REGULARITY OF OPTIMAL TRANSPORT IN CURVED GEOMETRY: THE NONFOCAL CASE

G. LOEPER AND C. VILLANI

Abstract. We explore some geometric and analytic consequences of a curvature condition introduced by Ma, Trudinger and Wang in relation to the smoothness of optimal transport in curved geometry. We discuss a conjecture according to which a strict version of the Ma–Trudinger–Wang condition is sufficient to prove regularity of optimal transport on a Riemannian manifold. We prove this conjecture under a somewhat restrictive additional assumption of nonfocality; at the same time, we establish the striking geometric property that the tangent cut locus is the boundary of a convex set. Partial extensions are presented to the case when there is no "pure focalization" on the tangent cut locus.

Contents

1. Introduction

This paper has two sides: on the one hand, it is a work on the smoothness of optimal transport; on the other hand, it is a work on the structure of the cut locus. The latter could be discussed independently of the former, but since the initial motivation was in optimal transport theory, and since both features are intimately entangled, we shall present both problematics together. Our introduction is reduced to the minimum that the reader should know to understand the paper; but much more information can be found in the books [29, 30]; especially [30, Chapter 12] is a long and self-contained introduction to the regularity of optimal transport.

1.1. **Regularity of optimal transport: background and main result.** After Caffarelli [2, 3, 4] and Urbas [28] studied the smoothness of optimal transport maps for the quadratic cost function in $\mathbb{R}^n$, the problem naturally arose to extend these results to more general cost functions [29, Section 4.3]. In this paper, we shall only consider the important case when the cost is the squared geodesic distance on a Riemannian manifold M; this cost function, first studied by McCann [24], has many applications in Riemannian geometry [30, Part II].

There was almost no progress on the smoothness issue before the introduction of the Ma–Trudinger–Wang tensor [22]. Let M be a Riemannian manifold, which as in the rest of this paper will implicitly be assumed to be smooth, connected and complete. Let $TM = \cup(\{x\} \times T_xM)$ stand for the tangent bundle over M, and let $\mathrm{cut}(M) = \cup(\{x\} \times \mathrm{cut}(x))$ denote the cut locus of M. The Ma–Trudinger–Wang (MTW) tensor $\mathfrak{S}$ can be defined on $T(M \times M \setminus \mathrm{cut}(M))$ as follows [30, Definition 12.26]. Let $(x, y) \in M \times M \setminus \mathrm{cut}(M)$, take coordinate systems $(x_i)_{1\leq i\leq n}$, $(y_j)_{1\leq j\leq n}$ around x and y respectively; set $c(x', y') = d(x', y')^2/2$, where d is the geodesic distance on M, and note that c is C^∞ around (x, y). Write c_i (resp. $c_{,j}$) for the partial derivative with respect to x_i (resp. y_j), evaluated at (x, y); $c_{i,j}$ for the mixed second derivative with respect to x_i and y_j, etc.; and write $(c^{i,j})$ for the components of the inverse of $(c_{i,j})$, always evaluated at (x, y). Then for any $\xi \in T_xM$, $\eta \in T_yM$,

$$\mathfrak{S}(x, y) \cdot (\xi, \eta) := \frac{3}{2} \sum_{ijk\ell rs} \Big(c_{ij,r}\, c^{r,s}\, c_{s,k\ell} - c_{ij,k\ell} \Big)\, \xi^i\, \xi^j\, \eta^k\, \eta^\ell. \tag{1.1}$$

According to Loeper [20], Kim and McCann [16], this formula defines a covariant tensor. Moreover, as noted in [20], if ξ and η are orthogonal unit vectors in T_xM, then $\mathfrak{S}(x, x) \cdot (\xi, \eta)$ coincides with the sectional curvature at x along the plane generated by ξ and η [30, Particular Case 12.29].

The main assumption used in [22, 26, 27] is that

$$\mathfrak{S}(x,y) \geq K\,|\xi|^2\,|\eta|^2 \qquad \text{whenever } \sum_{ij} c_{i,j}\,\xi^i\,\eta^j = 0, \tag{1.2}$$

where K is a positive constant (strong MTW condition) or $K = 0$ (weak MTW condition). Condition (1.2) implies that the sectional curvature of M is bounded below by K. Loeper [21] showed that the round sphere $\mathbb{S}^n$ satisfies (1.2) for some $K > 0$ (see also [32]).

There is by now plenty of evidence that these conditions, complicated as they seem, are natural assumptions to develop the regularity theory of optimal transport. In particular, Loeper [20] showed how to construct counterexamples to the regularity if the weak MTW condition is not satisfied. The following precise statement is proven in [30, Theorem 12.39]; vol stands for the Riemannian volume measure.

Theorem 1.1 (Necessary condition for the regularity of optimal transport)**.** *Let M be a Riemannian manifold such that $\mathfrak{S}(x,y)\cdot(\xi,\eta) < 0$ for some x, y, ξ, η. Then there are C^∞ positive probability densities f and g on M such that the optimal transport map from $\mu(dx) = f(x)\,\mathrm{vol}\,(dx)$ to $\nu(dy) = g(y)\,\mathrm{vol}\,(dy)$, with cost function $c = d^2$, is discontinuous.*

(For the sake of presentation, this theorem is stated in [30] under a compactness assumption, but the proof goes through easily to noncompact manifolds.)

Conversely, smoothness results have been obtained under various sets of assumptions including either the weak or the strong MTW condition [10, 16, 19, 20, 22, 27]; such results are reviewed in [30, Chapter 12]. For instance, [22] furnishes interior a priori regularity estimates (say C^1) on the optimal transport map, provided that the optimal transport plan is supported in a set $D \subset M \times M$ such that (a) c is uniformly smooth (say C^4) in D; (b) all sets $(\exp_x)^{-1}(D_x)$ and $(\exp_y)^{-1}(D_y)$ are convex (in T_xM and T_yM respectively), where $D_x = \{y;\ (x,y) \in D\}$, $D_y = \{x;\ (x,y) \in D\}$, and exp stands for the Riemannian exponential. (The meaning of the notation $(\exp_x)^{-1}$ will be recalled after Definition 1.2.) But so far (a) and (b) have been proven only in particular cases such as the sphere $\mathbb{S}^n$, or its quotients like the real projective space $\mathbb{RP}^n = \mathbb{S}^n/\{\pm\mathrm{Id}\,\}$ [16, 20]. There is also a partial result by Delanoë and Ge [6] working on perturbations of the sphere and assuming certain restrictions on the size of the data.

In this paper we suggest that a (possibly slightly modified) strict form of the MTW condition *alone* is a natural sufficient condition for regularity. We shall prove this conjecture only under a simplifying nonfocality assumption which we now explain. To begin with, let us introduce some notation:

Definition 1.2 (injectivity domain, tangent cut and focal loci)**.** Let M be a Riemannian manifold and $x \in M$. For any $\xi \in T_x M$, $|\xi| = 1$, let $t_C(\xi)$ be the first time t such that $(\exp_x(s\xi))_{0 \le s \le t'}$ is not minimizing for $t' > t$; and let $t_F(\xi) \ge t_C(\xi)$ be the first time t such that $d_{t\xi} \exp_x$ (the differential of $\exp_x$ at $t\xi$) is not one-to-one. We define

$$\mathrm{I}(x) = \Big\{ t\xi;\ 0 \le t < t_C(\xi) \Big\} \qquad\qquad = \text{injectivity domain at } x;$$

$$\mathrm{TCL}(x) = \Big\{ t\xi;\ t = t_C(\xi) \Big\} \quad = \partial \mathrm{I}(x) \qquad = \text{tangent cut locus at } x;$$

$$\mathrm{TFL}(x) = \Big\{ t\xi;\ t = t_F(\xi) \Big\} \qquad\qquad = \text{(first) tangent focal locus.}$$

Let further $\overline{\mathrm{I}}(x) = \mathrm{I}(x) \cup \mathrm{TCL}(x)$.

Then we define $\mathrm{I}(M) = \cup(\{x\} \times \mathrm{I}(x))$, $\mathrm{TCL}(M) = \cup(\{x\} \times \mathrm{TCL}(x))$, $\mathrm{TFL}(M) = \cup(\{x\} \times \mathrm{TFL}(x))$, and equip these sets with the topology induced by TM.

The denomination of (tangent) injectivity domain is justified by the fact that $\exp_x$ is one-to-one $\mathrm{I}(x) \to M \setminus \mathrm{cut}(x)$. We denote its inverse by $(\exp_x)^{-1} : M \setminus \mathrm{cut}(x) \to \mathrm{I}(x)$. Explicitly, $(\exp_x)^{-1}(y)$ is the unique velocity $v \in T_x M$ such that $(\exp_x tv)_{0 \le t \le 1}$ is minimizing and $\exp_x v = y$. By extension, if $y \in \mathrm{cut}(x)$, we denote by $(\exp_x)^{-1}(y)$ the set of all velocities v satisfying the latter properties. Basic properties of the injectivity domain and tangent cut locus are reviewed in Appendix C.

Definition 1.3 (nonfocality)**.** We say that the cut locus of M is *nonfocal* (or just that M is nonfocal) if $\mathrm{TCL}(M) \cap \mathrm{TFL}(M) = \emptyset$; or equivalently if $t_F(\xi) > t_C(\xi)$ for all (x, ξ) in the unit tangent bundle of M.

In this paper, we prove the following regularity result:

Theorem 1.4 (Sufficient condition for the regularity of optimal transport)**.** *Let M be a Riemannian manifold satisfying the strong MTW condition, and whose cut locus is nonfocal. Then for any two C^∞ positive probability densities f and g on M, the optimal transport map from $\mu(dx) = f(x)\,\mathrm{vol}\,(dx)$ to $\nu(dy) = g(y)\,\mathrm{vol}\,(dy)$, with cost function $c = d^2$, is C^∞.*

Before going on, let us pause to remark the spectacular contrast between Theorem 1.1 and Theorem 1.4: depending on just the tuning of the Ma–Trudinger–Wang condition, a "generic" solution of the optimal transport problem with smooth data may be either C^∞, or not even continuous.

Now let us comment on the assumptions of Theorem 1.4. The nonfocality assumption may seem ridiculous at first sight, since it is never satisfied by compact simply connected manifolds with positive curvature, at least in even dimension. (This result is due to Klingenberg, with ancestors as old as Poincaré; Weinstein [33, Section 6] collects various sufficient conditions so that the cut locus *is* focal.) Thus our assumptions basically need nontrivial topology — something which is very uncommon in optimal transport theory. In fact, the archetype of a manifold satisfying the assumptions of Theorem 1.4 is the real projective space.

However, to advocate for Theorem 1.4, let us point out that

(a) As noted in [6], it follows from known results in Riemannian geometry that any compact manifold with nontrivial topology, satisfying a strong enough (positive) curvature pinching assumption, has nonfocal cut locus.

(b) Theorem 1.4 is a particular case of a more general result (Theorem 1.8) which covers all known (non-flat) manifolds for which there is a C^∞ regularity theory of optimal transport.

(c) Theorem 1.4 is also the first result of its kind to allow for *perturbations*: if M satisfies the assumptions of Theorem 1.4, then any C^4 perturbation of M will also satisfy them (for instance, any C^4 perturbation of $\mathbb{RP}^n$).

Remark 1.5. There has been intense activity to find examples of manifolds satisfying MTW conditions. New examples can be found in [17], but at the time of writing they are still not many. Already, showing that the sphere satisfies these conditions was not a trivial problem [20, 32].

Remark 1.6. In connection with comment (c) above, let us record the following open problem: *Is the strong MTW condition stable under perturbations of the Riemannian metric?* What makes this question nontrivial is the fact that the MTW condition is nonlocal and should hold arbitrarily close to the cut locus, even though the dependence of the distance upon the Riemannian metric may become very wild as one approaches the cut locus. At a nonfocal cut point this problem is not serious (which explains comment (c)), but at a focal point this becomes nontrivial. In [6] the Ma–Trudinger–Wang is controlled near the sphere by two derivatives of the sectional curvatures, rather than four derivatives of the metric; but the focality problem is left unsolved. Using a clever strategy, Figalli and Rifford [10] managed to answer our question positively when $M = \mathbb{S}^2$.

1.2. **Cut locus: main result.** The proof of Theorem 1.4 is based on a striking geometric property which has interest on its own, and seems to be the first of its kind:

Theorem 1.7. *Let M be a Riemannian manifold with nonfocal cut locus, satisfying the strong Ma–Trudinger–Wang condition. Then there is $\kappa > 0$ such that all injectivity domains* $\mathrm{I}(x)$ *of M are κ-uniformly convex.*

To put this result in perspective with more familiar results, recall that the strong MTW condition is a reinforcement of the condition of uniformly positive sectional curvature, which implies an upper bound on the diameter of injectivity domains (this is just an awkward way to reformulate the Bonnet–Myers theorem). To summarize, if Sect stands for sectional curvature,

$$\begin{array}{ccc} \text{strong MTW} & \Longrightarrow & \text{Sect} \geq \kappa > 0 \\ \Downarrow & & \Downarrow \\ \text{uniform convexity of } \mathrm{I}(x) & \Longrightarrow & \text{bound on diameter} \end{array}$$

Apart from this link with the Bonnet–Myers theorem, Theorem 1.7 is substantially different from all previously known results or conjectures in the field: it does not bear on the size or dimension or topological structure of the cut locus, but on its global geometric shape. It also displays a "positive effect" of positive curvature; this was somewhat unexpected, since it is usually negative curvature which has a good impact on the structure of the cut locus (by preventing focalization).

This theorem will be proven in Section 5. The key step in our proof is a kind of "continuity method" set in the injectivity domains, where the norm plays the role of "ordering parameter", and the *strict* convexity allows to keep on increasing the parameter. A more general variant of Theorem 1.7, allowing for some sort of focalization (under assumptions which in particular include the sphere), will be proven in Theorem 1.8.

1.3. **Outline of proof of Theorem 1.4.** We now explain the plan of the proof of Theorem 1.4, and the role of Theorem 1.7 therein. The proof is divided into five steps, of variable difficulty.

1. According to McCann [24], the optimal transport map between $f(x)\,\mathrm{vol}\,(dx)$ and $g(y)\,\mathrm{vol}\,(dy)$ takes the form

$$T(x) = \exp_x(\nabla\psi(x)), \tag{1.3}$$

where each geodesic $(\exp_x(t\nabla\psi(x)))_{0\leq t\leq 1}$ is minimizing, ∇ stands for gradient, and the semiconvex function ψ solves a weak form of the Monge–Ampère type equation

$$\det\Big(\nabla^2\psi(x) + \nabla^2_{xx}c\big(x, \exp\nabla\psi(x)\big)\Big) = \Big|\det\nabla^2_{xy}c\big(x, \exp\nabla\psi(x)\big)\Big|\,\frac{f(x)}{g(\exp\nabla\psi(x))}. \tag{1.4}$$

(Here $\exp \nabla\psi(x)$ is a shorthand for $\exp_x \nabla\psi(x)$, ∇^2 stands for Hessian, ∇_x^2 for the Hessian with respect to the x variable, etc.)

2. The strong MTW condition implies certain inequalities between distances (Theorem 3.1), and the uniform convexity of all injectivity domains (Theorem 1.7). The combination of both implies a property of M which we call *uniform regularity* (Theorem 4.4); it is an intrinsic and global reformulation of similar conditions introduced earlier in the regularity theory of optimal transport.

3. From the uniform regularity follows the continuity of optimal transport, and in fact the C^1 regularity of ψ (Theorem 6.1). This step is based on the strategy of Loeper [20], simplified by Kim and McCann [16, Appendices], further simplified and extended in the present work.

4. The C^1 regularity of ψ (and the assumption on M) implies that the optimal transport stays away from the cut locus (Theorem 7.1), so it takes place in a domain where c is C^∞, with uniform bounds.

5. Steps 2 and 4 make it possible to apply the local a priori estimates of Ma, Trudinger and Wang [22] in $C^{k,\beta}$ (Hölder) spaces, where $\beta \in (0,1)$, and $k \in \mathbb{N}$ is arbitrarily large. (These a priori estimates are established for a smooth cost function defined in a domain of $\mathbb{R}^n \times \mathbb{R}^n$; but by the intrinsic nature of $\mathfrak{S}$ [16], [30, Remark 12.30] they also apply to a curved geometry.) Then one may conclude, using arguments similar to those in [22], that ψ is C^∞ if f and g are. This concludes the proof.

In the sequel, we shall only treat Steps 2 to 4 of the above outline of proof, since these are the only novel steps. This, together with the proof of Theorem 1.7, will occupy Sections 3 to 7.

Then in Section 8, we shall establish the $C^{1,\alpha}$ regularity of ψ without any smoothness assumption on the probability densities, in the style of [20].

We shall also establish a more general version of Theorems 1.4 and 1.7, which has the merit to cover at the same time the case of the sphere $\mathbb{S}^n$. Let us define

$$\delta(M) = \inf_{(x,v)\in \mathrm{TCL}(M)} \mathrm{diam}\big((\exp_x)^{-1}(\exp_x v)\big). \tag{1.5}$$

Theorem 1.8. *Let M be a Riemannian manifold satisfying condition* $\mathrm{MTW}(K_0, C_0)$ *of Section 2, for some $K_0, C_0 > 0$, such that $\delta(M)$ defined in* (1.5) *is positive, and such that for any $x \in M$,* $\mathrm{TFL}(x)$ *has nonnegative second fundamental form near* $\mathrm{TCL}(x) \cap \mathrm{TFL}(x)$*. Then* *(a) there is $\kappa > 0$ such that all injectivity domains of M are κ-uniformly convex;*

(b) for any two C^∞ positive probability densities f and g on M, the optimal transport map from $\mu(dx) = f(x)\,\mathrm{vol}\,(dx)$ to $\nu(dy) = g(y)\,\mathrm{vol}\,(dy)$, with cost function $c = d^2$, is C^∞.

Here are some comments on Theorem 1.8:

• In the notation of Proposition C.1, the assumption $\delta(M) > 0$ stands between $J = \emptyset$ (no focal cut velocity) and $J \setminus \Sigma = \emptyset$ (no purely focal cut velocity).

• Proposition C.5(a) and Lemma 2.3 show that Theorem 1.8 generalizes Theorems 1.4 and 1.7; in a way this result seems to be the best that one can hope for with the techniques of this paper. However, the assumptions of Theorem 1.8, unlike those of Theorems 1.4 and 1.7, are not stable under perturbation; this is the reason why we chose not to present it as our main result.

• The nonnegativity of the second fundamental form of $\mathrm{TFL}(x)$ is to be understood in weak sense (see the reminders in Appendix A). The extra assumption put on the focal cut locus is not so bad as one may think, because the focal locus is a much less mysterious object than the cut locus.

The proof of Theorem 1.8 follows the same general lines as the proofs of Theorems 1.4 and 1.7, but details are much more tricky. We shall sketch the arguments at the end of Sections 4, 5 and 7.

There are four Appendices. The first two are devoted to various notions related to convexity. In the third one, we gather some technical results about the structure of the tangent cut locus. (Hopefully our problems will constitute a motivation to push the study of this topic.) In the fourth one we construct a counterexample showing that positive sectional curvature alone does not guarantee the convexity of injectivity domains.

Acknowledgement: This work was started during a stay of the second author in Canberra (Summer 2007), funded by a FAST research grant coordinated by Philippe Delanoë and Neil Trudinger. This was a golden opportunity for him to learn the subject of regularity of optimal transport from Neil Trudinger and Xu-Jia Wang. We are grateful to Robert McCann and Young-Heon Kim for exchanging preprints and ideas; special thanks are due to Young-Heon for spotting a hole in a preliminary version of this paper. We warmly thank Ludovic Rifford and Alessio Figalli for a careful reading and precious comments and contributions about the focal case; we gladly note that our discussions led to the genesis of [10]. Further enlightening discussions with Philippe Delanoë, Étienne Ghys and Bruno Sevennec are acknowledged. These results were presented by the second author at Duke University, as part of the Gergen Lectures, in May 2008; it is a pleasure for him to thank Jonathan Mattingly for his kind invitation, and the whole mathematics department for their wonderful hospitality. Finally, anonymous referees are thanked for their valuable comments.

2. Various forms of the Ma–Trudinger–Wang condition

Let M be a Riemannian manifold, with Riemannian metric $\langle \cdot\,,\,\cdot \rangle_x$ at x. For any $(x,y) \in (M \times M) \setminus \mathrm{cut}(M)$ we define a bilinear form on $T_xM \times T_yM$, also denoted like a scalar product by abuse of notation, as follows:

$$\forall\,(\xi,\eta) \in T_xM \times T_yM, \qquad \begin{aligned} \langle \xi, \eta \rangle &= \langle \xi, (d_v \exp_x)^{-1}\eta \rangle_x \\ &= \langle (d_w \exp_y)^{-1}\xi, \eta \rangle_y, \end{aligned} \tag{2.1}$$

where $v = (\exp_x)^{-1}(y)$, $w = (\exp_y)^{-1}(x)$. When $x = y$, $\langle \xi, \eta \rangle$ coincides with $\langle \xi, \eta \rangle_x$. The bilinear form defined by (2.1) also coincides, up to a sign, with the pseudo-Riemannian metric introduced in [22] and further studied in [16]:

$$\langle \xi, \eta \rangle = \big\langle (-\nabla^2_{x,y} c)(x,y) \cdot \xi, \eta \big\rangle = -\sum_{ij} c_{i,j}\, \xi^i\, \eta^j.$$

Whenever $K_0, C_0 \geq 0$, we introduce the following curvature conditions:

$$\mathrm{MTW}(K_0) \qquad \begin{aligned} &\forall (x,y) \in (M \times M) \setminus \mathrm{cut}(M),\ \forall (\xi,\eta) \in T_xM \times T_yM, \\ &\big[\, \langle \xi, \eta \rangle = 0 \,\big] \implies \mathfrak{S}(x,y) \cdot (\xi,\eta) \geq K_0\, |\xi|^2\, |\widetilde{\eta}|^2; \end{aligned}$$

$$\mathrm{MTW}(K_0, C_0) \qquad \begin{aligned} &\forall (x,y) \in (M \times M) \setminus \mathrm{cut}(M),\ \forall (\xi,\eta) \in T_xM \times T_yM, \\ &\mathfrak{S}(x,y) \cdot (\xi,\eta) \geq K_0\, |\xi|^2\, |\widetilde{\eta}|^2 - C_0\, \big|\langle \xi, \eta \rangle\big|\, |\xi|\, |\widetilde{\eta}|; \end{aligned}$$

where $\widetilde{\eta} = (d_v \exp_x)^{-1}\eta$, and $v = (\exp_x)^{-1}(y)$. Note that $\langle \xi, \eta \rangle = \langle \xi, \widetilde{\eta} \rangle_x \leq |\xi|\, |\widetilde{\eta}|$; and $|\eta| \leq |\widetilde{\eta}|$ by nonnegative curvature. Further note that $\mathrm{MTW}(K_0, \infty)$ coincides with $\mathrm{MTW}(K_0)$.

Remark 2.1. Condition $\mathrm{MTW}(K_0, C_0)$ is stronger than $\mathrm{MTW}(K_0)$ in that it also controls from below (by a possibly negative number) the behavior of $\mathfrak{S}$ in the direction $\xi = \eta$. The identity $\mathfrak{S}(x,x) \cdot (\xi,\xi) = 0$ implies $C_0 \geq K_0$.

Remark 2.2. Condition $\mathrm{MTW}(K_0)$ is stronger than the original condition (1.2) introduced by Ma, Trudinger and Wang in [22], since $|\widetilde{\eta}| \geq |\eta|$. The replacement of η by $\widetilde{\eta}$ was suggested by a discussion with Figalli; it is somewhat natural because with this convention $\mathrm{MTW}(K_0, K_0)$ implies the nonnegativity of $\mathfrak{S}$ over all directions. Another option would have been to keep η (instead of $\widetilde{\eta}$) in both $\mathrm{MTW}(K_0)$ and $\mathrm{MTW}(K_0, C_0)$; for the purposes of this paper this would have worked just the same, because we mainly care for nonfocal situations. There is also a third possible condition:

$$\mathfrak{S}(x,y)\cdot(\xi,\eta) \geq K_0\,|\xi|^2\,|\eta|\,|\widetilde{\eta}| - C_0\,\big|\langle\xi,\eta\rangle\big|\,|\xi|\,|\eta|. \tag{2.2}$$

It was checked numerically by Figalli and Rifford [10] that the sphere $\mathbb{S}^n$ satisfies the optimal condition $\mathrm{MTW}(1,1)$, which implies (2.2), which in turns implies the modified $\mathrm{MTW}(1,1)$ with η in place of $\widetilde{\eta}$. Figalli and Rifford also prove that $\mathbb{S}^n$ satisfies $\mathrm{MTW}(K,K)$ for some $K \in (0,1)$. These estimates obviously imply the nonnegativity of the tensor $\mathfrak{S}$, which was noticed by Kim and McCann [16, Example 3.7].

In the sequel we shall assume $\mathrm{MTW}(K_0, C_0)$ for some $K_0 > 0$, $C_0 \in [K_0, \infty)$. This seems stronger than the assumption $\mathrm{MTW}(K_0)$ considered in [20, 22], but in fact these conditions are equivalent as long as y stays away from the focal locus of x, as the following result shows.

Lemma 2.3. *If a compact Riemannian manifold M has nonfocal cut locus, then*
(a) Condition (1.2) *with $K > 0$ implies* $\mathrm{MTW}(K_0)$ *for some $K_0 > 0$;*
(b) Condition $\mathrm{MTW}(K_0)$ *implies* $\mathrm{MTW}(K_0, C_0)$ *for some $C_0 \in [K_0, +\infty)$.*

Proof of Lemma 2.3. First, the nonfocality implies that

$$|\eta| \leq |\widetilde{\eta}| \leq C\,|\eta|,$$

for some finite constant $C = \sup_{(x,v)\in \mathrm{I}(M)} \|(d_v \exp_x)^{-1}\|$. So (1.2) does imply $\mathrm{MTW}(K_0)$ with $K_0 = K/C^2$.

By Proposition C.5(a), there is a bound on all derivatives of $c = d^2/2$ up to order 4, uniformly over $(M \times M) \setminus \mathrm{cut}(M)$; and there are upper bounds on the operator norms $\|d_v \exp_x\|$ and $\|(d_v \exp_x)^{-1}\|$, uniformly over $(x,v) \in \mathrm{I}(M)$.

Let U_xM stand for the space of unit tangent vectors at x. Whenever x and y are given in $M \times M \setminus \mathrm{cut}(M)$, choose a system of geodesic coordinates centered at x, and let $\mathcal{Q}$ be the space of all quadrilinear forms Q on $T_xM \times T_xM$, taking the form $Q(\xi,\eta) = \sum q_{ijk\ell}\,\xi^i\,\xi^j\,\eta^k\,\eta^\ell$, such that all coefficients $q_{ijk\ell}$ are bounded by 1 and $Q(\xi,\eta) \geq K_0$ for all $(\xi,\eta) \in U_xM \times U_xM$ satisfying $\langle\xi,\eta\rangle = 0$ (Note carefully: the

η appearing in the present argument is the $\widetilde{\eta}$ used before.) The condition $\langle \xi, \eta \rangle = 0$ defines the vanishing of the first coordinate of η in a certain orthogonal basis. Since functions in $\mathcal{Q}$ are uniformly Lipschitz in η, uniformly in ξ, there is a constant C_1, only depending on n, such that

$$\forall Q \in \mathcal{Q},\ \forall (x,y)\ \forall (\xi,\eta) \in U_xM \times U_xM \qquad Q(\xi,\eta) \geq K_0 - C_1\,|\langle \xi, \eta \rangle|. \tag{2.3}$$

Now if $\mathfrak{S}$ satisfies MTW(K_0), then $\mathfrak{S}/C_2$, viewed as a function on $T_xM \times T_xM$, belongs to $\mathcal{Q}$, where $C_2 = \max(1, \|c\|^3_{C^4})$ (the C^4 norm is over the whole of $M \times M \setminus \mathrm{cut}(M)$). Then by homogeneity, (2.3) implies

$$\mathfrak{S}(x,y)\cdot(\xi,\widetilde{\eta}) \geq K_0\, C_2\, |\xi|^2\, |\widetilde{\eta}|^2 - C_1 C_2\, |\langle \xi, \widetilde{\eta}\rangle|\, |\xi|\, |\widetilde{\eta}|$$

for all $(x,y) \in M \times M \setminus \mathrm{cut}(M)$ and $(\xi,\eta) \in T_xM \times T_yM$, $\widetilde{\eta} = (d_v \exp_x)^{-1}(\eta)$. $\square$

3. Metric consequences of the Ma–Trudinger–Wang condition

For any x in M we may introduce geodesic coordinates centered at x, or equivalently parameterize $y \in M \setminus \mathrm{cut}(x)$ by $p = (\exp_x)^{-1}(y)$. This operation transforms a tangent vector $\eta \in T_yM$ into $\widetilde{\eta} = (d_p \exp_x)^{-1} \cdot \eta \in T_xM$. Then for any two tangent vectors $\xi \in T_xM$, $\eta \in T_yM$,

$$\mathfrak{S}(x,y)\cdot(\xi,\eta) = -\frac{3}{2}\,\frac{\partial^2}{\partial p_\eta^2}\frac{\partial^2}{\partial x_\xi^2}\, c(x,y). \tag{3.1}$$

The meaning of (3.1) is as follows: first freeze y and differentiate $c(x,y) = d(x,y)^2/2$ twice with respect to x, in direction ξ; this gives a function of y. Then freeze x, parameterize y by geodesic coordinates p, and differentiate twice with respect to p, in direction $\widetilde{\eta}$. See [30, Chapter 12] for more information.

Trudinger and Wang [26] and Loeper [20] first explored the relations between lower bounds on $\mathfrak{S}$, and certain inequalities involving distance functions along c-segments; then Kim and McCann [15] [16, Appendices] introduced a more geometric and flexible method to study the same inequalities. By definition, a c-segment $(y_t)_{0\leq t\leq 1}$ with base x is the image by $\exp_x$ of a plain line segment included in $\mathrm{I}(x)$; in other words, it is a curve which appears as a straight line when written in geodesic coordinates centered at x. A c-segment is uniquely determined by its endpoints y_0, y_1 and the basepoint x, and will be denoted by $[y_0, y_1]_x$. Of course, if y_0 and y_1 are given in $\mathrm{I}(x)$, the c-segment $[y_0, y_1]_x$ does not necessarily exist, since there is no a priori reason why $[(\exp_x)^{-1}(y_0), (\exp_x)^{-1}(y_1)] \subset T_xM$ would be contained in $\mathrm{I}(x)$.

The following statement can be found in [30, Proof of Theorem 12.34], where it is proven with a slight variant of the Kim–McCann technique: *Assume the weak MTW* *condition. Let $\overline{x}, y_0, y_1 \in M$ be such that $(y_t)_{0\le t\le 1} = [y_0, y_1]_{\overline{x}}$ is well-defined, and that $[\overline{x}, x]_{y_t}$ is well-defined for any $t \in (0,1)$. Then*

(3.2)
$$\forall t \in (0,1) \qquad d(x,y_t)^2 - d(\overline{x},y_t)^2 \ge \min\Big(d(x,y_0)^2 - d(\overline{x},y_0)^2,\ d(x,y_1)^2 - d(\overline{x},y_1)^2\Big).$$

When the *strict* conditions $\mathrm{MTW}(K_0)$ or $\mathrm{MTW}(K_0, C_0)$ are enforced, the inequality (3.2) can be refined in two ways, as soon as x goes slightly away from $\overline{x}$:

(a) one may deform the c-segment $[y_0, y_1]_{\overline{x}}$ into a path with small acceleration (in geodesic coordinates centered at $\overline{x}$);

(b) away from $t = 0$ and $t = 1$, the inequality becomes strict; this property will use the finiteness of the constant C_0.

Here is a precise statement, where the dot as usual stands for the time-derivative (by convention, time means the variable parameterizing paths), and $z_+ = \max(z, 0)$.

Theorem 3.1. *Let M be a Riemannian manifold satisfying $\mathrm{MTW}(K_0, C_0)$ for some $K_0 \ge 0$, $C_0 \in [\max(K_0, 1), \infty]$. Let $\overline{x} \in M$ and let $(p_t)_{0\le t\le 1}$ be a C^2 curve drawn in $\mathrm{I}(\overline{x})$. For any $t \in (0,1)$, define $y_t = \exp_{\overline{x}} p_t$. Let $x \in \bigcap_{0\le t\le 1} M \setminus \mathrm{cut}(y_t)$; for each t define $\overline{q}_t = (\exp_{y_t})^{-1}(\overline{x})$, $q_t = (\exp_{y_t})^{-1}(x)$. Assume that*

(3.3)
$$\forall t \in (0,1), \qquad \begin{cases} [\overline{q}_t, q_t] \subset \mathrm{I}(y_t) \\ \\ -\big\langle \ddot{p}_t,\ (d_{p_t}\exp_{\overline{x}})^{-1}(q_t - \overline{q}_t)\big\rangle \le \varepsilon_0 \left|(d_{p_t}\exp_{\overline{x}})^{-1}(q_t - \overline{q}_t)\right|^2 |\dot{p}_t|^2, \end{cases}$$

for some $\varepsilon_0 \le K_0/8$; then inequality (3.2) still holds, and can be reinforced into

$$d(x,y_t)^2 - d(\overline{x},y_t)^2 \ge \min\Big(d(x,y_0)^2 - d(\overline{x},y_0)^2,\ d(x,y_1)^2 - d(\overline{x},y_1)^2\Big) + 2\,\lambda\, t(1-t)\left(\inf_{0\le s\le 1} |q_s - \overline{q}_s|^2\right) |p_1 - p_0|^2, \tag{3.4}$$

for any $t \in (0,1)$, where

(3.5)
$$\lambda = \frac{K_0\,\big(1 - 2\sigma\,\eta_0\ \mathrm{diam}(M)\big)_+^2}{12\,C_0\big(1 + \sigma\ \mathrm{diam}(M)\big)}, \qquad \sigma := \max|\dot{p}|, \quad \eta_0 \ge \sup_{0\le t\le 1}\left(\frac{|\ddot{p}_t|}{|q_t - \overline{q}_t|\,|\dot{p}_t|^2}\right).$$

Remark 3.2. This result extends and simplifies the results of Kim and McCann [16, Proposition B.1].

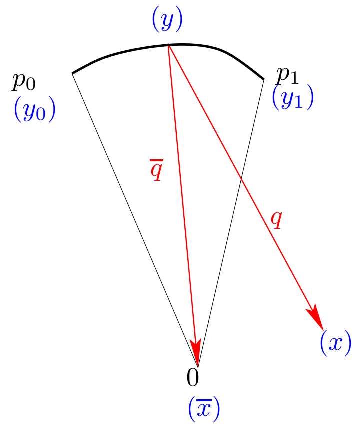

FIGURE 1. Setting of Theorem 3.1. The path joining p_0 to p_1 is drawn in $T_{\overline{x}}M$, while $\overline{q}$, q belong to T_yM. Points are displayed inside curved brackets near the corresponding tangent vectors.

Remark 3.3. Since $\mathrm{MTW}(K_0, C_0)$ implies nonnegative sectional curvature, all maps $\exp_{y_t}$ are 1-Lipschitz, so one could replace $|(d_{p_t}\exp_{\overline{x}})^{-1}(q_t - \overline{q}_t)|$ and $|q_t - \overline{q}_t|$ in Theorem 3.1 by the time-independent smaller quantity $d(\overline{x}, x)$. However, to study the implications of the MTW condition on the cut locus it will be crucial to keep $|q_t - \overline{q}_t|$ in the conclusion and not replace it by $d(\overline{x}, x)$.

Proof of Theorem 3.1. In the sequel we shall assume $K_0 > 0$; the case $K_0 = 0$ is treated similarly, and anyway is already considered in [30, Theorem 12.34]. The core of the method goes back to Kim and McCann [15].

Let $c = d^2/2$. If the assumptions of Theorem 3.1 are satisfied, we have $\nabla_x c(\overline{x}, y_t) + p_t = 0$, so $\nabla^2_{x,y} c(\overline{x}, y) \cdot \dot{y} + \dot{p} = 0$. Choose an orthonormal basis $(e_i)_{1\leq i\leq n}$ of $T_{\overline{x}}M$ and let $\zeta_i = \dot{p}_i$ $(1 \leq i \leq n)$ stand for the coordinates of the velocity of the path (p_t) in this basis. By direct calculation, using the convention of summation over repeated indices, we have

$$(\dot{y})^i = -c^{i,r}(\overline{x}, y_t)\,\zeta_r =: \zeta^i, \tag{3.6}$$

$$(\ddot{y})^i = -c^{i,k}\,c_{k,\ell j}\,c^{\ell,r}\,c^{j,s}\,\zeta_r\,\zeta_s - c^{i,r}\dot{\zeta}_r, \tag{3.7}$$

where all functions are evaluated at $(\overline{x}, y_t)$ and at time t.

Let then

$$h(t) = -c(x, y_t) + c(\overline{x}, y_t) + \delta\, t(1-t),$$

where δ will be chosen later on. For any $t \in (0,1)$ we have

$$\begin{aligned}\dot{h}(t) &= \big[-c_{,j}(x,y_t) + c_{,j}(\overline{x},y_t)\big]\,\zeta^j + \delta\,(1-2t)\\ &= -c_{i,j}(\overline{x},y_t)\,\eta^i\,\zeta^j + \delta\,(1-2t)\end{aligned} \tag{3.8}$$

where $\eta_j = -c_{,j}(x,y_t) + c_{,j}(\overline{x},y_t)$ and $\eta^i = -c^{j,i}(\overline{x},y_t)\,\eta_j$ (equivalently, $\eta = q_t - \overline{q}_t$). Next (using (3.6) and (3.7) and a bit of juggling with indices),

$$\ddot{h}(t) = -\Big(\big[c_{,ij}(x,y_t) - c_{,ij}(\overline{x},y_t)\big] - \eta^k\,c_{k,ij}(\overline{x},y_t)\Big)\,\zeta^i\,\zeta^j \,+ c_{i,j}\,\eta^i\,c^{j,r}\,\dot{\zeta}_r - 2\delta, \tag{3.9}$$

where again the summation over repeated indices is implicit, and expressions are evaluated at $(\overline{x},y_t)$ by default.

Now freeze t, y_t and ζ, and let $\Phi(z) = c_{,ij}(z,y_t)\,\zeta^i\,\zeta^j$. This can be seen as a function of $(\exp_{y_t})^{-1}(z) \in T_{y_t}M$, and then its second derivative in direction η is given by $-(2/3)\mathfrak{S}(z,y_t)\cdot(\zeta,\eta)$. Starting from (3.9), a Taylor formula for Φ on the segment $[\overline{q}_t, q_t]$ yields

$$\begin{aligned}\ddot{h}(t) &= \frac{2}{3}\int_0^1 \mathfrak{S}\big(y_t, [\overline{x},x]_{y_t}(s)\big)\cdot(\eta,\zeta)\,(1-s)\,ds + c_{i,j}(\overline{x},y_t)\,\eta^i\,c^{j,r}(\overline{x},y_t)\,\dot{\zeta}_r\, - 2\delta\\ &= \frac{2}{3}\int_0^1 \mathfrak{S}\big(y_t, \exp_{y_t}((1-s)\overline{q}_t + sq_t)\big)\cdot(\eta,\zeta)\,(1-s)\,ds + \eta^r\dot{\zeta}_r\, - 2\delta.\end{aligned}$$

(Note that η and ζ depend on t but not on s.)

Let $t \in [0,1]$ be a maximum of h. If t belongs to $(0,1)$, then $\dot{h}(t) = 0$ and $\ddot{h}(t) \le 0$. In particular, from (3.8),

$$\big|\langle \xi,\eta\rangle\big| = \big|-c_{i,j}\,\eta^i\,\zeta^j\big| = \delta\,|1-2t| \le \delta.$$

Then, by MTW(K_0,C_0) and $\int(1-s)\,ds = 1/2$,

$$\begin{aligned}\ddot{h}(t) &\ge \frac{K_0}{3}\,|\widetilde{\eta}|^2\,|\zeta|^2 - \frac{C_0}{3}\,|\langle\zeta,\eta\rangle||\zeta|\,|\widetilde{\eta}| + \eta^r\dot{\zeta}_r - 2\delta\\ &\ge \frac{K_0}{3}\,|\widetilde{\eta}|^2\,|\zeta|^2 - \delta\left(2 + \frac{C_0}{3}\,|\widetilde{\eta}|\,|\zeta|\right) + \langle\ddot{p}_t, \widetilde{\eta}\rangle,\end{aligned} \tag{3.10}$$

where $\eta = q_t - \overline{q}_t$ and (as in the previous section) $\widetilde{\eta} = (d_{p_t}\exp_{\overline{x}})^{-1}(q_t - \overline{q}_t)$.

If $\eta \ne 0$ and

$$-\langle\ddot{p}_t,\widetilde{\eta}\rangle \le \frac{K_0}{8}\,|\widetilde{\eta}|^2\,|\zeta|^2, \qquad \delta \le \frac{K_0}{2}\left(\frac{|\widetilde{\eta}|^2\,|\zeta|^2}{6 + C_0\,|\zeta|\,|\widetilde{\eta}|}\right), \tag{3.11}$$

then the right-hand side of (3.10) is positive, in contradiction with $\ddot{h}(t) \le 0$. (This argument also works when $C_0 = \infty$ and $\delta = 0$.)

The assumption (3.3) implies the first inequality in (3.11) by means of Cauchy–Schwarz inequality. Recalling that $C_0 \geq 1$, and noting that $x \to x^2/(1+x)$ is nondecreasing on $\mathbb{R}_+$ and $|\widetilde{\eta}| \geq |\eta|$ (by nonnegative curvature), we see that the second inequality in (3.11) is satisfied as soon as

$$\delta \leq \frac{K_0}{12\,C_0}\left(\frac{|\eta|^2\,|\zeta|^2}{1+|\eta|\,|\zeta|}\right). \tag{3.12}$$

We claim that (3.12) is true if $\delta \leq \lambda\,|q_t - \overline{q}_t|^2\,|p_1 - p_0|^2$, and (3.5) holds. Indeed, first note that

$$\begin{aligned}\int_0^1 |\zeta_s|\,ds &\leq |\zeta_t| + \int_0^1 |\dot{\zeta}_s|\,ds \\ &\leq |\dot{p}_t| + \int_0^1 \eta_0\,|q_s - \overline{q}_s|\,|\zeta_s|^2\,ds \\ &\leq |\dot{p}_t| + (2\eta_0\ \mathrm{diam}(M))\,(\sup|\zeta_s|)\int_0^1 |\zeta_s|\,ds,\end{aligned}$$

so $\int_0^1 |\zeta_s|\,ds \leq (1 - 2\eta_0\,\sigma\ \mathrm{diam}(M))_+^{-1}\,|\dot{p}_t|$. In particular,

$$|p_1 - p_0| \leq \int_0^1 |\zeta_s|\,ds \leq (1 - 2\eta_0\,\sigma\ \mathrm{diam}(M))_+^{-1}\,|\dot{p}_t|.$$

So

$$\begin{aligned}\delta &\leq \frac{K_0\,(1 - 2\eta_0\,\sigma\ \mathrm{diam}(M))_+^2}{12\,C_0\,(1+\sigma\ \mathrm{diam}(M))}\,|\eta|^2\,|p_1 - p_0|^2 \\ &\leq \frac{K_0\,|\zeta|^2\,|\eta|^2}{12\,C_0\,(1+|\eta|\,|\zeta|)},\end{aligned}$$

which is the desired (3.12).

In the end we are left with only three possibilities: either $\eta = 0$ (i.e. $x = \overline{x}$), or $t = 0$, or $t = 1$. In either case, we conclude that $h(t) \leq \max(h(0), h(1))$, which is equivalent to (3.4) (recall that $c = d^2/2$). □

4. Uniform regularity

The Ma–Trudinger–Wang condition is a differential version of the "regularity property" discussed in [30, Chapter 12], which underlies the regularity theory of optimal transport. If we want to adapt that theory to the setting of Riemannian manifolds, using Theorem 3.1, we immediately stumble upon the too severe requirement that x should not belong to $\mathrm{cut}(y_t)$ in (3.3). This cut locus issue is one of the main new problems appearing in the context of Riemannian manifolds.

This leads us to propose the following definition:

Definition 4.1 (Uniform regularity)**.** A Riemannian manifold M is said to be uniformly regular if there are $\varepsilon_0, \kappa, \lambda > 0$ such that

(a) $\mathrm{I}(x)$ is κ-uniformly convex for all x;

(b) For any $\overline{x} \in M$, let $(p_t)_{0 \le t \le 1}$ be a C^2 curve drawn in $\mathrm{I}(\overline{x})$, and let $y_t = \exp_{\overline{x}} p_t$; let further $x \in M$. If

$$\forall t \in (0,1), \qquad |\ddot{p}_t| \le \varepsilon_0 \, d(\overline{x}, x) \, |\dot{p}_t|^2, \qquad |\dot{p}_t| \le 2 \, \mathrm{diam}(M), \tag{4.1}$$

then for any $t \in (0,1)$,

$$d(x, y_t)^2 - d(\overline{x}, y_t)^2 \ge \min\Big(d(x, y_0)^2 - d(\overline{x}, y_0)^2, \; d(x, y_1)^2 - d(\overline{x}, y_1)^2 \Big) + 2 \, \lambda \, t(1-t) \, d(\overline{x}, x)^2 \, |p_1 - p_0|^2. \tag{4.2}$$

Remark 4.2. Reminders about the notion of uniform convexity are given in Appendix A. A κ-uniformly convex set in $\mathbb{R}^n$ has diameter bounded above by C/κ for some universal constant C; so uniform regularity automatically implies a uniform bound on the diameter of injectivity loci, and therefore on $\mathrm{diam}(M)$.

Example 4.3. $\mathbb{S}^n$ is uniformly regular.

Proof. The injectivity domains of $\mathbb{S}^n$ are spheres of radius π, so they are uniformly convex. In particular, if $x \in \bigcap_t (\mathbb{S}^n \setminus \mathrm{cut}(y_t))$ then $[\overline{q}_t, q_t] \subset \mathrm{I}(y_t)$, where the notation is the same as in Theorem 3.1. If moreover (p_t) satisfies (4.1), then

$$|\ddot{p}_t| \le \varepsilon_0 \, d(\overline{x}, x) \, |\dot{p}_t|^2 \; \le \varepsilon_0 \, |q_t - \overline{q}_t| \, |\dot{p}_t|^2 \le \varepsilon_0 \, |(d_{p_t} \exp_{\overline{x}})^{-1} (q_t - \overline{q}_t)| \, |\dot{p}_t|^2, \tag{4.3}$$

where both inequalities follow from the nonnegative curvature.

By [10], $\mathbb{S}^n$ satisfies $\mathrm{MTW}(K,K)$ for some $K > 0$; so all the assumptions of Theorem 3.1 are satisfied if $\varepsilon_0 \le K/8$. Moreover (4.1) also implies $|\ddot{p}_t| \le \varepsilon_0 \, |q_t - \overline{q}_t| \, |\dot{p}_t|^2$, so by choosing ε_0 small enough we can ensure that the constant λ in (3.5) (with $\eta_0 = \varepsilon_0$) will be positive. Then (4.2) follows by (3.4) and again $d(\overline{x}, x) \le |q_t - \overline{q}_t|$.

It remains to relax the assumption $x \in \bigcap_t (\mathbb{S}^n \setminus \mathrm{cut}(y_t))$. This is done by density as in [16]: it suffices to note that $\bigcap_t (\mathbb{S}^n \setminus \mathrm{cut}(y_t))$ coincides with $\mathbb{S}^n \setminus \{-y_t, \; 0 \le t \le 1\}$, which is dense in $\mathbb{S}^n$. $\square$

The density argument used in the above proof is very particular: for more general manifolds we do not expect $\bigcap(M \setminus \mathrm{cut}(y_t))$ to be dense; this is false even for a nonspherical ellipsoid! To convince the reader of the interest of Definition 4.1 we'd better find more "generic" examples. The following result allows to do so:

Theorem 4.4. *Let M be a Riemannian manifold with nonfocal cut locus, satisfying $\mathrm{MTW}(K,C)$ for some $K > 0$, such that all its injectivity domains are κ-uniformly convex, for some $\kappa > 0$. Then M is uniformly regular in the sense of Definition 4.1.*

The technical core of the proof of Theorem 4.4 is the following auxiliary result:

Lemma 4.5. *Let M be a Riemannian manifold with nonfocal cut locus. Let $\overline{x}, x \in M$, let $(p_t)_{0\le t\le 1}$ be a C^2 path drawn in $\mathrm{I}(\overline{x})$ and let $y_t = \exp_{\overline{x}} p_t$. Then for any $\eta > 0$ there is a path $(\widehat{p}_t)_{0\le t\le 1}$ such that $\|p_\cdot - \widehat{p}_\cdot\|_{C^2(0,1)} \le \eta$ and, if $\widehat{y}_t = \exp_{\overline{x}} \widehat{p}_t$, then $(\exp_x)^{-1}(\widehat{y}_t)$ is a piecewise C^2 path which always leaves and reenters $\mathrm{I}(x)$ transversally at discontinuity times.*

Proof of Lemma 4.5. First, since (p_t) stays away from the cut locus of $\overline{x}$, there is a neighborhood of the trajectory (y_t) on which the exponential map $\exp_{\overline{x}}$ is smoothly invertible; then it is equivalent to perturb the path (y_t) in C^2 topology, or to perturb the path (p_t) in C^2 topology.

Next, $\mathrm{cut}(x)$ is of dimension $n-1$, so its complement is dense. By perturbing (p_t) in C^2 topology, we can make sure that $y_0, y_1 \notin \mathrm{cut}(x)$.

Next, by nonfocality and inverse function theorem, for any $y \in M$ there are a neighborhood U of y, finitely many distinct velocities $v_1, \ldots, v_m$ in $\overline{\mathrm{I}}(M)$, and disjoint neighborhoods W_j of v_j, such that $(\exp_x)^{-1}(U) \subset W_1 \cup \ldots \cup W_m$ and the exponential map is smoothly invertible $U \to W_j$. We can cover M by a finite number of such neighborhoods U, extract a finite covering $\mathcal{U}$, and divide $[0,1]$ into subintervals $[t_j, t_{j+1}]$ such that the restriction of (y_t) to each $[t_j, t_{j+1}]$ is valued in some $U \in \mathcal{U}$. By a density argument again, we can ensure that $y_{t_j} \notin \mathrm{cut}(x)$. All in all, it suffices to prove the statement of the lemma when the path (y_t) is represented by finitely many smooth paths $p_t^{(1)}, \ldots, p_t^{(m)}$ with $\exp_{\overline{x}} p_t^{(j)} = y_t$, each of them defined in a neighborhood of some element of $\overline{\mathrm{I}}(x)$.

By Proposition C.5(c) in Appendix C, the set $\mathcal{N}$ of nondifferentiability points of $\mathrm{TCL}(x)$ is included in a finite union of smooth manifolds of dimension $n-2$, so the same is true of $\exp_x(\mathcal{N})$. By slightly perturbing the path (y_t), we can make sure that (y_t) stays a positive distance away from $\exp_x(\mathcal{N})$. On the other hand, if $\mathcal{D}$ stands for the set of differentiability points in $\mathrm{TCL}(x)$, any point in $\exp_x(\mathcal{D})$ has exactly two pre-images in $\overline{\mathrm{I}}(x)$; then the construction above shows that we just have to work with $m = 1$ (in which case the path (y_t) stays nicely in $M \setminus \mathrm{cut}(x)$) or $m = 2$, which we shall now assume.

So the problem is as follows: we have two paths $(p_t^{(1)})_{0\le t\le 1}$ and $(p_t^{(2)})_{0\le t\le 1}$, drawn in a neighborhood of some velocities v_1 and v_2 in $\overline{\mathrm{I}}(x)$, such that $\exp_x p_t^{(1)} = \exp_x p_t^{(2)}$. When $|p_t^{(1)}| < |p_t^{(2)}|$ we have $(\exp_x)^{-1}(y_t) = p_t^{(1)}$, and vice versa. Perturbing either of these paths in C^2 topology leads to a C^2 perturbation of (y_t).

As in the proof of Proposition C.5(c), we have a smooth hypersurface H in the neighborhood of v_1, delimiting two open "curved half-spaces" H_1 and H_2, such that $|p_t^{(1)}| < |p_t^{(2)}|$ if and only if $p_t^{(1)}$ belongs to H_1. The problem is to show that we can perturb $(p_t^{(1)})$ in C^2 topology, in such a way that it always crosses H transversally (that is, with a velocity which is not tangent to H) and finitely many times. This is done by an elementary argument after straightening up H by a smooth diffeomorphism. □

Proof of Theorem 4.4. Let $\overline{x}$, x, $(p_t)_{0\le t\le 1}$ and $(y_t)_{0\le t\le 1}$ be as in Definition 4.1. By density and Lemma 4.5, we may assume that $y_0, y_1 \notin \mathrm{cut}(x)$ and that $(\exp_x)^{-1}(y_t)$ always leaves and reenters $\mathrm{I}(x)$ transversally, all the other conditions in Definition 4.1 being unchanged (apart from a slight increase in ε_0, but we can slightly reduce it from the beginning). We may also assume $x \neq \overline{x}$, otherwise everything is trivial.

So the picture is as follows: there are finitely many times $t_0 = 0, t_1, \ldots, t_N = 1$ such that $y_t \in \mathrm{cut}(x)$ only for $t \in \{t_1, \ldots, t_{N-1}\}$, and $(\exp_x)^{-1}(y_t)$ enters $\mathrm{I}(x)$ transversally at t_j^+, leaves it transversally again at t_{j+1}^-.

Now we repeat the proof of Theorem 3.1 by studying the function $h(t) = -c(x, y_t) + c(\overline{x}, y_t) + \delta\, t(1-t)$, where $c = d^2/2$ and $\delta > 0$.

On each time-interval (t_i, t_{i+1}) we have $y_t \notin \mathrm{cut}(x)$, so $q_t = (\exp_{y_t})^{-1}(x)$ is well-defined, h is a smooth function of t, and by convexity of $\mathrm{I}(y_t)$ we have $[\overline{q}_t, q_t] \subset \mathrm{I}(y_t)$, where $\overline{q}_t = (\exp_{y_t})^{-1}(\overline{x})$. The second assumption in (3.3) is also satisfied by assumption (as in (4.3)). So we may indeed repeat the proof of Theorem 3.1, and conclude that if δ is defined as in that proof, the function h cannot have any maximum on (t_i, t_{i+1}). Since h is continuous on $[0, 1]$, it achieves its maximum at one of the times t_j.

Let us consider the behavior of h near t_j, for $j \in \{1, \ldots, N-1\}$. As in the proof of Proposition C.5(c), the function $c(x, y_t)$ can be written as a function of $p_t^{(1)}$ near t_j, where $\exp_x p_t^{(1)} = y_t$ and $p_t^{(1)}$ varies smoothly; moreover, $c(x, y_t) = \inf(E_1, E_2)$, where E_1 and E_2 are smooth functions, $\nabla E_1 \neq \nabla E_2$ and the equation of $\mathrm{TCL}(x)$ is given (locally) by $E_1 = E_2$. It follows that

$$\frac{d}{dt}\bigg|_{t=t_j^+} c(x, y_t) < \frac{d}{dt}\bigg|_{t=t_j^-} c(x, y_t);$$

so the graph of $t \to c(x, y_t)$ presents an upper spike at t_j. Since $c(\overline{x}, y_t)$ is a smooth function of t, the graph of h presents a downward spike at t_j, so t_j cannot be a maximum of h. The only possibility left out for h is to achieve its maximum at $t = 0$ or $t = 1$. Then we can write down the conclusion of Theorem 3.1, and by using the inequality $d(\overline{x}, x) \leq |q_t - \overline{q}_t|$ we obtain (4.2). □

To conclude this section, we note that Theorem 4.4 still holds true if the nonfocality condition is replaced by the more general assumptions of Theorem 1.4. Since the approximation Lemma 4.5 is not available any longer, to prove this we can invoke an improvement which is obtained from a more elaborate "probabilistic" argument developed by Figalli and the second author in the short paper [11]:

Lemma 4.6. *Let* $\overline{x}, x \in M$, *and let* $(p_t)_{0\leq t\leq 1}$ *be a smooth path not intersecting* $\mathrm{TFL}(\overline{x})$. *Then for any* $\varepsilon > 0$ *there is a path* $(\widetilde{p}_t)_{0\leq t\leq 1}$ *such that*

$$\|p_\cdot - \widetilde{p}_\cdot\|_{C^2(0,1)} \leq \varepsilon$$

and

$$\Big\{t;\ \exp_{\overline{x}} \widetilde{p}_t \in \mathrm{cut}(x)\Big\} \qquad \textit{is finite.}$$

We refer to [11] for the proof, which is based on the estimates in [13, 18].

5. Convexity of injectivity domains

The main goal of this section is the proof of Theorem 1.7. Before going into the proof, let us give some explanations. First, here is a heuristic reasoning suggesting a relation between the MTW condition and the convexity of $\mathrm{I}(x)$. Write $c = d^2/2$ and replace for a while condition $\mathrm{MTW}(K_0, C_0)$ by $\mathrm{MTW}(K_0, K_0)$, so that $\mathfrak{S} \geq 0$ on the whole of its domain of definition; equivalently, $\nabla_x^2 c(x, y)$ is a ("matrix"-valued) *concave* function of $p = (\exp_x)^{-1}(y)$. It is shown in [5, Proposition 2.5] that $\mathrm{cut}(x)$ is exactly the set of points y such that $c(\,\cdot\,, y)$ fails to be semiconvex at x; which means that formally $\nabla_x^2 c(x, y)$ admits a $-\infty$ eigenvalue. (Recall that all eigenvalues are bounded above since the distance is semiconcave.) Now if $p_0, p_1 \in \mathrm{I}(x)$, let $y_i = \exp_x(p_i)$ $(i = 0, 1)$, and $y_t = \exp_x((1-t)p_0 + tp_1)$, then $\nabla_x^2 c(x, y_0)$ and $\nabla_x^2 c(x, y_1)$ have only finite eigenvalues, and $\nabla_x^2 c(x, y_t)$ is a concave function of t, so $\nabla_x^2 c(x, y_t)$ remains bounded below for all t — so we expect y_t to be outside the cut locus of x.

However convincing this argument may look, it has two major shortcomings:

(a) In general the matrix $\nabla_x^2 c(x,y)$ may very well be uniformly bounded (from above and below) on its domain of definition. (Think of $M = \mathbb{S}^1$ where this Hessian is always equal to 1, except when $|x-y| = \pi$, and then it is not defined.) Making sense of the fact that $\nabla_x^2 c$ has $-\infty$ eigenvalues might require a distribution-type argument, which is not obvious at all to implement.

(b) Going from $\mathrm{MTW}(K_0, C_0)$ to $\mathrm{MTW}(K_0, K_0)$ we have added one direction of convexity, which in general is not in the original problem, and helps quite a bit in the heuristics. (Note: the $-\infty$ eigenvectors at a focal uniquely minimizing geodesic are always orthogonal to this geodesic. Bearing this in mind, the reader will easily become convinced that the missing direction is often the most interesting.)

For these reasons, we did not manage to make any sense of the heuristic argument. We shall now present a proof based on a radically different line of argumentation. It will rely on three main ideas: First, the use of "slightly curved" paths to explore $\mathrm{TCL}(x)$; secondly, the use of the norm of the velocity as an "ordering parameter" for a sort of continuity method; thirdly, the use of *uniform* convexity to continue the procedure. It should be noted that we will not take one x and prove the convexity of $\mathrm{I}(x)$; instead, we shall consider all x at the same time. (The uniform convexity of $\mathrm{I}(x)$ up to a certain norm will depend on convexity properties of $\mathrm{I}(y)$ for some $y \neq x$.) The technical core of the proof is Theorem 3.1. This may seem paradoxical, since the first condition in (3.3) needs some kind of convexity property of $\mathrm{I}(y_t)$, and we are precisely seeking to establish this property! But this problem will be resolved by some bootstrap argument.

Before starting the proof, we introduce a bit of notation, and two simple lemmas. Whenever $(x,v) \in TM$, we define

$$\phi(v) = -d_v \exp_x(v). \tag{5.1}$$

(So if $\gamma : [0,1] \to M$ is a constant-speed minimizing geodesic going from x to y, with initial velocity v_0 and final velocity v_1, the map ϕ is defined by $v_0 \longmapsto -v_1$.)

Lemma 5.1. *The function ϕ is a smooth norm-preserving involution $TM \to TM$, sending tangent cut locus into tangent cut locus and injectivity domain into injectivity domain:*

$$v \in \mathrm{TCL}(x) \Longrightarrow \phi(v) \in \mathrm{TCL}(\exp_x v); \qquad v \in \mathrm{I}(x) \Longrightarrow \phi(v) \in \mathrm{I}(\exp_x v).$$

Moreover, there is a constant K such that if $v, w \in \mathrm{TCL}(x)$ satisfy $\exp_x v = \exp_x w$, then

$$K^{-1}\,|v-w| \le |\phi(v) - \phi(w)| \le K\,|v-w|. \tag{5.2}$$

Lemma 5.2. *Let M be a compact Riemannian manifold with nonfocal cut locus, let $\delta > 0$ be given by Proposition C.5(a), and let $\delta' \in (0, \delta)$. Let $\overline{x} \in M$ and $v \in \mathrm{TCL}(\overline{x})$ such that $\mathrm{TCL}(\overline{x})$ admits a tangent hyperplane at v; then there are a neighborhood W of v, and a sequence $(x_k)_{k\in\mathbb{N}}$ converging to $\overline{x}$, such that for any $w \in \mathrm{TCL}(\overline{x}) \cap W$ and any $k \in \mathbb{N}$, one has*

(a) $x_k \notin \mathrm{cut}(y)$, *where* $y = \exp_{\overline{x}} w$;

(b) $\left|(\exp_y)^{-1}(x_k) - \phi(w)\right| > \delta'$.

The interpretation of Lemma 5.2 is as follows: $y = \exp_{\overline{x}} w \in \mathrm{cut}(\overline{x})$ sees $\overline{x}$ from two directions, $\phi(w)$ and, say, $\phi(\widehat{w})$, which are away from each other. By perturbing $\overline{x}$ into x_k we can make sure that $\phi(\widehat{w})$ is more favorable than $\phi(w)$.

Proof of Lemma 5.1. The first statement is obvious from the definition of ϕ. The second one is an immediate consequence of the observation that if γ is a geodesic joining x to y, then y is a cut point of x along γ if and only if x is a cut point of y along the time-reversed γ. As for the third statement, the inequality on the right is obtained with K equal to the supremum of all Lipschitz constants of $v \longmapsto -d_v \exp_x v$, and this implies the inequality on the left since ϕ is an involution on $\mathrm{TCL}(M)$. □

Proof of Lemma 5.2. By Proposition C.5(c), there is exactly one velocity $\widehat{v} \in \mathrm{TCL}(\overline{x})$ such that $\widehat{v} \neq v$ and $\exp_{\overline{x}} \widehat{v} = \exp_{\overline{x}} v$. By the same reasoning as in the proof of Proposition C.5, there are neighborhoods U of $\overline{x}$, V of $y = \exp_{\overline{x}} v$, W of v (in TM) and $\widehat{W}$ of $\widehat{v}$, such that any $x \in U$ and any $y \in V$ can be joined by exactly one geodesic with initial velocity $w \in W$, say $\gamma_{x,y}$; and one with initial velocity $\widehat{w} \in \widehat{W}$, say $\widehat{\gamma}_{x,y}$; and any minimizing geodesic $U \to V$ has to be one of these two geodesics.

Then $y \in \mathrm{cut}(x) \cap V$ if and only if $E(\gamma_{x,y}) = E(\widehat{\gamma}_{x,y})$, where $E(\gamma) = (1/2)\int |\dot{\gamma}(t)|^2\, dt = L(\gamma)^2/2$, and L is the length. In particular, if $w \in \mathrm{TCL}(\overline{x}) \cap W$ and $y = \exp_{\overline{x}} w$, then

$$E(\gamma_{\overline{x},y}) = E(\widehat{\gamma}_{\overline{x},y}). \tag{5.3}$$

Let $w_{x,y}$ (resp. $\widehat{w}_{x,y}$) stand for the initial velocity of $\gamma_{x,y}$ (resp. $\widehat{\gamma}_{x,y}$); this is a smooth function of $(x, y) \in U \times V$. Let then $x_k = \exp_{\overline{x}}(t_k \widehat{v})$ with $t_k > 0$, $t_k \to 0$ as $k \to \infty$. By the formula of first variation,

$$\begin{aligned} E(\gamma_{x_k,y}) &= E(\gamma_{\overline{x},y}) - t_k \langle \widehat{v}, w_{\overline{x},y}\rangle + O(t_k^2) \\ &= E(\gamma_{\overline{x},y}) - t_k \langle \widehat{v}, v\rangle + O(t_k \eta) + O(t_k^2), \end{aligned} \tag{5.4}$$

where $\eta = \mathrm{diam}(W \cap T_{\overline{x}} M)$.

On the other hand, taking into account (5.3),

$$E(\widehat{\gamma}_{x_k,y}) = (1-t_k)^2\,|\widehat{v}|^2 = E(\gamma_{\overline{x},y}) - t_k\,|\widehat{v}|^2 + O(t_k^2). \tag{5.5}$$

Since $|\widehat{v}| = |v|$ and $\widehat{v} \neq v$, we have $\langle \widehat{v}, \widehat{v}\rangle > \langle \widehat{v}, v\rangle$; so if η is small enough, we will have

$$E(\widehat{\gamma}_{x_k,y}) < E(\gamma_{x_k,y})$$

for all k large enough. This ensures that $x_k \notin \mathrm{cut}(y)$, and $(\exp_{x_k})^{-1}(y) = \widehat{w}_{x_k,y}$.

As we let $U, V, W, \widehat{W}$ shrink to $x, y, v, \widehat{v}$ respectively, $(\exp_y)^{-1}(x_k)$ becomes uniformly close to $\phi(\widehat{v})$. So if these neighborhoods are chosen small enough at the beginning, we have

$$|(\exp_y)^{-1}(x_k) - \phi(v)| \geq (1-\varepsilon)\,|\phi(\widehat{v}) - \phi(v)| \geq (1-\varepsilon)\,\delta,$$

where we used Lemma 5.1 and Proposition C.5(a); and ε is arbitrarily small. This implies the claim. □

Now we can enter the bulk of the proof of Theorem 1.7.

Proof of Theorem 1.7. First, $\mathrm{MTW}(K_0)$ implies that all sectional curvatures of M are bounded below by K_0, so M is compact. By Lemma 2.3, M satisfies $\mathrm{MTW}(K_0, C_0)$ for some $C_0 < \infty$. Also, by the nonfocality assumption, we do not lose anything replacing (3.3) by the stronger condition

$$|\ddot{p}_t| \leq \eta_0\,|q_t - \overline{q}_t|\,|\dot{p}_t|^2. \tag{5.6}$$

(The notation η_0 is consistent with (3.5).) We shall use it under the assumption that $\sup|\dot{p}_t|$ is so small, that λ in (3.5) is strictly positive.

From Proposition C.5(c), $\mathrm{TCL}(x)$ is semiconvex; so it is sufficient to get a lower bound on the second fundamental form of $\mathrm{TCL}(x)$ at each differentiability point $v \in \mathrm{TCL}(x)$. We shall prove

$$\mathrm{I\!I}_{\mathrm{TCL}(x)}(v) \geq \min\left(\eta_0\,\delta, \frac{1}{\mathrm{diam}(M)}\right), \tag{5.7}$$

where δ is given by Proposition C.5(a). It will be convenient to interpret the lower bound in (5.7) in the sense of (A.3); the smoothness of $\mathrm{TCL}(x)$ around v allows us to consider paths γ (denoted (p_t) in the sequel) drawn directly in $\mathrm{TCL}(x)$.

Let $\delta' \in (0, \delta)$, let $\kappa' < \min(\eta_0\,\delta', 1/(\mathrm{diam}\,M))$ and let κ be such that any set $C \subset B[0, \mathrm{diam}(M)]$ (the closed ball centered at 0 with radius $\mathrm{diam}(M)$) satisfying $\mathrm{I\!I}_C \geq \kappa'$ is κ-uniformly convex. The goal is to show that $\mathrm{I\!I}_{\mathrm{TCL}(x)} \geq \kappa'$. (Then (5.7) follows by a limiting procedure.)

The argument which we shall present is quite involved, and for pedagogical reasons, we shall first treat the case when there are no "concave parts" in the injectivity domains; more precisely we shall assume that

$$\text{There is no } (x, v) \in \mathrm{TCL}(M) \text{ with } \mathrm{I\!I}_{\mathrm{TCL}(x)}(v) \leq -\eta_0\, \delta'. \tag{5.8}$$

By this assumption we mean that for any $(x, v) \in \mathrm{TCL}(M)$ there is at least one eigenvalue of $\mathrm{I\!I}_{\mathrm{TCL}}(x)$ which is greater than $-\eta_0\, \delta'$; this should be understood in weak sense, as in Appendix A.

Let $P(r)$ and $Q(r)$ be the following two properties depending on M and r (varying in $(0, \mathrm{diam}(M)]$):

$P(r)$ *For any* $x \in M$, $v, w \in \mathrm{I}(x)$,

$$\Big(|v|, |w| \leq r,\ |v - w| \geq \delta'\Big) \implies [v, w] \subset \mathrm{I}(x).$$

$Q(r)$ *For any* $x \in M$, $\mathrm{I}(x) \cap B(0, r)$ *is* κ*-uniformly convex.*

Claim: *For any* $r \in (0, \mathrm{diam}(M)]$, *the properties* $P(r)$ *and* $Q(r)$ *are equivalent.*

Proof of Claim. It is obvious that $Q(r) \Rightarrow P(r)$ for all r; the problem is to show the converse. This will be achieved with the help of Theorem 3.1.

So let $r \in (0, \mathrm{diam}(M)]$, assume that M satisfies $P(r)$, let $\overline{x}$ be an arbitrary point in M, the goal is to prove that $\mathrm{I}(\overline{x}) \cap B(0, r)$ is κ-convex.

Since $B(0, r)$ is λ-uniformly convex for $\lambda = (\mathrm{diam}\, M)^{-1}$, the set $C = B(0, r) \cap \mathrm{TCL}(x)$ is semiconvex (in the sense of Definition B.1) and it is sufficient to establish

$$\mathrm{I\!I}_C(v) \geq \kappa' \tag{5.9}$$

at each point of differentiability v of ∂C. This is obvious if $v \in \partial B(0, r)$, so we shall assume $|v| < r$, $v \in \mathrm{TCL}(x)$ and $\mathrm{TCL}(\overline{x})$ differentiable at v.

Let W (neighborhood of v in TM) and $(x_k)_{k\in\mathbb{N}}$ be given by Lemma 5.2. Without loss of generality, each $w \in W$ satisfies $|w| < r$. Recall that $\mathrm{TCL}(\overline{x})$ is smooth around v; taking into account the simplifying assumption (5.8), we see that if (5.9) is violated, there is a C^2 path $(p(t))_{0\leq t\leq 1}$, with constant nonzero speed, drawn in $\mathrm{TCL}(\overline{x})$, such that

$$p\Big(\frac{1}{2}\Big) = v; \qquad \Big|\ddot{p}\Big(\frac{1}{2}\Big)\Big| < \eta_0\, \delta' \Big|\dot{p}\Big(\frac{1}{2}\Big)\Big|^2. \tag{5.10}$$

Restricting the path p and reparameterizing it if necessary, we may assume that

$$|\ddot{p}(t)| < \eta_0 \, \delta' \, |\dot{p}(t)|^2 \qquad \forall t \in (0,1). \tag{5.11}$$

We may also assume that $p(t) \in W$ for all $t \in [0,1]$, in particular $|p(t)| < r$.

Let then $y_t = \exp_{\overline{x}} p(t)$. By Lemma 5.2, for any t we have $x_k \notin \mathrm{cut}(y_t)$ and $|q_t - \overline{q}_t| > \delta'$, where $q_t = (\exp_{y_t})^{-1}(x_k)$, $\overline{q}_t = \phi(p(t)) \in W$. In particular, (5.11) implies (5.6).

The goal now is to apply Theorem 3.1. As $k \to \infty$, $|q_t| = d(x_k, y_t)$ approaches $d(\overline{x}, y_t) = |p(t)| < r$. So if k is large enough we have $|q_t| < r$ for all t. Further, we may perturb the path $(p(t))$ very slightly, say replace it by $(1-\alpha)p(t)$ for α very small, which ensures that $p(t) \in \mathrm{I}(\overline{x})$, and still all the previous (strict) inequalities are still satisfied with y_t replaced by $\exp_{\overline{x}}((1-\alpha)p(t))$. Then both $\overline{x}$ and x_k belong to $M \setminus \bigcup \mathrm{cut}(y_t)$; $|(\exp_{y_t})^{-1}(\overline{x})| < r$; $|(\exp_{y_t})^{-1}(x_k)| < r$; and $|(\exp_{y_t})^{-1}(\overline{x}) - (\exp_{y_t})^{-1}(x_k)| > \delta'$; so property $P(r)$ implies that $[(\exp_{y_t})^{-1}(\overline{x}), (\exp_{y_t})^{-1}(x_k)] \subset \mathrm{I}(y_t)$. Thus all the assumptions of Theorem 3.1 are fulfilled. Writing the conclusion of this theorem for $t = 1/2$, and then passing to the limit as $\alpha \to 0$, we obtain

$$d(x_k, y_{1/2})^2 - d(\overline{x}, y_{1/2})^2 \geq \min\Big(d(x_k, y_0)^2 - d(\overline{x}, y_0)^2,\ d(x_k, y_1)^2 - d(\overline{x}, y_1)^2\Big) + \frac{\lambda}{2} \, (\delta')^2 \, |p(0) - p(1)|^2,$$

where λ is determined by (3.5).

Passing to the limit $k \to \infty$, we deduce

$$0 \geq \frac{\lambda}{2} \, (\delta')^2 \, |p(0) - p(1)|^2, \tag{5.12}$$

which is a contradiction. The conclusion is that there is no nontrivial constant-speed path drawn in $\mathrm{TCL}(\overline{x})$ satisfying (5.10); therefore $Q(r)$ is satisfied. □

Back to the proof of Theorem 1.7, we note that $P(r)$ is obviously true for r small enough, so we may define

$$r_1 = \sup \big\{ r \in (0, \mathrm{diam}(M)];\ P(r) \text{ is true} \big\}.$$

Passing to the limit in $Q(r)$ as $r \to r_1$, we see that $P(r_1)$ and $Q(r_1)$ hold true. The goal is to show that $r_1 = \mathrm{diam}(M)$.

Should this be false, we would find a sequence $(\overline{x}_k)_{k \in \mathbb{N}}$ in M, $v_k, w_k \in \mathrm{I}(\overline{x}_k)$, $t_k \in (0,1)$ such that $|v_k|, |w_k| < r_1 + 1/k$, $|v_k - w_k| \geq \delta'$ and $z_k := (1-t_k)v_k + t_k w_k \notin \mathrm{I}(\overline{x}_k)$. Extracting subsequences if necessary, we may assume $\overline{x}_k \to \overline{x} \in M$, $v_k \to v \in \overline{\mathrm{I}}(\overline{x}_k)$, $w_k \to w \in \overline{\mathrm{I}}(\overline{x})$, $z_k \to z \in T_{\overline{x}} M$, $t_k \to \theta \in [0,1]$.

Then if $\theta \in (0,1)$ we have

$$v, w \in \overline{\mathrm{I}}(\overline{x}), \qquad |v|, |w| \le r_1, \qquad |v-w| \ge \delta', \qquad (1-\theta)v + \theta w \notin \mathrm{I}(\overline{x}).$$

Since $|(1-\theta)v + \theta w| < r_1$, this contradicts the uniform convexity of $\mathrm{I}(\overline{x}) \cap B(0, r_1)$; so this case is impossible.

The only case in which we cannot assume $\theta \in (0,1)$ is if $(1-t)v_k + tw_k \in \mathrm{I}(\overline{x}_k)$ for all $t \in [\theta_k, 1-\theta_k]$, where $\theta_k \to 0$; in particular $t_k \in [0, \theta_k] \cup [1-\theta_k]$, without loss of generality $t_k \in [0, \theta_k]$. Then since $v_k \in \mathrm{I}(\overline{x}_k)$ and $(1-t_k)v_k + t_k w_k \notin \mathrm{I}(\overline{x}_k)$, there is u_k arbitrarily close to $w_k - v_k$ such that $v_k + t_k u_k \notin \overline{\mathrm{I}}(\overline{x}_k)$, so there is a largest $t \le t_k$ such that $v_k + tu_k \in \mathrm{TCL}(\overline{x}_k)$, call it τ_k; then u_k belongs to $T^{\mathrm{ext}}_{v_k + \tau_k u_k} \mathrm{I}(\overline{x}_k)$, the exterior tangent cone to $\mathrm{I}(\overline{x}_k)$ at $v_k + \tau_k u_k$. By Proposition C.5(c), this cone is upper semicontinuously differentiable, in the sense of Definition C.4; so we may pass to the limit as $u_k \to w_k - v_k$, and deduce that $w_k - v_k \in T^{\mathrm{ext}}_{(1-\tau_k)v_k + \tau_k w_k} \mathrm{I}(\overline{x}_k)$. Since $\tau_k \to 0$, we may further pass to the limit as $k \to \infty$ to recover

$$w - v \in T^{\mathrm{ext}}_v \mathrm{I}(\overline{x}).$$

But since $v, w \in \overline{\mathrm{I}}(\overline{x})$ and $|v - w| \ge \delta' > 0$, this again contradicts the uniform convexity of $\mathrm{I}(\overline{x}) \cap B(0, r_1)$. So this case also is impossible, and we conclude that $r_1 = \mathrm{diam}(M)$, finishing the proof of the uniform convexity of all injectivity domains.

Now let us see how to do without the assumption (5.8). The problem is that if $\mathrm{TCL}(x)$ is "too much concave" at v, we may not find a path satisfying (5.10). The idea to resolve this problem is the following: if $B(0, r)$ touches a concavity region of $\mathrm{TCL}(x)$ as r increases, this contact will be tangential, and there the normal to $\mathrm{TCL}(x)$ will have a well-specified direction, which will allow to get the signs right at the level of the second condition in (3.3).

So from now on we shall not assume (5.8). The equivalence between $P(r)$ and $Q(r)$ remains true as long as $\partial(B(0,r) \cap \mathrm{I}(x))$ does not have a region of strict concavity. When we run the continuity procedure by letting r increase as before, we have two possibilities: either $r_1 = \mathrm{diam}(M)$, or there are $r \le r_1$, and $(x, v) \in \mathrm{TCL}(M)$ with $|v| = r$, such that $\mathrm{TCL}(x)$ is twice differentiable at v and $\mathrm{II}_{\mathrm{TCL}(x)}(v) \le -\eta_0\, \delta'$. Let r_2 be the infimum of such r; then we have a sequence $(x_\ell, v_\ell) \in \mathrm{TCL}(M)$ such that $\mathrm{II}_{\mathrm{TCL}(x_\ell)}(v_\ell) \le -\eta_0\, \delta'$ and $|v_\ell| \to r_2$. Without loss of generality, (x_ℓ, v_ℓ) converges in $\mathrm{TCL}(M)$ to some $(\overline{x}, \overline{v})$. If $\mathrm{TCL}(M)$ is smooth around $(\overline{x}, \overline{v})$ then necessarily $\mathrm{II}_{\mathrm{TCL}(\overline{x})}(\overline{v}) \le -\eta_0\, \delta'$; otherwise, $\mathrm{TCL}(M)$ presents an outward pointing singularity at $(\overline{x}, \overline{v})$, and this is incompatible with the property that $\overline{v}$ should be norm-minimizing. Thus $\mathrm{TCL}(\overline{x})$ is uniformly concave in a ball $B[\overline{v}, \alpha]$ centered at $\overline{v}$ ($\alpha > 0$), and the uniform convexity of $B(0,r_2)$ implies that $\overline{v}$ is the unique minimizer of $w \longmapsto |w|$ in $\mathrm{TCL}(\overline{x}) \cap B[\overline{v},\alpha]$. In particular, $B(0,r_2)$ is tangent to $\mathrm{TCL}(\overline{x})$ at $\overline{v}$, and both sets admit a common normal there, which is $\overline{v}/|\overline{v}|$.

It follows from formula (C.5) in Appendix C that the outward normal to $\mathrm{TCL}(\overline{x})$ at $\overline{v}$ is proportional to $(d_{\overline{q}}\exp_y)(q-\overline{q})$, where $\overline{q}=\phi(\overline{v})$ and $q=\phi(\widehat{v})$, with $\exp_{\overline{x}}\widehat{v}=\exp_{\overline{x}}\overline{v}$, $\widehat{v}\neq\overline{v}$ in $\mathrm{TCL}(\overline{x})$. So that normal vector is positively colinear to $\overline{v}$. Since $(d_{\overline{q}}\exp_y)(\overline{q}) = -\overline{v}$, we deduce that also $(d_{\overline{q}}\exp_y)(q)$ is proportional to $\overline{v}$, so by nonfocality q is proportional to $\overline{q}$, and by equality of norms the only possibility is $q=-\overline{q}$. (This is more or less a variant of the proof of [8, Chapter 13, Proposition 2.12], a step in Klingenberg's proof of the sphere theorem.) So we also have

$$(d_{\overline{v}}\exp_{\overline{x}})^{-1}(q-\overline{q}) = 2\,\overline{v}. \tag{5.13}$$

By the same reasoning as before, $Q(r_2)$ is true, and $P(r)$ is satisfied for $r\leq r_2+\varepsilon$, for some $\varepsilon>0$. So we can do as in the proof of the equivalence of $P(r)$ and $Q(r)$, with a path $(p(t))_{0\leq t\leq 1}$ drawn in $\mathrm{TCL}(\overline{x})\cap B(0,r_2+\varepsilon)$, satisfying

$$p\Big(\frac{1}{2}\Big)=\overline{v};\qquad \Big\langle \ddot{p}\Big(\frac{1}{2}\Big),\overline{v}\Big\rangle>0,$$

where the second inequality comes from the concavity and the fact that $\overline{v}$ is the normal to $\mathrm{TCL}(\overline{x})$. In view of (5.13), this is the same as

$$\Big\langle \ddot{p}\Big(\frac{1}{2}\Big),\,(d_{\overline{v}}\exp_{\overline{x}})^{-1}(q-\overline{q})\Big\rangle>0.$$

Restricting and reparameterizing the path $(p(t))$ if necessary, we may assume that it satisfies

$$\big\langle \ddot{p}(t),\,(d_v\exp_x)^{-1}(q_t-\overline{q}_t)\big\rangle>0$$

for all $t\in(0,1)$; then the second condition in (3.3) is obviously true, and we arrive at a contradiction by repeating the reasoning used in the proof of equivalence of $P(r)$ and $Q(r)$. □

In the end of this section we sketch the proof of Theorem 1.8(a).

Sketch of proof of Theorem 1.8(a). Let M satisfy the condition $\mathrm{MTW}(K_0,C_0)$ of Section 2, for some $K_0,C_0>0$, such that $\delta(M)$ defined in (1.5) is positive, and such that for any $x\in M$, $\mathrm{TFL}(x)$ has nonnegative second fundamental form near $\mathrm{TCL}(x)\cap\mathrm{TFL}(x)$.

The main step is to show that the injectivity domains are uniformly convex. For this the proof of Theorem 1.7 can be repeated, with the following modifications.

• By Proposition C.6, whose proof was communicated to us by Rifford, $\mathrm{I}(M)$ is semiconvex (in a sense which is made precise in Appendix C.3).

• When we establish the equivalence between properties $P(r)$ and $Q(r)$, we cannot any longer draw C^2 paths in $\mathrm{TCL}(\overline{x})$, since these sets are not necessarily smooth at differentiability points. However, $\mathrm{TCL}(\overline{x})$ is the boundary of a semiconvex set, so (by Alexandrov's theorem) it can be osculated at almost all of its points, from the inside (i.e. from $\mathrm{I}(\overline{x})$) and from the outside, by two spheres of arbitrarily close radii. If $\mathrm{I}(\overline{x})$ is not uniformly convex there, and not uniformly concave either, then we can find a C^2 path $(p(t))_{0\le t\le 1}$ drawn in $\overline{\mathrm{I}}(\overline{x})$ with $p(1/2) \in \mathrm{TCL}(\overline{x})$, $|\ddot{p}(t)| \le \varepsilon\,|\dot{p}(t)|^2$, $0 < |\dot{p}(t)| < \eta$, and $\mathrm{dist}(p(t), \mathrm{TCL}(\overline{x})) \le \gamma\,|p(0) - p(1)|^2$, where $\varepsilon, \eta, \gamma$ are arbitrarily small.

• In presence of focalization, (5.6) is a priori strictly stronger than (3.3), but this goes in the right direction for our purpose.

• Lemma 5.2 a priori does not apply. However, by Lemma 5.3 below, we can find $(x_k)_{k\in\mathbb{N}}$ such that x_k belongs to $\mathrm{cut}(y_t)$ only for finitely many times t, where $y_t = \exp_{\overline{x}} p(t)$; and $|(\exp_{y_t})^{-1}(x_k) - \phi(p(t))| > \delta'$ (depending only on M); and $d(x_k, \overline{x}) \le C\,\mathrm{dist}(p(t), \mathrm{TCL}(\overline{x})) + \alpha_k$, $\alpha_k \to 0$. The fact that y_t crosses $\mathrm{cut}(x_k)$ only finitely many times makes it possible to apply Theorem 3.1, thanks to a reasoning similar to the one in [30, Proof of Theorem 12.36]. Then the inequality

$$d(x_k, y_{1/2})^2 - d(\overline{x}, y_{1/2})^2 \ge \min\big(d(x_k, y_0)^2 - d(\overline{x}, y_0)^2,\ d(x_k, y_1)^2 - d(\overline{x}, y_1)^2\big) + \frac{\lambda}{2}\,(\delta')^2\,|p(0) - p(1)|^2$$

implies

$$\frac{\lambda}{2}\,(\delta')^2\,|p(0) - p(1)|^2 \le C'\,d(x_k, \overline{x}) \le C''\sup_{0\le t\le 1}\Big(\mathrm{dist}(p(t), \mathrm{TCL}(\overline{x})), + o(1)\Big) \le C''\,\gamma\,|p(0) - p(1)|^2 + o(1).$$

Taking $k \to \infty$ and γ small enough leads to a contradiction.

• The semiconvexity of $\mathrm{I}(M)$ implies the u.c.s.d.e. property used in the end of the proof of the first part of Theorem 1.7.

• Then we should rule out the possibility of $\mathrm{TCL}(x)$ becoming strictly concave. The reasoning used in the proof of Theorem 1.8 works away from $\mathrm{TCL} \cap \mathrm{TFL}$; then the convexity assumption in the statement of Theorem 1.8 rules out the possibility of this event happening in the neighborhood of $\mathrm{TCL} \cap \mathrm{TFL}$. (Recall that $\mathrm{I}(x)$ is included in the region interior to TFL(x); so if $v \in$ TCL(x) $\cap$ TFL(x) and TCL(x) is concave at v, then necessarily v is a concavity point of TFL(x).)

□

The following technical lemma was used in the proof of Theorem 1.8(a):

Lemma 5.3. *Let M be a compact Riemannian manifold satisfying $\delta(M) > 0$. Then there are $\delta', C, \eta > 0$ with the following property. For any $\overline{x} \in M$ and any $v \in$ TCL($\overline{x}$) there is a neighborhood W of v, of size η, such that for any (Lipschitz) path $(w_t)_{0\leq t\leq 1}$ drawn in $\overline{\mathrm{I}}(\overline{x}) \cap W$, there is a sequence $(x_k)_{k\in\mathbb{N}}$ of points in M such that, for all $k \in \mathbb{N}$,*

(a) x_k belongs to $\mathrm{cut}(y_t)$ only for a finite set of times t, where $y_t = \exp_{\overline{x}} w_t$;

(b) for all $t \in [0,1]$, $\mathrm{dist}((\exp_{y_t})^{-1}(x_k), \phi(w_t)) > \delta'$ (where $(\exp_{y_t})^{-1}(x_k)$ is a set if $x_k \in \mathrm{cut}(y_t)$);

(c) $d(x_k, \overline{x}) \leq C \sup_{0\leq t\leq 1} \mathrm{dist}(w_t, \mathrm{TCL}(\overline{x})) + \alpha_k$, where $\alpha_k \to 0$ as $k \to \infty$.

Recall that by convention the notation $(\exp_y)^{-1}(x)$ always stands for one or several *minimizing* velocities v such that $\exp_x v = y$. The proof of Lemma 5.3 uses the co-area formula in the same way as [11]. Another idea is to perturb $\overline{x}$ in the direction opposite to v, rather than in the direction of a minimizing geodesic as in Lemma 5.2.

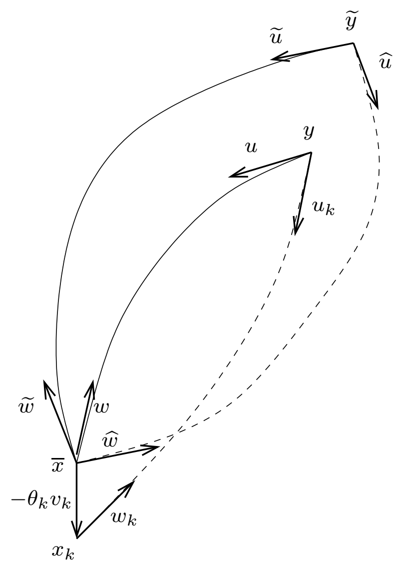

FIGURE 2. Picture of Lemma 5.3

Proof of Lemma 5.3. Let $\overline{x} \in M$, $v \in \mathrm{TCL}(\overline{x})$, $W = B(v, \eta)$ (ball of radius η around v, where η will be determined later). Let $(w_t)_{0\le t\le 1}$ be drawn in $\mathrm{TCL}(\overline{x}) \cap W$, and $y_t = \exp_{\overline{x}} w_t$. Let $\mathcal{H}^k$ stand for the k-dimensional Hausdorff measure. The set

$$\Sigma = \bigcup \{t\} \times \mathrm{cut}(y_t) \quad \subset \; [0,1] \times M$$

has finite $\mathcal{H}^n$ measure in view of the results in [13, 18]. By localization, this remains true in a system of Euclidean coordinates around $\overline{x}$, so Σ can be seen as a subset of $[0,1] \times B(0,1) \subset \mathbb{R}^{n+1}$, and the Hausdorff measure can be thought of as the Hausdorff measure in $\mathbb{R}^{n+1}$. Applying the co-area formula [9, Sections 2.10.25 and 2.10.26] with the map $f(t,x) = x$, we have

$$\mathcal{H}^n[\Sigma] \ge \int_{f(\Sigma)} \mathcal{H}^0[\Sigma \cap f^{-1}(z)]\, \mathcal{H}^n(dz).$$

Since the left-hand side is finite, $\mathcal{H}^0[\Sigma \cap f^{-1}(z)] < \infty$ for almost all z. In particular, if U is a small neighborhood of $\overline{x}$, of size r, then

$$Z = \Big\{ z \in U;\ z \notin \mathrm{cut}(y_t) \text{ except for finitely many times } t \Big\}$$

is dense in U. So we may choose $x_k = \exp_{\overline{x}}(-\theta_k\, v_k) \in Z$, where $\theta_k \in (0, r/(2\,\mathrm{diam}(M)))$ and $|v_k - v| \le \eta$.

For simplicity we shall now write $w = w_t$. Let $\widetilde{w} \in \mathrm{TCL}(\overline{x})$ such that $|w - \widetilde{w}| = \mathrm{dist}(w, \mathrm{TCL}(\overline{x}))$. By assumption there is $\widehat{w} \in \mathrm{TCL}(\overline{x})$ such that $|\widehat{w}| = |\widetilde{w}|$, $\exp_{\overline{x}} \widehat{w} = \exp_{\overline{x}} \widetilde{w}$ and $|\widehat{w} - \widetilde{w}| \ge \delta$. Let $u = \phi(w)$, $\widehat{u} = \phi(\widehat{w})$, $\widetilde{u} = \phi(\widetilde{w})$. (See Figure 2.) Let further $u_k \in (\exp_y)^{-1}(x_k)$, the goal is to show that $|u_k - u| \ge \delta'$ if δ' and θ_k are well-chosen. By Lemma 5.1 it suffices to show that $|w - w_k| \ge \delta'\, K$, where $w_k = \phi(u_k)$. We shall do this by contradiction. Assume that $|w - w_k| \le \eta$ (so $|v_k - w| \le |v_k - v| + |v - w| \le 2\eta$). Since $d(x_k, y)^2 - d(\overline{x}, y)^2 \le d(x_k, \widetilde{y})^2 - d(\overline{x}, \widetilde{y})^2 + O(d(y, \widetilde{y}))$, the formula of first variation for $(\,\cdot\,, \widetilde{y})$ implies

$$\begin{aligned} d(x_k, y)^2 &\le d(\overline{x}, y)^2 + 2\,\langle \theta_k v_k, \widehat{w}\rangle + O(\theta_k^2) + O(d(y,\widetilde{y})) \\ &\le d(\overline{x}, y)^2 + 2\theta_k\, \langle w, \widehat{w}\rangle + O(\theta_k\,\eta) + O(\theta_k^2) + O(d(y,\widetilde{y})). \end{aligned}$$

Let $\gamma_k(s) = \exp_{x_k}(s w_k)$, and let $L(\gamma_k)$ stand for the length of the path γ_k. Also by the formula of first variation,

$$\begin{aligned} L(\gamma_k)^2 &= d(\overline{x}, y)^2 + 2\,\langle \theta_k v_k, w_k\rangle + O(\theta_k^2) \\ &= d(\overline{x}, y)^2 + 2\theta_k\, |w|^2 + O(\theta_k\,\eta) + O(\theta_k^2). \end{aligned}$$

By assumption $L(\gamma_k)^2 = d(x_k, y)^2$; moreover $d(y, \widetilde{y}) \le K' \, |w - \widetilde{w}| = K' \operatorname{dist}(w, \mathrm{TCL}(\overline{x}))$ for some constant K' ($K' = 1$ will do in nonnegative curvature). So the above formulae imply

$$\theta_k \, |w|^2 \le \theta_k \, \langle w, \widehat{w} \rangle + O(\theta_k \, \eta) + O(\theta_k^2) + O\big(\operatorname{dist}(w, \mathrm{TCL}(\overline{x}))\big).$$

Since $|\widetilde{w}| = |\widehat{w}|$ and $|w - \widetilde{w}| \le \operatorname{dist}(w, \mathrm{TCL}(\overline{x}))$ this implies

$$\theta_k \, |\widetilde{w} - \widehat{w}|^2 \le O(\theta_k \, \eta) + O(\theta_k^2) + O\big(\operatorname{dist}(w, \mathrm{TCL}(\overline{x}))\big).$$

Since $|\widetilde{w} - \widehat{w}| \ge \delta$ this leads to

$$\theta_k \, \delta^2 \le \text{const.} \, \big(\theta_k \, \eta + \theta_k^2 + \operatorname{dist}(w, \mathrm{TCL}(\overline{x}))\big), \tag{5.14}$$

where the constant in the right-hand side depends only on M.

Inequality (5.14) is impossible if η is much smaller than δ^2 and θ_k is much smaller than δ^2 and much larger than $\operatorname{dist}(w, \mathrm{TCL}(\overline{x}))/\delta^2$. More formally, we can find a constant K, depending only on M, such that if $K \operatorname{dist}(w, \mathrm{TCL}(\overline{x}))/\delta^2 \le \theta_k \le \delta^2/K$ and $\eta \le \delta^2/K$, then $|w - w_k| > \eta$.

So we may choose, say, $\eta = \min(\delta^4/K^2, \delta^2/K)/2$, and then we can choose θ_k satisfying the above inequalities for all $w = w_t$, still $\theta_k \le K' \sup_t \operatorname{dist}(w_t, \mathrm{TCL}(\overline{x})) + \epsilon_k$, where ϵ_k goes to zero (the introduction of ϵ_k is necessary only in the case when $\operatorname{dist}(w_t, \mathrm{TCL}(\overline{x})) = 0$). Then the preceding reasoning shows that $|w - w_k| > \eta$, in particular $|u - u_k|$ is bounded below by means of Lemma 5.1. This concludes the proof of Lemma 5.3. □

6. From c-convexity to C^1 regularity

From now on, our main results do not explicitly need the nonfocality, but just the property of uniform regularity. In this section we denote by $\mathcal{B}(M)$ the Borel σ-algebra in M, and write $B_r(x) = B(x, r)$.

Theorem 6.1. *Let M be a uniformly regular Riemannian manifold, in the sense of Definition 4.1. Let $\mu(dx)$ and $\nu(dy)$ be two probability measures on M. If*

$$\lim_{\varepsilon \to 0} \left(\sup_{x \in M} \frac{\mu[B_\varepsilon(x)]}{\varepsilon^{n-1}} \right) = 0; \qquad \inf_{A \in \mathcal{B}(M)} \frac{\nu[A]}{\operatorname{vol}[A]} > 0, \tag{6.1}$$

then the optimal transport between μ and ν, for the cost function d^2, takes the form $T(x) = \exp_x \nabla\psi(x)$, where ψ is a C^1 semiconvex function. In particular T is continuous.

Corollary 6.2. *The conclusion of Theorem 6.1 holds true if* μ *and* ν *satisfy*

$$\sup_M \left(\frac{d\mu}{d\mathrm{vol}}\right) < +\infty; \qquad \inf_M \left(\frac{d\nu}{d\mathrm{vol}}\right) > 0;$$

where the second inequality means $\nu \geq a\,\mathrm{vol}$ *for some* $a > 0$.

Remark 6.3. The optimal transport map T is only defined μ-almost surely; on the other hand, the function ψ appearing in Theorem 6.1 is well-defined and C^1 *everywhere* in M.

Proof of Theorem 6.1. The assumption implies that μ does not charge sets of σ-finite $(n-1)$-dimensional Hausdorff measure. From McCann's theorem, in the form of [30, Theorem 10.41], there is a semiconvex function ψ such that $T = \exp(\nabla\psi)$. In particular, ψ is semiconvex, and $\nabla\psi(x) \in \overline{\mathrm{I}}(x)$ whenever x is a point of differentiability of ψ. A limiting argument using the semiconvexity of ψ shows that the subdifferential $\nabla^-\psi(x)$ is also valued in the convex hull of $\overline{\mathrm{I}}(x)$ (see [30, Remark 10.27]). Since $\overline{\mathrm{I}}(x)$ is convex, in fact $\nabla^-\psi(x) \subset \overline{\mathrm{I}}(x)$, for all $x \in M$. The problem is to show that ψ is differentiable in M, or equivalently that $\nabla^-\psi$ is *everywhere* single-valued; then by semiconvexity ψ will be continuously differentiable.

The proof of differentiability will be based on a strategy introduced by the first author in [20], which will be used again in Sections 7 and 8. It consists in combining the Kantorovich duality [30, Chapter 5], inequalities between distances in the style of (4.2), and mass comparison between certain well-chosen sets.

Reasoning by contradiction, assume that there are $\overline{x} \in M$ and $v_0 \neq v_1$ in $\overline{\mathrm{I}}(\overline{x})$ such that $v_0, v_1 \in \nabla^-\psi(\overline{x})$. Let $y_0 = \exp_{\overline{x}} v_0$, $y_1 = \exp_{\overline{x}} v_1$. (At this stage we do not exclude that $y_0 = y_1$.) By [30, Theorem 12.34], $c = d^2/2$ is a regular cost function, and by [30, Proposition 12.14(ii)] both y_0 and y_1 belong to $\partial_c\psi(\overline{x})$, which means

$$\psi(\overline{x}) + \frac{d(\overline{x}, y_i)^2}{2} = \inf_{x\in M} \left\{\psi(x) + \frac{d(x, y_i)^2}{2}\right\}, \qquad i = 0, 1. \tag{6.2}$$

In particular,

$$\frac{d(x, y_i)^2}{2} - \frac{d(\overline{x}, y_i)^2}{2} \geq \psi(\overline{x}) - \psi(x), \qquad i = 0, 1. \tag{6.3}$$

Now for $\varepsilon \in (0, 1)$ we define a region D_ε in $\overline{\mathrm{I}}(\overline{x})$ as follows: D_ε is swept by paths $(p_t)_{0\leq t\leq 1}$ drawn in $\overline{\mathrm{I}}(\overline{x})$, evaluated at $t \in [1/4, 3/4]$, with $p_0 = v_0$, $p_1 = v_1$, $|\dot{p}| \leq 2\,\mathrm{diam}(M)$, $\ddot{p}_t = 0$ for $t \notin [1/4, 3/4]$, $|\ddot{p}| \leq \eta_0\,\varepsilon\,|\dot{p}|^2$ for $t \in [1/4, 3/4]$.

By uniform convexity of $\overline{\mathrm{I}}(\overline{x})$, if η_0 is small enough then D_ε lies a positive distance σ away from $\mathrm{TCL}(\overline{x})$, where σ is proportional to $|v_0 - v_1|^2$. It follows that the paths

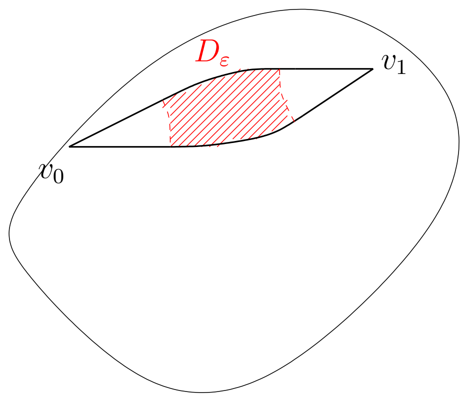

FIGURE 3. Construction of the set $D_\varepsilon \subset \bar{\mathrm{I}}(\overline{x})$.

(p_t) used to construct D_ε satisfy

$$\|(d_{p_t} \exp_{\overline{x}})^{-1}\| \leq C(M, \eta_0, |v_0 - v_1|)$$

for $t \in [1/4, 3/4]$, and condition (4.1) is satisfied as soon as η_0 is small enough and $d(\overline{x}, x) \geq \varepsilon$.

In the sequel, the notation $f = \Omega(g)$ means that f/g is bounded *from below* by a positive constant. Elementary geometric considerations show that D_ε contains a parallelepiped centered at $(v_0 + v_1)/2$, with one side of length $\Omega(|v_0 - v_1|)$ and all other sides of length $\Omega(\varepsilon)$; in particular

$$\mathcal{L}^n[D_\varepsilon] = \Omega(\varepsilon^{n-1}), \tag{6.4}$$

where $\mathcal{L}^n$ stands for the Lebesgue measure in $T_x M$. (The constants in (6.4) may depend on v_0, v_1, but here we only care about the dependence in ε.) Since D_ε is away from $\mathrm{TCL}(\overline{x})$, the exponential map is bi-Lipschitz in a neighborhood of D_ε, so the interpretation of the volume measure as a Hausdorff measure yields

$$\mathrm{vol}\,[Y_\varepsilon] = \Omega(\varepsilon^{n-1}), \qquad Y_\varepsilon = \exp_{\overline{x}}(D_\varepsilon). \tag{6.5}$$

If $x \in M$ satisfies $d(\overline{x}, x) \geq \varepsilon$, the paths $(p_t)_{0 \leq t \leq 1}$ used in the construction of D_ε satisfy all the assumptions in Definition 4.1, except that they are not necessarily valued in $\mathrm{I}(\overline{x})$. (Think that v_0 or v_1 may belong to $\mathrm{TCL}(\overline{x})$.) Replacing (p_t) by $(\theta\, p_t)_{0 \leq t \leq 1}$, where $\theta \in (0, 1)$ is very close to 1, we may apply Definition 4.1 to this modified path and then pass to the limit as $\theta \to 1$. In the end we conclude that for any $y \in Y_\varepsilon$, and any $x \in M \setminus B_\varepsilon(\overline{x})$,

$$d(x,y)^2 - d(\overline{x},y)^2 \geq \min\Big[d(x,y_0)^2 - d(\overline{x},y_0)^2,\, d(x,y_1)^2 - d(\overline{x},y_1)^2\Big] + \alpha\,\varepsilon^2\,|v_1 - v_0|^2 \tag{6.6}$$

for some $\alpha > 0$ ($\alpha = 3\lambda/8$ will do, where λ is defined by (3.5)).

Inequality (6.6) combined with (6.3) implies

$$\frac{d(x,y)^2}{2} - \frac{d(\overline{x},y)^2}{2} > \psi(\overline{x}) - \psi(x),$$

for all $y \in Y_\varepsilon$. So $y \notin \partial_c\psi(x)$, and by optimal transport theory [30, Chapter 5], the pair (x,y) cannot belong to the support of the optimal transport plan. Thus all the mass which is brought into Y_ε by the optimal transport map has to come from $B_\varepsilon(\overline{x})$. In particular

$$\mu[B_\varepsilon(\overline{x})] \geq \nu[Y_\varepsilon]. \tag{6.7}$$

By letting $\varepsilon \to 0$, we obtain a contradiction between (6.1), (6.5) and (6.7). The conclusion is that $\nabla^-\psi(\overline{x})$ is single-valued. □

7. Stay-away property

We now give sufficient conditions for the optimal transport map to stay away from the cut locus. Let $\delta(M)$ be defined as in (1.5). Recall from Proposition C.5(a) that $\delta(M) > 0$ if M has nonfocal cut locus.

Theorem 7.1. *Let M be a uniformly regular manifold, in the sense of Definition 4.1, such that $\delta(M) > 0$. Let μ and ν be two probability measures on M; assume that*

$$\begin{cases} \forall \varepsilon \in (0,1) \quad \mu[B_\varepsilon(x)] \leq m(\varepsilon)\,\varepsilon^{n-1} \\ \\ \forall A \in \mathcal{B}(M), \quad \nu[A] \geq a\,\mathrm{vol}\,[A], \end{cases} \tag{7.1}$$

where $m(\varepsilon) \to 0$ as $\varepsilon \to 0$, and $a > 0$. Then there is $\sigma > 0$ depending on μ,ν only via m and a, such that

$$\forall x \in M, \qquad d\big(T(x), \mathrm{cut}(x)\big) \geq \sigma.$$

We shall give two proofs of Theorem 7.1. The first one will be very simple but not constructive; the second one will be constructive and in particular provide estimates on σ. Apart from the very particular case of the sphere, treated in [7], these seem to be the first constructive lower bounds on the distance of optimal transport to the cut locus.

Nonconstructive proof of Theorem 7.1. From Section 6 we know that the optimal transport map T is continuous. Since the cut locus is compact, to prove that T stays away from the cut locus it suffices to check that $T(\overline{x}) \notin \mathrm{cut}(\overline{x})$, for all $\overline{x}$.

The argument, based on super- and subdifferentiability, is the same as in the proof of McCann's theorem, see [30, Theorem 10.28]. Let $\overline{x}$ be a point in M. Since ψ is $d^2/2$-convex and ψ is differentiable, we have

$$\frac{d(x,T(\overline{x}))^2}{2} - \frac{d(\overline{x},T(\overline{x}))^2}{2} \geq \psi(\overline{x}) - \psi(x) = \langle -\nabla\psi(\overline{x}), x-\overline{x}\rangle + o(d(x,\overline{x})); \tag{7.2}$$

here x is an arbitrary point in a neighborhood of $\overline{x}$, and $x-\overline{x}$ is an abuse of notation for $\dot{\gamma}(0)$, where γ is the unique (constant-speed, minimizing) geodesic going from $\overline{x}$ to x.

By (7.2), $-\nabla\psi(\overline{x})$ is a subgradient of $d(\cdot\,,T(\overline{x}))^2/2$ at $\overline{x}$. Being superdifferentiable and subdifferentiable at $\overline{x}$, this function is plainly differentiable, and from our nonfocality assumption this implies that $T(\overline{x}) \notin \mathrm{cut}(\overline{x})$ (Proposition C.5(b)).

This argument shows the existence of $\sigma > 0$ such that $d(x,T(x)) \geq \sigma$; it remains to check that these bounds are uniform in μ,ν. Should this be false, we would have sequences of probability measures $(\mu_k)_{k\in\mathbb{N}}$, $(\nu_k)_{k\in\mathbb{N}}$, satisfying the assumptions of the theorem, and associated functions ψ_k such that $d(\exp_{x_k}\nabla\psi(x_k),\mathrm{cut}(x_k)) \leq 1/k$ for some x_k. As $k\to\infty$, we may extract subsequences such that μ_k and ν_k converge weakly to μ and ν, satisfying the same assumptions (that is, (7.1)). Moreover, without loss of generality ψ_k converges uniformly to some $d^2/2$-function ψ, and x_k converges to x. Passing to the limit in the dual Kantorovich problem [30, Chapter 5], we see that ψ is optimal, and Theorem 6.1 implies that ψ is C^1, while the previous reasoning shows that $\exp_x\nabla\psi(x)\notin\mathrm{cut}(x)$. Next, the ψ_k are uniformly semiconvex and converge uniformly to the C^1 function ψ; it follows (by a localization argument and the corresponding result for convex functions in $\mathbb{R}^n$; a precise statement is in [31, Lemma 5.4]) that $\nabla\psi_k$ converges locally uniformly to $\nabla\psi$. So one can pass to the limit in the equation $d(\exp_{x_k}\nabla\psi(x_k),\mathrm{cut}(x_k))\leq 1/k$, to get $\exp_x\nabla\psi(x)\in\mathrm{cut}(x)$. But this is a contradiction. □

Constructive proof of Theorem 7.1. Let $\overline{x}$ be arbitrary in M, let $y_0 = T(\overline{x})$ and let $p_0 \in \overline{\mathrm{I}}(\overline{x})$ such that $\exp_{\overline{x}} p_0 = y_0$. Let y_c be the cut point of $\overline{x}$ along the geodesic $(\exp_{\overline{x}}(t\,p_0))_{t\geq 0}$, and let p_c be the corresponding velocity. The problem is to get a lower bound on $|p_c-p_0| = d(y_0,y_c)$; once this is done, a lower bound on $d(y_0,\mathrm{cut}(\overline{x}))$ follows by the same argument as in the proof of Proposition A.5 in Appendix A.

By assumption there are $\delta = \delta(M) > 0$ and $p_1 \in \bar{\mathrm{I}}(\overline{x})$ such that $\exp_{\overline{x}} p_1 = \exp_{\overline{x}} p_c$ and $|p_c - p_1| \geq \delta > 0$. Let $y_1 = \exp_{\overline{x}} p_1 = y_c$ (see Figure 4). We define the regions D_ε and Y_ε in the same way as in Section 6. There is a lower bound on $|p_0 - p_1|$ (depending on $\mathrm{diam}(M)$, $\mathrm{inj}(M)$ and $\delta(M)$), so that estimate (6.5) also holds in the present case, uniformly in $\overline{x}$ and p_0.

Comparing masses as in the proof of Theorem 6.1, we find that for ε small enough (much smaller than a/A), the optimal transport has to send some mass from $M \setminus B_\varepsilon(\overline{x})$ to Y_ε; in other words there is (x, y) in the support of the optimal transport plan such that $d(\overline{x}, x) \geq \varepsilon$ and $y \in Y_\varepsilon$. Then by Kantorovich duality,

$$\frac{d(x,y)^2}{2} - \frac{d(\overline{x},y)^2}{2} \leq \psi(\overline{x}) - \psi(x). \tag{7.3}$$

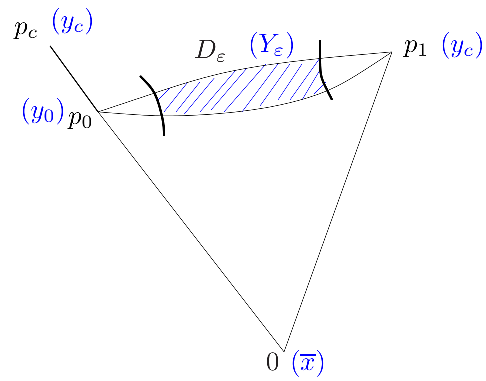

FIGURE 4. Constructive proof of Theorem 7.1. The problem is to show that p_0 does not approach the cut velocity p_c.

On the other hand, the same estimates as in the proof of Theorem 6.1 show that for all $y \in Y_\varepsilon$ and $x \notin B_\varepsilon(\overline{x})$

$$\begin{aligned}\frac{d(x,y)^2}{2} - \frac{d(\overline{x},y)^2}{2} &\geq \min\Big[\frac{d(x,y_0)^2}{2} - \frac{d(\overline{x},y_0)^2}{2},\ \frac{d(x,y_1)^2}{2} - \frac{d(\overline{x},y_1)^2}{2}\Big] + \lambda\,\varepsilon^2 \\ &\geq \min\Big[\psi(\overline{x}) - \psi(x),\ \psi(\overline{x}) - \psi(x) - C\,d(y_0, y_c)\Big] + \lambda\,\varepsilon^2 \\ &= \psi(\overline{x}) - \psi(x) - C\,d(y_0,y_c) + \lambda\,\varepsilon^2, \qquad (7.4)\end{aligned}$$

where λ is a positive number and C is a constant depending only on the diameter of M. (Here we have used the Lipschitz continuity of the squared distance.)

Comparison with (7.3) shows that

$$d(y_0, y_c) \geq \frac{\lambda\, \varepsilon^2}{C},$$

which is the desired lower bound (with ε of the order of a/A). □

We conclude by noting that in Sections 6 and 7 we never used the nonfocality assumption, but just the positivity of $\delta(M)$ as defined in (1.5). This combined with the remark at the end of Section 4 allows to establish Theorem 1.8(b).

8. Hölder continuity of optimal transport

In this section we shall use the stay-away property to derive $C^{1,\alpha}$ regularity estimates on ψ, again with the same method as in [20]. For simplicity we shall only consider the case when the densities f and g are bounded from above and below, respectively.

Theorem 8.1. *Let M be a Riemannian manifold satisfying* $\mathrm{MTW}(K_0)$ *for some $K_0 > 0$, and whose injectivity domains are all convex. Let μ and ν on M, with $\mu \leq A\,\mathrm{vol}\,,\ \nu \geq a\,\mathrm{vol}\ \ (A, a > 0)$. Assume that the optimal transport map $T = \exp(\nabla\psi)$ between μ and ν, for the cost d^2, satisfies* $\inf_x d(T(x), \mathrm{cut}(x)) \geq \sigma > 0$. *Then*

$$\psi \in C^{1,\alpha}(M), \qquad \alpha = \frac{1}{4n-1};$$

moreover, $\|\psi\|_{C^{1,\alpha}} \leq C(M, A, a, \sigma)$.

Remark 8.2. A recent result by Liu [19] (published when the present work was almost completed) shows that the optimal exponent is $\alpha = 1/(2n-1)$.

By putting together Theorems 1.7, 6.1 and 7.1, we immediately obtain the following corollaries:

Corollary 8.3. *Let M be a Riemannian manifold whose cut locus is nonfocal, satisfying* $\mathrm{MTW}(K_0)$ *for some $K_0 > 0$. Let μ, ν be two probability measures on M with $\mu \leq A\,\mathrm{vol}\,,\ \nu \geq a\,\mathrm{vol}\ \ (A, a > 0)$, and let $T = \exp(\nabla\psi)$ be the optimal transport map, where ψ is $d^2/2$-convex. Then $\psi \in C^{1,\alpha}$ and $\|\psi\|_{C^{1,\alpha}} \leq C(M, A, a)$.*

Corollary 8.4. *Let M be a uniformly regular Riemannian manifold satisfying $\delta(M) > 0$. Let μ, ν be two probability measures on M with $\mu \leq A\,\mathrm{vol}\,,\ \nu \geq a\,\mathrm{vol}\ \ (A, a > 0)$, and let $T = \exp(\nabla\psi)$ be the optimal transport map, where ψ is $d^2/2$-convex. Then $\psi \in C^{1,\alpha}$ and $\|\psi\|_{C^{1,\alpha}} \leq C(M, A, a)$.*

Corollary 8.3 has the advantage to be stable under C^4 perturbations of the metric, while Corollary 8.4 has the advantage to be more general and to cover the case of the sphere.

The proof of Theorem 8.1 is quite close to the proof of similar results in [20] or [16, Appendix E], with just simplifications in the presentation.

Proof of Theorem 8.1. First, since $T(x)$ stays a positive distance away from $\mathrm{cut}(x)$, we can do as if the distance function is uniformly smooth, and in particular replace $\mathrm{MTW}(K_0)$ by $\mathrm{MTW}(K_0, C_0)$, just as in Lemma 2.3.

Let $\widetilde{x}$ be an arbitrary point in M, and let $\widetilde{y} = T(\widetilde{x})$. If $\overline{x}$ is close enough to $\widetilde{x}$ we have $\mathrm{dist}(\overline{x}, \mathrm{cut}(\widetilde{y})) \geq \sigma/2$; in particular $\widetilde{v} = (\exp_{\overline{x}})^{-1}(\widetilde{y})$ is well-defined.

Let then $\overline{y} = T(\overline{x})$ and $\overline{v} = (\exp_{\overline{x}})^{-1}(\overline{y})$; also $\overline{y}$ lies a positive distance away from $\mathrm{cut}(\overline{x})$. To prove Theorem 8.1 it suffices to get an estimate like

$$|\overline{v} - \widetilde{v}| = O\big(d(\overline{x}, \widetilde{x})^\alpha\big); \tag{8.1}$$

then the Hölder regularity of the optimal transport map $T = \exp \nabla\psi$ will follow since (by nonnegative curvature) $d(\overline{y}, \widetilde{y}) \leq |\overline{v} - \widetilde{v}|$; and the regularity of $\nabla\psi$ will follow as well because the map $(x, y) \longmapsto (x, (\exp_x)^{-1}(y))$ is smooth in $\{d(y, \mathrm{cut}(x)) \geq \sigma\}$. (In fact, since we are away from the cut locus, there is a constant C, possibly depending on M and σ, such that $d(\overline{y}, \widetilde{y}) \leq |\overline{v} - \widetilde{v}| \leq C\, d(\overline{y}, \widetilde{y})$; so (8.1) is equivalent to the desired Hölder continuity property.)

Let us proceed to the proof of (8.1). As in Section 6, we start by noting that

$$\begin{cases} \psi(\overline{x}) + \dfrac{d(\overline{x}, \overline{y})^2}{2} = \inf\limits_{x \in M} \left(\psi(x) + \dfrac{d(x, \overline{y})^2}{2}\right) \\[2ex] \psi(\widetilde{x}) + \dfrac{d(\widetilde{x}, \widetilde{y})^2}{2} = \inf\limits_{x \in M} \left(\psi(x) + \dfrac{d(x, \widetilde{y})^2}{2}\right). \end{cases} \tag{8.2}$$

Next we define the sets D_ε and Y_ε as in Section 6, with $v_0 = \overline{v}$ and $v_1 = \widetilde{v}$. Since v_0 and v_1 lie away from $\mathrm{TCL}(\overline{x})$, D_ε also lies uniformly away from $\mathrm{TCL}(\overline{x})$, so $\mathrm{vol}\,[Y_\varepsilon]$ is comparable to $\mathrm{vol}\,[D_\varepsilon]$. For $d(\overline{x}, x) \geq \varepsilon$, D_ε contains a parallelepiped with one side of length $\Omega(|\overline{v} - \widetilde{v}|)$, and all other sides of length $\Omega(\varepsilon\, |\overline{v} - \widetilde{v}|^2)$ (since an acceleration of order $|q - \overline{q}| \geq \varepsilon$ is allowed). So

$$\mathrm{vol}\,[Y_\varepsilon] = \Omega(\mathrm{vol}\,[D_\varepsilon]) = \Omega\big(\varepsilon^{n-1}\, |\overline{v} - \widetilde{v}|^{2(n-1)+1}\big). \tag{8.3}$$

Let $y \in Y_\varepsilon$, and $x \in M \setminus B_\varepsilon(\overline{x})$. If $T(x) = y$, then $d(y, \mathrm{cut}(x)) \geq \sigma$. By choosing $|\overline{v} - \widetilde{v}|$ much smaller than σ and ε small enough, we can make sure that $\mathrm{cut}(x)$ does not meet Y_ε, so $q = (\exp_y)^{-1}(x)$ is well-defined for all $y \in Y_\varepsilon$, and the convexity of

$\mathrm{I}(y)$ implies $[\overline{q}, q] \subset \mathrm{I}(y)$, with $\overline{q} = (\exp_y)^{-1}(\overline{x})$. Thus we can apply Theorem 3.1. Combining this with (8.2), we find that, for some $\beta > 0$,

$$\begin{aligned} \frac{d(x,y)^2}{2} - \frac{d(\overline{x},y)^2}{2} &\geq \min\Big[\frac{d(x,\overline{y})^2}{2} - \frac{d(\overline{x},\overline{y})^2}{2},\ \frac{d(x,\widetilde{y})^2}{2} - \frac{d(\overline{x},\widetilde{y})^2}{2}\Big] + \beta\,\varepsilon^2\,|\overline{v}-\widetilde{v}|^2 \\ &\geq \min\Big(\psi(\overline{x}) - \psi(x),\ \psi(\widetilde{x}) - \psi(x) + \frac{d(\widetilde{x},\widetilde{y})^2}{2} - \frac{d(\overline{x},\widetilde{y})^2}{2}\Big) + \beta\,\varepsilon^2\,|\overline{v}-\widetilde{v}|^2. \end{aligned}$$

The latter expression is strictly greater than $\psi(\overline{x}) - \psi(x)$ as soon as

$$\psi(\widetilde{x}) - \psi(\overline{x}) + \frac{d(\widetilde{x},\widetilde{y})^2}{2} - \frac{d(\overline{x},\widetilde{y})^2}{2} + \beta\,\varepsilon^2\,|\overline{v}-\widetilde{v}|^2 > 0. \tag{8.4}$$

When the latter condition is satisfied, (x, y) cannot belong to the support of the transport plan; so all the mass in Y_ε has to come from $B_\varepsilon(\overline{x})$, and the bounds from above and below on μ and ν imply

$$\mathrm{vol}\,[Y_\varepsilon] \leq \frac{A}{a}\,\mathrm{vol}\,[B_\varepsilon(\overline{x})]. \tag{8.5}$$

Inequality (8.5) combined with (8.3) implies

$$\varepsilon^{n-1}\,|\overline{v}-\widetilde{v}|^{2(n-1)+1} \leq C'\,\varepsilon^n,$$

for some constant $C' > 0$, hence

$$|\overline{v}-\widetilde{v}|^{2n-1} = O(\varepsilon). \tag{8.6}$$

Condition (8.4) takes the form $h(\widetilde{x}) - h(\overline{x}) + \beta\,\varepsilon^2\,|\overline{v}-\widetilde{v}|^2 > 0$, where $h(x) = \psi(x) + d(x,\widetilde{y})^2/2$. The function ψ is semiconvex and $d^2(\cdot,\widetilde{y})/2$ is smooth around $\overline{x}$, if $\overline{x}$ is close enough to $\widetilde{x}$. (That is, if $d(\overline{x},\widetilde{x}) \leq \eta$, where η only depends on M and on a lower bound on $d(T(\widetilde{x}), \mathrm{cut}(\widetilde{x}))$.) Writing $\widetilde{x} = \exp_{\overline{x}}(\delta x)$ and noting that $\nabla\psi(\overline{x}) = \overline{v}$, we deduce

$$\begin{aligned} h(\widetilde{x}) &\geq h(\overline{x}) + \nabla h(\overline{x})\cdot(\delta x) - O\big(d(\overline{x},\widetilde{x})^2\big) \\ &= h(\overline{x}) + (\overline{v}-\widetilde{v})\cdot(\delta x) - O\big(d(\overline{x},\widetilde{x})^2\big) \\ &\geq h(\overline{x}) - |\overline{v}-\widetilde{v}|\,d(\overline{x},\widetilde{x}) - O\big(d(\overline{x},\widetilde{x})^2\big). \end{aligned}$$

If $K > 0$ is well-chosen then either the above expression is bounded below by $h(\overline{x}) - 2|\overline{v}-\widetilde{v}|\,d(\overline{x},\widetilde{x})$ or $|\overline{v}-\widetilde{v}| \leq K\,d(\overline{x},\widetilde{x})$. In the latter case our job is finished; in the former case we conclude that (8.4) holds true as soon as

$$|\overline{v}-\widetilde{v}|\,d(\overline{x},\widetilde{x}) \leq \gamma\,\varepsilon^2\,|\overline{v}-\widetilde{v}|^2$$

for some $\gamma > 0$ small enough, or equivalently

$$d(\overline{x}, \widetilde{x}) \leq \gamma\, \varepsilon^2\, |\overline{v} - \widetilde{v}|. \tag{8.7}$$

Also ε should be smaller than some $\overline{\varepsilon}$ for the whole argument to work.

If $d(\overline{x}, \widetilde{x}) \geq \gamma\, \overline{\varepsilon}^2\, |\overline{v} - \widetilde{v}|$ then we are done. Otherwise, choose $\varepsilon^2 = d(\overline{x}, \widetilde{x})/(\gamma\, |\overline{v} - \widetilde{v}|)$ and plug this in (8.6); it follows

$$|\overline{v} - \widetilde{v}|^{4n-1} = O\big(d(\overline{x}, \widetilde{x})\big),$$

which is the desired conclusion. □

Remark 8.5. The argument presented above is robust enough to allow for perturbations of various kind. For instance, it is adapted in [31] to prove the stability of the Hölder continuity at a given "mesoscopic" scale under C^2 (not C^4) perturbations of the metric.

9. Final comments and open problems

In this paper, we have mostly worked under a technical assumption of nonfocality (or at least, that no cut point is purely focal). One may however conjecture that the property of convexity of injectivity domains holds under $\mathrm{MTW}(K_0)$ also in presence of focalization; and maybe even under $\mathrm{MTW}(0)$. (The example of the flat torus shows that $\mathrm{MTW}(0)$ implies at best convexity of injectivity domains, not uniform convexity.)

It is also interesting to investigate the stability of this convexity property under small perturbations of the metric. The tangent cut locus is stable under C^2 perturbations of the metric, so it is natural to conjecture that the uniform convexity of tangent interior loci is preserved under C^4 perturbations. This conjecture and the previous one would be compatible with a third plausible conjecture, namely that $\mathrm{MTW}(K_0)$ is stable under C^4 perturbations of the metric, even in the focal case.

Similarly, one could conjecture that $\mathrm{MTW}(K_0)$ automatically implies the uniform regularity property of Definition 4.1; or even that $\mathrm{MTW}(0)$ implies the regularity property, which would be the same as in Definition 4.1 but with $\kappa = \lambda = \varepsilon_0 = 0$.

When we first formulated them, all these conjectures looked like distant dreams requiring new ideas. At the time of writing, they have been made a little less distant by a new work of Figalli and Rifford [10] who proved the stability of the Ma–Trudinger–Wang condition on $\mathbb{S}^2$, and deduced from this the convexity of injectivity domains of C^4 perturbations of $\mathbb{S}^2$. One of their main ideas was to extend the condition up to the *focal* locus, whose variation under deformations of the metric is much better understood than the variation of the cut locus. In view of this work, we believe that Theorem 1.7 is the first of a series of results relating Ma–Trudinger–Wang type curvature conditions with the convexity of injectivity domains.

Appendix A. Uniform convexity

In all this appendix, E is an n-dimensional Euclidean space with scalar product $\langle \cdot, \cdot \rangle$ and norm $|\cdot|$.

There is a natural notion of uniformly convex function $E \to \mathbb{R}$:

Definition A.1. A function $\Phi : E \to \mathbb{R}$ is said to be κ-uniformly convex if

$$\forall x, y \in E, \quad \forall t \in [0,1], \quad \Phi\big((1-t)x + ty\big) \leq (1-t)\,\Phi(x) + t\,\Phi(y) - \frac{\kappa\, t(1-t)}{2}\,|x-y|^2.$$

This notion is local and equivalent, modulo smoothness issues, to $\nabla^2\Phi \geq \kappa\,\mathrm{Id}$.

On the other hand, there are several possible definitions for a uniformly convex set. One of them is the following, where $d(x, A) = \inf\{|x-a|;\ a \in A\}$:

Definition A.2. A set $C \subset E$ is said to be κ-uniformly convex if

$$\text{(A.1)} \qquad \forall x, y \in C, \quad \forall t \in [0,1], \qquad d\big((1-t)x + ty,\, \partial C\big) \geq \frac{\kappa\, t(1-t)}{2}\,|x-y|^2.$$

Remark A.3. If C is κ-uniformly convex and $x, y \in C$ satisfy $|x-y| = \mathrm{diam}(C)$, then

$$\frac{|x-y|}{2} \geq d\left(\frac{x+y}{2},\, \partial C\right) \geq \frac{\kappa}{8}\,|x-y|^2,$$

so $\mathrm{diam}(C) = |x-y| \leq 4\,\kappa^{-1}$. (The constant is not optimal.)

If x, y approach a point z in ∂C and $(y-x)/|y-x|$ converges to a tangent vector ξ, the "global" inequality (A.1) reduces to the "local" inequality

$$\text{(A.2)} \qquad \big\langle \mathrm{I\!I}_C(x)\cdot\xi, \xi\big\rangle \geq \kappa\,|\xi|^2,$$

where $\mathrm{I\!I}_C(x)$ (second fundamental form at x) is the "covariant" gradient $\nabla_x N$ of the exterior normal vector N to C at x (extend N into a smooth function and restrict $\nabla_x N$ to $T_xC \times T_xC$).

A reformulation of (A.2) is

$$\text{(A.3)} \quad \text{For any } C^2 \text{ path } \gamma : (-\tau, \tau) \to \overline{C} \text{ with } \gamma(0) = x, \quad -\langle \ddot\gamma(0),\, N\rangle_x \geq \kappa\,|\dot\gamma(0)|^2.$$

If ∂C is C^2 at x, (A.2) and (A.3) are equivalent; but (A.3) has the advantage to make sense even when the boundary is not smooth, but only, say, Lipschitz. Note in particular that if C presents an inner spike at x then (A.3) is always violated. If on the contrary C presents an outer spike at x then (A.3) is automatically satisfied since there is no admissible path γ.

The next proposition relates the various notions of uniform convexity described above:

Proposition A.4. *Let C be a bounded connected closed subset of E with Lipschitz boundary. Then for any $\kappa > 0$ the following three properties are equivalent:*

(i) C is κ-uniformly convex;

(ii) $\forall x \in \partial C,\ \ \mathrm{I\!I}_C(x) \geq \kappa$ in the sense of (A.3);

(iii) $C = \{x;\ \Phi(x) \leq 0\}$ for some $\Phi : E \to \mathbb{R}$ which is λ-uniformly convex on E and L-Lipschitz on C, with $\lambda/L = \kappa$.

Proof of Proposition A.4. The proof is based on standard arguments; we only sketch it. (i) $\Rightarrow$ (ii) was explained above.

(ii) $\Longrightarrow$ (iii): Without loss of generality we look for a 1-Lipschitz function. If C is the ball $B(c, R)$ of radius R and center c, then $\kappa = 1/R$ and $\Phi_{c,R}(x) = |x-c|^2/(2R) - R/2$ will do. In the general case, for any $a \in \partial C$, there is a ball $B(c_a, R)$ of radius $R = 1/\kappa$ such that C is contained in B_a and $a \in \partial C \cap \partial B$. Then $\Phi = \sup \Phi_{c_a,R}$ does the job.

(iii) $\Rightarrow$ (i): if $\Phi(x) \leq 0$, $\Phi(y) \leq 0$, $\Phi(z) = 0$, then, with $L = \|\Phi\|_{\mathrm{Lip}(C)}$,

$$\begin{aligned}\big|(1-t)x + ty - z\big| &\geq -\frac{1}{L}\,\Phi\big((1-t)x + ty\big)\\ &\geq \frac{\lambda\, t(1-t)}{L}|x-y|^2 - \Big[(1-t)\Phi(x) + t\Phi(y)\Big] \geq \frac{\lambda\, t(1-t)}{L}|x-y|^2.\end{aligned}$$

□

If the set C is *starshaped* with respect to 0 and has nonempty interior, there is yet another natural notion of uniform convexity, which is closer to the usual notion of "uniformly convex norm", maybe more natural in the study of the cut locus. First define

$$\mathrm{inj}(C) = \sup\big\{r > 0;\ B(0,r) \subset C\big\},$$

where $B(0,r)$ is the open ball of radius r centered at 0. For any ξ in the unit sphere S of E, define

$$t_C(\xi) = \sup\,\{t \geq 0;\ t\xi \in C\} > 0$$

(this is the time at which ones gets out of C if ones travels at constant speed in direction ξ, starting from 0). Then for any $x \in C \setminus \{0\}$, define

$$t_C(0,x) = t_C\left(\frac{x}{|x|}\right) - |x| \geq 0$$

(this is the time one has to wait before getting out of C if one has been travelling from 0 and currently stands at x). Further set $t_C(0,0) = \mathrm{inj}(C)$.

Proposition A.5. *Let C be a bounded closed subset of E with nonempty interior and Lipschitz boundary, starshaped with respect to 0. Then the following two properties are equivalent:*

(i) C is κ_0-uniformly convex for some $\kappa_0 > 0$;

(ii) There is $\kappa > 0$ such that

$$\forall x, y \in C, \quad t_C\big(0,\, (1-t)x + ty\big) \geq \frac{\kappa\, t(1-t)}{2}\, |x-y|^2.$$

More precisely, the implication (i) ⇒ (ii) holds with $\kappa = \kappa_0$; and (ii) ⇒ (i) with $\kappa_0 = \kappa_0(n, \kappa, \|\partial C\|_{\mathrm{Lip}}, \mathrm{diam}(C), \mathrm{inj}(C))$.

Proof of Proposition A.5. It suffices to prove

$$d(x, \partial C) \leq t_C(0,x) \leq \frac{\kappa}{\kappa_0}\, d(x, \partial C). \tag{A.4}$$

The inequality on the left is obvious since $t_C(0,x) = |x-z|$, where z is the intersection of ∂C with the half-line starting from 0 and going through x. To prove the inequality on the right of (A.4), first note that, since ∂C is Lipschitz, there is a neighborhood U of z and an open angular sector Σ with apex z, such that $U \cap \Sigma$ is entirely contained in C. Since C is starshaped, the cone with apex 0 and basis $\Sigma \cap U$ is contained in C. Since also the ball $B(0, \mathrm{inj}(C))$ is contained in C, we deduce that C contains an open cone with apex z and axis $[0,z]$, whose angle at z is bounded below. It follows that $d(x, \partial C)/d(x,z)$ is bounded from below. □

Appendix B. Semiconvexity

Let U be an open subset of $\mathbb{R}^n$. Following common terminology, we say that a function $\Phi : U \to \mathbb{R}$ is semiconvex if it is locally the sum of a smooth and a convex function. More general notions of semiconvexity are reviewed in [30, Chapter 10].

We now introduce the notion of semiconvex *set*:

Definition B.1 (semiconvex set)**.** A set $S \subset \mathbb{R}^n$ is said semiconvex if it is locally defined by an equation $\{\Phi \leq a\}$, where Φ is a semiconvex function without critical point, and $a \in \mathbb{R}$.

Note that Φ is not necessarily smooth, so the noncriticality should be defined in a weak formulation: the subgradient of Φ (as defined e.g. in [30, Definition 10.5]) should stay a positive distance away from 0. This noncriticality assumption allows to locally change Φ into a genuinely convex function by a local change of coordinates. In particular, a semiconvex set S enjoys the same regularity properties as a convex set.

One may also give a weak definition of second fundamental form inequalities at the boundary of S, as in (A.3), except that now the lower bound κ may be negative.

Appendix C. Differential structure of the tangent cut locus

C.1. **Generalities.** Let M be a smooth, compact, connected n-dimensional Riemannian manifold equipped with a Riemannian metric $\langle \cdot, \cdot \rangle$ and a geodesic distance d. Recall that $d(x,y)^2 = \inf\{\int_0^1 |\dot\gamma(t)|^2\,dt;\ \gamma \in \mathrm{Lip}([0,1];M),\ \gamma(0)=x,\ \gamma(1)=y\}$. Let $\Gamma \subset \mathrm{Lip}([0,1];M)$ be the set of constant-speed minimizing geodesics $[0,1] \to M$ (equivalently, action-minimizing curves, where the action is $\int |\dot\gamma|^2$).

If x is given in M, we define

$$\mathcal{V}(x) = \Big\{\dot\gamma(0);\ \gamma \in \Gamma;\ \gamma(0) = x\Big\} \subset T_xM.$$

This is the set of all optimal velocities at x, that is, velocities which are used to go from x to some $y \in M$ along a path of least action (that is, a constant-speed minimizing geodesic). Let further $\mathcal{V} = \cup(\{x\} \times \mathcal{V}(x))$. If $\mathcal{V} \ni (x_k, v_k) \to (x,v)$ then up to extraction of a subsequence the minimizing geodesics $(\exp_{x_k}(tv_k))_{0\leq t\leq 1}$ converge to some minimizing geodesic, necessarily $(\exp_x(tv))_{0\leq t\leq 1}$, so $(x,v) \in \mathcal{V}$; in other words, $\mathcal{V}$ is closed (and in particular $\mathcal{V}(x)$ is closed for every x). Also, since the restriction of a minimizing geodesic is still minimizing, $\mathcal{V}(x)$ is starshaped with respect to 0.

Let $(x,v) \in \mathcal{V}$. If the geodesic joining x to $y = \exp_x v$ is unique, there are neighborhoods U of x, V of y and W of v, such that any minimizing geodesic starting in U and arriving in V has an initial velocity in W. (Otherwise there are $x_k \to x$, $y_k \to y$, $v_k \notin W$ such that $(\exp_{x_k}(tv_k))$ is minimizing and $y_k = \exp_{x_k}(v_k)$; passing to the limit as $k \to \infty$ gives a contradiction.) If in addition $d_v \exp_x$ is invertible, the implicit function theorem shows that the equation $y = \exp_x v$ is uniquely solvable in v for $(x,y) \in U \times V$; so (x,v) lies in the *interior* of $\mathcal{V}$.

Conversely, if (x,v) belongs to the interior of $\mathcal{V}$, the geodesic $\gamma = (\exp_x(tv))_{0\le t\le 1}$ can be extended to times $t \in (1, 1+\varepsilon)$ for some $\varepsilon > 0$; then by a well-known but fundamental result (see e.g. [12, Corollary 3.77]), γ is the unique minimizing geodesic between its endpoints and $d_v \exp_x$ is invertible.

This discussion shows that (with the notation of Definition 1.2) $\mathcal{V}$ coincides with $\overline{\mathrm{I}}(M)$, its interior with $\mathrm{I}(M)$, and its boundary with $\mathrm{TCL}(M)$. (And also $\mathcal{V}(x) = \overline{\mathrm{I}}(x)$ for each x, etc.) Moreover, it provides the following classical decomposition of the tangent cut locus:

Proposition C.1. *If M is a compact Riemannian manifold, then*

$$\mathrm{TCL}(M) = \Sigma \cup J,$$

where $\Sigma = \bigcup(\{x\} \times \Sigma_x)$, $J = \bigcup(\{x\} \times J_x)$, *and*

$$\Sigma_x = \Big\{ v \in \mathrm{TCL}(x);\ \exists w \in \mathrm{TCL}(x);\ w \neq v;\ \exp_x v = \exp_x w \Big\};$$

$$J_x = \Big\{ v \in \mathrm{TCL}(x);\ d_v \exp_x \textit{ is singular} \Big\}.$$

Moreover, $\mathrm{cut}(x) = \exp_x(\mathrm{TCL}(x))$ *and* $\mathrm{TCL}(x) = (\exp_x)^{-1}(\mathrm{cut}(x)) \cap \overline{\mathrm{I}}(x)$.

In the sequel, $(\exp_x)^{-1}(y)$ will always stand for $(\exp_x)^{-1}(y) \cap \overline{\mathrm{I}}(x)$, or for the element constituting this set if it is a singleton.

Remark C.2. A classical theorem by Bishop [1] states that Σ_x is dense in $\mathrm{TCL}(x)$; but this does not prevent J_x from being "large", as the example of the sphere shows. Moreover, if M is even-dimensional, simply connected and positively curved, then there is $x \in M$ such that $J_x \neq \emptyset$. (See [33, Section 6] for a short list of conditions guaranteeing the nonemptyness of J.)

Remark C.3. When one is interested in the regularity of the tangent cut locus, the simplest part usually consists in "purely cut" velocities, that is $\Sigma \setminus J$; while the most tricky consists in "purely focal" velocities, that is $J \setminus \Sigma$.

One can show (for instance by pushing the method in [23]) that $\mathrm{TCL}(x)$ is included in a countable union of smooth $(n-1)$-dimensional Lipschitz manifolds, and this union is locally finite on $\Sigma \setminus J$. However, some proofs in the present paper need a more precise description. The following issues (some of them very simple, some other very tricky) are of interest:

(1) *Relation with the smoothness of distance.* If $v \in \mathrm{TCL}(x)$ and $y = \exp_x v$, it is known that

$$\begin{cases} v \in \Sigma_x \iff \dfrac{d(\,\cdot\,,y)^2}{2} \text{ is not differentiable at } x; \\ \\ v \in J_x \setminus \Sigma_x \iff \quad \nabla_x \dfrac{d(\,\cdot\,,y)^2}{2} = v \ \text{ and } \dfrac{d(\,\cdot\,,y)^2}{2} \text{ is not semiconvex at } x; \end{cases}$$

the latter statement is proven in [5, Proposition 2.5].

(2) *Lipschitz regularity.* The question is whether $\mathrm{TCL}(M)$ is Lipschitz continuous, or equivalently whether the function $(x,\xi) \to t_C(x,\xi)$ (cut time in the direction ξ), defined on the unit tangent bundle UM, is Lipschitz continuous. The answer is affirmative according to the results of Itoh and Tanaka [13], Li and Nirenberg [18].

(3) *Upper semicontinuous differentiability from the exterior.* This property would be a way to express the fact that $\mathrm{I}(M)$ can have only outer spikes, not inner ones. Here is a precise definition:

Definition C.4 (upper semicontinuous differentiability from the exterior)**.** Let $B = \bigcup(\{x\} \times B_x)$ be a fibered bundle with projection π, let O be an open subset of B, with fibers $O_x = \pi^{-1}(x)$. Whenever $p \in \partial O_x$, we define the exterior tangent cone to O at p by

$$T_p^{\mathrm{ext}} O_x = \left\{ \lim_{\ell \to \infty} \frac{p_\ell - p}{t_\ell}, \qquad p_\ell \notin O_x,\ t_\ell > 0,\ p_\ell \to p \right\}.$$

We say that O is *upper semicontinuously differentiable from the exterior* (u.s.c.d.e.) if, for any sequence $(x_k, p_k) \to (x,p)$ in B, with $p_k \in \partial O_{x_k}$,

$$T_p^{\mathrm{ext}} O_x \supset \limsup_{k \to \infty} T_{p_k}^{\mathrm{ext}} O_{x_k}.$$

It is not known whether $\mathrm{I}(M)$ is u.s.c.d.e. in general, although this seems to be a natural conjecture.

(4) *Semiconcavity of the cut time.* A natural question is whether the cut time t_C is semiconcave on UM. If true, this property implies the u.s.c.d.e. property via classical stability properties of the superdifferential of (semi)concave functions. A related open problem evoked by Itoh and Tanaka [13, Problem 3.4] is whether $\mathrm{TCL}(x)$ is an *Alexandrov space* of curvature bounded below.

(5) *Study of differentiability points of* TCL*.* The issue is first to get sufficient conditions for the existence of a tangent (or an osculating circle) at $v \in \mathrm{TCL}(x)$, and to determine the normal vector. In the nonfocal case, we shall see that $v \in \mathrm{TCL}(x)$ is

a differentiability point (that is, a point where $\mathrm{TCL}(x)$ admits a tangent hyperplane) if and only if v is in competition with exactly one other optimal velocity.

Questions 3, 4 and 5 seem to be still open for general manifolds. In the sequel of this Appendix, we shall see that all these questions can be answered precisely in the nonfocal case, that is when $J = \emptyset$; and that some of these answers still hold under the weaker assumption that one is away from the purely focal tangent cut locus (in particular $J \setminus \Sigma = \emptyset$). For the convenience of the reader, we shall provide complete proofs.

C.2. **The nonfocal case.** In this subsection we shall assume that the cut locus of M is nonfocal, i.e. $J = \emptyset$ in the notation of Proposition C.1. In the sequel, a "curved half-space" is a set of the form $\{\Phi < 0\}$, where Φ is smooth and $d\Phi$ is invertible on $\{\Phi = 0\}$. (The function Φ might be defined only in the neighborhood of a given point.)

Proposition C.5. *If M is a compact Riemannian manifold with nonfocal cut locus, then*

(a) There is $\delta > 0$ such that for any $x \in M$ and any $v, w \in \mathrm{TCL}(x)$,

$$\big[v \neq w, \quad \exp_x v = \exp_x w\big] \implies \quad |v - w| > \delta. \tag{C.1}$$

In particular, $N(x, v) := \#(\exp_x)^{-1}(\exp_x v)$ is an upper semicontinuous function on $\mathrm{TCL}(M)$, bounded below by 2 and bounded above by some finite number.

(b) For any $x \in M$, $\mathrm{cut}(x)$ is exactly the set of points $y \in M$ such that $d(\,\cdot\,, y)^2$ fails to be differentiable at x. Moreover, for any $k \in \mathbb{N}$, $(x, y) \to (\exp_x)^{-1}(y)$ and $(x, y) \to d(x, y)^2$ are uniformly C^k in the open set $(M \times M) \setminus \mathrm{cut}(M)$; and $(x, v) \to (d_v \exp_x)^{-1}$ is uniformly C^k in $\mathrm{I}(M)$.

(c) Let $(x_0, v_0) \in \mathrm{TCL}(M)$; then for x close to x_0, the open set $\mathrm{I}(x)$ is locally around v_0, the intersection of $N(x_0, v_0) - 1$ transversal curved half-spaces varying smoothly with x. As a consequence,

- *$\mathrm{I}(M)$ is semiconvex (and in particular satisfies the u.s.c.d.e. property).*
- *v is a differentiability point of $\mathrm{TCL}(x)$ if and only if $N(x, v) = 2$.*
- *nondifferentiability points of $\mathrm{TCL}(x)$ are included in finitely many smooth manifolds of dimension at most $n - 2$.*

To summarize, the picture of the tangent cut locus in the nonfocal case is very simple: $\mathrm{TCL}(M)$ is smooth away from "multiple points" where three or more families of minimizing geodesics meet; and at such points there are outer spikes in certain directions.

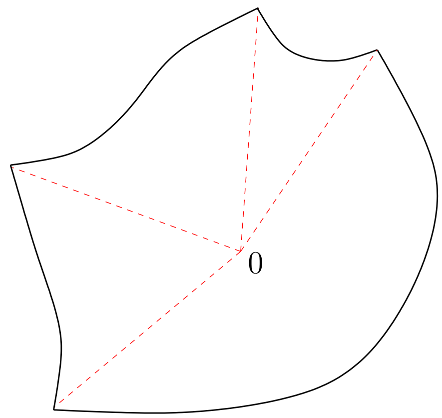

FIGURE 5. Possible shape of TCL(x) for a nonfocal manifold. The four velocities connected in dotted lines correspond to the same point y in M.

Proof of Proposition C.5. From the nonfocality assumption, $\mathrm{cut}(x) = \exp_x(\Sigma_x)$: in particular $v \in \overline{\mathrm{I}}(x)$ belongs to TCL(x), if and only if there is $w \in \overline{\mathrm{I}}(x)$, $w \neq v$, such that $\exp_x w = \exp_x v$.

(a) Both the tangent cut locus TCL(M) and the tangent focal locus TFL(M) are closed, so the assumption $\mathrm{TCL}(M) \cap \mathrm{TFL}(M) = \emptyset$ implies the existence of $\eta > 0$ such that

$$\mathrm{dist}\big(\mathrm{I}(x), \mathrm{TFL}(x)\big) = \mathrm{dist}\big(\mathrm{TCL}(x), \mathrm{TFL}(x)\big) \geq \eta > 0.$$

Also, $\det(d_v \exp_x)$ does not vanish on $\overline{\mathrm{I}}(M)$, so there is $\lambda > 0$ such that $\det(d_v \exp_x) \geq \lambda$ for all $(x, v) \in \mathrm{I}(M)$.

Next we claim that there is $r = r(M) > 0$ such that for any $x \in M$ and any $v_0 \in \mathrm{I}(x)$, the exponential map $\exp_x$ is one-to-one on a ball $B_r(v_0)$. This can be seen either by using a quantitative version of the implicit function theorem (see for instance [25]) or by a contradiction argument: If the claim were false, there would be $x_k \in M$ and $v_k, u_k, w_k \in T_{x_k}M$ such that $|v_k - u_k| \leq 1/k$, $|w_k - u_k| \leq 1/k$ and $\exp_{x_k} u_k = \exp_{x_k} w_k$; extracting a subsequence if necessary, we would have $x_k \to x$, $v_k \to v$, and the identity $\exp_{x_k} u_k = \exp_{x_k} w_k$ contradicts the fact that exp is a diffeomorphism near (x, v).

Then the statement about $N(x, v)$ being upper semicontinuous is a consequence of the stability of minimizing geodesics.

(b) If $v \in \overline{\mathrm{I}}(M)$ then v is a supergradient of $d(\,\cdot\,, \exp_x v)^2/2$ at x. So $d(\,\cdot\,, \exp_x v)^2/2$ cannot be dfferentiable at x if $N(x, v) > 1$. Conversely, if $N(x, v) = 1$ then $\nabla_x d(\cdot, \exp_x v)^2/2 = -v$. This proves the claim on the differentiability. Next, the claim about the regularity of $(d_v \exp_x)^{-1}$ follows by the implicit function theorem again. Finally, if $(x, y) \in M \times M \setminus \mathrm{cut}(M)$, and $\gamma = (\exp_x tv)_{0 \le t \le 1}$, then γ is the unique geodesic joining x to y, so there are neighborhoods U of x, V of y and W of v such that any geodesic curve from U to V is a perturbation of γ. So all derivatives of d^2 at x, y are expressed in terms of derivatives of $d \exp$ and $(d \exp)^{-1}$, which are all uniformly bounded.

(c) Let $F(x, v) = \exp_x v$. Let $x_0 \in M$, $v_0 \in \mathrm{TCL}(x_0)$, $y_0 = F(x_0, v_0)$, and let $v_1, \ldots, v_m$ be all solutions of $F(x, v_i) = y_0$, $v_i \neq v_0$. (So $m = N(x_0, v_0) - 1$.) By the implicit function theorem, F is locally invertible in a neighborhood B_i of each (x_0, v_i), so for any (x, w) in a neighborhood of (x_0, v_0) there are exactly m velocities $w_1(x, v), \ldots, w_m(x, v)$ in $T_x M$ such that $F(x, w_i) = F(x, w)$ and $(x, w_i) \in B_i$. These equations define local inverses $w_i = w_i(x, F(x, w))$. Then the equation of $\mathrm{I}(M)$ around (x_0, v_0) takes the form

$$E(x, w) < \min_{1 \le j \le m} E\Big(x, w_j\big(x, F(x, w)\big)\Big), \tag{C.2}$$

where $E(x, v) = (1/2) \int_0^1 |(d/dt) \exp_x(tv)|^2 \, dt = |v|^2/2$.

Let $\gamma_j(z, t) = \exp_x(t w_j(x, z))$ denote the unique geodesic joining x to z with initial velocity close to v_j.

For simplicity we shall now drop the x-dependence from the notation, keeping in mind that all functions involved depend smoothly on x. Let $\mathcal{E}_j(z) = E(w_j(z)) = |w_j(z)|^2/2$; then (C.2) can be reformulated as

$$\mathcal{E}_0(F(w)) - \mathcal{E}_j(F(w)) < 0 \qquad \forall j \in \{1, \ldots, m\}. \tag{C.3}$$

Since $\mathrm{grad}_z(\mathcal{E}_0 - \mathcal{E}_j) = \dot\gamma_0(z, 1) - \dot\gamma_j(z, 1)$, the function in (C.3) has differential

$$q \longmapsto \Big\langle dF(w) \cdot q,\, \dot\gamma_0(y, 1) - \dot\gamma_j(y, 1) \Big\rangle_y = \langle q, p_j \rangle, \tag{C.4}$$

where $y = F(w)$,

$$p_j = dF(\phi(w)) \cdot \big(\dot\gamma_0(y, 1) - \dot\gamma_j(y, 1)\big), \tag{C.5}$$

and $\phi(w) = -d_w F(w) = -\dot\gamma_0(y, 1)$ is the initial velocity to go from $F(w)$ back to w. (The proof of (C.5) will be given later; anyway (C.4) is sufficient to conclude, without the explicit form of p_j.) Since all velocities $\dot\gamma_j(z, 1)$ are distinct, and distinct from $\dot\gamma_0(z, 1)$ (otherwise γ_j and γ_i would coincide for all times for some $i \neq j$), all the vectors $\dot\gamma_0(z,1) - \dot\gamma_j(z,1)$ are nonzero and distinct; this and the invertibility of $dF(\phi(w))$ guarantee that all vectors p_j are nonzero and distinct. It follows that

- each equation $\{\mathcal{E}_0(F(w)) - \mathcal{E}_j(F(w)) < 0\}$ defines a curved half-space;
- any two of these curved half-spaces are transversal to each other;
- the set $\mathrm{I}(M)$ is semiconvex around (x_0, v_0);
- it is differentiable at (x_0, v_0) if and only if there is only one curved half-space, i.e. $m = 1$, i.e. $N(x_0, v_0) = 2$;
- nondifferentiability points are included in intersections of two transversal submanifolds of dimension $n-1$, which are submanifolds of dimension $n-2$.

To conclude, let us give a short proof of (C.5). It suffices to establish the following identity (which is interesting for other purposes):

$$\text{(C.6)} \quad \forall (x,y) \in M \times M \setminus \mathrm{cut}(M), \forall (\xi, \eta) \in T_xM \times T_yM,$$
$$\big\langle (d_v \exp_x) \cdot \xi, \eta \big\rangle_y = \big\langle \xi, (d_{\phi(v)} \exp_y) \cdot \eta \rangle_x, \qquad v = (\exp_x)^{-1}(y).$$

(By a limiting argument, this formula remains true for $y = \exp_x v$, $v \in \overline{\mathrm{I}}(x)$.) Let us introduce Jacobi fields X and Y along γ defined by $X(0) = 0$, $\dot X(0) = \xi$, $Y(1) = 0$, $\dot Y(1) = -\eta$. From the properties of Jacobi fields we have

$$\frac{d}{dt}\langle X(t), \dot Y(t)\rangle = \langle X(t), \ddot Y(t)\rangle + \langle \dot X(t), \dot Y(t)\rangle = -\langle X, R(Y, \dot\gamma)\dot\gamma\rangle + \langle \dot X, \dot Y\rangle,$$

where R is the Riemann curvature tensor. By the properties of the Riemann curvature, this quantity is symmetric in X and Y, so

$$\frac{d}{dt}\Big(\langle X(t), \dot Y(t)\rangle - \langle \dot X(t), Y(t)\rangle\Big) = 0,$$

in other words $\langle X(t), \dot Y(t)\rangle - \langle \dot X(t), Y(t)\rangle$ is independent of t. It follows that

$$\langle X(1), \dot Y(1)\rangle - \langle X(0), \dot Y(0)\rangle = \langle Y(1), \dot X(1)\rangle - \langle Y(0), \dot X(0)\rangle,$$

or equivalently $\langle X(1), \dot Y(1)\rangle = -\langle Y(0), \dot X(0)\rangle$, which is the same as (C.6). □

C.3. **The case $\delta > 0$.** In this subsection we shall see that some results of the previous subsection remain true at least under an assumption which is weaker than the nonfocality. For any $(x, v) \in \mathrm{TCL}(M)$, let

$$\delta(x,v) = \mathrm{diam}\big((\exp_x)^{-1}(\exp_x v)\big).$$

The function δ is easily shown to be upper semicontinuous. Let then

$$\widetilde{J} = \bigcap_{\delta_0 > 0} \overline{\{\delta \le \delta_0\}}. \tag{C.7}$$

The "purely focal" tangent locus being defined by $\delta = 0$, the set $\widetilde{J}$ can be thought of as the set of velocities which are "almost purely focal". The sphere is an example of manifold with $J \neq \emptyset$ but $\widetilde{J} = \emptyset$.

An element (x, v) of TM can be thought of as a triple (x, t, ξ), where $t = |v|$ and $\xi = v/|v|$. (If $t = 0$ there is no need to define ξ.) Then $\mathrm{I}(M) = \{\Phi < 0\}$, where $\Phi(t, x, \xi) = t - t_C(x, \xi)$. If $t_C(x, \cdot)$ is a semiconcave function of $\xi \in U_x M$, then $\mathrm{I}(x)$ is a semiconvex set in the sense of Definition B.1. By abuse of language, we shall say that $\mathrm{I}(M)$ is semiconvex if t_C is a semiconcave function of $(x, \xi) \in UM$. This is a more precise property than just requiring $\mathrm{I}(x)$ to be semiconvex for all x.

Proposition C.6. *With the above notation, let M satisfy $\widetilde{J} = \emptyset$. Then*

(a) There is $\delta_0 > 0$ such that for any $(x, v) \in \mathrm{TCL}(M)$, $\delta(x, v) \ge \delta_0$.

(b) $\mathrm{I}(M)$ is semiconvex.

Proof of Proposition C.6. (a) If the conclusion is false, there is a sequence $(x_k, v_k)_{k \in \mathbb{N}}$ in $\mathrm{TCL}(M)$ such that $\delta(x_k, v_k) \le 1/k$. Without loss of generality (x_k, v_k) converges to $(\overline{x}, \overline{v}) \in \mathrm{TCL}(M)$, which belongs to the closure of $\{\delta \le 1/\ell\}$, for all $\ell \in \mathbb{N}$. So $(\overline{x}, \overline{v}) \in \widetilde{J}$, which is impossible.

(b) The problem is to show that t_C inherits the semiconcavity property of the distance function, even though we cannot directly apply the implicit function theorem (because we might be considering focal points). The argument below was communicated to us by Ludovic Rifford. It shows that property (b) is true away from $\widetilde{J}$ even if the latter is not empty.

Let $\delta_0 > 0$, let $(\overline{x}, \overline{\xi})$ be such that $\mathrm{dist}((\overline{x}, \overline{\xi}), \{\delta < \delta_0\}) \ge \eta > 0$, where the distance is in both variables x and ξ (it is defined in local charts for instance). The goal is to prove the semiconcavity of t_C around $(\overline{x}, \overline{\xi})$. To do this, let us choose (x_0, ξ_0) close to $(\overline{x}, \overline{\xi})$; we shall show that there is a smooth function τ, defined in a neighborhood V_0 of (x_0, ξ_0), touching the graph of t_C at (x_0, ξ_0) from above. If the C^2 norm of τ is controlled from above and the size of V_0 is controlled from below (independently of (x_0, ξ_0) in a neighborhood of $(\overline{x}, \overline{\xi})$), then we will be done.

Let $y_0 = \exp_{x_0}(t_C(x_0, \xi_0)\, \xi_0)$. By definition of δ there is $\widetilde{\xi_0} \neq \xi_0$ such that $t_C(x_0, \widetilde{\xi_0}) = t_C(x_0, \xi_0) =: \tau_0$, and $y_0 = \exp_{x_0}(\tau_0\, \widetilde{\xi_0})$, and $|\tau_0\, \xi_0 - \tau_0\, \widetilde{\xi_0}| \ge \delta_0$, so $|\xi_0 - \widetilde{\xi_0}| \ge \delta_0/\theta$, where θ is a lower bound for τ_0.

Since d is semiconcave and $\exp_x(\tau_0\,\widetilde{\xi}_0) = y_0$, we have

$$d(x,y) \le d(x_0,y_0) - \langle \widetilde{\xi}_0, \delta x\rangle_x + \langle \widetilde{\xi}_1, \delta y\rangle_y + O(|\delta x|^2 + |\delta y|^2), \tag{C.8}$$

where δx and δy are small enough, $x' = \exp_{x_0}(\delta x)$, $y' = \exp_{y_0}(\delta y)$, and $\widetilde{\xi}_1$ is the velocity at time τ_0 of the geodesic curve $(\exp_{x_0}(t\widetilde{\xi}_0))_{t\ge 0}$. Moreover, the constants in the "error term" are uniform in (x_0, y_0).

From (C.8) there is a C^2 function $h = h(x,y)$, defined in a neighborhood U_0 of (x_0,y_0) (of size controlled from below), such that

$$\begin{cases} h(x_0,y_0) = d(x_0,y_0) = \tau_0; \\ \\ \nabla_{x,y} h(x_0,y_0) = (-\widetilde{\xi}_0, \widetilde{\xi}_1); \\ \\ h(x,y) \ge d(x,y) \quad \text{for } (x,y) \in U_0. \end{cases} \tag{C.9}$$

Perturbing h just a little bit, we might assume that the last inequality in (C.9) is strict unless $(x,y) = (x_0,y_0)$.

Let now $F : UM \times \mathbb{R}_+ \to \mathbb{R}$ be defined by

$$F(x,\xi,t) = h(x, \exp_x(t\xi)) - t.$$

This is a smooth function, satisfying

$$F(x_0,\xi_0,\tau_0) = h(x_0,\xi_0) - \tau_0 = 0;$$

$$\frac{\partial F}{\partial t}(x_0,\xi_0,\tau_0) = \xi_1 \cdot \nabla_y h(x_0,y_0) - 1 = \xi_1 \cdot \widetilde{\xi}_1 - 1 = -\frac{|\xi_1 - \widetilde{\xi}_1|^2}{2} \le -\frac{\delta^2}{2\theta} < 0.$$

Since $\partial F/\partial t(x_0,\xi_0,\tau_0) < 0$, we may apply the implicit function theorem to uniquely solve the equation

$$F(x,\xi,\tau(x,\xi)) = 0,$$

thus defining a C^2 function τ in a neighborhood V_0 of (x_0,ξ_0). Since all constants (and in particular the negative lower bound for $\partial F/\partial t$) are uniform, the size of this neighborhood is controlled from below, and the C^2 norm of τ is controlled from above.

Let $(x,\xi) \in V_0$ and let $y = \exp_x(\tau(x,\xi)\xi)$. If $(x,y) \ne (x_0,y_0)$ then we have $d(x,y) < h(x,y) = \tau(x,\xi)$, so necessarily $t_C(x,\xi) < \tau(x,\xi)$. If on the other hand $(x,y) = (x_0,y_0)$, this means that $t_C(x,\xi) = d(x_0,y_0) = \tau_0$, but this is the same as $\tau(x,\xi)$ since $F(x,\xi,t_C(x,\xi)) = 0$. In either case t_C is bounded above by τ in V_0, and this concludes the argument. □

Appendix D. A counterexample

In this section we sketch the construction of a compact surface S embedded in $\mathbb{R}^3$, with strictly positive curvature, such that $\mathrm{I}(x)$ is nonconvex for some $x \in S$. The construction is reminiscent to that in [14] showing that positive curvature does not imply the weak MTW condition.

In the three-dimensional Euclidean space with coordinates x_1, x_2, x_3, let P stand for the paraboloid $P = \{x_3 = (x_1^2 + x_2^2)/2\}$. Choose $A = (0, a_2, a_3)$ in the exterior of P with $0 < a_3 < a_2^2$. Let C be the cone made of all tangent lines passing through A, which are tangent to P. By direct computation, the volume enclosed by C is a connected component of $\{Q > 0\}$, where Q is a $(1,2)$ quadratic form, in particular it is convex. An elementary argument shows that the set $\mathcal{E}$ of points which are either in the interior side of P, or squeezed between C and P, is convex (see Fig. 6). Let $\widetilde{S}$ be the boundary of this set: this is a nonnegatively curved surface which takes the form of a cone sitting on top of a paraboloid.

Let $O = (0,0,0)$ and $\xi = (0,1,0) \in T_O\widetilde{S}$. The geodesic starting from O with velocity ξ, drawn on $\widetilde{S}$, ceases to be minimizing before it reaches A (indeed, no minimizing geodesic can pass through the apex of a cone); let t_0 be the maximal time interval on which this geodesic is minimizing.

Next let $\xi_\pm = (1/\sqrt{2})(\pm 1, 1, 0) \in T_O\widetilde{S}$. If A is close enough to P, the geodesics starting from $\xi_\pm$, drawn on $\widetilde{S}$, are entirely drawn on P and remain minimizing up to infinity. If $z > 0$ is large enough, these geodesics have existed for a time $t_1 \geq 2(t_0+1)$ when they reach the height z. We cut $\widetilde{S}$ above $\{x_3 = z\}$, close it by symmetry, and flatten a bit the surface so that it is smooth at $\{x_3 = z\}$. This defines a positively curved surface S on which $t_1\xi_\pm \in \mathrm{I}(O)$, but $(1/2)(t_1\xi_+ + t_1\xi_-) = \lambda\,\xi$ with $\lambda > \sqrt{2}t_0$, so $\lambda\,\xi \notin \mathrm{I}(O)$. Thus $\mathrm{I}(O)$ is not convex.

To complete the construction it suffices to flatten a bit the cone C into a smooth and positively curved surface C_ε, while keeping $C \cap P$ unchanged, such that C_ε converges to C in Hausdorff distance (and in C^1 away from A) as $\varepsilon \to 0$. Let S_ε be obtained as before by considering the cone C_ε sitting on top of the paraboloid P, cutting at height z, closing by symmetry and flattening. If ε is small enough, the cut time of the geodesic starting from O with initial velocity ξ will be no more than $t_0 + 1$, and the conclusion is unchanged.

Remark D.1. It would be more striking to have a counterexample whose cut locus is nonfocal. The construction might not be very difficult, but requires more care.

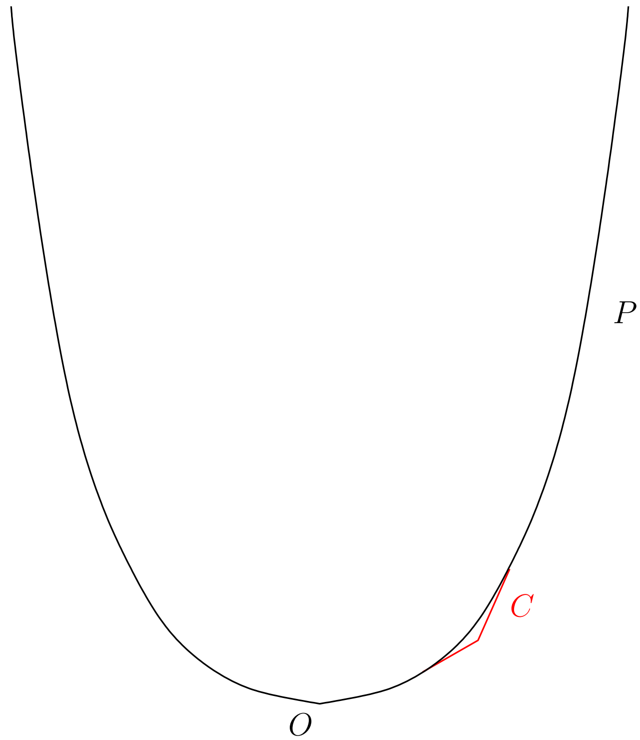

Figure 6. The small cone C touches the paraboloid P tangentially. For the resulting surface, the injectivity domain at O is not convex.

References

[1] Bishop, R. L. Decomposition of cut loci. *Proc. Amer. Math. Soc. 65*, 1 (1977), 133–136.

[2] Caffarelli, L. A. The regularity of mappings with a convex potential. *J. Amer. Math. Soc. 5*, 1 (1992), 99–104.

[3] Caffarelli, L. A. Boundary regularity of maps with convex potentials. *Comm. Pure Appl. Math. 45*, 9 (1992), 1141–1151.

[4] Caffarelli, L. A. Boundary regularity of maps with convex potentials. II. *Ann. of Math. (2) 144*, 3 (1996), 453–496.

[5] Cordero-Erausquin, D., McCann, R. J., and Schmuckenschläger, M. A Riemannian interpolation inequality à la Borell, Brascamp and Lieb. *Invent. Math. 146*, 2 (2001), 219–257.

[6] Delanoë, Ph., and Ge, Y. Regularity of optimal transportation maps on compact, locally nearly spherical, manifolds. Preprint, 2008.

[7] Delanoë, Pj., and Loeper, G. Gradient estimates for potentials of invertible gradient-mappings on the sphere. *Differential Geom. Appl. 20*, 2 (2004), 145–165.

[8] do Carmo, M. P. *Riemannian Geometry*. Mathematics: Theory & Applications. Birkhaüser Boston Inc., Boston, 1992.

[9] Federer, H. *Geometric measure theory*. Die Grundlehren der mathematischen Wissenschaften, Band 153. Springer-Verlag New York Inc., New York, 1969.

[10] Figalli, A., and Rifford, L. Continuity of optimal transport maps on small deformations of $\mathbb{S}^2$. Preprint, 2008.

[11] Figalli, A., and Villani, C. An approximation lemma about the geometry of the cut locus, with applications in optimal transport theory. To appear in *Meth. Appl. Anal.*

[12] Gallot, S., Hulin, D., and Lafontaine, J. *Riemannian geometry*, second ed. Universitext. Springer-Verlag, Berlin, 1990.

[13] Itoh, J.-i., and Tanaka, M. The Lipschitz continuity of the distance function to the cut locus. *Trans. Amer. Math. Soc. 353*, 1 (2001), 21–40.

[14] Kim, Y.-H. Counterexamples to continuity of optimal transportation on positively curved Riemannian manifolds. Preprint, 2007.

[15] Kim, Y.-H., and McCann, R. J. On the cost-subdifferentials of cost-convex functions. Preprint, 2007. Archived online at `arxiv.org/abs/0706.1266.` Included, with modifications, to Ref. [16].

[16] Kim, Y.-H., and McCann, R. J. Continuity, curvature, and the general covariance of optimal transportation. To appear in *J. Eur. Math. Soc.*

[17] Kim, Y.-H., and McCann, R. J. Towards the smoothness of optimal maps on Riemannian submersions and Riemannian products (of round spheres in particular). Preprint, 2008.

[18] Li, Y., and Nirenberg, L. The distance function to the boundary, Finsler geometry, and the singular set of viscosity solutions of some Hamilton–Jacobi equations. *Comm. Pure Appl. Math. 58* (2004), 85–146.

[19] Liu, J. Hölder regularity of optimal mapping in optimal transportation. To appear in *Calc. Var. Partial Differential Equations*.

[20] Loeper, G. On the regularity of maps solutions of optimal transportation problems. To appear in *Acta Math.*

[21] Loeper, G. On the regularity of maps solutions of optimal transportation problems II. Preprint, 2007.

[22] Ma, X.-N., Trudinger, N. S., and Wang, X.-J. Regularity of potential functions of the optimal transportation problem. *Arch. Ration. Mech. Anal. 177*, 2 (2005), 151–183.

[23] Mantegazza, C., and Mennucci, A. C. Hamilton-Jacobi equations and distance functions on Riemannian manifolds. *Appl. Math. Optim. 47*, 1 (2003), 1–25.

[24] McCann, R. J. Polar factorization of maps on Riemannian manifolds. *Geom. Funct. Anal. 11*, 3 (2001), 589–608.

[25] Rifford, L. Notes on the local inversion theorem. Personal communication.

[26] Trudinger, N. S., and Wang, X.-J. On strict convexity and continuous differentiability of potential functions in optimal transportation. Preprint, 2006.

[27] Trudinger, N. S., and Wang, X.-J. On the second boundary value problem for Monge–Ampère type equations and optimal transportation. Preprint, 2006.
Archived online at `arxiv.org/abs/math.AP/0601086`.

[28] Urbas, J. On the second boundary value problem for equations of Monge–Ampère type. *J. Reine Angew. Math. 487* (1997), 115–124.
[29] Villani, C. *Topics in optimal transportation*, vol. 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003.
[30] Villani, C. Optimal transport, old and new. *Grundlehren der mathematischen Wissenschaften*, Vol. 338, Springer-Verlag, 2009.
[31] Villani, C. Stability of a 4th-order curvature condition arising in optimal transport theory. *J. Funct. Anal. 255*, 9 (2008), 2683–2708.
[32] Von Nessi, G. T. Regularity results for potential functions of the optimal transportation problem on spheres and related Hessian equations. PhD Thesis of the Australian National University, 2008.
[33] Weinstein, A. The cut locus and conjugate locus of a Riemannian manifold. *Ann. of Math. (2) 87* (1968), 29–41.

Grégoire Loeper

Université Claude Bernard
ICJ, UMR CNRS 5208
43 Boulevard du 11 Novembre 1918
69622 Villeurbanne Cedex
FRANCE

e-mail: loeper@math.univ-lyon1.fr

Cédric Villani

ENS Lyon & Institut Universitaire de France
UMPA, UMR CNRS 5669
46 allée d'Italie
69364 Lyon Cedex 07
FRANCE

e-mail: cvillani@umpa.ens-lyon.fr