ECE531 Lecture 2a: A Mathematical Model for Hypothesis Testing (Finite Number of Possible Observations)

D. Richard Brown III

Worcester Polytechnic Institute

26-January-2011

Hypothesis Testing Basics

Examples of hypotheses:

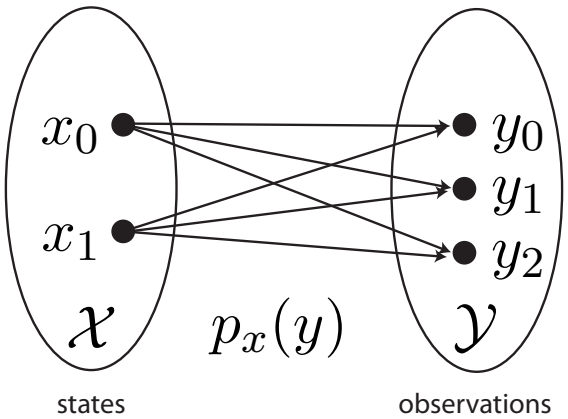
- The coin is fair ($\mathcal{H}_0$) or not fair ($\mathcal{H}_1$).
- The approaching airplane is friendly ($\mathcal{H}_0$) or unfriendly ($\mathcal{H}_1$).
- This email is spam ($\mathcal{H}_1$) or not spam ($\mathcal{H}_0$).
- The medical treatment is effective ($\mathcal{H}_1$) or ineffective ($\mathcal{H}_0$).
- Lance Armstrong used performance enhancing drugs ($\mathcal{H}_1$) or didn't ($\mathcal{H}_0$).
- Communication receiver: Given a codebook with M codewords, which codeword was sent ($\{\mathcal{H}_0, \ldots, \mathcal{H}_{M-1}\}$)?

Given a "noisy" observation, we want to decide among two or more possible underlying statistical situations ("hypotheses").

More generally, we want to specify a decision rule that maps observations to decisions optimally in some sense.

States and Observations

- Let $x \in \mathcal{X} = \{x_0, \ldots, x_{N-1}\}$ denote the **state**, a hidden variable about which we wish to make an inference.
- The available **observation** is modeled as a random variable Y taking on values in the set $\mathcal{Y} = \{y_0, \ldots, y_{L-1}\}$ (we will generalize to infinite $\mathcal{Y}$ later).
- For each state $x \in \mathcal{X}$, we assume that we are given a **probabilistic description** of the random variable Y when the state is x. The notation $p_x(y) = p_Y(y|x)$ means either the probability mass function (pmf) or the probability density function (pdf) of the random variable Y when the state is x.



Example

An unknown coin is fair (HT) or double-headed (HH). We want to determine which it is. We can flip the coin three times and record each outcome (heads or tails).

- What are the possible states $\mathcal{X}$? $\mathcal{X} = \{\mathsf{HT},\mathsf{HH}\}$.
- What are the possible observations $\mathcal{Y}$? $\mathcal{Y} = \{\mathsf{HHH},\mathsf{HHT},\ldots,\mathsf{TTT}\}$.
- What is $p_{\mathsf{HT}}(y)$? $p_{\mathsf{HT}}(y = \mathsf{HHH}) = \cdots = p_{\mathsf{HT}}(y = \mathsf{TTT}) = \frac{1}{8}$.
- What is $p_{\mathsf{HH}}(y)$? $p_{\mathsf{HH}}(y = \mathsf{HHH}) = 1$, $p_{\mathsf{HH}}(y \neq \mathsf{HHH}) = 0$.

Remark:

- Even though we don't know the state, we always assume a known probabilistic model for the observations. This assumption is critical for hypothesis testing.

Hypotheses and Decisions

- **Hypotheses** can be represented as a partition of $\mathcal{X}$, denoted by $\mathcal{H} = \{\mathcal{H}_0, \mathcal{H}_1, \ldots, \mathcal{H}_{M-1}\}$ where

$$\begin{aligned} \mathcal{H}_i &\subseteq \mathcal{X} \\ \mathcal{H}_i &\neq \emptyset \\ \mathcal{H}_i \bigcap \mathcal{H}_j &= \emptyset \text{ for } i \neq j \text{ and} \\ \bigcup_i \mathcal{H}_i &= \mathcal{X} \end{aligned}$$

- The set of possible **decisions** is then $\mathcal{Z} = \{0, 1, \ldots, M-1\}$ where decision i indicates the selection of hypothesis $\mathcal{H}_i$. In other words, decision i is the decision that $x \in \mathcal{H}_i$.
- If $\mathcal{X}$ is finite, then we must have $M \leq N$.

Types of Hypothesis Testing Problems

Recall $N = |\mathcal{X}|$ is the number of states (assume $\mathcal{X}$ is finite for now) and $M = |\mathcal{H}|$ is the number of hypotheses.

- If $M = 2$, then we have a **binary** hypothesis testing problem.
- If $M = N$, then we seek to decide the actual state. In this case we can take $\mathcal{H}_i = \{x_i\}$ and we have a **simple** hypothesis testing problem.
- If $M < N$ or $\mathcal{X}$ is infinite, then we have a **composite** hypothesis testing problem. At least one hypothesis contains more than one state.

Unlike a simple hypothesis with underlying distribution $p_x(y)$, a composite hypothesis does not completely specify the underlying distribution.

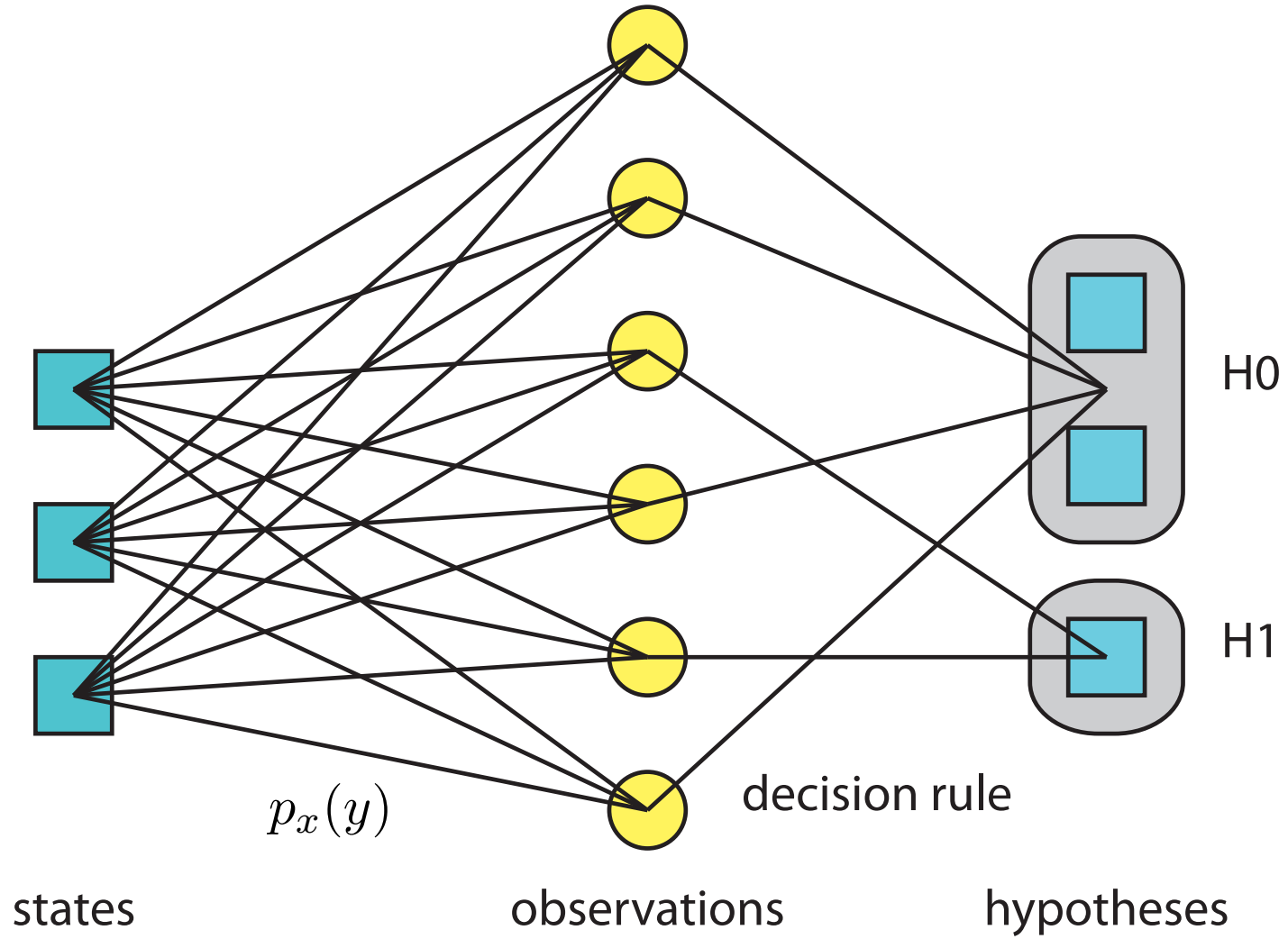
Our focus will be on simple hypothesis testing problems for now, but we will return to composite hypothesis testing in a few weeks.

Examples

We have a coin with $\mathrm{Prob}(\mathrm{H}) = q$ unknown.

1. Suppose q can only take on two values: q_0 or q_1. What kind of hypothesis testing problem is this? Binary, simple.
2. Suppose q can take on any value in the set $\{q_0, q_1, \ldots, q_{M-1}\}$ and we wish to determine which value it is. What kind of hypothesis testing problem is this? M-ary, simple.
3. Suppose q can take on any value in the set $\{q_0, q_1, \ldots, q_{N-1}\}$ but only wish to know if it is q_0 or not (e.g. $q_0 = 0.5$ "is the coin fair?"). What kind of hypothesis testing problem is this? Binary, composite $M = 2 < N$.
4. Suppose q can be any value in $[0, 1]$ and we want to determine this value. What kind of problem is this? Estimation.

Model Summary



Finite Observation Sets: Conditional Probability Matrix

When $\mathcal{X}$ and $\mathcal{Y}$ are finite with $|\mathcal{X}| = N$ and $|\mathcal{Y}| = L$, we can conveniently represent the conditional probabilities $p_x(y)$ in matrix form:

$$P = \begin{bmatrix} p_{x=x_0}(y = y_0) & \cdots & p_{x=x_{N-1}}(y = y_0) \\ \vdots & \ddots & \vdots \\ p_{x=x_0}(y = y_{L-1}) & \cdots & p_{x=x_{N-1}}(y = y_{L-1}) \end{bmatrix} \in \mathbb{R}^{L \times N}$$

Decision Rules

- We can think of a decision rule as a mapping from observations to hypotheses. Specifically, given observation index $\ell \in \{0, \ldots, L-1\}$, our decision rule tells us how to decide the hypothesis index $m \in \{0, \ldots, M-1\}$.
- Deterministic decision rules partition the observation space into subsets $\mathcal{Y}_0, \ldots, \mathcal{Y}_{M-1}$ such that

$$y \in \mathcal{Y}_i \quad \Rightarrow \quad \text{decide } \mathcal{H}_i$$

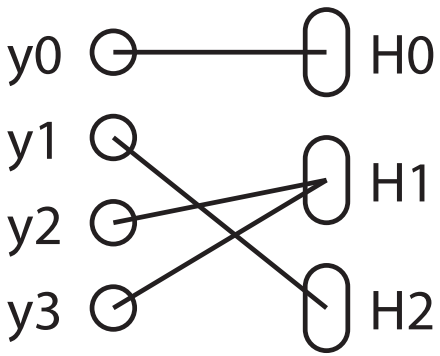
  with $\mathcal{Y}_i \subseteq \mathcal{Y}$, $\mathcal{Y}_i \bigcap \mathcal{Y}_j = \emptyset$ for $i \neq j$, and $\bigcup_i \mathcal{Y}_i = \mathcal{Y}$.
- There are lots of ways of specifying decision rules.

Decision Matrices

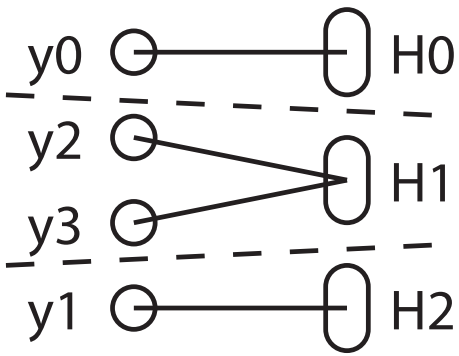
When we have a finite number of possible observations, one way to specify a decision rule is a **decision matrix** $D \in \mathbb{R}^{M \times L}$, e.g.

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

We can think of this graphically as



or



Finite Observation Sets: Conditional Decision Probabilities

Let

$$T = DP \in \mathbb{R}^{M \times N}$$

Note that

$$\begin{aligned} T_{ij} &= \sum_{k=0}^{L-1} D_{ik} P_{kj} \\ &= \sum_{k=0}^{L-1} D_{ik} \text{Prob}(y = y_k | x = x_j) \end{aligned}$$

Interpretation: T_{ij} is the probability of deciding $\mathcal{H}_i$ when the state is x_j.

Finite Observation Sets: Decision Costs

- Our goal is to specify a decision rule that is optimum in some sense.
- To do this, we specify a matrix C of **decision costs** where C_{ij} is the cost of deciding $\mathcal{H}_i$ when the state is x_j.

Examples:

1. Uniform cost assignment (UCA)

$$C_{ij} = \begin{cases} 0 & \text{if } i = j \\ 1 & \text{if } i \neq j \end{cases}$$

2. Quadradic cost assignment ($M = N$ and $\mathcal{X}$ is a subset of $\mathbb{R}$)

$$C_{ij} = (x_i - x_j)^2$$

Finite Observation Sets: Conditional Risks

Notation:

- $t_j \in \mathbb{R}^M$ = jth column of $T = DP$. This column contains the probabilities of deciding $\mathcal{H}_0, \ldots, \mathcal{H}_{M-1}$ when the state is x_j.
- $c_j \in \mathbb{R}^M$ = jth column of cost matrix C. This column contains the costs of deciding $\mathcal{H}_0, \ldots, \mathcal{H}_{M-1}$ when the state is x_j.
- $p_j \in \mathbb{R}^L$ = jth column of conditional probability matrix P. This column contains the probabilities of observing $y_0, \ldots, y_{L-1}$ when the state is x_j.

Note that the inner product

$$R_j(D) = c_j^\top t_j = c_j^\top D p_j \qquad j \in \{0, \ldots, N-1\}$$

gives the expected cost (also called the **conditional risk**) of using the decision matrix D when the state is x_j.

Working Example: Part 1

Scenario

We have a scenario with n i.i.d. coin flips where a H occurs with probability q and a T occurs with probability $1-q$. The parameter q takes one of two possible values $0 \leq q_0 < q_1 \leq 1$.

- The observation is the number of heads.
- We want to decide between $\mathcal{H}_0 : q = q_0$ or $\mathcal{H}_1 : q = q_1$.

- The set of states $\mathcal{X} = \{x_0 : q = q_0, x_1 : q = q_1\}$. $N = |\mathcal{X}| = 2$.
- The observation space $\mathcal{Y} = \{0, \ldots, n\}$ with

$$p_j(y = k) \quad = \quad \binom{n}{k} q_j^k (1 - q_j)^{n-k}$$

$L = |\mathcal{Y}| = n + 1$.

- This is a simple binary hypothesis testing problem since $M = N = 2$.

Working Example: Part 2

Suppose we have $n = 3$ coin flips. Then

$$P = \begin{bmatrix} (1-q_0)^3 & (1-q_1)^3 \\ 3q_0(1-q_0)^2 & 3q_1(1-q_1)^2 \\ 3q_0^2(1-q_0) & 3q_1^2(1-q_1) \\ q_0^3 & q_1^3 \end{bmatrix}$$

Suppose also that we use the uniform cost assignment

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

Note that there are a finite number of (deterministic) decision matrices that we can consider:

$$D \in \left\{ \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}, \ldots, \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \right\}$$

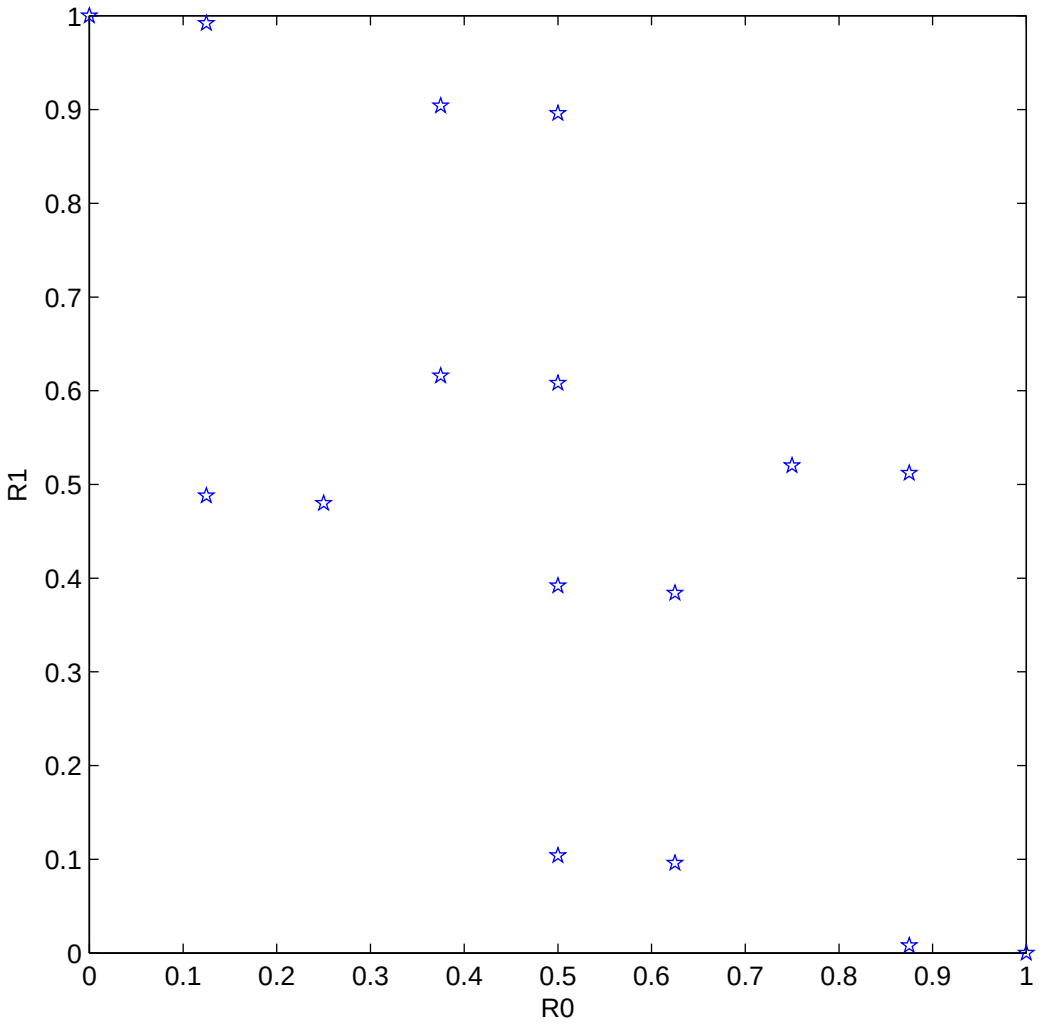
Working Example: Part 3

We can group the conditional risks $R_j(D)$ into an N-vector

$$R(D) = \begin{bmatrix} R_0(D) \\ R_1(D) \end{bmatrix} = \begin{bmatrix} c_0^\top D p_0 \\ c_1^\top D p_1 \end{bmatrix}$$

- $R(D) \in \mathbb{R}^N$ is called the **conditional risk vector** (CRV).
- Ideally, we would like both $R_0(D)$ and $R_1(D)$ to be small. It is usually not possible, however, to find a D that minimizes both simultaneously.
- To see this, we can plot the coordinates of these vectors in $\mathbb{R}^2$ for each of the (deterministic) decision rules...

Working Example: Risk Vectors [$q_0 = 0.5$ and $q_1 = 0.8$]

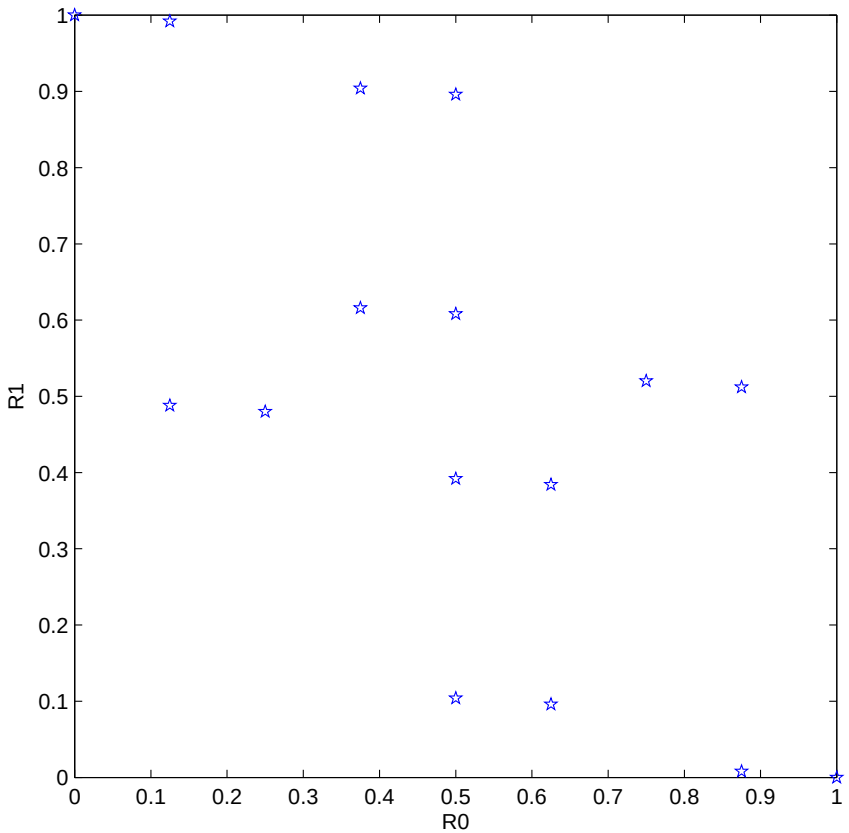


```
% ECE531 DRB 25-Jan-2011
% Plot the conditional risk vectors for a simple binary HT problem
N = 2;                  % number of hypotheses
M = 2;                  % number of states
n = 3;                  % number of flips
q0 = 0.5;               % prob heads under H0
q1 = 0.8;               % prob heads under H1
C = [0 1 ; 1 0];        % UCA
L = n+1;                % number of possible observations
totD = M^L;             % total number of decision matrices
B = makebinary(L,1);

% make conditional probability matrix
P0 = zeros(L,1);
P1 = zeros(L,1);
for i = 0:(L-1),
    P0(i+1) = nchoosek(n,i) * q0^i * (1 - q0)^(n-i);
    P1(i+1) = nchoosek(n,i) * q1^i * (1 - q1)^(n-i);
end
P = [P0 P1];

% compute CRVs for all possible deterministic decision matrices
for i = 0:(totD-1),
    D =  [ B(:,i+1)' ; 1-B(:,i+1)' ];    % decision matrix
    for j=0:N-1,
        R(j+1,i+1) = C(:,j+1)'*D*P(:,j+1);
    end
end

% plot
plot(R(1,:),R(2,:),'p'); xlabel('R0'); ylabel('R1');
axis square; grid on
```

```
function y=makebinary(K,unipolar)

  y=zeros(K,2^K);                    % all possible bit combos
  for index=1:K,
    y(K-index+1,:)=(-1).^ceil([1:2^K]/(2^(index-1)));
  end
  if unipolar>0,
    y = (y+1)/2;
  end
```

The Problem With Deterministic Decision Rules

- When the observation space is finite, there are only a finite number of deterministic decision matrices and achievable CRVs. How many? M^L.
- In our working example, what if we wanted to balance the risk such that $R_0(D) = R_1(D) = 0.4$?



Randomized Decision Rules

- So far, we have considered only deterministic decision rules. Given an observation $y \in \mathcal{Y}$, a deterministic decision rule is a map from $\mathcal{Y}$ directly to $\mathcal{Z}$ (the indices of the hypotheses).
- A generalization of this idea is a **randomized decision rule**. Given an observation $y \in \mathcal{Y}$, a randomized decision rule is a mapping from $\mathcal{Y}$ to a distribution (a pmf) on $\mathcal{Z}$. The set of valid pmfs on $\mathcal{Z}$ is denoted as $\mathcal{P}_M$.
- Examples of random decision matrices:

$$D = \begin{bmatrix} 0.9 & 0.9 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.9 & 0.9 \end{bmatrix} \text{ or } D = \begin{bmatrix} 0.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 & 0.5 \end{bmatrix}$$

- Note that the elements of D must be non-negative and the columns must sum to one.
- Note that the deterministic decision rules are special cases in the family of randomized decision rules $\mathcal{D}$.

Other Ways of Specifying Decision Rules (1 of 3)

Recall the deterministic **decision matrix** $D : \mathbb{R}^L \mapsto \mathbb{R}^M$

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

+ easily generalizable to random decision rules
+ convenient for generating conditional risk vectors in Matlab
- doesn't work for infinite observations spaces

Another way of specifying a deterministic decision rule is $\delta : \mathcal{Y} \mapsto \mathcal{Z}$

$$\delta(y) = m \text{ if we decide } \mathcal{H}_m \text{ when we observe } y$$

The D above is equivalent to $\delta(y_0) = 0$, $\delta(y_1) = 2$, and $\delta(y_2) = \delta(y_3) = 1$.

+ will work for infinite observations spaces
- not generalizable to random decision rules

Other Ways of Specifying Decision Rules (2 of 3)

A third way of specifying deterministic decision rules is $\delta : \mathcal{Y} \mapsto \mathbb{R}^M$ where

$$\delta_i(y) = \begin{cases} 1 & \text{if we decide } \mathcal{H}_i \text{ when we observe } y \\ 0 & \text{if we don't decide } \mathcal{H}_i \text{ when we observe } y \end{cases}$$

for $i = 0, \ldots, M-1$. Example:

$$D = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix} \Leftrightarrow \delta_i(y_\ell) = \begin{cases} 1 & i = 0 \text{ and } \ell = 0, \text{ or } i = 2 \text{ and } \ell = 1, \\ & \text{or } i = 1 \text{ and } \ell = 2, \text{ or } i = 1 \text{ and } \ell = 3 \\ 0 & \text{otherwise} \end{cases}$$

This generalizes to random decisions, except that we usually use the notation $\rho_i(y)$ to denote a random decision rule, e.g.

$$D = \begin{bmatrix} 0.7 & 0.4 & 0.5 & 0 \\ 0.2 & 0.4 & 0.2 & 0.9 \\ 0.1 & 0.2 & 0.3 & 0.1 \end{bmatrix} \Leftrightarrow \rho_0(y_0) = 0.7,\ \rho_1(y_0) = 0.2, \ldots$$

This is probably the most general way of specifying decision rules, but it can be notationally cumbersome.

Other Ways of Specifying Decision Rules (3 of 3)

In **binary** hypothesis testing problems, there are only two possible decisions: $\mathcal{H}_0$ and $\mathcal{H}_1$. It is convenient in this case to use the more compact notation:

$$\delta(y) = \begin{cases} 1 & \text{if we decide } \mathcal{H}_1 \text{ when we observe } y \\ 0 & \text{if we decide } \mathcal{H}_0 \text{ when we observe } y \end{cases}$$

Since there are only two possibilities, randomized decision rules can be written as

$$\rho(y) = \begin{cases} 1 & \text{if we always decide } \mathcal{H}_1 \text{ when we observe } y \\ \gamma & \text{if we decide } \mathcal{H}_1 \text{ with probability } \gamma \text{ when we observe } y \\ 0 & \text{if we always decide } \mathcal{H}_0 \text{ when we observe } y \end{cases}$$

Advantages and limitations:

+ works for random decision rules
+ work for infinite observations spaces
+ not cumbersome
- only applicable to binary hypothesis testing problems

Why We Like Randomized Decision Rules

Theorem

The family $\mathcal{D}$ of randomized decision rules is a compact, convex set.

Compact: Bounded and closed.
Convex: For each $\theta_1, \theta_2 \in \Theta$ and each $\gamma \in [0, 1]$,

$$\theta_{1,2,\gamma} = (1 - \gamma)\theta_1 + \gamma\theta_2 \in \Theta.$$

Proof.

$\mathcal{D} \subset \mathbb{R}^{M \times L}$. Since, for each $D \in \mathcal{D}$, $0 \leq D_{ij} \leq 1$, $\mathcal{D}$ is a bounded set. $\mathcal{D}$ is also closed because $D_{ij} = 0$ and $D_{ij} = 1$ are included in $\mathcal{D}$. Finally, for any $D, D' \in \mathcal{D}$ and $\gamma \in [0, 1]$

$$D'' = (1 - \gamma)D + \gamma D'$$

satisfies the properties that $0 \leq D''_{ij} \leq 1$ and $\sum_i D''_{ij} = 1$. Hence $D'' \in \mathcal{D}$ and $\mathcal{D}$ is convex. □

Linearity of the Risk Function

Theorem

The function $R : \mathbb{R}^{M\times L} \mapsto \mathbb{R}^N$ that maps a decision rule D to its conditional risk vector $R(D)$ is linear.

Proof.

For any $\gamma_1, \gamma_2 \in \mathbb{R}$ and decision rules $D_1, D_2 \in \mathbb{R}^{M\times L}$

$$\begin{aligned} R_j(\gamma_1 D_1 + \gamma_2 D_2) &= c_j^\top(\gamma_1 D_1 + \gamma_2 D_2)p_j \\ &= \gamma_1 c_j^\top D_1 p_j + \gamma_2 c_j^\top D_2 p_j \\ &= \gamma_1 R_j(D_1) + \gamma_2 R_j(D_2) \end{aligned}$$

Thus $R(\gamma_1 D_1 + \gamma_2 D_2) = \gamma_1 R(D_1) + \gamma_2 R(D_2)$. □

A linear map between finite dimensional vector spaces is continuous.

Achievable Conditional Risk Vectors

As D ranges over all possible decision rules in $\mathcal{D}$, $R(D)$ traces out a set $\mathcal{Q}$ of **achievable conditional risk vectors**. What does $\mathcal{Q}$ look like?

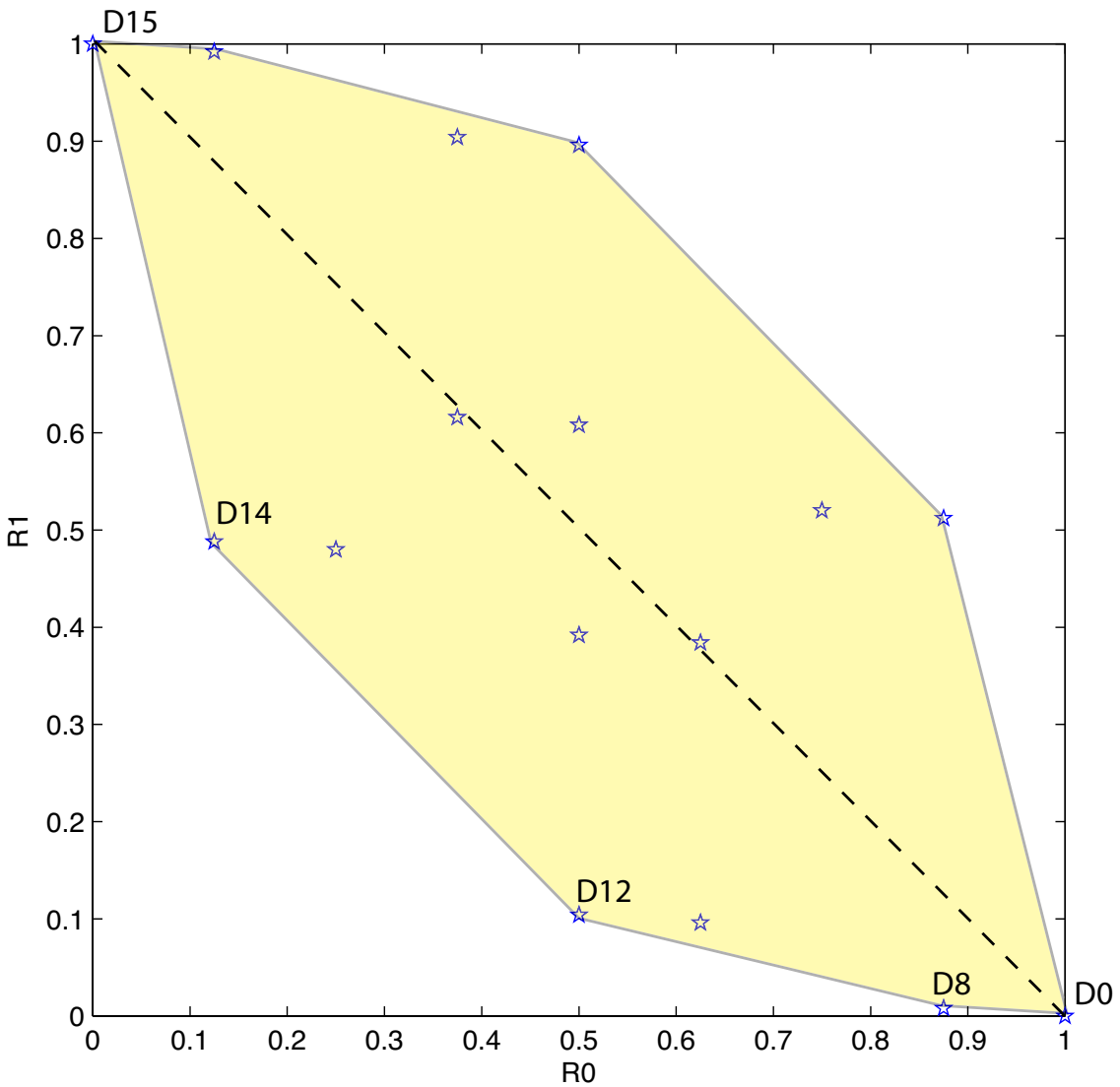
Theorem

$\mathcal{Q}$ is a closed and bounded polytope in $\mathbb{R}^N$

Proof.

$\mathcal{D}$ is a compact, convex polytope in $\mathbb{R}^{M\times L}$. We have $\mathcal{Q} = R(\mathcal{D})$. The map $R : \mathbb{R}^{M\times L} \mapsto \mathbb{R}^N$ is linear. Hence $\mathcal{Q}$ is a polytope since it is the image of a polytope under a linear map. The image of a compact set under a continuous map is compact. Thus $\mathcal{Q}$ is compact and hence closed and bounded. □

Working Example: Risk Vectors [$q_0 = 0.5$ and $q_1 = 0.8$]



- Can we now balance the risk $R_0 = R_1 = 0.4$?
- What does the line $R_0 + R_1 = 1$ represent? Random guessing.
- Where are the "good" decision rules? Southwest of the random guess line.
- What point on the Southwest boundary of $\mathcal{Q}$ corresponds to the best decision rule?

Pareto Optimal Decision Rules

A decision rule D dominates D' if for each $x_j \in \mathcal{X}$, $R_j(D) \leq R_j(D')$ and, for at least one j, the inequality is strict. Dominance is denoted as

$$R(D) \quad \prec \quad R(D')$$

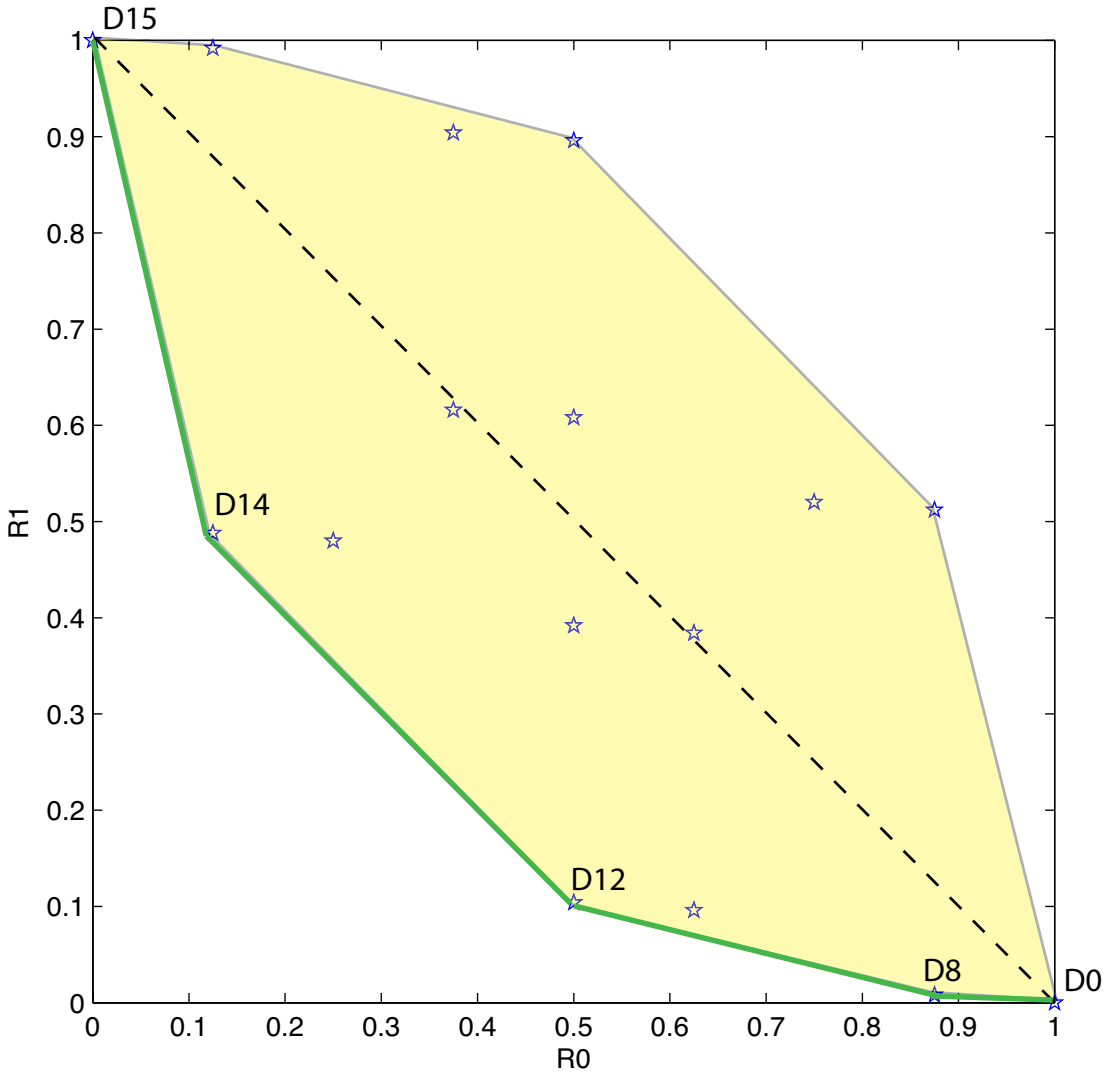
A decision rule D is Pareto optimal if no decision rule dominates it.
In our working example, the decision rules

$$D_0 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}, D_8 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix}, D_{12} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix},$$
$$D_{14} = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \text{ and } D_{15} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

are all Pareto optimal, as are all of the randomized decision rules $D_{0,8,\gamma}$, $D_{8,12,\gamma}$, $D_{12,14,\gamma}$, and $D_{14,15,\gamma}$ for $\gamma \in [0, 1]$.

Optimal Tradeoff Surface of $\mathcal{Q}$

The **optimal tradeoff surface** of $\mathcal{Q}$ is the set of all $R(D)$ for D Pareto optimal. Any "best" decision rule must have a CRV on this optimal tradeoff surface.

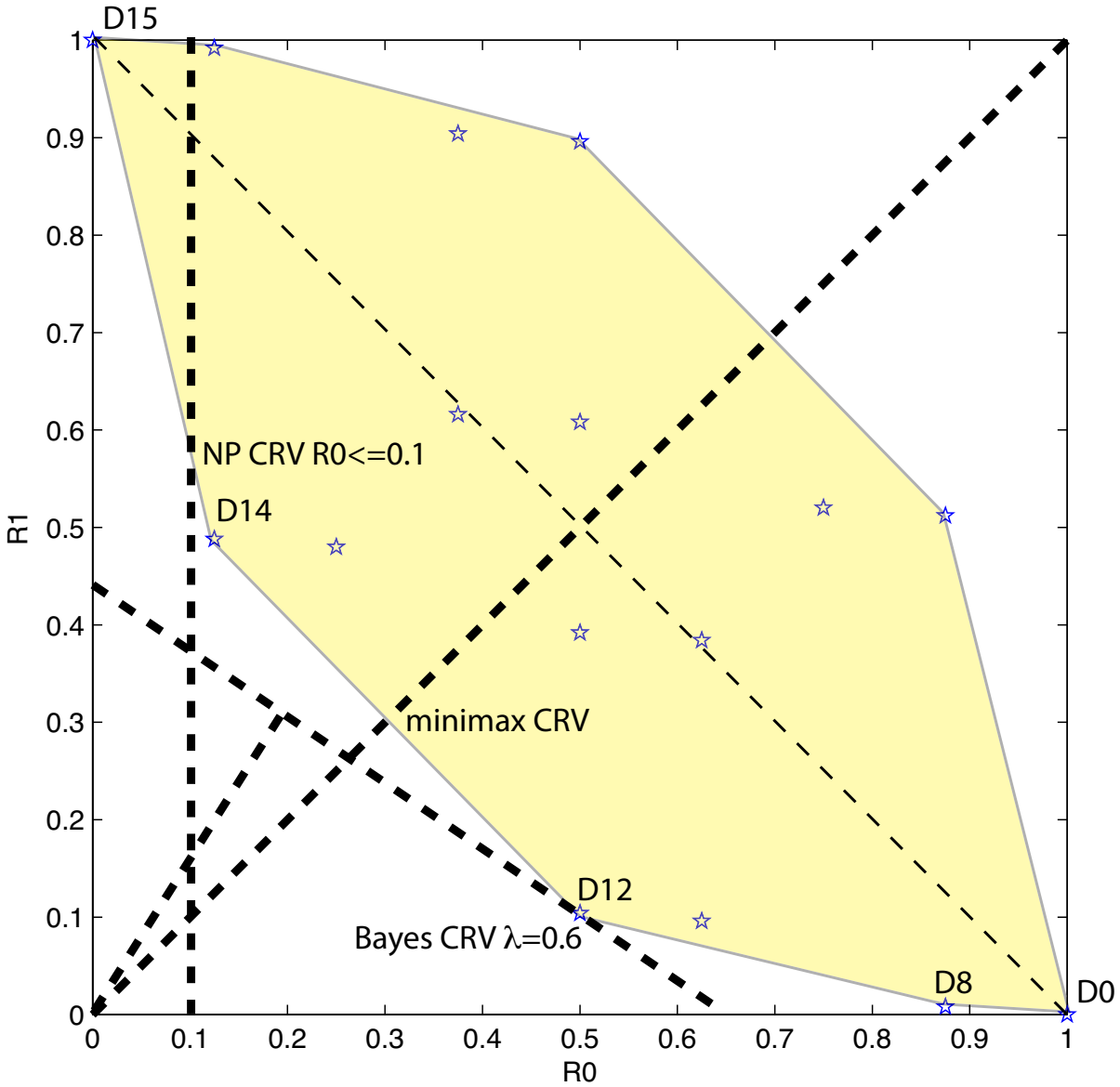


Specifying a Unique Decision Rule

Note that the optimal tradeoff surface does not specify a unique best decision rule. An additional criterion is needed.

1. **Neyman Pearson criterion**: Find D that minimizes $R_1(D)$ subject to an upper bound on $R_0(D)$.
2. **Bayes criterion**: Fix some $\lambda \in [0, 1]$ and define the weighted Bayes risk $r(D, \lambda) = (1 - \lambda)R_0(D) + \lambda(R_1(D))$. Find D that minimizes $r(D, \lambda)$.
3. **Minimax criterion**: Find D that minimizes $\max\{R_0(D), R_1(D)\}$.

Working Example: Risk Vectors [$q_0 = 0.5$ and $q_1 = 0.8$]



Summary of Main Results

We have introduced the notion of **conditional risks** as a way of quantifying the performance/consequences of a decision rule when the state is x_j:

$$R_j(D) \quad = \quad c_j^\top D p_j \text{ (finite observation spaces)}$$

We would like a decision rule that minimizes all conditional risks R_j for $j \in \{0, \ldots, N-1\}$ simultaneously. This is a **multi-objective optimization problem**.

Minimizing all conditional risks simultaneously is impossible, in general, since the conditional risks must be traded off against each other on the optimal tradeoff surface.