Given names, we want to make more names.

If we study ML for 10 hours, it will take

10 hrs * x days = 10,000

1000 days of study

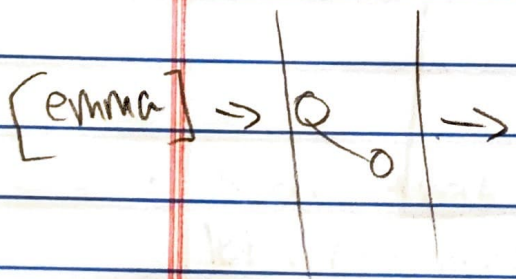
≈ 2.73 years

≈ 3 years of serious commitment to become an "ML expert"

| . | e | m | m | a | . |
|---|---|---|---|---|---|
| 0 | 1 | 2 | 3 | 4 | 5 |
| 0 | 5 | 13 | 13 | 1 | 0 |

| | |
|---|---|
| . | 0 |
| a | 1 |
| b | 2 |
| c | 3 |
| d | 4 |
| e | 5 |
| f | 6 |
| g | 7 |
| h | 8 |
| i | 9 |
| j | 10 |
| k | 11 |
| l | 12 |
| m | 13 |
| n | 14 |
| o | 15 |
| p | 16 |
| q | 17 |
| r | 18 |
| s | 19 |
| t | 20 |
| u | 21 |
| v | 22 |
| w | 23 |
| x | 24 |
| y | 25 |
| z | 26 |

27 Characters

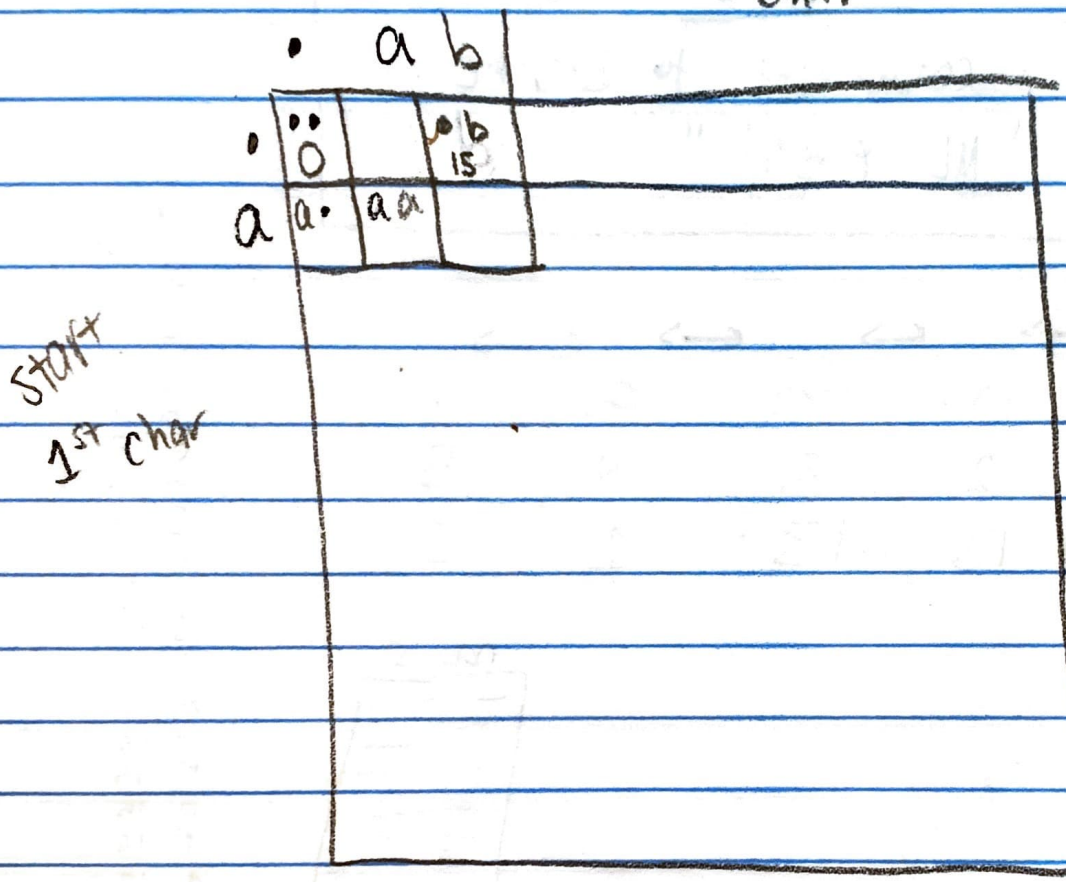




How do we define loss?

- create bigram count model
- Convert counts into probabilities
- Use $P_r$ in loss fn

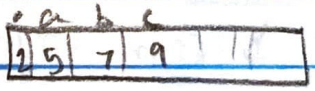
# Bigram Count Model




ob: How ma
does a w
w/ b

What does a row vs column tell us?

i.e. Start_char = a




$\frac{9}{23}$

$aa = \frac{5}{23}$

$a\left[\frac{2}{23}, \frac{5}{23}, \frac{7}{23}, \frac{9}{23}\right]$

Pr of end char given a

e m m a

. .2 .3 .9

A:

loss = $(1 - P_r)$

[aa  aa : 5
ab : 7

a ← aa
a ← ab

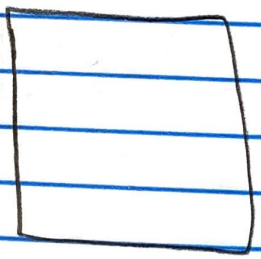
Given a, I should be able to 'index' into the count of an arbitrary 2nd char

Goal: Calculate loss for 'emma'
(1+2+1)

em → count → $P_r$

e m
↑
$P_r$

ea  $P_r$ ?
eb
:
em



Now that we have the counts, lets construct the probabilities
$P_r$(em) => Given e, $P_r$(m) ?

e

| a | b | Σ |
|---|---|---|
| 1 | 2 | 15 |

e_
↑
sum of all counts

$\frac{\text{count}(\_)}{\text{sum(counts)}}$

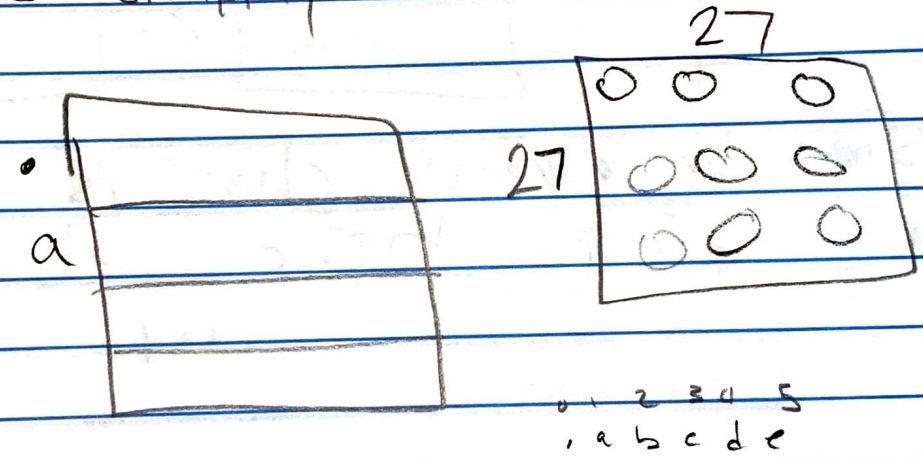
char
sum = 0

for each 1st char
  for each 2nd char
    sum[char] += char_pairs[1st char] + char_pairs[2nd char]

char_sum['e'] = 15

OK I think we're double counting the . character

Too much complexity w/ these loops, lets simplify the data structure




. e m m a .

↑ ↑

.e += 1

char_counts[0, 5] += 1

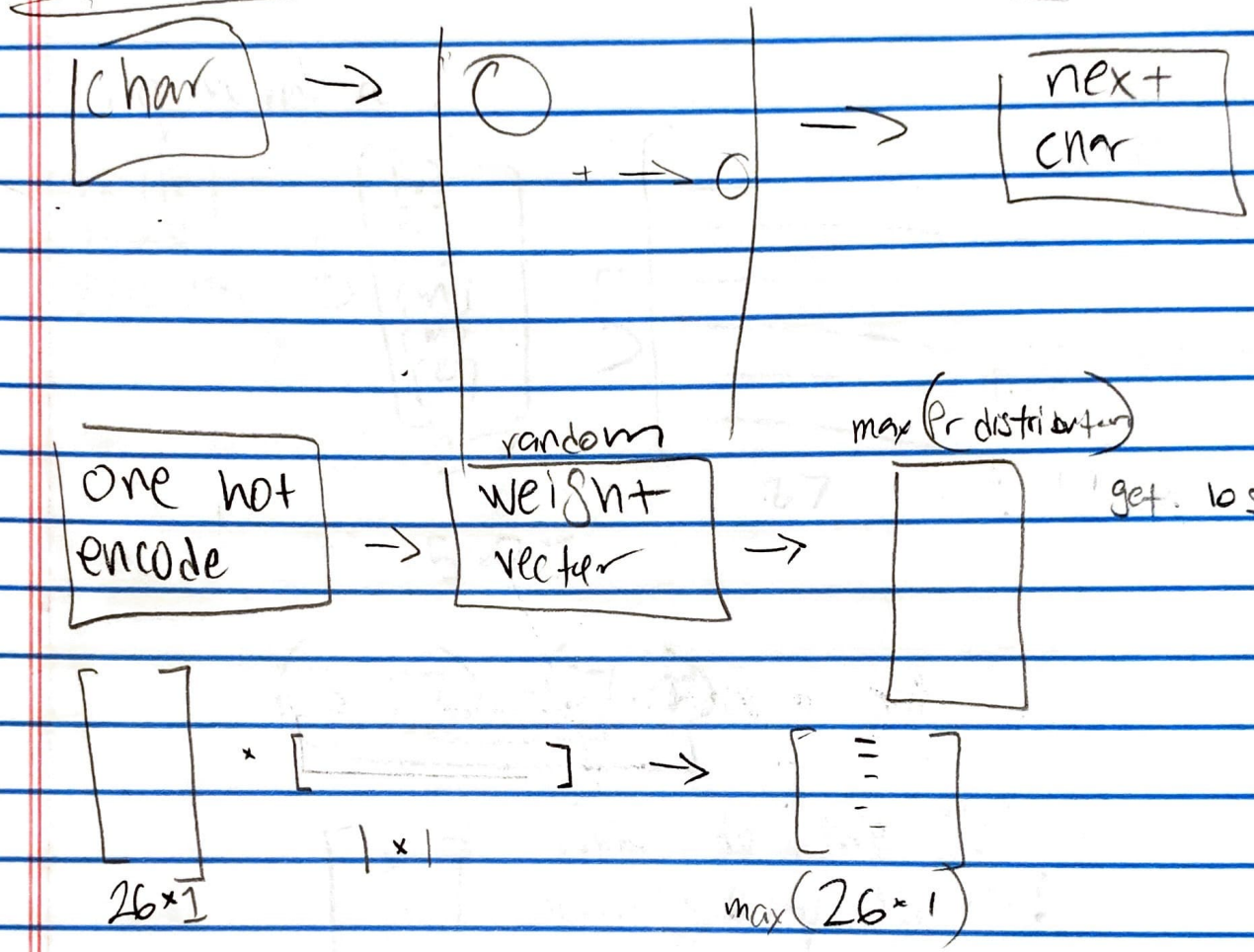
[0, first char]
[last char, 27]

How would it handle a. ?

A: manually accounted for this edge case

My loss fn looks ass

- lets normalize over name length



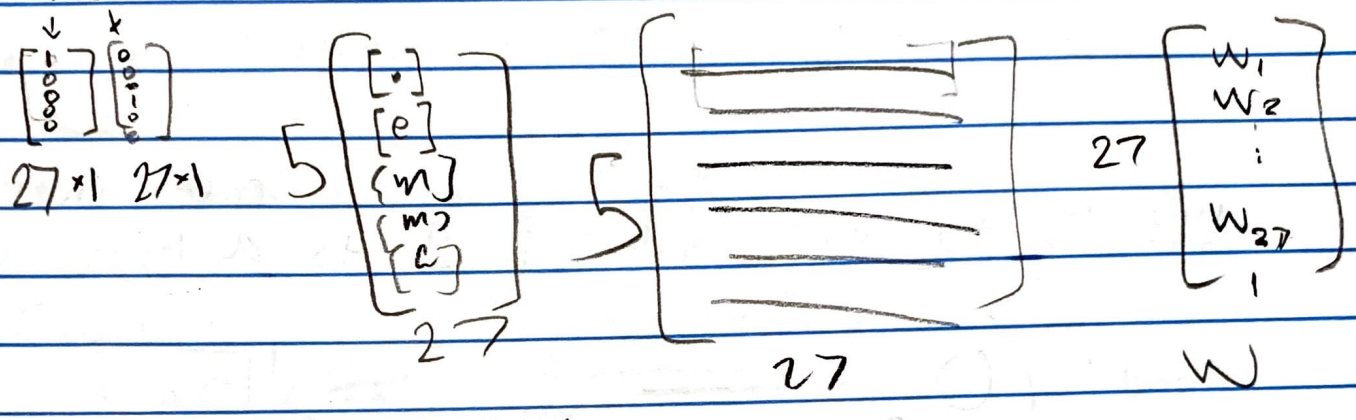

| input | output vector |
|---|---|
| . | e |
| e | m |
| m | m |
| m | a |
| a | . |

- One hot : Popular way to encode integers encode

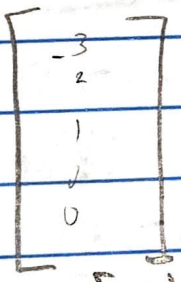
[ ]

5×27

.emma.




(5×27) × (27×27)




- What do the #'s in the resulting array tell us?

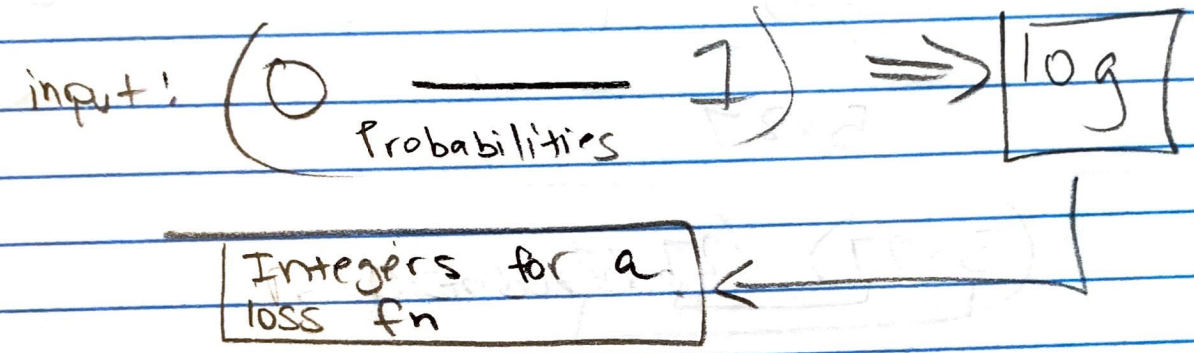
A: Not much, we need to increase W

(5×27) @ (27×27)

5×27 Pr likelihood of next Char

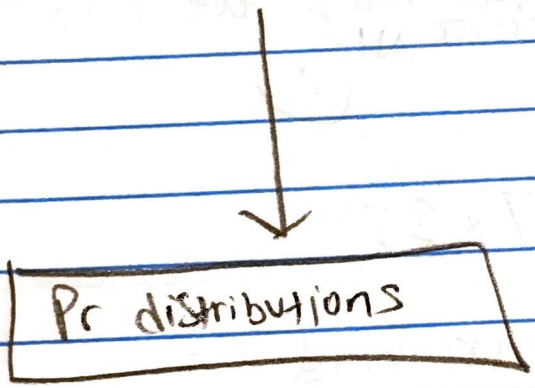
Pr

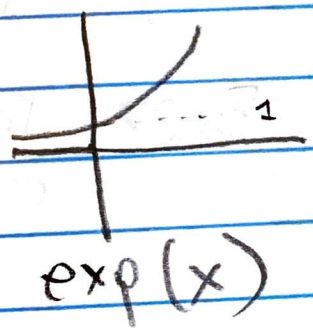
• Use log to squash values [0,1] to a nice integers for a loss fn





I) Fn → Pr needs to
- Make input numbers
  - all positive
  - [0,1]
  - sum to 1



(?) Why do we assume

xenc @ W = log counts

- Previously log took us from a Pr distribution to integer values for a loss fn

What do logits represent?

Counts → Pr → loss

counts / counts.sum()  log

Math: Fn(Pr) → ℝ
logits

Softmax (ℝ) → Pr

counts = exp(ℝ)  # make ℝ in positive range only

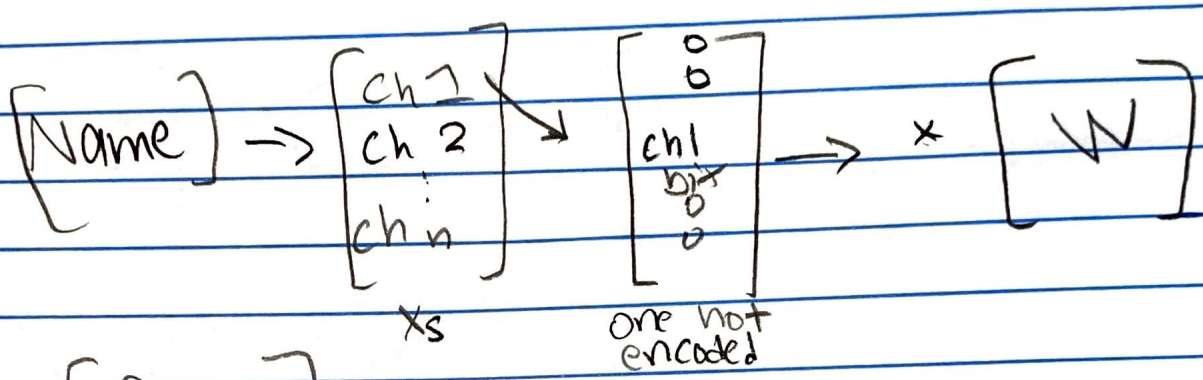
Pr = counts / counts.sum()  # squash to [0,1] & sum to one

return Pr

- Activation / normalization fn

- We have Pr matrix for the name Emma

- We're not taking the max Pr

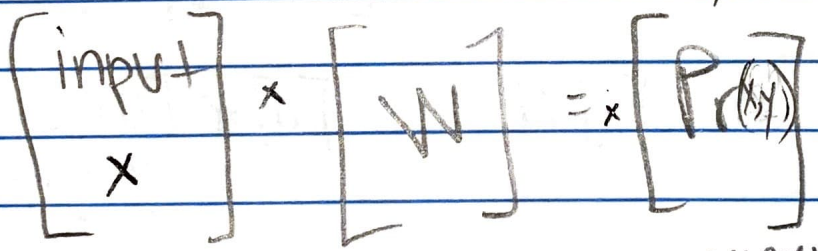



$=$ [PROBabilities of Bigram]

(We are deriving Pr from W instead of counting)

X = ch1 # input

Y = ch2 # output



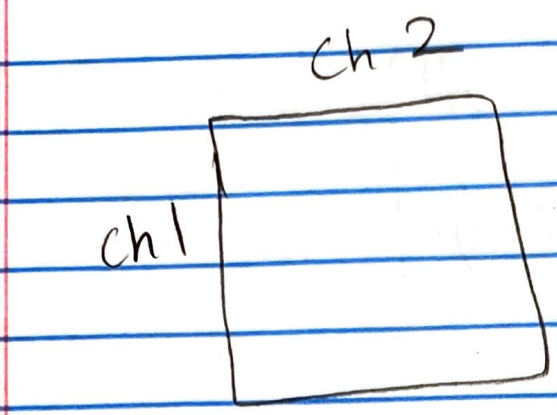

now we use y to index into our Pr matrix

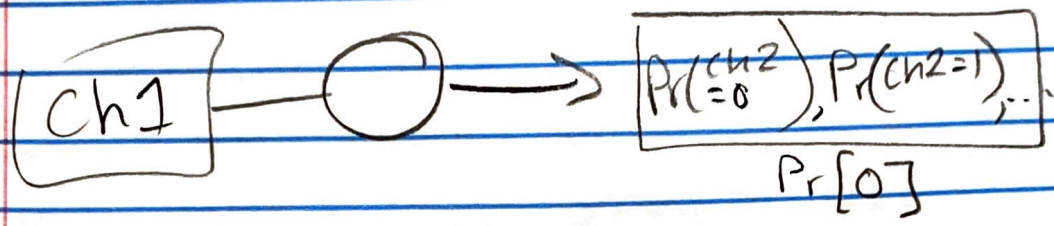
- $P_r$ = Represents the activation or output layer

$X_s$ = 1st char in bigram

$Y_s$ = 2nd char in bigram

Pr(X,Y) => Loss




$Pr(y|s) = Pr(y) = Pr \quad i=0$

$y = 'e'$ # truth used in loss

$y\_i = 4$ $\quad xs \rightarrow xs[i] \rightarrow xs[0] = 26$

$ys \rightarrow ys[i] \quad 4 \rightarrow e$

where $i=0$

$Pr[0][ys[i]]$

$= Pr[0] \; ys[0] \qquad Pr[i]$

$= [Ch1] \quad 4$

$5 \times 1$

$Pr =$ Ch1 | Pr(Ch2=a given Ch1=•) ⋮ | i=0

Given $\rightarrow Pr$ Given Ch1 = •, pr_dist ch[0] of name

predict ch[1] of name

$[Pr]$ Given • $[Pr(a)\ Pr(b)]$ Ch1 $[Pr(a=Ch2)]$ (Ch1, Ch2) Bigram for all chars in name

Pr_dist = a [ $P_r(ch2=a \mid ch1=\cdot$ ] ... $P_r(ch2=y_i \mid ch1=\cdot$ ]

$P_r(ch2=b \mid ch1=\cdot$

a

b

| | a | b |
|---|---|---|
| a | $P_r(ch2=a \mid ch1=a)$ | $P_r(ch2=b \mid ch1=a)$ |
| b | $P_r(ch2=a \mid ch1=b)$ | $P_r(ch2=b \mid ch1=b)$ |

- Funny the diagnol of pr-dist is the $P_r(ch1=ch2 \mid ch2)$

$[X] \times [W] = \times [\quad]$ (Y)

ch1 weight

$P_r(y \mid x)$ $y=4$

$P_r(y_i=ch_i \mid x_i)$ $P_r(y=4 \mid x)$

- Use y to calculate loss

$\text{Loss} = \sum 1 - P_r(y \mid x)$