Posterior vs. Parameter Sparsity in Latent Variable Models
Supplementary Material

João V. Graça
L^{2}F INESC-ID
Lisboa, Portugal

Kuzman Ganchev **Ben Taskar**
University of Pennsylvania
Philadelphia, PA, USA

Fernando Pereira
Google Research
Mountain View, CA, USA

1 Dual Derivation

1.1 Derivation of the ℓ_1/ℓ_∞ dual program

We want to optimize the objective:

$$\begin{aligned} \min_{q, c_{wt}} \quad & \mathrm{KL}(q||p) + \sigma \sum_{wt} c_{wt} \\ \text{s. t.} \quad & \mathbf{E}_q[f_{wti}] \leq c_{wt} \\ & 0 \leq c_{wt} \end{aligned} \tag{1}$$

The Lagrangian becomes:

$$\boxed{L(q, c, \alpha, \lambda) = \mathrm{KL}(q||p) + \sigma \sum_{wt} c_{wt} + \sum_{wti} \lambda_{wti}(\mathbf{E}_q[f_{wti}] - c_{wt}) - \alpha \cdot c} \tag{2}$$

where we are maximizing with respect to $\lambda \geq 0$ and $\alpha \geq 0$. Taking the derivative with respect to $q(z)$ we have:

$$\frac{\partial L(q, c, \alpha, \lambda)}{\partial q(z)} = \log q(z) + 1 - \log p(z) - f(z) \cdot \lambda \tag{3}$$

Setting this to zero and ensuring q normalizes we get:

$$q(z) = \frac{p(z) \exp(-f(z) \cdot \lambda)}{Z_\lambda} \tag{4}$$

Taking the derivative with respect to c_{wt} we have:

$$\frac{\partial L(q, c, \alpha, \lambda)}{\partial c_{wt}} = \sigma - \sum_i \lambda_{wti} - \alpha_{wt} \tag{5}$$

setting this to zero gives us $\alpha_{wt} = \sigma - \sum_i \lambda_{wti}$. Knowing that $\alpha_{wt} \geq 0$ we will have to introduce the constraint $\sigma \geq \sum_i \lambda_{wti}$. Substituting into the KL term we have: yields:

$$\begin{aligned} \mathrm{KL}(q||p) &= \sum_z \frac{p(z) \exp(-f(z) \cdot \lambda)}{Z_\lambda} \log \frac{p(z) \exp(-f(z) \cdot \lambda)}{Z_\lambda p(z)} \\ &= -\log(Z_\lambda) - \mathbf{E}_q[\lambda \cdot f] \end{aligned} \tag{6}$$

The second part of this will cancel with $\sum_{wti} \lambda_{wti} \mathbf{E}_q[f_{wti}]$ leaving us with:

$$\begin{aligned} L(q, c, \alpha, \lambda) &= -\log(Z_\lambda) + \sigma \sum_{wt} c_{wt} + \sum_{wti} \lambda_{wti}(-c_{wt}) - \alpha \cdot c \\ &= -\log(Z_\lambda) + \sigma \sum_{wt} c_{wt} - \sum_{wti} \lambda_{wti} c_{wt} - \sum_{wt} c_{wt}(\sigma - \sum_i \lambda_{wti}) \\ &= -\log(Z_\lambda) \end{aligned} \tag{7}$$

So our objective becomes very simple:

$$\boxed{\begin{array}{ll} \max\limits_{\lambda \geq 0} & -\log(Z_\lambda) \\ \text{s.t.} & \sum\limits_i \lambda_{wti} \leq \sigma \end{array}} \tag{8}$$

This can be done via projected gradient. The projection can be done in a way described in [2]. The basic idea is to use the fact that the solution will be of the form $\max(0, \lambda + \theta)$ for some θ.

2 Corpora

This section presents extra information about the four different corpora used in the main paper: the Wall Street Journal portion of the Penn treebank [3] using all 45 tags (PTB45) and with a tag set reduced to 17 tags [5] (PTB17); the Bosque subset of the Portuguese Floresta Sinta(c)tica Treebank [1][1] used for the ConLL X shared task on dependency parsing (PT-CoNLL)[2]; and the Bulgarian BulTreeBank [4] (BulTree) with the 12 coarse tags. All words that occurred only once were replaced by the token "unk". To measure model sparsity, we compute the average ℓ_1/ℓ_∞ norm over words occurring more than 10 times; the label 'L1LMax' denotes this measure in figures. Table 1 gives statistics for each corpus as well as the sparsity for a first-order HMM trained on the labeled data.

	Types	Tokens	Unk	Tags	Sup. ℓ_1/ℓ_∞	EM ℓ_1/ℓ_∞
PT-Conll	11293	206678	8.5%	22	1.14	4.57
BulTree	12177	174160	10%	12	1.04	3.51
PTB17	23768	950028	2%	17	1.23	3.97
PTB45	23768	950028	2%	45	1.37	5.43

Table 1: Corpus statistics. All words with only one occurrence where replaced by the 'unk' token. The third column shows the percentage of tokens replaced. Sup. ℓ_1/ℓ_∞ is the value of the sparsity measure for a fully supervised HMM trained on all available data and EM ℓ_1/ℓ_∞ is the value of the sparsity measure for a fully unsupervised HMM trained using standard EM on all available data.

2.1 Tag Sets

The tag sets along with their meaning is described in Tables 2 (Portuguese), 3 and 4 (English) and 5 (Bulgarian). Table 3 has the full Penn Treebank tag set, while Table 4 has shows the coarse version.

Name	Category	Name	Category
n	noun	prop	proper noun
adj	adjective	v-fin	finite verb
v-inf	infinitive	v-pcp	participle
v-ger	gerund	art	article
pron-det	determiner pronoun	pron-indp	independent pronoun
pron-pers	personal pronoun	adv	adverb
num	numeral	prp	preposition
in	interjection	conj-s	subordinating conjunction
conj-c	coordinating conjunction	ec	prefixes
pp	prepositional phrase	?	??
vp	??		

Table 2: Portuguese Conll Tag Set. There are 22 tags, from which only 20 are described in the guidelines[4]. We could not find a description for tags "?" and "vp". These tags only occur 1 and 2 times in the corpus, respectively.

[1] `http://www.linguateca.pt/Floresta/`
[2] `http://nextens.uvt.nl/~conll/`

Name	Category	Name	Category
$	dollar	`	opening quotation mark
”	closing quotation mark	(	opening parenthesis
)	closing parenthesis	,	comma
–	dash	.	sentence terminator
:	colon or ellipsis	CC	conjunction, coordinating
CD	numeral, cardinal	DT	determiner
EX	existential there	FW	foreign word
IN	preposition or conjunction, subordinating	JJ	adjective or numeral, ordinal
JJR	adjective, comparative	JJS	adjective, superlative
LS	list item marker	MD	modal auxiliary
NNPS	noun, proper, plural	NNS	noun, common, plural
NN	noun, common, singular or mass	NNP	noun, proper, singular
PDT	pre-determiner	POS	genitive marker
PRP	pronoun, personal	PRP$	pronoun, possessive
RB	adverb	RBR	adverb, comparative
RBS	adverb, superlative	RP	particle
SYM	symbol	TO	"to" as preposition or infinitive marker
UH	interjection	VB	verb, base form
VBD	verb, past tense	VBG	verb, present participle or gerund
VBN	verb, past participle	VBP	verb, present tense, not 3rd person singular
VBZ	verb, present tense, 3rd person singular	WDT	WH-determiner
WP	WH-pronoun	WP$	WH-pronoun, possessive
WRB	Wh-adverb		

Table 3: Penn tree bank full tag set [5]. There are a total of 45 tags.

Coarse Tag	Treebank tag	Coarse Tag	Treebank tag
ADJ	CD JJ JJR JJS PRP$	ADV	RB RBR RBS
CONJ	CC	DET	DT PDT
ENDPUNC	.	INPUNC	, : LS SYM UH
LPUNC	“ -LRB	N	EX FW NN NNP NNPS NNS PRP
POS	POS	PREP	IN
PRT	RP	RPUNC	” -RRB-
TO	TO	W	WDT WP$ WP WRB
V	MD VBD VBP VB VBZ	VBN	VBN
VBG	VBG		

Table 4: Penn tree bank reduced tag set. There are 17 tags total.

2.2 Function Words

Table 6 contains a list of the tags we considered closed class in each of the languages.

3 Experiments

This section contains extra experimental results that did not fit in the paper.

Table 7 contains the complete results set for the different corpus/models. For PTB45 Sparse performs better than all other models using the 1-Many mapping, while VEM performs better as observed before. Note that, as described in the paper different values for the transition prior on VEM do not significantly change the results.

Figure 1 shows scatter plots of the performance of each method with respect to the ℓ_1/ℓ_∞ value they achieve. We see that all languages follow the pattern described in the main paper, both Sparse and VEM achieve similar sparsity values, but the accuracies are very different. Sparse always achieves a better performance according to 1-Many mapping, while VEM achieves a better performance on 1-1 mapping in half the cases.

Name	Category	Name	Category
N	noun	A	adjective
H	hybrid	P	pronoun
M	numeral	V	verb
D	adverb	C	conjunction
T	article	R	Preposition
I	Interjection	Punct	punctuation

Table 5: Bulgarian BulTree tag set. There are 12 tags total. This is a reduced tag set where only the first letter of each tag was used which describes its main syntactic category. This results in 11 syntactic tags and an extra tag for punctuation.

PT-Conll	BulTree	PTB17
art	C	CONJ
punc	M	DET
prp	P	ENDPUNC
conj-c	Punct	INPUNC
conj-s	R	LPUNC
ec	T	POS
pp		PREP
pron-det		PRT
pron-indp		RPUNC
pron-pers		TO
		W

Table 6: Function word classes that were provided as input in the weakly-supervised setting described in the paper.

Figure 2 compares the different cluster sizes obtained by different models for all corpora. We see that all corpus follow the same pattern as PT-Conll (described in the paper). VEM with a prior on state emission of 10^{-1} achieves a distribution much more similar to the gold labeling, and with a prior on state emission of 10^{-3} still achieves a distribution closer to EM. These values have a relationship with the accuracy metrics. VEM(10^{-1}) has a better 1-1 mapping but a worse 1-many mapping.

Table 8 shows the mutual information in bits between gold tag distribution and hidden state distribution, for all models and corpus. We see that Sparse always produces a higher values then the competing methods.

Figure 3 shows performance as a function of the number of training iterations. For readability, the candlesticks have been shifted so as not to overlap; evaluations are performed every 10 iterations. Because of the warmup phase, the first 30 iterations of EM are identical to those of Sparse. Soon after this warmup period Sparse starts to perform much better than both baselines.

3.1 Weakly-Supervised Learning

We now consider the case where some supervision has been given in the form of a partial dictionary. As we note in the introduction, by setting to zero the probability of disallowed emissions, this partial supervision ensures sparsity for some words. If the dictionary is small, our method might still achieve some improvement over EM by ensuring sparsity over word types not in the dictionary. As the size of the dictionary increases, we would expect the benefits from our method to decrease, since the dictionary already ensures sparsity. Figure 4 shows the performance of the three learning methods as we increase the size of the dictionary. In all cases the dictionary contains the most common word types, that is, an x-axis value of n corresponds to a dictionary containing all possible POS tags of the n most common word types. We see that EM converges with our method for large dictionaries, but small improvements can be seen even with a dictionary of 10k words. Table 9 shows numerical results for a dictionary of 100 word types.

Estimator	PT-Conll		BulTree		PTB17		PTB45	
	1-Many	1-1	1-Many	1-1	1-Many	1-1	1-Many	1-1
EM	64.0(1.2)	40.4(3)	59.4(2.2)	42.0(3.0)	67.5(1.3)	46.4(2.6)	63.1(1.0)	44.2(3.0)
VEM(10^{-1},10^{-1})	60.4(0.6)	**51.1(2.3)**	54.9(3.1)	46.4(3.0)	68.2(0.8)*	**52.8(3.5)**	54.6(1.7)	**46.0(2.4)***
VEM(10^{-1},10^{-4})	63.2(1.0)*	48.1(2.2)	56.1(2.8)	43.3(1.7)*	67.3(0.8)*	49.6(4.3)	59.0(0.8)	43.3(1.6)*
VEM(10^{-4}, 10^{-1})	60.4(0.6)	50.8(2.5)	54.8(3.1)	46.2(3.0)	68.3(0.9)*	52.8(3.5)	54.3(1.6)	45.8(2.6)*
VEM(10^{-4}, 10^{-4})	63.2(1)*	48.1(2.2)	56.2(2.8)	43.3(1.7)*	67.3(0.8)*	49.6(4.3)	59.0(0.7)	43.6(1.5)*
Sparse (10)	68.5(1.3)	43.3(2.2)	65.1(1.0)	48.0(3.3)	69.5(1.6)	50.0(3.5)	64.2(1.0)	44.3(2.8)*
Sparse (32)	**69.2(0.9)**	43.2(2.9)	**66.0(1.8)**	48.7(2.2)	**70.2(2.2)**	49.5(2.0)	**65.4(1.0)**	44.5(2.7)*
Sparse (100)	68.3(2.1)	44.5(2.4)	65.9(1.6)	**48.9(2.8)**	68.7(1.1)	47.8(1.5)*	64.7(1.2)	42.4(2.5)

Table 7: Average accuracy (standard deviation in parentheses) over 10 different runs (same seeds used for each model) for 200 iterations. 1-Many and 1-1 are the two hidden-state to POS mappings described in the text. All models are first order HMMs: EM trained using expectation maximization, VEM trained using variational EM transition priors and observation priors shown in parentheses; Sparse trained using PR where the constraint strength (σ) in parentheses. **Bold** indicates the best value for each column. Under a paired t-test all results are significant(p=0.005) against the EM model. Exceptions are marked with a star.

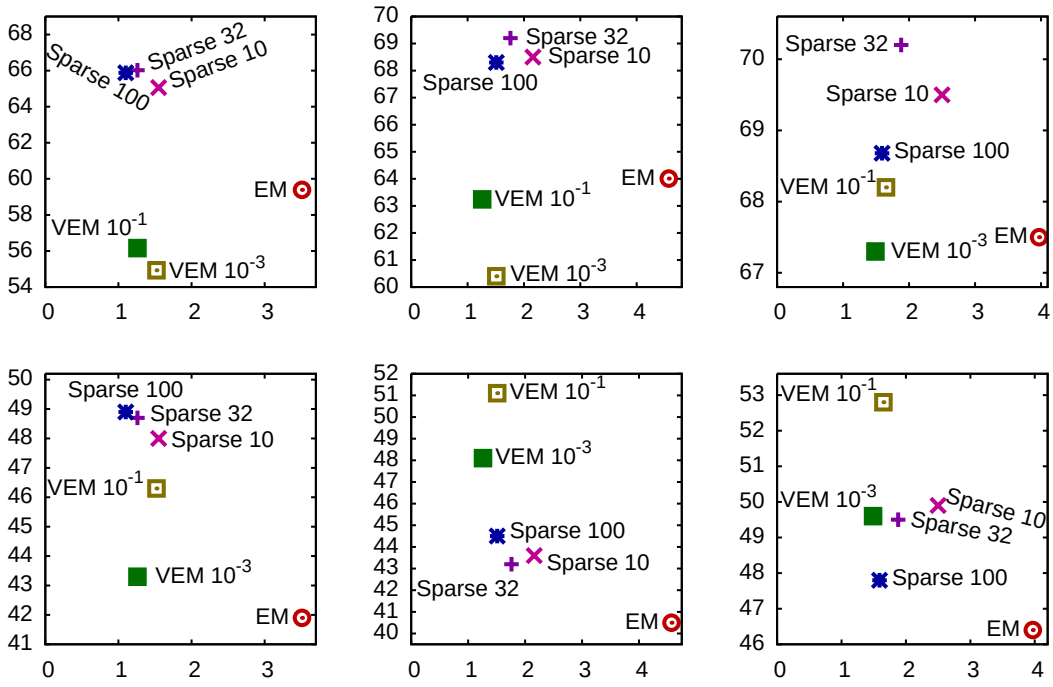

Figure 1: L1LMax vs Accuracy. Top 1-Many mapping, Bottom 1-1 mapping. Left: BulTree, Middle: PT-Conll Right: PTB17

References

[1] S. Afonso, E. Bick, R. Haber, and D. Santos. Floresta Sinta(c)tica: a treebank for Portuguese. In *In Proc. LREC*, pages 1698–1703, 2002.

[2] John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the L1-ball for learning in high dimensions. In *In Proc. ICML*, pages 272–279. ACM, 2008.

[3] M.P. Marcus, M.A. Marcinkiewicz, and B. Santorini. Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*, 19(2):313–330, 1993.

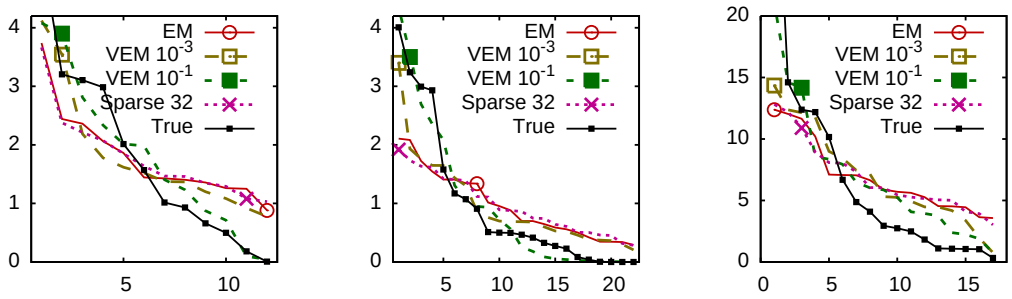

Figure 2: token distribution per hidden state. left: BulTree, middle: PT-Conll, right: PTB17. vertical axis: number of tokens in tens of thousands.

	EM	VEM 0.001	VEM 0.1	Sparse 10	Sparse 32	Sparse 100	Max
BG	1.32	1.30	1.24	1.50	**1.57**	1.55	3.05
PT-Conll	1.80	1.86	1.55	1.97	**2.07**	2.07	3.49
PTB17	1.75	1.70	1.72	1.91	**1.94**	1.89	3.22

Table 8: Mutual information in bits between gold tag distribution and hidden state distribution. Max is the mutual information for a hypothetical optimal distribution that completely matches the gold distribution.

[4] Kiril Simov, Petya Osenova, Milena Slavcheva, Sia Kolkovska, Elisaveta Balabanova, Dimitar Doikoff, Krassimira Ivanova, Alexander Simov, Er Simov, and Milen Kouylekov. Building a linguistically interpreted corpus of bulgarian: the bultreebank. In *In Proc. LREC*, page pages, 2002.

[5] N.A. Smith and J. Eisner. Contrastive estimation: Training log-linear models on unlabeled data. In *In Proc. ACL*, pages 354–362, 2005.

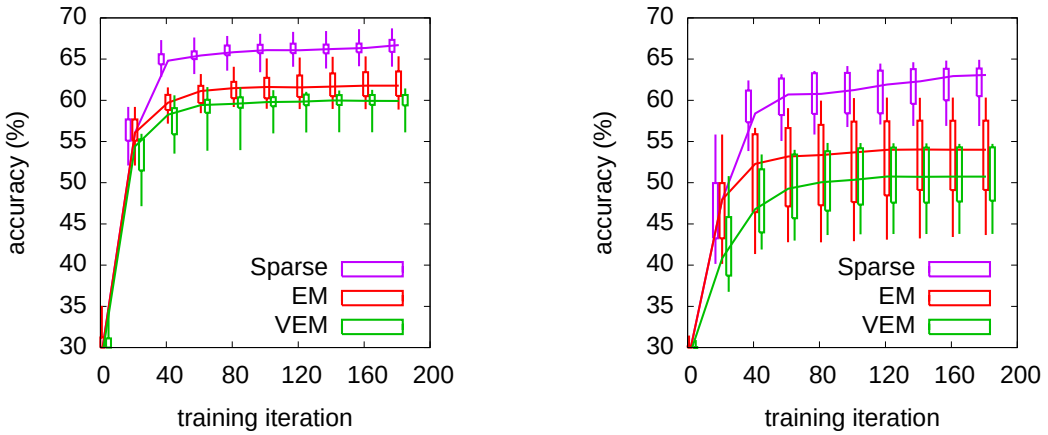

Figure 3: 1-Many learning curves for PT-Conll (left panel) and BulTree (right panel). Best parameter values where chosen from Table 7. Boxes extend from 1^{st} to 3^{rd} quartile and whiskers give minimum and maximum values.

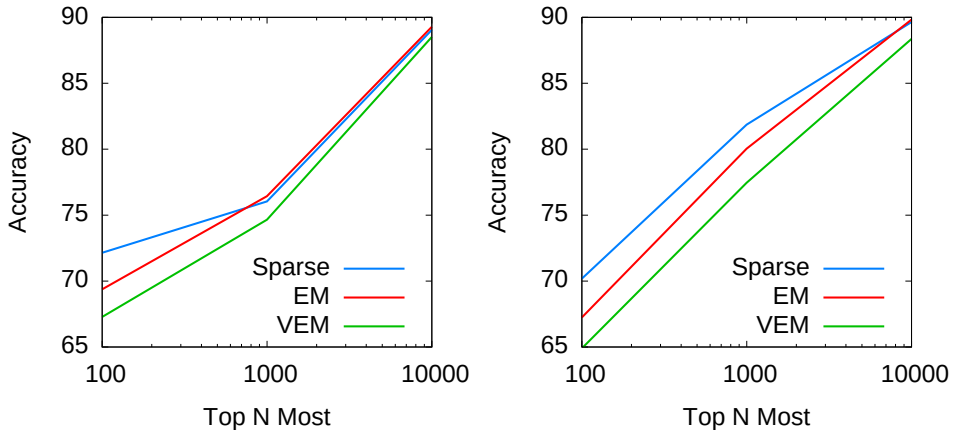

Figure 4: Performance of the three systems for different dictionary sizes on the PTB17 corpus (left) and the PT-CoNLL corpus (right).

Estimator	PT-Conll		BulTree		PTB17	
	1-Many	1-1	1-Many	1-1	1-Many	1-1
EM	69.4 (1.2)	49.6 (3.5)	73.2 (1.7)	60.4 (1.4)	67.2 (3.0)	53.7 (4.0)
VEM (10^{-1})	67.0 (0.9)	53.6 (2.7)	68.1 (3.5)	57.9 (3.0)	64.9 (3.9)	50.8 (2.0)
VEM (10^{-3})	67.3 (0.8)	48.5 (2.9)	68.8 (3.2)	57.2 (3.2)	64.9 (4.0)	46.0 (2.2)
VEM (10^{-4})	67.3 (0.7)	48.5 (2.9)	68.8 (3.2)	57.2 (3.3)	64.9 (4.0)	46.0 (2.3)
Sparse (10)	72.0 (1.1)	**52.3** (4.7)	76.2(1.0)	63.7 (1.3)	70.6(1.8)	**54.5 (2.2)**
Sparse (32)	**72.1 (1.1)**	52.5 (4.1)	**76.4 (1.1)**	**63.8 (0.9)**	**71.4 (2.2)**	53.8 (3.0)

Table 9: Weakly-supervised condition with a dictionary for the 100 most common words in the corpus. Average accuracy (standard deviation in parentheses) over 5 different runs (same seeds used for each model) for 200 iterations. Row and column headings and conventions are as in Table **??**.