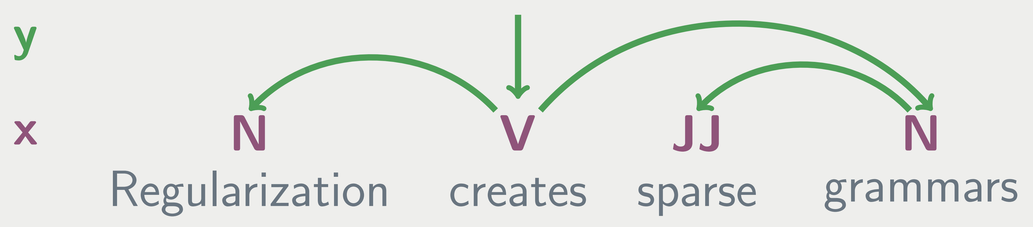
Sparsity in Dependency Grammar Induction

Jennifer Gillenwater[1], Kuzman Ganchev[1], João Graça[2], Fernando Pereira[3], Ben Taskar[1]

[1]University of Pennsylvania
[2]L[2]F INESC-ID
[3]Google

Motivation

Dependency Model with Valence (Klein and Manning, ACL 2004)

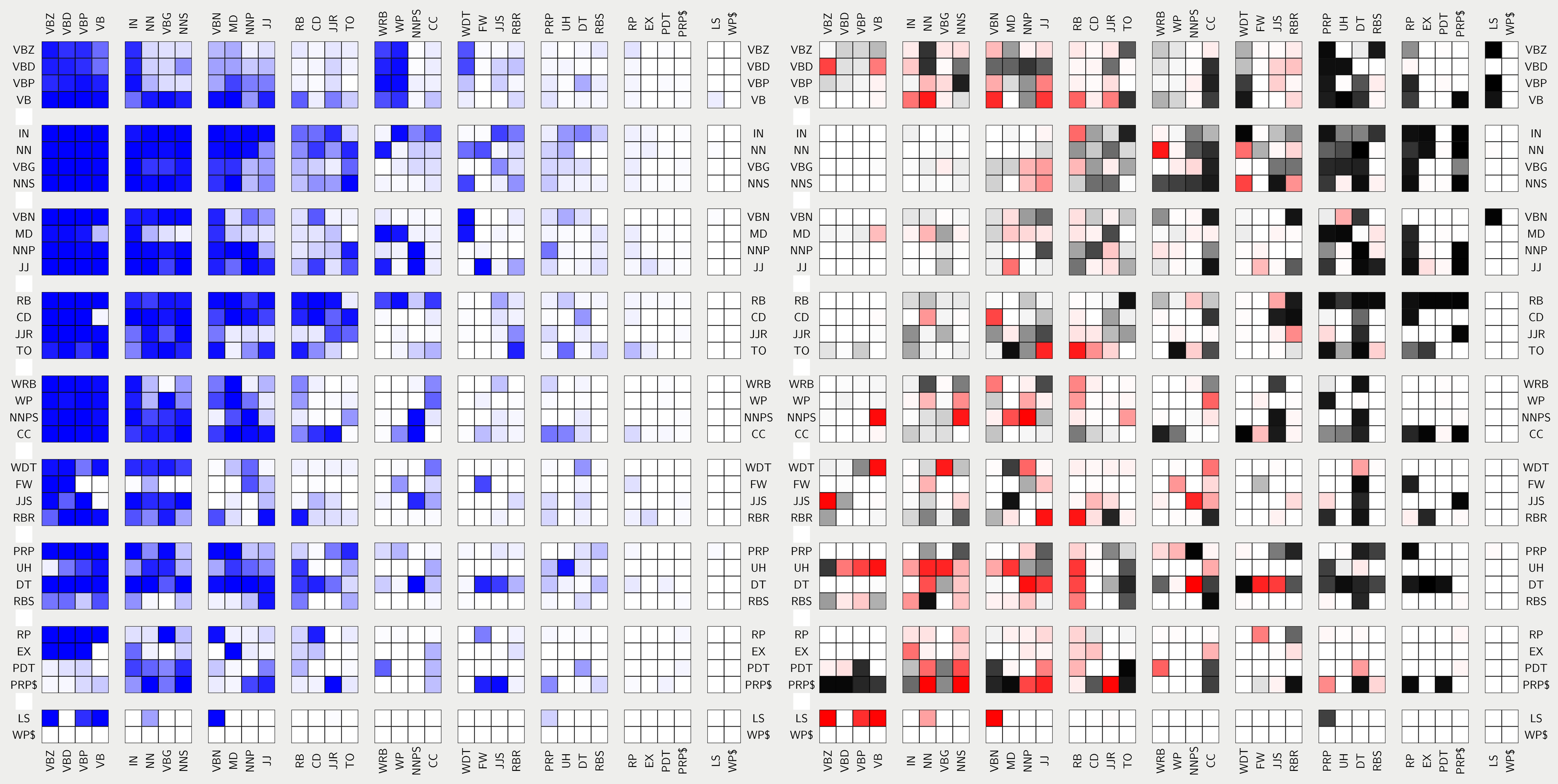


$$p_\theta(x, y) = \theta_{root(V)} \cdot \theta_{stop(nostop|V,right,false)} \cdot \theta_{child(N|V,right)} \cdot \theta_{stop(stop|V,right,true)} \cdot \theta_{stop(nostop|V,left,false)} \cdots$$

- **Task**: Unsupervised dependency grammar induction
- **Problem**: Model is simple, but still too permissive; most relations (e.g. DET $\longrightarrow$ V, N, JJ, etc.) should not occur
- **Solution**: Posterior constraints to limit grammar ambiguity during learning

Traditional Objective Optimization

- **Traditional objective**: marginal log likelihood $\max_\theta \mathcal{L}(\theta) = \hat{E}_X[\log \sum_y p_\theta(x, y)]$
- **Optimization method**: Expectation maximization (EM)
- **Figures**: Parent tags across, child tags down
- **Left:** Blank squares have max posterior **0**; many parent-child relations don't occur
- **Right**: Red have max < supervised, black have max > supervised; many dark squares implies model assigns non-zero probability to too many pairs

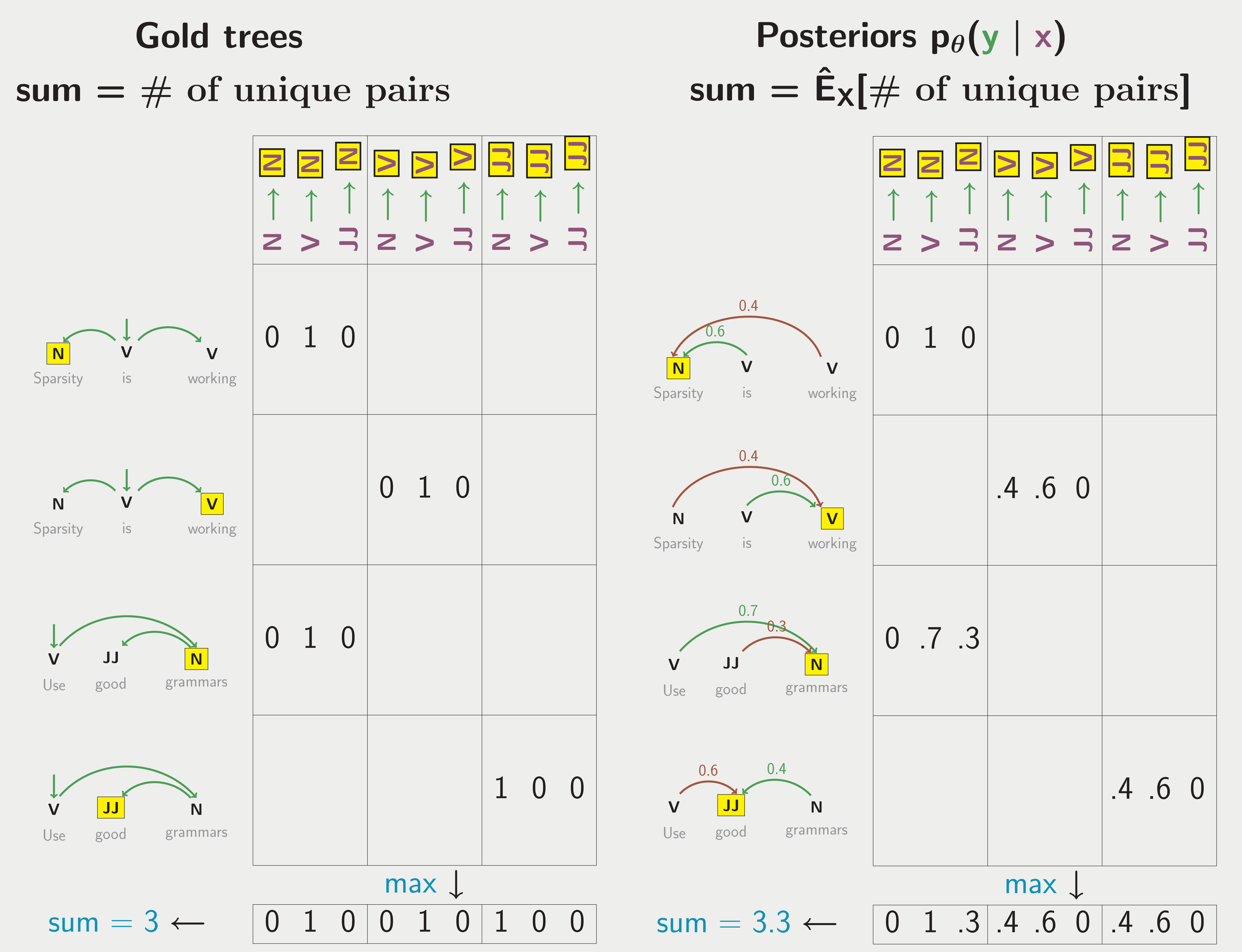


Max likelihood **EM**

Parameter Regularization: $\mathcal{L}(\theta) + \log p(\theta)$

- Hierarchical Dirichlet processes (Liang et al., EMNLP 2007; Johnson et al., NIPS 2007)
- Discounting Dirichlet prior (Headden et al., ACL 2009)
- Logistic normal prior (Cohen et al., NIPS 2008; Cohen and Smith, NAACL 2009)
- All of these tend to reduce unique # of children per parent, rather than directly reducing # of unique parent-child pairs: $\theta_{child(Y|X)} \neq$ **posterior**$(X \rightarrow Y)$

Ambiguity Measure Using Posteriors: $L_{1/\infty}$

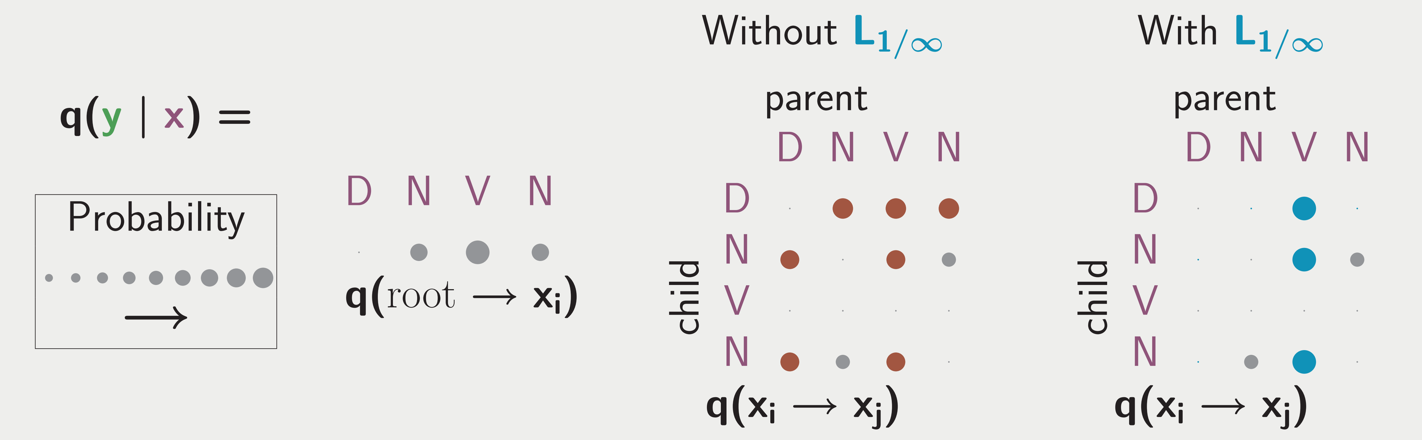


Posterior Regularization

Minimize # of unique pairs through E-step penalty, $L_{1/\infty}$ on the posteriors $q(y \mid x)$ (Graca et al., NIPS 2007 & 2009)

$$\textbf{M-Step} \quad \theta^{t+1} = \arg\max_\theta \hat{E}_X \left[\sum_y q^t(y \mid x) \log p_\theta(x, y) \right]$$

$$\textbf{E-Step} \quad q^t(y \mid x) = \arg\min_{q(y|x)} KL(q(y \mid x) \parallel p_{\theta^t}(y \mid x)) + \sigma L_{1/\infty}(q(y \mid x))$$



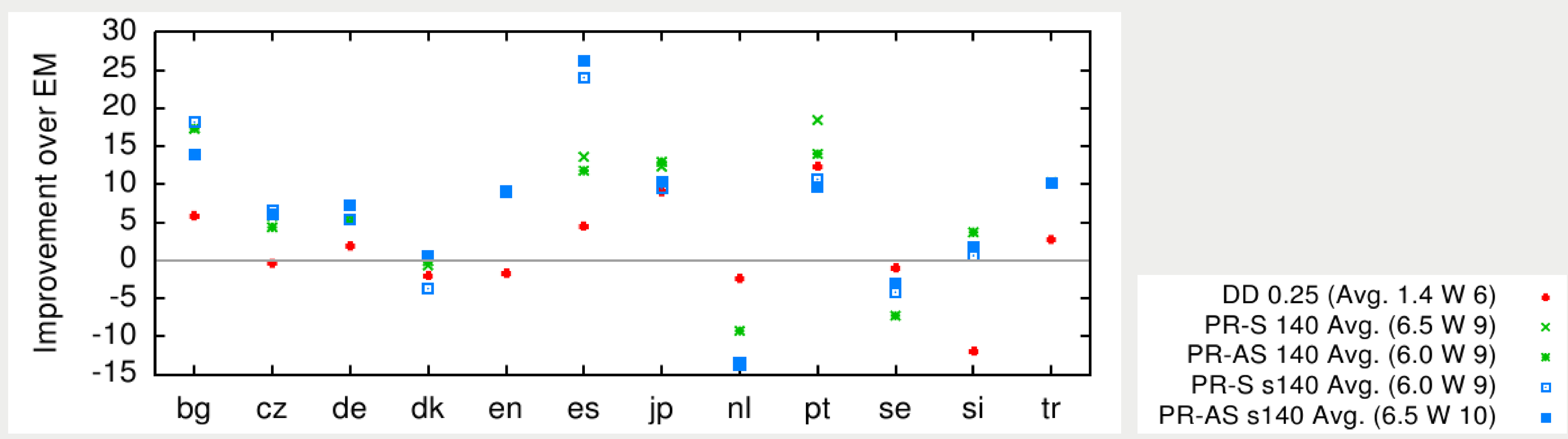
Experiments on English

- Penn Treebank data, strip punctuation, consider separately sentences of length ≤ 10 and 20, initialize model "harmonically", try $\sigma \in \{80, 100, 120, 140, 160, 180\}$
- **Table**: top — experiments on the basic dependency model with valence: our method and two parameter priors (Cohen et al., NIPS 2008; Cohen and Smith, NAACL 2009); bottom — extended version of the model with parameters that use more valence information: our method, and a non-sparsifying ($\alpha = 1$) discounting Dirichlet prior (DD) with random pools initialization and learned backoff weight λ (Headden et al., NAACL 2009).
- PR with random pools would likely produce best result of all

Learning Method	Accuracy		
	≤ 10	≤ 20	all
PR ($\sigma = 140$)	**62.1**	**53.8**	**49.1**
LN families	59.3	45.1	39.0
SLN TieV & N	61.3	47.4	41.4
PR ($\sigma = 140$, $\lambda = 1/3$)	64.4	55.2	50.5
DD ($\alpha = 1$, λ learned)	**65.0 (±5.7)**		

Experiments on 11 Other Languages

Figure: Relative error with respect to EM on the extended model; DD = discounting Dirichlet prior, PR = posterior regularization (σ = 160 chosen on English), PR-S = symmetric version of constraints, PR-AS = asymmetric version; Avg = average improvement over EM, W = # of languages better than EM



Parse Analysis

Parse	Unique parent-child pairs
d Una, nc papelera, v es	(v, nc); (nc, d)
d Una, nc papelera, v es	(v, d); (d, nc)
v Lleva, nc tiempo, v entenderlos	(v, nc); (v, v)

- Parses 1 and 3: 3 unique pairs total
- Parses 2 and 3: 4 unique pairs total

Conclusion

- For the basic model, average improvements over EM are 1.6% for DD, 6.7% for PR
- For the extended model, average improvements over EM are 1.4% for DD, 6.4% for PR
- Using posterior regularization significantly improves parsing accuracy