

The Quantitative Toolkit

Study Guide

by Richard J. Young, Ph.D.

richard@deepneuro.ai

A cross-cutting concept guide · The NeuroAI Handbook

The four math pillars the whole book stands on: information, inference, models, and cause.

One-sentence thesis. Four questions run underneath every chapter of this book, how much a signal tells you, how a belief should change, how a hypothesis becomes a testable model, and whether acting on it would change anything, and information theory, Bayesian inference, the GLM, and causal inference are the tools that answer them.

PART 1: Information Theory – How Much Does a Signal Tell You?

Information Is Surprise

Start with the question, not the equation. A message that you could have predicted (the sun rose today) tells you almost nothing; a message you could not (I won the lottery) tells you a great deal. Claude Shannon's 1948 insight was that this intuition can be measured: information is the resolution of uncertainty, and its unit is the bit. Shannon's **entropy** is the average surprise of a source, and it has a beautifully concrete reading, the minimum number of yes/no questions you need on average to pin down the outcome. A fair coin is **1 bit** (one question), a fair eight-sided die is 3 bits, an English letter about 4.2 bits. Entropy is largest when every outcome is equally likely and drops to zero when one outcome is certain.

In symbols, for outcomes with probabilities $p(x_i)$:

$$H(X) = - \sum_i p(x_i) \log_2 p(x_i)$$

Shared Knowledge and the Cost of Being Wrong

Two more quantities do most of the work in the book. **Mutual information** $I(X; Y)$ is the overlap of two uncertainties, how much learning Y reduces your surprise about X . It runs from 0 (independent) up to $\min(H(X), H(Y))$, and unlike correlation it catches nonlinear dependence. **KL divergence** $D_{KL}(P \parallel Q)$ is the penalty in extra bits for modeling the world with the wrong distribution Q when the truth is P ; it is not symmetric.

$$I(X; Y) = H(X) - H(X \mid Y)$$

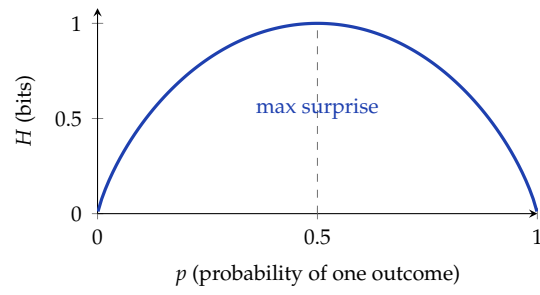
Watch out. These are not interchangeable with the statistics you already know. Two variables can share high mutual information while showing near-zero linear correlation, so a Pearson r of 0 does not license the word “independent.”

ANALOGY

KL divergence is a tourist navigating with a slightly wrong map Q of a city whose true layout is P . The divergence is the average number of extra wrong turns the bad map costs. Swap the maps and the cost changes, which is why $D_{KL}(P \parallel Q) \neq D_{KL}(Q \parallel P)$.

Why Brains and Nets Both Care

This is not decoration; it is the connective tissue of the book. Training a neural network by cross-entropy loss is exactly minimizing D_{KL} between the model's predictions and the data. Horace Barlow's **efficient coding hypothesis** (1961) says sensory systems evolved to strip redundancy and maximize information under a tight energy budget, which shows up as sparse coding and as predictive coding (transmit only the prediction error). Tishby's **information bottleneck** frames a deep network's layers as compressing the input while keeping what predicts the target. A single neuron carries only **2 to 3 bits** per spike, so the optic nerve's million axons are a genuine channel-capacity bottleneck, $C = \max_{p(x)} I(X; Y)$.



Binary entropy: uncertainty peaks at $p = 0.5$ (a coin you cannot call) and vanishes at $p = 0$ or 1 (a foregone conclusion).

The World Is Ambiguous

You see a shadowy figure in the dark. Is it a person, a tree, or a trick of the light? The same sensation can arise from different causes, sensory signals are corrupted by noise, and you never have complete information. The **Bayesian brain hypothesis** proposes that the brain copes by holding beliefs as probability distributions and updating them the moment evidence arrives. The engine of that update is Bayes' rule, which combines what you already knew (the prior) with what the data say (the likelihood) into a revised belief (the posterior).

$$P(H | D) = \frac{P(D | H) P(H)}{P(D)}$$

The prior dominates when data are scarce and is overwhelmed as evidence piles up. **Conjugate priors** (Beta-Binomial, Gamma-Poisson for spike counts, Normal-Normal) keep the posterior in the same family as the prior, which is what makes sequential updating cheap.

Weigh Evidence by Its Reliability

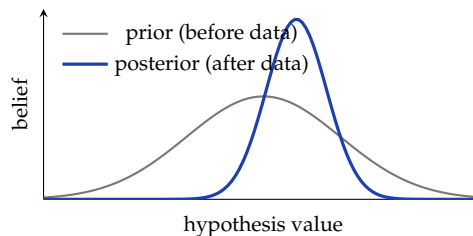
The framework makes a sharp, testable prediction about perception. When the brain fuses two noisy

cues, it should not average them equally; it should weight each by its **precision** (inverse variance), letting the more reliable cue speak louder.

$$\hat{x} = w_1 x_1 + w_2 x_2, \quad w_i \propto \frac{1}{\sigma_i^2}$$

Ernst and Banks (2002) showed humans combine visual and haptic size cues almost exactly this way. To turn a belief into an action, you maximize **expected utility** $\mathbb{E}[U(a)] = \sum_s U(s, a) P(s | D)$, weighting each outcome by its posterior probability. The **drift-diffusion model**, evidence accumulating to a threshold, captures both choice and reaction time and matches firing in areas LIP and FEF (Gold & Shadlen, 2007).

Watch out. Exact Bayesian inference is intractable for real problems, so the brain approximates, and people visibly deviate, neglecting base rates. On a condition with **1 percent** prevalence, a test that is 95 percent sensitive and 90 percent specific still leaves a positive result more likely wrong than right, because the prior is doing most of the work.



Bayesian updating: evidence pulls the belief toward the data and sharpens it. A broad prior becomes a narrow, confident posterior.

Fit, but Do Not Fool Yourself

You have a hypothesis about how a system works, say that a neuron's firing rate depends on stimulus contrast. Model fitting is the discipline of turning that hypothesis into numbers you can test: choose a model (the hypothesis), choose a cost function (how wrong a fit is), fit it (optimize the parameters), and, crucially, evaluate how well it predicts data it never saw. **Linear regression**, $y = \mathbf{X}\beta + \epsilon$ with Gaussian noise, is the starting point and is fine for something like an fMRI BOLD signal. But it is wrong for spike counts, which are non-negative integers, not bell-curved.

One Framework, Many Data Types

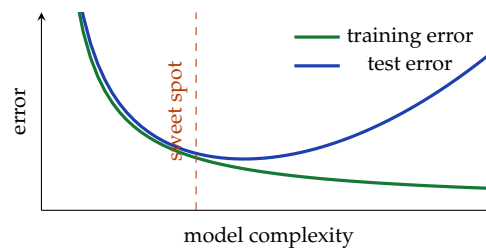
The **generalized linear model** fixes this with three parts: a linear predictor $\eta = \mathbf{X}\beta$, a link function $g(\mu) = \eta$, and a noise family matched to the data. Choose Gaussian with an identity link for continuous data, Bernoulli with a logit link for choices, and Poisson with a log link for spikes, which is the workhorse encoding model of systems neuroscience. Parameters are found by **maximum likelihood**, the values that make the observed spikes most probable.

$$\eta = \mathbf{X}\beta, \quad g(\mu) = \eta$$

The Overfitting Trap

The central danger is the **bias-variance tradeoff**. Too simple a model underfits (high bias); too flexible a model chases noise and overfits (high variance). Training error always falls as you add complexity, so it cannot tell you when you have gone too far, only **cross-validation** on held-out data can. **Regularization** buys back generalization by penalizing extreme weights: L2 (ridge) shrinks all coefficients, L1 (lasso) drives some to exactly zero and so selects features. **AIC** and **BIC** ($\text{BIC} = k \ln n - 2 \ln \hat{L}$) trade fit against parameter count, lower is better.

Watch out. There is a hidden bridge to Part 2: L2 regularization is exactly a Gaussian prior on the weights and L1 is a Laplace prior. Regularization is Bayes in disguise, so “frequentist” model fitting already smuggles in prior beliefs.



Training error keeps falling with complexity; test error turns back up. Only the held-out curve reveals overfitting, so never trust training fit alone.

Seeing Is Not Doing

Two neurons, one in visual cortex and one in motor cortex, fire together trial after trial. Does the visual neuron drive the motor one? Not necessarily, a presented stimulus may drive both, a **confounder** that manufactures the correlation while no arrow runs between the cells. This is the fault line under the whole book: every supervised model in Part 3 learns association, and association answers “what tends to happen,” not “what happens if I intervene.” Judea Pearl’s ladder names three rungs, association (seeing), intervention (doing), and counterfactual (imagining), and the point is that seeing and doing are different quantities. The $do(\cdot)$ operator marks an intervention that surgically cuts the arrows into X .

$$P(Y | do(X=x)) \neq P(Y | X=x)$$

Reading a Graph, Running an Intervention

A directed acyclic graph makes the assumptions explicit, and three shapes decide what you may adjust for. In a **chain** ($X \rightarrow Z \rightarrow Y$) and a **fork/confounder** ($X \leftarrow Z \rightarrow Y$), conditioning on Z blocks the path; on

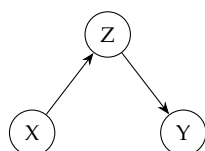
a **collider** ($X \rightarrow Z \leftarrow Y$), conditioning on Z opens a spurious one. In neuroscience the gold-standard intervention is **optogenetics**, using light to force a defined population on or off, a biological randomized trial that realizes the do-operator; its AI analog is **network ablation**, zeroing a unit to measure its causal contribution. When you cannot intervene, **instrumental variables** (a genetic variant affecting dopamine, say), regression with confounder control, and causal-discovery algorithms like **PC** recover effects from observational data.

Watch out. Adjusting for confounders only works for the confounders you measured; an unmeasured one biases the estimate in an unknown direction. And a causal-discovery algorithm outputs a hypothesis to test, never a proof.

ANALOGY

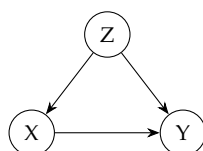
A rooster crows every morning just before dawn, perfectly correlated with sunrise. Silence the rooster (intervene) and the sun still rises. Correlation gave you the schedule; only the intervention

reveals there was never an arrow from bird to star.



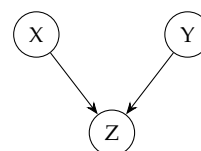
Chain

condition to block



Fork (confounder)

adjust for Z



Collider

never condition on Z

The three DAG shapes that govern adjustment. Conditioning on Z blocks a chain or fork but opens a collider, inventing a correlation from nothing.

Pillar	Question it answers	Neuro example	AI example
Information theory	How much does a signal carry?	Bits per spike; efficient coding	Cross-entropy loss; info bottleneck
Bayesian inference	How should belief change?	Cue integration; Bayesian brain	Priors; probabilistic models
GLM / fitting	What is the testable model?	Poisson GLM for spike trains	Regression, regularization, CV
Causal inference	Would acting change Y ?	Optogenetic intervention	Network ablation; IV, PC

Quick Self-Test

1. A colleague reports two brain signals have Pearson correlation near zero and concludes they are independent. Why is that conclusion unsafe, and what would you compute instead?
2. In one sentence, what does training a network by cross-entropy loss minimize, and what does that connect it to?
3. When the brain fuses a reliable visual cue and a noisy haptic cue, how should it weight them, and whose experiment confirmed this?
4. Training error keeps dropping as you add polynomial terms. Why does that not mean the bigger model is better, and what do you check instead?
5. State the difference between $P(Y | X)$ and $P(Y | do(X))$, and name the biological technique that realizes the do-operator.
6. You adjust for every confounder you measured and still worry the estimate is biased. What single unstated assumption could be failing?

If You Remember One Thing

These four pillars are one continuous argument, not four separate topics. Information theory says what a signal can carry, Bayesian inference says how to update belief about it, the GLM turns that belief into a model you can fit without fooling yourself, and causal inference tells you whether acting on the model would change anything. The seams between them are where the insight lives: cross-entropy is KL divergence, regularization is a prior, and an accurate predictive model is still the wrong tool for a question about intervention.

Richard J. Young, Ph.D. · richard@deepneuro.ai · Senior AI Research Scientist, UnitedHealth Group · UNLV Lee Business School · ORCID 0000-0002-1109-7552