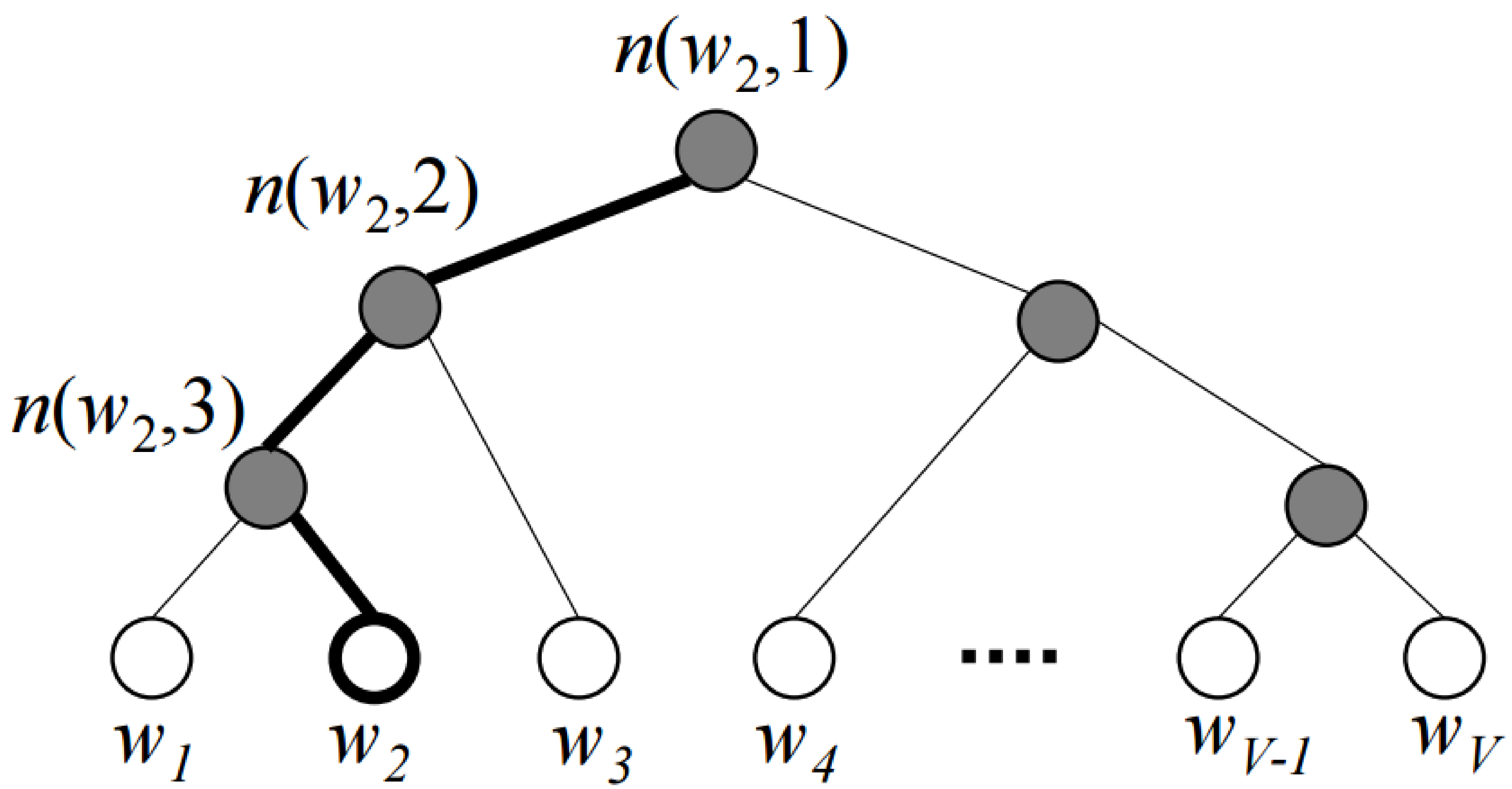
# Word Embeddings
# Negative Sampling

國立臺灣大學 資訊工程學系
陳縕儂 助理教授
http://vivianchen.idv.tw

# Hierarchical Softmax

- Idea: compute the probability of leaf nodes using the paths



$$O(N) \rightarrow O(\log N)$$

Mikolov et al., “Distributed representations of words and phrases and their compositionality,” in NIPS, 2013.

# Negative Sampling

- Idea: only update a sample of output vectors

$$C(\theta) = -\log \sigma({v'_{w_O}}^T v_{w_I}) + \sum_{w_j \in \mathcal{W}_{\text{neg}}} \log \sigma({v'_{w_j}}^T v_{w_I})$$

$${v'_{w_j}}^{(t+1)} = {v'_{w_j}}^{(t)} - \eta \cdot EI_j \cdot h$$

$$v_{w_I}^{(t+1)} = v_{w_I}^{(t)} - \eta \cdot EH^T$$

$$w_j \in \{w_O\} \cup \mathcal{W}_{\text{neg}}$$

$$EI_j = \sigma({v'_{w_j}}^T v_{w_I}) - t_j$$

$$EH = \sum_{w_j \in \{w_O\} \cup \mathcal{W}_{\text{neg}}} EI_j \cdot v'_{w_j}$$

Mikolov et al., "Distributed representations of words and phrases and their compositionality," in NIPS, 2013.

# Negative Sampling

- Sampling methods
  - Random sampling $w_j \in \{w_O\} \cup \mathcal{W}_{\text{neg}}$
  - Distribution sampling: $w_j$ is sampled from $P(w)$ What is a good $P(w)$?

Idea: less frequent words sampled more often

- Empirical setting: unigram model raised to the power of 3/4

| Word | Probability to be sampled for “neg” |
|---|---|
| is | $0.9^{3/4} = 0.92$ |
| constitution | $0.09^{3/4} = 0.16$ |
| bombastic | $0.01^{3/4} = 0.032$ |

Mikolov et al., “Distributed representations of words and phrases and their compositionality,” in NIPS, 2013.