Constructing Composite Likelihoods in General Random Fields

Sebastian Nowozin SEBASTIAN.NOWOZIN@MICROSOFT.COM
Microsoft Research Cambridge, 21 Station Road, Cambridge, UK

Abstract

We propose a simple estimator based on composite likelihoods for parameter learning in random field models. The estimator can be applied to all discrete graphical models such as Markov random fields and conditional random fields, including ones with higher-order energies. It is computationally efficient because it requires only inference over tree-structured subgraphs of the original graph, and it is consistent, that is, it asymptotically gives the optimal parameter estimate in the model class. We verify these conceptual advantages in synthetic experiments and demonstrate the difficulties encountered by popular alternative estimation approaches.

1. Introduction

Conditional random fields (CRF) (Lafferty et al., 2001; Sutton & McCallum, 2007a) are among the most popular statistical models in computer vision. Given an observation $\boldsymbol{x}$, a CRF specifies a conditional distribution $p(\boldsymbol{y}|\boldsymbol{x})$ over $\boldsymbol{y} \in \mathcal{Y}$, where the domain $\mathcal{Y}$ is usually a finite but very large set, such as the set of all possible binary image labelings. These models have been applied to a variety of vision tasks, such as image segmentation, scene understanding, and image denoising.

CRFs are typically *parameterized* by some weight vector $\boldsymbol{w} \in \mathbb{R}^d$, for example to specify the pairwise interactions. The parameterized distribution $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w})$ is most commonly given in terms of an *energy function*,

$$p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}) = \frac{1}{Z(\boldsymbol{x}, \boldsymbol{w})} \exp(-E(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{w})), \qquad (1)$$

where $Z(\boldsymbol{x}, \boldsymbol{w}) = \sum_{\boldsymbol{y} \in \mathcal{Y}} \exp(-E(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{w}))$ is the *partition function* that normalizes the distribution.

While early users have set these parameters manually, much recent research has been devoted to *learn* these parameters from annotated training data $\{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n=1,\ldots,N}$. In this paper we introduce two novel approximations for parameter learning; one for pairwise grid-structured graphs commonly encountered in computer vision problems, and one for general unstructured graphs. Our proposed methods are special cases of the general *composite likelihood* (Lindsay, 1988). They are simple to implement, computationally efficient, and provide accurate parameter estimates. Moreover, our method is the first to provide a composite likelihood for general unstructured input graphs. We place our method in a wider context by relating it to other popularly used methods, shown in Figure 1. In particular, we expose some of the deficits of learning approaches based on approximate inference.

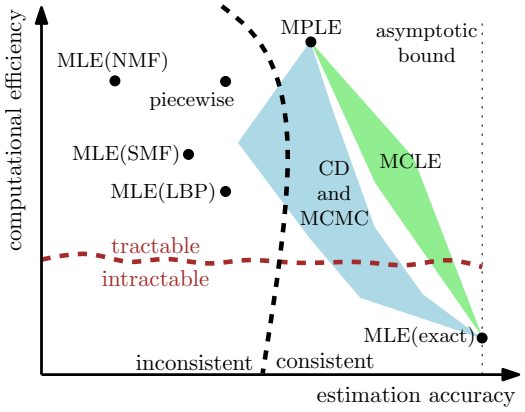

Figure 1. Simplified and schematic visualization of the estimation tradeoffs in random field models: exact maximum likelihood MLE(exact) is statistically efficient but intractable. Maximum likelihood estimators based on approximate inference such as *naive mean field* (NMF), *structured mean field* (SMF), or *loopy belief propagation* (LBP)—shown as MLE(NMF), MLE(SMF), and MLE(LBP)—are inconsistent, as is piecewise training. Stochastic approaches yield families of estimators (CD, MCMC-based, shaded blue) that can be consistent. Composite likelihoods (MCLE, shaded green), as advocated in this paper, are deterministic and combine favorable properties.

2. Parameter Estimation

In principle, the energy function $E : \mathcal{X} \times \mathcal{Y} \times \mathbb{R}^d \to \mathbb{R}$ can be arbitrary, but to encode independence assumptions and to remain computable it is chosen such that it additively decomposes over small subsets of all variables $\boldsymbol{y}$, i.e.,

$$E(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{w}) = \sum_{F \in \mathcal{F}} E_F(\boldsymbol{x}, y_F, \boldsymbol{w}), \tag{2}$$

where y_F is the subset of variables associated to the energy function E_F and we have a set $\mathcal{F}$ indexing these functions. This decomposition can be represented graphically by means of *factor graphs* (Kschischang et al., 2001; Koller & Friedman, 2009). Formally a factor graph is given as $G = (V, \mathcal{F}, \mathcal{E})$ with a set V of *variable nodes* (drawn "◯"), a set $\mathcal{F}$ of *factor nodes* (drawn "■"), and a set $\mathcal{E} \subseteq V \times \mathcal{F}$ of edges connecting variable and factor nodes. If we denote by $N(F) = \{i \in V : (i, F) \in \mathcal{E}\}$ the set of variables adjacent to a factor F, the so called *scope* of F, and write $y_F := y_{N(F)}$, we can directly read the energy function (2) from the graph (Nowozin & Lampert, 2011).

Approximate inference. In general, computing $Z(\boldsymbol{x}, \boldsymbol{w})$ for model (1) is intractable, as is computing expectations $\mathbb{E}_{\boldsymbol{y} \sim p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w})}[\cdot]$. While intractable to compute exactly, they can be approximated using *approximate inference* methods. Throughout this paper we use three popular approximate inference methods: naive mean field (NMF) (Wainwright & Jordan, 2008), structured mean field (SMF) (Bouchard-Côté & Jordan, 2009; Saul & Jordan, 1995), and loopy belief propagation (LBP) (Kschischang et al., 2001). Each method provides an approximate log-partition function and approximate marginal distributions for each factor, but the approximation quality can vary largely for different instances. NMF and SMF provide a lower bound $\tilde{Z}(\boldsymbol{x}, \boldsymbol{w}) \leq Z(\boldsymbol{x}, \boldsymbol{w})$ and realizable marginal distributions by finding a similar distribution within a tractable set of distributions; LBP is based on the *Bethe free energy* approximation of the distribution (Wainwright & Jordan, 2008).

2.1. Maximum Likelihood

For models of the form (1) we can use maximum likelihood estimation (MLE) to find an estimate of $\boldsymbol{w}$ from training data. Formally, this can be posed as finding the maximum aposteriori estimate of the posterior distribution $p(\boldsymbol{w}|\{(x_n, y_n)\}_{n=1,\dots,N})$, given a prior $p(\boldsymbol{w})$ on the parameters. This objective—the regularized likelihood—is defined as follows.

Definition 1 (Maximum Likelihood Estimation) *Given a family of model distributions $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w})$ parametrized by $\boldsymbol{w} \in \mathbb{R}^d$, a prior distribution $p(\boldsymbol{w})$, and a set of iid fully-observed training samples $\{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n=1,\dots,N}$, solve*

$$\boldsymbol{w}^* = \operatorname*{argmax}_{\boldsymbol{w} \in \mathbb{R}^d} \phi(\boldsymbol{w}), \tag{3}$$

$$\begin{aligned} \phi(\boldsymbol{w}) &= \log p(\boldsymbol{w}) + \sum_{n=1}^{N} \log p(\boldsymbol{y}_n|\boldsymbol{x}_n, \boldsymbol{w}) \\ &= \log p(\boldsymbol{w}) - \sum_{n=1}^{N} \Big[E(\boldsymbol{y}_n, \boldsymbol{x}_n, \boldsymbol{w}) + \log Z(\boldsymbol{x}_n, \boldsymbol{w}) \Big]. \end{aligned} \tag{4}$$

Numerical maximization of $\phi(\boldsymbol{w})$ requires the gradient in $\boldsymbol{w}$. The gradient $\nabla_{\boldsymbol{w}} \phi(\boldsymbol{w})$ is given by the difference of energy gradients between the *model expectation* and *sample expectation*, yielding the expression (Koller & Friedman, 2009)

$$\begin{aligned} \nabla_{\boldsymbol{w}} \phi(\boldsymbol{w}) = \nabla_{\boldsymbol{w}} \log p(\boldsymbol{w}) - \sum_{n=1}^{N} \Big(\nabla_{\boldsymbol{w}} E(\boldsymbol{y}_n, \boldsymbol{x}_n, \boldsymbol{w}) - \\ \mathbb{E}_{\boldsymbol{y} \sim p(\boldsymbol{y}|\boldsymbol{x}_n, \boldsymbol{w})} \left[\nabla_{\boldsymbol{w}} E(\boldsymbol{y}, \boldsymbol{x}_n, \boldsymbol{w}) \right] \Big). \end{aligned} \tag{5}$$

Consistency. Under technical assumptions (White, 1982) that can be easily satisfied[1], the MLE (3) applied to our discrete models is *consistent* in the sense defined by (White, 1982), that is, for almost every sequence of samples $(\boldsymbol{x}_n, \boldsymbol{y}_n)$ from the true distribution, we have $\boldsymbol{w}_N^* \to \hat{\boldsymbol{w}}$, where $\boldsymbol{w}_N^*$ is the estimate obtained by (3) from all samples up to and including the N'th sample, and $\hat{\boldsymbol{w}}$ is the parameter vector that minimizes the Kullback-Leibler divergence (Koller & Friedman, 2009) $D_{KL}(q(\boldsymbol{y}|\boldsymbol{x}) \| p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}))$ to the true distribution $q(\boldsymbol{y}|\boldsymbol{x})$. As the Kullback-Leibler divergence is a natural divergence measure between distributions, this intuitively means that a consistent estimator eventually recovers the "best possible" fit to the true distribution. **Inconsistent estimators.** If an estimator is not consistent it is said to be *inconsistent*. For such estimators, using more and more training data will not guarantee a better parameter estimate.

Does consistency matter? In the end, we care about a high accuracy on unseen test data. A method that is inconsistent but provides good estimates from few samples can be preferable over a method that is consistent but statistically inefficient, i.e. needing a large number of samples to produce reasonable estimates. That said, consistency is a desired property of

[1] For the models we consider the only condition that can be violated is the *identifiability* condition that ensures uniqueness of w^*.

an estimator and the decision to drop it in favor of something else should be based on empirical evidence.

2.2. Existing Methods and Literature Review

Unfortunately solving (3) is not tractable for all but the simplest models. Hence we now discuss some of the many available methods to solve (3) *approximately*.

Inference-based Likelihood Approximations. Arguably the most popular method to maximize (4) is to use the exact likelihood gradient expression (5) but to approximate $Z(\boldsymbol{x}_n, \boldsymbol{w})$ and the expectation $\mathbb{E}_{\boldsymbol{y}\sim p(\boldsymbol{y}|\boldsymbol{x}_n,\boldsymbol{w})}[\cdot]$ using an approximate inference method such as loopy belief propagation. Many approximate inference methods with different tradeoffs exist (Wainwright & Jordan, 2008; Koller & Friedman, 2009) but the effect of approximate inference when used for approximate parameter estimation is less clear.

The true log-likelihood function (4) is differentiable and if E_F is a linear function in $\boldsymbol{w}$, then (4) is a concave function in $\boldsymbol{w}$ (Koller & Friedman, 2009). When using approximate inference to compute the log-likelihood, the function is no longer guaranteed to be concave; it might have multiple modes and can be unbounded. Worse still, the gradient obtained from approximate marginals might not be the gradient of any differentiable function, preventing the principled use of line searches and quasi-Newton optimization methods for optimizing (4). This makes the optimization practically challenging, but even if an approximate maximum likelihood estimator is obtained heuristically it is typically *inconsistent*. In Section 3 we will demonstrate all these artifacts on a small toy experiment.

Piecewise training. The idea of *piecewise training* (Sutton & McCallum, 2005) is to decompose the model into tractable subgraphs, learning the weights for each part separately. This ignores interactions between parts during learning but is popular in computer vision, such as when a discriminative classifier is used to learn unary potentials separately (Shotton et al., 2007). The simplicity has motivated extensions to the approach; in (Sutton & McCallum, 2007b) the approach has been made more efficient by using conditioned factors, similar to the pseudolikelihood. Alahari et al. (2010) extend the piecewise training idea to max-margin training. Piecewise training is simple to implement and can be effective in practice; in general, however, piecewise training approaches are inconsistent (Sutton & McCallum, 2005). The spanning tree approximation of Pletscher et al. (2009) and surrogate likelihoods (Wainwright & Jordan, 2008) are also related to piecewise training.

Other objectives and methods. There exist a number of interesting alternative estimation methods some of which yield consistent estimators. *Score matching* (Hyvärinen, 2005) provides an alternative tractable objective function to estimate undirected models. For fully observed estimation the older method of *MCMC-MLE* (Descombes et al., 1999) combines the computational efficiency of deterministic maximization of an objective function with the asymptotic consistency of a sampling-based method. For partially observed data, *contrastive divergence* estimators (Hinton, 2002; Carreira-Perpiñán & Hinton, 2005; He et al., 2004) have become popular recently and attempts to generalize them have been made in the form of *contrastive objectives* (Vickrey et al., 2010).

As our methods are based on them, we discuss pseudo- and composite-likelihoods (Besag, 1977; Lindsay, 1988) separately and in more detail in Section 4.

2.3. Test-time Prediction

Once an approximation of $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w})$ has been obtained we apply statistical decision theory to solve the *structured prediction* task: given a sample $\boldsymbol{x}$ and a *structured loss* $\Delta : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$, predict $f(\boldsymbol{x}) \in \mathcal{Y}$ such that the expected loss $\mathbb{E}_{\boldsymbol{y}\sim p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w})}[\Delta(\boldsymbol{y}, f(\boldsymbol{x}))]$ is minimized. Depending on the structured loss Δ the prediction function $f : \mathcal{X} \to \mathcal{Y}$ differs; for example, the 0/1-loss $\Delta(\boldsymbol{y}, \boldsymbol{y}') = I(\boldsymbol{y} \neq \boldsymbol{y}')$ makes the MAP prediction $f(\boldsymbol{x}) = \operatorname{argmax}_{\boldsymbol{y}\in\mathcal{Y}} p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w})$ statistically optimal[2]; the Hamming loss $\Delta(\boldsymbol{y}, \boldsymbol{y}') = \sum_{i\in V} I(y_i \neq y'_i)$ makes the maximum posterior marginal (MPM) prediction rule $f(\boldsymbol{x}) = [\operatorname{argmax}_{y_i\in\mathcal{Y}_i} p(y_i|\boldsymbol{x}, \boldsymbol{w})]_{i\in V}$ statistically optimal.

In practice the issue of structured prediction in probabilistic models is often ignored and MAP inference is used without recognizing the implicit assumption of a 0/1-loss. However, typically the model is misspecified (White, 1982) and the best likelihood-based parameter estimate in the model class does not yield the best possible decisions as measured by a more general loss function (Pletscher et al., 2011). In this case as shown by Domke (2013) and Pletscher et al. (2011) a risk minimization approach based on the loss function of interest yields better predictive performance for the same class of energy functions.

Structured SVM. An alternative learning approach is empirical risk minimization in the form of the structured SVM (Tsochantaridis et al., 2005), recently popular in computer vision. While having distinct advantages in terms of tractability, one disadvantage is

[2] $I(\text{pred}) = 1$ if pred is true, 0 otherwise.

that there is no known method for risk minimization in random fields that is consistent – that is, eventually achieving the best possible generalization error with respect to the structured loss. In particular, the structured SVM is known to be inconsistent (McAllester, 2007). This is in contrast to likelihood-based parameter estimation, where a large number of consistent methods are available.

3. A Simple Experiment

To understand the different approximations to the likelihood function, we conduct the following simple experiment.

We define a simple pairwise d-by-d grid model as shown for $d = 3$ in Figure 3. All variables take one of three states, $\mathcal{Y}_i = \{1, 2, 3\}$ and the pairwise factors connecting them are the same throughout the model. The energies $E_F(y_i, y_j)$ are defined by the symmetric table

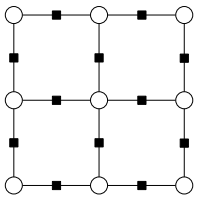

Figure 3. Grid graph.

$$\begin{bmatrix} 0 & a & b \\ a & 0 & a \\ b & a & 0 \end{bmatrix},$$

where $a = 0.5$ and $b = 0.7$ are fixed constants. The resulting model is homogeneous without unary interactions and moreover the pairwise interactions are submodular.

We obtain a fixed number of N samples from the model distribution and use different estimation methods to recover the parameters (a, b). Because the objective function is a function of only two arguments—a and b—we can visualize it. For maximum likelihood estimation based on exact and approximate inference we visualize the negative log-likelihood functions in Figure 2.

As can be seen from the results, using approximate inference methods to obtain an approximate value and gradient of the log-likelihood function is causing problems even on this small example with two parameters. For one, it most often yields inconsistent estimators: even with more and more training data we can not recover $\hat{\boldsymbol{w}}$. But it also makes optimization more challenging, namely, i) quasi-Newton methods such as L-BFGS break down in light of discontinuous gradients (NMF, SMF), ii) most line search methods rely on the assumption that the gradient and function values agree with each other, but this may no longer be the case, iii) in simple gradient methods the step size selection is harder to tune for non-convex problems (BP), and iv) being non-convex, only a local maximizer can be guaranteed.

4. Structural Likelihood Approximations

In light of the findings of the previous section, we take a step back and ask what properties we would like an *ideal* estimator to have. These are, 1. *Consistency*: asymptotic (in N) convergence to the best possible estimate, 2. *Computational efficiency*: efficiently computable, 3. *Statistical efficiency*: when consistent, having asymptotically the best possible rate of convergence, 4. *Determinism*: reproducible and deterministically computable.

Unfortunately, for general discrete graphical models no estimator satisfying all these properties can exist (Koller & Friedman, 2009). Therefore, to obtain a computable estimator we need to *give up* one or multiple of these desirable properties. In the following section we show that by giving up statistical efficiency we can obtain a practical estimator, in the form of composite likelihoods, that satisfies the other requirements.

4.1. Composite Likelihoods

Composite likelihoods (Lindsay, 1988; Dillon & Lebanon, 2009) are a family of estimation methods which we define as follows.

Definition 2 (Composite Likelihood) *Given a set $\{(A_j, B_j)\}_{j=1,\dots,k}$ of m-pairs (A_j, B_j) satisfying $A_j, B_j \subseteq V$, $A_j \cap B_j = \emptyset$, $A_j \neq \emptyset$ with weights $\beta_j > 0$, the* composite likelihood *of a fully observed dataset $D = \{(\boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)})\}_{n=1,\dots,N}$ is*

$$c\ell(\boldsymbol{w}; D) = \sum_{n=1}^{N} \sum_{j=1}^{k} \beta_j \log p(y_{A_j}^{(n)} | y_{B_j}^{(n)}, \boldsymbol{x}^{(n)}, \boldsymbol{w}). \quad (6)$$

The definition depends on "m-pairs": pairs of disjoint subsets of V. If these are chosen such that they define conditional distributions within the model and A_j contains all variables at least once, then it can be shown (Dillon & Lebanon, 2009) that maximizing (6) yields a consistent estimator. Note that when we choose $A_j = \{j\}$, $B_j = V \setminus \{j\}$, $\beta_j = 1$, we obtain the pseudolikelihood (Besag, 1977) as a special case. Likewise, for $A_1 = V$, $B_1 = \emptyset$ we obtain the exact likelihood. The composite likelihood family is therefore very general and also includes the inconsistent piecewise training objectives.

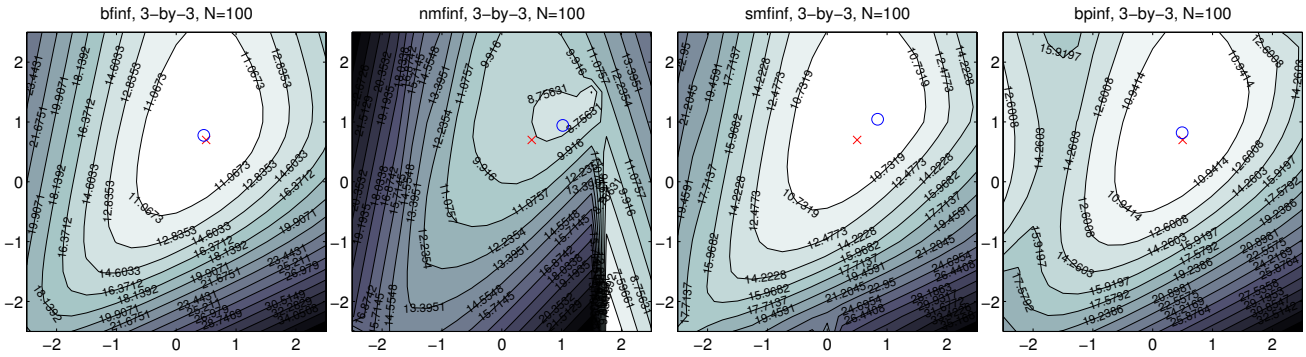

Figure 2. The drawbacks of inference-based likelihood approximations: the leftmost plot shows the true negative log-likelihood function for 100 samples on a 3-by-3 grid graph. The x/y-axis are the a and b parameter, respectively. The true parameters are marked by $\times$ in each plot, the (local) maximizer by $\circ$. The three following plots show the approximations provided by three different approximate inference methods: naive mean field (NMF), structured mean field (SMF), and sum-product belief propagation (BP). The plots are obtained numerically by evaluation on a 20-by-20 uniformly spaced mesh. Note that the true negative log-likelihood is convex; this is not the case for the naive mean field approximation which has a discontinuity arising from multimodality of the mean field objective. Here the structured mean field approximation yields an (almost) convex function but this is not the case in general. The belief propagation approximation is not convex but provides a good estimator locally.

Previous works using composite likelihood. Composite likelihoods have been introduced by Lindsay (Lindsay, 1988) as a generalization of the pseudo-likelihood (Besag, 1977). While popular in the statistics community (Varin et al., 2010), they have recently been used in machine learning (Liang & Jordan, 2008; Dillon & Lebanon, 2009; Asuncion et al., 2010). In particular, a stochastic version of (6) has been proposed by (Dillon & Lebanon, 2009), and a connection to contrastive divergence has been pointed out in (Asuncion et al., 2010).

We now propose two novel composite likelihoods, one applicable to grid graphs, and the other to general factor graphs.

4.2. Criss-cross Likelihood (MXXLE)

The first estimator we propose—named "criss-cross likelihood" (MXXLE)—is applicable to grid graphs with 4-neighborhood connectivity, as commonly used in computer vision. As shown in Figure 4 and 5 we propose to set A_j corresponding to horizontal and vertical chain-structured subgraphs of the original graph and $B_j = V \setminus A_j$, $\beta_j = 1$. For a w-$\times$-h graph we therefore have $w + h$ subgraphs, each being chain-structured. According to the previous requirements, maximizing (6) for this choice yields a consistent estimator. Moreover, to compute (6) and its gradient we only require inference over chains, which is tractable.

4.3. General factor graphs (MCLE)

For general factor graphs that do not follow a grid structure or contain higher-order interactions the above construction does not work. Instead we propose to construct composite likelihood objectives using so called "very acyclic" (*v-acyclic*) decompositions of the graph. The v-acyclic property has originally been proposed for structured mean field inference (Bouchard-Côté & Jordan, 2009). In essence, a v-acyclic decomposition is a set of subgraphs that contain blockwise conditionally independent variables. In (Bouchard-Côté & Jordan, 2009) the authors assumed a good v-acyclic decomposition is known and thus provided no algorithm to obtain a decomposition for a given graph.

To derive our proposed estimator we first need to extend the definition of v-acyclicity to higher-order factor graphs. In (Bouchard-Côté & Jordan, 2009) only pairwise interactions have been considered and the extension to the higher-order case was left ambiguous as there are multiple notions of acyclicity in hypergraphs.

Definition 3 (v-acyclic subgraph) *Given a factor graph $G = (V, \mathcal{F}, \mathcal{E})$, a factor graph $G' = (V, \mathcal{F}', \mathcal{E}')$ is called a* v-acyclic subgraph *of G, written $G' \subseteq_{vac} G$, if,*

1. *$\mathcal{F}' \subseteq \mathcal{F}$, $\mathcal{E}' = (V \times \mathcal{F}') \cap \mathcal{E}$, and*
2. *G' is acyclic (as an undirected graph), and*

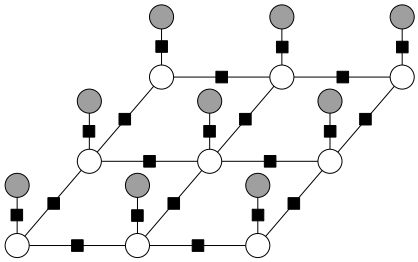

Figure 4. 3-by-3 grid-structured conditional random field with two factor types: unary observation factors and data-independent pairwise factors. For large grid graphs computing the likelihood function is intractable.

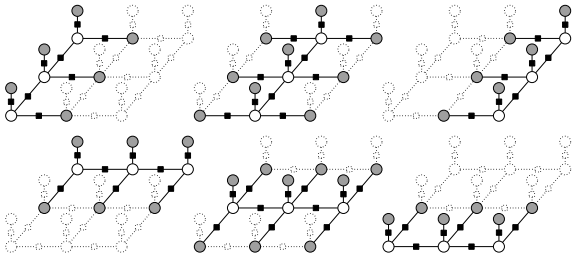

Figure 5. Crisscross likelihood approximation with six chain-structured subgraphs: three vertical and three horizontal chains. Factors and variable nodes drawn as dotted lines are not part of the component. Variable nodes that become shaded are instantiated with the observed ground truth.

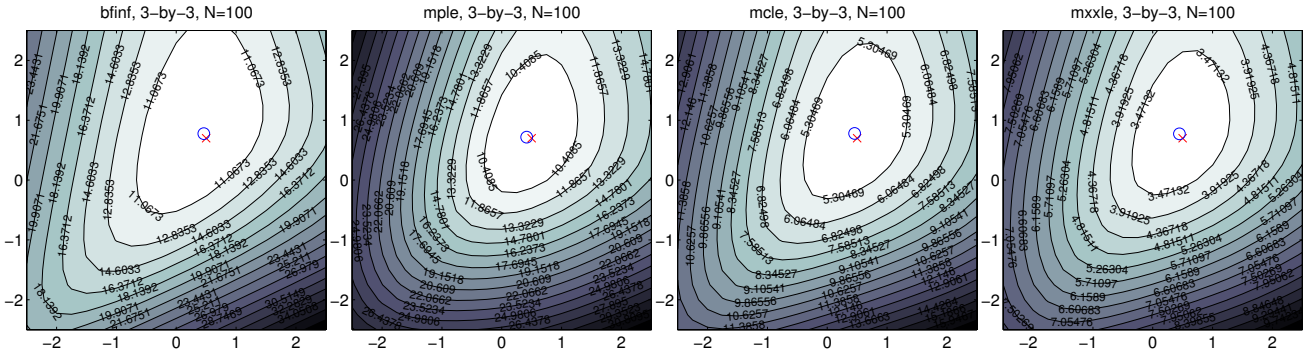

Figure 6. Both the exact and the approximate likelihood functions are convex and differentiable. From left to right: exact negative log-likelihood, pseudolikelihood (MPLE), v-acyclic composite likelihood (MCLE), and criss-cross likelihood (MXXLE).

3. *in $\mathcal{F}'$ at least two factors have been removed from each cycle in G.*

The first two properties merely state that G' is an acyclic subgraph of G. The last property is visualized in Figure 7. Assume we have somehow obtained a v-

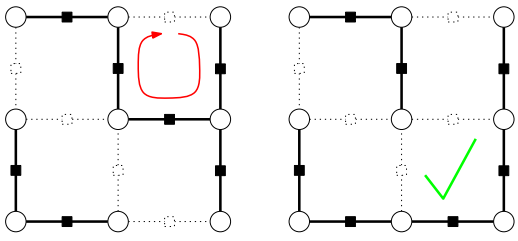

Figure 7. Two decompositions: the left decomposition is not v-acyclic because only one factor has been removed from the 4-cycle. The right decomposition is v-acyclic.

acyclic subgraph for a given input graph. Then we can use the connected components within this subgraph to define a composite likelihood: let A_j be the set of vertices contained in the j'th component and $B_j = V \setminus A_j$. Set $\beta_j = 1$. By constructing the components from the v-acyclic subgraph we automatically satisfy the conditional distribution assumption and therefore our composite likelihood produces a consistent estimator. The larger the components A_j, the more interactions between variables in $p(\boldsymbol{y})$ are retained, leading to a more informative composite likelihood (6). We now formalize the problem of how to obtain a v-acyclic subgraph with large components and then propose a greedy algorithm.

Problem 1 (Maximum v-acyclic subgraph) *Given a factor graph $G = (V, \mathcal{F}, \mathcal{E})$ and a weighting function $v : \mathcal{F} \to \mathbb{R}$, find*

$$G^*_{vac} = \underset{G' \subseteq_{vac} G}{argmax} \sum_{F \in \mathcal{F}} [\![F \in \mathcal{F}']\!] v(F),$$

where $[\![\cdot]\!]$ is one if its argument is true, zero otherwise.

The above problem has a straightforward intuitive meaning: find a v-acyclic subgraph that retains as

many factors as possible, as measured by the weighting function.

We tried three different algorithms to solve practical instances of the maximum v-acyclic subgraph problem; in the end we found the following simple greedy algorithm is competitive: 1. start with an empty subgraph (no factors) and order all factors by their weights in decreasing order, 2. iteratively select the factor with the largest weight and check whether adding it would violate v-acyclicity; if it does, discard the factor. If it does not, add the factor. Figure 8 shows a decomposition obtained by this algorithm on a real instance.

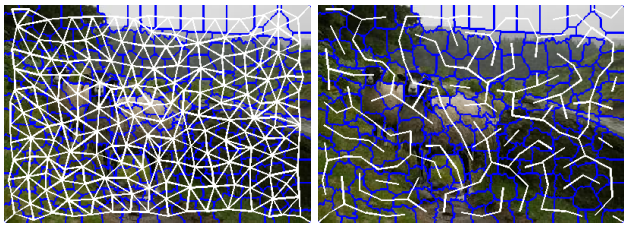

Figure 8. The superpixel segmentation of the image is used to define a pairwise graphical model, whose edges are shown in white (left). A *v-acyclic decomposition* of this graph, obtained by our algorithm, is shown (right). We use it to define the tractable and consistent composite likelihood approximation. Note that here the absence of an edge in the decomposition does not mean the interaction is ignored but that it is approximated. This is in contrast to popular piecewise training approaches (Sutton & McCallum, 2005; Pletscher et al., 2009; Alahari et al., 2010).

When learning the parameters of a model we do not know reasonable weights $v(F)$ to assign to each factor. In that case we simply retain as many factors as possible by setting all weights to one.

We now evaluate our two proposed estimators MXXLE for grid graphs and MCLE for general graphs.

5. Experiments and Results

We first revisit the simple experiment we used earlier. As a second experiment we use a more difficult synthetic task.

5.1. Revisiting the Simple Experiment

The experiment of Section 3 showed that good approximate inference does not automatically lead to good approximate estimators. Figure 6 shows the results of pseudolikelihood, our v-acyclic composite likelihood, and the criss-cross likelihood evaluated on the same experiment.

On this small experiment all three estimators do not exhibit the problems encountered earlier. In particular, the approximate negative log-likelihood functions are continuous, differentiable, and convex, and the estimated parameter vector is accurate.

5.2. Synthetic Estimation Efficiency Experiment

To understand the estimation accuracy of our proposed estimator we perform the following experiment on synthetic data. We create a 5-by-5 synthetic grid graph with 4-neighborhood connectivity. Each variable takes one of three states $\{1, 2, 3\}$ and there are no unary energy terms. The pairwise energies $E_{i,j}(y_i, y_j) = w_{i,j}(y_i, y_j) = w_{i,j}(y_j, y_i)$ are specific to each edge and symmetric; we sample $w_{i,j}(y_i, y_j) \sim \mathcal{N}(0, 1/6)$, but set $w_{i,j}(1, 1) = 0$ to ensure the parameters are identifiable. For this fixed model, we obtain a set of N samples from the model using careful Gibbs sampling.[3] From the sample set we attempt to recover the generating parameters $\hat{\boldsymbol{w}}$.

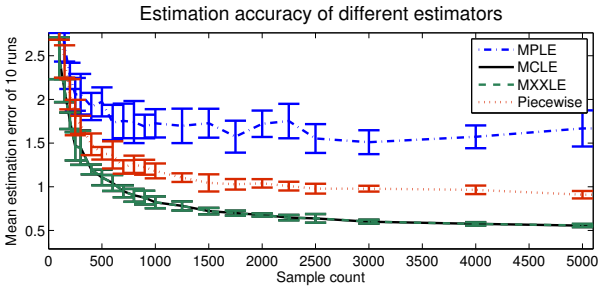

Figure 9. Estimation accuracy $\|\boldsymbol{w}_n^* - \hat{\boldsymbol{w}}\|/\|\hat{\boldsymbol{w}}\|$ of different estimators for a synthetic grid graph with random interactions. The sample count ranges from 50 to 5000 and the plots show the average and standard deviation of 10 replications for the same model.

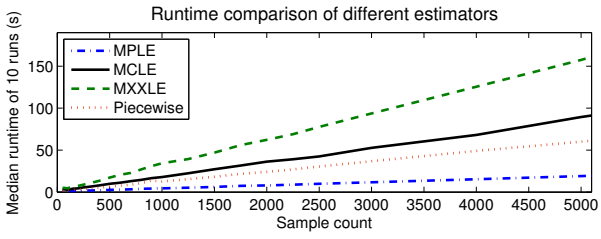

Figure 10. Estimator runtime to achieve an objective function gradient norm of 10^{-6} or smaller.

Figure 9 and 10 show the estimation accuracy and

[3] We use a single-site random sweep Gibbs sampler with 5000 burn-in sweeps and an inter-sample spacing of 100 sweeps.

runtime of our proposed MCLE and MXXLE estimators versus two popularly used estimators: pseudo-likelihood (MPLE) (Besag, 1977) and naive piecewise (NPW) (Sutton & McCallum, 2005) estimators. The differences between MCLE and MXXLE are very small, but both estimators show a clear improvement in the accuracy of recovering $\hat{\boldsymbol{w}}$ over the MPLE and NPW estimators. The MPLE estimator has a high variance. Regarding runtime MPLE is the fastest, MCLE requires about the same time as the piecewise estimator, and MXXLE is twice as expensive again.

6. Conclusion

We introduced a new method to construct composite likelihood estimators for general discrete graphical models. Our method is simple and deterministic and it combines favorable properties such as consistency and convexity.

There are two limitations of the proposed method: it does not apply to densely connected models, or to latent variable models. In densely connected models the proposed v-acyclic decomposition will eventually degenerate to the pseudo-likelihood approximation because there exist no v-acyclic decompositions except for the trivial one of removing all pairwise and higher-order factors. For latent variable models where we learn from partially observed instances the likelihood approximations no longer work; it is unclear whether composite likelihoods can be extended to this case, and what properties would continue to hold. For a recent discussion, see (Varin et al., 2010).

As models for many high-level computer vision tasks become more sophisticated and data-driven we believe that progress in parameter estimation is essential for these rich models to be successful. While there is no shortage of published estimation methods, to the best of our knowledge there are only very limited comparative studies examining their effectiveness.

The proposed estimators are available in the *grante* library at `http://www.nowozin.net/sebastian/grante/`. The library also includes a more recent exact solver for the maximum v-acyclic subgraph problem.

Acknowledgments. The author would like to thank Andrew Fitzgibbon for improving the writing of the paper and general feedback on the work. The author would also like to thank Jeremy Jancsary for discussing approximate likelihood methods.

References

Alahari, Karteek, Russell, Chris, and Torr, Phil H.S. Efficient piecewise learning for conditional random fields. In *CVPR*, 2010.

Asuncion, Arthur U., Liu, Qiang, Ihler, Alexander T., and Smyth, Padhraic. Learning in blocks: Composite likelihood and contrastive divergence. In *AISTATS*, 2010.

Besag, Julian. Efficiency of pseudolikelihood estimation for simple Gaussian fields. *Biometrica*, (64): 616–618, 1977.

Bouchard-Côté, Alexandre and Jordan, Michael I. Optimization of structured mean field objectives. In *UAI*, 2009.

Carreira-Perpiñán, Miguel Á. and Hinton, Geoffrey E. On contrastive divergence learning. In *AISTATS*, 2005.

Descombes, Xavier, Morris, Robin D., Zerubia, Josiane, and Berthod, Marc. Estimation of Markov random field prior parameters using Markov chain Monte Carlo maximum likelihood. *IEEE Transactions on Image Processing*, 8(7):954–963, 1999.

Dillon, Joshua and Lebanon, Guy. Statistical and computational tradeoffs in stochastic composite likelihood. In *AISTATS*, 2009.

Domke, Justin. Learning graphical model parameters with approximate marginal inference. *TPAMI*, pp. 1–1, 2013.

He, Xuming, Zemel, Richard S., and Carreira-Perpiñán, Miguel Á. Multiscale conditional random fields for image labeling. In *CVPR*, 2004.

Hinton, Geoffrey E. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

Hyvärinen, Aapo. Estimation of non-normalized statistical models by score matching. *JMLR*, 6:695–709, 2005. URL `http://www.jmlr.org/papers/v6/hyvarinen05a.html`.

Koller, Daphne and Friedman, Nir. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

Kschischang, Frank R., Frey, Brendan J., and Loeliger, Hans-Andrea. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2):498–519, February 2001.

Lafferty, John, McCallum, Andrew, and Pereira, F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001.

Liang, Percy and Jordan, Michael I. An asymptotic analysis of generative, discriminative, and pseudo-likelihood estimators. In *ICML*, 2008.

Lindsay, Bruce G. Composite likelihood methods. *Contemporary Mathematics*, 80, 1988.

McAllester, David. Generalization bounds and consistency for structured labeling. In Bakır, G., Hofmann, T., Schölkopf, B., Smola, A., Taskar, B., and Vishwanathan, S.V.N. (eds.), *Predicting Structured Data*. 2007.

Nowozin, Sebastian and Lampert, Christoph H. Structured learning and prediction in computer vision. *Foundations and Trends in Computer Graphics and Vision*, 6(3-4):185–365, 2011.

Pletscher, Patrick, Ong, Cheng Soon, and Buhmann, Joachim M. Spanning tree approximations for conditional random fields. In *AISTATS*, 2009.

Pletscher, Patrick, Nowozin, Sebastian, Kohli, Pushmeet, and Rother, Carsten. Putting MAP back on the map. In *DAGM*, 2011.

Saul, Lawrence K. and Jordan, Michael I. Exploiting tractable substructures in intractable networks. In *NIPS*, 1995.

Shotton, Jamie, Winn, John, Rother, Carsten, and Criminisi, Antonio. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *IJCV*, 81(1), January 2007.

Sutton, Charles and McCallum, Andrew. An introduction to conditional random fields for relational learning. In *Introduction to Statistical Relational Learning*, chapter 4. 2007a. URL http://www.cs.berkeley.edu/~casutton/publications/crf-tutorial.pdf.

Sutton, Charles A. and McCallum, Andrew. Piecewise training for undirected models. In *UAI*, pp. 568–575, 2005.

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

Tsochantaridis, Ioannis, Joachims, Thorsten, Hofmann, Thomas, and Altun, Yasemin. Large margin methods for structured and interdependent output variables. *JMLR*, 6:1453–1484, September 2005.

Varin, Cristiano, Reid, Nancy, and Firth, David. An overview of composite likelihood methods. *Statistica Sinica*, 2010.

Vickrey, David, Lin, Cliff Chiung-Yu, and Koller, Daphne. Non-local contrastive objectives. In *ICML*, 2010. URL http://www.icml2010.org/papers/592.pdf.

Wainwright, Martin J. and Jordan, Michael I. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2), 2008.

White, Halbert. Maximum-likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25, 1982.