Viterbi Training Improves Unsupervised Dependency Parsing

Valentin I. Spitkovsky

with **Hiyan Alshawi** (Google Inc.)
Daniel Jurafsky (Stanford University)
and **Christopher D. Manning** (Stanford University)

Outline

Outline

1. **Viterbi EM**

Outline

1. **Viterbi EM**
   — **faster**, **simpler** and more **accurate**

Outline

1. **Viterbi EM**
   — faster, simpler and more accurate
   — easy state-of-the-art results

Outline

1. **Viterbi EM**
   — faster, simpler and more accurate
   — easy state-of-the-art results
2. **Interpretation**

Outline

1. **Viterbi EM**
   — faster, simpler and more accurate
   — easy state-of-the-art results

2. **Interpretation**
   — machine learning and linguistic perspectives

Outline

1. **Viterbi EM**
   — faster, simpler and more accurate
   — easy state-of-the-art results

2. **Interpretation**
   — machine learning and linguistic perspectives
   — practical insights (some theoretical underpinning)

Outline

1. **Viterbi EM**
   — faster, simpler and more accurate
   — easy state-of-the-art results

2. **Interpretation**
   — machine learning and linguistic perspectives
   — practical insights (some theoretical underpinning)

3. **Core Issue**

Outline

1. **Viterbi EM**
   — faster, simpler and more accurate
   — easy state-of-the-art results
2. **Interpretation**
   — machine learning and linguistic perspectives
   — practical insights (some theoretical underpinning)
3. **Core Issue**
   — provably wrong objective functions

Outline

1. **Viterbi EM**
   — faster, simpler and more accurate
   — easy state-of-the-art results

2. **Interpretation**
   — machine learning and linguistic perspectives
   — practical insights (some theoretical underpinning)

3. **Core Issue**
   — provably wrong objective functions
   — theoretical insights (mathematically sound)

Problem: Unsupervised Learning of Parsing

Problem: Unsupervised Learning of Parsing

- **Input: Raw Text**

  ... By most measures, the nation's industrial sector is now growing very slowly — if at all. Factory payrolls fell in September. So did the Federal Reserve ...

Problem: Unsupervised Learning of Parsing

- **Input: Raw Text (Sentences, Tokens and POS-tags)**

  ... By most measures, the nation's industrial sector is now growing very slowly — if at all. Factory payrolls fell in September. So did the Federal Reserve ...

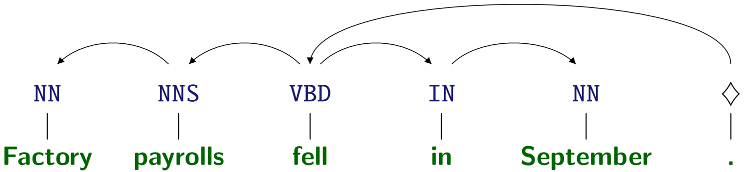
NN	NNS	VBD	IN	NN	♢
|	|	|	|	|	|
Factory	**payrolls**	**fell**	**in**	**September**	**.**

Problem: Unsupervised Learning of Parsing

- **Input: Raw Text (Sentences, Tokens and POS-tags)**

  ... By most measures, the nation's industrial sector is now growing very slowly — if at all. Factory payrolls fell in September. So did the Federal Reserve ...

- **Output: Syntactic Structures (and a Probabilistic Grammar)**



Disclaimer: Your Mileage May Vary...

Disclaimer: Your Mileage May Vary...

- **our scope is a *very* specific problem**

Disclaimer: Your Mileage May Vary...

- **our scope is a *very* specific problem**
- **but the high-level ideas *may* generalize**

Disclaimer: Your Mileage May Vary...

- **our scope is a *very* specific problem**
- **but the high-level ideas *may* generalize**

- **Classic EM: "focus across the board"**

(hard to see the trees for the forest)

Disclaimer: Your Mileage May Vary...

- **our scope is a *very* specific problem**
- **but the high-level ideas *may* generalize**

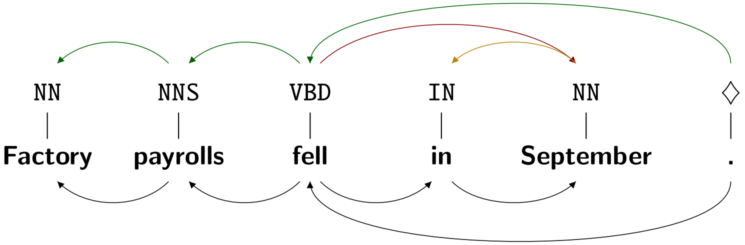
- **Classic EM: "focus across the board"**

  (hard to see the trees for the forest)

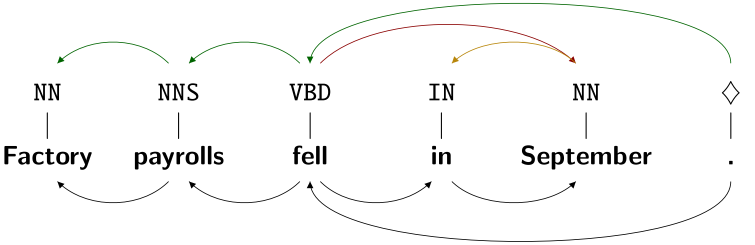
- **Viterbi EM: zoom in on likeliest tree**

Scoring: Directed Dependency Accuracy

Scoring: Directed Dependency Accuracy

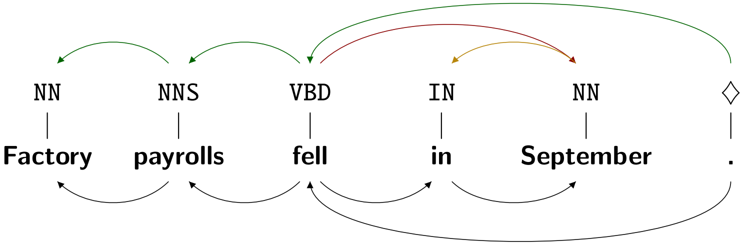


Scoring: Directed Dependency Accuracy



Directed score: $\frac{3}{5} = 60\%$

Scoring: Directed Dependency Accuracy



Directed score: $\frac{3}{5} = 60\%$ **(right/left-branching baselines:** $\frac{2}{5} = 40\%$**).**

State-of-the-Art: Dependency Model with Valence

State-of-the-Art: Dependency Model with Valence

- **a head-outward model, with word classes and valence/adjacency** **(Klein and Manning, 2004)**

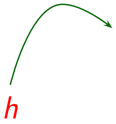
State-of-the-Art: Dependency Model with Valence

- **a head-outward model, with word classes and valence/adjacency** **(Klein and Manning, 2004)**

h

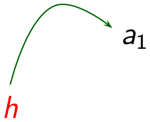
State-of-the-Art: Dependency Model with Valence

- **a head-outward model, with word classes and valence/adjacency** **(Klein and Manning, 2004)**



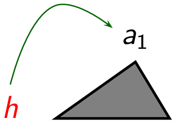
State-of-the-Art: Dependency Model with Valence

- **a head-outward model, with word classes and valence/adjacency** **(Klein and Manning, 2004)**



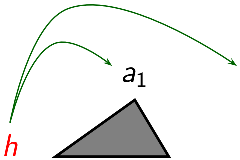
State-of-the-Art: Dependency Model with Valence

- a **head-outward** model, with **word classes** and **valence/adjacency** (Klein and Manning, 2004)



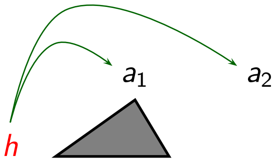
State-of-the-Art: Dependency Model with Valence

- a **head-outward** model, with **word classes** and **valence/adjacency** **(Klein and Manning, 2004)**



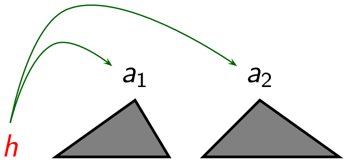
State-of-the-Art: Dependency Model with Valence

- **a head-outward model, with word classes and valence/adjacency** **(Klein and Manning, 2004)**



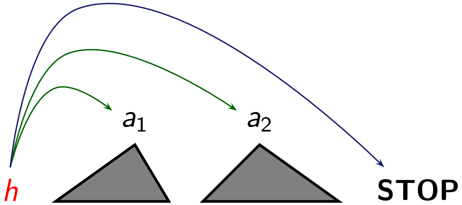
State-of-the-Art: Dependency Model with Valence

- a **head-outward** model, with **word classes** and **valence**/**adjacency** **(Klein and Manning, 2004)**



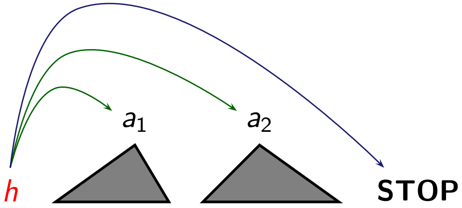
State-of-the-Art: Dependency Model with Valence

- **a head-outward model, with word classes and valence/adjacency** **(Klein and Manning, 2004)**



State-of-the-Art: Dependency Model with Valence

- **a head-outward model, with word classes and valence/adjacency** **(Klein and Manning, 2004)**



$$\mathbb{P}(t_h) = \prod_{dir \in \{\mathrm{L},\mathrm{R}\}} \left[\underline{\mathbb{P}_{\mathrm{STOP}}(c_h, dir, \overbrace{1_{n=0}}^{adj})} \prod_{i=1}^{n} \underline{\mathbb{P}(t_{a_i})}\, \mathbb{P}_{\mathrm{ATTACH}}(c_h, dir, c_{a_i}) \right.$$

$$\left. \underline{(1 - \mathbb{P}_{\mathrm{STOP}}(c_h, dir, \overbrace{1_{i=1}}^{adj}))} \right]_{n=|args(h,dir)|}$$

Learning: EM, via inside-outside re-estimation

Learning: EM, via inside-outside re-estimation

- **sentences** $\{s\}$

Learning: EM, via inside-outside re-estimation

- **sentences** $\{s\}$**, legal parse trees** $t \in T(s)$

Learning: EM, via inside-outside re-estimation

- **sentences $\{s\}$, legal parse trees $t \in T(s)$, and a gold t^***

Learning: EM, via inside-outside re-estimation

- **sentences $\{s\}$, legal parse trees $t \in T(s)$, and a gold t^***

- **non-convex objective — very sensitive to initialization**

Learning: EM, via inside-outside re-estimation

- **sentences $\{s\}$, legal parse trees $t \in T(s)$, and a gold t^***
- **non-convex objective — very sensitive to initialization**
- **maximizing the probability of data (sentence strings):**

$$\hat{\theta}_{\text{UNS}} = \arg\max_{\theta} \prod_{s} \underbrace{\sum_{t \in T(s)} \mathbb{P}_{\theta}(t)}_{\mathbb{P}_{\theta}(s)}$$

Learning: EM, via inside-outside re-estimation

- **sentences $\{s\}$, legal parse trees $t \in T(s)$, and a gold t^***
- **non-convex objective — very sensitive to initialization**
- **maximizing the probability of data (sentence strings):**

$$\hat{\theta}_{\text{UNS}} = \arg\max_{\theta} \prod_{s} \underbrace{\sum_{t \in T(s)} \mathbb{P}_{\theta}(t)}_{\mathbb{P}_{\theta}(s)}$$

- **supervised objective would be convex (counting):**

$$\hat{\theta}_{\text{SUP}} = \arg\max_{\theta} \prod_{s} \mathbb{P}_{\theta}(t^*(s))$$

Standard Corpus: WSJ*k*

Standard Corpus: WSJ*k*

- *The Wall Street Journal* **section of the Penn Treebank Project (Marcus et al., 1993)**

Standard Corpus: WSJk

- *The Wall Street Journal* **section of the Penn Treebank Project** **(Marcus et al., 1993)**
  - **... stripped of punctuation, etc.**

Standard Corpus: WSJk

- *The Wall Street Journal* **section of the Penn Treebank Project** **(Marcus et al., 1993)**
  - **... stripped of punctuation, etc.**
  - **... rid of sentences left with more than k POS tags;**

Standard Corpus: WSJk

- *The Wall Street Journal* **section of the Penn Treebank Project (Marcus et al., 1993)**
  - **... stripped of punctuation, etc.**
  - **... rid of sentences left with more than k POS tags;**
  - **... and converted to reference dependencies — $\{t^*\}$, using "head percolation rules" (Collins, 1999).**

Standard Corpus: WSJk

- *The Wall Street Journal* **section of the Penn Treebank Project** (Marcus et al., 1993)
  - **... stripped of punctuation, etc.**
  - **... rid of sentences left with more than k POS tags;**
  - **... and converted to reference dependencies — $\{t^*\}$, using "head percolation rules"** (Collins, 1999).

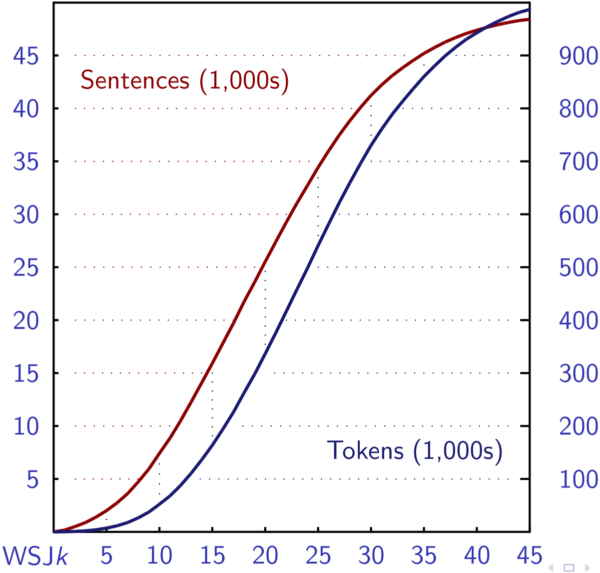
- **Training: traditionally, WSJ10** (Klein, 2005);

Standard Corpus: WSJk

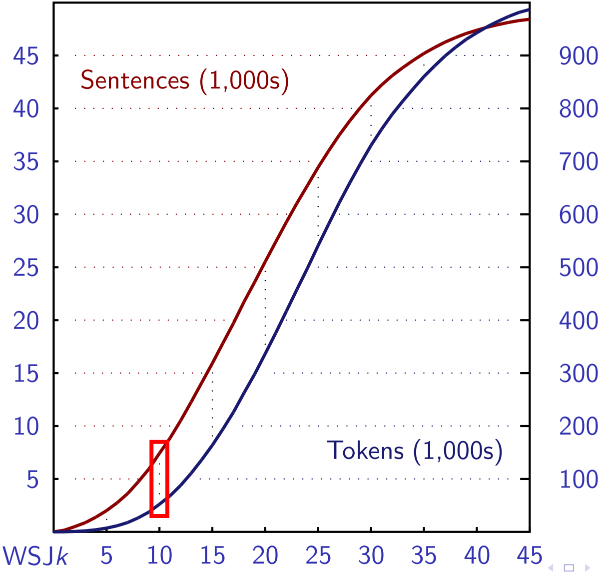
- *The Wall Street Journal* **section of the Penn Treebank Project (Marcus et al., 1993)**
  - **... stripped of punctuation, etc.**
  - **... rid of sentences left with more than k POS tags;**
  - **... and converted to reference dependencies — $\{t^*\}$, using "head percolation rules" (Collins, 1999).**

- **Training: traditionally, WSJ10 (Klein, 2005);**
- **Evaluation: Section 23 of WSJ$^\infty$ (all sentences).**

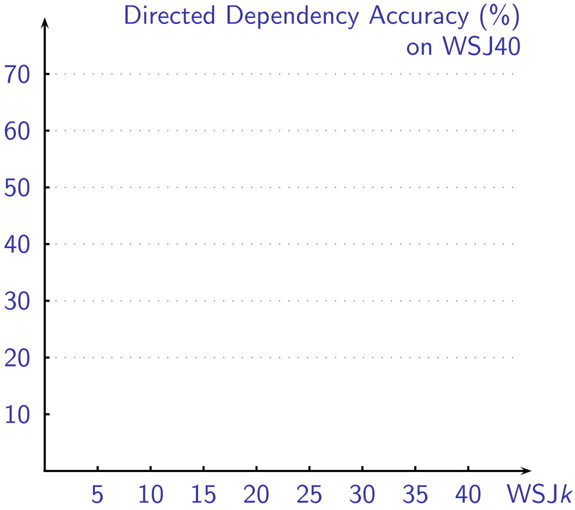
Standard Corpus: WSJk



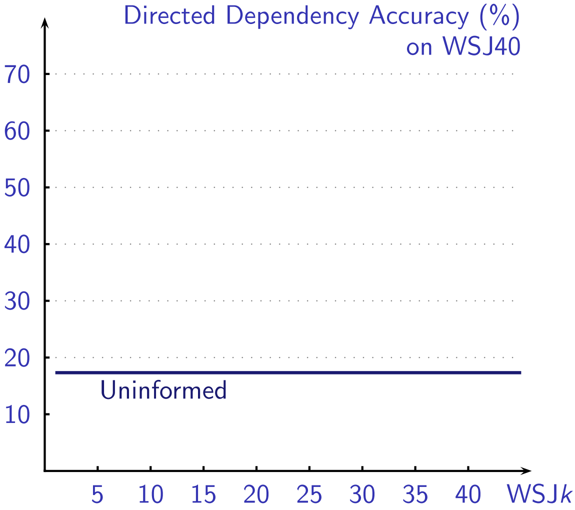
Standard Corpus: WSJk



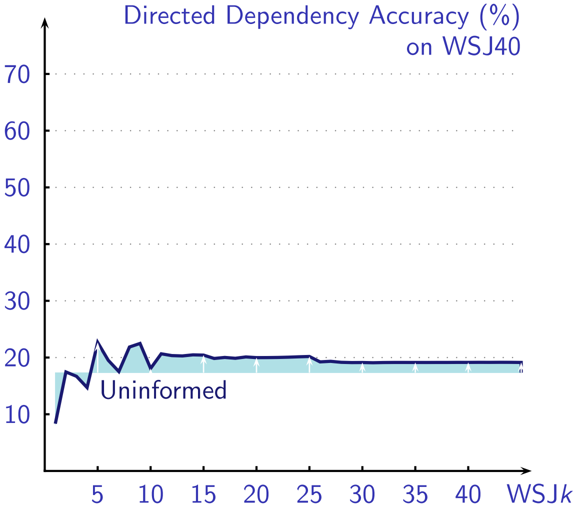
Classic EM: The Lay of the Land



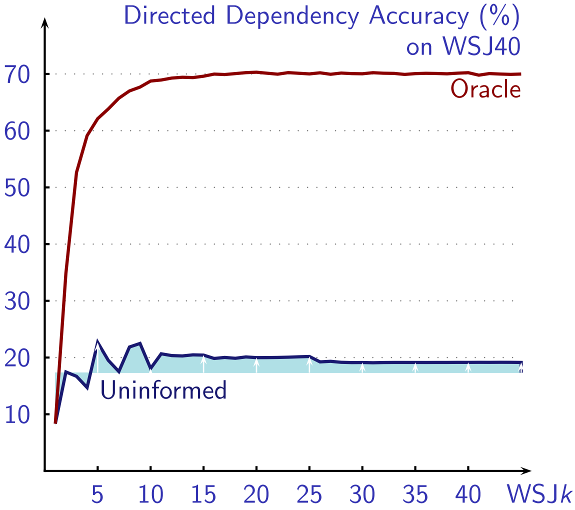
Classic EM: The Lay of the Land



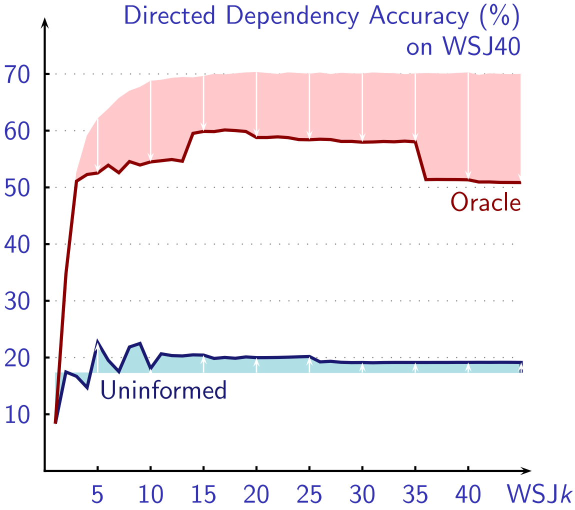
Classic EM: The Lay of the Land



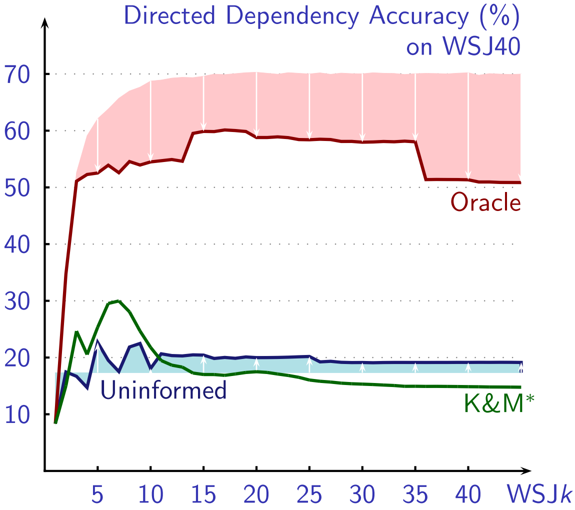
Classic EM: The Lay of the Land



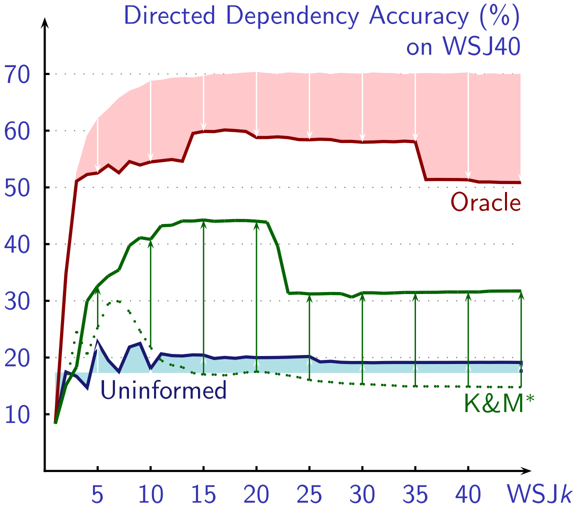
Classic EM: The Lay of the Land



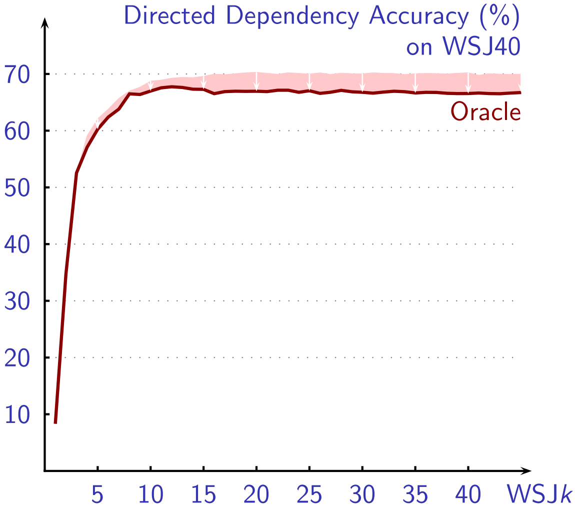
Classic EM: The Lay of the Land



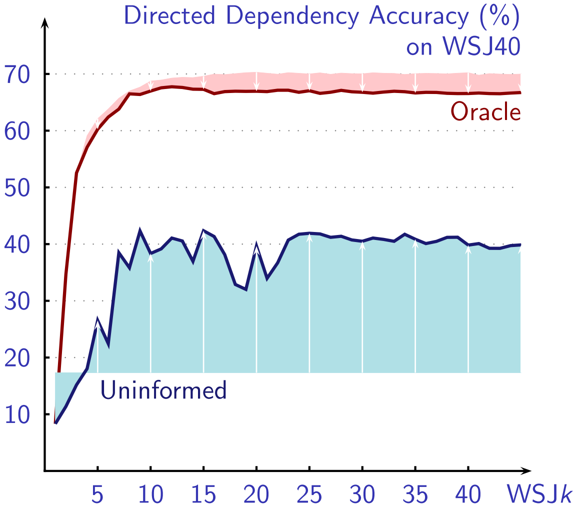
Classic EM: The Lay of the Land



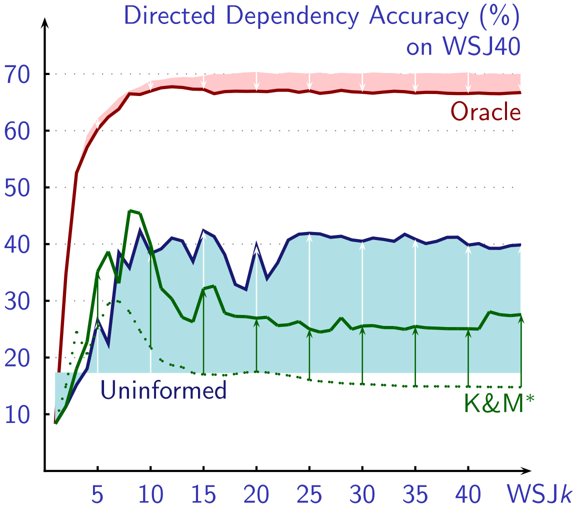
Viterbi EM: Results!



Viterbi EM: Results!



Viterbi EM: Results!



State-of-the-Art

State-of-the-Art

Section 23 of WSJ$^\infty$	
Right-Branching **Baseline**	
(Klein and Manning, 2004)	**32%**

State-of-the-Art

	Section 23 of WSJ$^\infty$
Right-Branching Baseline	
(Klein and Manning, 2004)	**32%**
DMV with Classic EM	
(Klein and Manning, 2004)	**34%**
(Spitkovsky et al., 2010)	**45%**

State-of-the-Art

Section 23 of WSJ$^\infty$

Right-Branching Baseline	
(Klein and Manning, 2004)	**32%**
DMV with Classic EM	
(Klein and Manning, 2004)	**34%**
(Spitkovsky et al., 2010)	**45%**
DMV with Viterbi EM	
with Smoothing	**45%**

State-of-the-Art

	Section 23 of WSJ$^\infty$	
Right-Branching Baseline		
(Klein and Manning, 2004)	**32%**	
DMV with Classic EM		
(Klein and Manning, 2004)	**34%**	
(Spitkovsky et al., 2010)	**45%**	
DMV with Viterbi EM		
with Smoothing	**45%**	
+ Clever Initialization	**48%**	

State-of-the-Art

	Section 23 of WSJ^{∞}	Brown100
Right-Branching Baseline		
(Klein and Manning, 2004)	**32%**	
DMV with Classic EM		
(Klein and Manning, 2004)	**34%**	
(Spitkovsky et al., 2010)	**45%**	**43%**
DMV with Viterbi EM		
with Smoothing	**45%**	**48%**
+ Clever Initialization	**48%**	**51%**

State-of-the-Art

	Section 23 of WSJ$^\infty$	Brown100	
Right-Branching Baseline			
(Klein and Manning, 2004)	32%		
DMV with Classic EM			
(Klein and Manning, 2004)	34%		
(Spitkovsky et al., 2010)	45%	43%	
DMV with Viterbi EM			
with Smoothing	45%	48%	($^+$5%)
+ Clever Initialization	48%	51%	

State-of-the-Art

	Section 23 of WSJ^{∞}	Brown100	
Right-Branching Baseline			
(Klein and Manning, 2004)	32%		
DMV with Classic EM			
(Klein and Manning, 2004)	34%		
(Spitkovsky et al., 2010)	45%	43%	
DMV with Viterbi EM			
with Smoothing	45%	48%	($^{+}$5%)
+ Clever Initialization	48%	51%	($^{+}$3%)

Interpretation: Why Does Viterbi EM Work?

Interpretation: Why Does Viterbi EM Work?

- **in theory, Viterbi is a quick-and-dirty approximation**

Interpretation: Why Does Viterbi EM Work?

- **in theory, Viterbi is a quick-and-dirty approximation**
- **in theory, Communism works...**

Interpretation: Why Does Viterbi EM Work?

- **in theory, Viterbi is a quick-and-dirty approximation**
- **in theory, Communism works...**

- **in practice, EM emulates supervised learning:**

Interpretation: Why Does Viterbi EM Work?

- **in theory, Viterbi is a quick-and-dirty approximation**
- **in theory, Communism works...**

- **in practice, EM emulates supervised learning:**

$$s \rightarrow \{t\} = T(s)$$

Interpretation: Why Does Viterbi EM Work?

- **in theory, Viterbi is a quick-and-dirty approximation**
- **in theory, Communism works...**

- **in practice, EM emulates supervised learning:**

$$s \to \{t\} = T(s)$$

Classic EM: $w_t = \mathbb{P}_\theta(t \mid s)$

Interpretation: Why Does Viterbi EM Work?

- **in theory, Viterbi is a quick-and-dirty approximation**
- **in theory, Communism works...**

- **in practice, EM emulates supervised learning:**

$$s \to \{t\} = T(s)$$

Classic EM: $w_t = \mathbb{P}_\theta(t \mid s)$

- **clearly, this is redistribution of ~~wealth~~ mass**

Interpretation: Why Does Viterbi EM Work?

- **in theory, Viterbi is a quick-and-dirty approximation**
- **in theory, Communism works...**

- **in practice, EM emulates supervised learning:**

$$s \to \{t\} = T(s)$$

$$\textbf{Classic EM: } w_t = \mathbb{P}_\theta(t \mid s)$$

- **clearly, this is redistribution of ~~wealth~~ mass**
  — also, resembles an omniscient central planner

Interpretation: Why Does Viterbi EM Work?

- in theory, Viterbi is a quick-and-dirty approximation
- in theory, Communism works...

- in practice, EM emulates supervised learning:

$$s \to \{t\} = T(s)$$

$$\text{Classic EM: } w_t = \mathbb{P}_\theta(t \mid s)$$

- clearly, this is redistribution of ~~wealth~~ mass
  — also, resembles an omniscient central planner
  (knows the true value of everything at all times)

Interpretation: Why Does Viterbi EM Work?

- **in theory, Viterbi is a quick-and-dirty approximation**
- **in theory, Communism works...**

- **in practice, EM emulates supervised learning:**

$$s \rightarrow \{t\} = T(s)$$

Classic EM: $w_t = \mathbb{P}_\theta(t \mid s)$

- **clearly, this is redistribution of ~~wealth~~ mass**
  — also, resembles an omniscient central planner
  (knows the true value of everything at all times)
  — could work, given a very powerful model θ**...**

Interpretation: How Does Classic EM Fail?

Interpretation: How Does Classic EM Fail?

- **our model is quite weak (e.g., doesn't handle agreement)**

Interpretation: How Does Classic EM Fail?

- **our model is quite weak (e.g., doesn't handle agreement)**
- **reserves a lot of mass for ludicrous parse trees...**

Interpretation: How Does Classic EM Fail?

- **our model is quite weak (e.g., doesn't handle agreement)**
- **reserves a lot of mass for ludicrous parse trees...**
  — each entitled to non-trivial support by the distribution

Interpretation: How Does Classic EM Fail?

- **our model is quite weak (e.g., doesn't handle agreement)**
- **reserves a lot of mass for ludicrous parse trees...**
  — each entitled to non-trivial support by the distribution
- **at small scales, this is not a problem (short sentences)**

Interpretation: How Does Classic EM Fail?

- **our model is quite weak (e.g., doesn't handle agreement)**
- **reserves a lot of mass for ludicrous parse trees...**
  — each entitled to non-trivial support by the distribution

- **at small scales, this is not a problem (short sentences)**
  — only so many possible parses → few free-loaders

Interpretation: How Does Classic EM Fail?

- **our model is quite weak (e.g., doesn't handle agreement)**
- **reserves a lot of mass for ludicrous parse trees...**
  — each entitled to non-trivial support by the distribution

- **at small scales, this is not a problem (short sentences)**
  — only so many possible parses $\rightarrow$ few free-loaders
- **eventually, exponentially many trees (unwashed masses)**

Interpretation: How Does Classic EM Fail?

- **our model is quite weak (e.g., doesn't handle agreement)**
- **reserves a lot of mass for ludicrous parse trees...**
  — each entitled to non-trivial support by the distribution

- **at small scales, this is not a problem (short sentences)**
  — only so many possible parses → few free-loaders
- **eventually, exponentially many trees (unwashed masses)**

result: a dog of a probability distribution...

Interpretation: How Does Classic EM Fail?

- our model is quite weak (e.g., doesn't handle agreement)
- reserves a lot of mass for ludicrous parse trees...
  — each entitled to non-trivial support by the distribution

- at small scales, this is not a problem (short sentences)
  — only so many possible parses → few free-loaders
- eventually, exponentially many trees (unwashed masses)

result: a dog of a probability distribution...

Interpretation: How Does Classic EM Fail?

- our **model** is quite **weak** (e.g., doesn't handle agreement)
- reserves a lot of mass for **ludicrous** parse **trees**...
  — each entitled to non-trivial support by the distribution

- at **small scales**, this is not a problem (short sentences)
  — only so many possible parses $\rightarrow$ few free-loaders
- eventually, **exponentially** many trees (unwashed masses)

result: a **dog** of a probability **distribution**...

... **wagged** by its very long **tail**

Interpretation: Idealogical Difference!

Interpretation: Idealogical Difference!

- **Viterbi EM is powered by greed** **(much like Capitalism)**

Interpretation: Idealogical Difference!

- **Viterbi EM is powered by greed (much like Capitalism)**
- **does not require ability to properly value all parse trees**

Interpretation: Idealogical Difference!

- Viterbi EM is powered by greed (much like Capitalism)
- does not require ability to properly value all parse trees
- so long as it can spot a decent one (winner-take-all)

Interpretation: Idealogical Difference!

- **Viterbi EM is powered by greed (much like Capitalism)**
- **does not require ability to properly value all parse trees**
- **so long as it can spot a decent one (winner-take-all)**

- **different (weaker?) requirement on models: (like IR)**

Interpretation: Idealogical Difference!

- **Viterbi EM is powered by greed (much like Capitalism)**
- **does not require ability to properly value all parse trees**
- **so long as it can spot a decent one (winner-take-all)**

- **different (weaker?) requirement on models: (like IR)**
  — θ needs to be just discriminative enough! (ranking)

Interpretation: Idealogical Difference!

- **Viterbi EM is powered by greed (much like Capitalism)**
- **does not require ability to properly value all parse trees**
- **so long as it can spot a decent one (winner-take-all)**

- **different (weaker?) requirement on models: (like IR)**
  — θ needs to be just discriminative enough! (ranking)

- **at small scales, data are too sparse (markets are illiquid)**

Interpretation: Idealogical Difference!

- **Viterbi EM is powered by greed** **(much like Capitalism)**
- **does not require ability to properly value all parse trees**
- **so long as it can spot a decent one** **(winner-take-all)**

- **different (weaker?) requirement on models:** **(like IR)**
  — θ needs to be just discriminative enough! **(ranking)**

- **at small scales, data are too sparse** **(markets are illiquid)**
- **improves with more data** **(statistics become efficient)**

Interpretation: Idealogical Difference!

- **Viterbi EM is powered by greed (much like Capitalism)**
- **does not require ability to properly value all parse trees**
- **so long as it can spot a decent one (winner-take-all)**

- **different (weaker?) requirement on models: (like IR)**
  — θ needs to be just discriminative enough! (ranking)

- **at small scales, data are too sparse (markets are illiquid)**
- **improves with more data (statistics become efficient)**
  — really, what we want from unsupervised learners!

Interpretation: Summary

Interpretation: Summary

- **Viterbi EM: focus on the individual best parse trees**

Interpretation: Summary

- **Viterbi EM: focus on the individual best parse trees — given a decent estimate,**
  makes rapid progress (the rich get richer)

Interpretation: Summary

- **Viterbi EM: focus on the individual best parse trees — given a decent estimate, makes rapid progress (the rich get richer)**
- **Classic EM: integrates over the collective forests**

Interpretation: Summary

- **Viterbi EM: focus on the individual best parse trees — given a decent estimate, makes rapid progress (the rich get richer)**
- **Classic EM: integrates over the collective forests — given a bad (uniform) estimate, makes little progress (all trees remain equally poor)**

Interpretation: Summary

- **Viterbi EM: focus on the individual best parse trees**
  — given a decent estimate,
  makes rapid progress (the rich get richer)

- **Classic EM: integrates over the collective forests**
  — given a bad (uniform) estimate,
  makes little progress (all trees remain equally poor)

  — given a great (supervised) estimate,
  cuts down the better trees (Dekulakization)


THE PARTY'S OVER

Interpretation: Connections

Interpretation: Connections

- **"learning by doing" — (unsupervised) self-training**
  (Clark et al., 2003; Ng and Cardie, 2003; McClosky et al., 2006)

Interpretation: Connections

- "learning by doing" — (unsupervised) self-training
  (Clark et al., 2003; Ng and Cardie, 2003; McClosky et al., 2006)
  — relevance to understanding language acquisition?

Interpretation: Connections

- **"learning by doing" — (unsupervised) self-training**
  (Clark et al., 2003; Ng and Cardie, 2003; McClosky et al., 2006)
  — relevance to understanding language acquisition?
  — human probabilistic parsing models massively pruned
  (Jurafsky, 1996; Chater et al., 1998; Lewis and Vasishth, 2005)

Interpretation: Connections

- “learning by doing” — (unsupervised) self-training
  (Clark et al., 2003; Ng and Cardie, 2003; McClosky et al., 2006)
  — relevance to understanding language acquisition?
  — human probabilistic parsing models massively pruned
  (Jurafsky, 1996; Chater et al., 1998; Lewis and Vasishth, 2005)

- synchronizing approximation across learning and inference
  — it’s a parser, not a language model! (Wainwright, 2006)

Interpretation: Connections

- "learning by doing" — (unsupervised) self-training
  (Clark et al., 2003; Ng and Cardie, 2003; McClosky et al., 2006)
  — relevance to understanding language acquisition?
  — human probabilistic parsing models massively pruned
  (Jurafsky, 1996; Chater et al., 1998; Lewis and Vasishth, 2005)

- synchronizing approximation across learning and inference
  — it's a parser, not a language model! (Wainwright, 2006)

- annealing of objective functions (Smith and Eisner, 2004)
  — $w_t \propto \mathbb{P}_\theta(t \mid s)^\beta$, $\beta \in [0, 1]$ (from Uniform to Classic EM)

Interpretation: Connections

- "learning by doing" — (unsupervised) self-training
  (Clark et al., 2003; Ng and Cardie, 2003; McClosky et al., 2006)
  — relevance to understanding language acquisition?
  — human probabilistic parsing models massively pruned
  (Jurafsky, 1996; Chater et al., 1998; Lewis and Vasishth, 2005)

- synchronizing approximation across learning and inference
  — it's a parser, not a language model! (Wainwright, 2006)

- annealing of objective functions (Smith and Eisner, 2004)
  — $w_t \propto \mathbb{P}_\theta(t \mid s)^\beta$, $\beta \in [0, 1]$ (from Uniform to Classic EM)
  — Viterbi EM: $\lim_{\beta \to \infty}$

Three Objective Functions

Three Objective Functions

- **supervised objective (convex):**

$$\hat{\theta}_{\text{SUP}} = \arg\max_{\theta} \prod_{s} \mathbb{P}_{\theta}(t^{*}(s))$$

Three Objective Functions

- **supervised objective (convex):**

$$\hat{\theta}_{\text{SUP}} = \arg\max_{\theta} \prod_{s} \mathbb{P}_{\theta}(t^*(s))$$

- **unsupervised objective (non-convex):**

$$\hat{\theta}_{\text{UNS}} = \arg\max_{\theta} \prod_{s} \underbrace{\sum_{t \in T(s)} \mathbb{P}_{\theta}(t)}_{\mathbb{P}_{\theta}(s)}$$

Three Objective Functions

- **supervised objective (convex):**

$$\hat{\theta}_{\text{SUP}} = \arg\max_{\theta} \prod_{s} \mathbb{P}_{\theta}(t^*(s))$$

- **unsupervised objective (non-convex):**

$$\hat{\theta}_{\text{UNS}} = \arg\max_{\theta} \prod_{s} \underbrace{\sum_{t \in T(s)} \mathbb{P}_{\theta}(t)}_{\mathbb{P}_{\theta}(s)}$$

- **another unsupervised objective (also non-convex):**

$$\hat{\theta}_{\text{VIT}} = \arg\max_{\theta} \prod_{s} \max_{t \in T(s)} \mathbb{P}_{\theta}(t)$$

Potential Disconnects

Potential Disconnects

- **classic unsupervised parsers:**

Potential Disconnects

- **classic unsupervised parsers:**
  — train with respect to sentence strings (learning)

Potential Disconnects

- **classic unsupervised parsers:**
  — train with respect to sentence strings (learning)
  — parse with respect to one-best trees (inference)

Potential Disconnects

- **classic unsupervised parsers:**
  — train with respect to sentence strings (learning)
  — parse with respect to one-best trees (inference)
  — judged against external references (evaluation)

Potential Disconnects

- **classic unsupervised parsers:**
  — train with respect to sentence strings (learning)
  — parse with respect to one-best trees (inference)
  — judged against external references (evaluation)

- **the true generative model θ^*:**

Potential Disconnects

- **classic unsupervised parsers:**
  — train with respect to sentence strings (learning)
  — parse with respect to one-best trees (inference)
  — judged against external references (evaluation)

- **the true generative model θ^*:**
  — may not yield the most discriminating parser

Potential Disconnects

- **classic unsupervised parsers:**
  — train with respect to sentence strings (learning)
  — parse with respect to one-best trees (inference)
  — judged against external references (evaluation)

- **the true generative model θ^*:**
  — may not yield the most discriminating parser
  — may assign suboptimal mass to strings

Potential Disconnects

- classic unsupervised parsers:
  — train with respect to sentence strings (learning)
  — parse with respect to one-best trees (inference)
  — judged against external references (evaluation)

- the true generative model θ^*:
  — may not yield the most discriminating parser
  — may assign suboptimal mass to strings

- Viterbi EM fixes one of these ...

Potential Disconnects

- **classic unsupervised parsers:**
  — train with respect to sentence strings (learning)
  — parse with respect to one-best trees (inference)
  — judged against external references (evaluation)

- **the true generative model θ^*:**
  — may not yield the most discriminating parser
  — may assign suboptimal mass to strings

- **Viterbi EM fixes one of these ...**
  — ... but both flavors of EM walk away from the supervised optimum

Reminder: Accuracy vs. $\theta^* \neq \hat{\theta}_{\text{SUP}}$

Reminder: Accuracy vs. $\theta^* \neq \hat{\theta}_{\text{SUP}}$

- **maximizing likelihood may degrade accuracy**

  (Pereira and Schabes, 1992; Elworthy, 1994; Merialdo, 1994)

Reminder: Accuracy vs. $\theta^* \neq \hat{\theta}_{\text{SUP}}$

- **maximizing likelihood may degrade accuracy**

  (Pereira and Schabes, 1992; Elworthy, 1994; Merialdo, 1994)
- **simple example: optimize the wrong model**

  (e.g., make incorrect independence assumptions)

Reminder: Accuracy vs. $\theta^* \neq \hat{\theta}_{\text{SUP}}$

- **maximizing likelihood may degrade accuracy**
  (Pereira and Schabes, 1992; Elworthy, 1994; Merialdo, 1994)
- **simple example: optimize the wrong model**
  (e.g., make incorrect independence assumptions)
- **fitting the (supervised) DMV to contrived symmetries:**

(i) a a a

(ii) a a a

Reminder: Accuracy vs. $\theta^* \neq \hat{\theta}_{\text{SUP}}$

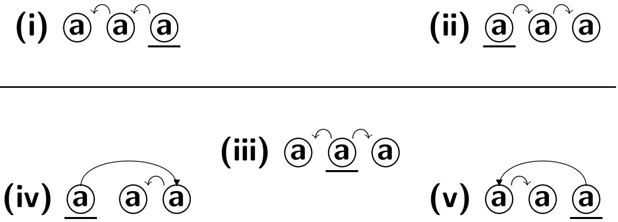
- **maximizing likelihood may degrade accuracy**
  (Pereira and Schabes, 1992; Elworthy, 1994; Merialdo, 1994)
- **simple example: optimize the wrong model**
  (e.g., make incorrect independence assumptions)
- **fitting the (supervised) DMV to contrived symmetries:**

(i) a a <u>a</u>   **(ii)** <u>a</u> a a

(iii) a <u>a</u> a

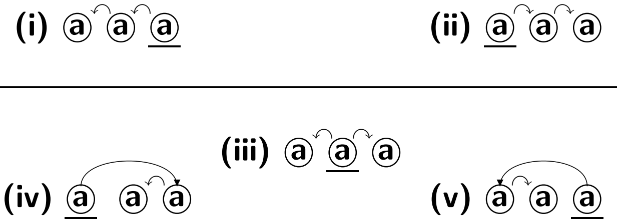
Reminder: Accuracy vs. $\theta^* \neq \hat{\theta}_{\text{SUP}}$

- maximizing likelihood may degrade accuracy
  (Pereira and Schabes, 1992; Elworthy, 1994; Merialdo, 1994)
- simple example: optimize the wrong model
  (e.g., make incorrect independence assumptions)
- fitting the (supervised) DMV to contrived symmetries:



Reminder: Accuracy vs. $\theta^* \neq \hat{\theta}_{\text{SUP}}$

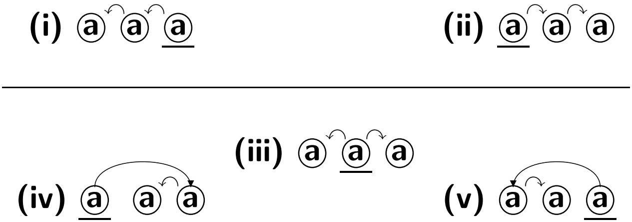
- **maximizing likelihood may degrade accuracy**
  (Pereira and Schabes, 1992; Elworthy, 1994; Merialdo, 1994)
- **simple example: optimize the wrong model**
  (e.g., make incorrect independence assumptions)
- **fitting the (supervised) DMV to contrived symmetries:**



- **expected accuracy for $\hat{\theta}_{\text{SUP}}$: 40%**

Reminder: Accuracy vs. $\theta^* \neq \hat{\theta}_{\text{SUP}}$

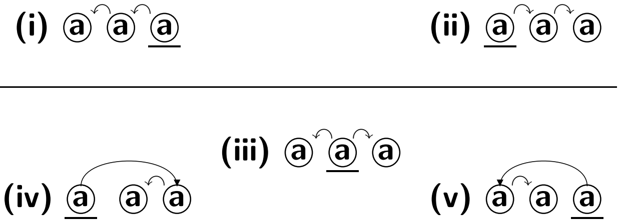
- **maximizing likelihood may degrade accuracy**
  (Pereira and Schabes, 1992; Elworthy, 1994; Merialdo, 1994)
- **simple example: optimize the wrong model**
  (e.g., make incorrect independence assumptions)
- **fitting the (supervised) DMV to contrived symmetries:**



- **expected accuracy for $\hat{\theta}_{\text{SUP}}$: 40% (20% for exact trees)**

Reminder: Accuracy vs. $\theta^* \neq \hat{\theta}_{\text{SUP}}$

- **maximizing likelihood may degrade accuracy**
  (Pereira and Schabes, 1992; Elworthy, 1994; Merialdo, 1994)
- **simple example: optimize the wrong model**
  (e.g., make incorrect independence assumptions)
- **fitting the (supervised) DMV to contrived symmetries:**



- **expected accuracy for $\hat{\theta}_{\text{SUP}}$: 40% (20% for exact trees) — yet could achieve 50% (for both) deterministically**

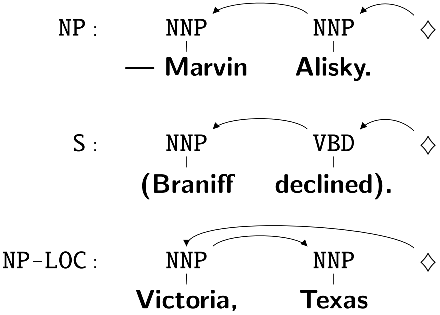
More Subtle: $\theta^* = \hat{\theta}_{\text{SUP}}$ vs. $\hat{\theta}_{\text{UNS}}$ vs. $\hat{\theta}_{\text{VIT}}$

More Subtle: $\theta^* = \hat{\theta}_{\text{SUP}}$ vs. $\hat{\theta}_{\text{UNS}}$ vs. $\hat{\theta}_{\text{VIT}}$

- **this time, an organic example:**

More Subtle: $\theta^* = \hat{\theta}_{\text{SUP}}$ vs. $\hat{\theta}_{\text{UNS}}$ vs. $\hat{\theta}_{\text{VIT}}$

- **this time, an organic example:**



More Subtle: $\theta^* = \hat{\theta}_{\text{SUP}}$ vs. $\hat{\theta}_{\text{UNS}}$ vs. $\hat{\theta}_{\text{VIT}}$

More Subtle: $\theta^* = \hat{\theta}_{\text{SUP}}$ vs. $\hat{\theta}_{\text{UNS}}$ vs. $\hat{\theta}_{\text{VIT}}$

- — **the right model**

More Subtle: $\theta^* = \hat{\theta}_{\text{SUP}}$ vs. $\hat{\theta}_{\text{UNS}}$ vs. $\hat{\theta}_{\text{VIT}}$

- **— the right model, DMV factors the parameters**

More Subtle: $\theta^* = \hat{\theta}_{\text{SUP}}$ vs. $\hat{\theta}_{\text{UNS}}$ vs. $\hat{\theta}_{\text{VIT}}$

- **— the right model, DMV factors the parameters**
  — no unwarranted independence assumptions

More Subtle: $\theta^* = \hat{\theta}_{\text{SUP}}$ vs. $\hat{\theta}_{\text{UNS}}$ vs. $\hat{\theta}_{\text{VIT}}$

- **— the right model, DMV factors the parameters**
  — no unwarranted independence assumptions
  — exact calculations (no numerical instabilities)

More Subtle: $\theta^* = \hat{\theta}_{\text{SUP}}$ vs. $\hat{\theta}_{\text{UNS}}$ vs. $\hat{\theta}_{\text{VIT}}$

- **— the right model, DMV factors the parameters**
  — no unwarranted independence assumptions
  — exact calculations (no numerical instabilities)
  — issue persists with infinite data

More Subtle: $\theta^* = \hat{\theta}_{\text{SUP}}$ vs. $\hat{\theta}_{\text{UNS}}$ vs. $\hat{\theta}_{\text{VIT}}$

- — the right model, DMV factors the parameters
  — no unwarranted independence assumptions
  — exact calculations (no numerical instabilities)
  — issue persists with infinite data
- can again find a more deterministic $\tilde{\theta}$ than θ^*:

More Subtle: $\theta^* = \hat{\theta}_{\text{SUP}}$ vs. $\hat{\theta}_{\text{UNS}}$ vs. $\hat{\theta}_{\text{VIT}}$

- — the right model, DMV factors the parameters
  — no unwarranted independence assumptions
  — exact calculations (no numerical instabilities)
  — issue persists with infinite data
- can again find a more deterministic $\tilde{\theta}$ than θ^*:
  — assigns zero probability to the truth

More Subtle: $\theta^* = \hat{\theta}_{\text{SUP}}$ vs. $\hat{\theta}_{\text{UNS}}$ vs. $\hat{\theta}_{\text{VIT}}$

- — the right model, DMV factors the parameters
  — no unwarranted independence assumptions
  — exact calculations (no numerical instabilities)
  — issue persists with infinite data

- can again find a more deterministic $\tilde{\theta}$ than θ^*:
  — assigns zero probability to the truth
  — attains higher likelihood on both unsupervised metrics

More Subtle: $\theta^* = \hat{\theta}_{\text{SUP}}$ vs. $\hat{\theta}_{\text{UNS}}$ vs. $\hat{\theta}_{\text{VIT}}$

- — the **right** model, DMV **factors** the parameters
  — no unwarranted independence assumptions
  — exact calculations (no numerical instabilities)
  — issue persists with infinite data

- can again find a more deterministic $\tilde{\theta}$ than θ^*:
  — assigns **zero probability** to the **truth**
  — attains **higher likelihood** on both unsupervised metrics
  — has the **same** expected (but lower variance) **accuracy**

More Subtle: $\theta^* = \hat{\theta}_{\text{SUP}}$ vs. $\hat{\theta}_{\text{UNS}}$ vs. $\hat{\theta}_{\text{VIT}}$

- — the right model, DMV factors the parameters
  — no unwarranted independence assumptions
  — exact calculations (no numerical instabilities)
  — issue persists with infinite data
- can again find a more deterministic $\tilde{\theta}$ than θ^*:
  — assigns zero probability to the truth
  — attains higher likelihood on both unsupervised metrics
  — has the same expected (but lower variance) accuracy
  — and is a fixed point for both flavors of EM

More Subtle: $\theta^* = \hat{\theta}_{\text{SUP}}$ vs. $\hat{\theta}_{\text{UNS}}$ vs. $\hat{\theta}_{\text{VIT}}$

- — the **right** model, DMV **factors** the parameters
  — no unwarranted independence assumptions
  — exact calculations (no numerical instabilities)
  — issue persists with infinite data

- can again find a more deterministic $\tilde{\theta}$ than θ^*:
  — assigns **zero probability** to the **truth**
  — attains **higher likelihood** on both unsupervised metrics
  — has the **same** expected (but lower variance) **accuracy**
  — and is a **fixed point** for both flavors of EM
  ... "fun" exercise, left to the readers! :)

More Subtle: $\theta^* = \hat{\theta}_{\text{SUP}}$ vs. $\hat{\theta}_{\text{UNS}}$ vs. $\hat{\theta}_{\text{VIT}}$

- — the **right** model, DMV **factors** the parameters
  — no unwarranted independence assumptions
  — exact calculations (no numerical instabilities)
  — issue persists with infinite data

- can again find a more deterministic $\tilde{\theta}$ than θ^*:
  — assigns **zero probability** to the **truth**
  — attains **higher likelihood** on both unsupervised metrics
  — has the **same** expected (but lower variance) **accuracy**
  — and is a **fixed point** for both flavors of EM
  ... "fun" exercise, left to the readers! :)

- Classic EM known for **local deterministic attractors**

More Subtle: $\theta^* = \hat{\theta}_{\text{SUP}}$ vs. $\hat{\theta}_{\text{UNS}}$ vs. $\hat{\theta}_{\text{VIT}}$

- — the right model, DMV factors the parameters
  — no unwarranted independence assumptions
  — exact calculations (no numerical instabilities)
  — issue persists with infinite data

- can again find a more deterministic $\tilde{\theta}$ than θ^*:
  — assigns zero probability to the truth
  — attains higher likelihood on both unsupervised metrics
  — has the same expected (but lower variance) accuracy
  — and is a fixed point for both flavors of EM
  ... "fun" exercise, left to the readers! :)

- Classic EM known for local deterministic attractors
  — Viterbi EM suggested as a remedy (de Marcken, 1995)

More Subtle: $\theta^* = \hat{\theta}_{\text{SUP}}$ vs. $\hat{\theta}_{\text{UNS}}$ vs. $\hat{\theta}_{\text{VIT}}$

- — the right model, DMV factors the parameters
  — no unwarranted independence assumptions
  — exact calculations (no numerical instabilities)
  — issue persists with infinite data

- can again find a more deterministic $\tilde{\theta}$ than θ^*:
  — assigns zero probability to the truth
  — attains higher likelihood on both unsupervised metrics
  — has the same expected (but lower variance) accuracy
  — and is a fixed point for both flavors of EM
  ... "fun" exercise, left to the readers! :)

- Classic EM known for local deterministic attractors
  — Viterbi EM suggested as a remedy (de Marcken, 1995)
  — but problem with objectives not confined to EM!

Conclusion

Conclusion

- **need stronger models and better objective functions**

Conclusion

- **need stronger models and better objective functions — but this pulls us back towards central planning...**

Conclusion

- **need stronger models and better objective functions — but this pulls us back towards central planning...**

- **grammar induction is inherently underdetermined**

Conclusion

- **need stronger models and better objective functions — but this pulls us back towards central planning...**

- **grammar induction is inherently underdetermined**
- **in general, unsupervised learning is underconstrained**

Conclusion

- need stronger models and better objective functions
  — but this pulls us back towards central planning...

- grammar induction is inherently underdetermined
- in general, unsupervised learning is underconstrained

- alternative: introduce application-specific constraints

Conclusion

- **need stronger models and better objective functions — but this pulls us back towards central planning...**

- **grammar induction is inherently underdetermined**
- **in general, unsupervised learning is underconstrained**

- **alternative: introduce application-specific constraints — encourage equilibria that share our values (regulation!)**

Conclusion

- **need stronger models and better objective functions — but this pulls us back towards central planning...**

- **grammar induction is inherently underdetermined**
- **in general, unsupervised learning is underconstrained**

- **alternative: introduce application-specific constraints — encourage equilibria that share our values (regulation!)**

1. **partial bracketings** **(Pereira and Schabes, 1992)**

Conclusion

- **need stronger models and better objective functions — but this pulls us back towards central planning...**

- **grammar induction is inherently underdetermined**
- **in general, unsupervised learning is underconstrained**

- **alternative: introduce application-specific constraints — encourage equilibria that share our values (regulation!)**

1. **partial bracketings** (Pereira and Schabes, 1992)
2. **synchronous grammars induction** (Alshawi and Douglas, 2000)

Conclusion

- **need stronger models and better objective functions — but this pulls us back towards central planning...**

- **grammar induction is inherently underdetermined**
- **in general, unsupervised learning is underconstrained**

- **alternative: introduce application-specific constraints — encourage equilibria that share our values (regulation!)**

1. **partial bracketings** (Pereira and Schabes, 1992)
2. **synchronous grammars induction** (Alshawi and Douglas, 2000)
3. **linear-time parsing, skewness, Zipf's Law...** (Seginer, 2007)

Conclusion

- need stronger models and better objective functions
  — but this pulls us back towards central planning...

- grammar induction is inherently underdetermined
- in general, unsupervised learning is underconstrained

- alternative: introduce application-specific constraints
  — encourage equilibria that share our values (regulation!)

1. partial bracketings (Pereira and Schabes, 1992)
2. synchronous grammars induction (Alshawi and Douglas, 2000)
3. linear-time parsing, skewness, Zipf's Law... (Seginer, 2007)
4. sparse posterior regularization (Ganchev et al., 2009)

Conclusion

- need stronger models and better objective functions
  — but this pulls us back towards central planning...

- grammar induction is inherently underdetermined
- in general, unsupervised learning is underconstrained

- alternative: introduce application-specific constraints
  — encourage equilibria that share our values (regulation!)

1. partial bracketings (Pereira and Schabes, 1992)
2. synchronous grammars induction (Alshawi and Douglas, 2000)
3. linear-time parsing, skewness, Zipf's Law... (Seginer, 2007)
4. sparse posterior regularization (Ganchev et al., 2009)
5. mining structure from web mark-up (Spitkovsky et al., 2010)

Summary

Summary

- **Viterbi EM well-suited to unsupervised parsing**

Summary

- **Viterbi EM well-suited to unsupervised parsing**
- **faster to run**

Summary

- **Viterbi EM well-suited to unsupervised parsing**
- **faster to run**
  — no outside charts (each iteration is faster)

Summary

- **Viterbi EM well-suited to unsupervised parsing**
- **faster to run**
  — no outside charts (each iteration is faster)
  — quicker to converge (4-10x fewer iterations)

Summary

- **Viterbi EM well-suited to unsupervised parsing**
- **faster to run**
  — no outside charts (each iteration is faster)
  — quicker to converge (4-10x fewer iterations)
- **scales better**

Summary

- **Viterbi EM well-suited to unsupervised parsing**

- **faster to run**
  — no outside charts (each iteration is faster)
  — quicker to converge (4-10x fewer iterations)

- **scales better**
  — efficiently handles larger data sets

Summary

- **Viterbi EM well-suited to unsupervised parsing**

- **faster to run**
  — no outside charts (each iteration is faster)
  — quicker to converge (4-10x fewer iterations)

- **scales better**
  — efficiently handles larger data sets
  — performs gracefully with more complex data

Summary

- **Viterbi EM well-suited to unsupervised parsing**

- **faster to run**
  — no outside charts (each iteration is faster)
  — quicker to converge (4-10x fewer iterations)

- **scales better**
  — efficiently handles larger data sets
  — performs gracefully with more complex data

- **simpler algorithm**

Summary

- **Viterbi EM well-suited to unsupervised parsing**

- **faster to run**
  — no outside charts (each iteration is faster)
  — quicker to converge (4-10x fewer iterations)

- **scales better**
  — efficiently handles larger data sets
  — performs gracefully with more complex data

- **simpler algorithm**
  — easier to code up, debug, and understand...

Summary

- **Viterbi EM well-suited to unsupervised parsing**

- **faster to run**
  — no outside charts (each iteration is faster)
  — quicker to converge (4-10x fewer iterations)

- **scales better**
  — efficiently handles larger data sets
  — performs gracefully with more complex data

- **simpler algorithm**
  — easier to code up, debug, and understand...
  — invites more flexible modeling techniques!

Summary

- **Viterbi EM well-suited to unsupervised parsing**
- **faster to run**
  — no outside charts (each iteration is faster)
  — quicker to converge (4-10x fewer iterations)
- **scales better**
  — efficiently handles larger data sets
  — performs gracefully with more complex data
- **simpler algorithm**
  — easier to code up, debug, and understand...
  — invites more flexible modeling techniques!
- **achieves state-of-the-art results!**

Thanks!

Questions?