Language and Statistics II

Lecture 9: (Mostly) Intro to Parsing

Noah Smith

Lecture Overview

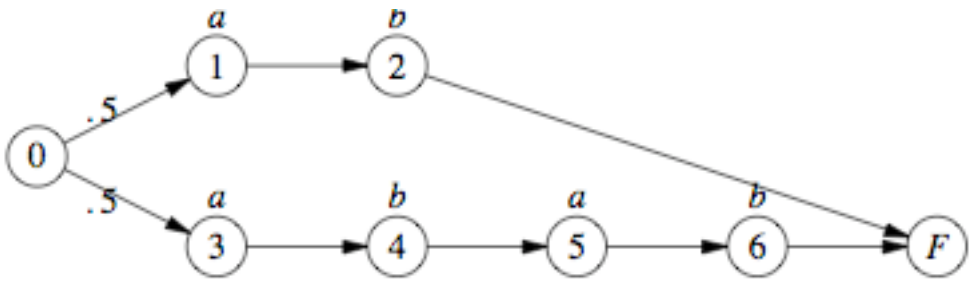
1. Finish up HMM topology learning.
2. Motivations for parsing.
3. A really simple model for parsing: PCFG
   - PCFGs as a stochastic process
   - Relationship to other models we know
   - Naïve training of PCFGs

Learning HMM Topologies

- The problem of **grammatical inference** is a big one.
  - Lots of theoretical work in automaton community.
  - Learnability results (often negative).
  - Not a lot of practical work
  - Not a lot of probabilistic work.
- One nice example: Stolcke and Omohundro (1993)
  - Later extended to PCFGs

HMM Topology Learning
(Stolcke and Omohundro, 1993)

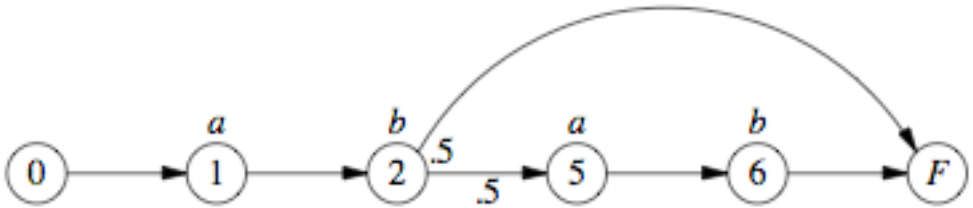
- Learning from emission symbol sequences only (don't know states).
- Start by building an HMM encoding **exactly** the data.
- In this example, we saw "ab" and "abab"

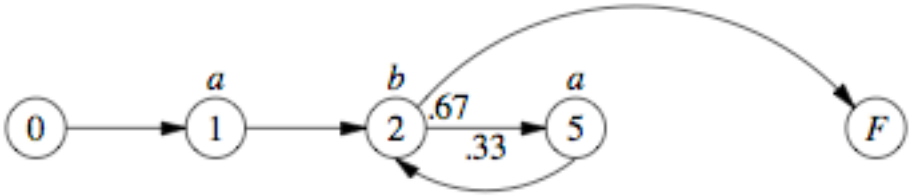


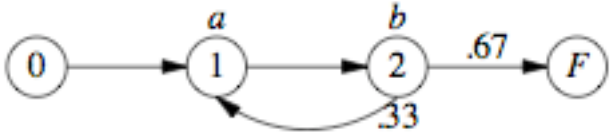
HMM Topology Learning
(Stolcke and Omohundro, 1993)

- Initialize as described.
- Iterate:
  - Consider merging each pair of states; compute p(data) under each merge.
    - Parameters under the merge are computed by merging the paths and re-normalizing the counts.
    - p(data) is estimated by taking the best path for each string.
  - Merge the two states that hurt p(data) the least.

Example







HMM Topology Learning
(Stolcke and Omohundro, 1993)

- Initialize as described.
- Iterate:
  - Consider merging each pair of states; compute p(data) under each merge.
    - Parameters under the merge are computed by merging the paths and re-normalizing the counts.
    - p(data) is estimated by taking the best path for each string.
  - Merge the two states that hurt p(data) the least.

- What's wrong here?

- Key idea: replace p(data) with p(data, model); factor in a prior.

HMM Topology Learning
(Stolcke and Omohundro, 1993)

- p(model structure) $\propto e^{-|Q|}$
- p(parameters | structure) ~ symmetric Dirichlet ("add 0.1" smoothing)

- The prior "makes up" for the decrease in likelihood as states are merged … but not forever!

HMM Topology Learning
(Stolcke and Omohundro, 1993)

- Initialize as described.
- Iterate:
  - Consider merging each pair of states; compute p(data, model) under each merge.
    - Parameters under the merge are computed by merging the paths and re-normalizing the counts.
    - p(data, model) is estimated by taking the best path for each string.
  - Merge the two states that increase p(data, model) the most; if none, then stop.

New topic …

Parsing

But first …

What languages are spoken
by people taking the class?

Parsing: Motivation

- Language modeling (Chelba & Jelinek, 1998) … predict next word given left syntactic context (syntax) instead of previous two words (trigram):

  John, who eats cookies, {love, loves} …

Or, transformations on data:

- Machine translation (Alshawi, 1996; Wu, 1997; Chiang, 2005 …)
- Information extraction (Hobbs, 1997; Viola & Narasimhan, 2005)
- Grammar checking (obviously!)
- NL interfaces to Databases (Collins and Zettlemoyer, 2005)
- Lexical learning (Lin, 1997)

Why are there so many parsing papers?

Why are we spending two weeks on parsing?

- Parsing is hard! (cf. tagging)
  - Low 90s right now
- So many theories …
- Easy to evaluate given annotated data
  - Evaluation is uncontroversial & automatic (cf. machine translation)
- Great problem for structured prediction
- Lots of room for magic: smoothing, search, model refinement
- History: parsing was **the** major problem in CL before the empirical paradigm shift - seen as crucial for "understanding"

Major Research Questions

- What's the right **representation**?
- What's the right **model**?
- How to learn to parse **empirically**?
- How to make parsers **fast**?
- How to incorporate structure **downstream**?

First Answers

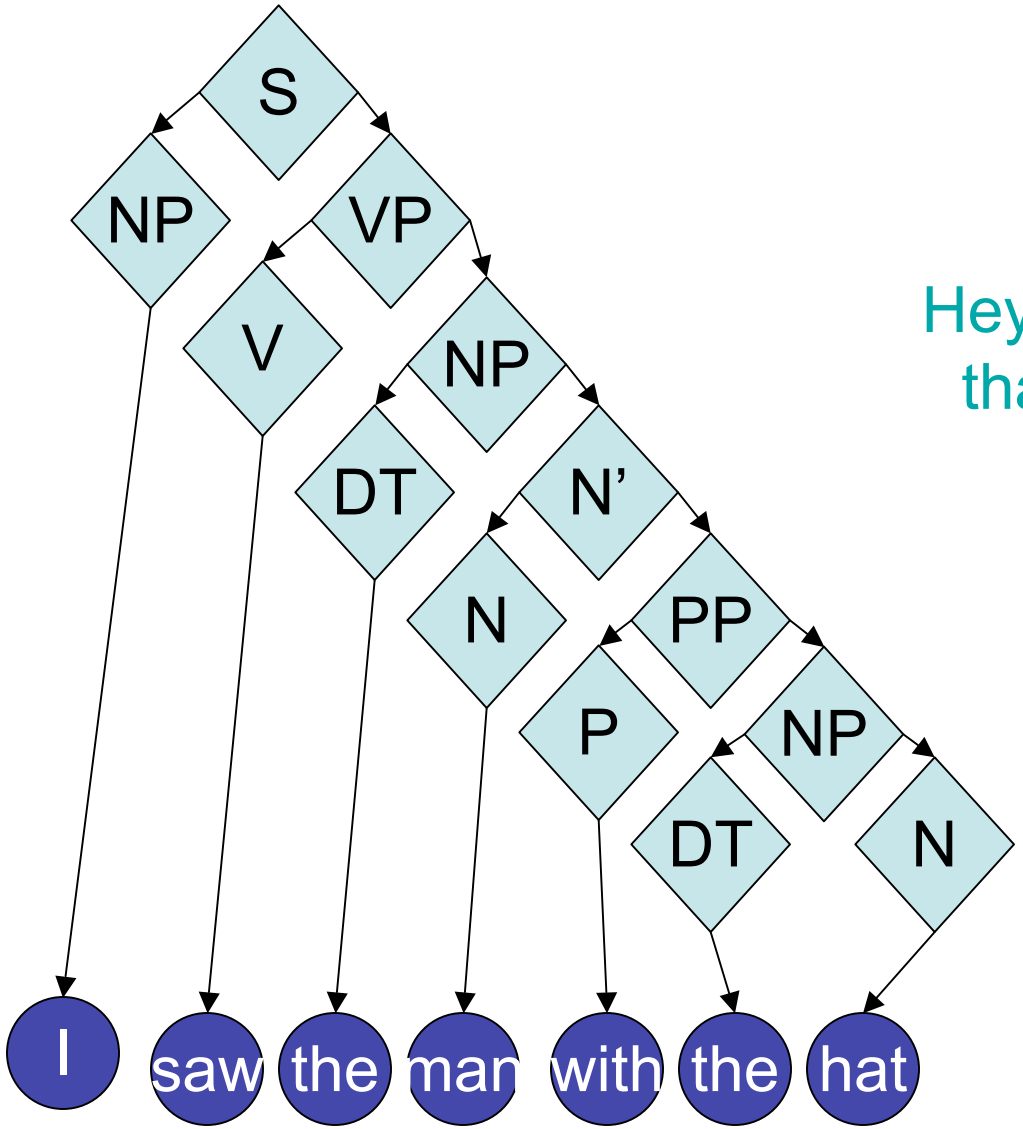
- What's the right **representation**? phrase structure
- What's the right **model**? PCFG
- How to learn to parse **empirically**? MLE/treebank
- How to make parsers **fast**? CKY/Earley's algorithm
- How to incorporate structure **downstream**? best parse

Context-Free Grammar

- Alphabet Σ
- Set of variables N
- Start symbol $S \in N$
- Rewrite rules: $X \rightarrow \alpha$, where $X \in N$ and $\alpha \in (N \cup \Sigma)^*$

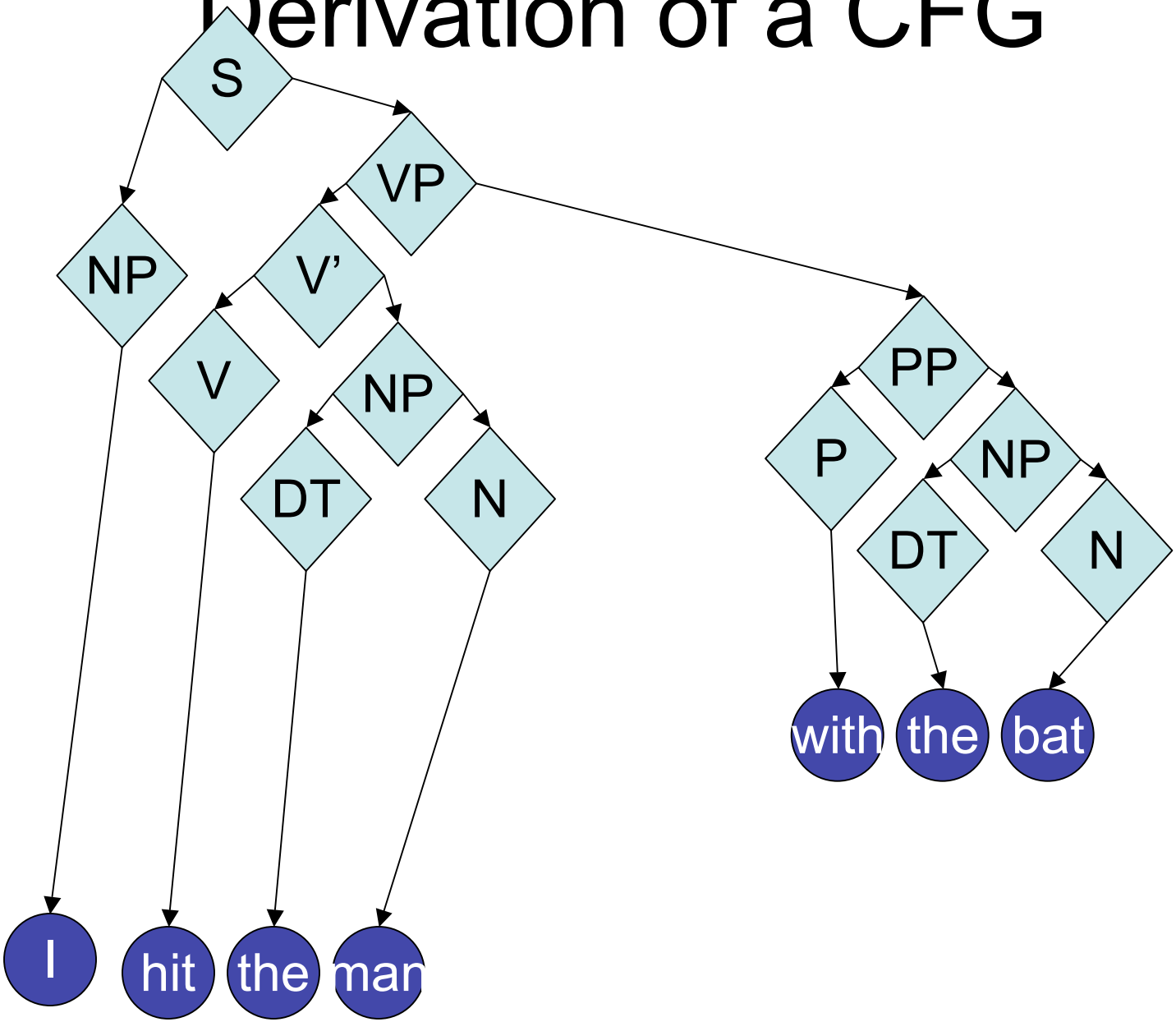
CNF: Assume $\alpha \in N^2 \cup \Sigma$.

Derivation of a CFG

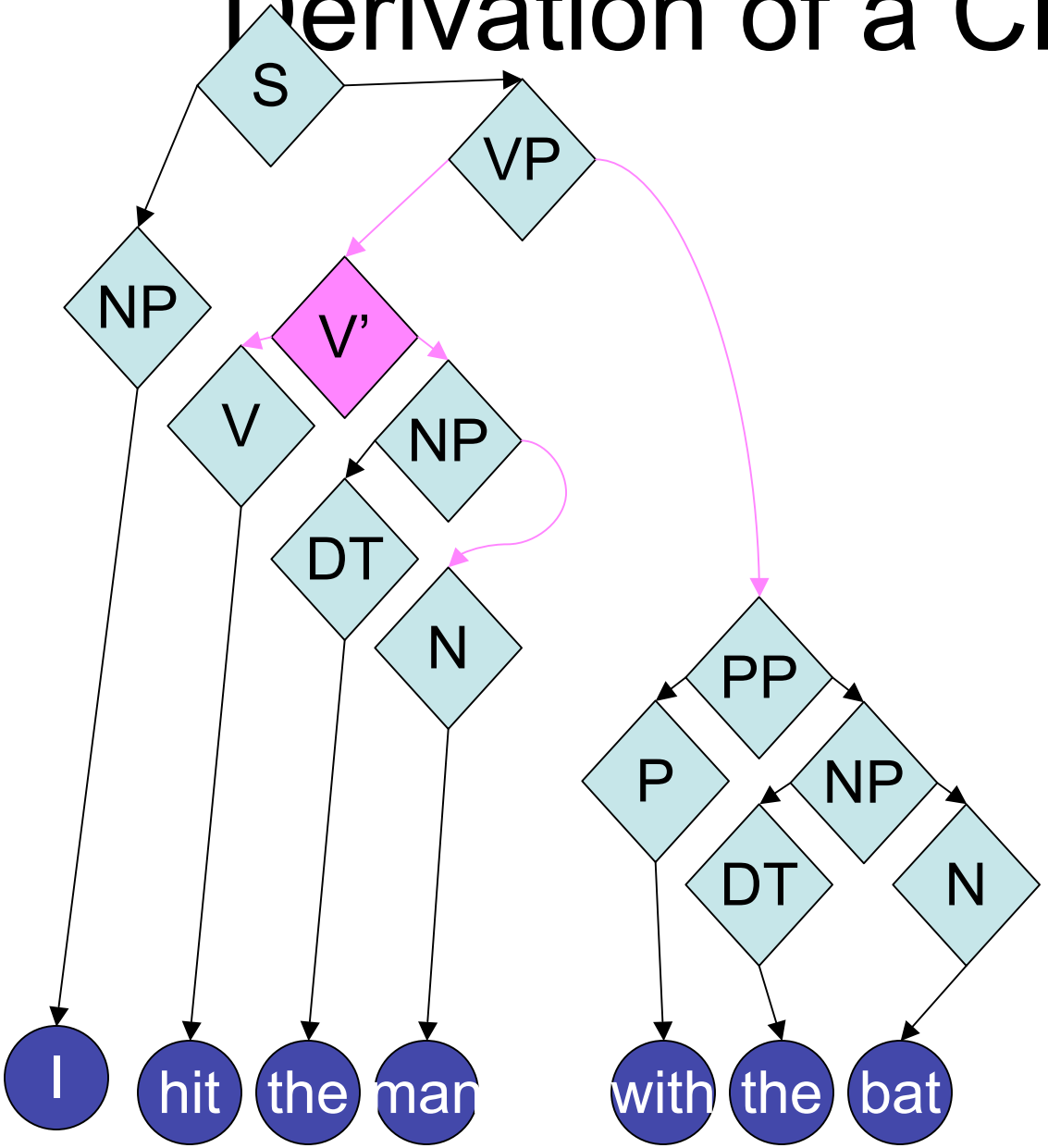


Hey, we could model that with an HMM!

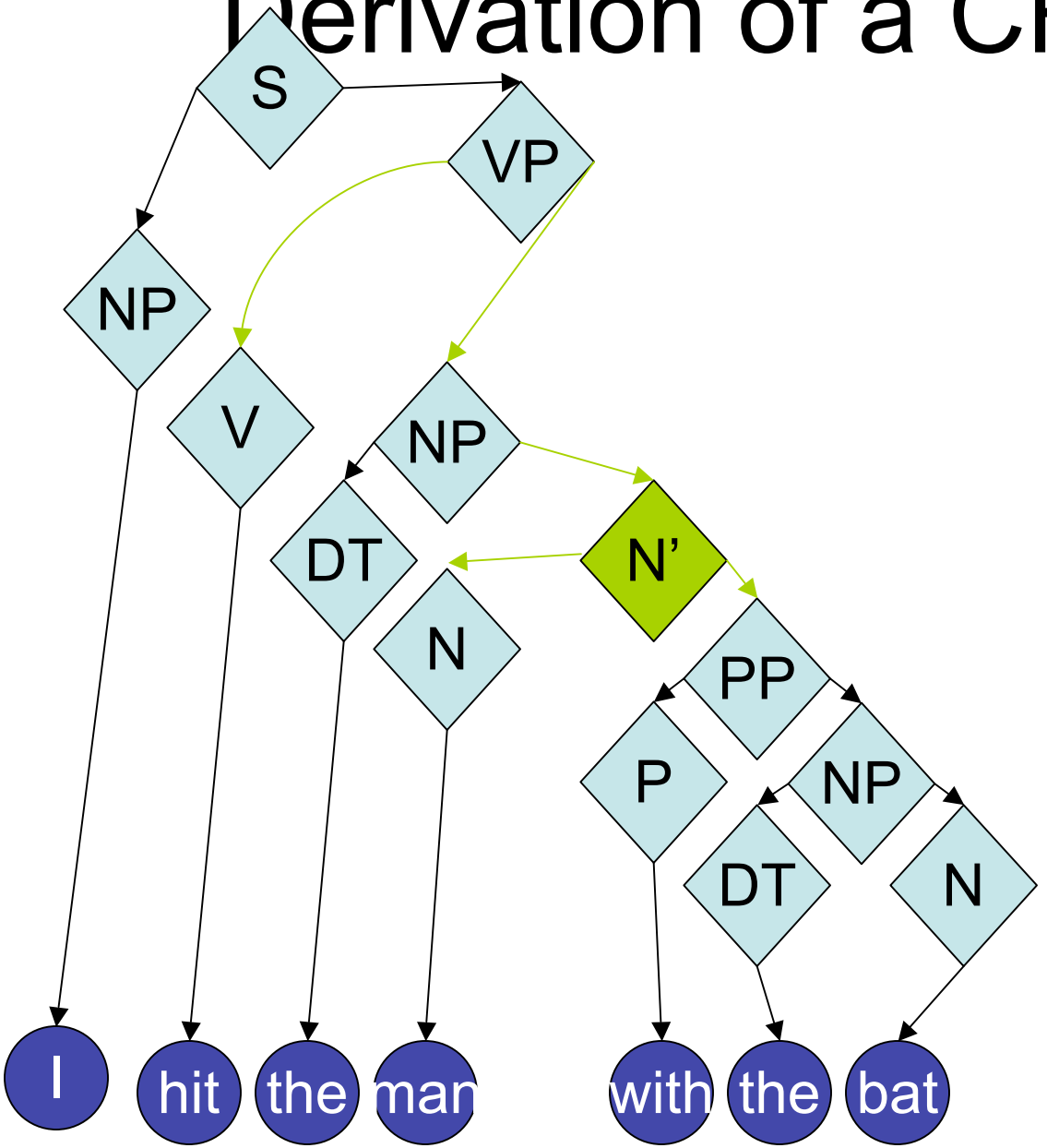
Derivation of a CFG



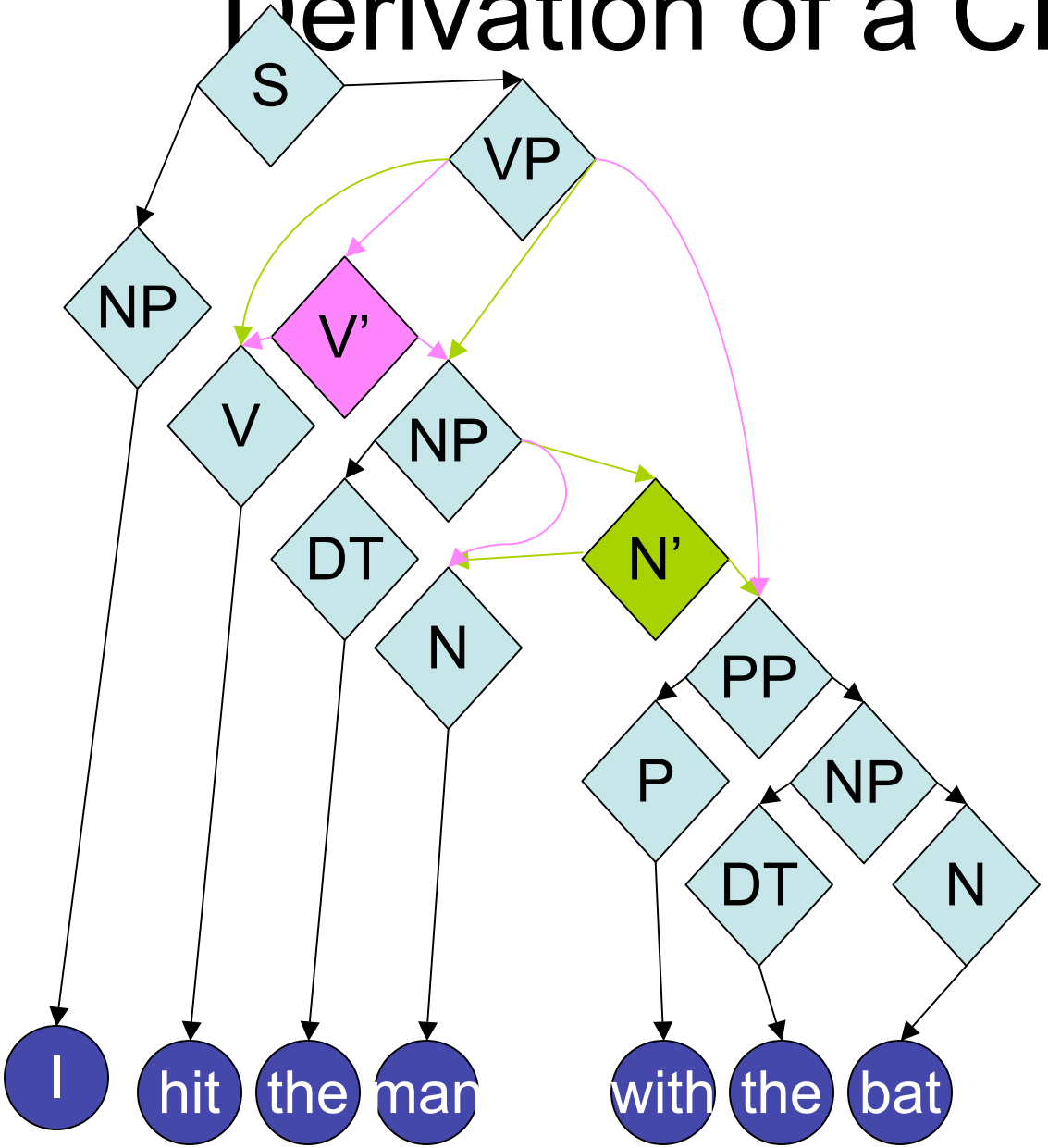
Derivation of a CFG



Derivation of a CFG



Derivation of a CFG



Disambiguation

S → NP VP

NP → I

V → hit

DT → the (2)

N → man

PP → P NP

P → with

NP → DT N

N → bat

VP → V NP

NP → DT N'

N' → N PP

VP → V' PP

NP → DT N

V' → V NP

Probabilistic CFG

- Alphabet Σ
- Set of variables N
- Start symbol $S \in N$
- Rewrite rules: $X \rightarrow^{p} \alpha$,
  where $X \in N$, $\alpha \in (N \cup \Sigma)^*$, and $p \in \mathbb{R}_{\geq 0}$.
  - All ps for a given X sum to one.

CNF: Assume $\alpha \in N^2 \cup \Sigma$.

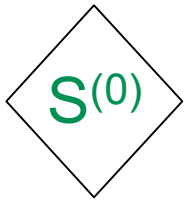
PCFG as a Generative Process

- Instantiate the start symbol S ($S^{(0)}$); $Q \leftarrow \{S^{(0)}\}$; $t \leftarrow 0$
- While $Q \neq \{\}$:
  - $t \leftarrow t + 1$
  - Pop a symbol $X^{(a)}$ off of Q.
  - Draw $\alpha = \langle \alpha_1, \alpha_2, \ldots, \alpha_k \rangle$ according to the distribution defined by $X \rightarrow^p \bullet$.
  - Add $\langle \alpha_1^{(t)}, \alpha_2^{(t)}, \ldots, \alpha_k^{(t)} \rangle$ to the tree as the sequence of children of $X^{(a)}$.
  - For each $\alpha_i^{(t)}$ that is a nonterminal, push $\alpha_k^{(t)}$ onto Q.

PCFG Example

$Q = \{S^{(0)}\}$

$t = 0$

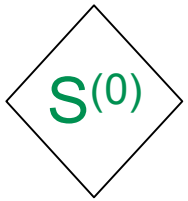


- **Instantiate the start symbol S ($S^{(0)}$); Q ← {$S^{(0)}$}; t ← 0**
- While Q ≠ {}:
  - t ← t + 1
  - Pop a symbol $X^{(a)}$ off of Q.
  - Draw $\alpha = \langle \alpha_1, \alpha_2, \ldots, \alpha_k \rangle$ according to the distribution defined by $X \rightarrow^p \bullet$.
  - Add $\langle \alpha_1^{(t)}, \alpha_2^{(t)}, \ldots, \alpha_k^{(t)} \rangle$ to the tree as the sequence of children of $X^{(a)}$.
  - For each $\alpha_i^{(t)}$ that is a nonterminal, push $\alpha_k^{(t)}$ onto Q.

PCFG Example

$Q = \{S^{(0)}\}$

$t = 0$

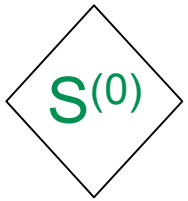


- Instantiate the start symbol S ($S^{(0)}$); $Q \leftarrow \{S^{(0)}\}$; $t \leftarrow 0$
- **While $Q \neq \{\}$:**
  - $t \leftarrow t + 1$
  - Pop a symbol $X^{(a)}$ off of Q.
  - Draw $\alpha = \langle \alpha_1, \alpha_2, \ldots, \alpha_k \rangle$ according to the distribution defined by $X \rightarrow^p \bullet$.
  - Add $\langle \alpha_1^{(t)}, \alpha_2^{(t)}, \ldots, \alpha_k^{(t)} \rangle$ to the tree as the sequence of children of $X^{(a)}$.
  - For each $\alpha_i^{(t)}$ that is a nonterminal, push $\alpha_k^{(t)}$ onto Q.

PCFG Example

$Q = \{S^{(0)}\}$

$t = 1$



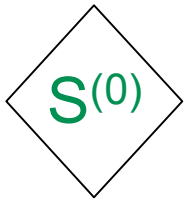
- Instantiate the start symbol S ($S^{(0)}$); Q ← $\{S^{(0)}\}$; t ← 0
- While Q ≠ {}:
  - **t ← t + 1**
  - Pop a symbol $X^{(a)}$ off of Q.
  - Draw $\alpha = \langle \alpha_1, \alpha_2, \ldots, \alpha_k \rangle$ according to the distribution defined by $X \rightarrow^p \bullet$.
  - Add $\langle \alpha_1^{(t)}, \alpha_2^{(t)}, \ldots, \alpha_k^{(t)} \rangle$ to the tree as the sequence of children of $X^{(a)}$.
  - For each $\alpha_i^{(t)}$ that is a nonterminal, push $\alpha_k^{(t)}$ onto Q.

PCFG Example

Q = {}

t = 1

$S^{(0)}$

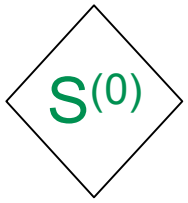


- Instantiate the start symbol S ($S^{(0)}$); Q ← {$S^{(0)}$}; t ← 0
- While Q ≠ {}:
  - t ← t + 1
  - **Pop a symbol $X^{(a)}$ off of Q.**
  - Draw $\alpha = \langle \alpha_1, \alpha_2, \ldots, \alpha_k \rangle$ according to the distribution defined by $X \rightarrow^p \bullet$.
  - Add $\langle \alpha_1^{(t)}, \alpha_2^{(t)}, \ldots, \alpha_k^{(t)} \rangle$ to the tree as the sequence of children of $X^{(a)}$.
  - For each $\alpha_i^{(t)}$ that is a nonterminal, push $\alpha_k^{(t)}$ onto Q.

PCFG Example

Q = {}

t = 1



$S^{(0)}$

$S \rightarrow^{0.84} NP\ VP$

- Instantiate the start symbol S ($S^{(0)}$); Q ← {$S^{(0)}$}; t ← 0
- While Q ≠ {}:
  - t ← t + 1
  - Pop a symbol $X^{(a)}$ off of Q.
  - **Draw $\alpha = \langle \alpha_1, \alpha_2, \ldots, \alpha_k \rangle$ according to the distribution defined by $X \rightarrow^{p} \bullet$.**
  - Add $\langle \alpha_1^{(t)}, \alpha_2^{(t)}, \ldots, \alpha_k^{(t)} \rangle$ to the tree as the sequence of children of $X^{(a)}$.
  - For each $\alpha_i^{(t)}$ that is a nonterminal, push $\alpha_k^{(t)}$ onto Q.

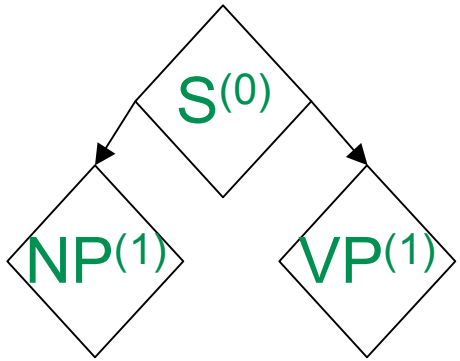
PCFG Example

Q = {}

t = 1

$S^{(0)}$

$S \rightarrow^{0.84} NP\ VP$



- Instantiate the start symbol S ($S^{(0)}$); Q ← {$S^{(0)}$}; t ← 0
- While Q ≠ {}:
  - t ← t + 1
  - Pop a symbol $X^{(a)}$ off of Q.
  - Draw $\alpha = \langle \alpha_1, \alpha_2, \ldots, \alpha_k \rangle$ according to the distribution defined by $X \rightarrow^p \bullet$.
  - **Add $\langle \alpha_1^{(t)}, \alpha_2^{(t)}, \ldots, \alpha_k^{(t)} \rangle$ to the tree as the sequence of children of $X^{(a)}$.**
  - For each $\alpha_i^{(t)}$ that is a nonterminal, push $\alpha_k^{(t)}$ onto Q.

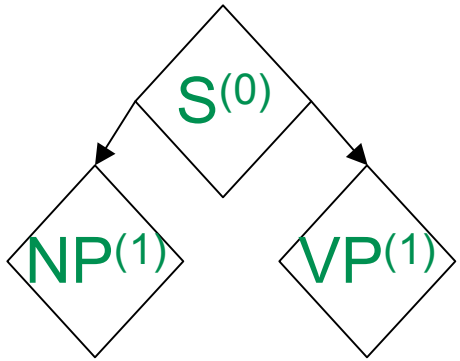
PCFG Example

$Q = \{NP^{(1)}, VP^{(1)}\}$

$t = 1$

$S^{(0)}$

$S \rightarrow^{0.84} NP\ VP$

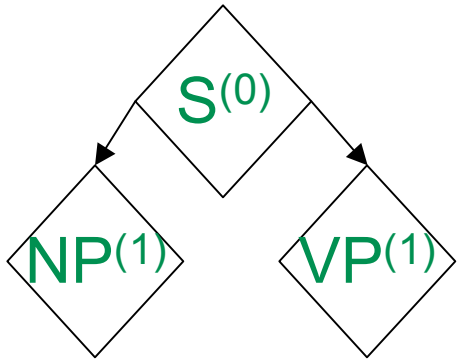


- Instantiate the start symbol S ($S^{(0)}$); $Q \leftarrow \{S^{(0)}\}$; $t \leftarrow 0$
- While $Q \neq \{\}$:
  - $t \leftarrow t + 1$
  - Pop a symbol $X^{(a)}$ off of Q.
  - Draw $\alpha = \langle \alpha_1, \alpha_2, \ldots, \alpha_k \rangle$ according to the distribution defined by $X \rightarrow^{p} \bullet$.
  - Add $\langle \alpha_1^{(t)}, \alpha_2^{(t)}, \ldots, \alpha_k^{(t)} \rangle$ to the tree as the sequence of children of $X^{(a)}$.
  - **For each $\alpha_i^{(t)}$ that is a nonterminal, push $\alpha_k^{(t)}$ onto Q.**

PCFG Example

$Q = \{NP^{(1)}, VP^{(1)}\}$

$t = 1$

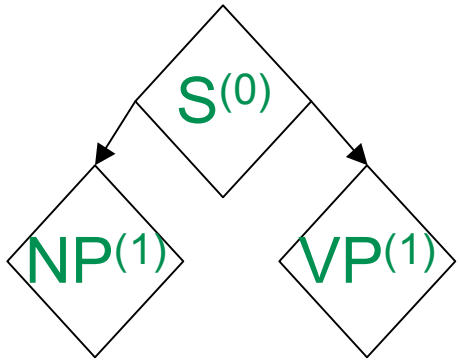


- Instantiate the start symbol S ($S^{(0)}$); $Q \leftarrow \{S^{(0)}\}$; $t \leftarrow 0$
- **While $Q \neq \{\}$:**
  - $t \leftarrow t + 1$
  - Pop a symbol $X^{(a)}$ off of Q.
  - Draw $\alpha = \langle \alpha_1, \alpha_2, \ldots, \alpha_k \rangle$ according to the distribution defined by $X \rightarrow^p \bullet$.
  - Add $\langle \alpha_1^{(t)}, \alpha_2^{(t)}, \ldots, \alpha_k^{(t)} \rangle$ to the tree as the sequence of children of $X^{(a)}$.
  - For each $\alpha_i^{(t)}$ that is a nonterminal, push $\alpha_k^{(t)}$ onto Q.

PCFG Example

$Q = \{NP^{(1)}, VP^{(1)}\}$

$t = 2$



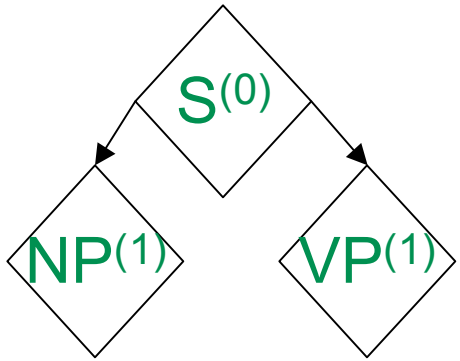
- Instantiate the start symbol S ($S^{(0)}$); Q ← $\{S^{(0)}\}$; t ← 0
- While Q ≠ {}:
  - **t ← t + 1**
  - Pop a symbol $X^{(a)}$ off of Q.
  - Draw $\alpha = \langle \alpha_1, \alpha_2, \ldots, \alpha_k \rangle$ according to the distribution defined by $X \rightarrow^p \bullet$.
  - Add $\langle \alpha_1^{(t)}, \alpha_2^{(t)}, \ldots, \alpha_k^{(t)} \rangle$ to the tree as the sequence of children of $X^{(a)}$.
  - For each $\alpha_i^{(t)}$ that is a nonterminal, push $\alpha_k^{(t)}$ onto Q.

PCFG Example

$Q = \{NP^{(1)}\}$

$t = 2$

$VP^{(1)}$



- Instantiate the start symbol S ($S^{(0)}$); $Q \leftarrow \{S^{(0)}\}$; $t \leftarrow 0$
- While $Q \neq \{\}$:
  - $t \leftarrow t + 1$
  - **Pop a symbol $X^{(a)}$ off of Q.**
  - Draw $\alpha = \langle \alpha_1, \alpha_2, \ldots, \alpha_k \rangle$ according to the distribution defined by $X \rightarrow^p \bullet$.
  - Add $\langle \alpha_1^{(t)}, \alpha_2^{(t)}, \ldots, \alpha_k^{(t)} \rangle$ to the tree as the sequence of children of $X^{(a)}$.
  - For each $\alpha_i^{(t)}$ that is a nonterminal, push $\alpha_k^{(t)}$ onto Q.

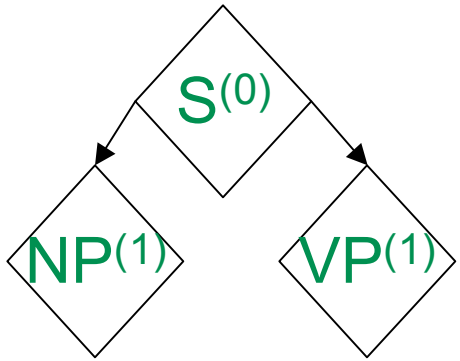
PCFG Example

$Q = \{NP^{(1)}\}$

$t = 2$

$VP^{(1)}$

$VP \rightarrow^{0.12} V$



- Instantiate the start symbol S ($S^{(0)}$); $Q \leftarrow \{S^{(0)}\}$; $t \leftarrow 0$
- While $Q \neq \{\}$:
  - $t \leftarrow t + 1$
  - Pop a symbol $X^{(a)}$ off of Q.
  - **Draw $\alpha = \langle \alpha_1, \alpha_2, \ldots, \alpha_k \rangle$ according to the distribution defined by $X \rightarrow^{p} \bullet$.**
  - Add $\langle \alpha_1^{(t)}, \alpha_2^{(t)}, \ldots, \alpha_k^{(t)} \rangle$ to the tree as the sequence of children of $X^{(a)}$.
  - For each $\alpha_i^{(t)}$ that is a nonterminal, push $\alpha_k^{(t)}$ onto Q.

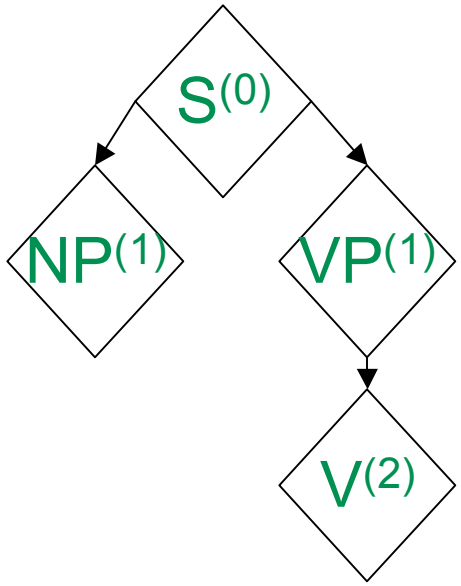
PCFG Example

$Q = \{NP^{(1)}\}$

$t = 2$

$VP^{(1)}$

$VP \rightarrow^{0.12} V$



- Instantiate the start symbol S ($S^{(0)}$); $Q \leftarrow \{S^{(0)}\}$; $t \leftarrow 0$
- While $Q \neq \{\}$:
  - $t \leftarrow t + 1$
  - Pop a symbol $X^{(a)}$ off of Q.
  - Draw $\alpha = \langle \alpha_1, \alpha_2, \ldots, \alpha_k \rangle$ according to the distribution defined by $X \rightarrow^p \bullet$.
  - **Add $\langle \alpha_1^{(t)}, \alpha_2^{(t)}, \ldots, \alpha_k^{(t)} \rangle$ to the tree as the sequence of children of $X^{(a)}$.**
  - For each $\alpha_i^{(t)}$ that is a nonterminal, push $\alpha_k^{(t)}$ onto Q.

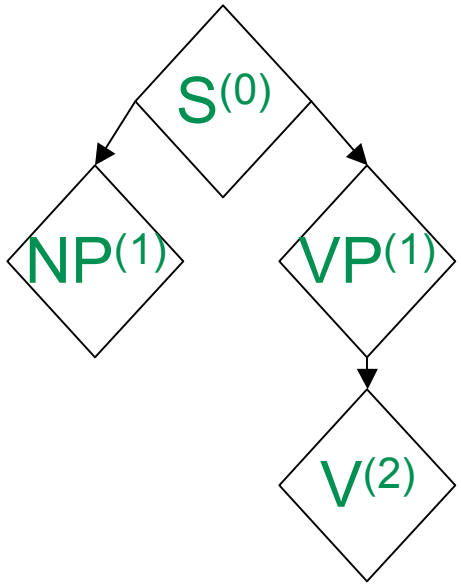
PCFG Example

$Q = \{NP^{(1)}, V^{(2)}\}$

$t = 2$

$VP^{(1)}$

$VP \rightarrow^{0.12} V$

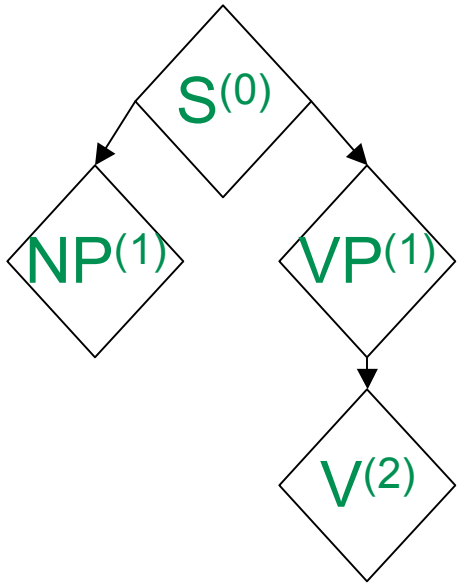


- Instantiate the start symbol S ($S^{(0)}$); $Q \leftarrow \{S^{(0)}\}$; $t \leftarrow 0$
- While $Q \neq \{\}$:
  - $t \leftarrow t + 1$
  - Pop a symbol $X^{(a)}$ off of Q.
  - Draw $\alpha = \langle \alpha_1, \alpha_2, \ldots, \alpha_k \rangle$ according to the distribution defined by $X \rightarrow^p \bullet$.
  - Add $\langle \alpha_1^{(t)}, \alpha_2^{(t)}, \ldots, \alpha_k^{(t)} \rangle$ to the tree as the sequence of children of $X^{(a)}$.
  - **For each $\alpha_i^{(t)}$ that is a nonterminal, push $\alpha_k^{(t)}$ onto Q.**

PCFG Example

$Q = \{NP^{(1)}, V^{(2)}\}$

$t = 2$



- Instantiate the start symbol S ($S^{(0)}$); $Q \leftarrow \{S^{(0)}\}$; $t \leftarrow 0$
- **While $Q \neq \{\}$:**
  - $t \leftarrow t + 1$
  - Pop a symbol $X^{(a)}$ off of Q.
  - Draw $\alpha = \langle \alpha_1, \alpha_2, \ldots, \alpha_k \rangle$ according to the distribution defined by $X \rightarrow^p \bullet$.
  - Add $\langle \alpha_1^{(t)}, \alpha_2^{(t)}, \ldots, \alpha_k^{(t)} \rangle$ to the tree as the sequence of children of $X^{(a)}$.
  - For each $\alpha_i^{(t)}$ that is a nonterminal, push $\alpha_k^{(t)}$ onto Q.

Mathematical Properties

- The probability of generating a tree is simply a product of the rule probabilities for all rule tokens in the tree.
- Given a tree, it's O(n) to compute the probability.
- The queueing policy doesn't matter as long as it's consistent. It doesn't affect the probabilities.
- Independence assumption?