Empirical Risk Minimization with Approximations of Probabilistic Grammars

Shay B. Cohen
Language Technologies Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
scohen@cs.cmu.edu

Noah A. Smith
Language Technologies Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
nasmith@cs.cmu.edu

A Supplementary Material

A longer version of the paper should be available soon [2]; we anticipate it will include:

1. Discussion of the relationship of our framework to Dirichlet smoothing.
2. Discussion of the relationship of our distributional assumptions to Tsybakov noise condition.
3. An explanation about the derivation of simpler complexity bounds in the supervised case.
4. An empirical exploration of the distributional assumptions.
5. A description of NP-hardness results for empirical risk minimization in the unsupervised setting.
6. A description of the variants of MLE algorithms needed to be used in our framework.

A.1 The $N_k = 2$ Assumption

When approximating a family of probabilistic grammars, it is convenient to assume the degree of the grammar is limited. We limit the degree of the grammar by making the assumption that $N_k \leq 2$. This assumption may seem, at first glance, somewhat restrictive, but we show next that for probabilistic context-free grammars (and as a consequence, other formalisms), this assumption does not restrict generative capacity.

We first show that any context-free grammar with arbitrary degree can be mapped to a new grammar that generates derivations that can be transformed back to derivations in the original grammar. Such a grammar is also called a "covering grammar." Let $\boldsymbol{G}$ be a CFG. Let A be the kth nonterminal. Consider the rules $A \rightarrow \alpha_i$ for $i \leq N_k$ where A appears on the left side. For each rule $A \rightarrow \alpha_i$, $i < N_k$, we create a new nonterminal in $\boldsymbol{G}'$ such that A_i has two rewrite rules: $A_i \rightarrow \alpha_i$ and $A_i \rightarrow A_{i+1}$. In addition, we create a rule $A \rightarrow A_1$ and $A_{N_k} \rightarrow \alpha_{N_k}$.

It is easy to verify that the resulting grammar $\boldsymbol{G}'$ has an equivalent capacity to the original CFG, $\boldsymbol{G}$. A simple transformation that converts each derivation in the new grammar to a derivation in the old grammar would be to collapse any path of nonterminals that were added to $\boldsymbol{G}'$ (i.e. all A_i for nonterminal A) such that we end up with nonterminals from the original grammar only. Similarly, any derivation in $\boldsymbol{G}$ can be converted to a derivation in $\boldsymbol{G}'$, by adding new nonterminals through unary application of rules of the form $A_i \rightarrow A_{i+1}$. Given a derivation z in $\boldsymbol{G}$, we denote by $\Upsilon_{\boldsymbol{G} \mapsto \boldsymbol{G}'}(z)$ the corresponding derivation in $\boldsymbol{G}'$ after adding the new non-terminals A_i to z. Throughout the paper, we will refer to the normalized form of $\boldsymbol{G}'$ as "binary normal form."[1]

[1]This notion of binarization is different from previous types of binarization for grammars. In most cases, previous work binarized grammars to have at most two nonterminals on the right side (i.e., Chomsky normal

Note that K', the number of multinomials in the binary normal form is a function of both the number of nonterminals in the original grammar and the number of rules in that grammar. More specifically, we have that $K' = \sum_{k=1}^{K} N_k + K$. To make the equivalence complete, we need to show that any *probabilistic* context-free grammar can be translated to a PCFG with $\max_k N_k \leq 2$ such that the two PCFGs induce equivalent distributions over derivations.

Lemma A.1. *Let $a_i \in [0,1]$, $i \in \{1,\ldots,N\}$ such that $\sum_i a_i = 1$. Define $b_1 = a_1$, $c_1 = 1 - a_1$, $b_i = \left(\frac{a_i}{a_{i-1}}\right)\left(\frac{b_{i-1}}{c_{i-1}}\right)$ and $c_i = 1 - b_i$ for $i \geq 2$. Then $a_i = \left(\prod_{j=1}^{i-1} c_j\right) b_i$.*

Proof. See [2]. □

Theorem A.2. *Let $\langle \boldsymbol{G}, \boldsymbol{\theta} \rangle$ be a probabilistic context-free grammar. Let $\boldsymbol{G}'$ be the transformation of $\boldsymbol{G}$ as defined above. Then, there exists $\boldsymbol{\theta}'$ for $\boldsymbol{G}'$ such that for any $z \in D(\boldsymbol{G})$ we have $\mathbb{Q}(z \mid \boldsymbol{\theta}, \boldsymbol{G}) = \mathbb{Q}(\Upsilon_{\boldsymbol{G} \mapsto \boldsymbol{G}'}(z) \mid \boldsymbol{\theta}', \boldsymbol{G}')$ where $\mathbb{Q}(\cdot \mid \boldsymbol{\theta}, \boldsymbol{G})$ (or $\mathbb{Q}(\cdot \mid \boldsymbol{\theta}', \boldsymbol{G}')$) is the probability distribution for $\langle \boldsymbol{G}, \boldsymbol{\theta} \rangle$ (or $\langle \boldsymbol{G}', \boldsymbol{\theta}' \rangle$).*

Proof. For the grammar $\boldsymbol{G}$, index the set $\{1, ..., K\}$ with nonterminals ranging from A_1 to A_K. Define $\boldsymbol{G}'$ as above. We need to define $\boldsymbol{\theta}'$. Index the multinomials in $\boldsymbol{G}'$ by (k, i), each having two events. Let $\mu_{(k,i),1} = \theta_{k,i}$, $\mu_{(k,i),2} = 1 - \theta_{k,i}$ for $i = 1$ and set $\mu_{k,i,1} = \theta_{k,i} / \mu_{(k,i-1),2}$, and $\mu_{(k,i-1),2} = 1 - \mu_{(k,i-1),2}$.

$\langle \boldsymbol{G}', \boldsymbol{\mu} \rangle$ is a *weighted* context-free grammar such that the $\mu_{(k,i),1}$ corresponds to the ith event in the k multinomial of the original grammar. Let z be a derivation in $\boldsymbol{G}$ and $z' = \Upsilon_{\boldsymbol{G} \mapsto \boldsymbol{G}'}(z)$. Then, from Utility Lemma A.3 and the construction of g', we have that:

$$
\begin{aligned}
\mathbb{Q}(z \mid \boldsymbol{\theta}, \boldsymbol{G}) &= \prod_{k=1}^{K} \prod_{i=1}^{N_k} \theta_{k,i}^{\psi_{k,i}(z)} && \text{(A1)} \\
&= \prod_{k=1}^{K} \prod_{i=1}^{N_k} \prod_{l=1}^{\psi_{k,i}(z)} \theta_{k,i} && \text{(A2)} \\
&= \prod_{k=1}^{K} \prod_{i=1}^{N_k} \prod_{l=1}^{\psi_{k,i}(z)} \left(\prod_{j=1}^{i-1} \mu_{(k,j),2}\right) \mu_{k,i,1} && \text{(A3)} \\
&= \prod_{k=1}^{K} \prod_{i=1}^{N_k} \left(\prod_{j=1}^{i-1} \mu_{(k,j),2}^{\psi_{k,i}(z)}\right) \mu_{k,i,1}^{\psi_{k,i}(z)} && \text{(A4)} \\
&= \prod_{k=1}^{K} \prod_{j=1}^{N_k} \prod_{i=1}^{2} \mu_{(k,j),i}^{\psi_{k,j}(z')} && \text{(A5)} \\
&= \mathbb{Q}(z' \mid \boldsymbol{\mu}, \boldsymbol{G}') && \text{(A6)}
\end{aligned}
$$

From [1], we know that the weighted grammar $\langle \boldsymbol{G}', \boldsymbol{\mu} \rangle$ can be converted to a probabilistic context-free grammar $\langle \boldsymbol{G}', \boldsymbol{\theta}' \rangle$, through a construction of $\boldsymbol{\theta}'$ based on $\boldsymbol{\mu}$, such that $\mathbb{Q}(z' \mid \boldsymbol{\mu}, \boldsymbol{G}') = \mathbb{Q}(z' \mid \boldsymbol{\theta}', \boldsymbol{G}')$. □

The proof for Lemma A.2 gives a construction the parameters $\boldsymbol{\theta}'$ of $\boldsymbol{G}'$ such that $\langle \boldsymbol{G}, \boldsymbol{\theta} \rangle$ is equivalent to $\langle \boldsymbol{G}', \boldsymbol{\theta}' \rangle$. The construction of $\boldsymbol{\theta}'$ can also be reversed: given $\boldsymbol{\theta}'$ for $\boldsymbol{G}'$, we can construct $\boldsymbol{\theta}$ for $\boldsymbol{G}$ so that again we have equivalence between $\langle \boldsymbol{G}, \boldsymbol{\theta} \rangle$ and $\langle \boldsymbol{G}', \boldsymbol{\theta}' \rangle$.

form). Another form of binarization for linear context-free rewriting systems is limiting the *fan-out* of the rules to two [5, 4]. We limit the number of *rules* for each nonterminal (or more generally, the number of elements in each multinomial).

A.2 Proof of Proposition 4.2

Proposition 4.2. There exists an M such that for any $m > M$ we have: $\mathbb{P}\left(\bigcup_{f\in\mathcal{F}}\{z \mid C_m(f)(z) - f(z) \geq \epsilon_{\text{tail}}(m)\}\right) \leq \epsilon_{\text{tail}}(m)$ for $\epsilon_{\text{tail}}(m) = \dfrac{N\log^2 m}{m^p - 1}$ and $C_m(f) = T(f, m^{-p})$.

Utility Lemma A.3. *(From [3].) Let $a \in [0,1]$ and let $b = a$ if $a \in [\gamma, 1-\gamma]$, $b = \gamma$ if $a \leq \gamma$ and $b = 1-\gamma$ if $a \geq 1-\gamma$. Then for any $\epsilon \leq 1/2$ such that $\gamma \leq \epsilon/(1+\epsilon)$ we have $\log a/b \leq \epsilon$.*

Proof of Proposition 4.2. Let $\mathcal{Z}(m)$ be the set of derivations of size bigger than $\log^2 m$. Let $f \in \mathcal{F}$. Define $f' = T(f, m^{-p})$. For any $z \notin \mathcal{Z}(m)$ we have that:

$$\begin{aligned} f'(z) - f(z) &= -\sum_{k=1}^{K}\left(\phi_{k,1}(z)\log\theta_{k,1} + \phi_{k,2}(z)\log\theta_{k,2} - \phi_{k,1}(z)\log\theta'_{k,1} - \phi_{k,1}(z)\log\theta'_{k,2}\right) \\ &\leq \sum_{k=1}^{K}\log^2 m\left(\max\{0, \log(\theta'_{k,1}/\theta_{k,1})\} + \max\{0, \log(\theta'_{k,2}/\theta_{k,2})\}\right) \end{aligned} \tag{A7}$$

Without loss of generality, assume $\epsilon_{\text{tail}}(n)/N\log^2 m \leq 1/2$. Let $\gamma = \dfrac{\epsilon_{\text{tail}}(m)/N\log^2 m}{1 + \epsilon_{\text{tail}}(m)/N\log^2 m} = 1/m^p$. From Utility Lemma A.3 we have that $\log(\theta'_{k,i}/\theta_{k,i}) \leq \epsilon_{\text{tail}}(m)/N\log m$. Plug this in into Eq. A7 ($N = 2K$) to get that for all $z \notin \mathcal{Z}(m)$ we have $f'(z) - f(z) \leq \epsilon_{\text{tail}}(m)$. It remains to show that the measure $\mathbb{P}(\mathcal{Z}(m)) \leq \epsilon_{\text{tail}}(m)$. Note that $\sum_{z\in\mathcal{Z}(m)}\mathbb{P}(z) \leq \sum_{k>\log^2 m} L\Lambda(k)r^k \leq L\sum_{k>\log^2 m} q^k = Lq^{\log^2 m}/(1-q) < \epsilon_{\text{tail}}(m)$ for $m > M$ where M is fixed. □

A.3 Proof of Lemma 4.5

Lemma 4.5. *Denote by $\mathcal{Z}_{\epsilon,n}$ the set $\bigcup_{f\in\mathcal{F}}\{z \mid C_n(f)(z) - f(z) \geq \epsilon\}$. Denote by $A_{\epsilon,n}$ the event "one of $z_i \in D$ is in $\mathcal{Z}_{\epsilon,n}$." Then if $\mathcal{F}_n$ properly approximates $\mathcal{F}$ then:*

$$\begin{aligned} &\mathbb{E}\left[R_{\text{emp},n}(g_n) - R_{\text{emp},n}(f_n^*)\right] \\ &\leq \left|\mathbb{E}\left[R_{\text{emp},n}(C_n(f_n^*)) \mid A_{\epsilon,n}\right]\right|\mathbb{P}(A_{\epsilon,n}) + \left|\mathbb{E}\left[R_{\text{emp},n}(f_n^*) \mid A_{\epsilon,n}\right]\right|\mathbb{P}(A_{\epsilon,n}) + \epsilon_{\text{tail}}(n) \end{aligned} \tag{A8}$$

where the expectations are taken with respect to the dataset D.

Proof. Consider the following:

$$\mathbb{E}[R_{\text{emp},n}(g_n) - R_{\text{emp},n}(f_n^*)] \tag{A9}$$
$$= \mathbb{E}[R_{\text{emp},n}(g_n) - R_{\text{emp},n}(C_n(f_n^*)) + R_{\text{emp},n}(C_n(f_n^*)) - R_{\text{emp},n}(f_n^*)] \tag{A10}$$
$$= \mathbb{E}[R_{\text{emp},n}(g_n) - R_{\text{emp},n}(C_n(f_n^*))] + \mathbb{E}[R_{\text{emp},n}(C_n(f_n^*)) - R_{\text{emp},n}(f_n^*)] \tag{A11}$$

Note first that $\mathbb{E}[R_{\text{emp},n}(g_n) - R_{\text{emp},n}(C_n(f_n^*))] \leq 0$, by the definition of g_n as the minimizer of $R_{\text{emp},n}$. We next bound $\mathbb{E}[R_{\text{emp},n}(C_n(f_n^*)) - R_{\text{emp},n}(f_n^*)]$. We know that from the requirement of proper approximation that we have:

$$\mathbb{E}[R_{\text{emp},n}(C_n(f_n^*)) - R_{\text{emp},n}(f_n^*)] \tag{A12}$$
$$= \mathbb{E}[R_{\text{emp},n}(C_n(f_n^*)) - R_{\text{emp},n}(f_n^*) \mid A_{\epsilon,n}]\mathbb{P}(A_{\epsilon,n}) + \tag{A13}$$
$$\mathbb{E}[R_{\text{emp},n}(C_n(f_n^*)) - R_{\text{emp},n}(f_n^*) \mid \neg A_{\epsilon,n}](1 - \mathbb{P}(A_{\epsilon,n}))$$
$$\leq |\mathbb{E}[R_{\text{emp},n}(C_n(f_n^*)) \mid A_{\epsilon,n}]|\mathbb{P}(A_{\epsilon,n}) + |\mathbb{E}[R_{\text{emp},n}(f_n^*) \mid A_{\epsilon,n}]|\mathbb{P}(A_{\epsilon,n}) + \epsilon_{\text{tail}}(n) \tag{A14}$$

and that equals the right side of Eq. A8. □

A.4 Explanation about Lemma 5.1

Lemma 5.1 is a variant of Theorem 24 in Pollard [6, chapter 2, pages 25–27]. We use the notation of Pollard in this section for compatibility. Pf is the expected value of f under distribution P and $P_n f$ is the expected function of f under the empirical distribution P_n.

Pages 30–31 in Pollard describe an extension of Theorem 24, where the empirical process in Theorem 24, $\sup_{f\in\mathcal{F}} |P_n f - Pf|$ is changed such that $\mathcal{F}$ become $\mathcal{F}_n$ (i.e., dependent on n). However, this extension assumes that $|f| \leq K$ for some constant K, which is not true for the proper approximations. We describe how to adapt the extension of Theorem 24 to get Lemma 5.1 with K_n.

We follow the first part of the proof of Theorem 24, and bound each of the quantities $PF_n\{F_n > K_n\}$ and $P_nF\{F_n > K_n\}$ by $\epsilon/2$. We know that the quantity $PF_n\{F_n > K_n\}$, for some n, becomes smaller than $\epsilon/2$ because of the requirement for the expected value of the truncated functions to be smaller than $\epsilon_{\text{bound}}(n)$. In addition, we can use Markov inequality and the fact that $\mathbb{E}[P_nF\{F_n > K_n\}] = PF_n\{F_n > K_n\}$ to bound the probability that $P_nF\{F_n > K_n\}$ is bigger than $\epsilon/2$. This is how we get the second summand in the right side of the probability term in Lemma 5.1. At that point, we can just follow the proof of Theorem 24 and its extension in pages 30–31 to get Lemma 5.1, using the truncated set of functions $\mathcal{F}_{\text{truncated},n}$.

A.5 Boundedness Property in the Unsupervised Case

To complete §5.2 in the manuscript, we show that the boundedness property holds for $\mathcal{F}'_n$.

Proposition A.4. *There exists a $\beta'(L,p,q,N) > 0$ such that $\mathcal{F}'_m$ has the boundedness property with $K_m = pN\log^3 m$ and $\epsilon_{\text{bound}}(m) = m^{-\beta'\log m}$.*

Proof. From the requirement of $\mathbb{P}$, we know that for any x we have a z such that $\text{yield}(z) = x$ and $|z| \leq \alpha|x|$. Therefore, if we let $\mathcal{X}(m) = \{x \mid |x| \leq \log^2 m/\alpha\}$, then we have for any $f \in \mathcal{F}'_m$ and $x \in \mathcal{X}(m)$ that $f(x) \leq pN\log^3 m = K_m$ (similarly to the proof of Proposition 4.1). Denote by $f_1(x,z)$ the function in $\mathcal{F}_m$ such that $f(x) = -\log\sum_z \exp(-f_1(x,z))$.

In addition, from the requirements on $\mathbb{P}$ and the definition of K_m we have:

$$\begin{aligned}
\mathbb{E}\left[|f| \times I(|f| \geq K_m)\right] &= \sum_x \mathbb{P}(x)f(x)I(f \geq K_m) && \text{(A15)}\\
&= \sum_{x:|x|>\log^2 m/\alpha} \mathbb{P}(x)f(x) && \text{(A16)}\\
&\leq \sum_{x:|x|>\log^2 m/\alpha} \mathbb{P}(x)f_1(x,z(x)) && \text{(A17)}
\end{aligned}$$

where $z(x)$ is some derivation for x. We have:

$$\begin{aligned}
\sum_{x:|x|>\log^2 m/\alpha} \mathbb{P}(x)f_1(x,z(x)) &\leq \sum_{x:|x|\geq\log^2 m/\alpha} \sum_{z\in D_x(\boldsymbol{G})} \mathbb{P}(x,z)f_1(x,z(x)) && \text{(A18)}\\
&\leq pN\log m \sum_{x:|x|>\log^2 m/\alpha} \sum_z \mathbb{P}(x,z)|z(x)| && \text{(A19)}\\
&\leq pN\log m \sum_{k>\log^2 m} \Lambda(k)r^k k && \text{(A20)}\\
&\leq pN\log m \sum_{k>\log^2 m} q^k k \leq \kappa\log m q^{log^2 m} && \text{(A21)}
\end{aligned}$$

for some constant $\kappa > 0$. Finally, for some $\beta'(L,p,q,N) = \beta' > 0$ and some constant M, if $m > M$ then $\kappa\log m\left(q^{\log^2 m}\right) \leq m^{-\beta'\log m}$. □

References

[1] Z. Chi. Statistical properties of probabilistic context-free grammars. *Computational Linguistics*, 25(1):131–160, 1999.

[2] S. B. Cohen and N. A. Smith. Empirical risk minimization for probabilistic grammars: Sample complexity and hardness of learning, in preparation.

[3] S. Dasgupta. The sample complexity of learning fixed-structure Bayesian networks. *Machine Learning*, 29(2-3):165–180, 1997.

[4] D. Gildea. Optimal parsing strategies for linear context-free rewriting systems. In *Proceedings of NAACL*, 2010.

[5] C. Gómez-Rodríguez and G. Satta. An optimal-time binarization algorithm for linear context-free rewriting systems with fan-out two. In *Proceedings of ACL-IJCNLP*, 2009.

[6] D. Pollard. *Convergence of Stochastic Processes*. New York: Springer-Verlag, 1984.