The DOP estimation method is biased and inconsistent

Mark Johnson *
Brown University

A "Data-Oriented Parsing" or DOP *model for statistical parsing associates fragments of linguistic representations with numerical weights, where these weights are estimated by normalizing the empirical frequency of each fragment in a training corpus (see Bod (1998) and references cited therein). This note observes that this estimation method is* biased *and* inconsistent*; i.e., that the estimated distribution does not in general converge on the true distribution as the size of the training corpus increases.*

1 Introduction

The "Data-Oriented Parsing" or DOP approach to statistical natural language analysis has attracted considerable attention recently and has been used to produce statistical language models based on various kinds of linguistic representation, as described in Bod (1998). These models are based on the intuition that statistical generalizations about natural languages should be stated in terms of "chunks" or "fragments" of linguistic representations. Linguistic representations are produced by combining these fragments, but unlike stochastic models such as PCFGs, a single linguistic representation may be generated by several different combinations of fragments. These fragments may be large, permitting DOP models to describe non-local dependencies. Usually the fragments used in a DOP model are themselves obtained from a training corpus of linguistic representations. For example, in DOP1 or Tree-DOP the fragments are typically all the connected multi-node trees that appear as subgraphs of any tree in the training corpus.

This note shows that the estimation procedure standardly used to set the parameters or fragment weights of a DOP model (see e.g., Bod (1998)) is biased and inconsistent. This means that as sample size increases the corresponding sequence of probability distributions estimated by this procedure does not converge to the true distribution that generated the training data. Consistency is usually regarded as the minimal requirement any estimation method must satisfy (Breiman, 1973; Shao, 1999), and the inconsistency of the standard DOP estimation method suggests it may be worth looking for other estimation methods. Note that while the bulk of DOP research uses the estimation procedure studied here, recently there has been research that has used other estimators for DOP models (Bonnema, Buying, and Scha, 1999; Bod, 2000) and it would be interesting to investigate the statistical properties of these estimators as well.

2 DOP1 models

For simplicity this note focuses on DOP1 or Tree-DOP models, in which linguistic representations are phrase-structure trees, but the results carry over to more complex models

* Cognitive and Linguistic Sciences, Providence, RI 02912. I would like to thank Rens Bod, Michael Collins, Eugene Charniak, David MacAllester and the anonymous reviewers for their excellent advice.

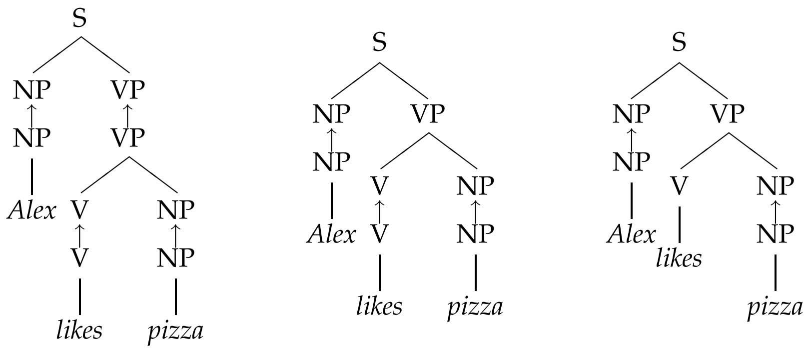

Figure 1
Depictions of three different derivations of the same tree representation of *Alex likes pizza*, with arrows indicating the sites of tree fragment substitutions.

which use attribute-value feature structure representations such as LFG-DOP. The fragments used in DOP1 are multi-node trees whose leaves may be labelled with nonterminals as well as terminals. A derivation starts with a fragment whose root is labelled with the start symbol, and proceeds by substituting a fragment for the leftmost nonterminal leaf under the constraint that the fragment's root node and the leaf node have the same label. The derivation terminates when there are no nonterminal leaves. Figure 1 depicts three different derivations which yield the same tree. The fragments used in these derivations could have been obtained from a training corpus of trees which contains trees for examples such as *Sasha likes motorcycles*, *Alex eats pizza*, etc.

In a DOP model each fragment is associated with a real-valued weight, and the weight of a derivation is the product of the weights of the tree fragments involved. The weight of a representation is the sum of the weights of its derivations, and a probability distribution over linguistic representations is obtained by normalizing the representations' weights.[1] Given a combinatory operation and a fixed set of fragments, a DOP model is a parametric model where the fragment weights are the parameters.

In DOP1 and DOP models based on it the weight associated with a fragment is estimated as follows (Bod, 1998). For each tree fragment f let $n(f)$ be the number of times it appears in the training corpus, and let F be the set of all tree fragments with the same root as f. Then the weight $w(f)$ associated with f is:

$$w(f) \;=\; \frac{n(f)}{\sum_{f' \in F} n(f')}.$$

This relative-frequency estimation method has the advantage of simplicity, but as shown in the following sections, it is biased and inconsistent.

3 Bias and Inconsistency

Bias and inconsistency are usually defined for parametric estimation procedures in terms that are not quite appropriate for evaluating the DOP estimation procedure, but their standard definitions (see Shao (1999) for a textbook exposition) will serve as the basis for the definitions adopted below. Let Θ be a vector space of real-valued parameters,

1 In DOP1 and similiar models it is not necessary to normalize the representations' weights if the fragments' weights are themselves appropriately normalized.

so that P_θ, $\theta \in \Theta$ is a probability distribution. In the DOP1 case, Θ would be the space of all possible weight assignments to fragments. An *estimator* ϕ is a function from a vector x of n samples to a parameter value $\phi(x) \in \Theta$, and an *estimation procedure* specifies an estimator ϕ_n for each sample size n.

Let X be a vector of n independent random variables distributed according to $\mathrm{P}_{\theta^\star}$ for some $\theta^\star \in \Theta$. Then $\phi(X)$ is also a random variable, ranging over parameter vectors Θ, with an expected value $\mathrm{E}_{\theta^\star}(\phi(X))$. The *bias* of the estimator ϕ at $\theta^\star$ is the difference $\mathrm{E}_{\theta^\star}(\phi(X)) - \theta^\star$ between its expected value and the "true" parameter value $\theta^\star$ that determines the distribution X. A biased estimator is one with nonzero bias for some value of $\theta^\star$.

A *loss function* $\mathcal{L}$ is a function from pairs of parameter vectors to the non-negative reals. Given a sample x drawn from the distribution $\theta^\star$, $\mathcal{L}(\theta^\star, \phi(x))$ measures the "cost" or the "loss" incurred by the error in the estimate $\phi(x)$ of $\theta^\star$. For example, a standard loss function is the Euclidean distance metric $\mathcal{L}(\theta^\star, \phi(x)) = ||\phi(X) - \theta^\star||^2$ (note that the results below do not depend on this choice of loss function). The *risk* of an estimator ϕ at $\theta^\star$ is its expected loss $\mathrm{E}_{\theta^\star}(\mathcal{L}(\theta^\star, \phi(X))$. An estimation procedure is *consistent* if and only if the limit of the risk of ϕ_n is 0 as $n \to \infty$ for all $\theta^\star$. (There are various different notions of consistency depending on how convergence is defined; however, the DOP1 estimator is not consistent with respect to any of the standard definitions of consistency).

Strictly speaking, the standard definitions of bias and loss function are not applicable to DOP estimation because there can be two distinct parameter vectors θ_1, θ_2 for which $\mathrm{P}_{\theta_1} = \mathrm{P}_{\theta_2}$ even though $\theta_1 \neq \theta_2$ (such a case is presented in the next section). Thus it is more natural to define bias and loss in terms of the probability distributions that the parameters specify, rather than in terms of the parameters themselves. In this paper, an estimator is unbiased iff $\mathrm{P}_{\mathrm{E}_{\theta^\star}(\phi(X))} = \mathrm{P}_{\theta^\star}$ for all $\theta^\star$, i.e., its expected parameter estimate specifies the same distribution as the true parameters. Similarly, the loss function is mean squared difference between the "true" and estimated distributions, i.e., if Ω is the event space (in DOP1, the space of all phrase-structure trees) then:

$$\mathcal{L}(\theta^\star, \phi(x)) = \sum_{\omega \in \Omega} \mathrm{P}_{\theta^\star}(\omega)(\mathrm{P}_{\theta^\star}(\omega) - \mathrm{P}_{\phi(x)}(\omega))^2.$$

As before, the risk of an estimator is its expected loss, and an estimation procedure is consistent iff the limit of the expected loss is 0 as $n \to \infty$.

4 A DOP1 example

This section presents a simple DOP1 model which only generates two trees with probability p and $1 - p$ respectively. The DOP relative frequency estimator is applied to a random sample of size n drawn from this population to estimate the tree weight parameters for the model. The bias and inconsistency of the estimator follows from the fact that these estimated parameters generate the trees with probabilities different to p and $1 - p$. The trees used and their DOP1 fragments are shown in Figure 2.

Suppose the "true" weights for the fragments $f_1, \ldots, f_7$ are zero except for the following fragments:

$$\begin{aligned} w^\star(f_4) &= p \\ w^\star(f_6) &= 1 - p \\ w^\star(f_7) &= 1 \end{aligned}$$

Then $\mathrm{P}_{w^\star}(t_1) = p$ and $\mathrm{P}_{w^\star}(t_2) = 1 - p$. (Note that exactly the same tree distribution could be obtained by setting $w^\star(f_1) = p$ and $w^\star(f_5) = 1 - p$ and all other weights to

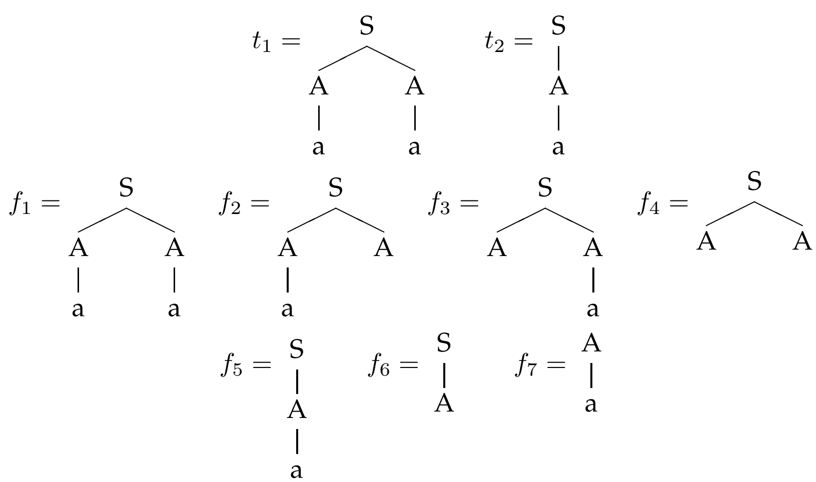

Figure 2
The trees t_1, t_2 and their associated fragments $f_1, \ldots, f_7$ in the DOP1 model.

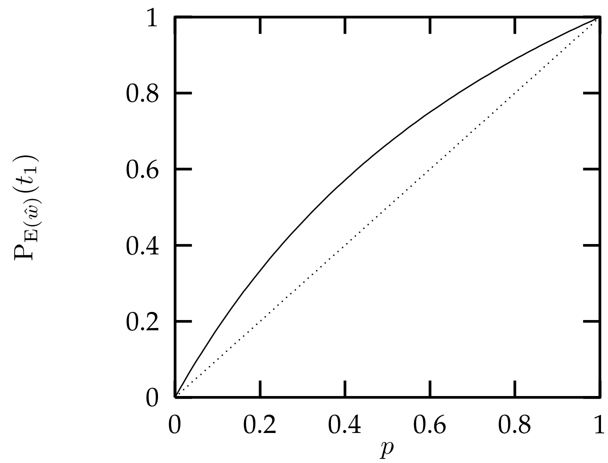

Figure 3
The value of $P_{E(\hat{w})}(t_1)$ as a function of $P_{w^\star}(t_1) = p$. The identity function p is also plotted for comparison.

zero; thus the tree weights are not identifiable). Then in a sample of size n drawn from the distribution $P_{w^\star}$ the expected number of occurences of tree t_1 is np and the expected number of occurences of tree t_2 is $n(1-p)$. Thus the expected number of occurences of the fragments in a sample of size n is:

$$\begin{aligned} E(n(f_i)) &= np && \text{for } i = 1, \ldots, 4. \\ E(n(f_i)) &= n(1-p) && \text{for } i = 5, 6. \\ E(n(f_7)) &= n + np. \end{aligned}$$

Thus after normalizing, the expected estimated weights for the fragments using the DOP estimator are:

$$\begin{aligned} E(\hat{w}(f_i)) &= \frac{p}{2+2p} && \text{for } i = 1, \ldots, 4. \\ E(\hat{w}(f_i)) &= \frac{1-p}{2+2p} && \text{for } i = 5, 6. \\ E(\hat{w}(f_7)) &= 1 \end{aligned}$$

Further calculation shows that:

$$\begin{aligned} P_{E(\hat{w})}(t_1) &= \frac{2p}{1+p} \\ P_{E(\hat{w})}(t_2) &= \frac{1-p}{1+p} \end{aligned}$$

Figure 3 shows how $P_{E(\hat{w})}(t_1)$ varies as a function of $P_{w^\star}(t_1) = p$. The difference $P_{E(\hat{w})}(t_1) - p$ reaches a maximum value of approximately 0.17 at $p = \sqrt{2} - 1$. Thus except for $p = 0$ and $p = 1$, $P_{E(\hat{w})} \neq P_{w^\star}$, i.e., the DOP1 estimator is biased.

Further, note that the estimated distribution $P_{E(\hat{w})}$ does not approach $P_{w^\star}$ as the sample size increases, so the expected loss does not converge to 0 as the sample size n increases. Thus the DOP1 estimator is also inconsistent.

5 Conclusion

The previous section showed that the relative frequency estimation procedure used in DOP1 and related DOP models is biased and inconsistent. Bias is not necessarily a defect in an estimator, and Geman, Bienenstock, and Doursat (1992) argue that it may be desirable to trade variance for bias. However, inconsistency is usually viewed as a fatal flaw of an estimator. Never the less, excellent empirical results have been claimed for the DOP1 model, so perhaps there are some circumstances in which inconsistent estimators perform well. Undoubtedly there are other estimation procedures for DOP models which are unbiased and consistent. For example, maximum likelihood estimators are unbiased and consistent across a wide class of models, including, it would seem, all reasonable DOP models (Shao, 1999). Bod (2000) describes a procedure for maximum likelihood estimation of DOP models based on an Expectation Maximumization-like algorithm. In addition, Rens Bod (p.c.) points out that because the set of fragments in a DOP1 model includes all of the trees in the training corpus, the maximum likelihood estimator will assign the training corpus trees their empirical frequencies, and assign zero weight to all other trees. However, this seems to be an overlearning problem rather than a problem with maximum likelihood estimation per se, and standard methods, such as cross-validation or regularization, would seem in principle to be ways to avoid such overlearning. Obviously empirical investigation would be useful here.

References

Bod, Rens. 1998. *Beyond grammar: an experience-based theory of language*. CSLI Publications, Stanford, California.

Bod, Rens. 2000. Combining semantic and syntactic structure for language modelling. In *Proceedings of the Eighth International Conference on Spoken Language Processing (ICSLP 2000)*, Beijing.

Bonnema, Remko, Paul Buying, and Remko Scha. 1999. A new probability model for data oriented parsing. In Paul Dekker and Gwen Kerdiles, editors, *Proceedings of the 12th Amsterdam Colloquium*, Amsterdam.

Breiman, Leo. 1973. *Statistics with a view toward applications*. Houghton Mifflin, Boston.

Geman, Stuart, Elie Bienenstock, and René Doursat. 1992. Neural networks and the bias/variance dilemma. *Neural Computation*, 4:1–58.

Shao, Jun. 1999. *Mathematical Statistics*. Springer Verlag, New York.