Probability & Statistics for Machine Learning

A Complete, Step-by-Step Treatment for the Quant / ML Researcher

Every result derived; every step justified; worked examples throughout

Abstract

Machine learning is applied probability: a model is a family of distributions, training is choosing one by an estimation principle, and prediction is computing an expectation. This document builds the probability and statistics that underpin learning, from the axioms to the limit theorems to the estimation theory, with the same discipline as its companions: *nothing asserted without explanation.* We prove why variance decomposes the way it does, why the central limit theorem makes Gaussians ubiquitous, why maximum likelihood is equivalent to minimizing KL divergence (and hence to cross-entropy loss), why a biased estimator can beat an unbiased one, and why multiple testing silently inflates false discoveries — the statistical disease at the heart of backtest overfitting. Worked examples and explicit ML/finance connections accompany every concept.

Contents

1 Orientation: Why Probability Is the Grammar of ML

Three roles probability plays, which structure this document:

- **Modeling uncertainty.** Data are noisy; a model that predicts a single number throws away the uncertainty that risk management and decision-making need. Probabilistic models predict *distributions*, and the language for that is random variables, densities, and expectations (Sections 2–5).
- **Justifying estimators.** Why does minimizing squared error make sense? Because it is maximum likelihood under Gaussian noise. Why does cross-entropy appear everywhere? Because it is the KL divergence between the data and the model. The estimation theory of Sections 7–8 supplies these justifications.
- **Quantifying confidence and avoiding self-deception.** How sure are we? The limit theorems (Section 6) and hypothesis-testing machinery (Section 9) answer this — and warn us, via multiple testing, about the ways we fool ourselves, which in quant finance is the difference between a real edge and an overfit backtest.

Intuition

A useful unifying view: a supervised model defines a conditional distribution $p_{\boldsymbol{\theta}}(y \mid \mathbf{x})$, and learning picks $\boldsymbol{\theta}$ to make the observed data probable. Regression assumes $y \mid \mathbf{x}$ is Gaussian; classification assumes it is Bernoulli/categorical. Once you see "the loss" as "the negative log-likelihood of an assumed noise model," the menagerie of loss functions collapses into one idea, and the role of probability becomes the organizing principle rather than a side topic.

2 Foundations: the Probability Space

2.1 Why we need a careful setup

It is tempting to define probability informally as "long-run frequency" or "degree of belief," but both run into trouble with infinite or continuous sample spaces (what is the probability a real-valued return is *exactly* 0.5? Zero, yet it can happen). The measure-theoretic foundation resolves this cleanly and is worth stating, because it underlies the rigorous treatment of continuous distributions, conditional expectation, and the convergence theorems that justify Monte Carlo and empirical risk minimization.

Definition 1 (Probability space)**.** A probability space is a triple $(\Omega, \mathcal{F}, \Pr)$: a sample space Ω of outcomes; a σ-algebra $\mathcal{F}$ of events (subsets of Ω closed under complement and countable union, containing Ω); and a probability measure $\Pr : \mathcal{F} \to [0,1]$ with $\Pr(\Omega) = 1$ and countable additivity — for disjoint events $A_1, A_2, \ldots$, $\Pr(\bigcup_i A_i) = \sum_i \Pr(A_i)$.

The σ-algebra is the collection of events we can assign probabilities to; countable additivity is the engine behind every limiting argument. From these axioms, the familiar rules follow as theorems, not assumptions: $\Pr(A^c) = 1 - \Pr(A)$ (since A and A^c are disjoint and union to Ω), monotonicity ($A \subseteq B \Rightarrow \Pr(A) \le \Pr(B)$), and inclusion–exclusion.

2.2 Random variables as measurements

Definition 2. A random variable is a (measurable) function $X : \Omega \to \mathbb{R}$ that assigns a number to each outcome.

The point of this definition is that it separates the underlying randomness ($\omega \in \Omega$) from the quantity we observe ($X(\omega)$). We rarely work with Ω directly; instead we work with the *law* (distribution) of X, summarized by the cumulative distribution function $F_X(x) = \Pr(X \le x)$, which

always exists. When F_X has a derivative we get a probability density $f_X = F_X'$ (continuous case); when X takes discrete values we get a probability mass function. The CDF is the unifying object because it exists in both cases and even for mixtures.

2.3 Conditioning and independence

Definition 3. $\Pr(A \mid B) = \Pr(A \cap B)/\Pr(B)$ (for $\Pr(B) > 0$) is the conditional probability of A given B. Events are independent if $\Pr(A \cap B) = \Pr(A)\Pr(B)$, equivalently $\Pr(A \mid B) = \Pr(A)$ — knowing B does not change the probability of A.

Conditioning is *learning*: it updates probabilities in light of information. Independence is the absence of informational content. The subtle and important refinement for ML is **conditional independence**: $X \perp Y \mid Z$ means $\Pr(X, Y \mid Z) = \Pr(X \mid Z)\Pr(Y \mid Z)$. This can hold even when X and Y are marginally dependent (they share a common cause Z; once Z is known they carry no further information about each other). Conditional independence is the structural assumption behind Naive Bayes and the entire theory of graphical models.

Worked Example

Two stocks' daily returns X, Y may be correlated (marginally dependent) simply because both respond to the market Z. Conditional on the market return, they might be independent: $X \perp Y \mid Z$. A naive model that ignores Z sees spurious dependence; a model that conditions on the common factor explains it away. This is exactly why factor models "residualize" returns against common factors before modeling idiosyncratic behavior — they are exploiting conditional independence.

2.4 Bayes' theorem: the logic of updating

Theorem 1 (Bayes)**.** $\Pr(\theta \mid x) = \dfrac{\Pr(x \mid \theta)\Pr(\theta)}{\Pr(x)}$, *i.e. posterior* $\propto$ *likelihood* $\times$ *prior.*

Proof. Both $\Pr(\theta \mid x)\Pr(x)$ and $\Pr(x \mid \theta)\Pr(\theta)$ equal $\Pr(\theta \cap x)$ by the definition of conditional probability. Equate them and divide by $\Pr(x)$. □

Trivial to prove, profound in use. It is the unique consistent way to update beliefs given evidence: start with a prior $\Pr(\theta)$, observe data x whose likelihood under each hypothesis is $\Pr(x \mid \theta)$, and obtain the posterior $\Pr(\theta \mid x)$. The denominator $\Pr(x) = \int \Pr(x \mid \theta)\Pr(\theta)\, d\theta$ (the *evidence*) is just the normalizer ensuring the posterior integrates to one. This single identity organizes the Naive Bayes classifier, Bayesian parameter estimation, Kalman filtering (Gaussian Bayes updates), and belief revision in trading.

The base-rate trap

A test for a rare condition (prevalence 1%) is 99% accurate. You test positive. What is the probability you have it? Bayes: $\Pr(\text{sick} \mid +) = \frac{0.99 \times 0.01}{0.99 \times 0.01 + 0.01 \times 0.99} = \frac{0.0099}{0.0198} = 0.5$. Only 50%, not 99%, because the rare prior pulls the posterior down — false positives from the large healthy population rival the true positives. This is *the* reason accuracy misleads on imbalanced classes, and why a quant "signal" that fires rarely needs a very low false-positive rate to be worth acting on.

3 Expectation, Variance, and Moments

3.1 Expectation and its linearity

The expectation $\mathbb{E}[X] = \int x\, dF_X(x)$ (sum for discrete, integral for continuous) is the probability-weighted average — the center of mass of the distribution. Its single most useful property:

Theorem 2 (Linearity of expectation)**.** $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$, for any X, Y, *whether or not they are independent.*

The "whether or not independent" is the powerful part: it lets us compute expected values of sums even under arbitrary dependence, which is why it is so often the right first tool. (The proof is linearity of the integral.)

3.2 Variance, covariance, and the diversification identity

Variance measures spread: $\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}X)^2] = \mathbb{E}[X^2] - (\mathbb{E}X)^2$ (the second form by expanding the square and using linearity). Covariance measures co-movement: $\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)] = \mathbb{E}[XY] - \mathbb{E}X\mathbb{E}Y$.

Theorem 3 (Variance of a sum)**.** $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y)$.

Proof. $\mathrm{Var}(X + Y) = \mathbb{E}[((X - \mathbb{E}X) + (Y - \mathbb{E}Y))^2]$; expand the square into $\mathbb{E}[(X - \mathbb{E}X)^2] + \mathbb{E}[(Y - \mathbb{E}Y)^2] + 2\mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)]$, the three terms by definition. □

Intuition

The cross term $2\,\mathrm{Cov}(X, Y)$ is the entire mathematical content of diversification. A portfolio $w_1X + w_2Y$ has variance $w_1^2 \mathrm{Var}\, X + w_2^2 \mathrm{Var}\, Y + 2w_1w_2\,\mathrm{Cov}(X, Y)$. If the assets are negatively correlated ($\mathrm{Cov} < 0$), the cross term *reduces* total variance below the weighted sum of individual variances — risk reduction with no sacrifice of expected return (which adds linearly regardless of correlation). Markowitz portfolio theory is, at its core, the minimization of this quadratic form $\mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w}$, a direct application of the variance-of-a-sum identity to many assets.

Worked Example

Two assets each with variance 1 and correlation ρ. Equal-weight portfolio ($w = 1/2$ each) variance: $\frac{1}{4}(1) + \frac{1}{4}(1) + 2 \cdot \frac{1}{4}\rho = \frac{1+\rho}{2}$. At $\rho = 1$ (identical): variance 1 (no diversification). At $\rho = 0$: variance $1/2$ (halved). At $\rho = -1$: variance 0 (risk eliminated). The benefit of diversification is precisely the gap between $\rho < 1$ and $\rho = 1$ — a quantity invisible without the covariance term.

3.3 The covariance matrix is PSD — revisited probabilistically

For a random vector $\boldsymbol{X}$, the covariance matrix $\boldsymbol{\Sigma} = \mathbb{E}[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^\top]$ collects all pairwise covariances. It is PSD because for any weights $\mathbf{w}$, $\mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w} = \mathrm{Var}(\mathbf{w}^\top \boldsymbol{X}) \geq 0$: the variance of *any* linear combination (portfolio) is nonnegative. A negative eigenvalue would mean a portfolio with negative variance — impossible — which is why estimated covariance matrices that come out non-PSD (from noise, missing data, or more assets than observations) must be repaired before use. This is the probabilistic source of the PSD-ness the linear algebra note treated algebraically.

3.4 Higher moments and why finance cares

Beyond mean and variance: **skewness** (third standardized moment) measures asymmetry, and **kurtosis** (fourth) measures tail heaviness. The Gaussian has skewness 0 and kurtosis 3; financial

returns are typically left-skewed (crashes are sharper than rallies) and leptokurtic (kurtosis > 3, fat tails — extreme moves far more frequent than Gaussian). This single fact invalidates the naive use of Gaussian risk models: value-at-risk computed under normality drastically underestimates tail risk. The moments are estimated from data and used to flag and model these departures.

3.5 The law of total expectation and variance

Theorem 4 (Tower rule and total variance)**.** $\mathbb{E}[X] = \mathbb{E}[\,\mathbb{E}[X \mid Y]\,]$, *and*

$$\operatorname{Var}(X) = \underbrace{\mathbb{E}[\operatorname{Var}(X \mid Y)]}_{\textit{within-group}} + \underbrace{\operatorname{Var}(\mathbb{E}[X \mid Y])}_{\textit{between-group}}.$$

The tower rule says you can average in stages. The total-variance decomposition splits variability into the part unexplained by Y (average within-group variance) and the part explained by Y (variance of group means). This is the probabilistic ancestor of the bias–variance decomposition, of R^2 (explained over total variance), and of the analysis-of-variance that underlies mixed-effects and hierarchical models.

Worked Example

Let returns depend on a hidden regime $Y \in \{\text{calm}, \text{crisis}\}$. Total return variance = average of the within-regime variances (volatility while in each regime) + variance of the regime mean returns (how different the regimes' average returns are). A regime-switching model that ignores Y attributes all variance to noise; decomposing it reveals how much risk is "which regime are we in" versus "noise within a regime" — directly actionable for risk budgeting.

4 Key Distributions and the Exponential Family

4.1 The distributions you must know cold

Each distribution is a model of a particular kind of randomness; knowing which to reach for is half of modeling.

- **Bernoulli**(p): a single yes/no with success probability p; mean p, variance $p(1-p)$ (maximized at $p = 1/2$, zero at the extremes — a certain outcome has no variance). The likelihood model behind logistic regression.
- **Binomial**(n, p): number of successes in n independent Bernoulli trials; mean np, variance $np(1-p)$.
- **Poisson**(λ): counts of rare events in a fixed interval; mean = variance = λ. The limit of Binomial as $n \to \infty$, $p \to 0$, $np \to \lambda$. Models arrivals, jumps, order flow.
- **Gaussian**(μ, σ^2): the bell curve; the limiting distribution of sums (CLT), the maximum-entropy distribution for fixed mean and variance, and the conjugate-friendly default.
- **Exponential**(λ): waiting time until a Poisson event; memoryless ($\Pr(X > s + t \mid X > s) = \Pr(X > t)$ — the past does not matter). Models durations between trades.
- **Student-t_ν**: Gaussian-like but with heavy tails controlled by ν; as $\nu \to \infty$ it becomes Gaussian. The honest default for financial returns.
- **Gamma, Beta**: the conjugate priors for Poisson/exponential rates and Bernoulli/binomial probabilities respectively — they make Bayesian updating closed-form.

4.2 The exponential family: one structure to unify them

Most of the above share a common form. A distribution is in the **exponential family** if its density can be written

$$p(x \mid \boldsymbol{\eta}) = h(x) \exp\big(\boldsymbol{\eta}^\top T(x) - A(\boldsymbol{\eta})\big),$$

where $\boldsymbol{\eta}$ are the natural parameters, $T(x)$ the sufficient statistics, $A(\boldsymbol{\eta})$ the log-partition function (the log-normalizer), and $h(x)$ a base measure. This is not bookkeeping; it has powerful consequences:

- **Sufficient statistics compress the data.** All the information the data carry about $\boldsymbol{\eta}$ is contained in $\sum_i T(x_i)$ — you can discard the raw data and keep only this sum without losing anything (the Fisher–Neyman factorization). For the Gaussian, the sufficient statistics are $\sum x_i$ and $\sum x_i^2$ — the sample mean and variance are all you need.
- **The log-partition generates moments.** $\nabla A(\boldsymbol{\eta}) = \mathbb{E}[T(X)]$ and $\nabla^2 A(\boldsymbol{\eta}) = \mathrm{Cov}[T(X)]$ — differentiating the normalizer yields the mean and covariance. (Proof: differentiate the identity $\int p\,dx = 1$ twice.)
- **It is the backbone of GLMs.** Generalized linear models attach a linear predictor to the natural parameter of an exponential-family response; logistic and Poisson regression are the Bernoulli and Poisson cases.

Worked Example

Write the Bernoulli in exponential-family form: $p(x \mid p) = p^x(1-p)^{1-x} = \exp\big(x \log \frac{p}{1-p} + \log(1-p)\big)$. So the natural parameter is the *log-odds* $\eta = \log \frac{p}{1-p}$, the sufficient statistic is $T(x) = x$, and $A(\eta) = \log(1 + e^\eta)$. The natural parameter being the log-odds is exactly why logistic regression models the log-odds as a linear function of features — the GLM is putting a linear predictor on the natural parameter. The appearance of $\log(1 + e^\eta)$ (the softplus) as the log-partition is why it shows up in the logistic loss.

5 The Multivariate Gaussian

This distribution deserves its own section because it is the workhorse of GMMs, Gaussian processes, Kalman filters, and factor models, and because its remarkable closure properties are what make those methods tractable.

5.1 Definition and geometry

$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ has density

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}(\det \boldsymbol{\Sigma})^{1/2}} \exp\Big(-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\Big).$$

Every piece has meaning. The exponent is a quadratic form $-\frac{1}{2}$ times the **Mahalanobis distance** $(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$ — the natural distance that accounts for correlation and scale (a point two standard deviations out along a low-variance direction is "far," even if its Euclidean distance is small). The level sets of constant density are ellipsoids whose axes are the eigenvectors of $\boldsymbol{\Sigma}$ and whose radii scale with $\sqrt{\lambda_i}$ (the eigenvalues) — the data "cloud" is an ellipsoid. The normalizer contains $(\det \boldsymbol{\Sigma})^{1/2}$, the volume scaling from the linear-algebra note: $\det \boldsymbol{\Sigma} = \prod \lambda_i$ is the squared volume of that ellipsoid.

5.2 The closure properties that make it tractable

Theorem 5 (Gaussians are closed under the operations we need)**.** *If* $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*:*

(a) ***Affine maps:*** $\mathbf{A}\boldsymbol{X} + \mathbf{b} \sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top)$.

(b) ***Marginals*** *are Gaussian: drop the irrelevant coordinates of* $\boldsymbol{\mu}$ *and* $\boldsymbol{\Sigma}$.

(c) ***Conditionals*** *are Gaussian: partition* $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$*; then* $\boldsymbol{X}_1 \mid \boldsymbol{X}_2 = \mathbf{x}_2$ *is Gaussian with mean* $\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$ *and covariance* $\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$ *(a Schur complement).*

The affine property (a) follows from the mean and variance transformation rules ($\mathbb{E}[\mathbf{A}\boldsymbol{X} + \mathbf{b}] = \mathbf{A}\boldsymbol{\mu} + \mathbf{b}$, $\mathrm{Cov}(\mathbf{A}\boldsymbol{X}) = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top$) plus the fact that linear combinations of jointly Gaussian variables are Gaussian. The conditional formula (c) is the most consequential single equation in this entire subject area:

Intuition

The conditional-mean formula $\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$ is *linear* in the observed $\mathbf{x}_2$ — it *is* linear regression, derived from the Gaussian. The coefficient $\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}$ is exactly the regression coefficient of $\boldsymbol{X}_1$ on $\boldsymbol{X}_2$. And the conditional covariance $\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$ does *not* depend on the observed value — observing $\mathbf{x}_2$ reduces uncertainty by a fixed amount regardless of what you observe. This is the engine of Gaussian-process regression (predict an unseen point by conditioning on observed ones), the Kalman filter (condition the state on each new measurement), and Bayesian linear regression. Master this one formula and a dozen methods become the same method.

5.3 Sampling correlated Gaussians

To sample $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$: Cholesky-factor $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^\top$, draw $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ (independent standard normals), and set $\boldsymbol{X} = \boldsymbol{\mu} + \mathbf{L}\mathbf{z}$. Then $\mathrm{Cov}(\boldsymbol{X}) = \mathbf{L}\,\mathrm{Cov}(\mathbf{z})\,\mathbf{L}^\top = \mathbf{L}\mathbf{L}^\top = \boldsymbol{\Sigma}$ by the affine property. This is how Monte Carlo risk simulations generate correlated asset returns — the Cholesky factor "paints" the desired correlation structure onto independent noise.

6 Limit Theorems: Why Averages Behave

6.1 The Law of Large Numbers

Theorem 6 (LLN). *For i.i.d.* $X_1, X_2, \ldots$ *with mean* μ*, the sample mean* $\bar{X}_n = \frac{1}{n}\sum_i X_i \to \mu$ *(in probability for the weak law; almost surely for the strong law).*

The LLN is the license for empirical estimation: averages converge to expectations, so estimating $\mathbb{E}[X]$ by $\bar{X}_n$ is sound. It justifies Monte Carlo integration (approximate an integral by a sample average) and empirical risk minimization (the average training loss approximates the true expected loss). The proof of the weak law is one line via Chebyshev: $\Pr(|\bar{X}_n - \mu| > \epsilon) \le \mathrm{Var}(\bar{X}_n)/\epsilon^2 = \sigma^2/(n\epsilon^2) \to 0$.

6.2 The Central Limit Theorem

Theorem 7 (CLT). *For i.i.d.* X_i *with mean* μ*, variance* $\sigma^2 < \infty$*,*

$$\sqrt{n}\,(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

The CLT is why the Gaussian is everywhere: *any* sum of many small independent effects is approximately Gaussian, regardless of the individual distributions. This is why measurement errors, aggregated noise, and many estimators are Gaussian, and why the $\sqrt{n}$ scaling appears in every standard error ($\widehat{\mathrm{se}} \propto 1/\sqrt{n}$ — to halve your error bars you need four times the data).

The CLT's fine print — critical for finance

The CLT requires *finite variance* and *independence*, and convergence is slow in the tails. Financial returns violate all three caveats: they have heavy tails (variance may be barely finite or, for some series, effectively infinite), they are dependent (volatility clustering — large moves follow large moves), and the rare extreme events live precisely in the tails the CLT approximates worst. So "returns are approximately normal by the CLT" is a dangerous half-truth: it holds in the bulk and fails exactly where risk lives. This is why tail risk is modeled with heavy-tailed distributions (Student-t, extreme value theory) rather than the Gaussian the CLT might naively suggest.

6.3 The delta method and Slutsky's theorem

The CLT plus a Taylor expansion gives the **delta method**: a smooth function of an asymptotically normal estimator is itself asymptotically normal, $\sqrt{n}(g(\hat{\theta}) - g(\theta)) \xrightarrow{d} \mathcal{N}(0, g'(\theta)^2\sigma^2)$. This propagates uncertainty through nonlinear transformations — from a log-odds to a probability, from an estimated rate to a price. **Slutsky's theorem** lets us replace unknown nuisance quantities (like σ) by consistent estimators without changing the limiting distribution — the formal justification for plugging in $\hat{\sigma}$ to form t-statistics.

7 Concentration Inequalities and Generalization

The limit theorems are asymptotic; concentration inequalities give *finite-sample*, often distribution-free, guarantees — the backbone of statistical learning theory.

Theorem 8 (A ladder of tail bounds).
- ***Markov*** *(nonnegative X):* $\Pr(X \geq a) \leq \mathbb{E}[X]/a$. *The crudest bound, from only the mean.*
- ***Chebyshev****:* $\Pr(|X - \mu| \geq k\sigma) \leq 1/k^2$. *Uses the variance; distribution-free.*
- ***Hoeffding*** *(bounded $X_i \in [a, b]$):* $\Pr(\bar{X}_n - \mathbb{E}\bar{X}_n \geq t) \leq \exp\big(-2nt^2/(b-a)^2\big)$. Exponential *concentration — the deviation probability shrinks exponentially in n.*

Markov follows from $\mathbb{E}[X] \geq \int_a^\infty x\, dF \geq a \Pr(X \geq a)$. Chebyshev is Markov applied to $(X - \mu)^2$. Hoeffding requires a more delicate moment-generating-function argument but delivers the exponential rate that makes learning possible.

Intuition

Concentration is why machine learning generalizes at all. Hoeffding says the training error of a *fixed* predictor is within $O(1/\sqrt{n})$ of its true error with high probability. The complication — and the subject of VC dimension and Rademacher complexity — is that we choose the predictor *after* seeing the data, from a whole class, so we need the bound to hold *uniformly* over the class. The richer the class (the more it can fit), the looser the uniform bound, which is the precise, quantitative statement of overfitting: capacity must be paid for in generalization gap. This is the theoretical counterpart to the bias–variance tradeoff.

8 Estimation Theory

8.1 Maximum likelihood and why it equals minimizing KL / cross-entropy

Given data $x_1, \ldots, x_n$ and a model family $p(x \mid \boldsymbol{\theta})$, the **maximum likelihood estimator** maximizes the probability of the observed data:

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \arg\max_{\boldsymbol{\theta}} \prod_i p(x_i \mid \boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \sum_i \log p(x_i \mid \boldsymbol{\theta}).$$

We take logs because the product underflows and the sum is easier to differentiate; the maximizer is unchanged since log is increasing. Now the conceptual payoff:

Theorem 9 (MLE minimizes KL divergence to the truth)**.** *Maximizing the average log-likelihood is equivalent to minimizing the Kullback–Leibler divergence $D_{KL}(p_{data} \,\|\, p_{\boldsymbol{\theta}})$ from the empirical data distribution to the model.*

Proof. $D_{\mathrm{KL}}(p_{\mathrm{data}} \| p_{\boldsymbol{\theta}}) = \mathbb{E}_{p_{\mathrm{data}}}[\log p_{\mathrm{data}}(x)] - \mathbb{E}_{p_{\mathrm{data}}}[\log p_{\boldsymbol{\theta}}(x)]$. The first term does not involve $\boldsymbol{\theta}$. The second is (the negative of) the average log-likelihood. So minimizing KL over $\boldsymbol{\theta}$ maximizes the average log-likelihood. □

Intuition

This is the bridge between probability and the loss functions you actually minimize. Minimizing negative log-likelihood *is* minimizing KL divergence *is* (for classification) minimizing cross-entropy. So "cross-entropy loss" is not an arbitrary choice — it is the unique loss that makes your model's distribution close to the data's in the KL sense, i.e. maximum likelihood. Likewise, squared-error loss is the negative log-likelihood under Gaussian noise (work it out: $-\log \mathcal{N}(y; \hat{y}, \sigma^2) = \text{const} + (y-\hat{y})^2/2\sigma^2$). Every common loss is a likelihood in disguise; choosing a loss is choosing a noise model.

Worked Example

Deriving the sample mean as the Gaussian MLE. For $x_i \sim \mathcal{N}(\mu, \sigma^2)$, the log-likelihood is $-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2$. Differentiate w.r.t. μ: $\frac{1}{\sigma^2}\sum_i (x_i-\mu) = 0 \Rightarrow \hat{\mu} = \bar{x}$. The sample mean is the MLE. Differentiate w.r.t. σ^2 and solve: $\hat{\sigma}^2 = \frac{1}{n}\sum_i (x_i-\hat{\mu})^2$ — the (biased) sample variance with divisor n, not $n-1$. The MLE is biased here; correcting the bias gives the familiar $n-1$ divisor. This tiny example contains the general lesson that MLEs can be biased.

8.2 Properties of the MLE and the Cramér–Rao bound

Under regularity conditions, the MLE is **consistent** ($\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}$), **asymptotically normal** ($\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}^{-1})$), and **asymptotically efficient** — it attains the smallest possible variance. That smallest variance is the **Cramér–Rao lower bound**: any unbiased estimator has $\mathrm{Var}(\hat{\theta}) \geq \mathcal{I}(\theta)^{-1}$, where $\mathcal{I}(\theta) = -\mathbb{E}[\partial^2 \log p / \partial \theta^2]$ is the **Fisher information** — the curvature of the log-likelihood, measuring how sharply the data pin down the parameter. High curvature (sharp peak) means low uncertainty. The MLE is optimal in the sense of saturating this bound asymptotically.

8.3 MAP estimation: where regularization comes from

The **maximum a posteriori** estimator maximizes the posterior:

$$\hat{\boldsymbol{\theta}}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{\theta}} \big[\log p(x \mid \boldsymbol{\theta}) + \log \pi(\boldsymbol{\theta})\big] = \text{MLE} + \text{a log-prior penalty.}$$

The log-prior is exactly a regularization term. This is the probabilistic origin of regularization:

- A **Gaussian prior** $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \tau^2 \mathbf{I})$ contributes $\log \pi \propto -\|\boldsymbol{\theta}\|^2 / 2\tau^2$ — an ℓ_2 penalty. **Ridge regression is MAP under a Gaussian prior.**
- A **Laplace prior** $\pi(\theta_j) \propto e^{-|\theta_j|/b}$ contributes $\log \pi \propto -\|\boldsymbol{\theta}\|_1 / b$ — an ℓ_1 penalty. **Lasso is MAP under a Laplace prior.**

The regularization strength λ is the ratio of noise variance to prior variance: a tighter prior (smaller τ) means stronger shrinkage. This is why regularization "encodes prior belief" — literally, the penalty *is* a log-prior.

8.4 Bias, variance, and why a biased estimator can win

The mean squared error of an estimator decomposes:

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2] = \underbrace{\mathrm{Var}(\hat{\theta})}_{} + \underbrace{\mathrm{Bias}(\hat{\theta})^2}_{}.$$

(Proof: add and subtract $\mathbb{E}[\hat{\theta}]$ inside the square; the cross term vanishes.) The crucial implication: minimizing MSE does *not* require unbiasedness. A biased estimator with much smaller variance can have lower MSE than the best unbiased one. This is the statistical justification for regularization (ridge is biased but lower-variance) and for shrinkage estimators like **James–Stein** (which provably dominates the sample mean in ≥ 3 dimensions) and **Ledoit–Wolf** covariance shrinkage (which pulls a noisy sample covariance toward a structured target, trading bias for a large variance reduction — essential when estimating the large covariance matrices of portfolio optimization).

Worked Example

Estimating a covariance matrix from n observations of p assets when $p \approx n$: the sample covariance is unbiased but extremely noisy (its smallest eigenvalues are badly underestimated, making $\boldsymbol{\Sigma}^{-1}$ explode — the same ill-conditioning from the linear-algebra note). Ledoit–Wolf shrinks it toward a scaled identity: $\hat{\boldsymbol{\Sigma}}_{\mathrm{LW}} = (1-\delta)\hat{\boldsymbol{\Sigma}} + \delta\,\bar{\lambda}\mathbf{I}$. This is biased but dramatically lower-variance and well-conditioned, yielding better out-of-sample portfolios. It is ridge regression's logic applied to covariance estimation, and it wins precisely because of the bias–variance MSE tradeoff.

9 Hypothesis Testing and the Multiple-Testing Trap

9.1 The framework

A hypothesis test pits a null H_0 against an alternative H_1, computes a test statistic, and asks how surprising the data are under H_0. Two error types: **Type I** (false positive, reject a true null, rate α) and **Type II** (false negative, fail to reject a false null, rate β); **power** $= 1-\beta$ is the probability of detecting a true effect. The p**-value** is $\Pr(\text{statistic at least as extreme} \mid H_0)$ — and the universal misinterpretation is worth stating: it is *not* $\Pr(H_0 \mid \text{data})$. A small p-value means the data are surprising under the null, not that the null is improbable.

A **confidence interval** is a procedure that, over repeated sampling, covers the true parameter $(1-\alpha)$ of the time. (Contrast the Bayesian **credible interval**, which is a direct posterior-probability statement about the parameter — the interpretation people *wish* a confidence interval had.)

9.2 Multiple testing: the statistical heart of backtest overfitting

Why testing many strategies manufactures false discoveries

Test one true null at $\alpha = 0.05$ and you have a 5% chance of a false positive. Test m independent true nulls and the chance of *at least one* false positive is $1-(1-0.05)^m$, which for $m = 20$ is 64% and for $m = 100$ is 99.4%. So if you try 100 random trading signals, you are almost certain to find one that looks "significant" purely by chance. The best in-sample Sharpe ratio across many tried strategies is *selection bias*, not skill.

Corrections control this inflation. **Bonferroni** tests each hypothesis at α/m to control the family-wise error rate (probability of *any* false positive) — conservative but simple. **Benjamini–Hochberg** controls the false discovery rate (expected *proportion* of false positives among rejections)

— more powerful, the right tool when you expect some true effects among many tests. In quant finance this is formalized as **backtest overfitting**: the more configurations you search, the more the apparent performance overstates reality. Remedies include the *deflated Sharpe ratio* (which discounts the observed Sharpe by the number of trials) and out-of-sample / walk-forward validation with a strict accounting of how many strategies were tried.

Intuition

Multiple testing is the same phenomenon as overfitting, viewed through inference rather than prediction. Overfitting: a flexible model fits noise because it has many degrees of freedom to try. Multiple testing: a researcher fits noise because they have many hypotheses to try. Both inflate apparent performance; both are corrected by penalizing the search (regularization / complexity penalties for models; α-adjustment / Sharpe-deflation for tests); both demand honest out-of-sample validation. A quant who internalizes this connection is inoculated against the most common way backtests lie.

10 Information Theory: the Quantities Behind the Losses

- **Entropy** $H(p) = -\sum_x p(x) \log p(x)$: the average surprise, or the minimum bits to encode samples from p. Maximal for the uniform distribution (maximum uncertainty), zero for a point mass (no uncertainty).
- **Cross-entropy** $H(p, q) = -\sum_x p(x) \log q(x)$: the average bits to encode samples from p using a code optimized for q. This *is* the classification loss: p is the true label distribution, q the predicted one.
- **KL divergence** $D_{\mathrm{KL}}(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = H(p, q) - H(p) \geq 0$, with equality iff $p = q$ (Gibbs' inequality, provable by Jensen). It measures the extra bits from using the wrong distribution. Asymmetric ($D_{\mathrm{KL}}(p\|q) \neq D_{\mathrm{KL}}(q\|p)$), so it is a divergence, not a distance.
- **Mutual information** $I(X; Y) = D_{\mathrm{KL}}(p_{XY}\|p_X p_Y)$: how much knowing Y reduces uncertainty about X — dependence beyond linear correlation. A feature-selection criterion that captures nonlinear relationships correlation misses.

The relationship D_{KL} = cross-entropy − entropy explains why minimizing cross-entropy minimizes KL (entropy of the data is constant) — closing the loop with the MLE result: **MLE = minimize KL = minimize cross-entropy**, three names for one principle.

11 Stochastic Processes for the Quant

A stochastic process is a collection of random variables indexed by time — the natural object for modeling price paths, order flow, and regimes.

- **Markov chains:** the future depends on the past only through the present ($\Pr(X_{t+1} \mid X_t, X_{t-1}, \dots) = \Pr(X_{t+1} \mid X_t)$). The long-run behavior is the stationary distribution $\boldsymbol{\pi}$ solving $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$ (a left eigenvector of the transition matrix with eigenvalue 1 — linear algebra again), and the second eigenvalue controls mixing speed. Models regime switching and underlies MCMC.
- **Martingales:** $\mathbb{E}[X_{t+1} \mid \mathcal{F}_t] = X_t$ — the best forecast of tomorrow is today. This is the mathematical form of the efficient-market hypothesis and the no-arbitrage condition: a fair game has no predictable drift. The optional stopping theorem (you cannot beat a martingale with a clever stopping rule) formalizes why simple timing strategies fail on a martingale.

- **Brownian motion and Itô calculus:** the continuous-time limit of a random walk, with independent Gaussian increments. The stochastic differential equation $dX = \mu\,dt + \sigma\,dW$ models asset prices; **Itô's lemma** (the chain rule for stochastic processes, with an extra second-order term $\frac{1}{2}\sigma^2\partial_{xx}$ because $(dW)^2 = dt$) is the tool that derives the Black–Scholes equation. The extra term is pure probability: Brownian motion's quadratic variation is nonzero, so the Taylor expansion keeps a second-order piece that ordinary calculus drops.
- **Poisson processes:** model the *timing* of discrete events (trades, jumps, defaults) with exponential inter-arrival times; the jump-diffusion extension adds Poisson jumps to Brownian motion to capture the discontinuities real markets exhibit.

12 Bayes' Theorem, Worked Several Ways

Bayesian updating is the logical core of probabilistic ML; working concrete examples builds the intuition that the formula alone does not.

12.1 The base-rate example

A screening test for a rare condition (prevalence 1%) has 90% sensitivity (detects the condition when present) and 95% specificity (5% false-positive rate). A test comes back positive — what is the probability the condition is present? By Bayes' theorem,

$$P(\text{cond} \mid +) = \frac{P(+ \mid \text{cond})P(\text{cond})}{P(+ \mid \text{cond})P(\text{cond}) + P(+ \mid \neg\text{cond})P(\neg\text{cond})} = \frac{0.90 \cdot 0.01}{0.90 \cdot 0.01 + 0.05 \cdot 0.99} = \frac{0.009}{0.009 + 0.0495} = \frac{0.0}{0.0}$$

Despite the test's high accuracy, a positive result implies only a 15.4% chance of the condition — because the condition is rare, false positives (from the 99% healthy) swamp true positives. This **base-rate fallacy** is the single most important and most misunderstood consequence of Bayes' theorem, and it recurs throughout ML: a fraud classifier, a rare-disease model, a default predictor all face the same arithmetic — a good classifier on a rare positive still produces mostly-false-positive alerts unless its specificity is extremely high.

12.2 Why the prior dominates when evidence is weak

Repeat with a more prevalent condition (30%): $P(\text{cond} \mid +) = \frac{0.90 \cdot 0.30}{0.90 \cdot 0.30 + 0.05 \cdot 0.70} = \frac{0.27}{0.27 + 0.035} = \frac{0.27}{0.305} = 0.885$. The *same test* now gives an 88.5% posterior — because the prior is higher. The posterior is the prior reweighted by the likelihood ratio, so when the prior is extreme (rare condition) it dominates unless the evidence is overwhelming. This is exactly the mechanism of regularization (a prior pulling the estimate) and of the shrinkage estimators later in this note: the prior matters most when the data are weak, and washes out as evidence accumulates.

Intuition

Bayes' theorem is the grammar of updating: posterior $\propto$ likelihood $\times$ prior. The worked base-rate example is the canonical lesson — a test's accuracy is not the probability you have the condition, because the base rate (prior) reweights everything. In ML this is the difference between a classifier's per-class accuracy and the actual reliability of its positive predictions, which is why precision (a posterior-like quantity) collapses under class imbalance even when sensitivity is high. The same structure — prior reweighted by evidence — underlies Bayesian inference, MAP regularization, naive Bayes, and the calibration of any probabilistic classifier. Internalizing the base-rate arithmetic immunizes you against the most common probabilistic error in applied work.

13 Expectation, Variance, and Covariance, Worked

The moments are the summary statistics every model consumes; we compute them concretely and prove the identities that ML relies on.

13.1 Linearity and the diversification identity

Expectation is linear unconditionally: $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$, regardless of dependence. **Variance is not**: $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y)$. Worked: two assets each with variance $\sigma^2 = 0.04$ (s.d. 20%) and correlation ρ. An equal-weight portfolio has variance

$$\mathrm{Var}\left(\tfrac{X+Y}{2}\right) = \tfrac{1}{4}[\sigma^2 + \sigma^2 + 2\rho\sigma^2] = \tfrac{\sigma^2}{2}(1 + \rho) = 0.02(1 + \rho).$$

At $\rho = 1$ (identical assets): 0.04 — no diversification, the portfolio is as risky as one asset. At $\rho = 0$ (independent): 0.02 — variance halved, s.d. down by $\sqrt{2}$. At $\rho = -1$ (perfect hedge): 0 — risk eliminated. This is the **diversification identity**: combining imperfectly-correlated risks reduces variance, and the benefit grows as correlation falls. It is the mathematical foundation of portfolio theory, of ensemble methods (averaging decorrelated models, the bagging formula of the trees note), and of why correlation, not just individual variance, governs aggregate risk.

13.2 The covariance matrix is positive semidefinite

For any random vector $\mathbf{X}$ and any fixed weights $\mathbf{w}$, the variance of the combination $\mathbf{w}^\top\mathbf{X}$ is $\mathrm{Var}(\mathbf{w}^\top\mathbf{X}) = \mathbf{w}^\top\boldsymbol{\Sigma}\mathbf{w}$, where $\boldsymbol{\Sigma}$ is the covariance matrix. Since a variance is non-negative, $\mathbf{w}^\top\boldsymbol{\Sigma}\mathbf{w} \geq 0$ for all $\mathbf{w}$ — which is exactly the definition of $\boldsymbol{\Sigma}$ being positive semidefinite. This is the probabilistic origin of the PSD property that the linear-algebra note treats abstractly: a covariance matrix *must* be PSD because portfolio variances cannot be negative. A zero eigenvalue corresponds to a portfolio with zero variance — a perfect linear dependence among the variables (redundant assets, collinear features), the same degeneracy that makes regression ill-posed.

Intuition

The moments connect probability to every other note. Linearity of expectation (always true) is what makes unbiased estimators and minibatch gradients work. The variance-of-a-sum identity, with its covariance cross-term, is the diversification principle that underlies portfolio risk and ensemble variance reduction alike. And the PSD-ness of the covariance matrix — forced by the non-negativity of variance — is the probabilistic root of positive definiteness, tying back to convex losses, valid kernels, and well-posed regression. The covariance matrix is the single object where probability, linear algebra, and finance meet, and its PSD structure is why all three behave well.

13.3 The law of total variance, worked

The **law of total variance** decomposes variance through a conditioning variable: $\mathrm{Var}(Y) = \mathbb{E}[\mathrm{Var}(Y \mid Z)] + \mathrm{Var}(\mathbb{E}[Y \mid Z])$ — "within-group" plus "between-group" variance. Worked: returns Y in two regimes Z (calm/crisis), each equally likely. Calm: mean 1%, variance 1. Crisis: mean -3%, variance 9. The within-regime (expected conditional) variance is $\frac{1}{2}(1) + \frac{1}{2}(9) = 5$. The between-regime variance is the variance of the regime means $\{1, -3\}$, which have mean -1 and variance $\frac{1}{2}(1 - (-1))^2 + \frac{1}{2}(-3 - (-1))^2 = \frac{1}{2}(4) + \frac{1}{2}(4) = 4$. Total variance $= 5 + 4 = 9$. The decomposition shows that nearly half the total return variance here comes from *regime switching* (the between term), not within-regime fluctuation — a quantitative argument for regime-aware models. This identity is also the exact algebraic skeleton of the bias–variance decomposition (the model-selection note) and of the explained/unexplained variance split in ANOVA (the regression note).

14 The Distributions You Must Know, with Their Roles

A compact catalog, each with the ML role that makes it worth memorizing.

- **Bernoulli/Binomial**: binary outcomes and their counts — the likelihood behind logistic regression and classification.
- **Categorical/Multinomial**: one-of-K outcomes — the softmax likelihood for multiclass.
- **Gaussian**: the limiting distribution of sums (CLT), the maximum-entropy distribution for fixed mean/variance, and the noise model behind squared-error regression.
- **Poisson**: counts of rare events in a fixed interval — trade arrivals, defaults, the likelihood behind Poisson regression.
- **Exponential/Gamma**: waiting times and positive continuous quantities — durations, the conjugate structure behind many Bayesian models.
- **Beta/Dirichlet**: distributions over probabilities — the conjugate priors for Bernoulli/categorical, hence the smoothing in naive Bayes and Bayesian updating of rates.
- **Student-t**: heavy-tailed, the realistic model for financial returns (the Gaussian understates tail risk).

14.1 The exponential family unifies them

Most of these belong to the **exponential family**, distributions of the form $p(x \mid \boldsymbol{\theta}) = h(x)\exp(\boldsymbol{\theta}^\top T(x) - A(\boldsymbol{\theta}))$, with natural parameter $\boldsymbol{\theta}$, sufficient statistic $T(x)$, and log-partition $A(\boldsymbol{\theta})$. This single structure unifies them: $A(\boldsymbol{\theta})$ generates the moments ($\nabla A = \mathbb{E}[T]$, $\nabla^2 A = \mathrm{Cov}(T)$), conjugate priors always exist, and maximum likelihood reduces to moment matching ($\mathbb{E}[T]$ = observed average of T). The entire GLM framework (the logistic-regression note) is built on this: each GLM picks an exponential-family response distribution, and the unified IRLS fitting algorithm follows from the shared structure. Recognizing a distribution as exponential-family immediately tells you its sufficient statistics, its conjugate prior, and how to fit it.

Intuition

The distribution catalog is not memorization for its own sake — each entry is the likelihood (noise model) of a specific ML method, so knowing the distribution is knowing the model. Gaussian noise gives squared-error regression; Bernoulli gives logistic regression; Poisson gives count regression; categorical gives softmax. The exponential family ties them together, explaining why one fitting algorithm (IRLS/Fisher scoring) handles them all and why conjugate priors and moment-matching MLEs appear everywhere. When you choose a loss function, you are implicitly choosing a noise distribution (the loss is the negative log-likelihood), so understanding the distributions is understanding what your model assumes about the world's randomness.

15 The Multivariate Gaussian, Worked

The Gaussian is the workhorse distribution of ML; its closure properties — worked concretely — are what make it tractable.

15.1 Conditioning, with numbers

Two jointly-Gaussian variables (e.g. two asset returns) with means $(0, 0)$, variances $(1, 1)$, and correlation $\rho = 0.6$, so covariance $\boldsymbol{\Sigma} = \left(\begin{smallmatrix} 1 & 0.6 \\ 0.6 & 1 \end{smallmatrix}\right)$. Suppose we observe $X_2 = 2$ (a 2-sigma move in asset 2). The conditional distribution of X_1 given $X_2 = 2$ is Gaussian with

$$\mathbb{E}[X_1 \mid X_2 = 2] = \mu_1 + \frac{\mathrm{Cov}(X_1, X_2)}{\mathrm{Var}(X_2)}(2 - \mu_2) = 0 + \frac{0.6}{1}(2) = 1.2,$$

$$\mathrm{Var}(X_1 \mid X_2 = 2) = \mathrm{Var}(X_1) - \frac{\mathrm{Cov}(X_1, X_2)^2}{\mathrm{Var}(X_2)} = 1 - \frac{0.36}{1} = 0.64.$$

Observing $X_2 = 2$ shifts our expectation of X_1 to 1.2 (the correlation pulls it up) and *reduces* its variance from 1 to 0.64 (the observation carries information). The conditional mean is a *linear* function of the observation — this linearity is exactly why linear regression is the optimal predictor under joint Gaussianity, and the variance reduction $1-\rho^2$ is the R^2 of that regression. The Gaussian's conditional-is-linear property is the probabilistic foundation of least squares.

15.2 Why the Gaussian is closed under the operations ML needs

Marginals of a Gaussian are Gaussian; conditionals are Gaussian (just shown); linear combinations are Gaussian; and sums of independent Gaussians are Gaussian. These closure properties mean that an entire pipeline of linear operations on Gaussian inputs stays Gaussian and remains analytically tractable — which is why Gaussian assumptions pervade ML (Kalman filters, Gaussian processes, Bayesian linear regression, variational approximations). No other continuous distribution combines this much closure with this much flexibility, and the closure is what lets the math stay in closed form rather than requiring simulation.

15.3 Sampling correlated Gaussians via Cholesky

To generate a draw from $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$: factor $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^\top$ (Cholesky, the linear-algebra note), draw independent standard normals $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and set $\mathbf{x} = \mathbf{L}\mathbf{z}$. Then $\mathrm{Cov}(\mathbf{x}) = \mathbf{L}\,\mathrm{Cov}(\mathbf{z})\mathbf{L}^\top = \mathbf{L}\mathbf{I}\mathbf{L}^\top = \mathbf{\Sigma}$ ✓. Worked for $\mathbf{\Sigma} = \left(\begin{smallmatrix} 1 & 0.6 \\ 0.6 & 1 \end{smallmatrix}\right)$: the Cholesky factor is $\mathbf{L} = \left(\begin{smallmatrix} 1 & 0 \\ 0.6 & 0.8 \end{smallmatrix}\right)$ (check: $\mathbf{L}\mathbf{L}^\top = \left(\begin{smallmatrix} 1 & 0.6 \\ 0.6 & 0.36+0.64 \end{smallmatrix}\right) = \mathbf{\Sigma}$ ✓). So $x_1 = z_1$, $x_2 = 0.6z_1 + 0.8z_2$ — the correlation is injected by mixing the independent noises. This is exactly how Monte Carlo risk simulations generate correlated scenarios, and how the reparameterization trick (the optimization note) samples a VAE's latent.

Intuition

The multivariate Gaussian is tractable because every operation ML performs — marginalize, condition, linearly transform, add — keeps it Gaussian and in closed form. The conditional-mean-is-linear property is the deep reason linear regression is optimal under joint normality, and the Cholesky sampling recipe is how correlated randomness is generated in every Monte Carlo engine. Conditioning reducing variance ($1 \to 1-\rho^2$) is the information content of an observation, quantified. These are not abstract facts: they are the machinery behind Gaussian processes, Kalman filtering, Bayesian linear models, VAEs, and risk simulation. The Gaussian earns its central place by being simultaneously flexible, closed under the right operations, and the CLT-mandated limit of aggregated noise.

16 Limit Theorems and Concentration, Worked

Why do averages behave predictably? The limit theorems and concentration inequalities answer this, and they underpin every generalization guarantee.

16.1 The CLT in action

The **Central Limit Theorem** says the standardized mean of n i.i.d. variables with mean μ and variance σ^2 converges to $\mathcal{N}(0,1)$: $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \to \mathcal{N}(0,1)$. The key consequence is the $\sqrt{n}$ rate: the standard error of the mean is $\sigma/\sqrt{n}$, shrinking as $1/\sqrt{n}$. Worked: to halve the uncertainty in an estimated mean, you need 4× the data; to get one more decimal digit (10× precision), 100× the data. This diminishing return governs everything from A/B test sample sizes to the precision of a

backtest's Sharpe estimate — and it is why more data helps but with steeply diminishing returns, a central fact of the bias–variance and learning-curve discussions (the model-selection note).

16.2 Concentration: finite-sample guarantees

The CLT is asymptotic; **concentration inequalities** give finite-n bounds. **Hoeffding's inequality**: for bounded i.i.d. variables in $[a, b]$, $P(|\bar{X}_n - \mu| \geq t) \leq 2\exp(-2nt^2/(b-a)^2)$ — the probability of the sample mean deviating from the truth by t decays *exponentially* in n. Worked: for variables in $[0, 1]$, the probability that the sample mean is off by more than $t = 0.1$ is at most $2e^{-2n(0.01)} = 2e^{-0.02n}$; at $n = 500$ this is $2e^{-10} \approx 9 \times 10^{-5}$ — a tiny chance, so 500 samples pin the mean within 0.1 with high confidence. This exponential concentration is the engine behind generalization bounds (the model-selection note): it bounds how far training error can stray from true error, and the union bound extends it across a hypothesis class to give uniform guarantees.

Intuition

The limit theorems and concentration inequalities are why learning is possible at all: they guarantee that averages over data (training error, estimated means, Monte Carlo estimates) converge to and concentrate around their true values. The CLT gives the $1/\sqrt{n}$ rate that sets every sample-size calculation and warns of diminishing returns. Concentration inequalities (Hoeffding, Bernstein, McDiarmid) give the finite-sample, exponential-decay guarantees that, combined with the union bound, produce the generalization bounds of statistical learning theory. Without concentration, there would be no reason training performance should predict test performance — these inequalities are the mathematical license to learn from finite data.

17 Maximum Likelihood and Fisher Information, Worked

Maximum likelihood is the estimation principle behind most of ML; working it on the Gaussian shows why it equals minimizing squared error, and the Fisher information quantifies how much the data tell us.

17.1 MLE for the Gaussian mean equals least squares

For i.i.d. data $x_i \sim \mathcal{N}(\mu, \sigma^2)$, the log-likelihood is $\ell(\mu) = -\frac{1}{2\sigma^2}\sum_i (x_i - \mu)^2 + \text{const}$. Maximizing over μ is *minimizing* $\sum_i (x_i - \mu)^2$ — least squares. Setting the derivative to zero: $\sum_i (x_i - \mu) = 0 \Rightarrow \hat{\mu} = \bar{x}$, the sample mean. So the MLE of a Gaussian mean is the sample average, and **maximizing a Gaussian likelihood is identical to minimizing squared error**. This is why squared-error regression is the maximum-likelihood estimator under Gaussian noise — the loss function *is* the negative log-likelihood, and the Gaussian's quadratic exponent is the squared-error loss. Choosing a loss is choosing a noise model.

17.2 MLE equals minimizing KL divergence / cross-entropy

More generally, maximizing the likelihood is equivalent to minimizing the KL divergence from the empirical data distribution to the model — equivalently, minimizing the cross-entropy. This is the unifying statement: squared error (Gaussian), cross-entropy/log-loss (Bernoulli/categorical), and Poisson deviance are all the cross-entropy / negative-log-likelihood for their respective noise models. The classification note's "minimize cross-entropy" and the regression note's "minimize squared error" are the same principle — maximum likelihood — instantiated with different distributions. This single idea connects nearly every loss function in the series.

17.3 Fisher information and the Cramér–Rao bound

The **Fisher information** $I(\theta) = -\mathbb{E}[\partial^2 \ell / \partial \theta^2]$ measures the curvature of the log-likelihood — how sharply the data pin down the parameter. The **Cramér–Rao bound** states that no unbiased estimator can have variance below $1/I(\theta)$: the Fisher information sets the precision ceiling. Worked for the Gaussian mean: $\ell = -\frac{1}{2\sigma^2}\sum(x_i - \mu)^2$, so $\partial^2 \ell / \partial \mu^2 = -n/\sigma^2$, giving $I(\mu) = n/\sigma^2$ and a variance bound σ^2/n — exactly the variance of the sample mean, so the sample mean is efficient (achieves the bound). The Fisher information also appears as the Hessian in IRLS (the logistic-regression note) and as the natural-gradient metric (the optimization note's NGBoost mention) — the same quantity governing estimation precision, optimization curvature, and the geometry of parameter space.

Intuition

Maximum likelihood is the thread connecting the entire series' loss functions: it equals minimizing squared error under Gaussian noise, cross-entropy under categorical/Bernoulli noise, and in general minimizing the KL divergence from data to model. So every loss is a negative log-likelihood, and choosing a loss is choosing a noise model — a unifying realization that turns a grab-bag of losses into one principle. The Fisher information then quantifies how much the data inform the parameters (the curvature of the likelihood), sets the precision ceiling (Cramér–Rao), and reappears as the Hessian in second-order fitting and as the metric in natural-gradient methods. MLE and Fisher information are the statistical engine under the optimization hood.

18 MAP Estimation: Where Regularization Comes From

Maximum likelihood plus a prior gives maximum a posteriori estimation — and this is precisely where regularization originates.

18.1 The derivation

Bayes' theorem gives the posterior $p(\boldsymbol{\theta} \mid \text{data}) \propto p(\text{data} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})$. Taking logs, the MAP estimate maximizes $\log p(\text{data} \mid \boldsymbol{\theta}) + \log p(\boldsymbol{\theta})$ — the log-likelihood *plus* a log-prior penalty. With a Gaussian prior $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \tau^2 \mathbf{I})$, the log-prior is $-\frac{1}{2\tau^2}\|\boldsymbol{\theta}\|^2 + \text{const}$, so MAP maximizes $\log p(\text{data} \mid \boldsymbol{\theta}) - \frac{1}{2\tau^2}\|\boldsymbol{\theta}\|^2$. For Gaussian-noise regression this is exactly **ridge regression**: the squared-error loss plus an ℓ_2 penalty with $\lambda = \sigma^2/\tau^2$. **So ridge regression is MAP estimation with a Gaussian prior**, and the penalty strength λ is the ratio of noise variance to prior variance — more regularization when the noise is large or the prior is tight.

18.2 The lasso as a Laplace prior

Replacing the Gaussian prior with a **Laplace** (double-exponential) prior $p(\theta_j) \propto e^{-|\theta_j|/b}$ gives a log-prior $-\frac{1}{b}\sum_j |\theta_j|$ — the ℓ_1 penalty. So **lasso is MAP estimation with a Laplace prior**, and its sparsity reflects the Laplace prior's sharp peak at zero (it places more prior mass exactly at zero than the Gaussian). This is the probabilistic origin of the regularizers the regression and optimization notes treat as penalties: every penalty is a log-prior, and the choice of penalty is the choice of prior belief about the parameters. Strong regularization is a confident prior that the parameters are small; weak regularization is a diffuse prior that lets the data speak.

Intuition

The MAP derivation reveals regularization as Bayesian inference in disguise: the penalty *is* a log-prior, λ encodes the prior's tightness relative to the noise, and the choice of penalty (ℓ_2 vs ℓ_1)

is the choice of prior (Gaussian vs Laplace). This unifies three views that recur across the series — the regression note's "shrink the coefficients," the optimization note's "constrained/penalized objective," and this note's "Bayesian posterior mode." It also explains *why* regularization helps: it injects prior information that stabilizes estimation when data are weak (the same lesson as the base-rate example), and it dissolves as data accumulate (the likelihood overwhelms the prior). Regularization is not an ad hoc trick; it is the rational use of prior belief.

19 Hypothesis Testing and the Multiple-Testing Trap

Statistical testing is the discipline that separates signal from noise, and its failure mode — multiple testing — is the statistical heart of backtest overfitting.

19.1 The framework, worked

A test statistic measures how far the data deviate from a null hypothesis, and the p-value is the probability of a deviation that large under the null. Worked: a strategy's mean daily return is 0.05% with daily s.d. 1% over $n = 252$ days. The t-statistic is $\frac{0.05}{1/\sqrt{252}} = \frac{0.05}{0.063} = 0.79$ — well below the ≈ 2 threshold for significance, so this single strategy's edge is not distinguishable from zero. To reach significance ($t = 2$) at this effect size would need $\frac{0.05}{1/\sqrt{n}} = 2 \Rightarrow \sqrt{n} = 40 \Rightarrow n \approx 1600$ days. The arithmetic shows how much data a small edge requires to confirm — and how easily noise masquerades as signal in short samples.

19.2 Multiple testing: how phantom alpha is manufactured

If you test *many* strategies, the chance that *at least one* looks significant by luck explodes. Under the null, each test has a 5% false-positive rate; testing m independent strategies, the probability of *no* false positive is 0.95^m, so the chance of at least one is $1 - 0.95^m$. Worked: with $m = 20$ strategies, $1 - 0.95^{20} = 1 - 0.36 = 0.64$ — a 64% chance of a spurious "significant" result. With $m = 100$, it is $1 - 0.95^{100} = 0.994$ — virtually certain. So a quant who backtests 100 ideas and reports the best one is almost guaranteed a great-looking backtest from pure noise. The corrections — **Bonferroni** (require $p < \alpha/m$), **false discovery rate** control (Benjamini–Hochberg), and the **deflated Sharpe ratio** (the evaluation note) — all adjust the significance bar for the number of trials. This is the statistical core of backtest overfitting and the reason honest trial-counting is non-negotiable in quant research.

Intuition

Multiple testing is the most important statistical pitfall in quantitative finance, and the arithmetic is unforgiving: test enough hypotheses and you will find "significant" results in pure noise with near-certainty. The single-test framework tells you how much data a real edge needs to confirm; the multiple-testing correction tells you how much higher the bar must rise once you account for how many ideas you tried. Every defense in the evaluation note — purged cross-validation, the deflated Sharpe ratio, honest trial accounting, out-of-sample holdouts — exists to counter this one trap. The discipline is to treat a beautiful backtest with suspicion proportional to the number of ideas searched, because the search itself manufactures the beauty.

20 Information Theory: the Quantities Behind the Losses

The losses ML minimizes are information-theoretic quantities; knowing them explains why the losses take the forms they do.

20.1 Entropy, cross-entropy, and KL, with numbers

Entropy $H(p) = -\sum_k p_k \log p_k$ measures a distribution's uncertainty. Worked: a fair coin has $H = -2(\frac{1}{2}\log_2\frac{1}{2}) = 1$ bit (maximum uncertainty); a biased coin with $p = 0.9$ has $H = -0.9\log_2 0.9 - 0.1\log_2 0.1 = 0.137 + 0.332 = 0.469$ bits (less uncertain). **Cross-entropy** $H(p,q) = -\sum_k p_k \log q_k$ measures the cost of coding data from p using a model q; it is exactly the log-loss of the classification note. **KL divergence** $D_{\mathrm{KL}}(p\|q) = \sum_k p_k \log(p_k/q_k) = H(p,q) - H(p)$ measures how far the model q is from the truth p — always ≥ 0, zero iff $p = q$. Worked: if the truth is $p = (0.9, 0.1)$ and the model predicts $q = (0.5, 0.5)$, the KL is $0.9\log_2(0.9/0.5) + 0.1\log_2(0.1/0.5) = 0.9(0.848) + 0.1(-2.322) = 0.763 - 0.232 = 0.531$ bits — the model wastes about half a bit per symbol by being wrong.

20.2 Why minimizing cross-entropy is the right objective

Since $H(p,q) = H(p) + D_{\mathrm{KL}}(p\|q)$ and $H(p)$ is fixed (a property of the data, not the model), minimizing cross-entropy is *identical* to minimizing the KL divergence from the model to the truth — i.e. making the model as close to the true distribution as possible. This is why cross-entropy is the universal classification loss: minimizing it is minimizing the (information-theoretic) distance between your predictions and reality, which is exactly maximum likelihood (shown earlier). The three views — minimize cross-entropy, minimize KL, maximize likelihood — are one objective, and information theory is what reveals their unity.

Intuition

Information theory supplies the meaning behind the losses. Entropy is irreducible uncertainty; cross-entropy is the coding cost of a wrong model and equals the log-loss; KL divergence is the distance from model to truth and equals cross-entropy minus entropy. Because the data's entropy is fixed, minimizing cross-entropy, minimizing KL, and maximizing likelihood are the same objective — the deepest unification in the series, explaining why log-loss is everywhere. The same quantities reappear as the ELBO's KL term (the unsupervised note), as mutual information in feature selection, and as the compression view of learning (MDL, the model-selection note). Information theory is the common language under the losses.

21 Further Worked Exercises

Worked Example

Exercise. Show that a biased estimator can have lower mean squared error than an unbiased one, and connect this to regularization.

Solution. Mean squared error decomposes as MSE = bias^2 + variance. An unbiased estimator has zero bias but may have high variance; a biased estimator trades a little bias for a large variance reduction, lowering total MSE. Concrete case: the James–Stein estimator shrinks the sample mean toward zero and, in dimension ≥ 3, has uniformly lower MSE than the (unbiased) sample mean — a famous and initially shocking result. Ridge regression is exactly this trade: it biases coefficients toward zero (introducing bias) but sharply reduces their variance (especially under collinearity), and for an appropriate λ the MSE is lower than unbiased OLS. So "unbiased" is not the goal — low MSE is, and accepting bias to cut variance is often optimal. This is the statistical justification for all shrinkage and regularization, and the bias–variance tradeoff (the model-selection note) is its general statement.

Worked Example

Exercise. A backtest of the single best of 50 tested strategies shows a t-statistic of 2.3 ($p \approx 0.02$). Should you trust it?
Solution. No — not without correcting for the 50 trials. The naive $p \approx 0.02$ looks significant, but with 50 tests the expected number of false positives at the 5% level is $50 \times 0.05 = 2.5$, so finding a couple of "significant" results in pure noise is expected. A Bonferroni correction requires $p < 0.05/50 = 0.001$ for genuine significance; the observed $p \approx 0.02$ fails this badly. Equivalently, the probability that the best of 50 noise strategies exceeds $t = 2.3$ by chance is substantial. The honest assessment is that this backtest is consistent with no real edge, and the deflated Sharpe ratio (the evaluation note) would formalize the deflation for 50 trials. The lesson: a p-value or t-statistic is meaningless without knowing how many hypotheses were tested to find it — the multiple-testing trap that manufactures phantom alpha.

Worked Example

Exercise. Explain why the sample covariance matrix is a poor estimate in high dimensions and what to do about it.
Solution. With p assets and n observations, the sample covariance has $p(p+1)/2$ parameters estimated from np numbers; when p is comparable to or larger than n, the estimate is noisy and, if $p > n$, *singular* (rank-deficient, non-invertible). Its small eigenvalues are especially unreliable — biased toward zero and dominated by noise — which is catastrophic for portfolio optimization, where the inverse covariance (dividing by those tiny, noisy eigenvalues) amplifies the error (the conditioning problem of the linear-algebra note). Remedies: **shrinkage** (Ledoit–Wolf) pulls the sample covariance toward a structured target (e.g. a scaled identity), trading bias for a huge variance reduction — exactly ridge regularization for covariance estimation; **factor models** impose low-rank-plus-diagonal structure (the PCA of the unsupervised note); and **hierarchical** methods (the trees note's HRP) avoid inversion entirely. All inject structure/prior information to stabilize an over-parameterized estimate — the same regularization principle threaded through the entire series.

22 Stochastic Processes for the Quant, Worked

Time-indexed randomness — prices, volatilities, order flow — is modeled by stochastic processes. We work the handful that matter most.

22.1 The random walk and why prices look like one

A **random walk** is $S_t = S_{t-1} + \epsilon_t$ with i.i.d. increments ϵ_t (mean 0, variance σ^2). Its defining features fall straight out of the moment rules. The mean is flat: $\mathbb{E}[S_t] = S_0$. But the variance *grows linearly in time*: since the increments are independent, $\mathrm{Var}(S_t) = \mathrm{Var}(\sum_{s \le t} \epsilon_s) = t\sigma^2$, so the standard deviation grows as $\sigma\sqrt{t}$ — the famous **square-root-of-time** scaling. Worked: if daily return volatility is 1%, the volatility over $t = 252$ trading days is $1\% \times \sqrt{252} \approx 15.9\%$, the standard annualization. This $\sqrt{t}$ law is why volatility scales with the square root of the horizon, why a longer backtest does not proportionally reduce noise, and why diffusion spreads as $\sqrt{t}$. The efficient-market intuition that price changes are unpredictable makes the random walk the natural null model for prices — and any strategy must beat this null to be real.

22.2 Martingales and the fair-game property

A **martingale** satisfies $\mathbb{E}[S_{t+1} \mid \text{past}] = S_t$ — the best forecast of tomorrow is today, a "fair game" with no exploitable drift. The random walk (with zero-mean increments) is a martingale. The

concept is central to finance: under the risk-neutral measure, discounted asset prices are martingales (the foundation of arbitrage-free pricing), and the martingale property formalizes "no predictable edge." For a quant, testing whether a price series is a martingale is testing whether it is predictable at all; a strategy's claimed alpha is a claim that some transformed price series is *not* a martingale — that there is a predictable component the market has not arbitraged away. The martingale is thus the precise statement of the null hypothesis every strategy is implicitly testing against.

22.3 Markov chains and regime models

A **Markov chain** has the memoryless property $P(\text{state}_{t+1} \mid \text{all past}) = P(\text{state}_{t+1} \mid \text{state}_t)$ — the future depends on the past only through the present state. Worked: a two-regime market (calm/crisis) with transition matrix $\mathbf{P} = \left(\begin{smallmatrix} 0.95 & 0.05 \\ 0.20 & 0.80 \end{smallmatrix}\right)$ (calm persists with prob 0.95, crisis with 0.80). The long-run (stationary) distribution $\boldsymbol{\pi}$ solves $\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}$: setting $\pi_{\text{calm}} \cdot 0.05 = \pi_{\text{crisis}} \cdot 0.20$ (balance of flows) gives $\pi_{\text{crisis}}/\pi_{\text{calm}} = 0.05/0.20 = 0.25$, so $\boldsymbol{\pi} = (0.8, 0.2)$ — the market is in crisis 20% of the time in the long run. The expected crisis duration is $1/(1 - 0.80) = 5$ periods (the mean of a geometric distribution). Markov chains underlie the regime-switching models (the regression note), hidden Markov models for state inference, and Markov-chain Monte Carlo for Bayesian computation — all exploiting the tractability the memoryless property provides.

22.4 Geometric Brownian motion: the continuous limit

Taking the random walk to continuous time with multiplicative (proportional) returns gives **geometric Brownian motion**, $dS = \mu S\,dt + \sigma S\,dW$ — the model behind Black–Scholes. The key features: prices stay positive (multiplicative, not additive, shocks), log-returns are normally distributed with variance $\sigma^2 t$ (the $\sqrt{t}$ law again), and the log-price is an ordinary Brownian motion with drift. The model's known flaws — constant volatility (real volatility clusters and spikes), Gaussian log-returns (real returns are heavy-tailed, the Student-t of the distributions catalog) — are exactly what more advanced models (stochastic volatility, jump-diffusions) repair, but GBM remains the baseline against which they are measured.

Intuition

Stochastic processes turn the static probability of the earlier sections into the dynamics that markets actually exhibit. The random walk's $\sqrt{t}$ volatility scaling is one of the most-used facts in finance (annualization, horizon scaling); the martingale property is the precise null hypothesis every alpha claim contests; the Markov chain is the engine of regime models and MCMC; and geometric Brownian motion is the continuous-time baseline for derivatives. Each is built from the same moment rules and limit theorems established earlier — independence makes variances add, the CLT makes the limit Gaussian, conditioning gives the Markov and martingale structure. For a quant, these processes are the vocabulary for describing how uncertainty evolves over time, and knowing their worked properties is knowing what "random" means for a price series.

23 Worked Bayesian Updating: Beta–Bernoulli

A fully worked conjugate update shows Bayesian learning in closed form and explains the smoothing that appears in naive Bayes and rate estimation.

23.1 The setup and the update

We estimate an unknown success probability θ (a fill rate, a default rate, a click-through rate). Start with a **Beta** prior $\theta \sim \text{Beta}(\alpha, \beta)$, whose mean is $\alpha/(\alpha + \beta)$ and which we read as "$\alpha - 1$

prior successes and $\beta - 1$ prior failures." Observe n trials with s successes (a Bernoulli/Binomial likelihood). Because the Beta is conjugate to the Bernoulli, the posterior is again Beta with updated counts:

$$\theta \mid \text{data} \sim \text{Beta}(\alpha + s,\ \beta + n - s).$$

The posterior mean is $\frac{\alpha+s}{\alpha+\beta+n}$ — the data counts simply add to the prior counts. Worked: a flat prior $\text{Beta}(1,1)$ (uniform, "one prior success and one prior failure") after observing $s = 7$ successes in $n = 10$ trials gives posterior $\text{Beta}(8,4)$, mean $8/12 = 0.667$ — pulled below the raw frequency $7/10 = 0.70$ toward the prior mean 0.5, because the prior carries a little weight. With a stronger prior $\text{Beta}(10,10)$ (worth ≈ 20 pseudo-observations), the same data give $\text{Beta}(17,13)$, mean $17/30 = 0.567$ — pulled much harder toward 0.5, because the prior is more confident.

23.2 Why this is exactly Laplace smoothing

The posterior-mean formula $\frac{\alpha+s}{\alpha+\beta+n}$ is precisely the **add-α smoothing** of naive Bayes (the kNN/Naive-Bayes note): the pseudo-counts α, β prevent a zero estimate when $s = 0$ or $s = n$. So Laplace smoothing is not an ad hoc fix — it is the posterior mean under a Beta prior, the Bayesian update made explicit. As n grows, the data counts swamp the prior counts and the estimate converges to the raw frequency s/n — the prior washes out, exactly as the base-rate discussion predicted. The prior matters most when data are scarce (few trials), which is precisely when an unsmoothed estimate is most dangerous.

Intuition

The Beta–Bernoulli update is the cleanest worked example of Bayesian learning: the prior and posterior share a form (conjugacy), the update is just "add the observed counts to the prior counts," and the posterior mean is a data/prior blend whose mix is governed by their relative counts. This single example illuminates several recurring themes — Laplace smoothing is its posterior mean, regularization is its MAP cousin, and the prior-washes-out-with-data behavior is the general fate of any prior. For a quant estimating a rate from limited data (a strategy's hit rate, a counterparty's default frequency), the Beta–Bernoulli update is the right tool: it gives an honest estimate that does not collapse to 0 or 1 on small samples and quantifies uncertainty through the full posterior, not just a point.

24 Consolidated Exercises: Probability Across the Series

Worked Example

Exercise. Show that under Gaussian noise, the maximum-likelihood regression coefficients equal the ordinary least-squares coefficients.

Solution. Model $y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d. The likelihood of the data is $\prod_i \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2}{2\sigma^2})$. The log-likelihood is $-\frac{1}{2\sigma^2} \sum_i (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2 + \text{const}$. Maximizing over $\boldsymbol{\beta}$ is minimizing $\sum_i (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2 = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$ — exactly the least-squares objective. So $\hat{\boldsymbol{\beta}}_{\text{MLE}} = \hat{\boldsymbol{\beta}}_{\text{OLS}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$. The Gaussian's quadratic exponent *is* the squared-error loss, which is why least squares and Gaussian-noise maximum likelihood coincide. Changing the noise distribution changes the loss: Laplace noise gives least-absolute-deviation, Bernoulli gives cross-entropy. This is the unifying "loss = negative log-likelihood" principle, here instantiated for the Gaussian.

Worked Example

Exercise. A classifier outputs probability 0.7 for an event. Over many such 0.7-predictions, the event occurs 50% of the time. What is wrong, and which quantity measures it?
Solution. The classifier is **miscalibrated**: its stated probabilities do not match observed frequencies (it says 0.7 but reality is 0.5, so it is overconfident). Calibration requires that among all predictions of probability p, the event occurs a fraction p of the time. The quantity that measures this is the **expected calibration error** (the evaluation note), the average gap between predicted probability and observed frequency across probability bins, and the **Brier score** (mean squared error of probabilities) penalizes it as part of its decomposition. The fix is post-hoc calibration (Platt scaling or isotonic regression, the SVM and classification notes), fit on held-out data. Miscalibration matters whenever the probability feeds a decision — position sizing, expected-loss pricing — because acting on 0.7 when the truth is 0.5 systematically misallocates. Good ranking (AUC) does not imply good calibration; they are distinct properties, and many strong classifiers rank well but need calibrating.

Worked Example

Exercise. Explain why independence of features makes the joint likelihood factorize, and why this both helps and hurts naive Bayes.
Solution. If features are conditionally independent given the class, the joint conditional density factorizes: $p(\mathbf{x} \mid c) = \prod_j p(x_j \mid c)$. This is the naive Bayes assumption, and it *helps* enormously: instead of estimating a high-dimensional joint density (infeasible, the curse of dimensionality), you estimate d one-dimensional densities — few parameters, low variance, fast, and robust with little data (the kNN/Naive-Bayes note's data-efficiency). It *hurts* when features are actually correlated: the factorization double-counts redundant evidence (correlated features each contribute as if independent), pushing the posterior probabilities toward 0/1 — the characteristic overconfidence. The saving grace is that classification needs only the correct argmax, not calibrated probabilities, so the *ranking* often survives the false assumption even when the probability magnitudes do not. Hence naive Bayes ranks well (good AUC) but is poorly calibrated (poor Brier) — a direct consequence of the factorization, and the reason it pairs naturally with a calibration step.

Worked Example

Exercise. Using the law of total variance, explain why a model that ignores regime structure underestimates risk.
Solution. The law of total variance gives $\mathrm{Var}(Y) = \mathbb{E}[\mathrm{Var}(Y \mid Z)] + \mathrm{Var}(\mathbb{E}[Y \mid Z])$, where Z is the regime. A model that ignores regimes captures only an average within-regime variance — roughly the first term, $\mathbb{E}[\mathrm{Var}(Y \mid Z)]$ — but *misses* the second term, $\mathrm{Var}(\mathbb{E}[Y \mid Z])$, the variance contributed by the regime means differing (calm vs. crisis having different average returns). As the worked example earlier showed, that between-regime term can be a large fraction of total variance (there, 4 of 9). So a single-regime model systematically *understates* total risk by omitting the between-regime component — it sees the day-to-day wiggle but not the regime-shift jumps. This is why risk models that ignore regime switching underestimate tail risk and why regime-aware models (the regression note's Markov-switching) matter for honest risk assessment. The decomposition quantifies exactly what is being left out.

25 Maximum Entropy and Why the Gaussian Is Special

Information theory not only measures losses; it explains *why* certain distributions are the natural defaults, through the maximum-entropy principle.

25.1 The principle

The **maximum-entropy principle** says: among all distributions consistent with what you know (your constraints), choose the one with the highest entropy — the least committal, assuming no structure beyond the constraints. This is the formal version of Occam's razor for distributions: do not assume more than your information justifies. The remarkable payoff is that the familiar distributions are each the maximum-entropy distribution under natural constraints:

- Fixed support $[a, b]$, nothing else $\Rightarrow$ the **uniform** distribution (maximum entropy is flatness).
- Fixed mean on $[0, \infty) \Rightarrow$ the **exponential** distribution.
- Fixed mean and variance on $\mathbb{R} \Rightarrow$ the **Gaussian**.
- Fixed mean count $\Rightarrow$ the **Poisson** (among count distributions).

So the Gaussian is not just the CLT limit — it is also the *most honest* distribution when all you are willing to specify is a mean and a variance, encoding no additional assumptions. This is a second, independent reason the Gaussian is the default noise model: choosing it commits to the least beyond the first two moments.

25.2 Why this matters for modeling

The maximum-entropy view tells you which distributional assumption is the most conservative given your knowledge, and it connects directly to the exponential family: maximum-entropy distributions under moment constraints are *exactly* the exponential family, with the constrained moments as sufficient statistics. So the exponential-family structure (the distributions section) and the maximum-entropy principle are two faces of one idea. Logistic regression itself is the maximum-entropy classifier subject to matching the observed feature–label correlations — which is why "maximum entropy model" (the NLP name) and "logistic regression" (the statistics name) are the same thing.

Intuition

Maximum entropy is the principle of least presumption: specify only what you know, and let the distribution be maximally noncommittal about everything else. It singles out the uniform, exponential, Gaussian, and Poisson as the honest defaults under their respective constraints, and it coincides with the exponential family and (for classification) with logistic regression. For a modeler, it answers "which distribution should I assume?" with a principled "the one that adds no unwarranted structure beyond your stated constraints." This is why Gaussian noise is the conservative default for a continuous quantity known only through its mean and variance — not because the world is Gaussian, but because assuming anything sharper would smuggle in unjustified information.

26 Mutual Information for Feature Selection, Worked

Mutual information measures how much knowing one variable reduces uncertainty about another: $I(X;Y) = H(Y) - H(Y \mid X) = D_{\mathrm{KL}}(p(x,y) \,\|\, p(x)p(y))$. It is zero iff X and Y are independent, and unlike correlation it captures *nonlinear* dependence — making it a powerful, model-agnostic feature-selection criterion.

26.1 A worked computation

A binary feature X and binary label Y with joint probabilities: $P(X{=}1, Y{=}1) = 0.4$, $P(X{=}1, Y{=}0) = 0.1$, $P(X{=}0, Y{=}1) = 0.1$, $P(X{=}0, Y{=}0) = 0.4$. Marginals: $P(X{=}1) = 0.5$, $P(Y{=}1) = 0.5$. Compute $I(X;Y) = \sum p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}$. Each "aligned" cell (0.4) contributes $0.4 \log_2 \frac{0.4}{0.25} = 0.4 \log_2 1.6 = 0.4(0.678) = 0.271$; each "misaligned" cell (0.1) contributes $0.1 \log_2 \frac{0.1}{0.25} = 0.1 \log_2 0.4 =$

$0.1(-1.322) = -0.132$. Summing all four: $2(0.271)+2(-0.132) = 0.542-0.264 = 0.278$ bits. So this feature carries 0.278 bits about the label (out of the label's 1 bit of total entropy) — a moderately informative feature. A feature independent of the label would give $I = 0$; a perfectly predictive one would give $I = H(Y) = 1$ bit.

26.2 Why mutual information beats correlation for selection

Correlation detects only *linear* association, so a feature with a strong nonlinear (e.g. U-shaped or threshold) relationship to the target can have zero correlation yet high mutual information. Ranking features by mutual information therefore catches predictive features that a correlation screen would discard — valuable in finance, where relationships are often nonlinear (volatility regimes, threshold effects). The cautions: mutual information must be *estimated* (binning or nearest-neighbor methods, which are noisy in high dimensions), and it scores features one at a time, missing interactions and redundancy (two features each informative but carrying the same information). So it is a strong univariate filter — a first-pass screen — best followed by a multivariate, model-based selection (the model-selection note's embedded methods) that accounts for redundancy.

Intuition

Mutual information is the information-theoretic measure of dependence, capturing nonlinear relationships that correlation misses and providing a model-agnostic feature-selection filter. The worked computation shows it quantifying, in bits, how much a feature tells you about the label. Its place in the toolkit is as a univariate screen — fast, model-free, nonlinear-aware — with the caveats that it is noisy to estimate and blind to interactions and redundancy, so it complements rather than replaces the embedded, multivariate selection of regularized models. It also reappears as the objective in information-bottleneck representations and as the KL term measuring dependence throughout the series. Knowing it gives a principled, distribution-level answer to "how informative is this feature?" beyond linear correlation.

27 KL Divergence in Variational Inference, Worked

The KL divergence is not only an evaluation quantity — it is the objective that variational inference optimizes, tying this note to the unsupervised-learning note's ELBO.

27.1 The variational objective

Exact Bayesian posteriors are usually intractable. **Variational inference** approximates the true posterior $p(\mathbf{z} \mid \mathbf{x})$ by a simpler distribution $q(\mathbf{z})$, chosen to minimize $D_{\mathrm{KL}}(q \,\|\, p(\mathbf{z} \mid \mathbf{x}))$. Since that KL contains the intractable evidence, the method instead maximizes the **ELBO**, and the algebra gives exactly

$$\log p(\mathbf{x}) = \mathrm{ELBO}(q) + D_{\mathrm{KL}}(q \,\|\, p(\mathbf{z} \mid \mathbf{x})).$$

Because the left side is fixed and the KL is ≥ 0, maximizing the ELBO *minimizes* the KL gap — pushing q toward the true posterior. This is the same decomposition as cross-entropy = entropy + KL: a fixed total split into a tractable objective plus a nonnegative divergence, so optimizing the tractable part closes the divergence.

27.2 The forward/reverse KL distinction

Variational inference minimizes the *reverse* KL $D_{\mathrm{KL}}(q\|p)$, which is **mode-seeking**: q concentrates on one mode of p and avoids regions where p is small (because q large where p small incurs huge

penalty). Maximum likelihood / cross-entropy minimizes the *forward* KL $D_{\mathrm{KL}}(p\|q)$, which is **mean-seeking**: q must cover all of p's support (it is penalized for being small where p is large), so it spreads to cover every mode. This asymmetry — KL is not a distance, $D_{\mathrm{KL}}(p\|q) \neq D_{\mathrm{KL}}(q\|p)$ — explains why variational approximations tend to be too *narrow* (mode-seeking, underestimating uncertainty), a known limitation, while maximum-likelihood fits tend to over-spread. Knowing which KL direction a method uses predicts how its approximation will fail.

Intuition

KL divergence is the connective tissue between this note and probabilistic ML. It is the gap in Jensen's inequality (the optimization note), the difference between cross-entropy and entropy (so minimizing log-loss minimizes KL to the truth), and the objective of variational inference (maximizing the ELBO minimizes KL to the posterior). Its asymmetry is not a defect to ignore but a design choice: reverse KL (variational) is mode-seeking and yields confident-but-narrow approximations; forward KL (maximum likelihood) is mean-seeking and yields broad ones. The same quantity thus measures model error, drives the training loss, and defines approximate inference — and understanding its direction tells you whether your method will under- or over-estimate uncertainty. KL is the single most important non-distance in machine learning.

28 Sufficient Statistics and Why They Matter

A **sufficient statistic** captures everything in the data relevant to a parameter — once you know it, the raw data carry no further information about that parameter. Formally, $T(\mathbf{x})$ is sufficient for θ if the conditional distribution of the data given $T(\mathbf{x})$ does not depend on θ.

28.1 The factorization criterion, worked

The Fisher–Neyman factorization theorem says T is sufficient iff the likelihood factors as $p(\mathbf{x} \mid \theta) = g(T(\mathbf{x}), \theta)\, h(\mathbf{x})$ — the parameter interacts with the data *only* through T. Worked for the Gaussian with known variance: the likelihood's θ-dependence is $\exp(\frac{\mu}{\sigma^2}\sum_i x_i - \frac{n\mu^2}{2\sigma^2})$, which depends on the data only through $\sum_i x_i$. So the sample sum (equivalently the sample mean) is sufficient for μ — to estimate the mean, you need only the sum, not the individual data points. For the Bernoulli, the count of successes $\sum_i x_i$ is sufficient for p; for the exponential family in general, the sufficient statistic is exactly the $T(x)$ in its canonical form $\exp(\theta^\top T(x) - A(\theta))$.

28.2 Why this is the deep reason behind so much structure

Sufficiency explains a remarkable amount. It is why exponential-family MLEs reduce to **moment matching** (set the sufficient statistic's model expectation equal to its observed value). It is why those distributions admit compact summaries — you can throw away the raw data and keep only T without losing estimation power, the foundation of online and streaming estimation (update the running sufficient statistics, discard the data). And it underlies the **Rao–Blackwell theorem**: any estimator can be improved by conditioning on a sufficient statistic, which is why estimators that depend on the data only through sufficient statistics are preferred. Sufficiency is the principle that tells you what to remember and what you can safely forget.

Intuition

Sufficient statistics answer "what part of the data actually matters for this parameter?" — and the answer, for the exponential family, is the handful of moments in the distribution's canonical form. This is why fitting those models reduces to moment matching, why they support streaming estimation (keep the running sufficient statistics, forget the raw data), and why Rao–

Blackwellization improves estimators by conditioning on them. For a quant processing high-frequency data, sufficiency is practical: you can maintain a few running aggregates (sums, sums of squares, counts) and estimate parameters exactly as if you had stored every tick. The concept ties together the exponential family, maximum likelihood, and efficient computation — a unifying thread of the estimation theory in this note.

29 A Final Synthesis: Probability as the Common Language

Every method in the series rests on the probability assembled in this note. The losses are negative log-likelihoods (Gaussian gives squared error, Bernoulli gives cross-entropy), so choosing a loss is choosing a noise model. Regularization is a log-prior (Gaussian gives ridge, Laplace gives lasso), so penalizing is encoding belief. Evaluation is calibration and proper scoring rules, grounded in the information theory above. The covariance matrix's PSD-ness — forced by variances being nonnegative — is the probabilistic root of the positive definiteness behind convex losses, valid kernels, and well-posed regression. The limit theorems license learning from finite data, and concentration inequalities turn that license into quantitative generalization bounds. Bayes' theorem is the engine of updating, from the base-rate arithmetic that governs imbalanced classification to the conjugate updates behind smoothing.

Intuition

Probability is the grammar that makes the rest of machine learning a single language rather than a list of tricks. Maximum likelihood unifies the losses; Bayesian updating unifies the regularizers; information theory explains why those losses take their forms; the moment rules and PSD covariance connect to linear algebra and finance; and the limit theorems and concentration inequalities are the reason finite data can teach us anything. A practitioner fluent in this grammar reads any new method by asking three questions — what is its likelihood (loss), what is its prior (regularizer), and how is it evaluated (proper scoring) — and the method's assumptions, failure modes, and connections to everything else become legible. That fluency, built on the worked foundations of this note, is the real prerequisite the series was written to supply.

30 Consolidated Exercises

1. **(Bayes)** Re-derive the base-rate example for a test with sensitivity 0.95, specificity 0.90, prevalence 2%; compute the posterior after a positive and a second independent positive.

2. **(Variance)** For a three-asset equal-weight portfolio with given covariance matrix, compute the variance and identify the contribution of the off-diagonal (covariance) terms.

3. **(Exponential family)** Write the Poisson in exponential-family form; identify the natural parameter and verify $A'(\eta) = \mathbb{E}[X] = \lambda$.

4. **(Gaussian conditioning)** For a bivariate normal with correlation ρ, derive $\mathbb{E}[X_1 \mid X_2 = x_2]$ and show it equals the regression line; interpret the shrinkage when $|\rho| < 1$.

5. **(MLE)** Derive the MLE for the rate λ of an exponential sample; show it is $1/\bar{x}$ and check consistency.

6. **(Bias–variance)** Show the n-divisor sample variance is biased and compute the bias; explain why the $(n-1)$ divisor corrects it (Bessel's correction).

7. **(MAP)** Show that a Gaussian prior on regression coefficients yields ridge and find the explicit relationship between the prior variance and λ.

8. **(Multiple testing)** You backtest 50 strategies; the best has a t-statistic of 2.3 ($p \approx 0.02$). Compute the Bonferroni-adjusted threshold and assess whether the result survives.

9. **(Information)** Prove $D_{\mathrm{KL}}(p\|q) \geq 0$ using Jensen's inequality applied to $-\log$.

10. **(Martingale)** Show a symmetric random walk is a martingale and explain why no stopping rule yields positive expected gain.

End of the Probability & Statistics prerequisite. Together with Linear Algebra and Calculus & Optimization, this grounds the machine-learning algorithm series that follows.