Efficient Optimization of an MDL-Inspired Objective Function for Unsupervised Part-of-Speech Tagging

Ashish Vaswani[1] **Adam Pauls**[2] **David Chiang**[1]

[1]Information Sciences Institute
University of Southern California
4676 Admiralty Way, Suite 1001
Marina del Rey, CA 90292
{avaswani,chiang}@isi.edu

[2]Computer Science Division
University of California at Berkeley
Soda Hall
Berkeley, CA 94720
adpauls@eecs.berkeley.edu

Abstract

The Minimum Description Length (MDL) principle is a method for model selection that trades off between the explanation of the data by the model and the complexity of the model itself. Inspired by the MDL principle, we develop an objective function for generative models that captures the description of the data by the model (log-likelihood) and the description of the model (model size). We also develop a efficient general search algorithm based on the MAP-EM framework to optimize this function. Since recent work has shown that minimizing the model size in a Hidden Markov Model for part-of-speech (POS) tagging leads to higher accuracies, we test our approach by applying it to this problem. The search algorithm involves a simple change to EM and achieves high POS tagging accuracies on both English and Italian data sets.

1 Introduction

The Minimum Description Length (MDL) principle is a method for model selection that provides a generic solution to the overfitting problem (Barron et al., 1998). A formalization of Ockham's Razor, it says that the parameters are to be chosen that minimize the description length of the data given the model plus the description length of the model itself.

It has been successfully shown that minimizing the model size in a Hidden Markov Model (HMM) for part-of-speech (POS) tagging leads to higher accuracies than simply running the Expectation-Maximization (EM) algorithm (Dempster et al., 1977). Goldwater and Griffiths (2007) employ a Bayesian approach to POS tagging and use sparse Dirichlet priors to minimize model size. More recently, Ravi and Knight (2009) alternately minimize the model using an integer linear program and maximize likelihood using EM to achieve the highest accuracies on the task so far. However, in the latter approach, because there is no single objective function to optimize, it is not entirely clear how to generalize this technique to other problems. In this paper, inspired by the MDL principle, we develop an objective function for generative models that captures both the description of the data by the model (log-likelihood) and the description of the model (model size). By using a simple prior that encourages sparsity, we cast our problem as a search for the maximum *a posteriori* (MAP) hypothesis and present a variant of EM to approximately search for the minimum-description-length model. Applying our approach to the POS tagging problem, we obtain higher accuracies than both EM and Bayesian inference as reported by Goldwater and Griffiths (2007). On a Italian POS tagging task, we obtain even larger improvements. We find that our objective function correlates well with accuracy, suggesting that this technique might be useful for other problems.

2 MAP EM with Sparse Priors

2.1 Objective function

In the unsupervised POS tagging task, we are given a word sequence $\mathbf{w} = w_1, \ldots, w_N$ and want to find the best tagging $\mathbf{t} = t_1, \ldots, t_N$, where $t_i \in \mathcal{T}$, the tag vocabulary. We adopt the problem formulation of Merialdo (1994), in which we are given a dictionary of possible tags for each word type.

We define a bigram HMM

$$P(w, t \mid \theta) = \prod_{i=1}^{N} P(\mathbf{w}, \mathbf{t} \mid \theta) \cdot P(t_i \mid t_{i-1}) \qquad (1)$$

In maximum likelihood estimation, the goal is to

find parameter estimates

$$\hat{\theta} = \arg\max_{\theta} \log P(\mathbf{w} \mid \theta) \quad (2)$$

$$= \arg\max_{\theta} \log \sum_{\mathbf{t}} P(\mathbf{w}, \mathbf{t} \mid \theta) \quad (3)$$

The EM algorithm can be used to find a solution. However, we would like to maximize likelihood and minimize the size of the model simultaneously. We define the size of a model as the number of non-zero probabilities in its parameter vector. Let $\theta_1, \ldots, \theta_n$ be the components of θ. We would like to find

$$\hat{\theta} = \arg\min_{\theta} \left(-\log P(\mathbf{w} \mid \theta) + \alpha \|\theta\|_0\right) \quad (4)$$

where $\|\theta\|_0$, called the L0 norm of θ, simply counts the number of non-zero parameters in θ. The hyperparameter α controls the tradeoff between likelihood maximization and model minimization. Note the similarity of this objective function with MDL's, where α would be the space (measured in nats) needed to describe one parameter of the model.

Unfortunately, minimization of the L0 norm is known to be NP-hard (Hyder and Mahata, 2009). It is not smooth, making it unamenable to gradient-based optimization algorithms. Therefore, we use a smoothed approximation,

$$\|\theta\|_0 \approx \sum_i \left(1 - e^{\frac{-\theta_i}{\beta}}\right) \quad (5)$$

where $0 < \beta \leq 1$ (Mohimani et al., 2007). For smaller values of β, this closely approximates the desired function (Figure 1). Inverting signs and ignoring constant terms, our objective function is now:

$$\hat{\theta} = \arg\max_{\theta} \left(\log P(\mathbf{w} \mid \theta) + \alpha \sum_i e^{\frac{-\theta_i}{\beta}}\right) \quad (6)$$

We can think of the approximate model size as a kind of prior:

$$P(\theta) = \frac{\exp \alpha \sum_i e^{\frac{-\theta_i}{\beta}}}{Z} \quad (7)$$

$$\log P(\theta) = \alpha \cdot \sum_i e^{\frac{-\theta_i}{\beta}} - \log Z \quad (8)$$

where $Z = \int_{d\theta} \exp \alpha \sum_i e^{\frac{-\theta_i}{\beta}}$ is a normalization constant. Then our goal is to find the maximum

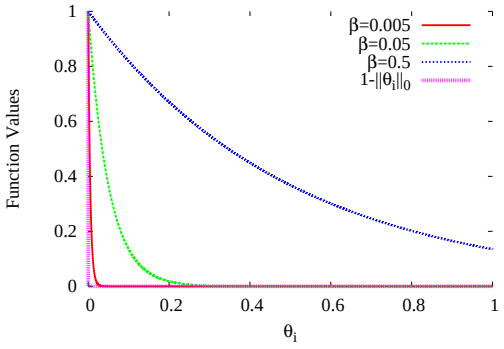

Figure 1: Ideal model-size term and its approximations.

a posteriori parameter estimate, which we find using MAP-EM (Bishop, 2006):

$$\hat{\theta} = \arg\max_{\theta} \log P(\mathbf{w}, \theta) \quad (9)$$

$$= \arg\max_{\theta} \left(\log P(\mathbf{w} \mid \theta) + \log P(\theta)\right) \quad (10)$$

Substituting (8) into (10) and ignoring the constant term $\log Z$, we get our objective function (6) again.

We can exercise finer control over the sparsity of the tag-bigram and channel probability distributions by using a different α for each:

$$\arg\max_{\theta} \left(\log P(\mathbf{w} \mid \theta) + \alpha_c \sum_{w,t} e^{\frac{-P(w|t)}{\beta}} + \alpha_t \sum_{t,t'} e^{\frac{-P(t'|t)}{\beta}}\right) \quad (11)$$

In our experiments, we set $\alpha_c = 0$ since previous work has shown that minimizing the number of tag n-gram parameters is more important (Ravi and Knight, 2009; Goldwater and Griffiths, 2007).

A common method for preferring smaller models is minimizing the L1 norm, $\sum_i |\theta_i|$. However, for a model which is a product of multinomial distributions, the L1 norm is a constant.

$$\sum_i |\theta_i| = \sum_i \theta_i$$

$$= \sum_t \left(\sum_w P(w \mid t) + \sum_{t'} P(t' \mid t)\right)$$

$$= 2|\mathcal{T}|$$

Therefore, we cannot use the L1 norm as part of the size term as the result will be the same as the EM algorithm.

2.2 Parameter optimization

To optimize (11), we use MAP EM, which is an iterative search procedure. The E step is the same as in standard EM, which is to calculate $P(\mathbf{t} \mid \mathbf{w}, \theta^t)$, where the θ^t are the parameters in the current iteration t. The M step in iteration $(t+1)$ looks like

$$\theta^{t+1} = \arg\max_{\theta} \left(E_{P(\mathbf{t}|\mathbf{w},\theta^t)} \left[\log P(\mathbf{w}, \mathbf{t} \mid \theta) \right] + \alpha_t \sum_{t,t'} e^{\frac{-P(t'|t)}{\beta}} \right) \tag{12}$$

Let $C(t, w; \mathbf{t}, \mathbf{w})$ count the number of times the word w is tagged as t in $\mathbf{t}$, and $C(t, t'; \mathbf{t})$ the number of times the tag bigram (t, t') appears in $\mathbf{t}$. We can rewrite the M step as

$$\theta^{t+1} = \arg\max_{\theta} \left(\sum_t \sum_w E[C(t,w)] \log P(w \mid t) + \sum_t \sum_{t'} \left(E[C(t,t')] \log P(t' \mid t) + \alpha_t e^{\frac{-P(t'|t)}{\beta}} \right) \right) \tag{13}$$

subject to the constraints $\sum_w P(w \mid t) = 1$ and $\sum_{t'} P(t' \mid t) = 1$. Note that we can optimize each term of both summations over t separately. For each t, the term

$$\sum_w E[C(t,w)] \log P(w \mid t) \tag{14}$$

is easily optimized as in EM: just let $P(w \mid t) \propto E[C(t,w)]$. But the term

$$\sum_{t'} \left(E[C(t,t')] \log P(t' \mid t) + \alpha_t e^{\frac{-P(t'|t)}{\beta}} \right) \tag{15}$$

is trickier. This is a non-convex optimization problem for which we invoke a publicly available constrained optimization tool, ALGENCAN (Andreani et al., 2007). To carry out its optimization, ALGENCAN requires computation of the following in every iteration:

- **Objective function**, defined in equation (15). This is calculated in polynomial time using dynamic programming.
- **Constraints**: $g_t = \sum_{t'} P(t' \mid t) - 1 = 0$ for each tag $t \in \mathcal{T}$. Also, we constrain $P(t' \mid t)$ to the interval $[\epsilon, 1]$.[1]
- **Gradient of objective function**:

$$\frac{\partial F}{\partial P(t' \mid t)} = \frac{E[C(t,t')]}{P(t' \mid t)} - \frac{\alpha_t}{\beta} e^{\frac{-P(t'|t)}{\beta}} \tag{16}$$

- **Gradient of equality constraints**:

$$\frac{\partial g_t}{\partial P(t'' \mid t')} = \begin{cases} 1 & \text{if } t = t' \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

- **Hessian of objective function**, which is not required but greatly speeds up the optimization:

$$\frac{\partial^2 F}{\partial P(t' \mid t) \partial P(t' \mid t)} = -\frac{E[C(t,t')]}{P(t' \mid t)^2} + \alpha_t \frac{e^{\frac{-P(t'|t)}{\beta}}}{\beta^2} \tag{18}$$

  The other second-order partial derivatives are all zero, as are those of the equality constraints.

We perform this optimization for each instance of (15). These optimizations could easily be performed in parallel for greater scalability.

3 Experiments

We carried out POS tagging experiments on English and Italian.

3.1 English POS tagging

To set the hyperparameters α_t and β, we prepared three held-out sets H_1, H_2, and H_3 from the Penn Treebank. Each H_i comprised about 24,000 words annotated with POS tags. We ran MAP-EM for 100 iterations, with uniform probability initialization, for a suite of hyperparameters and averaged their tagging accuracies over the three held-out sets. The results are presented in Table 2. We then picked the hyperparameter setting with the highest average accuracy. These were $\alpha_t = 80, \beta = 0.05$. We then ran MAP-EM again on the test data with these hyperparameters and achieved a tagging accuracy of 87.4% (see Table 1). This is higher than the 85.2% that Goldwater and Griffiths (2007) obtain using Bayesian methods for inferring both POS tags and hyperparameters. It is much higher than the 82.4% that standard EM achieves on the test set when run for 100 iterations.

Using $\alpha_t = 80, \beta = 0.05$, we ran multiple random restarts on the test set (see Figure 2). We find that the objective function correlates well with accuracy, and picking the point with the highest objective function value achieves 87.1% accuracy.

[1] We must have $\epsilon > 0$ because of the $\log P(t' \mid t)$ term in equation (15). It seems reasonable to set $\epsilon \ll \frac{1}{N}$; in our experiments, we set $\epsilon = 10^{-7}$.

α_t	β								
	0.75	0.5	0.25	0.075	0.05	0.025	0.0075	0.005	0.0025
10	82.81	82.78	83.10	83.50	83.76	83.70	84.07	83.95	83.75
20	82.78	82.82	83.26	83.60	83.89	84.88	83.74	84.12	83.46
30	82.78	83.06	83.26	83.29	84.50	84.82	84.54	83.93	83.47
40	82.81	83.13	83.50	83.98	84.23	85.31	85.05	83.84	83.46
50	82.84	83.24	83.15	84.08	82.53	84.90	84.73	83.69	82.70
60	83.05	83.14	83.26	83.30	82.08	85.23	85.06	83.26	82.96
70	83.09	83.10	82.97	82.37	83.30	86.32	83.98	83.55	82.97
80	83.13	83.15	82.71	83.00	**86.47**	86.24	83.94	83.26	82.93
90	83.20	83.18	82.53	84.20	86.32	84.87	83.49	83.62	82.03
100	83.19	83.51	82.84	84.60	86.13	85.94	83.26	83.67	82.06
110	83.18	83.53	83.29	84.40	86.19	85.18	80.76	83.32	82.05
120	83.08	83.65	83.71	84.11	86.03	85.39	80.66	82.98	82.20
130	83.10	83.19	83.52	84.02	85.79	85.65	80.08	82.04	81.76
140	83.11	83.17	83.34	85.26	85.86	85.84	79.09	82.51	81.64
150	83.14	83.20	83.40	85.33	85.54	85.18	78.90	81.99	81.88

Table 2: Average accuracies over three held-out sets for English.

system	accuracy (%)
Standard EM	82.4
+ random restarts	84.5
(Goldwater and Griffiths, 2007)	85.2
our approach	87.4
+ random restarts	87.1

Table 1: MAP-EM with a L0 norm achieves higher tagging accuracy on English than (2007) and much higher than standard EM.

system	zero parameters	bigram types
maximum possible	1389	–
EM, 100 iterations	444	924
MAP-EM, 100 iterations	695	648

Table 3: MAP-EM with a smoothed L0 norm yields much smaller models than standard EM.

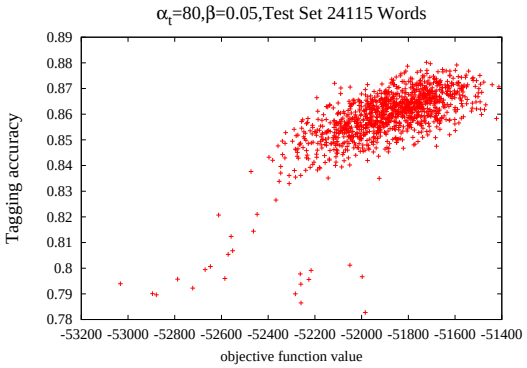

Figure 2: Tagging accuracy vs. objective function for 1152 random restarts of MAP-EM with smoothed L0 norm.

We also carried out the same experiment with standard EM (Figure 3), where picking the point with the highest corpus probability achieves 84.5% accuracy.

We also measured the minimization effect of the sparse prior against that of standard EM. Since our method lower-bounds all the parameters by ϵ, we consider a parameter θ_i as a zero if $\theta_i \leq \epsilon$. We also measured the number of unique tag bigram types in the Viterbi tagging of the word sequence. Table 3 shows that our method produces much smaller models than EM, and produces Viterbi taggings with many fewer tag-bigram types.

3.2 Italian POS tagging

We also carried out POS tagging experiments on an Italian corpus from the Italian Turin University Treebank (Bos et al., 2009). This test set comprises 21,878 words annotated with POS tags and a dictionary for each word type. Since this is all the available data, we could not tune the hyperparameters on a held-out data set. Using the hyperparameters tuned on English ($\alpha_t = 80, \beta = 0.05$), we obtained 89.7% tagging accuracy (see Table 4), which was a large improvement over 81.2% that standard EM achieved. When we tuned the hyperparameters on the test set, the best setting ($\alpha_t = 120, \beta = 0.05$ gave an accuracy of 90.28%.

4 Conclusion

A variety of other techniques in the literature have been applied to this unsupervised POS tagging task. Smith and Eisner (2005) use conditional random fields with contrastive estimation to achieve

α_t	β								
	0.75	0.5	0.25	0.075	0.05	0.025	0.0075	0.005	0.0025
10	81.62	81.67	81.63	82.47	82.70	84.64	84.82	84.96	84.90
20	81.67	81.63	81.76	82.75	84.28	84.79	85.85	88.49	85.30
30	81.66	81.63	82.29	83.43	85.08	88.10	86.16	88.70	88.34
40	81.64	81.79	82.30	85.00	86.10	88.86	89.28	88.76	88.80
50	81.71	81.71	78.86	85.93	86.16	88.98	88.98	89.11	88.01
60	81.65	82.22	78.95	86.11	87.16	89.35	88.97	88.59	88.00
70	81.69	82.25	79.55	86.32	89.79	89.37	88.91	85.63	87.89
80	81.74	82.23	80.78	86.34	89.70	89.58	88.87	88.32	88.56
90	81.70	81.85	81.00	86.35	90.08	89.40	89.09	88.09	88.50
100	81.70	82.27	82.24	86.53	90.07	88.93	89.09	88.30	88.72
110	82.19	82.49	82.22	86.77	90.12	89.22	88.87	88.48	87.91
120	82.23	78.60	82.76	86.77	**90.28**	89.05	88.75	88.83	88.53
130	82.20	78.60	83.33	87.48	90.12	89.15	89.30	87.81	88.66
140	82.24	78.64	83.34	87.48	90.12	89.01	88.87	88.99	88.85
150	82.28	78.69	83.32	87.75	90.25	87.81	88.50	89.07	88.41

Table 4: Accuracies on test set for Italian.

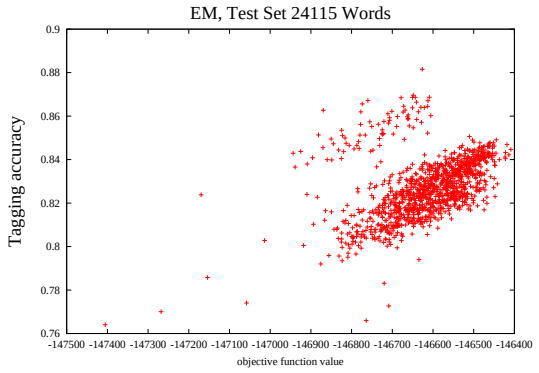

Figure 3: Tagging accuracy vs. likelihood for 1152 random restarts of standard EM.

88.6% accuracy. Goldberg et al. (2008) provide a linguistically-informed starting point for EM to achieve 91.4% accuracy. More recently, Chiang et al. (2010) use GIbbs sampling for Bayesian inference along with automatic run selection and achieve 90.7%.

In this paper, our goal has been to investigate whether EM can be extended in a generic way to use an MDL-like objective function that simultaneously maximizes likelihood and minimizes model size. We have presented an efficient search procedure that optimizes this function for generative models and demonstrated that maximizing this function leads to improvement in tagging accuracy over standard EM. We infer the hyperparameters of our model using held out data and achieve better accuracies than (Goldwater and Griffiths, 2007). We have also shown that the objective function correlates well with tagging accuracy supporting the MDL principle. Our approach performs quite well on POS tagging for both English and Italian. We believe that, like EM, our method can benefit from more unlabeled data, and there is reason to hope that the success of these experiments will carry over to other tasks as well.

Acknowledgements

We would like to thank Sujith Ravi, Kevin Knight and Steve DeNeefe for their valuable input, and Jason Baldridge for directing us to the Italian POS data. This research was supported in part by DARPA contract HR0011-06-C-0022 under subcontract to BBN Technologies and DARPA contract HR0011-09-1-0028.

References

R. Andreani, E. G. Birgin, J. M. Martnez, and M. L. Schuverdt. 2007. On Augmented Lagrangian methods with general lower-level constraints. *SIAM Journal on Optimization*, 18:1286–1309.

A. Barron, J. Rissanen, and B. Yu. 1998. The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, 44(6):2743–2760.

C. Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer.

J. Bos, C. Bosco, and A. Mazzei. 2009. Converting a dependency treebank to a categorical grammar treebank for italian. In *Eighth International Workshop on Treebanks and Linguistic Theories (TLT8)*.

D. Chiang, J. Graehl, K. Knight, A. Pauls, and S. Ravi. 2010. Bayesian inference for Finite-State transducers. In *Proceedings of the North American Association of Computational Linguistics*.

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Computational Linguistics*, 39(4):1–38.

Y. Goldberg, M. Adler, and M. Elhadad. 2008. EM can find pretty good HMM POS-taggers (when given a good start). In *Proceedings of the ACL.*

S. Goldwater and T. L. Griffiths. 2007. A fully Bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of the ACL.*

M. Hyder and K. Mahata. 2009. An approximate L0 norm minimization algorithm for compressed sensing. In *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*.

B. Merialdo. 1994. Tagging English text with a probabilistic model. *Computational Linguistics*, 20(2):155–171.

H. Mohimani, M. Babaie-Zadeh, and C. Jutten. 2007. Fast sparse representation based on smoothed L0 norm. In *Proceedings of the 7th International Conference on Independent Component Analysis and Signal Separation (ICA2007)*.

S. Ravi and K. Knight. 2009. Minimized models for unsupervised part-of-speech tagging. In *Proceedings of ACL-IJCNLP*.

N. Smith. and J. Eisner. 2005. Contrastive estimation: Training log-linear models on unlabeled data. In *Proceedings of the ACL.*