Preface

"*All roads lead to Rome.*"

Objective. This book reveals and studies a common and fundamental problem behind almost all modern practices of machine (artificial) intelligence. That is, *how to effectively and efficiently learn a low-dimensional distribution of data in a high-dimensional space and then transform the distribution to a compact and structured representation?* For any intelligent system, natural or man-made, such a representation can be generally regarded as a *memory* or (empirical) *knowledge* learned from data sensed from the external world. In recent years, people often informally refer to it as a "world model."

Intended Audience. This textbook aims to provide a systematic introduction to the mathematical and computational principles for learning (deep) representations of such data distributions, as a computable form of memory, for *senior undergraduate students and beginning graduate students*. The main prerequisites for this book are undergraduate linear algebra, probability/statistics, and optimization. Some familiarity with basic concepts from signal processing (sparse representation and compressed sensing in particular), information theory, and feedback control would enhance your appreciation.

Motivation. The main motivation for writing this book is that there have been tremendous developments in the past several years, by the authors and many colleagues, that aim to establish a principled and rigorous approach to understand deep neural networks and, more generally, intelligence itself. The deductive methodology advocated by this new approach is in direct contrast, and highly complementary, to the dominant methodology behind current practices of artificial intelligence, which is largely inductive and trial-and-error. The lack of understanding about such powerful AI models and systems has led to increasing hype and fears in society. We believe that a serious attempt to establish a principled approach to understand intelligence is more needed than ever. An overarching goal of this book is to provide solid theoretical and experimental evidence showing that it is now possible to study intelligence as a scientific and

mathematical subject. As we will argue that intelligence is the fundamental capability to develop new memory (or knowledge) or correct existing one. Hence, one may view this book as a first attempt to develop *a Mathematical Theory of Intelligence*, at the level of learning empirical knowledge as memory, as the subtitle of the book suggests.

At the technical level, the theoretical framework presented in this book helps reconcile a long-standing gap between the classical approach to modeling data structures that are mainly based on analytical geometric, algebraic, and probabilistic models (e.g., subspaces, Gaussians, and equations) and the "modern" approach that is based on empirically designed non-parametric data-driven models (e.g., deep networks). As it turns out, a unification of the two seemingly separate methodologies becomes possible and even natural if one realizes that they all try to learn and represent *low-dimensional* structures in the data distribution of interest. They are merely different ways to pursue, represent, and exploit the low-dimensional structures. From this perspective, even many seemingly unrelated computational techniques, developed independently in separate fields at different times, can now be better understood under a common theoretical and computational framework and probably can be studied together from now on. As we will see in this book, these techniques include but are not limited to:

1. Lossy compressive encoding-decoding developed in classic information theory and coding theory versus modern representation learning;
2. Denoising and dimensionality reduction in classical signal processing versus diffusion and denoising models for modern generative methods;
3. Bayesian inference via maximum a posteriori or conditioned sampling versus constrained optimization via continuation techniques[1].

Main Content. We believe that the unified conceptual and computational framework presented in this book will be of great value to readers who truly want to clarify mysteries and misunderstandings about deep neural networks and (artificial) intelligence. Furthermore, the framework is meant to provide readers with guiding principles for developing significantly better and truly intelligent systems in the future. More specifically, besides an informal introduction (chapter), the main technical content of the book will be organized as six closely related topics (chapters):

1. We will start with the classical and most basic models of Principal Component Analysis (PCA), Independent Component Analysis (ICA), and Dictionary Learning (DL), which assume that the low-dimensional distributions of interest have linear and independent structures. From these simple idealistic models that are well studied and understood in signal processing and compressed sensing, we will introduce the most basic and important ideas[2] for how to learn and represent low-dimensional distributions.

[1] such as the augmented Lagrangian method

[2] such as denoising and dimensionality reduction

2. To generalize these classical models and their solutions to low-dimensional distributions that may not be easily represented any simple analytic models, we introduce a universal computational principle for learning such distributions: *compression*. As we will see, data compression provides a unifying view of all seemingly different classic and modern approaches to distribution or representation learning, including entropy minimization, denoising via score-matching, lossy compression with rate distortion, and discriminative representation learning via maximizing information gain.

3. Within this unifying framework, modern Deep Neural Networks (DNNs), such as ResNet, CNN, and Transformer, can all be mathematically interpreted as (unrolled) optimization algorithms that iteratively achieve better compression and better representations by reducing coding length/rate or gaining information. Not only does this framework help explain empirically designed deep network architectures thus far, it also leads to new architecture designs that can be significantly simpler and more efficient.

4. Furthermore, to ensure that the learned representation for a data distribution is correct and consistent, the *auto-encoding* architectures that consist of both encoding and decoding become necessary. In order for a learning system to be fully automatic and continuous, we will introduce a powerful *closed-loop transcription framework* that enables an auto-encoding system to self-correct and thus self-improve via a minimax game between the encoder and decoder.

5. To connect theory to practice, we will then study how the learned data distribution and representation can be utilized as a powerful prior to conduct Bayesian inference or constrained optimization that manifests as almost all types of tasks and settings that are popular in the practice of modern artificial intelligence, including conditional estimation, completion, and generation of real-world high-dimensional data such as images and texts.

6. Last but not least, we will demonstrate step-by-step how to effectively and efficiently learn deep representations of low-dimensional data distributions with large-scale real-world datasets, including visual data, body motions, and text data, and use them in many practical applications such as image classification, image completion, image segmentation, image generation, motion estimation, and similar tasks for the text data.

Summary. To summarize, the technical content presented in this book establishes strong conceptual and technical connections between the classical analytical approach and the modern data-driven approach, between simple parametric models and deep non-parametric models, between diverse inductive practices and a unified deductive framework from first principles. We will reveal that many seemingly unrelated or even competing approaches, though developed in separate fields with different terminologies and at different times in history, yet all strive to achieve a common objective: