(B) = \mu(f^{-1}(B)) \quad \forall B \in \mathscr{B}(Y)$$

and call $f_\#\mu$ push forward measure.

It is a way to obtain a measure in a target set Y given a measure in a starting set X, through a function f.

Proposition 1.7 (I don't know how it's called). [1] *From the above definition we get that*

$$\int g \ \mathrm{d}f_\#\mu = \int g \circ f \ \mathrm{d}\mu$$

[1] EB: change of variables formula

We first prove the statement for characteristic functions. Indeed we have that if g is the characteristic function of a (Borel) set $B \subset X$, $g = \chi_B$, then the equation above develops as

$$
\begin{aligned}
RHS &= \int \chi_B \ \mathrm{d}f_{\#}\mu(B) = f_{\#}\mu(B) \\
LHS &= \int \chi_B(f(x)) \ \mathrm{d}\mu(x) = \mu(f^{-1}(B))
\end{aligned}
$$

*where the equality in RHS is because that integral means we are just measuring the set B; the LHS follows from $\chi_B(f(x)) = f^{-1}(B)$ as we are considering the image of f only for $x \in B$, hence the pre-image of B. Then (**??**)follows.*

Then, we have that by linearity, it is true also for simple functions, and by taking the limit for any function.

Theorem 1.8 (Disintegration theorem - high level). *Given a measure $\theta := f_{\#}\sigma$ there exists a family $\{\sigma_x\}_{x\in X}$ such that*

$$\sigma = \int_X \sigma_x \ \mathrm{d}\theta$$

and such family is θ-a.e. unique and we call σ_x conditional probabilities, $\sigma_x = E[\sigma|\{f = x\}]$.

The following notations are equivalent:

- $\sigma = \int_X \sigma_x \ \mathrm{d}\theta$
- $\sigma = \sigma_x \otimes \theta$
- $\sigma(\mathrm{d}x, \mathrm{d}y) = \sigma_x(\mathrm{d}y)\theta)\,\mathrm{d}x)$

where the last one follows from

$$\int_{X\times Y} f(x,y) \ \mathrm{d}\sigma(x,y) = \int_X \Big(\int_Y f(x,y) \ \mathrm{d}\sigma_x(y)\Big) \ \mathrm{d}\theta(x)$$

2. Formulations of optimal transport

- This and the next two sections are all based on Ambrosio et al. (2021) -

Consider probability measures $\mu \in \mathscr{P}(X), \nu \in \mathscr{P}(Y)$ and a Borel cost function $c(x,y) : X \times Y \to [0,\infty]$ that represents the cost of shipping a unit of mass from x to y.

First, here is Monge's fromulation of the optimal transport problem.

$$\inf\left\{\int_X c(x,T(x))\ \mathrm{d}\mu(x) : T : X \to Y \text{ Borel}, T_{\#}\mu = \nu\right\} \tag{M}$$

where T is a ***transport map*** and it is a way of carrying mass from X to Y.

Remark. Transport maps that bring μ in ν exist whenever μ has no atom. That is, if μ is a Dirac measure concentrated in one point x_0, and ν is not Dirac or it is Dirac concentrated in > 1 points y_0, y_1, then there is no function that brings the mass in x_0 into y_0, y_1.

Now, to look at the Kantorovich problem we first give the definition of ***transport plan***.

Definition 2.1 (Transport plan). *Given $\mu \in \mathscr{P}(X)$ and $\nu \in \mathscr{P}(Y)$, define*

$$\begin{aligned}\Gamma(\mu,\nu) := \{\pi \in \mathscr{P}(X \times Y) : \pi(A \times Y) = \mu(A),\\ \pi(X \times B) = \nu(B) \quad \forall \text{ Borel } A, B\}\end{aligned}$$

Transport plans are another way of carrying mass from X to Y, and in particular $\pi(A \times B)$ is the mass that was in A and has been sent to B.

The definition speaks well with the joint probability distribution of two random variables, whose laws are μ and ν. Indeed, the requirements in the definition are like marginalizations of π. Also, given coordinate projections of π, $p_X : X \times Y \to X$, $(x,y) \mapsto x$ and $p_Y : X \times Y \to Y$, $(x,y) \mapsto y$, being a transport plan is equivalent to

$$p_{X\#}\pi = \mu, \quad p_{Y\#}\pi = \nu$$

Kantorovich formulation of the optimal transport problem is then:

$$\inf\left\{\int_{X\times Y} c(x,y)\ \mathrm{d}\pi(x,y) : \pi \in \Gamma(\mu,\nu)\right\} \tag{K}$$

The optimal transport cost is denoted as

$$\mathscr{C}(\pi) := \int_{X\times Y} c(x,y)\ \mathrm{d}\pi(x,y) \tag{5}$$

Now, we turn to comparing the two formulations.

The relationship between transport maps and transport plans is summarized as follows. Given a transport map T, we can define the transport plan

$$\pi_T := (\mathrm{id} \times T)_{\#}\mu \tag{6}$$

where $(\mathrm{id} \times T) : X \to X \times Y$ is the map $x \mapsto (x, T(x))$. Then, by indicating with $\mathscr{C}$ the cost attained by a transport map/plan, we have

$$\mathscr{C}(\pi_T) \ = \int_X c(x, T(x)) \, \mathrm{d}\mu(x) = \ \mathscr{C}(T)$$

Thus,

$$\inf\nolimits_{(M)} \geqslant \inf\nolimits_{(K)}$$

meaning that the Kantorovich formulation is more general than the Monge one. This is due to the fact that a transport plan always exists, instead a transport map could not exist as we said above. Indeed, transport plans are a more flexible tool than transport maps.

Here is a list of the main advantages of (K) formulation:

(i) The class of transport plans is not empty (since $\mu \times \nu$ is always an option) and convex (can be checked with elementary operations on functions). Then, since the map $\pi \mapsto \mathscr{C}(\pi)$ is affine (it's an operation on the differential, *the Lebesgue integral is linear in the measure by definition*, can be checked through simple functions), also the set of plans that attain the minimum is convex.

(ii) The operation between plans is clearly more symmetric. In fact, the switching map $(x, y) \mapsto (y, x)$ induces a mapping from $\Gamma(\mu, \nu)$ to $\Gamma(\nu, \mu)$, giving an easy way to invert a plan. Instead, we need invertibility of T to ensure that if we have a map from μ to ν we also have the inverse map.

(iii) Existence of minimizers is ensured under a mild condition on the cost function (lower semicontinuity), whereas for (M) optimality required c to be convex and nondecreasing.

Finally, we add a third formulation, the *dynamic formulation.* As usual, some preliminary definitions first. The *length* of a curve for $t \in [0, 1]$ is

$$l(\gamma) = \int_0^1 |\gamma'(t)| \, \mathrm{d}t$$

which resembles "velocità * tempo".

Then, given a distance d we can characterize geodesics as curves $\gamma(t)$ such that $l(\gamma(t)) = d(\gamma(0), \gamma(1))$.

Definition 2.2 (Action of a curve). *The (possibly infinite) action of a curve* $\gamma : [0, 1] \to X$ *is*

$$\mathcal{A}(\gamma) = |\gamma(t)'|^2$$

Also, by Holder we have that $\mathcal{A}(\gamma) \geqslant l^2(\gamma) \geqslant d^2(\gamma(1), \gamma(0))$ *and* γ *is a geodesic if and only if* $\mathcal{A}(\gamma) = d^2(\gamma_0, \gamma_1)$[2]

In brief, the dynamic formulation of OT is the following problem

$$\min\left\{ \int_X \mathcal{A}(\gamma)\, \mathrm{d}\eta(\gamma) : \eta \in \mathscr{P}(X), \gamma(0)_{\#}\eta = \mu, \gamma(1)_{\#}\eta = \nu \right\} \tag{dyn}$$

The main difference with previous formulations is that Monge and Kantorovich formulations can be referred to as *static* ones, instead this is dynamic. This is because in (K) and (M), we only care of where we bring each mass from X to Y (which is modeled either through a map or a plan), here, we also care of how we do it, and we look for the optimal way. Hence if in the examples before we looked at the space X and at how mass was moved there, here I imagine it better on $\mathscr{P}(X)$, looking at geodesics joining two probability measures.

A *geodesic metric space* is a space such that for each pair of points in it there exists a geodesic joining them. If X is a geodesic metric space, then $\min_{(dyn)} = \min_{(K)}$.

3. Metric side of optimal transport

Here we want to *use* the optimal transport problem to endow the space of measures with a natural metric structure, once we have a metric space (X, d). This will be done by "lifting" metric properties from X to $\mathscr{P}_2(X)$. The main thing we need to do so is to define a distance on $\mathscr{P}_2(X)$, which will be the Wasserstein distance.

First, we put ourselves in the following setting. We consider the space of measures $\mathscr{P}_2(X)$, defined as follows.

Definition 3.1 (The Wasserstein space $\mathscr{P}_p$). *Given a metric space* (X, d) *and* $p \in [1, \infty)$*, we define*

$$\mathscr{P}_p(X) := \left\{ \mu \in \mathscr{P}(X) : \int_X d^p(x, x_0)\ \mathrm{d}\mu(x) < \infty \right. \\ \left. \textit{for some } x_0 \in X \right\}$$

This is hence the *set of measures* for which the integral of the squared distance between two points is bounded.

[2] Let us recap equivalent characterizations of geodesics: geodesics are curves that *have 0 acceleration; locally minimize length; whose length is equal to the distance between extremes; whose action is equal to the squared distance between extremes.*

Now we define the ***Wasserstein distance***, a core concept in optimal transport. Looking at the transport cost (5), we have seen it can be interpreted as the road we need to walk from each point mass of μ to ν and the cost we pay. It seems similar to a distance, however, that object does not satisfy the axioms of a distance function. By defining the cost function $c(x, y)$ of in terms of a distance d, the transport cost (5) becomes a *distance* between μ and ν.

Definition 3.2 (Wasserstein distance in $\mathscr{P}_2(X)$). *Given two measures $\mu, \nu \in \mathscr{P}_2(X)$, we define*

$$W_2^2(\mu, \nu) := \min \left\{ \int_{X \times X} d^2(x, y) \ \mathrm{d}\pi(x, y) : \pi \in \Gamma(\mu, \nu) \right\}$$

First, we look at the following theorem.

Theorem 3.3. *$(\mathscr{P}_2(X), W_2)$ is a metric space and the embedding $X \ni x \mapsto \delta_x \in \mathscr{P}_2(X)$ is isometric, i.e. $d(x, y) = W_2(\delta_x, \delta_y)$.*

We prove this theorem, i.e. prove that W_2 is a distance.

(i) **Finiteness of W_2.** Choose $\pi = \mu \times \nu$ and consider

$$\begin{aligned} d^2(x, y) &\leqslant (d(x, x_0) + d(y, x_0))^2 \\ &\leqslant d^2(x, x_0) + d^2(y, x_0) + 2d(x, x_0)d(y, x_0) \\ &\leqslant 2(d^2(x, x_0) + d^2(y, x_0)) \end{aligned}$$

where the first line is by triangle inequality (d is a distance) and the second by simple calculus. Then,

$$\begin{aligned} W_2^2(\mu, \nu) &\leqslant \int_{X \times X} d^2(x, y) \ \mathrm{d}\pi(x, y) \\ &\leqslant \int_{X \times X} 2(d^2(x, x_0) + d^2(y, x_0)) \ \mathrm{d}\pi(x, y) \\ &= 2 \int_X d^2(x, x_0) \ \mathrm{d}\mu(x) + 2 \int_X d^2(y, x_0) \ \mathrm{d}\nu(y) < \infty \end{aligned}$$

(ii) **Symmetry.** From the proof of Brenier, Knott-Smith theorem (theorem 5.2 in Ambrosio et al. (2021)): we define the function $i : (x, y) \mapsto (y, x)$ and we have that $d(x, y) = d(y, x)$. We can check that $i_\# : \Gamma(\mu, \nu) \to \Gamma(\nu, \mu)$, indeed

$$\Gamma(\mu, \nu) \ni \pi(x, y) \mapsto i_\#\pi(x, y) = \pi(i^{-1}(x, y)) = \pi(y, x) \in \Gamma(\nu, \mu)$$

by definition of push forward measure. Hence, starting from the other problem $W_2(\nu, \mu)$ and applying equation (1.7)

$$W_2^2(\nu, \mu) = \min \int d^2(y,x)\ \mathrm{d}\pi(y,x) = \min \int d^2(y,x)\, d\, i_{\#}\pi(x,y) = \min \int d^2(x,y)\, \mathrm{d}\pi(x,y) = W_2^2(\mu,$$

(*iii*) **Non-degeneracy.** One implication follows from non-degeneracy of the distance d, with $\pi = (\mathrm{id} \times \mathrm{id})_{\#}\mu$. We get the other implication by considering that $W_2(\mu, \nu) = 0$ implies $x = y$ for π a.e. (x, y), and thus

$$\int_X f(x)\ \mathrm{d}\mu(x) = \int_{X \times X} f(x)\ \mathrm{d}\pi(x,y) = \int_{X \times X} f(y)\ \mathrm{d}\pi(x,y) = \int_X f(y)\ \mathrm{d}\nu(y)$$

the first equality follows from the definition of π, and from $\int f\ \mathrm{d}\mu = \int f\ \mathrm{d}\nu$ we get $\mu = \nu$.

(*iv*) **Triangle inequality.** To prove triangle inequality, namely that $W_2(\mu, \nu) \leqslant W_2(\mu, \sigma) + W_2(\sigma, \nu)$, we consider the Dudley lemma stated in the appendix (A) and the triangle inequality property for the L^2 norm.

We consider 3 spaces X_1, X_2, X_3 and $\mu \in \mathscr{P}(X_1), \sigma \in \mathscr{P}(X_2), \nu \in \mathscr{P}(X_3)$. Given any optimal plan from μ to σ π^{12} and from σ to ν π^{23}, there exists an optimal plan π^{13} from μ to ν such that its marginalizations are the other 2 optimal plans. Then

$$\begin{aligned} W_2(\mu, \nu) &= ||d(x_1, x_3)||_{L^2(\pi^{13})} \\ &\leqslant ||d(x_1, x_2)||_{L^2(\pi^{13})} + ||d(x_2, x_3)||_{L^2(\pi^{13})} \\ &= ||d(x_1, x_2)||_{L^2(\pi^{12})} + ||d(x_2, x_3)||_{L^2(\pi^{23})} \\ &= W_2(\mu, \sigma) + W_2(\sigma, \nu) \end{aligned}$$

where the first inequality follows from triangle ineq. property of the L^2 norm, the equality from the fact that projecting p^{ij} is the identity map for (x_i, x_j) + def. of push forward.

Now, let us make some remarks and intuition on the Wasserstein distance.

We notice that by setting $c(x, y) = d^2(x, y)$, and considering two probability measures with the same support, the plan π attaining W_2^2 is exactly the plan solving the optimal transport problem K. The cost function is of course continuous, hence such a plan exists. Indeed, we can think of the formula for the W_2 distance as follows:

- The plan π tells us *which percentage of the mass x goes into the set y*. It gives us a rule to send mass from μ to ν. This is clear when π is induced by a transport map (maps each x into a y). In this case, we have

$$\pi_x = \delta_{T(x)}$$

- The distance d tells us how much we pay for the transport of each x into its y.

Indeed, in general, we can look at π under the lens of the disintegration theorem 1.8.

Given π such that $\mu = p_{\#}\pi$ (where p is the projection map) there exists a family of conditional probabilities π_x such that

$$\pi = \int \pi_x \, \mathrm{d}\mu$$

and

$$\int f(x,y)\, \mathrm{d}\pi(x,y) = \int \Big(\int f(x,y)\, \mathrm{d}\pi_x(y) \Big)\, \mathrm{d}\mu(x)$$

Also, π_x are supported on $p^{-1}(x) = \{x\} \times Y$ and can be viewed as probability measures on Y.

It follows that can represent π as the **measure-valued transport map** $x \mapsto \pi_x$.

Also, here are some examples of optimal transport problems.

Example. If $\mu = \frac{1}{n}\sum_i \delta_{x_i}$ and $\nu = \frac{1}{n}\sum_i \delta_{y_i}$, then we send each x_i into the closest y_i (recall drawing by Vitillaro) and pay a cost $|x_i - y_i|^2$. Hence $W_2^2(\mu,\nu) = \frac{1}{n}\sum_i |x_i - y_i|^2$.

Example. If μ is concentrated in x and ν in two atoms y_1 and y_2, then $\pi_x = \frac{1}{2}\delta_{y_1} + \frac{1}{2}\delta_{y_2}$.

Example. If instead the mass of ν is a segment S or a bounded surface S, we set $\pi_x = \mathscr{H}^1|_S$ or $\pi_x = \mathscr{H}^2|_S$ respectively, where $\mathscr{H}$ is the Hausdorff measure.

3.1. Completeness and convergence in $(\mathscr{P}_2(X), W_2)$. Here we look at completeness and the notion of convergence in $\mathscr{P}_2(X)$, two other notions that are lifted from X. First, let us look at completeness, by assuming (X,d) is complete.

In a nutshell, we define our Cauchy sequence of measures (μ_n) as marginalizations of a plan π_∞ (which exists by Dudley's lemma) and get convergence from the convergence of the sequence (p_n) of projection functions, which are Cauchy in a complete space.

Let (μ_n) be a Cauchy sequence with respect to the W_2 distance. We consider an iterated version of Dudley's lemma (see proposition 8.6 in Ambrosio et al. (2021)), which allows to define our sequence of measures as marginals of a measure $\pi_n \in \mathscr{P}(X_1 \times X_2 \times \dots X_n)$ for $1 \leqslant n \leqslant N$. When $N = \infty$, we generalize to the case $n = \infty$ by considering

$$\pi_\infty \in \mathscr{P}(\mathbb{X}) \quad \text{such that} \quad p_{\#}^{1,\dots,n}\pi_\infty = \pi_n$$

Then the lemma lets us define

$$\mu_n = (p_n)_\# \pi_\infty$$
$$\theta^{n,n+1} = (p_n, p_{n+1})_\# \pi_\infty$$

where μ_n are "simple" measures and $\theta^{n,n+1}$ are plans.

Then, we observe that p_n is a Cauchy sequence (non ho ben capito perché). The space in which p_n lives, $L^2(\mathbb{X}, \mathscr{B}_\infty, \pi_\infty, X)$ is complete by completeness of (X, d), hence p_n converges. We call p_∞ its limit and define

$$\mu_\infty := (p_\infty)_\# \pi_\infty$$

Then, we have

$$W_2^2(\mu_n, \mu_\infty) \leqslant \int d^2(p_n, p_\infty) \, \mathrm{d}\pi_\infty \to 0$$

i.e. μ_n converges to μ_∞ in W_2, hence $\mathscr{P}_2(X)$ is complete.

Now, we want to characterize convergence in $\mathscr{P}_2(x)$ according to W_2. We consider the following definition of weak convergence of measures

$$\int f \ \mathrm{d}\mu_n \to \int f \ \mathrm{d}\mu$$

for f continuous and bounded. Then, the result is that

(i) $\mu_n \xrightarrow{W_2} \mu \implies$ convergence of moments $\forall x$ and $\mu_n \to \mu$ weakly

(ii) convergence of moments for some x and $\mu_n \to \mu$ weakly $\implies \mu_n \xrightarrow{W_2} \mu$

The proof uses that the definition of weak convergence above is equivalent to the same statement for Lipschitz functions, and weak convergence as defined by the $\liminf$ ($\limsup$) of the measure of open (closed) sets, and for moments it uses the definition of W_2 of course.

3.2. Benamou-Brenier formula. Related to the dynamic formulation of ot is this formula which expresses the Wasserstein distance in terms of action of a curve. Indeed, we consider the quadratic action

$$\mathcal{A}(\mu, v) = \int_{\mathbb{R}^n} |v|^2 \ \mathrm{d}\mu(t)$$

and it turns out that

$$W_2^2(\mu, \mu) = \min \left\{ \int_0^1 |v(t)| \, \mathrm{d}\mu(t) : \frac{\mathrm{d}}{\mathrm{d}} \mu(t) + \operatorname{div}(v(t)\mu(t)) = 0 \right\} \tag{7}$$

Wasserstein distance under this formula can be interpreted as follows. It considers the problem of minimizing the action of a curve, i.e. the squared norm of its velocity, under a constraint. The idea is essentially that the squared velocity, times the mass (represented here by the density/measure of x), gives the kinetic energy of moving along the curve ($E = mc^2$). We are minimizing such energy; in other words, we are looking for the lowest possible effort to move from μ_0 to μ_1. The constraint is given by the *continuity equation*, which ensures that mass is preserved when moving from one extreme to the other. When the minimum is 0, the curve attaining it is the geodesic from μ_0 to μ_1, the constant speed curve (with 0 acceleration) that joins the two extremes. It is the length minimizing curve (at least locally), as we have seen. In fact, clearly, minimizing length is the same as minimizing energy (action), as is clear from their definitions (that geodesics can be characterized in this way is due to Otto, then Ambrosio, Gigli and Savaré pointed out that a metric notion of geodesics was more practical).

In this case, transporting mass from μ_0 to μ_1 is seen as a *flow* over time. As proved in proposition 17.9 of Ambrosio et al. (2021), there is a two-way correspondence between solutions to the continuity equation and absolutely continuous curves wrt W_2.

Remark (Continuity equation). We just want to point out that the continuity equation asks that mass is preserved during transportation. Also, proposition 16.3 in Ambrosio et al. (2021) states that a measure μ solves the continuity equation iff given any curve φ, the map $t \mapsto \int \varphi \, \mathrm{d}\mu_t$ is absolutely continuous and its derivative is $\int \nabla \varphi v_t \, \mathrm{d}\mu_t$ it is equivalent to absolute continuity of a curve, with v being its derivative.

Refer to section **??** for an explanation of the continuity equation.

4. Convexity of the entropy

This section is devoted to proving the following statement: the relative entropy is convex in $(\mathscr{P}_2(\mu_t), W_2)$.

First, we define the entropy.

Definition 4.1 (Relative entropy). *Given (X, g) with a canonical reference measure vol_g and a measure $\mu \in \mathscr{P}(X)$ such that $\mu = \varrho \, \mathrm{vol}_g$, we define the function $\mathcal{H}_g : \mathscr{P}(X) \to [0, \infty]$ as*

$$\mathcal{H}_g(\mu) := \int_X \varrho \log \varrho \, \mathrm{dvol}_g$$

(entropy is set to $+\infty$ if $\mu \neq \varrho \, \mathrm{vol}_g$).

We can check that $\mathcal{H}(\mu) \geqslant 0$ by considering that the function $h(x) = x \log x$, the *density energy*, is convex and applying Jensen.

Remark (Relative entropy and KL divergence). Given strict convexity of the integrand $h(\varrho)$, we have $\mathcal{H}_g(\mu) = 0 \iff \varrho = 1$, which is like saying that the distribution induced by μ is a Uniform distribution. Indeed, the relative entropy is equivalent to the Kullback-Leibler (KL) divergence between a probability measure μ and the Uniform one.

Now, let us restrict ourselves to the Euclidean case $X = \mathbb{R}^n$ for a moment. We can set

$$\mathrm{vol}_g = \mathscr{L}^n$$

Remark. From now on, our manifold is $\mathscr{P}_2(X)$, a space of probability measures. So our "points" are probability measures and we look at functions taking measures as input, like $\mathcal{H}$.

To study convexity of the entropy on a Riemannian manifold, we consider geodesics and study the convexity of a function *along geodesics*, just like in the Euclidean case we consider convex combinations (segments) of two points.

Let us define what convexity along geodesics means.

The main conceptual point is that until now we have talked about two probability measures (μ and ν then, μ_0 and μ_1 now), a transport map and a transport plan between them. Now, we introduce a third object, a *curve* between them. In particular, we look at the curve that is the (unique constant speed) geodesic between them. This is just a measure-valued function.

It follows that the notions of transport map and transport plan can be extended to *interpolated* maps and plans which depend on the time $t \in [0, 1]$.

More precisely, we consider two probability measures $\mu_0 = \varrho \mathscr{L}^n$ to μ_1. By Brenier theorem (theorem 5.2 in Ambrosio et al. (2021)), there exists an optimal transport map $T = \nabla \phi$ for ϕ lower semicontinuous, convex and μ_0-a.e. differentiable.

Moreover, we can define the *interpolated transport map*

$$T_t := (1-t)\,\mathrm{id} + tT \tag{8}$$

with

$$T_t = \nabla \phi_t, \quad \phi_t = (1-t)\frac{|\cdot|^2}{2} + t\phi$$

which we don't care about now.

Notice that this definition speaks well with the definition of a transport plan π_T induced by T (equation 6). Indeed, we extend π_T to vary along t, i.e. we define an *interpolated transport plan*

$$\pi_{T,t} = (\mathrm{id} \times T_t)_{\#}\mu_0$$

and we notice that

$$\pi_{T,0} = \mu_0$$

$$\pi_{T,1} = T_{\#}\mu_0 = \mu_1$$

(by definition of T).

By corollary 10.10 in Ambrosio et al. (2021), we have that

$$\mu_t := (T_t)_{\#}\mu_0$$

is the unique constant speed geodesic between μ_0 and μ_1.

The main result is that that the function $\mathcal{H}(\mu_t)$ is convex, i.e. **the relative entropy is convex along geodesics**. This is stated formally in theorem 15.16 in Ambrosio et al. (2021).

In brief, the proof uses mainly these ingredients: we define the entropy of $\mu_t = \varrho_t \mathscr{L}^n$ by considering the interpolated density ϱ_t, which is equal to a function of ϱ and T_t

$$\varrho_t = \frac{\varrho}{\det \nabla T_t} \circ (T_t)^{-1}$$

We can manipulate the entropy as follows

$$\mathcal{H}(\mu_t) = \int_{T_t(D_0)} U\left(\frac{\varrho}{\det \nabla T_t} \circ (T_t)^{-1}\right) \,\mathrm{d}y = \int_{D_0} U\left(\frac{\varrho}{\det \nabla T_t}\right) \det \nabla T_t \,\mathrm{d}x$$

Now, ignoring how D_0 is defined, we have that geodesic convexity of $\mathcal{H}(\mu_t)$ follows from convexity of the function

$$f(t) := U\left(\frac{\varrho(x)}{\det \nabla T_t(x)}\right) \det \nabla T_t(x)$$

This function can be seen as the composition $b \circ a$, with

$$a(t) := (det \nabla T_t(x))^{\frac{1}{n}}, \quad b(z) := U\left(\frac{\varrho(x)}{z^n}\right) z^n \tag{9}$$

a is concave by lemma 15.15 of Ambrosio et al. (2021), which is a pure algebraic result; b is convex thanks to an assumption on U (that it is $(MC)_n$, we ignore this here). Thus, $b \circ a$ is convex.

Then, geodesic convexity of the entropy follows from convexity of this integrand f.

Additionally, we can relate the above theory to these definitions (I don't remember where I took them). We can say that μ_t is the unique *Wasserstein geodesic*, i.e. the unique constant speed geodesic satisfying the following.

Definition 4.2 (W geodesic). [3] *Given two measures θ, σ, a curve γ_t is the (Wasserstein) geodesic from θ to σ if*

$$\gamma_0 = \theta,\ \gamma_1 = \sigma$$
$$W_2(\gamma_s, \gamma_t) = |t - s|\, W_2(\theta, \sigma) \quad \forall t, s \in [0, 1]$$

And geodesic convexity is also called *Wasserstein convexity* and can be stated as follows.

Definition 4.3 (W convexity). *Given a function f, and two measures θ, σ and their*[4] *geodesic γ_t, f is* ***convex along*** *γ_t if*

$$f(\gamma_t) \leqslant (1 - t) f(\theta) + t f(\sigma) \quad \forall t \in [0, 1]$$

Indeed, theorem 15.16 in Ambrosio et al. (2021) shows this result for $f = h(\varrho_t)$ integrand of $\mathcal{H}$. Hence a more visual formulation of the result is that

$$\mathcal{H}_g(\mu_t) \leqslant (1 - t)\mathcal{H}_g(\mu_0) + t\mathcal{H}_g(\mu_1) \quad \forall t \in [0, 1]$$

obtaining geodesic or W convexity of the entropy.

EB: di solito W convexity significa che f è convessa lungo ogni geodetica che connette le due misure. Come ribadito sopra, spesso la geodetica non è unica.

Remark. I would like to point out how some of the concepts we discussed so far are related to one another. In a HIGH LEVEL manner.

The problem of optimal transport, especially under the dynamic formulation (Benamou-Brenier), allows us to endow the space of probability measures with a Riemannian structure, namely geometric notions, with the W distance. This allows to talk about geodesics, indeed the essence of the dynamic formulation of the problem. This allows also to get the result that entropy is convex along geodesics. Moreover, since the seminar work of Otto and others, it is known the result discussed below, i.e. that the heat flow is the gradient flow of the entropy in the W_2 space of measures, which "has been the starting point

[3] EB: giusto un chiarimento (probabilmente inutile): γ_t è una curva di misure, ovvero una mappa $[0, 1] \to P_2$ che ad ogni t associa una misura γ_t

[4] EB: non è nesessariamente unica la geodetica che connette due misure

for many developments in evolution equations, probability theory and geometry" Maas (2011). These are all related because in the optimal transport problem we are minimizing the energy, entropy is an energy, and convexity of the entropy ensure the uniqueness of a minimizer. this is clearly related to the gradient flow of the entropy. We could conclude that the heat semigroup is the geodesic the ot problem looks for under the W_2 distance.

Later, we will see how this are related to the Ricci curvature which is somehow implicit in $\mathscr{P}_2(\mathbb{R}^n)$ as $\mathrm{Ric} = 0$, and becomes explicit on the Riemannian manifold. Another punto di collegamento è che la convessità dell'entropia è legata al Jacobiano, che ha a che fare con come il volume viene distorto quando la massa è preservata, cioè sotto all'equzione di continuità, ed è quindi chiaro che Ricci, che regola la distorsione del volume, sia legato a tutto ciò.

CHAPTER 4

Flows

Here we go over the prerequisits for chapter 19 of Ambrosio et al. (2021).

1. Notation

$x(t)$ represents a curve $x : [0, T] \to X$, $x_t \in X$ are its images or in general points of X where we want to emphazise the possible parametrization with t. $x'(t)$ and $\frac{\mathrm{d}}{\mathrm{d}t}x(t)$ are derivatives of the curve wrt time, where in the first case notation is less precise but it is generally used interchangeably, when there is only a time variable. Let me also clarify that $\nabla f(x, t)$, instead, refers to the *spatial* derivative of the function, wrt x. It is a vector indicating direction of ascent of f in the space X.

2. Heat equation

Let us look at what divergence and the Laplacian operator are first.

From an high level perspective, the divergence of a vector field is the extent to which the vector field is locally behaving as a source or a sink, a measure of its “outgoingness”. For instance, consider the velocity of air at each point, defining a vector field. While air is heated in a region, it expands in all directions, and thus the velocity field points outward from that region. The divergence of the velocity field in that region would thus have a positive value.

Definition 2.1 (Laplacian operator). *The Laplacian operator Δ is a second-order differential operator defined as the divergence of the gradient of f,*

$$\Delta f = \nabla^2 f = \nabla \cdot \nabla f = div_g \nabla f$$

Hence if $f = f(x)$ it is the second derivative, if $f = f(x, y)$ it is the sum of second derivatives

$$\Delta f = \sum_i \frac{\partial^2}{\partial x_i \partial x_i} f$$

We state the heat equation on the lines of these notes, looking at the heat flow.

Consider heat propagating in a region X and let $u(x, t)$ be the temperature at position $x \in X$ and time t. Let $H(t)$ be the total heat contained in X,

$$H(t) = \int_X c \cdot u(t, x) \, \mathrm{d}x$$

where c is some constant containing properties of the material of diffusion.

Then,

$$\frac{\mathrm{d}}{\mathrm{d}t} H(t) = \int_{\partial X} \kappa \nabla u \cdot n \, \mathrm{d}S$$

This is based on Fourier law, which says that heat flows from hot to cold regions proportional to the gradient of the temperature. In fact, κ is this proportion coefficient, n is the outward normal unit vector (which allows to entail where heat is entering or exiting), $\mathrm{d}S$ is the measure of the surface of ∂X, which is the boundary. This equation reads that *the change in heat is given by the integral of heat exiting the boundary of our space.*

Therefore,

$$\int_X c \cdot \frac{\mathrm{d}}{\mathrm{d}t} u(t, x) \, \mathrm{d}x = \int_{\partial X} \kappa \nabla u \cdot n \, \mathrm{d}S \tag{10}$$

Now, we apply the divergence theorem, which says that the total outward flow of a function from a region is given by the sum/integral of all the sources and sinks in the region. in fact, it is 0 when the sources and sinks within the region annhilate each other. Formally,

$$\int_{\partial X} f \cdot n \, \mathrm{d}S = \int_X \nabla \cdot f \, \mathrm{d}x$$

Then, equation 10 above can be improved as

$$\int_X c \cdot \frac{\mathrm{d}}{\mathrm{d}t} u(t, x) \, \mathrm{d}x = \int_{\partial X} \kappa \nabla u \cdot n \, \mathrm{d}S = \int_X \nabla \cdot (\kappa \nabla u) \, \mathrm{d}x = \int_X k \Delta u \, \mathrm{d}x$$

Hence we obtain the heat equation

$$\frac{\mathrm{d}}{\mathrm{d}t} u = \Delta u \tag{11}$$

for $k = 1$. It says that the change in temperature over time is equal to the divergence of the gradient of the temperature, its Laplacian, namely because the change over time is equal to the flow out of the boundary.

2.1. Heat semigroup. What follows in Ambrosio et al. (2021) is a proposition on some properties of the *Riemannian heat semigroup* (family of operators that solve the heat equation on a Riemannian manifold). We merely state them, without going into depth for now:

- the heat semigroup is stochastically complete and self adjoint
- it commutes with the laplacian operator
- we have a regularization estimate on it
- $\int_M g(P_t) f \,\mathrm{dvol}_g \leqslant \int_M g(f)\,\mathrm{dvol}_g$
- if $f \leqslant C$ then $P_t f \leqslant C$ and also for $\geqslant$

Another proposition states that the heat semigroup is smooth in space and time, i.e.

$$||P_t f||_{C^k} \leqslant C||f||_{L^2}$$

and the map $(t, x) \mapsto P_t f(x)$ is smooth in $(0, \infty) \times M$. It follows that $P_t f$ solves the heat equation “in the classical sense”.

Then, let us define the following to get convergence properties of the heat semigroup.

We can identify $P_t f$ with its space-time continuous representative $P_t^* \mu$. We define it by duality:

$$\int_M f \,\mathrm{d}P_t^* \mu := \int_M P_t f \,\mathrm{d}\mu \quad \text{for any } f \in C(M)$$

Notice $P_t^* : \mathscr{P}(M) \to \mathscr{P}(M)$, it is a monotone functional on $C(M)$.

We can apply Riesz representation theorem and state the following.

Proposition 2.2. *There exists a smooth function* $p.(0, \infty) \times M \times M \to [0, \infty)$ *such that*

$$P_t f(x) = \int_M p_t(x, y) f(y) \,\mathrm{dvol}_g(y)$$

and $p_t(x, y) = p_t(y, x)$. *Moreover,*

$$P_t^* \delta_x = p_t(x, \cdot)\,\mathrm{vol}_g$$

This is to say that P_t^* *represents* $P_t f$ *because given a* y *and a function* f, $P_t f(y)$ *is entirely determined by* $P_t^*(y)\delta_x = p_t(x, y)\,\mathrm{vol}_g$.

EB: cos’è $p_t(x, y)$ su $\mathbb{R}^n$?

3. Continuity equation

The continuity equation is a condition we will encounter often in this study. It is a principle of *mass preservation.*

Let ρ_t be a density (or a measure) and v_t a vector field. The continuity equation is

$$\frac{\mathrm{d}}{\mathrm{d}t}\mu_t + \nabla \cdot (\mu_t v_t) = 0 \tag{CE}$$

in a physical sense, it states that the change in density (mass * volume) over time is equal to how much stuff flows in (minus the divergence), in a very high level manner.

The velocity field v is not directly related to the curve μ, I just think of it as *a* velocity field for the objects living on X.

We have used and will use the continuity equation in the following ways:

- Benamou-Brenier formula for W_2: geodesics in the W_2 space of measures are solutions to the continuity equation;
- if a curve μ_t is absolutely continuous (not concentrated), there exists an "optimal" vector field v_t such that (CE) is satisfied and the action of v_t relative to μ_t equals the metric derivative of μ_t
- therefore, in general, we will look for AC curves and for the vector field that equals their metric derivative and satisfies (CE), giving geodesics, solutions to (dyn).
- in view of the paper, we define the tangent space at ρ as the set of all $\nabla\psi$ and we find that there is a unique $\nabla_0\psi$ that satisfies (CE) with ρ!

4. Gradient flow

Gradient flows are based on the notion of subdifferentials. *Subdifferentials* are a way to generalize the concept of gradients for non smooth functions. We indicate them as ∂f. When f is smooth, $\partial f = \{\nabla f\}$. In particular, the Gateaux subdifferential generalizes this concept for directional derivatives. Consider that

$$p \in \partial f \quad \iff \quad \liminf_{t\to 0^+} \frac{f(x+tv) - f(x)}{t} \geqslant (p, v) \quad \forall v \in H$$

Then,

Definition 4.1 (Gateaux subdifferential). *For $x \in Dom(f) \subset H$, the Gateaux subdifferential is*

$$\partial_G f := \left\{ p \in H : \liminf_{t\to 0^+} \frac{f(x+tv) - f(x)}{t} \geqslant (p, v) \quad \forall v \in H \right\}$$

We can look at the set $\partial_G f$ as the set of all those operators that underestimate the directional derivative of f along v, with (v, p) being only the Riesz representation of such operators.

Then, we define the ***gradient flow*** of a function as the set of curves $x(t)$ such that for $g \in \partial_G f, g + x(t) = 0.$.

Definition 4.2 (Gradient flow). *We say that $x : (0, \infty) \to Dom(f)$ is a gradient flow of f if $x \in AC_{loc}((0, \infty); H)$ and*

$$x'(t) \in -\partial_G f(x(t)) \quad \textit{for } \mathscr{L}^1\textit{-a.e.} t \in (0, \infty) \tag{12}$$

We also say that $x(t)$ starts from $\bar{x}$ if $\lim_{t \to 0} x(t) = \bar{x}$

Another useful result is Brézis-Komura theorem (theorem 11.7 in Ambrosio et al. (2021)), which states the following on the existence and shape of gradient flows.

Theorem 4.3 (Brézis-Komura). *Let f be λ-convex, lower semicontinuous and with dense domain. Then, a gradient flow $x(t)$ exists for any $\bar{x} \in Dom(f)$, which are the starting points of $x(t)$. In particular,*

$$x(t) = S_t \bar{x}$$

The family of operators $S_t : Dom(f) \to Dom(f)$ satisfies the following properties

$$\begin{aligned} S_{t+s} =& S_t \circ S_s && \textit{(semigroup property)} \\ |S_t \bar{x} - S_t \bar{y}| \leqslant & \, e^{-\lambda t} |\bar{x} - \bar{y}| && \textit{(contractivity property)} \end{aligned}$$

This means that the gradient flow of f defines a *semigroup* of the function, i.e. that f defines a family of operators S_t that is close under composition (and, in this case, contractive). Think of a semigroup as a family of operators with certain properties. In general, operators that are solutions to an ODE define a group (uc davis notes).

CHAPTER 5

Heat flow, optimal transport and Ricci curvature

1. Heat flow as grad flow of energy functionals

Define the *Dirichlet energy* as

Definition 1.1 (Dirichlet energy on M). *Consider a function $u \in L^2(\mathrm{vol}_g)$, $u : R^n \to \mathbb{R}$. The Dirichlet energy is a* real-valued quadratic functional $D : L^2(\mathrm{vol}_g) \to [0, \infty]$ *defined by*

$$D_g(u) := \frac{1}{2} \int_M |\nabla u|_g^2 \,\mathrm{dvol}_g$$

when u is a section of TM and $D_g(u) = +\infty$ otherwise.

The Dirichlet energy is "a measure of how variable a function is".

EB: potrebbe essere un buon esercizio verificare che l'energia di DIrichlet in $\mathbb{R}$ è un funzionale convesso e lower-semicontinuous in $L^2(\mathbb{R})$

It is convex, lower semicontinuous and with a dense domain, hence the theory of gradient flows (theorem 4.3) applies.

The gradient flow $P_t u$ solves the heat equation, i.e.

$$\frac{\mathrm{d}}{\mathrm{d}t} P_t u = \Delta_g P_t u \quad \text{for} \mathscr{L}^1\text{-a.e. } t \in (0, \infty)$$

Here, the LHS is the derivative of the map $t \mapsto P_t u$ with values on $L^2(\mathrm{vol}_g)$; the RHS is the Laplacian operator on $P_t u$.

Hence, we have found one of the interpretations of the heat semigroup, namely that the heat semigroup is the gradient flow of the Dirichlet energy on L^2.

This follows from the fact that the only possible element in the subgradient of D_g is $-\Delta_g$ (I don't know why), and the gradient flow is by definition x_t such that $x_t' = -\partial_g D_g$. Hence our gradient flow is such that $\frac{\mathrm{d}}{\mathrm{d}t} P_t f = -\partial_g f = \Delta_g f$.

Now, recall the dual of the heat semigroup P_t^*. There are two main results stated in theorem 19.4 for which we are interested in P_t^*. One is that it is an EVI gradient, and we discard it for now as I see it as only a useful tool connecting the other statements. The other interesting result, instead, is Kuwada equivalence, which states the following.

P_t^* is K-*contractive*, i.e.

$$W_2^2(P_t^*\mu, P_t^*\nu) \leqslant e^{-2Kt} W_2^2(\mu, \nu)$$

EB: prova a dimostrare questa disuguaglianza su $\mathbb{R}^n$

if and only if the gradient P_t is K-contractive, i.e.

$$|\nabla P_t f|^2 \leqslant e^{-2Kt} P_t |\nabla f|^2 \tag{13}$$

Notice that equation 13 states that for t big enough, the gradient of the heat semigroup goes to 0, i.e. heat converges to a constant!

Remark. Here we have seen the theory for the Dirichlet energy. In chapter 18, which I have yet to dive into, it is shown that the same theory holds on the *space of measures* $(\mathscr{P}_2 M, W_2)$ for the relative entropy. This is thanks to Otto, who shows that on the space of measures $\mathscr{P}_2(M)$, the heat semigroup can be defined as the gradient flow of the logarithmic entropy. This follows from mainly two passages: first, we consider that any gradient flow solves the continuity equation for some vector v,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mu_t + \mathrm{div}(v_t \mu_t) = 0$$

[1] where we set $v_t = -\nabla^W \mathcal{H}(\mu_t)$, so defining the gradient flow of the entropy in the space of measures. Then, we observe that

$$\nabla \mathcal{H}'(\varrho_t) = \frac{\nabla \varrho}{\varrho}$$

. Finally, we consider the heat semigroup ϱ_t and get:

$$\frac{\mathrm{d}}{\mathrm{d}t}\varrho_t = \Delta \varrho_t = \mathrm{div}(\nabla \varrho_t) = \mathrm{div}(\frac{\nabla \varrho_t}{\varrho_t}\varrho_t) = \mathrm{div}(\nabla^W \mathcal{H}(\varrho_t)\varrho_t)$$

i.e. the heat flow semigroup is the gradient flow of the entropy with respect to W_2. In fact, these are two (among more) interpretations of the heat flow: as the gradient flow of the Dirichlet energy in $\mathscr{L}^2$ and as the gradient flow of the entropy in $\mathscr{P}_2$. Also, on the space of measures or for metric spaces in general gradient flows are defined differently (EDE and EDI), but we do not dive into this specific definition and keep the functional one for intuition.

[1] we can ignore what the CE is for now

2. Relation to Ricci curvature

Now, we make some considerations on how Ricci relates to these and link the three concepts altogether.

Let us go back to the proof of geodesic convexity of the entropy in section. At some point, we looked at concavity of the map

$$t \mapsto (\det \nabla T_t(x))^{\frac{1}{n}}$$

which we denoted by a then.

Here, we try to extend this proof to Riemannian manifolds.

The expression for the optimal map T_t on Riemannian manifolds takes the form

$$T_t(x) = \exp(-t\nabla\phi(x)) =: \gamma_t$$

where we $\exp$ is the exponential map (definition 2.3).

Trying to reproduce the proof of convexity of the entropy, we study the quantity

$$\mathscr{J}(t) := \det(\nabla_x \exp(-t\nabla\phi))$$

and it turns out that

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathscr{J}_t^{\frac{1}{n}} + \frac{\mathrm{Ric}(\gamma', \gamma')}{n}\mathscr{J}_t^{\frac{1}{n}} \leqslant 0 \tag{14}$$

This equation related Ricci to the Jacobian[2], which governs how volumes change over the curve. Indeed, my intuition is that applying the measure valued function of the curve to a volume means moving the volume from μ_0 to μ_1 (gradually in t), and the determinant of our operator tells us exactly how volumes get distorted by the operator. In fact, it was clear from the initial proof that convexity of the entropy is related to the Jacobian of the gradient of the geodesic.

Alternatively, we can interpret Ricci curvature by looking at Bochner identity. In the Euclidean case $f : \mathbb{R}^n \to R$, for f sufficiently smooth, we have

$$\Delta\frac{|\nabla f|^2}{2} = ||\,\mathrm{Hess}\, f||^2 + (\nabla f, \nabla\Delta f)$$

[2] determinant of the Jacobian matrix aka gradient

Extending this to the Riemannian case, we get

$$\Delta_g \frac{|\nabla f|^2}{2} = ||\operatorname{Hess} f||^2 + (\nabla f, \nabla \Delta_g f) + \operatorname{Ric}(\nabla f, \nabla f)$$

Hence we get the Bochner inequality

$$\Delta_g \frac{|\nabla f|^2}{2} - (\nabla f, \nabla \Delta_g f) \geqslant \operatorname{Ric}(\nabla f, \nabla f) \tag{15}$$

To conclude, (14) and (15) give two alternative interpretations of the Ricci curvature. The first i related to how Ric controls volumes distortion along curves, referred to as Lagrangian interpretation. The second one shows, by choosing $f = \phi$, how Ric controls the behavior of the velocity vector field $\nabla\phi$, and it is referred to as the Euler interpretation.

Now we put it all together, stating the very last result relating Ricci to convexity of the entropy and the heat flow. I will state it in a high-level manner for now, mainly considering first some considerations we have gone over so far and then looking at how they relate to each other.

The following considerations on Ricci, the heat flow and convexity of the entropy are true

(i) the gradient flow of the logarithmic entropy solves the heat equation (eqv, is the heat semigroup)

(ii) we can represent the operator of the gradient flow with its dual P_t^*

(iii) the Kuwada equivalence, i.e. that P_t converges iff P_t^* is K-contractive

(iv) Ricci curvature is related to convexity of the entropy through inequality 14, and hence also to volume distortion

(v) Ricci curvature is related to the behavior of velocity fields through Bochner inequality 15

The main result that follows from all these considerations (and is partly already included), stated in theorem 19.4 of Ambrosio et al. (2021), is that **a lower bound on Ricci, K-convexity of entropy, and convergence of the heat flow are equivalent**.

The relations between the main statements passes through Bochner and through Pt* (being an EVI gradient flow of the entropy (which we did not define) and being K contractive).

While we do not go in depth in each claim for now, we look at how the statements above relate from an high-level perspective.

The explicit relation is from lower bound on Ricci to convexity of entropy and the other way around through eq (14), which also makes explicit that Ricci is related to volume distortion.

Convergence of the heat semigroup (gradient contractivity) is related to the above statements in a more indirect way. It is equivalent to P_t^* being K-contractive; EVI gradients of the entropy are K-contractive, and that the entropy has an EVI gradient is equivalent to being K-convex. This proves one side of the relation (from K-convexity to convergence of heat semigroup). On the other hand, gradient contractivity of P_t is equivalent to Bochner (I do not dive into this now). Then, it is quite easy to verify that $\mathrm{Ric} \geqslant K$ iff Bochner, one implication is immediate, the other comes from choosing a function f such that $\nabla f = v$ and $\mathrm{Hess}\, f = 0$.

CHAPTER 6

Paper: *Ricci curvature of finite Markov chains*

Reference: Erbar and Maas (2012), find it here

The main of the paper is to develop a variant notion of Ricci curvature lower bound for discrete spaces.

This notion builds upon the definition of a Ricci curvature lower bound for metric measure spaces in general, by Lott and Villani (and Sturm independently) (see the background section 1). Then, it is shown that a lower bound on the Ricci curvature is equivalent to geodesic convexity of the entropy, by Von Renesse and Sturm, as we have seen. This is useful because the notion of convexity does not appeal to the geometric structure of the space (differentiability) hence can be applied to more general spaces. Another core result known since the seminar work of Jordan, Kinderlehrer and Otto is that the heat flow is the gradient flow of the entropy in W_2 spaces, as we have seen. This has many implications (see background section below).

In fact, this paper wants to use the notion of convexity of the entropy to define a notion of Ric for discrete spaces. Moreover, we would like to interpret the heat flow as the gradient flow of some energy functional also in the discrete space. However, the W_2 distance is not appropriate for this purpose, because the discrete space does not have geodesics according to the W geometry (they are all constant), and we do not have the identification between heat flow and gradient flow of any entropy functional. This is because the metric derivative (the limit of the distance $d(\gamma(t+h), \gamma(t))$ as h goes to 0) of the Wasserstein distance of the heat flow can be infinite in discrete settings, hence we can not regard it as the gradient flow of some energy functional. See Maas (2011) for a more precise discussion of this incompatibility (remark 2.1).

Hence, we define a new metric $\mathcal{W}$ for which *time-continuous Markov chains are the gradient flow of the entropy*, analogously to the condition on W_2. In particular, the connection between Markov chains and the heat flow is that *the heat flow associated with a Markov kernel on a finite set is the gradient flow of the entropy with respect to* $\mathcal{W}$. Thus, we define convexity along $\mathcal{W}$ geodesics and we show that one such geodesic exists between each pair of probability measures on our discrete space. Then, we characterize the notion of a lower bound on the Ricci curvature with this notion of $\mathcal{W}$ convexity. Moreover, some examples are given on lower bounding the Ricci curvature of some discrete spaces, including the discrete circle and the discrete hypercube.

Lastly, the last sections of the paper include a tensorisation result (commonly given for this kind of objects) and show some well-known functional inequalities, which we do not dive into for now.

Hence, here we first give some more background, then we follow the same organization of the paper for defining $\mathcal{W}$ and giving the new definition of Ricci.

1. Background

Here we look at some background on extending the notion of Ricci curvature to other spaces, also looking at how it is related to other aspects. This part is based on Lott and Villani (2009).

Our aim is to define Ricci curvature for measure metric spaces in general (extending definitions from the Euclidean space to Riemannian manifolds).

The study of optimal transport has been proven (by McCann) to be closely related with convexity along geodesics of some functionals (on $P(\mathbb{R}^n)$), called *displacement convexity*. Recently, some regularity conditions have been extended from $\mathbb{R}^n$ to M. For instance, Hessian computations for some functions on $P_2(M)$ have been done. The result is a relationship between the Hessian of entropy function and Ricci (Otto Villani), as we have seen (without proving why, equation 14). Later, Cordero-Erausquin, McCann and Schmuckenschlager have found some rigorous results on the relationship between nonnegative Ricci and convexity of some functions on $P^{ac}(M)$. This work has been extended by Von Renesse and Sturm, who we cited above.

In this context, Lott and Villani *define* a new notion of Ricci curvature for metric measure spaces based on optimal transport and displacement convexity, which is what we referred to above.

2. Prerequisits

Here we define some prerequisits at the basis of the setting studied in the paper.

We have a finite space $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ and we can think of the elements of it as *states*, or nodes. We also have a Markov process of random variables that take values in $\mathcal{X}$.

We will deal with *continuous-time Markov chains* (CTMC), which are Markov chains where the time spent in each state is continuous. For comparison, in discrete-time Markov chains, a unit time is spent on each state. In our case instead, to respect the Markov property, the time spent in each state is modeled by an exponential random variable. This is because the exponential distribution only depends on some constant (rate) a_i for each state i, hence the time spent in each state will not depend on the past (states or the time spent on that state until now) (see these notes on CTMCs).[1] Also, we assume that the transition probabilities from one state to the other do not depend on the current time, but only on the current state (*homogeneous* Markov chain).

Given the homogeneity property, the transition probabilities can be described by a $|\mathcal{X}| \times |\mathcal{X}|$ matrix K,

$$K(x_i, x_j) = K_{ij} = \Pr(X_t = x_j \mid X_{t-1} = x_i)$$

$$\sum_{y \in \mathcal{X}} K(x, y) = 1$$

K is also called ***Markov kernel***. In this view, if we have an initial distribution over the states $p^0 \in \mathbb{R}^n$, the operation $K^t p^0$ gives the distribution over the states at time $t = 1$. Hence $((K^t)^T p^0)_i$ is the probability of living in state x_i after T iterations. Given $\mathcal{A}$ σ-algebra of $\mathcal{X}$, we can give another definition of K, as a function:

$$\begin{aligned} &K : \mathcal{X} \times \mathcal{A} \to [0, 1] \\ &(x_i, A) \mapsto \Pr(X_{t+1} \in A | X_t = x_i) \end{aligned}$$

also, $\Pr(X_{t+1} \in A | X_t = x_i) = \sum_{x_j \in A} K_{ij}$. A can be a singleton $\{y\}$, hence $K(x_i, x_j) = \Pr(x_j | x_i) = K_{ij}$, the probability of going to state x_j from state x_i. Moreover, in this view, $K(x, \cdot)$ becomes a probability measure on $\mathcal{A}$,

$$K(x, A) = \mu_x(A)$$

This is coherent with the general correspondence between probability measures and random variables, i.e. that $\Pr(X \in A) = \mu(A)$ when the probability measure μ is the *law* of the random variable X.

[1] Recall the density of $x \sim \exp(a)$ is $f(x; a) = ae^{-ax}$ for $x \geqslant 0$ and = 0 otherwise.

We assume that our Markov kernel is ***irreducible***, i.e. that for every $x, y \in \mathcal{X}$ there exists a sequence $\{x_i\}_{i=0}^{n}$ such that $x_0 = x$, $x_n = y$ and $K(x_{i-1}, x_i) > 0$ for all $i \in [1, n]$. Irreducibility means that we will explore all our space $\mathcal{X}$, and it implies the existence of a unique *steady state*.

A steady state is a stationary or invariant probability measure π such that

$$\sum_{x \in \mathcal{X}} \pi(x) = 1 \qquad \text{and} \qquad \pi(y) = \sum_{x \in \mathcal{X}} \pi(x) K(x, y)$$

It is a left eigenvector of the operator K. Its existence is guaranteed by Perron-Frobenius, as K is a stochastic matrix. The invariant measure is a probability distribution over the states (just like the previous example of p^0) such that once we arrive at this distribution, we always remain here. Indeed, we would like some property of convergence to this distribution.

We also assume our kernel is ***reversible***, meaning that the detailed balance condition holds,

$$K(x, y)\pi(x) = K(y, x)\pi(y)$$

This means that the reversed kernel that goes over the chain backwards is equal to the original one.

Look at these notes for more on the definition and properties of a Markov kernel.

Given a steady state π, we can look at its associated ***Markov semigroup***.

Doing a little step back, a *Markov operator* is an operator acting on the space of densities $\mathscr{P}(\mathcal{X})$ (see below) which describes the evolution of my chain. Indeed, we can look at our Markov chain of random variables as $X_1 \to X_2 \to \ldots \to X_n$, or equivalently as $\rho_1 \to \rho_2 \to \ldots \to \rho_n$ (when the laws of the random variables are absolutely continuous probability measures), and hence we can look at a transformation P that brings $\rho_t into \rho_{t+1}$. When this transformation is determined by a transition matrix, as in our case (K), P is linear, it is a map between Dirac measures δ_x for $x \in \mathcal{X}$ [source notes]. So we have

$$P_t\rho = \mathbb{E}[\rho_t | \rho_0 = \rho]$$

and call P_t a Markov operator. Then, we define a Markov semigroup as a family of Markov operators $\{P_t\}_{t \geqslant 0}$ such that

- $P_0 = \mathrm{id}$
- $P_t\mathbf{1} = \mathbf{1}$
- If $f \geqslant 0$, then $P_t f \geqslant 0$

- $P_{t+s} = P_t \circ P_s$
- the function $t \mapsto P_t\rho$ is continuous for each ρ

Consider our steady state π for our Markov chain. It is natural (says Chat) to look at this π as the measure of our space $\mathcal{X}$, and hence at the semigroup definition as

$$\int P_t f \,\mathrm{d}\pi = \int f \,\mathrm{d}\pi$$

This means that *the semigroup is self-adjoint with respect to the invariant measure*. Moreover, if we consider the dual of P_t as in the theory above, we have that

$$\int P_t f \,\mathrm{d}\mu = \int f \,\mathrm{d}P_t^* \mu$$

for every μ by definition, hence

$$\int P_t f \,\mathrm{d}\pi = \int f \,\mathrm{d}\pi = \int f \,\mathrm{d}P_t^* \pi$$

hence

$$P_t^* \pi = \pi$$

where P_t^* is an operator acting on measures, $P_t^* : \mathscr{P}(\mathcal{X}) \to \mathscr{P}(\mathcal{X})$ (instead $P_t : \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^{\mathcal{X}}$). Since we have that $\pi K = \pi$, I think that P_t^* and K might be equivalent, but K acts on $\mathcal{X}$ and P_t^* on $\mathscr{P}(\mathcal{X})$.

Also, we generally look at semigroups as exponentials, and in this case as a matrix exponential where the matrix is our *generator*, in our case:

$$P_t := e^{t(K-I)}$$

as the definition of generator is the matrix Q such that

$$Qf = \lim_{t \to 0} \frac{P_t f - f}{t}$$

Indeed

$$\pi P_t = \pi e^{t(K-I)} = \pi \sum_{n=0}^{\infty} \frac{1}{n!} t^n (K-I)^n = \sum_{n=0}^{\infty} \frac{1}{n!} t^n \pi (K-I)^n = \pi$$

as $\pi(K-I)(K-I)^{n-1} = 0$ as $\pi(K-I) = 0$, so the only remaining term of the sum is for $n = 0$ which gives π. Also we have

$$\rho_t = P_t \rho_0$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho_t = \frac{\mathrm{d}}{\mathrm{d}t} e^{t(K-I)} \rho_0 = (K-I) e^{t(K-I)} \rho_0 = (K-I)\rho_t \tag{16}$$

Remark. The Markov semigroup P_t is directly associated with the Markov kernel K: indeed, they are respectively the continuous and discrete faces of the same moon, our Markov chain. In fact, K tells me the evolution for a discrete time of my chain, P_t the evolution for a continuous time, and by simple statistical computations we can get the exponential form of P_t applying the formulas for a Poisson with rate = 1.

Let $\mathscr{P}(\mathcal{X})$ be the set of probability densities over $\mathcal{X}$, namely

$$\mathscr{P}(\mathcal{X}) := \Big\{\rho : \mathcal{X} \to \mathbb{R}^+ : \sum_x \rho(x)\pi(x) = 1\Big\}$$

and let $\mathscr{P}_*(\mathcal{X})$ be the set of strictly positive probability densities.

3. The metric $\mathcal{W}$

First, there are some objects to be defined.

Given our set $\mathcal{X}$, we define two kinds of operators:

(i) $\varphi \in \mathbb{R}^{\mathcal{X}}, \quad \varphi : \mathcal{X} \to \mathbb{R}$

(ii) $\Psi \in \mathbb{R}^{\mathcal{X}\times\mathcal{X}}, \quad \Psi : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$

namely functions on $\mathcal{X}$ taking either one or two variables as inputs. Then, we define two functionals acting on each of the two functions. The first is the *discrete gradient* ∇,

$$\nabla : \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^{\mathcal{X}\times\mathcal{X}},$$

$$\varphi(x) \mapsto \varphi(y) - \varphi(x) \ =: \nabla\varphi(x,y)$$

the second one is the *discrete divergence* $\nabla\cdot$,

$$\nabla\cdot : \mathbb{R}^{\mathcal{X}\times\mathcal{X}} \to \mathbb{R}^{\mathcal{X}},$$

$$\Psi(x,y) \mapsto \frac{1}{2}\sum_{y\in\mathcal{X}}(\Psi(x,y) - \Psi(y,x))K(x,y) \ =: \nabla\cdot\Psi(x)$$

These allow us to define a third operator, the Laplacian $\Delta : \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^{\mathcal{X}}$

$$\Delta := \nabla\cdot\nabla = K - I$$

where the composition of $\nabla\cdot$ and ∇ makes sense (input and target sets are respected) and equality to $K - I$ can be easily verified (quaderno). Crucially, applying this result to (16) we see that ***the Markov semigroup solves the heat equation***.[2]

[2] We say, equivalently, that the Markov chain can be interpreted as the heat flow, that there is an analogy between Markov semigroup and heat flow and that the Markov semigroup solves the heat equation

Finally, we have the following *integration by parts formula*

$$(\nabla\psi, \Psi)_\pi = -\left(\psi, \nabla \cdot \Psi\right)_\pi$$

where for $\varphi, \psi \in \mathbb{R}^{\mathcal{X}}$ and $\Psi, \Phi \in \mathbb{R}^{\mathcal{X}\times\mathcal{X}}$ we have

$$\begin{aligned}(\varphi, \psi)_\pi &= \sum_{x \in X} \varphi(x)\, \psi(x)\, \pi(x), \\ (\Phi, \Psi)_\pi &= \frac{1}{2} \sum_{x,y \in X} \Phi(x,y)\, \Psi(x,y)\, K(x,y)\, \pi(x).\end{aligned}$$

where π is our steady state. It follows that $\nabla\cdot$ *is the negative adjoint of* ∇.

Now, we want to re-define these functionals on a more specific setting, i.e. the Hilbert space $\mathcal{G}_\rho$ for a fixed density $\rho \in \mathscr{P}(\mathcal{X})$.

To do so, we consider the *logarithmic mean* function $\theta : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}^+$, defined by

$$\theta(s,t) := \int_0^1 s^{1-p} t^p \ \mathrm{d}p = \frac{s-t}{\log s - \log t}$$

for $s, t > 0$.

This function θ is one instance of a more general class of functions that satisfy the properties of regularity (continuity and smoothness), symmetry, positivity and normalization, monotonicity, positive homogeneity, zero boundary ($\theta(0,t) = 0$) and concavity. Moreover, they are bounded by above by the arithmetic mean,

$$\theta(s,t) \leqslant \frac{s+t}{2} \tag{17}$$

and satisfy the following conditions (lemma 2.2 in the paper).

$$s\,\partial_1(s,t) + t\,\partial_2\theta(s,t) = \theta(s,t) \tag{18}$$

$$s\,\partial_1\theta(u,v) + t\,\partial_2\theta(u,v) - \theta(s,t) \geqslant 0 \tag{19}$$

where ∂_i is the derivative with respect to the i-th entry. These will be useful later on.

Now, considering θ as the log mean, and fixing a density $\rho \in \mathscr{P}(\mathcal{X})$, we define

$$\hat{\rho}(x,y) := \theta(\rho(x), \rho(y))$$

which is a logarithmic mean between densities (I don't know what it means precisely for now).

Then, $\mathcal{G}_\rho$ is the space of all equivalence classes of functions $\Psi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, where the equivalence relation is identity on the set $\{(x,y) : \hat{\rho}(x,y)K(x,y) > 0\}$. We endow our space with the inner product:

$$(\Psi,\Phi)_\rho := \frac{1}{2}\sum_{x,y\in\mathcal{X}} \Psi(x,y)\Phi(x,y)\hat{\rho}(x,y)K(x,y)\pi(x)$$

The discrete gradient operator ∇ can be seen as $\nabla : L^2(\mathcal{X}) \to \mathcal{G}_\rho$, and its negative adjoint $\nabla\cdot_\rho$, the *ρ-divergence operator*, is defined as

$$\nabla\cdot_\rho \Psi(x) := \frac{1}{2}\sum_{y\in\mathcal{X}}(\Psi(x,y)-\Psi(y,x))\hat{\rho}(x,y)K(x,y)$$

3.1. Defining $\mathcal{W}$. Let us use the following short hand notation for the squared norm of the gradient of a function ψ, i.e. its *action* under ρ,

$$\mathcal{A}(\rho,\psi) := ||\nabla\psi||_\rho^2 = (\nabla\psi,\nabla\psi)_\rho = \frac{1}{2}\sum_{x,y\in\mathcal{X}}((\psi(y)-\psi(x))^2\hat{\rho}(x,y)K(x,y)\pi(x)$$

for $\psi\in\mathbb{R}^{\mathcal{X}}$ and $\rho\in\mathscr{P}(\mathcal{X})$.

Definition 3.1 (Distance $\mathcal{W}$). *For $\rho_0,\rho_1\in\mathscr{P}(\mathcal{X})$ we define*[3]

$$\mathcal{W}(\rho_0,\rho_1) := \inf\left\{\int_0^1 \mathcal{A}(\rho_t,\psi_t)\ \mathrm{dt} : (\rho,\psi)\in\mathcal{CE}_1(\rho_0,\rho_1)\right)$$

where the infimum runs over all pairs of curves. For $T>0$, $\mathcal{CE}_T$ is the set of pairs (ρ,ψ) being sufficiently regular and satisfying a continuity equation. In particular we ask:

(i) ρ *is* C^∞, $\rho:[0,T]\to\mathbb{R}^{\mathcal{X}}$

(ii) $\rho_t|_{t=0}=\rho_0$, $\rho_t|_{t=1}=\rho_1$

(iii) $\rho_t\in\mathscr{P}(\mathcal{X})$ *for all* $t\in[0,T]$

(iv) ψ *is measurable,* $\psi:[0,T]\to\mathbb{R}^{\mathcal{X}}$

(v) For all $x\in\mathcal{X}$ *and* $t\in(0,T)$ *we have*

$$\rho_t'(x)+\sum_{y\in\mathcal{X}}(\psi_t(y)-\psi_t(x))\hat{\rho}(x,y)K(x,y)=0 \qquad \textit{(continuity equation)}$$

Where the continuity equation in (v) can be re-written as

$$\rho_t'+\nabla\cdot_\rho(\nabla\psi)=0 \tag{20}$$

[3] I use ρ_0 although it can be confused with $\rho_t|_{t=0}$ because they must be equal in the end and I hate the notation with the bar

Hence in $\mathcal{CE}$ we look at ψ and ρ as curves, as we did in section 4.

Notice that we define $\mathcal{W}$ along the lines of the dynamic formulation of ot/ Benamou-Brenier formula (7). This implies looking for a geodesic that joins two measures. In fact, such a definition of the W_2 (here $\mathcal{W}$) metric has been referred to by Otto as giving a Riemannian structure to the space of measures. This allows to study geodesics, convexity of entropy and the heat flow as gradient flow.

3.2. Heat flow on $\mathcal{X}$.

The following theorem provides a crucial result on $\mathcal{W}$. Let us use the notation $P(\mathcal{X})$ for the space $\mathscr{P}_*(\mathcal{X})$ henceforth.

THEOREM 3.2. *(1) The space $(P(\mathcal{X}), \mathcal{W})$ is a complete metric space, compatible with the Euclidean topology*

(2) $\mathcal{W}$ ***restricted to*** *$P(\mathcal{X})$* ***is the Riemannian distance induced by the following Riemannian structure****:*

- *the tangent space at $\rho \in P(\mathcal{X})$ can be identified as*

$$T_\rho P(\mathcal{X}) := \{\nabla\psi : \psi \in \mathbb{R}^{\mathcal{X}}\}$$

with the identification that there exists a unique element $\nabla\psi_0$ such that the continuity equation (v) hols at $t = 0$, for a curve $(-\varepsilon, \varepsilon) \ni t \mapsto \rho_t \in P(\mathcal{X})$ such that $\rho_0 = \rho$. Cioè per ogni densità ρ, cioè un elemento che vive nel mio spazio $P(\mathcal{X})$, ho un insieme intero di elementi $\nabla\psi$ che gli sono tangenti, ma solo uno, che indichiamo come $\nabla\psi_0$, è tale che l'equazione di continuità è soddisfatta.

- *The Riemannian metric on T_ρ is given by the inner product (already defined by the definitions we gave until now)*

$$(\nabla\varphi, \nabla\psi) = \frac{1}{2}\sum_{x,y\in\mathcal{X}} (\varphi(y) - \varphi(x))(\psi(y) - \psi(x))\hat{\rho}(x,y)K(x,y)\pi(x)$$

(3) The heat flow is the gradient flow of the entropy, in the sense that for $\rho \in P(\mathcal{X})$ and $t \geqslant 0$ we have $\rho_t := P_t\rho \in P(\mathcal{X})$ and

$$D_t\rho_t = -grad\,\mathcal{H}(\rho_t)$$

where with D_t we mean in general the derivation of ρ_t.

PROOF. For the proof we are referred to the paper by Maas and I think I should look at it. □

The conceptual point here is the following: among the elements in the tangent space at ρ, we choose the one that solves the continuity equation. The continuity equation has $\hat{\rho}$ in the formula, hence the choice of the logarithmic mean is relevant, and we see why now.

Consider the heat equation $\rho_t' = \Delta\rho_t = \nabla\cdot\nabla\rho_t$. Consider also the continuity equation (20). The heat equation can be re-written as a continuity equation if

$$\hat{\rho}_t\,\nabla\psi_t = -\nabla\rho_t; \quad \nabla\psi_t = -\frac{\nabla\rho_t}{\hat{\rho}_t}$$

At the same time, we have that the gradient of the entropy under the identification above is

$$\text{grad}_{\mathcal{W}}\mathcal{H}(\rho_t) = \nabla\log\rho_t$$

$$\mathcal{H}(\rho_t) = \sum_{x\in\mathcal{X}}\rho_t(x)\log\rho_t(x)\pi(x)$$

We would like that the heat flow was the gradient flow of the entropy, i.e. that the derivative of ρ_t, which is $\nabla\psi_t$, was equal to minus the gradient of the entropy, hence

$$\nabla\psi_t = -\frac{\nabla\rho_t}{\hat{\rho}_t} = -\nabla\log\rho_t;$$

$$\frac{\nabla\rho_t}{\hat{\rho}_t} = \nabla\log\rho_t \tag{21}$$

but condition 21 is asking precisely that $\hat{\rho}_t$ is the logarithmic mean. Hence for θ being the logarithmic mean it is satisfied. It follows that the continuity equation can be seen as the heat equation, and the heat flow is the gradient flow of the entropy, because the derivation of ρ_t satisfies the continuity equation, hence it is the heat flow, and with the logarithmic mean ut follows that it is equal to the gradient flow of the entropy.

What we have found is that **the heat equation with a Markov kernel**, which is equivalent to considering the *flow* of a Markov chain, like instead of heat propagating we look at a Markov chain evolving, **is the gradient flow of the entropy**. Hence all the results we get in the final theorem of Ambrosio et al. (2021) hold :) !!!

Another way to look at this (which is what I wanted to say above) is by substituting in the heat flow our continuous time Markov semigroup $P_t = e^{t(K-I)}$, i.e. using this as the heat semigroup. We get that under $\mathcal{W}$, it is the gradient flow of the entropy (in fact we defined $\rho_t = P_t\rho$).

3.3. Alternative definition for $\mathcal{W}$. In the paper, an alternative definition for $\mathcal{W}$ is given. This is because it allows to show the equivalence with Maas previous definition, it is more practical to use and it is in fact used in many subsequent proofs (which I actually don't look at).

Namely, we define the following function α

$$\alpha(x,s,t)=\begin{cases}0, & \theta(s,t)=x=0\\ \frac{x^2}{\theta(s,t)}, & \theta(s,t)\neq 0\\ +\infty, & \theta(s,t)=0, x\neq 0\end{cases}$$

This is lower semicontinuous and convex, by concavity of θ and convexity of $x\mapsto\frac{x^2}{y}$.

Then, for $\rho\in\mathscr{P}(\mathcal{X})$ and a new function $V\in\mathbb{R}^{\mathcal{X}\times\mathcal{X}}$ we define

$$\mathcal{A}'(\rho,V):=\frac{1}{2}\sum_{x,y\in\mathcal{X}}\alpha(V(x,y),\rho(x),\rho(y))K(x,y)\pi(x)$$

and we set

$$\mathcal{CE}'_T(\rho_0,\rho_1):=\{(\rho,V):(i'),(ii),(iii),(iv'),(v')\text{ hold }\}$$

with

(i') ρ is continuous (before C^∞)

(iv') V is locally integrable (before ψ measurable)

(v') For all $x\in\mathcal{X}$ we have

$$\rho'_t(x)+\frac{1}{2}\sum_{y\in\mathcal{X}}(V_t(x,y)-V_t(y,x))K(x,y)=0;$$

$$\rho'_t(x)+\nabla\cdot V=0$$

continuity equation.

$\mathcal{A}'$ is convex by convexity of α. Now we can state the following equivalent version of $\mathcal{W}$.

Lemma 3.3. *For $\rho_0,\rho_1\in\mathscr{P}(\mathcal{X})$ we have*

$$\mathcal{W}^2(\rho_0,\rho_1)=\inf\left\{\int_0^1\mathcal{A}'(\rho_t,V_t):(\rho,V)\in\mathcal{CE}'(\rho_0,\rho_1\right\}$$

Moreover, for $\rho_0,\rho_1\in P(\mathcal{X})$, (iv) becomes that ψ is C^∞. (?)

We just notice that the infimum here is taken over a smaller set than before. Indeed, given a pair (ρ, ψ), we can set $V(x,y) = (\psi(y) - \psi(x))\hat{\rho}(x,y)$ to get the two definitions coincide. Hence the "$\geqslant$" proof is trivial. As for the other direction, the proof is slightly more complicated.

We disregard it as we will not be so interested in the quantity $\mathcal{A}'$ and in the alternative definition of $\mathcal{W}$.

3.4. Properties of $\mathcal{W}$. The metric $\mathcal{W}$ has the following properties:

- *Convexity*: consider two curves $\rho_t, \rho'_t \in \mathscr{P}(\mathcal{X})$. Then

$$\mathcal{W}^2\big((1-\tau)\rho_0 + \tau\rho'_0, (1-\tau)\rho_1 + \tau\rho'_1\big) \leqslant (1-\tau)\mathcal{W}^2(\rho_0, \rho_1) + \tau\mathcal{W}^2(\rho'_0, \rho'_1)$$

- *Lower bounds*:

$$\frac{1}{\sqrt{2}} d_{TV}(\rho_0, \rho_1) \leqslant \sqrt{2} W_{1,g}(\rho_0, \rho_1) \leqslant \mathcal{W}(\rho_0, \rho_1)$$

  where $W_{1,g}$ is the 1-Wasserstein distance induced by the graph distance and d_{TV} is the total variation metric;

- *Upper bounds*:

$$\mathcal{W}(\rho_0, \rho_1) \leqslant W_{2,\mathcal{W}}(\rho_0, \rho_1) \leqslant c \cdot W_{2,g}(\rho_0, \rho_1)$$

  where $W_{2,\mathcal{W}}$ is the Wasserstein distance induced by the metric obtained restricting $\mathcal{W}$ to $\mathcal{X}$, c is a constant depending on the Markov kernel K.

I disregard also these proofs.

4. Geodesics in $\mathcal{W}$ space

In this section of the paper they deal with geodesics in $(\mathscr{P}(\mathcal{X}), \mathcal{W})$.

First, it is shown that there always exists a geodesic connecting two densities $\rho_0, \rho_1 \in \mathscr{P}(\mathcal{X})$ that attains $\mathcal{W}$. In particular, there is a pair (ρ, V) that attains $\mathcal{W}$ as defined in lemma 3.3, hence $\mathcal{A}'(\rho_t, V_t) = \mathcal{W}^2(\rho_0, \rho_1)$ and ρ_t is a geodesic connecting ρ_0 and ρ_1.

Remember that such a geodesic ρ_t is such that

$$\mathcal{W}(\rho_s, \rho_t) = |s-t|\mathcal{W}(\rho_0, \rho_1)$$

I disregard also these proofs for now.

Then, we define absolutely continuous curves on $(\mathscr{P}(\mathcal{X}), \mathcal{W})$. A curve $(\rho_t)_{t\in[0,T]}$ is *absolutely continuous w.r.t* $\mathcal{W}$ if there is a function $m \in L^1(0,T)$ such that

$$W(\rho_s, \rho_t) \leqslant \int_s^t m(r)\,\mathrm{d}r \quad \forall\ 0 \leqslant s \leqslant t \leqslant T$$

and if $(\rho_t)_{t\in[0,T]}$ is absolutely continuous, then its metric derivative exists:

$$|\rho_t'| := \lim_{h\to 0} \frac{\mathcal{W}(\rho_{t+h}, \rho_t)}{|h|}$$

and $|\rho_t'| \leqslant m(t)$.

Now, we relate the length of a curve $\int_0^T |\rho_t'|\,\mathrm{d}t$ to its minimal action, and also give a characterization of absolute continuity.

Proposition 4.1 (Absolute continuity in $\mathcal{W}$). *(i) A curve $(\rho_t)_{t\in[0,T]}$ is absolutely continuous w.r.t. $\mathcal{W}$ if there exists a $V : [0,T] \to \mathbb{R}^{\mathcal{X}\times\mathcal{X}}$ such that $(\rho, V) \in \mathcal{CE}_T'(\rho_0, \rho_T)$ and*[4]

$$\int_0^T \sqrt{\mathcal{A}'(\rho_t, V_t)}\,\mathrm{d}t < \infty$$

(ii) In this case, we have $|\rho_t'|^2 \leqslant \mathcal{A}'(\rho_t, V_t)$ for a.e. $t \in [0,T]$

(iii) Also, there exists an a.e. uniquely defined function $\tilde{V} : [0,1] \to \mathbb{R}^{\mathcal{X}\times\mathcal{X}}$ such that $|\rho_t'|^2 = \mathcal{A}'(\rho_t, \tilde{V}_t)$ for a.e. t.

Non ho ben capito cosa ci stia dicendo questa proposition!

Finally, we state the following result on geodesics, which will be useful later on.

Recall that $(P(\mathcal{X}), \mathcal{W})$ is a Riemannian space, hence local existence and uniqueness of geodesics is guaranteed.

Proposition 4.2 (Formulas for geodesics on $P(\mathcal{X})$). *Let $\rho \in P(\mathcal{X})$ and $\psi \in \mathbb{R}^{\mathcal{X}}$. On a sufficiently small time interval around 0, the unique geodesic (ρ_t) with $\rho_0 = \rho$ and initial speed $\nabla\psi_0 = \nabla\psi$ satisfies the following equations:*

$$\partial_t \rho_t(x) + \sum_{y\in\mathcal{X}} (\psi_t(y) - \psi_t(x)) \hat{\rho}_t(x,y) K(x,y) = 0 \tag{22}$$

$$\partial_t \psi_t(x) + \frac{1}{2} \sum_{y\in\mathcal{X}} (\psi_t(x) - \psi_t(y))^2 \partial_1 \theta(\rho_t(x), \rho_t(y)) K(x,y) = 0 \tag{23}$$

[4] The integral with the sqrt of the action is equivalent to the integral in $\mathcal{W}$, up to reparametrization

where the first equation is a the continuity equation

$$\Delta_\rho \psi_t = \nabla \cdot_\rho \nabla \psi_t = \nabla \cdot (\nabla \psi_t \hat{\rho}_t)$$

$$(23) \text{ becomes } \frac{\mathrm{d}}{\mathrm{d}t}\rho_t + \Delta_\rho \psi_t = 0$$

For the second one I still don't have a nice intuition.

5. Defining Ricci for discrete spaces

After having seen in that the space $(\mathscr{P}(\mathcal{X}), \mathcal{W})$ geodesics always exists, it is natural to extend the celebrated definition of Ricci lower bound in terms of convexity of the entropy to discrete spaces.

Definition 5.1. *We say that K has non-local Ricci curvature bounded by below by $\kappa \in \mathbb{R}$ and write $\mathrm{Ric}(K) \geqslant \kappa$ if for every geodesic $(\rho_t)_{t\in[0,1]} \in (\mathscr{P}(\mathcal{X}), \mathcal{W})$ we have*

$$\mathcal{H}(\rho_t) \leqslant (1-t)\mathcal{H}(\rho_0) + t\mathcal{H}(\rho_1) - \frac{\kappa}{2}t(1-t)\mathcal{W}^2(\rho_0, \rho_1)$$

i.e. entropy is κ-convex along geodesics

Remark. Notice that equation (16) gives a precious insight now: we study a lower bound on Ricci *of* K, because K is the law that moves probability mass around our manifold of measures. In fact, curvature in the continuous case is related to the notion of connections, which give rules to move around our manifold, here it is related to K. That equation shows that the derivative of a curve is indeed governed by K, maybe analogously to connections. The discrete manifold of measures would be too flat, sparse and disconnected to talk about its curvature, there is no such concept (which is why we made all this mess).

Another result is a characterization of a lower bound on Ricci, by means of a lower bound on the Hessian of the entropy. Indeed, we show that κ-convexity of $\mathcal{H}$ along geodesics is equivalent to a lower bound on the Hessian of $\mathcal{H}$ on $P(\mathcal{X})$. Actually, the following theorem (the last of this section) shows many equivalent notions to $\mathrm{Ric}(K) \geqslant \kappa$.

Let us first define the following quantity, useful for the theorem.

For $\rho \in P(\mathcal{X})$ and $\psi \in \mathbb{R}^{\mathcal{X}}$ we define

$$\mathcal{B}(\rho, \psi) := \frac{1}{2}(\hat{\Delta}\rho \cdot \nabla\psi, \nabla\psi)_\pi - (\hat{\rho} \cdot \nabla\psi, \nabla\Delta\psi)_\pi \tag{24}$$

with

$$\hat{\Delta}\rho(x,y) := \partial_1\, \theta(\rho(x), \rho(y))\Delta\rho(x) + \partial_2\theta(\rho(x), \rho(y))\Delta\rho(y)$$

The main result on $\mathcal{B}$ is that $(\mathrm{Hess}\,\mathcal{H}(\rho_t), \nabla\psi_t, \nabla\psi_t)_\rho = \mathcal{B}(\rho, \psi)$. The proof of this is based on the geodesic equation we provided above (Proposition 4.2), I do not look at it.

The whole point of defining this quantity is that it is a discrete analogue of the Bochner quantity

and therefore $\mathcal{B}(\rho, \psi) \geqslant \kappa\mathcal{A}(\rho, \psi)$ can be seen as the Bochner inequality for a discrete setting.

Now we move to the main theorem, characterizing a lower bound on Ricci with many equivalent notions. Let us use the following notation

$$\frac{\mathrm{d}^+}{\mathrm{d}t} f(t) = \limsup_{h\downarrow 0} \frac{f(t+h) - f(t)}{h}$$

Theorem 5.2 (Ricci lower bound equivalent statements). *Let $\kappa \in \mathbb{R}$. For an irreducible and reversible Markov kernel $(\mathcal{X}, K)$ the following assertions are equivalent:*

(1) $\mathrm{Ric}(K) \geqslant \kappa$*;*

(2) For all $\rho, \nu \in \mathscr{P}(\mathcal{X})$ the following 'evolution variational inequality' holds for all $t \geqslant 0$:

$$\frac{1}{2}\frac{\mathrm{d}^+}{\mathrm{d}t}\mathcal{W}^2(P_t\rho, \nu) \;+\; \frac{\kappa}{2}\mathcal{W}^2(P_t\rho, \nu) \;\leqslant\; \mathcal{H}(\nu) \;-\; \mathcal{H}(P_t\rho);$$

(3) For all $\rho, \nu \in P(\mathcal{X})$, the above inequality holds for all $t \geqslant 0$;

(4) For all $\rho \in P(\mathcal{X})$ and $\psi \in \mathbb{R}^{\mathcal{X}}$

$$\mathcal{B}(\rho, \psi) \geqslant \kappa\mathcal{A}(\rho, \psi);$$

(5) For all $\rho \in P(\mathcal{X})$

$$\mathrm{Hess}\,\mathcal{H}(\rho) \geqslant \kappa;$$

(6) For all $\rho_0, \rho_1 \in P(\mathcal{X})$, there exists a constant speed geodesic $(\rho_t)_{t\in[0,T]}$ that connects them along which entropy is κ-convex.

Proof. I leave also this here for now. □

An interesting property that is implied by the evolution variational inequality in point (2) is κ-contractivity of the gradient flow.

Proposition 5.3. *Let $(\mathcal{X}, K)$ be as before, with $\operatorname{Ric}(K) \geqslant \kappa$. Then the associated Markov semigroup $(P_t)_{t\geqslant 0}$ satisfies*

$$\mathcal{W}(P_t\rho, P_t\sigma) \leqslant e^{-\kappa t}\mathcal{W}(\rho, \sigma)$$

for all $\rho, \sigma \in \mathscr{P}(\mathcal{X})$ and $t \geqslant 0$.

Proof. We refer to the application of proposition 3.1 of Daneri and Savaré (2008) to $\mathcal{H}$. As usual. □

This is convergence of the Markov chain !!

6. Applications

Some examples are given. One is very straightforward, on the complete graph with n nodes $\mathcal{K}^n$, and it is just a matter of calculations.

For the other examples, we do not employ the standard point of view of a Markov chain as jumps between states. Instead, we consider *moves and move rates*: let G be a set of functions from $\mathcal{X}$ to itself, the allowed moves between states, and let $c : \mathcal{X} \times G \to \mathbb{R}^+$ be a function representing how often a certain move $g \in G$ is made on a certain state $x \in \mathcal{X}$.

Definition 6.1 (Mapping representation). *A mapping representation of K is a pair (G, c) such that:*

1. *The generator che è $\Delta = K - I$ si può scrivere come*
$$\Delta\psi(x) = \sum_{g\in G} \nabla_g\psi(x)c(x, g)$$
where
$$\nabla_g\psi(x) = \psi(gx) - \psi(x)$$
2. *For every $g \in G$ there is a unique $g^{-1} \in G$ satisfying $g^{-1}(g(x)) = x$ for all $x : c(x, g) > 0$*
3. *For every $F : \mathcal{X} \times G \to \mathbb{R}$ we have*
$$\sum_{x\in\mathcal{X},\, g\in G} F(x, g)c(x, g)\pi(x) = \sum_{x\in\mathcal{X},\, g\in G} F(gx, g^{-1})c(x, g)\pi(x)$$

Pensavo di avere un insight ma non ci capisco un cazzzzz

Remark 5.3 → non ha senso

Every irreducible and reversible Markov chain has a mapping representation. This can be explicitly built as follows: consider the bijection g_{xy} that brings x in y and set G to be the set of all this maps; then consider $c(x, g_{xy}) = K(x, y)$ and $c(x, g_{zy}) = 0$ if $x \notin \{y, z\}$. Then G, c is a mapping representation, because (consider $g = g_{xy}$:

(1) $\Delta\psi(x) = \sum_{g\in G} \nabla_g \psi(x) c(x, g) = \sum_{g\in G}(\psi(gx) - \psi(x))K(x, y) = \sum_{y\in\mathcal{X}}(\psi(y) - \psi(x))K(x, y)$ which is the Laplacian operator applied to ψ by definition.
(2) trivial
(3) I don't know how to verify it but seems ok

Then the paper gives explicit expressions for $\mathcal{A}$ and $\mathcal{B}$, but I skip them for now.

What follows is a criterion to derive a bound on Ricci based on this mapping representation.

Proposition 6.2 (Criterion for Ricci via mapping reps). *Let K be an irreducible and reversible kernel on a finite set $\mathcal{X}$ and let (G, c) be a mapping representation. Consider the following conditions:*

(i) $g \circ l = l \circ g$, *for all* $g, l \in G$,
(ii) $c(gx, l) = c(x, l)$, *for all* $x \in \mathcal{X}$, $g, l \in G$,
(iii) $g \circ g = \mathrm{id}$, *for all* $g \in G$.

We have that if (i) and (ii) are satisfied, then $\mathrm{Ric}(K) \geqslant 0$. *If also (iii) is satisfied, then* $\mathrm{Ric}(K) \geqslant 2C$, *with* $C := \min\{c(x, g) : x \in \mathcal{X}, g \in G$ *such that* $c(x, g) > 0\}$

Then, this proposition is proven, which seems doable (as usual we skip it for now), and finally applied to two examples. The second one is more interesting, as the bound is optimal!

Example (Discrete circle). Consider the discrete circle $C_n = \mathbb{Z}/n\mathbb{Z}$ of n sites, and a random walk on it defined by the kernel $K(i, i-1) = K(i, i+1) = 1/2$ for $i \in C_n$. Think of this as n nodes put in a circle, and each time we have equal probability (1/2) of moving to the right or to the left. Then the possible moves we can make are $G = \{l, r\}$ and $l(i) = i+1, r(i) = i-1$ (by convention we number nodes in senso orario). Intuitively $c(i, l) = c(i, r) = 1/2$. By proposition 6.2, $\mathrm{Ric}(K) \geqslant 0$.

Example (Discrete hypercube). Let $\mathcal{Q}^n = \{0, 1\}^n$ be the n-dimensional hypercube and let K_n be the kernel defining a simple random walk on it. The natural mapping representation is given by $G = \{g_1, \ldots, g_n\}$ with g_i is the map that flips the i-th coordinate. For instance, for $n = 3$ the hypercube is $\mathcal{Q}^3 = \{(0, 0, 0), (0, 0, 1), (0, 1, 0), \ldots, (1, 1, 1)\}$ and $g_1 : (0, 0, 0) \mapsto (1, 0, 0)$, $g_2 : (1, 1, 1) \mapsto (1, 0, 1)$. Also, $c(x, g_i) = 1/n$ for each i and $x \in \mathcal{Q}^n$. All three conditions in proposition 6.2 are satisfied, hence $\mathrm{Ric}(K_n) \geqslant 2/n$.

REMARK. These examples shed light on some insights on the Ricci curvature of discrete spaces. For instance, if you think about them a priori, it is puzzling to define the curvature of a discrete circle. Or of a graph. In fact, we talk about curvature of such spaces once we are given a *rule* to move from state to state, namely the kernel. Indeed, as we already mentioned, in a continuous manifold its curvature, or shape in general, determines how I can move from point to point, how much time it takes me to go from point to point (given an initial velocity). Here, what determines this is the kernel or the mapping representation of it.

7. Ricci of transformations of Markov chains

In this section (section 7) they give two results on how Ricci transforms under transformation of the Kernel.

First, we look at *laziness*. Given a kernel K, the associated lazy kernel is $K_\lambda := (1-\lambda)I + \lambda K$. It has the same steady state π.

PROPOSITION 7.1 (Lazy Ricci). *Let* $\lambda \in (0,1)$. *If* $\mathrm{Ric}(K) \geqslant \kappa$, *then*

$$\mathrm{Ric}(K_\lambda) \geqslant \lambda\kappa$$

PROOF. A direct calculation shows that the action and Bochner quantities under laziness are

$$\mathcal{A}_\lambda = \lambda\mathcal{A}; \quad \mathcal{B}_\lambda = \lambda^2\mathcal{B}$$

Recall then that $\mathrm{Ric}(K) \geqslant \kappa$ iff $\mathcal{B} \geqslant \kappa\mathcal{A}$. Then $\mathcal{B} \geqslant \kappa\mathcal{A}$ implies $\lambda\mathcal{B} \geqslant \kappa\mathcal{A}_\lambda$ and $\mathcal{B}_\lambda \geqslant \kappa\lambda\mathcal{A}_\lambda$, hence $\mathrm{Ric}(K_\lambda) \geqslant \kappa\lambda$. □

The next result is a *tensorisation* result on Ricci, i.e. Ricci of the product of Markov chains. We first define the setting to talk about product of Markov chains.

The Cartesian product between two spaces A, B is the set of all *ordered pairs* $A \times B := \{(a,b) : a \in A, b \in B\}$. Then the *product chain* on the cartesian product space is $\mathbf{X} = (X_A, X_B)$ for X_A, X_B Markov chains on each set.

For $i = 1, \dots, n$ let $(\mathcal{X}_i, K_i)$ be irreducible and reversible Markov kernels with steady state π_i and let α_i be non-negative numbers that sum to 1. The product chain on the space $\mathcal{X} = \prod_i \mathcal{X}_i$ for $\mathbf{x} = (x_1, \dots, x_n)$ and $\mathbf{y} = (y_1, \dots, y_n)$ is $X_\alpha = (X_1, \dots, X_n)$[5] and has

[5] Although I'm not sure whether this notation is used for independent chains in the simple spaces. Here we do not assume independence, otherwise the rules for K_α would be different.

kernel K_α defined as

$$K_\alpha(x,y) = \begin{cases} \sum_i \alpha_i K_i(x_i,x_i), & x_i = x_i \, \forall i \\ \alpha_i K_i(x_i,y_i) & x_i \neq y_i \\ 0 & \text{otherwise} \end{cases}$$

for $x = \{x_1, \dots, x_n\}$ and similarly y. The steady state is the product $\otimes$ of steady states.

This read like this: take $\mathbf{x}$ and $\mathbf{y}$ defined with one coordinate per set $\mathcal{X}_i$. We change state by changing at most one coordinate of $\mathbf{x}$. Hence, if they have more than 1 coordinate different, the probability of going from $\mathbf{x}$ to $\mathbf{y}$ is 0. When exactly one coordinate is different (say x_i), we look at the Markov kernel defined in the corresponding space $(\mathcal{X}_i, K_i)$ weighted by α_i. Finally, the probability of remaining in the same state is a weighted average of the probabilities of remaining on x_i for each space $(\mathcal{X}_i, K_i)$. It is easily checked that these sum to 1, for instance for the hypercube case where $\mathcal{X}_i = \{0,1\}$ and arbitrary weights summing to 1. Assuming that in each state, for $x_i \neq y_i$, we have that $K(x_i,x_i) + K(x_i,y_i) = 1$,

$$\sum_{\mathbf{y}} K_\alpha(\mathbf{x},\mathbf{y}) = 0 + \sum_i \alpha_i K(x_i,x_i) + \sum_i \alpha_i K(x_i,y_i) = \sum_i \alpha_i(1) = 1$$

where the first summation term is for the one case where $\mathbf{x} = \mathbf{y}$, the second summation is for the n cases where they differ for one coordinate (one for each i).

Then, the result is the following.

THEOREM 7.2 (Tense Ricci). *Assume that* $\mathrm{Ric}(K_i) \geqslant \kappa_i$ *for* $i = 1, \dots, n$. *Then we have*

$$\mathrm{Ric}(K_\alpha) \geqslant \min_i \alpha_i \kappa_i$$

PROOF. hihihi □

Applications of the tensorisation result are two, to assymetric random walks on the discrete hypercube and to the discrete torus. In the first case, we study the hypercube as the tensor product of $\{0,1\}$ spaces, with parameters $p = K(0,1)$ and $q = K(1,0)$. For the discrete torus, we look at it as the product of random walks on d discrete circles, hence we get that Ricci is non-negative from the example above and the tensorisation theorem.

8. Motivation of paper

Lott: manuscript on Ricci bound for metric spaces

Ollivier: other notion of Ricci for discrete. He says that it has many powerful implications in terms of these inequalities that have probabilistic interest

- concentration of measure and Lévy–Gromov theorem
- bounding the diamater of the space (Bonnet–Myers)
- Brownian motion and Lichnerowicz's theorem

Ollivier and Villani (2011): they look for a lower bound on Ricci of the discrete hypercube. there are two possible notions: (1) given by Ollivier, related to "spheres being closer between eachother than their centers when curv > 0", (2) uses Brunn-Minkowski inequality (κ-concavity of the logarithmic volume), which leads to studying displacement convexity of the entropy *because that volumes spread out when mass is preserved implies that density decreases and is equivalent to entropy increasing...?*

Comunque, le applicazioni sono principalmente due:

- studiare la Ricci di spazi come l'ipercubo (e perché questo è importante non lo so ancora, ma Ollivier e Villani ci hanno fatto un paper Ollivier and Villani (2011) + il mio paper dice hypercube fundamental building block in math physicis e comp sci)
- ottenere versione discreta di certe "extremely powerful inequalities" nel mondo continuo (e.g. Ollivier Ollivier (2009) riporta quelle scritte sopra. poi le log-Sobolov inequality si sentono sempre. poi bo)

hypercube è importante in cs per esempio perché poi lo usi per modellare diverse situazioni e.g. lattici in crittografia e studiare diversi algoritmi su di esso. (source)

Tutti i paper di oggi sono Maas Ollivier Ollivier-Villani Otto-Villani (inequalities) Bakry-Emery !!

Markov operators by Bakry et al ma bo!

CHAPTER 7

Paper: *Functional Inequalities for Markov Chains with Non-Negative Ricci Curvature*

Reference: Erbar and Fathi (2018), find it here

1. Keywords and sumup

Spectral graph, Cheeger isoperimetric constant, modified logarithmic Sobolev constant *of the chain*

Bound depends only on the **diameter of the space**

I think this makes sense because Markov chains are finite so we can play with a bounded space (?)

We talk about an *entropic* Ricci curvature bound, i.e. based on entropy

From the previous paper: a positive entropic Ricci curvature lower bound implies

- a spectral gap estimate,
- a modified logarithmic Sobolev inequality,
- and an analogue of Talagrand's transport cost inequality

Here: Ric bounded by below, not necessarily $\mathrm{Ric} > 0 \longrightarrow$ only weak additional information is needed, e.g. on a bound on the diameter of the space, to still establish strong functional inequalities.

Results (on $\mathrm{Ric} \geqslant 0$):

- isoperimetric inequality (bound on the Cheeger constant, discrete analogous to Buser theorem),
- estimate on the spectral gap (negative bound generalization is expressed in terms of Poincaré inequality)
- modified logarithmic Sobolev inequality (implies Talagrand)

Additionally:

- Section 5 → troubles with high dimensions, so they try using measure concentration bounds
- New technical tool → equivalent characterization of entropic Ricci curvature lower bounds in terms of gradient estimates
- Application: interacting particle system, namely the zero-range process on the complete graph with constant rates

2. Recap and additions

The setup is the same as the last paper. In this section we discuss the divergences or new things this paper introduces.

Something new is how they introduce the Markov generator, as an operator L acting on functions $\psi : \mathcal{X} \to \mathbb{R}$ defined by

$$L(\psi(x)) = \sum_{y\in\mathcal{X}} (\psi(y) - \psi(x))Q(x, y) \tag{25}$$

where Q is "a collection of transition rates", i.e. our Markov kernel, previously K. This gives us the chance to do the following recap.

2.1. Markov kernel, semigroup and generator. The **Markov kernel** is the matrix in $\mathbb{R}^{\mathcal{X}\times\mathcal{X}}$ with transition probabilities describing the discrete-time evolution of our Markov chain. Previously it was K, here we'll use $\mathbf{Q}$. Its properties are that it describes an homogeneous Markov chain (does not depend on time), it is irreducible and reversible (see more below). Moreover, Q has a unique steady state π, which is such that $\pi Q = \pi$.

Remark. We right multiply probability measures by Q. It can be shown that this gives the right result. The intuitive reason is that K_{ij} is the probability of moving $x_i \to x_j$, and we care about the arriving position x_j when updating the Markov chain. Indeed, $\rho_{t+1}(x_j)$ will be a weighted average of the probability of transiting to x_j.

We can look at our Markov chain as a continuous-time evolution of probability measures in the space of measures $\mathscr{P}(\mathcal{X})$(curves of measures/densities parametrized by t), describing the evolution of the distribution of points in $\mathcal{X}$. Recall that if there is a reference measure on $\mathcal{X}$, which in our case is the invariant measure π, we can work with probability densities instead of measures, which is what we do here. Then, the operator describing the evolution of our chain is called **Markov operator** and we refer to them as $\mathbf{P_t}$. The family of operators they form is called **Markov semigroup**, indicated as $(\mathbf{P_t})_{\mathbf{t}\geqslant\mathbf{0}}$.

Often, we will use the term semigroup to refer to a generic element of the family P_t. The Markov semigroup has the main property that $P_{t+s} = P_t \circ P_s$.

There is an exponential form for the Markov operator which is generally used as a definition, and which relates it to the discrete kernel. This is

$$P_t := e^{t(Q-I)} \tag{26}$$

Next, we show how this is obtained.

The matrix defining this matrix exponential is called **Markov generator**, it is $Q - I$ from the previous paper, and indicated as $\mathbf{L}$ in this paper.

Now, we recall some properties of the Markov kernel. We have irreducibility that implies uniqueness of the steady state. And we have reversibility, also called detailed balance.

2.2. Induced distance on $\mathcal{X}$. Restriction of $\mathcal{W}$ to Dirac masses gives rise to a new distance on $\mathcal{X}$,

$$d_{\mathcal{W}}(x, y) := \mathcal{W}(\delta_x, \delta_y) \tag{27}$$

and also an upper bound on $d_{\mathcal{W}}$ is given, in terms of a weighted graph distance d_Q.

2.3. Ricci in terms of $\mathcal{A}$ and $\mathcal{B}$. This is not new, but I disregarded it in the last paper.

2.4. Gradient estimates.

3. Inequalities

The inequalities we look at are an isoperimetric inequality (IPI), a Poincaré inequality (PI), and a log-Sobolev inequality (LSI).

Motivation. The reason why Bakry and Emery related this theoretical framework to functional inequalities in the first place is that "*Nous allons maintenant indiquer diverses formulations équivalentes à l'hyper- contractivité lorsque le processus est une diffusion. La plus importante d'entre elles est l'inégalité (ou plutôt les inégalités) de Sobolev logarithmique, due à Gross [2] (Gross se place dans un cadre bien plus général que celui des diffusions markoviennes)*" Bakry and Émery (2006). Also, Boudou et al. (2006) says "One aim of ergodic theory for Markov process is to understand whether Ttf converges, and in which sense, to the equilibrium average $\int f \, \mathrm{d}\pi$ and, if this is the case, to give quantitative estimates on the rate of convergence. One of the main tools in this context is provided by functional inequalities, in particular Poincar ́e inequality, logarithmic-Sobolev inequality and modified logarithmic-Sobolev inequality".

So, the reason why we're interested in studying them is that they are equivalent to the deacy of energy functional along the evolution of our diffusion process, implying convergence to equilibrium. Moreover, they have also other implications, i.e. the LSI yields hypercontractivity of the semigroup.

As we will see, their discovery is entrenched with a deeper meaning in the continuous setting, where there are a number of reasons why they're useful (hypercontractivity, concentration of measure, etc.). This makes it a good result to have an analogous of them for the discrete setting. Here, they give results on the eigenvalues of the Laplacian, describing how the process evolves. Apart from implying convergence to equilibrium, they yield combinatorial properties and quantities of graphs, e.g. solutions to the sparsest cut problems, hence clustering, or number of loops.

Random facts. (1.1)=PI, (1.2)=LSI, (1.3)=MLSI:

These three inequalities are hierarchically ordered in the following sense: if (1.2) holds with $s > 0$, then (1.3) holds with $\alpha \geqslant s/4$; if (1.3) holds with $\alpha > 0$, then (1.1) holds with $k \geqslant \alpha/2$ Boudou et al. (2006).

Apparently, it is well known that (LSI) is equivalent to hypercontractivity of the semigroup and (PI) is equivalent to exponential convergence to equilibrium in L2 Caputo et al. (2009).

3.1. Gauss theorem (divergence theorem). This theorem is good to know as it justifies the heat equation, and the equivalence of a Dirichlet form and the integral of the squared gradient, used in the inequalities below. It states that given a vector field **f**

$$\int_S \mathbf{f} \, \mathrm{d}S = \int_V \nabla \cdot \mathbf{f} \, \mathrm{d}V \tag{28}$$

where V is a bounded region whose boundary $\partial V = S$ is a piecewise smooth closed surface. The integral on the right-hand side is taken with the normal n pointing outward. We can build an intuition behind this theorem as follows. When considering a small enough region V, we have

$$\int_V \nabla \cdot \mathbf{f} \, \mathrm{d}V \approx V \cdot \nabla \cdot \mathbf{f}$$

with equality being exact as V shrinks to zero size. By taking the limit as $V \to 0$ and multiplying by $1/V$, the theorem gives us a coordinate independent definition of divergence. In fact, **the right way to think about divergence is as the net flow into, or out of, a region**. [1]

[1] Cambridge notes

Together with this theorem, we consider the product rule for the divergence: given a scalar-valued function g and a vector (field) $\mathbf{f}$

$$\nabla \cdot (\mathbf{f}g) = (\nabla g)\mathbf{f} + g(\nabla \cdot \mathbf{f})$$

Let us apply the Gauss theorem and this decomposition together:

$$\int_V \nabla \cdot (\mathbf{f}g)\,\mathrm{d}V = \int_V \nabla g \cdot \mathbf{f}\,\mathrm{d}V + \int_V g(\nabla \cdot \mathbf{f})\,\mathrm{d}V = \int_S \mathbf{f} \cdot g\,\mathrm{d}S$$

This gives us the integration by parts in higher dimension formula

$$\int_V (\nabla \cdot \mathbf{f}) \cdot g\,\mathrm{d}V = \int_S \mathbf{f} \cdot g\,\mathrm{d}S - \int_V \nabla g \cdot \mathbf{f}\,\mathrm{d}V \tag{29}$$

We can apply this rule to the vector field ∇f, getting the Laplacian $\Delta f = \nabla \cdot \nabla f$, and to the scalar function f. Letting the boundary term vanish, we get that the Dirichlet form $\mathcal{E}(f, f)$(as used, for instance, by Boudou et al. (2006)) is the integral of the squared gradient of f, as used here théoriques et appliquées. This connects the two notations.

3.2. Spectral graph theory. We can define two relevant quantities for a graph, in particular a d-regular graph $G = (V, E)$.

Definition 3.1 (Sparsest cut). *The sparsity of a partition $(S, V - S)$ of a graph G (a cut) is*

$$\sigma(S) := \frac{\mathbb{E}_{(u,v)\sim E}\left|\mathbf{1}_S(u) - \mathbf{1}_S(v)\right|}{\mathbb{E}_{(u,v)\sim V^2}\left|\mathbf{1}_S(u) - \mathbf{1}_S(v)\right|} \tag{30}$$

i.e. the fraction of edges broken by the cut over the total fraction of vertices separated by the cut.

Then, the sparsest cut problem is to find the set of minimal sparsity, and the solution defines the sparsity of the whole graph G:

$$\sigma(G) := \min_{S\subset V; S\neq\varnothing} \sigma(S) \tag{31}$$

Remark. This terminology is quite confusing and entails an inversion. A higher value of σ implies that cuts are dense, in the sense that the number of edges they cut is large compared to the number of edges they separate. The terminology comes from optimization and approximation algorithms and the sparsest cut problem: from the need to find the cut that is as sparse as possible, they named the quantity that is optimized (minimized) *sparsity*.

DEFINITION 3.2 (Edge expansion). *In a d-regular graph, the edge expansion of a subset of vertices $S \subseteq V$ is*

$$\phi(S) := \frac{E(S, V-S)}{d \cdot |S|} \tag{32}$$

i.e. the ration between the number of edges between S and $V - S$ and the number of edges incident to any vertex in S. It is the fraction of edges incident to vertices in S which "go out" of S.

Similarly, we define the edge expansion of a graph as

$$\phi(G) := \min_{S:|S| \leqslant |V|/2} \phi(S) \tag{33}$$

where the condition on $|S|$ is equivalently $|S| \leqslant |V - S|$

These two quantities stay in the following relation:

$$\frac{1}{2}\sigma(S) \leqslant \phi(S) \leqslant \sigma(S)$$

and since $\sigma(S) = \sigma(V - S)$, it also holds

$$\frac{1}{2}\sigma(G) \leqslant \phi(G) \leqslant \sigma(G) \tag{34}$$

Then, we call a family of constant degree ***expanders*** a family of d-regualar graphs $\{G_n\}_{n \geqslant d}$ such that there is an absolute constant $\phi > 0$ such that $\phi(G_n) \geqslant \phi$ for every n (number of vertices).

This condition requires the minimal expansion of the graph to be high enough. This means, intuitively, that an adversary will have to cut a larger fraction of edges to separate a subset of vertices from the rest of the graph. In fact, ***expanders are sparse graphs with good connectivity properties***.

Here *sparse* should be interpreted in the vanilla way, i.e. in the sense that the number od edges is fixed $= d$, whereas the number of vertices $n \geqslant d$ can increase. This is not necessarily related to *cut sparsity* as defined above - a stupid example is a very dense graph in the vanilla sense, with 2 separated components. Indeed, also for expander, cut sparsity is high (equation 34).

Moreover, in other literatures, the edge expansion of a graph is referred to as *Cheeger constant* and indicated as h_G department.

Attempt: by looking at the graph in the Markov chain sense, vanilla sparsity implies that the adjacency matrix is sparse, which may be relaxed to a probabilistic point of view by having many transitions that happen with low probability. Cut sparsity and expansion, instead, may have to do with the change over time of the mass distribution, I mean that high values may imply an high probability that the Markov chain takes values in a limited area of the graph for a long time, which means the distribution of mass remains more or less the same, which may have to do with convergence.

Now, let us consider a d-regular graph $G = (V, E)$ and its adjacency matrix A. We can define the Laplacian operator of G as[2]

$$L := 1 - \frac{1}{d}A$$

where we notice that it is the very same definition of the generator of a Markov chain we had before, as our Markov kernel can be interpreted as $K = \frac{1}{d}A$!! Then, the following theorem is a fundamental result in spectral graph theory, relating the spectrum of L to combinatorial properties of our graph.

THEOREM 3.3. *Let G be a d-regular undirected graph, and $L = I - \frac{1}{d}\,A$ be its normalized Laplacian matrix. Let $\lambda_1 \leqslant \lambda_2 \leqslant \cdots \leqslant \lambda_n$ be the real eigenvalues of L with multiplicities. Then*

(1) $\lambda_1 = 0$ and $\lambda_n \leqslant 2$.
(2) $\lambda_k = 0$ if and only if G has at least k connected components.
(3) $\lambda_n = 2$ if and only if at least one of the connected components of G is bipartite.

PROOF. The proof makes repeated use of the following identity, whose proof is immediate: if L is the normalized Laplacian matrix of a d-regular graph G, and $\mathbf{x}$ is any vector, then

$$\mathbf{x}^{\mathsf{T}} L \mathbf{x} = \frac{1}{d} \sum_{\{u,v\} \in E} \left(x_u - x_v\right)^2. \tag{35}$$

Hence, by the variational characterization of eigenvectors

$$\lambda_1 \;=\; \min_{\mathbf{x} \in \mathbb{R}^n \setminus \{0\}} \frac{\mathbf{x}^{\mathsf{T}} L \mathbf{x}}{\mathbf{x}^{\mathsf{T}} \mathbf{x}} \;\geqslant\; 0.$$

[2] In the physicists' style

If we take $\mathbf{1} = (1, \dots, 1)$ to be the all-ones vector, we see that $\mathbf{1}^\mathsf{T} L \mathbf{1} = 0$, and so 0 is the smallest eigenvalue of L, with $\mathbf{1}$ being one of the vectors in the eigenspace of 1.

[etc etc]

□

This tells us that the first (lowest) eigenvalue is always 0 and is related with the existence of an invariant measure. Also, the first two properties imply that the multiplicity of 0 as an eigenvalue is precisely the number of connected components of G.

Therefore, what is of interest for us is the second eigenvalue, that is related with connectivity and the sparsest cut problem. We have

$$\lambda_2 = 0 \iff G \text{ has} \geqslant 2 \text{ connected components} \iff h_G = 0$$

because it means that there are two subgraphs in G with no edges between them (definition of connected components), hence they give a trivial cut with sparsity 0 and a trivial subset of vertices with expansion 0. The Cheeger inequality relaxes the previous result by saying that "λ_2 small $\Leftrightarrow$ h_G small":

$$\frac{\lambda_2}{2} \leqslant h_G \leqslant \sqrt{2 \cdot \lambda_2} \tag{Cheeger}$$

3.3. Isoperimetric inequality. The paper we're studying finds a bound on the Cheeger constant in terms of the spectral gap of the generator L. This is referred to as isoperimetric inequality, and it is formulated as

$$h \geqslant \frac{1}{3}\sqrt{Q_* \lambda_1} \tag{IPI}$$

where λ_1 indicates the spectral gap (change of notation w.r.t. before) and $Q_* \in \mathbb{R}$ being the minimal transition rate in the Markov kernel Q.

Notice also that here the Cheeger constant is expressed differently, i.e. as

$$h = \max_{A \subset \mathcal{X}} \frac{\pi^+(\partial A)}{\pi(A)(1 - \pi(A))}$$

Here, the numerator is the perimeter measure of A, $\pi^+(\partial A) = \sum_{x \in A, y \in A^c} Q(x,y)\pi(x)$. We can look at it as a probabilistic version of the edge count $E(S, V - S)$ we had before. If we consider that $\pi(A)$ as a probability of being in A, the denominator is the expectation of the measure of A. In the context of graphs, we can think of π as just a measure we want on our discrete space, and the measure of a vertex is generally its degree department. This is consistent with the denominator $d|A|$ we had above.

As said above, this is a bound on the fraction of edges that one needs to cut to isolate a portion of vertices of the graph. This inequality in fact has important implications in the problem of graph cuts, which is related to problems like clustering, shortest paths, belief propagation and optimal hypersurface separation, with applications in statistical mechanics, computer vision, and others.

3.4. Poincaré Inequality. The Poincaré inequality yields a bound on the first eigenvalue, via the spectral gap, of the Laplacian.

It is a bound on the variance functional:

$$\mathrm{Var}(f) \leqslant \int_{\mathbb{R}^n} \left|\nabla f\right|^2 \mathrm{d}\pi \tag{36}$$

3.5. Logarithmic Sobolev Inequality. The log-Sobolev inequalities were discovered by Gross in the context of constructive quantum field theory Gross (1975), which, in simple words, has to do with diffusion processes in a more general sense than our specific case of Markovian processes. What the inequalities do is they describe the distribution of the eigenvalues of the Laplacian operator.

It is a bound on the entropy functional:

$$\mathcal{H}(f) \leqslant \frac{1}{2} \int_{\mathbb{R}^n} \frac{\left|\nabla f\right|^2}{f} \mathrm{d}\pi \tag{37}$$

4. Applications

Given the work by Erbar and Maas, which also shows implications of a lower bound on Ricci for a (modified) logarithmic Sobolev inequality, a Talagrand transportation inequality, and a Poincaré inequality, it makes sense to try and obtain sharp Ricci curvature bounds in concrete discrete examples. These are a few results up to today

- Erbar and Maas: tensorization principle, crucial for application to high dimensions (discrete hypercube)
- Erbar, Maas and Tetali: Bernoulli–Laplace model and for the random transposition model on the symmetric group (sempre high-dimensional result)
- Mielke: one-dimensional birth and death processes

In spite of these, a **systematic approach for obtaining discrete Ricci bounds has been lacking**. Here in Fathi and Maas (2016) they propose this method and allow to recover the bound for many interacting particle systems on the complete graph, among which the above examples and the zero-range process Fathi and Maas (2016).

4.1. The zero-range process with constant rates.

Why the zero-range process? *This paper Fathi Erbar ← Fathi and Maas (2016) Fathi Maas ← Caputo et al. (2009) Caputo Da Pra Posta ← Boudou et al. (2006) Boudou Caputo Da Pra Posta*: The study of functional inequalities for interacting particle systems has been motivated by both **theoretical and computational purposes**, and has led to the development of a **rather sophisticated mathematical technology**. For instance, such inequalities on interacting particle systems have been used in statistical mechanics and physics, to model things like the Glauber dynamics (for magnetism) or spin models. Moreover, there is this wide range of literature each piece of which contributes a little to the bigger aim adapting the continuous definitions and approaches to discrete examples to obtain the same functional inequalities results. There are two foundational papers concerning functional inequalities and Ricci:

- Bochner who first approaches the study of estimates on the spectral gap of the Laplacian on Riemannian manifolds through Ricci (1946)
- Bakry and Emery who, in addition to estimates on spectral gap, estimates on the LSI, for diffusion processes in a more general context (remember LSI is equivalent to hypercontractivity of the semigroup) (1985)

The work by Bakry and Emery has inspired several developments, especially concerning diffusion models motivated by statistical mechanics. Among these, a possible timeline on papers tackling the zero-range example is:

Boudou et al. (2006) applies the Bochner identity to estimate the spectral gap of Markov chains;

Caputo et al. (2009) develops a method to obtain estimates on the exponential rate of decay of the relative entropy from equilibrium of Markov processes in discrete settings, again based on Bochner;

Fathi and Maas (2016) is based on the result of Erbar and Maas studied in july, and proposes a systematic method for obtaining discrete Ricci curvature bounds, which is based on Caputo et al. (2009) and yields new curvature bounds for zero-range processes on the complete graph;

[This paper]finally, the present paper, apart the results on functional inequalities with non-neg Ricci in general, use the bound in Caputo et al. (2009) to apply the results to the zero-range process.

domanda dal pubblico: ma c'è proprio bisogno di passare per Ricci per dimostrare queste inequalities per questi processi così semplici?

What is the zero-range process? Consider a conservative interacting system of finitely many particles moving with jumps in a finite set of states. At each iteration, at most one particle can move from one state to another, and the rate at which particles jump only depends on the number of particles in the initial state. The continuous-time Markov chain generator of this process is the zero-range process on the complete graph. The rate can be modeled as a function of number of particles at the initial node. The name *zero-range* comes from the idea of *range of an interaction* in physics. Among the four **fundamental interactions** (gravitational, electromagnetic, weak nuclear and strong nuclear), gravity and electromagnetism are interactions that produce long-range forces, which we experience everyday. In our particle system, the interaction has a spatial range equal to 0, because the mutual stochastic influence of particles is just local, as the jump rate depends only on the particles that are in the initial present state. Thus, we can look at this "influence" as a zero-range interaction, an interaction that has just a local effect. It is called an *interacting* particle system because the jump process of a particle depends stochastically on the others.

The setup is the following: consider the state space $\mathcal{X}_{K,L} := \{\eta \in \mathbb{R}^L : \sum_l (\eta)_l = K\}$, which are vectors describing the situation, i.e. the number of particles in each node. Then, we formalize the movement of particles as follows. Pick a node l uniformly at random and observe η_l, if the number of particles is positive, pick another node m uniformly at random where one particle will move. Then, $n^{l,m}$ indicates the new configuration. This process can be represented by a Markov chain (on the state space $\mathcal{X}_{K,L}$) with kernel

$$Q_{K,L}(\eta,\eta') = \begin{cases} \frac{1}{L} & \eta' = \eta^{l,m} \text{ for some } l, m, \\ 0 & \text{otherwise} \end{cases} \tag{38}$$

Notice that here we consider a degenerate process where jumping rates are constant at $1/L$. Constant rates imply that the invariant measure of $Q_{K,L}$ is the uniform distribution denoted by $\pi_{K,L}$.

Then, the ingredients are the following:

(1) positive bound on Ricci from Fathi and Maas (2016) for increasing jump rates $\longrightarrow$ easily applicable to get a non-negative bound for constant jump rates
(2) lemma on a diameter estimate
(3) Ricci bound + diameter estimate → MLSI for degenerate zero range process

What is the MLSI and what does it tell us. We will actually look at Poincaré inequality and logaritmic Sobolev together.

(1) The first formulation and interpretation of these inequalities is probabilistic, and they express a bound on the variance and the entropy of an observable function f via a norm of the gradient of f.
(2) These inequalities imply and enable us to quantify the exponential decay of variance and entropy of f along the evolution of the Markov semigroup.
(3) These results imply the convergence to equilibrium of the chain.
(4) Moreover, they have deeper implications regarding the spectral graph (PI) or a hypercontractivity property (LSI).

to do: look at the notes to expand on these facts above; conclude the application part saying what this tells us for the process (slides included); look at the rest of the paper

To understand the PI e LSI we start from the OU process. This represents the heat flow in $\mathbb{R}^n$ with the Gaussian measure γ_n. The Gaussian measure is in fact its stationary measure, so that if P_t is the OU semigroup,

$$\int_{\mathbb{R}^n} P_t f \,\mathrm{d}\gamma_n = \int_{\mathbb{R}^n} f \,\mathrm{d}\gamma_n$$

where $f : \mathbb{R}^n \to I, I \subset \mathbb{R}$ is our observable.

Then, we consider ϕ-entropies of f for a convex $\phi : I \to \mathbb{R}$,

$$\mathbb{E}^{\phi}_{\gamma_n}(f) = \int_{\mathbb{R}^n} \phi(f) \,\mathrm{d}\gamma_n - \phi\left(\int_{\mathbb{R}^n} f \,\mathrm{d}\gamma_n\right)$$

this generalizes the variance for $\phi(x) = x^2$ and Shannon's entropy for $\phi(x) = x \log x$.

The first result is that

...

This is proven by considering

$$\alpha(t) := \dots$$

and checking that $\alpha(0) = \dots$, $\alpha(\infty) = \dots$ By applying ... (change of variables and integrating by parts) we get the result.

Then, ...

4.1.1. *Getting convergence to equilibrium.* The MLSI estimate for the ZRP is then used to get converegence to equilibrium of the process. In particular, this convergence is expressed in terms of the **mixing time**.

The mixing time of a Markov chain is, in general, the smallest time in which the Markov semigroup is ε-close to the invariant measure. We consider in particular the total variation mixing time.

Definition 4.1 (Total variation mixing time). *The total variation mixing time is defined for $\varepsilon > 0$ as*

$$\tau_{\text{mix}}(\varepsilon) := \inf \left\{ t > 0 : \|P_t^* \delta_x - \pi\|_{TV} < \varepsilon \quad \forall x \in \mathcal{X} \right\}$$

Next we state the main result and clarify all the steps to obtain it.

Theorem 4.2 (Bound on mixing time of ZRP).

$$\tau_{\text{mix}}(\varepsilon) \leqslant KL \log L \left(\frac{1}{8} - \frac{\log \varepsilon}{c} \right) \tag{39}$$

We defer the specific definition of total variation distance below.

EB: E' necessario chiarire come tutti questi concetti e queste disuguaglianze si combinano per ottenere la stima sul tempo di mixing.

(1) Cosa centra MLSI con il fatto che il semigruppo è il flusso dell'entropia? Questo punto è talmente fondamentale che potresti considerare di far vedere il calcolo per ottenere entropy decay

(2) Capito (1), come si combina entropy decay e Pinsker per concludere? Anche questo è da esplicitare

(1) **How MLSI + P_t being the heat semigroup + Fisher being the derivative of the entropy imply entorpy decay?**

[Continuous case]

That P_t is the heat semigroup follows by construction of $\mathcal{W}$, and its relation to the gradient flow of the entropy was derived in Section 1. We show how Fisher is the derivative of the entropy.

Recall that for a probability densty $\rho \in \mathscr{P}(X)$,

$$\mathcal{H}(\rho) := \int \rho \log \rho \, \mathrm{d}\pi$$
$$\mathcal{I}(\rho) := \int \rho |\nabla \log \rho|^2 \, \mathrm{d}\pi$$

Then, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{H}(P_t\rho) &= \int \frac{\mathrm{d}}{\mathrm{d}t} P_t\rho \log(P_t\rho) \, \mathrm{d}\pi \\ &= \int \left(\frac{\mathrm{d}}{\mathrm{d}t}P_t\rho\right) \log P_t\rho + \frac{1}{P_t\rho} P_t\rho \left(\frac{\mathrm{d}}{\mathrm{d}t}P_t\rho\right) \mathrm{d}\pi \\ &= \int \frac{\mathrm{d}}{\mathrm{d}t}P_t\rho\Big(\log P_t\rho + 1\Big) \, \mathrm{d}\pi = \int \Delta P_t\rho\Big(\log P_t\rho + 1\Big) \, \mathrm{d}\pi\end{aligned}$$

where in the last passage Δ is the Laplacian operator and equality follows from P_t being the heat semigroup. Then, integrating by parts and usnig that $\nabla \log f = \nabla f / f$ we get

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{H}(P_t\rho) = \int \nabla P_t\rho \cdot \frac{\nabla P_t\rho}{P_t\rho} \, \mathrm{d}\pi = \mathcal{I}(P_t\rho)$$

Remark. Consider the generator of the chain $\mathcal{L}$, the invariant measure π and functions $f : X \to \mathbb{R}$. For consistency with most of the literature, define the Dirichlet form as

$$\mathcal{E}(f, g) := \int_{\mathcal{X}} f \mathcal{L} g \, \mathrm{d}\pi$$

Then, we define Fisher infromation of a probability denisty $\rho \in \mathcal{P}(X)$ as

$$\mathcal{I}(\rho) = \int_X \|\nabla \log \rho\|^2 \, \mathrm{d}\pi$$

Then, integrating by parts using Gaussian formula we see $\mathcal{E}(\rho, \log \rho) = \mathcal{I}(\rho)$.

Now, recall the MLSI:

$$\mathcal{H}(\rho) \leqslant \frac{1}{2\lambda}\mathcal{I}(\rho) \tag{MLSI}$$

by using the result above we get that the entropy is bounded by its derivative as

$$\mathcal{H}(P_t\rho) \leqslant -\frac{1}{2\lambda}\mathcal{H}'(P_t\rho)$$

Notice that this implies

$$\mathcal{H}'(P_t\rho) \leqslant -2\lambda\mathcal{H}(P_t\rho)$$

Then, we can apply Gronwall's inequality to obtain entropy decay:

$$\mathcal{H}(P_t\rho) \leqslant e^{-2\lambda t}\mathcal{H}(\rho) \tag{40}$$

THEOREM 4.3 (Gronwall's inequality). *Gronwall's inequality is a fundamental result in evolution problems. It states that a nonnegative function $\phi(t)$ such that there exists a constant C such that $\phi'(t) \leqslant C(t)\phi(t)$ satisfies*

$$\phi(t) \leqslant \phi(0)\exp\left(\int_0^t C(\tau)\,\mathrm{d}\tau\right)$$

PROOF. Consider the quantity

$$v(t) := \exp\left(\int_a^t C(\tau)\,\mathrm{d}\tau\right)$$

and observing that $v'(t) = v(t)C(t)$, $v(a) = 1$. Then using the quotient formula we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\phi(t)}{v(t)} = \frac{\phi'(t)v(t) - v'(t)\phi(t)}{v(t)^2} = \frac{\phi'(t)v(t) - C(t)\phi(t)v(t)}{v^2(t)} \leqslant 0$$

by assumption. Then the function $\phi(t)/v(t)$ is decreasing and hence bounded by its value at the lower extreme a. This yields

$$\frac{\phi(t)}{v(t)} \leqslant \frac{\phi(a)}{v(a)} = \phi(a)$$

which by setting $a = 0$ yields the result. □

I also think that entropy decay is, in general, equivalent to convergence to equilibrium because entropy going to 0 means that the evolution of the chain is converging to a known quantity, and that can only be the invariant measure.

However, next we demonstrate how entropy decay and the pinsker inequality allow us to express convergence to equilibrium in terms of a bound on the mixing time.

(2) **How entropy decay + Pinsker inequality imply bound on mixing time?**

DEFINITION 4.4 (Total variation distance). *Given two probability measures $\mu, \nu \in \mathscr{P}(\mathcal{X})$, the total variation (TV) distance between them is:*

$$\|\mu - \nu\|_{TV} := \max_{A\subseteq\mathcal{X}} |\mu(A) - \nu(A)|$$

The TV distance is attained taking the subset of events where one probability distribution dominates the other, otherwise contribution would cancel out and the total sum of differences would be lower.

Moreover, it stays in the following relation with the L_1 distance $\|\mu-\nu\|_1 := \sum_{\mathcal{X}} |\mu(x)-\nu(x)|$,

$$\|\mu-\nu\|_{TV} = \frac{1}{2}\|\mu-\nu\|_1$$

To see this, just notice that the probability mass in which $\mu \geqslant \nu$ must be equal to the probability mass in which the contrary happens, for the two to sum to 1.

Theorem 4.5. *Pinsker's inequality Let μ and ν be two probability distributions over $\mathcal{X}$. Then*

$$\|\mu-\nu\|_{TV}^2 \leqslant \frac{1}{2}D_{KL}(\mu||\nu)$$

Proof. By what said above, we can equivalently prove that

$$\frac{1}{2}\|\mu-\nu\|_1^2 \leqslant D_{KL}(\mu||\nu)$$

Now, let $A = \{\mu(x) \geqslant \nu(x)\}$ and define the two Bernoulli random variables

$$\mu_A := \begin{cases} 1 & \text{w.p. } \sum_{x\in A}\mu(x) \\ 0 & \text{w.p. } \sum_{x\notin A}\mu(x) \end{cases}$$

$$\nu_A := \begin{cases} 1 & \text{w.p. } \sum_{x\in A}\nu(x) \\ 0 & \text{w.p. } \sum_{x\notin A}\nu(x) \end{cases}$$

Consider the following function

$$f(\mu_A,\nu_A) := D_{KL}(\mu_A||\nu_A) - \frac{1}{2}\|\mu_A-\nu_A\|_1{}^2$$

Let us define $p := \sum_{x\in A}\mu(x)$ and $q = \sum_{x\in A}\nu(x)$ to study f. Clearly, if $p = q$ then $f = 0$. Assuming w.l.o.g. $p \geqslant q$, we have that f decreases with q: this is verifiable by making f explicit and studying the derivative, seeing that the term in D_{KL} (=$-\frac{1}{q(1-q)}$) leads the derivative over the term in the 1-norm (=4).

Then, we have that the function is decreasing in q and =0 if $p = q$. Hence for $p \geqslant q$ we have $f \geqslant 0$ hence

$$D_{KL}(\mu_A||\nu_A) \geqslant \frac{1}{2}\|\mu_A-\nu_A\|_1^2$$

We have

$$D_{KL}(\mu||\nu) \geqslant D_{KL}(\mu_A||\nu_A)$$
$$\|\mu_A - \nu_A\|_1 = \|\mu - \nu\|_1$$

where first inequality follows from the data processing inequality: the intuition is that the distance between μ_A and ν_A depends on the true distributions but also on the probability that $x \in A$, hence there is an additional potential cofounding variable between the variables μ_A, ν_A w.r.t. μ, ν; the second one is easily observable by computing the 1-norm splitting the summation into $x \in A$ and $x \notin A$. Putting together the three inequalities yields the result:

$$D_{KL}(\mu||\nu) \geqslant D_{KL}(\mu_A||\nu_A) \geqslant \frac{1}{2}\|\mu - \nu\|_1{}^2 = 2\|\mu - \nu\|_{TV}^2$$

□

For our aims it makes sense to consider $\nu = \pi$ where π is the unifrom measure, and Pinsker's inequality becomes

$$\|\mu - \pi\|_{TV}^2 \leqslant \frac{1}{2}\mathcal{H}(\mu) - \log p_\pi \tag{Pinsker}$$

where p_π is the mass assigned by the uniform distribution on each x. In our specific case, we look at the entropy of probability *densities*, and the denisty of mass assigned by the uniform distribution is 1 everywhere, so the log term cancels.

Recall also the entropy decay result

$$\mathcal{H}(P_t\rho) \leqslant e^{-2\lambda t}\mathcal{H}(\rho) \tag{Entropy decay}$$

Next we show how (Pinsker) and (Entropy decay) combine to obtain inequality (39), our main result.

Proof of theorem (4.2). First, we use a result that comes from the old paper, sepcifically from the evolutional variational inequality:

$$\mathcal{H}(P_t\rho) \leqslant \frac{\mathcal{W}(\rho, 1)}{4t}$$

This allows us to bound the entropy, $\mathcal{H}(P_t\rho) \leqslant 2$ for

$$t \geqslant \frac{D^2}{8} \tag{Bound #1}$$

Setting $t_0 = \frac{D^2}{8}$, we apply the entropy decay inequality considering this initial time. Also, we consider the evolution of the chain in terms of probability *measures*, accoridng to the dual semigroup P_t^*, to study convergence to the invariant measure π.

$$\|P_t^*\delta_x - \pi\|_{TV} \leqslant \sqrt{\frac{1}{2}\mathcal{H}(P_t^*\delta_x)} \leqslant \sqrt{\frac{1}{2}e^{-2\lambda t}\mathcal{H}(P_{t_0}^*\delta_x)} \leqslant \sqrt{\frac{1}{2}e^{-2\lambda t}2} = e^{\lambda t}$$

Then, we study for which t we get ε-close to the invariant measure:

$$e^{\lambda t} \leqslant \varepsilon; \quad -\lambda t \leqslant \log\varepsilon; \quad t \geqslant \frac{\log\varepsilon}{\lambda}$$

Substituting our constant $\lambda = \frac{c}{D^2}$ for $\mathrm{Ric} \geqslant 0$, we get the second bound

$$t \geqslant D^2\frac{\log\varepsilon}{c} \tag{Bound \#2}$$

Then, we sum the two bounds to obtain our resulting bound:

$$t \geqslant D^2\left(\frac{1}{8} - \frac{\log\varepsilon}{c}\right)$$

As τ_{mix} is defined as the infimum of such times t, we get an upper bound on the mixing time as in (39).

□

Bibliography

Luigi Ambrosio, Elia Brué, Daniele Semola, et al. *Lectures on optimal transport*, volume 130. Springer, 2021.

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

Anne-Severine Boudou, Pietro Caputo, Paolo Dai Pra, and Gustavo Posta. Spectral gap estimates for interacting particle systems via a bochner-type identity. *Journal of Functional Analysis*, 232(1):222–258, 2006.

Pietro Caputo, Paolo Dai Pra, and Gustavo Posta. Convex entropy decay via the bochner-bakry-emery approach. In *Annales de l'IHP Probabilités et statistiques*, volume 45, pages 734–753, 2009.

Sara Daneri and Giuseppe Savaré. Eulerian calculus for the displacement convexity in the wasserstein distance. *SIAM Journal on Mathematical Analysis*, 40(3):1104–1122, 2008.

UCSD Math department. Isoperimetric problems. URL https://fanchung.ucsd.edu/research/cb/ch2.pdf.

Matthias Erbar and Max Fathi. Poincaré, modified logarithmic sobolev and isoperimetric inequalities for markov chains with non-negative ricci curvature. *Journal of Functional Analysis*, 274(11):3056–3089, 2018. ISSN 0022-1236. doi: https://doi.org/10.1016/j.jfa.2018.03.011. URL https://www.sciencedirect.com/science/article/pii/S0022123618301101.

Matthias Erbar and Jan Maas. Ricci curvature of finite markov chains via convexity of the entropy. *Archive for Rational Mechanics and Analysis*, 206(3):997–1038, 2012.

Max Fathi and Jan Maas. Entropic ricci curvature bounds for discrete interacting systems. 2016.

Leonard Gross. Logarithmic sobolev inequalities. *American Journal of Mathematics*, 97(4): 1061–1083, 1975. ISSN 00029327, 10806377. URL http://www.jstor.org/stable/2373688.

John M Lee. *Smooth manifolds.* Springer, 2003.

John M Lee. *Introduction to Riemannian manifolds*, volume 2. Springer, 2018.

John Lott and Cédric Villani. Ricci curvature for metric-measure spaces via optimal transport. *Annals of Mathematics*, pages 903–991, 2009.

Jan Maas. Gradient flows of the entropy for finite markov chains. *Journal of Functional Analysis*, 261(8):2250–2292, 2011.

Shin-Ichi Ohta. *Ricci curvature, entropy, and optimal transport*, page 145–200. London Mathematical Society Lecture Note Series. Cambridge University Press, 2014.

Yann Ollivier. Ricci curvature of markov chains on metric spaces. *Journal of Functional Analysis*, 256(3):810–864, 2009.

Yann Ollivier and Cédric Villani. A curved brunn–minkowski inequality on the discrete hypercube, 2011. URL `https://arxiv.org/abs/1011.4779`.

Michael Spivak. *Calculus on manifolds: a modern approach to classical theorems of advanced calculus*. CRC press, 2018.

Université PSL Master Mathématiques théoriques et appliquées. Logarithmic sobolev inequalities essentials. URL `https://djalil.chafai.net/docs/M2/chafai-lehec-m2-lsie-lecture-notes.pdf`.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2008.

APPENDIX A

More Math

LEMMA 0.1 (Extension lemma for smooth functions). *Suppose M is a smooth manifold with or without boundary, $A \subseteq M$ is a closed subset, and $f : A \to \mathbb{R}^k$ is a smooth function. For any open subset U containing A, there exists a smooth function $\tilde{f} : M \to \mathbb{R}^k$ such that $\tilde{f}|_A = f$ and supp$\tilde{f} \subseteq U$.*

Lebesgue vs. Hausdorff La Lebesgue measure calcola solo un volume della dimensione dello spazio, e si usa nel caso Euclideo. La Hausdorff measure generalizza Lebesgue perché $\mathcal{H}_n = \mathcal{L}_n$, ma posso avere anche la misura di dimensione minore (e.g. se $n = 3$, Hausdorff mi definisce un volume ma anche una 2d distanza, quindi posso misurare le lunghezze di una superficie immersa in $\mathbb{R}^3$).

Variabili aleatorie e misure di probabilità Puoi vedere una variablile aleatoria come sullo stesso piano concettuale della misura di probabilità. Il passaggio più semplice è da variabile aleatoria $X \longrightarrow$ a misura di probabilità μ:

$$\text{Dato un insieme F,} \quad \mu(F) = \Pr(X \in F)$$

Diciamo che *μ è la legge della variabile aleatoria X* oppure *μ è la distribuzione di probabilità di X.*

Def: μ è assolutamente continua quando data una misura di volume vol indotta dalla metrica g, μ si può scrivere come una densità per la misura,

$$\mu = \rho \cdot \mathrm{vol}_g$$

e scriviamo $\mu << \mathrm{vol}_g$. Questo vuol dire che μ non è molto concentrata rispetto alla misura di volume. Esempio stupido è quando la misura μ è tutta concentrata in un punto, allora se io voglio trasportarla in una misura ν che è concentrata su due punti, non lo potrò mai fare con una funzione. Infatti usiamo misure AC per definire il problema di trasporto ottimale.

Quando μ è anche assolutamente continua, allora X ha una funzione di densità.

Inoltre, quando definiamo un piano di trasporto, lo possiamo vedere come una distribuzione di probabilità congiunta:

$$\pi(A \times Y) = \mu(A) \quad \longleftrightarrow \quad p(x) = \int p(x,z)\,dz$$

dove x è una realizzazione della variabile X e p la sua pdf.

A metric induces a distance and a volume measure Once we have chosen a metric, we get a ***distance*** and a ***volume measure*** that are the two canonical objects induced by the metric.

The distance is denoted here as d_g or just d and it is defined by Ambrosio et al. (2021) as

$$d_g^2(x,y) := \min\left\{ \int_0^1 g_{\gamma(t)}(\gamma'(t),\gamma'(t))\,dt : \gamma(0) = x, \gamma(1) = y, \gamma \in AC([0,1];M)\right\}$$

To grasp the idea, consider that in the Euclidean case the distance is defined as

$$\int_0^1 |\gamma'(t)|\,dt$$

where $|\gamma'(t)| = \sqrt{(\gamma'(t),\gamma'(t))} = \sqrt{g_{\gamma(t)}(\gamma'(t),\gamma'(t))}$ where here g is the standard Euclidean metric. In plain English, we are integrating velocity against time, hence computing the distance as velocita × tempo.

Then, we define the volume measure vol_g in either of these ways:

$$\mathrm{vol}_g := \varphi_{\#}\sqrt{\det(g_x)_{ij}}\,\mathcal{L}^n$$

where φ is a chart and $\mathcal{L}^n$ is the Lebesgue measure. $\varphi_{\#}$ is the *push forward* according to the chart. [1]

The idea is to multiply the Lebesgue measure by a quantity that indicates how much volumes are deformed on our manifold (according to g).

Alternatively, we define vol_g as the Hausdorff measure induced by d_g, i.e.

$$\mathrm{vol}_g = \mathcal{H}^n(S) := \inf\left\{ \sum_{i=0}^{\infty}\big(\mathrm{diam}\,U_i\big)^n : S \subseteq \bigcup_{i=0}^{\infty} U_i\right\}$$

where [...]

[1] This means that the volume is defined as the sqrt times the Lebesgue measure on the subset of the Euclidean space that is the image of our chart, and we push forward this through φ to get a definition of volume on the manifold. The push forward of a measure is exactly an operation that allows to use a measure defined on an input space to define a measure on another space through a map.

APPENDIX B

Extra

0.1. Flat connection. EXERCISE 1.1:

Definition 0.1 (Flat connection). *A flat connection is a connection whose Riemannian curvature is 0, i.e. $R(X, Y)Z = 0$ for any vector fields X, Y, Z.*

Remark. In fact, the curvature is dependent on the choice of connection. I think that this is why we generally choose the Levi Civita connection (and the metric induced by the Euclidean inner product) to get generalizable results.

We now want to show that the connection induced by the inner product in $\mathbb{R}^3$ - equivalently, the LC connection in $\mathbb{R}^3$ - is flat.

This follows directly from the definition given by Lee (Lee (2018)) of Euclidean connection and from the fact that the LC connection on an Euclidean space is the Euclidean connection. Indeed, we take for granted that $\nabla_X Y = XY$ and by applying the definition of Lie bracket we get flatness. In particular, by following the steps performed by Lee (p.194 onwards), we have that by definition of Euclidean connection

$$\begin{aligned}
&\nabla_X\nabla_Y Z = \nabla_X Y(Z^k)\partial_k = XY(Z^k)\partial_k;\\
&\nabla_Y\nabla_X Z = \nabla_Y X(Z^k)\partial_k = YX(Z^k)\partial_k;\\
&\nabla_X\nabla_Y Z - \nabla_Y\nabla_X Z = (XY(Z^k) - YX(Z^k))\partial_k = \nabla_{[X,Y]}Z
\end{aligned}$$

****e penso che non si possa dire che vale per tutte le torsion free perché si passa sempre per applicare a qualcosa ???*** Then, Christoffel symbols are 0 by definition of the connection and by equation 3.*

Also, we get that

$$\nabla_{e_j} e_i = 0 \quad \forall i, j$$

where with e_i we indicate in compact notation the standard basis vector fields $\frac{\partial}{\partial e_i}$, $e_i \in \mathbb{R}^n$, that apply the standard basis vectors to each point.

This directly follows by definition of Christoffel symbols.

EXERCISE 1.2: This allows us to express in an explicit form the connection of two vector fields X, Y.

Recall that

$$X = (X^1, \dots, X^n) = \sum_i X^i e_i$$
$$Y = (Y^1, \dots, Y^n) = \sum_i Y^i e_i$$

Then,

$$\nabla_X Y = \nabla_{\sum_i X^i e_i} Y = \sum_i X^i \nabla_{e_i} Y$$

using the property of linearity over $C^\infty(M)$ in the first component of connections and that X^i are functions. Then,

$$\begin{aligned} &= \sum_i X^i \nabla_{e_i} \sum_j Y^j e_j \\ &= \sum_i X^i \Big(\sum_j Y^j \nabla_{e_i} e_j + e_i \sum_j Y^j e_j\Big) \\ &= \sum_i X^i \Big(e_i \sum_j Y^j e_j\Big) = \sum_i X^i \sum_j e_i Y^j e_j \end{aligned}$$

where we use the product rule for connections, with $f = \sum_j Y^j$. Hence, to compute $\nabla_X Y$ we can just take the ordinary derivative of each Y^j in the direction of e_i, apply the obtained function to the vector field e_j and sum, then apply each function X^i to the obtained vector field $e_i Y^j e_j$ and sum (the sum of vector fields is a vector field as the space of vector fields is a vector space).

EXERCISE 1.3: We can also use this to compute the connections in a 2-dim manifold immersed in $\mathbb{R}^3$. We can check that this is equivalent to computing the Jacobian matrix of X and apply it to Y.

This is done by computing the tangential derivative as done above and (maybe) checking that

$$\nabla^t_X Y = \pi_{TM}(\nabla_{\tilde{X}}\tilde{Y}|_M)$$

is equivalent to computing the connection as

$$\nabla_X Y = \sum_i X^i \sum_j e_i Y^j e_j$$

0.2. Introducing curvatures with Lee's reasoning. Read this section referring to Lee, Intro to Riemannian Manifolds, Chapter 7 pp.193-195.

We build a vector field Z that is parallel wrt the equator (x_1 axis) and all the meridians (every x_2 coordinate line). Then we want to check if it is parallel also along *all* the x_1 coordinate lines, i paralleli. Recall that if X is parallel along a curve γ it means $\nabla_\gamma X = 0$. Hence we have that $\nabla_{\partial_2} Z = 0$ and want to check if $\nabla_{\partial_1} Z = 0$, where ∂_i represents the vector field that associates the vector x_1 to all points, that "extends" all the curves that describe paralleli.

We get that this is true if

$$\nabla_{\partial_2}\nabla_{\partial_1} Z = 0$$

and hence if

$$\nabla_{\partial_2}\nabla_{\partial_1} Z = \nabla_{\partial_1}\nabla_{\partial_2} Z$$

This condition is always true for the Euclidean connection, because ordinary (Euclidean) second partial derivatives commute. Hence what we wanted to show is confirmed for the **standard basis vector fields under the Euclidean connection**. If we extend this to **generic vector fields** X, Y we get that the condition doesn't hold anymore, and the difference between the two terms depends on the commutator:

$$\nabla_X\nabla_Y Z - \nabla_Y\nabla_X Z = \nabla_{[X,Y]} Z \tag{41}$$

This results from how the Euclidean connection is defined ($[\nabla_X Y]_i = [XY]_i$). Hence the original thing we wanted to show doesn't always hold (only when X and Y commute locally, i.e. $[X, Y] = 0$).

Then, since we know that a space is flat when locally isometric[1] to the Euclidean space, and we have result 41 for the Euclidean space, we take this as a ***flatness criterion*** and say that a space is flat, or has curvature 0, when that holds. Hence the definition of Riemannian curvature!

[1] Two metric spaces $(X, d), (Y, d)$ are isometric when there exists a bijective isometry between them, i.e. a function $f : X \to Y$ s.t. $d(a, b) = d(f(a), f(b))\, \forall a, b \in X$.

0.3. Gradient flows & Benamou-Brenier. I was interested in this paper. To understand it I looked at

- definition of gradient flows: as those curves $x(t)$, $x : I \to \mathrm{Dom}(f)$ such that they are descent directions as defined by the sub-differential(?)
  - -¿ definition of sub-differential -¿ how is p an object living on H that is equivalent to the gradient of f in the definitions fo convexity?
- Benamou-Brenier formula, related to continuity equation: alternative definition of W_2 distance
- Fisher-Rao come definita nel paper ???