

Session 3: Independence, Conditional Independence, and Acting on Belief

Movement I: Belief and Conditioning

Instructor: John Lipor

Sessions 1 and 2 built the machine that moves belief: conditioning, and its industrial form, Bayes' rule. One assumption rode along underneath the repeated-search formula without inspection. When we said “ m failed looks at a cell multiply its odds by α^m ,” we quietly treated those looks as *independent given the target's location*. Today we ask when that is legitimate, discover that the same two observations can be independent in one breath and dependent in the next, and then close the loop the course keeps promising: we turn a posterior into an *action*. The session thesis has two clauses: *independence is a modeling assumption, not a default, and a posterior is only as valuable as the action it improves*.

1 Launch

[**Thread A**] Two sonar passes sweep the same seafloor region; both come back empty. Were those two pieces of evidence independent? The honest answer is *no and yes*, and the gap between those two answers is the most important modeling idea in this course. *Yes*, in the sense that once we fix where the wreck actually is, the second sonar ping knows nothing about the first: the sensor's noise is fresh each pass. *No*, in the sense that we do *not* know where the wreck is, and the first empty pass has already nudged our belief, so it changes what we expect from the second. By the end of the hour we will make both answers precise, name the property that reconciles them (conditional independence), and then use the belief they produce to *decide where to look next*, at which point Movement I is complete: a belief, moved by evidence, converted into an action.

2 Development

2.1 Independence of events

Conditioning measures how much one event tells us about another. When it tells us *nothing*, the events are independent.

Definition 1 (Independence). *Two events A and B are independent if*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \mathbb{P}(B). \quad (1)$$

A collection $\{A_i\}_{i \in I}$ is **(mutually) independent** if for every finite subset $J \subseteq I$,

$$\mathbb{P}\left(\bigcap_{i \in J} A_i\right) = \prod_{i \in J} \mathbb{P}(A_i). \quad (2)$$

When $\mathbb{P}(B) > 0$, independence is equivalent to $\mathbb{P}(A|B) = \mathbb{P}(A)$: learning B leaves belief in A untouched. That is the reading to carry around, but the product form above is the definition because it is symmetric in A and B and survives $\mathbb{P}(B) = 0$.

Pitfall. Do not conflate *disjoint* with *independent*. Disjoint events satisfy $A \cap B = \emptyset$, a statement about the sets alone; independence is a statement about the sets *and* the probability measure. In fact two events with positive probability that are disjoint are as far from independent as possible: $\mathbb{P}(A \cap B) = 0$ while $\mathbb{P}(A)\mathbb{P}(B) > 0$, and observing one rules the other out.

Example 1 (Testing independence numerically). *Roll a fair die. Let $A = \{1, 2\}$ and $B = \{2, 4, 6\}$, so $\mathbb{P}(A) = 1/3$ and $\mathbb{P}(B) = 1/2$. Their intersection is $A \cap B = \{2\}$, so $\mathbb{P}(A \cap B) = 1/6 = (1/3)(1/2) = \mathbb{P}(A)\mathbb{P}(B)$: the events are independent, even though nothing about “low” and “even” looks obviously unrelated. Independence is a numerical fact to be checked, not an intuition to be trusted.*

2.2 Pairwise is not mutual

Mutual independence asks more than independence of every pair. The gap is easiest to see with two fair coins. Toss them and define

$$A = \{\text{coin 1 is heads}\}, \quad B = \{\text{coin 2 is heads}\}, \quad C = \{\text{the two coins agree}\}. \quad (3)$$

Each has probability $1/2$. Every *pair* is independent:

$$\mathbb{P}(A \cap B) = \frac{1}{4} = \mathbb{P}(A)\mathbb{P}(B), \quad \mathbb{P}(A \cap C) = \mathbb{P}(\text{both heads}) = \frac{1}{4} = \mathbb{P}(A)\mathbb{P}(C), \quad (4)$$

and $\mathbb{P}(B \cap C) = 1/4 = \mathbb{P}(B)\mathbb{P}(C)$ by the same token. Yet the three are *not* mutually independent: knowing A and B pins down C (if both coins are heads they certainly agree), so

$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}(\text{both heads}) = \frac{1}{4} \neq \frac{1}{8} = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C). \quad (5)$$

Pairwise $\neq$ mutual. Checking independence one pair at a time is not enough: any two of A, B, C carry no information about each other, but the three together are rigidly linked. Mutual independence must be verified on every subset, not just on pairs. The lesson generalizes: independence is a property of a *whole* collection under *one* probability measure, and it can fail at the joint level while holding at every pairwise level.

2.3 Conditional independence

We can now settle the launch question. The reconciliation of “no and yes” is that independence can hold *once a third event is fixed* without holding on its own.

Definition 2 (Conditional independence). *Events A and B are **conditionally independent given C** (with $\mathbb{P}(C) > 0$) if*

$$\mathbb{P}(A \cap B | C) = \mathbb{P}(A | C) \mathbb{P}(B | C). \quad (6)$$

Return to two looks at cell 1. Let θ be the target’s location, and take the prior $\mathbb{P}(\text{cell 1}) = p_1 = 0.7$ (so $\mathbb{P}(\text{not cell 1}) = 0.3$). Each look at cell 1 fails with the overlook probability $\alpha = 0.5$ if the target is there, and *always* fails if it is not. Let F_1, F_2 be the events “the first (second) look fails.”

Conditionally on the location, the looks are independent. Given that the target is in cell 1, the two looks are separate coin flips of the sensor:

$$\mathbb{P}(F_1 \cap F_2 | \text{cell 1}) = \alpha^2 = 0.25 = \alpha \cdot \alpha = \mathbb{P}(F_1 | \text{cell 1}) \mathbb{P}(F_2 | \text{cell 1}). \quad (7)$$

Given that the target is *not* in cell 1, both looks fail for certain, $\mathbb{P}(F_1 \cap F_2 | \text{not cell 1}) = 1 = 1 \cdot 1$. So F_1 and F_2 are conditionally independent given θ .

Marginally, the looks are dependent. Average over the two locations with the law of total probability:

$$\mathbb{P}(F_1) = p_1 \alpha + (1 - p_1) \cdot 1 = (0.7)(0.5) + 0.3 = 0.65, \quad (8)$$

$$\mathbb{P}(F_1 \cap F_2) = p_1 \alpha^2 + (1 - p_1) \cdot 1 = (0.7)(0.25) + 0.3 = 0.475. \quad (9)$$

By symmetry $\mathbb{P}(F_2) = 0.65$ as well, so

$$\mathbb{P}(F_2 | F_1) = \frac{\mathbb{P}(F_1 \cap F_2)}{\mathbb{P}(F_1)} = \frac{0.475}{0.65} \approx 0.731 \neq 0.65 = \mathbb{P}(F_2). \quad (10)$$

The first failure makes the second failure *more* likely. The mechanism is exactly Session 2: a failed look in cell 1 lowered our belief that the target is there (from 0.70 to $0.35/0.65 \approx 0.538$) and correspondingly raised our belief that it is elsewhere, where a second look is guaranteed to fail. The two observations are entangled not through any direct link but through the *common cause* θ they both depend on.

Conditionally i.i.d. given θ is the default sensor model. Repeated measurements of one latent quantity are the workhorse of the rest of the course: given the truth θ , the measurements are independent and identically distributed; marginally they are not, precisely because each one carries information about the shared θ . Every fusion rule, detector, and regression method ahead is built on this “conditionally i.i.d. given θ ” assumption. It is what made Session 2’s α^m factor legitimate.

2.4 Acting on belief: expected utility

A posterior is not the end of the story; it is an input. The course thesis asks how observation should change belief *and action*, and we now supply the second half. Fix a finite set of possible actions $\mathcal{A}$ and a **utility** $u(a, \theta)$ recording the payoff of taking action a when the truth is θ . We do not know θ ; we have a belief about it. So we score each action by its average payoff under that belief.

Definition 3 (Expected utility and the optimal action). *Under a belief $\mathbb{P}$ over θ , the expected utility of action a is*

$$\mathbb{E}[u(a, \theta)] = \sum_{\theta} u(a, \theta) \mathbb{P}(\theta), \quad (11)$$

and an **optimal action** is any $a^* \in \arg \max_{a \in \mathcal{A}} \mathbb{E}[u(a, \theta)]$.

Belief in, action out. The belief is the input and the action is the output. Change the belief and the optimal action can change, with the utility held fixed. This is why the loop matters: every observation updates the belief through Bayes' rule, and the updated belief may call for a different action than the one before it.

A one-shot illustration, continuing the two-cell search. After a failed look in cell 1 our belief that the target is there is 0.538. A recovery vehicle can be sent now (action “commit”) or held back (action “defer”). Committing pays $u(\text{commit}, \text{present}) = +8$ if the target is really in cell 1 and $u(\text{commit}, \text{absent}) = -12$ if it is not (a wasted, expensive deployment); deferring pays 0 either way. Then

$$\mathbb{E}[u(\text{commit})] = (0.538)(+8) + (0.462)(-12) \approx -1.23 < 0 = \mathbb{E}[u(\text{defer})], \quad (12)$$

so the optimal action is to *defer*. Note the dependence on belief: with the *prior* 0.70 (before any look) the same utilities give $\mathbb{E}[u(\text{commit})] = (0.7)(8) + (0.3)(-12) = +2.0 > 0$, and committing would be optimal. One failed look flipped the decision. The utility never changed; the belief did.

2.5 From belief to the next look

When the action is “which cell to search next,” expected utility takes a concrete and useful form. Suppose a single look at cell i costs c_i and, if the target is there, detects it with probability q_i (so $q_i = 1 - \alpha_i$). A natural objective is detection probability per unit cost.

Proposition 1 (Single-look allocation). *Under the current belief $(p_1, \dots, p_n)$, the single look maximizing the expected detection probability per unit cost is the cell maximizing*

$$\frac{p_i q_i}{c_i}. \quad (13)$$

The proof is one line, and it is deliberately yours: *Proof: HW1, Problem 4*. The lecture’s job is the modeling, that three ingredients, belief p_i , sensor quality q_i , and cost c_i , all enter the decision, and that none of them alone decides it.

Remark 1 (The greedy rule, repeated, is optimal). *For equal costs there is a striking fact we will use but not prove: repeating the greedy rule, look where $p_i q_i$ is largest, then apply Bayes' rule, then look again, minimizes the expected number of looks to find the target. This is the opening theorem of Stone's Theory of Optimal Search. We will not prove it, but Homework 1 will make you feel why it is plausible. Anything more, sequencing many looks under a budget, belongs to sequential decision theory and its later course, not here.*

Posterior $\rightarrow$ Action. The whole of Movement I is one loop. A *belief* (p_i) selects an *action*, the look maximizing $p_i q_i / c_i$; the action produces an *observation*, here a failure F_i ; the observation drives an *update*, Bayes' rule, which returns a new belief; and the cycle repeats with the quantities recomputed. Figure 1 draws the loop with the search quantities on each edge. Belief, action, observation, update: every session in this course is a new set of labels on these four nodes.

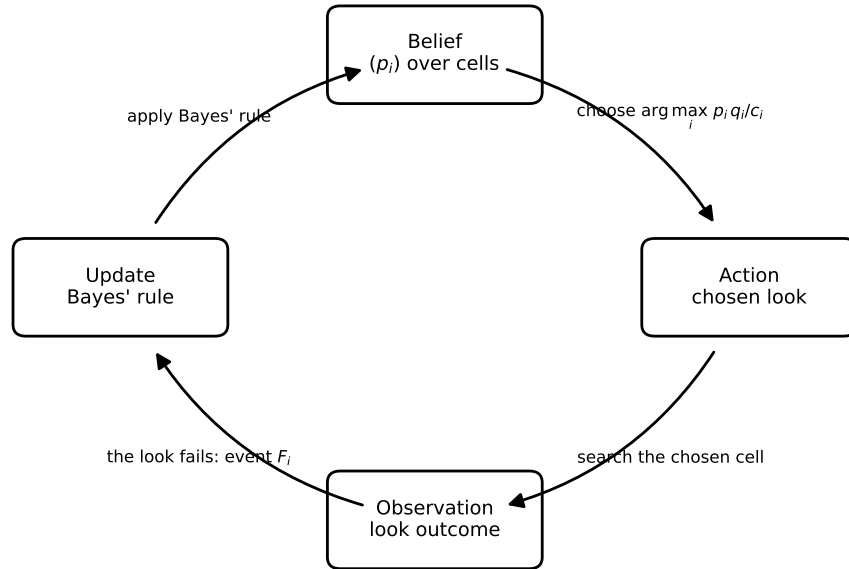


Figure 1: The belief–action–observation–update loop of Movement I. A posterior selects the next look through Proposition 1; the look's outcome updates the posterior through Bayes' rule; the cycle repeats.

3 Worked examples

3.1 Where should we look? Belief, sensor, and cost together

Four cells carry the prior $(p_1, p_2, p_3, p_4) = (0.4, 0.3, 0.2, 0.1)$. The sensors and search costs differ by cell: detection probabilities $q = (0.5, 0.9, 0.9, 0.9)$ and costs $c = (2, 1, 1, 3)$. Which

cell should we look in first? Score each by Proposition 1:

$$\frac{p_i q_i}{c_i} : \quad \underbrace{\frac{(0.4)(0.5)}{2} = 0.10}_{\text{cell 1}}, \quad \underbrace{\frac{(0.3)(0.9)}{1} = 0.27}_{\text{cell 2}}, \quad \underbrace{\frac{(0.2)(0.9)}{1} = 0.18}_{\text{cell 3}}, \quad \underbrace{\frac{(0.1)(0.9)}{3} = 0.03}_{\text{cell 4}}. \quad (14)$$

The optimal first look is cell 2, not cell 1, even though cell 1 holds the *largest* prior. Cell 1 is doubly penalized: its sensor is weak ($q_1 = 0.5$) and its search is expensive ($c_1 = 2$). A large prior is not a mandate to look; belief, sensor quality, and cost all vote.

One failed look, then re-rank. Suppose we look in cell 2 and fail. Its overlook probability is $\alpha_2 = 1 - q_2 = 0.1$, so by the failed-search update of Session 2 the evidence is $\mathbb{P}(F_2) = \alpha_2 p_2 + (1 - p_2) = (0.1)(0.3) + 0.7 = 0.73$, and the posterior is

$$p' = \left(\frac{0.4}{0.73}, \frac{(0.1)(0.3)}{0.73}, \frac{0.2}{0.73}, \frac{0.1}{0.73} \right) \approx (0.548, 0.041, 0.274, 0.137). \quad (15)$$

Re-score with the same q and c :

$$\frac{p'_i q_i}{c_i} \approx \underbrace{0.137}_{\text{cell 1}}, \quad \underbrace{0.037}_{\text{cell 2}}, \quad \underbrace{0.247}_{\text{cell 3}}, \quad \underbrace{0.041}_{\text{cell 4}}. \quad (16)$$

Now cell 3 leads and the next look moves there: the failed look in cell 2 drained its score and promoted the runners-up. This is the loop of Figure 1 turning once, with every number on the page.

4 Consolidation

AI corner: a language model is a conditional pmf. A large language model is, stripped to its mathematics, a single object: $\mathbb{P}(\text{next token} \mid \text{all tokens so far})$, a conditional distribution over a finite alphabet, learned from data. Generating text is Session 1’s chain rule run forward. For a four-token string,

$$\mathbb{P}(w_1 w_2 w_3 w_4) = \mathbb{P}(w_1) \mathbb{P}(w_2 \mid w_1) \mathbb{P}(w_3 \mid w_1 w_2) \mathbb{P}(w_4 \mid w_1 w_2 w_3), \quad (17)$$

and the model emits one token at a time by sampling from each conditional in turn. Two takeaways. First, everything such a model “knows” lives in conditional distributions, this course’s central object. Second, a “hallucination” is not a failed database lookup but confident sampling from a conditional distribution that the context has conditioned badly; the machine is doing exactly what its mathematics says, which is why the remedy is better conditioning, not better memory. One look ahead: every homework in this course closes by auditing such a machine’s probability work, and the toolkit for those audits is this course.

Work these in class.

1. (*Mechanical, conditional independence.*) Two looks at a cell that holds the target with prior $p_1 = 0.8$; each look overlooks a present target with probability $\alpha = 0.25$ and

always fails otherwise. Compute $\mathbb{P}(F_1)$, $\mathbb{P}(F_1 \cap F_2)$, and $\mathbb{P}(F_2 | F_1)$, and say whether F_1, F_2 are marginally independent.

Answer: $\mathbb{P}(F_1) = (0.8)(0.25) + 0.2 = 0.4$; $\mathbb{P}(F_1 \cap F_2) = (0.8)(0.0625) + 0.2 = 0.25$; $\mathbb{P}(F_2 | F_1) = 0.25/0.4 = 0.625 \neq 0.4 = \mathbb{P}(F_2)$, so they are *not* marginally independent (they are conditionally independent given the location).

2. (*Mechanical, allocation.*) Three cells with belief $(0.5, 0.3, 0.2)$, detection probabilities $q = (0.4, 0.9, 0.9)$, and costs $c = (1, 1, 2)$. Which cell is the optimal single look?

Answer: scores $p_i q_i / c_i = (0.20, 0.27, 0.09)$; cell 2 wins, though cell 1 has the largest prior.

3. (*Conceptual.*) A colleague fuses two passes of the *same* sonar over one region by multiplying their likelihoods as if the passes were independent, without conditioning on the target's location. In one sentence, what goes wrong?

Answer: the passes are only *conditionally* independent given the location; treating them as marginally independent double-counts the shared latent cause and overstates the evidence.

5 Synthesis

Movement I is now complete, and the course slogan has all three of its clauses answered for the search problem. *What is uncertain?* A latent θ , the target's cell. *What is observed?* Looks, whose outcomes depend on θ . *How should observation change belief and action?* Belief moves by conditioning and Bayes' rule; action follows by maximizing expected utility, which for search is the rule $\arg \max_i p_i q_i / c_i$. Independence turned out to be the subtle hinge: not a default to assume but a modeling choice to verify, and the right version of it, conditional independence given θ , is what lets us treat repeated sensor readings as clean, i.i.d. evidence while still letting each reading teach us about the shared cause. The four nodes, belief, action, observation, update, are the shape of everything ahead. Movement II keeps the loop and changes the objects inside it: the latent θ and the observations become *numbers*, random variables, and we will need expectation, distributions, and densities to run the same machine on them.

Self-check questions

1. In the two-coin example, every pair of A, B, C is independent, yet the three are not mutually independent. In one sentence, why?

Answer: any two of the events leave the third genuinely uncertain, but A and B together determine C , so $\mathbb{P}(A \cap B \cap C) = 1/4 \neq 1/8 = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$.

2. Two looks at one cell are conditionally independent given the target's location but marginally dependent. What couples them marginally?

Answer: the shared latent cause θ ; the first outcome shifts belief about θ , which changes the probability of the second (here a first failure raises the chance of a second).

3. In the allocation rule $\arg \max_i p_i q_i / c_i$, give one reason the cell with the largest prior can be the wrong place to look.

Answer: a weak sensor (small q_i) or a high cost c_i can pull that cell's score below a rival's, as cell 1 fell behind cell 2 in the worked example.