# Acknowledgment

**Funding Support.** This book is primarily based on research results that have been developed since 2018. Thanks to the generous startup funds from UC Berkeley (2018) and the University of Hong Kong (2023), Yi Ma was able to embark and focus on this new exciting research direction in the past eight years or so. Through these years, for research in this direction, Yi Ma and his research team at Berkeley have been supported by the following research grants:

- The multi-university *THEORINET* project for the Foundations of Deep Learning, jointly funded by the Simons Foundation and the National Science Foundation (DMS grant #2031899);
- The *Closed-Loop Data Transcription via Minimaxing Rate Reduction* project funded by the Office of Naval Research (grant N00014-22-1-2102);
- The *Principled Approaches to Deep Learning for Low-dimensional Structures* project funded by the National Science Foundation (CISE grant #2402951).

This book would have not been possible without the financial support for these research projects. The authors have drawn tremendous inspiration from research results by colleagues and students who have been involved in these projects.

# Notation

In this book we employ notation from a variety of mathematical disciplines, summarized below. The reader is encouraged to consult this list as a reference as they read the various chapters of the book. We have tried to keep their usage as coherent as possible across chapters.

As a first example, for a natural number $n$ the set $\{1, 2, \ldots, n\}$ is denoted $[n]$.

## Linear Algebra Notation

- Scalars (either deterministic or random) are non-bold, e.g., $x$, $X$.
- Vectors (either deterministic or random) are lower-case bold, e.g., $\boldsymbol{x}$, $\boldsymbol{\pi}$. Entries of vectors are denoted $x_i$ or $\pi_i$, alternatively $(\boldsymbol{x})_i$. Generic terms (which can be scalars, vectors, matrices, or higher-order tensors) are also lower-case bold.
- Matrices (either deterministic or random) are capital bold, e.g., $\boldsymbol{X}, \boldsymbol{\Pi}$. Entries of matrices are denoted $X_{ij}$ or $\Pi_{ij}$, alternatively $(\boldsymbol{X})_{ij}$. Columns of matrices are lower-case bold vectors, e.g., $\boldsymbol{x}_i$ is the $i^{\text{th}}$ column of $\boldsymbol{X}$, unless otherwise defined.
- The transpose of a matrix is $\top$, e.g., $\boldsymbol{A}^\top$. The adjoint is $^*$, e.g., $\boldsymbol{A}^*$. The pseudo-inverse is $\dagger$, e.g., $\boldsymbol{A}^\dagger$.
- The rank of a matrix is $\operatorname{rank}(\boldsymbol{A})$, the trace is $\operatorname{tr}(\boldsymbol{A})$, the determinant is $\det(\boldsymbol{A})$, and the log-determinant is $\operatorname{logdet}(\boldsymbol{A})$.
- The element-wise multiplication of matrices is $\boldsymbol{A} \odot \boldsymbol{B}$. Element-wise squaring is $\boldsymbol{A}^{\odot 2}$, etc. Similarly, the Kronecker product of matrices is $\boldsymbol{A} \otimes \boldsymbol{B}$. The iterated Kronecker product is $\boldsymbol{A}^{\otimes 2}$, etc.
- For a function $f\colon \mathbb{R} \to \mathbb{R}$, its element-wise application to a matrix $\boldsymbol{X}$ is $f[\boldsymbol{X}]$, i.e., with square brackets. For a function $f\colon \mathbb{R}^d \to \mathbb{R}^k$ and a matrix $\boldsymbol{X} \in \mathbb{R}^{d \times n}$, we may broadcast $f$ to apply to each column without special notation, i.e., $f(\boldsymbol{X}) = [f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n)] \in \mathbb{R}^{k \times n}$.

- The (Euclidean) unit sphere is $\mathbb{S}^{n-1} \subseteq \mathbb{R}^n$. The (Euclidean) unit ball is $\mathbb{B}^n \subseteq \mathbb{R}^n$.
- The set of orthogonal matrices are $\mathsf{O}(m,n) \subseteq \mathbb{R}^{m\times n}$ or $\mathsf{O}(n) = \mathsf{O}(n,n)$.
- Symmetric matrices are $\mathsf{Sym}(n)$, symmetric PSD matrices are $\mathsf{PSD}(n)$, symmetric PD matrices are $\mathsf{PD}(n)$.
- The eigenvalues of symmetric matrix $\boldsymbol{S} \in \mathsf{Sym}(n)$ are all real by the spectral theorem; they are denoted $\lambda_i(\boldsymbol{S})$ and ordered such that $\lambda_1(\boldsymbol{S}) \geq \cdots \geq \lambda_n(\boldsymbol{S})$.
- The singular values of $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ with $\operatorname{rank}(\boldsymbol{A}) = r$ are denoted $\sigma_i(\boldsymbol{A})$ and ordered such that $\sigma_1(\boldsymbol{A}) \geq \cdots \geq \sigma_r(\boldsymbol{A}) > 0$. By convention we set $\sigma_{r+1}(\boldsymbol{A}) = \sigma_{r+2}(\boldsymbol{A}) = \cdots = 0$.
- The projection onto a set $\mathcal{K}$ is denoted $\operatorname{proj}_{\mathcal{K}}$.
- The $\ell^p$ norms of vectors are denoted by $\|\boldsymbol{x}\|_p$, $p \in [0,\infty]$.
- The Euclidean operator norm on matrices is $\|\boldsymbol{A}\| = \sigma_1(\boldsymbol{A})$.
- The Frobenius norm on matrices is $\|\boldsymbol{A}\|_F = \operatorname{tr}(\boldsymbol{A}^\top \boldsymbol{A})$.
- The Euclidean inner product of vectors is $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{y}^\top \boldsymbol{x}$.
- The Frobenius inner product on matrices is $\langle \boldsymbol{X}, \boldsymbol{Y} \rangle = \operatorname{tr}(\boldsymbol{Y}^\top \boldsymbol{X})$.
- The all-ones vector/matrix is denoted $\mathbf{1}$ (the shape should be obvious from context). Similarly, the all-zeros vector/matrix is denoted $\mathbf{0}$. The identity matrix is denoted $\boldsymbol{I}$.

## Probability Notation

- The base probability measure is $\mathbb{P}$, i.e., the probability of an event $A$ occurring is $\mathbb{P}[A]$. We may specify the distribution of a random variable using a subscript, i.e., $\mathbb{P}_{\boldsymbol{x}\sim\mu}[\boldsymbol{x} \in S]$ describes the probabilities associated to a random variable $\boldsymbol{x}$ with distribution (measure) $\mu$. If two random variables $\boldsymbol{x}$ and $\boldsymbol{x}'$ have the same distribution, we write $\boldsymbol{x} \overset{\mathrm{d}}{=} \boldsymbol{x}'$.
- The expected value operator is $\mathbb{E}$, with the same subscript caveat.
- The covariance operator is Cov, with the same subscript caveat.
- The correlation operator is Corr, with the same subscript caveat.
- In certain probabilistic modeling contexts where a density is assumed (especially in Chapters 3 and 7), we also use the notation $p_{\boldsymbol{x}}$ for the density of a random variable $\boldsymbol{x}$. This extends to conditional densities, e.g. $p_{\boldsymbol{x}|\boldsymbol{y}}$ for the density of $\boldsymbol{x}$ conditioned on $\boldsymbol{y}$.

- We generally use Greek letters to denote realizations of random variables (in contrast to the random variables themselves). For example, $p_{\boldsymbol{x}}(\boldsymbol{\xi})$, $p_{\boldsymbol{x}|\boldsymbol{y}}(\boldsymbol{\xi} \mid \boldsymbol{\nu})$, $\mathbb{E}[\boldsymbol{x} \mid \boldsymbol{y} = \boldsymbol{\nu}]$, etc.
- The set of probability distributions on a finite set $\mathcal{X}$ is denoted $\Delta(\mathcal{X})$. (For $\mathcal{X} = [n]$, this is the probability simplex which can be embedded into $\mathbb{R}^n$).
- The Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ is denoted $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
- The uniform distribution over a compact set $\mathcal{X}$ is denoted $\mathcal{U}(\mathcal{X})$.

## Machine Learning Notation

- Encoders and denoisers are usually denoted by $f$ and usually have parameters $\theta \in \Theta$. We usually write the features $\boldsymbol{z} = f_\theta(\boldsymbol{x})$.
- Decoders are usually denoted by $g$ and usually have parameters $\eta$. We usually write the auto-encoding $\hat{\boldsymbol{x}} = g_\eta(\boldsymbol{z})$ corresponding to $\boldsymbol{x}$.
- When $f$ and $g$ are implemented as neural networks, they will have the same number of layers without loss of generality; we write them as $f = f^L \circ f^{L-1} \circ \cdots \circ f^2 \circ f^1 \circ f^{\text{pre}}$ and $g = g^{\text{post}} \circ g^L \circ g^{L-1} \circ \cdots \circ g^2 \circ g^1$.
- We write the features at the input to layer $\ell$ as $\boldsymbol{z}^\ell$ such that $\boldsymbol{z}^{\ell+1} = f^\ell(\boldsymbol{z}^\ell)$ with $\boldsymbol{z}^1 = f^{\text{pre}}(\boldsymbol{x})$ and $\boldsymbol{z} = \boldsymbol{z}^{L+1}$. Similarly the autoencoding features are $\hat{\boldsymbol{x}}^\ell$ with $\hat{\boldsymbol{x}}^{\ell+1} = g^\ell(\hat{\boldsymbol{x}}^\ell)$ with $\hat{\boldsymbol{x}}^1 = \boldsymbol{z}$ and $\hat{\boldsymbol{x}} = g^{\text{post}}(\hat{\boldsymbol{x}}^{L+1})$.
- For denoising, optimization, and processes which have a continuous time index, the time is almost always a subscript, e.g., $\boldsymbol{x}_t$. Discrete sequences can have the index be a superscript or subscript.

## Modeling and Optimization Notation

- As in the Probability and Machine Learning Notation sections above, when we have a model, say for the realizations of a data distribution supported on $\mathbb{R}^D$, we always denote it with "plain" accented variables, say $\boldsymbol{x}$.
- In cases where we assume our data $\boldsymbol{x}$ is generated by a model from a restricted class of *parametric models*, say with a parameter $\boldsymbol{\mu}$, we also denote the *true parameter* with plain accenting.[3]

[3] This could arise either due to us making an assumption about our data, or due to our data being synthetically generated. This convention allows us to adopt a concise notation across both 'probabilistic' and 'analytical' modeling setups. Readers more familiar with the latter kinds of frameworks may be used to using notation such as $\boldsymbol{\mu}_\sharp$ or $\boldsymbol{\mu}_o$ for this case; contrast with our accented decision variable notation.

- *Decision variables* (or learnable parameters) of a model fit to observed data are denoted with "tilde" accenting, e.g. $\tilde{\boldsymbol{x}}$, in cases where there is an underlying model $\boldsymbol{x}$, or with plain notation in cases where no confusion is possible (e.g., if an empirically-fit model for *nonparametric* data $\boldsymbol{x}$ involves "mean" parameters $\boldsymbol{\mu}$).
- *Optimal solutions* to optimization problems are denoted with "star" accenting, either as a superscript or a subscript (say $\boldsymbol{x}_\star$ or $\boldsymbol{x}^\star$).
- *Estimators* for data that correspond to a specific statistical model, especially for the mean of that statistical model, are denoted with "bar" accenting, say $\bar{\boldsymbol{x}}$ (especially for minimum mean-squared error denoising in Chapter 3).
- *Approximations* associated to computational procedures for modeling data are denoted with "hat" accenting, say $\hat{\boldsymbol{x}}$ for the output of an autoencoder or a generative model that approximates data $\boldsymbol{x}$.[4]

[4] We generally distinguish these cases from optimal solutions $\boldsymbol{x}_\star$, although in some special cases they are equal (e.g., in several examples in Chapter 2). Think of the distinction between the parameter $\boldsymbol{\mu}$ of a model, which might be fit with optimization, and the results of sampling with that learned parameter.

# Contents