

Notation Decoder

The symbols used in optimization and machine learning, in plain English

Computational Labs · Michael Carter, Ph.D. · Free to share

The symbols that stop people reading — Greek letters, subscript conventions, gradient and probability operators — collected in one place with plain-English translations. Written for anyone who can follow a machine learning explanation right up until the notation starts.

Where a symbol carries several meanings, all of them are listed. That ambiguity is not a flaw in your understanding; it is a real property of the literature, and knowing which meanings compete is most of the battle. The last page collects the ten confusions that cause the most trouble.

Greek letters

One letter, many jobs. The meaning is set by the field and often by the single paper you are reading — always check whether the author defined it.

Symbol	Read as	What it means
α	alpha	Learning rate (step size). Also: significance level in statistics; the pheromone weight in ant colony; the best-ranked wolf in grey wolf optimization; a momentum term.
β	beta	Momentum coefficient. Also: regression coefficients ($\beta_0, \beta_1, \dots$); the heuristic weight in ant colony; the second-ranked wolf.
γ	gamma	Discount factor in reinforcement learning. Also: the RBF kernel coefficient; a learning-rate decay factor.
δ	delta	An error signal, especially in backpropagation. Also: the third-ranked wolf; a small perturbation.
ϵ	epsilon	A deliberately small positive number: numerical-stability padding, a convergence tolerance, or the exploration rate in ϵ -greedy.
η	eta	Learning rate — the older convention, still standard in neural-network texts. In ant colony, the heuristic desirability of a move.
θ	theta	Model parameters, collectively. When a paper writes $f(x; \theta)$, θ is everything the model learned.
κ	kappa	Condition number of a matrix. Also Cohen's kappa, an inter-rater agreement statistic.
λ	lambda	Regularization strength. Also: a Lagrange multiplier; an eigenvalue. All three appear in ML papers.

Symbol	Read as	What it means
μ	mu	A mean. Also a momentum coefficient, and in evolution strategies the number of parents.
ν	nu	Degrees of freedom; the ν parameter in ν -SVM. Easily misread as v.
ξ	xi	Slack variable — the amount by which a point is allowed to violate the margin in a soft-margin SVM.
π	pi	The policy in reinforcement learning, $\pi(a s)$. Also a mixing coefficient. Occasionally 3.14159.
ρ	rho	Correlation coefficient. In ant colony, the pheromone evaporation rate.
σ	sigma	Standard deviation (σ^2 is variance). Also the sigmoid function $\sigma(z)$, and the singular values of a matrix.
τ	tau	Pheromone level on an edge in ant colony. Also a temperature or time constant; Kendall's tau.
ϕ	phi	A feature map — the function that lifts x into a higher-dimensional space in kernel methods.
ω	omega	Inertia weight in particle swarm. The omega wolves — the non-leaders — in grey wolf. Sometimes weights.
Δ	capital delta	A change or difference. ΔE is the change in objective value, the quantity simulated annealing decides whether to accept.
Σ	capital sigma	Summation. BUT also the covariance matrix. Context decides — a Σ with sub- and superscripts is a sum; a bare bold Σ is almost always covariance.
Π	capital pi	Product over a sequence, the multiplicative counterpart of Σ .
Λ	capital lambda	A diagonal matrix of eigenvalues.
Φ	capital phi	The cumulative distribution function of the standard normal. Also the matrix of feature-mapped inputs.
Ω	capital omega	Asymptotic lower bound, $\Omega(n)$. Also the sample space in probability.

Sets, spaces and logic

Symbol	Read as	What it means
$\mathbb{R}$	the reals	All real numbers.
$\mathbb{R}^n$	R-n	The space of real vectors with n components. ' $x \in \mathbb{R}^n$ ' means x is a list of n real numbers.
$\mathbb{R}^{m \times n}$	R m-by-n	The space of real matrices with m rows and n columns.
$\mathbb{Z}, \mathbb{N}, \mathbb{Q}$	integers, naturals, rationals	Whole numbers (positive and negative), counting numbers, and fractions.

Symbol	Read as	What it means
$\in \notin$	is / is not an element of	Membership. ' $x \in A$ ' reads 'x is in A'.
$\subset \subseteq$	subset / subset or equal	Containment between sets.
$\cup \cap$	union / intersection	Everything in either set / everything in both.
$\emptyset$	empty set	The set containing nothing.
$\times$	Cartesian product	$[-10, 10] \times [-10, 10]$ is the square search space: every pair (x, y) with both coordinates in $[-10, 10]$.
$[a, b]$	closed interval	All values from a to b, endpoints included. (a, b) excludes them.
$\{x : P(x)\}$	set-builder	The set of all x satisfying condition P. The colon reads 'such that'.
$\forall \exists$	for all / there exists	Quantifiers. ' $\forall x \in A$ ' means the statement holds for every x in A.
$\Rightarrow \Leftrightarrow$	implies / if and only if	One-way and two-way logical consequence.
$\triangleq :=$	is defined as	Introduces a definition rather than asserting an equality.
s.t.	subject to / such that	In optimization, everything after it is a constraint the solution must satisfy.

Vectors, matrices and norms

Convention: bold lowercase is a vector, bold uppercase a matrix, plain italic a scalar. Many papers drop the bold entirely and expect you to infer it.

Symbol	Read as	What it means
$\mathbf{x}^T$	x transpose	Flip rows and columns. Turns a column vector into a row vector.
$\mathbf{A}^{-1}$	A inverse	The matrix that undoes A. Exists only if A is square and non-singular.
$\mathbf{A}^+$	A pseudoinverse	The generalization of the inverse to non-square or singular matrices.
$\mathbf{I}$	identity matrix	1s on the diagonal, 0s elsewhere. $\mathbf{A}\mathbf{I} = \mathbf{A}$.
$\text{tr}(\mathbf{A})$	trace	Sum of the diagonal entries.
$\det(\mathbf{A})$	determinant	A scalar summarizing the matrix; zero means non-invertible. Sometimes written $ \mathbf{A} $.
$\text{diag}(\mathbf{v})$	diagonal matrix	A square matrix with the entries of $\mathbf{v}$ on its diagonal.
$\odot$	Hadamard product	Element-by-element multiplication — not matrix multiplication.
$\otimes$	Kronecker product	A block-structured product that builds a large matrix from two smaller ones.

Symbol	Read as	What it means
$\langle x, y \rangle$	inner product	Also written $x \cdot y$ or $x^T y$. Multiply matching components and add. Result is a single number.
$\ x\ $	norm of x	The length of x . Unsubscripted, assume the Euclidean (L2) norm.
$\ x\ _1$	L1 norm	Sum of absolute values. The penalty that drives Lasso to produce zeros.
$\ x\ _2$	L2 norm	Square root of the sum of squares — ordinary straight-line distance.
$\ x\ _\infty$	infinity norm	The largest absolute component.
$\ A\ _F$	Frobenius norm	The L2 norm applied to a matrix as if it were one long vector.

Calculus and gradients

Symbol	Read as	What it means
d/dx	derivative	Rate of change of a single-variable function.
$\partial/\partial x$	partial derivative	Rate of change with respect to one variable, all others held fixed. The curly ∂ signals 'partial'.
∇f	grad f / del f	The gradient: the vector of all partial derivatives. Points in the direction of steepest increase — which is why gradient descent moves in the direction $-\nabla f$.
$\nabla^2 f$	Hessian	The matrix of second partial derivatives. Describes curvature. Also written H .
J	Jacobian	The matrix of first partial derivatives of a vector-valued function.
$\int$	integral	Area under a curve; in probability, how you get from a density to a probability.
$O(\cdot)$	big-O	How cost grows with problem size. $O(n^2)$ means quadrupling when n doubles.

Optimization

Symbol	Read as	What it means
$\min f(x)$	minimize	The smallest value the function attains.
$\operatorname{argmin} f(x)$	arg min	The x that produces that smallest value — the input, not the output. Easily and consequentially confused with min.
argmax	arg max	The input producing the largest value. What a classifier returns when it picks a class.
$\inf / \sup$	infimum / supremum	Greatest lower bound and least upper bound. Used when a minimum or maximum may not actually be attained.
x^*	x star	The optimal solution. $f(x^*)$ is the optimal value.

Symbol	Read as	What it means
$\nabla f(x) = 0$	stationarity	The first-order condition for an unconstrained optimum. Necessary, not sufficient — saddle points satisfy it too.
$\mathcal{L}$	Lagrangian	The objective with constraints folded in via multipliers: $\mathcal{L}(x, \lambda) = f(x) + \lambda g(x)$. Also used for a loss function — check which.
KKT	the KKT conditions	Karush–Kuhn–Tucker: the conditions a constrained optimum must satisfy. The generalization of 'set the derivative to zero'.
s.t.	subject to	Everything following is a constraint.

Probability and statistics

Symbol	Read as	What it means
$P(A)$	probability of A	For an event. Capital P.
$p(x)$	density or mass at x	Lowercase p is a function over values. For continuous x this is a density and can exceed 1.
$P(A \mid B)$	probability of A given B	Conditional probability. The bar reads 'given'.
$E[X]$	expectation	The long-run average value of X. Often printed in blackboard-bold as a double-struck E, or written μ .
$\text{Var}(X)$	variance	$E[(X - \mu)^2]$. Its square root is the standard deviation σ .
$\text{Cov}(X, Y)$	covariance	How two variables move together. The matrix version is Σ .
$X \sim D$	X is distributed as D	The tilde assigns a distribution, it does not mean 'approximately'.
$N(\mu, \sigma^2)$	normal distribution	Gaussian with mean μ and variance σ^2 . Note: variance, not standard deviation.
i.i.d.	independent and identically distributed	The assumption that samples are drawn separately from the same distribution. Frequently false, usually assumed.
$\propto$	proportional to	Equal up to a constant factor. Common in Bayes' rule when the normalizer is ignored.
$D_{\text{KL}}(P \parallel Q)$	KL divergence	How much information is lost approximating P by Q. Not symmetric — the order matters.
$H(X)$	entropy	Average uncertainty in X. The quantity a decision tree reduces at each split.

Indexing, sums and products

Where most misreadings actually happen.

Symbol	Read as	What it means
$\sum_{i=1}^n$	sum from i equals 1 to n	Add up the expression for each value of i. The index below, the limit above.
$\prod$	product	The same, but multiply.
x_i	x sub i	Ambiguous by itself: either the i-th component of a single vector x, or the i-th vector in a collection. Context, and only context, resolves it.
$x^{(i)}$	x superscript i, parenthesized	The i-th training example. The parentheses exist specifically to distinguish this from a power.
$x_j^{(i)}$	combined	The j-th feature of the i-th training example.
$W^{[l]}$	square-bracketed superscript	The weights of layer l in a neural network.
θ_t	theta at t	The parameter values at iteration or time step t.
$1\{\text{condition}\}$	indicator function	1 when the condition holds, 0 otherwise. Turns a logical test into arithmetic. Usually printed as a blackboard-bold 1, and sometimes written $I\{\cdot\}$.

Machine learning conventions

Symbol	Read as	What it means
X	design matrix	Training inputs, one row per example, one column per feature. Usually $n \times d$.
y	targets	The true labels or values, one per example.
$\hat{y}$	y-hat	The prediction. A hat means estimated rather than observed.
n, d	sample and feature counts	n examples, d features. Some texts use m and n — check.
$h(x), f(x; \theta)$	hypothesis / model	The function mapping input to prediction, parameterized by θ .
$\mathcal{L}, J$	loss / cost	What training minimizes. $\mathcal{L}$ often per-example, J often averaged over the dataset.
w, b	weights and bias	The linear parameters. Together they are usually what θ refers to.
$\sigma(z)$	sigmoid	$1/(1 + e^{-z})$. Squashes any real number into (0, 1).
softmax	softmax	Turns a vector of scores into a probability distribution that sums to 1.
$\phi(x)$	feature map	The transformation into a space where the problem becomes linearly separable.
$k(x, x')$	kernel	The inner product in that transformed space, computed without performing the transformation.

Symbol	Read as	What it means
D	dataset	The collection $\{(x^{(i)}, y^{(i)})\}$ of labelled examples. Usually set in a script or calligraphic D to distinguish it from a matrix.

Population and metaheuristic notation

Notation specific to search methods — grey wolf, particle swarm, ant colony, differential evolution, simulated annealing, genetic algorithms.

Symbol	Read as	What it means
$x_i(t)$	position of individual i at iteration t	The standard way to write a candidate solution in a population method. t indexes iterations, not time.
$v_i(t)$	velocity	Particle swarm only. The step a particle carries into the next iteration.
p_i, g	personal best, global best	Particle swarm. p_i is where particle i has personally done best; g is the swarm's best.
w, c_1, c_2	inertia, cognitive, social	Particle swarm coefficients: how much a particle keeps its own momentum, trusts itself, and trusts the swarm.
$X_\alpha, X_\beta, X_\delta$	alpha, beta, delta positions	Grey wolf. The three best solutions found so far, which every other wolf moves toward.
A, C	coefficient vectors	Grey wolf. A controls how far a wolf may overshoot; C perturbs the target position.
a	control parameter	Grey wolf. Decreases linearly from 2 to 0 across the run — the single number that shifts the search from exploration to exploitation.
τ_{ij}	pheromone on edge (i, j)	Ant colony. The colony's shared memory of which routes worked.
η_{ij}	heuristic desirability	Ant colony. A problem-specific hint, typically the inverse of distance.
ρ	evaporation rate	Ant colony. How fast pheromone decays. Effectively the forgetting rate, and the main exploration control.
$T, \Delta E$	temperature, energy change	Simulated annealing. A worse move by ΔE is accepted with probability $\exp(-\Delta E/T)$.
F, CR, NP	scaling factor, crossover rate, population size	Differential evolution. F scales the difference vector that drives mutation.
p_c, p_m	crossover and mutation probability	Genetic algorithm.

Ten confusions worth memorizing

Almost every notation misreading in practice is one of these.

The trap	What is actually going on
A superscript can be a power or an index.	x^2 is x squared. $x^{(2)}$ — with parentheses — is the second training example. The parentheses are the only signal, and they are easy to miss.
Σ is two different things.	With limits above and below, it is a summation. Standing alone, bold, it is a covariance matrix.
$\cdot$ is three different things.	Absolute value of a number, cardinality of a set, determinant of a matrix. Read what is inside the bars.
$\mathbf{x} \in \mathbb{R}^n$ is a column by default.	This is why $\mathbf{x}^T \mathbf{y}$ is a scalar and $\mathbf{xy}^T$ is an $n \times n$ matrix. Getting this backwards produces dimension errors that look like bugs in your code.
log means natural log in machine learning.	Base e, essentially always — except in information theory, where entropy is usually base 2. Neither field states which.
α is a learning rate, a wolf, and a significance level.	Three fields, three meanings, sometimes in one literature review. The same is true of β , δ , λ and ρ . Never carry a symbol's meaning across papers.
∇f and $\partial f / \partial \mathbf{x}$ are the same information.	The gradient is just the partials collected into a vector. If ∇ looks unfamiliar, expand it into the list of partials you already understand.
A hat means estimated.	$\hat{y}$ is what the model predicts; y is what actually happened. $\hat{\theta}$ is a parameter estimated from data rather than a true underlying value.
argmin is not min.	min returns the lowest value; argmin returns the input that produced it. Confusing them is the most consequential notation error in optimization, because the two are different types of object.
A tilde assigns a distribution.	$X \sim N(0, 1)$ reads 'X is distributed as', not 'X is approximately'. Approximation is $\approx$.

A closing note. Notation is a compression scheme, not a test. Every symbol here stands for an operation you could carry out by hand on a small enough example — and the fastest way to make an unfamiliar equation concrete is to do exactly that: three data points, two dimensions, one iteration, on paper. If the arithmetic works, the notation was never the obstacle.

Free worked examples, complete with the hand calculations, at computationallabs.com/resources