Grundlehren der mathematischen Wissenschaften 338

A Series of Comprehensive Studies in Mathematics

Cédric Villani

Optimal Transport

Old and New

Springer

Cédric Villani

Unité de Mathématiques Pures et Appliquées (UMPA)
École Normale Supérieure de Lyon
46, allée d'Italie
69364 Lyon CX 07
France
cvillani@umpa.ens-lyon.fr

ISSN 0072-7830
ISBN 978-3-540-71049-3 e-ISBN 978-3-540-71050-9

Library of Congress Control Number: 2008932183

Mathematics Subject Classification Numbers (2000): 49-xx, 53-xx, 60-xx

Cover design: WMX Design GmbH, Heidelberg

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

Do mo chuisle mo chroí, Aëlle

Preface

When I was first approached for the 2005 edition of the Saint-Flour Probability Summer School, I was intrigued, flattered and scared.[1] Apart from the challenge posed by the teaching of a rather analytical subject to a probabilistic audience, there was the danger of producing a remake of my recent book *Topics in Optimal Transportation.*

However, I gradually realized that I was being offered a unique opportunity to rewrite the whole theory from a different perspective, with alternative proofs and a different focus, and a more probabilistic presentation; plus the incorporation of recent progress. Among the most striking of these recent advances, there was the rising awareness that John Mather's minimal measures had a lot to do with optimal transport, and that both theories could actually be embedded within a single framework. There was also the discovery that optimal transport could provide a robust synthetic approach to Ricci curvature bounds. These links with dynamical systems on one hand, differential geometry on the other hand, were only briefly alluded to in my first book; here on the contrary they will be at the basis of the presentation. To summarize: more probability, more geometry, and more dynamical systems. Of course there cannot be more of everything, so in some sense there is less analysis and less physics, and also there are fewer digressions.

So the present course is by no means a reduction or an expansion of my previous book, but should be regarded as a complementary reading. Both sources can be read independently, or together, and hopefully the complementarity of points of view will have pedagogical value.

[1] Fans of Tom Waits may have identified this quotation.

Throughout the book I have tried to optimize the results and the presentation, to provide complete and self-contained proofs of the most important results, and comprehensive bibliographical notes — a dauntingly difficult task in view of the rapid expansion of the literature. Many statements and theorems have been written specifically for this course, and many results appear in rather sharp form for the first time. I also added several appendices, either to present some domains of mathematics to non-experts, or to provide proofs of important auxiliary results. All this has resulted in a rapid growth of the document, which in the end is about six times (!) the size that I had planned initially. So **the non-expert reader is advised to skip long proofs at first reading**, and concentrate on explanations, statements, examples and sketches of proofs when they are available.

About terminology: For some reason I decided to switch from “transportation” to “transport”, but this really is a matter of taste.

For people who are already familiar with the theory of optimal transport, here are some more serious changes.

Part I is devoted to a qualitative description of optimal transport. The dynamical point of view is given a prominent role from the beginning, with Robert McCann’s concept of displacement interpolation. This notion is discussed before any theorem about the solvability of the Monge problem, in an abstract setting of “Lagrangian action” which generalizes the notion of length space. This provides a unified picture of recent developments dealing with various classes of cost functions, in a smooth or nonsmooth context.

I also wrote down in detail some important estimates by John Mather, well-known in certain circles, and made extensive use of them, in particular to prove the Lipschitz regularity of “intermediate” transport maps (starting from some intermediate time, rather than from initial time). Then the absolute continuity of displacement interpolants comes for free, and this gives a more unified picture of the Mather and Monge–Kantorovich theories. I rewrote in this way the classical theorems of solvability of the Monge problem for quadratic cost in Euclidean space. Finally, this approach allows one to treat change of variables formulas associated with optimal transport by means of changes of variables that are Lipschitz, and not just with bounded variation.

Part II discusses optimal transport in Riemannian geometry, a line of research which started around 2000; I have rewritten all these applications in terms of Ricci curvature, or more precisely curvature-dimension bounds. This part opens with an introduction to Ricci curvature, hopefully readable without any prior knowledge of this notion.

Part III presents a synthetic treatment of Ricci curvature bounds in metric-measure spaces. It starts with a presentation of the theory of Gromov–Hausdorff convergence; all the rest is based on recent research papers mainly due to John Lott, Karl-Theodor Sturm and myself.

In all three parts, noncompact situations will be systematically treated, either by limiting processes, or by restriction arguments (the restriction of an optimal transport is still optimal; this is a simple but powerful principle). The notion of approximate differentiability, introduced in the field by Luigi Ambrosio, appears to be particularly handy in the study of optimal transport in noncompact Riemannian manifolds.

Several parts of the subject are not developed as much as they would deserve. Numerical simulation is not addressed at all, except for a few comments in the concluding part. The regularity theory of optimal transport is described in Chapter 12 (including the remarkable recent works of Xu-Jia Wang, Neil Trudinger and Grégoire Loeper), but without the core proofs and latest developments; this is not only because of the technicality of the subject, but also because smoothness is not needed in the rest of the book. Still another poorly developed subject is the Monge–Mather–Mañé problem arising in dynamical systems, and including as a variant the optimal transport problem when the cost function is a distance. This topic is discussed in several treatises, such as Albert Fathi's monograph, *Weak KAM theorem in Lagrangian dynamics*; but now it would be desirable to rewrite everything in a framework that also encompasses the optimal transport problem. An important step in this direction was recently performed by Patrick Bernard and Boris Buffoni. In Chapter 8 I shall provide an introduction to Mather's theory, but there would be much more to say.

The treatment of Chapter 22 (concentration of measure) is strongly influenced by Michel Ledoux's book, *The Concentration of Measure Phenomenon*; while the results of Chapters 23 to 25 owe a lot to the monograph by Luigi Ambrosio, Nicola Gigli and Giuseppe Savaré, *Gradient flows in metric spaces and in the space of probability measures.* Both references are warmly recommended complementary reading. One can also consult the two-volume treatise by Svetlozar Rachev and Ludger Rüschendorf, *Mass Transportation Problems*, for many applications of optimal transport to various fields of probability theory.

While writing this text I asked for help from a number of friends and collaborators. Among them, Luigi Ambrosio and John Lott are the ones whom I requested most to contribute; this book owes a lot to their detailed comments and suggestions. Most of Part III, but also significant portions of Parts I and II, are made up with ideas taken from my collaborations with John, which started in 2004 as I was enjoying the hospitality of the Miller Institute in Berkeley. Frequent discussions with Patrick Bernard and Albert Fathi allowed me to get the links between optimal transport and John Mather's theory, which were a key to the presentation in Part I; John himself gave precious hints about the history of the subject. Neil Trudinger and Xu-Jia Wang spent vast amounts of time teaching me the regularity theory of Monge–Ampère equations. Alessio Figalli took up the dreadful challenge to check the entire set of notes from first to last page. Apart from these people, I got valuable help from Stefano Bianchini, François Bolley, Yann Brenier, Xavier Cabré, Vincent Calvez, José Antonio Carrillo, Dario Cordero-Erausquin, Denis Feyel, Sylvain Gallot, Wilfrid Gangbo, Diogo Aguiar Gomes, Nathaël Gozlan, Arnaud Guillin, Nicolas Juillet, Kazuhiro Kuwae, Michel Ledoux, Grégoire Loeper, Francesco Maggi, Robert McCann, Shin-ichi Ohta, Vladimir Oliker, Yann Ollivier, Felix Otto, Ludger Rüschendorf, Giuseppe Savaré, Walter Schachermayer, Benedikt Schulte, Theo Sturm, Josef Teichmann, Anthon Thalmaier, Hermann Thorisson, Süleyman Üstünel, Anatoly Vershik, and others.

Short versions of this course were tried on mixed audiences in the Universities of Bonn, Dortmund, Grenoble and Orléans, as well as the Borel seminar in Leysin and the IHES in Bures-sur-Yvette. Part of the writing was done during stays at the marvelous MFO Institute in Oberwolfach, the CIRM in Luminy, and the Australian National University in Canberra. All these institutions are warmly thanked.

It is a pleasure to thank Jean Picard for all his organization work on the 2005 Saint-Flour summer school; and the participants for their questions, comments and bug-tracking, in particular Sylvain Arlot (great bug-tracker!), Fabrice Baudoin, Jérôme Demange, Steve Evans (whom I also thank for his beautiful lectures), Christophe Leuridan, Jan Obłój, Erwan Saint Loubert Bié, and others. I extend these thanks to the joyful group of young PhD students and maîtres de conférences with whom I spent such a good time on excursions, restaurants, quantum ping-pong and other activities, making my stay in Saint-Flour truly wonderful (with special thanks to my personal driver, Stéphane Loisel, and my table tennis sparring-partner and adversary, François Simenhaus). I will cherish my visit there in memory as long as I live!

Typing these notes was mostly performed on my (now defunct) faithful laptop Torsten, a gift of the Miller Institute. Support by the Agence Nationale de la Recherche and Institut Universitaire de France is acknowledged. My eternal gratitude goes to those who made fine typesetting accessible to every mathematician, most notably Donald Knuth for TeX, and the developers of LaTeX, BibTeX and XFig. Final thanks to Catriona Byrne and her team for a great editing process.

As usual, I encourage all readers to report mistakes and misprints. **I will maintain a list of errata, accessible from my Web page**.

Lyon, June 2008 *Cédric Villani*

Contents

Part II Optimal transport and Riemannian geometry 353

Part III Synthetic treatment of Ricci curvature 731

Conventions

Axioms

I use the classical axioms of set theory; not the full version of the axiom of choice (only the classical axiom of "countable dependent choice").

Sets and structures

Id is the identity mapping, whatever the space. If A is a set then the function 1_A is the indicator function of A: $1_A(x) = 1$ if $x \in A$, and 0 otherwise. If F is a formula, then 1_F is the indicator function of the set defined by the formula F.

If f and g are two functions, then (f, g) is the function $x \longmapsto (f(x), g(x))$. The composition $f \circ g$ will often be denoted by $f(g)$.

$\mathbb{N}$ is the set of *positive* integers: $\mathbb{N} = \{1, 2, 3, \ldots\}$. A sequence is written $(x_k)_{k\in\mathbb{N}}$, or simply, when no confusion seems possible, (x_k).

$\mathbb{R}$ is the set of real numbers. When I write $\mathbb{R}^n$ it is implicitly assumed that n is a positive integer. The Euclidean scalar product between two vectors a and b in $\mathbb{R}^n$ is denoted interchangeably by $a \cdot b$ or $\langle a, b\rangle$. The Euclidean norm will be denoted simply by $|\cdot|$, independently of the dimension n.

$M_n(\mathbb{R})$ is the space of real $n \times n$ matrices, and I_n the $n \times n$ identity matrix. The trace of a matrix M will be denoted by $\operatorname{tr} M$, its determinant by $\det M$, its adjoint by M^*, and its Hilbert–Schmidt norm $\sqrt{\operatorname{tr}(M^*M)}$ by $\|M\|_{\mathrm{HS}}$ (or just $\|M\|$).

Unless otherwise stated, Riemannian manifolds appearing in the text are finite-dimensional, smooth and complete. If a Riemannian manifold M is given, I shall usually denote by n its dimension, by d the geodesic distance on M, and by vol the volume ($= n$-dimensional

Hausdorff) measure on M. The tangent space at x will be denoted by T_xM, and the tangent bundle by TM. The norm on T_xM will most of the time be denoted by $|\cdot|$, as in $\mathbb{R}^n$, without explicit mention of the point x. (The symbol $\|\cdot\|$ will be reserved for special norms or functional norms.) If S is a set without smooth structure, the notation T_xS will instead denote the tangent cone to S at x (Definition 10.46).

If Q is a quadratic form defined on $\mathbb{R}^n$, or on the tangent bundle of a manifold, its value on a (tangent) vector v will be denoted by $\langle Q\cdot v,\, v\rangle$, or simply $Q(v)$.

The open ball of radius r and center x in a metric space $\mathcal{X}$ is denoted interchangeably by $B(x,r)$ or $B_r(x)$. If $\mathcal{X}$ is a Riemannian manifold, the distance is of course the geodesic distance. The closed ball will be denoted interchangeably by $B[x,r]$ or $B_{r]}(x)$. The diameter of a metric space $\mathcal{X}$ will be denoted by $\operatorname{diam}(\mathcal{X})$.

The closure of a set A in a metric space will be denoted by $\overline{A}$ (this is also the set of all limits of sequences with values in A).

A metric space $\mathcal{X}$ is said to be *locally compact* if every point $x \in \mathcal{X}$ admits a compact neighborhood; and *boundedly compact* if every closed and bounded subset of $\mathcal{X}$ is compact.

A map f between metric spaces $(\mathcal{X}, d)$ and $(\mathcal{X}', d')$ is said to be C-Lipschitz if $d'(f(x), f(y)) \le C\, d(x,y)$ for all x, y in $\mathcal{X}$. The best admissible constant C is then denoted by $\|f\|_{\mathrm{Lip}}$.

A map is said to be locally Lipschitz if it is Lipschitz on bounded sets, *not necessarily compact* (so it makes sense to speak of a locally Lipschitz map defined almost everywhere).

A curve in a space $\mathcal{X}$ is a continuous map defined on an interval of $\mathbb{R}$, valued in $\mathcal{X}$. For me the words “curve” and “path” are synonymous. The time-t evaluation map e_t is defined by $e_t(\gamma) = \gamma_t = \gamma(t)$.

If γ is a curve defined from an interval of $\mathbb{R}$ into a metric space, its length will be denoted by $\mathcal{L}(\gamma)$, and its speed by $|\dot{\gamma}|$; definitions are recalled on p. 119.

Usually geodesics will be *minimizing*, *constant-speed* geodesic curves. If $\mathcal{X}$ is a metric space, $\Gamma(\mathcal{X})$ stands for the space of all geodesics $\gamma : [0,1] \to \mathcal{X}$.

Being given x_0 and x_1 in a metric space, I denote by $[x_0, x_1]_t$ the set of all t-barycenters of x_0 and x_1, as defined on p. 393. If A_0 and A_1 are two sets, then $[A_0, A_1]_t$ stands for the set of all $[x_0, x_1]_t$ with $(x_0, x_1) \in A_0 \times A_1$.

Function spaces

$C(\mathcal{X})$ is the space of continuous functions $\mathcal{X} \to \mathbb{R}$, $C_b(\mathcal{X})$ the space of bounded continuous functions $\mathcal{X} \to \mathbb{R}$; and $C_0(\mathcal{X})$ the space of continuous functions $\mathcal{X} \to \mathbb{R}$ converging to 0 at infinity; all of them are equipped with the norm of uniform convergence $\|\varphi\|_\infty = \sup |\varphi|$. Then $C_b^k(\mathcal{X})$ is the space of k-times continuously differentiable functions $u : \mathcal{X} \to \mathbb{R}$, such that all the partial derivatives of u up to order k are bounded; it is equipped with the norm given by the supremum of all norms $\|\partial u\|_{C_b}$, where ∂u is a partial derivative of order at most k; $C_c^k(\mathcal{X})$ is the space of k-times continuously differentiable functions with compact support; etc. When the target space is not $\mathbb{R}$ but some other space $\mathcal{Y}$, the notation is transformed in an obvious way: $C(\mathcal{X};\mathcal{Y})$, etc.

L^p is the Lebesgue space of exponent p; the space and the measure will often be implicit, but clear from the context.

Calculus

The derivative of a function $u = u(t)$, defined on an interval of $\mathbb{R}$ and valued in $\mathbb{R}^n$ or in a smooth manifold, will be denoted by u', or more often by $\dot{u}$. The notation d^+u/dt stands for the upper right-derivative of a real-valued function u: $d^+u/dt = \limsup_{s\downarrow 0}[u(t+s) - u(t)]/s$.

If u is a function of several variables, the partial derivative with respect to the variable t will be denoted by $\partial_t u$, or $\partial u/\partial t$. *The notation u_t does not stand for $\partial_t u$, but for $u(t)$.*

The gradient operator will be denoted by grad or simply ∇; the divergence operator by div or $\nabla\cdot$; the Laplace operator by Δ; the Hessian operator by Hess or ∇^2 (so ∇^2 *does not* stand for the Laplace operator). The notation is the same in $\mathbb{R}^n$ or in a Riemannian manifold. Δ is the divergence of the gradient, so it is typically a nonpositive operator. The value of the gradient of f at point x will be denoted either by $\nabla_x f$ or $\nabla f(x)$. The notation $\widetilde{\nabla}$ stands for the approximate gradient, introduced in Definition 10.2.

If T is a map $\mathbb{R}^n \to \mathbb{R}^n$, ∇T stands for the Jacobian matrix of T, that is the matrix of all partial derivatives $(\partial T_i/\partial x_j)$ $(1 \le i, j \le n)$.

All these differential operators will be applied to (smooth) functions but also to measures, by duality. For instance, the Laplacian of a measure μ is defined via the identity $\int \zeta\, d(\Delta\mu) = \int (\Delta\zeta)\, d\mu$ $(\zeta \in C_c^2)$. The notation is consistent in the sense that $\Delta(f\mathrm{vol}) = (\Delta f)\,\mathrm{vol}$. Similarly, I shall take the divergence of a vector-valued measure, etc.

$f = o(g)$ means $f/g \longrightarrow 0$ (in an asymptotic regime that should be clear from the context), while $f = O(g)$ means that f/g is bounded.

log stands for the natural logarithm with base e.

The positive and negative parts of $x \in \mathbb{R}$ are defined respectively by $x_+ = \max(x, 0)$ and $x_- = \max(-x, 0)$; both are nonnegative, and $|x| = x_+ + x_-$. The notation $a \wedge b$ will sometimes be used for $\min(a, b)$. All these notions are extended in the usual way to functions and also to signed measures.

Probability measures

δ_x is the Dirac mass at point x.

All measures considered in the text are Borel measures on **Polish spaces**, which are complete, separable metric spaces, equipped with their Borel σ-algebra. I shall usually not use the completed σ-algebra, except on some rare occasions (emphasized in the text) in Chapter 5.

A measure is said to be finite if it has finite mass, and locally finite if it attributes finite mass to compact sets.

The space of Borel probability measures on $\mathcal{X}$ is denoted by $P(\mathcal{X})$, the space of finite Borel measures by $M_+(\mathcal{X})$, the space of signed finite Borel measures by $M(\mathcal{X})$. The total variation of μ is denoted by $\|\mu\|_{\mathrm{TV}}$.

The integral of a function f with respect to a probability measure μ will be denoted interchangeably by $\int f(x)\, d\mu(x)$ or $\int f(x)\, \mu(dx)$ or $\int f\, d\mu$.

If μ is a Borel measure on a topological space $\mathcal{X}$, a set N is said to be μ-negligible if N is included in a Borel set of zero μ-measure. Then μ is said to be concentrated on a set C if $\mathcal{X} \setminus C$ is negligible. (If C itself is Borel measurable, this is of course equivalent to $\mu[\mathcal{X} \setminus C] = 0$.) By abuse of language, I may say that $\mathcal{X}$ has full μ-measure if μ is concentrated on $\mathcal{X}$.

If μ is a Borel measure, its support $\operatorname{Spt} \mu$ is the smallest *closed* set on which it is concentrated. The same notation Spt will be used for the support of a continuous function.

If μ is a Borel measure on $\mathcal{X}$, and T is a Borel map $\mathcal{X} \to \mathcal{Y}$, then $T_\# \mu$ stands for the image measure[2] (or push-forward) of μ by T: It is a Borel measure on $\mathcal{Y}$, defined by $(T_\#\mu)[A] = \mu[T^{-1}(A)]$.

The law of a random variable X defined on a probability space $(\Omega, \mathbb{P})$ is denoted by $\operatorname{law}(X)$; this is the same as $X_\# \mathbb{P}$.

The weak topology on $P(\mathcal{X})$ (or topology of weak convergence, or narrow topology) is induced by convergence against $C_b(\mathcal{X})$, i.e. *bounded*

[2] Depending on the authors, the measure $T_\#\mu$ is often denoted by $T\#\mu$, $T_*\mu$, $T(\mu)$, $T\mu$, $\int \delta_{T(a)}\, \mu(da)$, $\mu \circ T^{-1}$, μT^{-1}, or $\mu[T \in \cdot\,]$.

continuous test functions. If $\mathcal{X}$ is Polish, then the space $P(\mathcal{X})$ itself is Polish. Unless explicitly stated, I do not use the weak-$*$ topology of measures (induced by $C_0(\mathcal{X})$ or $C_c(\mathcal{X})$).

When a probability measure is clearly specified by the context, it will sometimes be denoted just by $\mathbb{P}$, and the associated integral, or expectation, will be denoted by $\mathbb{E}$.

If $\pi(dx\,dy)$ is a probability measure in two variables $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, its marginal (or projection) on $\mathcal{X}$ (resp. $\mathcal{Y}$) is the measure $X_{\#}\pi$ (resp. $Y_{\#}\pi$), where $X(x,y) = x$, $Y(x,y) = y$. If (x,y) is random with law $(x,y) = \pi$, then the conditional law of x given y is denoted by $\pi(dx|y)$; this is a measurable function $\mathcal{Y} \to P(\mathcal{X})$, obtained by disintegrating π along its y-marginal. The conditional law of y given x will be denoted by $\pi(dy|x)$.

A measure μ is said to be absolutely continuous with respect to a measure ν if there exists a measurable function f such that $\mu = f\,\nu$.

Notation specific to optimal transport and related fields

If $\mu \in P(\mathcal{X})$ and $\nu \in P(\mathcal{Y})$ are given, then $\Pi(\mu,\nu)$ is the set of all joint probability measures on $\mathcal{X} \times \mathcal{Y}$ whose marginals are μ and ν.

$C(\mu,\nu)$ is the optimal (total) cost between μ and ν, see p. 80. It implicitly depends on the choice of a cost function $c(x,y)$.

For any $p \in [1,+\infty)$, W_p is the Wasserstein distance of order p, see Definition 6.1; and $P_p(\mathcal{X})$ is the Wasserstein space of order p, i.e. the set of probability measures with finite moments of order p, equipped with the distance W_p, see Definition 6.4.

$P_c(\mathcal{X})$ is the set of probability measures on $\mathcal{X}$ with compact support.

If a reference measure ν on $\mathcal{X}$ is specified, then $P^{\mathrm{ac}}(\mathcal{X})$ (resp. $P_p^{\mathrm{ac}}(\mathcal{X})$, $P_c^{\mathrm{ac}}(\mathcal{X})$) stands for those elements of $P(\mathcal{X})$ (resp. $P_p(\mathcal{X})$, $P_c(\mathcal{X})$) which are absolutely continuous with respect to ν.

$\mathcal{DC}_N$ is the displacement convexity class of order N (N plays the role of a dimension); this is a family of convex functions, defined on p. 443 and in Definition 17.1.

U_ν is a functional defined on $P(\mathcal{X})$; it depends on a convex function U and a reference measure ν on $\mathcal{X}$. This functional will be defined at various levels of generality, first in equation (15.2), then in Definition 29.1 and Theorem 30.4.

$U^\beta_{\pi,\nu}$ is another functional on $P(\mathcal{X})$, which involves not only a convex function U and a reference measure ν, but also a coupling π and a distortion coefficient β, which is a nonnegative function on $\mathcal{X} \times \mathcal{X}$: See again Definition 29.1 and Theorem 30.4.

The Γ and Γ_2 operators are quadratic differential operators associated with a diffusion operator; they are defined in (14.47) and (14.48).

$\beta_t^{(K,N)}$ is the notation for the distortion coefficients that will play a prominent role in these notes; they are defined in (14.61).

CD(K,N) means "curvature-dimension condition (K,N)", which morally means that the Ricci curvature is bounded below by Kg (K a real number, g the Riemannian metric) and the dimension is bounded above by N (a real number which is not less than 1).

If $c(x,y)$ is a cost function then $\check{c}(y,x) = c(x,y)$. Similarly, if $\pi(dx\,dy)$ is a coupling, then $\check{\pi}$ is the coupling obtained by swapping variables, that is $\check{\pi}(dy\,dx) = \pi(dx\,dy)$, or more rigorously, $\check{\pi} = S_\#\pi$, where $S(x,y) = (y,x)$.

Assumptions **(Super)**, **(Twist)**, **(Lip)**, **(SC)**, **(locLip)**, **(locSC)**, **(H∞)** are defined on p. 234, **(STwist)** on p. 299, **(Cut^{n-1})** on p. 303.

Introduction

To start, I shall recall in Chapter 1 some basic facts about couplings and changes of variables, including definitions, a short list of famous couplings (Knothe–Rosenblatt coupling, Moser coupling, optimal coupling, etc.); and some important basic formulas about change of variables, conservation of mass, and linear diffusion equations.

In Chapter 2 I shall present, without detailed proofs, three applications of optimal coupling techniques, providing a flavor of the kind of applications that will be considered later.

Finally, Chapter 3 is a short historical perspective about the foundations and development of optimal coupling theory.

1

Couplings and changes of variables

Couplings are very well-known in all branches of probability theory, but since they will occur again and again in this course, it might be a good idea to start with some basic reminders and a few more technical issues.

Definition 1.1 (Coupling). *Let $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$ be two probability spaces. Coupling μ and ν means constructing two random variables X and Y on some probability space $(\Omega, \mathbb{P})$, such that* $\mathrm{law}\,(X) = \mu$, $\mathrm{law}\,(Y) = \nu$*. The couple (X, Y) is called a coupling of (μ, ν). By abuse of language, the law of (X, Y) is also called a coupling of (μ, ν).*

If μ and ν are the only laws in the problem, then without loss of generality one may choose $\Omega = \mathcal{X} \times \mathcal{Y}$. In a more measure-theoretical formulation, coupling μ and ν means constructing a measure π on $\mathcal{X} \times \mathcal{Y}$ such that π admits μ and ν as **marginals** on $\mathcal{X}$ and $\mathcal{Y}$ respectively. The following three statements are equivalent ways to rephrase that marginal condition:

- $(\mathrm{proj}_{\mathcal{X}})_{\#}\pi = \mu$, $(\mathrm{proj}_{\mathcal{Y}})_{\#}\pi = \nu$, where $\mathrm{proj}_{\mathcal{X}}$ and $\mathrm{proj}_{\mathcal{Y}}$ respectively stand for the projection maps $(x, y) \longmapsto x$ and $(x, y) \longmapsto y$;
- For all measurable sets $A \subset \mathcal{X}$, $B \subset \mathcal{Y}$, one has $\pi[A \times \mathcal{Y}] = \mu[A]$, $\pi[\mathcal{X} \times B] = \nu[B]$;
- For all integrable (resp. nonnegative) measurable functions φ, ψ on $\mathcal{X}, \mathcal{Y}$,

$$\int_{\mathcal{X}\times\mathcal{Y}} \big(\varphi(x) + \psi(y)\big)\, d\pi(x, y) = \int_{\mathcal{X}} \varphi\, d\mu + \int_{\mathcal{Y}} \psi\, d\nu.$$

C. Villani, *Optimal Transport*, Grundlehren der mathematischen
Wissenschaften 338,

A first remark about couplings is that they always exist: at least there is the **trivial coupling**, in which the variables X and Y are **independent** (so their joint law is the tensor product $\mu \otimes \nu$). This can hardly be called a coupling, since the value of X does not give any information about the value of Y. Another extreme is when all the information about the value of Y is contained in the value of X, in other words Y is just a function of X. This motivates the following definition (in which X and Y do not play symmetric roles).

Definition 1.2 (Deterministic coupling). *With the notation of Definition 1.1, a coupling (X, Y) is said to be deterministic if there exists a measurable function $T : \mathcal{X} \to \mathcal{Y}$ such that $Y = T(X)$.*

To say that (X, Y) is a deterministic coupling of μ and ν is strictly equivalent to any one of the four statements below:

- (X, Y) is a coupling of μ and ν whose law π is concentrated on the *graph* of a measurable function $T : \mathcal{X} \to \mathcal{Y}$;
- X has law μ and $Y = T(X)$, where $T_{\#}\mu = \nu$;
- X has law μ and $Y = T(X)$, where T is a **change of variables** from μ to ν: for all ν-integrable (resp. nonnegative measurable) functions φ,
$$\int_{\mathcal{Y}} \varphi(y)\, d\nu(y) = \int_{\mathcal{X}} \varphi\big(T(x)\big)\, d\mu(x); \tag{1.1}$$
- $\pi = (\mathrm{Id}\,, T)_{\#}\mu$.

The map T appearing in all these statements is the same and is uniquely defined μ-almost surely (when the joint law of (X, Y) has been fixed). The converse is true: If T and $\widetilde{T}$ coincide μ-almost surely, then $T_{\#}\mu = \widetilde{T}_{\#}\mu$. It is common to call T the **transport map**: Informally, one can say that T transports the mass represented by the measure μ, to the mass represented by the measure ν.

Unlike couplings, deterministic couplings do not always exist: Just think of the case when μ is a Dirac mass and ν is not. But there may also be infinitely many deterministic couplings between two given probability measures.

Some famous couplings

Here below are some of the most famous couplings used in mathematics — of course the list is far from complete, since everybody has his or her own preferred coupling technique. Each of these couplings comes with its own natural setting; this variety of assumptions reflects the variety of constructions. (This is a good reason to state each of them with some generality.)

1. The **measurable isomorphism.** Let $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$ be two Polish (i.e. complete, separable, metric) probability spaces without atom (i.e. no single point carries a positive mass). Then there exists a (nonunique) measurable bijection $T : \mathcal{X} \to \mathcal{Y}$ such that $T_{\#}\mu = \nu$, $(T^{-1})_{\#}\nu = \mu$. In that sense, all atomless Polish probability spaces are isomorphic, and, say, isomorphic to the space $\mathcal{Y} = [0, 1]$ equipped with the Lebesgue measure. Powerful as that theorem may seem, in practice the map T is very singular; as a good exercise, the reader might try to construct it "explicitly", in terms of cumulative distribution functions, when $\mathcal{X} = \mathbb{R}$ and $\mathcal{Y} = [0, 1]$ (issues do arise when the density of μ vanishes at some places). Experience shows that it is quite easy to fall into logical traps when working with the measurable isomorphism, and my advice is to never use it.

2. The **Moser mapping.** Let $\mathcal{X}$ be a smooth compact Riemannian manifold with volume vol, and let f, g be Lipschitz continuous positive probability densities on $\mathcal{X}$; then there exists a deterministic coupling of $\mu = f\,\mathrm{vol}$ and $\nu = g\,\mathrm{vol}$, constructed by resolution of an elliptic equation. On the positive side, there is a somewhat explicit representation of the transport map T, and it is as smooth as can be: if f, g are $C^{k,\alpha}$ then T is $C^{k+1,\alpha}$. The formula is given in the Appendix at the end of this chapter. The same construction works in $\mathbb{R}^n$ provided that f and g decay fast enough at infinity; and it is robust enough to accommodate for variants.

3. The **increasing rearrangement** on $\mathbb{R}$. Let μ, ν be two probability measures on $\mathbb{R}$; define their cumulative distribution functions by

$$F(x) = \int_{-\infty}^{x} d\mu, \qquad G(y) = \int_{-\infty}^{y} d\nu.$$

   Further define their right-continuous inverses by

$$F^{-1}(t) = \inf\Big\{x \in \mathbb{R};\ \ F(x) > t\Big\};$$
$$G^{-1}(t) = \inf\Big\{y \in \mathbb{R};\ \ G(y) > t\Big\};$$

and set

$$T = G^{-1} \circ F.$$

If μ does not have atoms, then $T_{\#}\mu = \nu$. This rearrangement is quite simple, explicit, as smooth as can be, and enjoys good geometric properties.

4. The **Knothe–Rosenblatt rearrangement** in $\mathbb{R}^n$. Let μ and ν be two probability measures on $\mathbb{R}^n$, such that μ is absolutely continuous with respect to Lebesgue measure. Then define a coupling of μ and ν as follows.

   Step 1: Take the marginal on the first variable: this gives probability measures $\mu_1(dx_1)$, $\nu_1(dy_1)$ on $\mathbb{R}$, with μ_1 being atomless. Then define $y_1 = T_1(x_1)$ by the formula for the increasing rearrangement of μ_1 into ν_1.

   Step 2: Now take the marginal on the first two variables and disintegrate it with respect to the first variable. This gives probability measures $\mu_2(dx_1\, dx_2) = \mu_1(dx_1)\, \mu_2(dx_2|x_1)$, $\nu_2(dy_1\, dy_2) = \nu_1(dy_1)\, \nu_2(dy_2|y_1)$. Then, for each given $y_1 \in \mathbb{R}$, set $y_1 = T_1(x_1)$, and define $y_2 = T_2(x_2; x_1)$ by the formula for the increasing rearrangement of $\mu_2(dx_2|x_1)$ into $\nu_2(dy_2|y_1)$. (See Figure 1.1.)

   Then repeat the construction, adding variables one after the other and defining $y_3 = T_3(x_3; x_1, x_2)$; etc. After n steps, this produces a map $y = T(x)$ which transports μ to ν, and in practical situations might be computed explicitly with little effort. Moreover, the Jacobian matrix of the change of variables T is (by construction) upper triangular with positive entries on the diagonal; this makes it suitable for various geometric applications. On the negative side, this mapping does not satisfy many interesting intrinsic properties; it is not invariant under isometries of $\mathbb{R}^n$, not even under relabeling of coordinates.

5. The **Holley coupling** on a lattice. Let μ and ν be two discrete probabilities on a finite lattice Λ, say $\{0,1\}^N$, equipped with the natural partial ordering ($x \le y$ if $x_n \le y_n$ for all n). Assume that

$$\forall x, y \in \Lambda, \qquad \mu[\inf(x,y)]\ \nu[\sup(x,y)] \ \ge\ \mu[x]\, \nu[y]. \tag{1.2}$$

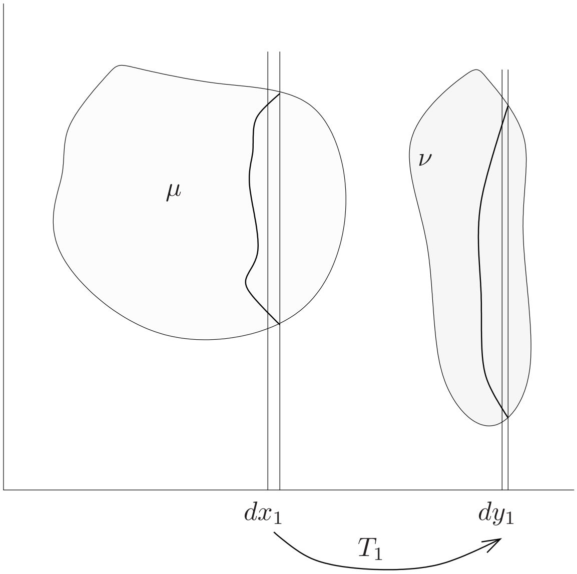

Fig. 1.1. Second step in the construction of the Knothe–Rosenblatt map: After the correspondance $x_1 \to y_1$ has been determined, the conditional probability of x_2 (seen as a one-dimensional probability on a small "slice" of width dx_1) can be transported to the conditional probability of y_2 (seen as a one-dimensional probability on a slice of width dy_1).

Then there exists a coupling (X, Y) of (μ, ν) with $X \leq Y$. The situation above appears in a number of problems in statistical mechanics, in connection with the so-called FKG (Fortuin–Kasteleyn–Ginibre) inequalities. Inequality (1.2) intuitively says that ν puts more mass on large values than μ.

6. **Probabilistic representation formulas** for solutions of partial differential equations. There are hundreds of them (if not thousands), representing solutions of diffusion, transport or jump processes as the laws of various deterministic or stochastic processes. Some of them are recalled later in this chapter.

7. The **exact coupling** of two stochastic processes, or Markov chains. Two realizations of a stochastic process are started at initial time, and when they happen to be in the same state at some time, they are merged: from that time on, they follow the same path and accordingly, have the same law. For two Markov chains which are started independently, this is called the **classical coupling**. There

are many variants with important differences which are all intended to make two trajectories close to each other after some time: the **Ornstein coupling**, the **ε-coupling** (in which one requires the two variables to be close, rather than to occupy the same state), the **shift-coupling** (in which one allows an additional time-shift), etc.

8. The **optimal coupling** or **optimal transport**. Here one introduces a **cost function** $c(x,y)$ on $\mathcal{X} \times \mathcal{Y}$, that can be interpreted as the work needed to move one unit of mass from location x to location y. Then one considers the **Monge–Kantorovich minimization problem**
$$\inf\ \mathbb{E}\ c(X,Y),$$
where the pair (X,Y) runs over all possible couplings of (μ,ν); or equivalently, in terms of measures,
$$\inf \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\, d\pi(x,y),$$
where the infimum runs over all joint probability measures π on $\mathcal{X}\times\mathcal{Y}$ with marginals μ and ν. Such joint measures are called **transference plans** (or transport plans, or transportation plans); those achieving the infimum are called **optimal transference plans**.

Of course, the solution of the Monge–Kantorovich problem depends on the cost function c. The cost function and the probability spaces here can be very general, and some nontrivial results can be obtained as soon as, say, c is lower semicontinuous and $\mathcal{X},\mathcal{Y}$ are Polish spaces. Even the apparently trivial choice $c(x,y) = 1_{x\neq y}$ appears in the probabilistic interpretation of total variation:
$$\|\mu-\nu\|_{TV} = 2\ \inf\Big\{\mathbb{E}\, 1_{X\neq Y};\quad \text{law}\,(X)=\mu,\ \text{law}\,(Y)=\nu\Big\}.$$
Cost functions valued in $\{0,1\}$ also occur naturally in Strassen's duality theorem.

Under certain assumptions one can guarantee that the optimal coupling really is deterministic; the search of deterministic optimal couplings (or Monge couplings) is called the **Monge problem**. A solution of the Monge problem yields a plan to transport the mass at minimal cost with a recipe that associates to each point x a single point y. ("*No mass shall be split.*") To guarantee the existence of solutions to the

Monge problem, two kinds of assumptions are natural: First, c should "vary enough" in some sense (think that the constant cost function will allow for arbitrary minimizers), and secondly, μ should enjoy some regularity property (at least Dirac masses should be ruled out!). Here is a typical result: If $c(x,y) = |x-y|^2$ in the Euclidean space, μ is absolutely continuous with respect to Lebesgue measure, and μ, ν have finite moments of order 2, then there is a unique optimal Monge coupling between μ and ν. More general statements will be established in Chapter 10.

Optimal couplings enjoy several nice properties:

(i) They naturally arise in many problems coming from economics, physics, partial differential equations or geometry (by the way, the increasing rearrangement and the Holley coupling can be seen as particular cases of optimal transport);

(ii) They are quite stable with respect to perturbations;

(iii) They encode good geometric information, if the cost function c is defined in terms of the underlying geometry;

(iv) They exist in smooth as well as nonsmooth settings;

(v) They come with a rich structure: an **optimal cost** functional (the value of the infimum defining the Monge–Kantorovich problem); a **dual variational problem**; and, under adequate structure conditions, a continuous **interpolation**.

On the negative side, it is important to be warned that optimal transport is in general not so smooth. There are known counterexamples which put limits on the regularity that one can expect from it, even for very nice cost functions.

All these issues will be discussed again and again in the sequel. The rest of this chapter is devoted to some basic technical tools.

Gluing

If Z is a function of Y and Y is a function of X, then of course Z is a function of X. Something of this still remains true in the setting of nondeterministic couplings, under quite general assumptions.

Gluing lemma. *Let $(\mathcal{X}_i, \mu_i)$, $i = 1,2,3$, be Polish probability spaces. If (X_1, X_2) is a coupling of (μ_1, μ_2) and (Y_2, Y_3) is a coupling of (μ_2, μ_3),* *then one can construct a triple of random variables* (Z_1, Z_2, Z_3) *such that* (Z_1, Z_2) *has the same law as* (X_1, X_2) *and* (Z_2, Z_3) *has the same law as* (Y_2, Y_3)*.*

It is simple to understand why this is called "gluing lemma": if π_{12} stands for the law of (X_1, X_2) on $\mathcal{X}_1 \times \mathcal{X}_2$ and π_{23} stands for the law of (X_2, X_3) on $\mathcal{X}_2 \times \mathcal{X}_3$, then to construct the joint law π_{123} of (Z_1, Z_2, Z_3) one just has to *glue* π_{12} and π_{23} along their common marginal μ_2. Expressed in a slightly informal way: Disintegrate π_{12} and π_{23} as

$$\begin{aligned} \pi_{12}(dx_1\, dx_2) &= \pi_{12}(dx_1|x_2)\, \mu_2(dx_2), \\ \pi_{23}(dx_2\, dx_3) &= \pi_{23}(dx_3|x_2)\, \mu_2(dx_2), \end{aligned}$$

and then reconstruct π_{123} as

$$\pi_{123}(dx_1\, dx_2\, dx_3) = \pi_{12}(dx_1|x_2)\, \mu_2(dx_2)\, \pi_{23}(dx_3|x_2).$$

Change of variables formula

When one writes the formula for change of variables, say in $\mathbb{R}^n$ or on a Riemannian manifold, a Jacobian term appears, and one has to be careful about two things: the change of variables should be *injective* (otherwise, reduce to a subset where it is injective, or take the multiplicity into account); and it should be somewhat smooth. It is classical to write these formulas when the change of variables is continuously differentiable, or at least Lipschitz:

Change of variables formula. *Let* M *be an* n*-dimensional Riemannian manifold with a* C^1 *metric, let* μ_0*,* μ_1 *be two probability measures on* M*, and let* $T : M \to M$ *be a measurable function such that* $T_\#\mu_0 = \mu_1$*. Let* ν *be a reference measure, of the form* $\nu(dx) = e^{-V(x)}\, \mathrm{vol}(dx)$*, where* V *is continuous and* vol *is the volume (or* n*-dimensional Hausdorff) measure. Further assume that*

(i) $\mu_0(dx) = \rho_0(x)\, \nu(dx)$ *and* $\mu_1(dy) = \rho_1(y)\, \nu(dy)$*;*

(ii) T *is injective;*

(iii) T *is locally Lipschitz.*

Then, μ_0*-almost surely,*

$$\rho_0(x) = \rho_1(T(x))\, \mathcal{J}_T(x), \tag{1.3}$$

where $\mathcal{J}_T(x)$ is the Jacobian determinant of T at x, defined by

$$\mathcal{J}_T(x) := \lim_{\varepsilon \downarrow 0} \frac{\nu[T(B_\varepsilon(x))]}{\nu[B_\varepsilon(x)]}. \tag{1.4}$$

The same holds true if T is only defined on the complement of a μ_0-negligible set, and satisfies properties (ii) and (iii) on its domain of definition.

Remark 1.3. When ν is just the volume measure, $\mathcal{J}_T$ coincides with the usual Jacobian determinant, which in the case $M = \mathbb{R}^n$ is the absolute value of the determinant of the Jacobian matrix ∇T. Since V is continuous, it is almost immediate to deduce the statement with an arbitrary V from the statement with $V = 0$ (this amounts to multiplying $\rho_0(x)$ by $e^{V(x)}$, $\rho_1(y)$ by $e^{V(y)}$, $\mathcal{J}_T(x)$ by $e^{V(x)-V(T(x))}$).

Remark 1.4. There is a more general framework beyond differentiability, namely the property of **approximate differentiability**. A function T on an n-dimensional Riemannian manifold is said to be approximately differentiable at x if there exists a function $\widetilde{T}$, differentiable at x, such that the set $\{\widetilde{T} \neq T\}$ has zero density at x, i.e.

$$\lim_{r \to 0} \frac{\operatorname{vol}\left[\left\{x \in B_r(x);\ T(x) \neq \widetilde{T}(x)\right\}\right]}{\operatorname{vol}\left[B_r(x)\right]} = 0.$$

It turns out that, roughly speaking, an approximately differentiable map can be replaced, up to neglecting a small set, by a Lipschitz map (this is a kind of differentiable version of Lusin's theorem). So one can prove the Jacobian formula for an approximately differentiable map by approximating it with a sequence of Lipschitz maps.

Approximate differentiability is obviously a local property; it holds true if the distributional derivative of T is a locally integrable function, or even a locally finite measure. So it is useful to know that the change of variables formula still holds true if Assumption (iii) above is replaced by

(iii') T is approximately differentiable.

Conservation of mass formula

The single most important theorem of change of variables arising in continuum physics might be the one resulting from the **conservation of mass** formula,

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho\,\xi) = 0. \tag{1.5}$$

Here $\rho = \rho(t,x)$ stands for the density of a system of particles at time t and position x; $\xi = \xi(t,x)$ for the velocity field at time t and position x; and $\nabla\cdot$ stands for the divergence operator. Once again, the natural setting for this equation is a Riemannian manifold M.

It will be useful to work with particle densities $\mu_t(dx)$ (that are not necessarily absolutely continuous) and rewrite (1.5) as

$$\frac{\partial \mu}{\partial t} + \nabla \cdot (\mu\,\xi) = 0,$$

where the time-derivative is taken in the weak sense, and the divergence operator is defined by duality against continuously differentiable functions with compact support:

$$\int_M \varphi\,\nabla \cdot (\mu\,\xi) = -\int_M (\xi \cdot \nabla\varphi)\,d\mu.$$

The formula of conservation of mass is an **Eulerian** description of the physical world, which means that the unknowns are fields. The next theorem links it with the **Lagrangian** description, in which everything is expressed in terms of particle trajectories, that are integral curves of the velocity field:

$$\xi\big(t, T_t(x)\big) = \frac{d}{dt}\,T_t(x). \tag{1.6}$$

If ξ is (locally) Lipschitz continuous, then the Cauchy–Lipschitz theorem guarantees the existence of a flow T_t locally defined on a maximal time interval, and itself locally Lipschitz in both arguments t and x. Then, for each t the map T_t is a local diffeomorphism onto its image. But the formula of conservation of mass also holds true without any regularity assumption on ξ; one should only keep in mind that if ξ is not Lipschitz, then a solution of (1.6) is not uniquely determined by its value at time 0, so $x \longmapsto T_t(x)$ is not necessarily uniquely defined. Still it makes sense to consider *random* solutions of (1.6).

Mass conservation formula. *Let M be a C^1 manifold, $T \in (0,+\infty]$ and let $\xi(t,x)$ be a (measurable) velocity field on $[0,T) \times M$. Let*