Supplementary Material for Online Adaptor Grammars with Hybrid Inference

Ke Zhai
Computer Science and UMIACS
University of Maryland
College Park, MD USA
zhaike@cs.umd.edu

Jordan Boyd-Graber
Computer Science
University of Colorado
Boulder, CO USA
jordan.boyd.graber@colorado.edu

Shay B. Cohen
School of Informatics
University of Edinburgh
Edinburgh, Scotland, UK
scohen@inf.ed.ac.uk

In this document, we outline the definition of an adaptor grammar and the generative process of our model (Section 1). This allows us to define a joint distribution over adapted grammars and observations (Section 2).

Uncovering the latent variables of the model (grammars and productions) requires posterior inference. We use online hybrid variational inference, which requires three components: positing a variational distribution (Section 3), deriving the mean-field updates (Section 4), and then adapting those updates into the online setting (Section 5).

Section 6 serves as a reference to review all of the notation used in this document.

1 Definition of Adaptor Grammar

An adaptor grammar is defined by a tuple $\boldsymbol{A} = \langle \boldsymbol{G}, M, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\alpha} \rangle$, containing

1. a grammar—$G = \langle W, N, \boldsymbol{R}, S \rangle$—defined by a set of terminals W, a set of nonterminals N, productions R, and start symbol $S \in N$;
2. a set of adapted nonterminal $M \subseteq N$; and
3. Pitman-Yor parameters a_c, b_c and symmetric Dirichlet parameter α_c (specific to each nonterminal c).

1.1 Generative Process: Adapting nonterminals to Create a Distribution over Trees

This section reviews the distribution over trees produced by adaptor grammars. This distribution is defined for each nonterminal c in two parts; first, the unadapted distribution G_c and the adapted distribution H_c. If a nonterminal is unadapted these are the same. We will first describe the unadapted distribution; however, because this is a recursive process, it will depend on the adapted distributions over other non-terminals. We will define these later.

Each nonterminal c induces a distribution over trees parametrized by the PCFG,[1]

$$G_c \equiv \sum_{c \to \beta \in R_c} \theta_{a \to \beta} \textsc{TreeDist}_c \left(H_{b_1}, H_{b_2}, \ldots \right); \tag{1}$$

where $\textsc{TreeDist}_c \left(\right)$ is a distribution over the subtrees rooted at nonterminal c,

$$\textsc{TreeDist}_c \left(H_{b_1}, H_{b_2}, \ldots, H_{b_n} \right) \left(\begin{array}{c} \text{a} \\ t_1 \quad \ldots \quad t_n \end{array} \right) = \prod_{i=1}^{n} H_{b_n}(t_n), \tag{2}$$

which states that for unadapted rules, each element of the right hand side of the rule is expanded independently.

Adapted rules relax this independence assumption via Bayesian nonparametric. For the nonterminals that are **adapted**, in addition to the PCFG, for each symbol a in the set of adapted nonterminals $\boldsymbol{M}$, we have a

[1] In the case of a terminal $b \in W$, G_b will be a distribution that puts all of its mass on the unit tree labeled as the terminal b.

set of weights

$$\pi'_{c,k} \sim \text{Beta}(1 - b_c, a_c + kb_c) \tag{3}$$

$$\pi_{c,k} \equiv \pi'_{c,k} \prod_{l=1}^{k-1} (1 - \pi'_{c,l}) \tag{4}$$

which assign probability mass to a countably infinite set of atoms $k = 1, \ldots \infty$,

$$z_{c,k} \sim G_c; \tag{5}$$

which are subtrees rooted at a. Together, these define a new distribution over yields for a nonterminal,

$$H_c(\cdot) \equiv \sum_i^{\infty} \pi_{c,i} \delta_{z_{c,i}}(\cdot), \tag{6}$$

where δ is a Dirac delta that is one if the argument matches the atom, zero otherwise. This nonparametric distribution defines the distribution over yields for that nonterminal.

1.2 A Distribution Over Observed Sentences

We assume that the above model produces sentences. We have D observations, where observations $d = 1, \ldots D$ come from:

$$t_d \sim H_S, \tag{7}$$

where S is the designated start state. The trees t_d are latent, but we do observe the terminals associated with each sentence, x_d.

However, writing a tree as t_d obscures the internal structure of this recursive object, so we will need to be able to recover the productions that created the tree. Thus, we will use $M(t_d)$ to encode the multiset of all of the adapted productions that generated the tree and $N(t_d)$ to denote the multiset of non-adapted productions that generated the tree.

2 Joint Likelihood

Given a set of sentences $\boldsymbol{X}$ and a grammar $\boldsymbol{G}$, the joint probability of a collection of parses $\boldsymbol{T}$, PCFG probabilities θ, and adapted grammar $\boldsymbol{H}$ is

$$p(\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\pi}, \boldsymbol{\theta} \mid \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\alpha}) = \prod_{c \in \boldsymbol{N}} \underbrace{p(\boldsymbol{\theta}_c \mid \boldsymbol{\alpha})}_{\text{PCFG Dirichlet prior}} \cdot \prod_{c \in \boldsymbol{M}} \prod_{i=1}^{\infty} \left(\underbrace{p(\pi_{c,i} \mid a_c, b_c) \cdot p(z_{c,i} \mid G_c)}_{\text{adapted stick breaking process}} \right)$$
$$\cdot \prod_{d \in \boldsymbol{D}} \underbrace{p(x_d, t_d \mid \boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{Z})}_{\text{parse trees}} \tag{8}$$

The probability of an observed string and a tree from an adaptor grammar is based on the productions used to form the tree t_d, both adapted productions $c \Rightarrow z_{(a,i)} \in M(t_d)$ and unadapted productions $b \to \beta \in N(t_d)$,

$$p(x_d, t_d \mid \boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{Z}) = \prod_{c \Rightarrow z_{c,i}} \pi_{c \Rightarrow z_{c,i}}^{|c \Rightarrow z_{c,i} \in M(t_d)|} \prod_{b \to \beta} \theta_{b \to \beta}^{|b \to \beta \in N(t_d)|} \mathbb{I}\left[\textsc{Yield}\left(t_d\right) = x_d\right]. \tag{9}$$

Similarly, the probability of observing an atom $z_{c,i}$ from nonterminal c is

$$p(z_{c,i} \mid G_c) = \prod_{d \Rightarrow z_{d,j}} \pi_{d \Rightarrow z_{d,j}}^{|d \Rightarrow z_{d,j} \in M(z_{c,i})|} \prod_{b \to \beta} \theta_{b \to \beta}^{|b \to \beta \in N(z_{c,i})|} \tag{10}$$

3 Variational Distribution

Our goal is to apply variational inference to the problem of uncovering the latent variables of the adaptor grammar problem.

After positing a variational distribution, this will induce a variational objective, which is a lower bound of the likelihood. We optimize this objective to obtain a distribution over the latent variables that will approximate the true posterior.

We posit a mean-field variational distribution that breaks the dependencies by assuming the following variational distribution q on the variables.

1. $\pi'_{c,i}$ is drawn from a variational Beta distribution $\pi'_{c,i} \sim \text{Beta}(\nu^1_c, \nu^2_c)$ Atom weights are deterministically defined by $\pi_{c,i} = \pi'_{c,i} \prod_{j=1}^{i}(1 - \pi'_{c,j})$.

2. This distribution is truncated, so that $\pi'_{c,K} \equiv 1$ for some index K_c. This implies that $\pi_{c,i}$ is zero beyond index K.

3. Because the distribution is truncated, we keep a set of adapted yields, which we call the **truncated nonterminal grammar** (TNG). For an adapted nonterminal, TNG_c represents a finite subset of $\boldsymbol{z}_c$, the trees weighted by the variational distribution. The size of TNG_c is K_c, a user-defined truncation parameter (larger K_c results in a higher-fidelity variational approximation).

4. θ_c, the multinomial over a nonterminal's unadapted rules, is governed by a variational Dirichlet distribution $\boldsymbol{\theta}_c \sim \text{Dir}(\boldsymbol{\gamma}_c)$.

5. We assume that the set of all trees that can produce a sentence can be modeled as a multinomial distribution. Thus, a tree t_d can be associated with a multinomial distribution $\phi_d(t_d)$ for the entire tree.

Given these distributions for each of the latent variables, we assume a mean-field distribution

$$q(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T} | \boldsymbol{\gamma}, \boldsymbol{\nu}) = \prod_{c \in \boldsymbol{N}} q(\boldsymbol{\theta}_c | \boldsymbol{\gamma}_c) \prod_{c \in \boldsymbol{M}} \prod_{i=1}^{\infty} q(\pi_{c,i} | \nu^1_{c,i}, \nu^2_{c,i}) \prod_{d}^{|\boldsymbol{D}|} q(t_{d,i} | \phi_{d,i}) \tag{11}$$

4 Inference

The evidence lower bound (ELBO) of the model is[2]

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{D}) = & \sum_{c \in \boldsymbol{N} \setminus \boldsymbol{M}} \underbrace{\mathbb{E}_q \left[\log p(\boldsymbol{\theta}_c \mid \boldsymbol{\alpha}_c)\right]}_{\text{PCFG rules}} + \sum_{c \in \boldsymbol{M}} \sum_{i=1}^{\infty} \underbrace{\mathbb{E}_q \left[\log p(\pi'_{c,i} \mid a_c, b_c)\right]}_{\text{PY stick}} \\
& + \sum_{c \in \boldsymbol{M}} \sum_{i=1}^{\infty} \underbrace{\mathbb{E}_q \left[\log p(z_{c,i} \mid \boldsymbol{\pi}, \boldsymbol{\theta})\right]}_{\text{PY atoms}} + \sum_{d \in \boldsymbol{D}} \underbrace{\mathbb{E}_q \left[\log p(x_d, t_d \mid \boldsymbol{\pi}, \boldsymbol{\theta})\right]}_{\text{Observations}} \\
& + \underbrace{\sum_{c \in \boldsymbol{M}} \mathbb{H}_q \left[q(\boldsymbol{\theta}_c)\right] + \sum_{c \in \boldsymbol{M}} \sum_{i=1}^{\infty} \mathbb{H}_q \left[q(\pi'_{c,i})\right]}_{\text{Entropy Terms}}
\end{aligned} \tag{12}$$

In this section, we further expand both the likelihood terms and the entropy terms to create a full objective function. Next, we derive coordinate-ascent updates for each of the variables in the objective that will optimize the objective function.

[2]For the sake of clarity, we have omitted the hyperparameters, which would make the complete expression $\mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{D}; \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\alpha})$

Utility Expectations Many of the expectations involve well-known log expectations of exponential family distributions. We state them here without explicit derivation.

$$\mathbb{E}_q[\log \pi'_{c,i}] = \Psi\left(\nu^1_{c,i}\right) - \Psi\left(\nu^1_{c,i} + \nu^2_{c,i}\right) \tag{13}$$

$$\mathbb{E}_q[\log(1 - \pi'_{c,i})] = \Psi\left(\nu^2_{c,i}\right) - \Psi\left(\nu^1_{c,i} + \nu^2_{c,i}\right) \tag{14}$$

$$\begin{aligned}\mathbb{E}_q\left[\log \pi_{c,i}\right] &= \sum_{j=1}^{i-1} \mathbb{E}_q[\log(1 - \pi'_{c,j})] + \mathbb{E}_q[\log \pi'_{c,i}] \\ &= \sum_{j=1}^{i-1} \left(\Psi\left(\nu^2_{c,j}\right) - \Psi\left(\nu^1_{c,j} + \nu^2_{c,j}\right)\right) + \Psi\left(\nu^1_{c,i}\right) - \Psi\left(\nu^1_{c,i} + \nu^2_{c,i}\right)\end{aligned} \tag{15}$$

4.1 Expanding Expectations

PCFG Rule Each nonterminal has a distribution over rules θ_c; the ELBO term associated with this multinomial is

$$\begin{aligned}\mathbb{E}_q\left[\log p(\boldsymbol{\theta}_c|\boldsymbol{\alpha}_c)\right] = &\log \Gamma\left(\textstyle\sum_{c\to\beta\in\boldsymbol{R}_c} \alpha_{c\to\beta}\right) - \sum_{c\to\beta\in\boldsymbol{R}_c} \log\Gamma\left(\alpha_{c\to\beta}\right) \\ &+ \sum_{c\to\beta\in\boldsymbol{R}_c} (\alpha_{c\to\beta} - 1)\mathbb{E}_q\left[\log\theta_{c\to\beta}\right],\end{aligned} \tag{16}$$

which can be further expanded using the expectation of a Dirichlet,

$$\mathbb{E}_q\left[\log\theta_{c\to\beta}\right] = \Psi\left(\gamma_{c\to\beta}\right) - \Psi\left(\textstyle\sum_{c\to\beta'\in\boldsymbol{R}_c} \gamma_{c\to\beta'}\right) \tag{17}$$

PY Stick The Pitman-Yor distribution has two components, a weighting over atoms and the atoms themselves. The ELBO term corresponding to the distribution over atom weights, π, is

$$\begin{aligned}\mathbb{E}_q\left[\log p(\pi'_{c,i} \mid a_c, b_c)\right] = &\log\Gamma\left(1 - b_c + a_c + ib_c\right) - \log\Gamma\left(1 - b_c\right) - \log\Gamma\left(a_c + ib_c\right) \\ &- b_c\mathbb{E}_q\left[\log\pi'_{c,i}\right] + (a_c + ib_c - 1)\mathbb{E}_q\left[\log(1 - \pi'_{c,i})\right]\end{aligned} \tag{18}$$

PY Atoms The atoms weighted by the Pitman-Yor distribution appear (Equation 10) in the ELBO term

$$\begin{aligned}\mathbb{E}_q\left[\log p(z_{c,i}|\boldsymbol{\pi},\boldsymbol{\theta})\right] = &\sum_{b\to\beta\in N(z_{c,i})} g(b\to\beta, z_{c,i})\mathbb{E}_q\left[\log\theta_{b\to\beta}\right] \\ &+ \sum_{b\Rightarrow z_{b,k}\in M(z_{c,i})} f(b\Rightarrow z_{b,k}, z_{c,i})\mathbb{E}_q\left[\log\pi_{b,k}\right].\end{aligned} \tag{19}$$

Observations Finally, observed trees are described by both adapted and unadapted rules (Equation 9) which contribute to the ELBO,

$$\mathbb{E}_q\left[\log p(x_d, t_d \mid \boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{Z})\right] = \sum_{b\Rightarrow z_{b,k}} \mathbb{E}_q\left[\log p(\pi_{b,k})\right] + \sum_{b\to\beta} \mathbb{E}_q\left[\log p(\theta_{b\to\beta})\right] \tag{20}$$

Entropy Terms Entropy terms for Dirichlet distribution

$$\begin{aligned}\mathbb{H}_q\left[\theta_c \mid \boldsymbol{\gamma}_c\right] = &-\log\Gamma\left(\textstyle\sum_{c\to\beta\in\boldsymbol{R}_c} \gamma_{c\to\beta}\right) + \sum_{c\to\beta\in\boldsymbol{R}_c} \log\Gamma\left(\gamma_{c\to\beta}\right) \\ &- \sum_{c\to\beta\in\boldsymbol{R}_c} (\gamma_{c\to\beta} - 1)\mathbb{E}_q\left[\log\theta_{c\to\beta}\right]\end{aligned} \tag{21}$$

Entropy term for Pitman-Yor process

$$\begin{aligned}\mathbb{H}_q\left[\pi'_{c,i} \,|\, \nu_{c,i}\right] = &-\log\Gamma\left(\nu^1_{c,i}+\nu^2_{c,i}\right)+\log\Gamma\left(\nu^1_{c,i}\right)+\log\Gamma\left(\nu^2_{c,i}\right)\\ &-(\nu^1_{c,i}-1)\mathbb{E}_q\left[\log\pi'_{c,i}\right]-(\nu^2_{c,i}-1)\mathbb{E}_q\left[\log(1-\pi'_{c,i})\right]\end{aligned} \tag{22}$$

4.2 Update for ϕ

Our distribution over trees for an observation is governed by a multinomial variational parameter ϕ. The coordinate ascent should update this equation based on the expectation of the other variational parameters excluding ϕ_d (we denote this expectation using $\mathbb{E}_{\neg\phi_d}[\cdot]$)

$$\begin{aligned}\log\phi_{d,i} =&\mathbb{E}_{\neg\phi_d}\left[\log p(x_d, t_{d,i})\right]+\text{CONST}\\ \phi_{d,i} \propto&\exp\left\{\mathbb{E}_{\neg\phi_d}\left[\log p(x_d, t_{d,i})\right]\right\}\\ =&\exp\left\{\mathbb{E}_{\neg\phi_d}\left[\log p(t_{d,i}|x_d)+\log p(x_d)\right]\right\}\\ \propto&\exp\left\{\mathbb{E}_{\neg\phi_d}\left[\log p(t_d|x_d)\right]\right\}.\end{aligned} \tag{23}$$

Instead of explicitly computing this expectation, we use the hybrid-MCMC approach (Mimno et al., 2012) to sample from $\exp\{\mathbb{E}_{-T_{d,i}}[\log p(\boldsymbol{T}|\boldsymbol{D})]\}$ to obtain a sparse estimate of this expectation. We approximate the above equation by sampling from the distribution over trees given a yield (Johnson et al., 2006).

We can create a PCFG that approximates an adaptor grammar(Cohen, 2011). This PCFG has a set of productions

$$R' = R \cup_{c\in N}\left(A \to \textsc{Yield}\left(x\right) : x \in \boldsymbol{x}_c\right) \tag{24}$$

weighted by

$$\log\theta'_{c\to\beta} = \begin{cases}\mathbb{E}_q\left[\log\pi_i\right], & \text{if } \text{TNG}_c(i) = \beta\\ \mathbb{E}_q\left[\log\pi_{K_c}\right]+\mathbb{E}_q\left[\log\theta_{c,\beta}\right], & \text{otherwise}\end{cases}$$

and then build a collection of sampled trees $S_d \equiv s_{d,1}, \ldots, s_{d,k}$. We then approximate our variational distribution over trees as:

$$\phi_d(t) = \begin{cases}\dfrac{|s_d \in S_d : s_d = t|}{|S_d|} & t \in S_d\\ 0 & \text{otherwise}\end{cases} \tag{25}$$

4.3 Total Counts

A sampled tree $s_{d,k}$ has three types of productions, whose overall prevalence in the corpus we will represent using counts f (adapted rules), g (unadapted rules), h (potentially adapted):

- f is the expected number of productions that are represented within the TNG. It sums over all trees, weighting the count of the productions in a tree by the probability of the tree under the variational distribution,

$$f_d(c \Rightarrow z_{c,i}) = \sum_t \phi_{d,t} \underbrace{|c \Rightarrow z_{c,i} : c \Rightarrow z_{c,i} \in M_{\text{TNG}_c}(t)|}_{\text{Count of production in tree } t_{d,t}} \tag{26}$$

- g is the expected number of productions used by the base distribution

$$g_d(b \to \beta) = \sum_t \phi_{d,t} |b \to \beta : b \to \beta \in N_R(t)| \tag{27}$$

- Finally, there is a third set of productions that could be adapted, but are not because they are not represented in our TNG. These are subtrees not in TNG_c rooted at an adapted nonterminal c,

$$h_d(c \Rightarrow z_{c,i}) = \sum_t \phi_{d,t} |c \Rightarrow z_{c,i} : c \Rightarrow z_{c,i} \in M_{\neg\text{TNG}_c}(t)|, \tag{28}$$

where $\neg\text{TNG}_c$ represents subtrees rooted at a **not** present in the TNG and $M_{\neg\text{TNG}_c}(t)$ represents the multi-set of subtrees rooted at a not present in the TNG_c but that appeared in tree derivation t.

4.4 Batch Updates for Global Parameters

For reference, we give the closed-form updates for the variational distributions for the nonterminal productions γ and for the stick breaking weights ν; these rely on optimizing $\mathcal{L}$ with respect to a single variational parameter, replacing that variational distribution with its update. While we will not use them for our algorithm, we write them here for comparison with previous variational inference algorithms (Cohen et al., 2010; Cohen, 2011) and for comparison against the online updates, which are discussed in the next section, Section 5.

4.4.1 Optimize γ

The update for the variational parameter governing the probability over unadapted rules is

$$\gamma_{c\to\beta} = \underbrace{\alpha_{c\to\beta}}_{\text{prior}} + \underbrace{\sum_{d\in\boldsymbol{D}} g_d(c\to\beta)}_{\text{rules in data}} + \underbrace{\sum_{b\in\boldsymbol{M}}\sum_{i=1}^{K_b} |c\to\beta : c\to\beta \in N(z_{b,i})|}_{\text{rules in adapted rules}}. \tag{29}$$

4.4.2 Optimize ν

The update for the variational parameter governing the stick-breaking weight for the i-th atom associated with nonterminal a is

$$\nu^1_{c,i} = \underbrace{\sum_{b\in\boldsymbol{M}}\sum_{k=1}^{K_b} |c\Rightarrow z_{c,i} : c\Rightarrow z_{c,i}\in M(z_{b,k})|}_{\text{Adapted rules of nonterminal } c \text{ used in } b\text{'s rules}} + \underbrace{\sum_{d\in\boldsymbol{D}} f_d(c\Rightarrow z_{c,i}) - b_c + 1}_{\text{Adapted rules in corpus}} \tag{30}$$

$$\nu^2_{c,i} = \sum_{b\in\boldsymbol{M}}\sum_{k=1}^{K_b}\sum_{j=1}^{K_c} |c\Rightarrow z_{c,j} : c\Rightarrow z_{c,j}\in M(z_{b,k})| \tag{31}$$

$$+ \sum_{d\in\boldsymbol{D}}\sum_{j=1}^{K_c} f_d(c\Rightarrow z_{c,j}) + a_c + ib_c. \tag{32}$$

5 Online Inference

Instead of processing all of the data B in a single batch, we will instead break our data into minibatches B^e. We will consistently use the superscript e to denote the current epoch of data we have observed. Each epoch e processes a minibatch and creates a set of candidate parses, creating the variational distribution ϕ over the possible interpretation of parses. The statistics from these parses then determine the global parameter updates after each minibatch.

The remainder of this section details some of the steps referenced in the algorithm.

5.1 Online Parameter Updates

We use the incremental EM method (Neal & Hinton, 1998). We accumulate sufficient statistics, and update the variational parameters γ, ϕ, ν. We denote $\tilde{f}^e$ and $\tilde{g}^e$ as the accumulated count of f and g values at iteration e,

$$\tilde{f}^e(c\Rightarrow z_{c,i}) = (1-\epsilon)\tilde{f}^{e-1}(c\Rightarrow z_{c,i}) + \epsilon\sum_d f_d(c\Rightarrow z_{c,i})$$
$$\tilde{g}^e(b\to\beta) = (1-\epsilon)\tilde{g}^{e-1}(b\to\beta) + \epsilon\sum_d g_d(b\to\beta) \tag{33}$$

where ϵ is a scaling parameter between 0 to 1. In this paper, we choose it as

$$\epsilon = (\tau + e)^{-\kappa} \tag{34}$$

where τ is a user defined inertia, e is the epoch counter and κ is scaling rate. So that in the online case, the update for epoch e is,

$$\gamma_{c\to\beta} = \alpha_{c\to\beta} + \tilde{g}^e(c \to \beta) + \sum_{b\in\boldsymbol{M}} \sum_{i=1}^{K_b} n(c \to \beta, z_{b,i}). \tag{35}$$

$$\nu^1_{c,i} = \sum_{b\in\boldsymbol{M}} \sum_{k=1}^{K_b} n(c \Rightarrow z_{c,i}, z_{b,k}) + \tilde{f}^e(c \Rightarrow z_{c,i}) - b_c + 1 \tag{36}$$

$$\nu^2_{c,i} = \sum_{b\in\boldsymbol{M}} \sum_{k=1}^{K_b} \sum_{j=1}^{K_c} n(c \Rightarrow z_{c,j}, z_{b,k}) + \sum_{j=1}^{K_c} \tilde{f}^e(c \Rightarrow z_{c,j}) + a_c + ib_c. \tag{37}$$

5.2 Refining TNG

After processing a minibatch of sentences (creating expected counts f, g, h), we must update our truncation set TNG. The process of updating the truncation set occurs in two stages: adding rules and reordering rules.

Adding Rules We add potentially adapted nonterminals after each minibatch. For a nonterminal c, we sort productions by the candidate count h and then add all of those productions to the TNG_c. After adding rules to TNG_c, the production counts (Equations 26, 27, and 28), e.g. counts previously associated with a candidate production (h) could now be associated with an adapted production (f).

Reordering Rules After every U minibatches we prune our TNG. After adding new rules seen in this minibatch (which happens after every minibatch), we sort the TNG by the number of times the rules have been seen and remove all but the top K_c rules.

6 Notation List

This section serves as a reference for all of the notation used in this document.

PCFG Notation

- $\boldsymbol{N}$ all the nonterminals
- $\boldsymbol{M}$ all the adapted nonterminals
- $a \in \boldsymbol{N}$ a nonterminal node
- $a \in \boldsymbol{M}$ an adapted nonterminal
- $a \to \beta$: a PCFG rule.
- $\boldsymbol{R}$: all grammar rules.
- $\boldsymbol{R}_c$: the subset of grammar rules with LHS a
- $\boldsymbol{\theta}_c$: the multinomial distribution over unadapted rules with LHS a
- α_c: the Dirichlet distribution prior parameter over θ_c

Adaptor Grammar Notation

- $\pi'_{c,i}$: the i-th draw from Beta distribution used in a stick breaking process for Pitman-Yor process.
- $\pi_{c,i}$: the weight assigned to the i-th stick in a stick breaking process, also known as the probability assigned to the i-th possible tree of the adapted nonterminal c.
- $\boldsymbol{z}_c$: the set of all possible trees rooted at an adapted nonterminal c. The i-th tree $z_{c,i}$ will have a corresponding stick breaking weight $\pi_{c,i}$.
- $c \Rightarrow z_{c,i}$: the adapted rule associated to the i-th stick in the stick breaking process associated with nonterminal c.

Data Notation

- t_d the tree generate the d^{th} string in the dataset.
- $N_R(t)$ the multi-set (may contain duplicates) of unadapted productions used in tree t; the set of productions from the original grammar rules R.
- $M_{\boldsymbol{Z}}(t)$ the multi-set (may contain duplicates) of adapted productions used in tree t; the set of productions encoded by some atom in $\boldsymbol{Z}$
- $x_d \in \boldsymbol{X}$ the observation associated with the d^{th} sentence.

Variational Distribution

- $q(\pi \mid \nu)$, a variational Beta for each stick weight for each nonterminal
- $q(\theta_c \mid \gamma_c)$: a variational Dirichlet for each multinomial distribution over rules for a nonterminal
- $q(t_d \mid \phi)$: a multinomial over the possible parse trees of a sentence
- TNG_c: The truncated nonterminal grammar, keeping a set of the sentences we can expand from nonterminal c
- $S_d \equiv \{s_{d,1}, \ldots s_{d,k}\}$: The set of parses we use to approximate ϕ

Expected Production Counts

- $f_d(c \Rightarrow z_{c,i})$: The expected number of adapted rule i from nonterminal a in document d
- $g_d(c \rightarrow \beta)$: The expected number of unadapted rule from nonterminal b in document d
- $h_d(c \Rightarrow z_{c,i})$: The expected number of candidate adapted rules from nonterminal c in document d

Online Update Parameters

- e: index of an epoch
- X^e: a minibatch of documents, the input to an epoch
- $\epsilon_e = (\tau + e)^\kappa$ the scaling parameter for epoch e
- τ: inertia
- κ: scaling rate
- U: reordering delay, how many epochs pass before sorting and truncating the TNG

References

Cohen, Shay B. Computational Learning of Probabilistic Grammars in the Unsupervised Setting. PhD thesis, Carnegie Mellon University, 2011.

Cohen, Shay B., Blei, David M., and Smith, Noah A. Variational inference for adaptor grammars. In NAACL, 2010.

Johnson, Mark, Griffiths, Thomas L., and Goldwater, Sharon. Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models. In NIPS, 2006.

Mimno, David, Hoffman, Matthew, and Blei, David. Sparse stochastic inference for latent Dirichlet allocation. In ICML, 2012.

Neal, Radford and Hinton, Geoffrey E. A view of the em algorithm that justifies incremental, sparse, and other variants. In Learning in Graphical Models, pp. 355–368. Kluwer Academic Publishers, 1998.