6 - Classification

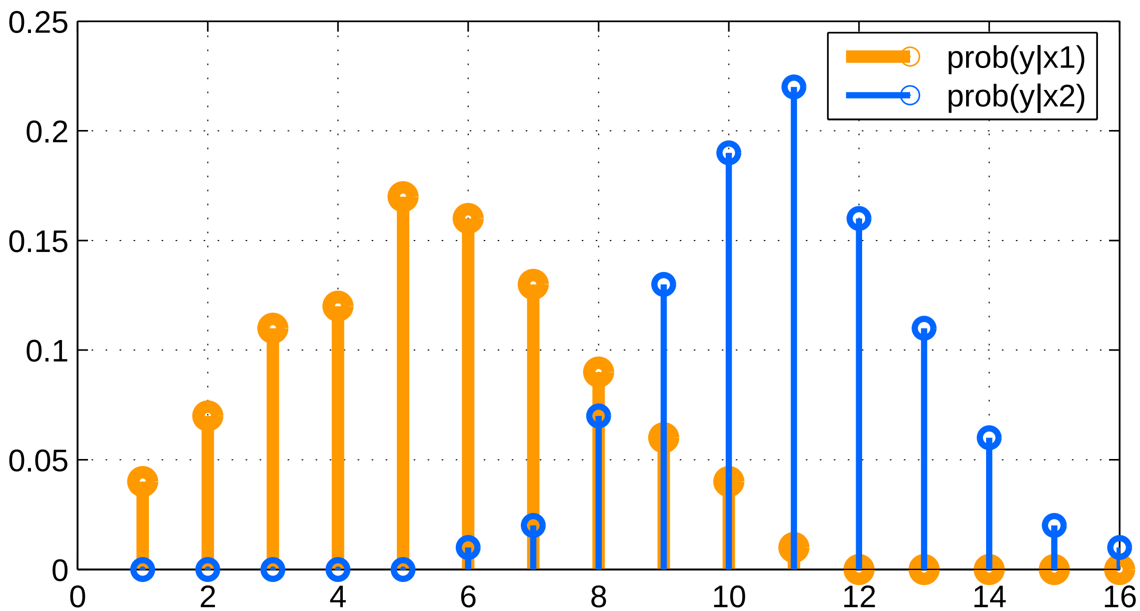
- Examples: radar system, binary transmission, OCR , spam filtering
- The classification problem
- Transition matrices and Bayes rule
- The importance of prior probabilities
- The MAP classifier and example
- Decision regions
- Example: gold coins
- Error analysis and the MAP classifier
- Cost functions and example
- Trade-offs and the Neyman-Pearson cost function
- Example: weighted-sum objective
- The operating characteristic
- Conditional errors and maximum likelihood

Example: Radar System

A radar system sends out n pulses, and receives y reflections, where $0 \leq y \leq n$.
Ideally, $y = n$ if an aircraft is present, and $y = 0$ otherwise.
In practice, reflections may be lost, or noise may be mistaken for reflections.

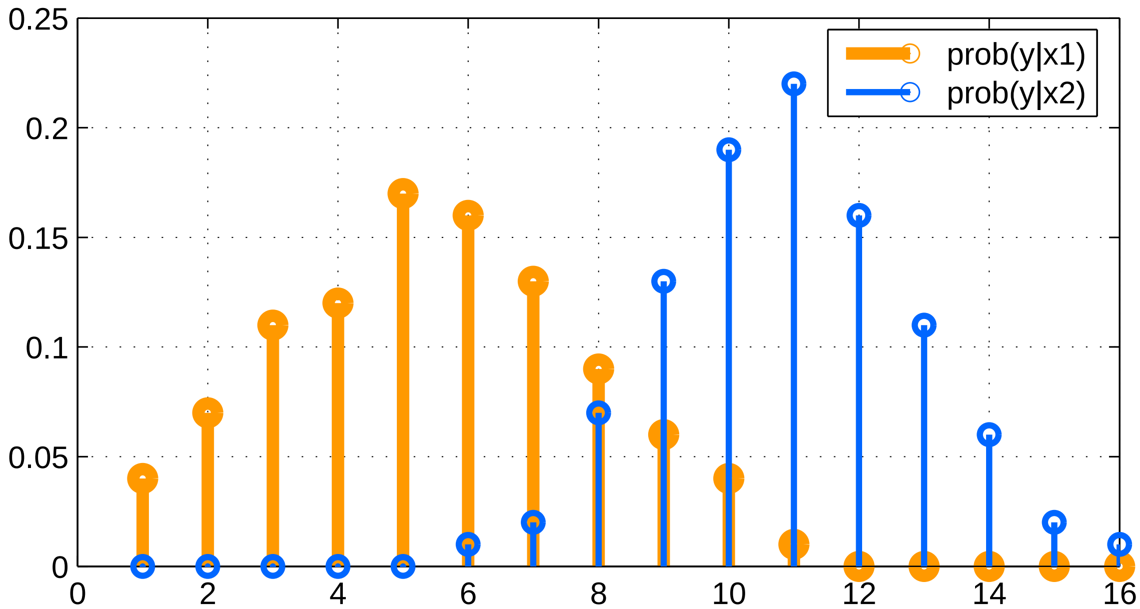
So we have two probability mass functions

$p_1(y) =$ the probability of receiving y reflections when there are no aircraft present
$p_2(y) =$ the probability of receiving y reflections when there is an aircraft present



If we measure y_{meas} reflections, how do we decide if an aircraft is present?

Example: Radar System



If there are fewer than 6 reflections, an aircraft is not present. If there are more than 11 reflections, an aircraft is present.

We would like to choose a *threshold* value, based on

- probabilities of errors; false-positives and false-negatives
- Costs assigned to these events

Other Examples

- *Binary transmission channel:* A binary bit is sent to us across a communication channel.
  - If a 1 is sent, then with probability 0.8 a 1 is received, and probability 0.2 a 0 is received
  - If a 0 is sent, then with probability 0.1 a 1 is received, and probability 0.9 a 0 is received

  We measure the received bit, and would like to determine which bit was sent.

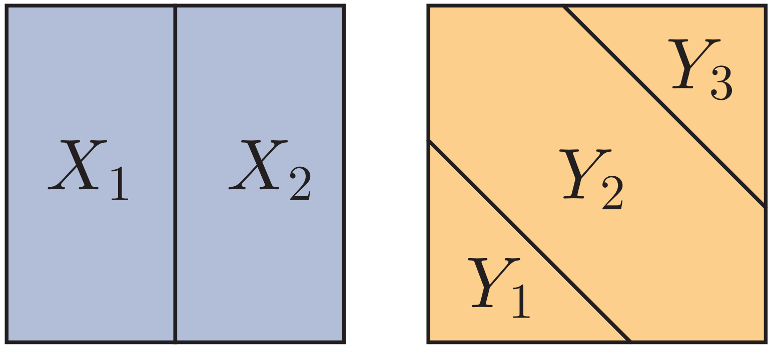
- *Optical character recognition:* We measure various features of a character in an optical system, such as
  - the width of the character
  - the ratio of black pixels to white pixels

  Which of the characters $A, B, \ldots, Z$ is it?

- *Spam filtering:* we measure which words are contained in the email. We would like to determine if the email is spam or not.

The Classification Problem

- $X_1, \ldots, X_n$ are events that partition Ω, called *hypotheses*
- $Y_1, \ldots, Y_m$ are events that partition Ω, called *observations*



The *outcome* of the experiment is $\omega \in \Omega$

- ω lies in *exactly one* of the events X_j and exactly one of the events Y_i
- In other words, exactly one 'hypothesis is true' and exactly one observation occurs

The *decision* or *classification* problem is as follows:

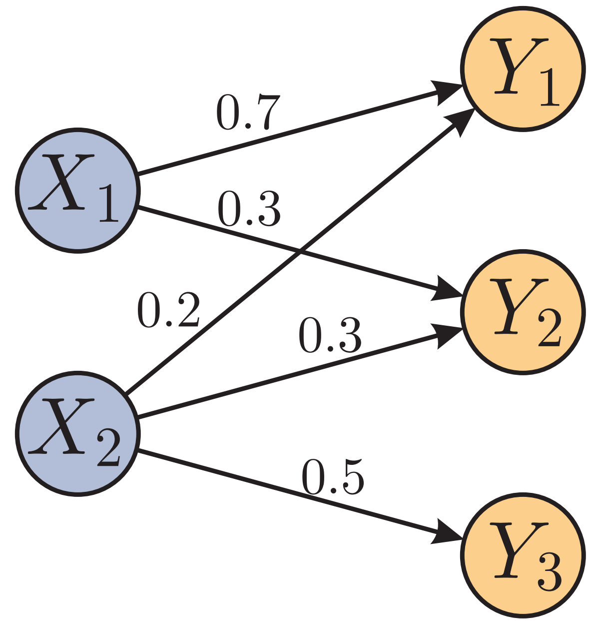
- We *measure* which of the Y_i the outcome lies in, say $Y_{i_{\text{meas}}}$
- We would like to pick j_{est} to *estimate* which X_j contains ω

Transition Matrices

We have a *transition matrix* $A \in \mathbb{R}^{m \times n}$

$$A_{ij} = \mathbf{Prob}(Y_i \,|\, X_j)$$

The matrix A is also called the *likelihood matrix*.



We can represent it as a *bipartite graph*, e.g.,

$$A = \begin{bmatrix} 0.7 & 0.2 \\ 0.3 & 0.3 \\ 0 & 0.5 \end{bmatrix}$$

- A is elementwise nonnegative and the sum of each column is one, i.e.,

$$A \succeq 0 \qquad \text{and} \qquad \mathbf{1}^T A = \mathbf{1}^T$$

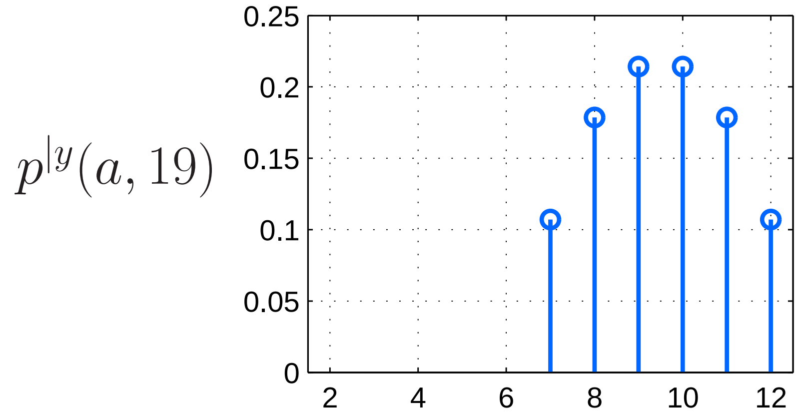
A matrix with these properties is called *column stochastic*.

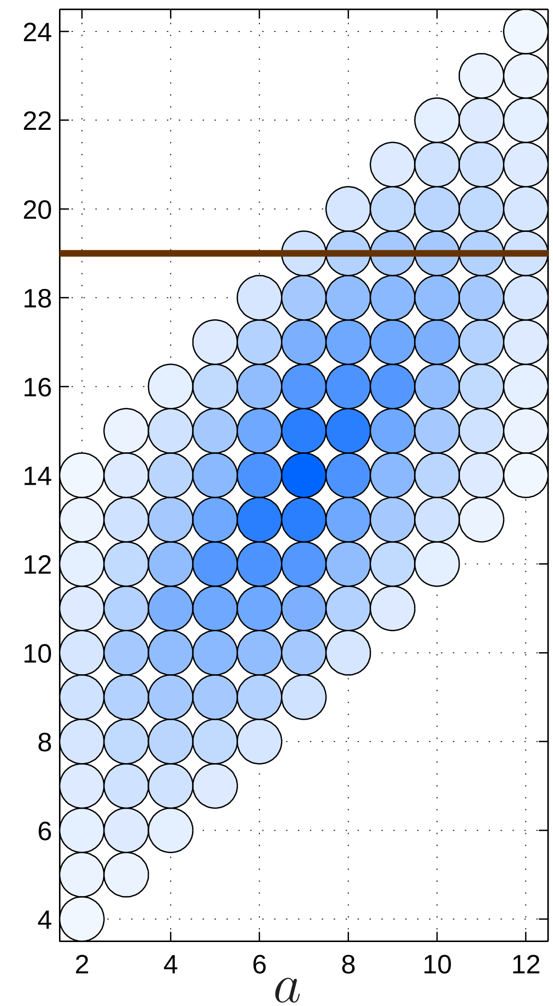
Conditional Probability

We would like to know

$$B_{i_{\text{meas}},j} = \mathbf{Prob}(X_j \mid Y_{i_{\text{meas}}})$$

- $\mathbf{Prob}(X_j \mid Y_{i_{\text{meas}}})$ is called the *a-posteriori probability*
- We will have a *different pmf* for each value of i_{meas}
- Once we have computed the a-posteriori pmf, we can pick an *estimate*, i.e., a value for j_{est}
- The estimate is usually chosen to minimize a *cost function*





Bayes Rule

For any events $A, B \subset \Omega$ with $\mathbf{Prob}(B) \neq 0$, Bayes rule is

$$\boxed{\mathbf{Prob}(A \mid B) = \frac{\mathbf{Prob}(B \mid A)\,\mathbf{Prob}(A)}{\mathbf{Prob}(B)}}$$

Because if $\mathbf{Prob}(B) \neq 0$, then

$$\mathbf{Prob}(A \mid B) = \frac{\mathbf{Prob}(A \cap B)}{\mathbf{Prob}(B)}$$

and so

$$\mathbf{Prob}(A \mid B)\,\mathbf{Prob}(B) = \mathbf{Prob}(B \mid A)\,\mathbf{Prob}(A)$$

Bayes Rule

The *Law of Total Probability* says that since $X_1, \dots, X_m$ partition Ω, we have for any event A

$$\mathbf{Prob}(A) = \sum_{j=1}^{m} \mathbf{Prob}(A \cap X_j)$$

Now by Bayes rule, we have

$$\begin{aligned}\mathbf{Prob}(X_j \,|\, Y_i) &= \frac{\mathbf{Prob}(Y_i \,|\, X_j)\,\mathbf{Prob}(X_j)}{\mathbf{Prob}(Y_i)} \\ &= \frac{\mathbf{Prob}(Y_i \,|\, X_j)\,\mathbf{Prob}(X_j)}{\sum_{k=1}^{m} \mathbf{Prob}(Y_i \cap X_k)}\end{aligned}$$

and therefore the *a-posteriori probability* is

$$\boxed{\mathbf{Prob}(X_j \,|\, Y_i) = \frac{\mathbf{Prob}(Y_i \,|\, X_j)\,\mathbf{Prob}(X_j)}{\sum_{k=1}^{m} \mathbf{Prob}(Y_i \,|\, X_k)\,\mathbf{Prob}(X_k)}}$$

Problem Data

We start with

- the *prior distribution* $x_j = \mathbf{Prob}(X_j)$ for $j = 1, \ldots, n$
- the *transition probabilities* $A_{ij} = \mathbf{Prob}(Y_i \mid X_j)$ for $i = 1, \ldots, m$ and $j = 1, \ldots, n$

From these we can find

- the *a-posteriori probabilities* $B_{ij} = \mathbf{Prob}(X_j \mid Y_i)$
- the *marginal pmf* $y_i = \mathbf{Prob}(Y_i)$
- and the *joint distribution* $J_{ij} = \mathbf{Prob}(Y_i \cap X_j)$

We have

$$y = Ax \qquad B_{ij} = \frac{J_{ij}}{y_i} \qquad J_{ij} = A_{ij} x_j$$

Example: Prior Probabilities

Why do we need prior probabilities? The following is the standard example.

Suppose we have a test for cancer, which has the following *accuracy*

- if the patient does not have cancer, then the probability of a negative result is 0.97, and of positive result is 0.03.
- if the patient has cancer, then the probability of a negative result is 0.02, and of a positive result is 0.98.

These are the *transition probabilities*

Suppose a patient takes this test. The probability of not having cancer is 0.992, and hence the probability of having cancer is 0.008.

These are the *prior probabilities*.

Example: Prior Probabilities

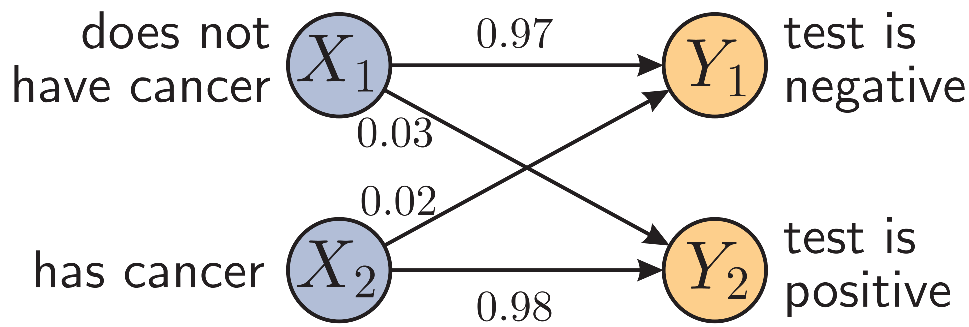
Imagine $10,000$ patients take this test.

- On average, 80 of these people will have cancer (0.008 probability) and since 98% of them will test positive, we will have 78 positive tests
- Of the 9,920 cancerless patients, 3% of them will test positive, giving a further 297 positive tests
- Hence of the total 375 positive tests, most (297) are false positives.
- The conditional probability of having cancer given that one tests positive is $78/375 = 0.208$

Example: Prior Probabilities

The transition matrix is

$$A = \begin{bmatrix} 0.97 & 0.02 \\ 0.03 & 0.98 \end{bmatrix}$$



The joint probabilities are

$J =$

	no cancer	cancer
test is negative	0.96224	0.00016
test is positive	0.02976	0.00784

But the conditional probabilities are

$B =$

	no cancer	cancer
test is negative	0.999834	0.000166251
test is positive	0.791489	0.208511

So given that the patient tests positive, the chances of having cancer are only 20%

Without a prior, one cannot draw any conclusion.

Classifiers

We would like to find a *classifier*, that is a map $f_{\text{est}} : \{1, \ldots, m\} \to \{1, \ldots, n\}$ which

> if we observe event Y_i, then we estimate that event X_j occurred, where $j = f_{\text{est}}(i)$

- Notice that *classification* is deliberately throwing away information, since we have the conditional probabilities $\mathbf{Prob}(X_j \mid Y_i)$.
- That is, the summary that *the patient does not have cancer* is less informative than *the patient has 20.8% chance of having cancer*

Classifiers

We will specify the estimator via a matrix $K \in \mathbb{R}^{m \times n}$, where

$$K_{ij} = \begin{cases} 1 & \text{if } j = f_{\text{est}}(i) \\ 0 & \text{otherwise} \end{cases}$$

- there is exactly one 1 in every row of K
- $K\mathbf{1} = \mathbf{1}$, i.e., K is *row stochastic*

The MAP Classifier

The *maximum a-posteriori probability (MAP)* classifier is

$$\boxed{f_{\mathsf{map}}(i_{\mathsf{meas}}) = \arg\max_j \ \mathbf{Prob}(X_j \,|\, Y_{i_{\mathsf{meas}}})}$$

If we measure that event $Y_{i_{\mathsf{meas}}}$ occurred, then we estimate which event $X_1, \dots, X_n$ occurred by picking the one which has the highest conditional probability

- We pick j to maximize the *conditional probability*

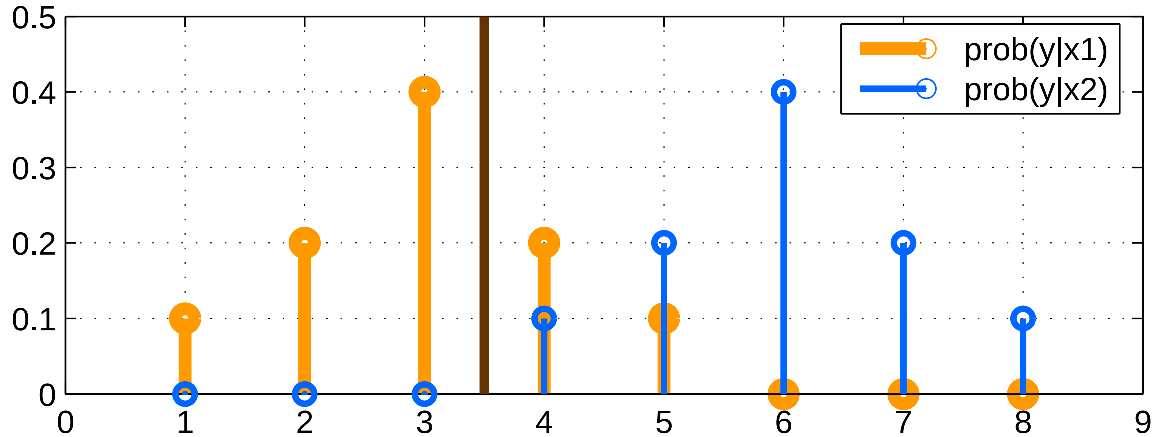
$$\mathbf{Prob}(X_j \,|\, Y_i) = \frac{\mathbf{Prob}(Y_i \,|\, X_j)\, \mathbf{Prob}(X_j)}{\mathbf{Prob}(Y_i)}$$

- This is the same as picking j to maximize the *joint probability*

$$\mathbf{Prob}(Y_i \,|\, X_j)\, \mathbf{Prob}(X_j)$$

Example

Here $n = 2$ and $m = 8$.



We have transition, prior, joint and conditional probabilities

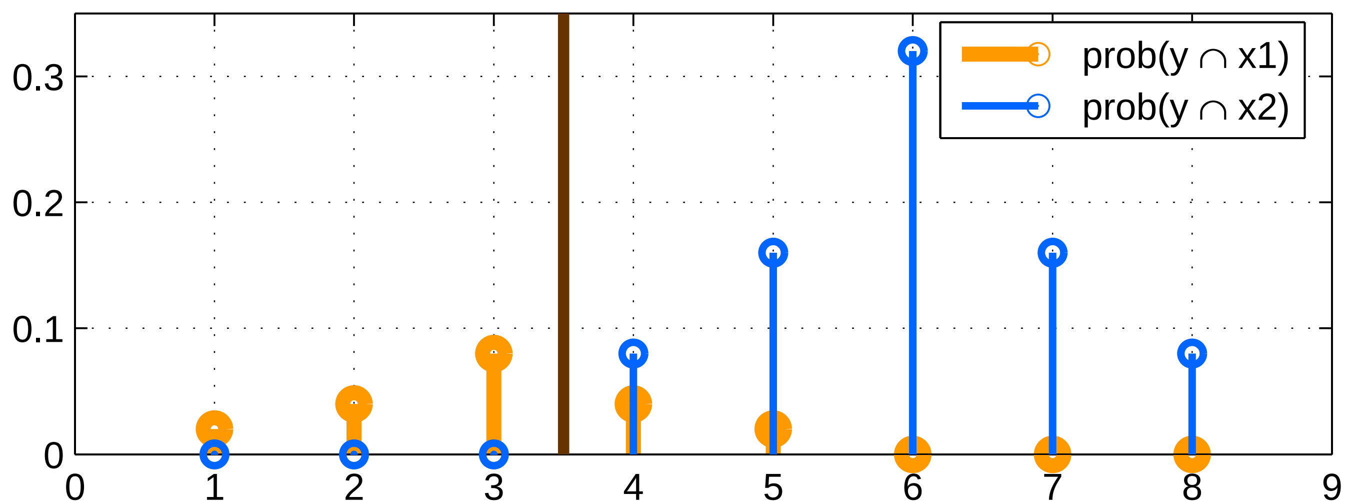
$$A = \begin{bmatrix} 0.1 & 0 \\ 0.2 & 0 \\ 0.4 & 0 \\ 0.2 & 0.1 \\ 0.1 & 0.2 \\ 0 & 0.4 \\ 0 & 0.2 \\ 0 & 0.1 \end{bmatrix} \qquad x = \begin{bmatrix} 0.2 \\ 0.8 \end{bmatrix} \qquad J = \begin{bmatrix} \boxed{0.02} & 0 \\ \boxed{0.04} & 0 \\ \boxed{0.08} & 0 \\ 0.04 & \boxed{0.08} \\ 0.02 & \boxed{0.16} \\ 0 & \boxed{0.32} \\ 0 & \boxed{0.16} \\ 0 & \boxed{0.08} \end{bmatrix} \qquad B = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1/3 & 2/3 \\ 1/9 & 8/9 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

The MAP Classifier

In terms of B and J, the MAP estimator is

$$\boxed{\text{pick } j \text{ corresponding to the largest element in row } i_{\text{meas}} \text{ of } B}$$

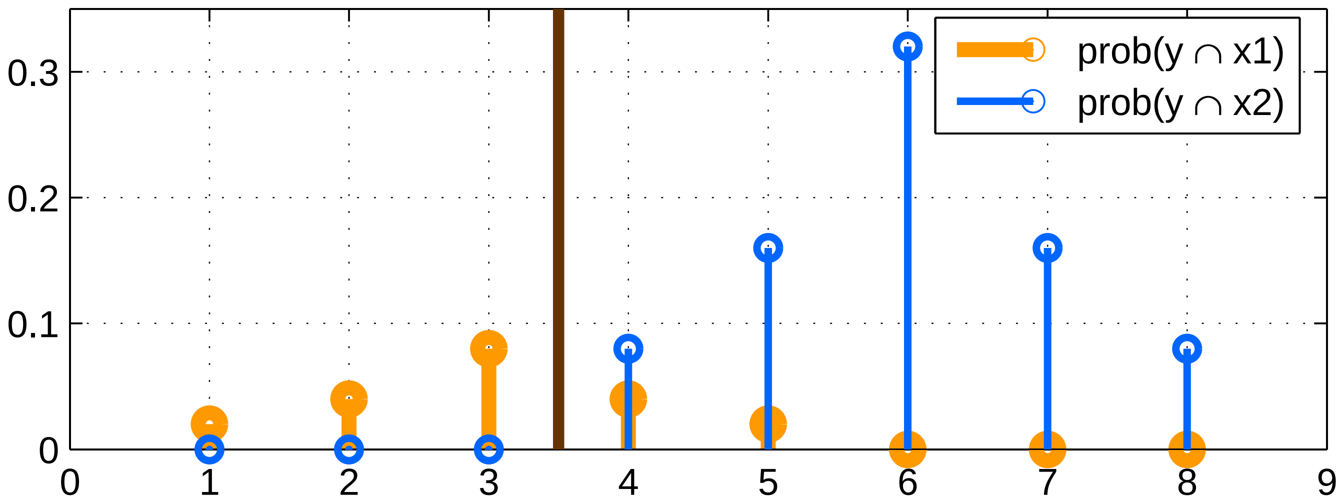
Equivalently, we can use J instead of B; the columns of J are plotted below.



in words: *scale* the transition pdf $\mathbf{Prob}(Y_i \,|\, X_j)$ by the prior pdf $\mathbf{Prob}(X_j)$, and pick the largest evaluated at $Y_{i_{\text{meas}}}$.

Decision Regions

The classifier splits the set of observations into *decision regions*

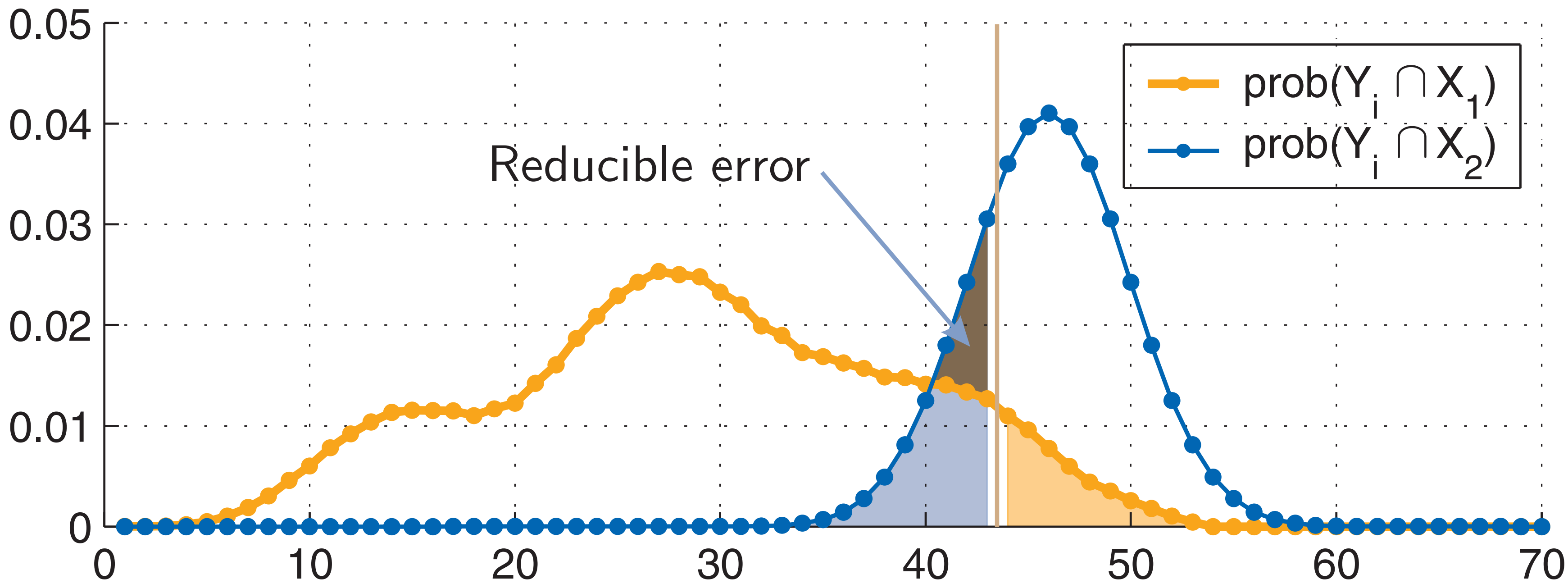


The decision regions are

$$R_1 = \{ Y_i \mid i \leq 3 \}$$
$$R_2 = \{ Y_i \mid i > 3 \}$$

- if $Y_{i_{\text{meas}}} \in R_i$, then we estimate that X_i occurred.
- We will see that this idea is useful when estimating in continuous probability spaces

Reducible Error



- The area (probability mass) under both curves sums to 1.
- If we choose the decision boundary shown at $i = 43$, then the error probability is the area of the three shaded regions
- By moving the decision boundary to 40, we can remove the *reducible error*

Example

Suppose there are four coins in a bag, some gold and some silver. Let

$$X_j = \mathbf{Prob}(j-1 \text{ of the coins in the bag are gold}) \qquad i = 1, \dots, 5$$

We have the prior pdf $x_j = \mathbf{Prob}(X_j)$

$$x = \begin{bmatrix} 0.05 & 0.15 & 0.15 & 0.6 & 0.05 \end{bmatrix}^T$$

We draw two coins at random from the bag. Let

$$Y_i = \mathbf{Prob}(i-1 \text{ of the coins drawn are gold})$$

Example

The transition matrix is

$$A = \begin{bmatrix} 1 & 1/2 & 1/6 & 0 & 0 \\ 0 & 1/2 & 2/3 & 1/2 & 0 \\ 0 & 0 & 1/6 & 1/2 & 1 \end{bmatrix}$$

As usual, $A_{ij} = \mathbf{Prob}(Y_i \mid X_j)$

Because, if there are q gold coins in the bag, then

- the probability of drawing 0 gold coins is $(4-q)(3-q)/12$
- the probability of drawing 1 gold coin is $q(4-q)/6$
- the probability of drawing 2 gold coins is $q(q-1)/12$

Example

The joint probability matrix is

$$J = \begin{bmatrix} 0.05 & \boxed{0.075} & 0.025 & 0 & 0 \\ 0 & 0.075 & 0.1 & \boxed{0.3} & 0 \\ 0 & 0 & 0.025 & \boxed{0.3} & 0.05 \end{bmatrix}$$

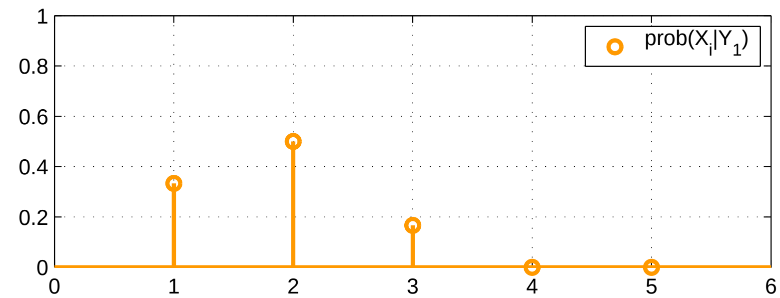
The map estimator is

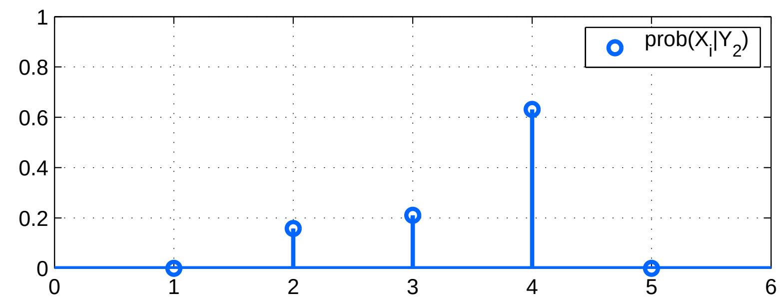
$$K = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

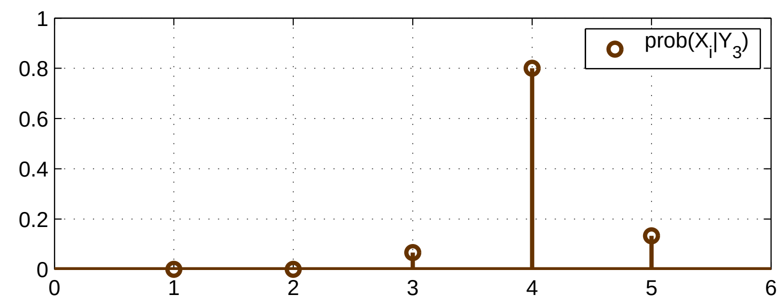
So, using the MAP estimator, we conclude

- if we draw no gold coins, we estimate there was 1 gold coin in the bag
- if we draw 1 or 2 or gold coins, we estimate there were 3 gold coins in the bag

The a-posteriori probabilities are shown on the right for each of the three possible measurements







Error Analysis

The *unconditional error matrix* $E \in \mathbb{R}^{n\times n}$ is

$$\begin{aligned}
E_{jk} &= \text{probability that } X_j \text{ is estimated and } X_k \text{ occurs} \\
&= \mathbf{Prob}(j_{\text{est}} = j \text{ and } X_k) \\
&= \sum_{i=1}^m \mathbf{Prob}\Big(j_{\text{est}} = j \text{ and } Y_i \text{ and } X_k\Big) \qquad \text{since the } Y_i \text{ partition } \Omega \\
&= \sum_{i=1}^m \mathbf{Prob}\Big(\bigcup \{\, Y_p \mid f_{\text{est}}(p) = j \,\} \cap Y_i \text{ and } X_k\Big)
\end{aligned}$$

Now notice that

$$\begin{aligned}
\bigcup \{\, Y_p \mid f_{\text{est}}(p) = j \,\} \cap Y_i &= \begin{cases} Y_i & \text{if } f_{\text{est}}(i) = j \\ \emptyset & \text{otherwise} \end{cases} \\
&= \begin{cases} Y_i & \text{if } K_{ij} = 1 \\ \emptyset & \text{otherwise} \end{cases}
\end{aligned}$$

Error Analysis

Therefore we have

$$\begin{aligned} E_{jk} &= \text{probability that } X_j \text{ is estimated and } X_k \text{ occurs} \\ &= \sum_{i=1}^{m} K_{ij}\, \mathbf{Prob}(Y_i \cap X_k) \\ &= \sum_{i=1}^{m} K_{ij} J_{ik} \end{aligned}$$

That is, $E = K^T J$.

Notice that $\mathbf{1}^T E \mathbf{1} = 1$.

Example: Error Analysis

For the coins example, we have

$$E = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0.05 & 0.075 & 0.025 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0.075 & 0.125 & 0.6 & 0.05 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

- Some rows are zero, since, e.g., we never estimate that are no coins in the bag.
- Ideally, we would have E zero on the off-diagonal elements.
- Notice that each column j sums to the prior probability $\mathbf{Prob}(X_j)$

Error Analysis

The probability that the estimate is correct is

$$\sum_{j=1}^{n} E_{jj} = \mathbf{trace}\, E$$

$$= \sum_{j=1}^{n} \sum_{i=1}^{m} K_{ij} J_{ij}$$

Hence to maximize the probability of a correct estimate, we pick K so that

$$K_{ij} = \begin{cases} 1 & \text{if } J_{ij} \text{ is the largest element of row } i \text{ of } J \\ 0 & \text{otherwise} \end{cases}$$

This is exactly the MAP classifier; i.e.,

$$\boxed{\text{The MAP classifier maximizes the probability of a correct estimate}}$$

Cost Functions

Suppose we now assign *costs* to errors

$$C_{jk} = \text{cost when } X_j \text{ is estimated and } X_k \text{ occurs}$$

The *expected cost* is

$$\begin{aligned}
\mathbf{E}\, C &= \sum_{j=1}^{n} \sum_{k=1}^{n} C_{jk}\, \mathbf{Prob}(j_{\text{est}} = j \text{ and } X_k) \\
&= \sum_{j=1}^{n} \sum_{k=1}^{n} C_{jk} E_{jk} \\
&= \mathbf{trace}(EC^T) \\
&= \mathbf{trace}(K^T J C^T)
\end{aligned}$$

This is called the *Bayes risk*

Cost Functions

Suppose we assign cost

$$C_{jk} = \begin{cases} 1 & \text{if } j \neq q \qquad \text{i.e., the estimate is wrong} \\ 0 & \text{otherwise} \end{cases}$$

That is

$$C = \begin{bmatrix} 0 & 1 & \dots & & 1 \\ 1 & 0 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & & \\ & & & 0 & 1 \\ 1 & \dots & & 1 & 0 \end{bmatrix} = \mathbf{1}\mathbf{1}^T - I$$

Then the Bayes risk is

$$\mathbf{E}\, C = \mathbf{trace}\big(E(\mathbf{1}\mathbf{1}^T - I)\big) = 1 - \mathbf{trace}\, E$$

- Hence minimizing this cost function maximizes the probability of a correct estimate.
- So the MAP classifier minimizes *this* cost function.

Choosing a Cost Function

Suppose we consider the radar example, where

$$\begin{aligned} X_1 &= \text{the event that there are no aircraft present} \\ X_2 &= \text{the event that there is an aircraft present} \end{aligned}$$

Then we may significantly prefer false positives to false negatives.

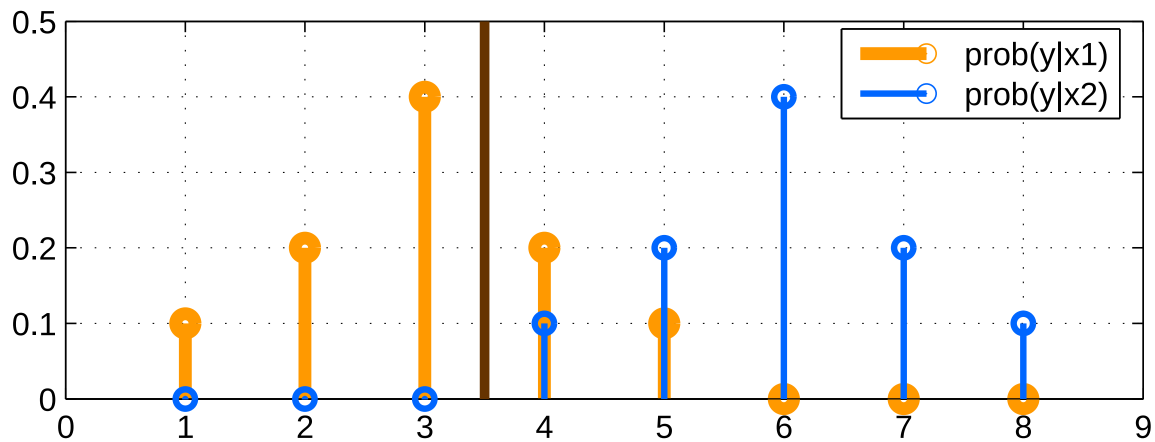
In that case we could choose, for example

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

- C_{21} is the cost for estimating X_2 when X_1 occurs
  i.e., the cost for *false positives*

- C_{12} is the cost for estimating X_1 when X_2 occurs
  i.e., the cost for *false negatives*

Example: Choosing a Cost Function



We would like to minimize $\mathbf{E}\, C = \mathbf{trace}(K^T J C^T)$, so we pick the smallest element in each row of JC^T

$$A = \begin{bmatrix} 0.1 & 0 \\ 0.2 & 0 \\ 0.4 & 0 \\ 0.2 & 0.1 \\ 0.1 & 0.2 \\ 0 & 0.4 \\ 0 & 0.2 \\ 0 & 0.1 \end{bmatrix} \qquad C = \begin{bmatrix} 0 & 100 \\ 1 & 0 \end{bmatrix} \qquad x = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} \qquad JC^T = \begin{bmatrix} \boxed{0} & 0.05 \\ \boxed{0} & 0.1 \\ \boxed{0} & 0.2 \\ 5 & \boxed{0.1} \\ 10 & \boxed{0.05} \\ 20 & \boxed{0} \\ 10 & \boxed{0} \\ 5 & \boxed{0} \end{bmatrix}$$

Trade-offs

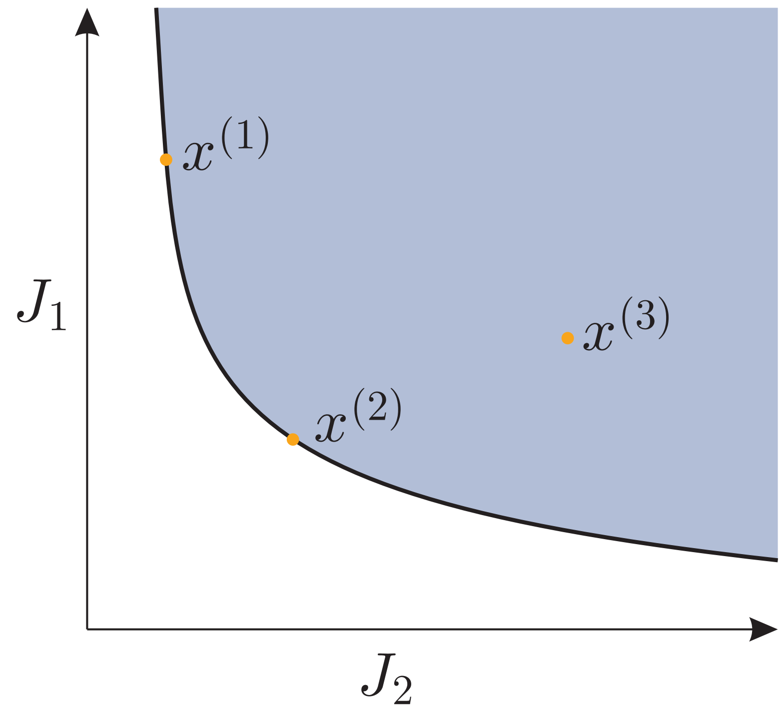
Often we would like to examine the trade off between

- $J_1 =$ the probability of making a false positive error.
- $J_2 =$ the probability of making a false negative error.

- usually the objectives are *competing*
- we can make one smaller at the expense of making the other larger

Trade-off Curve

- shaded area shows (J_2, J_1) achieved by some $x \in \mathbb{R}^n$
- clear area shows (J_2, J_1) not achieved by any $x \in \mathbb{R}^n$
- boundary of region is called *optimal trade-off curve*
- corresponding x called *Pareto optimal*



three example choices of x: $x^{(1)}$, $x^{(2)}$, $x^{(3)}$

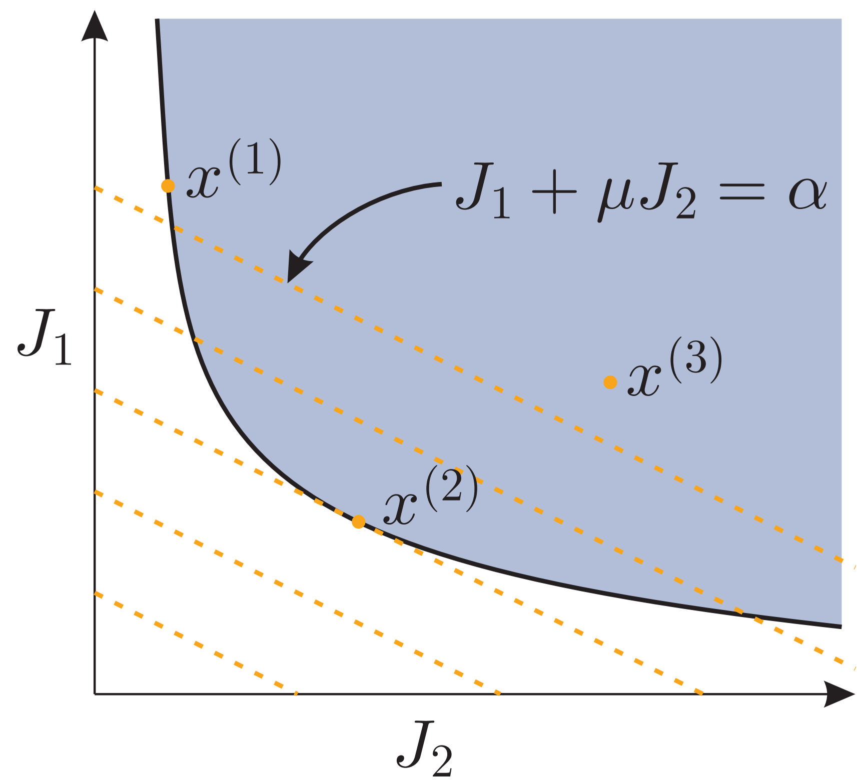
- $x^{(3)}$ is worse than $x^{(2)}$ on both counts (J_2 and J_1)
- $x^{(1)}$ is better than $x^{(2)}$ in J_2, but worse in J_1

Weighted-Sum Objective

to find Pareto optimal points, i.e. x's on optimal trade-off curve, we minimize the *weighted-sum* objective:

$$J_1 + \mu J_2$$

parameter $\mu \geq 0$ gives relative weight between J_1 and J_2



points where weighted sum is constant, $J_1 + \mu J_2 = \alpha$ correspond to line with slope $-\mu$

- $x^{(2)}$ minimizes the weighted-sum objective for μ shown
- by varying μ from 0 to $+\infty$, we can sweep out the entire *optimal trade-off curve*
- In some cases, the trade-off curve may not be *convex*; then there are Pareto points that are not found by minimizing a weighted sum.

Weighted-Sum Objective

We have

$$J_1 = \mathbf{Prob}(j_{\text{est}} = 2 \cap X_1)$$
$$J_2 = \mathbf{Prob}(j_{\text{est}} = 1 \cap X_2)$$

and we would like to minimize $J_1 + \mu J_2$

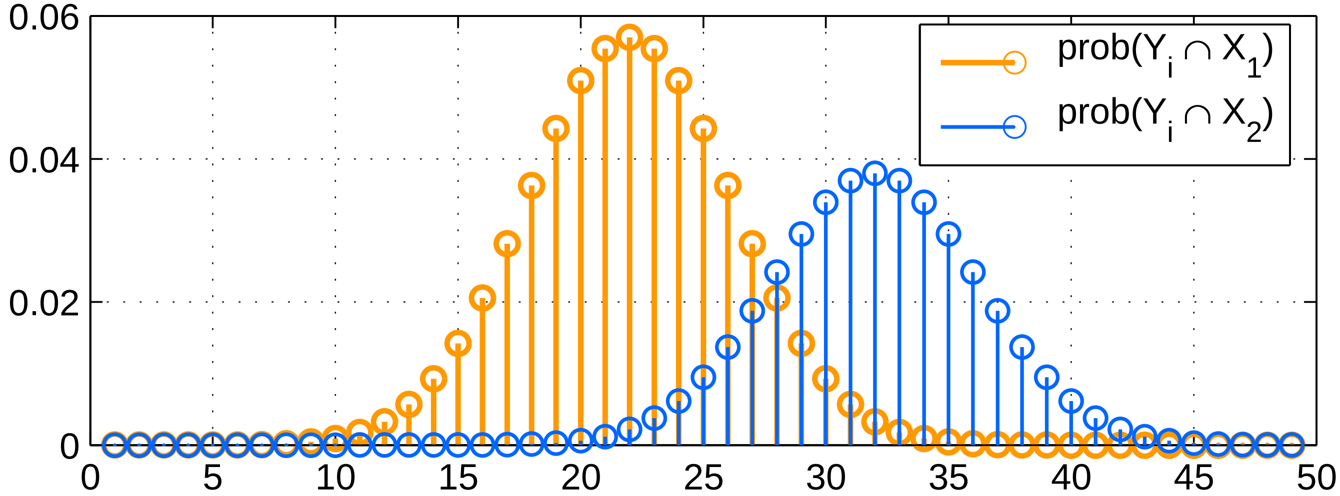
This is the same as picking cost matrix

$$C = \begin{bmatrix} 0 & \mu \\ 1 & 0 \end{bmatrix}$$

This is called the *Neyman-Pearson* cost function.

Example: Weighted-Sum Objective

Consider the joint probabilities

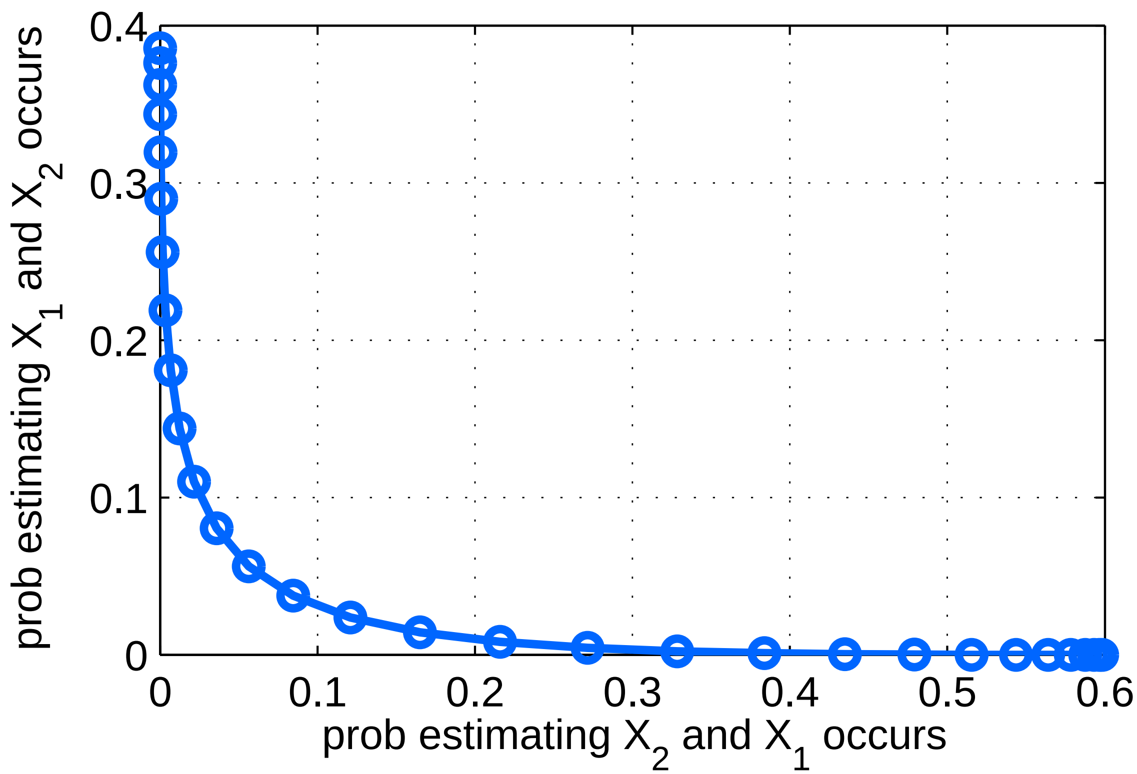


with prior probabilities

$$\mathbf{Prob}(X_1) = 0.6 \qquad \mathbf{Prob}(X_2) = 0.4$$

Example: Weighted-Sum Objective

The trade-off curve is below



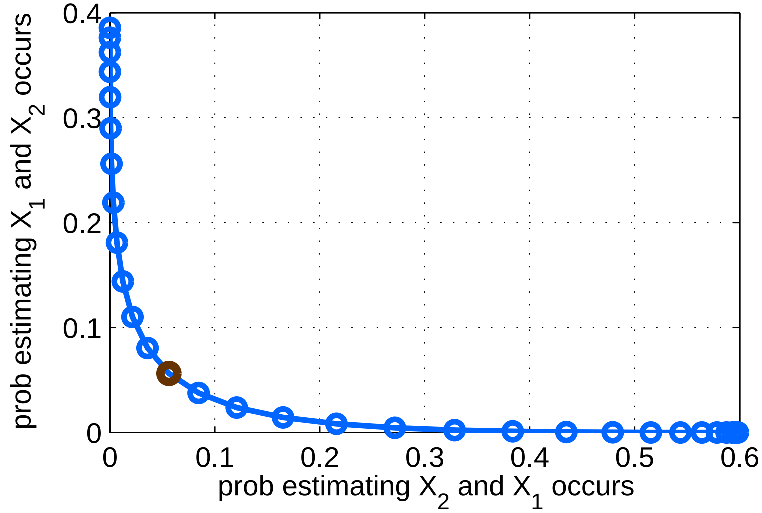
- This curve is called the *operating characteristic*
- Note intersections with axes at prior probabilities
- The pareto-optimal points are a *finite set*, not a continuous curve, since there are only a few choices for threshold value.

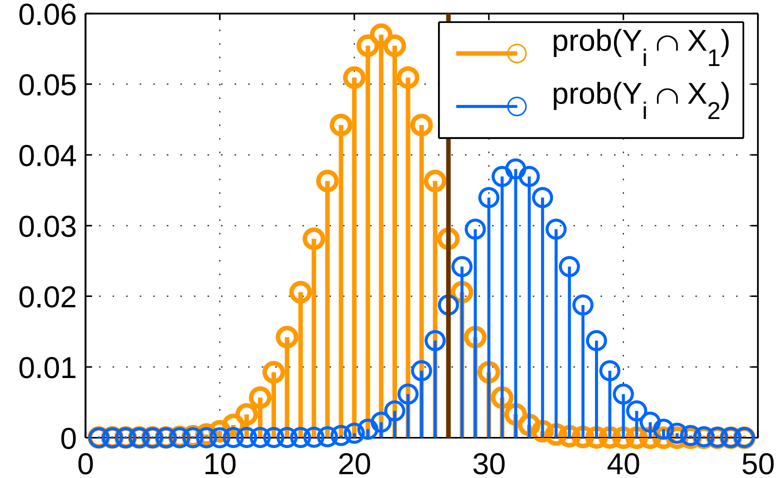
Operating characteristic

- Also called the *receiver operating characteristic* or ROC.
- Often plotted other way up

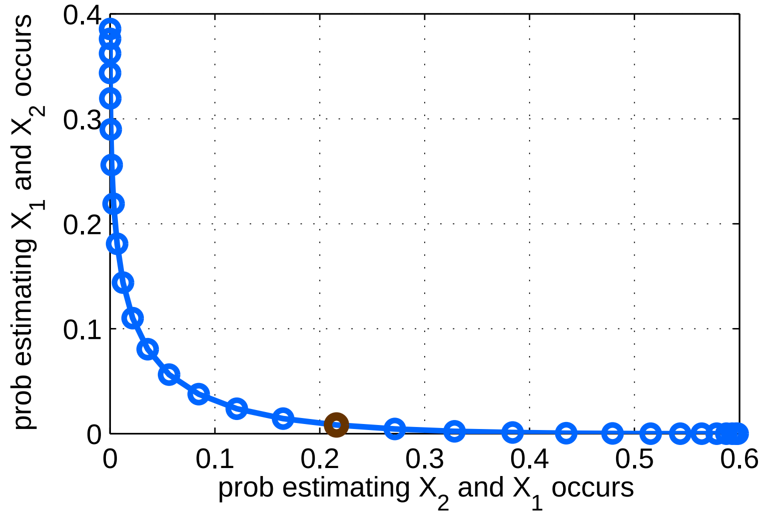
Example: Trading off Errors

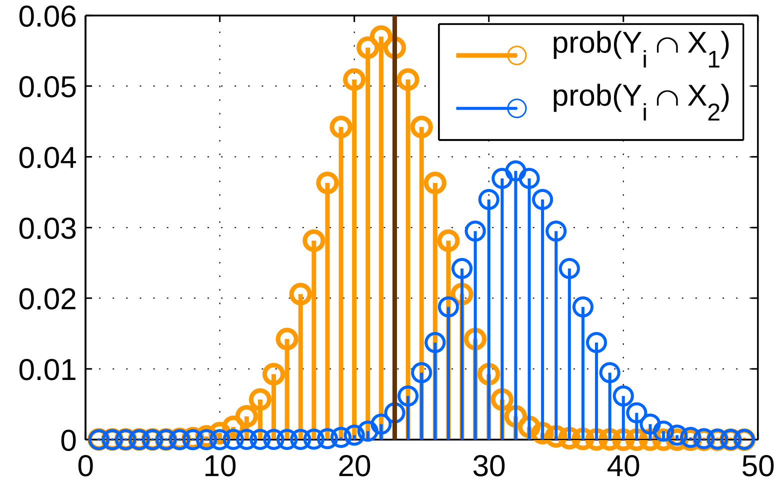
With $\mu = 1$





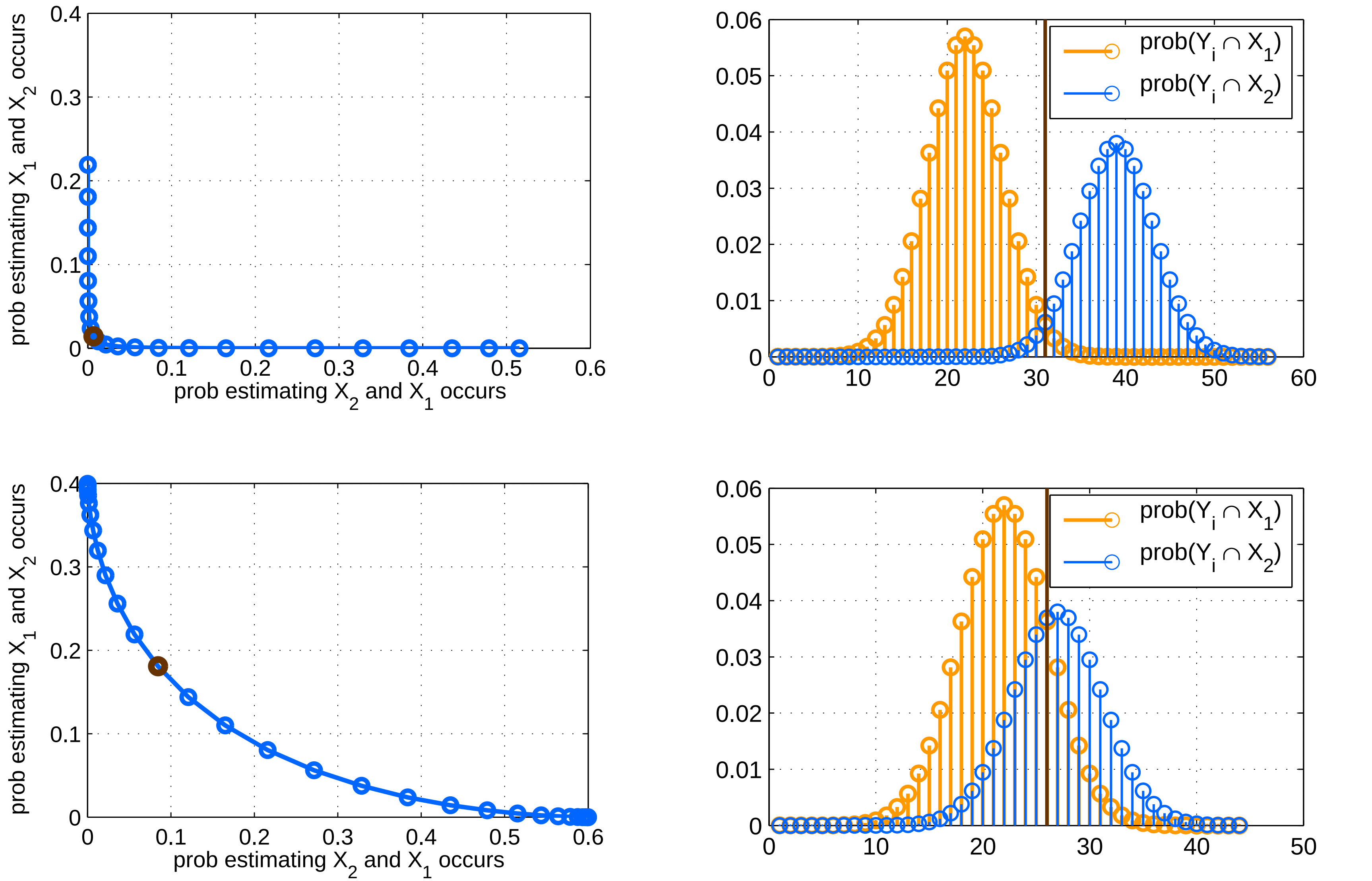
With $\mu = 10$





Example: Trading off Errors

The operating characteristic becomes gentler when it is hard to *distinguish* X_1 from X_2



Conditional Errors

The *conditional error matrix* $E^{\text{cond}} \in \mathbb{R}^{n\times n}$ is

$$
\begin{aligned}
E^{\text{cond}}_{jk} &= \text{probability that } X_j \text{ is estimated given that } X_k \text{ occurred} \\
&= \mathbf{Prob}(j_{\text{est}} = j \mid X_k) \\
&= \sum_{i=1}^{m} \mathbf{Prob}\Big(j_{\text{est}} = j \text{ and } Y_i \mid X_k\Big) \text{ since the } Y_i \text{ partition } \Omega \\
&= \sum_{i=1}^{m} \mathbf{Prob}\Big(\bigcup \{\, Y_p \mid \phi(p) = j \,\} \cap Y_i \mid X_k\Big)
\end{aligned}
$$

Now notice that

$$
\begin{aligned}
\bigcup \{\, Y_p \mid \phi(p) = j \,\} \cap Y_i &= \begin{cases} Y_i & \text{if } \phi(i) = j \\ \emptyset & \text{otherwise} \end{cases} \\
&= \begin{cases} Y_i & \text{if } K_{ij} = 1 \\ \emptyset & \text{otherwise} \end{cases}
\end{aligned}
$$

Conditional Errors

Therefore we have

$$\begin{aligned} E^{\text{cond}}_{jk} &= \sum_{i=1}^{m} K_{ij}\, \mathbf{Prob}(Y_i \mid X_k) \\ &= \sum_{i=1}^{m} K_{ij} A_{ik} \end{aligned}$$

That is

$$\boxed{E^{\text{cond}} = K^T A}$$

Conditional Errors

For the coins example, we have

$$E^{\text{cond}} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0.5 & 1/6 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0.5 & 5/6 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

E^{cond}_{jk} is the probability that X_j is estimated given that X_k occurred

- $\mathbf{1}^T E^{\text{cond}} = \mathbf{1}^T$, i.e., the column sums are one

  Because, when X_k occurs, some X_j is always estimated

- Ideally we would like $E^{\text{cond}} = I$

Maximum-Likelihood

When we do not have any prior probabilities, a commonly used heuristic is the method of *maximum likelihood*.

- MAP estimate: pick j to maximize the *joint probability*

$$\mathbf{Prob}(Y_i \,|\, X_j)\,\mathbf{Prob}(X_j)$$

- *Max Likelihood:* pick j to maximize the *a-priori probability*

$$\mathbf{Prob}(Y_i \,|\, X_j)$$

- We can also minimize costs associated with errors. In this case we minimize $\mathbf{trace}(E^{\text{cond}}C^T)$ instead of $\mathbf{trace}(EC^T)$.
- Similarly, we can construct a trade-off curve using these costs.
- The estimates are identical to those obtained when *all prior probabilities are equal.*