CS447: Natural Language Processing

http://courses.grainger.illinois.edu/cs447

Lecture 07: Lexical Semantics

Julia Hockenmaier

juliahmr@illinois.edu

ILLINOIS

Lecture 6 Part 3: Training Logistic Regression Models with (Stochastic) Gradient Descent

$P(Y \mid \mathbf{X})$ with Logistic Regression: Binary Classification

Task: Model $P(y \in \{0,1\} \mid \mathbf{x})$
for any input (feature) vector $\mathbf{x} = (x_1, \ldots, x_n)$

Idea: Learn feature weights $\mathbf{w} = (w_1, \ldots, w_n)$ (and a bias term b)
to capture how important each feature x_i is for predicting $y = 1$

For **binary** classification ($y \in \{0,1\}$),
(standard) logistic regression uses the sigmoid function:

$$P(Y{=}1 \mid \mathbf{x}) = \sigma(\mathbf{wx} + b) = \frac{1}{1 + \exp(-(\mathbf{wx} + b))}$$

Parameters to learn: one feature weight vector $\mathbf{w}$ and one bias term b

Learning parameters **w** and *b*

Training objective: Find parameters **w** and *b* that "capture the training data D_{train} as well as possible"

More formally (and since we're being probabilistic):

Find **w** and *b* that assign the largest possible conditional probability to the labels of the items in D_{train}

$$(\mathbf{w}^*, b^*) = \text{argmax}_{(\mathbf{w},b)} \prod_{(\mathbf{x}_i, y_i) \in D_{train}} P(y_i \mid \mathbf{x}_i)$$

⇒ Maximize $P(1 \mid \mathbf{x}_i)$ for any ($\mathbf{x}_i$,1) with a *positive* label in D_{train}

⇒ Maximize $P(0 \mid \mathbf{x}_i)$ for any ($\mathbf{x}_i$,0) with a *negative* label in D_{train}

Since $y_i \in \{0,1\}$ we can rewrite this to:

$$(\mathbf{w}^{w}, b^*) = \text{argmax}_{(\mathbf{w},b)} \prod_{(\mathbf{x}_i, y_i) \in D_{train}} P(1 \mid \mathbf{x}_i)^{y_i} \cdot [1 - P(1 \mid \mathbf{x}_i)]^{1-y_i}$$

For $y_i = 1$, this comes out to: $P(1 \mid \mathbf{x}_i)^1 (1 - P(1 \mid \mathbf{x}_i))^0 = P(1 \mid \mathbf{x}_i)$

For $y_i = 0$, this is: $P(1 \mid \mathbf{x}_i)^0 (1 - P(1 \mid \mathbf{x}_i))^1 = 1 - P(1 \mid \mathbf{x}_i) = P(0 \mid \mathbf{x}_i)$

Learning = Optimization = Loss Minimization

Learning = parameter estimation = optimization:

Given a particular class of model (logistic regression, Naive Bayes, …) and data D_{train},
find **the *best* parameters** for this class of model on D_{train}

If the model is a probabilistic classifier, think of optimization as Maximum Likelihood Estimation (**MLE**)

*"Best" = return (among all possible parameters for models of this class) parameters that assign the **largest probability** to* D_{train}

In general (incl. for probabilistic classifiers), think of optimization as **Loss Minimization**:

*"Best" = return (among all possible parameters for models of this class) parameters that have the **smallest loss** on* D_{train}

"Loss": how bad are the predictions of a model?

*The **loss function** we use to measure loss depends on the class of model*
$L(\hat{y}, y)$: how bad is it to predict $\hat{y}$ if the correct label is y ?

Conditional MLE ⟹ Cross-Entropy Loss

Conditional MLE: *Maximize probability* of labels in D$_{train}$

$$(\mathbf{w}^*, b^*) = \text{argmax}_{(\mathbf{w},b)} \prod_{(\mathbf{x}_i, y_i)\in D_{train}} P(y_i \mid \mathbf{x}_i)$$

⇒ Maximize $P(1 \mid \mathbf{x}_i)$ for any ($\mathbf{x}_i$,1) with a *positive* label in D$_{train}$

⇒ Maximize $P(0 \mid \mathbf{x}_i)$ for any ($\mathbf{x}_i$,0) with a *negative* label in D$_{train}$

Equivalently: *Minimize negative log prob.* of correct labels in D$_{train}$

$P(y_i \mid \mathbf{x}) = 0 \Leftrightarrow -\log(P(y_i \mid \mathbf{x})) = +\infty$ if y_i is the correct label for $\mathbf{x}$, this is the worst possible model

$P(y_i \mid \mathbf{x}) = 1 \Leftrightarrow -\log(P(y_i \mid \mathbf{x})) = 0$ if y_i is the correct label for $\mathbf{x}$, this is the best possible model

The negative log probability of the correct label is a **loss** function:

$-\log(P(y_i \mid \mathbf{x}_i))$ is **smallest** (0) when we assign **all** probability to the **correct** label

$-\log(P(y_i \mid \mathbf{x}_i))$ is **largest** $(+\infty)$ when we assign **all** probability to the **wrong** label

This negative log likelihood loss is also called **cross-entropy loss**

From **loss** to per-example **cost**

Let's define the "**cost**" of our classifier on the whole dataset as its **average loss** on each of the *m* training examples:

$$\text{Cost}_{CE}(D_{\text{train}}) = \frac{1}{m} \sum_{i=1..m} -\log P(y_i \mid \mathbf{x}_i)$$

For each example:

$$-\log P(y_i \mid \mathbf{x}_i)$$

$$= -\log(\ P(1 \mid \mathbf{x}_i)^{y_i} \cdot P(0 \mid \mathbf{x}_i)^{1-y_i}\ )$$

[either $y_i = 1$ or $y_i = 0$]

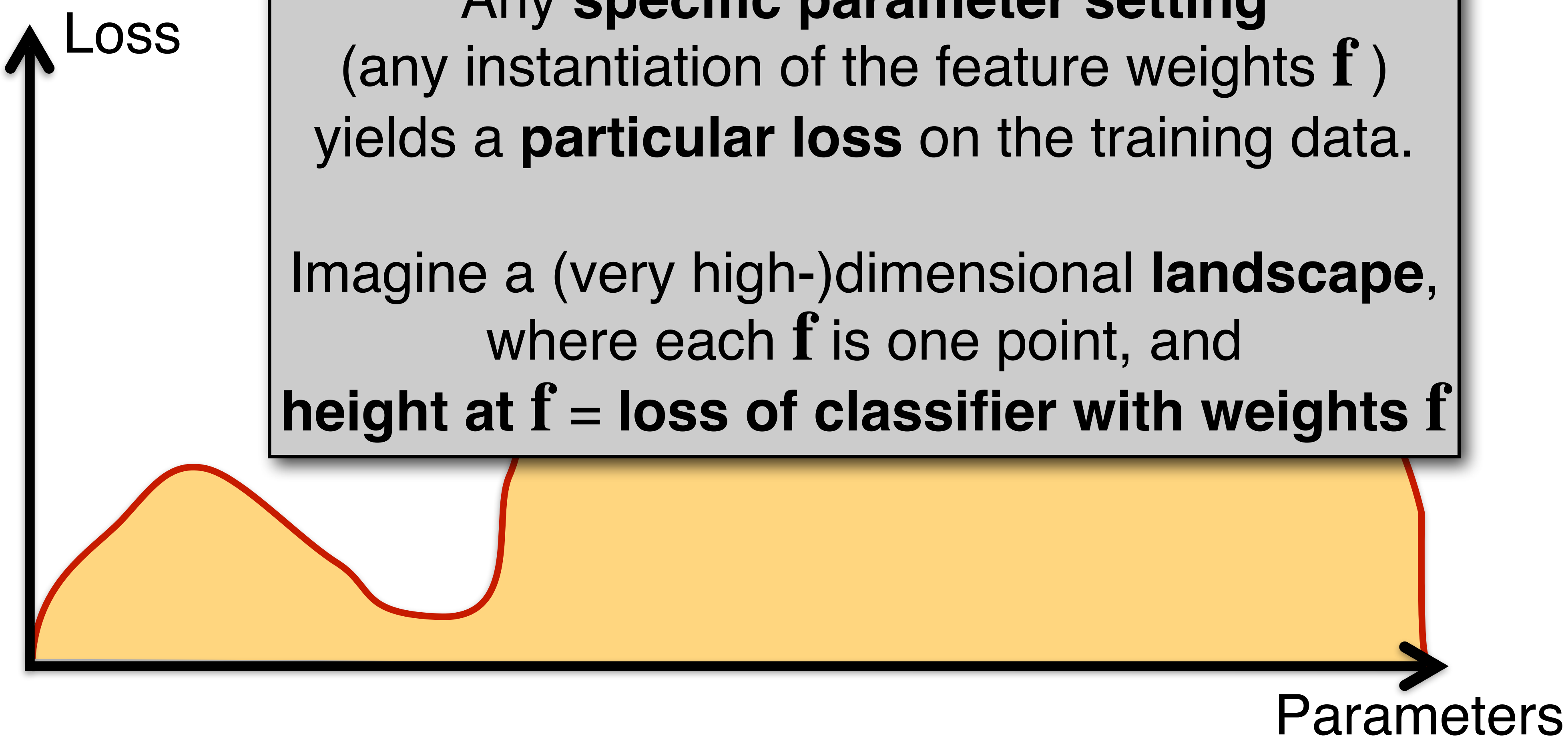
$$= -[\, y_i \log(\ P(1 \mid \mathbf{x}_i)) + (1 - y_i)\log(P(0 \mid \mathbf{x}_i))]$$

[moving the log inside]

$$= -[\, y_i \log(\sigma(\mathbf{w}\mathbf{x}_i + b)) + (1 - y_i)\log(1 - \sigma(\mathbf{w}\mathbf{x}_i + b))]$$

[plugging in definition of $P(1 \mid \mathbf{x}_i)$]

The loss surface

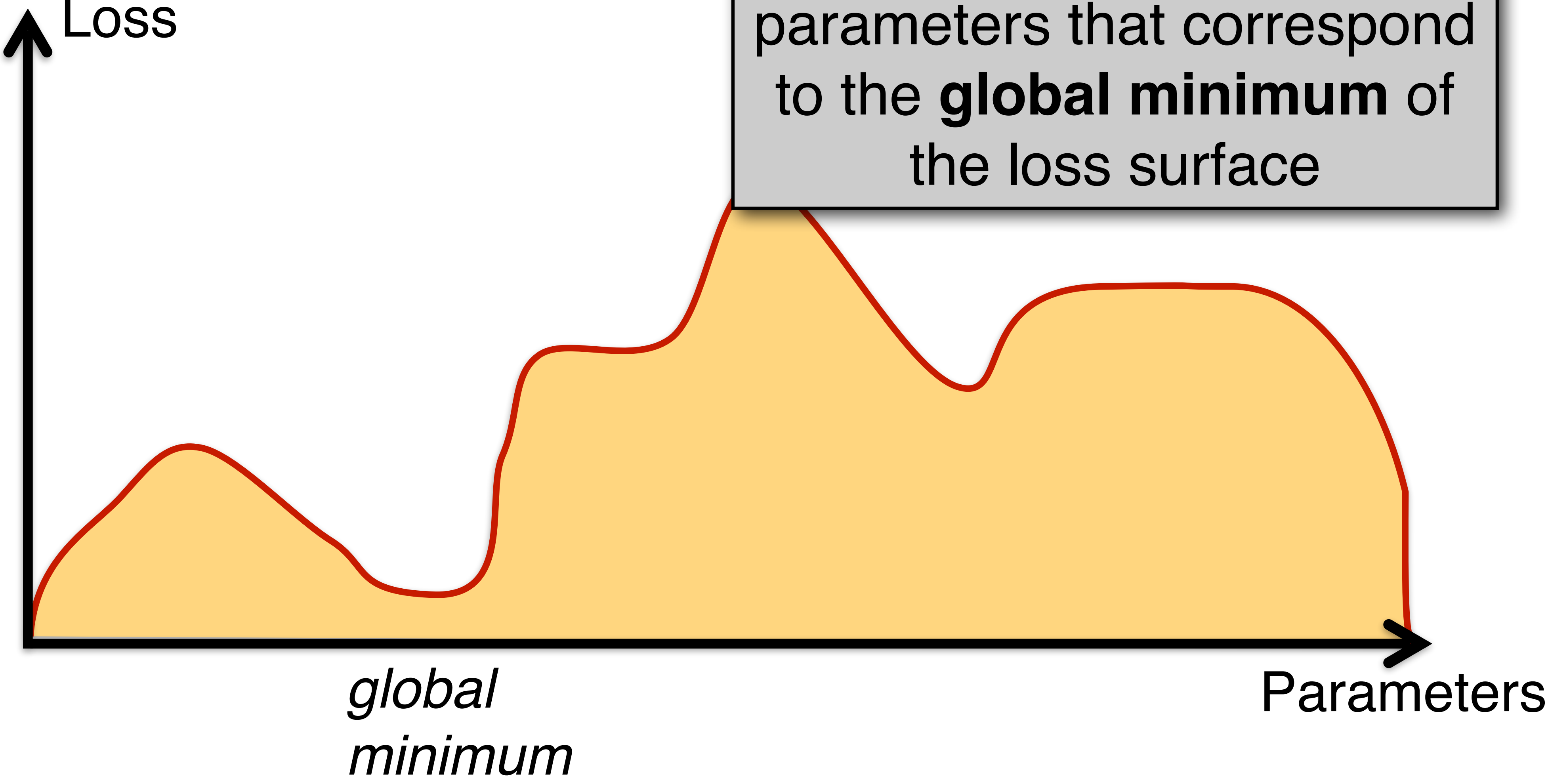


Any **specific parameter setting**
(any instantiation of the feature weights $\mathbf{f}$)
yields a **particular loss** on the training data.

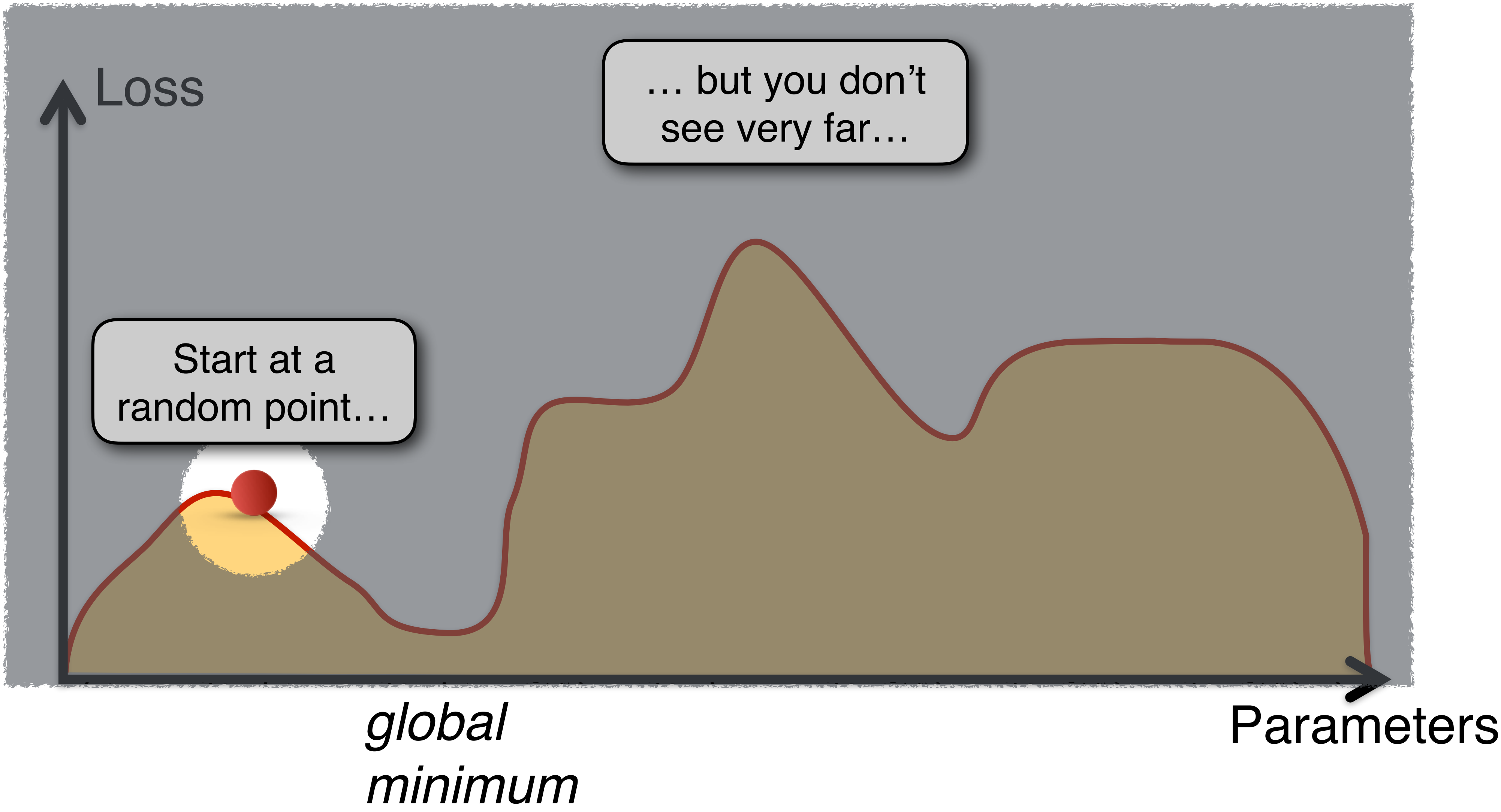
Imagine a (very high-)dimensional **landscape**,
where each $\mathbf{f}$ is one point, and
height at $\mathbf{f}$ = loss of classifier with weights $\mathbf{f}$

Learning = Moving in this landscape

Learning = finding the parameters that correspond to the **global minimum** of the loss surface



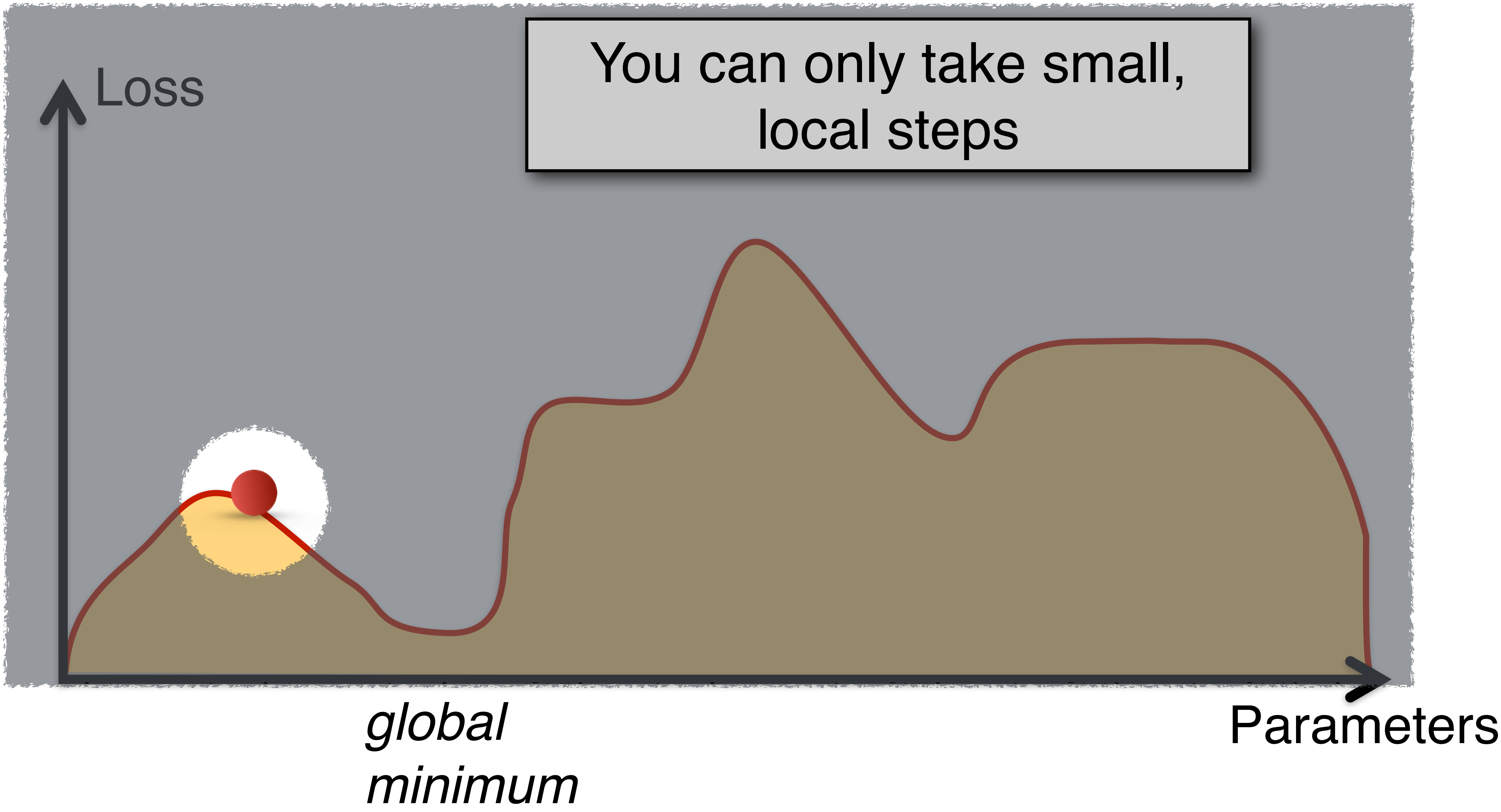
Learning = Moving in this landscape



Start at a random point…

… but you don’t see very far…

Learning = Moving in this landscape



You can only take small, local steps

Moving with Gradient Descent

How do you know *where* and *how much* to move?

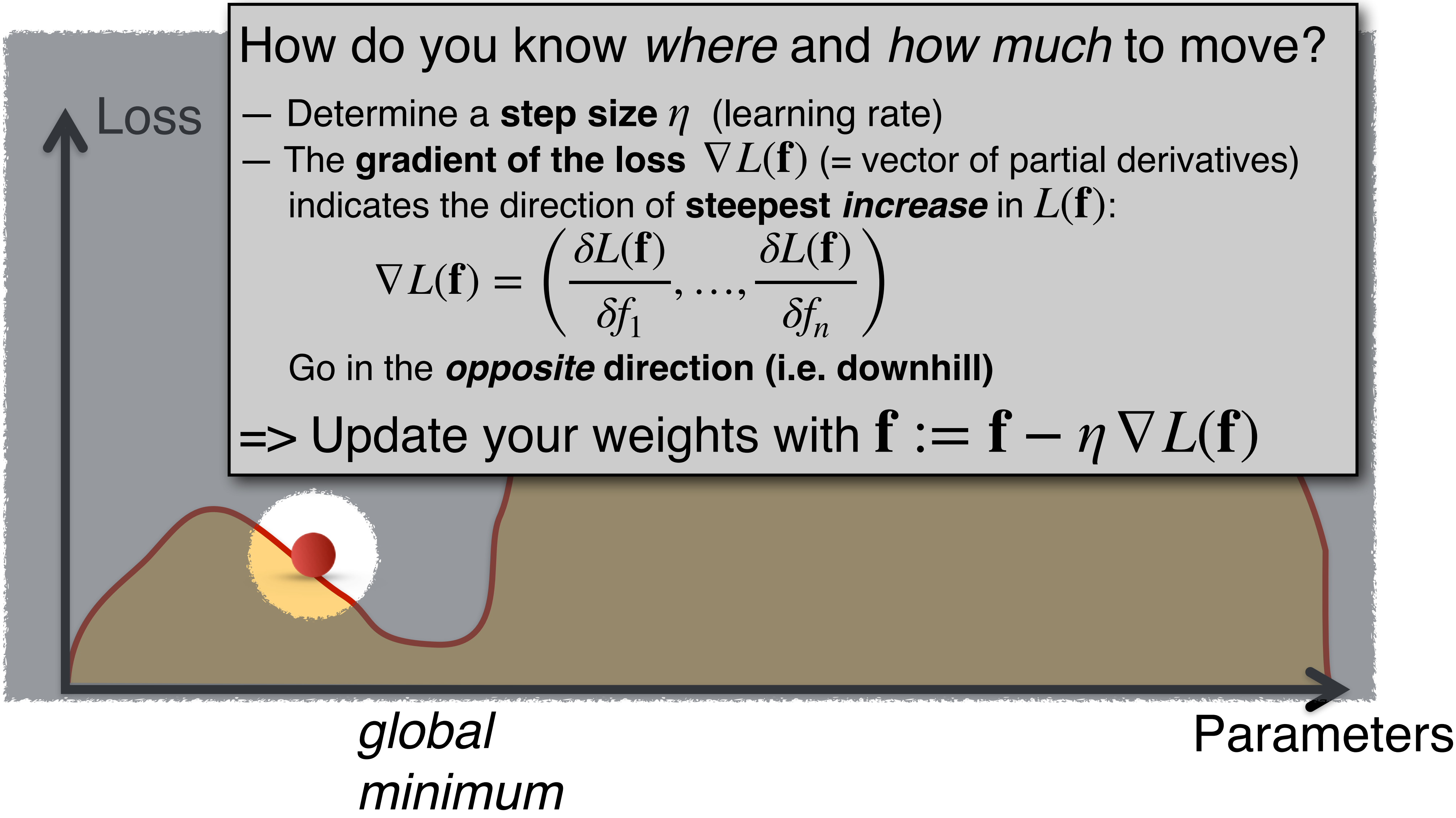
— Determine a **step size** η (learning rate)

— The **gradient of the loss** $\nabla L(\mathbf{f})$ (= vector of partial derivatives) indicates the direction of **steepest *increase*** in $L(\mathbf{f})$:

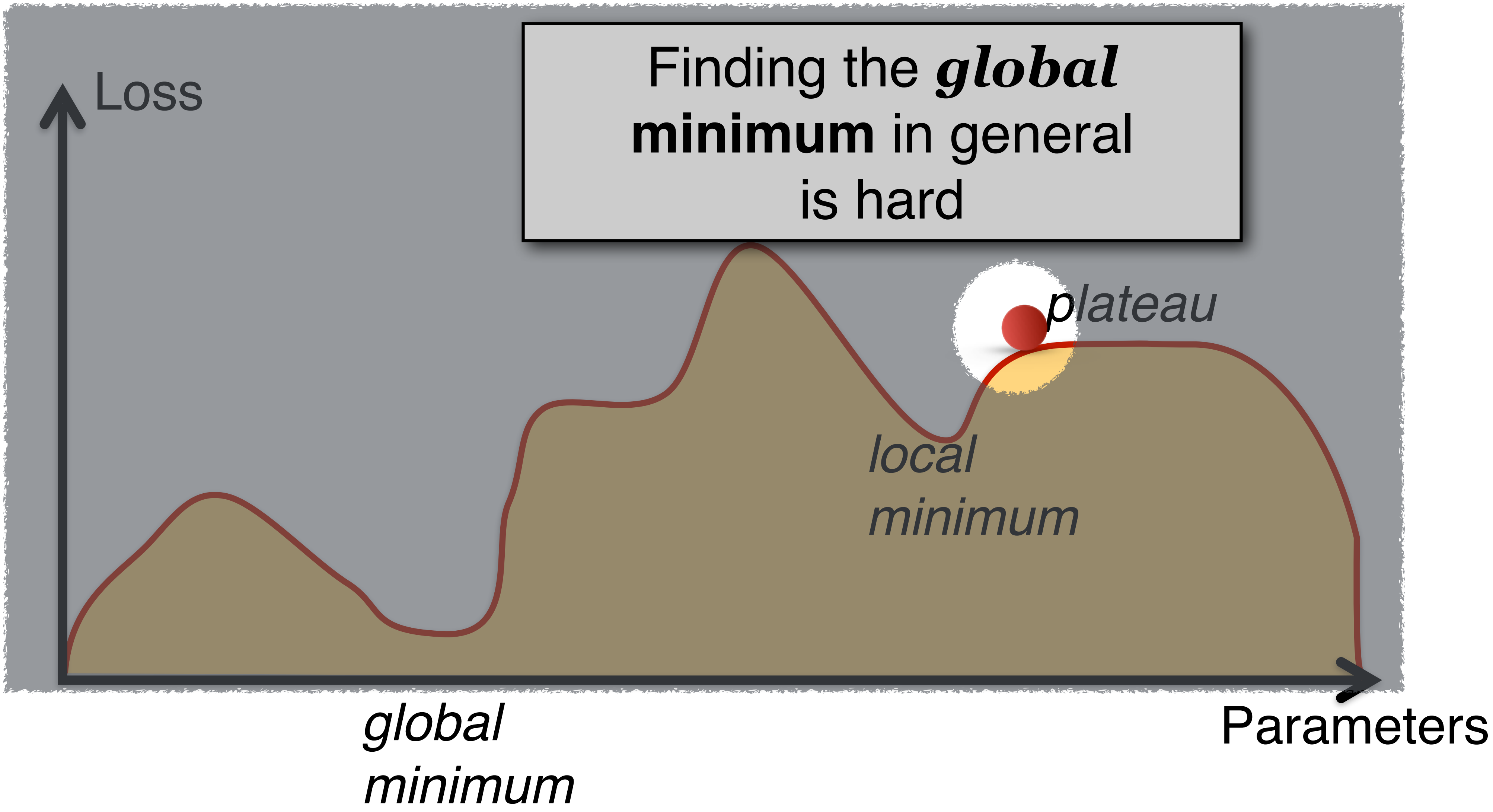
$$\nabla L(\mathbf{f}) = \left(\frac{\delta L(\mathbf{f})}{\delta f_1}, \ldots, \frac{\delta L(\mathbf{f})}{\delta f_n} \right)$$

Go in the ***opposite* direction (i.e. downhill)**

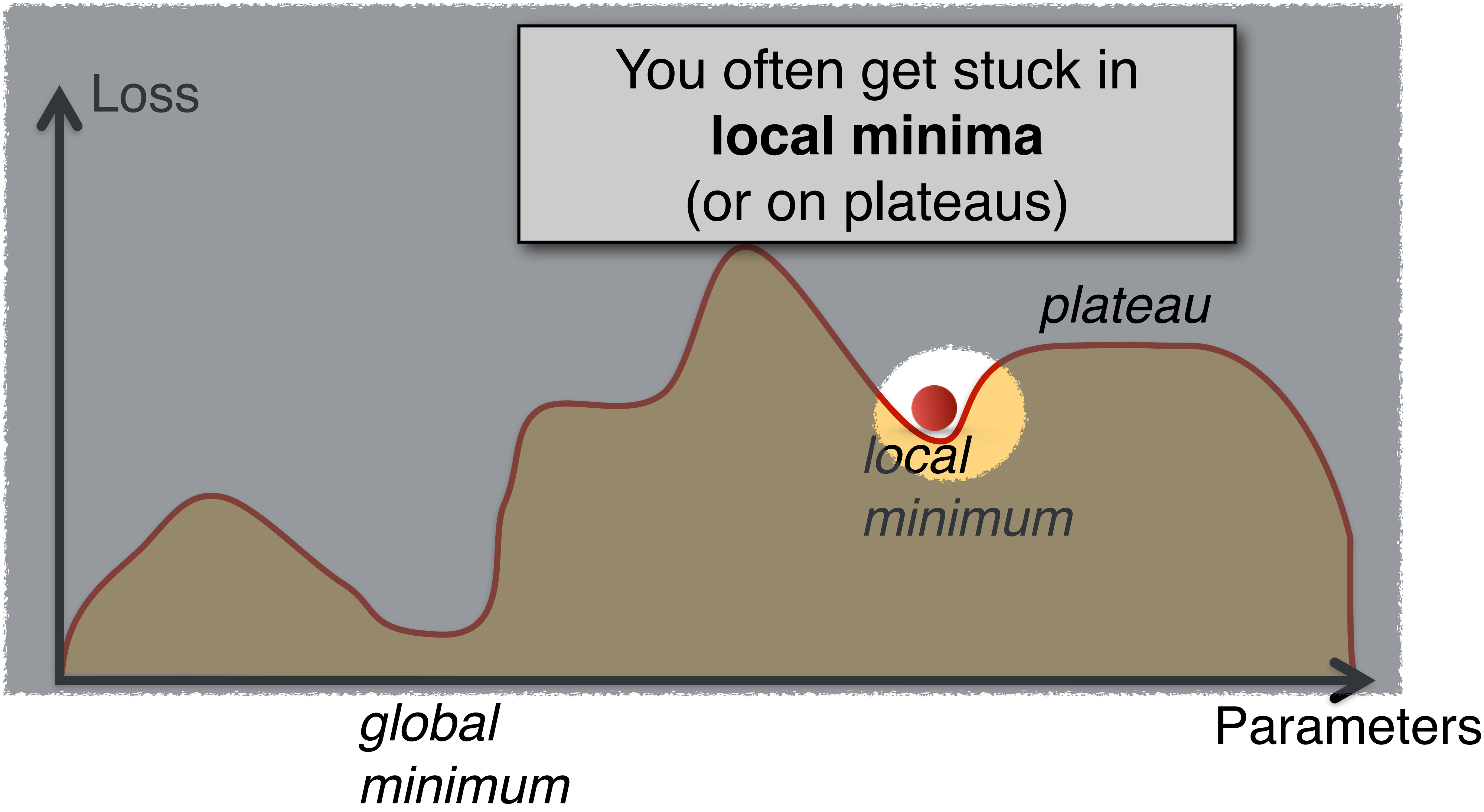
=> Update your weights with $\mathbf{f} := \mathbf{f} - \eta \nabla L(\mathbf{f})$



Gradient Descent finds *local* optima



Gradient Descent finds *local* optima



You often get stuck in
local minima
(or on plateaus)

(Stochastic) Gradient Descent

We want to find parameters that have **minimal cost** (loss) on our training data.

We don't know the shape of the whole loss surface.

Each setting of the model parameters corresponds to **one point on the loss surface**.

The **gradient** of the loss of our current parameters tells us the slope of the loss surface at the current point

And we can take a **(small) step** in the right (downhill) direction (to update our parameters)

Gradient descent:
Compute loss for entire dataset before updating weights

Stochastic gradient descent:
Compute loss for one (randomly sampled) training example before updating weights

Stochastic Gradient Descent

```
function STOCHASTIC GRADIENT DESCENT($L()$, $f()$, $x$, $y$) returns $\theta$
    # where: L is the loss function
    #        f is a function parameterized by $\theta$
    #        x is the set of training inputs $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$
    #        y is the set of training outputs (labels) $y^{(1)}, y^{(2)}, \ldots, y^{(n)}$

$\theta \leftarrow 0$
repeat T times
  For each training tuple $(x^{(i)}, y^{(i)})$ (in random order)
  Compute $\hat{y}^{(i)} = f(x^{(i)};\theta)$     # What is our estimated output $\hat{y}$?
  Compute the loss $L(\hat{y}^{(i)}, y^{(i)})$   # How far off is $\hat{y}^{(i)})$ from the true output $y^{(i)}$?
  $g \leftarrow \nabla_\theta L(f(x^{(i)};\theta), y^{(i)})$        # How should we move $\theta$ to maximize loss ?
  $\theta \leftarrow \theta - \eta g$           # go the other way instead
return $\theta$
```

Gradient for Logistic Regression

Computing the gradient of the loss for example $\mathbf{x}_i$ and weight $\mathbf{w}_j$ is very simple (x_{ji}: j-th feature of $\mathbf{x}_i$)

$$\frac{\delta L(\mathbf{w}, b)}{\delta w_j} = [\sigma(\mathbf{w}\mathbf{x}_i + b) - y_i]x_{ji}$$

More details

The **learning rate** η affects **convergence**

There are many options for setting the **learning rate**: fixed, decaying (as a function of time), adaptive,…

Often people use more complex schemes and optimizers

Mini-batch training computes the gradient on a small batch of training examples at a time.

Often more stable than SGD.

Regularization keeps the size of the weights under control

L1 or L2 regularization

Lexical Semantics and the Distributional Hypothesis

Let’s look at words again….

So far, we’ve looked at…

… the **structure** of words (**morphology**)

… the **distribution** of words (**language modeling**)

Today, we’ll start looking at the **meaning** of words (**lexical semantics**).

We will consider:

… the **distributional hypothesis** as a way to identify words with similar meanings

… two kinds of **vector representations** of words that are inspired by the distributional hypothesis

Today’s lecture

Part 1: Lexical Semantics
and the Distributional Hypothesis

Part 2: Distributional similarities
(from words to sparse vectors)

Part 3: Word embeddings
(from words to dense vectors)

Reading: Chapter 6, Jurafsky and Martin (3rd ed).

What do words **mean**, and how do we **represent** that?

Do we want to represent that…
… "cassoulet" is a French dish?
… "cassoulet" contains meat?
… "cassoulet" is a stew?

What do words mean, and how do we represent that?

Do we want to represent…

… that a "bar" is a place to have a drink?

… that a "bar" is a long rod?

… that to "bar" something means to block it?

Different approaches to lexical semantics

Roughly speaking, NLP draws on two different types of approaches to capture the meaning of words:

The lexicographic tradition aims to capture the information represented in lexicons, dictionaries, etc.

The distributional tradition aims to capture the meaning of words based on large amounts of raw text

The lexicographic tradition

Uses resources such as lexicons, thesauri, ontologies etc. that capture explicit knowledge about word meanings.

Assumes words have *discrete* word senses:

bank1 = financial institution; bank2 = river bank, etc.

May capture *explicit relations* between word (senses):
“*dog*” is a “*mammal*”, “*cars*” have “*wheels*” etc.

The Distributional Tradition

Uses large corpora of raw text to learn the meaning of words from the contexts in which they occur.

Maps words to (sparse) vectors that capture corpus statistics

Contemporary variant: use neural nets to learn dense vector "embeddings" from very large corpora

(this is a prerequisite for most neural approaches to NLP)

If each word type is mapped to a single vector, this ignores the fact that words have multiple senses or parts-of-speech

Lexicographic approaches to word meaning

Where we're at

We have looked at how to represent the meaning of sentences based on the meaning of their words (using predicate logic).

Now we will get back to the question of how to represent the meaning of words
(although this won't be in predicate logic)

We will look at lexical resources (WordNet)

We will consider two different tasks:

- Computing word similarities
- Word sense disambiguation

Different approaches to lexical semantics

Lexicographic tradition (today's lecture)

- Use lexicons, thesauri, ontologies
- Assume words have discrete word senses:

bank1 = financial institution; bank2 = river bank, etc.

- May capture explicit relations between word (senses): "dog" is a "mammal", etc.

Distributional tradition (earlier lectures)

- Map words to (sparse) vectors that capture corpus statistics
- Contemporary variant: use neural nets to learn dense vector "embeddings" from very large corpora

(this is a prerequisite for most neural approaches to NLP)

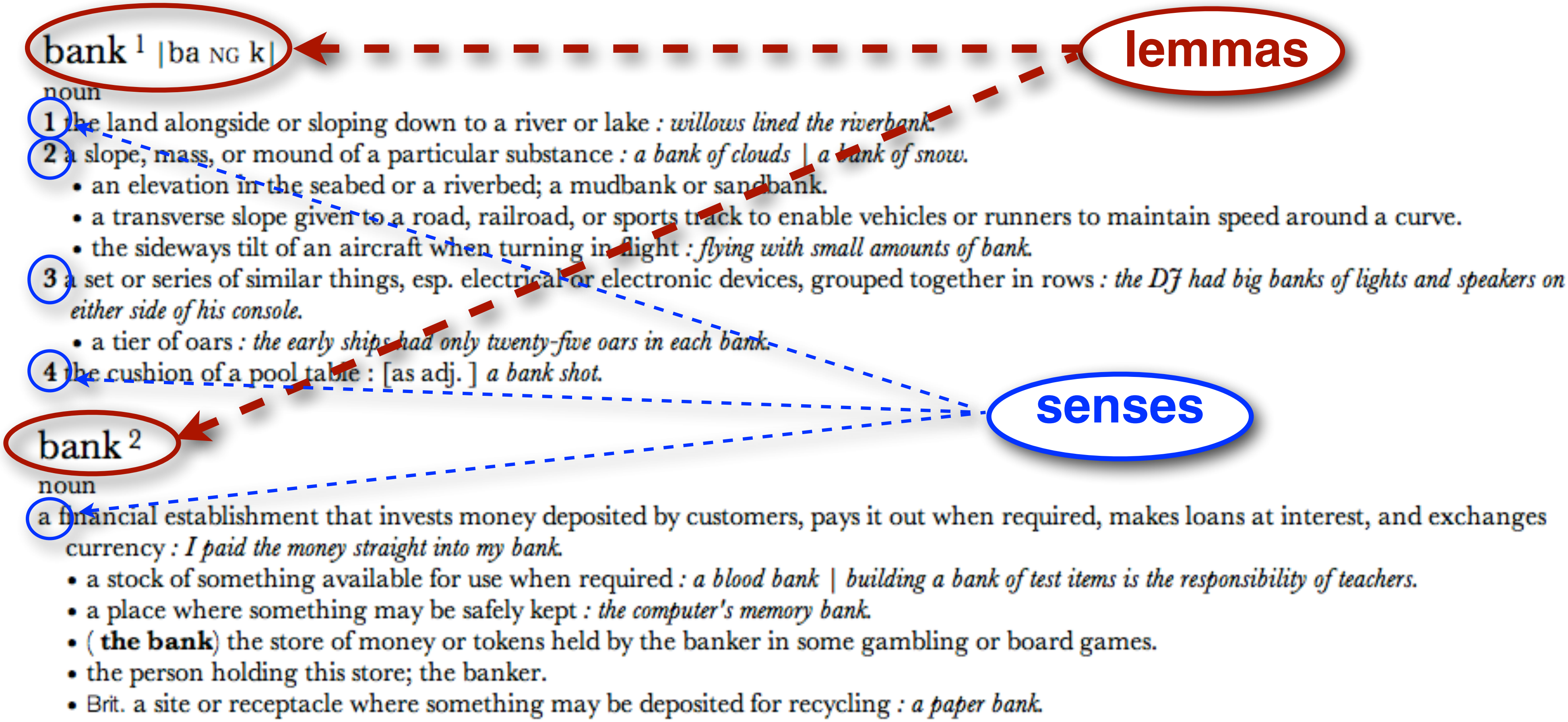
- This line of work often ignores the fact that words have multiple senses or parts-of-speech

Word senses

What does *‘bank’* mean?

- **a financial institution**
  (US banks have raised interest rates)
- **a particular branch of a financial institution**
  (the bank on Green Street closes at 5pm)
- **the bank of a river**
  (In 1927, the bank of the Mississippi flooded)
- **a ‘repository’**
  (I donate blood to a blood bank)

Lexicon entries



bank ^{1} |baNGk|

noun

1 the land alongside or sloping down to a river or lake : *willows lined the riverbank.*

2 a slope, mass, or mound of a particular substance : *a bank of clouds | a bank of snow.*

- an elevation in the seabed or a riverbed; a mudbank or sandbank.
- a transverse slope given to a road, railroad, or sports track to enable vehicles or runners to maintain speed around a curve.
- the sideways tilt of an aircraft when turning in flight : *flying with small amounts of bank.*

3 a set or series of similar things, esp. electrical or electronic devices, grouped together in rows : *the DJ had big banks of lights and speakers on either side of his console.*

- a tier of oars : *the early ships had only twenty-five oars in each bank.*

4 the cushion of a pool table : [as adj.] *a bank shot.*

bank ^{2}

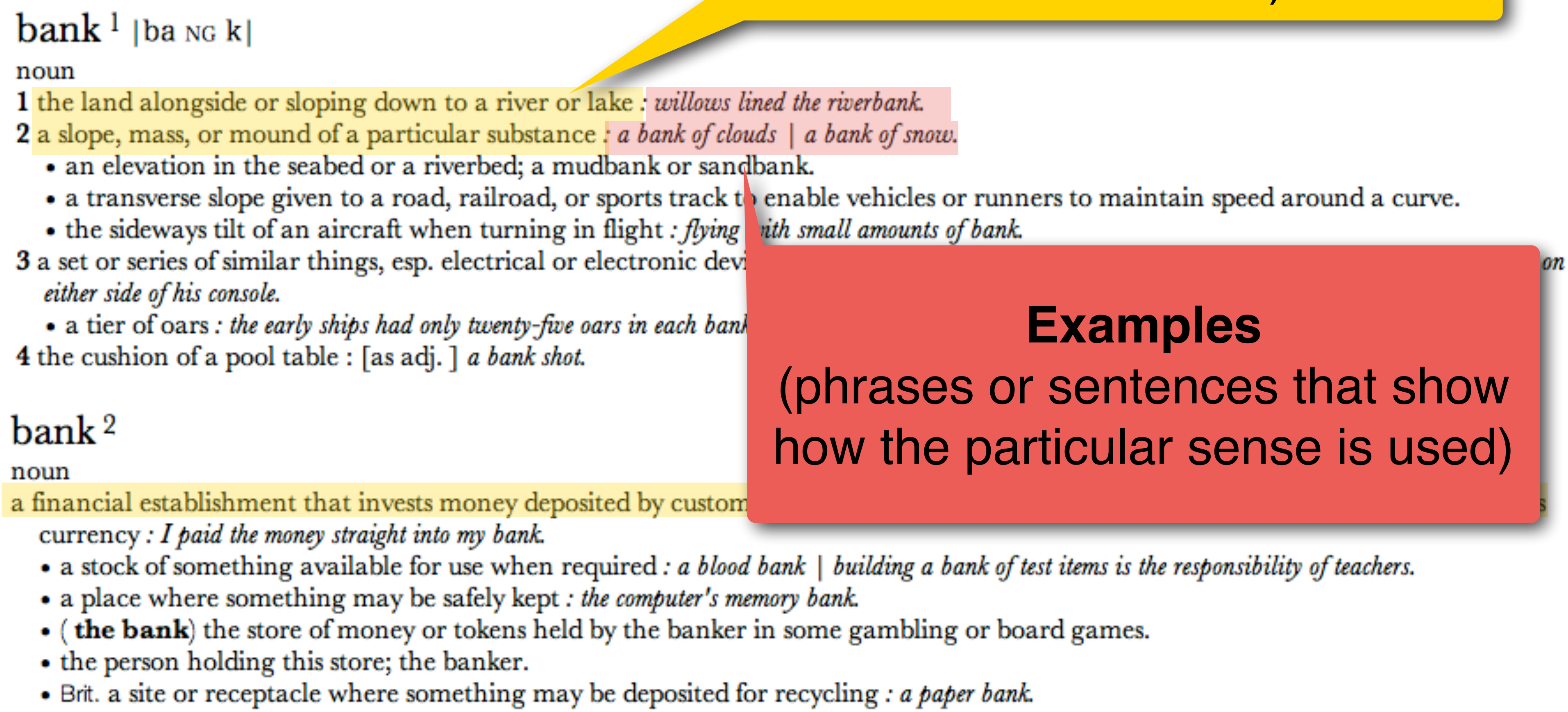
noun

a financial establishment that invests money deposited by customers, pays it out when required, makes loans at interest, and exchanges currency : *I paid the money straight into my bank.*

- a stock of something available for use when required : *a blood bank | building a bank of test items is the responsibility of teachers.*
- a place where something may be safely kept : *the computer's memory bank.*
- (**the bank**) the store of money or tokens held by the banker in some gambling or board games.
- the person holding this store; the banker.
- Brit. a site or receptacle where something may be deposited for recycling : *a paper bank.*

Lexicon entries

Glosses
(definitions intended for human readers)



bank [1] |ba NG k|

noun

1 the land alongside or sloping down to a river or lake : *willows lined the riverbank.*

2 a slope, mass, or mound of a particular substance : *a bank of clouds | a bank of snow.*

- an elevation in the seabed or a riverbed; a mudbank or sandbank.
- a transverse slope given to a road, railroad, or sports track to enable vehicles or runners to maintain speed around a curve.
- the sideways tilt of an aircraft when turning in flight : *flying with small amounts of bank.*

3 a set or series of similar things, esp. electrical or electronic devi *on either side of his console.*

- a tier of oars : *the early ships had only twenty-five oars in each bank*

4 the cushion of a pool table : [as adj.] *a bank shot.*

bank [2]

noun

a financial establishment that invests money deposited by custom s currency : *I paid the money straight into my bank.*

- a stock of something available for use when required : *a blood bank | building a bank of test items is the responsibility of teachers.*
- a place where something may be safely kept : *the computer's memory bank.*
- (**the bank**) the store of money or tokens held by the banker in some gambling or board games.
- the person holding this store; the banker.
- Brit. a site or receptacle where something may be deposited for recycling : *a paper bank.*

Examples
(phrases or sentences that show how the particular sense is used)

Some terminology

Word forms: *runs, ran, running; good, better, best*

Any, possibly inflected, form of a word
(i.e. what we talked about in morphology)

Lemma (citation/dictionary form): *run*

A basic word form (e.g. infinitive or singular nominative noun) that is used to represent all forms of the same word.
(i.e. the form you'd search for in a dictionary)

Lexeme: RUN(V), GOOD(A), BANK^1(N), BANK^2(N)

An abstract representation of a word (and all its forms), with a part-of-speech and a set of related word senses.
(Often just written (or referred to) as the lemma, perhaps in a ***different*** **FONT**)

Lexicon:

A (finite) list of lexemes

Trying to make sense of senses

Polysemy:

A lexeme is polysemous if it has different *related senses*

bank = financial institution or building

Homonyms:

Two lexemes are homonyms if their *senses are unrelated*, but they happen to have the **same spelling and pronunciation**

bank = (financial) bank or (river) bank

Relations between senses

Symmetric relations:

Synonyms: *couch/sofa*

Two lemmas with the **same** sense

Antonyms: *cold/hot, rise/fall, in/out*

Two lemmas with the **opposite** sense

Hierarchical relations:

Hypernyms and **hyponyms**: *pet/dog*

The hyponym *(dog)* is **more specific** than the hypernym *(pet)*

Holonyms and **meronyms:** *car/wheel*

The meronym *(wheel)* is a **part of** the holonym *(car)*

Metonymy

Some senses of a word may be related in a systematic way, e.g. …

… organizations and buildings:

I see you in front of the bank on Green Street.

… cars and their drivers:

This Camry looks new. vs. *The Camry honked at me.*

… authors and their works:

Jane Austen wrote Emma. vs *I really like Austen*

… plants and the food derived from them:

Plums have beautiful blossoms. vs *I ate a plum*

WordNet and WordNet-based Word Similarity

WordNet

Very large, publicly available **lexical database** of English:

110K nouns, 11K verbs, 22K adjectives, 4.5K adverbs
(WordNets for many other languages exist or are under construction)

Each word has a POS tag and one or more **word senses**.

Avg. # of senses: 1.23 nouns, 2.16 verbs, 1.41 adj, 1.24 adverbs

Word senses are grouped into synonym sets (“**synsets**”)

81K noun synsets, 13K verb synsets, 19K adj. synsets, 3.5K adverb synsets

Synsets are connected in a hierarchy/network defined via **conceptual-semantic relations**

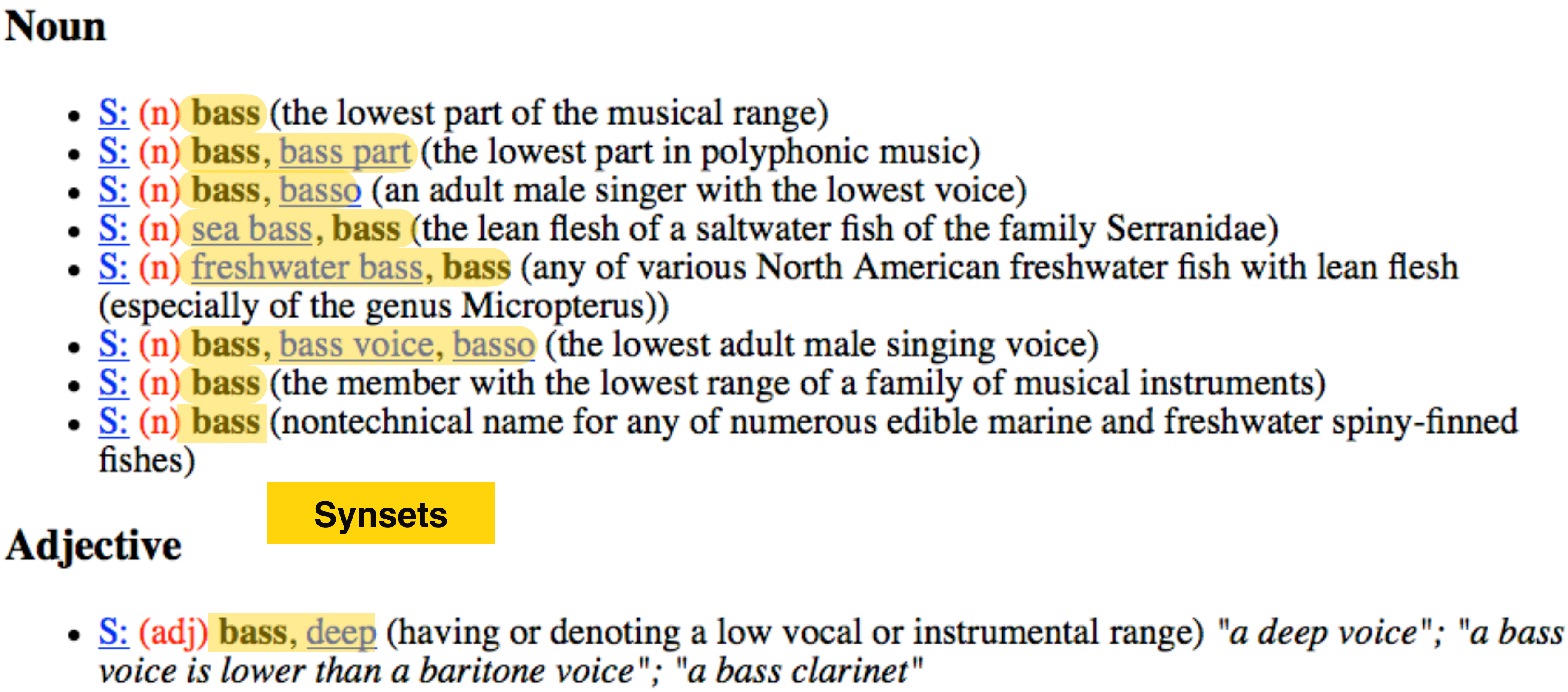
— hypernym/hyponym relation (IS-A)
— holonym/meronym relation (HAS-A)

Also lexical relations (derivational morphology), and lemmatization

Available at http://wordnet.princeton.edu

A WordNet example

Searching for “bass” returns



Noun

- S: (n) **bass** (the lowest part of the musical range)
- S: (n) **bass**, bass part (the lowest part in polyphonic music)
- S: (n) **bass**, basso (an adult male singer with the lowest voice)
- S: (n) sea bass, **bass** (the lean flesh of a saltwater fish of the family Serranidae)
- S: (n) freshwater bass, **bass** (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
- S: (n) **bass**, bass voice, basso (the lowest adult male singing voice)
- S: (n) **bass** (the member with the lowest range of a family of musical instruments)
- S: (n) **bass** (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

Synsets

Adjective

- S: (adj) **bass**, deep (having or denoting a low vocal or instrumental range) *"a deep voice"; "a bass voice is lower than a baritone voice"; "a bass clarinet"*

Hierarchical synset relations: Nouns (I)

IS-A relations (hyponymy):

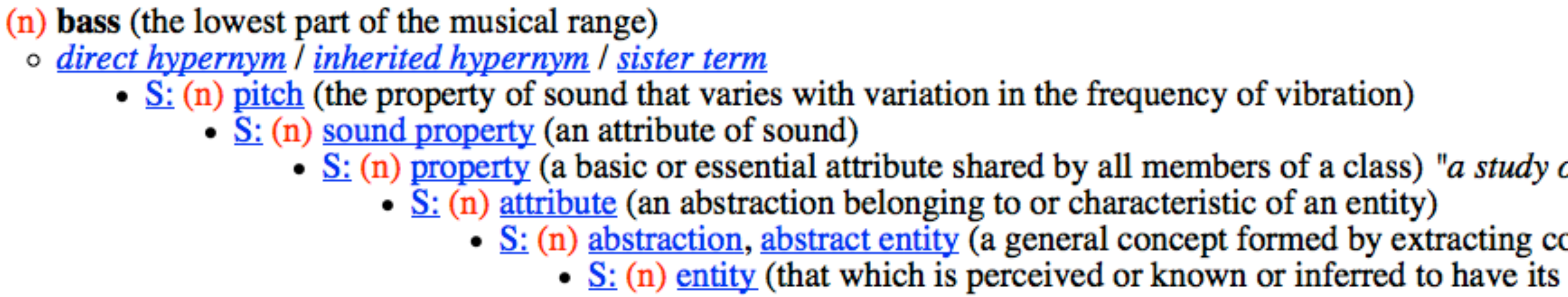
Hypernym/**hyponym** (between concepts)

meal is a hypernym (superordinate) of *breakfast*
breakfast is a hyponym (subordinate) of *meal*
dog is a hypernym (superordinate) of *poodle*
poodle is a hyponym (subordinate) of (IS-A) *dog*

Instance hypernym/**hyponym** (concepts and instances)

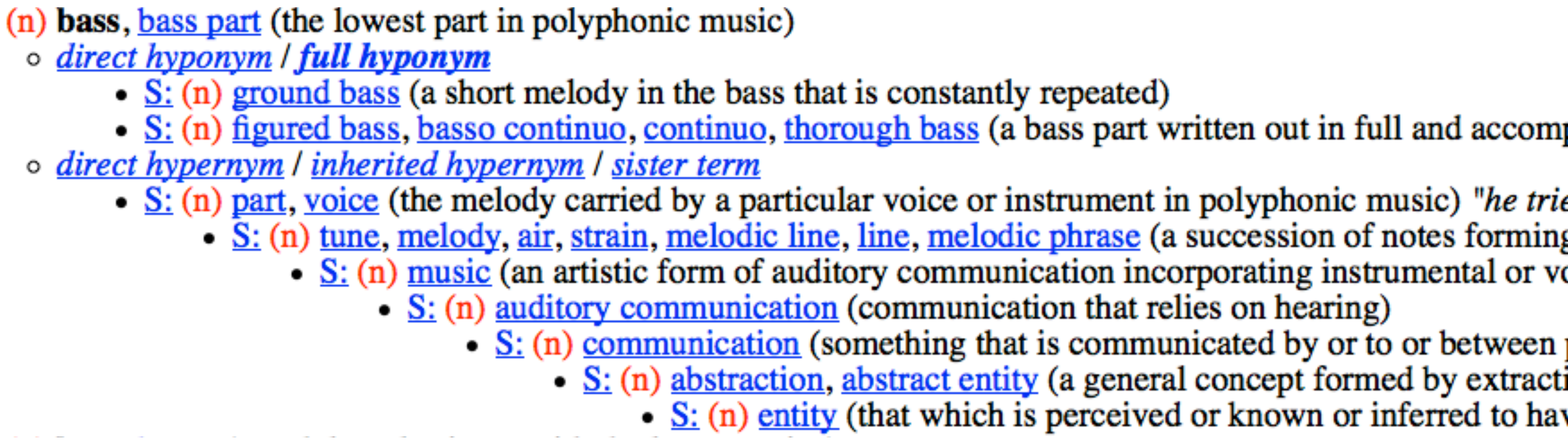
composer is the instance hypernym of (HAS-INSTANCE) *Bach*
Bach is an instance hyponym of (IS-INSTANCE-OF) *composer*

WordNet Hypernyms and Hyponyms



(n) **bass** (the lowest part of the musical range)

- *direct hypernym* / *inherited hypernym* / *sister term*
  - S: (n) pitch (the property of sound that varies with variation in the frequency of vibration)
    - S: (n) sound property (an attribute of sound)
      - S: (n) property (a basic or essential attribute shared by all members of a class) *"a study o*
        - S: (n) attribute (an abstraction belonging to or characteristic of an entity)
          - S: (n) abstraction, abstract entity (a general concept formed by extracting co
            - S: (n) entity (that which is perceived or known or inferred to have its



(n) **bass**, bass part (the lowest part in polyphonic music)

- *direct hyponym* / ***full hyponym***
  - S: (n) ground bass (a short melody in the bass that is constantly repeated)
  - S: (n) figured bass, basso continuo, continuo, thorough bass (a bass part written out in full and accomp
- *direct hypernym* / *inherited hypernym* / *sister term*
  - S: (n) part, voice (the melody carried by a particular voice or instrument in polyphonic music) *"he trie*
    - S: (n) tune, melody, air, strain, melodic line, line, melodic phrase (a succession of notes forming
      - S: (n) music (an artistic form of auditory communication incorporating instrumental or vo
        - S: (n) auditory communication (communication that relies on hearing)
          - S: (n) communication (something that is communicated by or to or between p
            - S: (n) abstraction, abstract entity (a general concept formed by extracti
              - S: (n) entity (that which is perceived or known or inferred to hav

Hierarchical synset relations: Nouns (II)

Part-Whole relations (meronymy):

Member holonym/meronym (groups and members)

crew is a member holonym of (HAS-MEMBER) co-pilot

co-pilot is a member meronym of (IS-MEMBER-OF) *crew*

Part holonym/meronym (wholes and parts)

car is a part holonym of (HAS-PART) *wheel*

wheel is a part meronym of (IS-PART-OF) *car*

Substance holonym/meronym (substances and components)

bread is a substance holonym of (HAS-COMPONENT) *flour*

flour is a substance meronym of (IS-COMPONENT-OF) *bread*

Hierarchical synset relations: Verbs

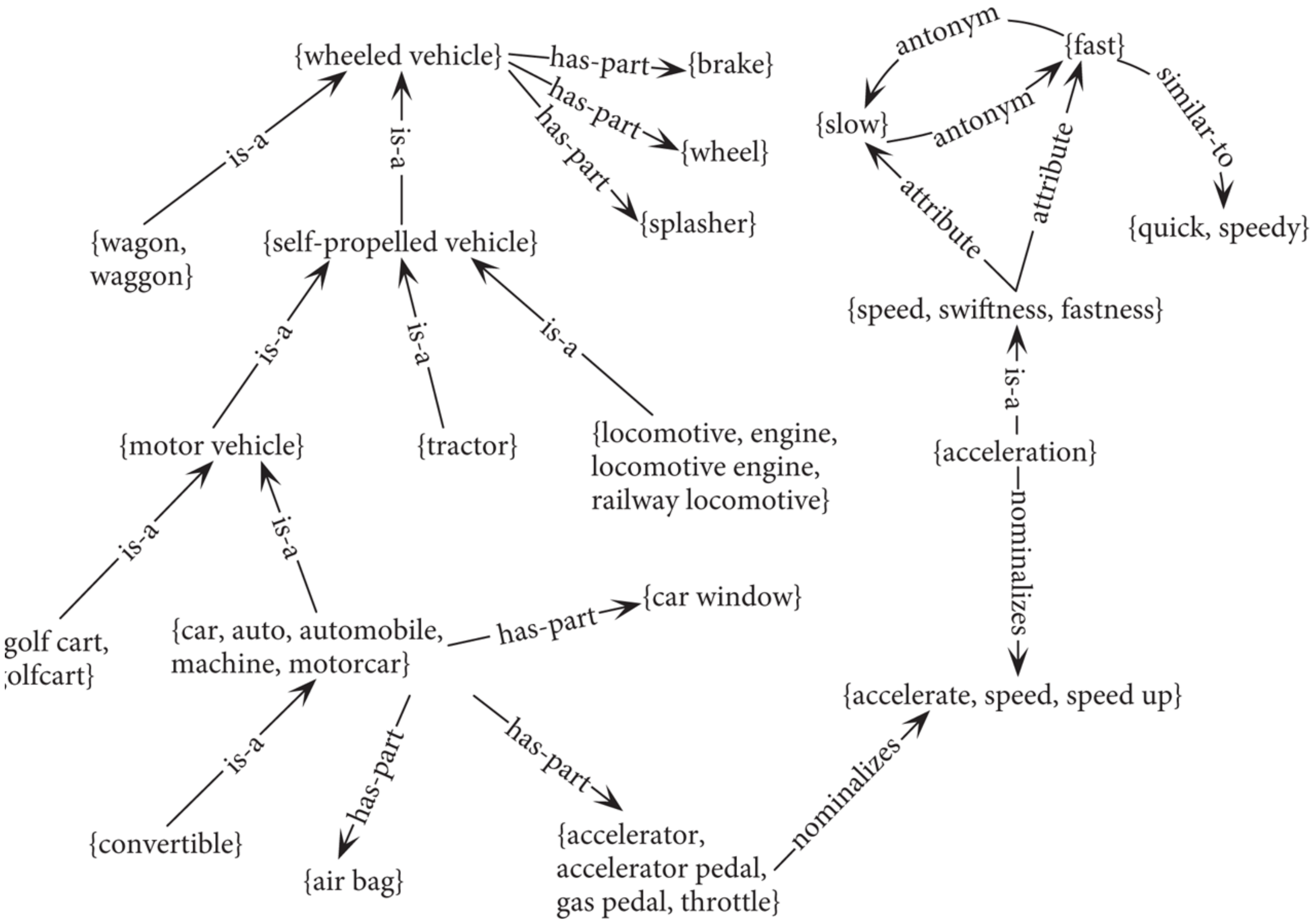
Hypernym/troponym (between events):

travel/fly, walk/stroll
Flying is a troponym of *traveling:*
it denotes **a specific manner** of *traveling*

Entailment (between events):

snore/sleep
Snoring **entails (presupposes)** *sleeping*
(if somebody is snoring, they have to be sleeping)

WordNet relations as a graph



(Figure from Jurafsky & Martin, 3rd Edition, and Navigli 2016)

WordNet as a semantic network

The **Hypernym/hyponym** relations (IS-A) and **holonym/meronym** relations (HAS-A) in WordNet capture some important world knowledge, e.g.:

car IS-A motor-vehicle IS-A… IS-A wheeled-vehicle

wheeled-vehicle HAS-A brake

→ car IS-A wheeled-vehicle

→ car HAS-A brake

We can interpret WordNet as a simple "semantic network" (for semantic networks in AI see e.g. http://www.jfsowa.com/pubs/semnet.htm)

WordNet-based word similarity

There have been many attempts to exploit resources like WordNet to compute word (sense) similarities.

Classic approaches use the distance (**path length**) between synsets (these paths typically only consider hypernym/hyponym relations), possibly augmented with corpus statistics

More recent (neural) approaches aim to learn (non-Euclidean) embeddings that capture the hierarchical hypernym/hyponym structure of WordNet.

What do we mean by “word (sense) similarity”?

There are many aspects to “similarity”:

— **Similarity as synonymy:**

sim(*couch*, *sofa*)> sim(*poodle*, *dog*) > sim(*poodle*, *pug*), …

Do the two words/senses have the same meaning?

(WordNet: synsets are synonyms (similarity=1), but hypernym/hyponyms (*dog*/*poodle*) are also more similar to each other than unrelated words)

— **Similarity as association:**

How related are the two words/senses to each other?

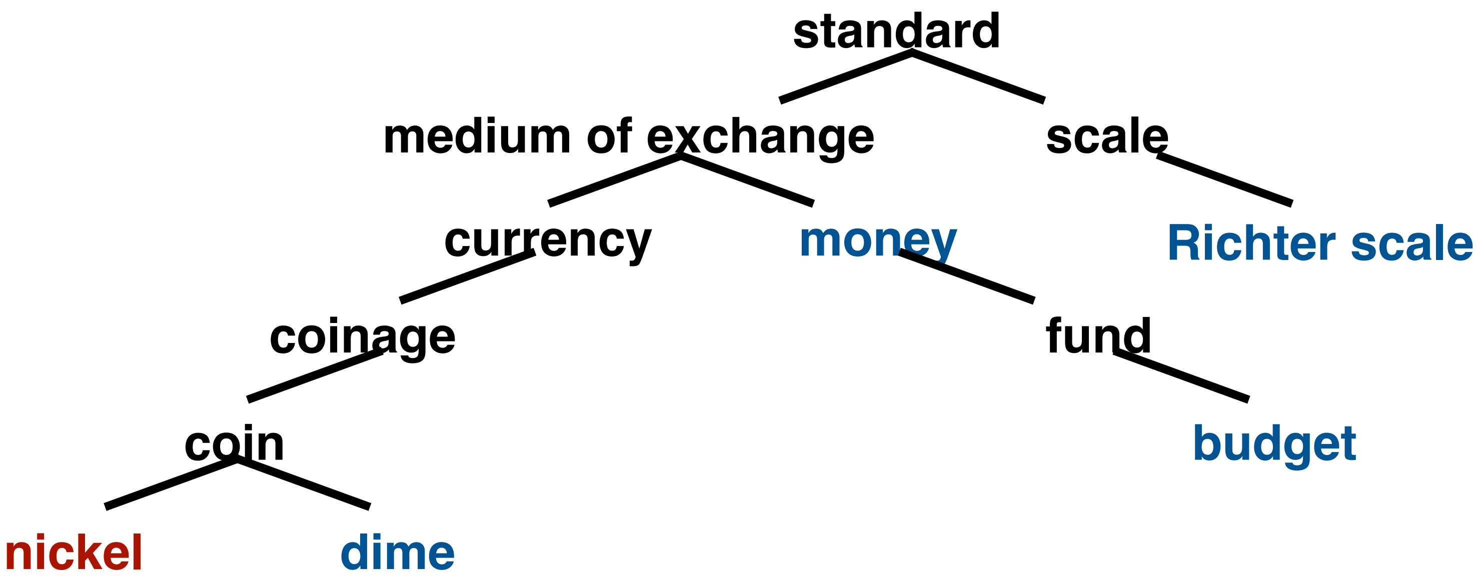
coffee and *cup* are strongly associated, but not synonyms

“**Semantic fields**”: sets of words that are topically related

(WordNet: holonyms/meronyms etc. capture some associations)

Earlier metrics of similarity in NLP often conflate both notions, but see e.g. SimLex-999 https://www.aclweb.org/anthology/J15-4004.pdf

WordNet path lengths: examples and problems



Path length is just the distance between synsets

pathlen(nickel, dime) = 2 (nickel—coin—dime)
pathlen(nickel, money) = 5 (nickel—…—medium of exchange—money)
pathlen(nickel, budget) = 7 (nickel—…—medium of exchange—…–budget)

But do we really want the following?

pathlen(nickel, coin) < pathlen(nickel, dime)
pathlen(nickel, Richter scale) = pathlen(nickel, budget)

Problems with thesaurus-based similarity

We need to have a thesaurus!
(not available for all languages)

We need to have a thesaurus that contains the words we’re interested in.

We need a thesaurus that captures a rich hierarchy of hypernyms and hyponyms.

Most thesaurus-based similarities depend on the specifics of the hierarchy that is implement in the thesaurus.

Learning hyponym relations

If we don't have a thesaurus, can we *learn* that Corolla is a kind of car from text?

Certain **phrases and patterns** indicate hyponym relations:

Hearst(1992) [Hearst patterns]

Enumerations: *cars* ***such as*** *the Corolla, the Civic, and the Vibe,*
Appositives: *the Corolla* ***,*** *a popular car…*

We can also **learn these patterns** if we have some **seed examples of hyponym relations** (e.g. from WordNet):

1. *Take all hyponym/hypernym pairs from WordNet (e.g. car/vehicle)*
2. *Find all sentences that contain both, and identify patterns*
3. *Apply these patterns to new data* to get new hyponym/hypernym pairs