Higher-Order Inference for Multi-class Log-supermodular Models

Jian Zhang, Josip Djolonga and Andreas Krause
Department of Computer Science, ETH Zurich
jizhang@student.ethz.ch {josipd,krausea}@inf.ethz.ch

Abstract

Higher-order models have been shown to be very useful for a plethora of computer vision tasks. However, existing techniques have focused mainly on MAP inference. In this paper, we present the first efficient approach towards approximate Bayesian marginal inference in a general class of high-order, multi-label attractive models, where previous techniques slow down exponentially with the order (clique size). We formalize this task as performing inference in log-supermodular models under partition constraints, and present an efficient variational inference technique. The resulting optimization problems are convex and yield bounds on the partition function. We also obtain a fully factorized approximation to the posterior, which can be used in lieu of the true complicated distribution. We empirically demonstrate the performance of our approach by comparing it to traditional inference methods on a challenging high-fidelity multi-label image segmentation dataset. We obtain state-of-the-art classification accuracy for MAP inference, and substantially improved ROC curves using the approximate marginals.

1. Introduction

Dealing with uncertainty is a central challenge in computer vision and pattern recognition. Probabilistic modeling and inference have consequently received much attention in these fields. However, general purpose approximate inference algorithms such as belief propagation, mean-field [1] and variants *slow down exponentially* with the size of the largest interaction (number of variables per clique) in the model. Hence, their application has been mostly restricted to low-order models of limited expressive power. For the purpose of *MAP (Maximum A Posteriori) inference*, capturing richly parameterized interactions between large sets of variables has been shown to be very beneficial to computer vision tasks, e.g., in semantic segmentation [2]. The resulting energy minimization problems – even though involving large cliques – remain tractable, as long as the energy functions are *submodular*. The downside of MAP inference

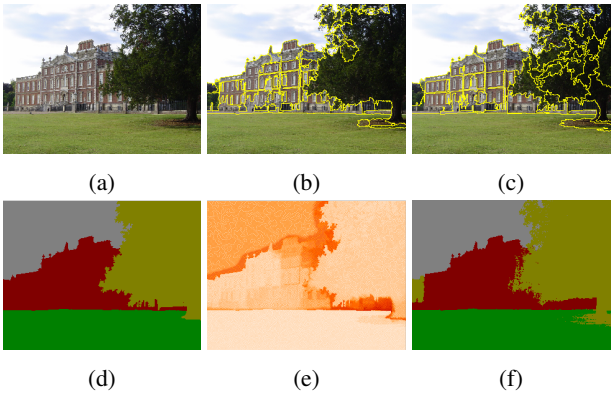


Figure 1: Results on a sample image. (a) Original image. (b,c) Two superpixel layers that are used to define the higher-order potentials. (d) The MAP estimate using our formulation. (e) Entropy of the approximated marginals. (f) Ground truth. Note that our MAP estimate closely matches the ground truth, while capturing uncertainty in difficult regions.

is that all uncertainty is collapsed into a single configuration. A variety of scenarios, such as computing gradients for conditional random field training, integrating estimation into more complex probabilistic models and quantifying the uncertainty in estimations, require (approximated) marginal inference. *Given that submodularity has profound implications for MAP inference, what are its consequences for marginal inference?*

Related work. Most methods for inference in high-order models focus on MAP inference. For example, Kohli et al. [2] show how to minimize the energy of robust P^n potentials by move-making using graph cuts. Tarlow et al. [3] demonstrate how to compute the max-product messages for a family of potentials. Zhang et al. [4] utilize parallelization to achieve constant acceleration for both MAP and marginal inference, but with exponential complexity in terms of model order. A recent primal-dual method [5] employs move-making-like MAP inference for arbitrary higher-order potentials while the iteration-wise complex-

ity is still prohibitively exponential in the clique size. In the case of nested cardinality-based potentials, Tarlow et al. [6] develop a fast algorithm for the exact computation of the factor-to-variable messages for marginal inference. The strategy of Krähenbühl et al. [7] to speed up the mean-field updates by filtering has been extended to higher-order models by Vineet et al. [8], for a family of potentials different from the ones considered in this paper. However, the two families of potentials partially overlap on representative vision models such as higher order Potts model.

Submodularity, as a discrete analogue of convexity, has profound consequences for optimization, with many applications in computer vision, machine learning and beyond [9, 10, 11]. In vision, submodular functions have been primarily utilized for efficient MAP inference as they have the remarkable property that they can be minimized in polynomial time. The most notable case is MAP inference in the attractive Ising model using graph cuts [12], but the problem remains tractable regardless of the cardinality of the potentials. Jegelka et al. [13] use submodularity to model cooperative behavior between edges for image segmentation. The consequences of submodularity for (approximate) marginal inference, however, have not been investigated until very recently. Djolonga et al. [14] present a variational inference approach applicable to arbitrary log-supermodular models. Their approach, however, is limited to binary variables. A different class of submodular-based models (also for binary variables) is studied by Iyer et al. [15].

Our contributions. We introduce a class of *multi-label log-supermodular* models, which generalizes the Potts model [16], but allows to capture attractive interactions of arbitrary order. Extending the approach of [14] on inference in probabilistic submodular models, we propose a novel variational inference scheme, which yields a bound on the partition function. The resulting optimization problem is convex, and can be solved efficiently using the Frank-Wolfe (conditional gradient) algorithm with low per-iteration complexity. At every step, it produces a completely factorized distribution that approximates the true posterior. For the special case of two labels, thresholding this approximate distribution recovers the true MAP assignment. Furthermore, our approach can be interpreted as – and used for – smoothing a convex relaxation to the MAP inference problem. In summary, our main contributions are:

- A novel Bayesian modeling framework for multi-class log-supermodular distributions.
- An easy-to-implement technique for approximate marginal inference with a guaranteed convergence rate.
- A clear connection between our inference scheme and smoothed MAP inference.

Figure 2: The binary (two-label) case. The set consisting of the green elements correspond to the configuration $X_1 = X_3 = 2$ and $X_2 = X_4 = X_5 = X_6 = 1$.

- Experiments demonstrating the scalability and effectiveness on a natural scene segmentation task.

2. Review: Binary Models & Submodularity

We now review relevant background for the *binary* (two-label) case and defer the discussion for the multi-label case to section 3. Formally stated, we seek to model a joint distribution for a random vector $\mathbf{X} = [X_1, X_2, \ldots, X_N]$ such that each component X_i takes on values in $\{1, 2\}$.

Random vectors vs. sets. To draw connections to combinatorial optimization, it will be very important to state the distribution as being defined over *subsets* of a ground set $V = \{1, 2, \ldots, N\}$. We can construct a bijection between the set of all possible states $\{1, 2\}^N$ and the set of all subsets of V by identifying any configuration $\mathbf{x} = (x_1, x_2, \ldots, x_N) \in \{1, 2\}^N$ with the set $A_\mathbf{x} = \{i \mid x_i = 2\}$. Please see figure 2 for an illustration. With this perspective, we will consider distributions of the form

$$P(A) = \frac{1}{\mathcal{Z}} \exp(-F(A)),$$

where $F\colon 2^V \to \mathbb{R}$ and $\mathcal{Z} = \sum_{A \subseteq V} \exp(-F(A))$ is the so-called *partition function* that normalizes the distribution. Hereby, F is often called the *energy function*.

Submodular minimization and MAP. A very important problem is to compute the *MAP configuration*, i.e., finding a set A^* such that $F(A^*)$ is minimal (and hence $P(A^*)$ is maximal). While generally very challenging, a well-studied class of functions for which this can be done in polynomial time are *submodular* functions [17]. A set function $F\colon 2^V \to \mathbb{R}$ is said to be submodular with ground set V if it satisfies the following property

$$F(i \mid A) \geq F(i \mid B)$$

for all sets $A \subseteq B$ and $i \notin B$, where we define the marginal gain $F(i \mid A)$ as

$$F(i \mid A) = F(A \cup \{i\}) - F(A).$$

This definition states the *diminishing returns property* — the benefit of any item i decreases as the context A grows. We will also assume that F is normalized so that $F(\emptyset) = 0$. However, note that this is no restriction as the distributions

$P(A) \propto \exp(-F(A))$ and $P(A) \propto \exp(-F(A)+\beta)$ are identical for any constant β.

Very often the energy function $F(\cdot)$ decomposes and can be written as sum of simpler functions, i.e., $F(A) = \sum_{i=1}^{R} F_i(A \cap U_i)$, where $U_i \subseteq V$ is the domain of the i-th component F_i. For these functions, $\max_i |U_i|$ is called the *order* of the model. As an important example, the widely used *cut function* is of this form:

$$F(A) = \sum_{u \in A, v \notin A} w_{u,v} = \sum_{u,v \in V} w_{u,v}[|A \cap \{u,v\}| = 1].$$

Hereby, $w_{u,v} \geq 0$ are non-negative weights, and $[\cdot]$ denotes the indicator function of the event encoded by its argument. The cut function can be used to obtain the attractive Ising model when the ground set can be organized on a grid and we add weights between immediate neighbors. Then, the resulting probability distribution is

$$P(A) \propto \prod_{u \in V, v \in V} \exp(-w_{u,v}[|A \cap \{u,v\}| = 1]).$$

We can immediately see the attractive behavior as it prefers neighbors to be assigned to the same class (so that the above intersections are of size 0 or 2). The cut function is of order 2. In this paper, we will consider functions of much higher order, discussed in more details in section 7.

A special family of submodular functions that we will make extensive use of are *modular* functions (which are of order 1). A function $s\colon 2^V \to \mathbb{R}$ is modular if $s(A) = \sum_{i \in A} s(\{i\})$. Note that their corresponding distributions completely factorize as $P(A) \propto \prod_{i \in A} \exp(-s(\{i\}))$, and the factors are often called *unary potentials*. Modular functions are essentially linear functions in the discrete world and can be represented by the values $s_i = s(\{i\})$ for $i \in V$. We will thus treat modular functions $s(\cdot)$ as vectors $\mathbf{s} \in \mathbb{R}^N$ with coordinates $s_i = s(\{i\})$.

3. Modeling for the Multi-label Case

We will now show how to handle the more general case where each variable can take one of L different values $\{1, 2, \ldots, L\}$[1]. We will represent each random variable X_i by L distinct elements $V_i = \{v_{i,1}, v_{i,2}, \ldots, v_{i,L}\}$ corresponding to the values it can take. The idea is that if for example $v_{i,5}$ is chosen, then this corresponds to X_i taking on the value 5. This is also known as the 1-of-L encoding. To make sure that the variable can take only a single value, we will add a constraint $\mathcal{M}_i$ that forces the distribution to assign non-zero mass only to those configurations that select *exactly one* element from V_i. Formally stated, $\mathcal{M}_i = \{A\colon |A \cap V_i| = 1\}$ and our final constraint is $\mathcal{M} = \cap_{i=1}^{N} \mathcal{M}_i$. In the combinatorial optimization literature, these constraints are known as the bases of the partition matroid [18][§2.1] and we will use the shorter name *partition constraints*. Note that the final ground set is equal to $V = \cup_{i=1}^{N} V_i$ and has a total of NL elements. We depict this with an example in figure 3. In our experiments we will define one energy function for each set of elements of the ground set that correspond to the same label. We thus also define the sets $V^j = \{v_{i,j} \mid i = 1, 2, \ldots, N\}$ for $j = 1, 2, \ldots, L$.

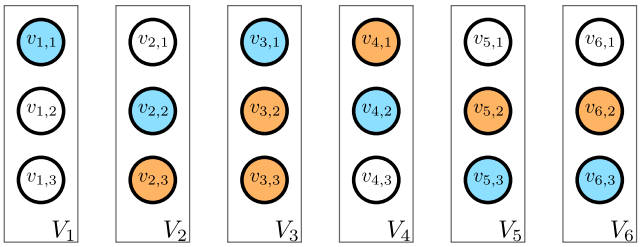


Figure 3: An illustration of the ground set when we have 6 random variables and three possible values/labels for each. The set corresponding to the blue elements satisfies the partition constraints and specifies a feasible assignment. The orange set does not as it picks two values from V_3 and no value from V_1.

With this set-based view at hand, we will consider distributions of the form

$$P(A) = \begin{cases} \frac{1}{Z} \exp(-F(A)) & \text{if } A \in \mathcal{M} \\ 0 & \text{otherwise} \end{cases}$$

for some submodular function $F\colon 2^V \to \mathbb{R}$. We call such models *multi-label log-supermodular models*.

Using our modeling methodology we can also obtain[2] the Potts model [16], which extends the Ising model to multiple labels. The idea is to use one cut function F_j as in the Ising model for each subset of elements that belong to some label, i.e., F_j is a cut on the elements V^j. Then, the Potts model corresponds to $P(A) \propto \exp(-\sum_{j=1}^{L} F_j(A))$.

Another family of submodular functions that are useful in modeling are *concave-over-modular* functions, i.e., those of the form $F(A) = h(s(A))$, where s is a non-negative modular function and h is concave. We will make use of (sums of) such functions in section 7 for modeling label-consistency over superpixels. Note that models of this form have very high order (usually $O(N)$).

4. Probabilistic Inference

Given the log-supermodular model we have just introduced, we are interested in *marginal inference*, i.e., computing the marginal probabilities $P(X_i = j) = P(v_{i,j} \in A)$

[1] Our approach can be easily extended to the case where each random variable takes on a different set of values.

[2] Here we just sketch the construction and a formal proof can be found in the appendix.

for every $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, L$. The central challenge here is the computation of the normalizer

$$\mathcal{Z} = \sum_{A \in \mathcal{M}} \exp(-F(A))$$

without exhaustive enumeration. Unfortunately, even for the attractive Ising model, which has only pairwise interactions and a tractable MAP problem, computing the normalization $\mathcal{Z}$ constant is known to be #P-hard [19]. It is also hard to even approximate it, as shown in [20].

In this paper, we pursue a variational approach that optimizes a bound on $\mathcal{Z}$ by approximating the distribution P by simpler distributions with analytical normalization constants. Before we present our approximation scheme, we have to introduce the *base polytope* $B(F)$ of the submodular function F, which is defined as

$$B(F) = \{\mathbf{s} \in \mathbb{R}^{NL} \mid \forall A \subseteq V \colon s(A) \leq F(A)\} \\ \cap \{\mathbf{s} \in \mathbb{R}^{NL} \mid s(V) = F(V)\},$$

where $\mathbf{s}$ denotes the modular function $s(\cdot)$ when seen as a vector. It is exactly the set of all modular *lower bounds* of F that are tight at the ground set V. We build on the approach of [14] (who addressed the case of two labels) and bound the partition function as follows. For any $\mathbf{s} \in B(F)$ we have that for all $A \subseteq V$ it holds that $s(A) \leq F(A)$, which implies the following inequality

$$\mathcal{Z} = \sum_{A \in \mathcal{M}} e^{-F(A)} \leq \sum_{A \in \mathcal{M}} e^{-s(A)}. \tag{1}$$

Hence, we can upper-bound the partition function by the partition function of the distribution $Q(A) \propto \exp(-s(A))$. Moreover, because $\mathbf{s} \in B(F)$ is a free parameter, we have a variational bound that we can optimize. For any $\mathbf{s} \in B(F)$ we also have an approximative distribution $Q(A) \propto \exp(-s(A))$, which can be easily used in lieu of $P(A)$ as it fully factorizes and has an analytical partition function. Specifically, its partition function is $\prod_{i=1}^{N} \sum_{j=1}^{L} \exp(-s_{i,j})$, which we can plug into (1) to arrive (after taking logs) at the following convex problem

$$\underset{\mathbf{s} \in B(F)}{\text{minimize}} \sum_{i=1}^{N} \log \sum_{j=1}^{L} \exp(-s_{i,j}). \tag{2}$$

The question that remains is how to tackle the above problem. The most remarkable fact about $B(F)$ is that even though it is defined by exponentially many inequalities, we can efficiently optimize linear functions over it due to the celebrated result of Edmonds [17]. To maximize a linear function $\langle \mathbf{w}, \mathbf{s} \rangle$ over $\mathbf{s} \in B(F)$, we first have to sort the elements of $\mathbf{w}$. Specifically, let $\sigma : \{1, 2, \ldots, NL\} \to V$ be a bijection so that $w_{\sigma(1)} \geq w_{\sigma(2)} \geq \ldots \geq w_{\sigma(NL)}$, and construct the sets $S_0 = \emptyset$ and $S_i = \{\sigma(1), \sigma(2), \ldots, \sigma(i)\}$ for $i = \{1, 2, \ldots, NL\}$. Then, the optimizer is the vector $\mathbf{s}^*$ with coordinates $s^*_{\sigma(i)} = F(S_i) - F(S_{i-1})$. Hence, we need $O(NL \log NL)$ operations to sort the vector $\mathbf{w}$ and at most NL function evaluations to obtain the function differences. Below, we show how this result can be used as a key subroutine for efficient variational inference.

As linear programming over the base polytope $B(F)$ is very cheap, a natural candidate for solving eq. (2) is the Frank-Wolfe algorithm, which only needs access to such a procedure. The algorithm has a $O(1/k)$ convergence rate [21], requires no parameter tuning and provides an easily computable duality gap at each iteration.

In the common case of decomposable functions, i.e., $F(A) = \sum_{r=1}^{R} F_r(A)$, we can easily parallelize the computation of the linear optimization oracle, due to the fact that $B(F) = \sum_{r=1}^{R} B(F_r)$ [22][§4.2]. The complete procedure is provided as algorithm 1, where we have denoted the objective of eq. (2) as $g(\cdot)$. Note that computing $\nabla g(\mathbf{s})$ is very efficient, as it amounts to the computation of the marginal probabilities under the distribution $Q(A) \propto \exp(-s(A))$. Specifically, $[\nabla g(\mathbf{s})]_{i,j} = -e^{-s_{i,j}} / \sum_{k=1}^{L} e^{-s_{i,k}}$.

Algorithm 1 Inference with Frank-Wolfe

1: Initialize $\boldsymbol{s} = \boldsymbol{s}_0 \in B(F)$
2: **for** $k = 1$ to `MAX_STEPS` **do**
3: $\quad \boldsymbol{x}_r = \arg\min_{\boldsymbol{y} \in B(F_r)} \langle \nabla g(\boldsymbol{s}), \boldsymbol{y} \rangle$ in parallel for r
4: $\quad \boldsymbol{x} = \sum_{r=1}^{R} \boldsymbol{x}_r$
5: $\quad$ **if** $\langle \mathbf{x} - \mathbf{s}, \nabla g(\mathbf{s}) \rangle \leq \epsilon$ **then**
6: $\quad\quad$ **break**
7: $\quad$ **end if**
8: $\quad \boldsymbol{s} = \boldsymbol{s} + \gamma\,(\boldsymbol{x} - \boldsymbol{s})$ with $\gamma = 2/\,(k + 2)$
9: **end for**
10: **return** $Q(s_{i,j} \in A) \propto \exp(-s_{i,j})$

5. Special Properties in the Binary Case

For the binary case there are in fact two approaches one could take. First, we can simply use a ground set of size N without any constraints, as discussed in section 2. Then, the distribution will be equal to

$$P(A) \propto \exp(-G(A)) \text{ for } A \subseteq \{1, 2, \ldots, N\}. \tag{3}$$

This class of models has been considered by [14], who presented a variational inference scheme for it. Alternatively, we can use our partition constraint construction from section 3 with a ground set of size $2N$. Then, the model is

$$P(A) = \begin{cases} \exp(-F(A)) & \text{if } A \in \mathcal{M} \\ 0 & \text{otherwise} \end{cases}, \tag{4}$$

defined on the ground set

$$V = \{v_{i,j} \mid i = 1, 2, \ldots, N \text{ and } j = 1, 2\}.$$

The two models can be seen to be equivalent if in the latter case the function F splits across the labels, which can be formalized as follows. Note that we can write V as $V = V^1 \cup V^2$, where V^j are the elements corresponding to label j. What we mean by the splitting condition is that the energy can be written as $F(A) = F_1(A \cap V^1) + F_2(A \cap V^2)$ for some functions $F_1 : 2^{V^1} \to \mathbb{R}$ and $F_2 : 2^{V^2} \to \mathbb{R}$. To define the equivalent model of the form eq. (3) we define for $k = 1, 2$

$$G_k : 2^{\{1,2,\ldots,N\}} \to \mathbb{R} \text{ as } G_k(A) = F_k(\{v_{i,k} \mid i \in A\}).$$

Then, the equivalence of these models is immediate if we use $G(A) = G_1(A) + G_2(\overline{A})$, where we have denoted by $\overline{A}$ the complement of A. Note that it is easy to go in the other direction (from eq. (3) to eq. (4)) by defining

$$F(A) = G(\{i \mid v_{i,2} \in A\}).$$

The natural question that arises if the two inference procedures, the one of [14] and ours from eq. (2) will yield the same result when ran on their respective models. The problem from [14] is equal to

$$\underset{\mathbf{s} \in B(G)}{\text{minimize}} \sum_{i \in V} \log(1 + e^{-s_i}), \tag{5}$$

which is not immediately equivalent to eq. (2). However, in the setting we have outlined above they do yield exactly the same probabilities.

Claim 1. *The inference problems eq. (2) and eq. (5) will yield the same marginal probabilities if F and G are in the above correspondence.*

The proof of claim 1 is provided in the supplement. This result has two important consequences. First, it shows that our approach strictly generalizes that of [14]. Second, it proves that the theoretical guarantees from [23] also hold for our procedure in the binary case. For example, thresholding the approximate marginals at $1/2$ will give us an *exact MAP solution*. Furthermore, the authors in [23] show that eq. (5) is equivalent to minimizing the following divergence over all factorized distributions Q

$$D_\infty(P \parallel Q) = \log \sup_{A \subseteq V} \frac{P(A)}{Q(A)}.$$

This divergence-centric view gives us a hint of the qualitative behavior of the approximative distribution. Namely, due to the penalization by the worst case ratio, we would expect approximative distributions Q that try to avoid being over-confident and spread the probability over a larger family of sets. In section 7, we will see how this results in expressive approximate marginals with rich dynamic range.

6. Probabilistic Inference as Smoothed MAP

In many applications, marginals are important and useful. In other settings, a single MAP configuration might suffice. In this section, we discuss how one can interpret our marginal inference technique as solving a smoothed MAP problem which provides computational benefits. Smoothing (using the ℓ_2 norm) has been used in the case of submodular optimization by [24], and entropy-based smoothing has been used for LP-based MAP inference, e.g., [25, 26].

The MAP problem, i.e. finding the most probable configuration for our model is easily seen to be equivalent to the following discrete optimization problem

$$\begin{aligned} \underset{A \subseteq V}{\text{minimize}} \quad & F(A) \\ \text{s.t.} \quad & |A \cap V_i| = 1, \quad \forall i = 1, \ldots, N. \end{aligned} \tag{6}$$

To explain the connection with our approach we need the *Lovász extension* f of F [27], which is defined as

$$f(\mathbf{p}) = \sup_{\mathbf{s} \in B(F)} \langle \mathbf{p}, \mathbf{s} \rangle.$$

It is called an extension as it agrees with F on the vertices of the unit cube. Formally, for any $A \subseteq V$ we have that $F(A) = f(\mathbf{1}_A)$, where $\mathbf{1}_A$ is the characteristic vector of A with 1 in those positions corresponding to the elements of A and 0 elsewhere. Thus, we can see the above problem as the following binary minimization problem

$$\begin{aligned} \underset{\mathbf{p} \in \{0,1\}^N}{\text{minimize}} \quad & f(\mathbf{p}) \\ \text{s.t.} \quad & \mathbf{1}^T \mathbf{p}_i = 1, \quad \forall i = 1, \ldots, N, \end{aligned} \tag{7}$$

where $\mathbf{p}_i \in \{0, 1\}^L$ are the components of $\mathbf{p}$ corresponding to the i-th random variable[3]. A natural approach (c.f. [28]) to solve this problem is to relax $\mathbf{p}_i$ to the probability simplex (the convex hull of the feasible vectors)

$$\Delta = \{\mathbf{p}_i \in \mathbb{R}^L \mid \mathbf{p}_i \geq 0 \text{ and } \mathbf{1}^T \mathbf{p}_i = 1\}.$$

Now the connection is clear from the following claim.

Claim 2. *The Fenchel dual of the problem in eq. (2) is equal to the following.*

$$\underset{\mathbf{p}_i \in \Delta}{\textit{minimize}} \quad f(\mathbf{p}) - \sum_{i=1}^{N} \mathbb{H}[\mathbf{p}_i], \tag{8}$$

where $\mathbb{H}[\mathbf{p}]$ denotes the Shannon entropy. There is zero duality gap and the pair $(\mathbf{s}^*, \mathbf{p}^*)$ is primal-dual optimal iff $\langle \mathbf{s}^*, \mathbf{p}^* \rangle = f(\mathbf{p}^*)$ and

$$p^*_{i,j} = \exp(-s^*_{i,j}) / \sum_{j=1}^{L} \exp(-s^*_{i,j}).$$

[3] Specifically, those entries correspond to V_i

A natural step is to instead smooth the objective using $\epsilon\mathbb{H}[\cdot]$ for some *smoothing parameter* $\epsilon > 0$, which allows to control the level of smoothness [4]. More precisely, we can solve the following problem for some $\epsilon \geq 0$

$$\underset{\mathbf{p}_i \in \Delta}{\text{minimize}} \quad f(\mathbf{p}) - \epsilon \sum_{i=1}^{N} \mathbb{H}[\mathbf{p}_i]. \tag{9}$$

If we reduce ϵ the objective will get closer to the MAP relaxation (which corresponds to $\epsilon = 0$), but the optimization procedure will become more challenging, since the Lovász extension is non-smooth. We formalize this relationship in the following claims, while in section 7 we conduct experiment with different smoothing strengths ϵ.

Claim 3 (from [21])**.** *For smoothing parameter $\epsilon > 0$ the Frank-Wolfe algorithm converges at a rate of $O(\frac{1}{\epsilon k})$.*

Claim 4. *Let us denote with $\mathbf{p}^*(\epsilon)$ the optimum of eq. (9) when using smoothing parameter $\epsilon \geq 0$. Then, we have that*

$$f(\mathbf{p}^*(\epsilon)) - f(\mathbf{p}^*(0)) \leq \epsilon N \log L.$$

Moreover, for any ϵ, N and L, we can always construct a submodular function F so that

$$f(\mathbf{p}^*(\epsilon)) - f(\mathbf{p}^*(0)) = \frac{1}{2}\epsilon N \log(L-1).$$

The first part of claim 4 has been used before, see e.g., [26]. Hence, the bound on the convergence rate grows linearly with $1/\epsilon$ and the suboptimality decreases linearly with ϵ. Moreover, this convergence bound can be very close to tight for an adversely chosen F.

7. Experiments

We now report our experimental setup and results on a challenging semantic segmentation task.

Higher-order modeling. As standard components, our energy function consists of unary terms specified as a modular function $m(A)$ and pairwise terms (a Potts model) $F_{\text{Potts},\,\theta}$ which capture pairwise interactions of neighboring pixels[5]. We first generate multiple layers of superpixels $\mathcal{P}_\ell \subseteq 2^{\{1,\dots,N\}}$ with the mean-shift algorithm [29]. I.e., for each layer ℓ, $\mathcal{P}_\ell$ is a collection of connected regions of pixels with homogeneous appearance. See figure 1b and 1c for an illustration. For every generated superpixel $S \in \mathcal{P}_\ell$, $S \subseteq \{1, 2, \dots, N\}$ and label $j \in \{1, 2, \dots, L\}$, we define $V_S^j = \{v_{i,j} \mid i \in S\}$. Note that V_S^j are the elements of the ground set that have to be chosen if the variables in S take on value j. To encourage label consistency in each superpixel S, we introduce a concave-of-cardinality potential

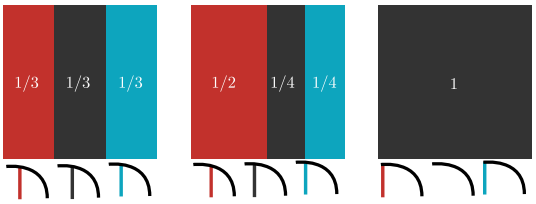


Figure 4: An illustration of the potentials. Every square represents a superpixel with a different proportion of labels. Underneath we show for every label class its contribution to the energy by the length of the corresponding bar (the value of the potential for that label). The final energy is the sum of these three contributions and is minimal when all labels are assigned to a single class.

function of the form $F_S(A) = \sum_{j=1}^{L} |V_S^j \backslash (A \cap V_S^j)|^\alpha$ for some $\alpha \in (0, 1)$, which is illustrated in figure 4. The overall energy is then given by

$$F(A) = m(A) + \beta_1 F_{\text{Potts},\theta}(A) + \beta_2 \sum_{\ell} \sum_{S \in \mathcal{P}_\ell} F_S(A),$$

with β_2 and α as the higher-order parameters.

We consider the following log-supermodular models.

- SUBMOD$_{\text{PAIR}}$: Potts model on a grid ($\beta_2 = 0$).
- SUBMOD$_{\text{2-LAYER}}$: Higher-order with 2 layers ($\beta_1 = 0$).
- SUBMOD$_{\text{3-LAYER}}$: Higher-order with 3 layers ($\beta_1 = 0$).
- SUBMOD$_{\text{2-LAYER-PAIR}}$: SUBMOD$_{\text{2-LAYER}}$ plus pairwise interactions (both $\beta_1 > 0$ and $\beta_2 > 0$).

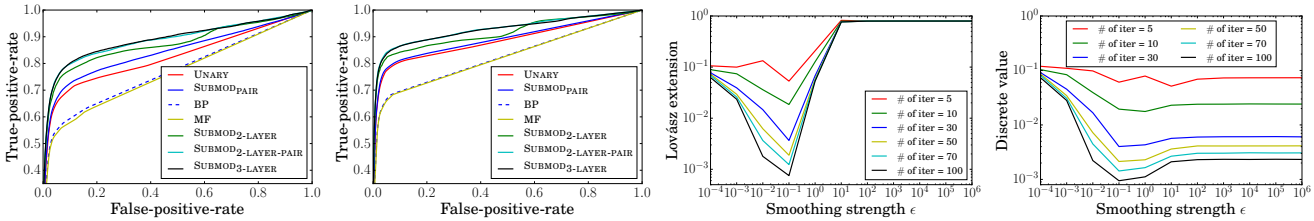
To concretely specify the super pixels, denote by s_p, s_r and m_r the spatial bandwidth, the range bandwidth and the minimum size of regions for mean-shift segmentation respectively. We generate superpixels with $(s_p, s_r, m_r) = (7, 4, 500)$, $(7, 7, 300)$ and $(7, 10, 100)$. Models with two layers of superpixels use the first and last configurations.

Experimental setup. We evaluate our approach on the MSRC-21 dataset. As the original dataset has only coarse-grain annotations, we use the fine-grain annotations for a subset of 93 images by [7]. For all the experiments we use the *TextonBoost* unary features [7]. We report the results using 5-fold cross-validation with the parameters chosen using grid search (exact numbers provided in the appendix). For SUBMOD$_{\text{2-LAYER-PAIR}}$, which uses both pairwise and higher-order interactions, we set the pairwise parameters to those selected for SUBMOD$_{\text{PAIR}}$, and cross-validate the higher-order parameters.

In addition to our models, we do experiments on the Potts model with the mean-field and belief-propagation algorithms from `libDAI` [30]. We also report the performances of ROBUST-PN [2] and the fully connected pairwise model CRF$_{\text{FULLY}}$ [7] as computed by [7]. Our inference procedure runs on 4 threads on an Intel Core-i5 quad-core 3.2 GHz processor. The algorithms using `libDAI` run on a single core of the same processor.

[4] Can be also seen as changing the temperature, shown in the appendix.

[5] We discuss the details of the pairwise potentials in the appendix.



(a) (b) (c) (d)

Figure 5: (a,b) ROC curves over trimaps with bandwidths 0 (i.e. only boundary pixels) and 10 respectively. (c,d) shows for $\text{SUBMOD}_{\text{3-LAYER}}$ the values of the Lovász extension and the discrete energy after a fixed number of iterations.

	Area under ROC curve				KL-divergence				Pixel-wise accuracy	Running time
bandwidth	0	5	10	20	0	5	10	20		
UNARY	0.8254	0.8602	0.8841	0.9112	3.01	2.39	2.00	1.55	$83.71\% \pm 1.81\%$	–
$\text{SUBMOD}_{\text{PAIR}}$	0.8443	0.8708	0.8905	0.9119	2.82	2.22	1.86	1.45	$83.92\% \pm 1.81\%$	12.58
BP_{PAIR}	0.7727	0.8034	0.8245	0.8504	12.03	9.66	8.14	6.34	$83.91\% \pm 1.81\%$	25.64
MF_{PAIR}	0.7663	0.8006	0.8226	0.8499	9.71	8.62	7.37	5.77	$83.83\% \pm 1.82\%$	203.53
$\text{SUBMOD}_{\text{2-LAYER}}$	0.8735	0.9035	0.9132	0.9233	1.68	1.16	0.99	0.79	$88.55\% \pm 1.80\%$	12.53
$\text{SUBMOD}_{\text{2-LAYER-PAIR}}$	0.8886	**0.9184**	**0.9278**	**0.9371**	1.62	1.14	0.97	0.78	$88.48\% \pm 1.68\%$	20.10
$\text{SUBMOD}_{\text{3-LAYER}}$	**0.8904**	0.9173	0.9264	0.9355	**1.57**	**1.11**	**0.95**	**0.77**	$\mathbf{88.61\% \pm 1.70\%}$	15.86
$\text{CRF}_{\text{FULLY}}$	–	–	–	–	–	–	–	–	$88.2\% \pm 0.7\%$	0.2
ROBUST-PN	–	–	–	–	–	–	–	–	$86.5\% \pm 1.0\%$	30

Table 1: Results for the AUC and KL divergence metrics across several trimaps. The pixel-wise accuracies are computed across the full image. The running times and accuracies of the last two rows are as reported by [7].

7.1. Evaluating inference

Estimation of marginals. Because computing the true marginals is intractable due to the size and order of the model, we use the area under the ROC curve (AUC) as a proxy to quantitatively measure the quality of marginals. As we work in a multi-class setting, we first generate one ROC curve per class using the 1-vs-all strategy, which are then averaged to yield the overall curve, as implemented in `scikit-learn` [31]. In addition to the AUC, we evaluate the pixel-wise average KL divergence $\mathbb{KL}(q \,\|\, p)$ between the estimated marginals q and the ground truth labelling p (which is a deterministic 0-1 distribution). We also report numbers on the most challenging part of the image — those pixels around the semantic boundaries. Following [2], a trimap with bandwidth h is defined as the union of $(2h+1) \times (2h+1)$ neighborhoods of all boundary pixels. In figure 5a and figure 5b, we show the ROC curves and the corresponding areas under them are shown in table 1. For different trimap bandwidths, either $\text{SUBMOD}_{\text{3-LAYER}}$ or $\text{SUBMOD}_{\text{2-LAYER-PAIR}}$ achieves the best AUC. We also observe that the higher-order models dramatically improve the AUC over UNARY, while the pairwise models have at most minimal benefits. In table 1, we observe that the higher-order models achieve a much lower KL divergence than the competing models. As the marginals from BP_{PAIR} and MF_{PAIR} are substantially more concentrated near 0 or 1, these can contribute large values to the KL divergence if they are concentrated at the wrong class. Thus we obtain much worse values from BP_{PAIR} and MF_{PAIR} than those from UNARY in table 1.

MAP Estimation. As shown in table 1, $\text{SUBMOD}_{\text{3-LAYER}}$ achieves the best result of 88.61%, followed by $\text{CRF}_{\text{FULLY}}$. As $\text{CRF}_{\text{FULLY}}$ is a pairwise model, the runtime is not directly comparable to our higher-order approach. Hence, at a speed similar to ROBUST-PN we obtain approximate marginals in addition to a high-quality MAP solution.

Efficiency-accuracy trade-off. To better understand the effects of the smoothing parameter ϵ (see section 6), we plot on figures 5c and 5d the values of the Lovász extension and the discrete energy after a fixed number of iterations. First, it is evident from both plots that with weak smoothing one obtains very bad results, as postulated by claim 3. Large values of ϵ seem to hurt the optimization of the Lovász extension, but the effect on the discrete energy seems much more benign. We believe that this is due to the fact that to minimize the discrete energy you just need the right marginal to be largest, which can be achieved even when distribution is close to uniform (as preferred by large values of ϵ). Moreover, note that using our inference approach ($\epsilon = 1$) we do minimally worse in terms of MAP performance, but we also obtain a bound on the partition function and well-motivated approximate marginals.

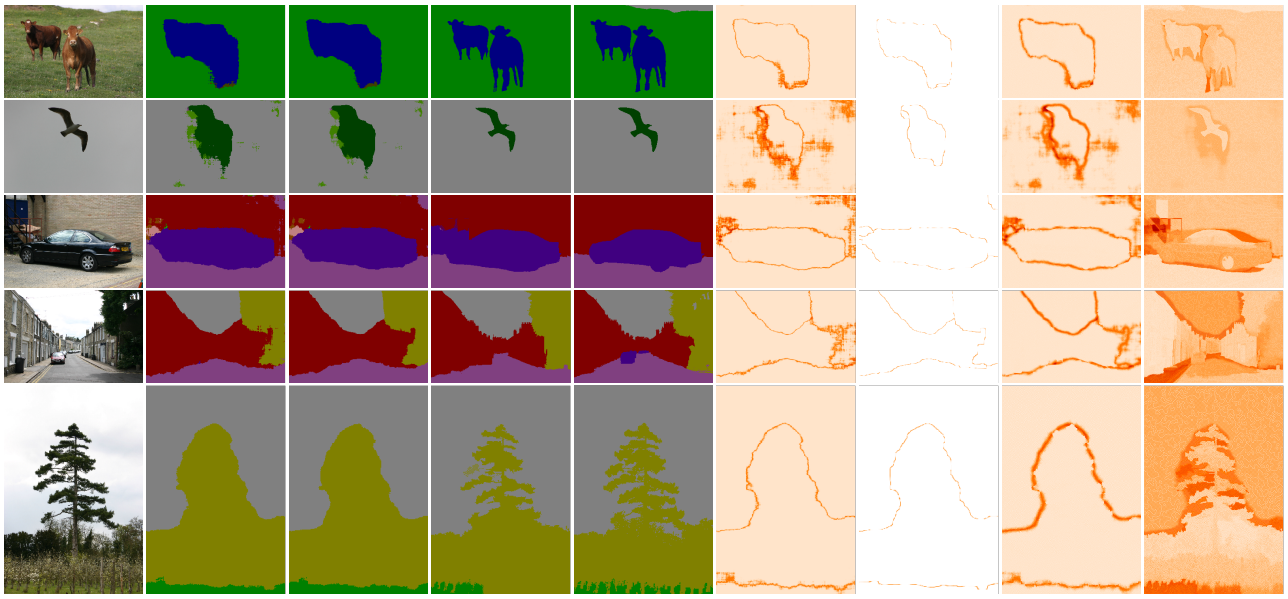


Figure 6: Qualitative results — MAP estimates and the entropy of the estimated marginals (higher entropy for darker regions.)

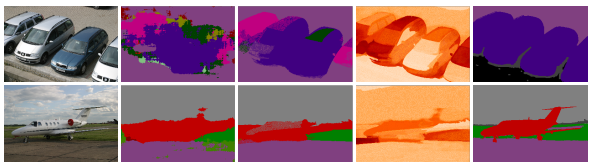


Figure 7: Hard examples for SUBMOD$_{3\text{-LAYER}}$.

7.2. Qualitative results

To better understand the resulting segmentations, we visualize several examples in figure 6. Comparing with UNARY, the pairwise model SUBMOD$_{\text{PAIR}}$ is able to only eliminate small noisy spots, while the higher-order factors in SUBMOD$_{3\text{-LAYER}}$ also smooth out relatively larger noisy regions. As the generated superpixels usually preserve the fine semantic boundaries, our higher-order prior preserves these region boundaries better than simple pairwise smoothing. As shown on the figure, SUBMOD$_{3\text{-LAYER}}$ can indeed produce good segmentations even if the unary potentials do not align well with the true boundaries.

To understand the behavior of the resulting approximate posterior, we show in figure 6 the entropies of the estimated marginals. In our visualization darker colors correspond to higher entropy (higher uncertainty). BP$_{\text{PAIR}}$ is rather overconfident and produces high uncertainty only in a very narrow range while SUBMOD$_{\text{PAIR}}$ reports uncertainty with much higher dynamic range. Both pairwise models reduce the entropy in uncertain regions resulting from the noisy unary features. The uncertainty in the estimates from SUBMOD$_{3\text{-LAYER}}$ follows a different pattern and similar uncertainty levels are typically observed within single regions. In addition, and most importantly, we can also see from the hard examples in figure 7 that our approach is *uncertain* in the regions where it produces wrong estimates, which is what is desired when the data is noisy.

8. Conclusion

In this paper, we proposed a log-supermodular model for multi-class probabilistic modeling with higher-order interactions. We posed a variational inference procedure as a convex optimization problem over the base polytope, and showed how to solve it efficiently using the Frank-Wolfe algorithm. We also made a connection to a smoothed convex MAP relaxation and discussed the trade-off obtained by changing the smoothing parameter. In comparison with multiple pairwise and higher-order baselines, our model achieved state-of-the-art performance for both estimating the MAP and the marginals on the challenging MSRC-21 dataset. We believe that our multi-class framework includes a large family of richly parameterized models, and our very easy-to-implement inference algorithm makes them highly accessible. Together, these results present a step towards building a more powerful toolbox for modeling and quantifying uncertainty under complex data dependencies.

Acknowledgements. This research was supported in part by SNSF grant 200021_137528, ERC StG 307036, Microsoft Research Faculty Fellowship and a Google European Doctoral Fellowship.

References

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

[2] Pushmeet Kohli, Philip HS Torr, et al. Robust higher order potentials for enforcing label consistency. *International Journal of Computer Vision*, 82(3):302–324, 2009. 1, 6, 7

[3] Daniel Tarlow, Inmar E Givoni, and Richard S Zemel. Hop-map: Efficient message passing with high order potentials. In *International Conference on Artificial Intelligence and Statistics*, pages 812–819, 2010. 1

[4] Jian Zhang, Alex Schwing, and Raquel Urtasun. Message passing inference for large scale graphical models with high order potentials. In *Advances in Neural Information Processing Systems*, pages 1134–1142, 2014. 1

[5] Alexander Fix, Chen Wang, and Ramin Zabih. A primal-dual algorithm for higher-order multilabel markov random fields. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1138–1145. IEEE, 2014. 1

[6] Daniel Tarlow, Kevin Swersky, Richard S Zemel, Ryan P Adams, and Brendan J Frey. Fast exact inference for recursive cardinality models. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, 2012. 2

[7] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *arXiv preprint arXiv:1210.5644*, 2012. 2, 6, 7

[8] Vibhav Vineet, Jonathan Warrell, and Philip HS Torr. Filter-based mean-field inference for random fields with higher-order terms and product label-spaces. *International Journal of Computer Vision*, 110(3):290–307, 2014. 2

[9] Mukund Narasimhan, Nebojsa Jojic, and Jeff A Bilmes. Q-clustering. In *Advances in Neural Information Processing Systems*, pages 979–986, 2005. 2

[10] Andreas Krause and Carlos Guestrin. Near-optimal nonmyopic value of information in graphical models. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, July 2005. 2

[11] Francis R Bach. Structured sparsity-inducing norms through submodular functions. In *Advances in Neural Information Processing Systems*, pages 118–126, 2010. 2

[12] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(11):1222–1239, 2001. 2

[13] Stefanie Jegelka and Jeff Bilmes. Submodularity beyond submodular energies: coupling edges in graph cuts. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1897–1904. IEEE, 2011. 2

[14] Josip Djolonga and Andreas Krause. From MAP to marginals: Variational inference in bayesian submodular models. In *Advances in Neural Information Processing Systems*, pages 244–252, 2014. 2, 4, 5, 11

[15] Rishabh Iyer and Jeff Bilmes. Submodular point processes. In *18th International Conference on Artificial Intelligence and Statistics (AISTATS-2015)*, May 2015. 2

[16] Renfrey Burnard Potts. Some generalized order-disorder transformations. In *Mathematical proceedings of the cambridge philosophical society*, volume 48, pages 106–109. Cambridge Univ Press, 1952. 2, 3

[17] Jack Edmonds. Submodular functions, matroids, and certain polyhedra. *Edited by G. Goos, J. Hartmanis, and J. van Leeuwen*, page 11, 1970. 2, 4

[18] Dominic James Anthony Welsh. *Matroid theory*. Courier Corporation, 2010. 3

[19] Mark Jerrum and Alistair Sinclair. Polynomial-time approximation algorithms for the ising model. *SIAM Journal on computing*, 22(5):1087–1116, 1993. 4

[20] Leslie Ann Goldberg and Mark Jerrum. The complexity of ferromagnetic ising with local fields. *Combinatorics, Probability and Computing*, 16(01):43–61, 2007. 4

[21] Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning*, 2013. 4, 6, 13

[22] Satoru Fujishige. *Submodular functions and optimization*, volume 58. Elsevier, 2005. 4, 12

[23] Josip Djolonga and Andreas Krause. Scalable variational inference in log-supermodular models. *arXiv preprint arXiv:1502.06531*, 2015. 5

[24] Peter Stobbe and Andreas Krause. Efficient minimization of decomposable submodular functions. In *Proc. Neural Information Processing Systems (NIPS)*, 2010. 5

[25] Tamir Hazan and Amnon Shashua. Norm-product belief propagation: Primal-dual message-passing for approximate inference. *Information Theory, IEEE Transactions on*, 56(12):6294–6316, 2010. 5

[26] Ofer Meshi, Amir Globerson, and Tommi S Jaakkola. Convergence rate analysis of map coordinate minimization algorithms. In *Advances in Neural Information Processing Systems*, pages 3014–3022, 2012. 5, 6

[27] László Lovász. Submodular functions and convexity. In *Mathematical Programming The State of the Art*, pages 235–257. Springer, 1983. 5

[28] Chandra Chekuri and Alina Ene. Submodular cost allocation problem and applications. In *Automata, Languages and Programming*, pages 354–366. Springer, 2011. 5

[29] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(5):603–619, 2002. 6

[30] Joris M. Mooij. libDAI: A free and open source C++ library for discrete approximate inference in graphical models. *Journal of Machine Learning Research*, 11:2169–2173, August 2010. 6

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 7

[32] Francis Bach. Convex analysis and optimization with submodular functions: a tutorial. *arXiv preprint arXiv:1010.4207*, 2010. 11, 12

[33] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004. 12

[34] Angelia Nedic, DP Bertsekas, and AE Ozdaglar. Convex analysis and optimization. *Athena Scientific*, 2003. 12

[35] R Tyrrell Rockafellar et al. Extension of fenchelduality theorem for convex functions. *Duke mathematical journal*, 33(1):81–89, 1966. 12

[36] Sham Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization. 13

[37] R Tyrrell Rockafellar. *Convex analysis*. Number 28. Princeton university press, 1970. 13

A. Proof of ClaimClaim 1

We want to show an equivalence between the inference procedures for the models

$$P(A) = \begin{cases} \exp(-F_1(A \cap V^1) - F_2(A \cap V^2)) & \text{if } A \in \mathcal{M} \\ 0 & \text{otherwise} \end{cases}, \tag{10}$$

and

$$P(A) \propto \exp(-G_1(A) - G_2(\overline{A})), \tag{11}$$

where

$$G_k : 2^{\{1,2,\dots,N\}} \to \mathbb{R} \text{ as } G_k(A) = F_k(\{v_{i,k} \mid i \in A\}).$$

We will denote the Lovász extensions of F_1, F_2, G_1 and G_2 by f_1, f_2, g_1 and g_2 respectively (they are all maps $\mathbb{R}^N \to \mathbb{R}$). Note that $f_1 = g_1$ and $f_2 = g_2$ if we assume the natural ordering of the respective ground sets. The Lovász extension of $F(A) = F_1(A \cap V^1) + F_2(A \cap V^2)$ is equal to $f(\mathbf{p}) = f_1(\mathbf{p}^1) + f_2(\mathbf{p}^2)$, where $\mathbf{p}^k \in [0,1]^N$ are the entries corresponding to $(v_{1,k}, v_{2,k}, \dots, v_{N,k})$, and $\mathbf{p}$ is the concatenation of $\mathbf{p}^1$ and $\mathbf{p}^2$. Hence, the marginals that we obtain using our method are the optimum of the following problem.

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & f_1(\mathbf{p}^1) + f_2(\mathbf{p}^2) - \sum_{j=1}^{N} \mathbb{H}[(p_j^1, p_j^2)] \\ \text{subject to} \quad & \mathbf{p}^1 + \mathbf{p}^2 = 1 \\ & \mathbf{p}^1 \geq 0 \\ & \mathbf{p}^2 \geq 0 \end{aligned}$$

This problem is equivalent to (substituting $\mathbf{p}^2 = \mathbf{1} - \mathbf{p}^1$)

$$\underset{\mathbf{p}^1 \in [0,1]^N}{\text{minimize}} \; f_1(\mathbf{p}^1) + f_2(\mathbf{1} - \mathbf{p}^1) - \sum_{i=1}^{N} \mathbb{H}[(p_i^1, 1 - p_i^1)].$$

Moreover, because $f_2(\mathbf{1} - \mathbf{p}^1) = f_2(-\mathbf{p}^1) + \text{const.}$ [32][Prop. 3.1 (d)], we have that the above is equivalent to

$$\underset{\mathbf{p}^1 \in [0,1]^N}{\text{minimize}} \; f_1(\mathbf{p}^1) + f_2(-\mathbf{p}^1) - \sum_{i=1}^{N} h(p_i^1),$$

where $h(z) = -z \log z - (1 - z) \log(1 - z)$ is the binary entropy. On the other hand, if we apply the technique from [14] to eq. (11), we obtain as marginals the optimizer $\mathbf{q}^*$ of the problem

$$\underset{\mathbf{q} \in [0,1]^N}{\text{minimize}} \; g(\mathbf{q}) - \sum_{i=1}^{N} h(q_i), \tag{12}$$

where g is the Lovász extension of $G(A) = G_1(A) + G_2(\overline{A}) - G_2(\{1, 2, \dots, N\})$. We have to subtract the constant $G_2(\{1, 2, \dots, N\})$ to make sure that G is normalized, but this has no effect on the distribution. To show the equivalence between these two approaches, we have to prove that $g(\mathbf{q}) = f_1(\mathbf{q}) + f_2(-\mathbf{q})$. As we have already observed that $g_1 = f_1$, it remains to be shown that the Lovász extension of $\overline{G_2}(A) = G_2(\overline{A}) - G_2(\{1, 2, \dots, N\})$ is equal to $f_2(-\mathbf{q})$. This can be immediately seen as $B(\overline{G_2}) = -B(G_2)$ [32][Appendix B] and $f_2 = g_2$, which completes the proof.

B. Proof of ClaimClaim 2

Before showing the lemma, we would like to point out an alternative way of looking at the objective in eq. (9). Note that the Lovász extension of $\overline{F}(A) = \frac{1}{\epsilon} F(A)$ is equal to $\overline{f}(\mathbf{p}) = \frac{1}{\epsilon} f(\mathbf{p})$ [32][Prop. 3.1 (a)]. Hence, we would have obtained the same objective if we had simply changed the temperature of the model and used our inference approach on that, i.e. if we used the energy $\frac{1}{\epsilon} F(A)$ instead of $F(A)$. Thus, varying the smoothing strength is equivalent to changing the temperature.

We will now prove a stronger result, from which the claim follows by setting $\epsilon = 1$.

Claim 5. *For any $\epsilon > 0$, the objective in eq. (9) is the Fenchel dual of*

$$\underset{\mathbf{s}\in B(F)}{\text{minimize}} \sum_{i=1}^{N} \epsilon \log \sum_{j=1}^{L} e^{-\frac{s_{i,j}}{\epsilon}}. \tag{13}$$

There is zero duality gap at the primal-dual pair $(\mathbf{s}^*, \mathbf{p}^*)$ if and only if $p^*_{i,j} = \frac{\exp(-s^*_{i,j}/\epsilon)}{\sum_{j=1}^{L}\exp(-s^*_{i,j}/\epsilon)}$ and $\langle \mathbf{p}^*, \mathbf{s}^*\rangle = f(\mathbf{p}^*)$.

Proof. Let us denote the objective in eq. (13) as $h(\mathbf{s})$. From [33][Ex. 3.25], the convex conjugate of $h_i(\mathbf{s}_i) = \log \sum_{j=1}^{L} e^{s_{i,j}}$ is

$$h^*_i(\mathbf{w}_i) = -\mathbb{H}(\mathbf{w}_i) = \begin{cases} \sum_{j=1}^{L} w_{i,j} \log w_{i,j} & \text{if } \mathbf{w}_i \in \Delta \\ +\infty & \text{otherwise} \end{cases}. \tag{14}$$

Hence $\epsilon h^*_i(-\mathbf{w}_i) = -\epsilon\mathbb{H}(-\mathbf{w}_i)$ and $\epsilon h_i(-\frac{\mathbf{s}_i}{\epsilon})$ are conjugate. As h_i and h_j are independent (share no variables) for $i \neq j$,

$$h^*(\mathbf{w}) = \sum_{i=1}^{N} \epsilon h^*_i(-\mathbf{w}_i) = -\epsilon \sum_{i=1}^{N} \mathbb{H}\left(-\mathbf{w}_i\right).$$

In addition, the convex conjugate of the indicator $I_{B(F)}(\mathbf{s})$ is the Lovász extension $f(\mathbf{w})$ [32][§3]. Thus $g(\mathbf{s}) = -I_{B(F)}(\mathbf{s})$ is the concave conjugate of $g^*(\mathbf{w}) = -f(-\mathbf{w})$. From the Fenchel duality theorem [34][Prop. 7.2.2], we have that

$$\underset{-\mathbf{w}_i\in\Delta}{\text{maximize}} \underbrace{-f(-\mathbf{w})}_{g^*(\mathbf{w})} + \underbrace{\epsilon \sum_{i\in\mathcal{I}} \mathbb{H}(-\mathbf{w}_i)}_{-h^*(\mathbf{w})} \tag{15}$$

is the Fenchel dual problem of

$$\underset{\mathbf{s}\in\mathbb{R}^{NL}}{\text{minimize}} \underbrace{\epsilon \sum_{i=1}^{N} h_i(-\frac{\mathbf{s}_i}{\epsilon})}_{h(\mathbf{s})} + \underbrace{I_{B(F)}(\mathbf{s})}_{-g(\mathbf{s})}. \tag{16}$$

According to Theorem 1 in [35], as $h(\mathbf{s})$ is continuous on the whole domain, strong duality holds at some pair $(\mathbf{s}^*,\ \mathbf{w}^*)$. From Theorem 2 in [35], zero duality gap is achieved if and only if $\mathbf{w}^* \in \partial h(\mathbf{s}^*) \cap \partial\left(-g(\mathbf{s}^*)\right)$. As $h(\mathbf{s})$ is differentiable, we have $w^*_{i,j} = -\exp(-\frac{s^*_{i,j}}{\epsilon})/\sum_{j=1}^{L}\exp(-\frac{s^*_{i,j}}{\epsilon})$. To ensure that $\mathbf{w}^* \in \partial\left(-g(\mathbf{s}^*)\right)$, we also need $\langle \mathbf{w}^*, \mathbf{s}^*\rangle = g(\mathbf{s}^*) + g^*(\mathbf{w}^*)$, i.e. $\langle \mathbf{w}^*, \mathbf{s}^*\rangle = -f(-\mathbf{w}^*)$.

By reparameterizing with $\mathbf{p} = -\mathbf{w}$, we can verify the equivalence of the problem in eq. (9) and the one in eq. (15). Then, the optimality condition is $\langle \mathbf{p}^*,\ \mathbf{s}^*\rangle = f(\mathbf{p}^*)$ and $p^*_{i,j} = \exp(-\frac{s^*_{i,j}}{\epsilon})/\sum_{j=1}^{L}\exp(-\frac{s^*_{i,j}}{\epsilon})$. □

C. Parallelization of Algorithm 1

In algorithm 1 we solve $\min_{\mathbf{y}\in B(F_r)}\langle \mathbf{c}, \mathbf{y}\rangle$ in parallel. As $F(A) = \sum_{i=1}^{R} F_i(A)$ we have that $B(F) = \sum_{i=1}^{R} B(F_i)$ is the Minkowski sum of the base polytopes $B(F_i)$ [22][§4.2], we can write any $\mathbf{s} \in B(F)$ as $\mathbf{s} = \sum_{i=1}^{R} \mathbf{s}_i$, where $\mathbf{s}_i \in B(F_i)$. Conversely, for any $\mathbf{s}_i \in B(F_i)$, we have $\sum_{i=1}^{R} \mathbf{s}_i \in B(F)$. Thus we can conclude

$$\min_{\mathbf{s}\in B(F)} \langle \mathbf{c}, \mathbf{s}\rangle = \min_{\mathbf{s}_i\in B(F_i)} \sum_{i=1}^{R} \langle \mathbf{c}, \mathbf{s}_i\rangle = \sum_{i=1}^{R} \min_{\mathbf{s}_i\in B(F_i)} \langle \mathbf{c}, \mathbf{s}_i\rangle,$$

so that we can solve each of the R optimization problems separately and combine their solutions.

D. Proof of ClaimClaim 3

Claim 3. *For smoothing parameter $\epsilon > 0$, Frank-Wolfe converges at a rate of $O(\frac{1}{\epsilon k})$.*

Proof. Let us denote the relative interior of the domain by $\hat{\Delta} = \left\{\mathbf{p} \mid \mathbf{1}^T \mathbf{p}_i = 1, \mathbf{p}_i > 0\right\}$. On $\hat{\Delta}$ the Hessian $H_\epsilon(\mathbf{p})$ of the entropy term $h_\epsilon(\mathbf{p}) = -\epsilon \sum_{i=1}^N \mathbb{H}(\mathbf{p}_i) = \sum_{i=1}^N \sum_{j=1}^L p_{i,j} \log p_{i,j}$ is diagonal with entries $\epsilon \frac{1}{p_{i,j}}$ and is thus ϵ-strongly convex. From Theorem 6 in [36], we know that the conjugate h_ϵ^* is $1/\epsilon$-smooth. This implies that the objective (eq. (13)) of the Frank-Wolfe problem is thus $1/\epsilon$-smooth, as it is the conjugate of h_ϵ when restricted to $\hat{\Delta}$[6]. Lemma 7 in [21] implies the curvature parameter is $C_{h_\epsilon^*} \leq \text{diam}_{\|\cdot\|_2} \left(B(F)\right)^2 / \epsilon$, and thus the convergence rate is $O(C_{h_\epsilon^*}/k) = O(\frac{1}{\epsilon k})$. □

E. Proof of ClaimClaim 4

Remember that we have defined $\mathbf{p}^*(\epsilon)$ to be the optimal solution of the program in eq. (9) with $\epsilon \geq 0$. In order to prove claim 4 we first construct a family of functions (one for each triplet (ϵ, N, L)) that satisfy the provided lower bound.

Lemma 1. *Given $\epsilon > 0$, N and L, we define the modular function $F(A) = s(A)$ with entries*

$$s_{i,j} = \begin{cases} -\epsilon \log(L-1) & \text{if } j = 1 \\ 0 & \text{otherwise} \end{cases}.$$

Then $f(\mathbf{p}^*(\epsilon)) - f(\mathbf{p}^*(0)) = \frac{1}{2}\epsilon N \log(L-1)$.

Proof. As $\mathbf{s}$ is the only feasible point in $B(F)$ it has to be optimal, and according to claim 5 the optimal $\mathbf{p}^*(\epsilon)$ is

$$p^*_{i,j}(\epsilon) = \frac{\exp(-\frac{s_{i,1}}{\epsilon})}{\sum_{k=1}^L \exp(-\frac{s_{i,k}}{\epsilon})} = \begin{cases} \frac{1}{2} & \text{if } j = 1 \\ \frac{1}{2L-2} & \text{otherwise} \end{cases}.$$

As the Lovász extension is equal to $f(\mathbf{p}) = \langle \mathbf{s}, \mathbf{p} \rangle$, we have that $f(\mathbf{p}^*(\epsilon)) = \sum_{i=1}^N s_{i,1} p^*_{i,1}(\epsilon)$. Because $\mathbf{p}_i \in \Delta$, we easily see that $f(\mathbf{p}^*(0)) = \sum_{i=1}^N \min_{j=1,2,\ldots,L} s_{i,j} = \sum_{i=1}^N s_{i,1} = -N\epsilon \log(L-1)$, which gives

$$f(\mathbf{p}^*(\epsilon)) - f(\mathbf{p}^*(0)) = \sum_{i=1}^N s_{i,1} p^*_{i,1}(\epsilon) - \sum_{i=1}^N s_{i,1} = \frac{1}{2}\epsilon N \log(L-1).$$

□

Claim 4. *Let us denote with $\mathbf{p}^*(\epsilon)$ the optimum of eq. (9) when using smoothing parameter $\epsilon \geq 0$. Then, we have that*

$$f(\mathbf{p}^*(\epsilon)) - f(\mathbf{p}^*(0)) \leq \epsilon N \log L. \tag{17}$$

Moreover, for any ϵ, N and L, we can always construct a submodular function F so that

$$f(\mathbf{p}^*(\epsilon)) - f(\mathbf{p}^*(0)) = \frac{1}{2}\epsilon N \log(L-1). \tag{18}$$

Proof. We have just proved the second part (18) in lemma 1. To prove the first part (17), note that

$$f(\mathbf{p}^*(0)) \geq f(\mathbf{p}^*(0)) - \epsilon \sum_{i=1}^N \mathbb{H}(\mathbf{p}_i^*(0)) \geq f(\mathbf{p}^*(\epsilon)) - \epsilon \sum_{i=1}^N \mathbb{H}(\mathbf{p}_i^*(\epsilon)) \geq f(\mathbf{p}^*(\epsilon)) - \epsilon N \log L.$$

The last inequality follows from the fact that the entropy for discrete distributions over L atoms is maximal for the uniform distribution, and equals to $\log L$. □

[6] According to Theorem 12.2 in [37], the conjugate is the same whether we restrict h to its interior or not, as the later function is the closure of the former.

Pairwise		Higher-order	
θ	β_1	α	β_2
$(0.01, 0.1, 1, 10, 100)$	$(0.1, 0.5, 1, 5)$	$(0.8, 0.9)$	$(25, 37.5, 50, 62.5, 75)$

Table 2: The parameter grid used. The best parameters were chosen using cross-validation.

F. Experiments

F.1. Representing the Potts model

In this subsection we show how to represent the Potts model using our modelling approach. Assume the pairwise graphical model is represented by a graph $G(U, E)$ with random variables $X_u \in \{1, 2, ..., L\}$ for $u \in U = \{1, 2, \ldots, N\}$. Note that there is no assumption about the structure of the graph G. We define the pairwise Potts energy as

$$\phi(\boldsymbol{x}) = \sum_{\{u,w\}\in E} \phi_{u,w}(x_u, x_w) = \sum_{\{u,w\}\in E} \lambda_{u,w}[x_u \neq x_w],$$

for some weights $\lambda_{u,v} \geq 0$. For each factor $\phi_{u,w}(x_u, x_w)$, we define

$$F_{u,w}(A) = \sum_{j=1}^{L} \frac{1}{2}\lambda_{u,w} \left[|A \cap \{v_{u,j},\ v_{w,j}\}| = 1\right]. \tag{19}$$

We will show that the models $P(\mathbf{x}) \propto \exp(-\phi(\mathbf{x}))$ and

$$P(A) = \begin{cases} \frac{1}{Z}\exp(-F(A)) & \text{if } A \in \mathcal{M} \\ 0 & \text{otherwise} \end{cases}$$

give the same probabilities, where $F(A) = \sum_{(u,w)\in E} F_{u,w}(A)$ and $A_{\mathbf{x}} = \{v_{i,x_i} \mid i = 1, 2, \ldots, N\} \in \mathcal{M}$. Consider any $(u, w) \in E$. If $x_u = x_w = l$, then this edge will have no contribution in $\phi(\mathbf{x})$. Similarly $F_{u,w}(A_{\mathbf{x}})$ will equal to zero as all intersections in the sum are empty, except for $j = l$, in which case the intersection equals 2. If $x_u \neq x_w$, then the contribution in $\phi(\mathbf{x})$ will be equal to $\lambda_{u,w}$. In $F_{u,w}(A_{\mathbf{x}})$ all intersections will be empty, except for $j = x_u$ and $j = x_w$ in which case the intersections are of size 1. Hence, each of them will contribute $\frac{1}{2}\lambda_{u,w}$, so that the total contribution towards the energy is again $\lambda_{u,w}$. Finally, note that the sets of the form $A_{\mathbf{x}}$ for some $\mathbf{x} \in \{1, \ldots, L\}^N$ are exactly the elements in $\mathcal{M}$, which completes the proof.

F.2. Parameter grid

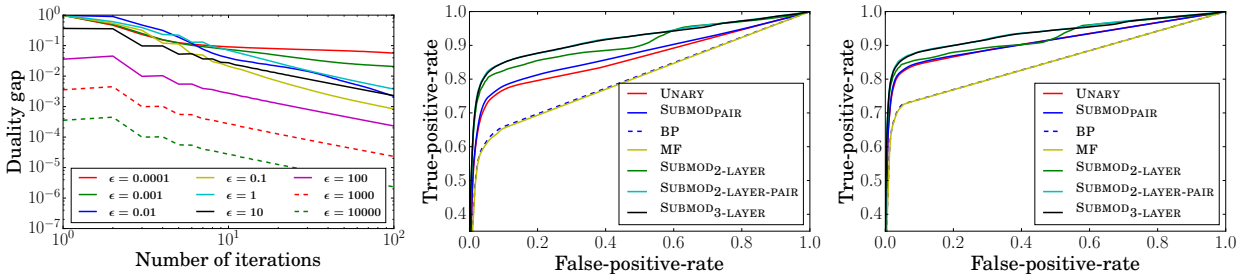
To model the multi-label pairwise smoothness for directly neighboring pixels, we use in eq. (19) the weights $w_{i,j} = \exp\left(-\theta\|\boldsymbol{I}_i - \boldsymbol{I}_j\|^2/255^2\right)$. The full parameter grid is shown in table 2.

F.3. Additional experimental results.

ROC curves over trimaps. Due to lack of space in the paper, we show the ROC curves for trimap bandwidths 5 and 20 in figures 8b and 8c. Note that the areas under them were reported in table 1.

Empirical analysis of ClaimClaim 3. We report in figure 8a the curves showing the duality gap averaged over all data samples using the model SUBMOD$_{3\text{-LAYER}}$. We have normalized the gaps, so that for every image the maximum observed gap across all runs (for different ϵ) is equal to 1. As expected from claim 3, the curves for different ϵ are roughly linear functions with different offsets. For relatively large ϵ we can see bigger values of ϵ do produce lower curves. However, for small values for ϵ the ordering of the curves is not as expected and thus the analysis on the convergence rate might not be always be tight.

Note that the curves in figure 5c and figure 5d are also averaged from 93 samples. To normalize the values of the Lovász extension, we pick for each sample the largest and smallest observed values and set them to 1 and 0 respectively. To compute the discrete energy, we consider the dual variables in claim 5 as approximated marginals and choose the label with highest marginal for each pixel. These values have been normalized and averaged in the same way as the Lovász extension.



(a) (b) (c)

Figure 8: Additional experimental results. (a) shows the duality gaps for different ϵ in a log-log scale. (b, c) show the ROC curves for trimap bandwidths 5 and 20 respectively.