1 List of Important Terminologies

1. Marginal likelihood $P(s|\alpha)$, under Bayesian framework, parameter θ being marginalized out.

2. Posterior distribution $P(\theta, t|s, \alpha)$

2 Starting from MLE over Hidden Parse Trees

Objective: We want to learn an optimal θ, a probability distribution over rewrite rules in PCFG, given observed sentences $\boldsymbol{s}$ i.e.

$$\arg\max_{\theta} P(\boldsymbol{s}, t|\theta)$$

where parse trees t are hidden. How do we do that? Let's start from not considering Bayesian framework. One approach is to use maximum likelihood estimation over hidden parse trees. We will try to learn θ that maximizes the marginal likelihood $P(s|\theta) = \sum_t P(s, t|\theta)$ with EM algorithm (known as "inside-outside algorithm").

1. Predict the parse trees given θ [4].

2. Maximize θ [4].

3. Repeat 1. and 2. till convergence.

3 Applying Bayesian Framework over Hidden Parse Trees

Then, let's consider applying Bayesian framework and try to aim for maximizing the posterior distribution $P(\theta|\boldsymbol{s})$:

$$P(\theta|\boldsymbol{s}) \propto P(\boldsymbol{s}|\theta)P(\theta)$$

Use Dirichlet prior on each θ_A (where A is a non-terminal):

$$\theta_A \sim Dirichlet(\alpha_A)$$

As a result of including Dirichlet prior, θ now depends on a hyper-parameter α of a Dirichlet distribution. Thus, the posterior distribution $P(\theta|\boldsymbol{s})$ can be rewritten as [3]:

$$P(\theta|\boldsymbol{s}, \alpha) \propto P(\boldsymbol{s}|\theta)P(\theta|\alpha)$$

Assume that s is distributed as $Multinomial$ i.e. $\boldsymbol{s} \sim Multinomial(\theta)$. Then, the posterior $P(\theta|\boldsymbol{s})$ or $P(\theta|\boldsymbol{s}, \alpha)$ is also distributed as $Dirichlet$:

$$\theta|\boldsymbol{s} \sim Dirichlet(f_{\boldsymbol{t}} + \alpha)$$

where $f_{\boldsymbol{t}}$ is the vector of production counts in $\boldsymbol{t}$ indexed by $r \in R$ [6]. This is intuitive result too since we can directly reflect the number of times we observed a rule in the training data.

4 Training Bayesian PCFG under supervised and unsupervised setting

Under supervised setting, since the data consists of parse trees $\boldsymbol{t}$ [3], simply replace $\boldsymbol{s}$ by $\boldsymbol{t}$:

$$P(\boldsymbol{s}|\theta) = P(\boldsymbol{t}|\theta) = \prod_i P(t_i|\theta)$$

Under unsupervised setting, we regard the parse tree t as a latent variable, $theta$ as a parameter. Since the data consists of sentences $\boldsymbol{s}$,

$$P(\boldsymbol{s}|\theta) = \prod_i P(s_i|\theta) = \prod_i \sum_{t_i \in T : yield(t_i) = s_i} P(t_i|\theta)$$

Moreover, in unsupervised setting, we are also interested in knowing the parse trees t for given sentences s. So the posterior $P(\theta|\boldsymbol{s}, \alpha)$ becomes:

$$P(\theta|\boldsymbol{s}, \alpha) = \sum_{\boldsymbol{t}} P(\boldsymbol{t}, \theta|\boldsymbol{s}, \alpha)$$

In conclusion, under both unsupervised setting with Bayesian framework, our objective is to compute the posterior $P(\boldsymbol{t}, \theta|\boldsymbol{s}, \alpha)$ using Dirichlet prior with hyper-parameter α.

5 Marginal Likelihood vs. Joint likelihood

Once again, $\theta \sim Dirichlet(\alpha)$. In general, marginal likelihood is a likelihood function in which some parameter variables are marginalized out. Note that marginal likelihood

when a variable θ is marginalized out.

$$\arg\max_\theta L(\theta|X) = P(X|\theta) = \arg\max_\theta \prod_{i=1}^{n} \sum_{y \in \Omega_Y} P(y|\theta) P(x_i|y, \theta).$$

5.1 Why Marginal Likelihood?

Since outputs (= labels) are not given in latent variable models [9], we cannot compute the joint likelihood of inputs and outputs. Therefore, we would like to compute the marginal likelihood (by summing over all outputs for a given input) [1]. If we are using EM to learn the rule probabilities, then it attempts to maximize the marginal likehood of inputs [1].

Note that marginal likelihood can appear in either Bayesian framework or non-Bayesian framework.

6 Summary of Parameter Learning [5]

Here, we assume that the prior is dirichlet distribution, likelihood is multinomial distribution, and therefore, the posterior distribution is also Dirichlet distribution.

	labeled data (=Parse trees known)	unlabeled data (= Parse trees unknown)
ML	frequency	EM
Bayesian	updated Dirichlet distribution	MCMC/VB

7 Do you understand the difference between MAP and Bayesian approach?

“Both ML and MAP return only single and specific values for the parameter θ. Bayesian estimation, by contrast, calculates fully the posterior distribution $P(\theta|X)$.” [7].

8 Key Features

1. PCFG
2. Bayesian Framework
3. Variational Bayes, Gibbs sampling, particle filter
4. Unsupervised (parse trees not observed)

9 Sparse Grammar = Sparse Dirichlet Prior

1. “Our results illustrate that Bayesian inference using a prior that favors sparsity can produce linguistically reasonable analyses in situations in which EM does not.” [6].
2. “This ability to bias the sampler toward sparse grammars (i.e., grammars in which many productions have probabilities close to 0) is useful **when we attempt to identify relevant productions from a much larger set of possible productions via parameter estimation**” [6].
3. “Thus in this application the ability to prefer sparse grammars enables us to find linguistically meaningful analyses. This ability to find linguistically meaningful structure is relatively rare in our experience with unsupervised PCFG induction” [6].
4. “We found that for $\alpha > 10^{-2}$ the samples produced by the Hastings algorithm were the same trivial analyses as those produced by the IO algorithm, but as α was reduced below this t began to exhibit nontrivial structure” [6].

10 Misc. Notes

1. In a supervised setting, a tree t and a sentence s are both observable.
2. Geometric distibution, Binomial disctribution all have Dirichelt distribution as its conjuguate priors (or multi-dimensional beta distributions)
3. Using sparse Dirichlet priors ($\alpha < 1$) is the key. (Encouraging majority of values will be concentrated in a few of the values)
4. Why VB instead of MCMC? 1) Scaling or training on large amount of data is easier. It can also be applied to two level hierarchical Bayesian models, but not to the multiple levels of hierarchical Bayesian models. In that case, MCMC is good at it.

11 Future Studies

1. Read [2] and [8].
2. Variational Inference, Gibbs sampling, and particle filter in general.
3. Start from assuming that the parse trees $\boldsymbol{t}$ are observed. Then, study about the Bayesian framework applied to PCFG.
4. Review the NLP lectures by Chris Manning on Coursera (Assignment PDFs are available at http://www.mohamedaly.info/teaching/cmp-462-spring-2013)
5. Why marginal likelihood is important when we want to estimate the posterior using Variational Bayes?

References

[1] Miguel Ballesteros. Probabilistic Context-Free-Grammars. `http://demo.clab.cs.cmu.edu/fa2015-11711/images/c/ce/Pcfgs.pdf/`. [Online; accessed 31-Jan-2016].

[2] Matthew James Beal. *Variational algorithms for approximate Bayesian inference*. PhD thesis, University of London, 2003.

[3] Shay B. Cohen and Mark Johnson. The effect of non-tightness on bayesian estimation of pcfgs. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1033–1041, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[4] Milo Ercegovevi. Learning Accurate, Compact, and Interpretable Tree Annotation. `https://www.coli.uni-saarland.de/~yzhang/rapt-ws1112/slides/ercegovcevic.ppt/`. [Online; accessed 4-Feb-2016].

[5] Zoubin Ghahramani. Machine Learning Summer School Lecture 3: Learning parameters and structure. `http://mlg.eng.cam.ac.uk/mlss09/mlss_slides/Ghahramani_3.pdf/`. [Online; accessed 29-Jan-2016].

[6] Mark Johnson, Thomas Griffiths, and Sharon Goldwater. Bayesian inference for PCFGs via Markov chain Monte Carlo. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 139–146, Rochester, New York, April 2007. Association for Computational Linguistics.

[7] Avinash Kak. ML, MAP, and Bayesian — The Holy Trinity of Parameter Estimation and Data Prediction. `https://engineering.purdue.edu/kak/Tutorials/Trinity.pdf/`. [Online; accessed 4-Feb-2016].

[8] Kenichi Kurihara and Taisuke Sato. An application of the variational bayesian approach to probabilistic context-free grammars. In *IJCNLP-04 Workshop beyond shallow analyses*, 2004.

[9] Dan Klein Percy Liang, Michael I. Jordan. Probabilistic Grammars and Hierarchical Dirichlet Processes. `http://www-cs.stanford.edu/~pliang/papers/hdppcfg-haba.pdf/`. [Online; accessed 31-Jan-2016].