Problem Set 07: Bayesian updating

Do this problem set *before* the lecture on Bayesian updating. You need no material from the lecture: each problem starts with a grey box that gives the tools it uses, and each question is cut into small steps. Problems 1 and 4 come first if time is short; the lecture will build on all five.

Toolbox 1 — Bayes' rule in five steps

A person does not know the **state** θ (ill or healthy, guilty or innocent, talented or not). She observes an **outcome** x (a test result, a piece of evidence) and asks: how likely is each state, now that I have seen x?

1. **Prior.** Write the probability of each state before the outcome: $P(\theta)$.
2. **Likelihood.** For each state, write the probability of the outcome you saw: $P(x \mid \theta)$.
3. **Joint.** Multiply: $P(x \text{ and } \theta) = P(x \mid \theta)\, P(\theta)$.
4. **Total.** Add the joint probabilities over all states: $P(x) = \sum_\theta P(x \text{ and } \theta)$.
5. **Posterior.** Divide each joint probability by the total: $P(\theta \mid x) = \dfrac{P(x \text{ and } \theta)}{P(x)}$.

Example. An urn is of type A (2 red balls out of 3) or of type B (1 red ball out of 3), each with probability $\frac{1}{2}$. You draw a red ball.

State	1. Prior	2. Likelihood of red	3. Joint	5. Posterior
A	$\frac{1}{2}$	$\frac{2}{3}$	$\frac{1}{2} \cdot \frac{2}{3} = \frac{1}{3}$	$\frac{1/3}{1/2} = \frac{2}{3}$
B	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}$	$\frac{1/6}{1/2} = \frac{1}{3}$
4. Total			$P(\text{red}) = \frac{1}{3} + \frac{1}{6} = \frac{1}{2}$	

The posteriors always add up to 1: a good check of your arithmetic.

Toolbox 2 — the same rule with odds

The **odds** of A against B are $P(A)/P(B)$. The **likelihood ratio** of an outcome x is $P(x \mid A)/P(x \mid B)$: how much more likely the outcome is under A than under B. With two states, Bayes' rule becomes one multiplication:

$$\underbrace{\frac{P(A \mid x)}{P(B \mid x)}}_{\text{posterior odds}} = \underbrace{\frac{P(x \mid A)}{P(x \mid B)}}_{\text{likelihood ratio}} \times \underbrace{\frac{P(A)}{P(B)}}_{\text{prior odds}}.$$

To go back from odds o to a probability: $P(A \mid x) = \dfrac{o}{1+o}$. **Example** (urn above): prior odds 1, likelihood ratio $\frac{2/3}{1/3} = 2$, posterior odds 2, so $P(A \mid \text{red}) = \frac{2}{1+2} = \frac{2}{3}$, as in the table.

Problem 1 A rare condition and repeated tests

Warm-up

A condition affects 2% of the population. A test is positive with probability 0.9 if the person has the condition, and with probability 0.1 if she does not. A person takes the test.

(a) **Prior and likelihoods.** Write $P(\text{ill})$, $P(\text{healthy})$, $P(+ \mid \text{ill})$ and $P(+ \mid \text{healthy})$.

(b) **The table.** Complete the table for a positive result, following Toolbox 1.

State	Prior	Likelihood of $+$	Joint	Posterior
ill	_____	_____	_____	_____
healthy	_____	_____	_____	_____
Total			$P(+) =$ _____	

(c) **The same with 1000 people.** Out of 1000 people, how many are ill? How many of them test positive? How many healthy people test positive? Among all who test positive, what share is ill? Compare with (b).

(d) **The same with odds.** Compute the prior odds of ill against healthy, the likelihood ratio of a positive result, the posterior odds, and the posterior probability (Toolbox 2). You should find the number of (b) again.

(e) **Base-rate neglect.** Many people answer "90%" to (b). Show that this answer amounts to using prior odds equal to 1, i.e. forgetting that the condition is rare.

(f) **A negative result.** Redo the table of (b) for a negative result. What is $P(\text{ill} \mid -)$?

(g) **On average, beliefs do not move.** Before the test, the person does not know which result she will get. Compute the average posterior,

$$P(+) \cdot P(\text{ill} \mid +) + P(-) \cdot P(\text{ill} \mid -),$$

and compare it with the prior. *Keep this result in mind: Problem 2 uses it.*

(h) **Two tests.** The person takes a second, independent test: given her state, the second result does not depend on the first. So $P(+,+ \mid \text{ill}) = 0.9 \times 0.9$.

(i) Write $P(+,+ \mid \text{healthy})$, then build the table for two positive results. What is $P(\text{ill} \mid +,+)$?

(ii) Do the same for one positive and one negative result. What do you notice? Would the order (first $+$ then $-$, or the reverse) change the answer?

(i) **How many tests?** Each positive result multiplies the odds by the likelihood ratio of (d).

(i) Write the odds after n positive results.

(ii) Show that $P(\text{ill}) \geq 0.95$ exactly when the odds are at least 19.

(iii) Compute the odds for $n = 1, 2, 3, 4$ and find how many positive results in a row are needed.

Problem 2 Designing an investigation

Core

What you need — signals and posteriors

A **signal** is a rule that produces a message x with a probability that depends on the state: $P(x \mid \theta)$. The test of Problem 1 is a signal with messages $+$ and $-$. Before the message arrives, the posterior is uncertain: it takes the value $P(\theta \mid x)$ with probability $P(x)$. In Problem 1(g) you found that **the average posterior equals the prior**. This holds for every signal. A list of posteriors with probabilities whose average equals the prior is called **Bayes plausible**. Conversely, every Bayes-plausible list can be produced by some signal; part (d) shows how.

A defendant is guilty or innocent. The judge and the prosecutor share the prior $P(\text{guilty}) = 0.3$. The judge convicts if and only if her posterior probability of guilt is at least $\frac{1}{2}$. Before anything is known, the prosecutor designs an investigation (a signal) and wants to maximise the probability of conviction. The judge knows how the investigation works and updates by Bayes' rule.

(a) **Which lists are possible?** Below, each list gives the possible posteriors of guilt and their probabilities. Compute the average posterior of each list, and say which lists are Bayes plausible.

List	Posteriors (probabilities)	Average posterior
τ_A	0.1 (prob. $\frac{1}{2}$), 0.6 (prob. $\frac{1}{2}$)	_____
τ_B	0.1 (prob. $\frac{2}{3}$), 0.7 (prob. $\frac{1}{3}$)	_____
τ_C	0.4 and 0.6, with any probabilities	_____
τ_D	0 (prob. 0.7), 1 (prob. 0.3)	_____

(b) **Conviction probabilities.** For each Bayes-plausible list, which posteriors lead to a conviction? What is the probability of conviction? What is it with no investigation at all?

(c) **An upper bound.** We show that no investigation convicts with probability above 0.6.

(i) Let $q \in [0,1]$ be a posterior, and let $c(q) = \frac{1}{2}$ if $q \geq \frac{1}{2}$ (conviction) and $c(q) = 0$ if $q < \frac{1}{2}$. Show that $q \geq c(q)$. *Hint: treat the two cases $q \geq \frac{1}{2}$ and $q < \frac{1}{2}$ separately.*

(ii) Average both sides of $q \geq c(q)$ over the posteriors of any investigation. Why does the left side average to 0.3? Show that the right side averages to $\frac{1}{2} \times P(\text{conviction})$.

(iii) Conclude that $P(\text{conviction}) \leq 0.6$.

(d) **The best investigation.** The bound is reached when every posterior satisfies $q = c(q)$, i.e. when the posteriors are 0 and $\frac{1}{2}$.

(i) Let the posterior $\frac{1}{2}$ (message "suspicious") have probability w and the posterior 0 (message "cleared") probability $1 - w$. Use Bayes plausibility to find w.

(ii) Fill in the joint probabilities, using $P(\text{message and state}) = P(\text{message}) \times P(\text{state} \mid \text{message})$:

	guilty	innocent	Total
cleared (posterior 0)	_____	_____	$1 - w$
suspicious (posterior $\frac{1}{2}$)	_____	_____	w
Total	_____	_____	1

Check that the column totals give back the prior.

(iii) Divide each column by its total to get $P(\text{message} \mid \text{state})$: this describes the investigation. Say in words what it does with a guilty and with an innocent defendant.

(iv) Check with Bayes' rule (Toolbox 1) that "suspicious" leads to the posterior $\frac{1}{2}$.

(e) **Is the judge fooled?** What share of innocent defendants is convicted? Among convicted defendants, what share is guilty? Is the judge mistaken about anything?

Problem 3 A credit score and two groups

Core

What you need — error rates and calibration

A lender approves ($S = 1$) or rejects ($S = 0$) applicants. An applicant would repay ($\theta = 1$) or not ($\theta = 0$), and belongs to group g_1 or g_2. The **base rate** $p_g = P(\theta = 1)$ is the share of good types in group g. Within a group, three numbers describe the score:

- false-positive rate $FP = P(S = 1 \mid \theta = 0)$: share of bad types wrongly approved;
- false-negative rate $FN = P(S = 0 \mid \theta = 1)$: share of good types wrongly rejected;
- positive predictive value $PPV = P(\theta = 1 \mid S = 1)$: share of approved applicants who repay.

The score is **calibrated** if $P(\theta = 1 \mid S = s)$ is the same in both groups, for $s = 0$ and $s = 1$: a score means the same thing for everyone. Two other fairness criteria ask for equal FP and equal FN across groups. The **law of total probability** gives $P(\theta = 1) = P(S = 1)\,P(\theta = 1 \mid S = 1) + P(S = 0)\,P(\theta = 1 \mid S = 0)$.

A score is calibrated: in both groups, $P(\theta = 1 \mid S = 1) = 0.8$ and $P(\theta = 1 \mid S = 0) = 0.1$. The base rates are $p_{g_1} = 0.5$ and $p_{g_2} = 0.2$.

(a) **Group g_1.**

(i) Let $s = P(S = 1)$ in group g_1. Use the law of total probability to write an equation in s, and solve it.

(ii) Fill in the joint probabilities, using $P(S \text{ and } \theta) = P(S) \times P(\theta \mid S)$, and check that the row of totals gives the base rate:

	$\theta = 1$	$\theta = 0$	Total
$S = 1$	_____	_____	s
$S = 0$	_____	_____	$1 - s$
Total	_____	_____	1

(iii) Compute $FN = P(S = 0 \text{ and } \theta = 1)/P(\theta = 1)$ and $FP = P(S = 1 \text{ and } \theta = 0)/P(\theta = 0)$.

(b) **Group g_2.** Repeat (a) for group g_2. Compare the error rates of the two groups. Who is hurt in each group?

(c) **Equal error rates instead.** A regulator requires group g_2 to have the error rates of group g_1, and the base rate stays 0.2. Fill in a new table for g_2, now using $P(S \text{ and } \theta) = P(\theta) \times P(S \mid \theta)$, and compute $P(\theta = 1 \mid S = 1)$ and $P(\theta = 1 \mid S = 0)$. Is the score still calibrated?

(d) **Why one cannot have everything.** Fix a group. We show that

$$FP = \frac{p}{1-p} \times \frac{1 - PPV}{PPV} \times (1 - FN). \qquad (*)$$

(i) Write $FP = \dfrac{P(S = 1 \text{ and } \theta = 0)}{1 - p}$.

(ii) Show that $\dfrac{P(S = 1 \text{ and } \theta = 0)}{P(S = 1 \text{ and } \theta = 1)} = \dfrac{1 - PPV}{PPV}$. *Hint: divide the numerator and the denominator by $P(S = 1)$.*

(iii) Show that $P(S = 1 \text{ and } \theta = 1) = p\,(1 - FN)$, and combine (i)–(iii) to get $(*)$.

(iv) Suppose FP, FN and PPV are equal in the two groups. What does $(*)$ imply for the base rates? Conclude.

(e) **How error rates move with the base rate** (practice with derivatives). Take any calibrated score with $r_1 = P(\theta = 1 \mid S = 1)$ and $r_0 = P(\theta = 1 \mid S = 0)$, $0 < r_0 < r_1 < 1$, and a base rate p with $r_0 < p < r_1$. The steps of (a) give

$$FN(p) = \frac{r_0}{r_1 - r_0}\Big(\frac{r_1}{p} - 1\Big), \qquad FP(p) = \frac{1 - r_1}{r_1 - r_0} \cdot \frac{p - r_0}{1 - p}.$$

(i) Check both formulas with the numbers of group g_1.

(ii) Differentiate FN with respect to p. *Hint: the derivative of $\frac{1}{p}$ is $-\frac{1}{p^2}$; constants stay in front.* Is FN increasing or decreasing in p?

(iii) Differentiate $\dfrac{p - r_0}{1 - p}$ with the quotient rule $\left(\dfrac{u}{v}\right)' = \dfrac{u'v - uv'}{v^2}$. What is its sign? Conclude for FP.

(iv) In words: under a calibrated score, which group's good types are more often rejected?

Problem 4 Hiring with normal beliefs

Warm-up

What you need — updating with normal distributions

The **precision** of a normal variable is one over its variance: $\tau = 1/\sigma^2$. A precise source has a small variance. Let the unknown θ be normal, $\theta \sim \mathcal{N}(\mu, \sigma_\theta^2)$, and let a signal add independent normal noise: $X = \theta + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$. After seeing $X = x$, the belief about θ is again normal, with

$$\text{mean} = \frac{\tau_\theta}{\tau_\theta + \tau_\varepsilon}\,\mu + \frac{\tau_\varepsilon}{\tau_\theta + \tau_\varepsilon}\,x, \qquad \text{variance} = \frac{1}{\tau_\theta + \tau_\varepsilon},$$

where $\tau_\theta = 1/\sigma_\theta^2$ and $\tau_\varepsilon = 1/\sigma_\varepsilon^2$. In words: the new mean is a **weighted average** of the prior mean and the signal, with weights proportional to the precisions, and the **precisions add up**. With several independent signals $X_i = \theta + \varepsilon_i$ of precisions τ_i, all precisions add:

$$\text{mean} = \frac{\tau_\theta \mu + \sum_i \tau_i x_i}{\tau_\theta + \sum_i \tau_i}, \qquad \text{variance} = \frac{1}{\tau_\theta + \sum_i \tau_i}.$$

Example. $\theta \sim \mathcal{N}(0, 1)$, $\sigma_\varepsilon^2 = 1$, $x = 2$: both precisions equal 1, the weights are $\frac{1}{2}, \frac{1}{2}$, the mean is 1 and the variance $\frac{1}{2}$.

A worker's productivity is $\theta \sim \mathcal{N}(10, 4)$. An employer reads a reference letter $X_1 = \theta + \varepsilon_1$, $\varepsilon_1 \sim \mathcal{N}(0, 4)$, and sees a work sample $X_2 = \theta + \varepsilon_2$, $\varepsilon_2 \sim \mathcal{N}(0, 1)$, all independent. The realisations are $x_1 = 14$ and $x_2 = 8$.

(a) **Precisions.** Compute τ_θ, τ_1 and τ_2. Which source is the most precise?

(b) **The letter alone.** Compute the weights, the mean and the variance of the belief after the letter.

(c) **Then the work sample.** The belief after the letter becomes the new prior. Update it with $x_2 = 8$: new precisions, weights, mean, variance.

(d) **Both at once.** Use the formula for several signals with x_1 and x_2 together. Compare with (c).

(e) **The other order.** Update first with the work sample, then with the letter. Does the order matter?

(f) **How much data?** How many independent letters (each with noise variance 4) would bring the variance of the belief down to $\frac{1}{2}$ or less? How many work samples? Complete: "one work sample is worth _____ letters".

Problem 5 Working hard to look talented

Core

What you need — maximising a function of effort

- **Derivatives used here.** For a constant k: the derivative of k is 0; of $k\,a$ is k; of $\frac{1}{2}a^2$ is a.

- **Maximum.** If $U'(a) = 0$ at $a = \hat{a}$ and $U''(a) < 0$ for all a, then $\hat{a}$ is the unique maximiser of U.
- **Expectations.** For numbers k, b: $\mathbb{E}[kX + b] = k\,\mathbb{E}[X] + b$.

A worker's talent $\theta \sim \mathcal{N}(\mu, \sigma_\theta^2)$ is unknown to everyone, the worker included. She chooses an effort $a \geq 0$ at cost $\frac{1}{2}a^2$. A manager does not see the effort; she sees the output $X = \theta + a + \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ independent of θ. The next employer pays a wage equal to the manager's expected talent given X. Effort does not change talent, so why work? Because extra output looks like extra talent. In equilibrium the manager correctly anticipates the effort, which we call a^* (Holmström, 1999).

(a) **The manager's inference.** Since the manager expects effort a^*, she reads $X - a^* = \theta + \varepsilon$ as a signal of θ. Use the box of Problem 4 to show that the wage is

$$w(X) = (1 - k)\,\mu + k\,(X - a^*), \qquad k = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\varepsilon^2}.$$

Hint: write the precisions as $1/\sigma^2$ and multiply the numerator and the denominator of the weight by $\sigma_\theta^2\sigma_\varepsilon^2$.

(b) **The worker's expected wage.** If the worker chooses a, what is $\mathbb{E}[X]$? Show that her expected wage is $\mu + k\,(a - a^*)$.

(c) **The worker's choice.** Her payoff is $U(a) = \mu + k\,(a - a^*) - \frac{1}{2}a^2$, where μ, k and a^* are constants for her.

(i) Compute $U'(a)$ and $U''(a)$.

(ii) Find the effort that maximises U. Does it depend on a^*?

(d) **Equilibrium.** In equilibrium the manager's guess is right: the effort of (c) equals a^*. Find a^*. Compute it when $\sigma_\theta^2 = \sigma_\varepsilon^2 = 1$. How does it change when output becomes noisier (σ_ε^2 larger)? Why?

(e) **A test score.** The manager now also sees a test score $Z = \theta + \eta$, $\eta \sim \mathcal{N}(0, \sigma_\eta^2)$, independent of everything else and not affected by effort.

(i) With the formula for several signals, write the weight k' that the wage puts on $X - a^*$, using precisions $\tau_\theta, \tau_\varepsilon, \tau_\eta$.

(ii) Repeat (b)–(d): show that the new equilibrium effort is k'. *Hint: the test score enters the wage, but its expectation does not depend on a.*

(iii) Compute k' when all three variances equal 1. Does the test score raise or lower effort? Why?

(f) **Who gains?** Use (b) to show that in equilibrium the expected wage equals μ, with or without the test score. Compute the worker's expected payoff $\mu - \frac{1}{2}(a^*)^2$ in both cases (all variances equal to 1). Does the test score help her?

References

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163.

Holmström, B. (1999). Managerial incentive problems: A dynamic perspective. *Review of Economic Studies*, 66(1):169–182.

Kamenica, E. and Gentzkow, M. (2011). Bayesian persuasion. *American Economic Review*, 101(6):2590–2615.

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In *Proceedings of the 8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, 43:1–43:23.