

Math refresher

Linear algebra

Why linear algebra? Represent a set of linear equations:

$$4x_1 + -5x_2 = -13 \quad (1)$$

$$-2x_1 + 3x_2 = 9 \quad (2)$$

can be represented as:

$$Ax = b, \quad A = \begin{bmatrix} 4 & -5 \\ -2 & 3 \end{bmatrix}, \quad b = \begin{bmatrix} 13 \\ -9 \end{bmatrix} \quad (3)$$

Now we can solve the set of equations by solving for x .

Additionally, provides convenient notation. Will often write $\sum_i w_i x_i$; it is shorter (and easier to work with) to write $w^\top x$ or $\langle w, x \rangle$ or $(w \cdot x)$ for this (we will stick with the first, but any is okay) ... Matlab: `w'x`.

Notation:

1. Matrices are capitalized, $A \in \mathbb{R}^{M \times N}$ means an M -row, N -column matrix.
2. Vectors will be column vectors, so that $x \in \mathbb{R}^M$ means $x \in \mathbb{R}^{1 \times M}$. To get a row vector, we write x^\top .
3. If x is a vector, then x_m is the m th element of x (a scalar).
4. If A is a matrix, then $A_{m,n}$ or A_{mn} (when clear) is the m th row, n th column scalar from A .
5. If A is a matrix, then $A_{:,n}$ is the column vector from the n th column of A and $A_{m,:}$ is the row vector from the m th row of A .

Simple operations:

1. $A \in \mathbb{R}^{M \times N}$ and $\alpha \in \mathbb{R}$, then $A + \alpha$ is defined by $(A + \alpha)_{mn} = A_{mn} + \alpha$ and $(\alpha A)_{mn}$ is αA_{mn} . Subtraction and division similar.
2. $A \in \mathbb{R}^{M \times N}$ and $B \in \mathbb{R}^{N \times P}$ means $C = AB \in \mathbb{R}^{M \times P}$, with $C_{mp} = \sum_n A_{mn} B_{np} = A_{m,:} B_{:,p}$.
3. Vectors $x \in \mathbb{R}^M$ and $y \in \mathbb{R}^M$, $x^\top y = \sum_m x_m y_m$ is a scalar.
4. Vectors $x \in \mathbb{R}^M$ and $y \in \mathbb{R}^N$, $xy^\top \in \mathbb{R}^{M \times N}$ defined by $(xy^\top)_{mn} = x_m y_n$.
5. Multiplication is associative and distributive but *not* commutative: $(AB)C = A(BC)$ and $A(B+C) = AB + AC$, but it is possible that $AB \neq BA$.

Additional definitions:

1. Identity matrix, $I \in \mathbb{R}^{N \times N}$ defined by $I_{mn} = 1_{m=n}$. For all A (of appr. size), $AI = IA = A$.
2. Diagonal matrix, $D = \text{diag}(d) \in \mathbb{R}^{N \times N}$ for $d \in \mathbb{R}^N$ defined by $D_{mn} = 1_{m=n}d_m$.
3. Transpose, if $A \in \mathbb{R}^{M \times N}$ then $A^\top \in \mathbb{R}^{N \times M}$ with $(A^\top)_{nm} = A_{mn}$. Holds: $(A^\top)^\top = A$, $(AB)^\top = B^\top A^\top$ and $(A+B)^\top = A^\top + B^\top$.
4. Symmetric, square A is symmetric if $A = A^\top$. Often $\mathbb{S}^N = \{A \in \mathbb{R}^{N \times N} : A \text{ is symmetric}\}$.
5. Trace, the sum of the diagonals: $\text{tr } A = \sum_n A_{nn}$ for $A \in \mathbb{R}^{N \times N}$.

Norms:

1. For a vector x , the norm of x , $\|x\|$ is typically the Euclidean, or ℓ_2 norm: $\|x\|_2^2 = x^\top x$.
2. ℓ_p norm is $\|x\|_p^p = \sum_n |x_i|^p$
3. Infinite norm: $\|x\|_\infty = \max_i |x_i|$.

Inverse:

1. $A \in \mathbb{R}^{N \times N}$, and A invertible, then A^{-1} is unique such that $A^{-1}A = AA^{-1} = I$.
2. A is invertible if all of its rows (or columns) are linearly independent.
3. $(A^{-1})^{-1} = A$, $(AB)^{-1} = B^{-1}A^{-1}$, $(A^{-1})^\top = (A^\top)^{-1}$.
4. If $Ax = b$ then $x = A^{-1}b$.

Determinant

1. $A \in \mathbb{R}^{N \times N}$, then $\det A \in \mathbb{R}$ (sometimes denoted $|A|$).
2. $\det I = 1$, $\det A = \det A^\top$, $\det AB = \det A \det B$
3. $\det A = 0$ implies A is singular (non-invertible); otherwise, $\det A^{-1} = 1/\det A$
4. If we multiply one row of A by α to get A' , then $\det A' = \alpha \det A$. If we swap two rows of A to get A' then $\det A' = -\det A$.
5. In general:

$$\det A = \sum_n (-1)^{m+n} A_{mn} \det A_{-m,-n} \quad (\text{for any } m \in 1, \dots, N) \quad (4)$$

$$= \sum_m (-1)^{m+n} A_{mn} \det A_{-m,-n} \quad (\text{for any } n \in 1, \dots, N) \quad (5)$$

6. If $A = \text{diag}(a)$ then $\det A = \prod_n a_n$ (other special cases exist).

Positive definite matrices

1. For a quadratic form, $x^\top Ax = \sum_m \sum_n A_{mn} x_m x_n$, we can assume A is symmetric.
2. If $A \in \mathbb{S}^N$, we say A is pd if for all $x \in \mathbb{R}^N$, $x^\top Ax > 0$ (we write $A \succ 0$). Set is \mathbb{S}_{++}^N .

3. If ... and $x^\top Ax \geq 0$ then A is psd ($A \succeq 0$) and set is \mathbb{S}_+^N .
 4. PD and ND matrices are always invertible
 5. If $B \in \mathbb{R}^{M \times N}$ then $G = A^\top A \in \mathbb{R}^{N \times N}$ is always psd.
-

Calculus

If $f : \mathbb{R} \rightarrow \mathbb{R}$ is well-behaved, then $\partial f / \partial x$ is the first derivative and $\partial^2 f / \partial x^2$ is the second derivative (etc.)

Gradient is multivariate extension. Suppose $f : \mathbb{R}^{M \times N} \rightarrow \mathbb{R}$ is well-behaved, then the *gradient* is defined by:

$$\nabla_A f = \begin{bmatrix} \frac{\partial f}{\partial A_{11}} & \frac{\partial f}{\partial A_{12}} & \cdots & \frac{\partial f}{\partial A_{1N}} \\ \frac{\partial f}{\partial A_{21}} & \frac{\partial f}{\partial A_{22}} & \cdots & \frac{\partial f}{\partial A_{2N}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial A_{M1}} & \frac{\partial f}{\partial A_{M2}} & \cdots & \frac{\partial f}{\partial A_{MN}} \end{bmatrix} \in \mathbb{R}^{M \times N} \quad (6)$$

$$(\nabla_A f)_{mn} = \frac{\partial f}{\partial A_{mn}} \quad (7)$$

The usual linearity rules hold.

As with standard calculus, $(\nabla f)(A) = 0$ means that A is an extremum of f .

The Hessian is a multivariate extension of the second derivative. Suppose $f : \mathbb{R}^N \rightarrow \mathbb{R}$, then:

$$\nabla_x^2 f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_N} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_N} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_M \partial x_1} & \frac{\partial^2 f}{\partial x_M \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_M^2} \end{bmatrix} \in \mathbb{R}^{N \times N} \quad (8)$$

$$(\nabla_x^2 f)_{mn} = \frac{\partial^2 f}{\partial x_m \partial x_n} \quad (9)$$

Examples ...

Probability and Statistics

A random variable (r.v.) is a value determined by chance (drawn by a probability distribution).

1. Input data
2. Output data

3. Noise
4. ...

Often will think in terms of a *data generating model*.

Discrete distributions take on discrete values:

1. Bernoulli (coin flipping): $p(1) = \pi$, $p(0) = 1 - \pi$ for $\pi \in [0, 1]$. Also $p(x) = \pi^x(1 - \pi)^{1-x}$.
2. Binomial (coin flipping, cont'd): what is the probability of k heads in a sequence of n trials ($n \geq k$)? $p(k | n, \pi) = \binom{n}{k} \pi^k (1 - \pi)^{n-k}$.
3. Multinomial (rolling dice): parameter vector $\theta \in \mathbb{R}^K \geq 0$ with $\sum_k \theta_k = 1$. Let x_k be the number of times k comes up in n rolls, then $p(x | \theta, n) = \binom{n}{\prod_k x_k} \prod_k \theta_k^{x_k}$.

Continuous distributions take on continuous values. Think of discrete distributions with number of values approaching infinity, and probability of each approaching zero.

1. Req: $\int_{-\infty}^{\infty} dx p(x) = 1$.
2. Events are: $p(a < x < b) = \int_a^b dx p(x)$

Examples:

1. Uniform distribution: $p(x | R) = 1_{0 \leq x \leq R} \frac{1}{R}$
2. Univariate normal distribution ($x, \mu, \sigma^2 \in \mathbb{R}$)

$$p(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{1}{2\sigma^2} (x - \mu)^2 \right] \quad (10)$$

Examples...

3. Multivariate normal distribution ($x, \mu \in \mathbb{R}^N$, $\Sigma \in \mathbb{S}_+^N$):

$$p(x | \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^N \det \Sigma}} \exp \left[-\frac{1}{2} (x - \mu)^\top \Sigma^{-1} (x - \mu) \right] \quad (11)$$

Measure distance with Mahalanobis, points of equal distance have constant density.

Examples...

Expectations:

1. For discrete p , $\mathbb{E}_{x \sim p}[x] = \sum_i p(x_i) x_i$
2. For continuous p , $\mathbb{E}_{x \sim p}[x] = \int dx p(x) x$
3. Called the mean, or center of mass; not the same as the *mode*.
4. Expectations of functions are obtained by replacing “ x ” in definition with $f(x)$; eg: $\mathbb{E}_{x \sim p}[f(x)] = \sum_i p(x_i) f(x_i)$.

5. If a independent of x , then $\mathbb{E}[ax] = a\mathbb{E}[x]$.
6. If x and y are independent then $\mathbb{E}[x + y] = \mathbb{E}[x] + \mathbb{E}[y]$.
7. Often $\mathbb{E}[xy] \neq \mathbb{E}[x]\mathbb{E}[y]$

The *variance* is a measure of dispersion: $\mathbb{V}[x] = \mathbb{E}[(x - \mathbb{E}[x])^2] = \mathbb{E}[x^2] - (\mathbb{E}[x])^2$.

Covariance: $Cov[x, y] = \mathbb{V}[x, y] = \mathbb{E}[(x - \mathbb{E}[x])(y - \mathbb{E}[y])]$

If p is a joint distribution ($p(x, y)$), then $\int dx \int dy p(x, y) = 1$ and can obtain marginal by $p(x) = \int dy p(x, y)$.

x and y are *independent* in $p(x, y)$ if $p(x, y) = p(x)p(y)$. Simplifies marginalization a lot!

Conditional: $p(y | x) = p(x, y)/p(x)$ and $p(x, y) = p(y | x)p(x) = p(x | y)p(y)$. If independent, $p(y | x) = p(y)$. Yields chain rule.

Bayes' rule: $p(y | x) = \frac{p(x | y)p(y)}{p(x)}$ as prior times likelihood over marginal.