Continuous Automata, Compactness, and Young Measures

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Scott Bryan Messick

May 2016

Abstract

I suggest a framework for thinking about continuous-time processes as examples of a general topological notion of *continuous automaton*. A continuous automaton is a continuous action of a topological monoid on a topological space. This framework includes the traditional theory of finite automata, by considering the free monoid on a finite alphabet and imbuing both the monoid and the set of states with the discrete topology. I show how to develop within this framework an analogous theory of continuous-time automata by replacing where appropriate criteria of *finiteness* with criteria of *compactness*.

A key difficulty to be overcome in this analogy is compactness of the space of possible inputs received over a fixed time interval. The difficulty arises because of the possibility of arbitrarily fast oscillations of input, and has no discrete-time analogue. I discuss the significance of this problem, and show how to solve the problem by using the space of Young measures.

I develop the theory of Young measures as an alternative (to that of Lebesgue) metric completion of the space of rational step functions on an interval. This approach is more constructive than the measure-theoretic approach, and helps to suggest that Young measures can be thought of as continuous-time "words", much as a finite automata theorist thinks about words over an alphabet. This method for developing Young measures is new and interesting in its own right. I believe that the utility and naturalness of Young measures in this context reflects a deeper connection between control theory and the theory of computation.

With continuous-time automata defined using Young measures, I prove several preliminary

theorems, including an existence theorem for automata defined by differential equations, and Myhill-Nerode theorems for transducers and deducers (continuous analogues of recognizers). These theorems begin, barely, to give shape to an abstract theory of continuous-time processes.

Finally, I discuss limits to the parallel between discrete-time and continuous-time automata, and outline several possible ways of continuing the research in the hope that it would ultimately lead to innovation in the actual practice of engineering continuous systems.

Biographical Sketch

Scott Messick completed his B.S. in Mathematics at the University of Chicago in 2010.

to future engineers

Acknowledgements

I am grateful to my friends, instructors, and others who have helped me develop this research into its current form. There are a few I want to thank by name. Jun Le Goh and Daoji Huang helped find and correct subtle gaps in various proofs. I also thank Daoji for being generally awesome. Iian Smythe gave me an invaluable mathematical pointer early on. Bob Constable taught me the foundations of constructive thought, which subtly permeate the ideas presented here. Anonymous reviewers of earlier versions of this work provided extremely useful feedback on wide-ranging matters, including among many other things the writing style and how to explain what the Young metric actually does. Lastly, I thank my advisor Anil Nerode deeply for his support and insight.

Contents

Preface

In my thesis work I have sought to initiate a general mathematical theory of continuous-time computation, or equivalently of continuous-time processes. Over time I have learned to appreciate the vast scope of such a project, and the work I actually did—though I believe it is good work and I am proud of it—barely scratches the surface.

Academic writing generally tends to become dry and formal as the author tries to make sure the tone is objective and the results rigorously justified. Here I have consciously chosen to reject these aims. I have fought to keep the writing lively and to include, at least in some cases, material which is subjective, not rigorously justified, or both. I believe that to make real intellectual progress on such a large project requires openness and honesty about the real state of the research, which is inherently incomplete. So for example, I have tried to convey tidbits which might look like the following:

- I tried this thing and it didn't work.
- I tried this thing and it seems like it's working but I didn't have time to take it very far.
- There's this body of research which I think might be relevant but I didn't have time to read it.

These sorts of information could be crucial clues to a possible future researcher who is reading this text—and I would really like to encourage anyone who is interested to help continue the work. There's plenty to go around!

With all that said, a casual writing style has its drawbacks. It can be distracting, and can confuse the reader as to what has really been demonstrated and what hasn't. (And I certainly wouldn't want the reader to think my mathematical proofs are not intended to meet the community's usual standards of rigor.) So for the sake of organization, I have concentrated most of the more subjective or speculative parts of the work into "Notes" sections at the end of each chapter.

The most important part of the work is philosophical. The overwhelming majority of formal mathematical tools which I'm about to describe are adapted versions of something which is already known to a different mathematical community, but from a different perspective. The most salient example is that of Young measures, but there are many more. In some cases I have extended or reinvented these tools to fit the needs of the project. But I must emphasize that in no way have I exhausted the produce of the existing mathematical literature on all related topics, which include dynamical systems, automata theory, control theory, and topology, among many others. On the contrary, continuing to engage with the experts and literature in these topics will be an essential part of future work in the area. I have therefore included in each "Notes" section as much useful knowledge about the literature as I can state with reasonable confidence, even when it is highly incomplete. Unavoidably, there will be mistakes or misunderstandings. Many of the works cited in the bibliography are ones that I have only skimmed, perhaps years ago.

Personal circumstances have been another factor in the conspiracy to prevent me from bringing this work at least to a natural resting point. But despite everything, I have tried my utmost to present it in a readable form, and even to make it enjoyable. And on that note, let us turn to the work itself.

1 Introduction

The discrete theory of computation has been a spectacular success. Basic notions such as Turing machines and finite automata have proved to be thoroughly mathematically robust and suitable for all sorts of theoretical investigation from different perspectives.

My purpose in carrying out the work for this thesis was to try to find equally robust notions which could serve as a foundation for a theory of continuous computation, especially computation over continous time. Better still, I hoped to find robust notions which could serve as a foundation for all forms of computation, with discrete and continuous-time notions each falling out naturally as special cases.

Inevitably, the setting of continuous time introduces many complications. I have focused mainly on finding a parallel for the classical theory of finite automata, because finite automata are the simplest notion of a computational device I know. Still, big questions arise almost immediately. For example, what is supposed to be the notion of a word in continuous time, to be fed into a continuous automaton as input? A natural suggestion would be to take some sort of functions $[0, \ell] \to \Sigma$ from an interval of time values into an alphabet, since finite words are functions $\{1, 2, \ldots, n\} \to \Sigma$. Unfortunately, the suggestion is not precise enough to resolve the issue. Would we be taking continuous functions? Measurable functions? 1-Lipschitz functions? Arbitrary set-theoretic functions?

Setting the question about words aside for a moment, I adopted two principles for how the parallelism with finite automata theory should proceed: finite sets are replaced by compact spaces, and all maps are required to be continuous. These principles work together in the

sense that when restricting to continuous functions, compact spaces behave somewhat as finite sets do with arbitrary functions. Especially, the continuous image of a compact space is compact.

My motivation for requiring all maps to be continuous stems from the insight of constructive analysis that for any totally defined function to be computable in any practical sense, it has to be continuous. Roughly, the idea is that we have no hope of exactly specifying a point in an infinite space (at least in general), so we are going to have to make do with approximations. A topology on a space specifies precisely what it means to approximate a point in the space, at least in the limit. A metric goes further and actually gives an absolute notion of quality of approximation. Either way, for a map from an input space to an output space to be computable, it is necessary that an approximation of the output can be determined from a sufficiently good approximation of the input.

However, I have not attempted to require that the objects and maps be effective in the sense of computable or constructive analysis. The reason is that this work is supposed to expand our understanding of computation into a setting where time simply *is* continuous, much as it usually seems to be in the real world. Though fundamental physics is not exactly solved, most physics models seem to treat time as continuous and require systems to evolve continuously. Moreover, we as human beings usually think of time as proceeding continuously, and in many application domains the most natural models do in fact have continuous time.[1]

The rest of this chapter contains a discussion of the basic problems in defining continuous automata at a more concrete level.

[1] A possible objection to the lack of effectiveness requirements is that some objects which mathematically satisfy the demands of the theory will be impossible to actually build or compute in the real world. However, this fact is already a consequence of the idealizing nature of mathematics. Certainly there are computable functions in the classical sense which could never be carried out in the real world because they require more steps than the number of particles in the observable universe. To put it differently, the mathematical notion of computability is an approximation of practical computability *from above*: every practically computable function is mathematically computable, but the converse does not hold. Experience suggests that practical computability is too finicky a concept to usefully pin down with mathematical precision, so we are happy to have a mathematically perspicuous approximation from above. And so it shall be with continuous computation. Now, to be clear, effectiveness results from the standpoint of computable analysis would be entirely welcome and are a worthy goal. I only mean to say why they are not part of the fundamental definitions.

1.1 A Sample Menu of Continuous Devices

Let us consider some concrete examples of mathematical behavior which might be desired from a continuous-time device (i.e. a continuous automaton, which we have yet to define). Not all of these examples will really constitute admissible behavior for a continuous automaton, but we include negative examples as we are trying to show what is really at stake in this definition.

Throughout the thesis we assume a deterministic, automatic style of computation, unless otherwise noted. By "automatic", we mean that the computation updates its state instanteously in response to ongoing input. There is no external memory and no waiting for the machine to halt. For the time being we treat the input as a function of time which takes values in some alphabet Σ. To emphasize that input is changing over time, we may use the phrase *input signal*.

Within these restrictions, the computation style of the examples still varies over a couple of dimensions:

Alphabet type. Discrete or continuous? Basic respective examples are $\Sigma = \{0, 1\}$ and $\Sigma = [0, 1]$.

Output style. Recognizer or transducer? Traditionally, an automaton would fit in one of these two categories. A recognizer has a set of accept states and there by computes some language (set of inputs). A transducer would actively put out an output signal of a similar kind as the input, perhaps over a different alphabet. Here we also allow another style which we call a *deducer*. A deducer is like a recognizer but with more than two, usually infinitely many, outcomes. It computes a function from all possible input signals into some static space of possible outcomes. In the discrete world all deducers could be reduced to a finite collection of recognizers working together; but there is no obvious analogue in the continuous case.

Input restrictions. As mentioned before, there is a priori no obvious choice of a set of

functions to serve as continuous-time words. We will later conclude that we really want are Young measures (a sort of generalized function).

Here are the examples.

Time counter. For any alphabet Σ, a deducer which tracks the amount of time the signal spends inside a fixed subset $A \subseteq \Sigma$.

Delay. A transducer which, for any alphabet, outputs a signal on the same alphabet delayed in time by a fixed value τ. The output for the first τ time-units is a fixed constant.

Integrator. $\Sigma = [0, 1]$ or more generally, Σ may be a subset of a topological vector space. An integrator would simply integrate the signal over time, providing its value as an output signal. The value of the integral up to the current time is the only state information needed, with some arbitrary initial state.

Differentiator. A transducer which outputs the derivative of the input signal. Some sort of state information is needed about the infinitesimal past.

Alternation counter. A deducer for a discrete alphabet, say $\Sigma = \{0, 1\}$, which counts the number of changes in the input value.

Switching controller. The alphabet is discrete, say $\Sigma = \{0, 1\}$. We imagine an object which can be switched between two different physical behaviors, each represented by a vector field on the state space. At any given time, the input value determines which behavior is in effect, the state moving along the flow curve for the corresponding vector field. This example may be generalized for a continuous alphabet by considering a continuously parametrized family of vector fields.

1.2 Continuity and Compactness Principles

Now we examine in more detail how to state effectiveness criteria in terms of continuity and compactness. To begin, what do we mean exactly when we say *all* maps should be continuous? Consider the following particular maps associated to a continuous process. Note that whenever we say a map should be continuous, we are also implicitly asserting that its domain and codomain should have a specified topology.

Dynamics of the computation. The update rule which determines a new state given an old state and the intervening input should be continuous in both arguments. This map should be continuous.

Outcome map for a deducer. The map which associates states to outcomes. This map should be continuous. Note that this criterion almost makes traditional recognizers impossible: the input space is likely connected, so any continuous recognizer to two outcomes would be trivial.

Overall input-output map. For a deducer with a given start state, the continuity of this map follows from the above by composition. For a transducer, the map from input to output signals (for any given time interval and start state) should be continuous. We will define deducers and transducers precisely in Chapter 5.

Input as a function of time. We do *not* require this function to be continuous. It is not part of the computation, but rather is given to us. It would also be inperspicuous in the following way: there would be implicit state information not attached to the state space, because if the input takes on a certain value at a certain time, input would be required to have a matching left one-sided limit at that time.

Similarly, let us consider exactly which spaces we are asserting should be compact. Since points in spaces have to be specified by approximation, to say a space is compact is to say

that for any given degree of approximation, specifying a point requires only finitely much information.

The alphabet Σ. The alphabet should be compact.

The fixed-interval input space. By this we mean the space of all possible input signals over a given time interval. This space should be compact. Note that the discrete analogue comes for free in finite automata theory: if Σ is finite, then so is Σ^n for any n.

The unrestricted input space. By this we mean the space of all possible input functions over any time interval. We do *not* require this space to be compact. It is analogous to Σ^*.

The state space. The state space should be compact.

In discussing the examples, we sometimes will gloss over relatively innocuous non-compact state spaces such as the non-negative reals. The reason is convenience; in real-world problems real variables come with bounds and the example is easily patched by restricting the variable not to leave these bounds; or to cycle through. Future researchers may also find it sensible to allow non-compact spaces which are required to be smaller or more tractible than the input space in some other ways, such as being finite dimensional, or having a vector space structure. We will for the present remain focused on developing a continuous-time automata theory in parallel with finite automata theory.

1.3 Abstract Definition of an Automaton

The discussion so far leads us to the following formal definition.

Definition 1.1. A *continuous automaton* is a topological space S, the *state space*, together with a topological monoid M, the *input monoid*, and a continuous right action of M on S,

the *update rule*. A continuous right action of M on S is a continuous map

$$S \times M \to S$$

typically written as $s \cdot m$ for $s \in S$, $m \in M$, satisfying the *action law* or *law of causality*:

$$\begin{aligned} s \cdot 1_M &= s \\ (s \cdot m_1) \cdot m_2 &= s \cdot m_1 m_2. \end{aligned} \tag{1.1}$$

We say a continuous automaton is *compact* if its state space is.

Note that a version of this definition appeared in [8] (which does not consider continuous time), and that special cases include discrete-time continuous-space automata, as in that paper, and the classical theory of finite automata, when all sets are given the discrete topology and $M = \Sigma^*$.

The definition leaves us some flexibility in how the input signals are defined. A prototypical input monoid may be defined as follows. An element $u \in M$ is a measurable function $u : [0, \ell(u)] \to \Sigma$ where Σ is a fixed compact alphabet. (The choice of measurable functions over some other class of functions is not important now.) The monoid operation is concatenation:

$$uv : [0, \ell(u) + \ell(v)] \to \Sigma$$

$$(uv)(t) = \begin{cases} u(t) & 0 \le t < \ell(u) \\ v(t - \ell(u)) & \ell(u) \le t \le \ell(u) + \ell(v). \end{cases}$$

Here we have just shifted v over and joined it with u. This monoid together with Definition 1.1 could serve to formalize what we so far have called a "continuous-time process". This monoid is intended to be analogous to the free monoid Σ^* under concatenation. However, it does not satisfy all the compactness principles.

1.4 Compactness of M_ℓ

Let M_ℓ be the set of input signals of some fixed length ℓ. In other words M_ℓ is the set of continuous-time words of length ℓ, for whatever we decide to call a word. Our second compactness principle should now say that M_ℓ is a compact subspace of M. Unfortunately, almost no function space is even locally compact, no matter how compact the domain and codomain might be. In particular, neither Lebesgue measurable functions nor any of the other classes of functions mentioned earlier yields a locally compact space when endowed with any standard topology, including uniform convergence, L^p-like metrics, or convergence in measure.

We mentioned that the discrete analogue of this principle, finiteness of Σ^n, came for free, and now it appears we have no obvious way to achieve it at all. Nonetheless, we really would like to have this compactness and will devote special effort to finding a suitable monoid which satisfies this property. We insist for two reasons. First, theoretically, a glance at standard automata theory shows that almost all interesting results ultimately depend on the fact that Σ^n is finite, not just that Σ is finite. Second, it makes physical sense to require compactness. We are finitary beings–even if a physical device somehow did correctly process what is essentially an infinite amount of information in a finite time, we would have no way to verify that it did so correctly.

One way to get M_ℓ to be compact is to really restrict the class of functions, say to Lipschitz functions with Lipschitz constant less than some fixed bound K. This set of functions is equicontinuous and thus the Arzelà-Ascoli theorem makes it compact. However, we earlier mentioned drawbacks to assuming continuous input functions, and it seems to us that in real-world problems it is seldom reasonable assume a uniform Lipschitz bound.

We will instead solve the problem by introducing a new metric on functions and take the completion with respect to this metric, which will turn out to give us the space of Young measures. We can prove that the metric is totally bounded, so the completion is compact.

To motivate the necessity of this construction, consider why function spaces are not

compact: functions can have oscillations of unboundedly high frequency. For example, if $f_n : [0,1] \to \{0,1\}$ is the function which alternates 2^n times between the values 0 and 1 on intervals of equal width, then (f_n) has no convergent subsequence. Problems with high frequency oscillations are not limited to the realm of pure mathematics; for example, electronic devices which are sensitive to the frequency of their input malfunction at some point if the frequency is too high. For a device to be have uniformly continuously with respect to the new metric will be essentially to require the device to tolerate high frequency noise by averaging it out.

Before proceeding to formally develop the theory, let us briefly revisit the examples from Section 1.1. The time counter is fine. It will give an approximately correct value even if the input is shifted around or averaged out. The integrator (which generalizes the time counter) handles high-frequency input just fine, and is a prototype for a large class of examples based on differential equations, including the switching controller, we will prove can be defined on the compactification (Corollary 4.6).

The differentiator and alternation counter do not work as described. (These are the only two which we have to reject.) They are sensitive to oscillations in the input. For example, the derivative of a function can change dramatically and become not well-defined even if the function is given only a very small uniform bump, if that bump happens to have high-frequency noise in it. Similarly, there is no way the alternation counter can be interpreted as needing only a finite amount of information from any interval of input, unless we somehow know that the input literally cannot oscillate at more than a certain frequency. Mathematically, these operations cannot be meaningfully extended to the space of Young measures, our compactification of the input space. See Section 6.2 for further discussion.

The delay device does make sense for high-frequency input. See Examples 4.2 and 5.7. The delay is an interesting case for two reasons. First, it cannot easily be analyzed in the same way as the class of examples based on differential equations. Second, the fact that M_ℓ is compact is what allows the delay to be computed with compact state space, since it must

keep track of part of the input signal as state information.

1.5 Notes

So far I haven't mentioned previous work on the subject of continuous-time computation. There has actually been quite a lot of such work, but it has been somewhat scattered, taking place in a number of distinct clusters. Each cluster roughly corresponds to a research community with its own perspective and goals, which all seem to be at least slightly different (and sometimes very different) from each other and from my own goals outlined above. See [4] for a good overall survey.

A few examples stand out. Rabinovich [11] (following Trakhtenbrot) considers automata over continuous time but which nonetheless have finite (discrete) sets of states. Note that in the present model the input monoid would typically be connected, so since the action is required to be continuous, the state could never move between different connected components. In other words, this model allows continuous time and continuous space, or discrete time and any kind of space, but not continuous time and discrete space. See Section 6.3 for further discussion. Jeandel [8] goes the other way and considers discrete-time, continuous-space computation. Such a possibility is included in the present formalism but not emphasized. A much older line of work is represented by [12], which views continuous semigroup actions as "topological machines" (in some cases the word "automata" was used).

The concept of hybrid automata [1, 9, 7, 10] includes a direct axiomatization of the possibility of both discrete and continuous state evolution, with various possible models considered for the continuous part. Hybrid automata theory has been very successful in practical applications to real-world hybrid systems (systems which include interacting analog and digital components).

I am quite curious about the question of whether and how the theory of hybrid automata can be connected to the present work. The situation is unclear because although hybrid

automata theory explicitly allows for discrete state-transitions, their discreteness doesn't quite seem essential because they really represent shifts between different dynamical modes, in each of which the state evolves continuously. See Section 6.3.

I mentioned that my motivation for requiring continuity comes from constructive analysis in its various forms, and particularly the ubiquitous observation that effectively computable (in whatever sense) functions have to be continuous. Possible sources for this include Bishop's classic purely constructive text [3] or the (classical) theory of computable analysis—a good recent reference is [16]. Brouwer controversially claimed, as part of his intuitionistic philosophy of mathematics, that *all* functions must be continuous. Taken together, the constructivist arguments for continuity—the notion that the idea of a total discontinuous function on e.g. the reals is rather dubious—have influenced me considerably. In the course of developing Young measures as a metric completion, I realized that probably many notions of analysis, such as measurable functions, can easily be recast in sensibly constructive fashion, never mentioning discontinuous functions, which provides insight into what role those objects really play in mathematics. For example, the point of a measurable function is largely to be integrated against continuous test functions—itself a continuous operation. This sort of insight seems useful even if one doesn't want to adopt the full baggage of a fully constructivist mathematical foundation.

2 Preliminaries

2.1 Spaces, Continuity, Completions, Compactness

Here we will review some of the basic definition of topology that allow mathematical understanding of the continuous world. This review will be incomplete and almost absurdly quick, but we can pause to point out some issues that will be especially pertinent to continuous automata theory.

A *topological space* consists of an underlying set X and a topology on X. A topology on X associates to each point $x \in X$ a filter of subsets of X, each containing the point x. (A *filter* is a non-empty collection of subsets of X closed upwards and closed under finite intersections.) Sets in this filter are called the *neighborhoods* of X (not necessarily open). The neighborhoods of x define a sort of floating standard of closeness-to-x. This definition allows us to make precise the idea that a continuous function maps nearby points to nearby points: given $f : X \to Y$, where X and Y are topological spaces, we say f is continuous provided that whenever $f(x) = y$, for every neighborhood $V \ni y$, there is a neigborhood $U \ni x$ such that $f(U) \subseteq V$.

An *open* set is one which includes a neighborhood of all of its members. Topological spaces and continuity are commonly defined referring only to the notion of open sets. We can recover the neighborhood structure: a neighborhood of x is any set A with an open set in between $x \in U \subseteq A$. Continuity takes a formally simple but unintuitive form now: preimages of open sets are open. These definitions may give the impression that the notion

of *closeness* formalized by a topological space is uniform, that it would make sense to speak generally of two points being close together. It is not so. We can talk about closeness to a point x by membership in neighborhoods of x. We cannot speak of x and y being close, unless one of them is fixed and we refer to its neighborhoods. A function is continuous if and only if it is continuous at each point.

A *metric space* is an underlying set X and a metric $d : X \times X \to [0, \infty)$ on X. That d is a *metric* means

- d is *positive definite*: $d(x, y) = 0 \iff x = y$
- d is *symmetric*: $d(x, y) = d(y, x)$
- d satisfies the *triangle inequality*:

$$d(x, z) \leq d(x, y) + d(y, z).$$

These axioms simply formalize the intuitive idea that $d(x, y)$ is "the distance between x and y". Given a point x we may speak of $B_r(x)$, the ball of radius r centered at x, which is the set of points at distance at most r from x. Considering the positive-radius balls about x as generating a filter of neighborhoods for x (take the upwards closure), we get a topology associated to the metric.

Unlike a topology however, a metric *does* give a uniform standard of closeness. We can describe, for example, the set of pairs (x, y) at distance less than $\frac{1}{2}$. Given a function on metric spaces $f : X \to Y$, we say that f is *uniformly continuous* if there is a a modulus m, i.e. a function such that $m(\delta)$ converges to 0 as $\delta \downarrow 0$, such that

$$d(f(x_1), f(x_2)) \leq m(d(x_1, x_2)).$$

We can also describe continuity in terms of a metric: f is continuous if for each x there is a

modulus m_x such that

$$d(f(x), f(x')) \leq m_x(d(x, x')).$$

A standard example of a continuous function which is not uniformly continuous is $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = x^2$. Note how m_x has to depend on the magnitude of x.

Subsets of $\mathbb{R}^n$ under the usual Euclidean metric provide a rich and plentiful set of examples of metric spaces. Also important are infinite-dimensional function spaces, which we will discuss more later. Metric spaces are of course also good examples of topological spaces, but we should note that they enjoy special properties not all topological spaces have. Every metrizable topology is *Hausdorff*, meaning every pair of points has a pair of disjoint neighborhoods, and *first-countable*, meaning every neighborhood filter is countably generated. That metric spaces are first-countable implies that their topology is determined entirely by convergence of sequences.

Topological spaces are *homeomorphic* if there is a bijection which is continuous both ways. Metric spaces are *equivalent* if there is a bijection uniformly continuous both ways. Two metrics on a set are *equivalent* if the identity function gives an equivalence of metric spaces.[1]

In many contexts we have a useful basic intuition that adding more elements to a set makes it larger. (Even for infinite sets, we at least think of it as larger in the sense of having more things in it.) In the context of topology we have to revise this intuition. To see why, it is easier first to think about the reverse situation where we remove points from a space.

Recall that any subset $A \subseteq X$ can be given the *subspace* topology: the neighborhoods of a point are the old neighborhoods intersected with the subset. This topology has the naturally desirable property that any continuous function remains continuous when restricted to a subspace. However, it is not true that any continuous function on the subspace can be

[1] Equivalence of metric spaces retains the uniform information (information about what functions are uniformly continuous) but forgets any additional geometric information of the metric. In this thesis that's all we really want. There is a notion of a *uniform space*, which mathematically is the simplest and most general possible formulation of uniform information. But the additional generality and clarity it buys us is modest, whereas metric spaces are widely known and easy to define.

extended to the whole space. For example, $f(x) = 1/x$ is continuous on the nonzero real numbers, but cannot be extended continuously to all real numbers.[2] The space is actually smaller in the sense that fewer continuous functions are defined on it. Intuitively we can think of it as smaller because its points are closer together—points on either side of 0 previously had no relationship whatsoever, lying in distinct connected components of the space, but now they are joined together by their membership in neighborhoods of 0.

The general point is that a space X is in many (but not all) ways smaller than a dense subspace $D \subseteq X$. A continuous function from D to a Hausdorff space can be extended in at most one way to all of X. The typical case is that such a continuous function could not be extended, like $1/x$ above. To actually *enlarge* a space, we need to either remove open sets (making points further apart) or add points which are not in the closure of the original space.

The reader should therefore not be surprised when we later find we sometimes need to *add* new states to an automaton in order to get a Nerode-minimum automaton. The issue also comes up when we consider completions of spaces. In general, if X is a space, any way we can densely embed X in another space might be said to give a completion. A priori there is no canonical way to do this, but in the case of metric spaces there is a notion of completeness and a unique completion for any space. (The completion is unique for the metric, not for the topology.)

Working in a metric space, we can define the *diameter* of a set as the supremal distance between any two points. A sequence is *Cauchy* if its tails have arbitrarily small diameter. Every convergent sequence is Cauchy. Conversely, we might expect a Cauchy sequence to converge; its elements are getting closer and closer together. In general, a Cauchy sequence may fail to converge; for example, consider the sequence $(\frac{1}{n})$ in the nonzero real numbers. It would converge if we hadn't removed its limit! A *complete* metric space is one in which every Cauchy sequence converges. Using equivalence classes of Cauchy sequences, we can

[2] Arguably, $1/x$ should be regarded as discontinuous even as a partial function on the real numbers, having no limit at 0.

abstractly construct a *completion* for any metric space, a complete metric space into which the original densely embeds. As an example, the space of real numbers is often constructed as the completion of the rational numbers under the usual metric. The completion of a metric space, with the embedding, is unique up to a unique isometry which commutes with the embeddings. In other words, it is as well-defined as we could possibly hope.

Consider the open interval $(0, 1)$ as topological space. We have given this space as a subspace of $[0, 1]$, but up to homeomorphism we could just as easily consider it to be a copy of $\mathbb{R}$ or as a circle with one point removed. Each of these three viewpoints suggests a different metric on $(0, 1)$ and all of them give the same topology. However they are not equivalent as metric spaces, and their completions are different. In the first case (copy of $\mathbb{R}$), the space is already complete. In the second case, the completion is $[0, 1]$, and in the third case the completion is the circle.

Uniformly continuous functions on a dense subset of a metric space have a unique one extension to the entire space. (For the extension to be defined, we may have to pass to the completion of the codomain if it is not already complete, but the construction is canonical.) Contrast with the situation described above for arbitrary continuous functions. For example in case we were completing the open interval to the circle, the metric *knew* that the ends of the interval were close together: $d(\frac{1}{n}, 1 - \frac{1}{n})$ converges to zero. So the function $f : (0, 1) \to [0, 1]$ defined by $f(x) = x$ could not be extended, but that is because it is not uniformly continuous. If we are concerned only with uniformly continuous functions defined on the space, the completion is not actually making the metric space smaller (and certainly not larger). The space is not really even changing; we are just calling attention to certain points in the space which were already implicitly present.

Compactness is a natural smallness criterion for topological spaces which is (imperfectly) analogous to finiteness in the discrete world. A topological space is *compact* if every open cover has a finite subcover. Compact spaces enjoy some properties which resemble finiteness: they are closed under taking finite unions, images by continuous functions, and finite prod-

ucts. Subspaces of a compact space are usually only compact if they are closed, however. To be precise, every closed subset of a compact space is compact, and the converse holds in a Hausdorff space. Also, *arbitrary* products of compact spaces are compact; to understand this fact requires understanding the nuances of the definition of the product topology in this case.

In a discrete setting, we also have the useful fact that the set of functions $A \to B$ is finite if A and B are finite. There is no easy way to lift this fact to a continuous setting. Merely defining a topology on a function space demands some effort and care, and the usual ways of doing so do not preserve compactness. Young measures represent a way to work around this problem, though Young measures are strictly speaking not functions.

Compact metric spaces are always complete. Furthermore, the concept can be usefully factored: a metric space is compact if and only if it is complete and totally bounded. A *totally bounded* space is one which for every $\varepsilon > 0$ is the ε-neighborhood of a finite subset. Equivalently: a metric space is totally bounded if and only if every sequence has a Cauchy subsequence. A metric space is compact if and only if every sequence has a convergent subsequence. Every continuous function defined on a compact metric space is uniformly continuous.

2.2 The Compact-Open Topology

The compact-open topology is often the appropriate topology to put on a set of continuous functions between two topological spaces.

Definition 2.1. Let X and Y be topological spaces, and let $C(X, Y)$ be the set of continuous functions $X \to Y$. The *compact-open topology* on $C(X, Y)$ is defined by the following subbasic open sets:

$$V(K, U) = \{f \in C(X, Y) : f(K) \subseteq U\}$$

where $K \subseteq X$ is compact and $U \subseteq Y$ is open.

In case Y has a metric or uniform structure, the compact-open topology may be thought of as the topology of uniform convergence on compact sets. It generalizes the product topology, which arises when X is discrete. The basic properties we need are summarized in the following proposition.

Proposition 2.2. Suppose X is a locally compact Hausdorff space and Y is any topological space. Let $C(X, Y)$ be the space of continuous functions $X \to Y$ in the compact-open topology.

- *Application.* The application map $C(X, Y) \times X \to Y$ given by $(f, x) \mapsto f(x)$ is continuous.

- *Parametrization.* Suppose P is a topological space and $g : P \times X \to Y$ is a continuous map. (We think of P as continuously parametrizing a family of functions in F.) Then the curried map $h : P \to C(X, Y)$ defined by $h(p)(x) = g(p, x)$ is continuous.

2.3 Measure Theory

In this section we give a similarly hasty review of measure-theoretic concepts needed. Our use of these concepts will revolve almost entirely around the KW metric to be defined in the next section.

Recall that the *Borel* subsets of a topological space X form the smallest class of subsets extending the open (and closed) subsets and closed under countable unions and complements. The class of Borel subsets of X we denote $\mathcal{B}(X)$. Virtually every set that can be explicitly described in practice is Borel.

A *positive Borel extended measure* or simply *positive measure* μ on a space X is a function $\mathcal{B}(X) \to [0, +\infty]$ such that $\mu(\emptyset) = 0$ and μ is countably additive:

$$\sum_{n=0}^{\infty} \mu(A_n) = \mu\left(\bigcup_{n=0}^{\infty} A_n\right).$$

In particular, $\mu(A \cup B) = \mu(A) + \mu(B)$ whenever A, B are disjoint Borel sets. ("Extended" clarifies that $+\infty$ is a possible value, while "positive" clarifies that a set cannot have negative measure.)

Intuitively, a measure is a distribution of mass over a set. The definition allows the possibility of non-atomic measures; these are measures which assign zero mass to each singleton, yet they may assign positive mass to the entire space. A fundamental example is Lebesgue measure λ on the real line; we have for each interval $\lambda([a, b]) = b - a$, so λ generalizes the notion of length. The *delta-mass* at a point $x \in X$ is the measure δ_x defined as follows: $\delta_x(A) = 1$ when $x \in A$, and $\delta_x(A) = 0$ otherwise.

A central use of positive measures is to define integration. A function $X \to \mathbb{R}$ is *Borel measurable* or simply *measurable* if the preimage of every interval (equivalently, every Borel subset of $\mathbb{R}$) is a Borel subset of X. Virtually every function that can be explicitly described in practice is Borel measurable. Integration can then defined for certain large classes of Borel measurable functions; these will certainly include nonnegative functions (if $+\infty$ is allowed as a value) and bounded functions supported on a set of finite measure. Integration will have its usual properties:

- Expected value on characteristic functions:

$$\int_X \chi_A \, d\mu = \mu(A)$$

  where $\chi_A(x) = 1$ when $x \in A$ and $\chi_A(x) = 0$ otherwise.

- Linearity:

$$\int_X (cf + g) \, d\mu = c \int_X f \, d\mu + \int_X g \, d\mu$$

  for $c \geq 0$ and f, g measurable functions of the specified type.

- Monotonicity: whenever $f \leq g$,

$$\int_X f \, d\mu \leq \int_X g \, d\mu.$$

and under suitable regularity conditions, integration is uniquely determined by these properties. (Then $\int_A f \, d\mu$ would be defined to be $\int_X f\chi_A \, d\mu$. Monotonicity plays a role similar to uniform continuity in ensuring uniqueness.)

A *bounded signed Borel measure* or simply *signed measure* μ on X is a function $\mathcal{B}(X) \to \mathbb{R}$ such that $\mu(\emptyset) = 0$ and μ is countably additive. ("Bounded" clarifies that $+\infty$ is not allowed as a value. Given σ-additivity, it can be proven μ must be bounded in the usual sense for a real-valued function.) Note that the signed measures on X form a vector space over $\mathbb{R}$ with regard to pointwise (i.e. setwise) addition and multiplication by constants.

Signed measures allow some sets to be assigned negative mass; an alternate intuition is that of a distribution of electric charge. One way to get examples of signed measures is by integration: if μ is a bounded positive measure, and $f : X \to \mathbb{R}$ is an integrable function, then $A \mapsto \int_A f \, d\mu$ is a signed measure. Many but not all concepts for positive measures may be defined similarly for signed measures.

A key fact is the Hahn decomposition theorem, which says that for any signed measure μ on X there exists Borel sets H_+ and H_- such that the restriction of μ is a positive or negative measure respectively (still bounded, of course). It follows that μ has a unique *Jordan decomposition* $\mu = \mu_+ - \mu_-$ where μ_+ and μ_- are each bounded positive measures. Using the Jordan decomposition we can easily define integration:

$$\int_X f = \int_{\mu_+} f - \int_{\mu_-} f. \tag{2.1}$$

Given a Jordan decomposition for μ, the *total variation measure* of μ is defined as $|\mu| = \mu_+ + \mu_-$. The *total variation* of μ is the number $|\mu|(X)$. Total variation defines a norm on the space of signed measures on X. The topology induced by the total variation norm is too

strong for our purposes. This topology ignores the underlying topology of X: for example, the delta-masses at each point of X would form a discrete subspace of the signed measures. (We would prefer for this subspace to be a homeomorphic copy of X.)

If Y is another topological space (possibly $Y = X$) and $f : X \to Y$ is a Borel measurable function (pre-image of each Borel set is Borel), then we can use f to push forward measures from X to Y. Suppose μ is a (signed or positive extended) measure on X. The pushforward $\nu = f_*(\mu)$ is a measure on Y defined by $\nu(B) = \mu(f^{-1}(B))$. Intuitively, we are reallocating the mass of μ according to the function f, to get ν.

2.4 The Kantorovich-Wasserstein Metric

Definition 2.3. Let X be a compact metric space. Let μ, ν be bounded signed Borel measures on X such that $\mu(X) = \nu(X)$. A *signed coupling* of μ and ν is a Borel measure on $X \times X$ such that $\pi_{1,*}(\gamma) = \mu$ and $\pi_{2,*}(\gamma) = \nu$, or equivalently:

$$\gamma(A \times X) = \mu(A)$$
$$\gamma(X \times B) = \nu(B)$$

for each $A, B \subseteq X$ Borel. The *cost* of a coupling γ is

$$\text{cost}(\gamma) = \int_{X \times X} \mathrm{d}_X d\,|\gamma|$$

where $|\gamma|$ is the total variation measure of γ. The *Kantorovich-Wasserstein metric* or *KW metric* is defined as $\mathrm{d}_{\mathrm{KW}}(\mu, \nu) = \inf_\gamma \text{cost}(\gamma)$, the infimal cost of couplings of μ and ν.

The KW metric is also called the "earth mover's metric": we can think about it intuitively as the cost of moving dirt around to get from one measure (distribution of dirt) to the other, if a particular movement costs the amount of dirt times distance moved. The reader should take care to remember that the KW metric is only a metric on the set of measures of a single

fixed total variation; that is, $\mathrm{d_{KW}}(\mu, \nu)$ is defined only if $\mu(X) = \nu(X)$. (Otherwise there are no couplings.)

In case X is a finite set with the discrete metric, signed measures on X correspond to functions $X \to \mathbb{R}$ and further to elements of $\mathbb{R}^n$, where $n = |X|$. In this case, the KW metric, where defined, is exactly one-half times the L^1 metric. If X is given any other metric, the KW metric will remain bilipschitz equivalent.

We will often consider cases where μ and ν are finitely supported (even if X itself is not finite). Each integral becomes a finite sum, so this case is relatively easy to visualize. To visualize a coupling, draw the complete bipartite graph between two copies of Y and label the edge (x, y) with $\gamma(\{(x, y)\})$, which we may abbreviate in this case as $\gamma(x, y)$. The vertices on either side may also be labeled by μ and ν respectively. Then the projection condition says that the numbers on all edges entering or leaving a vertex must equal the number on that vertex.

Most of the literature on optimal transport considers only positive measures and couplings. (However, the cost function, which for us is just the metric on X, is sometimes allowed to be nonpositive.) The following proposition shows that allowing couplings to be negative does not change too much.

Proposition 2.4. Given bounded signed measures μ, ν and a coupling γ between them, there is a coupling γ' of the same cost such that γ' is positive except possibly on the diagonal $\{(x, x) : x \in X\}$.

Proof. Intuitively, the idea of proof is that moving negative measure from A to B should be like moving positive measure from B to A. To make the idea precise, the values of the coupling on the diagonal have to be corrected.

As a warm-up case, assume that μ, ν, and γ are finitely supported, and assume γ never assigns nonzero measure to both (x, y) and (y, x). Then we can define γ' in the following way. Whenever $\gamma(x, y) > 0$, then γ' will agree with γ on (x, x), (x, y), (y, x), and (y, y). On the other hand if $\gamma(x, y) < 0$, then let

- $\gamma'(y,x) = -\gamma(x,y)$
- $\gamma'(x,y) = 0$
- $\gamma'(x,x) = \gamma(x,x) + \gamma(x,y)$
- $\gamma'(y,y) = \gamma(y,y) + \gamma(x,y)$.

Due to the correction on the diagonal, the measure assigned to $\{x\} \times \{x,y\}$ is unchanged, as is $\{y\} \times \{x,y\}$, so γ' remains a coupling between μ and ν. Furthermore, since $d(x,y) = d(y,x)$ and $d(x,x) = d(y,y) = 0$, the cost is unchanged.

For the general case we will use a similar idea. First we claim it is enough to prove the statement for couplings γ which are entirely negative. To prove this sufficiency, let μ, ν be arbitrary signed measures, $\mu(X) = \nu(X)$, and let γ be a coupling of μ with ν.

Let $\gamma = \gamma_+ - \gamma_-$ be the Jordan decomposition of γ. Then we have by (2.1)

$$\text{cost}(\gamma) = \int_{X\times X} \mathrm{d}_X d\,|\gamma| = \int_{X\times X} \mathrm{d}_X d\gamma_+ + \int_{X\times X} \mathrm{d}_X d\gamma_-.$$

On the other hand, if we replace $-\gamma_-$ in the sum $\gamma = \gamma_+ - \gamma_-$ by some new measure γ'_- as per our assumption, then the measure $\gamma_+ + \gamma'_-$ has the same cost. It also remains a coupling of μ and ν: the projections of γ'_- agree with those of γ (even though they likely are not μ and ν), and projections are linear. So the claim is proved.

It remains to prove the case where γ is a negative coupling of arbitrary signed measures μ and ν. Let $P : X \times X \to X \times X$ be defined by $P(x,y) = (y,x)$. Consider the measure $-P_*(\gamma)$: this measure is a positive coupling of $-\nu$ with $-\mu$. Now let α be the coupling of $\mu+\nu$ with itself by pushing $\mu+\nu$ forward onto the diagonal by the diagonal map $X \to X \times X$. In other words, α is the unique zero-cost coupling of $\mu + \nu$ with itself. Now let

$$\gamma' = \alpha - P_*(\gamma).$$

Then $\text{cost}(\gamma') = \text{cost}(-P_*(\gamma)) = \text{cost}(\gamma)$, and γ' has projections $\mu+\nu-\nu = \mu$ and $\mu+\nu-\mu = \nu$

respectively, as desired. □

This observation makes use of the fact that d is a metric, in the following way. Moving negative measure from x to $y \neq x$ is equivalent to moving positive measure y to x, *provided* we make the appropriate adjustment to the amount of meausure left at x, and at y (i.e., $\gamma(x,x)$ and $\gamma(y,y)$). As the cost of leaving measure in place is zero (d is a metric), this adjustment does not affect the cost. (This argument works in the non-finite case, if one is careful, by first taking the Hahn-Jordan decomposition of the off-diagonal part of the coupling and then proceeding as above on measurable rectangles which do not intersect the diagonal.)

A special kind of coupling is given by a transport function $f : X \to X$, which says for each x, in effect, "move all this measure to $f(x)$".

Observation 2.5. If $f : X \to X$ is a Borel function such that $\pi_*(\mu) = \nu$, then it generates a coupling γ_f of μ and ν:

$$\gamma_f(A \times B) = \mu(\{x \in A : f(x) \in B\}).$$

An equivalent description is given by considering the inclusion map induced by f of X into the graph of f. Then γ_f is the pushforward of μ along this inclusion.

Not all couplings are given by transport functions. Sometimes it is necessary to split measure up and move it to more than one place.

Couplings have a composition operation, which may be used to prove the triangle inequality for $\mathrm{d_{KW}}$.

Definition 2.6. Suppose γ_1 couples μ and ν, and γ_2 couples ν and ρ, all Borel measures on X. Then the *composition* $\gamma_1 \circ \gamma_2$ is defined as follows. First define a measure β on $X \times X \times X$:

$$\beta(A \times B \times C) = \frac{\gamma_1(A \times B)\gamma_2(B \times C)}{\nu(B)}.$$

Then

$$(\gamma_1 \circ \gamma_2)(A \times C) = \beta(A \times X \times C).$$

Proposition 2.7. Let γ_1, γ_2 be couplings as in Definition 2.6. Then

$$\text{cost}(\gamma_1 \circ \gamma_2) \leq \text{cost}(\gamma_1) + \text{cost}(\gamma_2).$$

In case the couplings which are being composed split up measure a lot, the composed coupling will smear over all branches of splittings of γ_1 and γ_2.

Observation 2.8. If γ_1 and γ_2 are given by functions, then composition of couplings agrees with composition of functions.

For more details on composition of couplings, see [2, Lemma 5.32 and Remark 5.3.3].

We will later need the following fact about the finite case.

Proposition 2.9. Suppose μ and ν are bounded integer-valued measures on a finite metric space Y, and γ is a coupling of μ, ν. Then there is an integer-valued coupling γ' with $\text{cost}(\gamma') \leq \text{cost}(\gamma)$.

Essentially, Proposition 2.9 says that there is no advantage to splitting up measure in a coupling beyond what is called for by the measures. The proposition follows from Birkhoff's Theorem that the extreme points of the set of doubly stochastic matrices are exactly the permutation matrices. See [14] (the exercise at the end of the introduction[3]).

2.5 Notes

For general topology background, see [17]. For measure theory, [6] is a possibility.

The KW metric goes back a long way, as does the associated topic of optimal transport. See Villani's book [15] for a general introduction. Regarding terminology, see particularly

[3] As suggested by [2].

the bibliographic notes in [15, Chapter 6]. The two standard names here are basically Wasserstein and Kantorovich; the former has the appeal of being already very widely used, while the latter gives credit to the most important contributor (and apparently the second earliest, after Gini). I chose the name Kantorovich-Wasserstein metric as a compromise between these considerations. The descriptive name "earth mover's metric" also appears, especially in computer science (especially for comparing images).

The definition here allowed arbitrary signed measures, by virtue of allowing signed couplings. So in the case of total variation zero, we have a norm. According to the same reference [15, Chapter 6] this norm has often been called the Kantorovich-Rubinstein norm, and Kantorovich himself was the one who pointed out it can be extended to a norm on the space of signed measures, a norm which very nicely metrizes the topology of narrow (also called weak or weak*) convergence of measures. I do not know of a good English-language exposition of this norm as a norm (not merely a metric on probability measures, as it is normally described). For the reader who wants to track this down (and does not speak Russian), the bibliography of [15] seems a plausible starting point.

3 Young Measures

In Section 3.1 we will give a leisurely conceptual introduction to Young measures, saving the formal development for the rest of the chapter. The Young measures on a fixed interval, say $[0, 1]$, form a compact space which may be regarded as a compactification of the set of measurable functions, as we alluded to in Section 1.4. Young measures were introduced as "generalized curves" by Young—see [18]—originally defined as weak limits of sequences of functions in an appropriate dual space. The modern approach is to define Young measures as certain measures on the product space, such as $[0, 1] \times \Sigma$.

Here we will adopt a variant of the measure-theoretic approach. We define the space of Young measures as the completion of the space of a function space in a certain metric. We will take advantage of a limited amount of measure theory in defining and understanding this metric. However, we will establish almost all results by using uniform continuity arguments on suitable dense subsets, like rational step words, that can be understood in a finite way. The approach has the advantage of being mostly self-contained and relatively concrete.

3.1 Introduction to Young Measures

Let Σ be a compact metric space. As a simple nontrivial example, we may imagine $\Sigma = \{0, 1\}$. We are going to build a notion of continuous-time words of length 1 by starting with a modest class of functions $[0, 1] \to \Sigma$, defining a metric which we call the *Young metric*, and taking the completion of that metric. The class of functions can be as small as the rational

step functions (functions which are constant on a partition of $[0,1]$ into rational intervals), though any L^1-dense subset of the measurable functions would work just as well. Recall that the completion of a metric space will be compact if and only if the space is totally bounded. So our goal is to ensure that the metric we define is totally bounded on the rational step functions.

Let us motivate the definition of the metric by considering particular examples of pairs of rational step functions. Given such a pair, there must exist N such that the functions are constant on intervals of the form $(k/N, (k+1)/N)$. These are just discrete words of length N, scaled down to have "letter width" $1/N$ and thereby fit in the interval $[0,1]$ as continuous-time words. How different are these words?

$$\texttt{011}\mathbf{1}\texttt{11001011100}\mathbf{0}\texttt{111010111011}$$

$$\texttt{011}\mathbf{0}\texttt{11001011100}\mathbf{1}\texttt{111010111011}$$

They differ in two places, so their Hamming distance is 2. Say we scale that down appropriately for the continuous-time words in M_1, and their distance would be $2/N$. (In this example $N = 32$.) If we defined our metric using a scaled Hamming distance like this, we would end up essentially with an L^1 metric. In fact, we need a weaker topology, which is to say that at least some words need to be closer together than they were before. Consider these:

$$\texttt{01010101010101010101010101010101}$$

$$\texttt{10101010101010101010101010101010}$$

These words have the maximum possible Hamming distance. But we want to say they're actually close together. The reason is this: we could easily have some noise in our measurements of time that would make these words totally indistinguishable. Or to put it another way, if we look at the values of these words on any subinterval of time, we are going to get about the same number of 0s and 1s regardless of which word we look at. To be precise, the difference in occurrence of any letter on any subinterval is at most 1, so let's say these words have distance $1/N$. A less obvious example:

0010010101000010100100100010100 0

0010100011001000010010010010 0

These words have a scaled Hamming distance of $13/N$, but the new proposal gives them distance just $2/N$: on any interval we can find, the difference in the number of 1s (equivalently, 0s) occuring in each word is at most 2. Note that on any subinterval we get about two-thirds 0s and one-third 1s.

We need to be precise about the "difference in occurrence" of values on a subinterval, but unfortunately this issue is more complicated in the case where Σ is not discrete. For example, if $\Sigma = [0, 1]$ and we have the constant functions $u(t) = 0.5$ and $u(t) = 0.501$, these functions should also be close to each other. At this point we will introduce some measure theory into the picture.

When comparing two functions u and v on an interval $[s, t]$, we only mean to look at their values on that interval, not where the values occur. So we look at the pushforward measures $u_*(\lambda_{[s,t]})$ and $v_*(\lambda_{[s,t]})$ where $\lambda_{[s,t]}$ is Lebesgue measure on $[s, t]$. If u and v are piecewise constant, these measures are a finite sum of point-masses at the values of the functions u and v. The mass at each value is the total width of intervals on which the function took that value. In symbols,

$$u_*(\lambda_{[s,t]})(\{\sigma\}) = \lambda(\{r \in [s, t] : u(r) = \sigma\}). \tag{3.1}$$

To compare these measures, particularly in case of non-discrete Σ, we will use the Kantorovich-Wasserstein or KW metric. The KW metric is the infimal cost of transforming one measure into another by moving measure around: to move m measure over a distance d costs $m \cdot d$. If the measures are finite sums of point masses, then only finitely many discrete moves are needed. (See Definition 2.3.)

To keep the formalism as simple as possible, note that a simple triangle-inequality argument shows that we can restrict our attention to initial subintervals $[0, s]$: if we want to

consider $[s,t]$ we can compare on both $[0,s]$ and $[0,t]$.

With that in mind, we can define the *accumulant* of a measurable function $u : [0,1] \to \Sigma$. The *accumulant* of u, denoted $\overline{u}$, is the measure-valued function defined by

$$\overline{u}(s) = u_*(\lambda_{[0,s]}) \tag{3.2}$$

or equivalently

$$\overline{u}(s)(X) = \lambda\left(\{t \in [0,s] : u(t) \in X\}\right). \tag{3.3}$$

(See Definition 3.1.) So the accumulant $\overline{u}$ is a measure-valued function of time. Roughly speaking, it tells us, up to this time, how much time the function u has spent at each value so far.

Then the Young metric will be defined as

$$d(u,v) = \sup_{s\in[0,1]} \mathrm{d_{KW}}(\overline{u}(s), \overline{v}(s)). \tag{3.4}$$

Note that this metric is technically only defined on equivalence classes of functions up to almost-everywhere equality. (We take this quotient even in the case of rational step functions, where it amounts to ignoring the endpoints.) We are quite happy with this, however, because it renders moot any concern of endpoint conventions. As mentioned in the introduction, to achieve compactness of M_ℓ we have to give up on the idea that the inputs are functions in a strict sense.

We will see (Proposition 3.3) that we that the Young metric is indeed totally bounded on the measurable functions (in particular on the rational step functions).

3.2 The Young Metric

For the remainder of the chapter, fix a compact metric space Σ. A *word* will be an equivalence class of Lebesgue measurable functions $[0,\ell] \to \Sigma$ for some $\ell \geq 0$. After this section, we may

also use the word *word* to refer to a Young measure on $[0, \ell]$. Often we consider the case $\ell = 1$ for simplicity.

Definition 3.1. Let $u : [0, \ell] \to \Sigma$ be a word, and let $S \subseteq [0, \ell]$ be a measurable set. The *accumulation* of u on S is

$$\mathrm{acc}_S(u) = u_*(\lambda \restriction S),$$

that is,

$$\mathrm{acc}_S(u)(A) = \lambda(\{t \in S : u(t) \in A\}).$$

The *accumulant* of u is the function $\overline{u}$ defined by

$$\overline{u}(t) = \mathrm{acc}_{[0,t]}(u).$$

Definition 3.2. Given words $u, v : [0, \ell] \to \Sigma$, the Young metric is defined by

$$\mathrm{d}(u, v) = \sup_{t \in [0,1]} \mathrm{d}_{\mathrm{KW}}(\overline{u}(t), \overline{v}(t)).$$

Proposition 3.3. The set of words of length ℓ is totally bounded in the Young metric.

Proof. Assume $\ell = 1$; routine modifications will cover the general case. Let $\varepsilon > 0$. We will choose some large $N > 0$ and a finite approximating set C whose words are constant on intervals of the form $[k/N, (k+1)/N]$, and whose values lie in a finite approximating set $A \subseteq \Sigma$. We can assume the $\varepsilon/3$-neighborhood of A is all of Σ, and then it will turn out that $N > \frac{3|A| \operatorname{diam}(\Sigma)}{\varepsilon}$ is large enough. (Assume without loss of generality that we are in the case $\operatorname{diam}(\Sigma) \geq 1$.) Each element $a \in A$ will be responsible for approximating all values in some measurable set $B_a \subseteq \Sigma$, contained in the ball $B_{\varepsilon/3}(a)$, and such that together the B_a's partition Σ.

Now given a word $u \in F_0$, let us approximate it by some word $v \in C$. While u can change value in any way at any time, v can only take on values in A and can only change values at multiples of $1/N$. Proceeding inductively from left to right, we track the amount of time

u has spent in each B_a, and choose the value of v on any interval to correct the biggest shortfall of v relative to u. That is, we choose the values of v to minimize

$$\max_{a \in A} \overline{v}(k/N)(B_a) - \overline{u}(k/N)(B_a)$$

for each k in order. Because $\overline{v}(k/N)$ and $\overline{u}(k/N)$ have the same total measure, we will never be forced to choose a value such that this difference is greater than $1/N$. (It could only happen if somehow $\overline{v}((k-1)/N)$ was bigger than $\overline{u}(k/N)$ on every set B_a, which is impossible.)

Now we just need to show that v is indeed close to u, that is, $\mathrm{d_{KW}}(\overline{u}(s), \overline{v}(s)) \leq \varepsilon$ for any s. We can achieve this bound in three steps:

1. Round s down to the nearest k/N. We are ignoring at most $1/N$ measure here.

2. Move all measure in each B_a to a. We are moving at most 1 total measure by at most distance $\varepsilon/3$.

3. Anywere there is excess positive measure at a point a, move it to some other point a' which has the opposite discrepancy. By the observation, we are moving at most $1/N$ measure for each a here, so a total of $\leq |A|/N$ measure by a distance at most $\mathrm{diam}(\Sigma)$.

In summary,

$$\begin{aligned} \mathrm{d_{KW}}(\overline{u}(s), \overline{v}(s)) &\leq \frac{1}{N} + 1 \cdot \frac{\varepsilon}{3} + \frac{|A| \operatorname{diam}(\Sigma)}{N} \\ &\leq \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} \leq \varepsilon. \end{aligned}$$

□

Since the Young metric is totally bounded, its completion is compact and represents a suitable candidate for the space of possible inputs over a fixed time interval. However, we are not done with our study of this metric. The criterion that two words are nearby in the Young metric as defined above turns out to be difficult to use in practice for uniform

continuity arguments. We can address that issue by introducing equivalent variants of the Young metric without that defect. Exhibiting variants of the metric should also help with general understanding. First we will show that it doesn't really matter whether we use a supremum or integral in the definition of the Young metric. (In the next section, we will introduce the chain Young metric, which is more substantially different and will be important later.)

Definition 3.4. Let $u, v : [0, \ell] \to \Sigma$ be measurable functions. Then the *area Young metric* $\mathrm{d_A}(u, v)$ is

$$d(u, v) = \int_0^\ell \mathrm{d_{KW}}(\overline{u}(s), \overline{v}(s))ds.$$

Recall that two metrics are equivalent if the identity map is uniformly continuous both ways. If the identity is uniformly continuous only one way, we say that the latter metric is (nonstrictly) *weaker*, as points may be (in some topological sense independent of the actual values of the metric) closer together.

Proposition 3.5. Given ℓ and Σ, the Young metric is equivalent to the area Young metric.

Proof. For one direction, note that $\mathrm{d_A}(u, v) \leq \ell \mathrm{d}(u, v)$:

$$\int_0^\ell \mathrm{d_{KW}}(\overline{u}(s), \overline{v}(s))ds \leq \ell \cdot \sup_{s\in[0,1]} \mathrm{d_{KW}}(\overline{u}(s), \overline{v}(s)).$$

For the other direction, first observe that since accumulants add new measure at a rate of 1, for any measurable functions u, v, the function $\alpha_{u,v}$ given by $\alpha_{u,v}(s) = \mathrm{d_{KW}}(\overline{u}(s), \overline{v}(s))$ must be $2D$-Lipschitz where $D = \mathrm{diam}(\Sigma)$. Then we claim the following inequality:

$$\int_0^\ell \alpha_{u,v}(s)ds \geq \frac{1}{4D} \left(\sup \alpha_{u,v}(s)\right)^2.$$

To prove the inequality: suppose $\alpha_{u,v}$ achives its maximum $M = \sup \alpha_{u,v}(s)$ at time r. Picture the graph of $\alpha_{u,v}$. By the Lipschitz restriction, $\alpha_{u,v}$ must lie above the right triangle between $(r, 0)$, (r, M), and $(r - \frac{M}{2D}, 0)$.

Rewriting the inequality, we have

$$\sup \alpha_{u,v}(s) \leq 2D^{1/2} \left(\int_0^1 \alpha_{u,v}(s) ds \right)^{\frac{1}{2}},$$

that is,

$$d(u,v) \leq 2D^{1/2} \mathrm{d_A}(u,v)^{\frac{1}{2}}. \qquad \square$$

3.3 The Chain Young Metric

The chain Young metric is more combinatorial than the Young metric and is defined only for rational step words.

Definition 3.6. A *rational step word* $[0,\ell] \to \Sigma$ is a word which is constant on each interval of a rational partition of $[0,\ell]$.

By taking a common denominator, we may assume the partition is equal-length: $0 < 1/N < 2/N < \cdots < \ell - 1/N < \ell$ for some large N. So rational step words are like usual discrete words in Σ^* that have been scaled down by a precise amount.

Definition 3.7. Let $u, v : [0,\ell] \to \Sigma$ be rational step words. The *chain Young metric* $\mathrm{d_C}(u,v)$ is defined as follows. $\mathrm{d_C}(u,v)$ is the infimal total cost of a finite chain connecting u to v by push and pull moves, defined as follows. Before the move, we have a rational step word $w : [0,\ell] \to \Sigma$; after, we will have $w' : [0,\ell] \to \Sigma$. All intervals are rational.

- *Vertical push.*[1] On interval $I = [t, t+a]$ where w is constant, change its value: $w \restriction I = \sigma_1$, $w' \restriction I = \sigma_2$. Cost: $a\mathrm{d}_\Sigma(\sigma_1, \sigma_2)$.

- *Neighbor swap.* On consecutive intervals $I_1 = [t, t+a]$, $I_2 = [t+a, t+2a]$, swap the

[1]The "horizontal" vs. "vertical" terminology in this chapter comes from pictures I often draw of $[0,\ell] \times \Sigma$ in which Σ is represented by the vertical axis. So a "vertical push" means pushing something inside Σ.

values of w:

$$w \restriction I_1 = \sigma_1 \qquad w \restriction I_2 = \sigma_2$$
$$w' \restriction I_1 = \sigma_2 \qquad w' \restriction I_2 = \sigma_1.$$

Cost: a^2.

Checking that the chain Young metric is indeed a metric is routine. The triangle inequality comes for free (more or less) by concatenating chains. The only slightly tricky point is to verify that distinct points have positive distance: consider that there is a minimum cost required to make progress either vertically or horizontally on a single interval. We make no guarantee that the infimum is realized, but we will be content to work with chains which approximate the distance to within epsilon.[2]

Theorem 3.8. *On rational step words, the chain Young metric is equivalent to the Young metric.*

We will devote most of the rest of the section to proving Theorem 3.8. The chain Young metric exhibits in a particularly clear manner the difference between the Young metric and the L^1 metric: the L^1 metric on rational step words would be obtained exactly if we omitted neighbor swaps from the definition. The following proposition establishes the easier of the two directions.

Proposition 3.9. For rational step words $u, v : [0, \ell] \to \Sigma$, the area Young metric is weaker than the chain Young metric.

Proof. We can actually prove a Lipschitz relationship. It is enough to find a constant K such that whenever w and w' are separated by one step of cost C, $\mathrm{d_A}(w, w') \leq KC$.

[2]If values in Σ have irrational distances, it is possible the infimum is not realized. If not, we conjecture that it is realized. But it does not seem worth the effort to make that change since it makes no difference for our goals.

- Suppose the step is a vertical push on interval $[t, t+a]$ from σ_1 to σ_2. Then the cost is $C = a\mathrm{d}_\Sigma(\sigma_1, \sigma_2)$, which is also the KW-difference in accumulation on any interval containing $[t, t+a]$: the mass times the distance. So, $\mathrm{d}_{\mathrm{KW}}(\overline{w}(s), \overline{w'}(s)) \leq C$ for all $s \in [0, \ell]$. Considering this step, a constant of ℓ is sufficient.

- Suppose the step is a neighbor swap of σ_1 and σ_2 on $[t, t+a], [t+a, t+2a]$. Then $\overline{w}(s) = \overline{w'}(s)$ except for $s \in [t, t+2a]$, where the graph of $\mathrm{d}_{\mathrm{KW}}(\overline{w}(s), \overline{w'}(s))$ forms a triangle of width $2a$ and height $a\mathrm{d}(\sigma_1, \sigma_2)$, with an area of $a^2\mathrm{d}(\sigma_1, \sigma_2) = \mathrm{d}(\sigma_1, \sigma_2)C$. Considering this step, a constant of $\mathrm{diam}(\Sigma)$ is sufficient.

So we can take $K = \max(\ell, \mathrm{diam}(\Sigma))$. □

To go the other direction, we have to use the fact that the accumulants are always nearby in the KW metric to find an efficient chain of neighbor swaps and pushes. In case $\overline{u}(\ell) \neq \overline{v}(\ell)$, it is obvious that we have to do some vertical pushes. When we have $\overline{u}(\ell) = \overline{v}(\ell)$ after doing these pushes, we could in principle find a chain using only neighbor swaps, but in general this is too expensive. It depends on how far we have to look to find the correct value to swap with. Ultimately, we will carry out a divide-and-conquer approach where we shrink our working intervals in half each time, balancing them as we go using whichever method is more efficient at the moment.

The main difficulty of choosing between horizontal and vertical movement does not actually occur in the discrete case. If we assume d_Σ is always 0 or 1, it makes sense to do the minimum number of vertical pushes to make the word balanced, and then use only neighbor swaps. Every neighbor swap should decrease the area Young metric by its cost. Contrast with the non-discrete case where we cannot say that a neighbor swap improves the area Young metric much if the values were nearby in Σ. In the non-discrete case, the problem of finding an optimal chain is essentially an optimal transport problem involving the arbitrary compact metric space Σ, which is quite complex. We will only try to find a chain that's good enough to prove equivalence.

Next we observe that when considering a chain between rational step words u and v, we can simplify our lives by considering all the steps to take place on intervals of the same width a, which is a common divisor of all interval endpoints occurring in the chain.

Observation 3.10. Using multiple neighbor swaps, we can "horizontally push" a value σ on some interval $[ka, (k+1)a]$ to a distant interval $[k'a, (k'+1)a]$ with every interval in between being shifted back by a.

$$1000 \longrightarrow 0100 \longrightarrow 0010 \longrightarrow 0001$$

The cost is $|k' - k|\, a^2$.

Observation 3.11. A vertical push can be decomposed into vertical pushes on smaller intervals, without affecting the total cost. A neighbor swap can be similarly decomposed, using the previous observation.

$$111000 \longrightarrow 110001 \longrightarrow 100011 \longrightarrow 000111$$

To take full advantage of Observation 3.10 and make our balancing argument work, we will first introduce another intermediate metric which is defined on pseudo-words.

Definition 3.12. A *rational step pseudo-word* is a function on $[0, \ell]$ constant on each interval of a rational partition, whose values are finitely supported signed measures on Σ.

If u is a rational step word, the corresponding pseudo-word replaces each value σ of u by the measure δ_σ.

Definition 3.13. If u is a rational step pseudo-word, its *accumulation* $\mathrm{acc}_S(u)$ on a set $S \subseteq [0, \ell]$ is

$$\mathrm{acc}_S(u) = \int_S u(t)dt.$$

The *net mass* of u on S is $\mathrm{acc}_S(u)(\Sigma)$. The *total net mass* of u is $\mathrm{acc}_{[0,\ell]}(u)(\Sigma)$.

Note that the integral here is actually a finite sum of measures over each rational interval, weighted by the length of that interval. For example, on an interval $[t, t + a]$ where u

constantly has measure μ, the accumulation is the measure $a\mu$. To avoid confusion, keep in mind that the measure contributed to the accumulation by an interval is not literally the value of the pseudo-word—it is weighted by the length of the interval. So an actual word, viewed as a pseudo-word, always has delta-masses (measure 1) on every interval, however small. Note that the total net mass of any actual word on $[0, \ell]$ is ℓ.

One motivation for introducing pseudo-words is to make the idea of horizontal pushes official. For example, we can now think of a neighbor swap as happening in two steps, where in between we have a pseudo-word:

$$\underline{\sigma_1}\,\underline{\sigma_2} \quad \longrightarrow \quad \begin{array}{c} \sigma_1 \\ \sigma_2 \\ \hline\hline \end{array} \quad \longrightarrow \quad \underline{\sigma_2}\,\underline{\sigma_1}.$$

The intermediate pseudo-word takes two values: the 0 measure on the left interval, and $\delta_{\sigma_1} + \delta_{\sigma_2}$ on the right interval.

Definition 3.14. Let u, v be rational step pseudo-words with the same total net mass. The *pseudo-word chain Young metric* $\mathrm{d}_\psi(u, v)$ is defined as a chain metric with the following moves. We move w to w'. All intervals are rational, but r may be any real number.

- *Vertical push.* On an interval $I = [t, t + a]$:

$$w \restriction I = \mu$$
$$w' \restriction I = \mu + r\delta_{\sigma_2} - r\delta_{\sigma_1}.$$

  Cost: $a r \mathrm{d}_\Sigma(\sigma_1, \sigma_2)$.

- *Horizontal push.* Move value from $I_1 = [t, t + a]$ to $I_2 = [t', t' + a]$:

$$w \restriction I_1 = \mu \qquad\qquad w' \restriction I_1 = \mu - r\delta_\sigma$$
$$w \restriction I_2 = \nu \qquad\qquad w' \restriction I_2 = \nu + r\delta_\sigma.$$

Cost: $ar\,|t - t'|$.

Note that the moves do not change the total net mass. The steps in pseudo-word chains are more flexible in how and when they can be applied: they allow splitting up mass and going negative. In fact, given values I, r, σ_1, σ_2, we can apply the corresponding vertical push to any rational step pseudo-word w to get w', and similarly for a horizontal step given I_1, I_2, r, σ. More precisely, each of these functions consists of adding a certain pseudo-word, zero except on the intervals involved, to w. So vertical pushes and horizontal pushes can be applied at any time and they all commute with each other.

Given our identification of actual words with certain pseudo-words, $\mathrm{d}_\psi(u, v)$ makes sense (as an induced metric) for the rational step words. We would like to show that it is equivalent to d_C on actual words. First, the easy direction.

Proposition 3.15. On rational step words, the pseudo-word chain Young metric is weaker than the chain Young metric. In particular, for rational step words u, v, we have

$$\mathrm{d}_\psi(u, v) \leq 2\mathrm{d}_\mathrm{C}(u, v).$$

Proof. Given a d_C chain between u and v, note first that each vertical push corresponds to a d_ψ vertical push of the same cost, with $r = 1$, and σ_2 chosen to cancel out the original value. Secondly, each neighbor swap can be converted to two horizontal pushes as observed above, each with $r = 1$ and $|t' - t| = a$, each of which costs $ar\,|t - t'| = a^2$. So we get a d_ψ chain of at most twice the cost. □

Before proving the other direction, we note that d_ψ is almost a KW metric in its own right. The next lemma shows that we can always replace a chain by a coupling.

Lemma 3.16. Suppose $\mathcal{C}$ is a d_ψ-chain connecting rational step pseudo-words u and v. Let a be the corresponding fundamental width and let $\mathcal{I}$ be the set of fundamental intervals. Consider $\mathcal{I}$ to inherit the usual metric on $[0, \ell]$ by letting each interval be represented by

its center point. Let $\tilde{u}$ and $\tilde{v}$ be finitely supported measures on $\mathcal{I} \times \Sigma$ defined by assigning the accumulation of the original words on each interval to the corresponding vertical cross section $I \times \Sigma$. Let $\mathcal{I} \times \Sigma$ have the sum metric. Then there is a finitely supported signed coupling γ on $\mathcal{I} \times \Sigma$ which joins $\tilde{u}$ and $\tilde{v}$ such that

$$\text{cost}(\gamma) \leq \text{cost}(\mathcal{C}).$$

Conversely, given a coupling γ, there is a chain $\mathcal{C}$ such that $\text{cost}(\mathcal{C}) = \text{cost}(\gamma)$.

Proof. Each step in a d_ψ-chain corresponds to a coupling of the same cost.

- A vertical push moves ar mass from (I, σ_1) to (I, σ_2).
- A horizontal push moves ar mass from (I_1, σ) to (I_2, σ).

(The coupling is otherwise concentrated on the diagonal.) By composing these couplings, we get a coupling with cost at most that of the chain.

Conversely, given a coupling γ, since γ is finite, we can look at each distinct pair $((I_1, \sigma_1), (I_2, \sigma_2))$ with positive mass separately and add a single vertical and horizontal push, or only one if $I_1 = I_2$ or $\sigma_1 = \sigma_2$. In the sum metric, the contribution of this pair to the cost of the coupling agrees exactly with the cost of these one or two steps. □

The next lemma helps to address the issue of converting horizontal pushes to neighbor swaps.

Lemma 3.17. Let P be a permutation on $\{1, 2, \ldots, n\}$. Then P is a product of a number of neighbor swaps at most

$$\sum_{i=1}^{n} |i - P(i)|.$$

Proof. Let P' be the permutation resulting from restricting P to $\{1, 2, \ldots, n-1\}$ and sliding

to the left:

$$P'(i) = \begin{cases} P(i) & P(i) < P(n) \\ P(i) - 1 & P(i) \geq P(n). \end{cases}$$

By the induction hypothesis, we can write P' as a product of neighbor swaps of size $\leq$ cost(P'), where by cost(P) we mean $\sum_i |i - P(i)|$. We can then swap the last element n into position by using an additional $n - P(n)$ neighbor swaps. In total, we need cost$(P') + (n - P(n))$ neighbor swaps.

So what is cost(P')? If $P(i) > P(n)$, i.e. i will go to the right of where n does, then P' puts i one place to the left of where P does. The effect on the cost is ± 1 depending on the sign of $P(i) - i$, i.e., whether i was moving left or right. To be precise:

$$|P'(i) - i| = \begin{cases} |P(i) - i| & P(i) = P'(i) \\ |P(i) - i| - 1 & P(i) < P'(i) \text{ and } P(i) > i \\ |P(i) - i| + 1 & P(i) < P'(i) \text{ and } P(i) \leq i. \end{cases}$$

To summarize, if i is treated differently by P' than by P, then its cost is lower if it was moving to the right ($P(i) > i$) and higher otherwise. Because these i are the ones which end up in the far-right interval $[P(n), n-1]$, they are on net either stationary or moving to the right. Hence

$$\text{cost}(P') \leq \text{cost}(P) - (n - P(n))$$

from which the conclusion follows. □

Proposition 3.18. Let u, v be rational step words. Then $\mathrm{d_C}(u, v) \leq \mathrm{d}_\psi(u, v)$.

Proof. Consider a d_ψ chain and let a be the fundamental width. By Lemma 3.16, we can replace the chain with a coupling of cost ψ on the fundamental intervals. By Observation 2.4, we can assume the coupling is nonnegative. We can also apply Proposition 2.9 to get a coupling which does not split measure. Then the coupling must actually come from a

function, which assigns to each interval where the mass of that interval must go, whether it moves in $[0, \ell]$ or Σ or both.

If the mass of an interval needs to be moved vertically, a vertical push for a d_C-chain does the job at the same cost. If the mass needs to be moved horizontally, we need to achieve this using neighbor swaps. To each small interval inside $[0, \ell]$ we have assigned a target interval where that interval's mass will go. Now we have the combinatorial problem of effecting these horizontal pushes with neighbor swaps.

The cost of the horizontal pushes is $a^2\text{cost}(P)$ where P is the corresponding permutation on the fundamental intervals, numbered in order. By Lemma 3.17 we can achieve this by at most $\text{cost}(P)$ neighbor swaps, which also cost $a^2\text{cost}(P)$. □

Now we have shown $\mathrm{d_C}$ is equivalent to d_ψ, and it remains to show that d_ψ is equivalent to the Young metric itself, specifically that it is weaker.

Proposition 3.19. Given ℓ and Σ, the pseudo-word chain Young metric is equivalent to the Young metric on rational step words.

Proof. Assume $\ell = 1$ and $\text{diam}(\Sigma) = 1$; routine modifications will cover the general case. Suppose $d(u, v) \leq \delta$, where $u, v : [0, 1] \to \Sigma$ are rational step words. We will construct a chain to show that $\mathrm{d}_\psi(u, v)$ is bounded by some function of δ which converges to zero as $\delta \downarrow 0$. Throughout the construction, if w is an intermediate pseudo-word in the chain, we will maintain the invariant that $d(w, v) \leq \delta$. In fact we need to define what $d(w, v)$ means since w is not an actual word. We define $d(w, v)$ by extending the definition of an accumulant: if w is a pseudo-word, its accumulant $\overline{w}$ is defined so that $\overline{w}(s)$ is the accumulation of w on $[0, s]$. Then $d(w, v) = \sup_s \mathrm{d_{KW}}(\overline{w}(s), \overline{v}(s))$.

We will construct a chain with fundamental width a which is a common divisor of the partition points of u and v and of a sufficiently large power of 2.

1. Use vertical pushes to achieve balance on $[0, 1]$: $\overline{w}(1) = \overline{v}(1)$. To prevent from unbalancing any smaller intervals, distribute the vertical pushes evenly across the entire

word. For example, if a mass m of σ_1 needs to be pushed to σ_2, we do a vertical push of $am\delta_{\sigma_2} - am\delta_{\sigma_1}$ on each fundamental interval. Note that the ability to do a balanced vertical push like this was a primary reason for introducing pseudo-words. Note also that w may now have negative mass in some places.

The cost of this step is $\leq \delta$.

2. Divide $[0,1]$ into $[0,\frac{1}{2}]$ and $[\frac{1}{2},1]$, and rebalance the word (according to the following substeps) to have accumulations matching v on each of these intervals, using a combination of vertical and horizontal pushes. Then subdivide each of these intervals in half and repeat on the four intervals of width $\frac{1}{4}$. Continue until the intervals have width $2^{-N} \leq \delta$.

   To clarify terminology: at stage 0 we have the one ($2^0 = 1$) working interval $[0,1]$ and we balance its two halves. At stage $k < N$ we have 2^k working intervals of width 2^{-k}, and we balance the two halves of each working interval. To balance the two halves of a working interval, take the following steps:

   a) Fix an optimal coupling of the accumulation of the working word w on the left half interval with that of the target word v. Assume (Proposition 2.9) the coupling has values divisible by a. Note that since the working interval was already balanced by a previous stage, the reverse coupling works for the right half interval. Also fix a *realization* of this coupling: assign to each fundamental interval $[t, t+a]$ with value σ_1 a pair (σ_1, σ_2) (possibly $\sigma_1 = \sigma_2$) such that the total mass (a from each fundamental interval) contributed to any pair (σ_1, σ_2) is equal to the value of the coupling on that pair.

   b) For each fundamental interval, if the portion of the coupling assigned to this interval sends σ_1 to σ_2 with $\mathrm{d}_\Sigma(\sigma_1, \sigma_2) \leq \sqrt{\delta}$, then use a vertical push (possibly null) to accomplish the change. Perform an equal portion of this vertical push at every fundamental interval on the left half interval (a balanced vertical push).

Also perform the opposite push on the right half interval.

The cost of this step is $\leq 2^{-k}\sqrt{\delta}$.

c) On all other fundamental intervals, perform equal horizontal pushes from this one fundamental interval to each fundamental interval in the right half interval, to effect that portion of the coupling. (This step also effects the corresponding portion of the reverse coupling on the other half interval.)

To bound the cost for this step, note that the total width of fundamental intervals which contribute more than $a\sqrt{\delta}$ must be $\leq \delta/\sqrt{\delta} = \sqrt{\delta}$, by Markov's inequality on the cost of the coupling. So the cost is $\leq 2 \cdot 2^{-k}\sqrt{\delta}$.

d) The total cost for stage k, counting all intervals, is $\leq 3\sqrt{\delta}$. Also note that the invariant has been maintained because the pushes were done in a balanced way.

The number of stages is $N = \left\lceil \log_{\frac{1}{2}}(\delta) \right\rceil$ so the total cost of this part is $\leq 3\sqrt{\delta} \left\lceil \log_{\frac{1}{2}(\delta)} \right\rceil$.

3. Use horizontal pushes inside each interval of width 2^{-N} to complete the chain. The cost of this step is $\leq (2^{-N})^2 \cdot 2^N = 2^{-N} \leq \delta$.

The total cost of the chain is $\leq 3\sqrt{\delta} \left\lceil \log_{\frac{1}{2}}(\delta) \right\rceil + 2\delta$. □

3.4 The Graph Young Metric

Any measurable word induces a measure on $[0, \ell] \times \Sigma$ by pushing forward Lebesgue measure onto the graph of the word as a function. So we can consider the metric on measurable words induced by the KW metric for measures on $[0, \ell] \times \Sigma$ of total measure ℓ. We might call this the *graph Young metric*. The biggest reason for considering this metric is that it gives a proof that our definition of Young measures agrees (up to homeomorphism) with the usual definition in the literature.

Proposition 3.20. On rational step words, the Young metric is equivalent to the metric induced by the KW metric on $[0, \ell] \times \Sigma$.

Proof. Instead of the Young metric, we equivalently consider the pseudo-word chain Young metric. Each step in a d_ψ chain directly corresponds to a coupling on $[0, \ell] \times \Sigma$ of the same cost. (Direct computation verifies that with the mass spread out over an interval, the total cost of moving it to another interval is the same as if the mass were concentrated in the center.)

Conversely, suppose γ is a coupling between the measures induced by rational step words u and v. We will show how to transform γ into a coupling of the same or lesser cost which is given by a transport function which is affine on each interval of a finite partition. Then, after rational approximation, we have a coupling corresponding to a pseudo-word chain of the same cost.

Let $[0, \ell]$ be partitioned into intervals of equal length such that u and v are each constant an these intervals. Let's call these *constancy intervals*. Let $\gamma_{I,I'}$ be the restriction of γ to $(I \times \{\sigma_1\}) \times (I' \times \{\sigma_2\})$, where σ_1 and σ_2 are the values of u and v on constancy intervals I and I' respectively. This restriction is a coupling of some possbily messy measures on the intervals $I \times \{\sigma_1\}$ and $I' \times \{\sigma_2\}$. The total mass of the coupled measures is less than the size of the intervals, however. So we can replace this coupling by a coupling $\beta_{I,I'}$ of copies of Lebesgue measure on a subinterval of $I \times \{\sigma_1\}$ and a subinterval of $I' \times \{\sigma_2\}$. (We defer specifying which subinterval exactly until later.) Consider how $\beta_{I,I'}$ is different from $\gamma_{I,I'}$:

- $\beta_{I,I'}$ is given by a transport function
- the vertical component of the cost is unchanged
- the total amount of measure moved is unchanged
- the actual measures (i.e. what $\beta_{I,I'}$ is a coupling *of*) are changed
- the horizontal component of the cost is changed.

Despite the latter two changes, in considering all pairs of constancy intervals, we can choose $\beta_{I,I'}$ so that the sum β is in fact a coupling of u and v of the same cost, and has the

desired properties of being given by a transport function which is affine on each interval in a partition (which refines the constancy intervals). For example, if we take all pairs (I, I') in some order, we can assign the subintervals to use the cheapest space still available, so that when we are done, the subintervals chosen form a partition of each constancy interval. Then we can show that the total cost of β must be at most that of γ.

We prove that the cost does not increase by adding it up in a different way. Any coupling γ of u and v can be decomposed into three couplings $\gamma_0, \gamma_1, \gamma_2$ as follows. γ_0 operates within each interval $I \times \{\sigma\}$ by all the (Lebesgue) measure of that interval to the endpoints. γ_2 has the opposite property, moving measure from the endpoints to the whole interval. Then γ_1 is concentrated entirely on moving measure between endpoints—so in fact γ_1 is finitely supported. The sum of the costs of $\gamma_0, \gamma_1, \gamma_2$ is that of γ: the vertical cost occurs entirely in γ_1, and the horizontal cost associated with all mass is decomposed into the cost of moving from a point in an interval to one of its endpoints (γ_0), then to an endpoint of another interval (γ_1), then to the interior of that interval (γ_2). The total horizontal distance moved is the same either way.

If we similarly decompose β, we have that $\beta_1 = \gamma_1$, but β_0 and β_2 may differ from γ_0 and γ_2 respectively. Both β_0 and γ_0 move the mass on a constancy interval to its endpoints, but β_0 does so in an optimal way (by construction). Here optimality amounts to the following: if the interval is $[s, s']$ then the mass moved to s should be drawn from $[s, t]$ and that moved to s' should be drawn from $[t, s']$, for some t. Similarly γ_2 distributes mass from the endpoints to the interval, and β_2 does so in an optimal way. It follows that the cost of β is at most that of γ.

The minor issue remains that the intervals on which β is affine may not be rational. It costs us an arbitrarily small amount to approximate by a nearby coupling on rational intervals with the same properties. Then since this coupling is affine on each interval of a rational partition, we can replace it by a pseudo-word chain as desired. □

Corollary 3.21. *The completion of the space of rational step words in any of the equivalent*

variants of the Young metric is homeomorphic to the space of Young measures with domain $[0, \ell]$ *and codomain* Σ *under stable convergence.*

Proof. The stable topology on Young measures agrees with the topology they inherit as a subspace of $I \times \Sigma$ in the narrow (or weak or weak*) convergence. See [5, Theorem 2.1.13], with Theorem 2.1.3 and Section 1.3 of the same reference providing context. (All the technical conditions are easily satisfies in our case of compact metric spaces, and all of the different stable topologies are equivalent.) Furthermore, the KW metric metrizes this topology (when restricted to measures of a fixed total mass, here ℓ), see for example [15, Corollary 6.13]. The corollary follows from these facts and Proposition 3.20. □

3.5 The Young Monoid

Definition 3.22. The Young metric is extended to compare words of different length as follows. Assume u, v are words and $\ell(u) \leq \ell(v)$. Then

$$\mathrm{d}(u, v) = \mathrm{d}(u, v \restriction [0, \ell(u)]) + \mathrm{diam}(\Sigma)(\ell(v) - \ell(u)).$$

It is straightforward to check that this extension remains a metric. (The factor of $\mathrm{diam}(\Sigma)$ prevents the triangle inequality from being violated by snipping and extending to cheat around a vertical push.)

Remark 3.23. The area Young metric may be extended in the same manner. The chain and pseudo-word chain metrics may be extended by adding two additional type of moves:

- *Extension.* If w is defined on $[0, \ell]$, w' may be defined on $[0, \ell + q]$ where $q \in \mathbb{Q}$, with $w \restriction [0, \ell] = w' \restriction [0, \ell]$. Cost: $q \,\mathrm{diam}(\Sigma)$.
- *Retraction.* A retraction from w to w' is the same as an extension from w' to w, with the same cost.

Definition 3.24. The *concatenation* of words u and v is given by

$$(uv)(t) = \begin{cases} u(t) & 0 \le t < \ell(u) \\ v(t - \ell(u)) & \ell(u) \le t \le \ell(u) + \ell(v). \end{cases}$$

Proposition 3.25. The following operations are uniformly continuous in the Young metric:

- concatenation
- length
- for fixed $S \subseteq [0, \ell]$, accumulation on S.

Proof. Straightforward. □

Note that if we considered measurable words in the L^1 metric, concatenation would not be uniformly continuous, because the first word may change slightly in length while the second word has rapid oscillations.

Definition 3.26. The *Young monoid* $\mathrm{YM}(\Sigma)$ over a compact space Σ, here denoted simply M, is defined as follows. Let M_ℓ be the completion of rational words of length ℓ in the Young metric. Then the underlying set of M is

$$M = \bigcup_{\ell \in [0,\infty)} M_\ell$$

and the monoid operation is concatenation. The unique element of M_0 is the identity. Elements of $\mathrm{YM}(\Sigma)$ are called *Young measures* or *Young words* or simply *words* over Σ.

By Proposition 3.25, we observe that M_ℓ remains the set of words of length ℓ in the new definition, and we may define $M_{\le \ell}$ in the obvious way. Young words have accumulations on any set, though they needn't have a well-defined value at a point.

Note that these observations follow by applying uniform continuity on $M_{\leq \ell}$ for sufficiently large ℓ. We make no assertions here about the equivalence of metrics on all of M or about uniform continuity in general on all of M.

The following proposition summarizes what we know about compactness. We can say a set is *bounded* if its words have bounded length.

Proposition 3.27. A subset of $\mathrm{YM}(\Sigma)$ is compact if and only if it is closed and bounded.

Proof. Let $M = \mathrm{YM}(\Sigma)$ with M_ℓ defined as above. We have already shown that M_ℓ is compact. It follows that $M_{\leq \ell}$ is compact: consider a sequence and pass first to a subsequence which the lengths of the words converge, say to ℓ', replace this subsequence by an equivalent subsequence in $M_{\ell'}$, and apply compactness of $M_{\ell'}$.

Every bounded set is a subset of some $M_{\leq \ell}$, so it follows that every closed and bounded set is compact. Conversely, as length is continuous, every compact set must be bounded. □

To finish the section, we give without proof a few ways of thinking about Young words other than as elements of an abstract completion. The reader may find it instructive to consider the limit of the sequence of words on $[0, 1]$ which oscillate equally and ever more rapidly between 0 and 1, and describe this limit in terms of each of these constructs.

- A Young word corresponds to a *pseudo-accumulant*: a Borel measure-valued function α on $[0, \ell]$ such that
  - $\alpha(t)(\Sigma) = t$ for each $t \in [0, \ell]$
  - $t \mapsto \alpha(t)(A)$ is increasing for each Borel $A \subseteq \Sigma$.
- A Young word corresponds to a Borel measure on $[0, \ell] \times \Sigma$ whose projection onto the first coordinate is Lebesgue measure.
- A Young word corresponds to a suitably measurable probability measure-valued function $[0, \ell] \to \mathcal{P}(\Sigma)$. (This is the *disintegrated form* of the Young measure.) This

function is the time-derivative of the corresponding pseudo-accumulant. It generalizes the notion of pseudo-word from Definition 3.12.

3.6 Notes

The modern measure-theoretic approach to Young measures, according to [6, Section 3.10], was initiated by E. J. Balder and M. Valadier. A relatively short introduction is [13]. The recent book [6] introduces Young measures and includes a self-contained account of all the necessary measure theory as well.

This approach to defining Young measures is original to the best of my knowledge—the definitions of all the variants are my own. This remark applies even to the graph Young metric—the possibility of using the KW-metric (or any other metrization of the weak topology) as a natural metric on Young measures is implicit in standard results, but I have never seen attention called to it.

The way I have presented the approach is probably not optimal. If I wanted to be mathematically efficient, this entire metric approach would have been quite unnecessary, as I could quote the standard definition of Young measures as well as standard versions of the main existence result depending on them (Corollary 4.6 from the next chapter). On the other hand, I had discovered some of the ideas independently and grew attached to them.

The approach has the advantage that it is relatively concrete, does not rely as heavily on measure theory, and is more constructive. But my presentation is not optimal in this ense either. In retrospect, it would have been better to go even further in relying on metric space concepts rather than measure theory, given that I was basically reinventing Young measures from scratch anyway. I realized only very late—too late for an overhaul of my thesis—that the chain Young metric actually does not use the concept of a measure at all, is fairly intuitive, and could have been made primary. This metric is the one that is used in the major proof to come later (Theorem 4.5). And the chain Young metric also does an

even better job at explaining the introductory examples from Section 3.1. Supposing the compactness proof were adapted, the original definition of the Young metric here could be thrown out, and measure theory would then be needed only to prove equivalence with the standard definitions in the literature.

The reason in fact that Section 3.1 was based on the first definition of the Young metric is that it was the definition I discovered independently when trying to solve the compactness problem on my own, before I had any idea what Young measures were.[3] I will note that apart from making the compactness proof relatively straightforward, it remains possible that this definition will have an advantage for algorithms. For example, in the case where $\Sigma = \{0, 1\}$, computing the Young metric should require a single scan of the word, whereas computing the chain Young metric would seem to require a search for an optimal chain.

Presentation and algorithms aside, one reason the metric approach has not been attempted much elsewhere is undoubtedly that it limits the generality. Modern references all consider much more than the case of compact metric spaces–the theory (in its measure-theoretic form) seems to work well with the domain being any measure space (which for us was Lebesgue measure on $[0, \ell]$) and the codomain be a topological space with very mild regularity conditions, such as a Suslin space. Even when the codomain is not compact, there are good conditions known for which subsets of Young measures are compact (see the Prokhorov theorem). I'm sure such generality could be pulled into the continuous automata setting if it proves useful. For a reference on just how far the generality can be pushed (and how technical the theory becomes), see [5].[4]

[3] Sadly, despite the many applications that have been found for them and their fundamental conceptual role in the existence theory, Young measures seem to be regarded as a specialist tool which doesn't deserve mention in most standard textbooks on real analysis, calculus of variations, or optimal control. So it is not surprising that I had not heard of them. (Even though I had the good fortune to have an advisor who is very much aware!)

[4] Although the generality of that reference probably cannot be achieved, I would speculate that the approach here can likely still be extended to a rather general case: when the domain and the codomain are both uniform spaces, with the domain being equipped with some measure. It would take a lot of work though.

4 Examples of Continuous Automata

In this chapter we will seek to precisely describe natural examples of continuous automata over Young monoids. The most important result of this chapter will come in Section 4.2, when we will show that we can get a large class of examples by considering differential equations with an input parameter.

4.1 First Examples

Example 4.1. Let M be a topological monoid, and let M act on itself by right multiplication. Then (M, M) is a continuous automaton.

Example 4.2. Let $M = \mathrm{YM}(\Sigma)$ be a Young monoid. Fix $\tau \geq 0$ and let $S = M_\tau$, the space of words of length exactly τ. We may define an action of M on S by

$$s \cdot m = sm \restriction (\ell(s) + \ell(m) - \tau),$$

that is, we consider the final τ-length segment of the word sm. This action is continuous and defines a continuous automaton which is compact.

Example 4.2 may be described as the compact automaton whose state remembers exactly the last τ-part of the input. This automaton may be used to implement a delay device (Example 5.7); as we observed in the introduction, its compactness comes straight from the compactness of M_ℓ.

The next example is in some ways opposite.

Example 4.3. Let M, τ, and S be as before, but define a new action of M on S by

$$s \cdot m = \begin{cases} sm & \ell(sm) \leq \tau \\ sm \restriction \tau & \ell(sm) \geq \tau. \end{cases}$$

This automaton remembers only the *first* τ-part of the input. As such, it is eventually constant: after the input exceeds length τ, no further state change will occur.

4.2 Automata from Differential Equations

The following is another example of a continuous automaton.

Example 4.4. Let $S = S^1 \subset \mathbb{R}^2$ be the unit circle and let Σ be a compact subset of reals, for example $\Sigma = \{+1, -1\}$ or $\Sigma = [-1, 1]$. Then differential equation

$$\frac{d\theta}{dt} = \sigma$$

defines a continuous automaton with $\mathrm{YM}(\Sigma)$ acting on S.

This example can be justified as follows. Think of the circle as a $\mathbb{R}/2\pi\mathbb{Z}$. For any Young word we can take the accumulation (recall Proposition 3.25) of +1 or of -1 in a given word. Then the state simply tracks the difference of these accumulations, mod 2π. The goal of the rest of the section will be to generalize this example by allowing (almost) arbitrary differential equations in the definition of a continuous automaton.

Since we defined the Young monoid $M = \mathrm{YM}(\Sigma)$ as a metric completion, the natural way to construct a continuous automaton is by first determing its behavior on a nice dense subset of $M_{\leq 1}$, like the rational step words. If the behavior is uniformly continuous in the Young metric, then it extends to all of $M_{\leq 1}$. And if it is causal, it extends to all of M.

Note that the behavior of the entire automaton is uniquely determined by the dynamics associated with each individual letter $\sigma \in \Sigma$, since constant words suffice to generate the step words, which are dense. The following theorem gives sufficient conditions for such dynamics to produce a continuous automaton.

Theorem 4.5. *Let S be a complete metric space, Σ a compact metric space. Let $\varphi : \Sigma \times \mathbb{R}_{\geq 0} \times S \to S$ be a continuously Σ-parametrized family of flows in S, denoted as $(\sigma, t, x) \mapsto \varphi_{\sigma,t}(x)$. Assume furthermore that*

- *The family has uniformly bounded speed: there is a constant A such that*

$$d_S(x, \varphi_{\sigma,t}(x)) \leq At. \tag{4.1}$$

- *The family is uniformly Lipschitz in x: there is a constant K such that*

$$d_S(\varphi_{\sigma,t}(x), \varphi_{\sigma,t}(y)) \leq e^{Kt} d_S(x, y). \tag{4.2}$$

- *There is a constant B such that*

$$d_S(\varphi_{\sigma,t}(x), \varphi_{\tau,t}(x)) \leq B\, d_\Sigma(\sigma, \tau) t. \tag{4.3}$$

- *There is a constant C such that*

$$d_S\left(\varphi_{\sigma,t}(\varphi_{\tau,t}(x)), \varphi_{\tau,t}(\varphi_{\sigma,t}(x))\right) \leq Ct^2. \tag{4.4}$$

  for any $t \geq 0$, any $x \in S$, and any $\sigma, \tau \in \Sigma$.

Then the action of the constant words defined by φ_σ for $\sigma \in \Sigma$ extends uniquely to a continuous automaton over $YM(\Sigma)$.

Informally speaking, the inequalities amount to the following:

- Each flow has uniformly bounded speed (4.1).

- For each flow, (4.2) says that running the flow for any fixed amount of time t results in a Lipschitz map $\varphi_{\sigma,t} : S \to S$. Furthermore note that for a larger time such as $2t$, $\varphi_{\sigma,2t}$ would also be forced to be Lipschitz; the exponential in the inequality says that the Lipschitz constants are all compatible in this way. They are generated infinitesimally, so to speak.

- Equation (4.3) says that flows parametrized by different letters may differ locally according to the distance between those letters in Σ.

- Equation (4.4) says that any two of the flows commute with one another at second-order. This is necessary to show uniform continuity in the Young metric—think of neighbor swaps. The condition is true of any sufficiently smooth family of flows, as we will see.

Proof. Let $M = \mathrm{YM}(\Sigma)$. The flow $\varphi_{\sigma,t}$ defines an action of the submonoid of M consisting of words which are constantly σ. For words w which are finite products of constant words, we can define $\varphi_w : S \to S$ in an straightforward manner and check that this action satisfies the law of causality (1.1). Notationally, we will write $\varphi_{w,t}(x)$ to mean $\varphi_{w\restriction[0,t]}(x)$.

Now, let $u, v \in M_{\leq 1}$ be rational step words with $\mathrm{d_C}(u, v) \leq \delta$, and consider a chain of vertical pushes, neighbor swaps, and extensions/retractions between u and v of cost $\leq \delta$. Let w be an arbitrary intermediate word in the chain. For each type of move, we will show that the effect of the move on $\varphi_w(x)$ is at most a constant times the cost of the move.

- *Vertical push.* Say the vertical push is on $[s, s + a]$ from σ to τ. By (4.3) applied to $\varphi_{w,s}(x) = \varphi_{w',s}(x)$,

$$\mathrm{d}_S(\varphi_{w,s+a}(x), \varphi_{w',s+a}(x)) \leq Ba\mathrm{d}_\Sigma(\sigma, \tau).$$

So if $t \leq 1$, then $t - s \leq 1$ and by (4.2):

$$\mathrm{d}_S(\varphi_w(x), \varphi_{w'}(x)) \leq Be^K a \mathrm{d}_\Sigma(\sigma, \tau) = Be^K \mathrm{cost}(w, w').$$

- *Extension/restriction.* If w is extended to w' with $\ell(w') = \ell(w) + r$ then the cost is $r \operatorname{diam}(\Sigma)$ and by bounded speed (4.1), $\varphi_w(x)$ is moved at most by Ar. Hence

$$\mathrm{d}(\varphi_w(x), \varphi_{w'}(x)) \leq At = \frac{A}{\operatorname{diam}(\Sigma)} \mathrm{cost}(w, w').$$

- *Neighbor swap.* Say the neighbor swap occurred on $[s, s+a]$, $[s+a, s+2a]$ with values σ and τ respectively. Then, applying (4.4) to $y = \varphi_{w,s}(x) = \varphi_{w',s}(x)$, we get

$$\mathrm{d}_S(\varphi_{w,s+2a}(x), \varphi_{w',s+2a}(x)) \quad = \quad \mathrm{d}_S(\varphi_{\tau,a}(\varphi_{\sigma,a}(y)), \varphi_{\sigma,a}(\varphi_{\tau,a}(y))) \quad \leq \quad Ca^2$$

so by (4.2),

$$\mathrm{d}_S(\varphi_w(x), \varphi_{w'}(x)) \leq Ce^K a^2 = Ce^K \mathrm{cost}(w, w').$$

Summing up, we have

$$\mathrm{d}_S(\varphi_w(x), \varphi_{w'}(x)) \leq \max\left(Be^K, \frac{A}{\operatorname{diam}(\Sigma)}, Ce^K\right) \mathrm{d}_{\mathrm{C}}(w, w').$$

So the action of rational step words in $M_{\leq 1}$ is uniformly continuous in the chain Young metric. It follows that the action can be uniquely extended to all of $M_{\leq 1}$ by uniform continuity, and to all of M by causality. □

Corollary 4.6. *Suppose Σ is a compact metric space and $S \subseteq \mathbb{R}^n$ is compact. Consider a parametrized system of differential equations*

$$\frac{d}{dt}\vec{y} = f(\sigma, \vec{y}) \tag{4.5}$$

where $f : \Sigma \times S \to \mathbb{R}^n$ is Lipschitz and everywhere tangent to S. Then there is a unique continuous automaton over the YM(Σ) such that for each $\sigma \in \Sigma$, constant input of value σ causes the state to evolve according to (4.5).

Proof. By the Picard-Lindelöf theorem, the system has uniquely defined solutions for all t, for each fixed σ. Let $\varphi_{\sigma,t} : S \to S$ be the corresponding family of flows. We only need to show that the hypotheses of Theorem 4.5 are satisfied.

- Since f is bounded on S, bounded speed (4.1) follows immediately.
- Given $x, y \in S$, $\sigma \in \Sigma$, let $r(t) = |\varphi_{\sigma,t}(x) - \varphi_{\sigma,t}(y)|$. Let K be a Lipschitz bound for f. Then

  $$|r'(t)| \leq K\,|r(t)|$$

  and hence

  $$|r(t)| \leq |r(0)|\,e^{Kt}.$$

  Rewriting, we get

  $$|\varphi_{\sigma,t}(x) - \varphi_{\sigma,t}(y)| \leq e^{Kt}\,|x - y|$$

  which is exactly (4.2). (This estimate is Grönwall's inequality.)
- For this and the next estimate, we need to use Taylor's Theorem:

  $$\varphi_{\sigma,t}(x) = x + \dot{\varphi}_{\sigma,0}(x)t + O(t^2)$$

  where $\dot{\varphi}$ is the time-derivative of φ and $O(t^2)$ is some term which is bounded by a constant multiple of t^2. Importantly, the constant in $O(t^2)$ does not depend on σ or x: since $\dot{\varphi}$ is Lipschitz, it is also absolutely continuous so the Lebesgue integral remainder

applies, and $\dot{\varphi}$ is bounded as S is compact. We get

$$\varphi_{\sigma,t}(x) - \varphi_{\tau,t}(x) = (\dot{\varphi}_{\sigma,0}(x) - \dot{\varphi}_{\tau,0}(x))t + O(t^2)$$
$$|\varphi_{\sigma,t}(x) - \varphi_{\tau,t}(x)| \leq K d_\Sigma(\sigma,\tau)t$$

where again K is a Lipschitz constant for f (hence for $\dot{\varphi}$). We have proved (4.3).

- For (4.4), we again use the Taylor estimate and the fact that $\dot{\varphi}$ is Lipschitz:

$$\begin{aligned}
&\varphi_{\sigma,t}(\varphi_{\tau,t}(x)) - \varphi_{\tau,t}(\varphi_{\sigma,t}(x)) \\
&= \varphi_{\tau,t}(x) + \dot{\varphi}_{\sigma,0}(\varphi_{\tau,t}(x))t - \varphi_{\sigma,t}(x) - \dot{\varphi}_{\tau,0}(\varphi_{\sigma,t}(x))t + O(t^2) \\
&= x + \dot{\varphi}_{\tau,0}(x)t + \dot{\varphi}_{\sigma,0}(\varphi_{\tau,t}(x))t - x - \dot{\varphi}_{\sigma,0}(x)t - \dot{\varphi}_{\tau,0}(\varphi_{\sigma,t}(x))t + O(t^2) \\
&= t \cdot \Big(\big(\dot{\varphi}_{\tau,0}(x) - \dot{\varphi}_{\tau,0}(\varphi_{\sigma,t}(x))\big) + \big(\dot{\varphi}_{\sigma,0}(\varphi_{\tau,t}(x)) - \dot{\varphi}_{\sigma,0}(x)\big) \Big) + O(t^2) \\
&= t \cdot O(t) + O(t^2) \\
&= O(t^2).
\end{aligned}$$

□

In some cases we will want to be able to define similarly an automaton with a non-compact state space. Routine modifications of the above proofs yield the following results.

Theorem 4.7. *Let S, Σ, and φ be as in the statement of Theorem 4.5: S complete, Σ compact, φ a Σ-parametrized family of flows on S. Assume that*

- *φ has uniformly bounded speed (4.1), and*
- *on each bounded subset of S, φ satisfies (4.2), (4.3), and (4.4).*

Then the action of the constant words defined by φ_σ extends uniquely to a continuous automaton over $YM(\Sigma)$.

Corollary 4.8. *Suppose Σ is a compact metric space and $S \subseteq \mathbb{R}^n$ is arbitrary. Consider a parametrized system of differential equations*

$$\frac{d}{dt}\vec{y} = f(\sigma, \vec{y}) \tag{4.6}$$

where $f : \Sigma \times S \to \mathbb{R}^n$ is locally Lipschitz and everywhere tangent to S, and such that for each fixed $\sigma \in \Sigma$, the differential equations have global solutions in S for all time. Then there is a unique continuous automaton over $YM(\Sigma)$ such that for each $\sigma \in \Sigma$, constant input of value σ causes the state to evolve according to (4.6).

4.3 Notes

Corollary 4.6 has long been known in some form, and is known in somewhat more generality (more general than Cor. 4.8). Indeed, (4.5) is the prototypical controlled system and the original point of Young measures was to provide for the existence of optimal controls (without convexity assumptions). To see this theorem in the light of control theory, consider that if we had a cost function on trajectories through the state space and showed that to be uniformly continuous as well, then its infimum would be realized by some Young word (here a control function) as M_ℓ is compact.

So what then is the difference between a continuous automaton and a controlled system? It seems to me that there is none at this fundamental mathematical level. I suggest a duality between control theory and the theory of computation: both concern fundamentally a notion which might go under various names: automaton, controlled system, or simply *dynamical system with input*. But the different perspectives afforded by the different intended application domains lead to vastly different research questions about these objects, and of course different languages for talking about them, which in practice has caused the respective academic communities to be nearly disjoint. All three, in fact: dynamical systems, computation, and control. To spell out the differences explicitly:

- In dynamical systems, one studies the possible dynamics of actions of a given monoid (most often $\mathbb{N}$, $\mathbb{Z}$, or $\mathbb{R}$) on a given space, both individually and collectively.

- In control theory, one considers a dynamical system with input, taking the dynamics as given, and seeks to find input (controls) which produce desired behavior.[1]

- In automata theory, one takes the monoid as given and seeks to define dynamics which produce a desired input-output relation (for all possible inputs).

Let me emphasize that I expect no converse for Theorem 4.5 or Corollary 4.6 of any kind. There should exist plenty of non-smooth automata, not describable by differential equations because they do not satisfy (4.4). I haven't included such examples in the main text because I didn't find or develop enough tools to clearly justify their existence.

Nonetheless, here is a possible example. Let $S = [0, 1]$ (which I think of as being drawn vertically for some reason), and $\Sigma = \{\sigma_+, \sigma_-\}$, with the two letters respectively pushing the state up or down at constant speed 1, in the interior of the interval. When the state is $s = 1$, the letter σ_+ has no effect, and similarly for $s = 0$ and σ_-. (We could think of these as discontinuous differential equations.) It is easy to define a family of flows in this way, and there is an action of the rational step words. Note that (4.4) fails: consider starting at 0 and applying the words $\sigma_-^t\sigma_+^t$ or $\sigma_+^t\sigma_-^t$ for any $t > 0$. So some neighbor swaps would have an outsize effect; the proof given for (4.5) could not work. Nonetheless, I conjecture this action is uniformly continuous in the Young metric and extends to a continuous automaton over $\mathrm{YM}(\Sigma)$.

Here is another example. Let $S \subseteq [0, 1] \times [0, 1]$ be the filled-in triangle of pairs $(x, y) \in [0, 1] \times [0, 1]$ satisfying $y \geq x$. Let $\Sigma = \{\sigma_N, \sigma_E\}$; the letters "north" and "east" respectively push in the appropriate directions at constant speed, as long as the state lies in the interior of the triangle. On the top boundary, σ_N has no effect. What about the diagonal boundary?

[1] At my thesis defense, Professor Nerode counteered that this is not a correct view of how control works in practice. The reason is that feedback is virtually always necessary for effective control, for reasons of stability, and so in control practice it is also the case that one needs to produce a good input-output relation, as in my description of automata theory here.

- We could have σ_E have no effect on states on the diagonal boundary. Then the flows definitely fail to define a continuous automaton. Consider the unit-length words obtained by repeating many small copies of $\sigma_E\sigma_N$ or $\sigma_N\sigma_E$; in one case the state will crawl up the diagonal, while in the other case it will remain stationary. As these words are arbitrarily close in the Young metric, uniform continuity is violated.

- We could have σ_E push upward along the diagonal boundary in such a way that the x value changes at (the same) constant speed. Then I'm not sure and it may be that this defines a continuous automaton over $\mathrm{YM}(\Sigma)$.

Note that in these examples the action is not reversible: there are words whose action on the state-space is non-injective. Example 4.2 also had this property. On the other hand, it is a peculiarity of automata defined by differential equations (Corollary 4.6) that they are all reversible.

5 Input/Output and the Myhill-Nerode Theorems

Recall that in the introduction we discussed different ways of using an automaton to perform computation. A transducer continuously operates on an input signal, producing an output signal of a similar kind. A deducer operates on a completed input signal and returns some fixed piece of information about it. We will define these notions formally and then prove a Myhill-Nerode theorem for each one. The Myhill-Nerode theorem gives an exact condition for when a function can be computed by an automaton with a compact state space. Furthermore this condition takes a particularly simple form: given the function, there is a canonical automaton computing it; if any automaton computing the function is compact, the canonical one must be.

5.1 Automata with I/O: Deducers and Transducers

Recall (Definition 1.1) that a continuous automaton (M, S) a continuous action of an input monoid M on a state space S. The action itself is suppressed by the (M, S) notation but usually denoted by $(\cdot)$ as in $s \cdot m$.

Definition 5.1. A *deducer* (M, S, o, s_0) is a continuous automaton (M, S) together with a distinguished start state s_0 and a continuous map $o : S \to \mathcal{O}$, where $\mathcal{O}$ is a topological space, the *outcome space*.

The deducer (M, S, o, s_0) *computes* the map $c : M \to \mathcal{O}$, defined by $c(m) = o(s_0 \cdot m)$.

We may refer to a deducer simply as (S, o, s_0) if the monoid is understood from context.

Example 5.2. Let M and S be as in Example 4.4 (the circle automaton). Let $\mathcal{O}$ be another circle, thought of as $[0, 2\pi]/\sim$. Define $o : S \to \mathcal{O}$ by $o(\theta) = 2\theta \pmod{2\pi}$, so o is a double-covering of the circle $\mathcal{O}$ by the circle S. Let s_0 be arbitrary. Then (M, S, o, s_0) is a deducer.

Example 5.3. Let (M, S) be any continuous automaton and $s_0 \in S$ be arbitrary. Then $(M, S, \mathrm{id}_S, s_0)$ is a deducer, where id_S is the identity map on S.

Definition 5.4. A *transducer* (M, S, N, T, s_0) is a continuous automaton (M, S) together with a distinguished start state s_0 and a *transduction output*: a continuous map $T : S \times M \to N$, written $(s, m) \mapsto T_s(m)$, where N is another topological monoid, which satisfies the transduction law:

$$T_s(m_1 m_2) = T_s(m_1) T_{s \cdot m_1}(m_2). \tag{5.1}$$

The transducer (M, S, N, T, s_0) *computes* the map $C : M \to N$ defined simply as $C = T_{s_0}$.

We may refer to a transducer simply as (S, T, s_0). The transduction law (5.1) is in the same spirit as the law of causality (1.1) which defines monoid actions, and it ensures that the transducer does in fact compute a well-defined function C.

Example 5.5. Let (M, S, o, s_0) be a deducer, where $o : S \to \mathcal{O}$ with $\mathcal{O}$ compact, and M is a Young monoid. Let $N = \mathrm{YM}(\mathcal{O})$. Given $s \in S$, $m \in M$, let $T_s(m)$ be the function $f : [0, \ell(m)] \to \mathcal{O}$, regarded as a Young word, where

$$f(t) = o(s \cdot (m \restriction t)).$$

In words, the transduction output traces the trajectory through $\mathcal{O}$ as the input word is read.

We can start with an arbitrary automaton (M, S) with distinguished state s_0, consider the deducer from Example 5.3, and then apply Example 5.5. Then we get a transducer whose output simply records the state history when starting from s_0.

Note that the output is always continuous in Example 5.5, even when the input is an arbitrary Young word. By contrast, consider the following class of examples.

Example 5.6. Let $f : \Sigma \to \Sigma'$ be a continuous function of compact metric spaces. Then we can define a transducer (M, S, N, T, s_0) as follows. Let $S = \{s_0\}$ be a one-point space, $M = \mathrm{YM}(\Sigma)$, $N = \mathrm{YM}(\Sigma')$, and $T_{s_0}(m) = f \circ m$. (Note that $f \circ m$ makes sense as a new Young word; we can define it in case m is a function and then extend by uniform continuity.)

Example 5.6 shows what can be done with a transducer of a single state: namely, we can transform the input values as soon as they come in, according to a fixed function f. If f happens to be a homeomorphism, then the output will be continuous only if the input is. In particular, if f is the identity on Σ, then we have an identity transducer.

Example 5.7. We can now fully precisely describe the delay transducer mentioned in the introduction. Recall the automaton of Example 4.2. Let $s_0 \in S$ be an arbitrary word of length τ. We can define a transduction output by $T_s(m) = sm \restriction \ell(m)$.

It will be useful to have a basic notion of a map between continuous automata.

Definition 5.8. Let (M, S) and (M, S') be continuous automata. A *continuous automaton morphism* from (M, S) to (M, S') is a continuous map $\varphi : S \to S'$ which respects the action: $\varphi(s \cdot m) = \varphi(s) \cdot m$ for all $m \in M$, $s \in S$.

5.2 Myhill-Nerode Theorem for Deducers

From the standpoint of the map computed by a deducer, the relevant information contained in any given state s is what the outcome will be given any possible future input, a map $M \to \mathcal{O}$ defined by $m \mapsto o(s \cdot m)$. This observation inspires the following definition.

Definition 5.9. Let M be a locally compact Hausdorff monoid and $\mathcal{O}$ a topological space. The *universal deducer* for $M, \mathcal{O}$ is the automaton whose state space is $\mathscr{S} = C(M, \mathcal{O})$, the space of continuous functions $M \to \mathcal{O}$ in the compact-open topology, and whose action is defined by

$$(f \cdot m)(n) = f(mn)$$

for $f \in \mathscr{S}$, $m, n \in M$.

(Recall the compact-open topology from Section 2.2.) Note that the universal deducer has a natural map $\mathscr{S} \to \mathcal{O}$ given by $f \mapsto f(1_M)$, but is not itself a deducer because it doesn't have a distinguished start state.

Proposition 5.10. The universal deducer is well-defined; its action is continuous.

Proof. The action map

$$(\cdot) : \mathscr{S} \times M \to \mathscr{S},$$

that is,

$$(\cdot) : C(M, \mathcal{O}) \times M \to C(M, \mathcal{O})$$

is curried from the map

$$C(M, \mathcal{O}) \times M \times M \to \mathcal{O}$$

defined as $(f, m, n) \mapsto f(mn)$. This map is continuous by the application property and it follows that $(\cdot)$ is continuous by the parametrization property. □

Proposition 5.11. Let (S, o) be a $(M, \mathcal{O})$-deducer, where M is locally compact Hausdorff. The natural map $\varphi : S \to \mathscr{S}$ defined by $\varphi(s)(m) = o(s \cdot m)$ is a continuous automaton morphism.

Proof. The map $\varphi : S \to \mathscr{S}$ is curried from the map

$$S \times M \to \mathcal{O}$$

given by $(s, m) \to o(s \cdot m)$, and is therefore continuous. And φ respects the action:

$$\varphi(s \cdot m)(n) = o(s \cdot m \cdot n) = o(s \cdot mn) = \varphi(s)(mn) = (\varphi(s) \cdot m)(n). \qquad \square$$

Definition 5.12. Let $c : M \to \mathcal{O}$ be continuous, where M is locally compact Hausdorff. The *canonical deducer* for c is the deducer $(M, R_c, \hat{o}, f_0)$ defined as follows. The state space R_c will be a subspace of $\mathscr{S}$, the universal deducer's state space.

- The initial state f_0 is defined by $f_0(m) = c(m)$.
- The action is inherited from the universal deducer.
- The state space R_c is the closure of the states reachable from f_0.
- The outcome map $\hat{o}$ is defined by $\hat{o}(f) = f(1_M)$.

Proposition 5.13. Let $c : M \to \mathcal{O}$ be as above. The canonical deducer for c computes c.

Proof.

$$\hat{o}(f_0 \cdot m) = (f_0 \cdot m)(1_M) = f_0(m) = c(m). \qquad \square$$

Theorem 5.14. *Let $c : M \to \mathcal{O}$ be continuous, where M is a locally compact Hausdorff monoid and $\mathcal{O}$ is any topological space. There is a compact deducer computing c if and only if the canonical deducer for c is compact.*

Proof. It remains only to show that if there is a compact deducer (M, S, o, s_0) computing c, then the canonical deducer is also compact. Let $Q = s_0 \cdot M$ be the reachable states in S. Note that we may throw out unreachable states without affecting the map computed by the deducer. So we can replace S by $\mathrm{cl}(Q)$, the closure of the reachable states, which remains a compact set. (We can replace simply by Q, but that may fail to be compact.) Assume we have already done this, so Q is dense in S. Let $\varphi : S \to \mathscr{S}$ be the natural map given (as above) by $\varphi(s)(m) = o(s \cdot m)$. Proposition 5.11 showed φ is continuous. We claim the image

of S under φ in $\mathscr{S}$ is R_c. First, note that $\varphi(s_0) = f_0$:

$$\varphi(s_0)(m) = o(s_0 \cdot m) = c(m) = f_0(m).$$

Since φ is a morphism, the image of Q under φ is exactly $f_0 \cdot M$, the reachable states in the canonical automaton. Because the closure of Q is S and S is compact, the image of S under φ must agree with the closure of the image of Q. So the image of S is exactly R_c as desired.

As the continuous image of a compact set, R_c must itself be compact. □

Example 5.15. If o is the identity map, as in Example 5.3, and the reachable states are dense, then the natural morphism φ from the deducer to the corresponding canonical deducer will be an isomorphism. (It is a bijection whose inverse is also a continuous automaton morphism.) In other words, if every state is taken as a distinct possible outcome, then Nerode-minimization cannot have any effect.

Example 5.16. Consider Example 5.2. The function computed is $w \mapsto \operatorname{acc}_{\{+1\}}(w) - \operatorname{acc}_{\{-1\}}(w) \pmod{2\pi}$. (Recall Definition 3.1 and Proposition 3.25.) The canonical automaton which computes the same function will just be a copy of the circle automaton (Example 4.4), and the natural map is itself a double covering of one circle by another.

Example 5.17. Let M be as in the previous example. There is a natural automaton with state space $\mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0}$ which computes $w \mapsto \operatorname{acc}_{\{+1\}}(w) - \operatorname{acc}_{\{-1\}}(w)$. The outcome map is $o(x, y) = x - y$. The corresponding minimum automaton will have as its state space a copy of $\mathbb{R}_{\geq 0}$, so this function cannot be computed by a compact automaton. If the outcome map instead is $o(x, y) = x - y \pmod{2\pi}$, the function computed is the same as the previous example and the canonical automaton is compact.

In these examples so far, all states in the automata were reachable; nothing happened when we took the closure. Next we give an example where this is not the case.

Example 5.18. Let M be a Young Monoid and let $\mathcal{O} = \mathbb{R}_{\geq 0}$. The canonical automaton for the length map $\ell : M \to \mathcal{O}$ has as state space just another copy of $\mathbb{R}_{\geq 0}$.

Now change the function by defining $\mathcal{O} = \mathbb{R}_{\geq 0} \cup \{+\infty\}$ as the one-point compactification of $\mathbb{R}_{\geq 0}$. The length function remains defined $\ell : M \to \mathcal{O}$. Its canonical automaton has state space a copy of $\mathcal{O}$ (the one-point compactification).

So by "enlarging" the outcome space to its compactification, we made it so that the length function is computable by a compact automaton. The meaning of this example deserves careful consideration. The function $\ell : M \to \mathbb{R}_{\geq 0}$ is not computable by a compact automaton, but $\ell : M \to \mathbb{R}_{\geq 0} \cup \{+\infty\}$ is. Intuitively, the second function is easier to compute than the first because it demands less accuracy: any demanded degree of accuracy (open cover, which we can assume is finite) allows us to classify words beyond some fixed length as *approximately infinitely long*, and no further precision is needed.

We also remark that this example has nothing to do with continuous time, but the topological character of the state space. A similar example can be constructed with the length function on Σ^*, as long as we allow for compact but non-finite state spaces.

5.3 Transducers

Suppose M is a locally compact Hausdorff monoid and N is any topological monoid. We want to define a canonical transducer for a continuous function $F : M \to N$, but there is an additional difficulty: F has no hope of being computed by any automaton at all unless it is *causal*, meaning that $F(m_1 m_2)$ can always be written as $F(m_1)n$ for some n. To obtain a canonical automaton, we will need that this n is uniquely defined and obtained continuously from m_1 and m_2.

Convention 5.19. For the remainder of this section, we assume that M is a locally compact Hausdorff monoid, and N is any topological monoid which satisfies the following.

- N is left-cancellative: $ab = ac \implies b = c$. Consequently, we can define $a\backslash b$ as the unique n such that $an = b$, provided it exists. Then $a(a\backslash b) = b$.

- The map $(a, b) \mapsto a\backslash b$ is continuous on its domain.

Note that these conditions do hold in case M and N are Young monoids.

Definition 5.20. A map $f : M \to N$ is *causal* if $f(1_M) = 1_N$ and for every $m_1, m_2 \in M$, there exists $n \in N$ such that $f(m_1 m_2) = f(m_1) \cdot n$.

Definition 5.21. The *universal transducer* for M, N is the continuous automaton whose state space $\mathscr{S} \subset C(M, N)$ is the set of causal continuous functions $M \to N$, in the compact-open topology, with the action defined by

$$(f \cdot m_1)(m_2) = f(m_1)\backslash f(m_1 m_2).$$

Proposition 5.22. The universal transducer is well-defined; its action is continuous.

Proof. Let $C \subset C(M, N)$ be the space of causal continuous functions. The action map $C \times M \to C$ may be curried from the map

$$C \times M \times M \to N \qquad (f, m_1, m_2) \mapsto f(m_1)\backslash f(m_1 m_2) \qquad \square$$

which is continuous.

Next we show that the universal transducer has a natural transduction output. Note that it is still not actually a transducer because it lacks a distinguished starting state.

Definition 5.23. The *universal transduction output* is given by $\hat{T}_f(m) = f(m)$, i.e., $\hat{T}_f = f$.

Proposition 5.24. The universal transduction output satisfies the transduction law (5.1).

Proof. It is natural here to work through the transduction law backwards:

$$\begin{aligned}\hat{T}_f(m_1)\hat{T}_{f\cdot m_1}(m_2) &= f(m_1)(f \cdot m_1)(m_2) \\ &= f(m_1)(f(m_1)\backslash f(m_1 m_2)) = f(m_1 m_2) = \hat{T}_f(m_1 m_2).\end{aligned}$$

□

Proposition 5.25. Let (S, T, s_0) be a transducer. Let φ be the natural map $\varphi : S \to \mathscr{S}$ given by $\varphi(s) = T_s$. Then φ is an automaton morphism.

Proof. The map $\varphi : S \to C(M, N)$, given by $\varphi(s)(m) = T_s(m)$, is curried from T itself and therefore continuous. And φ respects causality:

$$(\varphi(s) \cdot m_1)(m_2) = \varphi(s)(m_1)\backslash\varphi(s)(m_1 m_2) = T_s(m_1)\backslash T_s(m_1 m_2)$$
$$= T_s(m_1)\backslash(T_s(m_1)T_{s\cdot m_1}(m_2)) = T_{s\cdot m_1}(m_2) = \varphi(s \cdot m_1)(m_2).$$

□

Definition 5.26. Let $F : M \to N$ be a continuous causal function. The *canonical transducer* for F is the transducer $(M, R_F, \hat{T}, f_0)$ defined as follows. Note that the state space R_F is defined to be a subspace of $\mathscr{S}$, the universal transducer's state space.

- The initial state f_0 is $f_0 = F$.
- The action is inherited from the universal transducer.
- The state space R_F is the closure of the states reachable from f_0.
- The transduction output is defined as $\hat{T}_f = f$.

Proposition 5.27. Let $F : M \to N$ be a continuous causal function. The canonical transducer for F computes F.

Proof. Note that by definition the canonical transducer for F computes the map $\hat{T}_{f_0} = f_0 = F$. □

Theorem 5.28. *Let $F : M \to N$ be a continuous causal function, where M and N are monoids satisfying the requirements of Convention 5.19. There is a compact transducer computing F if and only if the canonical transducer for F is compact.*

Proof. This argument is very similar to the deducer case (Theorem 5.14). It remains only to show that if there is a compact transducer (S, T, s_0) computing F, then the canonical transducer for F is compact. Let $Q = s_0 \cdot M$ be the reachable states in S. We may assume Q is dense. Let $\varphi : S \to \mathscr{S}$ be the natural map defined as in Proposition 5.25. Then we claim the image of φ in $\mathscr{S}$ is R_F. Note that $\varphi(s_0) = T_{s_0} = F = f_0$, so the initial state is respected. Since φ is a morphism, the image of Q is exactly $f_0 \cdot M$, the reachable states in the canonical automaton. Since the closure of Q is S and S is compact, the image of S must be the closure of image of Q, which is exactly R_F as desired.

As the continuous image of a compact set, R_F must itself be compact. □

Example 5.29. Let M be a Young monoid and let $w_0 \in M$ be a word of length τ. Define a causal continuous function F by $F(u) = w_0 u \restriction \ell(u)$. F is the delay function, the function computed by the delay transducer from Example 5.7. That transducer is isomorphic to the canonical transducer for F. The intrinsic state (state of the universal transducer) $F \cdot m$ may be described by

$$(F \cdot m)(u) = (w_0 m \restriction [\ell(m), \ell(m) + \tau])u \restriction \ell(u).$$

To prove isomorphism, it suffices to note that extracting the word $(w_0 m \restriction [\ell(m), \ell(m) + \tau])$ from the function $F \cdot m$ is continuous as a map from a subset of $C(M, M)$ into M. (Use the application property; evaluate the map on any fixed word of length τ.)

5.4 Topology from the Input

The topology of a canonical automaton is the weakest topology which makes the output map continuous. It is possible to go in approximately the opposite direction and ask for the strongest possible topology on an automaton. In this case, the output map is irrelevant because the downward pressure on the topology comes from the requirement that the action itself be continuous.

Proposition 5.30. Let M be a topological monoid, and suppose M acts on a set in such

a way that $S = s_0 \cdot M$ for some $s_0 \in S$ (every state is reachable from s_0). When S is equipped with the final (quotient) topology determined by the map $m \mapsto s_0 \cdot m$, (M, S) is a continuous automaton. Furthermore, this topology is the strongest topology on S which makes the action into a continuous automaton.

Proof. Let $i : M \to S$ be the given map, $i(m) = s_0 \cdot m$. With S given the final topology, we can regard it as a quotient of M with i as the quotient map. Correspondingly, $S \times M$ can be regarded as a quotient of $M \times M$ (by considering the corresponding equivalence relation, ignoring the second coordinate). The proof relies on the fact that we get the same topology if we take the product first or if we take the quotient first.

The action map $S \times M \to S$ is descended from the map $M \times M \to S$ given by $(m, m') \mapsto i(mm')$ and is therefore continuous.

Conversely, any topology which makes (M, S) a continuous automaton will make i a continuous map, so the quotient topology is the strongest possible. □

5.5 Notes

The reader must be wondering by now: so which functions can be computed by a compact automaton? In the classical finite theory we had a beautiful answer to this question, which was that long list of equivalent definitions of a regular language. Given that the Myhill-Nerode theorems worked out as nicely as they did, I would hope for something similar. I still have that hope, but there are many complications, so I have not been able to answer this question so far.

One issue should already be apparent: we had to replace recognizers with deducers in order to be compatible with the principles of the theory. Recognizers mean working with sets instead of functions, which is certainly more convenient in the finite theory. It is natural to ask whether there is in fact a good notion of continuous recognizer. It seems to me that any set recognized is going to have to be open: the mere fact that reading input could introduce

a slight error would prevent us from ever computationally verifying that a boundary point lies in a given set. So we could declare an open subset of S to be the accepting states. (We'd get the same effect if we started with an $\mathcal{O}$-deducer and considered an open set in $\mathcal{O}$.) Let's call this an *open recognizer*.

One of the basic facts about regular languages is that they are closed under complements, and this fact is very easy to prove when thinking of regular languages as those accepted by a DFA: we take the complement of the accepting states. However, it doesn't make sense to ask for the complement of the set recognized by an open recognizer because that set is not open. Note that recognizing an open set of Young words is somewhat like computing a c.e. set: we will certainly get positive confirmation if a word is in the set but we may or may not get confirmation if the word fails to be in the set–if it's on the boundary, we couldn't be sure.

The next natural alternative for complementation would be to ask about the interior of the complement of the set recognized. Perhaps with a positive answer to that question, one might be able to build a notion of intuitionistic regular expressions for continuous automata, a development which would fascinate me. Unfortunately, the obvious construction doesn't seem to work: given $U \subseteq S$ open, suppose $C \subseteq S$ is the interior of U's complement. Unfortunately, it does not follow that $i^{-1}(C)$ is the interior of the complement of $i^{-1}(U)$, where i is the input map $i(m) = s_0 \cdot m$. For example, i could collapse an open set in M to a point on the boundary of U. I don't know of a reason why this problem could not be overcome, but it doesn't seem too easy.

Setting the complementation issue aside, it makes sense to work with deducers and functions. In the finite case, we could perfectly well have considered functions from Σ^* into arbitrary finite sets, defining such a function to be regular if the preimage of each point is regular, and then these are exactly the functions computed by finite deducers. In the continuous case, I don't know if it's possible to encode all functions in terms of sets, but it *is* possible to go the other way. Let $\mathcal{O}$ be the two point space in which one point is open but

not closed as a singleton. (Then the reverse holds for the other point.) For any topological space X, continuous functions $X \to \mathcal{O}$ are in a natural one-to-one correspondence with the open subsets of X. So open recognizers are in a natural correspondence with $\mathcal{O}$-deducers, and the Myhill-Nerode theorem applies to them.

There remains a major issue: in the finite case, a lot of the work is really done by the notion of a nondeterministic automaton. I do not know of any way to prove that regular lanagues are closed under Kleene-* other than through nondeterministic automata and the fact that they can be determinized.

For continuous automata defining nondeterminism and obtaining a determinization theorem are necessarily more complicated, but I believe it is possible. A notion of continuous relation is needed, for which I suggest copying the definition of continuous function: we say a binary relation R is continuous if the preimage of every open set is open. (This notion is weaker than the notions which usually appear in the literature on set-valued analysis.) This definition sits well with the notion of an open recognizer defined above.

The reader will undoubtedly notice that I have not done much with transducers. The reason is that they are slightly more complicated, so it is more work to produce examples. Nonetheless, I think transducers are ultimately much more important than deducers for applications, as they represent parts which can be composed together to form more complex automata, much as one might built an analog electrical circuit out of basic components. An especially interesting question is under what circumstances it makes sense to plug a transducer's output into its own input and obtain a recursively defined automaton.

(Incidentally, the definition of transducers on the face of it is not original; see the old paper [12]. The perspective there seems quite different, however.)

6 Caveats

This chapter collects several miscellaneous results along the lines that continuous automata theory is not as nice as we might naively hope. The material here is even less developed than that of previous chapters, but we consider it important information. Of course, there is much room for future work.

6.1 Most Simulations Will Lose Accuracy

In this section we discuss the problem of error propagation, a difficulty which seems to affect any form of computation with a nondiscrete space of states.

We imagine that any implementation of a continuous automaton would only expected to be approximate, and more specifically that state information is not stored with absolutely perfect fidelity. (In fact, it is sufficient for our argument to allow any one of three types of errors: inaccuracy in the stored state, inaccuracy in state transitions, and inaccuracy in reading input. In practice, we'd expect all three.) So presumably it is possible, every so often for the state to move at least a little bit from what it was supposed to be. We might ask: given that we can't expect perfect fidelity, is there an imperfect degree of fidelity which would allow the computation to remain reasonably accurate (to some standard) for all time? And the answer is: no, of course not, sooner or later error propagation will catch up with us. However, there is an exceptional case, which is basically when the state information is going to be forgotten anyway.

Formally, let us describe the hoped-for condition that an automaton could be simulated indefinitely without loss of too much accuracy from state errors. Assume that M is a Young monoid. For this section we will only consider *metric automata*, by which we mean automata whose state spaces are metric spaces.

Definition 6.1. A metric automaton over M is *self-stabilizing* if for every $\varepsilon > 0$, there exists $\delta > 0$ and $\tau < +\infty$ such that for every $s \in S$ and every finite sequence of input words $m_1, \ldots, m_k$ such that $\ell(m_i) \geq \tau$ for each i,

$$B_\varepsilon(s \cdot m_1 m_2 \cdots m_k) \supseteq B_\delta(B_\delta(\cdots B_\delta(B_\delta(B_\delta(s) \cdot m_1) \cdot m_2) \cdot \cdots) \cdot m_k).$$

In words, given tolerance ε we must be able to specify a small enough error size δ and interval size τ such that if we run the automaton there is no way for an adversary to cause δ-perturbations at τ-intervals and make the perturbed state vary more than ε from the true state. Intuitively, as long as the state space is connected, we should not expect this condition to be true very often because no matter how small δ is, a finite number of δ-perturbations should easily be able to alter the state by more than ε. Nonetheless, there is a class of automata with connected state spaces which are self-stabilizing.

Definition 6.2. A metric automaton (M, S) is *forgetful* if there exists $T < +\infty$ such that for every $m \in M_{\geq T}$, $S \cdot m$ is a singleton; in other words, $s \cdot m$ does not depend on s.

Example 6.3. The automaton described in Example 4.2 is forgetful. Moreover any forgetful automaton can be simulated by an example in this class.

Definition 6.4. A metric automaton (M, S) is *asymptotically forgetful* if for every $\varepsilon > 0$ there exists $T < +\infty$ such that for every $m \in M_{\geq T}$, we have $\operatorname{diam}(S \cdot m) < \varepsilon$.

Automata in which the state converges to a certain particular state for all inputs would be asymptotically forgetful, for example, and further examples could be obtained by taking products of these with the forgetful automata already discussed.

Proposition 6.5. Asymptotically forgetful metric automata are self-stabilizing.

Proof. Given $\varepsilon > 0$, let T be the witness for asymptotic forgetfulness coresponding to $\varepsilon/2$. Then let $\delta = \varepsilon/2$ and $\tau = T$. Let $s, m_1, m_2, \ldots, m_k$ be given. Consider the set

$$B_\delta(B_\delta(\cdots B_\delta(B_\delta(B_\delta(s) \cdot m_1) \cdot m_2) \cdot \cdots) \cdot m_k).$$

Except for the last δ-perturbation, we have a set of points in the range of the action of m_k. So

$$B_\delta(B_\delta(\cdots B_\delta(B_\delta(B_\delta(s) \cdot m_1) \cdot m_2) \cdot \cdots) \cdot m_k) \subseteq B_\delta(S \cdot m_k) = B_{\varepsilon/2}(S \cdot m_k) \subseteq B_\varepsilon(s \cdot m_1 \cdots m_k)$$

since $S \cdot m_k$ has diameter $\leq \varepsilon/2$ and includes the element $s \cdot m_1 \cdots m_k$. □

The main reason for introducing asymptotically forgetful automata was that this class, in the case of a compact, connected state space, exactly characterizes the exceptional behavior needed to make a connected automaton self-stabilizing.

Theorem 6.6. *Suppose (M, S) is a metric automaton where S is compact and connected. Then (M, S) is self-stabilizing if and only if it is asymptotically forgetful.*

Lemma 6.7. Suppose (M, S) is a self-stabilizing metric automaton. Then for every $\varepsilon > 0$ there exists $\delta > 0$ such that $\mathrm{d}_S(s_1, s_2) < \delta$ implies that $\mathrm{d}_S(s_1 \cdot u, s_2 \cdot u) < \varepsilon$ for all words $u \in M$.

Proof. The property asserted is a weakened version of the definition of being self-stabilizing, where there is only one word and one perturbation allowed. □

Proof of Theorem 6.6. The backwards direction is a special case of Proposition 6.5. For the forwards direction, suppose (M, S) is self-stabilizing. Let $\varepsilon > 0$ and fix δ and τ as corresponding witnesses for being self-stabilizing. The idea of the proof will be to show that given any two states s_1, s_2 and infinite word $w \in M_\infty$ (which may be thought of as

an increasing path through M starting at the empty word), it is possible to hop from the w-trajectory of s_1 to that of s_2 using only δ-perturbations after intervals of length τ. If the amount of time it takes to do so does not depend on s_1, s_2, or w, then because (M, S) is self-stabilizing, we will have that $S \cdot u$ has diameter at most ε.

Let $\varepsilon' = \delta$ and fix δ' as a witness corresponding to Lemma 6.7. By compactness and connectedness, let N be large enough that every two states are connected by a δ'-chain of size at most N. Now let s_1 and s_2 be arbitrary states. Let $s_1 = r_0, r_1, \ldots, r_N = s_2$ be a δ'-chain connecting s_1 and s_2. Now given any word m with $\ell(m) \geq N\tau$, we can factor $m = m_1 \cdots m_N$ with $\ell(m_i) \geq \tau$ for each i. Consider the m-trajectory of each r_i. By the lemma, the trajectories of r_i and r_{i+1} remain ε'-close, i.e., δ-close for all time. We can hop from the trajectory of s_1 to that of s_2 as follows: first apply m_1 to $s_1 = r_0$, then perturb by δ to the trajectory of r_1 and apply m_2, perturb to the trajectory of r_2 and so on, until we apply m_N and perturb to the trajectory of $r_N = s_2$. The result is that

$$s_2 \cdot m \in B_\delta(B_\delta(\cdots B_\delta(B_\delta(B_\delta(s_1) \cdot m_1) \cdot m_2) \cdot \cdots) \cdot m_N) \subseteq B_\varepsilon(s_1 \cdot m)$$

where the last inclusion comes from being self-stabilizing. So $\mathrm{d}_S(s_1 \cdot m, s_2 \cdot m) \leq \varepsilon$. Since m was an arbitrary word of length at least $N\tau$, we have that $T = N\tau$ witnesses that (M, S) is asymptotically forgetful for this ε. □

6.2 No Differentiation

In the introduction, we noted that the alternation-counter and differentiator cannot tolerate high-frequency noise and therefore are not permissible in our framework. Crudely speaking, any sort of operation which is sensitive to exactly what values a function takes on at exactly what times will not extend. They are operations on functions that are not uniformly continuous in the Young metric; hence they are not operations on Young words.

Nonetheless, there are real devices referred to as "differentiators", for example. Naturally

these real devices do not really behave as such on arbitrarily high-frequency input. In this section we give as an example a continuous automaton over the Young monoid whose behavior resembles that of an alternation-counter. We do not attempt to define precisely the sense of this resemblance.

To make the behavior uniformly continuous in the Young metric, we will have the device measure alternations only relative to some specified time-scale constant $\varepsilon > 0$, by waiting for the input to accumulate value at a new alphabet letter.

Example 6.8. Let $\Sigma = \{0, 1\}$ and $M = \mathrm{YM}(\Sigma)$. Define a continuous automaton (M, S) as follows. The state space is $S = M_\varepsilon \times [0, \infty)$.

If the state is written (v, a), then v continuously updates to record the last ε time units of input, and a updates as

$$\frac{da}{dt} = \frac{2\overline{v}(\varepsilon)(1 - u(t))}{\varepsilon}$$

where $u(t)$ is the letter currently being read.

If the input contains alternations which are separated by at least ε, then the a part of the state simply counts their number, assuming it started at zero. (The initial state of v also matters for whether the first input read is considered an alternation.) The one exception is when an alternation happened more recently than ε time units ago, in which case it has not yet been fully counted, allowing a to change continuously. If ε is very small, we might realistically never see a taking a non-integer value. We do not know if this technique could be reasonably generalized to a systematic way to define approximations of differentiation-like behavior with continuous automata over a Young monoid.

6.3 No Discontinuous State Jumps

Since the action is continuous, the image under the action of any connected set of inputs from any state (or connected set of states) is connected. A similar statement applies for path-connectedness. Young monoids are always path-connected, so in that case we can summarize

by saying that discontinuous state jumps are disallowed. The situation should be contrasted with the definition of hybrid automata, where state jumps are explicitly allowed.

It would be wrong to suppose that a hybrid automaton—or other notion of dynamical system—could not be modelled in the continuous automaton framework just because it has discrete state jumps. The following example illustrates how a translation might occur.

A basic example of a hybrid automaton is the bouncing ball automaton. (This automaton has no input other than time, so we might just as well call it a dynamical system.) The state consists of a height y and velocity v, obeys the differential equations

$$\frac{dy}{dt} = v \qquad \frac{dv}{dt} = -1$$

and additionally undergoes a (forced) discrete state transition whenever $v < 0$ and $y = 0$: then $v := -cv$, where c is a constant, $0 < c \leq 1$.

If $c = 1$ the behavior is easy to understand as the ball returns to the same height after every bounce, and the dynamics are periodic. If $c < 1$ then the ball will undergo infinitely many bounces in a finite time interval—it returns to a smaller height after each bounce, and takes a smaller amount of time to do so, with each of these values decaying exponentially. The above definition does not seem to specify the behavior after the end of this interval. Let us assert that the ball should be considered to be at rest with $y = 0$, $v = 0$ for all remaining time.

Example 6.9. Let $M = \mathbb{R}_{\geq 0}$. Let $S = \mathbb{R}_{\geq 0} \times \mathbb{R}/\sim$, where $(y, v) \sim (y', v')$ when either $y = y'$ and $v = v'$ or $y = 0$ and $v' = -cv$. (So S is endowed with the quotient topology.) The dynamics can be defined for all time essentially as described above, except now the "discrete jumps" no longer correspond to a state change at all. We assert without proof that this action is continuous, so (M, S) is a continuous automaton.

Another manner in which seemingly discontinuous state jumps might be modelled in the setting of continuous automata is disconnectedness in the input monoid. For example, con-

sider the free product $M = \Sigma^* * \mathrm{YM}(\Sigma')$ where Σ is a finite set and Σ' is a compact metric space. (We assume Σ and Σ' are disjoint.) Elements of M may be described as words which alternate between letters in Σ and (possibly empty) Young words over Σ'. The connected components correspond exactly to elements of Σ^*: as a set, the component contains all possible ways of interspering Young words amongst the letters of the discrete word.

Now consider a continuous automaton over M. When a letter in Σ occurs in the input, the new state need not be close to the old one. In fact, we can obtain a continuous automaton over M by combining any action of Σ^* with any continuous action of $\mathrm{YM}(\Sigma')$. A notable special case occurs when Σ' is a singleton, so $\mathrm{YM}(\Sigma')$ is a copy of $\mathbb{R}_{\geq 0}$. A possible interpretation would be of a system that evolves continuously over time but also responds to discrete input events, which may never occur, or may occur extremely rapidly in quick succession.

6.4 Notes

It seems highly plausible that results similar to Theorem 6.6 might have been proved in other contexts, but I'm not quite sure where to look for references. Theorem 6.6 considers only the compact case, but I expect it wouldn't be too hard to extend the result by considering a local version of asymptotic forgetfulness. Then the assumption would probably be that the state space is locally compact.

Overall, Section 6.1 considers only worst-case analysis. It would be natural to ask how much accuracy loss is expected or typical, but these questions would need to be considered in more specific contexts which provide additional mathematical structure, and would probably have less general answers. Such theorems would live closer to particular applications of concern.

In Section 6.2, I mentioned the example of electrical circuits which are called differentiators. A natural research question for extending the ideas of that section would be to give a complete translation of the possible behaviors of electrical circuits to the language of

continuous automata. Such research might include extending the notion of continuous automata to describe an idealized differentiator automaton which would be the limit of some sequence of approximations. It's also possible that there is no natural way to carry out such an extension, which would be an even more interesting outcome worthy of study.

If the idea of Section 6.3 for translating hybrid automata could be made to work, it might also enable a new understanding of the problem of Zeno behavior, long pondered in the hybrid automaton literature (see for example [9]). Zeno behavior occurs when a hybrid automaton, like the bouncing ball automaton for $c < 1$ executes infinitely many discrete transitions in a finite amount of time. In some cases, the run cannot be extended further. Note that hybrid automata are normally defined in a nondeterministic manner, which seems more natural in that setting. A complete understanding of the connection between hybrid automata and continuous automata would likely require a theory of nondeterministic continuous automata.

Example 6.9 and the question of translating hybrid automata generally serve to illustrate that the existence theory from Chapter 4 is still insufficient for practical needs, as it cannot cover non-smooth behavior. I didn't try to prove continuity rigorously because I don't yet have the tools to give the right proof. A proof given now specifically for this example would likely be rather ad hoc and unilluminating.

7 Conclusions

Recall the purpose stated in the introduction: to find robust notions to serve as a foundation for the theory of continuous-time computation, or better yet of a general theory of computation including both discrete and continuous-time computations. It's time to reflect on the (rather modest) progress I've made toward this goal. The main idea here is that the notion of a continuous automaton over a Young monoid is a plausible candidate for such a foundational notion.

To be clear, the task of establishing a foundation is a complex and subtle one. For example, Definition 1.1 (which is not original to me) defined a continuous automaton simply as a continuous monoid action. As useful as this definition may be, we cannot just assert that we have a foundation for all automata theory and declare victory. A theoretical foundation requires some explanation of how it can be built on, how to express relevant ideas inside the theory, how it unifies understanding of issues arising from different application domains.

Given the vast scope a foundation of continuous-time computation hopes to cover, of course my work here would ultimately only play a small part in establishing and justifying a new foundation for continuous-time computation. Nonetheless, I will state the following conclusions.

- Compact automata over Young monoids can be developed in parallel with finite automata theory; this contributes to unified understanding.
  - The space of words of a fixed length is compact, much as it is finite in the discrete case.

  - A Myhill-Nerode theorem can be proven for transducers, deducers, and open recognizers. The theorem is about as similar to the discrete version as could be expected.
  - Corollary 4.6 shows that it is possible to construct continuous automata in a manner somewhat analogous to the often-seen picture of a labeled directed graph. In this case the arrows of the graph become vector fields—one for each label—and the labels may be taken from any compact metric space.
  - *However*, I haven't shown how to carry out analogues of regular expressions or nondeterminism, though I think it is possible to do so. And Theorem 6.6 suggests that non-discrete compact automata are inherently more difficult to simulate than their finite counterparts.

- Continuous automata over Young monoids have a strong and direct connection to control theory, as they can equally be thought of as controlled systems. (See Section 4.3.) This connection was a surprise to me. Yet in hindsight it seems natural, and it contributes to unified understanding.

- Compact automata over Young monoids allow other natural examples, particularly the delay (see Examples 4.2 and 5.7) which are not easily understood from the differential equations perspective. In contexts where transducers are the main object of study, the delay seems like a fundamental sort of behavior.

Overall, what evidence I have tentatively suggests that this notion—a compact automaton over a Young monoid—is a good formalization for the intuitive concept of a machine which continuously receives information and can store a finite amount of it at any one time. With that said, I've only given the most rudimentary examples and methods for continuous automata over Young monoids. Especially, I haven't demonstrated how to really connect this theory with the various practical applications it is intended to support. (An exception is control theory, but even there the information so far is quite limited. How many methods

and theorems in control theory can actually be extended if continuous automata are taken as the basic concept?)

My main proposal for future work is a research program which would really test this proposed foundation by giving a detailed description, for as many application domains as possible, of the basic concepts, methods, and results of that domain in terms of the foundation, i.e., in terms of continuous automata over Young monoids. For example, I would speculate, on purely conceptual grounds, that the following scenarios, among many others, could be usefully thought about in terms of continuous automata over Young monoids.

- *Electrical circuits and circuit components.*
- *Organisms and robots interacting with an environment.*
- *Autonomous vehicles.*
- *Software interfaces.*
- *Physics and PDEs.*

To really carry out this project well will require the contribution of researchers with expertise in these areas beyond what I have or can easily acquire. And of course, it will require expanding the foundation itself, both because the work is far from finished (as I have repeatedly made note of at the end of each chapter), and because each specialized area will have its own unique demands to be met. We will know the project has succeeded when a theorem or technique from one community of researchers becomes available to another by virtue of being described in these terms—available not merely in a formal sense but in the sense that it is understood, digested, and readily exploitable for solving new problems in different areas. Such is the benefit of a good foundation.

Bibliography

[1] Rajeev Alur, Costas Courcoubetis, Thomas A Henzinger, and Pei-Hsin Ho. *Hybrid automata: An algorithmic approach to the specification and verification of hybrid systems*. Springer, 1993.

[2] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer, 2006.

[3] Errett Bishop. *Foundations of constructive analysis*. McGraw-Hill Book Company, 1967.

[4] Olivier Bournez and Manuel L Campagnolo. A survey on continuous time computations. In *New Computational Paradigms*, pages 383–423. Springer, 2008.

[5] Charles Castaing, Paul Raynaud De Fitte, and Michel Valadier. *Young measures on topological spaces: with applications in control theory and probability theory*, volume 571. Springer Science & Business Media, 2004.

[6] Liviu C Florescu and Christiane Godet-Thobie. *Young measures and compactness in measure spaces*. Walter de Gruyter, 2012.

[7] Thomas A Henzinger. *The theory of hybrid automata*. Springer, 2000.

[8] Emmanuel Jeandel. Topological automata. *Theory of Computing Systems*, 40(4):397–407, 2007.

[9] Karl Henrik Johansson, Magnus Egerstedt, John Lygeros, and Shankar Sastry. On the regularization of zeno hybrid automata. *Systems & Control Letters*, 38(3):141–150, 1999.

[10] Nancy Lynch, Roberto Segala, and Frits Vaandrager. Hybrid i/o automata. *Information and Computation*, 185(1):105–157, 2003.

[11] Alexander Rabinovich. Automata over continuous time. *Theoretical Computer Science*, 300(1):331–363, 2003.

[12] Kripasindhu Sikdar. On topological machines. *Journal of Computer and System Sciences*, 14(1):17–48, 1977.

[13] Michel Valadier. *A course on Young measures*. Department des Sciences Mathématiques, 1994.

[14] Cédric Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.

[15] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2009.

[16] Klaus Weihrauch. *Computable analysis: an introduction*. Springer Science & Business Media, 2012.

[17] Stephen Willard. *General topology*. Courier Corporation, 1970.

[18] Laurence Chisholm Young. *Lectures on the calculus of variations and optimal control theory*, volume 304. American Mathematical Soc., 1980.