MTAT.07.003 Cryptology II

Theoretical Background

Sven Laur
University of Tartu

Probability Theory

What is a random variable?

A *discrete random variable* f is formally a function $f : \Omega \to \{0,1\}^*$ where Ω is a *sample space* that models non-deterministic behaviour. Now for each output y there is a corresponding *elementary event*

$$\Omega_y = \{\omega \in \Omega : f(\omega) = y\} .$$

A *probability measure* $\Pr : \mathcal{F}(\Omega) \to [0,1]$ describes relative likelihood of *observable events* $\mathcal{F}(\Omega) = \{\emptyset, \Omega_0, \Omega_1, \Omega_{00}, \Omega_{01}, \ldots, \Omega_0 \cup \Omega_1, \ldots, \Omega\}$:

$$\Pr\left[\omega \in \Omega : f(\omega) \in \mathcal{Y}\right] \doteq \sum_{y \in \mathcal{Y}} \Pr\left[\omega \in \Omega_y\right] \quad ,$$

where by convention the probability measure is normalised

$$\Pr\left[\omega \in \Omega\right] = \sum_{y \in \{0,1\}^*} \Pr\left[\omega \in \Omega_y\right] = 1 \quad .$$

Conditional probability

Often, the presence of one event is correlated with some other events. The corresponding influence is formally quantified by *conditional probability*

$$\Pr\left[f(\omega) = y | g(\omega) = x\right] \doteq \frac{\Pr\left[f(\omega) = y \wedge g(\omega) = x\right]}{\Pr\left[g(\omega) = x\right]}$$

Consequently, for any two events A and B:

$$\Pr\left[A \wedge B\right] = \Pr\left[A\right] \cdot \Pr\left[B|A\right] = \Pr\left[B\right] \cdot \Pr\left[A|B\right] \quad .$$

Two *events are independent* if $\Pr\left[A \wedge B\right] = \Pr\left[A\right] \cdot \Pr\left[B\right]$.

Total Probability Formula

Let $\mathcal{H}_1, \ldots, \mathcal{H}_n$ be mutually exclusive events such that

$$\Pr[\mathcal{H}_i \wedge \mathcal{H}_j] = 0 \qquad \text{and} \qquad \Pr[\mathcal{H}_1 \vee \ldots \vee \mathcal{H}_n] = 1 \enspace .$$

Then for any any event A we can express

$$\Pr[A] = \sum_{i=1}^{n} \Pr[\mathcal{H}_i] \cdot \Pr[A|\mathcal{H}_i] \enspace .$$

PDF and CDF. Theory

Discrete random variables do not have a classical *probability density function*. Instead, we can consider probabilities of the smallest observable events $\Omega_0, \Omega_1, \Omega_{00}, \Omega_{01}, \ldots$. Consider the corresponding pseudo-density function

$$p_x \doteq \Pr\left[\omega \in \Omega : f(\omega) = x\right] \quad .$$

Then we can express a *cumulative distribution function*

$$F(y) = \Pr\left[\omega \in \Omega : f(\omega) \leq y\right]$$

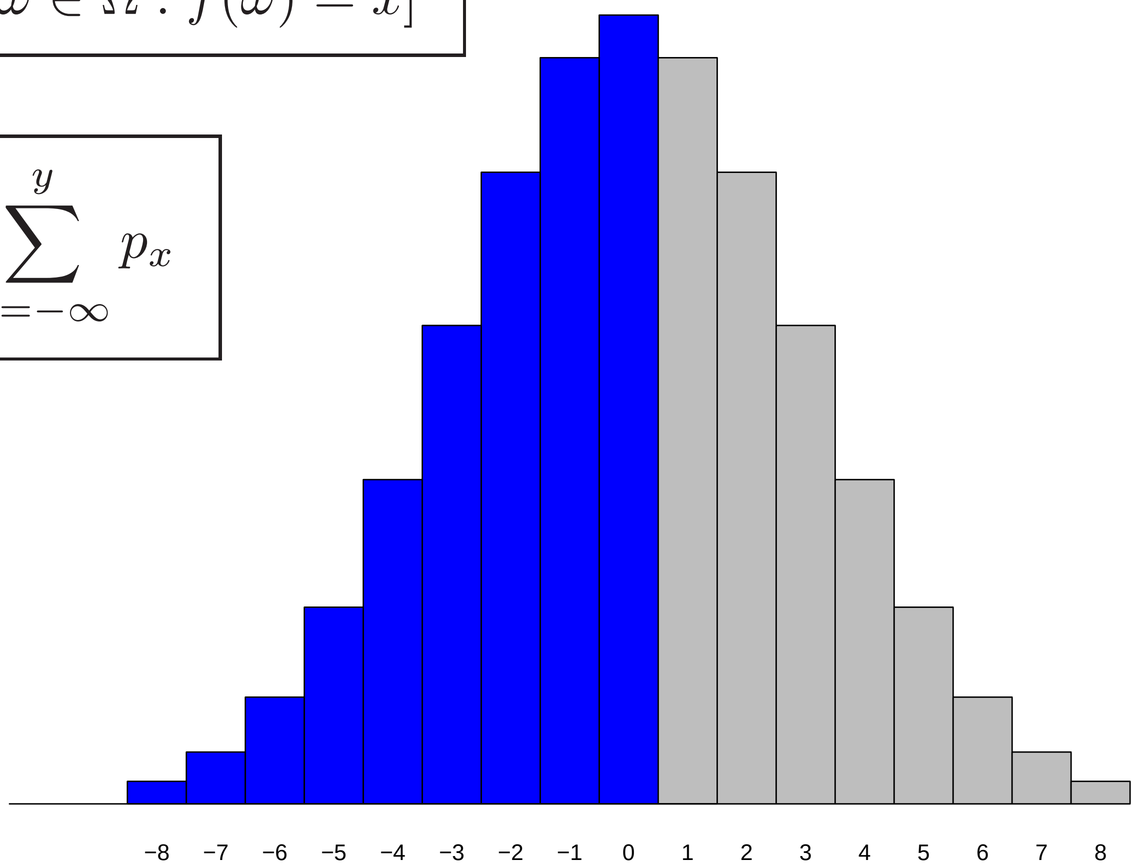
in terms of pseudo-density function

$$F(y) = \sum_{x=-\infty}^{y} \Pr\left[\omega \in \Omega : f(\omega) = x\right] = \sum_{x=-\infty}^{y} p_x \quad .$$

PDF and CDF. Illustration

$$p_x = \Pr\left[\omega \in \Omega : f(\omega) = x\right]$$

$$F(y) = \sum_{x=-\infty}^{y} p_x$$



Expected value

The *expected value* of a random variable f is defined as

$$\mathbf{E}[f] = \sum_{x \in \{0,1\}^*} x \cdot \Pr[\omega \in \Omega : f(\omega) = x] = \sum_{x \in \{0,1\}^*} p_x \cdot x \enspace .$$

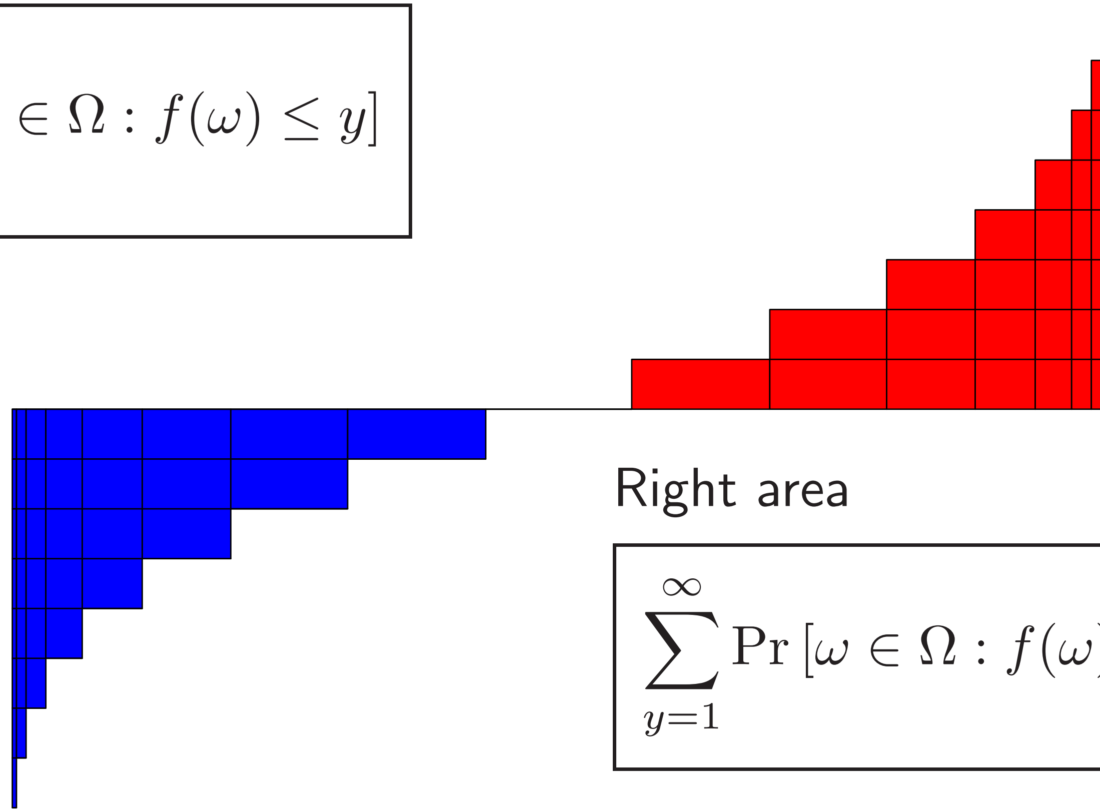
Alternatively, we can compute expected value as

$$\begin{aligned}\mathbf{E}[f] &= \sum_{y=1}^{\infty} \Pr[\omega \in \Omega : f(\omega) \geq y] - \sum_{y=-\infty}^{-1} \Pr[\omega \in \Omega : f(\omega) \leq y] \\ &= \sum_{y=0}^{\infty} (1 - F(y)) - \sum_{y=-\infty}^{-1} F(y) \enspace .\end{aligned}$$

Corresponding proof

Left area

$$\sum_{y=-\infty}^{-1} \Pr\left[\omega \in \Omega : f(\omega) \leq y\right]$$



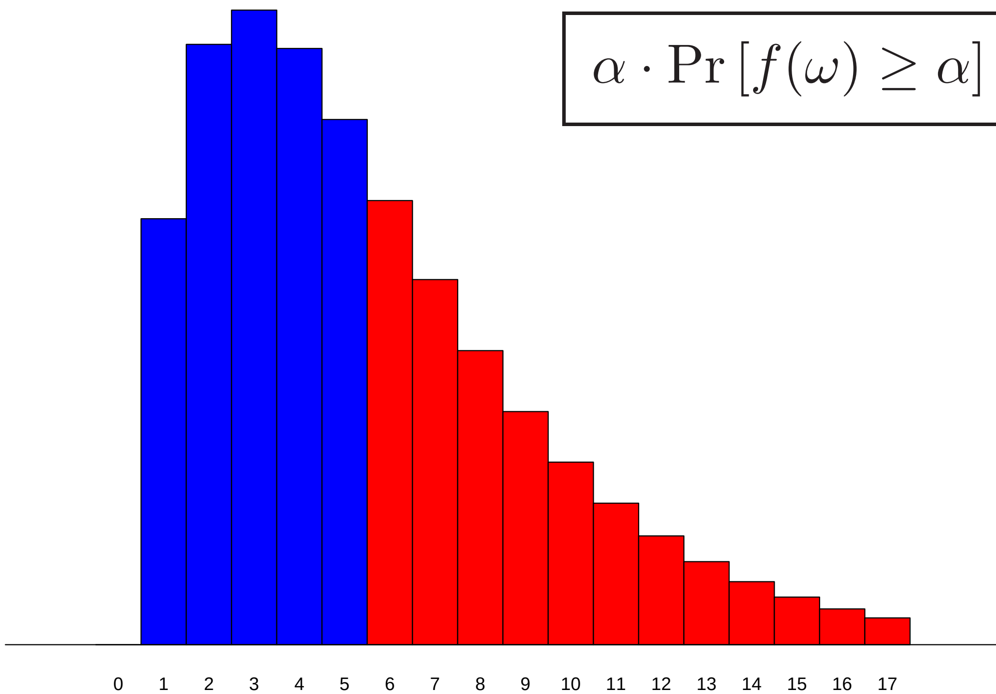
Right area

$$\sum_{y=1}^{\infty} \Pr\left[\omega \in \Omega : f(\omega) \geq y\right]$$

Markov's inequality

For every non-negative random variable $\Pr[f(\omega) \geq \alpha] \leq \frac{\mathbf{E}[f]}{\alpha}$.

$$\alpha \cdot \Pr[f(\omega) \geq \alpha] \leq \mathbf{E}[f]$$



Jensen's inequality

Let x be a random variable. Then for every convex-cup function f

$$\mathbf{E}\left[f(x)\right] \leq f(\mathbf{E}\left[x\right])$$

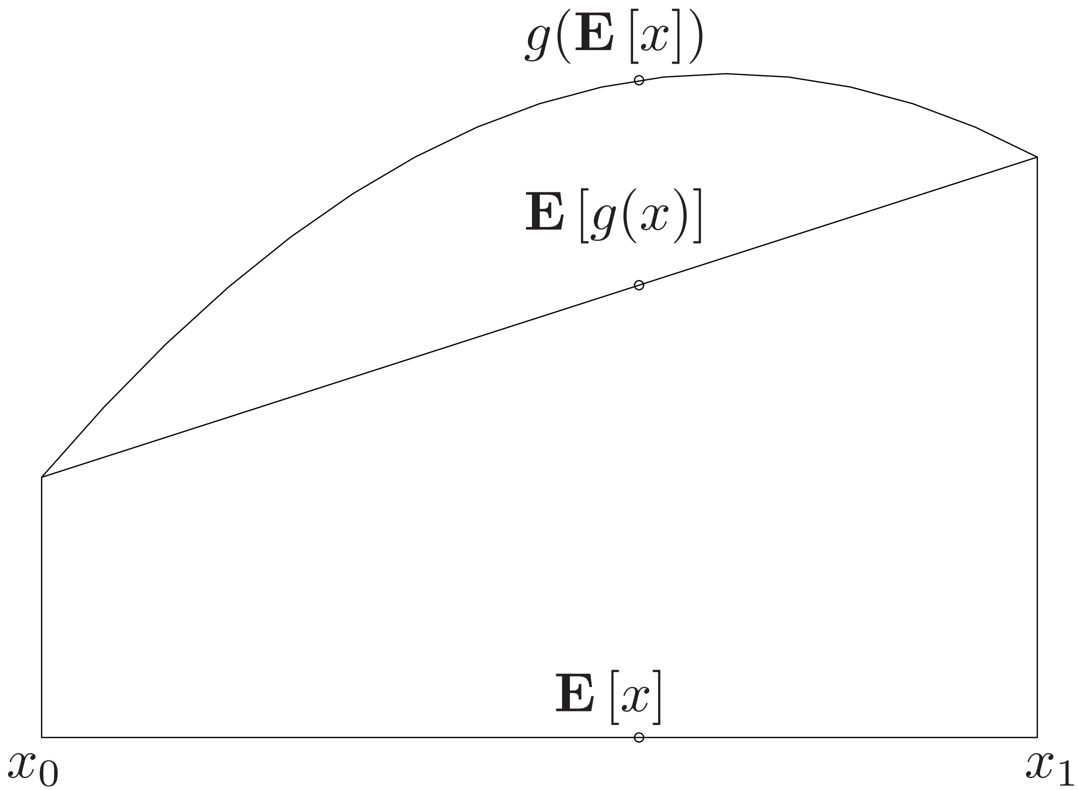
and for every convex-cap function g

$$\mathbf{E}\left[g(x)\right] \geq g(\mathbf{E}\left[x\right]) \enspace .$$

These inequalities are often used to get lower and upper bounds.

Corresponding proof

Note that it is sufficient to give a proof for sums with two terms.



Variance

Variance characterises how scattered are possible values

$$\mathbf{D}[f] = \mathbf{E}\left[(f - \mathbf{E}[f])^2\right] = \mathbf{E}\left[f^2\right] - \mathbf{E}[f]^2 \quad .$$

Usually, one also needs standard deviation

$$\boldsymbol{\sigma}[f] = \sqrt{\mathbf{D}[f]} \quad .$$

Chebyshev's inequality assures that

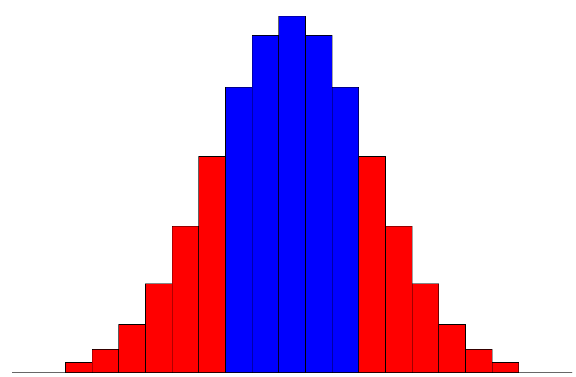
$$\Pr\left[|f(\omega) - \mathbf{E}[f]| \geq \alpha \cdot \boldsymbol{\sigma}[f]\right] \leq \frac{\mathbf{D}[f]}{\alpha^2}$$

Proof of Chebyshev's inequality

Let $g = (f - \mathbf{E}[f])^2$. Then by definition $\mathbf{D}[f] = \mathbf{E}[g]$ and we can apply Markov's inequality

$$\Pr\left[(f - \mathbf{E}[f])^2 > \alpha^2 \cdot \mathbf{E}[g]\right] \leq \frac{\mathbf{E}[g]}{\alpha^2}$$

$$\Pr\left[|f - \mathbf{E}[f]| > \alpha \cdot \boldsymbol{\sigma}[f]\right] \leq \frac{\mathbf{D}[f]}{\alpha^2}$$



Entropy

Shannon entropy

Entropy is another measure of uncertainty for random variables. Intuitively, it captures the minimal amount of bits that are needed on average to describe a value of a random variable X.

Shannon entropy is defined as follows

$$H(X) = -\sum_{x \in \{0,1\}^*} p_x \cdot \log_2 p_x = -\mathbf{E}\left[\log_2 \Pr\left[X = x\right]\right]$$

It is straightforward but tedious to prove

$$0 \leq H(X) \leq \log_2 |\mathrm{supp}(X)|$$

where the *support* of X is defined as $\mathrm{supp}(X) = \big\{x \in \{0,1\}^* : p_x > 0\big\}$.

Conditional of entropy

Conditional entropy is defined as follows

$$H(Y|X) = -\mathbf{E}_{X,Y}\left[\log_2 \Pr\left[Y|X\right]\right]$$

Now observe that

$$\begin{aligned}
H(X,Y) &= -\mathbf{E}_{X,Y}\left[\log_2 \Pr\left[X \wedge Y\right]\right] \\
&= -\mathbf{E}_{X,Y}\left[\log_2 \Pr\left[X\right] + \log_2 \Pr\left[Y|X\right]\right] \\
&= -\mathbf{E}_{X}\left[\log_2 \Pr\left[X\right]\right] - \mathbf{E}_{X,Y}\left[\log_2 \Pr\left[Y|X\right]\right] \\
&= H(X) + H(Y|X) \enspace .
\end{aligned}$$

Mutual information

Recall that entropy characterises the average length of minimal description. Now if we consider two random variables. Then we can describe them jointly or separately. *Mutual information* captures the corresponding gain

$$I(Y : X) = H(X) + H(Y) - H(X, Y)$$

Evidently, mutual information between independent variables is zero:

$$I(Y : X) = H(X) + H(Y) - H(X) - \underbrace{H(Y|X)}_{H(Y)} = 0 \ .$$

Similarly, if X and Y coincide then

$$I(Y : X) = H(X) + H(Y) - H(X) - \underbrace{H(Y|X)}_{0} = H(X) \ .$$

Min-entropy. Rényi entropy

Shannon entropy is not always descriptive enough for measuring uncertainty. For example, consider security of passwords.

▷ Obviously, we can just try the most probable password. The corresponding uncertainty measure is known as *min-entropy*

$$H_\infty(X) = -\log_2 \max_{x \in \{0,1\}^*} \Pr[X = x]$$

▷ Often, we do not want that two persons have coinciding passwords. The corresponding uncertainty measure is known as *Rényi entropy*

$$H_2(X) = -\log_2 \Pr[x_1 \leftarrow X, x_2 \leftarrow X : x_1 = x_2]$$

where x_1 and x_2 are independent draws from the distribution X.

Hypothesis Testing

Standard setting

The best way to model secrecy is hypothesis testing.

$s \leftarrow \mathcal{S}_0$

$x \leftarrow f(s)$ →

Given x and
- description of $\mathcal{S}_0$
- description of $f(\cdot)$

Charlie has to accept or reject
- a hypothesis $\mathcal{H}$

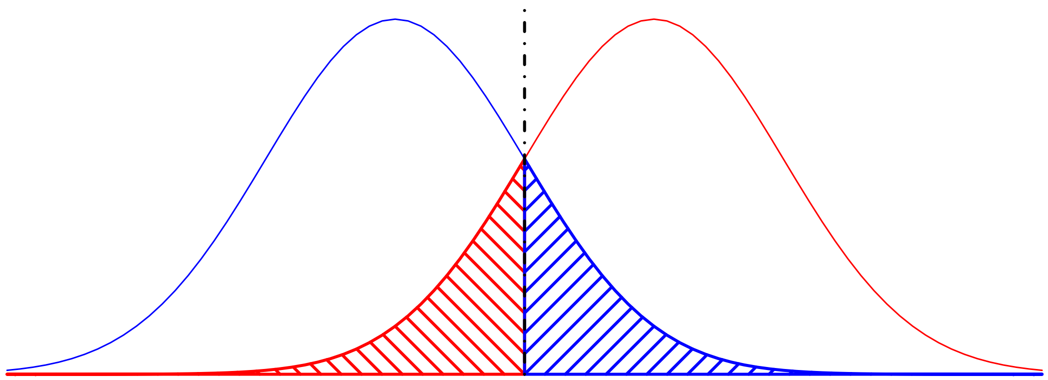
There are several types of hypotheses:

▷ *simple hypotheses* $\mathcal{H} = [s \stackrel{?}{=} s_0]$

▷ *complex hypotheses* $\mathcal{H} = [s \stackrel{?}{=} s_0 \vee s \stackrel{?}{=} s_1 \vee \ldots \vee s \stackrel{?}{=} s_k]$

▷ *trivial hypotheses* that always hold or never hold.

Simple hypothesis testing



Simple hypothesis $\mathcal{H}_0$ and $\mathcal{H}_1$ always determine the distribution of the observable variable $x \leftarrow f(s)$. Consequently, an adversary $\mathcal{A}$ that can choose between two hypothesis $\mathcal{H}_0$ and $\mathcal{H}_1$ can do two types of errors:

▷ probability of *false negatives* $\alpha(\mathcal{A}) \doteq \Pr\left[\mathcal{A}(x) = 1 | \mathcal{H}_0\right]$

▷ probability of *false positives* $\beta(\mathcal{A}) \doteq \Pr\left[\mathcal{A}(x) = 0 | \mathcal{H}_1\right]$

The corresponding aggregate error is $\gamma(\mathcal{A}) = \alpha(\mathcal{A}) + \beta(\mathcal{A})$.

Various trade-offs

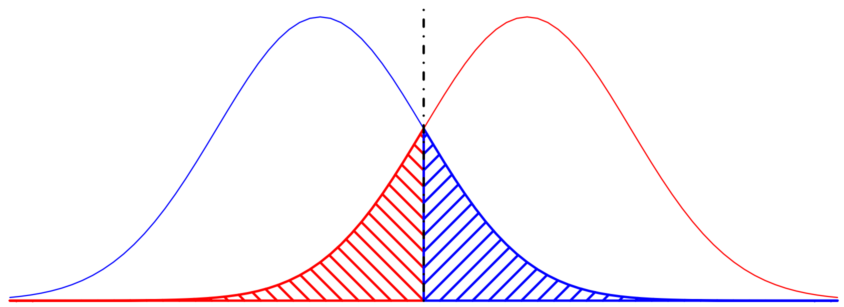
A reoccurring task in statistics is to minimise the probability of false positives $\beta(\mathcal{A})$ so that the probability of false negatives $\alpha(\mathcal{A})$ is bounded.

The most obvious strategy is to choose a trade-off point η and define

$$\mathcal{A}(x) = \begin{cases} 1, \text{if } \Pr\left[x|\mathcal{H}_0\right] < \eta \cdot \Pr\left[x|\mathcal{H}_1\right] \\ 0, \text{if } \Pr\left[x|\mathcal{H}_0\right] > \eta \cdot \Pr\left[x|\mathcal{H}_1\right] \\ \text{throw a } \rho\text{-biased coin, otherwise} \end{cases}$$

Neyman-Pearson Theorem. The likelihood ratio test described above achieves optimal $\beta(\mathcal{A})$ for any bound $\alpha(\mathcal{A}) \leq \alpha_0$. The aggregate error $\gamma(\mathcal{A})$ is minimised by choosing $\eta = 1$ and using a fair coin to break ties.

Statistical distance



Formally, statistical distance is defined as re-scaled ℓ_1-distance

$$\mathsf{sd}_x(\mathcal{H}_1, \mathcal{H}_1) = \frac{1}{2} \cdot \sum_x \left|\Pr\left[x|\mathcal{H}_0\right] - \Pr\left[x|\mathcal{H}_1\right]\right|$$

but it is straightforward to see

$$\mathsf{sd}_x(\mathcal{H}_0, \mathcal{H}_1) = \max_{\mathcal{A}} \Pr\left[\mathcal{A}(x) = 0|\mathcal{H}_0\right] - \Pr\left[\mathcal{A}(x) = 0|\mathcal{H}_1\right] \quad ,$$

$$\mathsf{sd}_x(\mathcal{H}_0, \mathcal{H}_1) = 1 - \min_{\mathcal{A}} \gamma(\mathcal{A}) \quad .$$

Computational distance

Although the best likelihood ratio test minimises the aggregate error $\gamma(\mathcal{A})$, it is often infeasible to use it:

- ▷ the description of the corresponding decision border is too complex,
- ▷ it is infeasible to compute $\Pr[x|\mathcal{H}_0]$ and $\Pr[x|\mathcal{H}_1]$.

Therefore, we must consider properties of optimal t-time test algorithms instead. The corresponding distance measure

$$\mathsf{cd}_x^t(\mathcal{H}_0, \mathcal{H}_1) = \max_{\mathcal{A} \text{ is } t\text{-time}} |\Pr[\mathcal{A}(x) = 0|\mathcal{H}_0] - \Pr[\mathcal{A}(x) = 0|\mathcal{H}_1]|$$

is known as *computational distance*. Evidently

$$\lim_{t\to\infty} \mathsf{cd}_x^t(\mathcal{H}_0, \mathcal{H}_1) = \mathsf{sd}_x(\mathcal{H}_0, \mathcal{H}_1) \ .$$