Notation and abbreviations

General notation

This book observes the following general mathematical notation to avoid any confusion arising from notation:

$\boxed{\mathbb{B} = \{\text{true, false}\}}$

Set of boolean values

$\boxed{\mathbb{Z}^+ = \{1, 2, \cdots\}}$

Set of positive integers

$\boxed{\mathbb{R}}$

Set of real numbers

$\boxed{\mathbb{R}_{>0}}$

Set of positive real numbers

$\boxed{\mathbb{R}^D}$

Set of D dimensional real numbers

$\boxed{\Sigma^*}$

Set of all possible strings composed of letters

$\boxed{\emptyset}$

Empty set

$\boxed{a}$

Scalar variable

$\boxed{\mathbf{a}}$

Vector variable

$$\mathbf{a} = \begin{bmatrix} a_1 & \cdots & a_N \end{bmatrix}^{\mathsf{T}} = \begin{bmatrix} a_1 \\ \vdots \\ a_N \end{bmatrix}$$

Elements of a vector, which can be described with the square brackets $[\cdots]$, $^{\mathsf{T}}$ denotes the transpose operation

$$\mathbf{A}$$

Matrix variable

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

Elements of a matrix, which can be described with the square brackets $[\cdots]$

$$\mathbf{I}_D$$

$D \times D$ identity matrix

$$|\mathbf{A}|$$

Determinant of square matrix

$$\mathrm{tr}[\mathbf{A}]$$

Trace of square matrix

$$A, \mathcal{A}$$

Set or sequential variable

$$A = \{a_1, \cdots, a_N\} = \{a_n\}_{n=1}^{N}$$

Elements in a set, which can be described with the curly brackets $\{\cdots\}$

$$A = \{a_n\}$$

Elements in a set, where the range of index n is omitted for simplicity

$$a_{n:n'} = \{a_n, \cdots, a_{n'}\} \quad n' > n$$

A set of sequential variables, which explicitly describes the range of elements from n to n' by using : in the subscript

$$|A|$$

The number of elements in a set A. For example $|\{a_n\}_{n=1}^{N}| = N$

$$f(x) \text{ or } f_x$$

Function of x

$\boxed{p(x) \text{ or } q(x)}$

Probabilistic distribution function of x

$\boxed{\mathcal{F}[f]}$

Functional of f. Note that a functional uses the square brackets $[\cdot]$ while a function uses the bracket $(\cdot)$.

$\boxed{\mathbb{E}_{p(x|y)}[f(x)|y] = \int f(x)p(x|y)dx}$

The expectation of $f(x)$ with respect to probability distribution $p(x|y)$

$\boxed{\mathbb{E}_{(x)}[f(x)|y] = \int f(x)p(x|y)dx \text{ or } \mathbb{E}_{(x)}[f(x)] = \int f(x)p(x|y)dx}$

Another form of the expectation of $f(x)$, where the subscript with the probability distribution and/or the conditional variable is omitted, when it is trivial.

$$\boxed{\delta(a,a') = \begin{cases} 1 & a = a' \\ 0 & \text{Otherwise} \end{cases}}$$

Kronecker delta function for discrete variables a and a'

$\boxed{\delta(x - x')}$

Dirac delta function for continuous variables x and x'

$\boxed{A^{\mathrm{ML}}, A^{\mathrm{ML2}}, A^{\mathrm{MAP}}, A^{\mathrm{DT}}, \ldots}$

The variables estimated by a specific criterion (e.g., Maximum Likelihood (ML)) are represented with the superscript of the abbreviation of the criterion.

Basic notation used for speech and language processing

We also list the notation specific for speech and language processing. This book tries to maintain consistency by using the same notation, while it also tries to use commonly used notation in each application. Therefore, some of the same characters are used to denote different variables, since this book needs to introduce many variables.

Common notation

$\boxed{\Theta}$

Set of model parameters

$\boxed{M}$

Model variable including types of models, structure, hyperparameters, etc.

Ψ

Set of hyperparameters

$Q(\cdot|\cdot)$

Auxiliary function used in the EM algorithm

$\mathbf{H}$

Hessian matrix

Acoustic modeling

$T \in \mathbb{Z}^+$

Number of speech frames

$t \in \{1, \cdots, T\}$

Speech frame index

$\mathbf{o}_t \in \mathbb{R}^D$

D dimensional feature vector at time t

$\mathbf{O} = \{\mathbf{o}_t | t = 1, \cdots, T\}$

Sequence of T feature vectors

$J \in \mathbb{Z}^+$

Number of unique HMM states in an HMM

$s_t \in \{1, \cdots, J\}$

HMM state at time t

$S = \{s_t | t = 1, \cdots, T\}$

Sequence of HMM states for T speech frames

$K \in \mathbb{Z}^+$

Number of unique mixture components in a GMM

$v_t \in \{1, \cdots, K\}$

Latent mixture variable at time t

$\boxed{V = \{v_t | t = 1, \cdots, T\}}$

Sequence of latent mixture variables for T speech frames

$\boxed{\alpha_t(j) \in [0, 1]}$

Forward probability of the partial observations $\{\mathbf{o}_1, \cdots, \mathbf{o}_t\}$ until time t and state j at time t

$\boxed{\beta_t(j) \in [0, 1]}$

Backward probability of the partial observations $\{\mathbf{o}_{t+1}, \cdots, \mathbf{o}_T\}$ from $t + 1$ to the end given state j at time t

$\boxed{\delta_t(j) \in [0, 1]}$

The highest probability along a single path, at time t which accounts for previous observations $\{\mathbf{o}_1, \cdots, \mathbf{o}_t\}$ and ends in state j at time t

$\boxed{\xi_t(i, j) \in [0, 1]}$

Posterior probability of staying state i at time t and state j at time $t + 1$

$\boxed{\gamma_t(j, k) \in [0, 1]}$

Posterior probability of staying at state j and mixture component k at time t

$\boxed{\pi_j \in [0, 1]}$

Initial state probability of state j at time $t = 1$

$\boxed{a_{ij} \in [0, 1]}$

State transition probability from state $s_{t-1} = i$ to state $s_t = j$

$\boxed{\omega_{jk} \in [0, 1]}$

Gaussian mixture weight at component k of state j

$\boxed{\boldsymbol{\mu}_{jk} \in \mathbb{R}^D}$

Gaussian mean vector at component k of state j

$\boxed{\boldsymbol{\Sigma}_{jk} \in \mathbb{R}^{D \times D}}$

Gaussian covariance matrix at component k of state j. Symmetric matrix

$\boxed{\mathbf{R}_{jk} \in \mathbb{R}^{D \times D}}$

Gaussian precision matrix at component k of state j. Symmetric matrix, and the inverse of covariance matrix $\boldsymbol{\Sigma}_{jk}$

Language modeling

$w \in \Sigma^*$

Category (e.g., word in most cases, phoneme sometimes). The element is represented by a string in Σ^* (e.g., “I” and “apple” for words and /a/ and /k/ for phonemes) or a natural number in $\mathbb{Z}^+$ when the elements of categories are numbered.

$\mathcal{V} \subset \Sigma^*$

Vocabulary (dictionary), i.e., a set of distinct words, which is a subset of Σ^*

$|\mathcal{V}|$

Vocabulary size

$v \in \{1, \cdots, |\mathcal{V}|\}$

Ordered index number of distinct words in vocabulary $\mathcal{V}$

$w_{(v)} \in \mathcal{V}$

Word pointed by an ordered index v

$\{w_{(v)}|v = 1, \cdots, |\mathcal{V}|\} = \mathcal{V}$

A set of distinct words, which is equivalent to vocabulary $\mathcal{V}$

$J \in \mathbb{Z}^+$

Number of categories in a chunk (e.g., number of words in a sentence or number of phonemes or HMM states in a speech segment)

$i \in \{1, \cdots, J\}$

ith position of category (e.g., word or phoneme)

$w_i \in \mathcal{V}$

Word at ith position

$W = \{w_i|i = 1, \cdots, J\}$

Word sequence from 1 to J

$w_{i-n+1}^{i} = \{w_{i-n+1} \cdots w_i\}$

Word sequence from $i - n + 1$ to i

$p(w_i|w_{i-n+1}^{i-1}) \in [0, 1]$

n-gram probability, which considers $n - 1$ order Markov model

$\boxed{c(w_{i-n+1}^{i-1}) \in \mathbb{Z}^{+}}$

Number of occurrences of word sequence w_{i-n+1}^{i-1} in a training corpus

$\boxed{\lambda_{w_{i-n+1}^{i-1}}}$

Interpolation weight for each w_{i-n+1}^{i-1}

$\boxed{M \in \mathbb{Z}^{+}}$

Number of documents

$\boxed{m \in \{1, \cdots, M\}}$

Document index

$\boxed{d_m}$

mth document, which would be represented by a string or positive integer

$\boxed{c(w_{(v)}, d_m) \in \mathbb{Z}^{+}}$

Number of co-occurrences of word $w_{(v)}$ in document d_m

$\boxed{K \in \mathbb{Z}^{+}}$

Number of unique latent topics

$\boxed{z_i \in \{1, \cdots, K\}}$

ith latent topic variable for word w_i

$\boxed{Z = \{z_j | j = 1, \cdots, J\}}$

Sequence of latent topic variables for J words

Abbreviations

AIC: Akaike Information Criterion (page 217)
AM: Acoustic Model (page 3)
ARD: Automatic Relevance Determination (page 194)
ASR: Automatic Speech Recognition (page 58)
BIC: Bayesian Information Criterion (page 8)
BNP: Bayesian Nonparametrics (pages 337, 345)
BPC: Bayesian Predictive Classification (page 218)
CDHMM: Continuous Density Hidden Markov Model (page 157)
CRP: Chinese Restaurant Process (page 350)
CSR: Continuous Speech Recognition (page 334)
DCLM: Dirichlet Class Language Model (page 326)

DHMM: Discrete Hidden Markov Model (page 62)

DNN: Deep Neural Network (page 224)

DP: Dirichlet Process (page 348)

EM: Expectation Maximization (page 9)

fMLLR: feature-space MLLR (page 204)

GMM: Gaussian Mixture Model (page 63)

HDP: Hierarchical Dirichlet Process (page 337)

HMM: Hidden Markov Model (page 59)

HPY: Hierarchical Pitman–Yor Process (page 383)

HPYLM: Hierarchical Pitman–Yor Language Model (page 384)

iid: Independently, identically distributed (page 216)

KL: Kullback–Leibler (page 79)

KN: Kneser–Ney (page 102)

LDA: Latent Dirichlet Allocation (page 318)

LM: Language Model (page 3)

LSA: Latent Semantic Analysis (page 113)

LVCSR: Large Vocabulary Continuous Speech Recognition (page 97)

MAP: Maximum A-Posteriori (page 7)

MAPLR: Maximum A-Posteriori Linear Regression (page 287)

MBR: Minimum Bayes Risk (page 56)

MCE: Minimum Classification Error (page 59)

MCMC: Markov Chain Monte Carlo (page 337)

MDL: Minimum Description Length (page 9)

MFCC: Mel-Frequency Cepstrum Coefficients (page 249)

MKN: Modified Kneser–Ney (page 111)

ML: Maximum Likelihood (page 77)

ML2: Type-2 Maximum Likelihood (page 188)

MLLR: Maximum Likelihood Linear Regression (page 200)

MLP: MultiLayer Perceptron (page 326)

MMI: Maximum Mutual Information (page 167)

MMSE: Minimum Mean Square Error (page 139)

MPE: Minimum Phone Error (page 167)

nCRP: nested Chinese Restaurant Process (page 360)

NDP: Nested Dirichlet Process (page 360)

NMF: Non-negative Matrix Factorization (page 124)

pdf: probability density function (page 63)

PLP: Perceptual Linear Prediction (page 54)

PLSA: Probabilistic Latent Semantic Analysis (page 113)

PY: Pitman–Yor Process (page 379)

QB: Quasi-Bayes (page 180)

RHS: Right-Hand Side (page 199)

RLS: Regularized Least-Squares (page 188)

RVM: Relevance Vector Machine (page 192)

SBL: Sparse Bayesian Learning (page 194)

SBP: Stick Breaking Process (page 348)
SMAP: Structural Maximum A-Posteriori (page 288)
SMAPLR: Structural Maximum A-Posteriori Linear Regression (page 288)
SVD: Singular Value Decomposition (page 114)
SVM: Support Vector Machine (page 188)
tf–idf: term frequency – inverse document frequency (page 113)
UBM: Universal Background Model (page 172)
VB: Variational Bayes (page 7)
VC: Vapnik–Chervonenkis (page 191)
VQ: Vector Quantization (page 62)
WB: Witten–Bell (page 102)
WER: Word Error Rate (page 56)
WFST: Weighted Finite State Transducer (page 60)
WSJ: Wall Street Journal (page 108)