Common Mathematical Notation in Machine Learning

Søren Hauberg
Section for Cognitive Systems
Technical University of Denmark
sohau@dtu.dk

February 5, 2019

1 Formatting and fonts

It is quite common to use various stylistic conventions when writing mathematics to give the reader visual hints as to what's going on. There are no official conventions regarding such hints but Bishop's textbook provides a commonly applied standard. We will also try to follow this convention in this course.

The most basic convention is that

- **Scalars** are written in lower case italics, i.e. $x = 7$, $y \in \mathbb{R}$, $z = x + y$.
- **Vectors** are written in lower case bold, i.e. $\mathbf{x} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$. In general, vectors are *column vectors*, so $\mathbf{x} \in \mathbb{R}^D$ is the same as writing $\mathbf{x} \in \mathbb{R}^{D \times 1}$. Elements of vectors are scalars and are, hence, not written in bold, e.g. the d^{th} element of $\mathbf{x}$ is written x_d.
- **Matrices** are written in upper case bold, i.e. $\mathbf{X} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$. Elements of matrices are usually written in upper case, but not in bold, e.g. X_{ij}.

2 Distributions

You will often find that the almost same notation is used to mean slightly different things, e.g. if $\mathbf{x}$ (by convention that's a vector) is distributed according to a normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ then we may write

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{2.1}$$

This reads "$\mathbf{x}$ *is distributed according to a normal distribution with mean* $\boldsymbol{\mu}$ *and covariance* $\boldsymbol{\Sigma}$*.*" We can choose to communicate the same information as

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{2.2}$$

which says *"The distribution of* $\mathbf{x}$ *is normal with parameters* $\boldsymbol{\mu}$ *and* $\boldsymbol{\Sigma}$*."*

This notation is often confusing at first sight. In Eq. 2.1 we write $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ while in Eq. 2.2 we write $\mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma})$. Note how in the latter case, $\mathcal{N}$ also include the argument $\mathbf{x}$. You should read that $\mathcal{N}$ is a function with input argument $\mathbf{x}$ that has additional parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. In Eq. 2.1 $\mathcal{N}$ is not a function *per se*, but rather a specification of random variable.

3 Identity matrices

The identity matrix (diagonal matrix with 1's along the diagonal) is written $\mathbf{I}$. This can be tricky as the $d \times d$ identity matrix is then written the same way as the $D \times D$ identity matrix. So technically $\mathbf{I} \neq \mathbf{I}$, which is a quite confusing notation. In practice, you will have to guess the dimension of $\mathbf{I}$ from its context.

4 Data matrices

We often deal with individual observations which are vectors $\mathbf{x} \in \mathbb{R}^D$. We usually have N observations, which can be written $\mathbf{x}_{1:N} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$. Many times it is convenient to have all observations jointly in one matrix, which can be done by stacking the data vectors row-wise or column-wise, i.e.

$$\mathbf{X} = \begin{pmatrix} — \mathbf{x}_1^\top — \\ \vdots \\ — \mathbf{x}_N^\top — \end{pmatrix} \in \mathbb{R}^{N \times D} \qquad \text{or} \qquad \mathbf{X} = \begin{pmatrix} | & & | \\ \mathbf{x}_1 & \cdots & \mathbf{x}_N \\ | & & | \end{pmatrix} \in \mathbb{R}^{D \times N}. \tag{4.1}$$

Different authors have different conventions, so always (always!) check if the dimensions assumed in the mathematical expressions match that of your code.

5 Matrix operators

Two of the most elementary properties of a matrix have changing notation.

- **The trace** (sum of diagonal elements) of a matrix $\mathbf{X}$ has many different notations; the list include at least

$$\text{Trace}(\mathbf{X}) = \text{trace}(\mathbf{X}) = \text{tr}(\mathbf{X}) = \text{Tr}(\mathbf{X}). \tag{5.1}$$

- **The determinant** of matrix $\mathbf{X}$ is either written

$$\det(\mathbf{X}) = \det\mathbf{X} = |\mathbf{X}|. \tag{5.2}$$

  Note that $|x|$ is the absolute value of scalar x, which also happens to be the determinant of the 1×1 matrix containing x as its sole value.