

# Mathematics for Machine Learning

Marc Peter Deisenroth, A. Aldo Faisal, Cheng Soon Ong



---

# Contents

<i>List of illustrations</i>	ii
<i>Foreword</i>	1
<b>Part I Mathematical Foundations</b>	<b>9</b>
<b>1 Introduction and Motivation</b>	<b>11</b>
1.1 Finding Words for Intuitions	12
1.2 Two Ways to Read this Book	13
1.3 Exercises and Feedback	16
<b>2 Linear Algebra</b>	<b>17</b>
2.1 Systems of Linear Equations	19
2.2 Matrices	22
2.3 Solving Systems of Linear Equations	27
2.4 Vector Spaces	35
2.5 Linear Independence	40
2.6 Basis and Rank	44
2.7 Linear Mappings	48
2.8 Affine Spaces	61
2.9 Further Reading	63
Exercises	63
<b>3 Analytic Geometry</b>	<b>70</b>
3.1 Norms	71
3.2 Inner Products	72
3.3 Lengths and Distances	75
3.4 Angles and Orthogonality	76
3.5 Orthonormal Basis	78
3.6 Orthogonal Complement	79
3.7 Inner Product of Functions	80
3.8 Orthogonal Projections	81
3.9 Rotations	91
3.10 Further Reading	94
Exercises	95
<b>4 Matrix Decompositions</b>	<b>98</b>

4.1	Determinant and Trace	99
4.2	Eigenvalues and Eigenvectors	105
4.3	Cholesky Decomposition	114
4.4	Eigendecomposition and Diagonalization	115
4.5	Singular Value Decomposition	119
4.6	Matrix Approximation	129
4.7	Matrix Phylogeny	134
4.8	Further Reading	135
	Exercises	137
<b>5</b>	<b>Vector Calculus</b>	<b>139</b>
5.1	Differentiation of Univariate Functions	141
5.2	Partial Differentiation and Gradients	146
5.3	Gradients of Vector-Valued Functions	149
5.4	Gradients of Matrices	155
5.5	Useful Identities for Computing Gradients	158
5.6	Backpropagation and Automatic Differentiation	159
5.7	Higher-order Derivatives	164
5.8	Linearization and Multivariate Taylor Series	165
5.9	Further Reading	170
	Exercises	170
<b>6</b>	<b>Probability and Distributions</b>	<b>172</b>
6.1	Construction of a Probability Space	172
6.2	Discrete and Continuous Probabilities	178
6.3	Sum Rule, Product Rule and Bayes' Theorem	183
6.4	Summary Statistics and Independence	186
6.5	Gaussian Distribution	197
6.6	Conjugacy and the Exponential Family	204
6.7	Change of Variables/Inverse Transform	214
6.8	Further Reading	220
	Exercises	221
<b>7</b>	<b>Continuous Optimization</b>	<b>225</b>
7.1	Optimization using Gradient Descent	227
7.2	Constrained Optimization and Lagrange Multipliers	233
7.3	Convex Optimization	236
7.4	Further Reading	246
	Exercises	247
	<b>Part II Central Machine Learning Problems</b>	<b>249</b>
<b>8</b>	<b>When Models meet Data</b>	<b>251</b>
8.1	Empirical Risk Minimization	258
8.2	Parameter Estimation	265
8.3	Probabilistic Modeling and Inference	272
8.4	Directed Graphical Models	277

<i>Contents</i>	iii
8.5 Model Selection	283
<b>9 Linear Regression</b>	289
9.1 Problem Formulation	291
9.2 Parameter Estimation	292
9.3 Bayesian Linear Regression	303
9.4 Maximum Likelihood as Orthogonal Projection	313
9.5 Further Reading	315
<b>10 Dimensionality Reduction with Principal Component Analysis</b>	317
10.1 Problem Setting	318
10.2 Maximum Variance Perspective	320
10.3 Projection Perspective	325
10.4 Eigenvector Computation and Low-Rank Approximations	333
10.5 PCA in High Dimensions	335
10.6 Key Steps of PCA in Practice	336
10.7 Latent Variable Perspective	339
10.8 Further Reading	343
<b>11 Density Estimation with Gaussian Mixture Models</b>	348
11.1 Gaussian Mixture Model	349
11.2 Parameter Learning via Maximum Likelihood	350
11.3 EM Algorithm	360
11.4 Latent Variable Perspective	363
11.5 Further Reading	368
<b>12 Classification with Support Vector Machines</b>	370
12.1 Separating Hyperplanes	372
12.2 Primal Support Vector Machine	374
12.3 Dual Support Vector Machine	383
12.4 Kernels	388
12.5 Numerical Solution	390
12.6 Further Reading	392
<i>References</i>	395
<i>Index</i>	407

---

## List of Figures

1.1	The foundations and four pillars of machine learning.	14
2.1	Different types of vectors.	17
2.2	Linear algebra mind map.	19
2.3	Geometric interpretation of systems of linear equations.	21
2.4	A matrix can be represented as a long vector.	22
2.5	Matrix multiplication.	23
2.6	Examples of subspaces.	39
2.7	Geographic example of linearly dependent vectors.	41
2.8	Two different coordinate systems.	50
2.9	Different coordinate representations of a vector.	51
2.10	Three examples of linear transformations.	52
2.11	Basis change.	56
2.12	Kernel and image of a linear mapping $\Phi : V \rightarrow W$ .	59
2.13	Lines are affine subspaces.	62
3.1	Analytic geometry mind map.	70
3.2	Illustration of different norms.	71
3.3	Triangle inequality.	71
3.4	$f(x) = \cos(x)$ .	76
3.5	Angle between two vectors.	77
3.6	Angle between two vectors.	77
3.7	A plane can be described by its normal vector.	80
3.8	$f(x) = \sin(x) \cos(x)$ .	81
3.9	Orthogonal projection.	82
3.10	Examples of projections onto one-dimensional subspaces.	83
3.11	Projection onto a two-dimensional subspace.	85
3.12	Gram-Schmidt orthogonalization.	89
3.13	Projection onto an affine space.	90
3.14	Rotation.	91
3.15	Robotic arm.	91
3.16	Rotation of the standard basis in $\mathbb{R}^2$ by an angle $\theta$ .	92
3.17	Rotation in three dimensions.	93
4.1	Matrix decomposition mind map.	99
4.2	The area of a parallelogram computed using the determinant.	101
4.3	The volume of a parallelepiped computed using the determinant.	101
4.4	Determinants and eigenspaces.	109
4.5	C. elegans neural network.	110
4.6	Geometric interpretation of eigenvalues.	113
4.7	Eigendecomposition as sequential transformations.	117

4.8	Intuition behind SVD as sequential transformations.	120
4.9	SVD and mapping of vectors.	122
4.10	SVD decomposition for movie ratings.	127
4.11	Image processing with the SVD.	130
4.12	Image reconstruction with the SVD.	131
4.13	Phylogeny of matrices in machine learning.	134
5.1	Different problems for which we need vector calculus.	139
5.2	Vector calculus mindmap.	140
5.3	Difference quotient.	141
5.4	Taylor polynomials.	144
5.5	Jacobian determinant.	151
5.6	Dimensionality of partial derivatives.	152
5.7	Gradient computation of a matrix with respect to a vector.	155
5.8	Forward pass in a multi-layer neural network.	160
5.9	Backward pass in a multi-layer neural network.	161
5.10	Data flow graph.	161
5.11	Computation graph.	162
5.12	Linear approximation of a function.	165
5.13	Visualizing outer products.	166
6.1	Probability mind map.	173
6.2	Visualization of a discrete bivariate probability mass function.	179
6.3	Examples of discrete and continuous uniform distributions.	182
6.4	Mean, Mode and Median.	189
6.5	Identical means and variances but different covariances.	191
6.6	Geometry of random variables.	196
6.7	Gaussian distribution of two random variables $x, y$ .	197
6.8	Gaussian distributions overlaid with 100 samples.	198
6.9	Bivariate Gaussian with conditional and marginal.	200
6.10	Examples of the Binomial distribution.	206
6.11	Examples of the Beta distribution for different values of $\alpha$ and $\beta$ .	207
7.1	Optimization mind map.	226
7.2	Example objective function.	227
7.3	Gradient descent on a two-dimensional quadratic surface.	229
7.4	Illustration of constrained optimization.	233
7.5	Example of a convex function.	236
7.6	Example of a convex set.	236
7.7	Example of a nonconvex set.	237
7.8	The negative entropy and its tangent.	238
7.9	Illustration of a linear program.	240
8.1	Toy data for linear regression	254
8.2	Example function and its prediction	255
8.3	Example function and its uncertainty.	256
8.4	$K$ -fold cross validation.	263
8.5	Maximum likelihood estimate.	268
8.6	Maximum a posteriori estimation.	268
8.7	Model fitting.	270
8.8	Fitting of different model classes.	271
8.9	Examples of directed graphical models.	278

8.10	Graphical models for a repeated Bernoulli experiment.	280
8.11	D-separation example.	281
8.12	Three types of graphical models.	282
8.13	Nested cross validation.	283
8.14	Bayesian inference embodies Occam's razor.	285
8.15	Hierarchical generative process in Bayesian model selection.	286
9.1	Regression.	289
9.2	Linear regression example.	292
9.3	Probabilistic graphical model for linear regression.	292
9.4	Polynomial regression.	297
9.5	Maximum likelihood fits for different polynomial degrees $M$ .	299
9.6	Training and test error.	300
9.7	Polynomial regression: Maximum likelihood and MAP estimates.	302
9.8	Graphical model for Bayesian linear regression.	304
9.9	Prior over functions.	305
9.10	Bayesian linear regression and posterior over functions.	310
9.11	Bayesian linear regression.	311
9.12	Geometric interpretation of least squares.	313
10.1	Illustration: Dimensionality reduction.	317
10.2	Graphical illustration of PCA.	319
10.3	Examples of handwritten digits from the MNIST dataset.	320
10.4	Illustration of the maximum variance perspective.	321
10.5	Properties of the training data of MNIST '8'.	324
10.6	Illustration of the projection approach.	325
10.7	Simplified projection setting.	326
10.8	Optimal projection.	328
10.9	Orthogonal projection and displacement vectors.	330
10.10	Embedding of MNIST digits.	332
10.11	Steps of PCA.	337
10.12	Effect of the number of principal components on reconstruction.	338
10.13	Squared reconstruction error versus the number of components.	339
10.14	PPCA graphical model.	340
10.15	Generating new MNIST digits.	341
10.16	PCA as an auto-encoder.	344
11.1	Dataset that cannot be represented by a Gaussian.	348
11.2	Gaussian mixture model.	350
11.3	Initial setting: GMM with three mixture components.	350
11.4	Update of the mean parameter of mixture component in a GMM.	355
11.5	Effect of updating the mean values in a GMM.	355
11.6	Effect of updating the variances in a GMM.	358
11.7	Effect of updating the mixture weights in a GMM.	360
11.8	EM algorithm applied to the GMM from Figure 11.2.	361
11.9	Illustration of the EM algorithm.	362
11.10	GMM fit and responsibilities when EM converges.	363
11.11	Graphical model for a GMM with a single data point.	364
11.12	Graphical model for a GMM with $N$ data points.	366
11.13	Histogram and kernel density estimation.	369
12.1	Example 2D data for classification.	371



12.2	Equation of a separating hyperplane.	373
12.3	Possible separating hyperplanes	374
12.4	Vector addition to express distance to hyperplane.	375
12.5	Derivation of the margin: $r = \frac{1}{\ w\ }$ .	376
12.6	Linearly separable and non linearly separable data.	379
12.7	Soft margin SVM allows examples to be within the margin.	380
12.8	The hinge loss is a convex upper bound of zero-one loss.	382
12.9	Convex hulls.	386
12.10	SVM with different kernels.	389



---

## Foreword

Machine learning is the latest in a long line of attempts to distill human knowledge and reasoning into a form that is suitable for constructing machines and engineering automated systems. As machine learning becomes more ubiquitous and its software packages become easier to use it is natural and desirable that the low-level technical details are abstracted away and hidden from the practitioner. However, this brings with it the danger that a practitioner becomes unaware of the design decisions and, hence, the limits of machine learning algorithms.

The enthusiastic practitioner who is interested to learn more about the magic behind successful machine learning algorithms currently faces a daunting set of pre-requisite knowledge:

- Programming languages and data analysis tools
- Large-scale computation and the associated frameworks
- Mathematics and statistics and how machine learning builds on it

At universities, introductory courses on machine learning tend to spend early parts of the course covering some of these pre-requisites. For historical reasons, courses in machine learning tend to be taught in the computer science department, where students are often trained in the first two areas of knowledge, but not so much in mathematics and statistics.

Current machine learning textbooks primarily focus on machine learning algorithms and methodologies and assume that the reader is competent in mathematics and statistics. Therefore, these books only spend one or two chapters of background mathematics, either at the beginning of the book or as appendices. We have found many people who want to delve into the foundations of basic machine learning methods who struggle with the mathematical knowledge required to read a machine learning textbook. Having taught undergraduate and graduate courses at universities, we find that the gap between high-school mathematics and the mathematics level required to read a standard machine learning textbook is too big for many people.

This book brings the mathematical foundations of basic machine learning concepts to the fore and collects the information in a single place so that this skills gap is narrowed or even closed.

### *Why Another Book on Machine Learning?*

Machine learning builds upon the language of mathematics to express concepts that seem intuitively obvious but which are surprisingly difficult to formalize. Once formalized properly, we can gain insights into the task we want to solve. One common complaint of students of mathematics around the globe is that the topics covered seem to have little relevance to practical problems. We believe that machine learning is an obvious and direct motivation for people to learn mathematics.

“Math is linked in the popular mind with phobia and anxiety. You’d think we’re discussing spiders.” (Strogatz, 2014)

This book is intended to be a guidebook to the vast mathematical literature that forms the foundations of modern machine learning. We motivate the need for mathematical concepts by directly pointing out their usefulness in the context of fundamental machine learning problems. In the interest of keeping the book short, many details and more advanced concepts have been left out. Equipped with the basic concepts presented here, and how they fit into the larger context of machine learning, the reader can find numerous resources for further study, which we provide at the end of the respective chapters. For readers with a mathematical background, this book provides a brief but precisely stated glimpse of machine learning. In contrast to other books that focus on methods and models of machine learning (MacKay, 2003; Bishop, 2006; Alpaydin, 2010; Rogers and Girolami, 2016; Murphy, 2012; Barber, 2012; Shalev-Shwartz and Ben-David, 2014) or programmatic aspects of machine learning (Müller and Guido, 2016; Raschka and Mirjalili, 2017; Chollet and Allaire, 2018) we provide only four representative examples of machine learning algorithms. Instead we focus on the mathematical concepts behind the models themselves. We hope that readers will be able to gain a deeper understanding of the basic questions in machine learning and connect practical questions arising from the use of machine learning with fundamental choices in the mathematical model.

We do not aim to write a classical machine learning book. Instead, our intention is to provide the mathematical background, applied to four central machine learning problems, to make it easier to read other machine learning textbooks.

### *Who is the Target Audience?*

As applications of machine learning become widespread in society we believe that everybody should have some understanding of its underlying principles. This book is written in an academic mathematical style, which enables us to be precise about the concepts behind machine learning. We encourage readers unfamiliar with this seemingly terse style to persevere and to keep the goals of each topic in mind. We sprinkle comments and remarks throughout the text, in the hope that it provides useful guidance with respect to the big picture.

*The book assumes the reader to have mathematical knowledge commonly*

*covered in high-school mathematics and physics.* For example, the reader should have seen derivatives and integrals before, and geometric vectors in two or three dimensions. Starting from there we generalize these concepts. Therefore, the target audience of the book includes undergraduate university students, evening learners and learners participating in online machine learning courses.

In analogy to music, there are three types of interaction, which people have with machine learning:

**Astute Listener** The democratization of machine learning by the provision of open-source software, online tutorials and cloud-based tools allows users to not worry about the specifics of pipelines. Users can focus on extracting insights from data using off-the-shelf tools. This enables non-tech savvy domain experts to benefit from machine learning. This is similar to listening to music; the user is able to choose and discern between different types of machine learning, and benefits from it. More experienced users are like music critics, asking important questions about the application of machine learning in society such as ethics, fairness, and privacy of the individual. We hope that this book provides a foundation for thinking about the certification and risk management of machine learning systems, and allows them to use their domain expertise to build better machine learning systems.

**Experienced Artist** Skilled practitioners of machine learning can plug and play different tools and libraries into an analysis pipeline. The stereotypical practitioner would be a data scientist or engineer who understands machine learning interfaces and their use cases, and is able to perform wonderful feats of prediction from data. This is similar to a virtuoso playing music, where highly skilled practitioners can bring existing instruments to life, and bring enjoyment to their audience. Using the mathematics presented here as a primer, practitioners would be able to understand the benefits and limits of their favorite method, and to extend and generalize existing machine learning algorithms. We hope that this book provides the impetus for more rigorous and principled development of machine learning methods.

**Fledgling Composer** As machine learning is applied to new domains, developers of machine learning need to develop new methods and extend existing algorithms. They are often researchers who need to understand the mathematical basis of machine learning and uncover relationships between different tasks. This is similar to composers of music who, within the rules and structure of musical theory, create new and amazing pieces. We hope this book provides a high-level overview of other technical books for people who want to become composers of machine learning. There is a great need in society for new researchers who are able to propose and explore novel approaches for attacking the many challenges of learning from data.

### Contributors

We are grateful to many people, who looked at early drafts of the book and suffered through painful expositions of concepts. We tried to implement their ideas that we did not violently disagree with. We would like to especially acknowledge Christfried Webers for his careful reading of many parts of the book, and his detailed suggestions on structure and presentation. Many friends and colleagues have also been kind enough to provide their time and energy on different versions of each chapter. We have been lucky to benefit from the generosity of the online community, who have suggested improvements via `github.com`, which greatly improved the book.

The following people have found bugs, proposed clarifications and suggested relevant literature, either via `github.com` or personal communication. Their names are sorted alphabetically.

Abdul-Ganiy Usman	Ellen Broad
Adam Gaier	Fengkuangtian Zhu
Aditya Menon	Fiona Condon
Adele Jackson	Georgios Theodorou
Aleksandar Krnjaic	He Xin
Alexander Makrigiorgos	Irene Raissa Kameni
Alfredo Canziani	Jakub Nabaglo
Ali Shafti	James Hensman
Alasdair Tran	Jamie Liu
Amr Khalifa	Jean Kaddour
Andrew Tanggara	Jean-Paul Ebejer
Antal A. Buss	Jerry Qiang
Antoine Toisoul Le Cann	Jitesh Sindhare
Angus Gruen	John Lloyd
Areg Sarvazyan	Jonas Ngnawe
Artem Artemev	Jon Martin
Artyom Stepanov	Justin Hsi
Bill Kromydas	Kai Arulkumaran
Bob Williamson	Kamil Dreczkowski
Boon Ping Lim	Lily Wang
Chao Qu	Lionel Tondji Ngoupeyou
Cheng Li	Lydia Knüfing
Chris Sherlock	Mahmoud Aslan
Christopher Gray	Markus Hegland
Daniel McNamara	Matthew Alger
Daniel Wood	Matthew Lee
Darren Siegel	Mark Hartenstein
David Johnston	Mark van der Wilk
Dawei Chen	Martin Hewing

Maximus McCann	Shawn Berry
Mengyan Zhang	Sheikh Abdul Raheem Ali
Michael Bennett	Sheng Xue
Michael Pedersen	Sridhar Thiagarajan
Minjeong Shin	Syed Nouman Hasany
Naveen Kumar	Szymon Brych
Nico Montali	Thomas Bühler
Oscar Armas	Timur Sharapov
Patrick Henriksen	Tom Melamed
Patrick Wieschollek	Vincent Adam
Pattarawat Chormai	Vincent Dutordoir
Paul Kelly	Vu Minh
Petros Christodoulou	Wasim Aftab
Piotr Januszewski	Wen Zhi
Pranav Subramani	Wojciech Stokowiec
Quyu Kong	Xiaonan Chong
Ragib Zaman	Xiaowei Zhang
Rui Zhang	Yazhou Hao
Ryan-Rhys Griffiths	Yicheng Luo
Samuel Ogunmola	Young Lee
Sandeep Mavadia	Yu Lu
Sarvesh Nikumbh	Yun Cheng
Sebastian Raschka	Yuxiao Huang
Senanayak Sesh Kumar Karri	Zac Cranko
Seung-Heon Baek	Zijian Cao
Shahbaz Chaudhary	Zoe Nolan
Shakir Mohamed	

Contributors through github, whose real names were not listed on their github profile, are:

SamDataMad	insad	empet
bumptiousmonkey	HorizonP	victorBigand
idoamihai	cs-maillist	17SKYE
deepakiim	kudo23	jessjing1995

We are also very grateful to the many anonymous reviewers organized by Cambridge University Press who read one or more chapters of earlier versions of the manuscript, and provided constructive criticism that led to considerable improvements. A special mention goes to Dinesh Singh Negi, our  $\text{\LaTeX}$  support person, for detailed and prompt advice about  $\text{\LaTeX}$ -related issues. Last but not least, we are very grateful to our editor Lauren Cowles, who has been patiently guiding us through the gestation process of this book.

## Table of Symbols

Symbol	Typical meaning
$a, b, c, \alpha, \beta, \gamma$	scalars are lowercase
$\mathbf{x}, \mathbf{y}, \mathbf{z}$	vectors are bold lowercase
$\mathbf{A}, \mathbf{B}, \mathbf{C}$	matrices are bold uppercase
$\mathbf{x}^\top, \mathbf{A}^\top$	transpose of a vector or matrix
$\mathbf{A}^{-1}$	inverse of a matrix
$\langle \mathbf{x}, \mathbf{y} \rangle$	inner product of $\mathbf{x}$ and $\mathbf{y}$
$\mathbf{x}^\top \mathbf{y}$	dot product of $\mathbf{x}$ and $\mathbf{y}$
$B = (\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3)$	(ordered) tuple
$\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3]$	matrix of column vectors stacked horizontally
$\mathcal{B} = \{\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3\}$	set of vectors (unordered)
$\mathbb{Z}, \mathbb{N}$	integers and natural numbers, respectively
$\mathbb{R}, \mathbb{C}$	real and complex numbers, respectively
$\mathbb{R}^n$	$n$ -dimensional vector space of real numbers
$\forall x$	universal quantifier: For all $x$
$\exists x$	existential quantifier: There exists $x$
$a := b$	$a$ is defined as $b$
$a =: b$	$b$ is defined as $a$
$a \propto b$	$a$ is proportional to $b$ , i.e., $a = \text{constant} \cdot b$
$g \circ f$	function composition: “ $g$ after $f$ ”
$\iff$	if and only if
$\implies$	implies
$\mathcal{A}, \mathcal{C}$	sets
$a \in \mathcal{A}$	$a$ is an element of the set $\mathcal{A}$
$\emptyset$	empty set
$D$	number of dimensions; indexed by $d = 1, \dots, D$
$N$	number of data points; indexed by $n = 1, \dots, N$
$\mathbf{I}_m$	identity matrix of size $m \times m$
$\mathbf{0}_{m,n}$	matrix of zeros of size $m \times n$
$\mathbf{1}_{m,n}$	matrix of ones of size $m \times n$
$\mathbf{e}_i$	standard/canonical vector (where $i$ is the component that is 1)
$\dim$	dimensionality of vector space
$\text{rk}(\mathbf{A})$	rank of matrix $\mathbf{A}$
$\text{Im}(\Phi)$	image of linear mapping $\Phi$
$\ker(\Phi)$	kernel (null space) of a linear mapping $\Phi$
$\text{span}[\mathbf{b}_1]$	span (generating set) of $\mathbf{b}_1$
$\text{tr}(\mathbf{A})$	trace of $\mathbf{A}$
$\det(\mathbf{A})$	determinant of $\mathbf{A}$
$ \cdot $	absolute value or determinant (depending on context)
$\ \cdot\ $	norm; Euclidean unless specified
$\lambda$	eigenvalue or Lagrange multiplier
$E_\lambda$	eigenspace corresponding to eigenvalue $\lambda$



Symbol	Typical meaning
$\boldsymbol{\theta}$	parameter vector
$\frac{\partial f}{\partial x}$	partial derivative of $f$ with respect to $x$
$\frac{df}{dx}$	total derivative of $f$ with respect to $x$
$\nabla$	gradient
$\mathcal{L}$	Lagrangian
$\mathcal{L}$	negative log-likelihood
$\binom{n}{k}$	Binomial coefficient, $n$ choose $k$
$V_X[\mathbf{x}]$	variance of $\mathbf{x}$ with respect to the random variable $X$
$E_X[\mathbf{x}]$	expectation of $\mathbf{x}$ with respect to the random variable $X$
$\text{Cov}_{X,Y}[\mathbf{x}, \mathbf{y}]$	covariance between $\mathbf{x}$ and $\mathbf{y}$ .
$X \perp\!\!\!\perp Y \mid Z$	$X$ is conditionally independent of $Y$ given $Z$
$X \sim p$	random variable $X$ is distributed according to $p$
$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$
$\text{Ber}(\mu)$	Bernoulli distribution with parameter $\mu$
$\text{Bin}(N, \mu)$	Binomial distribution with parameters $N, \mu$
$\text{Beta}(\alpha, \beta)$	Beta distribution with parameters $\alpha, \beta$

### Table of Abbreviations and Acronyms

Acronym	Meaning
e.g.	exempli gratia (Latin: for example)
GMM	Gaussian mixture model
i.e.	id est (Latin: this means)
i.i.d.	independent, identically distributed
MAP	maximum a posteriori
MLE	maximum likelihood estimation/estimator
ONB	orthonormal basis
PCA	principal component analysis
PPCA	probabilistic principal component analysis
REF	row echelon form
SPD	symmetric, positive definite
SVM	support vector machine