Basic counter

Given a bit sequence $y_1 y_2 \dots y_K,\ y_k \in \{0,\ 1\}$ of length K a series $c_k \in \mathbb{R}$ with $\forall k\ c_k > 0$ can be used to evaluate a probability estimation $p = P(y = 1)$.

$$H_K(p) = -\sum_{k=1}^{K} c_k y_k \log_2 p + c_k(1 - y_k) \log_2 p \tag{1}$$

More recent observations should receive higher attention, since commonly stationary sources cannot be assumed. To achieve that the series c_k must fulfill $c_l > c_k\ l > k,\ 1 \le l, k \le K$. Eq. (1) evaluates the entropy of the *already known* bit sequence: The coding cost of a one-bit is $-\log_2 p$ and $-\log_2 1 - p$ for a zero-bit, respectively. Merging both formulation yields the compact representation $-y \log_2 p - (1 - y) \log_2 1 - p$. An optimal estimation p minimizes H_K.

$$p_K = \arg\min_p H_K \tag{2}$$

Note that a single prediction p at a step K is evaluated using the whole, weighted, bit history. The first derivative of (1) and $\log_x y = \frac{\ln x}{\ln y}$ are utilized to get the optimal estimation p:

$$\begin{aligned}
\frac{\partial H_K}{\partial p} &= 0, && (3)\\
0 &= -\frac{1}{\ln 2} \sum_{k=1}^{K} c_k \left(\frac{y_k}{p} - \frac{1 - y_k}{1 - p} \right), && (4)\\
0 &= -\sum_{k=1}^{K} c_k \left(y_k(1 - p) - (1 - y_k)p \right), && (5)\\
0 &= -\sum_{k=1}^{K} c_k \left(y_k - (y_k + (1 - y_k))p \right), && (6)\\
p \sum_{k=1}^{K} c_k &= \sum_{k=1}^{K} c_k y_k, && (7)\\
p &= \frac{S_K}{T_K} = \frac{\sum_{k=1}^{K} c_k y_k}{\sum_{k=1}^{K} c_k}. && (8)
\end{aligned}$$

Optimality can be proven by deriving eq. (4) and examining the condition $\frac{\partial^2 H_K}{\partial p^2} > 0$.

$$\begin{aligned}
\frac{\partial^2 H_K}{\partial p^2} &= -\frac{1}{\ln 2} \frac{\partial}{\partial p} \sum_{k=1}^{K} c_k \frac{y_k - (y_k + (1 - y_k))p}{p(1 - p)} && (9)\\
&= -\frac{1}{\ln 2} \frac{\partial}{\partial p} \sum_{k=1}^{K} c_k \underbrace{\frac{y_k - p}{p(1 - p)}}_{q_k} && (10)
\end{aligned}$$

The quotient q_k can be reformulated:

$$q_k|_{y_k=0} = \frac{1}{p-1}, \tag{11}$$

$$q_k|_{y_k=1} = \frac{1}{p}, \tag{12}$$

$$q_k = \frac{1}{p+y_k-1}. \tag{13}$$

Joining equations (10) and (13) yields

$$\frac{\partial^2 H_K}{\partial p^2} = -\frac{1}{\ln 2}\sum_{k=1}^{K} c_k \frac{\partial}{\partial p}\frac{1}{p+y_k-1}, \tag{14}$$

$$= \frac{1}{\ln 2}\sum_{k=1}^{K} c_k \frac{1}{(p+y_k-1)^2}. \tag{15}$$

Eq. (15) exists and is positive, iff:

- $c_k > 0$ and
- $p \in (0,\ 1) \subset \mathbb{R}$, since there's no restriction to y_k.

Approximation

For practical implementations it is reasonable to chose $c_k = c^{K-k},\ c \in (0, 1] \subset \mathbb{R}$. The term T_K contained in equation (8) becomes a geometric series.

$$T_K = \sum_{k=1}^{K} c^{K-k} = \sum_{k=0}^{K-1} c^k \tag{16}$$

Comparing nominators and denominators of (8) for K and $K-1$ in conjunction with eq. (16) reveals an iterative formulation.

$$S_K = cS_{K-1} + y_k \tag{17}$$

$$T_K = cT_{K-1} + 1 \tag{18}$$

An alternative formulation arises after compareing the estimations p in steps $K-1$ and K.

$$p_K = p_{K-1} + \frac{1}{T_K}(y_K - p_{K-1}) \tag{19}$$

$$= \left(1 - \frac{1}{T_K}\right) p_{K-1} + \frac{1}{T_K} y_K \tag{20}$$

Further simplifications require a very long bit sequence ($K \rightarrow \infty$). Equation (16) can be replaced by its limit

$$T_K \xrightarrow{K\to\infty} \frac{1}{1-c} \tag{21}$$

and (20) is characterized by exponential smoothing:

$$p_K = p_{K-1} + (1-c)(y_K - p_{K-1}) \quad (22)$$
$$= cp_{K-1} + (1-c)y_K \quad (23)$$

Finally the result is the (in)famous linear counter used in almost every context mixing compression scheme.