

Recitation 2

Mathematical foundations

CS-UY 4563: Introduction to Machine Learning

Daniel Enriquez

Fall 2026

Probability and events

An **outcome** is one possible result of an experiment. An **event** is a set of outcomes whose probability we want to describe.

Draw one of ten marbles, each equally likely. Three are red and seven are blue. The event “blue” contains seven outcomes, so

$$\Pr(\text{blue}) = \frac{7}{10}.$$

Probabilities are nonnegative, total one over all outcomes, and add for events that cannot occur together.

Counting outcomes gives probabilities here because every marble is equally likely to be drawn.

Random variables

A **random variable** assigns a number to each outcome. Some marbles also have a painted stripe; record color and marking as

$$U = \begin{cases} 1 & \text{blue,} \\ 0 & \text{red,} \end{cases} \quad V = \begin{cases} 1 & \text{striped,} \\ 0 & \text{plain.} \end{cases}$$

Uppercase U, V name the variables; lowercase u, v name possible values. A probability mass function records

$$p_U(u) = \Pr(U = u), \quad p_U(1) = 0.70.$$

The same notation can describe a binary feature or class label. A variable names what is recorded, not a particular observed value.

Joint and marginal distributions

A **joint distribution** describes two variables together:

$$p_{U,V}(u, v) = \Pr(U = u, V = v).$$

	$V = 0$ (plain)	$V = 1$ (striped)	$p_U(u)$
$U = 0$ (red)	0.10	0.20	0.30
$U = 1$ (blue)	0.30	0.40	0.70
$p_V(v)$	0.40	0.60	1

A **marginal** adds over the other variable:

$$p_U(u) = \sum_v p_{U,V}(u, v).$$

Conditional probability

Conditioning restricts attention to an observed group and rescales its probabilities to total one. For an event B with $\Pr(B) > 0$,

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)}.$$

Among the six striped marbles, four are blue:

$$\Pr(U = 1 \mid V = 1) = \frac{0.40}{0.60} = \frac{2}{3}.$$

The group has changed, so the probability need not remain 0.70.

The bar means “given”; $A \cap B$ means both events occur. In classification, $\Pr(Y = 1 \mid X = x)$ conditions the label on observed features.

Expectation and variance

Expectation averages values using their probabilities:

$$\mathbb{E}[U] = \sum_u u p_U(u).$$

Variance measures average squared deviation from that mean:

$$\text{Var}(U) = \mathbb{E}[(U - \mu)^2] = \mathbb{E}[U^2] - \mu^2, \quad \mu = \mathbb{E}[U].$$

For the marble's color, $\Pr(U = 1) = 0.70$:

$$\mathbb{E}[U] = 0.70, \quad \mathbb{E}[U^2] = 0.70, \quad \text{Var}(U) = 0.21.$$

Use integrals for densities. Assume finite second moments for variance. Linearity of expectation does not require independence.

Why the mean minimizes squared distance

Choose a constant a to minimize $\mathbb{E}[(U - a)^2]$. Write $\mu = \mathbb{E}[U]$:

$$\begin{aligned}\mathbb{E}[(U - a)^2] &= \mathbb{E}[((U - \mu) + (\mu - a))^2] \\ &= \text{Var}(U) + 2(\mu - a)\mathbb{E}[U - \mu] + (\mu - a)^2 \\ &= \text{Var}(U) + (\mu - a)^2.\end{aligned}$$

Deviations from the mean average to zero, so the cross term disappears. Only the extra square depends on a ; choosing $a = \mu$ removes it and leaves the variance.

Within a group, use that group's probabilities and mean. The same proof then gives the best constant for that group.

Vectors, inner products, and matrices

A **vector** is an ordered list of coordinates, written here as a column. Its dimension is the number of coordinates.

$$x = \begin{bmatrix} 2 \\ 3 \end{bmatrix} \in \mathbb{R}^2, \quad w = \begin{bmatrix} 4 \\ 1 \end{bmatrix}, \quad w^T x = 4(2) + 1(3) = 11.$$

The **inner product** multiplies matching entries and adds them. This is how a linear model combines feature values with coefficients.

A **matrix** $X \in \mathbb{R}^{n \times d}$ has n rows and d columns. A dataset can store one observation per row and one feature per column.

The transpose T exchanges rows and columns, so w^T is a row.

Matrices as linear transformations

Multiplication combines the columns using the input coordinates:

$$\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = u_1 \begin{bmatrix} 1 \\ 2 \end{bmatrix} + u_2 \begin{bmatrix} 2 \\ 4 \end{bmatrix}.$$

For example, the input $(1, 1)^\top$ produces $(3, 6)^\top$.

If A has columns $a_1, \dots, a_d$, the same calculation is

$$Au = u_1 a_1 + \dots + u_d a_d.$$

The column space is the set of outputs these combinations can produce.

A linear transformation preserves sums and scaling: $A(au + bv) = aAu + bAv$.

Column space and null space

For $A = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$, every output lies on one line:

$$Au = (u_1 + 2u_2) \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

The column space has dimension one, so $\text{rank}(A) = 1$.

$$A \begin{bmatrix} -2 \\ 1 \end{bmatrix} = 0.$$

This nonzero vector lies in the null space. Adding any multiple of it to an input leaves the output unchanged.

Could an inverse recover which input produced the output?

Eigenvectors and eigenvalues

For $A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$, two directions are easy to follow:

$$A \begin{bmatrix} 1 \\ 1 \end{bmatrix} = 3 \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad A \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

An eigenvector satisfies $Aq = \lambda q$, with $q \neq 0$: it stays on its line, scaled by λ . Both displayed vectors have length $\sqrt{2}$, so their unit versions are

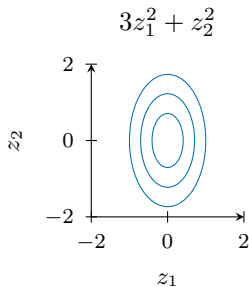
$$q_1 = \frac{1}{\sqrt{2}}(1, 1)^T, \quad q_2 = \frac{1}{\sqrt{2}}(1, -1)^T.$$

These perpendicular unit vectors form an orthonormal basis for the plane, with eigenvalues 3 and 1.

A quadratic in the new coordinates

Using the unit directions q_1, q_2 , write $u = z_1 q_1 + z_2 q_2$.

$$u^T A u = 2u_1^2 + 2u_1 u_2 + 2u_2^2 = 3z_1^2 + z_2^2.$$



Each coordinate $z_j = q_j^T u$ measures the component along q_j . The ellipses join points of equal value; they are narrower along the steeper direction, z_1 .

Here $u^T A u > 0$ for every $u \neq 0$: A is positive definite.

The gradient

The **gradient** is the vector of partial derivatives: slopes with respect to each coordinate, holding the others fixed.

$$f(u) = \frac{1}{2}(u_1^2 + 10u_2^2), \quad \nabla f(u) = \begin{bmatrix} u_1 \\ 10u_2 \end{bmatrix}.$$

At $u = (1, 2)^\top$, a step $v = (0.01, -0.01)^\top$ changes f by about

$$\nabla f(u)^\top v = 1(0.01) + 20(-0.01) = -0.19.$$

The negative gradient points along the steepest local decrease per unit Euclidean distance.

Direct substitution gives -0.18945 . The linear approximation improves as the displacement shrinks.

The Hessian and curvature

The **Hessian** is the matrix of second partial derivatives. It describes how the gradient changes as the point moves:

$$H = \nabla^2 f(u) = \begin{bmatrix} 1 & 0 \\ 0 & 10 \end{bmatrix} \quad \text{for } f(u) = \frac{1}{2}(u_1^2 + 10u_2^2).$$

The diagonal entries give coordinate curvatures 1 and 10. The zeros mean that changing one coordinate leaves the other slope unchanged.

$$f(u + v) - f(u) = \nabla f(u)^\top v + \frac{1}{2}v^\top H v.$$

For this quadratic, slope and curvature give the exact change.

Along a unit direction a , curvature is $a^\top H a$. Along a unit eigenvector, it equals that eigenvalue.

Convexity and a global minimum

A function is **convex** when its graph lies below each chord:

$$f((1-t)u + tv) \leq (1-t)f(u) + tf(v), \quad 0 \leq t \leq 1.$$

The domain must contain the segment between any two of its points. For a differentiable convex function, each tangent plane is a global lower bound:

$$f(v) \geq f(u) + \nabla f(u)^T(v - u).$$

At $\nabla f(u) = 0$, this says $f(v) \geq f(u)$ for every v , so u is a **global minimizer**.

A stationary point has zero gradient. Without convexity, that alone does not guarantee a minimum.

Gradient descent on a quadratic

To find a minimum, start at $u^{[0]}$ and repeatedly step against the current gradient:

$$u^{[t+1]} = u^{[t]} - \alpha \nabla f(u^{[t]}), \quad \alpha > 0.$$

For $f(u) = \frac{1}{2}(u_1^2 + 10u_2^2)$,

$$u_1^{[t+1]} = (1 - \alpha)u_1^{[t]}, \quad u_2^{[t+1]} = (1 - 10\alpha)u_2^{[t]}.$$

At $\alpha = 0.1$, a start at $(2, 1)^\top$ follows

$$(2, 1)^\top \longrightarrow (1.8, 0)^\top \longrightarrow (1.62, 0)^\top.$$

What would happen to the second coordinate at $\alpha = 0.3$?

Square brackets label iterations, not powers. All coordinates are updated from the same current point.

Curvature limits the step size

Along an eigenvector of curvature λ_j , each update multiplies the error measured from the minimizer by $1 - \alpha\lambda_j$. That error shrinks when

$$|1 - \alpha\lambda_j| < 1 \quad \Longleftrightarrow \quad 0 < \alpha\lambda_j < 2.$$

The largest curvature gives the tightest bound:

$$0 < \alpha < \frac{2}{\lambda_{\max}}.$$

For curvatures 1 and 10, the range is $0 < \alpha < 0.2$. At $\alpha = 0.1$, the error multipliers are 0.9 and 0.

For a fixed positive-definite quadratic, this condition ensures convergence from every starting point.

Rescaling the coordinates

To balance the two curvatures, substitute

$$u_1 = z_1,$$
$$u_2 = z_2/\sqrt{10}.$$

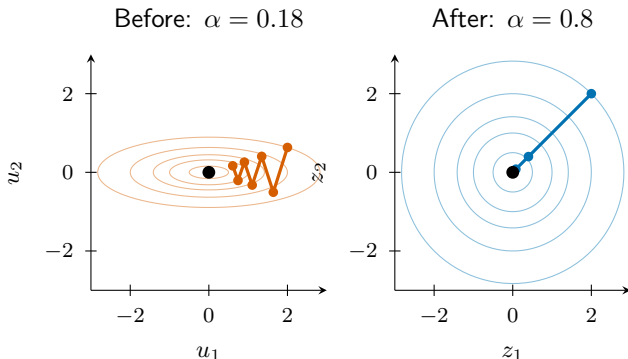
Then

$$f(u) = \frac{1}{2}(u_1^2 + 10u_2^2)$$

becomes

$$\tilde{f}(z) = \frac{1}{2}(z_1^2 + z_2^2).$$

Corresponding starting points and the same contour values. Six updates in each coordinate system, with the displayed step sizes.



Standardizing a feature

Standardization subtracts each feature's training mean and divides by its training standard deviation:

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, \quad s_j > 0.$$

Here i indexes observations and j feature columns. Each training column then has mean zero and variance one.

If trip times have mean 30 minutes and standard deviation 10, a 50-minute trip becomes 2. Using seconds gives the same value:

$$(3000 - 1800)/600 = 2.$$

Fit the means and scales on training data; reuse them for validation and test data. A constant column needs separate handling.

Why scale the dataset?

For squared-error regression, feature columns determine the curvature:

$$L(w) = \frac{1}{2n} \|Xw - y\|_2^2, \quad H = \frac{1}{n} X^\top X.$$

After centering the columns, H_{jj} is feature j 's empirical variance. Unequal variances can restrict the gradient step size. Standardization makes the diagonal entries one; correlations can still leave a narrow bowl.

For distances, scaling compares differences relative to each feature's training variation. It changes the geometry without adding information.

With a free intercept and no regularization, linear predictions can be preserved by changing coefficients. A fixed coefficient penalty need not be preserved.

Practice: predicting a trip duration

In this constructed distribution, a trip takes $Y = 20$ or 40 minutes:

Joint probability	$Y = 20$	$Y = 40$
Weekday	0.30	0.20
Weekend	0.10	0.40

- Which constant prediction minimizes expected squared loss?
- If the day type is available, which two predictions minimize it?
- How much expected squared loss does that extra information remove?

10 minutes. Work together for 4 minutes, then compare your calculations and explain the remaining error.

Practice: why should feature units matter?

A least-squares fit uses $w_1x_1 + w_2x_2$. Suppose its objective has Hessian

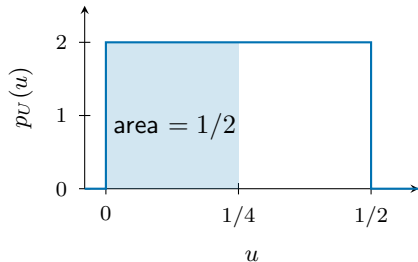
$$H = \begin{bmatrix} 1 & 0 \\ 0 & 100 \end{bmatrix}.$$

- Which fixed step sizes give convergence from every start? At $\alpha = 0.01$, which error component shrinks slowly?
- Replace x_2 by $x_2/10$. How must w_2 change to preserve predictions? What happens to the two curvatures?

If time permits: would scaling fix nearly identical features?

10 minutes. No intercept or regularizer. Use the diagonal quadratic as a guide.

Further reading: probability density



A uniform point on $[0, 1/2]$ has density 2 on that interval.

$$\Pr(U \leq 1/4) = \int_0^{1/4} 2 \, du = \frac{1}{2}.$$

The shaded area is the probability. The height is a density, so it can exceed one.

For a variable with a density, every exact point has probability zero. Interval probabilities come from integrals.

Further reading: Bayes' rule

Bayes' rule expresses a conditional probability through the reverse condition:

$$\Pr(Y = 1 \mid X = 1) = \frac{\Pr(X = 1 \mid Y = 1) \Pr(Y = 1)}{\Pr(X = 1)}.$$

For spam detection, let $Y = 1$ mean spam and $X = 1$ mean that a word is present. The prior $\Pr(Y = 1)$ gives the class frequency; the likelihood $\Pr(X = 1 \mid Y = 1)$ gives the word frequency within that class.

Their product gives the joint probability. Dividing by the word's overall frequency compares spam with all emails containing it.

Assume positive conditioning probabilities. Their values must be learned from data.

Further reading: covariance

With finite second moments, the covariance matrix averages products of centered coordinates:

$$\Sigma_{ij} = \mathbb{E}[(z_i - \mu_i)(z_j - \mu_j)], \quad \mu = \mathbb{E}[z].$$

The diagonal entries are variances. For a weighted combination $a^T z$, the full matrix gives

$$\text{Var}(a^T z) = a^T \Sigma a \geq 0.$$

Because a variance cannot be negative, Σ is positive semidefinite. Along a unit eigenvector q , this gives $\text{Var}(q^T z) = q^T \Sigma q = \lambda$: the eigenvalue is the variance in that direction.

Further reading: the coordinate change

Write $u = Sz$ with $S = \text{diag}(1, 1/\sqrt{10})$; diag places those entries on the diagonal. For $\tilde{f}(z) = f(Sz)$, the chain rule gives

$$\nabla \tilde{f}(z) = S^T \nabla f(Sz), \quad \tilde{H} = S^T H S.$$

For $H = \text{diag}(1, 10)$ and our chosen S , $\tilde{H} = I$.

A gradient step in z , mapped back to u , becomes

$$u^+ = u - \alpha S S^T \nabla f(u).$$

The minimizing points correspond, while the update directions can change.

Diagonal scaling can address unequal coordinate scales. It need not remove correlations or improve conditioning in every problem.