Detection Cost Function / empirical Bayes risk:

- **Define the costs of different kind of errors** (C_{fn}, C_{fp})
- **Define the class prior probability** (π_T, $\pi_F = 1 - \pi_T$)
- **Evaluate by computing empirical Bayes risk**

$$DCF_u(C_{fn}, C_{fp}, \pi_T) = \pi_T C_{fn} P_{fn} + (1 - \pi_T) C_{fp} P_{fp}$$

- P_{fn} **and** P_{fp} **are the false negative and false positive rates, and depend on the selected threshold** t.

The triplet (π_T, C_{fn}, C_{fp}) represents the working point of an application for a binary classification task.

Normalized DCF: we compare the system DCF w.r.t. the best dummy system

$$DCF(\pi_T, C_{fn}, C_{fp}) = \frac{DCF_u(\pi_T, C_{fn}, C_{fp})}{\min(\pi_T C_{fn}, (1-\pi_T) C_{fp})}$$

Note that the best dummy system corresponds to optimal Bayes decisions based on prior information alone, i.e., using the prior probability as if it was the recognizer posterior probability for any given sample.

In terms of normalized DCF, the applications (π_T, C_{fp}, C_{fn}) **and** $(\tilde{\pi}, 1, 1)$ **are again equivalent.**

We can interpret $\tilde{\pi}$ as an **effective** prior: if the class prior for H_T was $\tilde{\pi}$ and we assumed uniform costs, we would obtain the same normalized costs as for our original application.

For systems producing well-calibrated log–likelihood ratios

$$s = \log \frac{f_{X|C}(x|H_T)}{f_{X|C}(x|H_F)}$$

the optimal threshold (optimal Bayes decision) becomes, in terms of effective prior:

$$t = -\log \frac{\tilde{\pi}}{1 - \tilde{\pi}}$$

Well-Calibrated System: The LLRs accurately reflect the true probabilities, leading to reliable and optimal decision-making. The threshold calculation based on the prior probabilities results in correct classification decisions.

Not Well-Calibrated System: The LLRs do not accurately reflect the true probabilities, resulting in unreliable and often incorrect decisions. The threshold might be calculated correctly, but the LLRs themselves are misleading.

In summary, a well-calibrated system ensures that the log-likelihood ratios it produces are accurate reflections of the true underlying probabilities, leading to optimal decision thresholds and correct classifications. A not well-calibrated

system, on the other hand, provides misleading LLRs, resulting in suboptimal decision-making.

Miscalibration Means a Shift or Scaling Issue, Not a Prediction Issue

• If a model separates classes well but the scores are shifted or compressed, you can still find a threshold that minimizes errors effectively.

• The model may still rank samples correctly (e.g., higher scores for positive samples, lower for negatives), but the numerical values of the scores are not meaningful as probabilities.

In general, systems often do not produce well–calibrated LLRs

- Non-probabilistic scores (e.g. SVM)
- Mis-match between train and test populations
- Non-accurate model assumptions

In these cases, we say that scores are **mis-calibrated**

The theoretical threshold $-\log \frac{\tilde{\pi}}{1-\tilde{\pi}}$ is not optimal anymore.

We can define the **minimum** cost DCF_{min} corresponding to the use of the optimal threshold for a given evaluation set.

We consider varying the threshold t to obtain all possible combinations of P_{fn} and P_{fp} for the evaluation set.

We select the threshold corresponding to the lowest DCF.

We can also compute the **actual** DCF obtained using the threshold corresponding to the effective prior $\tilde{\pi}$.

The difference between the actual and minimum DCF represents the loss due to score mis-calibration.