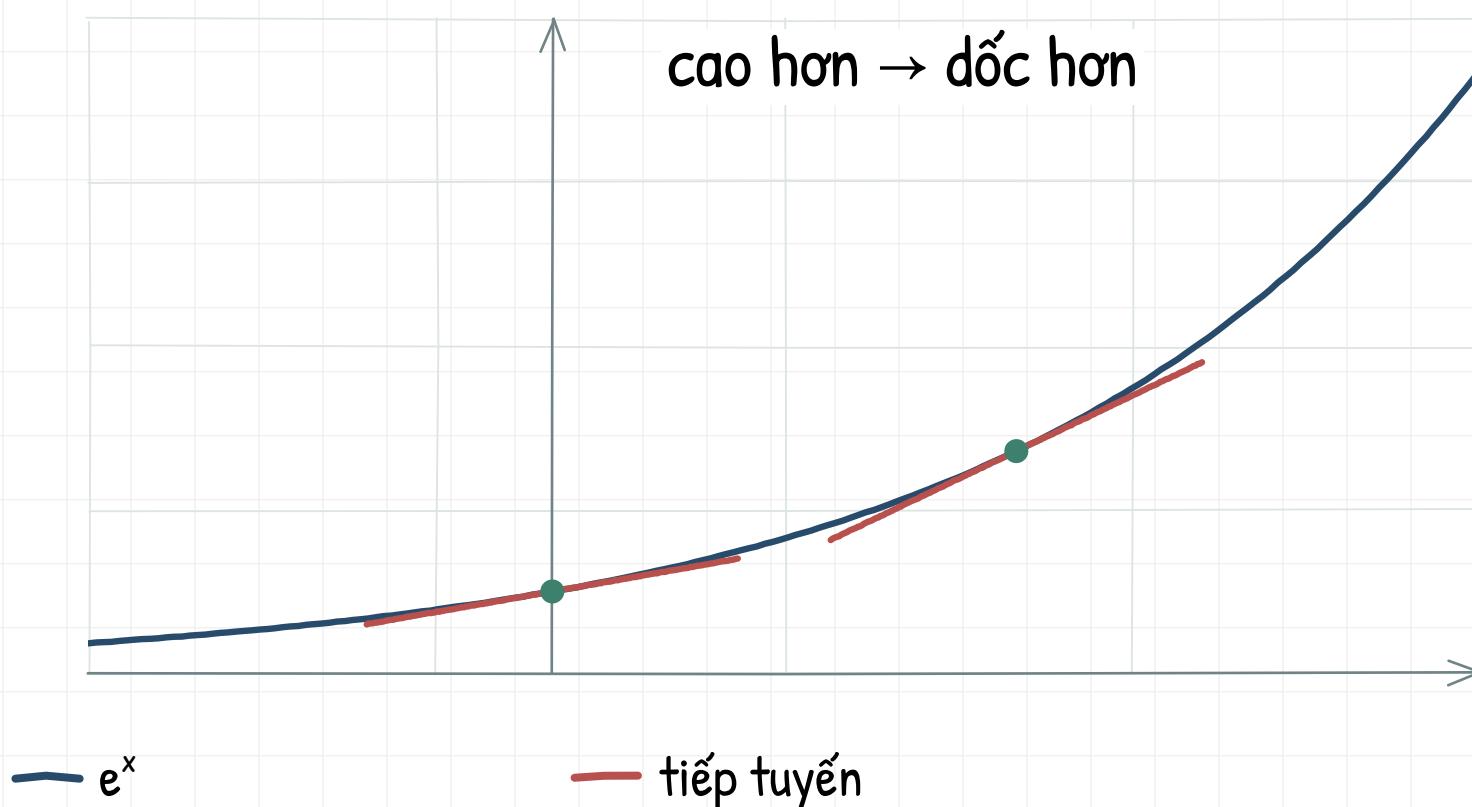


# Tăng theo lượng đang có

Một lượng càng lớn thì tăng càng nhanh, theo cùng một tỉ lệ.



## Từ hình đến công thức

Mô hình:  $y' = k \cdot y \rightarrow y(x) = C \cdot e^{(kx)}$ .

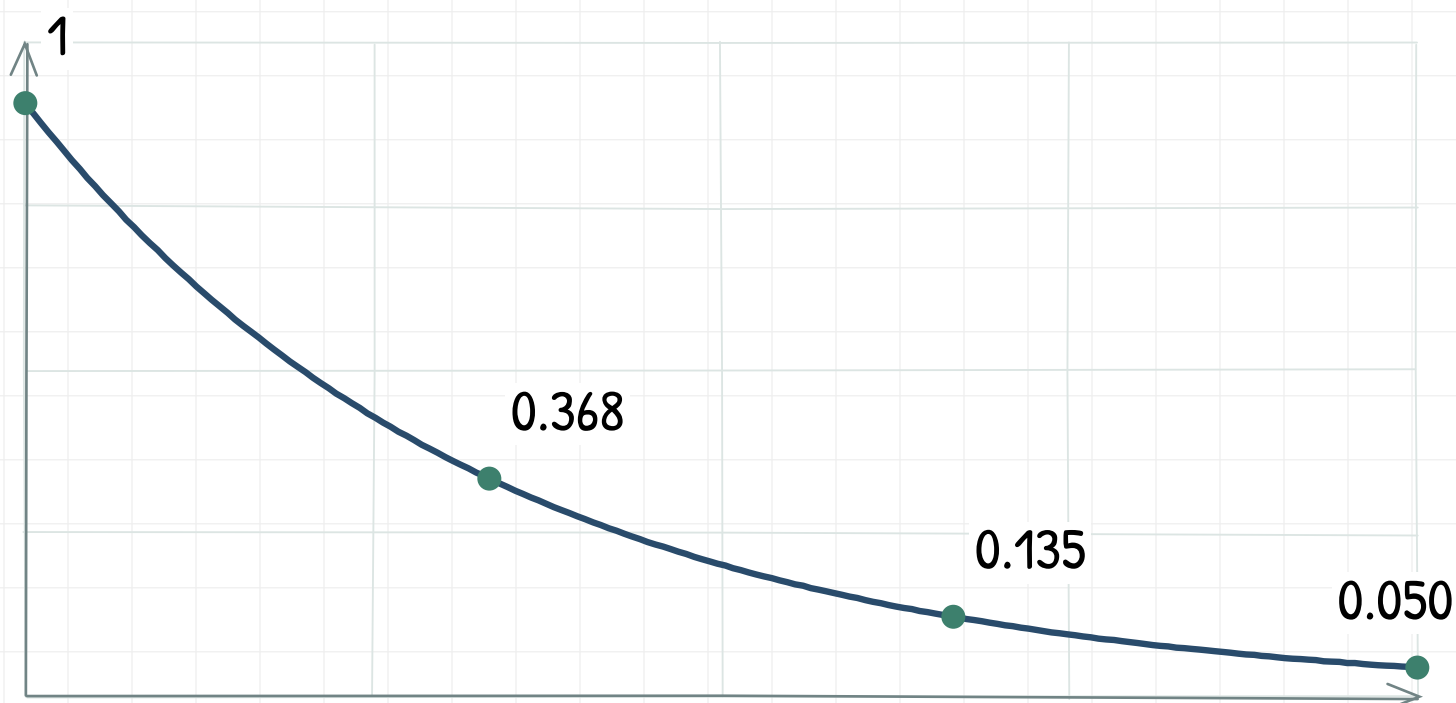
Với  $k=1$  và  $y(0)=1$ , nghiệm duy nhất là  $e^x$ . Nếu chưa biết  $y(0)$ , còn cả họ  $C \cdot e^x$ .

Với  $a>0$ :  $(a^x)' = \ln(a) \cdot a^x$ . Cơ số  $e$  làm  $\ln(a)=1$ .

“Duy nhất” cần điều kiện ban đầu, không chỉ  $y' = y$ .

# Mỗi chạng mắt cùng một tỉ lệ

Đường cong giảm nhanh lúc còn nhiều, chậm dần khi còn ít.



—  $R(t), S = 10$

— trục ngang: thời gian

## Từ hình đến công thức

$$R' = -R/S, S > 0 \rightarrow R(t) = R_0 \cdot e^{(-t/S)}.$$

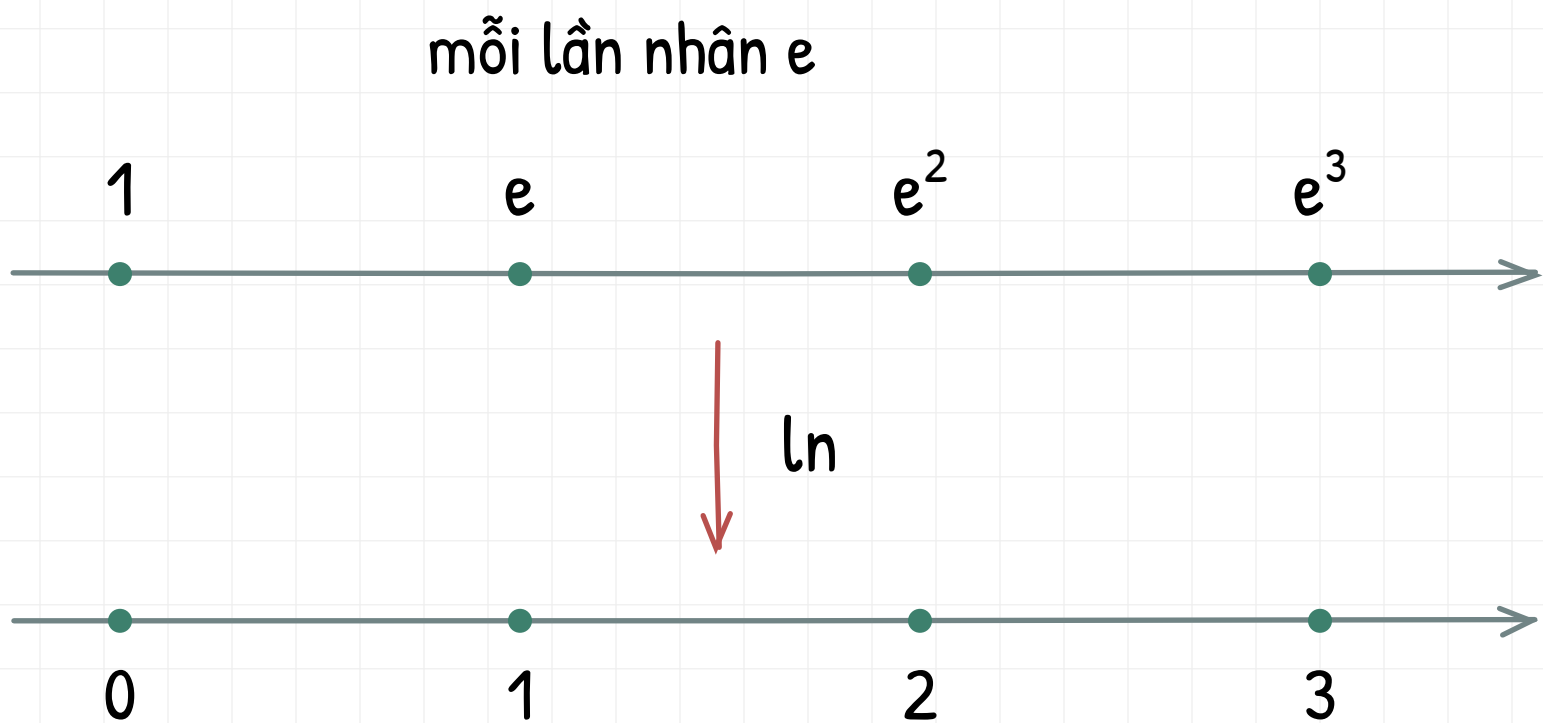
Sau mỗi  $S$  đơn vị thời gian:  $R(t+S) / R(t) = 1/e$ .

Nửa đời:  $R/R_0 = 1/2 \rightarrow t = S \cdot \ln 2$ . Với  $S=10$ :  $t \approx 6.931$ .

Đây là mô hình khi tốc độ giảm tỉ lệ với lượng còn lại.

# Log: đổi phép nhân thành cộng

Trên thước log, nhân cùng một hệ số là đi cùng một khoảng.



## Từ hình đến công thức

$\ln$  là hàm ngược của  $\exp$ :  $\ln(e^x) = x$ ;  $(\ln x)' = 1/x$ , với  $x > 0$ .

$\ln(ab) = \ln a + \ln b$ ;  $\ln(a^k) = k \cdot \ln a$ , với  $a, b > 0$ .

Vì  $\ln$  tăng nghiêm ngặt: tối đa likelihood dương cũng là tối đa log-likelihood.

Đổi thang đo giữ thứ tự, đồng thời biến tích thành tổng.

# Một nghìn xác suất nhỏ

Nhân liên tục làm số co lại; lấy log trước giúp giữ thông tin.

| Cách tính                       | Kết quả float64 |
|---------------------------------|-----------------|
| $0.01 \times \dots \times 0.01$ | 0.0             |
| $1000 \cdot \ln(0.01)$          | -4605.17        |
| $\ln(\text{tích đã thành } 0)$  | $-\infty$       |

## Từ hình đến công thức

Tích thật  $0.01^{1000} > 0$ , nhưng nằm ngoài dải float64.

Phải cộng log từng số trước:  $\log L = \sum \ln p_i$ .

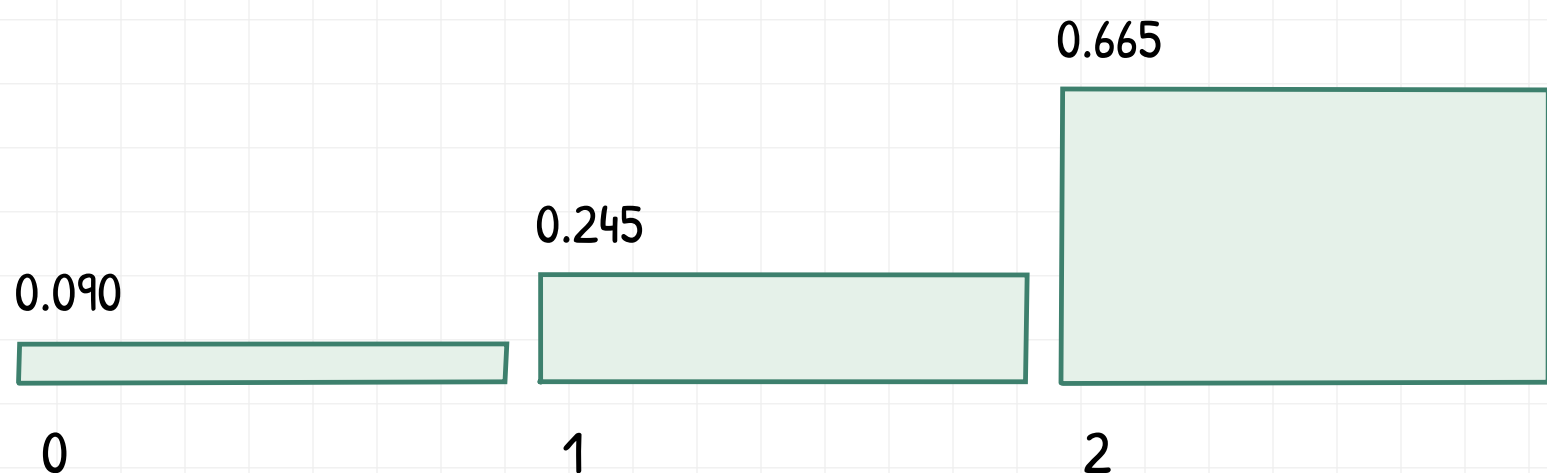
Đạo hàm của tổng cũng dễ xử lý hơn. Lấy log sau khi underflow không cứu lại được.

Giữ phép tính trong miền log càng lâu càng tốt.

# Softmax: điểm số thành xác suất

Ba cột cao thấp khác nhau; chia tổng để mọi phần cộng lại thành 1.

logits  $z = (0, 1, 2)$



## Từ hình đến công thức

$p_i = \exp(z_i) / \sum_j \exp(z_j)$ . Trên số thực:  $p_i > 0$  và  $\sum p_i = 1$ .

Cộng cùng  $c$  vào mọi logit:  $\exp(c)$  triệt tiêu giữa tử và mẫu.

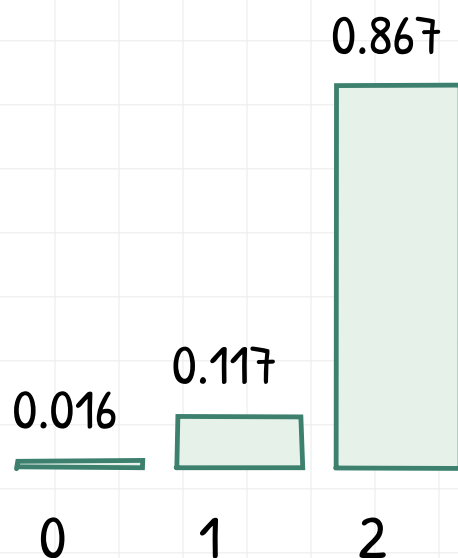
Do đó  $\text{softmax}(z+c) = \text{softmax}(z)$ . Chỉ chênh lệch giữa logits có ý nghĩa.

Trên máy, xác suất cực nhỏ vẫn có thể bị làm tròn về 0.

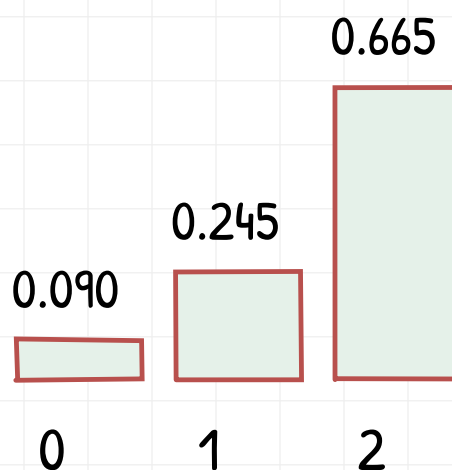
# Nhiệt độ đổi độ sắc nét

Giữ logits (0,1,2), chỉ đổi  $T > 0$  trong  $\text{softmax}(z/T)$ .

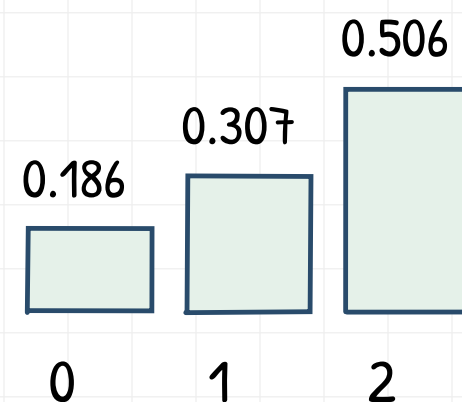
$T = 0.5$



$T = 1$



$T = 2$



## Từ hình đến công thức

$T$  nhỏ: tập trung vào logit lớn nhất.  $T$  lớn: phân phối gần đều hơn.

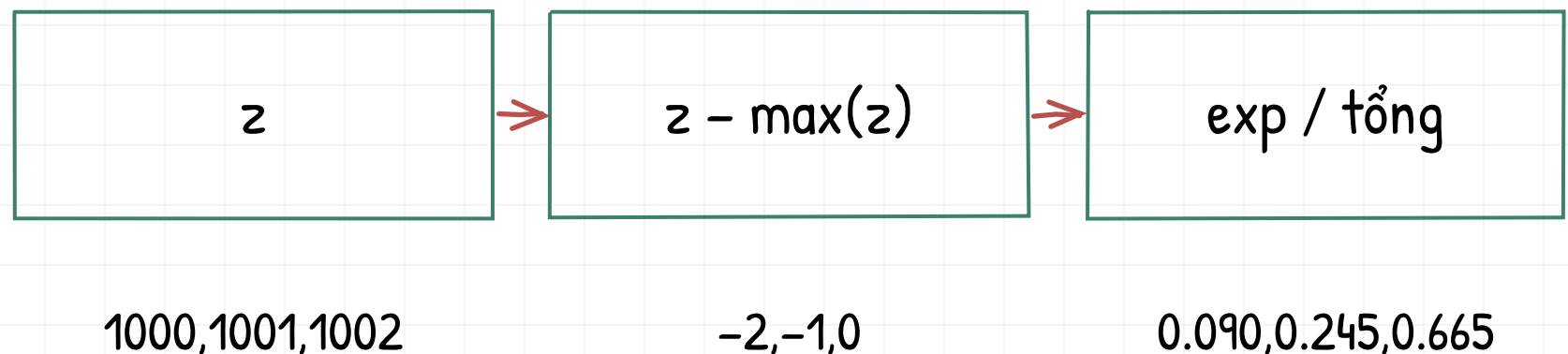
Đổi  $\exp(z)$  sang  $2^z$  tương đương  $\text{softmax}(z \cdot \ln 2)$ , tức  $T = 1/\ln 2$ .

Nếu nhiều logits cùng lớn nhất,  $T \rightarrow 0$  chia đều cho các vị trí đó.

Đổi cơ số không phá softmax; nó đổi nhiệt độ.

# Tính ổn định, rồi đo sai

Trừ max trước khi mũ hoá; lấy log trực tiếp để tránh xác suất thành 0.



## Từ hình đến công thức

Đặt  $a = \max(z)$ :  $\text{logsumexp}(z) = a + \ln \sum \exp(z_j - a)$ .

$\log p_i = z_i - \text{logsumexp}(z)$ ; cross-entropy  $L = -\log p_k$ .

Với  $z = (0, 1, 2)$ , nhãn  $k = 2$ :  $L \approx 0.4076$ ; gradient  $\nabla_z L = p - y$  ( $y$  one-hot).

Công thức tương đương trên giấy có thể khác độ ổn định trên máy.

# Tự kiểm tra khi đi tiếp

## Che đáp án, thử nói bằng lời

1.  $y' = y$  có buộc  $y = e^x$  không?
2.  $S=10$  thì nửa đời của decay là bao nhiêu?
3. Tại sao cộng log thay vì nhân xác suất?
4. Cộng 1000 vào mọi logit đổi softmax không?
5. Vì sao cần log-softmax thay vì  $\log(\text{softmax})$ ?

## Đáp nhanh

1. Chỉ khi  $y(0)=1$ ; tổng quát  $y=C \cdot e^x$ .
2.  $10 \cdot \ln 2 \approx 6.931$  đơn vị thời gian.
3. Tránh underflow và biến đạo hàm của tích thành của tổng.
4. Không về toán; cách tính trực tiếp có thể overflow.
5. Xác suất có thể làm tròn về 0 trước khi lấy log.