■ Day 2 – Data Science Math Gym

Gradient Descent & KL Divergence (Feynman Style + Visuals)

■ Concept 1: Gradient Descent

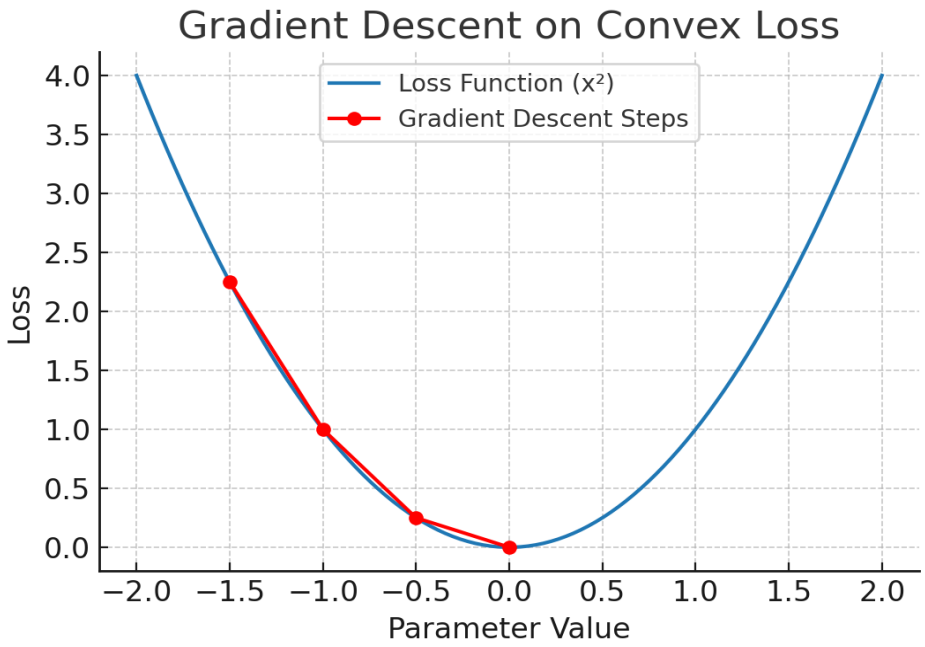
Gradient Descent is like rolling a ball downhill — the ball always rolls in the steepest direction. In machine learning, this "hill" is our loss function. Each step we take is guided by the derivative (gradient), telling us where to move to reduce the loss. A small step (learning rate) avoids overshooting, and each iteration gets us closer to the optimal model. This is how models like linear and logistic regression learn during training.

■ Concept 2: KL Divergence

KL Divergence measures how different two probability distributions are. It’s like asking, “How surprised would I be if I thought the world looked like Q, but it’s actually P?” KL(P■Q) = 0 means no surprise — they match. But the higher the KL divergence, the more "informational distance" there is between the truth (P) and your model (Q). It's widely used in variational inference, language modeling, and more.

■ Visual Examples

■ Gradient Descent Optimization Path



■ KL Divergence Between Two Distributions