ECE531 Lecture 2a: A Mathematical Model for Hypothesis Testing

D. Richard Brown III

Worcester Polytechnic Institute

29-January-2009

Hypothesis Testing Basics

Examples:

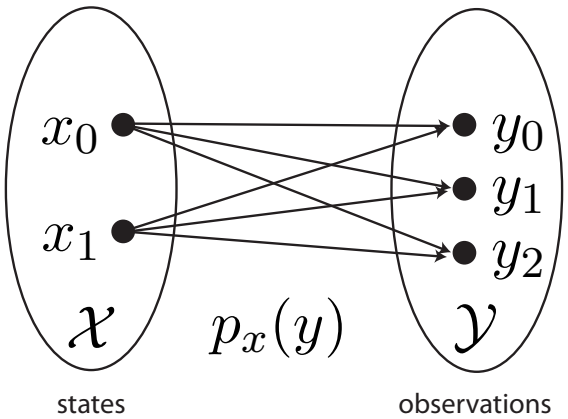
- The coin is fair or not fair.
- The approaching airplane is friendly or unfriendly.
- This email is spam or not spam.
- The medical treatment is effective or not effective.
- Which candidate will win the primary election?
- Communication receiver: Given a codebook with M codewords, which codeword was sent?

Given a "noisy" observation, we want to decide among two or more possible 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 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.

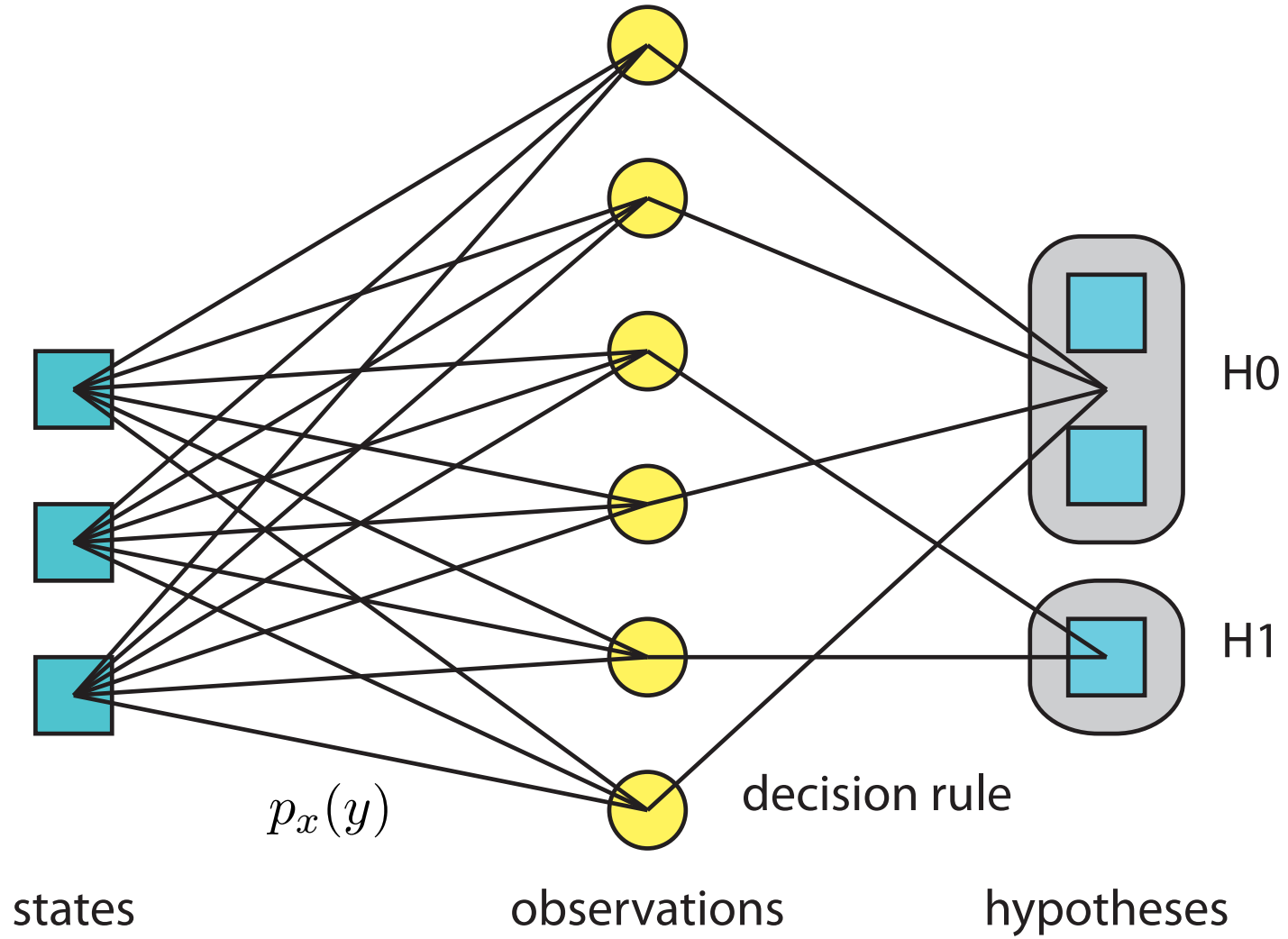
Our focus will be on simple hypothesis testing problems for now, but we will return to composite hypothesis testing soon.

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: Matrix Notation and Decisions

- 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:

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

- 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 $m \in \{0, \ldots, M-1\}$.
- It is convenient to write a decision matrix $D \in \mathbb{R}^{M \times L}$, e.g.

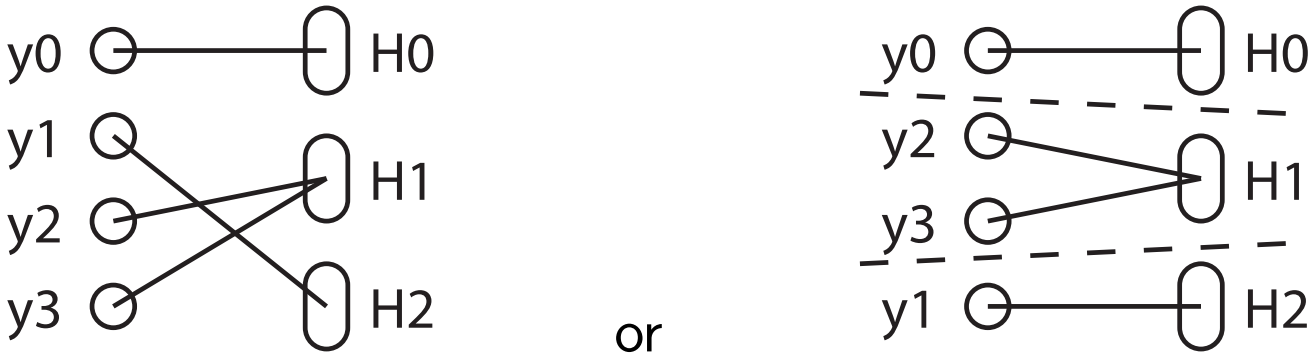
$$D = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

Decision Rules: Partitioning of the Observation Space

Suppose we had the decision matrix

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

we can think of this graphically as



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 $\bigcup_i \mathcal{Y}_i = \mathcal{Y}$.

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

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

- This is a simple binary hypothesis testing problem.

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 rules that we can consider:

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

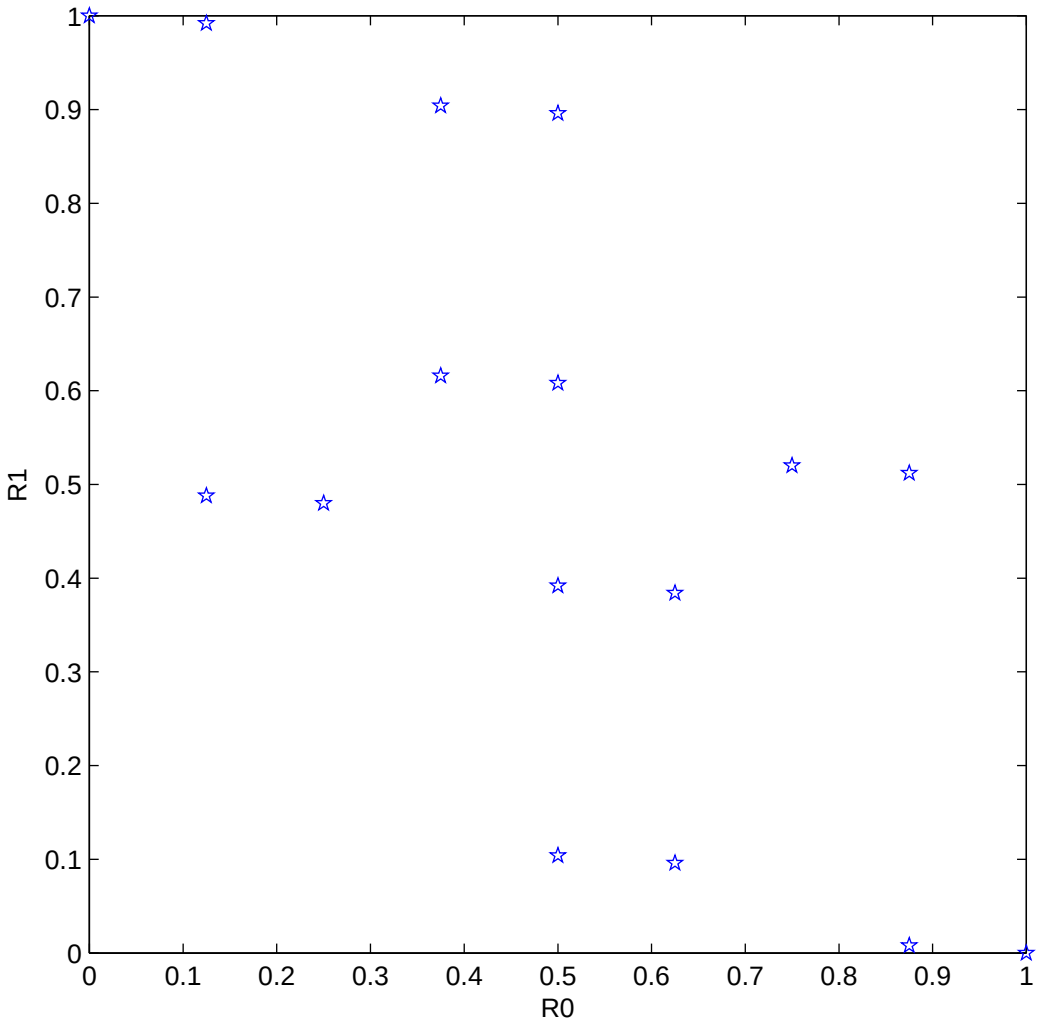
Working Example: Part 3

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

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

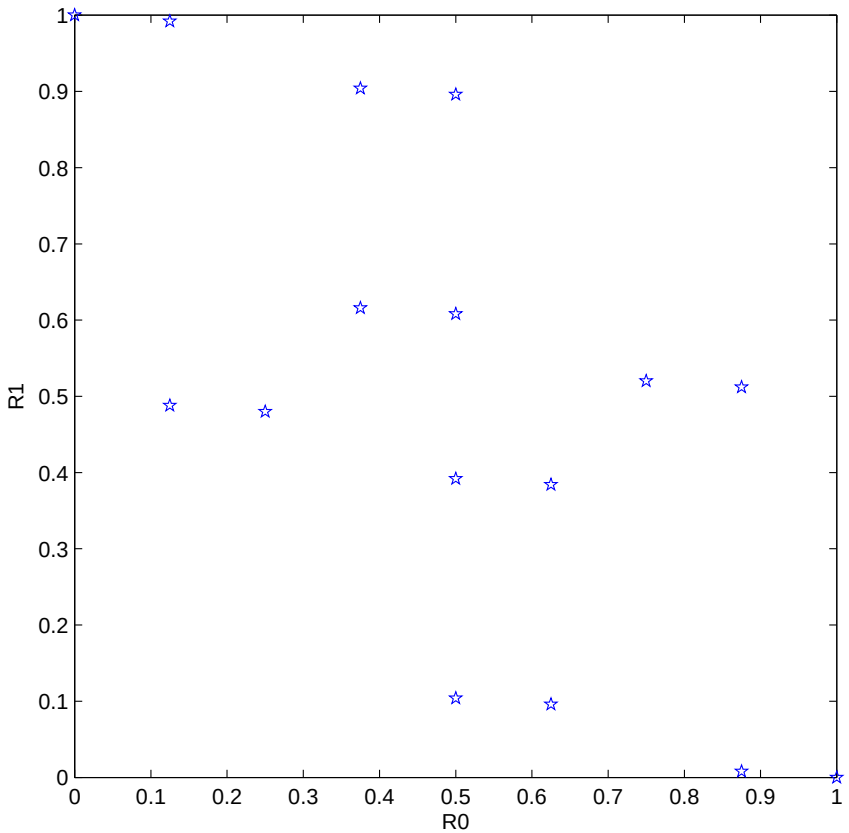
- $R(D) \in \mathbb{R}^N$ is called the **conditional risk vector** (CRV).
- Ideally, we would like both $R_0(D)$ and $R_1(D)$ to be small. It is usually not possible, however, to find a D that minimizes both simultaneously.
- To see this, we can plot the coordinates of these vectors in $\mathbb{R}^2$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

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



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:

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

- 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 $\mathcal{D}$.

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 $\alpha \in [0, 1]$,

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

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 $\alpha \in [0, 1]$

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

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 $\alpha_1, \alpha_2 \in \mathbb{R}$ and decision rules $D_1, D_2 \in \mathbb{R}^{M\times L}$

$$\begin{aligned} R_j(\alpha_1 D_1 + \alpha_2 D_2) &= c_j^\top(\alpha_1 D_1 + \alpha_2 D_2)p_j \\ &= \alpha_1 c_j^\top D_1 p_j + \alpha_2 c_j^\top D_2 p_j \\ &= \alpha_1 R_j(D_1) + \alpha_2 R_j(D_2) \end{aligned}$$

Thus $R(\alpha_1 D_1 + \alpha_2 D_2) = \alpha_1 R(D_1) + \alpha_2 R(D_2)$. □

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?

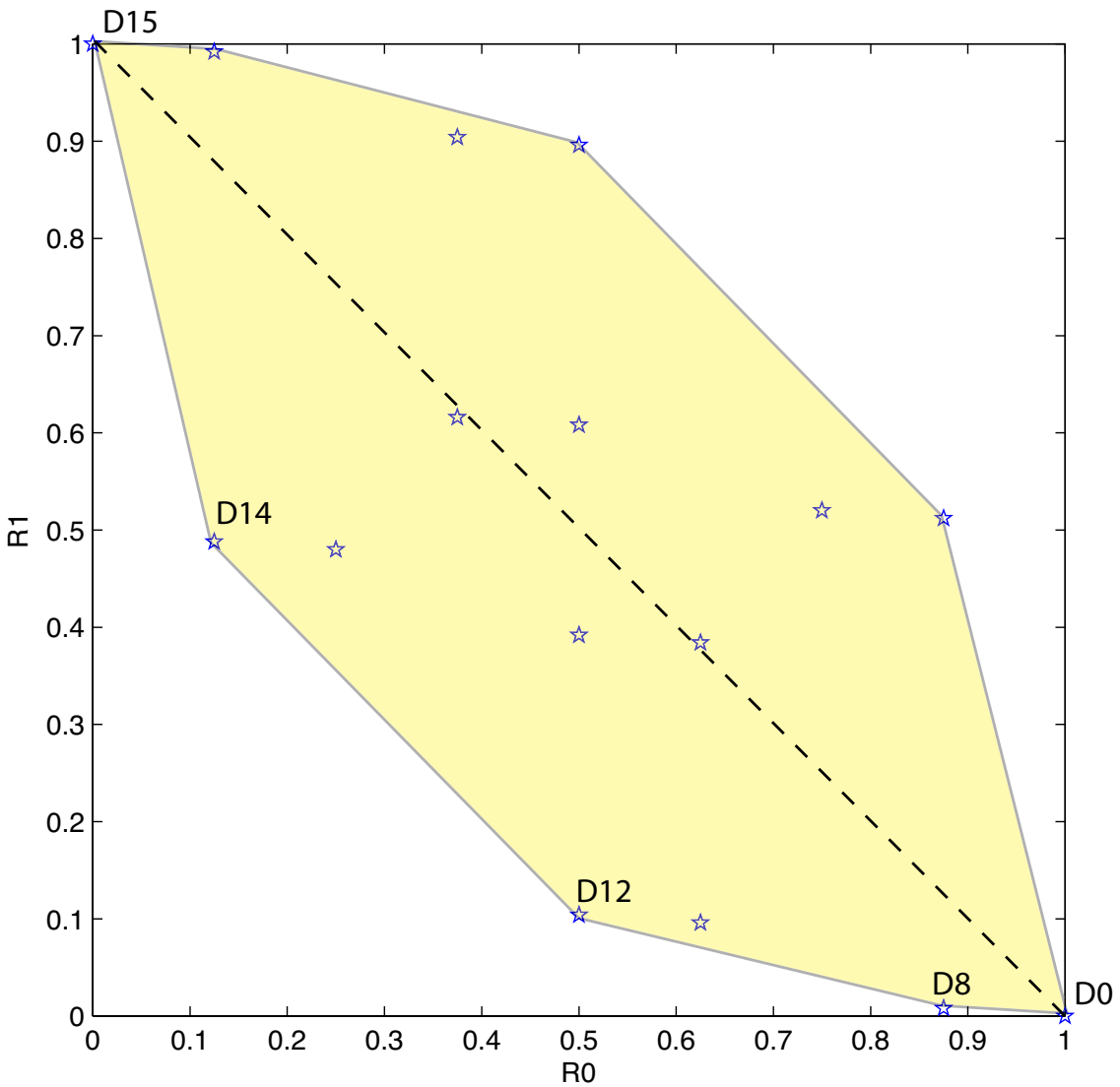
Theorem

$\mathcal{Q}$ 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 $D_{0,8,\alpha}$, $D_{8,12,\alpha}$, $D_{12,14,\alpha}$, and $D_{14,15,\alpha}$ for $\alpha \in [0, 1]$.

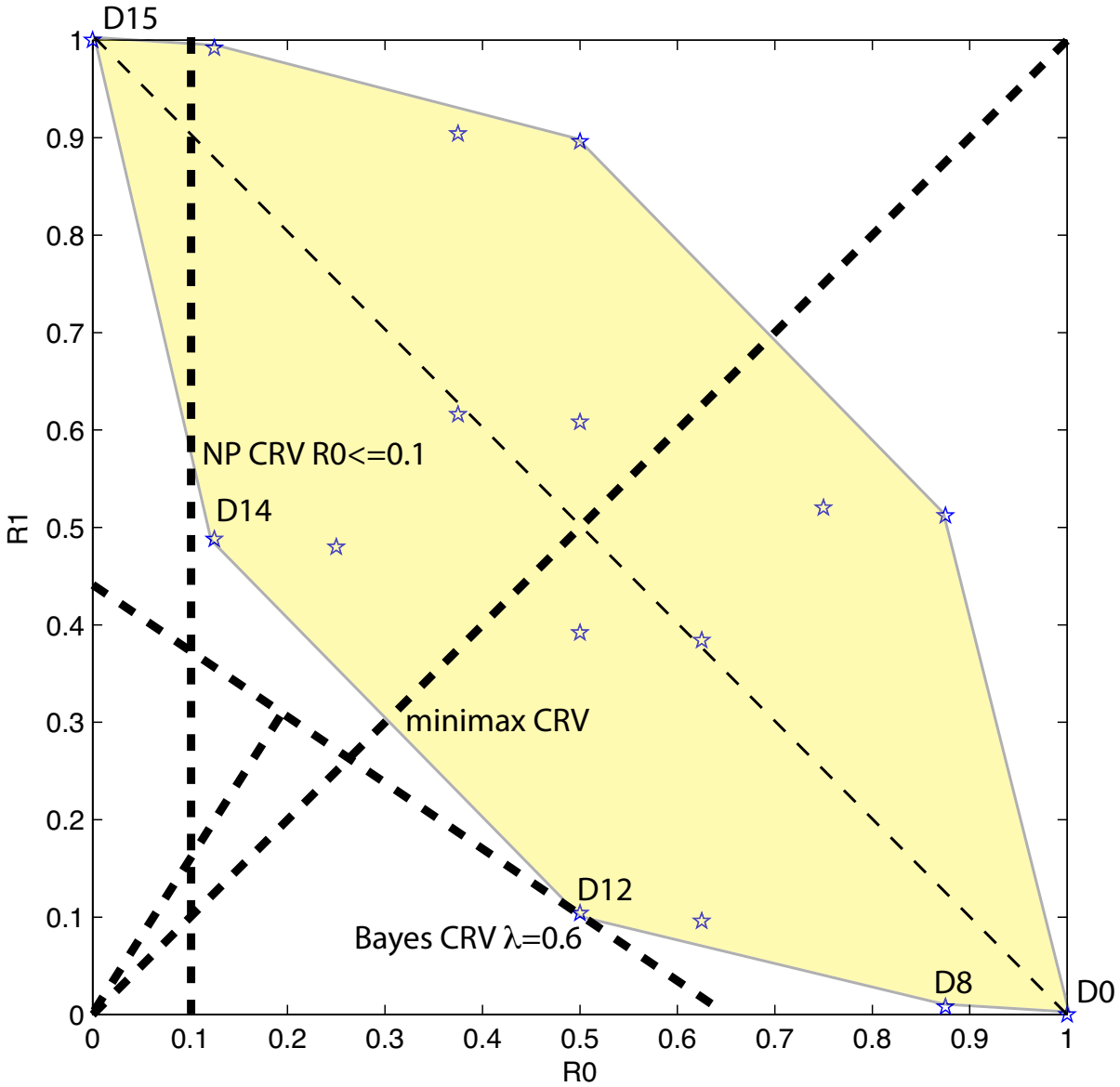
Optimal Tradeoff Surface of $\mathcal{Q}$

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 conditional risk vector on this optimal tradeoff surface.

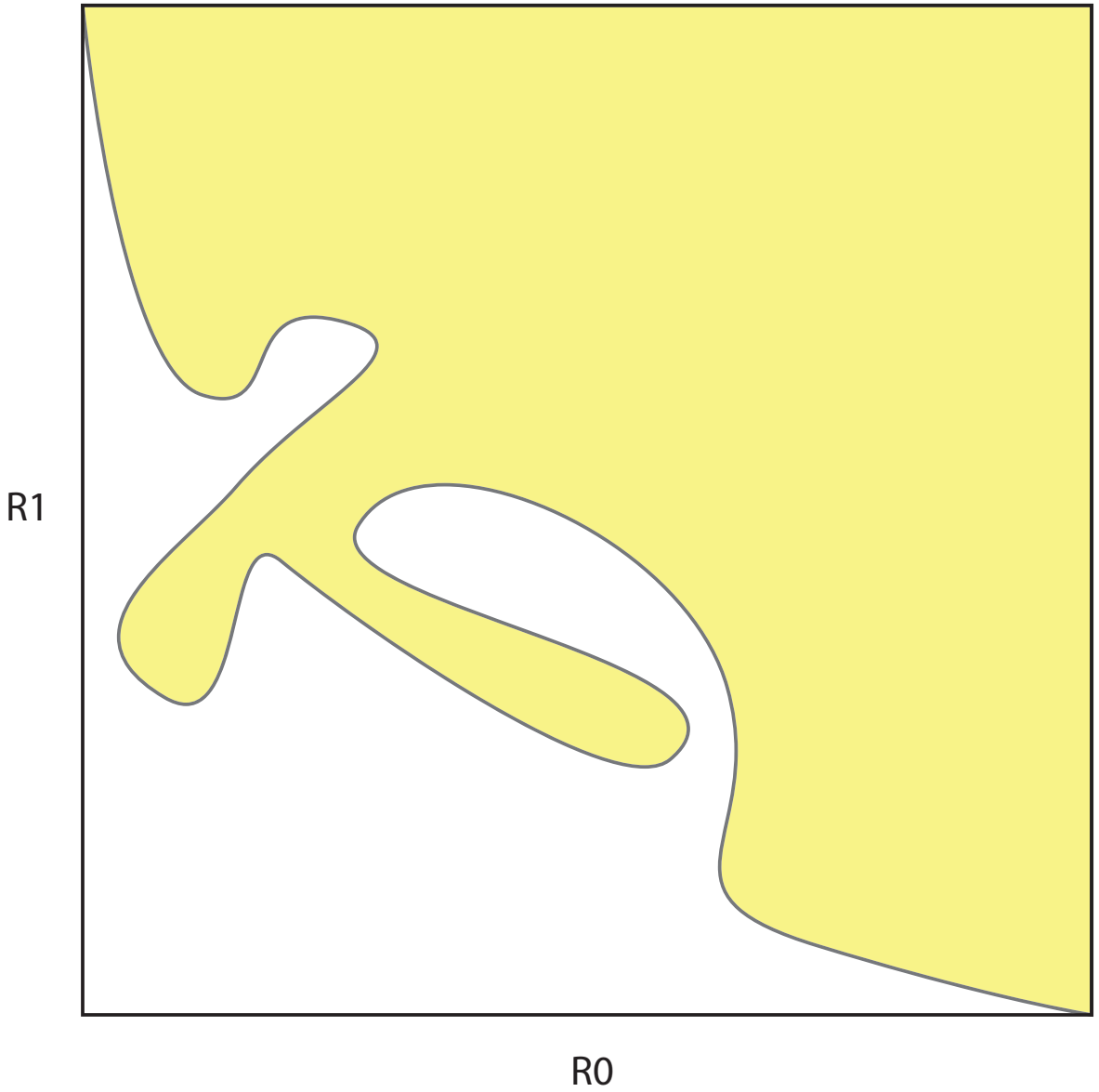
Note that the optimal tradeoff surface does not specify a unique best decision rule. An additional criterion is needed.

1. **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 $r(D, \lambda)$.
2. **Minimax criterion**: Find D that minimizes $\max\{R_0(D), R_1(D)\}$.
3. **Neyman Pearson criterion**: Find D that minimizes $R_1(D)$ subject to an upper bound on $R_0(D)$.

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



Why is Convexity Important?



Infinite Observation Sets: Part 1

In many problems of interest, the observation space is not finite. We can generalize our insight from the finite observation space as follows:

1. We can no longer use a decision matrix. Our randomized decision rule is denoted as $\rho : \mathcal{Y} \mapsto \mathcal{P}_M$ where $\mathcal{P}_M$ is the set of pmfs on $\mathcal{Z}$. We still use $\mathcal{D}$ to denote the set of decision rules $\rho \in \mathcal{D}$.
2. We denote $\rho_i(y)$ as the probability of deciding $\mathcal{H}_i$ when the observation is y.
3. The cost of deciding $\mathcal{H}_i$ when the state is x_j is still denoted as C_{ij}. Hence, for state x_j and the observation y, the cost of using decision rule ρ is

$$C_j(\rho) = \sum_{i=0}^{M-1} \rho_i(y) C_{ij}$$

Infinite Observation Sets: Part 2

The conditional risk for state x_j is then

$$R_j(\rho) = \int_{y \in \mathcal{Y}} C_j(\rho) p_j(y)\, dy = \int_{y \in \mathcal{Y}} \left[\sum_{i=0}^{M-1} \rho_i(y) C_{ij} \right] p_j(y)\, dy$$

where $p_j(y)$ is the known conditional density that probabilistically describes the relationship between state x_j and the observations.

As before, we can group these individual conditional risks into a conditional risk vector $R(\rho) \in \mathbb{R}^N$.

Theorem

The function $R : \rho \mapsto R(\rho)$ is linear.

Proof.

Same idea as the case with finite $\mathcal{Y}$. □

Infinite Observation Sets: Part 3

If we let the decision rule ρ range over all of $\mathcal{D}$, $R(\rho)$ traces out the set $\mathcal{Q}$ of achievable conditional risk vectors in $\mathbb{R}^N$.

Theorem

$\mathcal{Q}$ is a closed and bounded convex subset of $\mathbb{R}^N$.

The proof of this is omitted here since it is a bit messy and requires some understanding of topology, pointwise convergence, and the dominated convergence theorem.

The concepts of **Pareto optimal decision rules** and the **optimal tradeoff surface** of $\mathcal{Q}$ also apply to the case of infinite $\mathcal{Y}$.

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:

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

$$R_j(\rho) = \int_{y \in \mathcal{Y}} \left[\sum_{i=0}^{M-1} \rho_i(y) C_{ij} \right] p_j(y)\, dy \text{ (infinite observation spaces)}$$

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.

Achievable CRVs and Optimal Tradeoff Surface