



Algebra and Notation Refresher for ML

Maths for AI | Sets, functions, sums Σ and products (capital Pi), exponents and logarithms, indices, and a decoder for the Greek letters and symbols of ML papers

[#Maths](#) [#Notation](#) [#Logs](#) [#Sums](#) [#Beginner](#)

Notation in plain words

Mathematical notation

×1

×10

×100

Overview

Close up

Under the microscope

A compressed sentence: Σ means 'add up', capital Pi 'multiply', log 'how many times do I multiply', subscripts say 'which one'.

ML papers reuse a small vocabulary: bold lower-case for vectors, upper-case for matrices, Greek letters for parameters and constants, hats for estimates.

- x_i = the i -th example
- θ = all model parameters
- $\hat{y}$ = a prediction

Conventions vary by community; every good paper defines its symbols in a 'Notation' paragraph. When in doubt, check the shapes.

- $\log = \ln$ unless a base is written
- $\|x\|$ without subscript usually means the L2 norm

ANALOGY

Like SMS abbreviations: 'brb' saves letters once you know the code; the same message spelled out is longer, never harder.

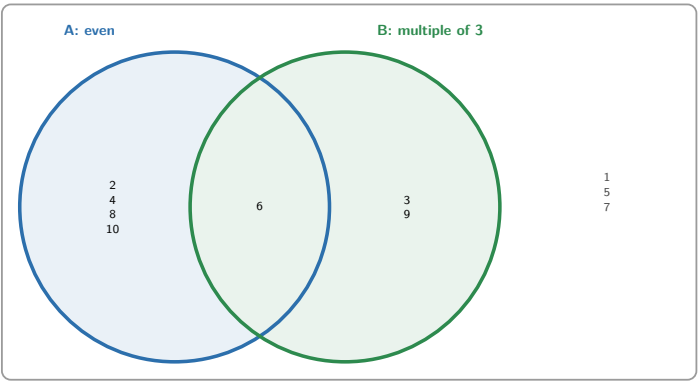
Notation is shorthand, not new mathematics. Three zoom levels on the same idea

Before Reading ML formulas, can you explain...

	Not yet	Partly	Yes
MUST KNOW Order of operations			
powers, then $\times$ and $/$, then $+$ and $-$: $2 + 3 \times 4^2 = 50$			
MUST KNOW Fractions and negative numbers			
$-(1/4) \times 8 = -2$			
MUST KNOW Solving for an unknown			
$2x - 5 = 9$ gives $x = 7$			
HELPFUL Powers with negative and fractional exponents			
$2^{-3} = 1/8$, $9^{1/2} = 3$; Learn: section 'Exponents and logarithms'			
HELPFUL Reading a function plot			
where it rises, falls, crosses zero			

Check before reading on. A 'not yet' on a must-know line: revise it first (any school algebra text)

Sets and functions



Two sets of whole numbers from 1 to 10: $A = \text{even}$, $B = \text{multiples of 3}$. $A \cap B = \{6\}$ (intersection: in both), $A \cup B = \{2, 3, 4, 6, 8, 9, 10\}$ (union: in either), $A \setminus B = \{2, 4, 8, 10\}$ (difference: in A, not in B)

Name	Symbol	Reads as, with an ML example
Membership	$x \in \mathbb{R}^d$	x belongs to the set of d-dimensional real vectors: one input with d features
Subset	$A \subseteq B$	every element of A is in B: the training set is a subset of all data
Set builder	$\{x : x > 0\}$	the set of all x such that x > 0
Cardinality	$ \mathcal{D} $	number of elements of the dataset D, often written N
Integer range	$[N] = \{1, \dots, N\}$	index set of the N training examples
Function	$f : \mathbb{R}^d \rightarrow \mathbb{R}^k$	f takes a d-vector and returns a k-vector: a classifier with k class scores
Maps to	$x \mapsto x^2$	the function that sends x to x squared
Composition	$(f \circ g)(x) = f(g(x))$	apply g, then f: layer 2 after layer 1
Inverse	f^{-1}	undoes f: exp is the inverse of ln
Indicator	$\mathbb{I}[\hat{y}_i = y_i]$	1 if prediction i is right, else 0; accuracy is its average

where x = an input, a number or a vector; d, k = number of input features, number of outputs; A, B = sets; $\mathcal{D}$ = the dataset, a set of N examples; $y_i, \hat{y}_i$ = true label and predicted label of example i
 Set and function symbols you will meet in every paper

min f(x)

the smallest value of f

vs

argmin f(x)

the x that gives it

Tell: Is the answer a value or a location?

f(x)

a number: f evaluated at x

vs

f

the whole rule

Tell: Is there an argument in brackets?

x_i (subscript)

element i of x

vs

x̂(i) (bracketed superscript)

example i of the dataset

Tell: Brackets around a superscript mean 'which example', not a power

A ∩ B

in both sets

vs

A ∪ B

in at least one

Tell: ∩ looks like an n for 'and both'

Four pairs readers confuse in their first papers

≡ Sums Σ and products (capital Pi)		
Rule	Formula	Example
Sum	$\sum_{i=1}^n a_i = a_1 + a_2 + \dots + a_n$	Σ from i=1 to 4 of i = 1+2+3+4 = 10
Constant out	$\sum_{i=1}^n c a_i = c \sum_{i=1}^n a_i$	Σ 3i for i=1..4 = 3 × 10 = 30
Split	$\sum_i (a_i + b_i) = \sum_i a_i + \sum_i b_i$	loss = data term + penalty term
Constant term	$\sum_{i=1}^n c = n c$	Σ 5 for i=1..10 = 50
Mean	$\bar{a} = \frac{1}{n} \sum_{i=1}^n a_i$	mean of 2, 4, 9 = 15/3 = 5
Double sum	$\sum_i \sum_j W_{ij}$ = sum of all entries	rows first or columns first: same total
Product	$\prod_{i=1}^n a_i = a_1 a_2 \dots a_n$	Product of i for i=1..4 = 24 = 4!
Log of product	$\log \prod_i p_i = \sum_i \log p_i$	turns a likelihood into a log-likelihood

where i, j = index variables (counters); the name does not matter; n = number of terms; a_i, b_i = the i-th term of two sequences; c = a constant that does not depend on the index; W_{ij} = entry in row i, column j of a matrix W; p_i = a probability, between 0 and 1
 Rules that let you move sums and products around. Every derivation in sheets 05-11 uses them

PROBLEM A model gives the true labels of three examples probabilities 0.5, 0.2 and 0.1. Compute the likelihood (product) and the log-likelihood (sum of logs)

GIVEN

- $p_1 = 0.5$, $p_2 = 0.2$, $p_3 = 0.1$ (probability given to the right answer)
- likelihood L = product of the p_i ; log-likelihood $LL = \sum \ln p_i$ ($\ln$ = natural log, base $e \approx 2.718$)

1 Multiply the probabilities

$$L = 0.5 \times 0.2 \times 0.1 = 0.01$$

Why: independent examples: probabilities multiply

2 Take each natural log

$$\ln 0.5 = -0.693; \ln 0.2 = -1.609; \ln 0.1 = -2.303$$

Why: logs of numbers below 1 are negative

3 Add the logs

$$LL = -0.693 - 1.609 - 2.303 = -4.605$$

Why: log of a product = sum of logs

4 Turn into a loss

$$NLL = -LL / 3 = 1.535 \text{ nats per example}$$

Why: negative log-likelihood, averaged: smaller is better

ANSWER $L = 0.01$, $LL = -4.605$, average loss 1.535 nats

CHECK $e^{-4.605} = 0.0100$ ✓. With 10,000 examples a product would underflow to 0 in floating point; the sum of logs stays a normal number

Why every ML loss is a sum of logs and not a product of probabilities

Expanding a sum by hand

$\sum$ from $i=1$ to 4 of $(2i - 1) = 1 + 3 + 5 + 7 = 16$. Check with the rules: $2 \times \sum i - \sum 1 = 2 \times 10 - 4 = 16$. Sums of the first n odd numbers always give n^2 .

Exponents and logarithms

Identity	Formula	Numeric check
Product of powers	$b^m \cdot b^n = b^{m+n}$	$2^3 \times 2^4 = 2^7 = 128$
Power of power	$(b^m)^n = b^{mn}$	$(2^3)^2 = 2^6 = 64$
Negative exponent	$b^{-n} = 1/b^n$	$10^{-3} = 0.001$
Fractional exponent	$b^{1/n} = \sqrt[n]{b}$	$8^{(1/3)} = 2$
Definition of log	$\log_b x = y \iff b^y = x$	$\log_2 1024 = 10$
Log of product	$\log(xy) = \log x + \log y$	$\ln 6 = \ln 2 + \ln 3 = 0.693 + 1.099 = 1.792$
Log of quotient	$\log(x/y) = \log x - \log y$	$\log\text{-odds} = \ln p - \ln(1 - p)$
Log of power	$\log(x^k) = k \log x$	$\log_{10} (10^{25}) = 25$
Change of base	$\log_b x = \ln x / \ln b$	$\log_2 10 = 2.303 / 0.693 = 3.32$
Inverse pair	$e^{\ln x} = x$, $\ln(e^x) = x$	exp and ln undo each other

where b = the base, a positive number other than 1 (2, e or 10 in practice); m , n , k = exponents (any real numbers); x , y = positive real numbers (log is undefined at 0 and below); e = Euler's number, about 2.71828; p = a probability strictly between 0 and 1

The identities used to simplify losses, scaling laws and floating-point formats

Base	Written	Where in AI	1 unit equals
$e \approx 2.718$	$\ln$, log (ML papers)	losses in PyTorch and JAX, log-likelihoods	1 nat
2	$\log_2$, lg	information in bits, tree depth, Big-O of binary search	1 bit = 0.693 nat
10	$\log_{10}$, log (engineering)	orders of magnitude, dB, log-scale plots of FLOPs	1 decade

Which log, where. Converting between them only multiplies by a constant

YOU MAY HEAR

" $\log(a + b) = \log a + \log b$ "

BUSTED

ACTUALLY

$\log(a \times b) = \log a + \log b$

There is no simple rule for the log of a sum; that is why log-sum-exp (sheet 12) exists.

YOU MAY HEAR

"log in a PyTorch loss is base 10"

BUSTED

ACTUALLY

torch.log is the natural log

Cross-entropy in PyTorch is in nats; divide by $\ln 2 = 0.693$ to get bits.

YOU MAY HEAR

" $(a + b)^2 = a^2 + b^2$ "

BUSTED

ACTUALLY

$(a + b)^2 = a^2 + 2ab + b^2$

With $a = 3$, $b = 4$: 49, not 25. The cross term $2ab$ appears in every variance derivation.

YOU MAY HEAR

"A product of many small probabilities is just a small number"

BUSTED

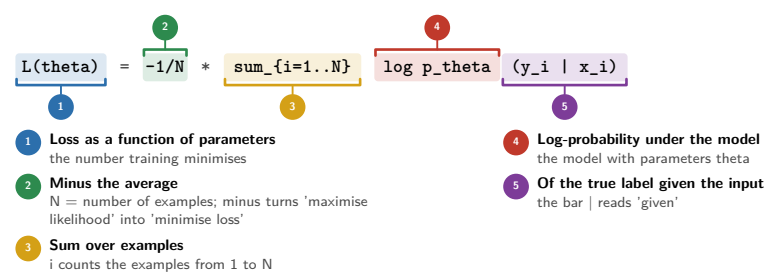
ACTUALLY

It underflows to exactly 0 in float32

1000 probabilities of 0.1 give 10^{-1000} ; float32 cannot go below about 10^{-45} . Work with logs.

Log and exponent traps that produce wrong derivations

Indices and the Greek alphabet of ML papers



A typical training objective, segment by segment. The same pattern (average over examples of a log-probability) is the cross-entropy loss of sheet 11

Letter	Symbol	Usual meaning
alpha	α	learning rate; significance level of a test; Dirichlet concentration
beta	β	Adam decay rates beta_1, beta_2; regression coefficients; KL weight in a VAE or DPO
gamma	γ	discount factor in RL (0.99); scale of a normalisation layer
delta	δ, Δ	small change or error signal; Delta = difference between two values
epsilon	ϵ	tiny constant for stability (1e-8); noise in diffusion; exploration rate
eta	η	learning rate (often used instead of alpha)
theta	θ	all trainable parameters of a model
lambda	λ	regularisation strength; eigenvalue; Lagrange multiplier; Poisson rate
mu	μ	mean of a distribution or of a batch
sigma	σ, Σ	standard deviation; sigmoid function; Sigma = covariance matrix or a sum
pi	π	policy in RL; mixture weights; the constant 3.14159
phi, psi	ϕ, ψ	parameters of a second network (encoder, critic); a feature map
tau	τ	temperature of a softmax; time step; target-network update rate
rho	ρ	correlation coefficient; density
kappa	κ	condition number of a matrix (sheet 12)
nabla	$\nabla_{\theta} L$	gradient of L with respect to theta (sheet 06)

Greek letters and their usual job in ML papers. The meaning is a convention: always check the paper's own notation section

Mark	Symbol	Meaning
hat	$\hat{y}$	estimate or prediction of y
bar	$\bar{x}$	average of x
star	θ^*	optimal value of theta
tilde	$x \sim p$	x is sampled from distribution p
expectation	$\mathbb{E}_{x \sim p}[f(x)]$	average of f(x) when x is drawn from p
proportional	$p(\theta \mid x) \propto p(x \mid \theta) p(\theta)$	equal up to a constant factor
definition	$z := Wx + b$	z is defined as W x + b
norm	$\ x\ _2 = \sqrt{\sum_i x_i^2}$	length of vector x (sheet 02)
transpose	$x^\top$	row version of column vector x (sheet 03)
element-wise	$a \odot b$	multiply entry by entry

where y , x = a target value; an input or a sample; p = a probability distribution; f = any function of x ;
 W , b = weight matrix and bias vector of a linear layer; a , b = two arrays of the same shape (in the last row)

Accents and operators: a hat is an estimate, a bar an average, a star an optimum

EXPERT SAYS

We minimise the expected negative log-likelihood under the empirical distribution.



IN PLAIN WORDS

We make the average of -log(probability of the right answer) over the training examples as small as possible.

expected = averaged, weighted by probability
negative log-likelihood = -log of the probability the model gives to the data
empirical distribution = the training set, each example with weight 1/N

EXPERT SAYS

Let $x \in R^d$ and $\theta^* = \operatorname{argmin}$ over θ of $L(\theta) + \lambda \|\theta\|^2$.



IN PLAIN WORDS

x is a list of d numbers; the best parameters are those that make the loss plus a size penalty smallest.

R^d = vectors of d real numbers
argmin = the input that minimises
 λ = how much the penalty counts
 $\|\theta\|^2$ = sum of squared parameters

Paper sentences, translated

Check yourself

```
Check any identity numerically (Python, NumPy)
import numpy as np
p = np.array([0.5, 0.2, 0.1])
print(np.prod(p))          # 0.01
print(np.log(p).sum())     # -4.605 (natural log)
print(np.log2(1024))       # 10.0
print(sum(2*i - 1 for i in range(1, 5))) # 16
```

Tick a box only when you could explain the point to someone else



Σ adds, capital Pi multiplies
constants come out, sums split



log turns products into sums
the reason losses are sums of log-probabilities



log means ln in ML
nats; divide by ln 2 for bits



argmin returns a location
min returns a value



Greek letters are conventions
theta parameters, eta or alpha learning rate, lambda penalty



Hat = estimate, bar = mean, star = optimum
and $x \sim p$ means sampled from p

Tick a line only when you could explain it to someone else

Q1 Σ from i=1 to 5 of 2?

A 10: a constant summed n times is $n \times c$.

Q2 Simplify $\log(a b^3)$.

A $\log a + 3 \log b$ (log of product, then log of power).

Q3 What is $\log_2$ of 0.25?

A -2, because $2^{-2} = 1/4$.

Q4 Why do ML losses use sums of logs instead of products of probabilities?

A A product of many probabilities underflows to 0 and is awkward to differentiate; the sum of logs is stable and the argmax is the same.

Q5 What does $\theta^* = \operatorname{argmin}$ over θ of $L(\theta)$ say?

A θ^* is the parameter value at which the loss L is smallest.

Q6 $x^{(i)}$ versus x_i ?

A $x^{(i)}$ is the i-th example of the dataset; x_i is the i-th component of a vector x.

Q7 What does the indicator $1[\hat{y} = y]$ average to over a test set?

A The accuracy: the share of predictions equal to the label.

Q8 Convert 1.386 nats to bits.

A $1.386 / 0.693 = 2$ bits.

Q9 Which Greek letter usually names the softmax temperature?

A tau; a large tau flattens the distribution.