

BELLTES ACADEMY

The Core Foundations of AI & ML

A Comprehensive Conceptual Reference Guide

Written by Belltes Khoza

Table of Contents

Topic	Page
1. Linear Algebra Foundations	3
2. Probability & Decision Making	4
3. Gradient Descent Optimization	5
4. Deep Learning & Neural Networks	6
5. Large Language Models (LLMs) & Transformers	7

1. Linear Algebra Foundations

Vectors

Definition: One-dimensional arrays of numbers representing magnitude and direction.

Example: Storing a user's age, income, and credit score as [30, 50000, 720].

Analogy: A GPS coordinate pointing exactly where to walk from a baseline position.

Matrices

Definition: Two-dimensional grids of numbers mapping multi-featured transformation spaces.

Example: An image grid of pixels where rows and columns store raw brightness variables.

Analogy: A giant Excel spreadsheet where every cell is wired to compute together.

Matrix Multiplication

Definition: The combined transformation operation mapping features from one space into another.

Example: Multiplying input features by structural weight parameters to calculate a raw layer score.

Analogy: A factory assembly system where components slide through specialized geometric filters.

Eigenvalues & Eigenvectors

Definition: Axes along which a linear transformation only scales without changing spatial direction.

Example: Used in Principal Component Analysis to find the core direction of maximum data spread.

Analogy: A wind tunnel testing a car; the main structural frame stays straight while wind scales past it.

Singular Value Decomposition (SVD)

Definition: Matrix factorization method breaking a dataset into core geometric components.

Example: Compressing heavy image files or driving low-rank recommendation system engines.

Analogy: Deconstructing a musical chord into its individual distinct elemental notes.

Dot Product

Definition: The sum of the products of corresponding elements across two equal sequences.

Example: Measuring raw item similarity scores within vector database searches.

Analogy: Checking how perfectly a custom template overlaps with a drawing pattern.

Determinants

Definition: A scaling factor representing how much a matrix changes spatial volume sizes.

Example: Testing whether a matrix inversion setup is computationally stable or completely impossible.

Analogy: A multiplier indicating how much a balloon expands or deflates when squeezed.

Vector Spaces

Definition: A clean mathematical closing space where vectors can be added and scaled properly.

Example: The underlying multi-dimensional semantic mapping space where word embeddings exist.

Analogy: A flat canvas where you can paint anywhere without running off the borders.

Orthogonality

Definition: The state where two vectors are strictly perpendicular, showing zero redundant correlation.

Example: Creating completely independent feature baselines to eliminate data overlap errors.

Analogy: Two lines on a floor meeting at exactly 90 degrees, moving in totally separate directions.

Inverse Matrices

Definition: A matrix that perfectly reverses the transformation steps of an original target matrix.

Example: Solving exact weight coefficients instantly within classical closed-form linear systems.

Analogy: Unlocking a padlock using the exact matching reverse key turn.

2. Probability & Decision Making

Bayes' Theorem

Definition: A formula updating the probability of a hypothesis as fresh validation evidence arrives.

Example: Filtering spam emails based on occurrences of highly suspicious target phrases.

Analogy: Updating your weather guess from sunny to rainy the second you hear thunder outside.

Conditional Probability

Definition: The likelihood of an outcome occurring given that another prior event is already true.

Example: Predicting a user buys a specific phone case given they just bought the base phone.

Analogy: The chance of slipping on ice given that it is currently freezing winter.

Probability Distributions

Definition: Mathematical functions showing the relative likelihood of tracking different possible outcomes.

Example: Mapping natural human height variations across a bell-shaped Gaussian distribution curve.

Analogy: A chart showing which slots on a roulette wheel collect the most dropped chips.

Expectation Value

Definition: The long-term average value of a random variable across infinite repeated trials.

Example: Calculating the mathematical risk weight of an algorithm approving a credit application.

Analogy: The average amount of cash you expect to win or lose per pull on a slot machine.

Maximum Likelihood Estimation

Definition: A strategy finding parameter values that maximize the probability of observing current data.

Example: Fitting a clean logistic regression line tightly across collected patient data points.

Analogy: Adjusting a camera focus ring until the blurred image looks as sharp as possible.

Entropy

Definition: The quantifiable measure of disorder, uncertainty, or random surprise in a dataset.

Example: Evaluating splits inside a Decision Tree to find the cleanest classification step.

Analogy: A box of perfectly sorted colorful beads vs a messy box shaken up completely.

Cross-Entropy Loss

Definition: A metrics penalty score measuring the divergence between true targets and predicted probabilities.

Example: The primary error function optimized during deep image classification training loops.

Analogy: A penalty fee charged for every single millimeter your dart misses the exact bullseye.

Markov Chains

Definition: Stochastic models where the next state relies strictly on the current present state.

Example: Predicting autocomplete text structures word-by-word based on the immediate word typed.

Analogy: A frog hopping across leaves, looking only at its current leaf to pick the next leap.

Joint Probability

Definition: The absolute likelihood of two independent events occurring simultaneously together.

Example: Calculating the chance that a visitor is both under 25 and browsing from a mobile device.

Analogy: The odds that it is simultaneously raining outside while your watch alarm sounds.

Central Limit Theorem

Definition: The rule stating sample means approximate a normal distribution as sample sizes scale up.

Example: Justifying the use of Gaussian error structures across complex real-world data tracking.

Analogy: A messy pile of sand naturally forming a smooth, predictable cone when poured out.

3. Gradient Descent Optimization

Cost Function

Definition: A mathematical formula measuring how poorly a model is performing against targets.

Example: Using Mean Squared Error to evaluate house price predictions against sale realities.

Analogy: A score counter tracking how many structural errors an apprentice makes during a build.

The Gradient

Definition: The vector vectoring towards the sharpest directional ascent of a multi-variable function.

Example: Computing partial derivatives across layer weights to find direction changes.

Analogy: A compass needle pointing directly up towards the peak of a steep mountain ridge.

Learning Rate

Definition: A tuning multiplier controlling the step size taken towards the error minimum.

Example: Setting an optimal scalar value like 0.001 to keep optimization runs steady.

Analogy: The stride size you take while hiking down a mountain path in pitch black darkness.

Stochastic Gradient Descent

Definition: Updating internal weights using only a single randomly selected training sample at a time.

Example: Training huge baseline models efficiently over massive streaming online datasets.

Analogy: A hiker guessing the way down a foggy hill by feeling the ground with one foot step.

Mini-Batch Gradient Descent

Definition: An optimization variant grouping training data into small batches for weight updates.

Example: Processing 32 or 64 images at once to maximize modern GPU matrix acceleration benefits.

Analogy: A team of scouts checking 32 local paths before agreeing on a collective direction down.

Local Minima

Definition: A point where the error looks lowest locally but is higher than the absolute global lowest point.

Example: A neural network getting stuck in a sub-optimal setup early during training runs.

Analogy: Getting trapped in a small valley puddle halfway up a mountain, thinking you reached bottom.

Momentum

Definition: An optimization enhancer accelerating updates by factoring in prior step directions.

Example: Using velocity histories to roll safely past noisy flat spots in loss landscapes.

Analogy: A heavy marble rolling down a bowl, gaining speed to clear small bumps easily.

Adam Optimizer

Definition: An adaptive algorithm calculating custom learning rates for every unique individual parameter.

Example: Standard default optimizer used to train modern deep architectures efficiently.

Analogy: A car with smart independent suspension adjusting every wheel to fit the terrain perfectly.

Vanishing Gradient

Definition: The issue where gradients shrink exponentially during backpropagation, stalling updates.

Example: Early deep networks failing to learn feature steps inside their front layer sequences.

Analogy: A long whisper chain where the original message completely fades out by person ten.

Exploding Gradient

Definition: The condition where error gradients scale exponentially, causing unstable weight collapses.

Example: Model parameters hitting numerical overflow errors, returning ruined NaN training results.

Analogy: A microphone feedback loop getting louder and louder until the speakers pop completely.

4. Deep Learning & Neural Networks

Artificial Neurons

Definition: The foundational node computing weighted feature sums before passing them through activation.

Example: A basic node layer element receiving token numbers to output a firing value.

Analogy: A small water valve that opens only when the incoming water pressure hits a limit.

Activation Functions

Definition: Mathematical mappings injecting non-linear properties into neural network paths.

Example: Utilizing Rectified Linear Units (ReLU) to clip negative feature values cleanly to zero.

Analogy: A light switch that clicks on only after toggle pressure crosses a specific point.

Backpropagation

Definition: The foundational algorithm using the chain rule to distribute error blame backwards.

Example: Calculating layer weight adjustments immediately following a bad output estimation.

Analogy: A director correcting the lighting crew because a stage scene looked too dark.

Convolutional Layers

Definition: Specialized feature filters sliding across inputs to extract localized spatial patterns.

Example: Detecting edges, corners, and textures inside early computer vision image tasks.

Analogy: Moving a small magnifying glass across a map to search exclusively for crossroad icons.

Pooling Layers

Definition: Downsampling operations reducing feature map dimensions to achieve spatial invariance.

Example: Using Max Pooling to extract the strongest activation signal from a 2x2 patch region.

Analogy: Summarizing a detailed text page into a single short sentence of main points.

Recurrent Connections

Definition: Architectural loops passing past activation states back into current data steps.

Example: Processing streaming sensor values or text tokens step-by-step over historical timelines.

Analogy: Keeping the plot of the previous chapter in mind while reading the current sentence.

Overfitting

Definition: When a network memorizes random training noise instead of learning general patterns.

Example: A model score hitting 100% on training data but dropping to 50% on test verification data.

Analogy: A student memorizing exact practice questions instead of learning the core formulas.

Dropout

Definition: A regularization technique randomly muting a percentage of neurons during training runs.

Example: Turning off 20% of layer nodes dynamically to build resilient multi-path networks.

Analogy: Forcing different workers to rotate shifts so no single person holds a monopoly on info.

Batch Normalization

Definition: Scaling activation outputs to zero mean and unit variance across mini-batches.

Example: Stabilizing deep layers to allow faster optimization speeds and higher learning rates.

Analogy: Adjusting everyone's walkie-talkies to the same volume baseline to avoid sudden screams.

Latent Space

Definition: A compressed hidden layer space where similar conceptual data points cluster together.

Example: The inner bottleneck layer of an autoencoder mapping images to small feature codes.

Analogy: Organizing library books by abstract genres instead of sorting strictly by jacket colors.

5. Large Language Models (LLMs) & Transformers

Self-Attention

Definition: A mechanism calculating mathematical relevance scores between all tokens in a sequence.

Example: Mapping that 'bank' relates to 'river' instead of 'money' based on surrounding words.

Analogy: A reader highlighting related nouns and verbs across a paragraph with colorful markers.

Tokenization

Definition: The process of splitting raw string sentences into small structured numeric pieces.

Example: Converting the phrase 'AI Models' into discrete dictionary ID numbers like [4502, 912].

Analogy: Slicing a long Lego structure into its individual block pieces to count them up.

Positional Encoding

Definition: Adding distinct mathematical vectors to embeddings to give tokens spatial sequence context.

Example: Letting a transformer process words in parallel while preserving exact word order meanings.

Analogy: Stamping an explicit delivery sequence number onto boxes loaded onto a truck together.

Multi-Head Attention

Definition: Running several attention loops in parallel to track diverse semantic relationships simultaneously.

Example: Tracking subject-verb links, tense changes, and pronouns all at the same moment.

Analogy: A film crew where one camera tracks the actor, another the background, and another lighting.

Retrieval-Augmented Generation

Definition: Querying external databases to inject factual real-time text context into prompts.

Example: Looking up a user's account history balance file before answering a customer support bot question.

Analogy: An open-book exam where you consult an encyclopedia before writing down your answer.

Vector Databases

Definition: Storage engines optimized for ultra-fast nearest-neighbor searches across multi-dimensional spaces.

Example: Indexing text chunk embeddings into a Pinecone instance for contextual retrieval.

Analogy: A massive warehouse where items with similar shapes are stored right next to each other.

Parameter-Efficient Fine-Tuning

Definition: Adapting massive models by training a small subset of extra auxiliary parameters.

Example: Using LoRA layers to fine-tune a 70-billion parameter model on consumer-grade hardware.

Analogy: Placing a small customized adapter plug on a standard garden hose to fit a special sink nozzle.

Temperature Scaling

Definition: A parameter adjusting the randomness of next-token generation probability maps.

Example: Setting temperature to 0.0 for deterministic coding tasks or 0.8 for creative writing.

Analogy: A dial controlling whether an artist sticks to the script or improvises wildly.

Context Window

Definition: The maximum sequence length limit an LLM can process in a single inference call.

Example: An engine comfortably ingestion 128,000 tokens of raw legal text files at once.

Analogy: The physical size of a desk determining how many open documents you can read at once.

Reinforcement Learning from Human Feedback

Definition: Aligning model generation styles with human preferences using reward scoring setups.

Example: Fine-tuning base models to avoid generating dangerous instructions or toxic text loops.

Analogy: A dog trainer giving treats when the puppy sits politely instead of barking wildly.