Lecture 1b: Probability Theory

Sum and product rules, Bayes, the Gaussian, and Bayesian curve fitting

Jue Guo

University at Buffalo

May 7, 2026

Outline

Where we are

Recap (1a). Polynomial curve fitting exposed two recurring objects:

- the noise on the targets,
- the uncertainty in the weights $\mathbf{w}$ we are forced to commit to.

Both are best handled by probability theory, the tool that lets us *quantify* uncertainty rather than wave it away.

Plan for this deck

Build the calculus of probability from two rules, develop the Gaussian as our workhorse distribution, and re-derive everything we did in 1a from a probabilistic viewpoint – arriving at full *Bayesian* curve fitting.

A motivating example – two boxes of fruit

Two boxes – red and blue.

- Red contains 2 apples, 6 oranges.
- Blue contains 3 apples, 1 orange.

Pick a box at random (red 40%, blue 60%) and draw a fruit uniformly.

Two random variables. $B \in \{r, b\}$ is the box, $F \in \{a, o\}$ is the fruit.

Read directly from the setup.

$$p(B = r) = \tfrac{4}{10}, \quad p(B = b) = \tfrac{6}{10}; \qquad p(F = a \mid B = r) = \tfrac{1}{4}, \quad p(F = a \mid B = b) = \tfrac{3}{4}.$$

Question. You are told only that an apple was drawn. What is the probability that the box was red?

Two rules from a counting picture

Two random variables $X \in \{x_1, \ldots, x_M\}$ and $Y \in \{y_1, \ldots, y_L\}$. Run N trials; let n_{ij} be the count with $X = x_i,\ Y = y_j$, $c_i = \sum_j n_{ij}$, $r_j = \sum_i n_{ij}$.

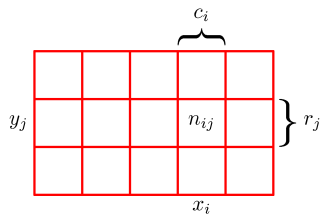

Figure 1.10

Define probabilities as fractions in the limit $N \to \infty$:

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}, \qquad p(X = x_i) = \frac{c_i}{N}, \qquad p(Y = y_j \mid X = x_i) = \frac{n_{ij}}{c_i}.$$

The sum and product rules

From $c_i = \sum_j n_{ij}$ and $n_{ij} = (n_{ij}/c_i) \cdot c_i$ we read off:

$$\text{sum rule:} \qquad p(X) = \sum_Y p(X, Y), \tag{1}$$

$$\text{product rule:} \qquad p(X, Y) = p(Y \mid X)\, p(X). \tag{2}$$

Vocabulary.

- $p(X, Y)$ – joint: "probability of X and Y".
- $p(X)$ – marginal: "probability of X", summing out Y.
- $p(Y \mid X)$ – conditional: "probability of Y given X".

Together with non-negativity and total probability one, these two rules build *everything* we will do this term.

Bayes' theorem

Apply the product rule symmetrically – $p(X, Y) = p(Y \mid X)\,p(X) = p(X \mid Y)\,p(Y)$:

$$p(Y \mid X) \;=\; \frac{p(X \mid Y)\,p(Y)}{p(X)}, \qquad p(X) \;=\; \sum_{Y} p(X \mid Y)\,p(Y). \tag{3}$$

Belief update. Given a *prior* $p(Y)$ (what we believe before seeing X) and a *likelihood* $p(X \mid Y)$ (how plausible X is under each Y), Bayes' theorem produces the *posterior* $p(Y \mid X)$. The denominator $p(X)$ is just the normalising constant – the *evidence* – chosen so the posterior sums to one.

Why this matters

Almost every probabilistic model in this course is one prior, one likelihood, and an application of this single equation.

Boxes and fruit, worked

Marginal probability of an apple, by the sum rule:

$$p(F = a) \;=\; p(F = a \mid B = r)\, p(B = r) + p(F = a \mid B = b)\, p(B = b) \;=\; \tfrac{1}{4} \cdot \tfrac{4}{10} + \tfrac{3}{4} \cdot \tfrac{6}{10} \;=\; \tfrac{11}{20}.$$

Posterior of the box, by Bayes' theorem:

$$p(B = r \mid F = a) \;=\; \frac{p(F = a \mid B = r)\, p(B = r)}{p(F = a)} \;=\; \frac{(1/4)(4/10)}{11/20} \;=\; \tfrac{2}{11}.$$

Prior $p(B = r) = 4/10$ shifts to posterior $p(B = r \mid F = a) = 2/11 \approx 0.18$ – seeing an apple makes the apple-poor red box *less* likely.

Independence

Two random variables are independent when the joint factorises:

$$p(X, Y) \;=\; p(X)\, p(Y). \tag{4}$$

By the product rule this is equivalent to

$$p(Y \mid X) = p(Y),$$

i.e. knowing X tells you nothing new about Y. **Where this is used.**

- Models of N i.i.d. data points: $p(\mathcal{D} \mid \theta) = \prod_n p(x_n \mid \theta)$.
- Naive Bayes classifiers: features assumed conditionally independent given class.
- Graphical models: missing edges encode conditional independencies.

Continuous variables: PDFs and CDFs

Replace probabilities with a probability density $p(x)$ such that

$$\Pr(x \in (a, b)) = \int_a^b p(x)\,dx, \qquad p(x) \geq 0, \qquad \int_{-\infty}^{\infty} p(x)\,dx = 1. \tag{5}$$

The cumulative distribution function is $P(z) = \int_{-\infty}^{z} p(x)\,dx$, and $P'(x) = p(x)$. **Sum and product rules carry over** – replace sums with integrals:

$$p(x) = \int p(x, y)\,dy, \qquad p(x, y) = p(y \mid x)\,p(x).$$

Mixed discrete/continuous works as you'd expect; for discrete x, $p(x)$ is a *probability mass function*.

Change of variables

Suppose $x = g(y)$ with g smooth and monotonic, and let p_x, p_y be the densities of x, y. The probability mass in matching infinitesimal intervals must agree, so

$$p_y(y) \;=\; p_x\big(g(y)\big)\left|\frac{dg(y)}{dy}\right|. \tag{6}$$

The Jacobian factor accounts for stretching $dy \to dx = g'(y)\,dy$. **Multivariate version.** For $\mathbf{x} = g(\mathbf{y})$,

$$p_y(\mathbf{y}) \;=\; p_x\big(g(\mathbf{y})\big)\left|\det \tfrac{\partial g_i}{\partial y_j}\right|.$$

This identity will reappear whenever we reparameterise distributions – e.g. in normalising flows and in deriving the multivariate Gaussian.

Expectations

The expectation of $f(x)$ under $p(x)$ is

$$\mathbb{E}[f] \;=\; \sum_{x} p(x) f(x) \qquad \text{or} \qquad \mathbb{E}[f] \;=\; \int p(x) f(x)\, dx. \tag{7}$$

Monte Carlo bridge. Drawing N i.i.d. samples $x_1, \ldots, x_N \sim p$,

$$\mathbb{E}[f] \;\simeq\; \frac{1}{N} \sum_{n=1}^{N} f(x_n),$$

exact in the limit $N \to \infty$. Underlies essentially every approximate-inference scheme we'll meet. **Notation for several variables.** $\mathbb{E}_x[f(x,y)] = \int p(x) f(x,y)\, dx$ averages over x for fixed y; conditional expectations replace the marginal with a conditional.

Variance and covariance

Variance of $f(x)$:

$$\text{var}[f] \;=\; \mathbb{E}\big[(f(x) - \mathbb{E}[f(x)])^2\big] \;=\; \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2. \tag{8}$$

Setting $f(x) = x$ gives $\text{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2$. **Covariance** of two scalars:

$$\text{cov}[x, y] \;=\; \mathbb{E}_{x,y}[(x - \mathbb{E}[x])(y - \mathbb{E}[y])] \;=\; \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\,\mathbb{E}[y]. \tag{9}$$

Independence $\Rightarrow$ zero covariance; the converse fails in general. **Vector form.** For $\mathbf{x} \in \mathbb{R}^D$, $\text{cov}[\mathbf{x}] = \mathbb{E}[\mathbf{x}\mathbf{x}^\top] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}]^\top$ is symmetric, positive semi-definite – the same matrix that parameterises a multivariate Gaussian.

Two readings of the same calculus

Up to now: probabilities are long-run frequencies of repeatable events.

The Bayesian move. Read probabilities as degrees of belief. Cox's axioms guarantee that any consistent calculus of belief obeys the same sum and product rules. *Only the interpretation changes.*

	Frequentist	Bayesian
What is random?	Data $\mathcal{D}$	Parameters $\mathbf{w}$
Parameters $\mathbf{w}$ are ...	fixed, unknown	uncertain, with a prior
Output	point estimate + error bars	full posterior
Error bars from	resampling (e.g. bootstrap)	posterior credible region
Workhorse	maximum likelihood	Bayes' theorem

Both share the likelihood $p(\mathcal{D} \mid \mathbf{w})$ – they differ in what is conditioned on.

Posterior $\propto$ likelihood $\times$ prior

Treat the weight vector $\mathbf{w}$ as a random variable.

- Prior $p(\mathbf{w})$: beliefs before seeing data.
- Likelihood $p(\mathcal{D} \mid \mathbf{w})$: plausibility of $\mathcal{D}$ at each $\mathbf{w}$.

$$p(\mathbf{w} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \mathbf{w})\, p(\mathbf{w})}{p(\mathcal{D})}, \qquad p(\mathcal{D}) = \int p(\mathcal{D} \mid \mathbf{w})\, p(\mathbf{w})\, d\mathbf{w}. \tag{10}$$

Often summarised as

$$\text{posterior} \;\propto\; \text{likelihood} \times \text{prior}.$$

Maximum likelihood: $\mathbf{w}_{\mathrm{ML}} = \arg\max_{\mathbf{w}} p(\mathcal{D} \mid \mathbf{w})$. Equivalent to minimising $-\ln p(\mathcal{D} \mid \mathbf{w})$ – which, under Gaussian noise (next), turns out to be exactly least squares.

Priors in practice

The prior is the strength *and* the weakness of the Bayesian view.

- Lets us inject knowledge: smoothness, sparsity, physical constraints.
- Can mislead if chosen badly – "garbage in, garbage out".

Common choices.

- Conjugate priors: chosen so that the posterior is in the same family as the prior. Closed-form updates; we'll see this for Gaussians shortly.
- Noninformative priors: used when little is known a priori (e.g. a flat or scale-invariant prior).

When closed forms fail. Outside conjugate cases, the integrals defining $p(\mathcal{D})$ and the posterior have no analytic solution. Two big tools:

- Markov chain Monte Carlo (MCMC) – sampling.
- Variational inference – optimisation.

Both are central later in the course.

Univariate Gaussian

For $x \in \mathbb{R}$,

$$\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \qquad \beta \equiv 1/\sigma^2 \text{ (precision).} \tag{11}$$

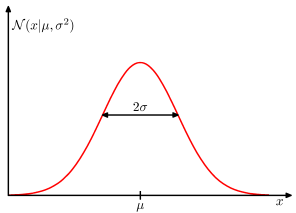

Figure 1.13

$\mathbb{E}[x] = \mu, \quad \mathbb{E}[x^2] = \mu^2 + \sigma^2, \quad \text{var}[x] = \sigma^2.$

Multivariate Gaussian

For $\mathbf{x} \in \mathbb{R}^D$, with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$,

$$\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}. \tag{12}$$

Geometry. Level sets are ellipsoids; the principal axes are the eigenvectors of $\boldsymbol{\Sigma}$, with semi-axis lengths $\sqrt{\lambda_i}$. **Why so central?**

- Closed under marginalisation, conditioning, and linear transformations.
- Conjugate to itself for unknown mean.
- Maximum-entropy distribution given mean and covariance.
- Central limit theorem – emerges as the limiting distribution of sums of i.i.d. variables.

Almost every closed-form Bayesian computation in this course rides on these properties.

Maximum likelihood for the Gaussian

Observe N i.i.d. samples $\mathcal{D} = \{x_1, \ldots, x_N\}$ from $\mathcal{N}(\mu, \sigma^2)$. Likelihood and log-likelihood:

$$p(\mathcal{D} \mid \mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n \mid \mu, \sigma^2),$$

$$\ln p(\mathcal{D} \mid \mu, \sigma^2) = -\tfrac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 - \tfrac{N}{2} \ln \sigma^2 - \tfrac{N}{2} \ln(2\pi).$$

Setting derivatives to zero:

$$\mu_{\mathrm{ML}} = \tfrac{1}{N} \sum_n x_n, \qquad \sigma^2_{\mathrm{ML}} = \tfrac{1}{N} \sum_n (x_n - \mu_{\mathrm{ML}})^2. \tag{13}$$

The sample mean and sample variance fall out; the appearance of $\sum_n (x_n - \mu)^2$ here is what will let us connect MLE to least squares on the next slide.

Bias of σ^2_{ML}

Taking expectations under the true distribution:

$$\mathbb{E}[\mu_{\mathrm{ML}}] = \mu, \qquad \mathbb{E}[\sigma^2_{\mathrm{ML}}] = \tfrac{N-1}{N}\,\sigma^2. \tag{14}$$

μ_{ML} is unbiased; σ^2_{ML} systematically *underestimates* the true variance because it measures spread around the sample mean rather than the true mean. **Unbiased correction.**

$$\widetilde{\sigma}^2 \;=\; \tfrac{N}{N-1}\,\sigma^2_{\mathrm{ML}} \;=\; \tfrac{1}{N-1}\sum_n (x_n - \mu_{\mathrm{ML}})^2.$$

Why this matters. ML's bias-toward-overconfidence is the simplest example of the overfitting that plagued the $M = 9$ polynomial in 1a: ML fits the data, not the truth. The Bayesian remedy – integrating over parameter uncertainty – is coming up.

Derivation – where the $(N-1)/N$ bias comes from

Fix true mean μ, variance σ^2; the ML estimators are $\mu_{\mathrm{ML}} = \frac{1}{N}\sum_n x_n$ and $\sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_n (x_n - \mu_{\mathrm{ML}})^2$.

Step 1. Write the sample-mean residual as a sum of two pieces:

$$x_n - \mu_{\mathrm{ML}} \;=\; (x_n - \mu) - (\mu_{\mathrm{ML}} - \mu).$$

Step 2. Square and sum over n. The cross term vanishes because $\sum_n (x_n - \mu) = N(\mu_{\mathrm{ML}} - \mu)$:

$$\sum_n (x_n - \mu_{\mathrm{ML}})^2 \;=\; \sum_n (x_n - \mu)^2 \;-\; N\,(\mu_{\mathrm{ML}} - \mu)^2.$$

Step 3. Take expectations. $\mathbb{E}[(x_n - \mu)^2] = \sigma^2$ and, by the i.i.d. Gaussian assumption, $\mathrm{var}[\mu_{\mathrm{ML}}] = \sigma^2/N$:

$$\mathbb{E}\left[\sum_n (x_n - \mu_{\mathrm{ML}})^2\right] \;=\; N\sigma^2 \;-\; N \cdot \tfrac{\sigma^2}{N} \;=\; (N-1)\,\sigma^2.$$

Step 4. Divide by N:

$$\mathbb{E}[\sigma^2_{\mathrm{ML}}] \;=\; \tfrac{N-1}{N}\,\sigma^2.$$

Reading. ML measures spread around μ_{ML}, which is *itself* drawn closer to the data than the true μ. One degree of freedom has been "spent" fitting the mean, leaving $N-1$ degrees of freedom for the variance.

Concrete example – Gaussian MLE on 4 points

Suppose the truth is $\mathcal{N}(\mu = 0, \sigma^2 = 1)$ and we observe four points:

$$\mathcal{D} = \{-0.8,\ 0.2,\ 1.0,\ -0.4\}.$$

Sample mean. $\mu_{\text{ML}} = \frac{1}{4}(-0.8 + 0.2 + 1.0 - 0.4) = 0.00$. Lucky here – in general it scatters around the true μ.

ML variance. $\sigma^2_{\text{ML}} = \frac{1}{4}\left[(-0.8)^2 + 0.2^2 + 1.0^2 + (-0.4)^2\right] = \frac{1}{4}(0.64 + 0.04 + 1.00 + 0.16) = 0.46$.

Unbiased estimate. $\widetilde{\sigma}^2 = \frac{N}{N-1}\sigma^2_{\text{ML}} = \frac{4}{3} \cdot 0.46 \approx 0.61$.

Diagnosis. The truth is $\sigma^2 = 1$; ML reports 0.46 (under by a factor of 2); the unbiased correction reports 0.61. The ratio of the two estimates is $(N-1)/N = 3/4$ exactly, as the bias formula predicts. The bias shrinks like $1/N$, but for small N it is severe – and "N small relative to model capacity" is exactly the regime where overfitting bites.

A probabilistic regression model

Place a Gaussian noise model on top of the polynomial of 1a:

$$p(t \mid x, \mathbf{w}, \beta) \;=\; \mathcal{N}\big(t \mid y(x, \mathbf{w}),\, \beta^{-1}\big), \tag{15}$$

where β is the noise precision. For an i.i.d. training set $\mathcal{D} = \{(x_n, t_n)\}$, the likelihood is

$$p(\mathcal{D} \mid \mathbf{w}, \beta) \;=\; \prod_{n=1}^{N} \mathcal{N}\big(t_n \mid y(x_n, \mathbf{w}),\, \beta^{-1}\big), \tag{16}$$

and the negative log-likelihood is

$$-\ln p(\mathcal{D} \mid \mathbf{w}, \beta) \;=\; \tfrac{\beta}{2} \sum_{n=1}^{N} \big\{y(x_n, \mathbf{w}) - t_n\big\}^2 \;-\; \tfrac{N}{2} \ln \beta + \tfrac{N}{2} \ln(2\pi).$$

Maximum likelihood = least squares

Maximising the log-likelihood over $\mathbf{w}$:

- the β-dependent terms drop out,
- what remains is exactly the sum-of-squares error from 1a.

$$\mathbf{w}_{\mathrm{ML}} \;=\; \arg\min_{\mathbf{w}} \tfrac{1}{2} \sum_{n} \big\{ y(x_n, \mathbf{w}) - t_n \big\}^2 .$$

Given $\mathbf{w}_{\mathrm{ML}}$, maximising over β gives

$$\tfrac{1}{\beta_{\mathrm{ML}}} \;=\; \tfrac{1}{N} \sum_{n=1}^{N} \big\{ y(x_n, \mathbf{w}_{\mathrm{ML}}) - t_n \big\}^2 , \tag{17}$$

the empirical residual variance. **Predictive distribution** (point-estimate version):

$$p(t \mid x, \mathbf{w}_{\mathrm{ML}}, \beta_{\mathrm{ML}}) \;=\; \mathcal{N}\big(t \mid y(x, \mathbf{w}_{\mathrm{ML}}),\, \beta_{\mathrm{ML}}^{-1}\big).$$

Adding a prior ⇒ MAP = ridge

Place a zero-mean isotropic Gaussian prior on the weights,

$$p(\mathbf{w} \mid \alpha) \;=\; \mathcal{N}\!\left(\mathbf{w} \mid \mathbf{0},\, \alpha^{-1}\mathbf{I}\right) \;\propto\; \exp\!\left\{-\tfrac{\alpha}{2}\,\mathbf{w}^{\top}\mathbf{w}\right\}. \tag{18}$$

The negative log-posterior, $-\ln p(\mathbf{w} \mid \mathcal{D}) = -\ln p(\mathcal{D} \mid \mathbf{w}) - \ln p(\mathbf{w}) + \text{const}$, becomes

$$\tfrac{\beta}{2}\sum_{n}\left\{y(x_n, \mathbf{w}) - t_n\right\}^2 \;+\; \tfrac{\alpha}{2}\,\mathbf{w}^{\top}\mathbf{w}. \tag{19}$$

The maximum-a-posteriori (MAP) estimate is exactly the ridge estimate of 1a, with regularisation parameter $\lambda = \alpha/\beta$.

Take-home

Regularisation isn't an ad hoc trick – it's a Gaussian prior on the weights, wearing a different hat.

Beyond point estimates

MAP still commits to a single $\mathbf{w}$. The fully Bayesian approach *integrates* over weights:

$$p(t \mid x, \mathcal{D}) \;=\; \int p(t \mid x, \mathbf{w})\, p(\mathbf{w} \mid \mathcal{D})\, d\mathbf{w}. \tag{20}$$

For our Gaussian-noise / Gaussian-prior model the posterior $p(\mathbf{w} \mid \mathcal{D})$ is itself Gaussian (conjugacy), and the predictive distribution comes out in closed form:

$$p(t \mid x, \mathcal{D}) \;=\; \mathcal{N}\big(t \mid m(x),\, s^2(x)\big), \tag{21}$$

with explicit formulas for the mean $m(x)$ and variance $s^2(x)$ (see PRML §1.2.6 / lecture notes). The mean $m(x)$ tracks the regularised least-squares fit; the variance $s^2(x)$ is the genuinely new object.

Two sources of predictive uncertainty

The predictive variance splits cleanly:

$$s^2(x) = \underbrace{\beta^{-1}}_{\text{noise on } t} + \underbrace{\boldsymbol{\phi}(x)^\top \mathbf{S}\, \boldsymbol{\phi}(x)}_{\text{uncertainty in } \mathbf{w}}, \tag{22}$$

where $\boldsymbol{\phi}(x) = (1, x, x^2, \ldots, x^M)^\top$ and $\mathbf{S}$ is the posterior covariance of $\mathbf{w}$.

- First term: irreducible noise, present even with infinite data.
- Second term: shrinks as we collect more data and $\mathbf{S}$ shrinks.
- Crucially, the second term depends on x: predictions are tighter where the training inputs concentrated and looser elsewhere.
- This honest report of where the model is and is not confident is precisely what point estimates cannot give us.

Bayesian predictive curve, sketched

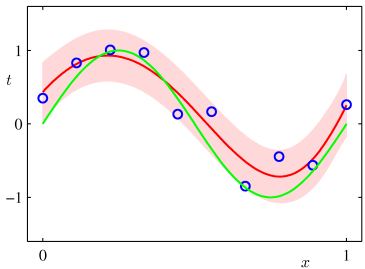

Figure 1.17

Red curve: posterior predictive mean $m(x)$. Pink band: one-standard-deviation region $m(x) \pm s(x)$. Compare with PRML Fig. 1.17.

Takeaways

- Two rules + Bayes are the entire calculus of probability, for both discrete and continuous variables.
- Bayesian probabilities reinterpret the same calculus as a consistent calculus of belief.
- The Gaussian is our workhorse: closed under marginalisation, conditioning, and linear transformations; conjugate to itself.
- Maximum likelihood under Gaussian noise = least squares; MAP under a Gaussian weight prior = ridge.
- Going fully Bayesian replaces a point estimate with a predictive distribution that splits irreducible noise from parameter uncertainty.

Reading

Bishop, PRML (2006), §1.2 in full.