

# Lecture 14: Review



Gabriel Poesia

Assistant Professor, CSE

University of Michigan, Ann Arbor

[poesia@umich.edu](mailto:poesia@umich.edu)

<https://gpoesia.com>

**Assignment 3 due tonight!**

**Midterm on Monday.**

**DOW 1010** (here) and **DOW 1014** (next door).

Exams will be on desks, facing down. Find an empty spot.

Logistics recap will be posted on Ed.

# Review quiz

Open **In-class quiz #2** on [Gradescope](#).

Graded for **completion only**.

We'll discuss solutions at **10:45**.

Quiz #1 from last week: if time. Solutions are in these slides.

- Quiz #2: problems and solutions
- Quiz #1: problems and solutions (if time)

# IN-CLASS QUIZ #2

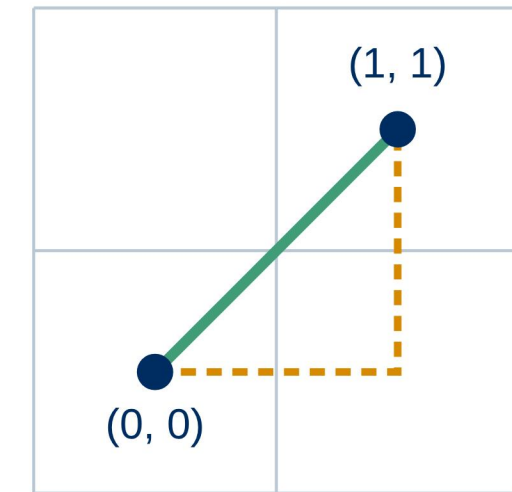
Problems and solutions



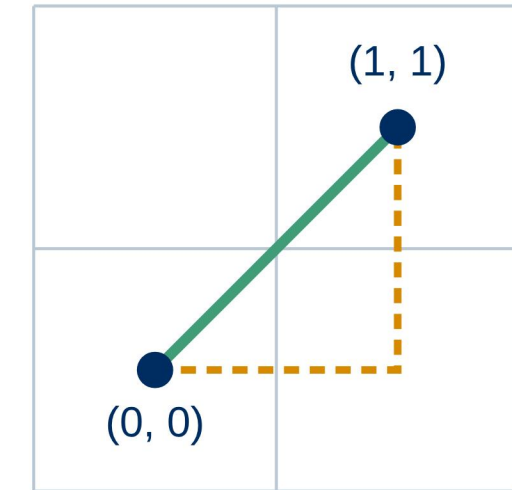
UNIVERSITY OF  
MICHIGAN

## Is Manhattan distance admissible?

1. **Q1.1:** Every move costs 1.
2. **Q1.2:** Diagonals cost 3; horizontal/vertical moves cost 1.



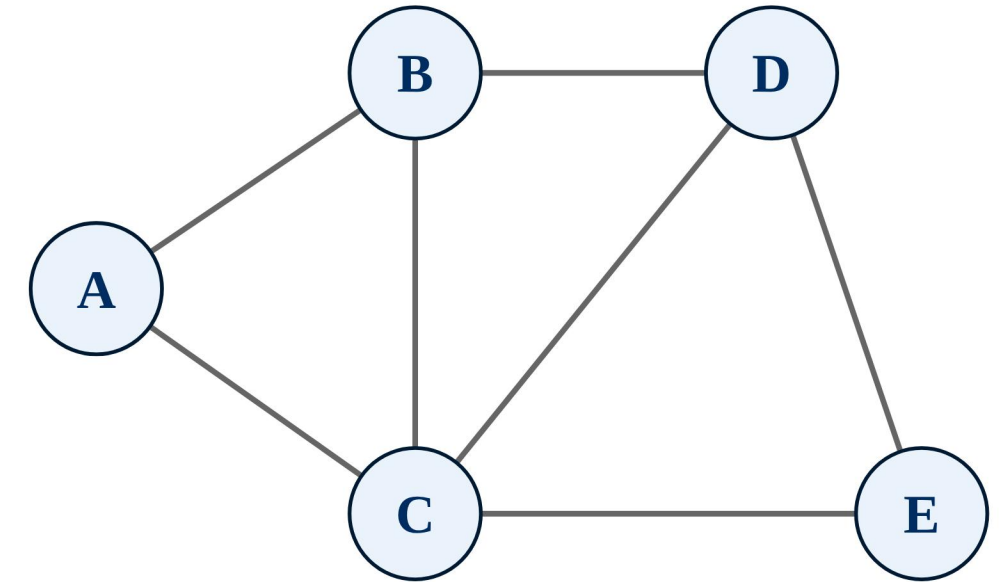
Relaxation: ignoring obstacles cannot increase the cost.



## Q2: CSPs

Schedule exams  $A$ – $E$ . Connected exams (see graph) need different time slots.

1. **Q2.1:** Set  $A = 1$ . Which unassigned variables does forward checking update?
2. **Q2.2:** What is their new domain?
3. **Q2.3:** Next choose  $C = 3$  (LCV). Finish with forward checking and MRV.



Slots:  $\{1, 2, 3\}$ . Also,  $E \neq 3$ .

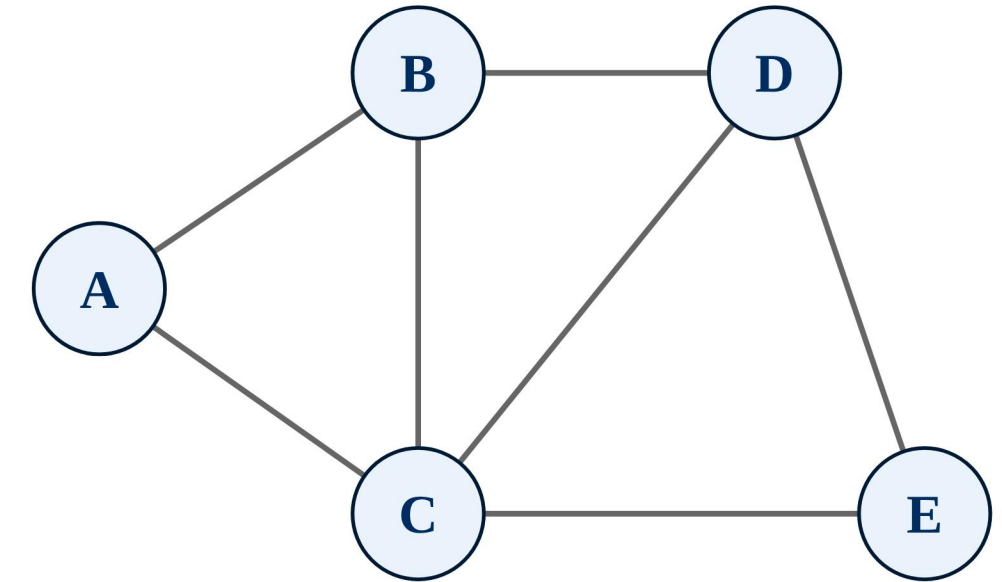
**Q2.1–2:**  $B$  and  $C$  get domain  $\{2, 3\}$ .

| Assign  | $A$ | $B$        | $C$        | $D$           | $E$        |
|---------|-----|------------|------------|---------------|------------|
| $A = 1$ | 1   | $\{2, 3\}$ | $\{2, 3\}$ | $\{1, 2, 3\}$ | $\{1, 2\}$ |
| $C = 3$ | 1   | $\{2\}$    | 3          | $\{1, 2\}$    | $\{1, 2\}$ |
| $B = 2$ | 1   | 2          | 3          | $\{1\}$       | $\{1, 2\}$ |
| $D = 1$ | 1   | 2          | 3          | 1             | $\{2\}$    |
| $E = 2$ | 1   | 2          | 3          | 1             | 2          |

MRV:  $B$ , then  $D$ , then  $E$ . Each has one value left.

## Q2.3: 12312.

$C = 3$  is chosen by LCV: removes 2 neighbor values, whereas  $C = 2$  removes 3.



Slots:  $\{1, 2, 3\}$ . Also,  $E \neq 3$ .

## Q3: DPLL

1. **Q3.1:** Which symbol is pure? Why can we set it without branching?
2. **Q3.2:** Then try  $P = 1$ . Which clause becomes unit? Propagate.
3. **Q3.3:** Backtrack; try  $P = 0$ . Which clause becomes unit? Propagate.
4. **Q3.4:** Resolve (1) and (2). What follows?

- (1)  $P \vee Q$
- (2)  $\neg P \vee Q$
- (3)  $\neg Q \vee R$
- (4)  $\neg Q \vee \neg R$
- (5)  $\neg R \vee S \vee P$

## Q3: solution

**Q3.1:**  $S$  is pure. Set  $S = 1$ .

Satisfies (5), cannot falsify another clause.

|             | Q3.2: $P = 1$            | Q3.3: $P = 0$            |
|-------------|--------------------------|--------------------------|
| Satisfied   | Clause (1)               | Clause (2)               |
| Unit clause | <b>(2)</b> becomes $Q$   | <b>(1)</b> becomes $Q$   |
| Set $Q = 1$ | (3): $R$ ; (4): $\neg R$ | (3): $R$ ; (4): $\neg R$ |

Propagate  $R$ : clause (4) becomes empty.

## Both branches fail: UNSAT.

- (1)  $P \vee Q$
- (2)  $\neg P \vee Q$
- (3)  $\neg Q \vee R$
- (4)  $\neg Q \vee \neg R$
- (5)  $\neg R \vee S \vee P$

Resolve (1) and (2) on  $P$ :

$$\frac{P \vee Q \quad \neg P \vee Q}{Q}$$

**Answer: Q.**

Resolve (3) and (4) on  $R$ :

$$\frac{\neg Q \vee R \quad \neg Q \vee \neg R}{\neg Q}$$

$Q$  and  $\neg Q$  resolve to the **empty clause**.

- (1)  $P \vee Q$
- (2)  $\neg P \vee Q$
- (3)  $\neg Q \vee R$
- (4)  $\neg Q \vee \neg R$
- (5)  $\neg R \vee S \vee P$





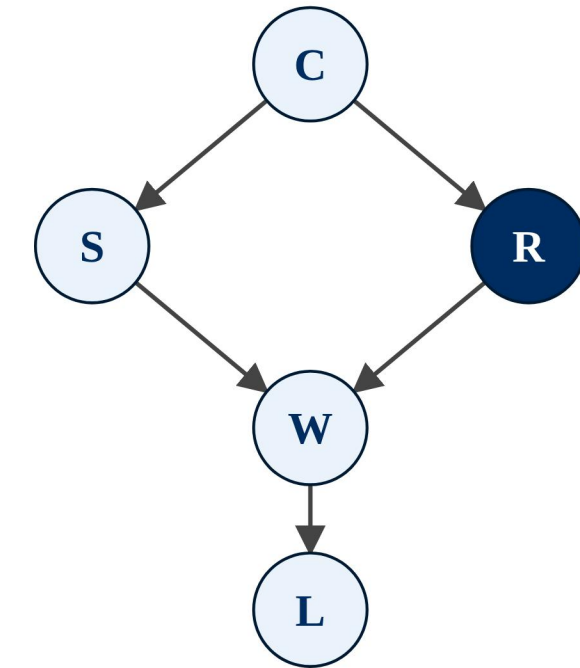
UNIVERSITY OF  
MICHIGAN

- Parent:  $C$ .
- Child:  $W$ .
- Child's other parent:  $S$ .

$$\text{MB}(R) = \{C, S, W\}$$

The only variable outside  $R$  and its blanket is  $L$ :

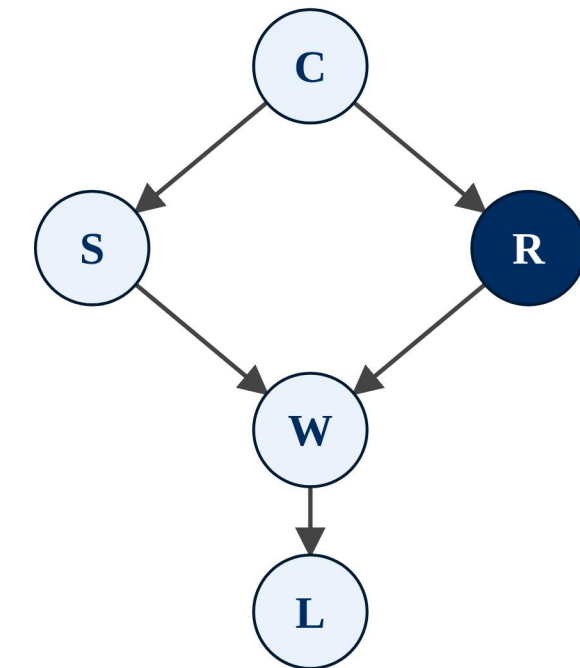
$$R \perp L \mid C, S, W$$



$C$ : cloudy ·  $S$ : sprinkler ·  $R$ : rain  
 $W$ : wet grass ·  $L$ : slippery lawn

# Q4: inference

Compute  $P(R = 1 \mid S = 1)$  as a fraction.  
Why don't we need the CPTs of  $W$  and  $L$ ?  
Compare with the prior  $P(R = 1)$ .



$C$ : cloudy ·  $S$ : sprinkler ·  $R$ : rain

$W$ : wet grass ·  $L$ : slippery lawn

$$P(C = 1) = \frac{1}{2}$$

| $c$ | $P(S = 1 \mid c)$ | $P(R = 1 \mid c)$ |
|-----|-------------------|-------------------|
| 1   | 1/4               | 3/4               |
| 0   | 3/4               | 1/4               |

# Q4: marginalization

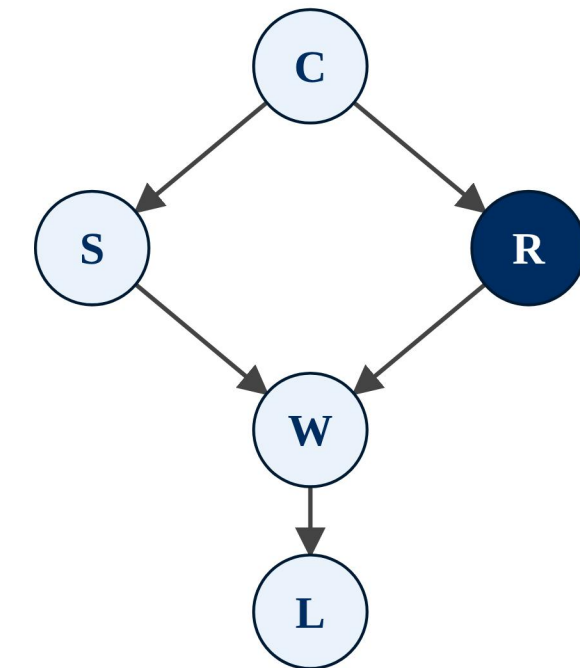
Neither  $W$  nor  $L$  is observed or queried.

Sum out  $L$ , then  $W$ :

$$\sum_l P(l \mid w) = 1$$
$$\sum_w P(w \mid s, r) = 1$$

Only the factors for  $C$ ,  $S$ , and  $R$  remain:

$$P(r, s) = \sum_c P(c)P(s \mid c)P(r \mid c)$$



$C$ : cloudy ·  $S$ : sprinkler ·  $R$ : rain

$W$ : wet grass ·  $L$ : slippery lawn

$$P(C = 1) = \frac{1}{2}$$

| $c$ | $P(S = 1 \mid c)$ | $P(R = 1 \mid c)$ |
|-----|-------------------|-------------------|
| 1   | 1/4               | 3/4               |
| 0   | 3/4               | 1/4               |

Sum over  $C$ :

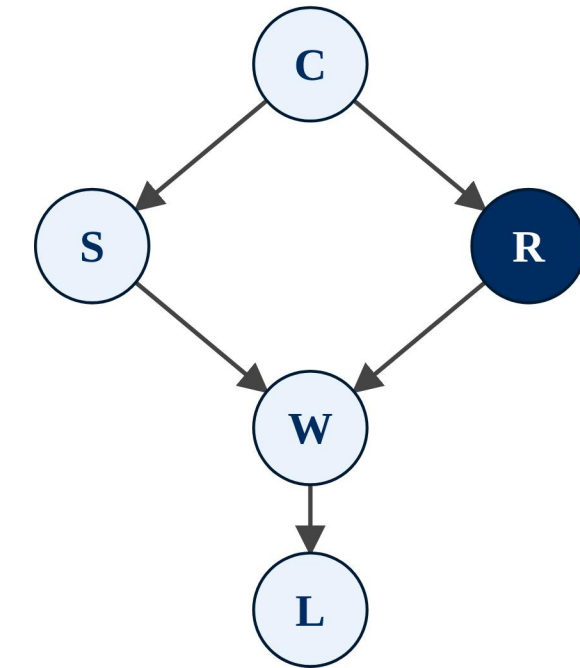
$$P(R = 1, S = 1) = \frac{1}{2} \cdot \frac{1}{4} \cdot \frac{3}{4} + \frac{1}{2} \cdot \frac{3}{4} \cdot \frac{1}{4} = \frac{3}{16}$$

$$P(S = 1) = \frac{1}{2} \cdot \frac{1}{4} + \frac{1}{2} \cdot \frac{3}{4} = \frac{1}{2}$$

$$P(R = 1 \mid S = 1) = \frac{3/16}{1/2} = \boxed{\frac{3}{8}}$$

$$P(R = 1) = \frac{1}{2} \cdot \frac{3}{4} + \frac{1}{2} \cdot \frac{1}{4} = \frac{1}{2}$$

Sprinkler on  $\rightarrow$  cloudy less likely  $\rightarrow$  rain less likely.



$C$ : cloudy ·  $S$ : sprinkler ·  $R$ : rain

$W$ : wet grass  $\cdot L$ : slippery lawn

$$P(C = 1) = \frac{1}{2}$$

| $c$ | $P(S = 1 \mid c)$ | $P(R = 1 \mid c)$ |
|-----|-------------------|-------------------|
| 1   | 1/4               | 3/4               |
| 0   | 3/4               | 1/4               |

# Q5: SGD

Start each example from  $w = 1$ .

1. **Q5.1:**  $(x, y) = (2, 1)$ . Correct prediction? Does SGD increase or decrease  $w$ ?
2. **Q5.2:** Same questions for  $(x, y) = (-2, 1)$ .
3. **Q5.3:** Which gives a larger update, with the same  $\eta$ ?

$$w = 1, \eta > 0.$$

$$p = \sigma(wx)$$

$$\text{Class 1 if } p \geq \frac{1}{2}.$$

$$L = -y \log p - (1 - y) \log(1 - p)$$

$$g = \frac{\partial L}{\partial w} = (p - y)x$$

$$w' = w - \eta g$$

$$\sigma(0) = \frac{1}{2}, \sigma \text{ increasing.}$$

$$\sigma(-z) = 1 - \sigma(z).$$

# Q5: gradients

| $(x, y)$   | $(2, 1)$                       | $(-2, 1)$                      |
|------------|--------------------------------|--------------------------------|
| $p$        | $\sigma(2) > 1/2$              | $\sigma(-2) < 1/2$             |
| Prediction | 1: <b>correct</b>              | 0: <b>incorrect</b>            |
| Gradient   | $2\sigma(2) - 2 < 0$           | $2 - 2\sigma(-2) > 0$          |
| Update     | <b>Increase <math>w</math></b> | <b>Decrease <math>w</math></b> |

A correct prediction still has a gradient ( $p < 1$  because of sigmoid).

$$w = 1, \eta > 0.$$

$$p = \sigma(wx)$$

Class 1 if  $p \geq \frac{1}{2}$ .

$$L = -y \log p - (1 - y) \log(1 - p)$$

$$g = \frac{\partial L}{\partial w} = (p - y)x$$

$$w' = w - \eta g$$

$$\sigma(0) = \frac{1}{2}, \sigma \text{ increasing.}$$

$$\sigma(-z) = 1 - \sigma(z).$$

# Q5: magnitudes

**Q5.3:**  $\Delta w = -\eta g$ , so  $|\Delta w| = \eta|g|$ .

For  $(2, 1)$ , the gradient is negative:

$$|\Delta w_+| = \eta|2\sigma(2) - 2| = 2\eta[1 - \sigma(2)]$$

For  $(-2, 1)$ , the gradient is positive:

$$|\Delta w_-| = \eta|2 - 2\sigma(-2)| = 2\eta[1 - \sigma(-2)]$$

Using  $1 - \sigma(z) = \sigma(-z)$ :

$$|\Delta w_+| = 2\eta\sigma(-2) < 2\eta\sigma(2) = |\Delta w_-|$$

**$(-2, 1)$  gives the larger update.**

$$w = 1, \eta > 0.$$

$$p = \sigma(wx)$$

Class 1 if  $p \geq \frac{1}{2}$ .

$$L = -y \log p - (1 - y) \log(1 - p)$$

$$g = \frac{\partial L}{\partial w} = (p - y)x$$

$$w' = w - \eta g$$

$\sigma(0) = \frac{1}{2}$ ,  $\sigma$  increasing.

$\sigma(-z) = 1 - \sigma(z)$ .



UNIVERSITY OF  
MICHIGAN

1. **Q6.1:** Compute  $c_1, c_2, c_3$  and  $z$ .
2. **Q6.2:** How many learnable parameters?
3. Compare with a fully connected layer: 4 inputs, 3 outputs, no bias.

$$x = (1, 2, 0, -1)$$

$$k = (1, -1)$$

Stride 1; no padding or bias.

$$c_i = k_1 x_i + k_2 x_{i+1}$$

$$i = 1, 2, 3$$

$$r_i = \max(0, c_i)$$

$$z = r_1 + r_2 + r_3$$

The final sum has fixed weights.

| $i$ | Window    | $c_i$          | $r_i$ |
|-----|-----------|----------------|-------|
| 1   | $(1, 2)$  | $1 - 2 = -1$   | 0     |
| 2   | $(2, 0)$  | $2 - 0 = 2$    | 2     |
| 3   | $(0, -1)$ | $0 - (-1) = 1$ | 1     |

$$c = (-1, 2, 1), \quad r = (0, 2, 1), \quad \boxed{z = 3}$$

ReLU **before** summing: the negative activation contributes 0.

$$x = (1, 2, 0, -1)$$

$$k = (1, -1)$$

Stride 1; no padding or bias.

$$c_i = k_1 x_i + k_2 x_{i+1}$$

$$i = 1, 2, 3$$

$$r_i = \max(0, c_i)$$

$$z = r_1 + r_2 + r_3$$

The final sum has fixed weights.

# Q6: parameters

Convolution here is equivalent to multiplying by this matrix:

$$\begin{bmatrix} k_1 & k_2 & 0 & 0 \\ 0 & k_1 & k_2 & 0 \\ 0 & 0 & k_1 & k_2 \end{bmatrix}$$

**CNN: 2 parameters**,  $k_1$  and  $k_2$ , reused at each position.

ReLU and the sum: no parameters.

**Fully connected:**  $3 \times 4 = 12$ . Each entry is independent.

$$x = (1, 2, 0, -1)$$

$$k = (1, -1)$$

Stride 1; no padding or bias.

$$c_i = k_1 x_i + k_2 x_{i+1}$$

$$i = 1, 2, 3$$

$$r_i = \max(0, c_i)$$

$$z = r_1 + r_2 + r_3$$

The final sum has fixed weights.

# IN-CLASS QUIZ #1

# Q1: receptive fields

Consider an activation near the center of layer 4.

## ? QUESTION

What size region of the input can affect it?

Answer as  $N \times N$ .

Layer 4 is after **three convolutions**.

$100 \times 100$  input.

$5 \times 5$  kernels, stride 1.

No pooling.

Input · Layer 1



Conv · Layer 2



Conv · Layer 3



Conv · Layer 4

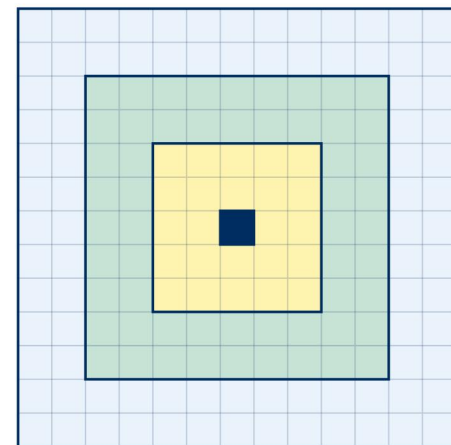
# Q1: solution

Start with  $r_0 = 1$ . Stride 1 gives:

$$r_{\ell+1} = r_{\ell} + (5 - 1)$$

$$1 \rightarrow 5 \rightarrow 9 \rightarrow 13$$

Same vertically:  $13 \times 13$ .



1 → 5 → 9 → 13 pixels per side

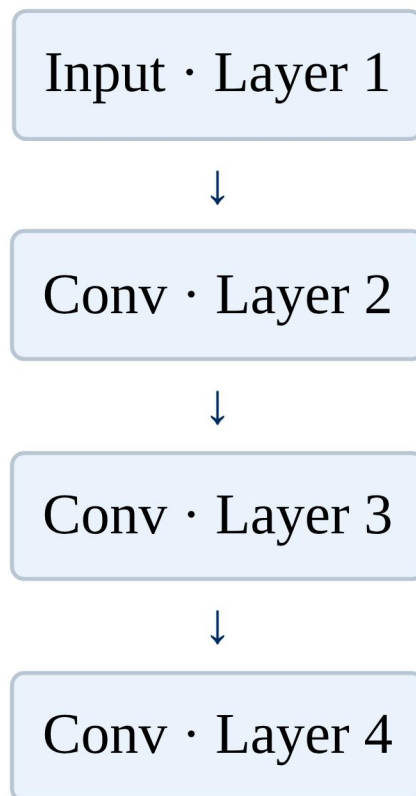
Interpreting the diagram: the dark blue cell (layer 4) is computed from convolution over the yellow region.

If you center the filter at each yellow cell, you get the green region (think "under" yellow), etc...

$100 \times 100$  input.

$5 \times 5$  kernels, stride 1.

No pooling.



# Q2: backprop

A scalar network with an “inverted ReLU”.

1. **Q2.1:** Find  $\partial L / \partial w$ .
2. **Q2.2:** Find  $\partial L / \partial b$ .
3. **Q2.3:** Which  $\eta$  gives a perfect prediction after one SGD step?

Update both parameters from their starting values.

$$w = 2, b = -1$$

$$(x, y) = (-1, -1)$$

$$z = wx + b$$

$$h = \min(0, z)$$

$$L = (h - y)^2$$

# Q2: gradients

Forward pass:

$$z = 2(-1) - 1 = -3, \quad h = -3, \quad L = 4$$

$z < 0$ , so  $\partial h / \partial z = 1$ .

$$\frac{\partial L}{\partial z} = 2(h - y) \cdot 1 = -4$$

$$\frac{\partial L}{\partial w} = \frac{\partial L}{\partial z} \cdot x = (-4)(-1) = \boxed{4}$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial z} \cdot 1 = \boxed{-4}$$

$$w = 2, b = -1$$

$$(x, y) = (-1, -1)$$

$$z = wx + b$$

$$h = \min(0, z)$$

$$L = (h - y)^2$$

## Q2: step size

Apply the gradients  $(4, -4)$ :

$$w' = 2 - 4\eta, \quad b' = -1 + 4\eta$$

$$z' = (2 - 4\eta)(-1) + (-1 + 4\eta) = -3 + 8\eta$$

To get  $h' = -1$ , we need  $z' = -1$ .

$$-3 + 8\eta = -1 \quad \implies \quad \boxed{\eta = \frac{1}{4}}$$

Double-checking:  $w' = 1, b' = 0 \rightarrow h' = -1, L' = 0$ .

$$w = 2, b = -1$$

$$(x, y) = (-1, -1)$$

$$z = wx + b$$

$$h = \min(0, z)$$

$$L = (h - y)^2$$

# Q3: initial loss

## ? QUESTION

Before training, what average per-token loss do we expect?

Transformer: 1B parameters.

Vocabulary:  $V = 100,000$ .

Per-token cross-entropy:

$$L = -\log p(\text{correct token})$$

Standard is natural log.

# Q3: solution

Randomly initialized model will give approximately uniform predictions, so:

$$p(\text{correct token}) \approx \frac{1}{V}$$

$$L \approx \log V = \log(100,000) \approx \boxed{11.51}$$

("nats per token", if you want a unit for this loss)

Depends on vocabulary size, not parameter count.

Transformer: 1B parameters.

Vocabulary:  $V = 100,000$ .

Per-token cross-entropy:

$$L = -\log p(\text{correct token})$$

Standard is natural log.