# Part & Clamp: Efficient Structured Output Learning

**Patrick Pletscher**
ETH Zurich
Zürich, Switzerland

**Cheng Soon Ong**
NICTA
Melbourne, Australia

## Abstract

Discriminative training for general graphical models is a challenging task, due to the intractability of the partition function. We propose a computationally efficient approach to estimate the partition sum in a structured learning problem. The key idea is a lower bound of the partition sum that can be evaluated in a fixed number of message passing iterations. The bound makes use of a subset of the variables, a feedback vertex set, which allows us to decompose the graph into tractable parts. Furthermore, a tightening strategy for the bound is presented, which finds the states of the feedback vertex set that maximally increase the bound, and clamps them. Based on this lower bound we derive batch and online learning algorithms and demonstrate their effectiveness on a computer vision problem.

## 1 Introduction

Discriminative structured output prediction (Bakir et al., 2007) is a popular approach for classification tasks with interdependent variables. Applications include multi-class and multi-label classification problems, gene finding in bioinformatics, object recognition in computer vision and part-of-speech tagging in natural language processing. In many computer vision tasks, the underlying structure is a graphical model containing a large number of loops, rendering inference and learning intractable. This work considers the problem of learning the parameters of the graphical model and formulates a novel lower bound on the structured output loss. The evaluation of the bound typically requires only a few iterations of a modified message passing algorithm, where the number of iterations is fixed and dependent on the specified budget. Our approach consists of two parts: First, a subset of the output variables is selected, a so called *feedback vertex set* (FVS), with the property that any cycle in the graph contains at least one variable in the FVS. Second, a conditioned partition sum for one state of the variables in the FVS is repeatedly computed for a few low-energy states, to successively tighten the lower bound. Each individual computation requires two message passes. We show that this lower bound is well-suited for structured output learning, especially in an online scenario.

The contributions of this paper are as follows: First, we propose a generalization of composite likelihood for computing a lower approximation of the structured partition sum and formulate a tightening strategy. We show that composite likelihood is a specific instance of this framework. Second, we introduce a forest decomposition and formulate it as a minimal feedback vertex set (FVS) problem. Third, a variational algorithm, MAX-TIGHTEN is introduced. The algorithm finds the states of the FVS which maximally increase the lower bound. We introduce batch and online algorithms for learning with the lower bound. The performance of the algorithms is demonstrated on a computer vision dataset.

## 2 Structured Output Prediction

In structured output prediction, the task is predicting interdependent output variables $\boldsymbol{y} \in \mathcal{Y}$ for a given input variable $\boldsymbol{x} \in \mathcal{X}$. An individual output variable $y_i$ has a discrete and finite output domain $\mathcal{Y}_i$. The dependencies between the variables are assumed to be specified by a parametrized factor graph; the parameters are denoted by $\boldsymbol{w}$. Assuming linearity of the parametrization, the score (i.e., negative energy) of an input/output configuration $(\boldsymbol{x}, \boldsymbol{y})$ can be written as an inner product $\langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y}) \rangle$. Here $\boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y}) : \mathcal{X} \times \mathcal{Y} \to \mathcal{H}$ is a mapping of the variables to a joint input/output space in correspondence to the factor graph. The map-

Appearing in Proceedings of the $15^{th}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume 22 of JMLR: W&CP 22. 

ping $\boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y})$ can also be thought of as sufficient statistics of the model. The energy $E(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{w})$ for a given input/output pair can be written as

$$\begin{aligned} E(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{w}) &= -\langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y}) \rangle \\ &= -\sum_{t \in \mathcal{T}} \sum_{\alpha \in \mathcal{C}(t)} \langle \boldsymbol{w}^t, \boldsymbol{\phi}_t(\boldsymbol{x}, \boldsymbol{y}_\alpha, \alpha) \rangle \\ &= \sum_{t \in \mathcal{T}} \sum_{\alpha \in \mathcal{C}(t)} E_\alpha(\boldsymbol{y}_\alpha, \boldsymbol{x}, \boldsymbol{w}^t) \end{aligned} \tag{1}$$

Here, $t$ runs over potentials that share the same parameter (factor templates $\mathcal{T}$) and $\alpha$ runs over the different factors of the factor graph $\mathcal{C}$. This work considers pairwise models, thus the factor graph can be represented as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$[1]. Finally, $E_\alpha$ denotes the energy of factor $\alpha$.

We review estimation of the parameters in Section 2.1 and prediction for such models in Section 2.2. These tasks are computationally intractable for general loopy graphical models due to the partition sum.

### 2.1 Learning

We consider the task of learning the parameters $\boldsymbol{w}$ of such a model from a given training set $\{(\boldsymbol{x}^n, \boldsymbol{y}^n)\}_{n=1}^N$. For discriminative models, the two dominant approaches in the literature are the Conditional Random Field (Lafferty et al., 2001) (CRF) and the Structured Support Vector Machine (Tsochantaridis et al., 2005; Taskar et al., 2003). The CRF is a log-linear model for the posterior distribution over outputs given an input:

$$P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}) = \frac{1}{Z(\boldsymbol{x}, \boldsymbol{w})} \exp(\langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y}) \rangle), \tag{2}$$

$$Z(\boldsymbol{x}, \boldsymbol{w}) = \sum_{\boldsymbol{y} \in \mathcal{Y}} \exp(\langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y}) \rangle). \tag{3}$$

The parameters $\boldsymbol{w}$ are generally estimated using maximum likelihood, and the main computational burden is in evaluating the partition sum $Z(\boldsymbol{x}, \boldsymbol{w})$. This is due to the fact that $\mathcal{Y}$ is an exponentially large set.

Taking the negative logarithm of the likelihood and including an $\ell_2$ regularizer, yields the structured output learning problem:

$$\min_{\boldsymbol{w}} \frac{1}{N} \sum_{n=1}^{N} -\langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x}^n, \boldsymbol{y}^n) \rangle + \log Z(\boldsymbol{x}^n, \boldsymbol{w}) + \frac{\lambda}{2} \|\boldsymbol{w}\|^2. \tag{4}$$

In this work we focus on accurate and efficient parameter estimation for CRFs. Nevertheless, the results are also relevant for the structured SVM, as recent work (Pletscher et al., 2010) shows deep connections between the two learning problems.

[1] We consider only the output variables to be part of the graphical model. The input variables are observed, and thus can always be absorbed into the factors.

### 2.2 Prediction

For a given parameter $\boldsymbol{w}$ and an error function $\Delta_{\boldsymbol{y}'}(\boldsymbol{y})$, which measures the loss incurred by predicting $\boldsymbol{y}$ instead of $\boldsymbol{y}'$, Bayesian decision theory predicts output variables $\boldsymbol{y}^\star$ according to

$$\boldsymbol{y}^\star = \operatorname*{argmin}_{\boldsymbol{y}} \sum_{\boldsymbol{y}'} \Delta_{\boldsymbol{y}'}(\boldsymbol{y}) P(\boldsymbol{y}'|\boldsymbol{x}, \boldsymbol{w}). \tag{5}$$

This is the best predictor possible for a given $\Delta$ under the assumption that the learned distribution is equal to the true underlying model. Taking $\Delta$ to be the zero-one loss on the full output, results in the maximum-a-posterior (MAP) predictor $\boldsymbol{y}^* = \operatorname{argmax}_{\boldsymbol{y}} P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w})$, whereas a zero-one loss on the individual output variables leads to the maximum posterior marginal (MPM) predictor $y_i^* = \operatorname{argmax}_{y_i} P(y_i|\boldsymbol{x}, \boldsymbol{w})$.

## 3 Composite Likelihood and Contrastive Divergence

Composite likelihood is a common approach for approximate parameter estimation in CRFs. Let $\mathcal{V}$ be the set of output variables. Furthermore, let $(\mathcal{A}, \mathcal{B})$ be a partition of $\mathcal{V}$ into two sets (i.e., $\mathcal{A} \cup \mathcal{B} = \mathcal{V}$ and $\mathcal{A} \cap \mathcal{B} = \emptyset$). In composite likelihood[2] (Lindsay, 1988) the intractable full likelihood in (2) is approximated by the conditional distribution of the variables in $\mathcal{A}$ given the ground-truth state of the variables in $\mathcal{B}$. Several decompositions $\{(\mathcal{A}_m, \mathcal{B}_m)\}_{m=1}^M$ are combined by multiplying their respective conditional likelihoods. The maximum composite likelihood estimator is then computed as:

$$\min_{\boldsymbol{w}} \frac{1}{N} \sum_{n=1}^{N} \frac{1}{M} \sum_{m=1}^{M} -\log P(\boldsymbol{y}^n_{\mathcal{A}_m} | \boldsymbol{x}^n, \boldsymbol{y}^n_{\mathcal{B}_m}, \boldsymbol{w}) + \frac{\lambda}{2} \|\boldsymbol{w}\|^2. \tag{6}$$

Note that in (6) the form of the decomposition are left unspecified. Pseudolikelihood (Besag, 1975) is a special case of composite likelihood that assumes particularly simple decompositions given by $\mathcal{A}_m = \{m\}, \mathcal{B} = \mathcal{V} \backslash \{m\}$ with $m$ running over all the output variables in $\mathcal{V}$. Dillon & Lebanon (2010) recently introduced a stochastic version of composite likelihood, where the decompositions are chosen stochastically. This allows for computationally more demanding decompositions, such as a small cyclic subgraph, to be included with some probability. Despite its simplicity, composite likelihood was shown to be consistent (Lindsay, 1988). Under weak regularity conditions, the composite likelihood estimate asymptotically converges to the maximum-likelihood estimate.

[2] We restrict ourselves to the *conditional* composite likelihood; the marginal conditional likelihood is not discussed.

An alternative method for approximate parameter learning is contrastive divergence (Hinton, 2000): A Markov Chain Monte Carlo (MCMC) sampler is run for few iterations (typically around five iterations) to compute an approximate gradient of (4). Stochastic gradient descent is then used for the minimization. The trick is that the MCMC sampler is initialized with the ground-truth label. Asuncion et al. (2010) point out the similarities between contrastive divergence and composite likelihood. While contrastive divergence gives satisfying results in practice, it has been shown that it does not converge in general (Sutskever & Tieleman, 2010).

In (Vickrey et al., 2010) a non-local contrastive divergence is introduced. Low energy configurations are computed using an approximate MAP inference algorithm. An approximation of the partition sum is then obtained by adding up the contributions of the generated states. Non-local contrastive divergence is shown to be consistent.

## 4 Lower Bounding the Structured Output Loss

In this section we introduce an extension of composite likelihood which can be understood as a lower bound of the partition sum. Given a partition of $\mathcal{V}$ into two sets, $\mathcal{A}$ and $\mathcal{B}$, the partition sum in (3) can be decomposed into two sums running over the states of the variables in $\mathcal{A}$ and $\mathcal{B}$. A trivial lower-bound is obtained by summing over only a (small) subset $\underline{\mathcal{Y}}_\mathcal{B} \subseteq \mathcal{Y}_\mathcal{B}$ of the large state space $\mathcal{Y}_\mathcal{B}$:

$$\begin{aligned} Z(\boldsymbol{x}, \boldsymbol{w}) &= \sum_{\boldsymbol{y}_\mathcal{B} \in \mathcal{Y}_\mathcal{B}} \sum_{\boldsymbol{y}_\mathcal{A} \in \mathcal{Y}_\mathcal{A}} \exp(\langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y}) \rangle) \\ &\geq \sum_{\boldsymbol{y}_\mathcal{B} \in \underline{\mathcal{Y}}_\mathcal{B}} \sum_{\boldsymbol{y}_\mathcal{A} \in \mathcal{Y}_\mathcal{A}} \exp(\langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y}) \rangle) \\ &=: Z(\boldsymbol{x}, \boldsymbol{w}, \mathcal{B}, \underline{\mathcal{Y}}_\mathcal{B}). \end{aligned} \tag{7}$$

The set $\mathcal{Y}_\mathcal{B}$ contains all possible states of the variables in $\mathcal{B}$ and is therefore exponentially large. Using a subset $\underline{\mathcal{Y}}_\mathcal{B}$ may result in a relatively poor approximation for high-entropy distributions. However, as we will show, for parameter learning this simple approach can be very effective. The choice of the decomposition, $(\mathcal{A}, \mathcal{B})$ as well as the states in the set $\underline{\mathcal{Y}}_\mathcal{B}$ are discussed in detail in Section 5.1 and Section 5.2 respectively. Actual learning algorithms are given in Section 5.3. The remainder of this section discusses extensions of the lower bound and its connection to previous work.

### 4.1 Several Decompositions

To decrease the effects of poor decompositions, several partitions $\mathcal{D} = \{(\mathcal{A}_1, \mathcal{B}_1), \dots, (\mathcal{A}_M, \mathcal{B}_M)\}$ and corresponding states $\mathcal{Z} = \{\underline{\mathcal{Y}}_{\mathcal{B}_1}, \dots, \underline{\mathcal{Y}}_{\mathcal{B}_M}\}$ can be combined. Let $Z^m := Z(\boldsymbol{x}, \boldsymbol{w}, \mathcal{B}_m, \underline{\mathcal{Y}}_{\mathcal{B}_m})$. The arithmetic and geometric mean as well as the maximum of all the bounds are also valid lower bounds:

$$Z^{\mathrm{a},\mathcal{D},\mathcal{Z}}(\boldsymbol{x}, \boldsymbol{w}) := \frac{1}{M} \sum_{m=1}^{M} Z^m, \tag{8}$$

$$Z^{\mathrm{g},\mathcal{D},\mathcal{Z}}(\boldsymbol{x}, \boldsymbol{w}) := \left( \prod_{m=1}^{M} Z^m \right)^{1/M}, \tag{9}$$

$$Z^{\mathrm{m},\mathcal{D},\mathcal{Z}}(\boldsymbol{x}, \boldsymbol{w}) := \max_m Z^m. \tag{10}$$

The maximum over the different decompositions results in the tightest bound, but has the disadvantage of being non-differentiable w.r.t. the parameters due to the maximum function. This can cause problems when minimizing (4) using quasi-Newton solvers that rely on the smoothness of the objective. The geometric mean is commonly used by composite likelihood approaches, and is obtained by considering the arithmetic average of the log partition sum. This has the advantage that it is smooth. However, the arithmetic mean actually provides a tighter lower bound while maintaining differentiability.

**Lemma 1.** *The relation between the different lower bound combinations is*

$$Z^{m,\mathcal{D},\mathcal{Z}}(\boldsymbol{x}, \boldsymbol{w}) \geq Z^{a,\mathcal{D},\mathcal{Z}}(\boldsymbol{x}, \boldsymbol{w}) \geq Z^{g,\mathcal{D},\mathcal{Z}}(\boldsymbol{x}, \boldsymbol{w}). \tag{11}$$

*Furthermore, all combination approaches lead to lower bounds on $Z(\boldsymbol{x}, \boldsymbol{w})$.*

### 4.2 Connection to Composite Likelihood

The following lemma draws the connection between the lower bound in (7) and composite likelihood.

**Lemma 2.** *Composite likelihood learning with decompositions $\mathcal{D}$ is equivalent to lower bounding the partition sum in (4) for each example $(\boldsymbol{x}^n, \boldsymbol{y}^n)$ by the geometric average lower bound $Z^{g,\mathcal{D},\mathcal{Z}^n}(\boldsymbol{x}^n, \boldsymbol{w})$ with $\underline{\mathcal{Y}}^n_{\mathcal{B}_m} = \{\boldsymbol{y}^n_{\mathcal{B}_m}\}$ and $\mathcal{Z}^n = \{\underline{\mathcal{Y}}^n_{\mathcal{B}_1}, \dots, \underline{\mathcal{Y}}^n_{\mathcal{B}_M}\}$.*

The lower bound in our work differs in two key aspects from the classic composite likelihood: First, we give a concrete choice of the decompositions by feedback-vertex sets, balancing computational tractability and accuracy of the estimator. Second, in addition to the ground-truth, several other low energy states of the variables in $\mathcal{B}$ are used. As will be shown in the experiments, especially the second contribution improves the results drastically.

### 4.3 Asymptotic Consistency

The bound in (7) can be understood as an efficient extension of the non-local contrastive objective introduced in (Vickrey et al., 2010). In their work a learning

objective is formulated where the partition sum is replaced by an exhaustive enumeration over low energy states which are computed using MAP inference. Contrary to our work, the non-local contrastive objective does however not consider the decomposition of the partition sum into two parts. Therefore much more states need to be considered explicitly. Nevertheless, the proof of asymptotic convergence in (Vickrey et al., 2010) also generalizes to our lower bound (when the geometric combination is used) subject to the same regularity assumptions.

### 4.4 Comparison to Upper Bounds

Finally, in contrast to our lower bound a considerable amount of work has investigated upper bounding the partition sum. Upper bounds are generally obtained using variational inference. Such an example is the tree reweighted belief propagation (Wainwright et al., 2002). A problem of learning with upper bounds such as (Hazan & Urtasun, 2010; Wainwright et al., 2002; Meshi et al., 2010) arises from the fact that convergence of the message passing algorithms used to compute the upper bound is generally slow, or sometimes not even guaranteed.

The general perception in the field is that upper bounds are superior to lower bounds, as intuitively less things can go wrong when minimizing an upper bound. This view is also supported by (Finley & Joachims, 2008). Our work questions this belief by showing that composite likelihood is in fact a lower bound. Furthermore, we give experimental support that our lower bound leads to state-of-the-art accuracy on a well-studied data set.

### 4.5 Connection to Cutset Conditioning

Our work is related to the relatively old idea of cutset conditioning in Bayesian networks, dating back to Pearl (1990), see also (Koller & Friedman, 2009, Section 9.5). For the task of probabilistic inference, a subset of the nodes is exhaustively enumerated over, whereas for the remaining variables a sum-product algorithm is used. We use a very similar idea in the next section. Cutset conditioning is generalized in (Horvitz et al., 1989) for approximate inference. However, to the best of our knowledge, cutset conditioning has not been used for learning undirected models, nor have its connection to composite likelihood been explored.

## 5 Part & Clamp

This section describes the details of the main parts of our proposed *part & clamp* algorithm.

### 5.1 Part: Finding a Minimum Feedback Vertex Set

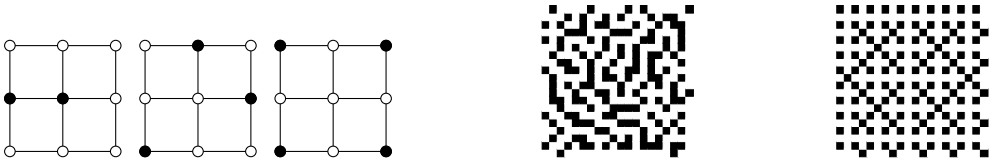
The lower bound in (7) is valid for any choice of partition $\mathcal{A}, \mathcal{B}$. For tractable computations, we consider $\mathcal{B}$ such that all loops in the graph $\mathcal{G}$ are blocked by at least one variable in $\mathcal{B}$. Such a subset of the output variables is called a feedback vertex set (Vazirani, 2001). To make this property explicit, we will use $\mathcal{F}$ to denote such a set $\mathcal{B}$ and $\mathcal{V}\backslash\mathcal{F}$ to denote its complement. Due to the FVS property (see Figure 1), $\mathcal{V}\backslash\mathcal{F}$ is a forest and hence conditioned on the state of the FVS, $\boldsymbol{y}_{\mathcal{F}}$, summation over $\mathcal{Y}_{\mathcal{V}\backslash\mathcal{F}}$ for the remaining variables, can be carried out exactly using the sum-product algorithm.

The decision variant of the minimal FVS is an NP-hard problem (Vazirani, 2001) and therefore one has to resort to approximate algorithms. In our work we consider the unweighted version, where the number of variables $|\mathcal{F}|$ in the FVS is minimized. This can be motivated by the principle of insufficient reason: all the variables are assumed to have the same contribution to the partition sum. Therefore, the minimum FVS results in the lowest approximation error. More complex selection criteria based on marginal variable weights would be possible. However, in our work we keep the decomposition fixed during learning and thus an initial estimate of the parameters would need to be obtained in order to make the selection based on marginals meaningful.

**Algorithm 1** Growing forests algorithm for the feedback vertex set problem.

**Require:** $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.
$\mathcal{F} = \emptyset, \mathcal{Q} = \emptyset, \forall i \in \mathcal{V} : \text{visited}(i) = 0$.
**while** not all vertices visited **do**
  Choose $i$ at random from $\{j \in \mathcal{V} : \text{visited}(j) = 0\}$.
  $i \to Q$.
  **repeat**
    $i \leftarrow Q$.
    **if** $|\{j \in \mathcal{N}(i) : \text{visited}(j) \wedge j \notin \mathcal{F}\}| \geq 2$ **then**
      $\mathcal{F} = \mathcal{F} + \{i\}$.
    **end if**
    visited$(i) = 1$.
    $\forall j \in \mathcal{N}(i) \wedge \neg\text{visited}(j) : j \to Q$ (random order).
  **until** $Q = \emptyset$.
**end while**
**return** $\mathcal{F}$

A series of papers (Becker & Geiger, 1996; Chudak et al., 1998; Bafna et al., 1999) gives 2-approximation algorithms for the minimum FVS problem. In our work we consider a simpler probabilistic algorithm (Chandrasekaran et al., 2011) based on a breadth-first exploration of the graph. The algorithm is shown in Algorithm 1; $\mathcal{N}(i)$ denotes the neighborhood of a vertex $i$. An algorithm with a depth-first exploration is



(a) example FVSs for $3 \times 3$ grid (b) DF (0.38) (c) BF (0.33)

Figure 1: Different feedback vertex sets (in black) of a grid-graph. (b) & (c) show small FVSs obtained using the algorithms described below. BF and DF denote the breadth-first and depth-first approach. We indicate the fraction of variables in the FVS in brackets. A checkerboard decomposition would have a fraction of around 0.5.

obtained by using a stack instead of the queue $\mathcal{Q}$. Generally the results with depth-first exploration were inferior to the ones obtained using breadth-first. For a grid graph the breadth-first approach leads to a close to optimal FVS ratio of around 1/3, which is in the order of the lower bound in (Luccio, 1998).

### 5.2 Clamp: Choosing the States of the Feedback Vertex Set

The set $\underline{\mathcal{Y}}_{\mathcal{F}}$ is initialized with the ground-truth label $\boldsymbol{y}^n_{\mathcal{F}}$, corresponding to the input $\boldsymbol{y}^n$ as the only state. For a given parameter $\boldsymbol{w}$ and feedback vertex set $\mathcal{F}$ let us consider the problem of finding a labeling $\boldsymbol{y}^\star_{\mathcal{F}}$ to be included in $\underline{\mathcal{Y}}_{\mathcal{F}}$. We here follow a greedy approach by including $\boldsymbol{y}^\star_{\mathcal{F}}$ to *maximally tighten* the lower bound in (7). We choose the states of the FVS according to:

$$\boldsymbol{y}^\star_{\mathcal{F}} = \operatorname*{argmax}_{\boldsymbol{y}_{\mathcal{F}} \in \mathcal{Y}_{\mathcal{F}}} \sum_{\boldsymbol{y}_{\mathcal{V}\setminus\mathcal{F}} \in \mathcal{Y}_{\mathcal{V}\setminus\mathcal{F}}} \exp(\langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y}) \rangle). \tag{12}$$

The maximization above is more complex than standard energy minimization problems arising from MAP inference. The task is sometimes described as marginal MAP (Koller & Friedman, 2009). Here, some variables $\boldsymbol{y}_{\mathcal{V}\setminus\mathcal{F}}$ are summed over and other variables $\boldsymbol{y}_{\mathcal{F}}$ are maximized over. We derive a variational approach for this problem, which we named MAX-TIGHTEN. To simplify the notation, we drop the dependence of the energy on $\boldsymbol{w}$ and $\boldsymbol{x}$. Furthermore, let $Z(\boldsymbol{y}_{\mathcal{F}})$ denote the partition sum for the FVS variables clamped to state $\boldsymbol{y}_{\mathcal{F}}$.

We follow the recent approach in (Liu & Ihler, 2011; Jiang et al., 2011) which formulates the marginal MAP as a variational problem over the marginal polytope (Wainwright & Jordan, 2008). Let us first rewrite the log partition sum for a given $\boldsymbol{y}_{\mathcal{F}}$ through its dual:

$$\begin{aligned} A(\boldsymbol{y}_{\mathcal{F}}) &:= \log \sum_{\boldsymbol{y}_{\mathcal{V}\setminus\mathcal{F}}} \exp(-E(\boldsymbol{y}_{\mathcal{F}}, \boldsymbol{y}_{\mathcal{V}\setminus\mathcal{F}})) \qquad (13) \\ &= \min_P \sum_{\boldsymbol{y}_{\mathcal{V}\setminus\mathcal{F}}} P(\boldsymbol{y}_{\mathcal{V}\setminus\mathcal{F}}|\boldsymbol{y}_{\mathcal{F}}) E(\boldsymbol{y}) \\ &\quad + \sum_{\boldsymbol{y}_{\mathcal{V}\setminus\mathcal{F}}} P(\boldsymbol{y}_{\mathcal{V}\setminus\mathcal{F}}|\boldsymbol{y}_{\mathcal{F}}) \log P(\boldsymbol{y}_{\mathcal{V}\setminus\mathcal{F}}|\boldsymbol{y}_{\mathcal{F}}) \end{aligned}$$

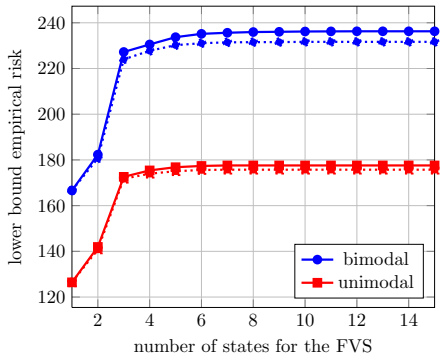
The full problem in (12) can then be rewritten as

$$\max_{\boldsymbol{y}_{\mathcal{F}}} A(\boldsymbol{y}_{\mathcal{F}}) = \min_P \sum_{\boldsymbol{y}} P(\boldsymbol{y}) E(\boldsymbol{y}) + \underbrace{\sum_{\boldsymbol{y}} P(\boldsymbol{y}) \log P(\boldsymbol{y}_{\mathcal{V}\setminus\mathcal{F}}|\boldsymbol{y}_{\mathcal{F}})}_{=:-H(\boldsymbol{y}_{\mathcal{V}\setminus\mathcal{F}}|\boldsymbol{y}_{\mathcal{F}})}. \tag{14}$$

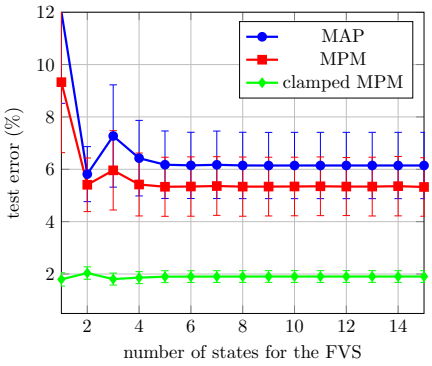
The first term in (14) is a standard average energy and the second term corresponds to the negative *conditional entropy* of $\boldsymbol{y}_{\mathcal{V}\setminus\mathcal{F}}$. Unfortunately, the variational problem above still remains intractable as the optimization problem has an exponential number of variables (or equivalently an exponential number of constraints if expressed using the marginal polytope). Furthermore, the conditional entropy does not factorize into marginals, which makes an approximation even more difficult. Both, (Liu & Ihler, 2011) and (Jiang et al., 2011) choose to approximate (14) by relaxing the constraint set to the local marginal polytope and replacing the entropy term with a unary and pairwise approximation. We here choose the approach from (Jiang et al., 2011) which reduces to a hybrid message-passing algorithm in which for variables in the FVS a max-product update is performed, whereas for the remaining nodes a sum-product update is used.

For obtaining a $\boldsymbol{y}^\star_{\mathcal{F}}$, the obtained pseudo-beliefs need to be rounded to integer values. The approach ignores the constraint that $\boldsymbol{y}^\star_{\mathcal{F}}$ should be different from all the states already in $\underline{\mathcal{Y}}_{\mathcal{F}}$ and thus in theory might generate a state that is already modeled. This could probably be improved with an approach akin to the $M$-best MAP algorithm (Fromer & Globerson, 2009), at the cost of an increased runtime. We observed empirically that states were rarely chosen repeatedly.

Alternatively, instead of using the MAX-TIGHTEN approach, one can simply compute a MAP labeling. While this approach does not guarantee the most effective tightening, it has the advantage that very efficient specialized solvers are available. In practice we have found that initializing the message-passing algorithm for MAX-TIGHTEN with a smoothed version of a MAP label consistently lead to the best results.




(a) Lower bound on the empirical risk.




(b) Prediction error on the bimodal test set.

Figure 2: Batch learning for the binary image denoising dataset. The lower bound and the prediction error are visualized when increasing $\underline{\mathcal{Y}}_{\mathcal{F}}$. We observe that the first couple of iterations are the most important. The dotted curve in (a) corresponds to tightening using MAP inference, the solid curve to MAX-TIGHTEN: The differences are more pronounced for the bimodal data set. The curves in (b) correspond to different prediction approaches using the same parameter estimate. The error bars show the standard deviation of the prediction errors. Note that clamped MPM is unrealistic (since it requires labels), and is shown to indicate the theoretical optimum.

### 5.3 Derived Learning Algorithms

Here we describe parameter learning with the proposed lower bound. The approximate objective is obtained by replacing the partition sum in (4) by $Z(\boldsymbol{x}, \boldsymbol{w}, \mathcal{F}, \underline{\mathcal{Y}}_{\mathcal{F}})$ which is defined in (7):

$$\min_{\boldsymbol{w}} \frac{1}{N} \sum_{n=1}^{N} -\langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x}^n, \boldsymbol{y}^n) \rangle + \log Z(\boldsymbol{x}, \boldsymbol{w}, \mathcal{F}, \underline{\mathcal{Y}}_{\mathcal{F}}) + \frac{\lambda}{2} \|\boldsymbol{w}\|^2. \quad (15)$$

To make learning efficient, the gradient of the approximate objective needs to be computed efficiently. The derivative of the approximate log partition sum w.r.t. $\boldsymbol{w}$ is the expected feature map. The expectation now only runs over the states in $\underline{\mathcal{Y}}_{\mathcal{F}}$. The expectation computation requires the marginals $P_\alpha(\boldsymbol{y}_\alpha | \underline{\mathcal{Y}}_{\mathcal{F}})$ of the factors. The overall gradient is therefore given by

$$\frac{\partial}{\partial \boldsymbol{w}} = \sum_{n=1}^{N} -\boldsymbol{\phi}(\boldsymbol{x}^n, \boldsymbol{y}^n) + \sum_{t \in \mathcal{T}} \sum_{\alpha \in \mathcal{C}(t)} P_\alpha(\boldsymbol{y}_\alpha | \underline{\mathcal{Y}}_{\mathcal{F}}) \phi_t(\boldsymbol{x}^n, \boldsymbol{y}_\alpha, \alpha) + \lambda \boldsymbol{w}. \quad (16)$$

All the required quantities can be computed by aggregating the results from simple sum-product runs for the different clamping configurations $\boldsymbol{y}_{\mathcal{F}} \in \underline{\mathcal{Y}}_{\mathcal{F}}$. We illustrate this for two configurations $\boldsymbol{y}^1_{\mathcal{F}}, \boldsymbol{y}^2_{\mathcal{F}}$ and the case of larger $\underline{\mathcal{Y}}_{\mathcal{F}}$ is straightforward. Given the marginals $P_\alpha(\boldsymbol{y}_\alpha | \boldsymbol{y}^1_{\mathcal{F}})$ and $P_\alpha(\boldsymbol{y}_\alpha | \boldsymbol{y}^2_{\mathcal{F}})$ for the two clamping states and the corresponding partition sums $Z^1, Z^2$, the combined quantities are obtained as follows:

$$\begin{aligned} Z^{1,2} &= Z^1 + Z^2, \\ P_\alpha(\boldsymbol{y}_\alpha | \underline{\mathcal{Y}}_{\mathcal{F}}) &= \frac{Z^1_\alpha(\boldsymbol{y}_\alpha)}{Z^{1,2}} + \frac{Z^2_\alpha(\boldsymbol{y}_\alpha)}{Z^{1,2}}, \\ Z^k_\alpha(\boldsymbol{y}_\alpha) &= Z^k P_\alpha(\boldsymbol{y}_\alpha | \boldsymbol{y}^k_{\mathcal{F}}) \quad \text{for } k \in \{1, 2\}. \end{aligned} \quad (17)$$

As described in Section 4.1, several decompositions can be used using different combination approaches. The marginal computations for the maximum and geometric average combinations are simple (marginals of the decomposition with the maximum value and an average of the different marginals, respectively). The arithmetic average combination turns out to be a partition sum computation of a special form, which is however also tractable.

**Batch Learning** Batch learning for the proposed lower bound computes $M$ decompositions of the graphical model for each example. The set $\underline{\mathcal{Y}}^n_{\mathcal{F}}$ for each of the examples $\boldsymbol{x}^n$ is initialized with the ground-truth label $\boldsymbol{y}^n_{\mathcal{F}}$. The resulting bound is then minimized using L-BFGS leading to a first parameter estimate. For this parameter, a tightening operation using MAX-TIGHTEN is performed for each decomposition and example. The tightening of the lower bound is followed by a minimization w.r.t. the parameters. This is repeated until convergence. The batch algorithm has close relationships to the cutting planes algorithm employed in the training of structured SVMs (Tsochantaridis et al., 2005).

**Online Learning** In order to solve large scale problems, we propose an online learning version of the algorithm using stochastic gradient descent (SGD). SGD

| | Prediction \ Train | Pseudo-likelihood | Composite likelihood | Contrastive divergence | Part & Clamp batch | Part & Clamp online |
|---|---|---|---|---|---|---|
| bimodal | MAP | $15.58 \pm 4.11$ | $12.02 \pm 3.50$ | $7.01 \pm 1.71$ | $6.14 \pm 1.27$ | $5.16 \pm 0.77$ |
| | MPM | $11.86 \pm 3.40$ | $9.33 \pm 2.69$ | $6.72 \pm 1.67$ | $5.32 \pm 1.12$ | $5.20 \pm 0.80$ |
| | clamped MPM | $1.77 \pm 0.25$ | $1.80 \pm 0.26$ | $1.96 \pm 0.22$ | $1.90 \pm 0.22$ | $2.23 \pm 0.25$ |
| unimodal | MAP | $5.28 \pm 1.47$ | $4.43 \pm 1.26$ | $2.39 \pm 0.47$ | $2.40 \pm 0.50$ | $2.40 \pm 0.46$ |
| | MPM | $4.13 \pm 1.18$ | $3.66 \pm 0.96$ | $2.37 \pm 0.45$ | $2.40 \pm 0.42$ | $2.42 \pm 0.43$ |
| | clamped MPM | $0.98 \pm 0.22$ | $1.01 \pm 0.21$ | $1.05 \pm 0.21$ | $1.03 \pm 0.22$ | $1.17 \pm 0.23$ |

Table 1: Test error for learning on the binary image denoising dataset. A single FVS decomposition per example is used, i.e. $M = 1$. Composite likelihood refers to the first iteration of the Part&Clamp algorithm (i.e. only the ground-truth state is used in $\underline{\mathcal{Y}}_{\mathcal{F}}$). For the SGD updates in contrastive divergence and the online variant of Part&Clamp, all the images were considered for a single update step to simplify the comparison to the batch learning algorithms. For contrastive divergence 5 Gibbs iterations were used, for the online version of Part&Clamp a budget of two labels, i.e. $|\underline{\mathcal{Y}}_{\mathcal{F}}| = 2$.

evaluates the loss for a subset of examples and takes a step in the direction of the gradient. In our implementation, a budget is specified for each example, this budget corresponds to the number of states in $\underline{\mathcal{Y}}_{\mathcal{F}}$. At each iteration the lower bound is tightened using MAX-TIGHTEN and in case the budget is exceeded, the highest energy state is pruned from $\underline{\mathcal{Y}}_{\mathcal{F}}$ (the ground-truth state is however never removed). Followed by an evaluation of the lower bound and its derivative.

## 6 Experiments

We evaluate the performance of the proposed approach on the application of binary image denoising. We use the dataset in (Kumar & Hebert, 2006) and follow their experimental settings. The dataset consists of two noise scenarios: a unimodal and bimodal noise model. 10 images are used for training and 150 images for testing. The graphical model is given by the standard four connected grid of size $64 \times 64$. In all of the experiments we use $\lambda = 1$, we observed little change when varying the regularization parameter. Figure 2 shows the development of the lower bound for batch learning. As expected, the lower bound increases as more labels are added to $\underline{\mathcal{Y}}_{\mathcal{F}}$. Furthermore, the more states considered for the FVS, the better the test error. Comparing the curves for the unimodal and bimodal noise datasets, we notice that the lower bound flattens off after fewer iterations of the algorithm in the easier unimodal dataset. For MAP prediction we used sequential tree-reweighted belief propagation (TRWS) (Kolmogorov, 2006) and for MPM we used Gibbs sampling (with 1000 sweeps).

Table 1 shows the image denoising results obtained using different learning and prediction algorithms. As expected, MPM prediction outperforms MAP prediction. We also include a prediction version ('clamped MPM') where the minimal FVS algorithm is used to find a FVS, the state of which is clamped to its *true value*. Only for the variables in $\mathcal{V} \backslash \mathcal{F}$ a label is predicted. We report the test error of the clamped MPM to give a rough estimate of the prediction error underlying the surrogate loss used in composite likelihood training. As can bee seen, the Part&Clamp approach, in both its online and batch version, performs well and does not exhibit the poor behavior of contrastive divergence on the more difficult bimodal dataset. The online version of Part&Clamp performs better than the batch version, as it seems to overfit less to the training data (the train error is however worse). The results improve slightly on (Kumar & Hebert, 2006) where a regularization heuristic for pseudo-likelihood was used. (Hazan & Urtasun, 2010) studies a different setting, which should have a diminishing test error as the output label is always the same in training and testing and the parametrization is powerful enough to simply remember this particular output label.

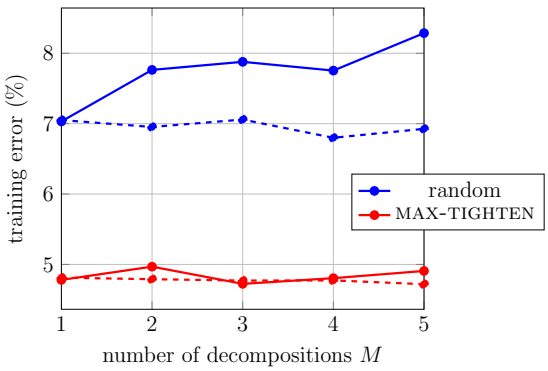



Figure 3: Training error for different number of decompositions for a clamping set of size five. We visualize the results when either using the max-tighten algorithm for tightening (red) or a random state for clamping (blue). The results with the geometric mean (solid line) are less robust w.r.t. the decompositions than the arithmetic mean (dashed line) for the random clamping.

Figure 3 reports results when several decompositions are considered. In each setting the batch algorithm is run for one to five different FVSs with a clamping set of size five, i.e. $|\underline{\mathcal{Y}}_{\mathcal{F}}| = 5$. MPM inference is used for prediction. It can be observed that increasing the number of FVSs has little influence on the results. The combination approach (arithmetic or geometric mean) behaved as predicted by Lemma 1; the arithmetic mean leads to slightly more robust results. This behavior is more pronounced if random states are used for clamping.

## 7 Summary

We propose a lower approximation of the partition sum which improves with increasing computational resources. Our method consists of finding good partitions of the graphical model (Part) and for those partitions find good states of the conditioning set (Clamp). We solve the first problem by finding a minimal FVS to obtain the largest possible tractable subgraph. Then we propose a variational approach MAX-TIGHTEN to optimize the states of the conditioning set. We demonstrate that the resulting learning algorithm has good performance in a computer vision task. Furthermore, the online version enables large scale learning of conditional random fields.

### References


Asuncion, A., Liu, Q., Ihler, A., and Smyth, P. Learning with Blocks: Composite Likelihood and Contrastive Divergence. In *AISTATS*, volume 9, pp. 33–40, 2010.

Bafna, V., Berman, P., and Fujito, T. A 2-Approximation Algorithm for the Undirected Feedback Vertex Set Problem. *SIAM Journal on Discrete Mathematics*, 12(3):289–297, 1999.

Bakir, G., Hofmann, T., Schölkopf, B., Smola, A., Taskar, B., and Vishwanathan, S. V. N. *Predicting Structured Data*. The MIT Press, 2007.

Becker, A. and Geiger, D. Optimization of Pearl's method and greedy-like approximation algorithms for the vertex feedback set problem. *Artificial Intelligence*, 83(1):167–188, 1996.

Besag, J. Statistical Analysis of Non-Lattice Data. *The Statistican*, 24(3):179–195, 1975.

Chandrasekaran, K., Karp, R., Moreno-Centeno, E., and Vempala, S. Algorithms for Implicit Hitting Set Problems. In *SODA*, pp. 614–629, 2011.

Chudak, F., Goemans, M., Hochbaum, D., and Williamson, D. A primal-dual interpretation of two 2-approximation algorithms for the feedback vertex set problem in undirected graphs. *Operations Research Letters*, 22(4):111–118, 1998.

Dillon, J. and Lebanon, G. Stochastic Composite Likelihood. *JMLR*, 11:2597–2633, 2010.

Finley, T. and Joachims, T. Training structural SVMs when exact inference is intractable. In *ICML*, pp. 304–311, 2008.

Fromer, M. and Globerson, A. An LP View of the M-best MAP problem. In *NIPS*, pp. 567–575, 2009.

Hazan, T. and Urtasun, R. A Primal-Dual Message-Passing Algorithm for Approximated Large Scale Structured Prediction. In *NIPS*, 2010.

Hinton, G. Training Products of Experts by Minimizing Contrastive Divergence. *Neural Computation*, 14(8):1771–1800, 2000.

Horvitz, E., Suermondt, J., and Cooper, G. Bounded Conditioning : Flexible Inference for Decisions Under Scarce Resources. In *Proceedings of Conference on Uncertainty in Artificial Intelligence*, 1989.

Jiang, Jiarong, Rai, Piyush, and Daumé III, Hal. Message-Passing for Approximate MAP Inference with Latent Variables. In *NIPS*, 2011.

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

Kolmogorov, V. Convergent tree-reweighted message passing for energy minimization. *PAMI*, 28(10): 1568–83, 2006.

Kumar, S. and Hebert, M. Discriminative Random Fields. *IJCV*, 68(2):179–201, 2006.

Lafferty, J., Mccallum, A., and Pereira, F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *ICML*, pp. 282–289, 2001.

Lindsay, B. Composite Likelihood Methods. *Contemporary Mathematics*, 80, 1988.

Liu, Q. and Ihler, A. Variational Algorithms for Marginal MAP. In *UAI*, 2011.

Luccio, F. Almost exact minimum feedback vertex sets in meshes and butter-flies. *Information Processing Letters*, pp. 59–64, 1998.

Meshi, O., Sontag, D., Jaakkola, T., and Globerson, A. Learning Efficiently with Approximate Inference via Dual Losses. In *ICML*, 2010.

Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. 1990.

Pletscher, P., Ong, C., and Buhmann, J. Entropy and Margin Maximization for Structured Output Learning. In *ECML*, 2010.

Sutskever, I. and Tieleman, T. On the Convergence Properties of Contrastive Divergence. In *AISTATS*, 2010.

Taskar, B., Guestrin, C., and Koller, D. Max-Margin Markov Networks. In *NIPS*, 2003.

Tsochantaridis, I., Joachims, T., Hofmann, T., and Altun, Y. Large Margin Methods for Structured and Interdependent Output Variables. *Journal of Machine Learning Research*, 6:1453–1484, 2005.

Vazirani, V. *Approximation Algorithms*. Springer, 2001.

Vickrey, D., Lin, C., and Koller, D. Non-Local Contrastive Objectives. In *ICML*, 2010.

Wainwright, M. and Jordan, M. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008.

Wainwright, M., Jaakkola, T., and Willsky, A. A new class of upper bounds on the log partition function. In *UAI*, pp. 536–543, 2002.