Lecture 10

2025-11-05

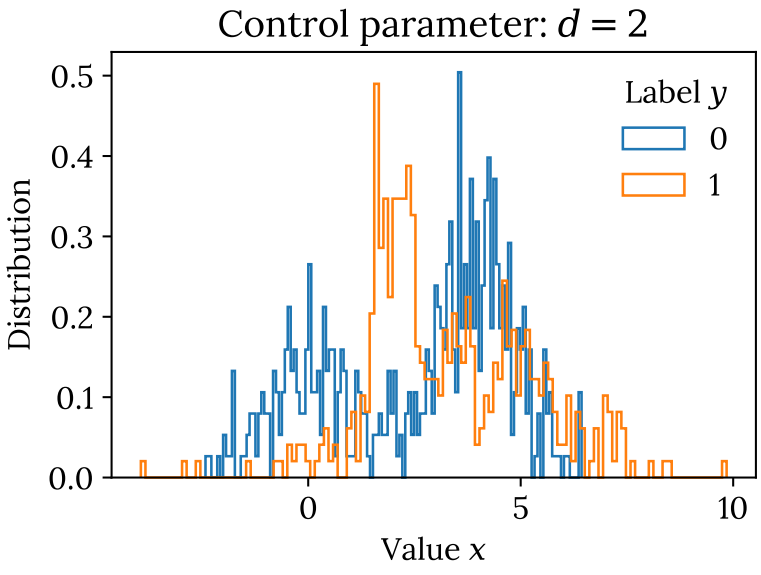
More information; Likelihood ratio test; Mutual information; Lagrange multiplier; Implications of constraints

- Online slides: lec10-mutual_information.html
- Code: code10.ipynb
- Homework: hw10.pdf, hw10-data.npz

Username: cns, Password: nycu2025

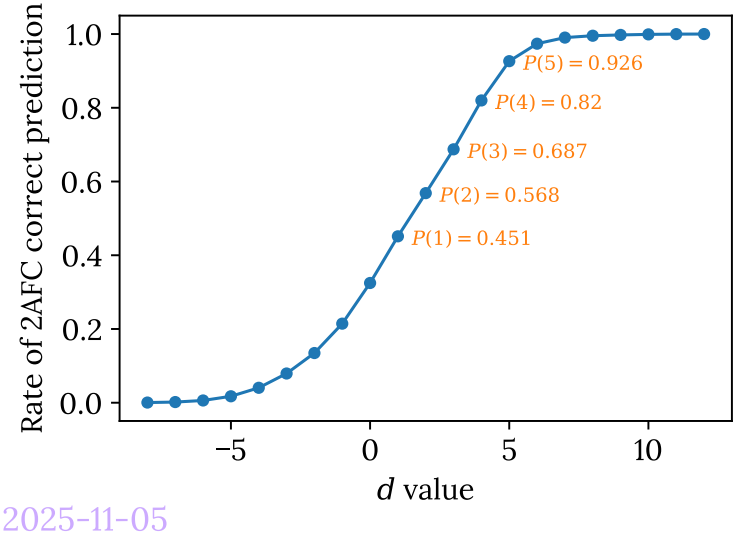
Laboratory of Computing Biological Networks

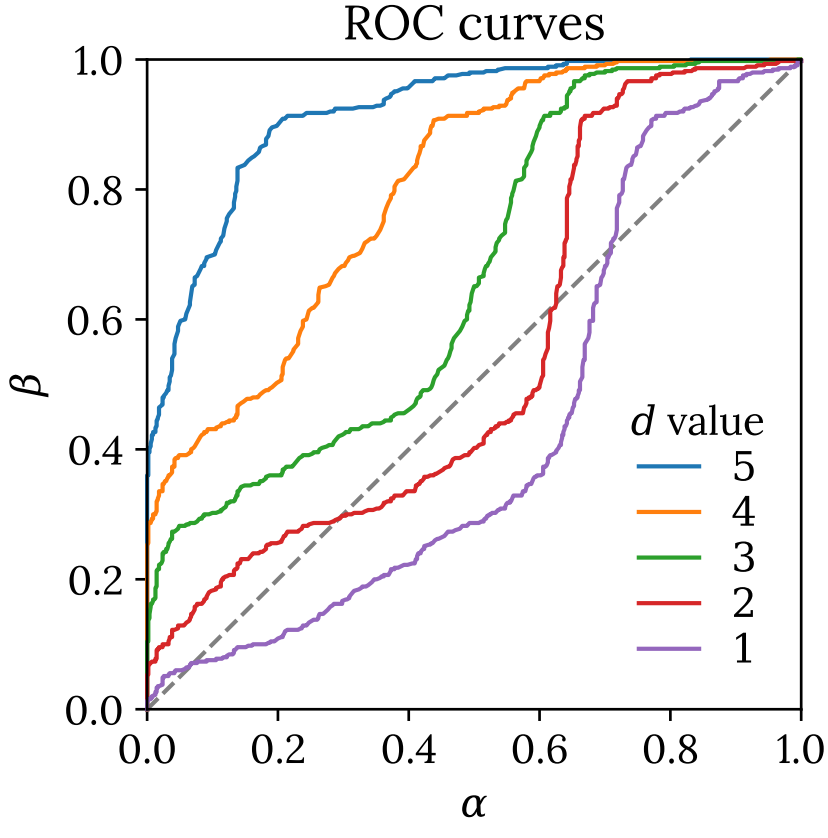
Some odd distribution



When the data distribution is not unimodal, applying a simple threshold may not gives the best estimates of the labels.

When the order of output values of two classes is reversed, the fraction of correct predicts can go below 50%.





Likelihood ratio test

— a more sophisticated strategy

Ratio between the outcome likelihoods of the two possibilities

$$l(r) = \frac{p[r|+]}{p[r|-]} \quad \text{— the likelihood ratio}$$

The slopes of $\alpha(z)$ and $\beta(z)$:

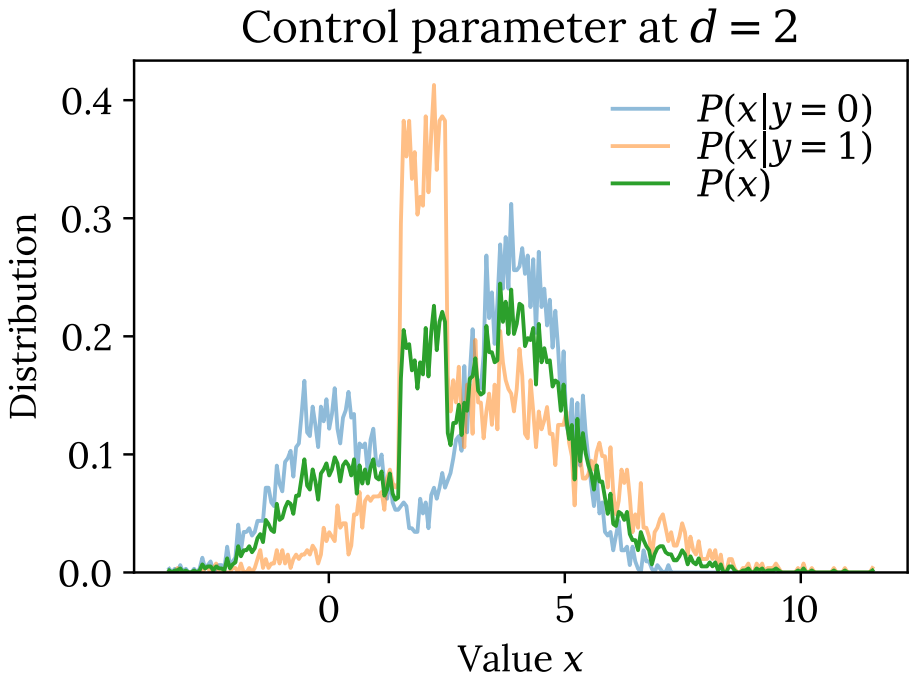
$$\beta'(z) = \frac{d}{dz}\int_z^\infty dr p[r|+] = -p[z|+].$$

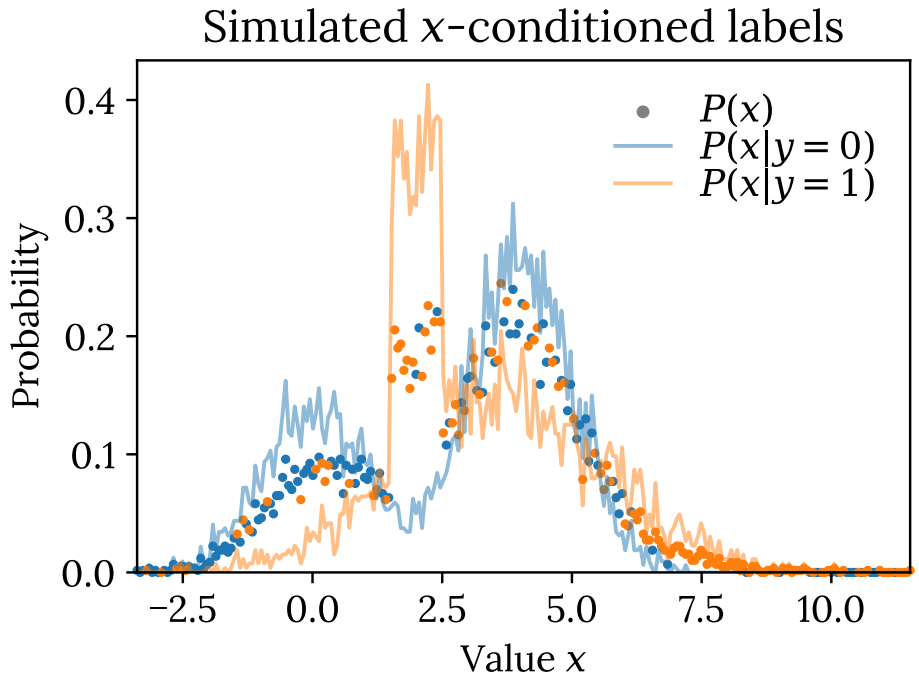
and, similarly, $\alpha'(z) = -p[z|-]$. The likelihood ratio

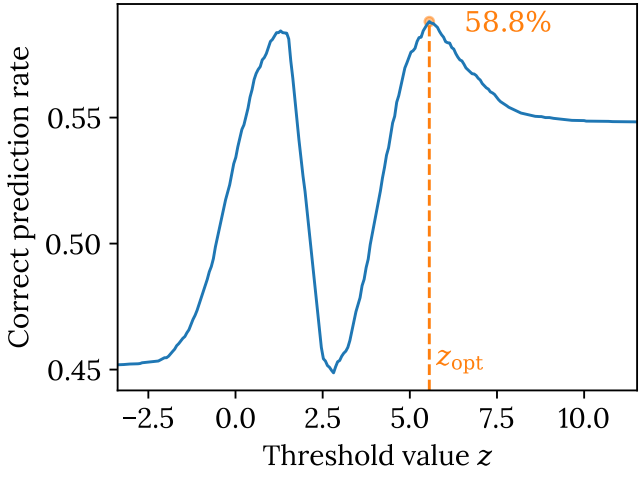
$$l(z) = \frac{p[z|+]}{p[z|-]} = \frac{\beta'(z)}{\alpha'(z)} = \frac{d\beta}{dz}\Big/\frac{d\alpha}{dz} = \frac{d\beta}{d\alpha}$$

is simply the slope $\beta'(\alpha)$ of the ROC curve.

Example: Prediction at $d = 2$

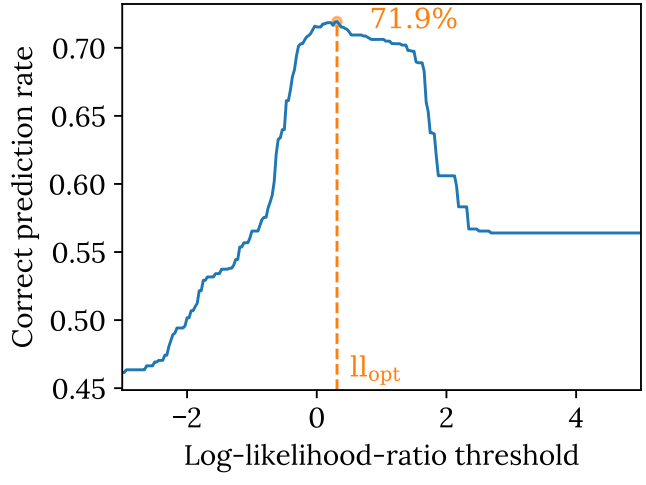




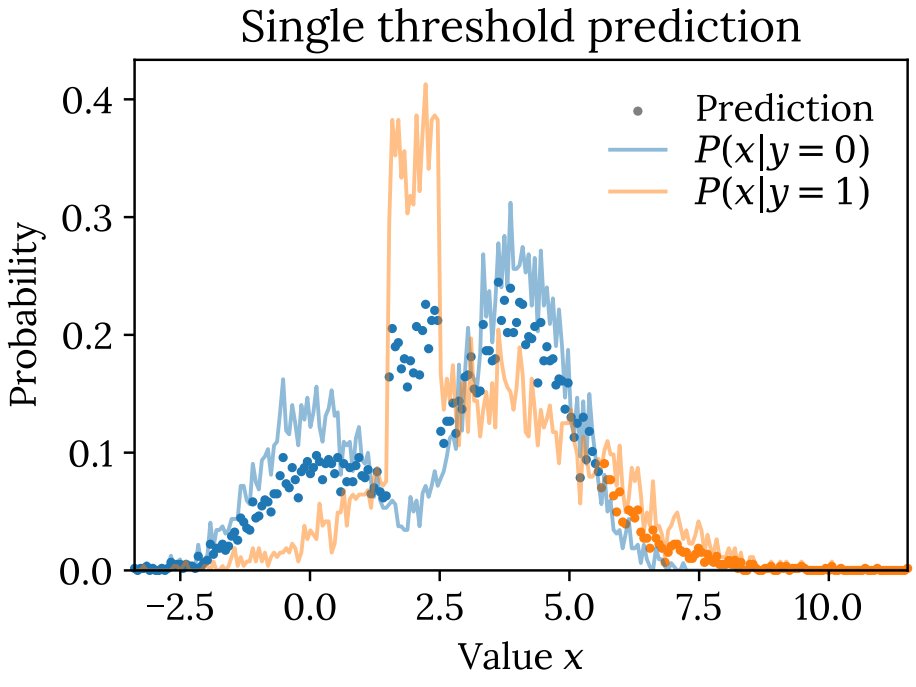


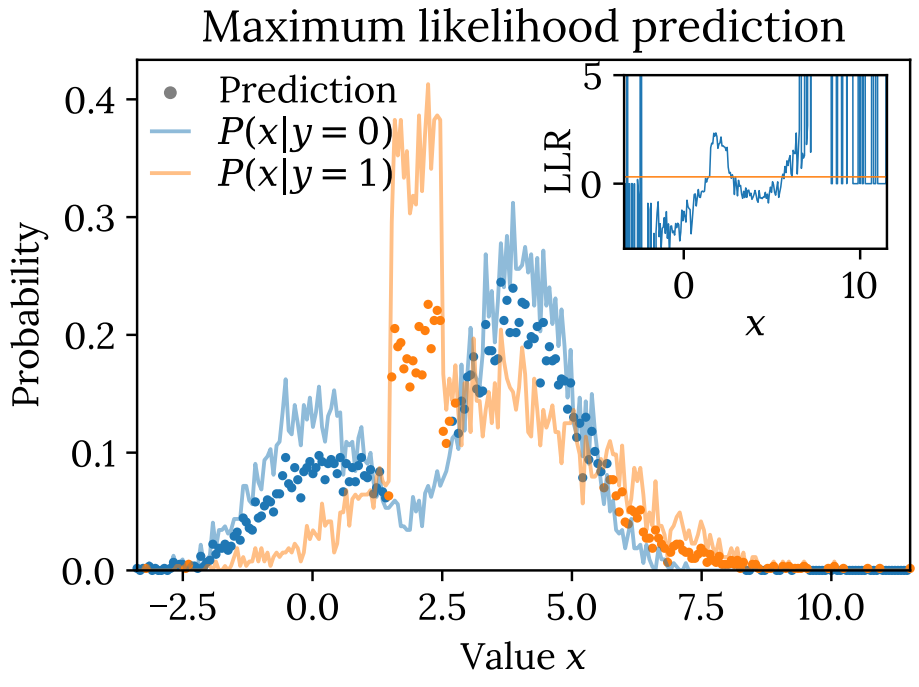
⇐ Decision threshold

Maximum likelihood ⇒



Optimal predictions





Without a prior or other incentives, the Maximum Likelihood estimation is the best we can do in predicting the labels of values from the learned distributions.

Bayesian approach with penalty for error

– when different mistakes carry different consequences

L_+ (L_-): loss for a wrong "+" ("−")

Given firing rate r, probability for $+$ is $P[+|r] = \dfrac{p[r|+]P[+]}{p[r]}$,

and similarly, $P[-|r] = \dfrac{p[r|-]P[-]}{p[r]}$. Meanwhile, the expected losses for a "+" and a "−" answer are $L_+P[-|r]$ and $L_-P[+|r]$ respectively. Choose "+" if $\text{Loss}_+ < \text{Loss}_-$, that is,

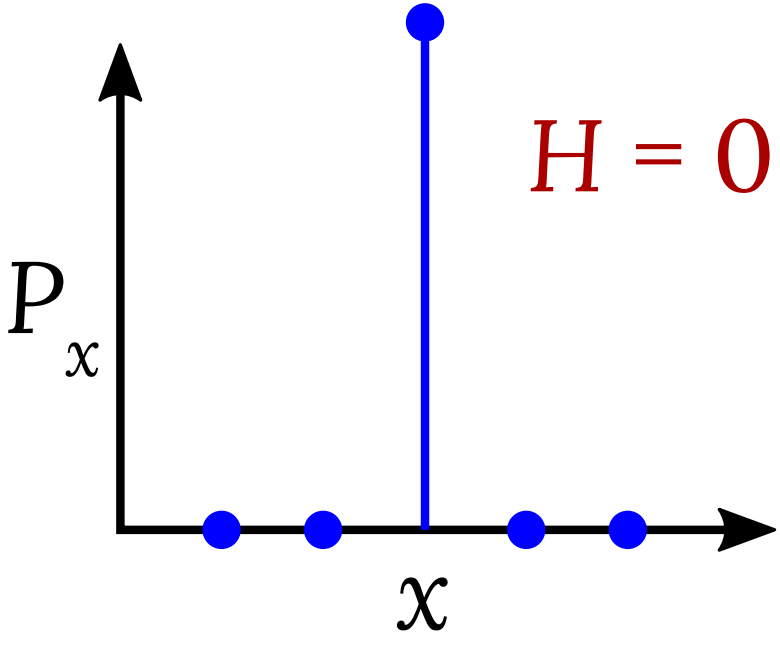
$$l(r) = \frac{p[r|+]}{p[r|-]} \geq \frac{L_+P[-]}{L_-P[+]}.$$

Threshold in likelihood ratio

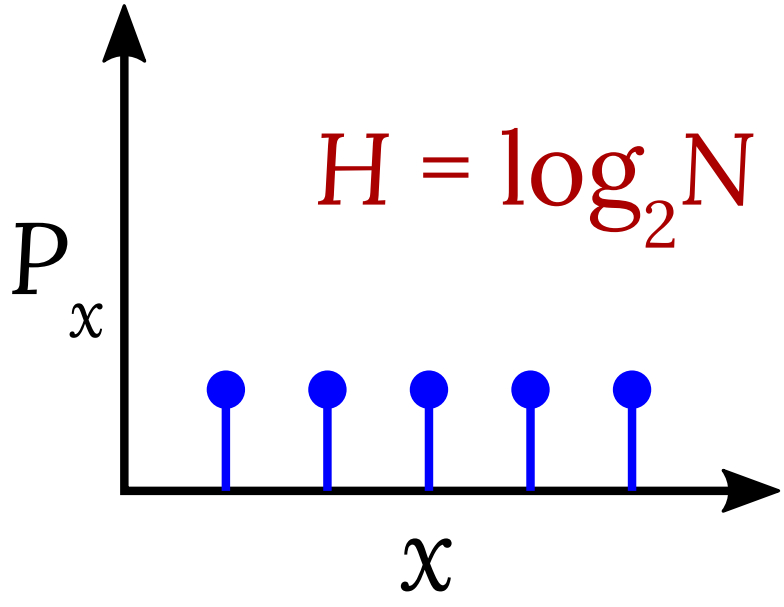
Both relative losses and prior probabilities are to be considered.

Entropy for discrete cases

Shannon entropy: $H = -\sum_x P_x \log_2 P_x$ (bits)



Minimal entropy



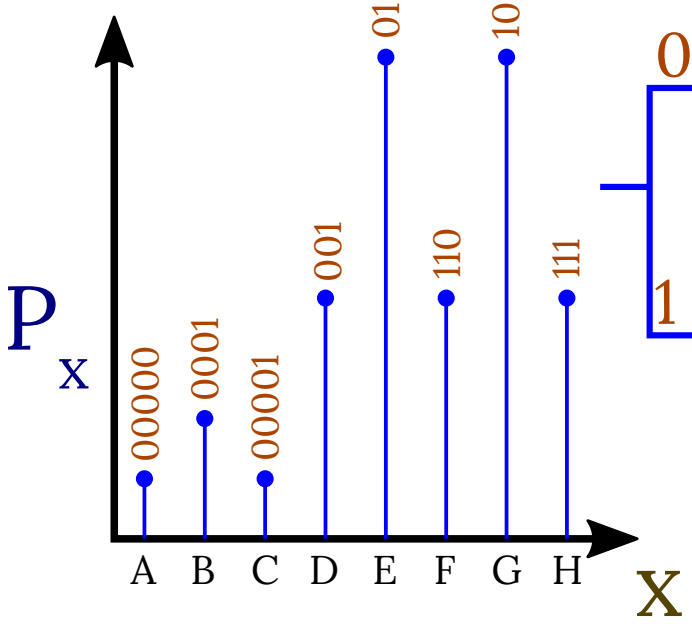
Maximal entropy

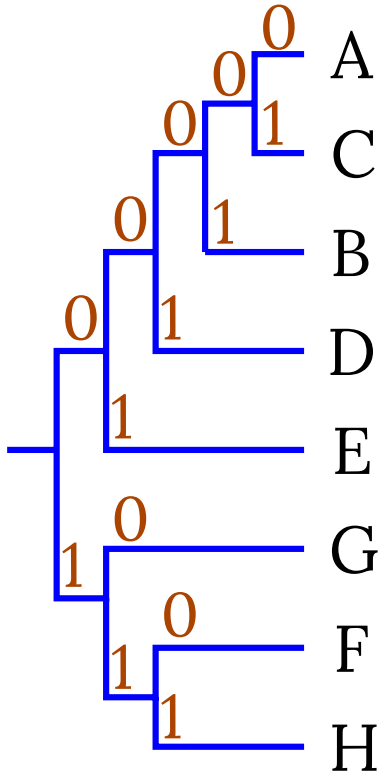
- Zero when there is only one possibility
- Maximized for a uniform distribution
- Invariant under a reversible change of variable: $x \to f(x)$

The Huffman code

– How we can have fractional bits of information

$N = 8, \log_2 N = 3$



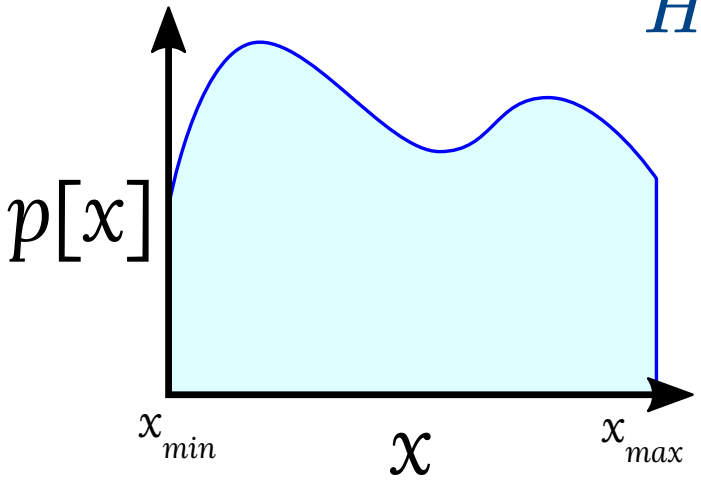


Fixed-length code takes $\log_2 N$ bits to encode event of N possibilities. By using variable-length code, we can shorten the expected number of bits required for each event realization. The minimum expected length of the code is just the entropy of the event.

Expectation value of code length

$$5\frac{1}{32} + 4\frac{1}{16} + 5\frac{1}{32} + 3\frac{1}{8} + 2\frac{1}{4} + 3\frac{1}{8} + 2\frac{1}{4} + 3\frac{1}{8} = 2.6875$$

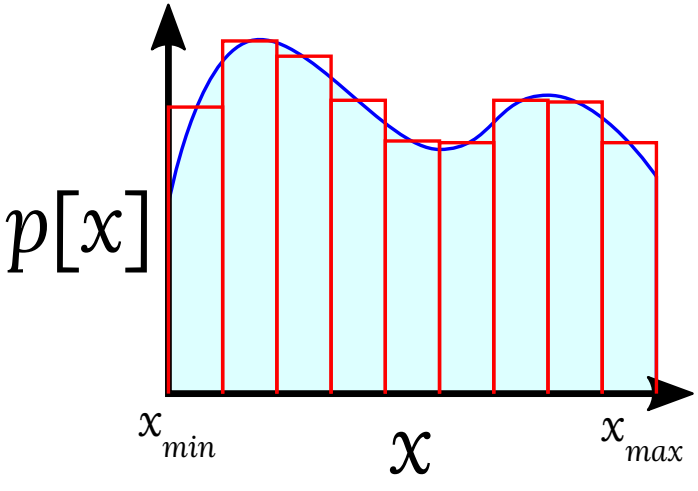
Entropy for continuous variables



$$H = -\sum_i p[x_i]\Delta x \log_2(p[x_i]\Delta x)$$

$$= -\sum_i p[x_i]\Delta x \log_2 p[x_i] - \log_2 \Delta x$$

$$= -\int dx p(x) \log_2 p(x) - \log_2 \Delta x$$



- Diverges when Δx goes to zero
- No longer invariant under a change of variable

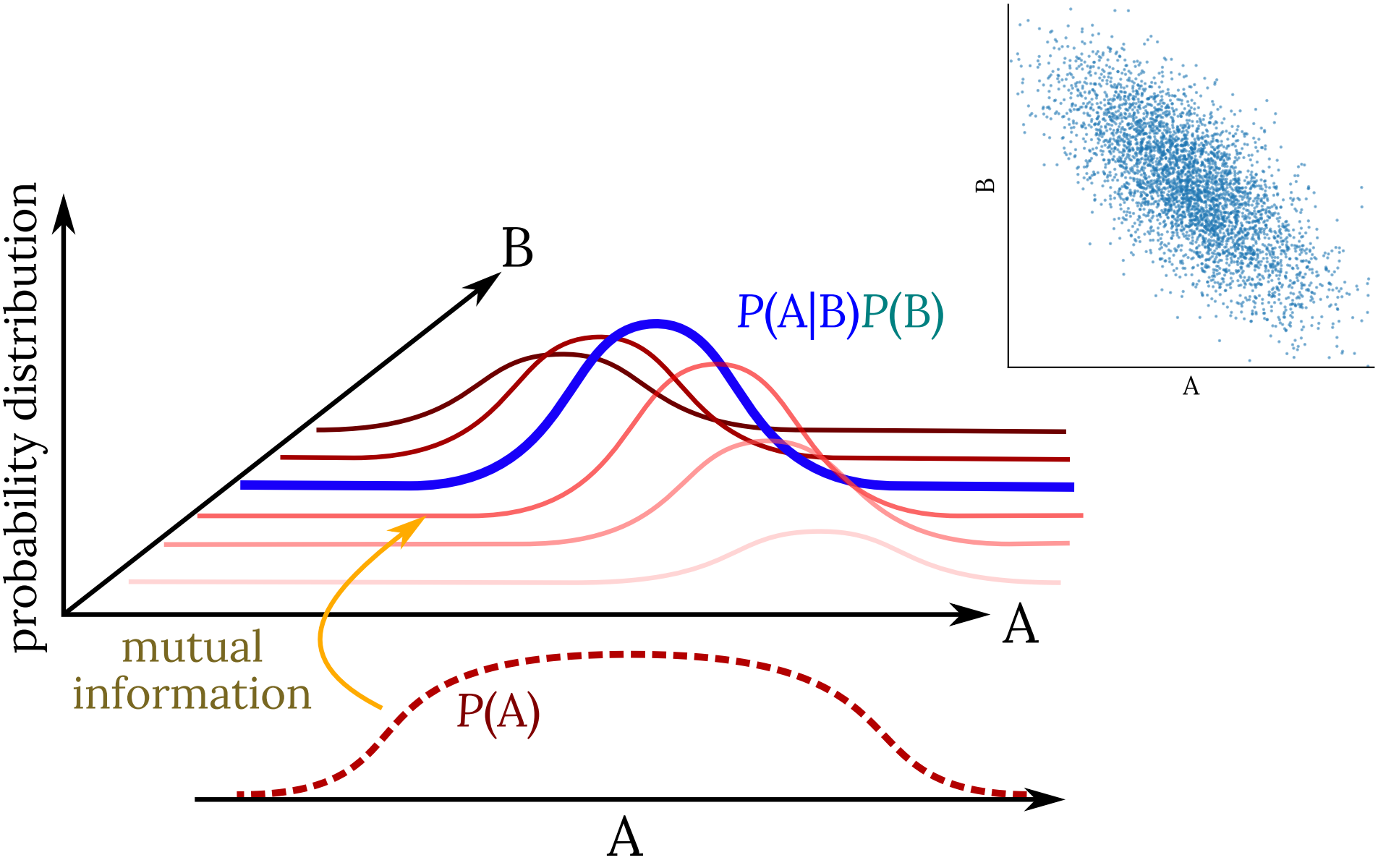
Mutual information for continuous variables

$$I_{\mathrm{m}}[x,y] = H[x] + H[y] - H[x,y]$$

$$\begin{aligned} I_m[x,y] = & -\int dx p[x] \log_2 p[x] - \log_2 \Delta x \\ & -\int dy p[y] \log_2 p[y] - \log_2 \Delta y \\ & +\int dx dy p[x,y] \log_2 p[x,y] + \log_2(\Delta x \Delta y) \end{aligned}$$

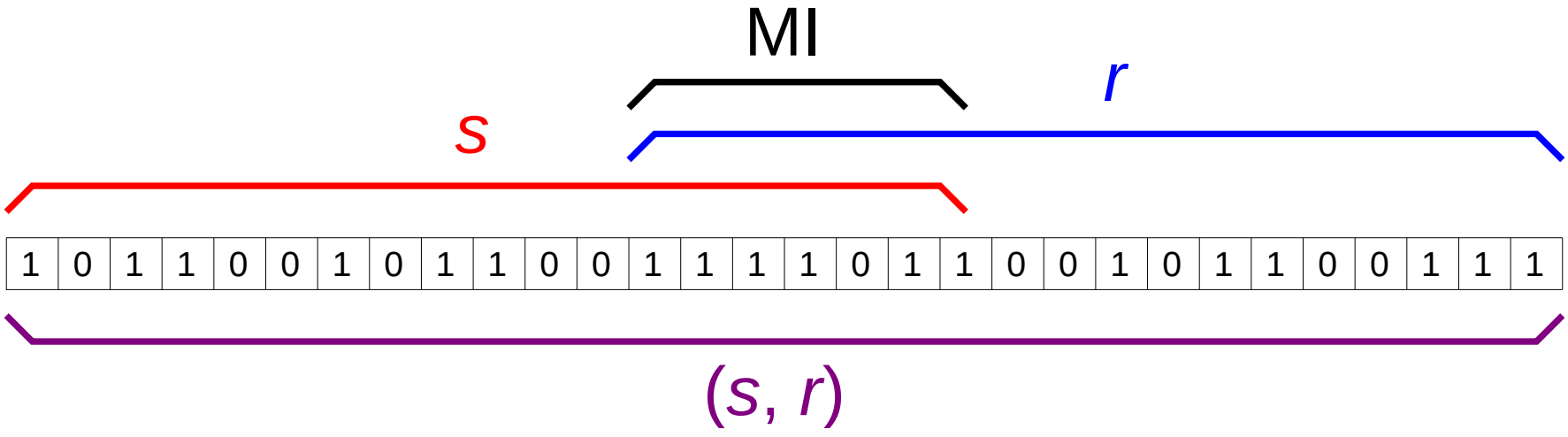
- Diverging terms cancel. → Remains a finite quantity!
- Invariant under a change of variables

Intuitive view of MI



Counting the bits for MI

Assume we can find a coding system where (a) noise in r; (b) mutual information between r, s; and (c) the rest of entropy in s are all encoded in different bits.



$$H(P[r]) + H(P[s]) - H(P[r, s])$$

Kullback-Leibler (KL) divergence

- Distance of one distribution from a reference distribution.

$$D_{\mathrm{KL}}(P, Q) = \sum_x P[x] \log_2 \frac{P[x]}{Q[x]}$$

- Asymmetric measure of the difference between two distributions
- KL divergence is zero if and only if the two distributions are the same.
- MI is KL divergence of the joint distribution $P[s, r]$ from the independent distribution $P[s]P[r]$. Thus, MI is never negative.
- Other popular test for difference between distributions: Kolmogorov-Smirnov Test

Entropy maximization

- Assume that biological system is optimized to convey as much information as possible
- Compute resulting response characteristics and compare with experimental observation
- Simplification: maximizing response entropy
- Different constraint observed in maximization leads to different form of response function

Maximization with limited range

Limited range of firing rate from 0 to $r_{\max}$

Maximize entropy

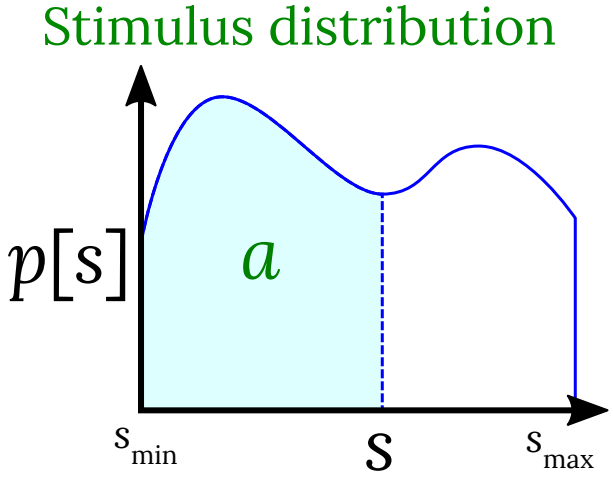
$$-\int_0^{r_{\max}} dr p[r] \log_2 p[r]$$

Under constraint

$$\int_0^{r_{\max}} dr\, p[r] = 1$$

Solution: $p[r] = \dfrac{1}{r_{\max}}$ (histogram equalization)

Application of histogram equalization principal

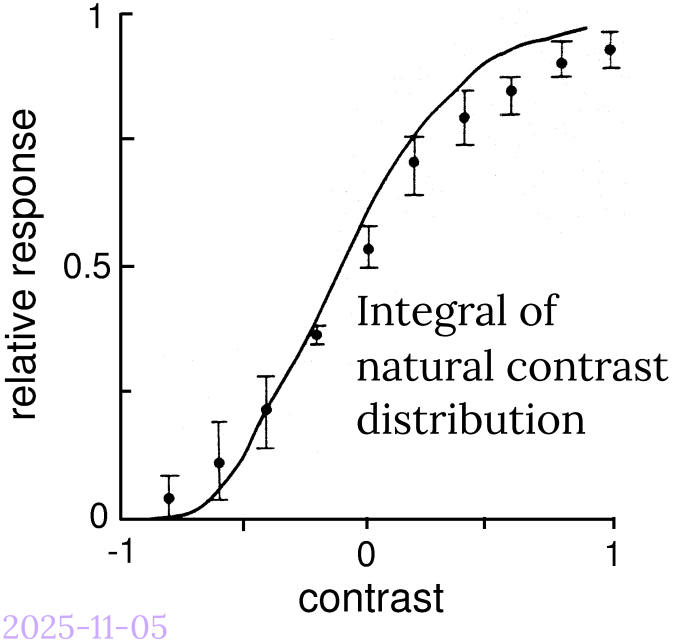


Probability is preserved in mapping from $s \rightarrow r$:

$$p[s]ds = p[r]dr = \frac{1}{r_{\max}}dr$$

Assuming $r = f(s)$, we thus have

$$\frac{dr}{ds} = r_{\max}p[s] \Rightarrow f(s) = r_{\max}\int_{s_{\min}}^{s} ds'$$



☆ The mapping function $f(s)$ can be predicted from the stimulus distribution.

Response of large monopolar cell in visual system of fly (Dayan & Abbott Fig. 4.2)

Constrained maximization

— *The method of Lagrange multiplier*

Derivative test for a maximum $x^\star$ (or minimum) of a function f:

$$f'(x^\star) = 0$$

With equality constraint $g(x) = 0$, additional parameter λ is introduced to construct the Lagrangian function:

$$\mathcal{L}(x, \lambda) = f(x) + \lambda g(x).$$

The new conditions for maximum (or a stationary point) are given by

$$\frac{\partial \mathcal{L}}{\partial x} = 0 \text{ and } \frac{\partial \mathcal{L}}{\partial \lambda} = 0$$

Maximization of entropy

— *constrained by probability rule two*

Lagrangian function: $\mathcal{L} = H + G$

$$\mathcal{L}\left[p[x], \lambda\right] = -\int dx\, p[x] \ln p[x] + \lambda \left(\int dx\, p[x] - 1 \right)$$

The stationary condition:

$$\frac{\delta H}{\delta p[x]} = -\int dx' \delta\left(p[x'] - p[x]\right) \ln p[x']$$

$$-\int dx' p(x') \frac{\delta\left(p[x'] - p[x]\right)}{p[x']}$$

$$= -\frac{1}{p'[x]} \left(\ln p[x] + 1\right)$$

$$\frac{\delta G}{\delta p[x]} = \lambda \frac{1}{p'[x]}$$

Extrema

$$\ln p[x] = \lambda - 1$$

$p[x]$ is a constant.

Different constraint

Fix mean firing rate additionally

$$\frac{\delta G}{\delta p[r]} = \frac{\delta}{\delta p[r]}\left(-\mu \int dr'\, r'\, p[r']\right) = -\frac{\mu r}{p'[r]}$$
$$\Rightarrow p[r] \propto \exp(-\mu r + \lambda)$$

Fix variance additionally

$$\frac{\delta G}{\delta p[r]} = \ldots - \frac{\delta}{\delta p[r]}\sigma \int dr' r'^2 p[r'] = \ldots - \sigma r^2$$
$$\Rightarrow p[r] \propto \exp(-\sigma r^2 + \ldots)$$