

## Random Matrix Methods for Machine Learning

This book presents a unified theory of random matrices for applications in machine learning, offering a large-dimensional data vision that exploits concentration and universality phenomena. This enables a precise understanding, and possible improvements, of the core mechanisms at play in real-world machine learning algorithms. The book opens with a thorough introduction to the theoretical basics of random matrices, which serves as a support to a wide scope of applications ranging from support vector machines, through semi-supervised learning, unsupervised spectral clustering, and graph methods, to neural networks and deep learning. For each application, the authors discuss small- versus large-dimensional intuitions of the problem, followed by a systematic random matrix analysis of the resulting performance and possible improvements. All concepts, applications, and variations are illustrated numerically on synthetic as well as real-world data, with MATLAB and Python code provided on the accompanying website.

**Romain Couillet** is a full professor at Grenoble Alpes University, France. Prior to that, he was a full professor at CentraleSupélec, University of Paris-Saclay, France. His research topics are in random matrix theory applied to statistics, machine learning, and signal processing. He is the recipient of the 2021 IEEE/SEE (Institute of Electrical and Electronics Engineers/Society of Electricity, Electronics and Information and Communication Technologies) Glavieux Prize, the 2013 CNRS (The National Center for Scientific Research) Bronze Medal, and the 2013 IEEE ComSoc Outstanding Young Researcher Award.

**Zhenyu Liao** is an associate professor with Huazhong University of Science and Technology (HUST), China. He is the recipient of the 2021 East Lake Youth Talent Program Fellowship of HUST, the 2019 ED STIC Ph.D. Student Award, and the 2016 Supélec Foundation Ph.D. Fellowship of University of Paris-Saclay, France.

# Random Matrix Methods for Machine Learning

ROMAIN COUILLET

*Grenoble Alpes University*

ZHENYU LIAO

*Huazhong University of Science and Technology*



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](http://www.cambridge.org)

Information on this title: [www.cambridge.org/9781009123235](http://www.cambridge.org/9781009123235)

DOI: 10.1017/9781009128490

© Romain Couillet and Zhenyu Liao 2022

This publication is in copyright. Subject to statutory exception and to the provisions of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press.

First published 2022

Printed in the United Kingdom by TJ Books Limited, Padstow Cornwall

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

ISBN 978-1-009-12323-5 Hardback

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

## Contents

	<i>Preface</i>	<i>page vii</i>
<b>1</b>	<b>Introduction</b>	<b>1</b>
	1.1 Motivation: The Pitfalls of Large-Dimensional Statistics	1
	1.2 Random Matrix Theory as an Answer	13
	1.3 Outline and Online Toolbox	31
<b>2</b>	<b>Random Matrix Theory</b>	<b>35</b>
	2.1 Fundamental Objects	36
	2.2 Foundational Random Matrix Results	42
	2.3 Advanced Spectrum Considerations for Sample Covariances	80
	2.4 Preliminaries on Statistical Inference	88
	2.5 Spiked Models	102
	2.6 Information-plus-Noise, Deformed Wigner, and Other Models	115
	2.7 Beyond Vectors of Independent Entries: Concentration of Measure in RMT	130
	2.8 Concluding Remarks	146
	2.9 Exercises	147
<b>3</b>	<b>Statistical Inference in Linear Models</b>	<b>155</b>
	3.1 Detection and Estimation in Information-plus-Noise Models	156
	3.2 Covariance Matrix Distance Estimation	173
	3.3 M-Estimators of Scatter	185
	3.4 Concluding Remarks	198
	3.5 Practical Course Material	200
<b>4</b>	<b>Kernel Methods</b>	<b>207</b>
	4.1 Basic Setting	208
	4.2 Distance and Inner-Product Random Kernel Matrices	211
	4.3 Properly Scaling Kernel Model	228
	4.4 Implications to Kernel Methods	242
	4.5 Concluding Remarks	272
	4.6 Practical Course Material	273

vi	<b>Contents</b>	
<b>5</b>	<b>Large Neural Networks</b>	277
	5.1 Random Neural Networks	277
	5.2 Gradient Descent Dynamics in Learning Linear Neural Nets	293
	5.3 Recurrent Neural Nets: Echo State Networks	300
	5.4 Concluding Remarks	307
	5.5 Practical Course Material	310
<b>6</b>	<b>Large-Dimensional Convex Optimization</b>	313
	6.1 Generalized Linear Classifier	314
	6.2 Large-Dimensional Support Vector Machines	326
	6.3 Concluding Remarks	331
	6.4 Practical Course Material	333
<b>7</b>	<b>Community Detection on Graphs</b>	337
	7.1 Community Detection in Dense Graphs	337
	7.2 From Dense to Sparse Graphs: A Different Approach	354
	7.3 Concluding Remarks	360
	7.4 Practical Course Material	361
<b>8</b>	<b>Universality and Real Data</b>	364
	8.1 From Gaussian Mixtures to Concentrated Random Vectors and GAN Data	364
	8.2 Wide-Sense Universality in Large-Dimensional Machine Learning	373
	8.3 Discussions and Conclusions	376
	<i>Bibliography</i>	378
	<i>Index</i>	401

## Preface

Numerous and large-dimensional data is now a default setting in modern machine learning (ML). Standard ML algorithms, starting with kernel methods such as support vector machines and graph-based methods like the PageRank algorithm, were however initially designed out of small-dimensional intuitions and tend to misbehave, if not completely collapse, when dealing with real-world large datasets. Random matrix theory has recently developed a broad spectrum of tools to help understand this new “curse of dimensionality,” to help repair or completely recreate the suboptimal algorithms, and most importantly, to provide new intuitions to deal with modern data mining.

This book primarily aims to deliver these intuitions, by providing a digest of the recent theoretical and applied breakthroughs of random matrix theory into ML. Targeting a broad audience, spanning from undergraduate students interested in statistical learning to artificial intelligence engineers and researchers alike, the mathematical prerequisites to the book are minimal (basics of probability theory, linear algebra, and real and complex analyses are sufficient): As opposed to introductory books in the mathematical literature of random matrix theory and large-dimensional statistics, the theoretical focus here is restricted to the *essential* requirements to ML applications. These applications range from detection, statistical inference, and estimation to graph- and kernel-based supervised, semisupervised, and unsupervised classification, as well as neural networks: For these, a precise theoretical prediction of the algorithm performance (often *inaccessible* when not resorting to a random matrix analysis), large-dimensional insights, methods of improvement, along with a fundamental justification of the wide-scope applicability of the methods to real data, are provided.

Most methods, algorithms, and figures proposed in the book are coded in MATLAB and Python and made available to the readers (<https://github.com/Zhenyu-LIAO/RMT4ML>). The book also contains a series of exercises of two types: short exercises with corrections available online to familiarize the reader with the basic theoretical notions and tools in random matrix analysis, as well as long guided exercises to apply these tools to further concrete ML applications.