Chapter 1
Introduction

Abstract In this chapter we introduce the reader to the problem addressed by this monograph. First we explain the main question at hand and its motivation in the context of the Wentzell-Freidlin theory of rare transition paths. We then summarize the main features of our existence theory, and the various approaches used in the literature. Finally, we explain the structure of this monograph and introduce some notation.

1.1 Geometric Action Functionals

A geometric action S is a mapping that assigns to every unparameterized oriented rectifiable curve γ in $\mathbb{R}^n$ a number $S(\gamma) \in [0, \infty)$. It is defined via a curve integral

$$S(\gamma) := \int_\gamma \ell(z, dz) := \int_0^1 \ell(\varphi, \varphi') \, \mathrm{d}\alpha, \tag{1.1}$$

where $\varphi\colon [0, 1] \to \mathbb{R}^n$ is any absolutely continuous parameterization of γ, and where the local action $\ell \in C(\mathbb{R}^n \times \mathbb{R}^n, [0, \infty))$ must have the properties

(i) $\forall x, y \in \mathbb{R}^n \ \ \forall c \geq 0\colon \ \ \ell(x, cy) = c\ell(x, y)$,

(ii) for every fixed $x \in \mathbb{R}^n$ the function $\ell(x, \cdot\,)$ is convex.

While (i) guarantees that the second integral in (1.1) is independent of the choice of φ, (ii) is necessary to ensure that S is lower semi-continuous in a certain sense. A trivial example is given by $\ell(x, y) = |y|$, in which case $S(\gamma)$ is just the Euclidean length of γ, or more generally, by $\ell(x, y) = |y|_{g_x}$ for any Riemannian metric g. In fact, ℓ generalizes the well-studied notion of a Finsler metric [2] in that (a) ℓ only needs to be continuous (no smoothness required), that (b) we do not require that $\ell(x, y) = \ell(x, -y)$, and that (c) ℓ^2 need not be *strictly* convex in y.

Now given two sets $A_1, A_2 \subset \mathbb{R}^n$, in this work we develop criteria under which there exists a minimum action curve $\gamma^\star$ leading from A_1 to A_2, i.e., under which

M. Heymann, *Minimum Action Curves in Degenerate Finsler Metrics*,
Lecture Notes in Mathematics 2134, DOI 10.1007/978-3-319-17753-3_1

$\exists \gamma^\star \in \Gamma_{A_1}^{A_2} := \{\gamma \mid \gamma \text{ starts in } A_1 \text{ and ends in } A_2\}$ such that

$$S(\gamma^\star) = \inf_{\gamma \in \Gamma_{A_1}^{A_2}} S(\gamma). \tag{1.2}$$

We then prove properties of the minimizer $\gamma^\star$ without finding $\gamma^\star$ explicitly.

Although our existence results can certainly be applied to the exemplary local actions given above, the present work was primarily motivated by a recently emerging problem from large deviation theory that is adding a considerable layer of difficulty: In contrast to Finsler metrics, in this example $\ell(x, y)$ vanishes in some direction $y = b(x) \neq 0$, which allows for curves γ (the flowlines of the vector field b) with positive Euclidean length but vanishing action $S(\gamma)$.

1.2 Example: Large Deviation Theory

Consider for some $b \in C^1(\mathbb{R}^n, \mathbb{R}^n)$ and some small parameter $\varepsilon > 0$ the stochastic differential equation (SDE)[1]

$$\mathrm{d}X_t^\varepsilon = b(X_t^\varepsilon)\,\mathrm{d}t + \sqrt{\varepsilon}\,\mathrm{d}W_t, \qquad X_{t=0}^\varepsilon = x_1, \tag{1.3}$$

where $(W_t)_{t\geq 0}$ is an n-dimensional Brownian motion, and where the zero-noise-limit, i.e., the ODE $\dot{x} = b(x)$, has two stable equilibrium points $x_1, x_2 \in \mathbb{R}^n$. The presence of the small noise allows for rare transitions from x_1 to x_2 that would be impossible without the noise (*green curve* in Fig. 1.1), and one is interested in

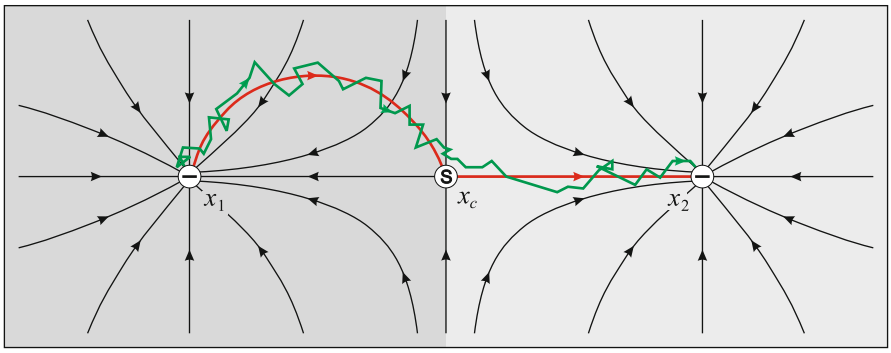

Fig. 1.1 Rare noise-induced transitions from one meta-stable state to another (*green curve*) stay near the minimum action curve $\gamma^\star$ (*red*) with high probability

[1]The reader with no background in probability theory should not feel discouraged here: No knowledge in that field will be required to understand the results or proofs in this monograph.

the frequency and the most likely pathway of these transitions. Both questions are answered within the framework of Wentzell-Freidlin Theory [8] (a subfield of large deviation theory), the key object being the quasipotential

$$V(x_1, x_2) = \inf_{\substack{T>0 \\ \chi \in \bar{C}^{x_2}_{x_1}(0,T)}} S_T(\chi), \tag{1.4}$$

$$\text{where} \qquad S_T(\chi) = \frac{1}{2} \int_0^T |b(\chi) - \dot{\chi}|^2 \, \mathrm{d}t, \tag{1.5}$$

and where $\bar{C}^{x_2}_{x_1}(0, T)$ denotes the space of all absolutely continuous functions $\chi\colon [0, T] \to \mathbb{R}^n$ fulfilling $\chi(0) = x_1$ and $\chi(T) = x_2$.

The idea behind this formula is that transitions have been shown to more likely occur in neighborhoods of paths χ with small action $S_T(\chi)$, and thus $V(x_1, x_2)$ is a measure for how likely it is to see *any* transition within some fixed observation time (with smaller values of V indicating a higher likelihood). Furthermore, the expected time until a transition to x_2 happens was shown to scale like $\mathrm{e}^{V(x_1,x_2)/\varepsilon}$ as $\varepsilon \searrow 0$ [16]. Observe that $S_T(\chi)$ cannot be made arbitrarily small, since paths χ that leave x_1 must deviate from the flowlines of b (which fulfill $\dot{\chi} = b(\chi)$).

An unpleasant feature of this formulation is that the minimization problem (1.4) does not have a minimizer $(T^\star, \chi^\star)$, i.e., a function $\chi^\star \in \bar{C}^{x_2}_{x_1}(0, T^\star)$, defined on some optimal finite time interval $[0, T^\star]$, at which the infimum (1.4) is achieved. The main reason for this is that by [8, Chap. 4, Lemma 3.1] $\dot{\chi}^\star$ would need to vanish at x_1 and x_2, and typically also at some critical point x_c along the way (see Sect. 4.4), so that $\chi^\star$ would need infinite time each to leave x_1, pass x_c and approach x_2. Therefore, in general it is not even possible to define a minimizer $\chi^\star\colon \mathbb{R} \to \mathbb{R}^n$ on an infinite time interval, but one would rather have to paste together two solutions $\chi_1^\star, \chi_2^\star\colon \mathbb{R} \to \mathbb{R}^n$ with

$$\lim_{t\to-\infty} \chi_1^\star(t) = x_1, \qquad \lim_{t\to\infty} \chi_1^\star(t) = \lim_{t\to-\infty} \chi_2^\star(t) = x_c, \qquad \text{and} \qquad \lim_{t\to\infty} \chi_2^\star(t) = x_2.$$

This is a major problem for both analytical and numerical work, and so in [9, 10] the use of the alternative representation

$$V(x_1, x_2) = \inf_{\gamma \in \Gamma^{x_2}_{x_1}} S(\gamma) \tag{1.6}$$

was suggested, where the geometric action $S(\gamma)$ is given by

$$\ell(x, y) = |b(x)||y| - \langle b(x), y\rangle, \qquad \text{(SDE)} \tag{1.7}$$

which can be seen as a degenerate version of a Randers metric [2, Chap. 11]. A minimizer $\gamma^\star$ of (1.6), i.e., a *maximum likelihood transition curve* (the *red curve* in Fig. 1.1), seems more feasible to exist in this formulation since the time parameterization has been eliminated from the problem.

This geometric reformulation of the quasipotential generalizes also to other types of Markovian time-homogeneous[2] stochastic dynamics, such as SDEs with multiplicative noise or continuous-time Markov jump processes [9, 10, 16], with modified (in the latter case not Randers-like) local action ℓ. It was shown to effectively remove the numerical difficulties [9–11, 19], and our goal in this monograph is now to demonstrate also its analytical advantages when addressing geometrical[3] questions.

1.3 Key Features of the Existence Theory

The goal of this monograph is to develop a comprehensive geometric theory for proving the existence of minimum action curves, the key features of which are the following:

(i) The theory can be applied to a large class of geometric actions, including those encountered in the context of large deviation theory. It also applies to Riemannian actions (as a trivial example), and in fact to actions that at different locations in space can have features of one or the other.
(ii) The minimization is carried out over the space of rectifiable curves with start and end points in some prescribed sets A_1 and A_2, respectively.
(iii) Curves can be constrained to only traverse points in a prescribed closed subset $\tilde{D} \subseteq \mathbb{R}^n$.
(iv) Whenever possible, minimizers $\gamma^\star$ are shown to be rectifiable as well.
(v) The conditions of the key theorems are non-technical and easy to check based on information that is explicitly available in practice.
(vi) Smoothness requirements on the local action ℓ and related functions are kept to a minimum.

In the process, the reader will be provided with the necessary basic definitions and concepts. The tools that we develop for our purposes have value in their own right, as they may be of use also in other problems related to geometric actions.

[2] That is, the definition of the dynamics via its drift and noise covariance matrix in the case of an SDE, or via its jump rates in the case of a jump process, cannot explicitly depend on time.

[3] See, however, [10, Sect. 2.4] for how the optimal time parameterization can be recovered from the minimum action curve $\gamma^\star$.

1.4 Techniques Used in the Literature

Let us take a look at some methods that have been used in the literature to prove the existence of optimal *time-dependent* curves, and let us understand why they either cannot be applied in the given geometric setting at all, or why they would only lead to partial results. The approaches fall into two categories:

(a) constructive techniques, which are based on the derivation of an ODE that minimizing curves need to fulfill, and which effectively transform the minimization problem into a boundary value problem with start point x_1 and end point x_2; and
(b) abstract techniques based on the lower semi-continuity of the action functional of interest.

1.4.1 Constructive Techniques

Two prominent examples of constructive techniques based on an ODE are the following:

(i) *First-Order ODE for Drift Vector Fields with a Gradient-Like Structure.* This technique can only be used for the specific action (1.5), where the drift vector field b must be of the form $b(x) = -\nabla V(x) + v^{\perp}(x)$ for some potential function $V: D \to [0, \infty)$, $D \subseteq \mathbb{R}^n$, and for some vector field $v^{\perp}$ perpendicular to ∇V. Under these assumptions, a simple estimate can show that any solution $\chi(t)$ of the ODE $\dot{\chi} = \nabla V(\chi) + v^{\perp}(\chi)$ minimizes the action between its start and end point [8, Chap. 4, Theorem 3.1]. Now assume that the given start point x_1 is the unique minimum of V and the only point at which ∇V vanishes, and that $V(x_2) \leq \inf_{x \in \partial D} V(x)$. Then since the solution of the above ODE with $\chi(t = 0) = x_2$ fulfills $\frac{\mathrm{d}}{\mathrm{d}t} V(\chi(t)) = |\nabla V(\chi(t))|^2 > 0$ for $\forall t \leq 0$ and therefore approaches x_1 as $t \to -\infty$, one can conclude that $\chi|_{(-\infty,0]}$ is a (generalized) minimizer of (1.4).

(ii) *The Euler-Lagrange Equation.* If the action S_T is not in the specific form (1.5) then there is no general first-order equivalent to the above ODE. Instead, one can derive a *second*-order ODE called the Euler-Lagrange equation for the extremals χ of S_T, by setting the variation δS_T equal to zero (this is the equivalent of finding the minimum of a function $f(x)$ by attempting to solve $f'(x) = 0$). For fixed T, one is then again left with the boundary problem that requires $\chi(0) = x_1$ and $\chi(T) = x_2$.

To obtain a more general theory that is not taylored to any specific action, one can write this ODE in the form of the $2n$-dimensional first-order ODE system $\dot{\chi} = \frac{\mathrm{d}H}{\mathrm{d}p}(\chi, p)$, $\dot{p} = -\frac{\mathrm{d}H}{\mathrm{d}x}(\chi, p)$, where the function $H(x, p)$ is the Hamiltonian associated to the action S_T (more precisely, it is the Legendre transform of its integrand). Necessarily, this reduction to a first-order system comes along with more relaxed boundary conditions: The solution $(\chi(t), p(t))$ must now lead from a point of the form (x_1, p_1) and to one of the form (x_2, p_2).

To minimize also over all $T > 0$ in (1.4), it turns out that we also need to ask that $H(x_1, p_1) = H(x_2, p_2) = 0$; if x_1 and x_2 are critical points of the system (i.e., if $\frac{\mathrm{d}H}{\mathrm{d}p}(x_{1,2}, 0) = 0$) then for a subclass of Hamiltonians ($\mathscr{H}_0$ in Definition 2.12 (iii)) this implies that $p_1 = p_2 = 0$.

The main problems with these constructive approaches are the following: First, the statement about the ODE in the first approach only holds for actions S_T in the given specific form, and its proof cannot be extended to general actions. Furthermore, if the point x_1 is not an attractor of b then the solution χ starting at x_2 will in general not lead to x_1 as $t \to -\infty$, and so the above statement ("if a solution of the ODE connects x_1 and x_2 then it is a minimizer") becomes worthless. The problem persists if x_1 and x_2 are replaced by sets A_1 and A_2, respectively.

The general Hamiltonian ODE still leaves us with the problem of showing that the derived boundary value problem actually has a solution, and it is unclear how this problem can be approached in our intended generality. Instead, this formulation is more useful in situations in which the existence of a minimizer can be *assumed*: For example, in [15] minimizers in $\mathbb{R}^2$ were computed numerically by solving the boundary value problem via the shooting method, and in [4, 5] the Hamiltonian formulation has turned out to be useful for proving properties of minimizers, addressing uniqueness questions, and investigating the regularity of the quasipotential.

The biggest two problems with *any* ODE-based constructive approach, however, are the following: First, minimizers $\gamma^\star$ of (1.2) have numerically been found to generally have cusps as they pass critical points (even in the basic case where ℓ is given by (1.7) with some smooth b, see Fig. 1.1 or [10, Fig. 4.1]). Therefore we know that there is no ODE that the arclength parameterization of $\gamma^\star$ could possibly fulfill throughout the entire curve.

Second, ODE-based approaches (both for geometric and for time-parameterized curves) would not allow us to constrain our curves to be contained in some given set $\tilde{D} \subseteq \mathbb{R}^n$ (point (1.3) in our wish list in Sect. 1.3), since such constraints can cause $\gamma^\star$ to become non-smooth when the curve reaches and then traces the (potentially also non-smooth) boundary $\partial\tilde{D}$.

For these reasons, such approaches are not an option for us.

1.4.2 The Lower Semi-Continuity Technique

The idea behind the lower semi-continuity approach is the following: As we know, any continuous function $f: I \to \mathbb{R}$ defined on a compact interval $I \subset \mathbb{R}$ obtains its infimum on I (i.e., $\exists x^\star \in I: f(x^\star) = \inf_{x \in I} f(x)$). However, it is not hard to see that we can in fact allow f to have jumps, as long as the function value at such points is not larger than any of the two one-sided limits. More generally, we only need to ask that $\forall x \in I: f(x) \leq \liminf_{y \to x} f(y)$. Functions with this property are called *lower semi-continuous*.

The proof that this property indeed still suffices is analogous to the continuous case: Take any minimizing sequence $(x_k)_{k \in \mathbb{N}}$ (i.e., $\lim_{k \to \infty} f(x_k) = \inf_{x \in I} f(x)$),

choose a converging subsequence $(x_{k_l})_{l\in\mathbb{N}}$ (this is possible since I is compact), and call its limit $x^\star \in I$. Then

$$f(x^\star) \le \liminf_{y\to x^\star} f(y) \le \lim_{l\to\infty} f(x_{k_l}) = \inf_{x\in I} f(x),$$

where we first used the lower semi-continuity of f, then the definition of lim inf, and finally the property of the minimizing sequence. This shows that $x^\star$ is a minimizer.

Now in our situation, in which the function $f(x)$ is replaced by the functional $S(\gamma)$, why would we not simply define ourselves a topology on the space of curves under which S is continuous, and then use the standard continuity result? The above proof shows that there is a fine trade-off to be made: If we choose the topology too fine (making it too hard for a sequence of curves to converge) then we may no longer be able to find a converging subsequence of our minimizing sequence of curves; if we choose the topology too coarse (making it too easy to converge) then our functional may no longer be continuous. It is for this reason that one commonly uses this weakened form of continuity—lower semi-continuity—when it comes to functionals: to ease this trade-off to the point that the existence proof can be completed.

Using this approach in our geometric context, one quickly arrives at the following first result (Proposition 3.8): *If there exists a minimizing sequence* $(\gamma_k)_{k\in\mathbb{N}}$ *of* (1.2) *whose curves* γ_k *are all contained in some compact set* $K \subset \mathbb{R}^n$ *and have uniformly bounded curve lengths, then there exists a minimizer* $\gamma^\star \in \Gamma_{A_1}^{A_2}$. (The conditions on $(\gamma_k)_{k\in\mathbb{N}}$ guarantee the existence of a converging subsequence, obtained by applying Arzelà-Ascoli's theorem.)

In practice, however, this criterion alone is of little use since minimizing sequences are not at our direct disposal, and so their curve lengths can be hard to control. What we need is an estimate that bounds the length of a curve γ in terms of its action $S(\gamma)$: since the curves in any minimizing sequence $(\gamma_k)_{k\in\mathbb{N}}$ have (converging and therefore) bounded actions, this would imply that the length condition in the statement above is fulfilled.

Now we see the challenge of our proof: The degeneracy of our local action $\ell(x, y)$ can allow a curve to move in a direction y ($= b(x)$ for the SDE geometric action (1.7)) at no cost, and so there can be arbitrarily long curves with small or zero action. Furthermore, at some critical points x_c (in the SDE case those points with $b(x_c) = 0$), $\ell(x_c, y)$ may even vanish for *every* direction y, which again allows for arbitrarily long curves near this point with arbitrarily small action. For this reason, the desired estimate described above (Lemma 6.13) and our resulting main existence criteria (Propositions 3.23 and 3.25) will be intimately tied to the flowline diagram of the drift vector field b, or of a generalized definition thereof for general geometric actions (Definition 2.7).

In [8, Chap. 4, Lemma 2.2], the existence of a (generalized) *time-parameterized* minimizer $\chi^\star\colon (-\infty, 0] \to \mathbb{R}^n$ of (1.4)–(1.5) is shown in the case where x_1 is an attractor of the vector field b and x_2 is a point in its basin of attraction (thus avoiding much of the problems caused by the time parametrization). Its proof suggests one

way of obtaining such an estimate away from critical points also for our geometric action $S(\gamma)$, based on the observation that there are no infinitely long flowlines or limit cycles in our region of interest. Following that specific route would however come at the cost that we would lose control over the minimizer's curve length near critical points, and so we would not be able to prove that our obtained minimizer $\gamma^\star$ stays rectifiable as it passes critical points. Our estimate in Lemma 6.13 instead, which carefully quantifies some decisive constants involved, does provide us with the desired extra amount of control near critical points, albeit at the cost of some extra work in our proofs.

1.5 Properties of Minimum Action Curves

Then turning our attention to the *properties* of minimizers, we consider a subclass of geometric actions that still contains the large deviation geometric actions mentioned above. For our main result, suppose that the drift b has two basins of attraction (see, e.g., Figs. 1.1, 3.4a,b, or 4.2), and let $\gamma^\star$ be the minimum action curve leading from one attractor to the other.

Since for the class of actions in question $\gamma^\star$ can follow the flowlines of b at no cost, it is not surprising that the second ("downhill") part of $\gamma^\star$ will be a flowline connecting a saddle point to the second attractor. In particular, the *last* hitting point of the separatrix is a point with zero drift (the saddle point). Here we prove also the non-obvious fact that also the *first* hitting point must have zero drift. In practice, such knowledge can be used either to gain confidence in the output of algorithms that compute $\gamma^\star$ numerically (such as the geometric minimum action method, gMAM, see [9, 10]), or to speed up such algorithms by restricting their search to only those curves with these properties.

Finally, we will demonstrate how the same result (Corollary 4.5) that is used to prove this property can also be used to prove the non-existence of minimizers in some situations.

1.6 The Structure of this Monograph

This monograph is split into two main parts and an appendix. In Part I we lay out all our results on the existence of minimum action curves, we demonstrate with several examples how to use our criteria in practice, we discuss when minimizers do *not* exist, and finally we prove the above-mentioned properties of minimum action curves. The reader who is only interested in gaining enough working knowledge to use our existence criteria in practice will find it sufficient to read only this first part.

Part II consists of two chapters: Chap. 6 contains the proofs of our key criteria (stated in Part I) under which a "local" existence property holds to which our global existence theorem has been reduced in Part I; the reader who wants to know why

these criteria work should also read this chapter. Chapter 7 contains the proof of a very technical lemma that was needed in Chap. 6 in order to deal with curves that are passing a saddle point; the reader can decide to skip this chapter without losing much insight.

Appendices A and B contain some of the more technical proofs that we have omitted in Parts I and II, respectively, in order to not interrupt the flow of the main arguments. While Appendix A can significantly contribute to the understanding of Part I, Appendix B is very technical in nature and can be skipped as well.

The suggested reading order is as follows: Part I, Appendix A, Part II, Appendix B.

1.7 Notation and Assumptions

For a point $x \in \mathbb{R}^n$ and a radius $r > 0$ we define the open and the closed balls

$$B_r(x) := \{w \in \mathbb{R}^n \,|\, |w - x| < r\} \qquad \text{and} \qquad \bar{B}_r(x) := \{w \in \mathbb{R}^n \,|\, |w - x| \leq r\}.$$

Similarly, for a set $A \subset \mathbb{R}^n$ and a distance $r > 0$ we define the open and the closed neighborhoods $N_r(A)$ and $\bar{N}_r(A)$ as

$$N_r(A) := \{w \in \mathbb{R}^n \,|\, \operatorname{dist}(w, A) < r\} \quad \text{and} \quad \bar{N}_r(A) := \{w \in \mathbb{R}^n \,|\, \operatorname{dist}(w, A) \leq r\}.$$

Furthermore, we denote by $\bar{A}$, by $A^c := \mathbb{R}^n \setminus A$, by $A^\circ := (\overline{A^c})^c$, and by $\partial A := \bar{A} \setminus A^\circ$ the closure, the complement, the interior, and the boundary of A in $\mathbb{R}^n$, respectively. For a point x on a C^1-manifold M we denote by $T_x M$ the tangent space of M at x.

For a function f and a subset A of its domain we denote by $f|_A$ the restriction of f to A, and we use the notation $f \equiv c$ to emphasize that f is constant. Expressions of the form $\mathbb{1}_{cond}$ denote the indicator function that returns the value 1 whenever the condition *cond* is fulfilled and 0 otherwise.

Finally, throughout this monograph we let $\tilde{D} \subseteq D \subseteq \mathbb{R}^n$ be two fixed connected sets, where D is open, and where $\tilde{D}$ is closed in D. An additional technical assumption on $\tilde{D}$ will be made at the beginning of Sect. 3.1. D will serve as our state space,[4] i.e., as the set that the curves γ live in, and $\tilde{D}$ will be used for an additional constraint on the curves γ during our minimization, i.e., we will in fact minimize over $\Gamma_{A_1}^{A_2} := \{\gamma \subset \tilde{D} \,|\, \gamma \text{ starts in } A_1 \text{ and ends in } A_2\}$. (For simplicity we suppress the dependence of $\Gamma_{A_1}^{A_2}$ on $\tilde{D}$ in our notation.) If no such constraint is desired, just choose $\tilde{D} := D$; the reader is encouraged to consider this simple unconstrained case whenever on first reading he may feel overwhelmed by some definition or statement involving $\tilde{D}$.

[4] Note that we may occasionally reuse the letter n of our state space dimension also for other purposes, e.g., as an index for sequences such as $(\gamma_n)_{n \in \mathbb{N}}$.

http://www.springer.com/978-3-319-17752-6

Minimum Action Curves in Degenerate Finsler Metrics
Existence and Properties
Heymann, M.
2015, XV, 186 p. 14 illus., 11 illus. in color., Softcover
ISBN: 978-3-319-17752-6