Learning Tree Conditional Random Fields

Joseph K. Bradley JKBRADLE@CS.CMU.EDU
Carlos Guestrin GUESTRIN@CS.CMU.EDU
Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA, USA (both authors)

Abstract

We examine maximum spanning tree-based methods for learning the structure of tree Conditional Random Fields (CRFs) $P(\mathcal{Y}|\mathcal{X})$. We use edge weights which take advantage of local inputs $\mathcal{X}$ and thus scale to large problems. For a general class of edge weights, we give a negative learnability result. However, we demonstrate that two members of the class–local Conditional Mutual Information and Decomposable Conditional Influence–have reasonable theoretical bases and perform very well in practice. On synthetic data and a large-scale fMRI application, our methods outperform existing techniques.

1. Introduction

The study of probabilistic graphical models (cf., Koller and Friedman (2009)) often focuses on Bayesian networks, Markov random fields, and other generative models of probability distributions $P(\mathcal{Y})$. However, *Conditional Random Fields (CRFs)* (Lafferty et al., 2001), which model conditional distributions $P(\mathcal{Y}|\mathcal{X})$, offer computational and statistical advantages when we require $P(\mathcal{Y}|\mathcal{X})$ but not the full joint distribution $P(\mathcal{Y}, \mathcal{X})$. CRFs have been successful in a variety of areas, from natural language processing tasks such as part of speech tagging (Lafferty et al., 2001) to activity recognition (Vail et al., 2007) and heart motion abnormality detection (Schmidt et al., 2008).

Most research on CRFs has focused on inference or parameter learning with fixed, expert-chosen structures. Reliance on hand-picked structures is often pragmatic, for structure learning can be very expensive. In fact, parameter learning, which is often a subroutine in structure learning methods, requires inference for each training example at each iteration for CRFs; inference, in turn, is intractable for general CRFs (Koller & Friedman, 2009). Thus, although structured models of $\mathcal{Y}$ are more expressive than unstructured ones, researchers often use unstructured models for simplicity, especially when expert knowledge does not specify a structure (e.g., Palatucci et al. (2009)).

Low-treewidth CRFs–CRFs for which the outputs $\mathcal{Y}$ form a low-treewidth graph–are a key exception, permitting tractable inference and parameter learning (Shahaf et al., 2009). In this work, we develop methods for efficiently learning tree structures for CRFs. We take advantage of *local inputs*, where outputs Y_i are only directly influenced by small subsets of $\mathcal{X}$. We advocate a simple maximum spanning tree-based algorithm, for which we examine a general class of easily estimated edge weights. In a negative result for this class, we prove a major distinction between learning tree CRFs $P(\mathcal{Y}|\mathcal{X})$ and learning trees $P(\mathcal{Y})$. However, for two weighting methods in this class, we demonstrate favorable theoretical properties and, via analysis on simple models, can expose the domains in which they succeed. Using synthetic data, we demonstrate empirically that our methods represent a significant improvement over previous methods in terms of recovering the true structure of tree CRFs. Finally, we apply our methods to a structured prediction problem involving fMRI data from Palatucci et al. (2009).

2. CRF Structure Learning

We first define CRFs and structure learning. Let $\mathcal{X}$ be a set of *input variables* and $\mathcal{Y}$ a set of *output variables*. A CRF models a conditional distribution as $P(\mathcal{Y}|\mathcal{X}) = \frac{1}{Z(\mathcal{X})} \prod_j \phi_j(Y_{C_j}, X_{C_j})$, where $Y_{C_j} \in \mathcal{Y}, X_{C_j} \in \mathcal{X}$ form the domain of factor ϕ_j and Z is a *partition function* dependent on $\mathcal{X}$. As for generative models, if the CRF's graph (over $\mathcal{Y}$, not $\mathcal{X}$) is *low-treewidth*, inference and parameter learning are tractable. Representing and learning $P(\mathcal{Y}|\mathcal{X})$ can be easier than working with the joint $P(\mathcal{Y}, \mathcal{X})$. Even if $P(\mathcal{Y}, \mathcal{X})$ is high-treewidth, the conditional $P(\mathcal{Y}|\mathcal{X})$ may not be. (If $P(\mathcal{Y}, \mathcal{X})$ is low-treewidth, so is $P(\mathcal{Y}|\mathcal{X})$.) We distin-

Appearing in *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, 2010.

Algorithm 1 Tree CRF Structure Learner

Input: Dataset D over $\mathcal{Y}, \mathcal{X}$; inputs X_i for each Y_i
Initialize G to be the complete graph over $\mathcal{Y}$.
for all (Y_i, Y_j) **do**
  In G, set $Weight(Y_i, Y_j) \longleftarrow Score(i, j)$.
end for
Return: $MaxSpanningTree(G)$

guish between three learning tasks: feature selection, structure learning, and parameter learning.

Feature selection means to select inputs X_{C_j} relevant to each set Y_{C_j}; i.e., a factor ϕ_j involving Y_{C_j} can only use the selected inputs: $\phi_j = \phi_j(Y_{C_j}, X_{C_j})$. In this work, we assume we are given an *input mapping* specifying inputs X_{C_j} for each potential factor involving Y_{C_j}. For many applications, expert knowledge can provide input mappings (such as in image segmentation, where $\mathcal{Y}$ are segment labels and $\mathcal{X}$ are pixels); in other cases, sparsistent methods (Ravikumar et al., 2008) can be used for feature selection.

Structure learning means to select outputs Y_{C_j} for each factor, thus choosing a graph over $\mathcal{Y}$ representing conditional independence in the distribution. Though expert knowledge can sometimes dictate structure, conditional independence can be less intuitive than non-conditional independence (such as in our fMRI application), making structure learning a valuable tool.

Parameter learning means choosing the values of the factors $\phi_j(Y_{C_j}, X_{C_j})$, where the factor domain $Y_{C_j} \cup X_{C_j}$ is fixed. This has been well-studied for both tractable structures (e.g., Lafferty et al. (2001)) and intractable (e.g., Schmidt et al. (2008)).

2.1. Related Work

Few papers address CRF structure learning. Torralba et al. (2004) proposed Boosted Random Fields, which select features and learn CRF structure using greedy steps to approximately maximize data log likelihood. Schmidt et al. (2008) proposed maximizing block-ℓ_1-regularized pseudolikelihood, which gives a convex program and tends to produce sparse models. However, both methods learn intractable (high-treewidth) models in general. Shahaf et al. (2009) learn low-treewidth CRFs by using ideas from graph cuts to maximize a conditional mutual information-based criterion. Like us, Schmidt et al. (2008) and Shahaf et al. (2009) assume pre-specified input mappings.

Like Shahaf et al. (2009), we consider learning tractable, low-treewidth models. They demonstrate that low-treewidth structures can perform better than more general structures which require approximate inference and parameter learning. After stating our approach, we compare it with theirs in Section 3.1.

We are not aware of learnability results for CRF structures other than learning being hard in general. This claim follows from the hardness of learning general generative models (e.g., Srebro (2003)).

3. Efficiently Recovering Tree CRFs

Since we only learn tree CRFs, we simplify notation. We write $Y_{ij} \equiv \{Y_i, Y_j\}$. The inputs for Y_i specified by the input mapping are X_i; likewise, X_{ij} corresponds to Y_{ij}. Our goal now is a scalable algorithm for learning tree CRF structures. We begin by proposing a gold standard and showing it is ideal but impractical.

3.1. A Gold Standard

We can define an algorithm analogous to Chow-Liu for generative models (Chow & Liu, 1968) by showing that the conditional log likelihood of a tree CRF *decomposes* over the edges (i, j) and vertices i in the tree T. Let Q be our model and P the true distribution. Using the (optimal) parameters $Q(A|B) = P(A|B)$,

$$\begin{aligned}\mathrm{E}_P\left[\log Q(\mathcal{Y}|\mathcal{X})\right] &= \textstyle\sum_{(i,j)\in T} \mathrm{E}_P\left[\log Q(Y_{ij}|\mathcal{X})\right] \\ &\quad - \textstyle\sum_i (deg_i - 1)\mathrm{E}_P\left[\log Q(Y_i|\mathcal{X})\right] \\ &= \textstyle\sum_{(i,j)\in T} I_P(Y_i; Y_j|\mathcal{X}) + C,\end{aligned}$$

where subscript P means w.r.t. P, deg_i is the degree of vertex i, and C is a constant w.r.t. the structure.

Assuming the *global Conditional Mutual Information (CMI)* $I(Y_i; Y_j|\mathcal{X})$ is easy to compute, we can recover the maximum likelihood model: $\forall(i, j)$, compute $I(Y_i; Y_j|\mathcal{X})$, and choose the maximum spanning tree. This method was proposed by Friedman et al. (1997) for learning Tree-Augmented Naive Bayes classifiers. Unfortunately, as the dimensionality of $\mathcal{X}$ grows, this method quickly becomes intractable. Computing $I(Y_i; Y_j|\mathcal{X})$ requires an accurate estimate of $P(Y_{ij}|\mathcal{X})$ (or similar quantities); $P(Y_{ij}|\mathcal{X})$ can be as expensive to compute and represent as $P(\mathcal{Y}|\mathcal{X})$ when the dimensionalities of $\mathcal{Y}$ and $\mathcal{X}$ are of the same order. This observation emphasizes the need to parametrize our model with *local inputs* $X_i \subset \mathcal{X}$, rather than *global inputs* $X_i = \mathcal{X}$, to ensure scalability w.r.t. $\mathcal{X}$. [1]

Ideally, we could retain the efficiency of spanning trees while making use of local inputs in $\mathcal{X}$, i.e., only calcu-

[1] If the true model were a tree CRF with local inputs and complexity were ignored, global CMI could recover the true structure, after which the model could be succinctly parametrized with local inputs. However, if the true CRF were not a tree, it is unclear whether global CMI would recover the optimal projection onto a tree with local inputs.

lating probabilities of the form $P(Y_{ij}|X_{ij})$ conditioned on small sets X_{ij}. With local inputs, though, the partition function prevents the conditional log probability from decomposing over edges and vertices:

$$\log P(\mathcal{Y}|\mathcal{X}) = -\log Z(\mathcal{X}) + \sum_{(i,j)\in T} \log \phi_{ij}(Y_{ij}, X_{ij}).$$

Our primary goal is to overcome the intractability of the partition function by deriving a meaningful local edge score $Score(i,j)$ usable in Algorithm 1.

Shahaf et al. (2009) took a similar approach, defining edge scores and maximizing the weight of edges chosen for a low-treewidth model. However, they used global inputs, with edge scores set to the global CMI $I(Y_i; Y_j|\mathcal{X})$. They focused on the second step–maximizing the weight of edges included in the model–while we focus on better methods for weighting edges. Our work is compatible with theirs; their algorithm could use our edge scores to learn treewidth-k models.

3.2. Score Decay Assumption

We phrase our analysis of edge scores in terms of recovering the structure of tree CRFs. I.e., we assume the true distribution is representable by a tree CRF. We begin by defining a desirable property for edge scores.

Definition: If the true model $P(\mathcal{Y}|\mathcal{X})$ is a tree T, the *Score Decay Assumption (SDA)* states that, for any edge $(i,j) \in T$, if a path in T of length > 1 between k, l includes (i,j), then $Score(i,j) > Score(k,l)$.

Intuitively, the SDA says the score between vertex pairs decays with distance. Yet this is less strict than requiring, e.g., that the assumption hold regardless of whether (i,j) is an edge in T; thus, the SDA does not rely on comparing pairs of edges not in T. This condition is necessary and sufficient for recovering trees.

Theorem 1 *Suppose $P(\mathcal{Y}|\mathcal{X})$ is representable by a tree CRF with structure T. The Score Decay Assumption holds for a score S w.r.t. P iff Algorithm 1 using score S can recover T.*

Proof: While building a maximum spanning tree over $\mathcal{Y}$ by Kruskal's Algorithm (Kruskal, Jr., 1956), say we add edge $(k,l) \notin T$. Let $path_{kl}$ be the path between k, l in T. There exists an edge $(i,j) \in path_{kl}$ not yet added, so $Score(i,j) < Score(k,l)$, violating the SDA. I.e., we add an edge not in T iff the SDA is violated, so Algorithm 1 recovers T iff the SDA holds. ■

Recall that we wish to use local scores $S(i,j) = f(Y_{ij}, X_{ij})$ for scalability. We make a simplifying assumption about the input mapping: a factor involving Y_{ij} depends on $X_i \cup X_j$ (not arbitrary X_{ij}). This assumption makes our methods less general but more practical. (E.g., feature selection requires inputs for each of $|\mathcal{Y}|$ outputs, rather than for $\binom{|\mathcal{Y}|}{2}$ possible edge factors.) We now define a general class of such scores.

Definition: The *Local Linear Entropy Scores* are scores $Score(i,j)$ representable as linear combinations of entropies over subsets of $\{Y_i, Y_j, X_i, X_j\}$.

This class of scores includes, for example, the local Conditional Mutual Information $I(Y_i; Y_j|X_{ij}) = H(Y_i|X_{ij}) + H(Y_j|X_{ij}) - H(Y_{ij}|X_{ij})$ we consider in Section 4.2. Unfortunately, the Local Linear Entropy Scores are insufficient in general for recovering tree CRFs, as the following theorem demonstrates.

Theorem 2 *Assume that the edge score S is symmetric, i.e., $S(i,j) = S(j,i)$. Even if we assume the class of distributions we are learning:*

- *are representable by tree CRFs,*
- *obey the input mapping, i.e., that if the true model has edge (i,j), a factor involving Y_{ij} involves no more inputs than X_{ij},*
- *and have no trivial potentials, i.e., that no potentials are deterministic or effectively absent,*

then every Local Linear Entropy Score S violates the Score Decay Assumption for some models from this class, even with exact entropy estimates.

Proof Sketch: Since conditional entropies equal a difference between non-conditional entropies, any Local Linear Entropy Score $S(i,j)$ may be written as

$$\mathbf{w} \cdot \big(H(Y_i) + H(Y_j), H(X_i) + H(X_j), H(Y_{ij}), H(X_{ij}), \\ H(Y_i, X_i) + H(Y_j, X_j), H(Y_i, X_j) + H(Y_j, X_i), \\ H(Y_i, X_{ij}) + H(Y_j, X_{ij}), H(Y_{ij}, X_i) + H(Y_{ij}, X_j), \\ H(Y_{ij}, X_{ij})\big)$$

(This has all non-conditional entropies, grouped since S is symmetric.) The proof considers a list of cases using these observations: **(1)** Since tree CRFs generalize trees (where $\mathcal{X} = \emptyset$), the score must (approximately) reduce to the mutual information $S(i,j) = I(Y_i; Y_j)$ when $\mathcal{X} = \emptyset$. **(2)** Since we can have arbitrary factors over $\mathcal{X}$, the score must not be exactly $S(i,j) = I(Y_i; Y_j)$. **(3)** We can introduce simplifying constraints by considering models for which the inputs $\mathcal{X}$ are conditionally independent given the outputs $\mathcal{Y}$.
These constraints allow us to prove that, for certain classes of distributions, we require $S(i,j) \approx I(Y_i; Y_j)$, but that such a score fails for other classes. ■

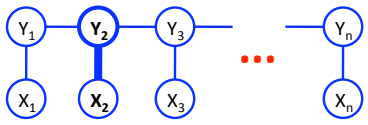

Figure 1. Counterexample for piecewise likelihood and local CMI: $P(Y, X)$ has a comb structure, with weak pairwise factors everywhere except for one strong factor $\phi(Y_2, X_2)$.

4. Heuristic Scores

Despite this negative result, we are able to identify certain Local Linear Entropy Scores which are intuitive and have desirable theoretical properties. Though the scores violate the Score Decay Assumption in general, they are very successful empirically (Section 5).

4.1. Piecewise Likelihood

We want to base our scores upon the conditional log likelihood, which does not decompose over edges because of the partition function $Z(\mathcal{X})$. Much research aims at handling partition functions tractably. We frame our analysis in terms of one such work: piecewise likelihood (Sutton & McCallum, 2005; 2007).

Piecewise likelihood approximates the log likelihood by upper-bounding $Z(\mathcal{X})$. The bound effectively divides $Z(\mathcal{X})$ into one term for each factor, permitting learning each factor's parameters independently. Sutton and McCallum (2005; 2007) apply this approximation to parameter learning for Markov random fields and CRFs, achieving fast learning and high accuracy. For pairwise models, if we combine factors with their Z terms, piecewise likelihood becomes $\sum_{(i,j)} \log P(Y_{ij}|X_{ij})$. This Local Linear Entropy Score is usable in Algorithm 1: $S(i,j) = \mathrm{E}_P[\log P(Y_{ij}|X_{ij})]$.

However, piecewise likelihood is a poor edge score. If a pair (Y_i, X_i) share a strong potential, piecewise likelihood weights will likely result in a star structure over $\mathcal{Y}$ with Y_i at the center, as in the following example.

Example 1: Consider Figure 1's model. (Y_2, X_2) has a strong potential so that $H(Y_2|X_2) \approx 0$. The score for any edge $(2, i)$ is:

$$\begin{aligned}\mathrm{E}[\log P(Y_{2,i}|X_{2,i})] &= -H(Y_{2,i}|X_{2,i}) \\ &= -H(Y_i|X_{2,i}) - H(Y_2|Y_i, X_{2,i}) \\ &\approx -H(Y_i|X_{2,i}) > -H(Y_i|X_i)\end{aligned}$$

All other X_j participate in weak potentials, so $-H(Y_i|X_i) \approx -H(Y_i|X_{ij}) > -H(Y_{ij}|X_{ij})$. Even if edge (Y_2, Y_i) is not in the true model and (Y_i, Y_j) is, we will score (Y_2, Y_i) above (Y_i, Y_j). This behavior appeared often in our synthetic experiments. ■

Nevertheless, piecewise likelihood is a useful approximation which helps tie our two proposed edge scores to the ideal but intractable global CMI.

4.2. Local CMI

The first Local Linear Entropy Score we select is a local version of global CMI $I(Y_i; Y_j|\mathcal{X})$. The *local CMI* is:

$$\begin{aligned}I(Y_i; Y_j|X_{ij}) = &\mathrm{E}_P\left[\log P(Y_{ij}|X_{ij})\right] \\ &-\mathrm{E}_P\left[\log P(Y_i|X_{i,j}) + \log P(Y_j|X_{i,j})\right].\end{aligned}$$

The local CMI score is interpretable as the piecewise likelihood (first line), minus correction terms for the edge's endpoints (second line). Intuitively, these corrections discount interactions between each Y_i and $\mathcal{X}$. We can also interpret local CMI as a bound on the likelihood gain of a tree CRF over a disconnected model.

Proposition 3 *Let $Q_T(\mathcal{Y}|\mathcal{X})$ be the projection of the true distribution $P(\mathcal{Y}|\mathcal{X})$ onto tree structure T w.r.t. $P(\mathcal{X})$, and let $Q_{disc}(Y|X) \equiv \prod_i P(Y_i|X_i)$ be the projection onto the disconnected model. The local CMI score $CMI(i,j)$ for Q_T bounds the likelihood gain* [2]*:*

$$\mathrm{E}_P\left[\log Q_T(\mathcal{Y}|\mathcal{X}) - \log Q_{disc}(\mathcal{Y}|\mathcal{X})\right] \geq \sum_{(i,j)\in T} CMI(i,j)$$

Proof: Choose a topological ordering of $\mathcal{Y}$ in T with root Y_1. Let Y_{Pa_i} be Y_i's parent. Entropies and expectations are defined w.r.t. $Q'(Y, X) \equiv Q_T(Y|X)P(X)$.

$$\begin{aligned}&E[\log Q_T(Y|X) - \log Q_{disc}(Y|X)] \\ &= \textstyle\sum_i H(Y_i|X_i) - H(Y|X) \\ &= \textstyle\sum_i H(Y_i|X_i) - H(Y_1|X) - \sum_{i>1} H(Y_i|Y_{Pa_i}, X) \\ &= H(Y_1|X_1) - H(Y_1|X) \\ &\quad + \textstyle\sum_{i>1} H(Y_i|X_i) - H(Y_i|Y_{Pa_i}, X) \\ &\geq H(Y_1|X_1) - H(Y_1|X_1) \\ &\quad + \textstyle\sum_{i>1} H(Y_i|X_{i,Pa_i}) - H(Y_i|Y_{Pa_i}, X_{i,Pa_i}) \\ &= \textstyle\sum_{i>1} H(Y_i|X_{i,Pa_i}) + H(Y_{Pa_i}|X_{i,Pa_i}) \\ &\qquad - H(Y_{i,Pa_i}|X_{i,Pa_i}) \\ &= \textstyle\sum_{(i,j)\in T} CMI(i,j) \quad ■\end{aligned}$$

Like piecewise likelihood, local CMI can perform poorly if a pair (Y_i, X_i) has a strong potential.

Example 2: Consider again Figure 1's model, where $\phi(Y_2, X_2)$ is strong enough that $H(Y_2|X_2)$ is small. Local CMI gives a small score to any edge with Y_2 since

$$I(Y_2; Y_j|X_{2,j}) = H(Y_2|X_{2,j}) - H(Y_2|Y_j, X_{2,j}) \approx 0.$$

Since the score for the false edge (Y_1, Y_n) does not condition on X_2, it could be much higher than the

[2] Both piecewise likelihood and local CMI have bounds relating them to the tree CRF log likelihood; we can show that neither bound is strictly better for all distributions.

score of the true edges $(Y_1, Y_2), (Y_2, Y_3)$. Note that this is a separate issue from identifiability; with local inputs, it is vital that we connect Y_1, Y_2 and Y_2, Y_3. ■ However, local CMI performs fairly well in practice.

4.3. Decomposable Conditional Influence

To overcome the above counterexample, we propose a final Local Linear Entropy Score dubbed the Decomposable Conditional Influence (DCI):

$$DCI(i,j) \equiv \mathrm{E}_P\left[\log P(Y_{ij}|X_{ij})\right] - \mathrm{E}_P\left[\log P(Y_i|X_i) + \log P(Y_j|X_j)\right].$$

The first expectation is the piecewise likelihood (as for local CMI), but the second has terms equal to the edge's endpoints' scores in the disconnected model $P_{disc}(Y|X) \equiv \prod_i P(Y_i|X_i)$, giving the following result:

Proposition 4 *When building a spanning tree T, if we add edge (i,j) and T does not yet contain edges adjacent to i, j, then DCI is an exact measure of the likelihood gain from adding edge (i,j).*

Moreover, DCI succeeds on Figure 1's counterexample for piecewise likelihood and local CMI.

Example 2, continued: The DCI edge scores are:

$$\begin{aligned} &DCI(1,2) \\ &\quad = -H(Y_{1,2}|X_{1,2}) + H(Y_1|X_1) + H(Y_2|X_2) \\ &\quad \approx -H(Y_1|X_{1,2}) + H(Y_1|X_1) \\ &DCI(1,n) \\ &\quad = -H(Y_{1,n}|X_{1,n}) + H(Y_1|X_1) + H(Y_n|X_n) \\ &\quad = [H(Y_1|X_1) - H(Y_1|X_{1,n})] \\ &\qquad + [H(Y_n|X_n) - H(Y_n|Y_1, X_{1,n})] \end{aligned}$$

Since Y_1, Y_n are far apart, the two terms in $DCI(1,n)$ are much closer to 0 than the sum in $DCI(1,2)$. ■ DCI performs very well in practice (Section 5.1).

4.4. Sample Complexity

Though Local Linear Entropy Scores violate the Score Decay Assumption in general, it is instructive to consider the sample complexity if the assumption is met.

Theorem 5 *Let $\mathcal{Y}, \mathcal{X}$ be discrete and $P(\mathcal{Y}|\mathcal{X})$ be representable by a CRF with tree structure T. (Assume w.l.o.g. that X_i is a single variable; we can merge multiple variables into a new variable of higher arity.) Let all variables have arity $\leq R$. Let $|\mathcal{Y}| = n$. Assume a Local Linear Entropy Score S meets the Score Decay Assumption by ϵ; i.e., for each edge $(i,j) \in T$ on the path between k, l, $S(i,j) - S(k,l) > \epsilon$.*
To recover the tree with probability at least $1-\gamma$, it suffices to train on a set of i.i.d. samples of size

$$O\left(\frac{R^8}{\epsilon^2}\log^2\left(\frac{R}{\epsilon}\right)\left(\log n + \log\frac{1}{\gamma} + \log R\right)\right)$$

Proof: We use this result on the sample complexity of estimating entropies from Hoffgen (1993): The entropy over k discrete variables with arity R may be estimated within absolute error Δ with probability $\geq 1-\gamma$ using $O\left(\frac{R^{2k}}{\Delta^2}\log^2\left(\frac{R^k}{\Delta}\right)\log\left(\frac{R^k}{\gamma}\right)\right)$ i.i.d. samples (and time).

Local Linear Entropy Scores may be represented by a constant number of entropies over $\leq k = 4$ variables. With $\binom{n}{2}$ potential edges, we must compute $O(n^2)$ entropies. To ensure estimates of scores order edges in the same way as the true scores, we must estimate scores within error $\Delta = \epsilon/2$. These values, Hoffgen's result, and a union bound complete the proof. ■

This theorem makes 2 key predictions: 1) sample complexity will increase only logarithmically in $n = |\mathcal{Y}|$, and 2) high arity R drastically increases sample complexity, indicating the importance of local inputs. Both predictions are born out by our empirical results.

5. Experiments

We first use synthetic data to compare local CMI and DCI against other CRF learning methods: piecewise likelihood, global CMI, and Schmidt et al. (2008)'s block-ℓ_1-regularized pseudolikelihood. (We omit results from piecewise likelihood since it does poorly.) We also tested two unstructured models: the *disconnected CRF with global inputs* $P(\mathcal{Y}|\mathcal{X}) = \prod_i P(Y_i|\mathcal{X})$ and the *disconnected CRF with local inputs* $P(\mathcal{Y}|\mathcal{X}) = \prod_i P(Y_i|X_i)$. Based on these results, we tested the best methods on the larger-scale fMRI application.

5.1. Synthetic Models

We tested a wide variety of synthetic models over binary variables; the models varied as follows:
Chains vs. trees: Chains have joint distributions $P(\mathcal{Y}, \mathcal{X})$ representable by ladders composed of cliques $(Y_i, Y_{i+1}), (X_i, X_{i+1}), (Y_i, X_i)$. Trees are the natural generalization, generated using non-preferential random attachment (Nakazatoa & Arita., 2007). We tested with and without *cross factors* $\phi(Y_i, X_{i+1})$.
Tractable vs. intractable joint models $P(\mathcal{Y}, \mathcal{X})$: For our tractable models, $P(\mathcal{Y}, \mathcal{X})$ may be sampled from directly. For our intractable models, $P(\mathcal{X})$ and $P(\mathcal{Y}|\mathcal{X})$ are tractable, but $P(\mathcal{Y}, \mathcal{X})$ is not (but may be sampled via $x \sim P(\mathcal{X}), y \sim P(\mathcal{Y}|\mathcal{X} = x)$).
Associative vs. random factors: Associative factors set $\phi(A, B) = \exp(s)$ if $A = B$ and $\phi(A, B) = 1$ if $A \neq B$, where s is a factor strength. Random factors have each value $\log\phi(a, b)$ sampled from $Uniform[-s, s]$. We set strengths s separately for Y-Y, Y-X, and X-X factors. For associative factors, we tried both fixed and alternating positive and negative strengths.

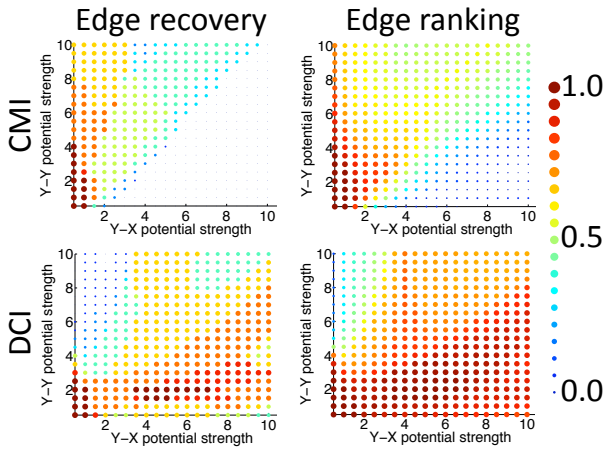

Figure 2. Local CMI vs. DCI with exact scores. Chains with associative factors, cross factors, alternating +/− strengths. Plots: varying magnitudes of Y-Y and Y-X strengths for fixed magnitude 2 for X-X (in log space). "Edge recovery" is the fraction of true edges recovered; "edge ranking" is the fraction of (true, false) edge pairs with the true edge ranked above the false. $|\mathcal{Y}| = |\mathcal{X}| = 10$.

These models have natural input mappings from Y_i to X_i which we gave our learners access to.

5.1.1. Exact Scores

We tested exact scores on simple models to illuminate domains in which local CMI and DCI succeed. Figure 2 shows model recovery results for length-10 chains. DCI outperforms local CMI when Y-X potentials are stronger (matching the counterexample in Figure 1) and when Y-Y potentials are weak. Peculiarly, local CMI does better when X-X potentials are strong. These trends were similar for most models we tested, though local CMI did significantly better without alternating positive and negative potentials.

5.1.2. Tests with Samples

Our next tests used samples from the synthetic models. For structure learning, we computed $P(Y_i, Y_j|X_C)$ where X_C is small using tables of counts. For large sets X_C, we used ℓ_2-regularized logistic regression. We chose regularization parameters separately for every regression during structure learning, which outperforms fixed regularization. We smoothed estimates of $P(A|B = b)$ with one extra example per $a \in Val(A)$.

For parameter learning, we used conjugate gradient to maximize the ℓ_2-regularized data log likelihood. For both structure and parameter learning, we chose regularization parameters via 10-fold cross-validation, testing 10 values between .001 and 50 (on a log scale). Global CMI has a natural parametrization with factors $P(Y_i, Y_j|\mathcal{X})$, but to be fair, we switched to factors with local inputs $\phi(Y_i, Y_j, X_{ij})$ during parameter learning, which gave higher performance.

We tested Schmidt et al. (2008)'s method using their implementation of structure and parameter learning and inference (via loopy belief propagation).

Figures 3 and 4 show results for tree CRFs with intractable joints $P(\mathcal{Y}, \mathcal{X})$, associative factors with alternating potentials (Y-Y, Y-X, X-X alternating between ± 4, ± 2, ± 1, respectively, in log space), with cross factors. "Test accuracy" is 0/1 (predicting all of $\mathcal{Y}$ or not). Training time includes cross-validation for choosing regularization for structure but not parameter learning. Figure 3 compares varying training set sizes, while Figure 4 varies the model size.

In both, DCI consistently outperforms other methods in recovering true edges, except for small sample sizes. Global CMI is the next most competitive, overtaking DCI in accuracy with enough training data. However, global CMI becomes prohibitively expensive as the training set and model sizes increase. Like DCI, local CMI is tractable, but it underperforms DCI.

These tests are difficult for the Schmidt et al. (2008) method, for it learns general CRFs, not tree CRFs. The plots omit log likelihood for Schmidt et al. since it is intractable to compute, though it could be approximated via a projection. We omit results from other models for lack of space. In general, DCI performs best, especially with large models and random factors.

Though local CMI and DCI do not obey the Score Decay Assumption in general, we observed that they approximately follow it. Figure 5 plots SDA violation for consecutive triplets (i, j, k) in the true CRF, measured as $(1/2)[(S(i,k) - S(i,j)) + (S(i,k) - S(j,k))]$. SDA violation and edge recovery are strongly anti-correlated.

5.2. fMRI

We next applied our CRF learning methods to an fMRI application from Palatucci et al. (2009). The learners takes inputs $\mathcal{X}$ which are voxels (3-D pixels) from fMRI images of test subjects' brains and predicts a vector $\mathcal{Y}$ of semantic features which describe what the test subject is thinking of (e.g., "Is it man-made?"; "Can you hold it?"). This application is much more challenging than our synthetic experiments. After preprocessing, their dataset has 60 examples (objects), with $|\mathcal{Y}| = 218$ and $|\mathcal{X}| = 500$, for each of 9 test subjects. Palatucci et al. (2009) gives more details.

Given the success of DCI in synthetic tests, we chose it for the fMRI data. $\mathcal{Y}$ and $\mathcal{X}$ are real-valued,

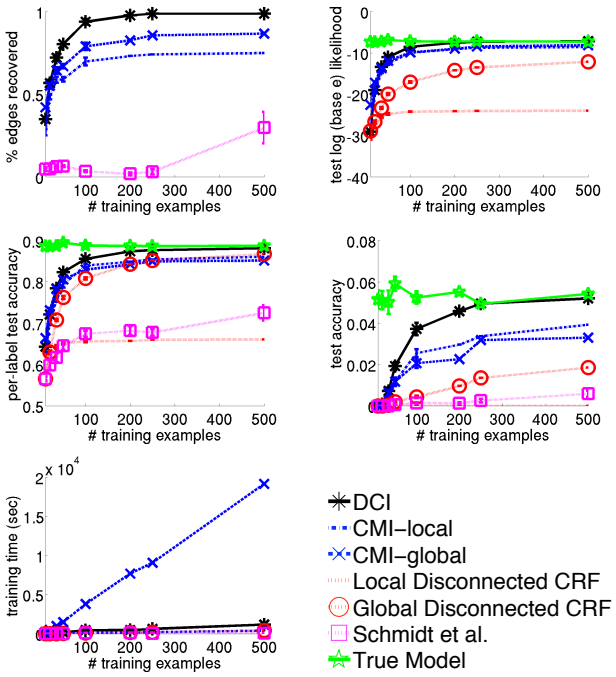

Figure 3. Synthetic data: Varying training set size. Tree CRFs $P(\mathcal{Y}|\mathcal{X})$ with associative factors. $|\mathcal{Y}| = |\mathcal{X}| = 40$. 1000 test examples. Averaged over 10 models/random samples; error bars (very small) show 2 standard errors.

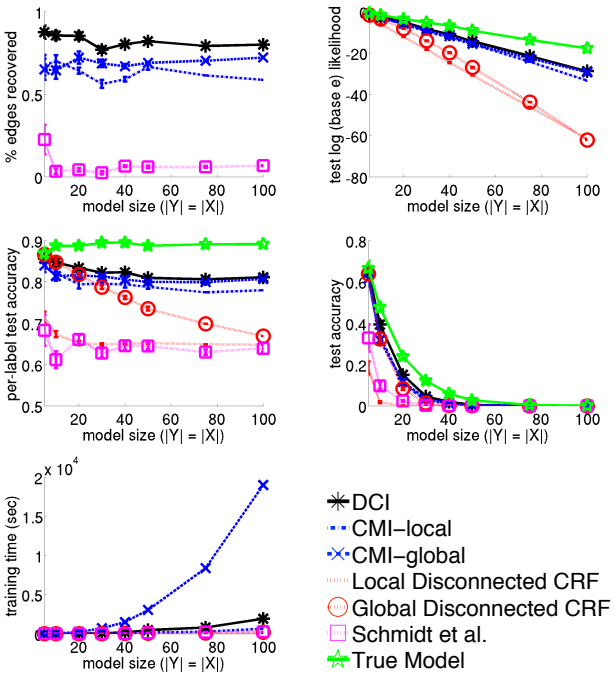

Figure 4. Synthetic data: Varying model size ($|\mathcal{Y}| = |\mathcal{X}|$). Tree CRFs $P(\mathcal{Y}|\mathcal{X})$ with associative factors. 50 training, 1000 test examples. Averaged over 10 models/random samples; error bars (very small) show 2 standard errors.

so we used conditional Gaussian factors $\phi(y, x) = \exp(-(1/2)(Ay - (Cx + b))^2$, where y, x are vectors. This parametrization is similar to that of Tappen et al. (2007), though they do not do general parameter learning, and it permits unconstrained optimization. [3] We regularized A and C, b separately, choosing regularization via 10-fold cross validation (CV) on values between .0001 and 30 (in a grid in log space). Because of the expense of CV, we ran CV on test subject 0 and used the chosen regularization for subjects 1-8.

With no natural input mapping, we used ℓ_1-regularized regression to do feature selection. To decrease the number of parameters (for both computational and statistical benefits), we tried two methods:
CRF 1: We chose ≤ 10 highest-weight inputs per Y_i, accounting for 1/5 of the regression weights on average. To use all of $\mathcal{X}$ without increasing complexity, we added fixed factors $\phi(Y_i, \mathcal{X}) = P(Y_i|\mathcal{X})$, $\forall i$.
CRF 2: We chose ≤ 20 inputs per Y_i; we added the same fixed factors. After structure learning, we parametrized edge factors to be independent of $\mathcal{X}$.

Palatucci et al. (2009) test *zero-shot learning*, which permits predictions about classes not seen during training. After predicting semantic features $\mathcal{Y}$ from images $\mathcal{X}$, they use hand-built "true" $\mathcal{Y}$ vectors to decode which object the test subject is thinking of. For testing, they use leave-2-out CV: Train on 58 objects; predict $\mathcal{Y}$ for 2 held-out objects i, j. Object i is classified correctly if its predicted $\widehat{\mathcal{Y}}^{(i)}$ is closer in ℓ_2 norm to its true $\mathcal{Y}^{(i)}$ than to the true $\mathcal{Y}^{(j)}$.

We used the same setup, scoring using their accuracy measure, squared error of predicted $\mathcal{Y}$, and log probability $\log P(\mathcal{Y}|\mathcal{X})$. Palatucci et al. (2009) use ridge regression $Y_i \sim \mathcal{X}, \forall i$, equivalent to a disconnected CRF with global inputs; we used this as a baseline.

Figure 6 compares disconnected and tree CRFs. The discrepancy between the 3 performance metrics is remarkable: Tree CRFs are best at predicting $\mathcal{Y}$ w.r.t. log likelihood and squared error (before decoding), but disconnected CRFs are best w.r.t. the accuracy metric (after decoding). This behavior could be caused by decoding via Euclidean distance and not accounting for the relative importance of each Y_i. We also tested decoding by predicting the more likely of the two held-out objects' $\mathcal{Y}$ vectors, but this performed worse with all learning methods. Learning this decoding $\mathcal{Y} \longrightarrow$ objects might avoid this problem and be a valuable addition to the zero-shot learning framework.

[3] We technically must constrain A so that $A^T A$ is invertible, but this was not a problem in our tests.

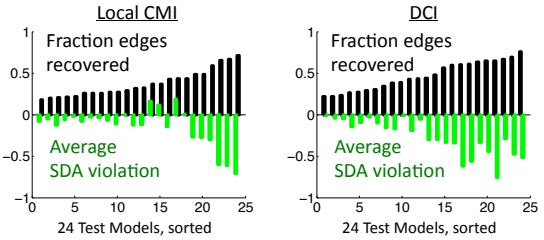

Figure 5. Score Decay Assumption violation vs. edge recovery on 24 models: ($|\mathcal{Y}| = 10, 15, 20$) $\times$ (with/out cross factors) $\times$ (2 associative factor types, 2 random). SDA violation averaged over consecutive triplets. For edge recovery, up=better; for SDA violation, down=better. 50 train exs. Tests averaged over 10 samples. Tractable $P(\mathcal{Y}, \mathcal{X})$.

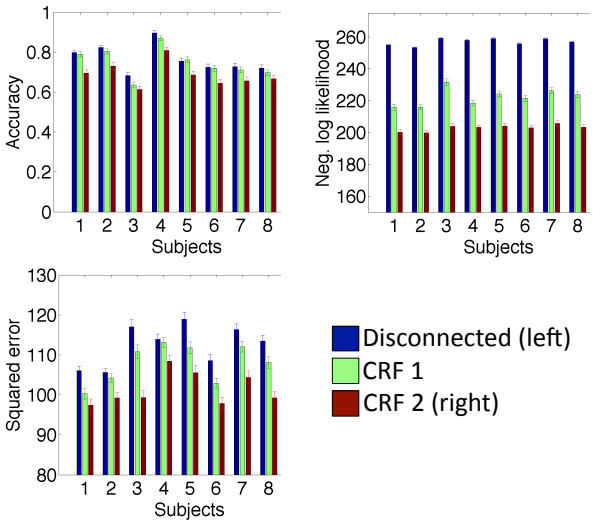

Figure 6. fMRI results. Error bars are 2 standard errors long. CRF structure and parameter learning with fixed regularization took only about 2-3 times as long as closed-form leave-one-out cross validation for ridge regression.

6. Discussion

Combining a maximum spanning tree algorithm with carefully chosen edge scores allows us to learn expressive models while avoiding the costliness of many structure learning methods. Despite our negative result for Local Linear Entropy Scores, local CMI and DCI scores can often recover the edges of tree CRFs.

Using our edge scores with Shahaf et al. (2009)'s generalization of the maximum spanning tree approach is a natural next step. Also, finding subclasses of tree CRFs which are recoverable via local scores would be worthwhile. We are currently applying our methods to templated models for relational data, a learning setting which naturally suggests local inputs.

Software: Our CRF learning code is available at `http://www.select.cs.cmu.edu/code/index.html`.

Acknowledgements

Many thanks to Mark Palatucci and Dean Pomerleau for the fMRI data and helpful advice and to the anonymous reviewers for their feedback. This work was supported by NSF Career IIS-0644225, ONR MURI N000140710747, and ARO MURI W911NF0810242.

References

Chow, C.K. and Liu, C.N. Approximating discrete probability distributions with dependence trees. *IEEE Trans. on Info. Theory*, 14:462–467, 1968.

Friedman, N., Geiger, D., and Goldszmidt, M. Bayesian network classifiers. *Machine Learning*, pp. 131–163, 1997.

Höffgen, K.-U. Learning and robust learning of product distributions. In *COLT*, 1993.

Koller, D. and Friedman, N. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

Kruskal, Jr., J.B. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proc. AMS*, 7(1):48–50, 1956.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001.

Nakazatoa, K. and Arita., T. A growth model of community graph with a degree distribution consisting of two distinct parts. *Physica A: Stat. Mech. and Apps.*, 376: 673–678, 2007.

Palatucci, M., Pomerleau, D., Hinton, G., and Mitchell, T. Zero-shot learning with semantic output codes. In *NIPS*, 2009.

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. Model selection in gaussian graphical models: High-dimensional consistency of l1-regularized mle. In *NIPS*, 2008.

Schmidt, M., Murphy, K., Fung, G., and Rosales, R. Structure learning in random fields for heart motion abnormality detection. In *CVPR*, 2008.

Shahaf, D., Chechetka, A., and Guestrin, C. Learning thin junction trees via graph cuts. In *AI-Stats*, 2009.

Srebro, N. Maximum likelihood bounded tree-width markov networks. *AI*, 143(1):123–138, 2003.

Sutton, C. and McCallum, A. Piecewise training of undirected models. In *UAI*, 2005.

Sutton, C. and McCallum, A. Piecewise pseudolikelihood for efficient training of conditional random fields. In *ICML*, 2007.

Tappen, M.F., Liu, C., Adelson, E.H., and Freeman, W.T. Learning gaussian conditional random fields for low-level vision. In *CVPR*, 2007.

Torralba, A., Murphy, K., and Freeman, W. Contextual models for object detection using boosted random fields. In *NIPS*, pp. 1401–1408, 2004.

Vail, D. L., Veloso, M. M., and Lafferty, J. D. Conditional random fields for activity recognition. In *Proc. 6th Intl. Joint Conf. on Autonomous Agents and Multiagent Systems*, pp. 1–8, 2007.