The Effect of Non-tightness on Bayesian Estimation of PCFGs

Shay Cohen (Columbia University, University of Edinburgh)
and
Mark Johnson (Macquarie University)

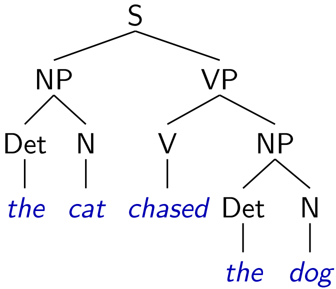
August, 2013

We thank the anonymous reviewers and Giorgio Satta for their valuable comments. Shay Cohen was supported by the National Science Foundation under Grant #1136996 to the Computing Research Association for the CIFellows Project, and Mark Johnson was supported by the Australian Research Council's Discovery Projects funding scheme (project numbers DP110102506 and DP110102593)

Probabilistic context-free grammars (PCFGs)

Probability	*Rule*
1.0	S → NP VP
1.0	NP → Det N
1.0	VP → V NP
0.7	Det → the
0.3	Det → a
0.4	N → cat
0.6	N → dog
0.2	V → chased
0.8	V → liked

Parse tree



Tree probability $= 1.0 \times 1.0 \times 0.7 \times 0.4 \times 1.0 \times 0.2 \times 1.0 \times 0.7 \times 0.6 = 0.02352$

PCFGs and tightness

- $\boldsymbol{p} \in [0,1]^{|\mathcal{R}|}$ is a vector of *rule probabilities* indexed by rules $\mathcal{R}$
- A PCFG associates each tree t with a *measure* $m_{\boldsymbol{p}}(t)$:

$$m_{\boldsymbol{p}}(t) \;=\; \prod_{A \to \alpha \in \mathcal{R}} p_{A \to \alpha}^{n_{A \to \alpha}(t)}, \text{ where:}$$

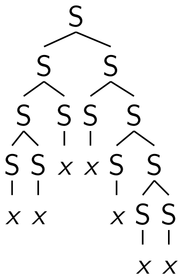
  $n_{A \to \alpha}(t)$ is the number of times rule $A \to \alpha$ is used in the derivation of t
- The *partition function* Z of a PCFG is:

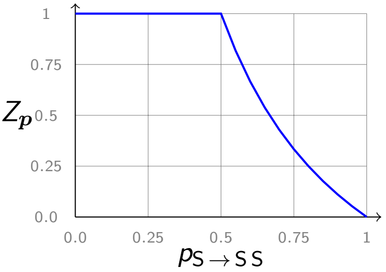
$$Z_{\boldsymbol{p}} \;=\; \sum_{t \in \mathcal{T}} m_{\boldsymbol{p}}(t)$$

- PCFGs require the rule probabilities expanding a non-terminal to be normalised, *but this does not guarantee that $Z_{\boldsymbol{p}} = 1$*
- When $Z_{\boldsymbol{p}} < 1$, we say the PCFG is "*non-tight*."

Catalan grammar: an example of a non-tight PCFG

- PCFG has two rules: $\mathsf{S} \rightarrow \mathsf{S}\,\mathsf{S}$ and $\mathsf{S} \rightarrow x$
- It generates strings of x of arbitrary length
- It generates all possible finite binary trees
  - or equivalently, all possible well-formed bracketings
  - called the *Catalan grammar* because the number of parses of x^n is Catalan number C_{n-1}
- The PCFG is non-tight when $p_{\mathsf{S} \rightarrow \mathsf{S}\,\mathsf{S}} > 0.5$





Why can the Catalan grammar be non-tight?

- Every binary tree over n terminals has $n - 1$ non-terminals
  - $\Rightarrow$ probability of a tree *decreases exponentially with length*
- The number of different binary trees with n terminals is C_{n-1}
  - $\Rightarrow$ number of trees grammar *grows exponentially with length*
- When $p_{\mathsf{S} \to \mathsf{S}\,\mathsf{S}} \geq 0.5$, the PCFG puts non-zero mass on non-terminating derivations
  - this grammar defines a *branching processes*
  - At each step, $p_{\mathsf{S} \to \mathsf{S}\,\mathsf{S}}$ is probability of reproducing, $p_{\mathsf{S} \to x}$ is probability of dying
  - $p_{\mathsf{S} \to \mathsf{S}\,\mathsf{S}} < 0.5 \Rightarrow$ population dies out (subcritical)
  - $p_{\mathsf{S} \to \mathsf{S}\,\mathsf{S}} > 0.5 \Rightarrow$ population grows unboundedly (supercritical)
- Mini-theorem: *every linear PCFG is tight* (except on cases of measure zero under continuous priors)
  - CFG is *linear* $\Leftrightarrow$ RHS of every rule contains at most one non-terminal
  - HMMs are linear PCFGs $\Rightarrow$ always tight

Bayesian inference of PCFGs

- Bayesian inference uses Bayes rule to compute a posterior over rule probability vectors $\boldsymbol{p}$

$$\underbrace{\mathrm{P}(\boldsymbol{p} \mid \boldsymbol{D})}_{\text{Posterior}} \propto \underbrace{\mathrm{P}(\boldsymbol{D} \mid \boldsymbol{p})}_{\text{Likelihood}} \underbrace{\mathrm{P}(\boldsymbol{p})}_{\text{Prior}}$$

where $\boldsymbol{D} = (D_1, \ldots, D_n)$ is the training data (trees or strings)

- Bayesians prefer the full posterior distribution $\mathrm{P}(\boldsymbol{p} \mid \boldsymbol{D})$ to a point estimate $\widehat{\boldsymbol{p}}$
- *If the prior assigns non-zero mass to non-tight grammars, in general the posterior will too*
- As the number of independent observations n in the training data grows, the posterior concentrates around the MLE
  - MLE is always a tight PCFG (Chi and Geman 1998)
  - *As $n \to \infty$ the posterior concentrates on tight PCFGs*

3 approaches to non-tightness in the Bayesian setting

- If the grammar is linear, then all continuous priors lead to tight PCFGs
- Three different approaches to Bayesian inference with non-tight grammars:
  1. **"Sink element"**: assign mass of "infinite trees" to a *sink element*, implicitly assumed by Johnson et al (2007)
  2. **"Only tight"**: redefine prior so it only places mass onto tight grammars
  3. **"Renormalisation"**: divide by partition function to ensure normalisation

Assume for now that trees and strings are observed in $\boldsymbol{D}$ (supervised learning)

“Only tight” approach

Let $\mathrm{I}(p)$ be 1 if p is tight and 0 otherwise.

Given a “non-tight prior” $P(p)$, define a new prior P' as:

$$P'(p) \;\propto\; P(p)\,\mathrm{I}(p)$$

If $P(p)$ is conjugate family of priors with respect to PCFG likelihood, then $P'(p)$ is also conjugate

We can draw samples from $P'(p \mid \boldsymbol{D})$ using *rejection sampling*:

- Draw PCFG parameters p from $P(p \mid \boldsymbol{D})$ until p is tight
  - $P(p \mid \boldsymbol{D})$ is a product of Dirichlets

  $\Rightarrow$ can use textbook algorithms for sampling from Dirichlets

Renormalisation approach

Renormalise the measure $\mu_p(t)$ over finite trees (Chi, 1999)

If $P(p \mid \alpha)$ is a product of Dirichlets, posterior is:

$$\mathrm{P}(p \mid D) \;=\; \prod_{i=1}^{n} \frac{\mu_p(t_i)}{Z_p}\, \mathrm{P}(p \mid \alpha) \;\propto\; \frac{1}{Z_p^n}\, \mathrm{P}(p \mid \alpha + n(D)).$$

where $n(D)$ is the count vector over all rules for the data $\boldsymbol{D}$

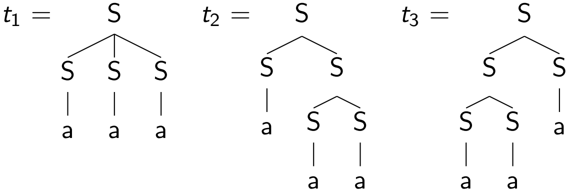
- Use a *Metropolis-Hastings sampler* to sample from $\mathrm{P}(p \mid \boldsymbol{D})$
  - proposal distribution is product of Dirichlets

Samplers for each approach can be used within a component-wise Gibbs sampler for the unsupervised case where only strings are observed.

Toy example

Consider the grammar $S \to S\,S\,S | S\,S | a$

Let $w = a\,a\,a$

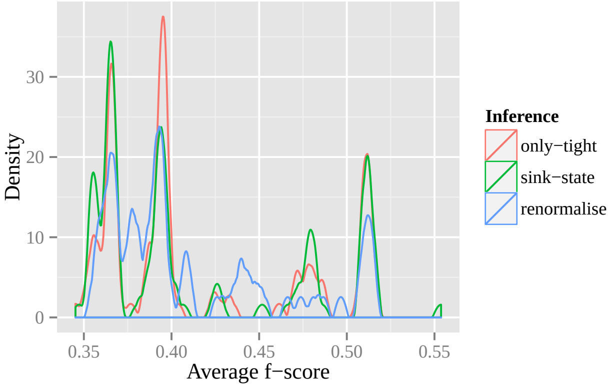


- Uniform prior ($\alpha = 1$)
- Sink-element approach: $\mathrm{P}(t_1 \mid w) = \frac{7}{11} \approx 0.636364$.
- Only-tight approach: $\mathrm{P}(t_1 \mid w) = \frac{11179}{17221} \approx 0.649149$.
- Renormalisation approach: $\mathrm{P}(t_1 \mid w) \approx 0.619893$.

$\Rightarrow$ All three approaches induce different posteriors from uniform prior

Experiments on WSJ10

- Task: unsupervised estimation of Smith et al (2006)'s PCFG version of the DMV (Klein et al 2004) from WSJ10
- 100 runs of each sampler for 1,000 MCMC sweeps
- Computed average F_1 score on every 10th sweep for last 100 sweeps
- Kolmogorov-Smirnov tests did not show a statistically significant difference



Conclusion

- *Linear* CFGs are tight regardless of the prior
- For non-linear CFGs, three approaches are suggested for handling non-tightness
- The three approaches are not mathematically equivalent, but experiments on WSJ Penn treebank showed that they behave similarly empirically

Open problem: are the approaches reducible in the following sense?

> *Given a prior P for one of the approaches, is there a prior P' for another approach such that for all data $\boldsymbol{D}$, the posteriors under both approaches are the same.*