DS 207: Introduction to Natural Language Processing

INDIAN INSTITUTE OF SCIENCE

भारतीय विज्ञान संस्थान

Text Classification

Danish Pruthi

Slides courtesy: Graham Neubig

Example: topic classification

- Sports: "Kohli scores another remarkable hundred"
- Politics: "Minister announces new metro plans ahead of elections"
- Entertainment: "A sleeper hit, 12th fail, is praised by many Bollywood actors"
- Finance: "Stocks for Indian delivery startups plummet"

Generative Naive Bayes

$$P(X^{(i)}, y^{(i)}) = P(y^{(i)})\ P(X^{(i)} | y^{(i)})$$

$$P(X | y) = P(w_1, w_2, w_3, \dots w_t | y)$$

$$P(X | y) = \prod_{i}^{t} P(w_i | y)$$

$$P(w_i = \text{" Kohli"} | y = \text{"sports"})$$

Estimating parameters

$$\mathbf{P}(w_i = \text{"Kohli"} \mid y = \text{"sports"})$$

$$= \frac{\text{count}(w_i = \text{Kohli} \in y = \text{sports})}{\sum_{w \in |V|} \text{count}(w \in y = \text{sports})}$$

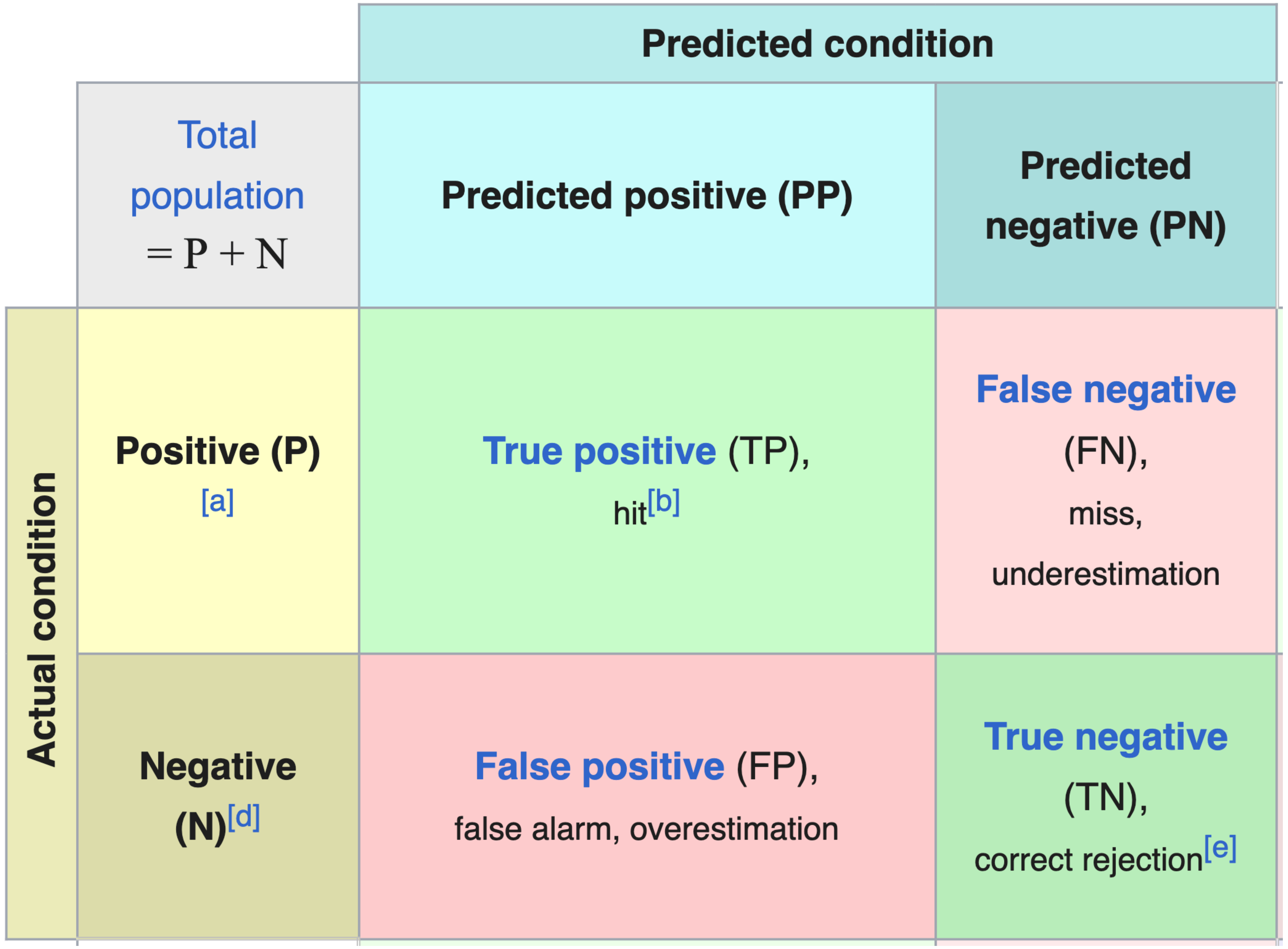
Add-α smoothing (or Laplace smoothing)

$$\mathbf{P}(\mathrm{w_i} = \text{"Kohli"} \mid y = \text{"finance"})$$

$$= \frac{\mathbf{count}(w_i = \mathbf{Kohli} \in y = \mathbf{finance}) + \alpha}{\sum_{w \in |V|} (\mathbf{count}(w \in y = \mathbf{finance}) + \alpha)}$$

Evaluation

- Accuracy $= \frac{TP + TN}{P + N}$
- Precision = TP / (Predicted) P
- Recall = TP / (Actual) P
- F1 score



		Predicted condition	
	Total population = P + N	Predicted positive (PP)	Predicted negative (PN)
Actual condition	Positive (P) [a]	**True positive** (TP), hit[b]	**False negative** (FN), miss, underestimation
	Negative (N)[d]	**False positive** (FP), false alarm, overestimation	**True negative** (TN), correct rejection[e]

What metrics are best suited

- Diagnosing rare type of cancer
- Criminal punishment
- Detecting spam
- Recruitment/Filtering based on text in the CVs
- Recommending products/songs/movies

Generative vs Discriminative models

- Generative model: a model that calculates the probability of the input data itself

$$P(X) \qquad P(X, Y)$$

stand-alone *joint*

- Discriminative model calculates the probability of a class (or trait) given the data

$$P(Y \mid X)$$

conditional

Rule based systems: is the headline sports or entertainment?

- **Feature extraction**: Extract the salient features for making the predictions from text

  1. Does the headline contain the word "Kohli"
  2. Does the headline contain the word "win"
  3. Does the headline contain the word "cricket"
  4. Does the headline contain the word "actor"
  5. Does the headline contain the word "show"
  6. Does the headline contain the word "blockbuster"

Rule based systems: is the headline sports or entertainment?

- **Feature extraction:** Extract the salient features for making the predictions from text

1. Does the headline contain the word "Kohli"
2. Does the headline contain the word "win"
3. Does the headline contain the word "cricket"
4. Does the headline contain the word "actor"
5. Does the headline contain the word "show"
6. Does the headline contain the word "blockbuster"

$$f(x) = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Kohli scores a ton, India win.

Rule based systems: is the headline sports or entertainment?

- **Feature extraction**: Extract the salient features for making the predictions from text

1. Does the headline contain the word "Kohli"
2. Does the headline contain the word "win"
3. Does the headline contain the word "cricket"
4. Does the headline contain the word "actor"
5. Does the headline contain the word "show"
6. Does the headline contain the word "blockbuster"

$$f(x) = \boxed{\begin{matrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{matrix}} \qquad W = \boxed{\begin{matrix} +1 \\ +1 \\ +1 \\ -1 \\ -1 \\ -1 \end{matrix}}$$

A three step process for making predictions

- **Feature extraction:** Extract the salient features for making the decision from text

$$\mathbf{h} = f(\mathbf{x})$$

- **Score calculation:** Calculate a score for one or more possibilities

$$s = \mathbf{w} \cdot \mathbf{h} \qquad\qquad \mathbf{s} = W\mathbf{h}$$

binary *multi-class*

- **Decision function:** Choose one of the several possibilities

$$\hat{y} = \mathrm{decide}(\mathbf{s})$$

Another discriminative classifier

Another discriminative classifier

- **Feature Extraction:** $h = f(x)$

One hot vectors ("bag of words")

- h is very sparse.
- Order ignored.
- "count of words" instead of "bag of words" could be used.

Another discriminative classifier

- **Feature Extraction:** $h = f(x)$

  One hot vectors ("bag of words")

- **Score Calculation:** binary or multi-class

  $s = \mathbf{w} \cdot \mathbf{h}$ $\mathbf{s} = W\mathbf{h}$

Another discriminative classifier

- **Feature Extraction:** $h = f(x)$

  One hot vectors ("bag of words")

- **Score Calculation:** binary or multi-class

$$s = \mathbf{w} \cdot \mathbf{h} \qquad \mathbf{s} = W\mathbf{h}$$

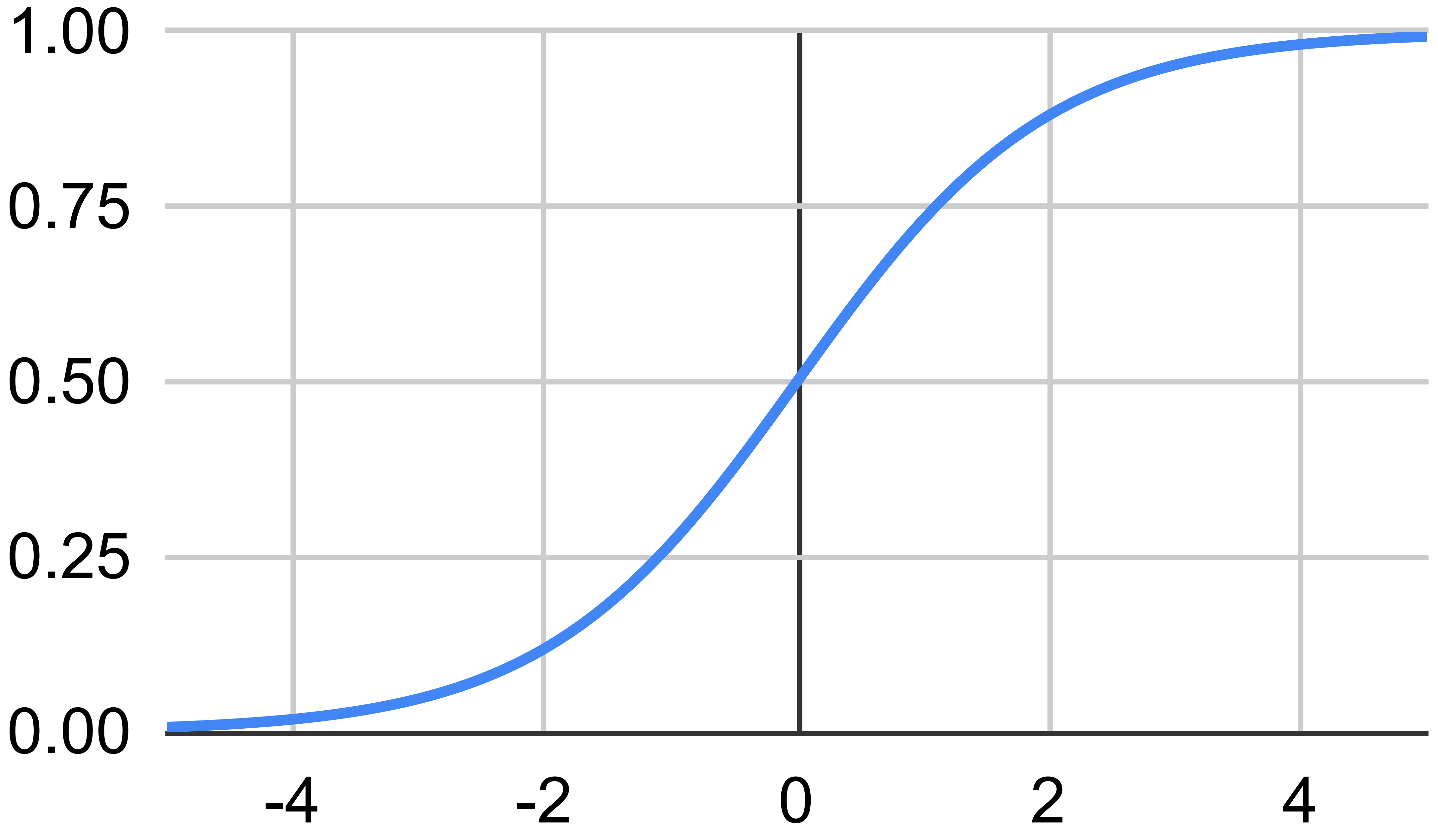
- **Decision:** Convert to a probability:

$$P(y|x) = \sigma(s) \text{ or softmax}(\mathbf{s})$$

Sigmoid function

- **Sigmoid can be used for binary decisions**

$$\sigma(s) = \frac{1}{1 + e^{-s}}$$



Softmax

- **Softmax is used for multi-label classification**

To make values positive.

$$\mathbf{softmax}(s) = \frac{\boxed{e^{s_i}}}{\sum_i^d e^{s_i}}$$

$$s = \begin{pmatrix} -3.2 \\ -2.9 \\ 1.0 \\ 2.2 \\ 0.6 \\ \dots \end{pmatrix} \longrightarrow p = \begin{pmatrix} 0.002 \\ 0.003 \\ 0.329 \\ 0.444 \\ 0.090 \\ \dots \end{pmatrix}$$

why not $\sigma(wx+b)$?

Softmax

- **Softmax is used for multi-label classification**

$$\textbf{soft}\underset{\uparrow}{\underline{\textbf{max}}}(s) = \frac{e^{s_i}}{\sum_i^d e^{s_i}} \qquad s = \begin{pmatrix} 1 \\ 6 \\ 2 \end{pmatrix} \longrightarrow p = \begin{pmatrix} 0.006 \\ 0.975 \\ 0.017 \end{pmatrix}$$

points to "max"

Recap: Discriminative models

- **Define a model that calculates probability directly based on parameters W**

$$P(Y|X;W)$$

Computing loss (or error/cost) function

Computing loss (or error/cost) function

Define a loss function (a function which is lower for better models):

Computing loss (or error/cost) function

Define a loss function (a function which is lower for better models):

Computing loss (or error/cost) function

Define a loss function (a function which is lower for better models):

$$L(W) = -\sum_{x, y \in D} \log P(y \mid x; W)$$ **[negative log likelihood]**

Train to find W what values of W gives the minimum loss ?

Computing loss (or error/cost) function

Define a loss function (a function which is lower for better models):

$$L(W) = -\sum_{x,\,y\ \in\ D} \log P(y\,|\,x; W)$$ **[negative log likelihood]**

Computing loss (or error/cost) function

Define a loss function (a function which is lower for better models):

$$L(W) = -\sum_{x,\, y \;\in\; D} \log P(y \,|\, x; W)$$ **[negative log likelihood]**

$$L(W) = -\sum_{i}^{n} \Big(y_i \;\log P(y = 1 \,|\, x_i; W) \quad + \quad (1 - y_i)\; \log(P(y = 0 \,|\, x_i; W) \Big)$$

[binary cross entropy]

Learn parameters through gradient descent

- Compute the gradient of the loss function with respect to the parameters

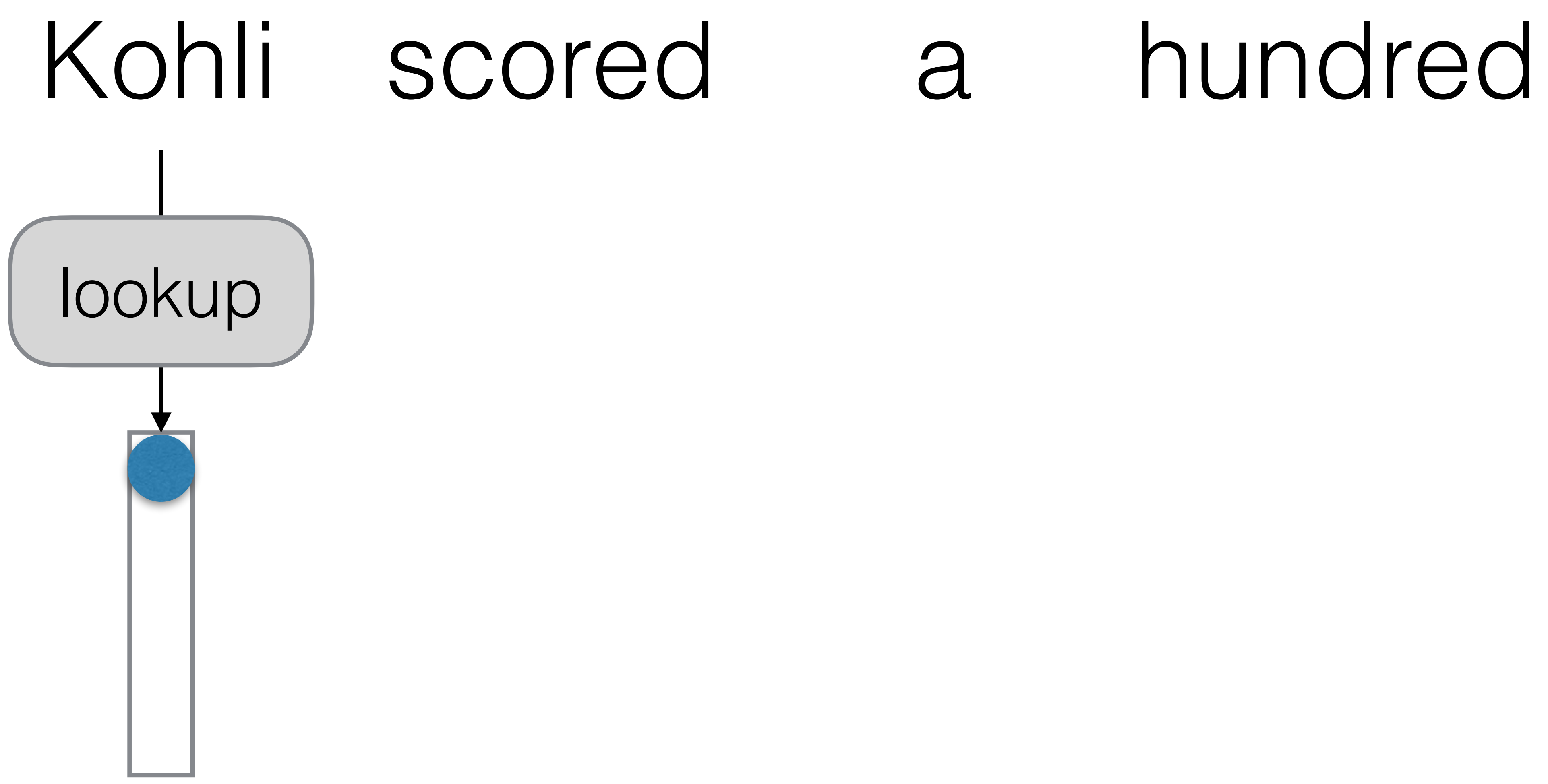
- Keep updating the parameters to move in a direction that decreases the loss

$$W_{t+1} = W_t - \alpha \nabla_{W_t} L(W)$$

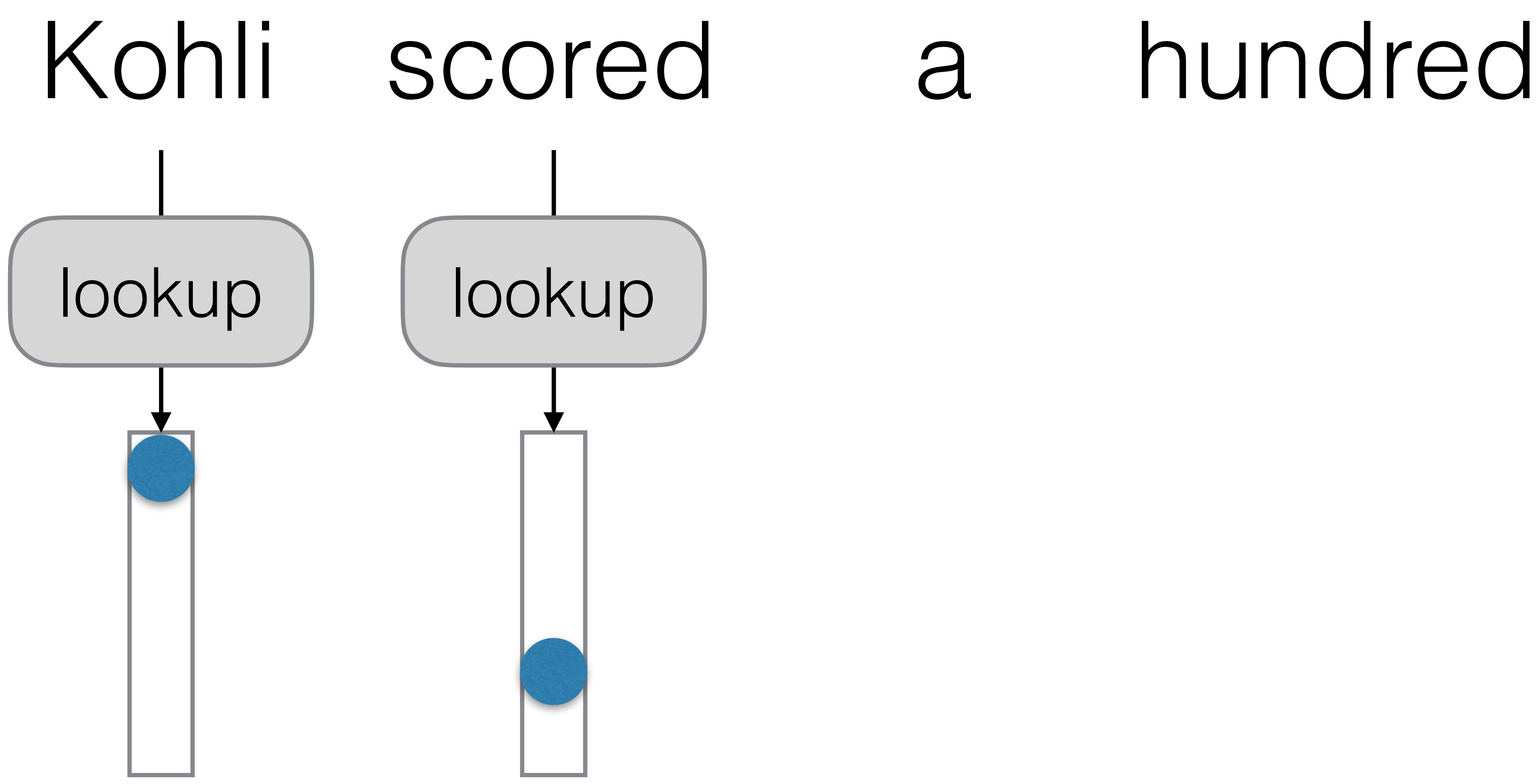
Computational Graph View

Kohli scored a hundred

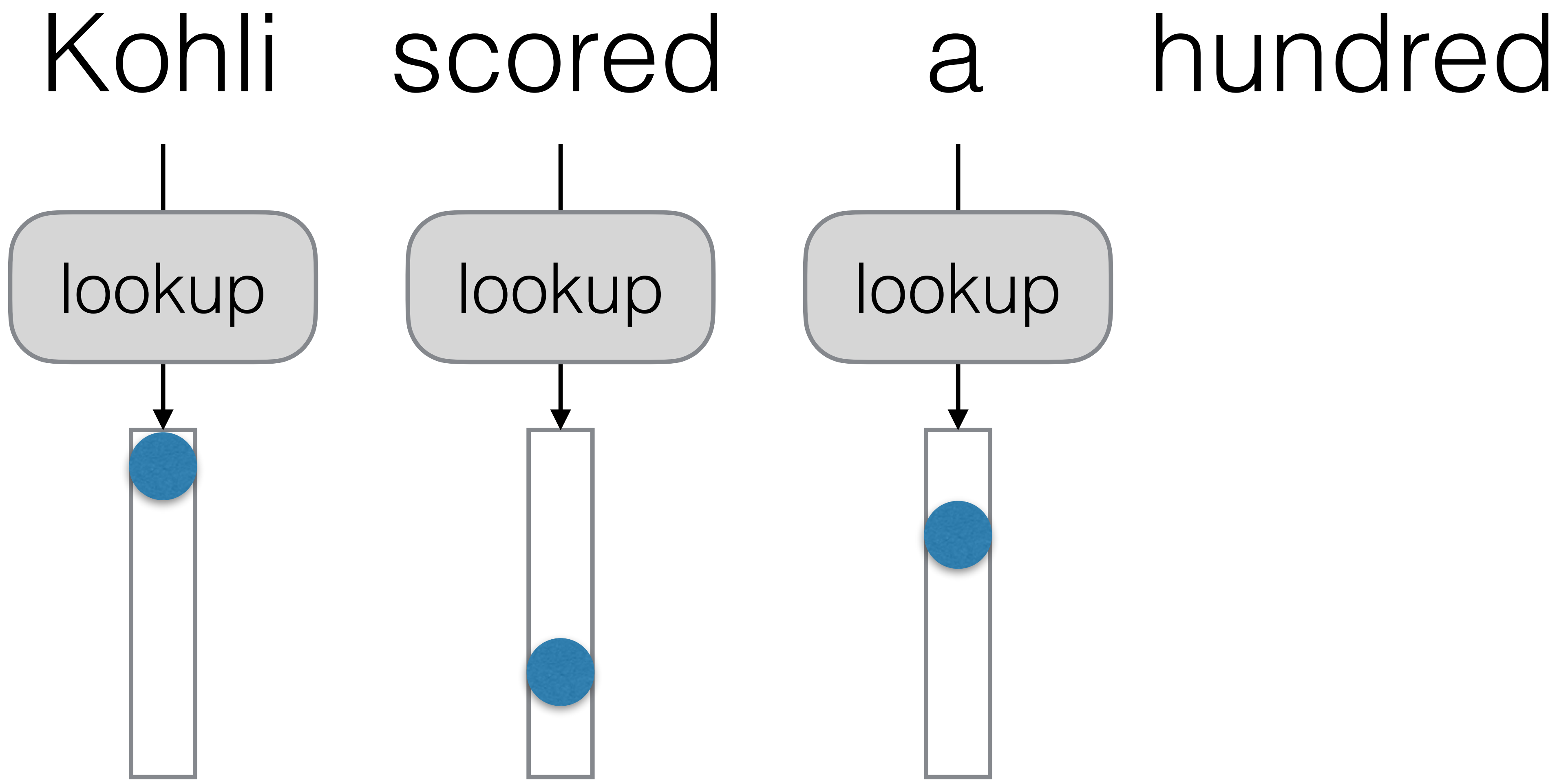
Computational Graph View



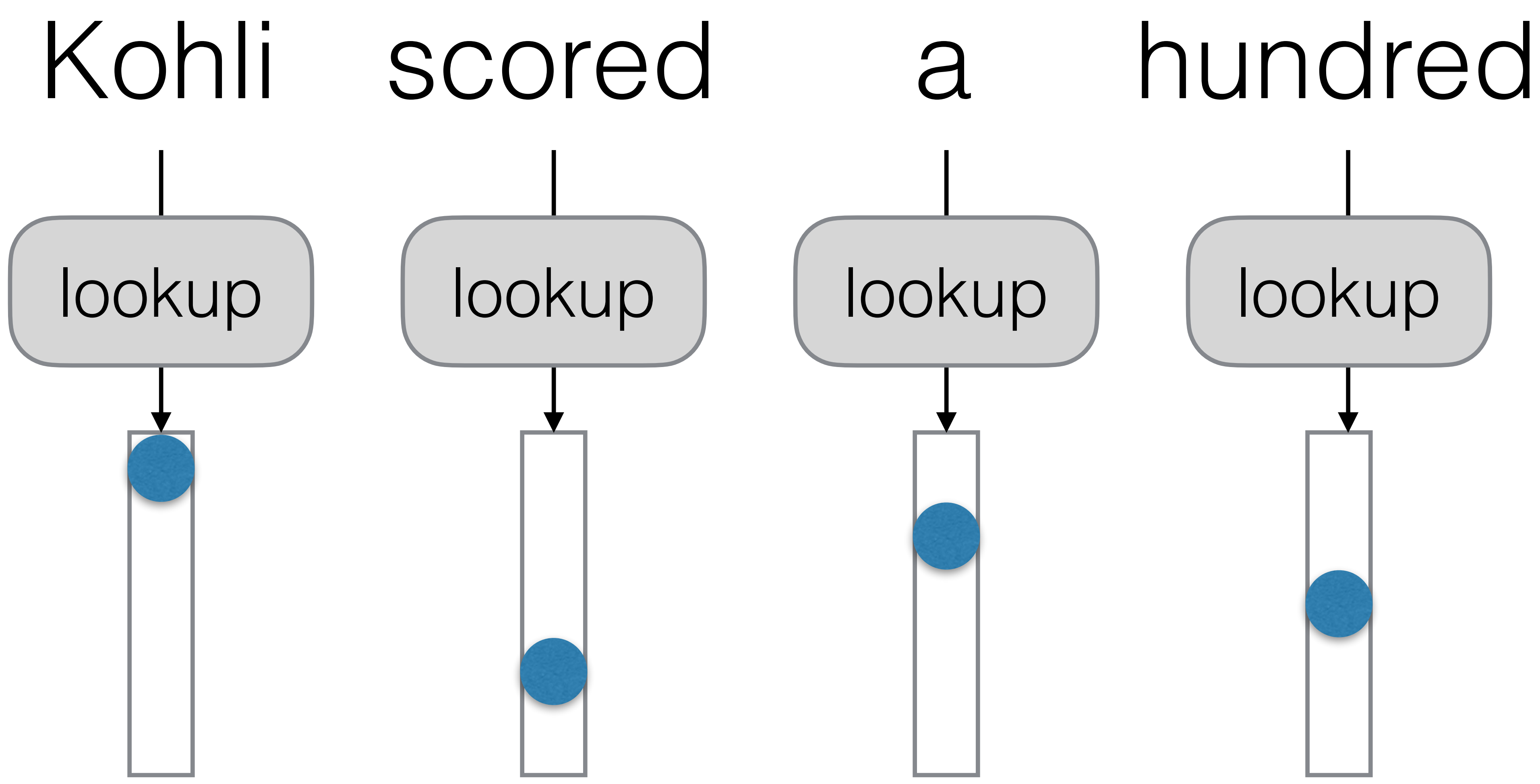
Computational Graph View



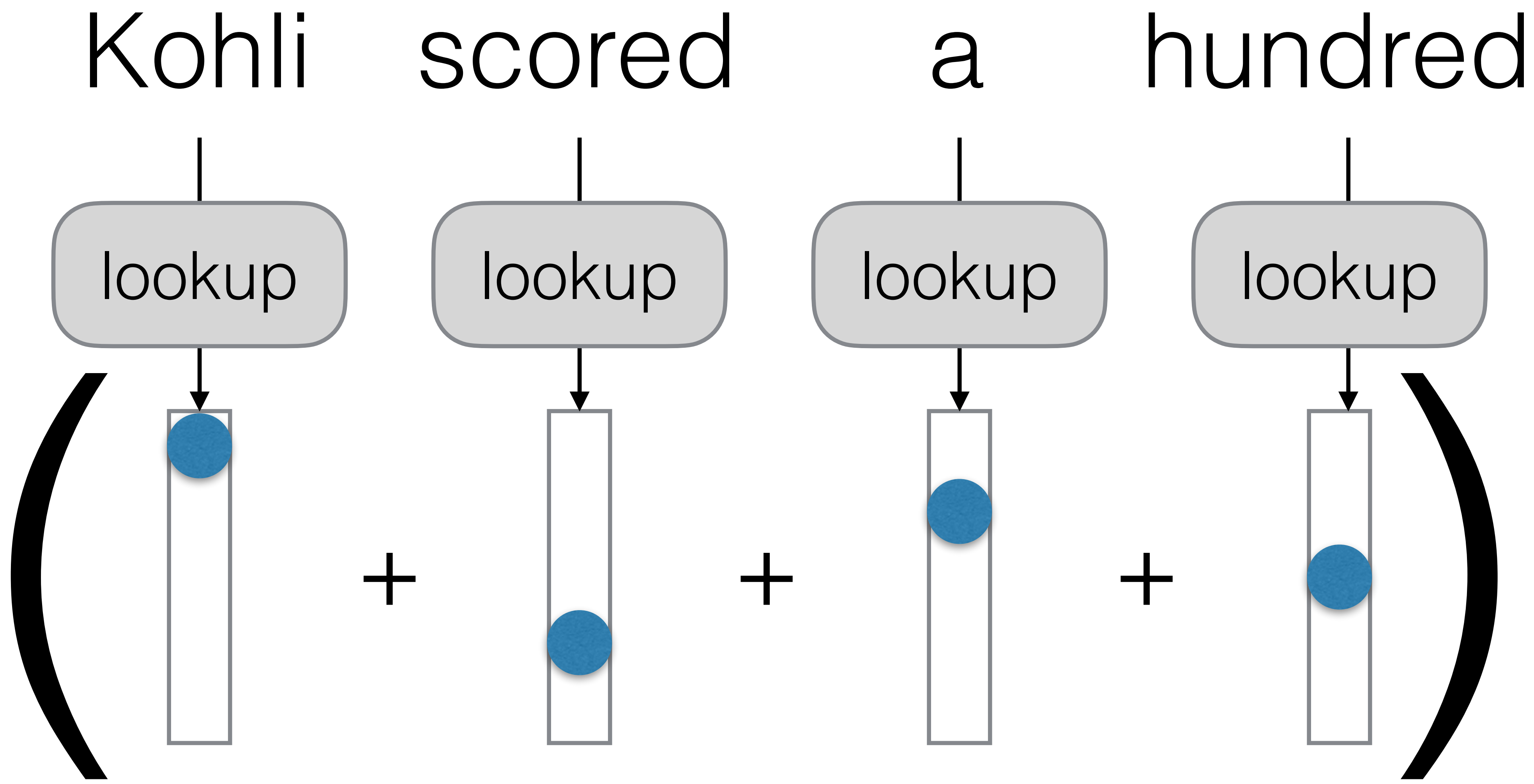
Computational Graph View



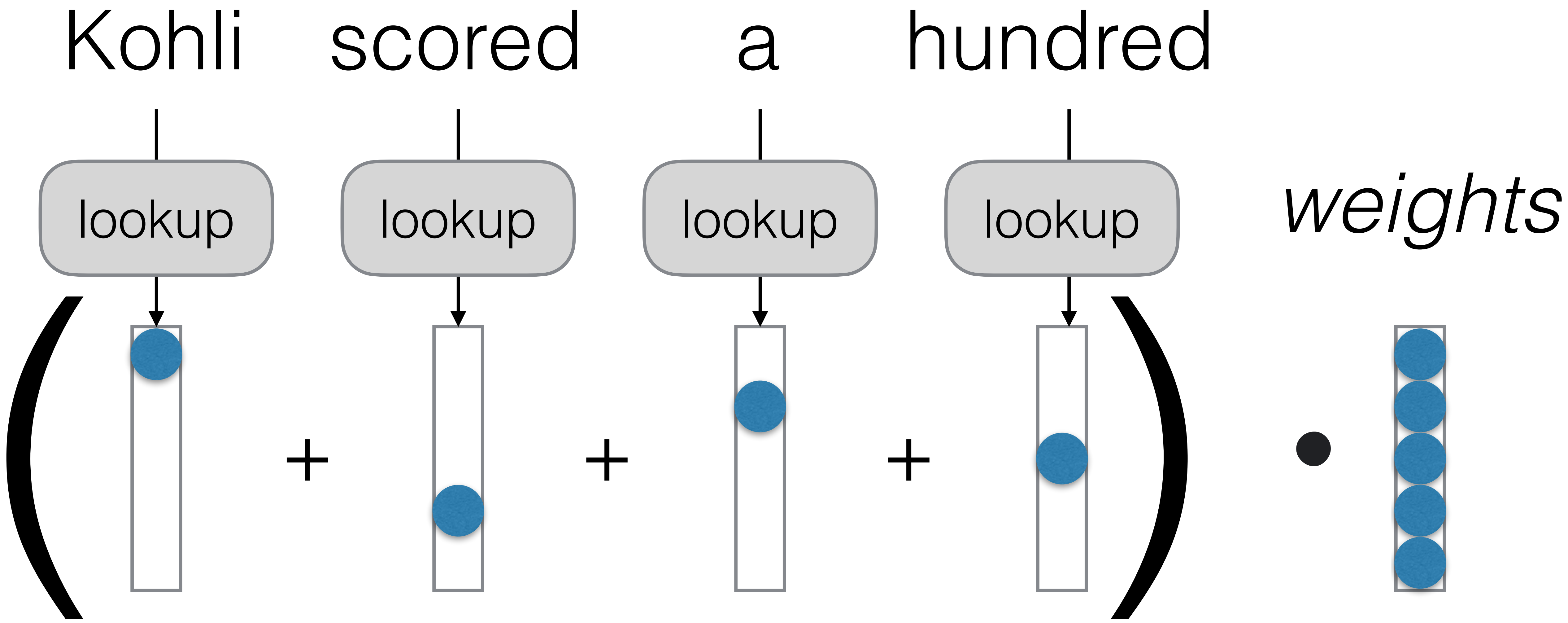
Computational Graph View



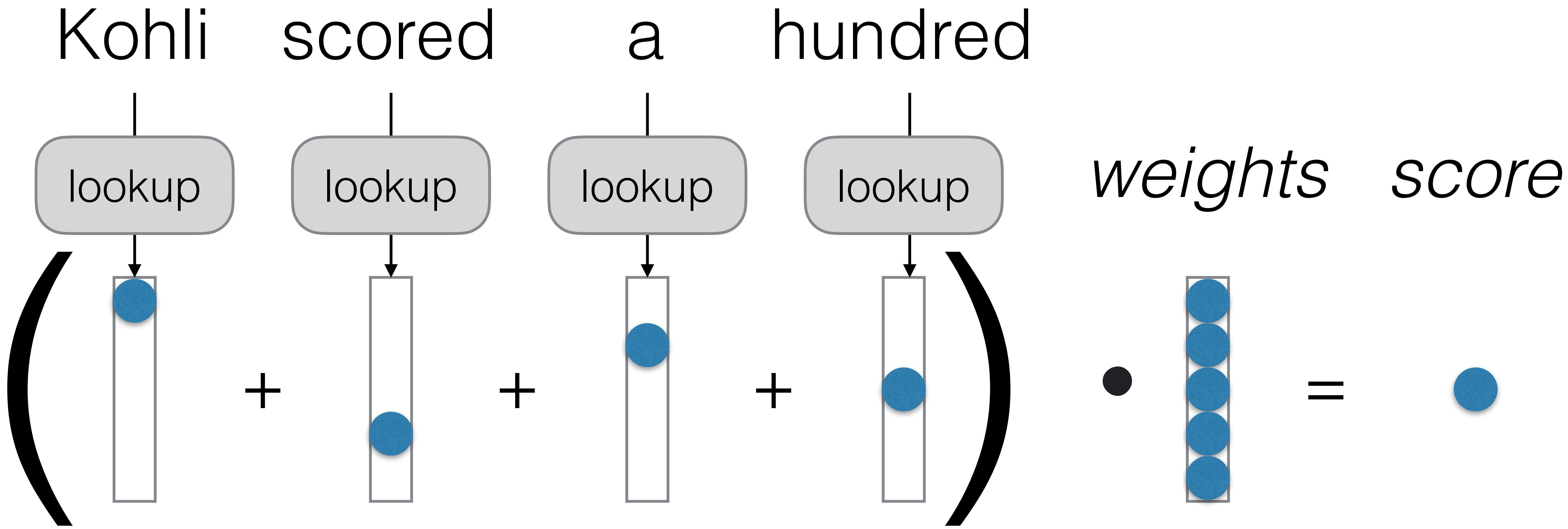
Computational Graph View



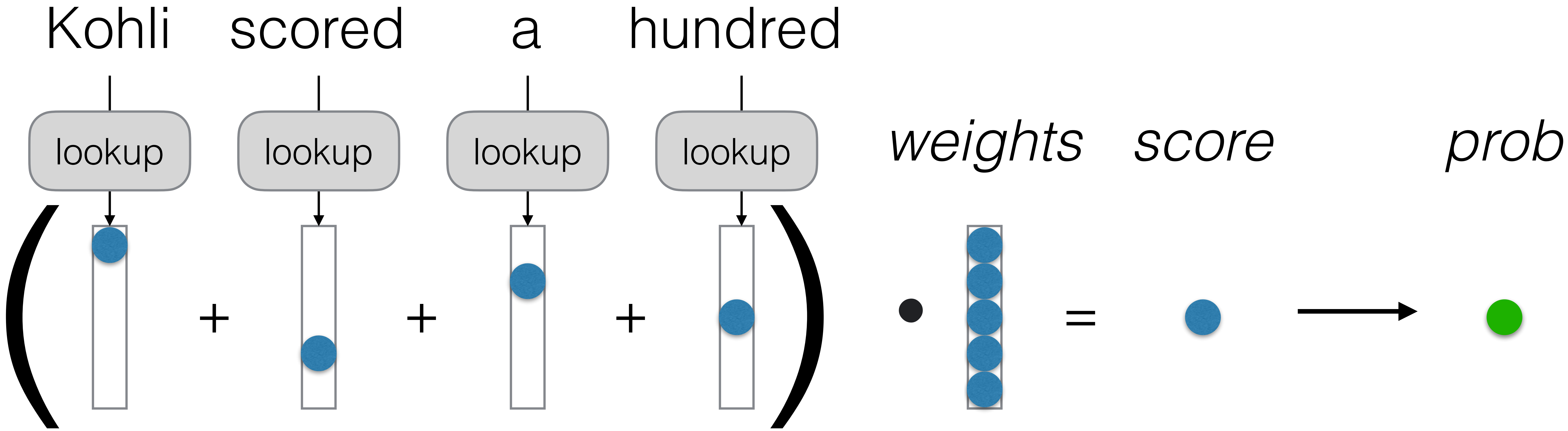
Computational Graph View



Computational Graph View



Computational Graph View



Questions?

Next class: Word2vec