Mathematical Pictures at a Data Science Exhibition

In the past few decades, heuristic methods adopted by big tech companies have complemented existing scientific disciplines to form the new field of Data Science. This text provides deep and comprehensive coverage of the mathematical theory supporting the field. Composed of 27 lecture-length chapters with exercises, it embarks the readers on an engaging itinerary through key subjects in data science, including machine learning, optimal recovery, compressive sensing (also known as compressed sensing), optimization, and neural networks. While standard material is covered, the book also includes distinctive presentations of topics such as reproducing kernel Hilbert spaces, spectral clustering, optimal recovery, compressive sensing, group testing, and applications of semidefinite programming. Students and data scientists with less mathematical background will appreciate the appendices that supply more details on some of the abstract concepts.

SIMON FOUCART is Professor of Mathematics at Texas A&M University, where he was named Presidential Impact Fellow in 2019. He has previously written, together with Holger Rauhut, the influential book *A Mathematical Introduction to Compressive Sensing* (2013).

Mathematical Pictures at a Data Science Exhibition

SIMON FOUCART
Texas A&M University

CAMBRIDGE UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre,
New Delhi – 110025, India

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781316518885
DOI: 10.1017/9781009003933

First published 2022

A catalogue record for this publication is available from the British Library.

ISBN 978-1-316-51888-5 Hardback
ISBN 978-1-009-00185-4 Paperback

Pour Jeanne, à nouveau,

and now also for Émile and Léopold

Cambridge University Press
978-1-316-51888-5 — Mathematical Pictures at a Data Science Exhibition
Simon Foucart
Frontmatter
More Information

Contents

Preface

Traditional scientific disciplines have lately been complemented by heuristics adopted in big tech companies to form the new field of Data Science. Now making its way into university curricula, this loosely defined field immediately brings computer science and statistics to mind. But mathematics, too, plays a central role by laying foundations and developing new theories. This book focuses on the subfield of Mathematical Data Science. Since its content is also loosely defined, a selection of topics was made to provide summaries only of Machine Learning, Optimal Recovery, Compressive Sensing (also known as Compressed Sensing), Optimization, and Neural Networks.

Audience: This book is intended for mathematicians who wish to know bits and pieces about Data Science. Ideally, it will convince them that there is some elegant theory behind this trendy field. Although the book may also be valuable for genuine data scientists in search of mathematical sophistication, it should primarily serve as a resource for a graduate course on Data Science given in a department of mathematics. In brief, the most important word of the title is the first one.

Theme: The common thread throughout this book is the processing of data given in the form

$$y_i = f(x^{(i)}), \qquad i \in [1:m],$$

toward the ultimate goal of learning/approximating/recovering[1] the unknown function f. In PART ONE (Machine Learning), one mostly thinks of the $x^{(i)}$ as independent realizations of a random variable and one targets results valid in expectation or with high probability. In PART TWO (Optimal Recovery), one thinks instead of the $x^{(i)}$ as fixed entities and one targets results valid

[1] The favored terminology seems to depend on one's inclination and training.

with certainty in a worst-case setting, given some side information about f. In PART THREE (Compressive Sensing), this side information conveys e.g. that f is a linear function depending on few variables. In this framework, it is actually possible to recover f exactly. A shared concern in these first three parts is complexity (sample or information complexity), i.e., the minimal number m of data that makes the learning/approximation/recovery task possible. The task almost invariably requires solving a minimization program, so PART FOUR (Optimization) reviews the necessary material. Finally, PART FIVE (Neural Networks) studies tools for the approximation of f that have recently proved very effective in Deep Learning.

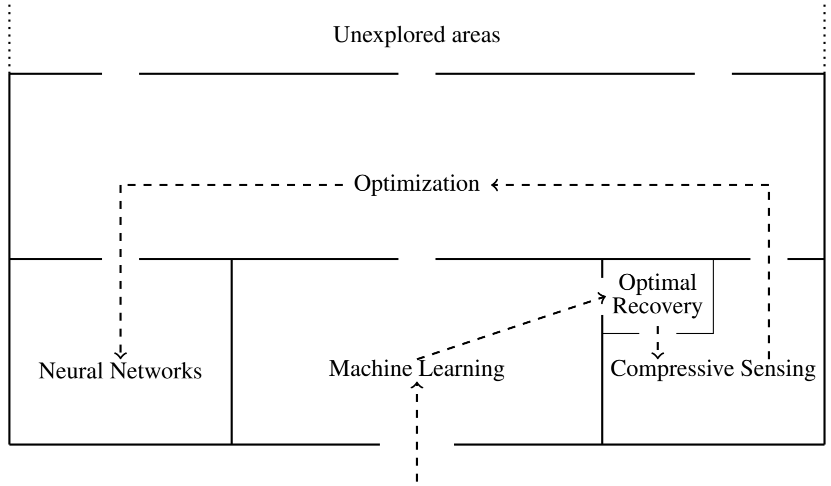

Figure 0.1 Map of the exhibition.

Content: The field of Data Science is too vast to be covered in a single book. As a matter of fact, each of the five parts picked out here is by itself worth a whole book, if not more. The executive summaries for each of the five parts suggest further readings that go into further detail. The selection of topics was merely dictated by my personal taste and interests. The route through the selected topics is metaphorically similar to an exhibition's itinerary; see Figure 0.1. Indeed, we enter through a hall (of Machine Learning) where the brightest lights lead us; continuing our path, we stumble upon a vestibule (of Optimal Recovery) filled with charming but neglected works, before arriving at a chamber (of Compressive Sensing) that we are particularly fond of; pausing for a while, we realize that the previous rooms were all connected to the large court (of Optimization) that we visit next; finally, as time runs out, we decide to come back later to see some unexplored areas, but we cannot avoid a stop at the most fashionable parlor (of Neural Networks).

Unexplored Areas: Many important topics are left out of this biased overview of mathematical Data Science. They include data assimilation (Law et al., 2015), data streams (Muthukrishnan, 2005), uncertainty quantification (Smith, 2013), reinforcement learning (Sutton and Barto, 2018), and topological data analysis (Dey and Wang, 2022), among others.

Novelty: Some of the topics covered here can already be found in book form in other places. This is particularly true for PART ONE, whose novelty lies mostly in the presentation. Other topics are unlikely to appear elsewhere. For instance, PART TWO is rather uncommon—the content of Chapter 10 there is found only in research articles. PART THREE is original as its presentation relies fully on a modified version of the standard restricted isometry property. This property plays the central role in the exposition of One-Bit Compressive Sensing offered in Chapter 17, which follows the survey article (Foucart, 2017). Most of PART FOUR is rather standard, except Chapter 22 where semidefinite programming techniques are applied to Optimal Recovery. The ingredients of PART FIVE are currently scattered around the literature. Finally, a sizable appendix is included in order to make the text almost self-contained, so that outside references are not required in the main text (with the exception of a few footnotes). It can serve as a toolkit for mathematical scientists who lack a formal training in high-dimensional geometry, probability theory, functional analysis, matrix analysis, or approximation theory. The results recalled in the appendix are of course not new, but some proofs may be innovative (e.g. the von Neumann trace inequality, the Birkhoff theorem). Some other results may not be very familiar (e.g. the Korovkin theorem, the Kolmogorov theorem).

Computational illustrations: Arguably, the field of Data Science would be inconsequential without computations. Although this book focuses on theory, most of its chapters are accompanied by unpretentious implementations, both in MATLAB and in Python. They can be found at

github.com/foucart/Mathematical_Pictures_at_a_Data_Science_Exhibition

Acknowledgment: This book originated from the lecture notes I wrote for a graduate course entitled Topics in Mathematical Data Science and delivered at Texas A&M University in Fall 2019 and in Fall 2020. Its completion was eased by a course development grant from the Texas A&M Institute of Data Science. The first bricks were actually laid while I was visiting the Institute for Foundations of Data Science at the University of Wisconsin–Madison during a sabbatical semester in Spring 2019. I am indebted to both these institutes for their support. I am also grateful to be associated with various grants from the NSF (DMS-1622134, DMS-1664803, CCF-1934904, DMS-2053172) and from the ONR (N00014-20-1-2787). Finally, I wish to thank a few colleagues for their feedback during the book's development, namely Radu Balan, Albert Cohen, Rémi Gribonval, Mark Iwen, Philipp Petersen, Sebastien Roch, Jan Vybíral, and Stephan Wojtowytsch.

Notation

Commonly Used Notation

$[i:j]$	the set $\{i, i+1, \ldots, j\}$ of integers from i to j
$\mathbb{N}$	the set $\{0, 1, 2, \ldots\}$ of natural numbers, including 0
$\mathbb{N}^*$	the set $\{1, 2, 3, \ldots\}$ of natural numbers, excluding 0
$\mathbb{Z}$	the set of integers
$\mathbb{Q}$	the set of rational numbers
$\mathbb{R}$	the set of real numbers
$\mathbb{C}$	the set of complex numbers
i	the imaginary unit $\sqrt{-1}$
$\mathcal{A}, \mathcal{S}, \mathcal{X}$	generic sets
$\mathcal{S}^c$	the complement of $\mathcal{S}$ (relative to $\mathcal{X} \supseteq \mathcal{S}$, i.e., $\mathcal{X} \setminus \mathcal{S}$)
$\mathcal{S} \Delta \mathcal{S}'$	the symmetric difference $(\mathcal{S} \cup \mathcal{S}') \setminus (\mathcal{S} \cap \mathcal{S}')$ of $\mathcal{S}$ and $\mathcal{S}'$
$\mathbb{1}_{\mathrm{event}}$	the number equal to 1 if event is true and to 0 otherwise
$\mathbb{1}_{\mathcal{S}}$	the indicator function of a set $\mathcal{S}$ (so that $\mathbb{1}_{\mathcal{S}}(x) = \mathbb{1}_{\{x \in \mathcal{S}\}}$); it can also represent a vector in $\{0, 1\}^n$ when $\mathcal{S} \subseteq [1:n]$
$\lvert\mathcal{S}\rvert$	the cardinality of a finite set $\mathcal{S}$
F, X	generic vector spaces
$\mathrm{span}(\mathcal{S})$	the linear subspace spanned by a set $\mathcal{S} \subseteq F$
$\mathrm{conv}(\mathcal{S})$	the convex hull of a set $\mathcal{S} \subseteq F$
$\mathrm{Ex}(\mathcal{S})$	the set of extreme points of a set $\mathcal{S} \subseteq F$
$\mathrm{cl}(\mathcal{X})$	the closure of a set $\mathcal{S} \subseteq F$
$\mathrm{vol}(\mathcal{S})$	the volume of a set $\mathcal{S} \subseteq F$
H	a Hilbert space
$\langle x, x' \rangle$	the inner product between two vectors $x, x' \in H$
$\mathcal{S}^\perp$	the linear space orthogonal to the set $\mathcal{S} \subseteq H$
$P_{\mathcal{V}}$	the orthogonal projector onto the linear subspace $\mathcal{V}$
T^*	the adjoint of a linear operator T defined on H

$\|x\|_F$	the norm of a vector $x \in F$
$\mathrm{dist}_F(x, \mathcal{S})$	the distance from $x \in F$ to a subset $\mathcal{S}$ of F
$B(c, r)$	the ball centered at $c \in F$ with radius $r \geq 0$
B_F	the unit ball $B(0, 1)$ of a normed space F
F^*	the dual space of a normed space F
ℓ_p^n	the space $\mathbb{R}^n$ or $\mathbb{C}^n$ normed with $\|x\|_p = \left[\sum_{i=1}^n \|x_i\|^p \right]^{1/p}$
B_p^n	the unit ball of the space ℓ_p^n
$(e_1, \ldots, e_n)$	the canonical basis for $\mathbb{R}^n$ or $\mathbb{C}^n$
$[x_1; \ldots; x_n]$	a column vector with entries $x_1, x_2, \ldots, x_n$
$F(\mathcal{X}, \mathcal{Y})$	the space of functions from a set $\mathcal{X}$ to a set $\mathcal{Y}$
$C(\mathcal{X})$	the space of continuous functions from $\mathcal{X}$ to $\mathbb{R}$
$C^k(\mathcal{X})$	the space of k-times continuously differentiable functions
$L_p(\mathcal{X})$	the space of functions with integrable pth power
$W_p^k(\mathcal{X})$	the Sobolev space of functions with kth derivative in $L_p(\mathcal{X})$
$\mathcal{K}_{\mathrm{Lip}}$	the set of functions f with Lipschitz constant $\|f\|_{\mathrm{Lip}} \leq 1$
δ_x	the evaluation functional at a point $x \in \mathcal{X}$
μ, ν	generic Borel measures
$\mathcal{M}(\mathcal{X})$	the set of Borel measures on $\mathcal{X}$
$\mathcal{M}_+(\mathcal{X})$	the set of nonnegative Borel measures on $\mathcal{X}$
$\mathcal{P}_n$	the space of algebraic polynomials of degree $\leq n$
$\mathcal{T}_n$	the space of trigonometric polynomials of degree $\leq n$
A, B, X	generic matrices
$A_{i,j}$	the entry of a matrix A on the ith row and jth column
A^*	the adjoint of a matrix A, defined by $A^*_{i,j} = \overline{A_{j,i}}$
$A^\top$	the transpose of a matrix A, defined by $A^\top_{i,j} = A_{j,i}$
$A^{-\top}$	the matrix $(A^{-1})^\top = (A^\top)^{-1}$
$\mathrm{diag}[x_1; \ldots; x_n]$	a diagonal matrix with diagonal entries $x_1, x_2, \ldots, x_n$
$\lambda_j(A)$	the jth eigenvalue of A (in nonincreasing order)
$\sigma_j(A)$	the jth singular value of A (in nonincreasing order)
$A \succeq 0$	means that the matrix A is positive semidefinite
$\langle A, B \rangle_F$	the Frobenius inner product between $A, B \in \mathbb{R}^{m \times n}$
$\|A\|_F$	the Frobenius norm of $A \in \mathbb{R}^{m \times n}$
$\|A\|_{2 \to 2}$	the operator norm of $A \in \mathbb{R}^{m \times n}$, i.e., $\sigma_1(A)$
$\|A\|_*$	the nuclear norm $A \in \mathbb{R}^{m \times n}$, i.e., $\sum_{i=1}^m \sigma_i(A)$
$\ker(A)$	the null space of (a matrix or linear map) A
$\mathrm{ran}(A)$	the range of (a matrix or linear map) A
$u * v$	the (discrete or continuous) convolution product of u and v
$\log_2(x)$	the logarithm in base 2 of $x \in (0, +\infty)$
$\ln(x)$	the natural logarithm (in base e) of $x \in (0, +\infty)$
$\exp(x)$	the exponential of $x \in \mathbb{R}$

$\lfloor x \rfloor$	the floor of $x \in \mathbb{R}$, i.e., the integer satisfying $x - 1 < \lfloor x \rfloor \leq x$
$\lceil x \rceil$	the ceiling of $x \in \mathbb{R}$, i.e., the integer satisfying $x \leq \lceil x \rceil < x + 1$
$\mathbb{P}[\mathcal{E}]$	the probability of an event $\mathcal{E}$
$\mathbb{E}[Z]$	the expectation of a random variable Z
$\mathbb{V}[Z]$	the variance of a random variable Z
$\mathcal{N}(0, \sigma^2)$	the normal distribution with mean zero and variance σ^2
g	a standard gaussian random variable, i.e., $g \sim \mathcal{N}(0, 1)$

Machine-Learning-Specific Notation

$\mathcal{X}$	the set where the instances $x^{(i)}$ (aka datapoints) live
$\mathcal{Y}$	the set where the targets y_i (aka observations) live, often $\mathcal{Y} = \mathbb{R}$, $\mathcal{Y} = \{0, 1\}$, or $\mathcal{Y} = \{-1, +1\}$
$\mathcal{H}$	a hypothesis class, i.e., a subset of $F(\mathcal{X}, \mathcal{Y})$
Loss	a function from $\mathcal{Y} \times \mathcal{Y}$ into $[0, +\infty)$ such that $L(y, y) = 0$
$\mathrm{Risk}_f(h)$	the risk of a predictor $h \in \mathcal{H}$ given $f \in F(\mathcal{X}, \mathcal{Y})$
$\mathcal{S}$	an element of $(\mathcal{X} \times \mathcal{Y})^m$ representing a sample
$\widehat{\mathrm{Risk}}_{\mathcal{S}}(h)$	the empirical risk of $h \in \mathcal{H}$ relative to $\mathcal{S}$
$\varepsilon_{\mathrm{app}}$	the approximation error
$\varepsilon_{\mathrm{est}}$	the estimation error
$m_{\mathcal{H}}(\varepsilon, \delta)$	the sample complexity
$\mathrm{vc}(\mathcal{H})$	the Vapnik–Chervonenkis dimension of $\mathcal{H} \subseteq F(\mathcal{X}, \{0, 1\})$
K	a kernel, i.e., a symmetric function defined on $\mathcal{X} \times \mathcal{X}$

Optimal-Recovery-Specific Notation

$\mathcal{K}$	the model set
Q	a quantity of interest, typically a linear map
λ_i	the ith observation functional, often equal to $\delta_{x^{(i)}}$
Λ	the observation map
Δ	a recovery map
$\mathrm{Err}_{\mathcal{K},Q}(\Lambda, \Delta)$	the worst-case error of Δ (for Q over $\mathcal{K}$ given Λ)
$\mathrm{Err}^*_{\mathcal{K},Q}(\Lambda)$	the intrinsic error (for Q over $\mathcal{K}$ given Λ)
$\mathrm{Err}^0_{\mathcal{K},Q}(\Lambda)$	the null error (for Q over $\mathcal{K}$ given Λ)
$\mathrm{Err}^*_{\mathcal{K},Q}(m)$	the mth minimal intrinsic error (for Q over $\mathcal{K}$)
$m_{\mathcal{K},Q}(\varepsilon, d)$	the information complexity
$\mathrm{Var}_{HK}(f)$	the variation of f (in the sense of Hardy and Krause)
$\mathrm{Disc}^*(\mathfrak{X})$	the star discrepancy of a finite set $\mathfrak{X}$

Compressive-Sensing-Specific Notation

Σ_s^N	the set of s-sparse vectors in $\mathbb{R}^N$ or $\mathbb{C}^N$
$\mathrm{supp}(x)$	the support of a vector x in $\mathbb{R}^N$ or $\mathbb{C}^N$
S	an index set, i.e., a subset of $[1:N]$
x_S	the vector x whose entries outside of S are zeroed out
H_s	the hard thresholding operator with parameter s
A	an $m \times N$ observation matrix
A_S	the submatrix of A with columns indexed by S^c removed
$\mathcal{A}$	an observation map (defined on a space of matrices)
μ	the coherence of a matrix
δ	ℓ_1-restricted isometry constant of a matrix or linear map
α	lower ℓ_1-restricted isometry constant of a matrix or linear map
β	upper ℓ_1-restricted isometry constant of a matrix or linear map
Δ	a recovery map
$d^m(\mathcal{K}, F)$	the mth Gelfand width of a set $\mathcal{K}$ in a space F
$\chi(v)$	the binary vector with ith entry given by $\mathbb{1}_{\{v_i>0\}}$

Optimization-Specific Notation

$(x^t)_{t\geq 0}$	a sequence of vectors produced by some iterative algorithm
∇f	the gradient of a multivariate function f
L	the Lagrangian of a minimization program
λ, ν	the dual optimization variables, aka Lagrange mutipliers
C^*	the dual cone of a set C
$\mathrm{Toep}_\infty(u)$	the infinite symmetric Toeplitz matrix built from $(u_n)_{n\geq 0}$
$\mathrm{Toep}_{N+1}(u)$	the finite symmetric Toeplitz matrix built from $(u_n)_{n=0}^N$

Neural-Networks-Specific notation

ϕ	a generic activation function
ReLU	the rectified linear unit defined by $\mathrm{ReLU}(x) = \max\{x, 0\}$
n_ℓ	the width of the ℓth layer
$x^{[\ell]}$	the state vector at the ℓth layer
$W^{[\ell]}$	the weight matrix (in $\mathbb{R}^{n_\ell \times n_{\ell-1}}$) producing the ℓth layer
$b^{[\ell]}$	the bias vector (in $\mathbb{R}^{n_\ell}$) producing the ℓth layer
$\mathcal{N}_\phi$	the linear space of functions generated by shallow networks
$\mathcal{N}_\phi^n$	the set of functions generated by width-n shallow networks
$\mathcal{N}_\phi^{n,L}$	the set of functions generated by width-n, depth-L networks