Scalable Variational Inference in Log-supermodular Models

Josip Djolonga JOSIPD@INF.ETHZ.CH
Andreas Krause KRAUSEA@ETHZ.CH
Department of Computer Science, ETH Zurich

Abstract

We consider the problem of approximate Bayesian inference in log-supermodular models. These models encompass regular pairwise MRFs with binary variables, but allow to capture high-order interactions, which are intractable for existing approximate inference techniques such as belief propagation, mean field, and variants. We show that a recently proposed variational approach to inference in log-supermodular models –L-FIELD– reduces to the widely-studied minimum norm problem for submodular minimization. This insight allows to leverage powerful existing tools, and hence to solve the variational problem orders of magnitude more efficiently than previously possible. We then provide another natural interpretation of L-FIELD, demonstrating that it *exactly* minimizes a specific type of Rényi divergence measure. This insight sheds light on the nature of the variational approximations produced by L-FIELD. Furthermore, we show how to perform parallel inference as message passing in a suitable factor graph at a linear convergence rate, without having to sum up over all the configurations of the factor. Finally, we apply our approach to a challenging image segmentation task. Our experiments confirm scalability of our approach, high quality of the marginals, and the benefit of incorporating higher-order potentials.

1. Introduction

Performing inference in probabilistic models is one of the central challenges in machine learning, providing a foundation for making decisions with uncertain data. Unfortunately, the general problem is intractable and one must resort to approximate inference techniques. The importance of this problem is witnessed by the amount of interest it has attracted in the research community, which has resulted in a large family of approximations, most notably the mean-field (Wainwright & Jordan, 2008) and belief propagation (Pearl, 1986) algorithms and their variants. One major drawback of these and many other techniques is the exponential dependence on the size of the largest factor which restricts the class of models one can use. In addition, these methods generally involve non-convex objectives, resulting in local optima (or even non-convergence).

We consider the problem of inference in distributions over sets, also known as point processes. Formally, we have some finite ground set V and a measure P that assigns some probability $P(A)$ to every subset $A \subseteq V$. We would like to point out that we can equivalently see such distributions as being defined over $|V|$ Bernoulli random variables $X_i \in \{0, 1\}$, one for every element in the ground set $i \in V$ indicating if element i has been selected. As a concrete example showing this equivalence consider the task of image segmentation in computer vision, where one wants to separate the foreground from the background pixels. Traditionally, one defines one random variable $X_p \in \{0, 1\}$ for each pixel p indicating if the pixel is in the foreground or the background. We can also isomorphically treat the distribution as being defined over *subsets* of the set of all pixels V. In this case, for any subset $A \subseteq V$ the quantity $P(A)$ is the probability of pixels A belonging to the foreground. In the remaining of the paper we will employ this latter view of distributions over sets. The additional assumption that we make is that the distribution is *log-supermodular*, i.e. can be written as $P(A) = \frac{1}{Z} \exp(-F(A))$, where F is some *submodular* function.

Related work. Submodular functions are a family of set functions exhibiting a natural diminishing returns property, originating first in the field of combinatorial optimization (Edmonds, 1970). They have been applied to many problems in machine learning, including clustering (Narasimhan et al., 2005), variable selection (Krause & Guestrin, 2005), structured norms (Bach, 2010), dictionary learning (Cevher & Krause, 2011), etc. Submodular functions have huge implications for the tractability of (ap-

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

proximate) optimization, akin to convexity and concavity in continuous domains. While the major emphasis has consequently been on optimization, submodular functions can be also employed to define probabilistic models. Special cases include Ising models used in computer models and the determinantal point process (DPP) (Kulesza & Taskar, 2012) used for modeling diversity. Alas, submodularity does not render the inference problem tractable, which remains extremely difficult even for the Ising model (Goldberg & Jerrum, 2007; Jerrum & Sinclair, 1993) which has only pairwise interactions.

The study of approximate Bayesian inference in general log-supermodular models has been recently initiated by Djolonga & Krause (2014). They provide a general variational approach –L-FIELD– that optimizes bounds on the partition function via the differentials of submodular functions. While their approach leads to optimization problems that can be solved exactly in polynomial time for arbitrary high order interactions, presently the approach is slow, and impractical for large scale inference tasks such as those arising in computer vision. Iyer & Bilmes (2015) study another family of submodular distributions.

Our contributions. We improve over their result in several ways. First, by showing an equivalence of L-FIELD with a classical problem in submodular minimization – the minimum norm point problem – we obtain access to a large family of specially crafted algorithms that can handle models with very large numbers of variables. In the experimental section we indeed perform inference over images, which have hundreds of thousands of variables. This insight also implies, for example, that the approximation agrees on the mode of the distribution, hence the MAP problem is solved for free. Secondly, by establishing another important connection, namely to a specific type of information divergence, we shed light on the type of approximations that result from this method. Thirdly, we show how special structure of many real-world log-supermodular models (such as those in image segmentation with high-order potentials) enable a highly efficient parallel message passing algorithm that converges to the global optimum at a linear rate. Lastly, we perform extensive experiments on a challenging image segmentation task, demonstrating that our approach is scalable, provides more accurate marginals than existing techniques, and provides evidence on the effectiveness of models using high-order interactions.

2. Background: Submodularity and log-supermodular models

Formally, a function $F\colon 2^V \to \mathbb{R}$ is said to be submodular if for any pair of sets $A \subseteq B$ and $x \notin B$ it holds that

$$F(\{x\} \cup A) - F(A) \geq F(\{x\} \cup B) - F(B).$$

Figure 1: The original image and two layers of superpixels to be used for defining attractive higher order potentials.

In other words, the *gain* of adding any element x decreases as the context grows, which is the diminishing returns property already mentioned. Without any loss of generality we assume that F is normalized so that $F(\emptyset) = 0$. We will consider Gibbs distributions that arise from these models, i.e. probability measures of the form

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

for some submodular $F\colon 2^V \to \mathbb{R}$. These models are called *log-supermodular* or *attractive* for reasons explained below.

Examples. A typical example of such models is the regular Ising model, which can be used for the image segmentation task from the introduction. Continuing with that example, we define a set of edges E that connect neighboring pixels, and for every pair of neighbors $\{p, p'\}$ we specify a weight $w_{\{p,p'\}} \geq 0$ that quantifies their similarity. To model the preference of neighbors to be assigned to the same segment, we use the *cut function*

$$\forall A \subseteq V\colon F(A) = \sum_{\{p,p'\} \in E} 1_{|A \cap \{p,p'\}|=1} w_{\{p,p'\}}.$$

Hence, if we place two neighboring pixels p and p' in different segments, we will cut the edge $\{p, p'\}$ and be "penalized" by the corresponding weight, which explains the attractive behavior of the model. We can go one step further and define regions $P_i \subseteq V$ which we would prefer to be in the same segment. One strategy to generate the regions, used by Kohli et al. (2009), is to generate superpixels, as illustrated on Figure 1. We can then modify the model to incorporate these higher order potentials by adding terms of the form $\phi(|P_i \cap A|/|P_i|)$ for some concave function ϕ. As a concrete example, consider $\phi(z) = z(1 - z)$, which assigns a value of 0 when the pixels in the superpixel are in the same segment, and assigns a larger penalty otherwise, which is maximal when the pixels are equally split between the two segments.

Modular functions. The simplest family of submodular functions are modular functions, which can be seen as the discrete analogue of linear functions. Namely, a function $s\colon 2^V \to \mathbb{R}$ is said to be modular if for all $A \subseteq V$ it holds that $s(A) = \sum_{i \in A} s(\{i\})$. The family of distributions that arise from these functions are exactly the family of completely factorized distributions[1], because

$$P(S) \propto \exp(-s(S)) = \prod_{i \in S} \exp(-s_i).$$

As evident from their definition, modular functions are uniquely defined through the quantities $s(\{i\})$ for all $i \in V$. It is very useful to view modular functions as vectors $\mathbf{s} \in \mathbb{R}^V$ with coordinates $s_i = s(\{i\})$.

Submodular polyhedra. There are several polyhedra that contain some of these modular functions (in their vectorial representation) that we will make use of. More specifically, we are interested in the submodular polyhedron $P(F)$ and the base polytope $B(F)$, which are defined as

$$P(F) = \{\mathbf{s} \in \mathbb{R}^V \mid \forall A \subseteq V\colon s(A) \leq F(A)\}, \quad (1)$$

$$B(F) = P(F) \cap \{\mathbf{s} \in \mathbb{R}^V \mid s(V) = F(V)\}. \quad (2)$$

In other words, $P(F)$ is the set of all modular *lower bounds* of the function F, while $B(F)$ adds the further restriction that the bound must be tight at the ground set V. It can be shown that these polyhedra are not empty and their geometry is also well understood (Fujishige, 2005; Bach, 2013). Moreover, what is especially surprising, is that even though $B(F)$ is defined with exponentially many inequalities, we can optimize linear functions over it in $O(|V| \log |V|)$ time with a simple greedy strategy (Edmonds, 1970).

MAP estimation and the minimum norm point. A very natural question that arises for any probabilistic model is that of finding a MAP configuration, which for log-supermodular distribution amounts to minimizing the function F. This is a problem that has been studied in much detail and has resulted in numerous approaches. The fastest known combinatorial algorithm due to Orlin (2009) has a bound of $O(n^6 + \tau n^5)$, where τ is the cost of evaluating the function, and can be prohibitively expensive to run for larger ground sets. An algorithm that performs better in practice, but only has a pseudopolynomial running time guarantee (Chakrabarty et al., 2014), is the Fujishige-Wolfe algorithm (Fujishige, 1980). This method approaches the problem by solving the following convex program, known as the *minimum norm problem*.

$$\underset{\mathbf{s} \in B(F)}{\text{minimize}} \ \|\mathbf{s}\|^2. \quad (3)$$

[1]Because we use Gibbs distributions, note that they can not assign zero probabilities.

One can extract the solution to the submodular minimization problem from the solution to the above problem by thresholding, which is formalized in the following theorem.

Theorem 1 (Fujishige (2005)). *If $\mathbf{s}^*$ is the optimal solution to problem (3), define the following sets*

$$A_- = \{v \mid v \in V \text{ and } s^*_v < 0\}, \text{and}$$
$$A_0 = \{v \mid v \in V \text{ and } s^*_v \leq 0\}.$$

Then A_- and A_0 are the unique minimal and maximal minimizers of F.

3. Variational inference with L-FIELD

The main barrier to performing inference in log-supermodular models is the computation of the normalizing factor $\mathcal{Z}$, also known as the partition function in the statistical physics literature. We cannot compute it directly as we have to sum up over all $S \subseteq V$, so we have to use approximative techniques. One common approach is to define an optimization problem over some variational parameter $\mathbf{q}$, so that we can compute the quantity of interest by optimizing this problem.

We now review the variational approximation technique recently introduced by Djolonga & Krause (2014). Their method relies on two main observations: *(i)* modular functions have analytical log-partition functions and *(ii)* submodular functions can be lower-bounded by modular functions. The main idea is the following: if it holds that $\forall A \subseteq V\colon s(A) \leq F(A)$, then it will certainly be the case that

$$\log \sum_{A \subseteq V} e^{-F(A)} \leq \log \sum_{A \subseteq V} e^{-s(A)} = \sum_{i \in V} \log(1 + e^{-s_i}).$$

We thus have a family of variational upper bounds on the partition function parametrized by the modular functions $\mathbf{s}$, over which we can optimize to minimize the right hand side of the inequality. As shown by Djolonga & Krause (2014) this variational problem can be reduced to the following convex separable optimization problem over the base polytope

$$\underset{\mathbf{s} \in B(F)}{\text{minimize}} \sum_{i \in V} \log(1 + \exp(-s_i)). \quad (4)$$

This problem – L-FIELD – can be then solved using the divide-and-conquer algorithm (Bach, 2013; Jegelka et al., 2013) by solving at most $O(\min\{|V|, \log \frac{1}{\epsilon}\})$ MAP problems, where ϵ is the tolerated error on the marginals. It can be also approximately solved using the Frank-Wolfe algorithm at a convergence rate of $O(1/k)$. While these results establish tractability of the variational approach, in general solving even one MAP problem requires submodular minimization – an expensive task, and repeated solution may be too costly. Convergence of the Frank-Wolfe method is empirically slow.

4. L-FIELD ≡ Minimum norm point.

Our first main contribution is the following, perhaps surprising, result:

Theorem 2. *Problems* (4) *and* (3) *have the same solution.*

The proof of this theorem (given in the appendix) crucially depends on the peculiar characteristics of the base polytope. Similar results have been shown (for other objectives) by Nagano & Aihara (2012). This theorem has three immediate, extremely important consequences. First, since the minimum-norm point approach is often the method of choice for submodular minimization anyway, this insight reduces the cost from solving many MAP problems to a *single* minimum norm point problem, which leads to substantial performance gains – a factor of $O(|V|)$ if we seek the optimal variational solution! Secondly, given this equivalence and Theorem 1, we can immediately see that we can in fact extract the MAP solution by thresholding the marginals at $1/2$.

Corollary 1. *We can extract the unique minimal and maximal MAP solutions by thresholding the optimal marginal vector at* $1/2$.

Thus, the L-FIELD approach results in the *exact* MAP solution in addition to approximate marginals and an upper bound on the partition function. Thirdly, since the minimum norm point problem is well studied, faster algorithms for important special cases become available. In particular, in §6, we demonstrate how certain types of log-supermodular distributions enable extremely efficient parallel inference.

5. The divergence minimization perspective

The L-FIELD approach attacks the partition function directly. One can of course employ the factorized distribution parametrized by the minimizer $\mathbf{s}^*$ of the upper bound to obtain approximate marginals. However, it is not immediately clear what properties the resulting distribution has, apart from agreeing on the mode (as shown by Corollary 1). To this end, we turn to the theory of divergence measures as that will enable us to understand the solutions preferred by the method. Divergence measures are functions $D(P \,\|\, Q)$ of two probability distributions P and Q that quantify the degree of dissimilarity between the arguments. Once we have picked a divergence measure D, we are interesting in minimizing $D(P \,\|\, Q)$ among some set of approximative distributions $Q \in \mathcal{Q}$. The family which is of particular interest to us is that of completely factorized distributions that assign positive probabilities, which we now formally define.

Definition 1. *We define the set* $\mathcal{Q}$ *of completely factorized positive distributions as*

$$\mathcal{Q} = \{Q \mid Q(S) \propto \prod_{i \in S} \exp(-q_i) \;\; \textit{for some } \mathbf{q} \in \mathbb{R}^V\}.$$

There are many choices for a divergence measure, the most prominent examples being the KL-divergence and the family of Rényi divergences (Rényi, 1961). Of particular interest for our analysis is the special infinite order of the Rényi divergence, defined as follows:

Definition 2 (Van Erven & Harremoës (2012))**.** *Define the Rényi divergence of infinite order between* $P(S)$ *and* $Q(S)$

$$D_\infty(P \,\|\, Q) = \log \sup_{S \subseteq V} \frac{P(S)}{Q(S)}. \tag{5}$$

In the terminology of Minka et al. (2005) we can see that the D_∞ divergence is *inclusive*, which means that it would try to "cover" as much as possible from the distribution: The variational approximation is conservative in the sense that it attempts to spread mass over all sets that achieve substantial mass under the true distribution. As we now show, it turns out that when we minimize this divergence for log-supermodular distributions we can focus our attention only on some specific factorized distributions.

Lemma 1. *When* P *is log-supermodular, to solve* $\text{minimize}_{Q \in \mathcal{Q}}\, D_\infty(P \,\|\, Q)$ *we have to only optimize over modular functions* q *that are global lower bounds of* F.

What this lemma essentially says, is that a minimizing distribution $\mathbf{q}^*$ can be always found in $P(F)$. This result also implies the central result of this section, that the variational approach we have considered essentially minimizes the infinite divergence.

Theorem 3. *When* P *is log-supermodular, the problem* $\text{minimize}_{Q \in \mathcal{Q}}\, D_\infty(P \,\|\, Q)$ *is equivalent to problem* (4).

This theorem has the following immediate consequence:

Corollary 2. *For log-supermodular models, problem* $\text{minimize}_{Q \in \mathcal{Q}}\, D_\infty(P \,\|\, Q)$ *is polynomial-time tractable via* $O(|V|)$ *MAP (submodular minimization) problems.*

Hence, any log-supermodular distribution has the property that we can find the closest factorized distribution to it w.r.t. this specific divergence in polynomial time, irrespective of whether the distribution factorizes into smaller factors or not. We would like to point out that the above criterion does not necessarily hold in general for non-log-supermodular distributions, which we formally show.

Lemma 2. *Lemma 1 does not hold for general point processes. Specifically, there exists a log-submodular counter example.*

The proofs of all claims are provided in the supplemental material.

6. Parallel inference for decomposable models

Very often the submodular function F has structure that one can exploit to procure faster inference algorithms. In particular, the function often *decomposes*, i.e., can be written as a sum of (simpler) functions as

$$F(S) = \sum_{i=1}^{R} F_i(S \cap V_i),$$

where $F_i : 2^{V_i} \to \mathbb{R}$ are submodular functions with ground sets V_i. This setting has been considered, e.g., by Stobbe & Krause (2010) and Jegelka et al. (2013). The decomposition implies that the corresponding distribution factorizes as follows

$$P(S) \propto \prod_{i=1}^{R} \exp(-F_i(S \cap V_i)). \tag{6}$$

In fact, the examples we discussed in §2 both have this form, factorizing either into pairwise potentials, or into the potentials defined by the superpixels. Such models can be naturally represented via a factor graph G that has as nodes the union of the ground sets V_i, and the factors $F_1, \dots, F_R$. We then add edges E in a bipartite way by connecting each factor F_i to the elements V_i that participate in it (e.g. F_i is connected to v iff $v \in V_i$). For any node w in the graph (either a factor, or variable node), we will denote its neighbors by $\delta(w)$.

When the function enjoys such a decomposition, the base polytope can be written as the Minkowski sum of the base polytopes of the summands, or formally [2]

$$B(F) = \sum_{i=1}^{R} B(F_i).$$

Hence, the minimum norm problem (3) that we are interested in can be rewritten as the following problem.

$$\underset{\mathbf{q}_i \in B(F_i)}{\text{minimize}} \sum_{v \in V} \Big(\sum_{F_i \in \delta(v)} q_{i,v} \Big)^2.$$

In the following, we discuss two natural message passing algorithms exploiting this structure.

Expectation propagation. A very natural approach would be to perform block coordinate descent one factor at a time. If we look through the lens of divergence measures, as introduced in §5, we can make a clear connection to (a variant of) *expectation propagation*[3], the message passing approach of Minka et al. (2005) specialized to minimizing the divergence $D_\infty(P \,\|\, Q)$, which we now briefly describe. The main idea is to approximate each factor $\exp(-F_i(S \cap V_i))$ with a completely factorized distribution $Q_i(S) \propto \exp(-q_i(S))$, such that the product $\prod_{i=1}^{R} Q_i$ is a good approximation to the true distribution in terms of the given divergence. Then, we optimize iteratively using the following procedure.

1. Pick a factor F_i.
2. Replace the other factors F_j for $j \neq i$ with their approximations Q_j and minimize

$$D_\infty\Big(\frac{1}{Z_i} \exp(-F_i(S)) \prod_{j \neq i} Q_j \,\Big\|\, \prod_{j=1}^{R} Q_j\Big)$$

over the factorized approximation Q_i.

In other words, we choose a factor and minimize the infinite divergence for that factor, but instead of using the true factors $\exp(-F_j(S))$ for $j \neq i$, we replace them with their current modular approximations Q_j.

A parallel approach. One downside of the approach presented above is that it has to be applied *sequentially*, i.e., one factor has to be updated at a time to ensure convergence. An alternative is to apply an approach used by Jegelka et al. (2013), which allows to perform message passing *in parallel* without losing the convergence guarantees. Jegelka et al. (2013) assume that all factors depend on *all* variables (i.e. $V_i = V$). In the following, we generalize their setting in order to allow $V_i \neq V$. By changing the dual problem they consider (shown in detail in the appendix) we arrive at a form that is more natural to our setting and can be seen as performing message passing in the factor graph. To describe the messages, we have to define the following pair of norms that arise from the structure of the factor graph.

Definition 3. *For any $\mathbf{x}_S \in \mathbb{R}^S$, where $S \subseteq V$, we define the following pair of norms.*

$$\|\mathbf{x}_S\|_G^2 = \sum_{v \in S} \frac{1}{|\delta(v)|} x_v^2, \text{ and } \|\mathbf{x}_S\|_{G*}^2 = \sum_{v \in S} |\delta(v)| x_v^2.$$

The messages from variables to factors are simple sums, similar to those in standard belief propagation

$$\mu_{v \to F_i}^{t+1} = \frac{1}{|\delta(v)|} \sum_{F_j \in \delta(v)} \mu_{F_j \to v}^{t}.$$

The factors always keep some vector on their base polyhedron, which at iteration t will be denoted by $\mathbf{q}_i^t \in B(F_i)$. Then, based on the incoming messages, they update this vector by solving a convex problem, which is much cheaper

[2] If $v \notin V_i$, then the elements from $B(F_i)$ are treated as having a zero for that coordinate.

[3] Typically, expectation propagation is defined w.r.t. the KL-divergence.

than the exhaustive computation one has to do for belief propagation (which is *exponential* in the factor size). We will denote the message sent from node u to node w at iteration t by $\mu^t_{u\to w}$. If $\mathbf{m}^t_i \in \mathbb{R}^{V_i}$ is the *vector* of messages received at iteration t at node F_i (one message from each $v \in V_i$), then the factor solves a projection problem parametrized by $(\mathbf{m}^t_i, \mathbf{x}^t_i)$, whose solution is assigned to $\mathbf{x}^{t+1}_i$. Written formally, we have

$$\mathbf{q}^{t+1}_i = \operatorname*{argmin}_{\mathbf{q}_i \in B(F_i)} \|\mathbf{q}_i - (\mathbf{q}^t_i - \mathbf{m}^t_i)\|^2_{G*}.$$

As this is a convex separable problem on the base polytope, it can be solved for example using the divide-and-conquer algorithm (Bach, 2013). Having solved this problem, the factor sends the following messages to its neighbours

$$\mu^{t+1}_{F_i\to v} = \mathbf{q}^{t+1}_v.$$

Stated differently, it will send to every variable node v the coordinate of the stored vector corresponding to that variable. At every iteration t we can extract the current factorized approximation to the full distribution by simply considering the incoming messages at the variable nodes. Specifically, the approximation $\mathbf{q}^t$ at time step has in the v-th coordinate the sum of incoming messages at the node v, or formally

$$q^t_v = \sum_{F_i \in \delta(F_j)} \mu^t_{F_i\to v}.$$

Because the algorithm can be seen as performing block coordinate descent on a specific problem (discussed in the appendix), the message passing algorithm described above possesses strong convergence guarantees that depend on the structure of the factor graph. These guarantees even hold if all messages from nodes to factors, and all messages from factors to nodes are each computed *in parallel*. An important quantity that appears in the convergence rate is the maximal variable connectivity $\Delta_V = \max_{v\in V} |\delta(v)|$. Based on recent new results by Nishihara et al. (2014) on block coordinate descent for a similar dual (assuming that all factors depend on all variables, as considered by Jegelka et al. (2013)), we extend their analysis to obtain a linear convergence rate for our message passing scheme.

Theorem 4 (Extension of Nishihara et al. (2014))**.** *If the graph is Δ_V-regular, s.t. every variable appears in exactly Δ_V factors, then the message passing algorithm converges linearly with rate $(1 - \frac{1}{|V|\Delta_V})^2$. More specifically*

$$\|\mathbf{q}^t - \mathbf{q}^*\| \le 2\|\mathbf{q}^0 - \mathbf{q}^*\|_\infty \sqrt{\Delta_V E}(1 - \frac{1}{|V|^2\Delta_V^2})^t,$$

where $\mathbf{q}^*$ is the optimal point, $\mathbf{q}^0$ is the initial point and E is the number of edges in the factor graph.

7. Experiments

We now report experimental results [4] on applying our parallel variational inference scheme to a challenging image segmentation problem as motivated in §2. The goal of our experiments is to test the scalability of our approach to large problems, and to evaluate the quality of the marginals both qualitatively and quantitatively. We used the data from Jegelka & Bilmes (2011), which contains a total of 36 images, each with a highly detailed (pixel-level precision) ground truth segmentation. Due to intractability, we cannot compute the exact marginals against which we would ideally wish to compare. As a proxy for measuring the quality of the approximations, we use the area under the ROC curve (AUC) as compared to the ground truth segmentation. We classify each pixel independently as fore- or background by comparing its approximate marginal against a threshold, which we vary to obtain the ROC curve. We have used the following model, which contains both pairwise and higher-order interactions.

$$F(A) = \alpha m(A) + \beta F_{\text{cut}}(A) + \gamma \sum_{P_i \in \mathcal{P}} |P_i| \phi\left(\frac{|A \cap P_i|}{|P_i|}\right),$$

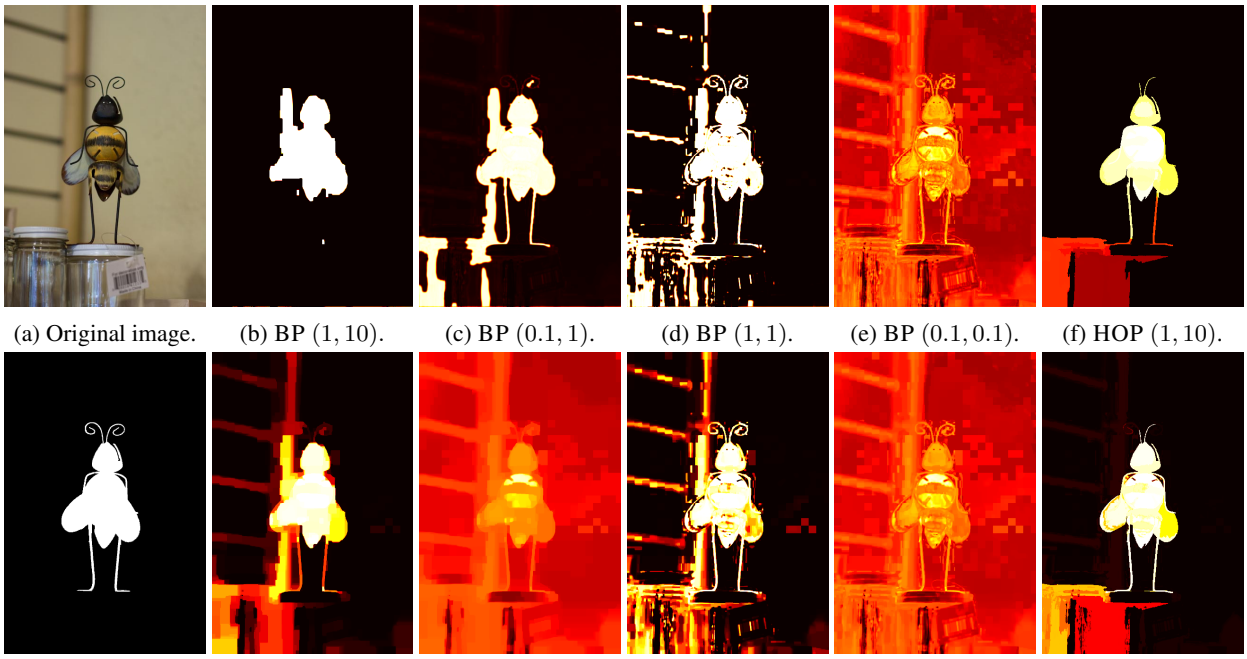
where

- the unary potentials $m(\cdot)$ were learned from labeled data using a 5 component GMM;
- the pairwise potentials F_{cut} connect neighboring pixels $\mathbf{x}$ and $\mathbf{x}'$ with weights $w(\mathbf{x}, \mathbf{x}') = \exp(-\theta\|\mathbf{x} - \mathbf{x}\|^2)$, where $\mathbf{x}$ and $\mathbf{x}'$ are the RGB values of the pixels;
- the higher order potentials were generated using the mean-shift algorithm of Comaniciu & Meer (2002). We have used two overlapping layers of superpixels, each layer with different granularity. The concave function was defined as $\phi(z) = z^{0.6}(1-z)^{0.6}$.

We compared the following inference techniques. The reported typical running times are for an image of size 427x640 pixels on a quad core machine and we report the wall clock time of the inference code (without setting up the factor graph or generating the superpixels).

- Unary potentials only with independent predictions, i.e., $\beta = \gamma = 0$.
- Belief propagation (BP) and mean-field (MF) for the pairwise model (i.e. $\gamma = 0$). We have used the implementation from `libDAI` (Mooij, 2010). The maximum number of iterations was set to 70. We note that this code is not parallelized. When we observe

[4] The code will be made available at `http://people.inf.ethz.ch/josipd/`.



(a) Original image. (b) BP $(1, 10)$. (c) BP $(0.1, 1)$. (d) BP $(1, 1)$. (e) BP $(0.1, 0.1)$. (f) HOP $(1, 10)$.

(g) Ground truth. (h) DR $(1, 10)$. (i) DR $(0.1, 1)$. (j) DR $(1, 1)$. (k) DR $(0.1, 0.1)$. (l) HOP $(1, 1)$.

Figure 2: Example marginals from the different approximation procedures for the original image (a) with ground truth segmentation (g). For the results comparing BP and DR (b-e,h-k) we have used the same pairwise weights and weights. The numbers in the parenthesis correspond to the constants by which the unaries and the prior are multiplied (i.e. to (α, β) for the pairwise models and (α, γ) for the higher-order model). Note how BP is overconfident, whereas our methods offer marginals with much higher dynamic range.

fast convergence, for example BP can converge in 3 iterations, it takes about 45 seconds. Even though we have set a relatively low number of iterations, the running times can be extremely slow if the methods do not converge. For example, running mean-field for 70 iterations can take more than 9 minutes.

- Our approach using only pairwise potentials ($\gamma = 0$), solved using the total variation Douglas-Rachford (DR) code from (Barbero & Sra, 2011; 2014; Jegelka et al., 2013). We ran for at most 100 iterations. The inference takes typically less than a second.
- Our approach with higher order potentials (HOP) only ($\beta = 0$). The inference takes less than 13 seconds.

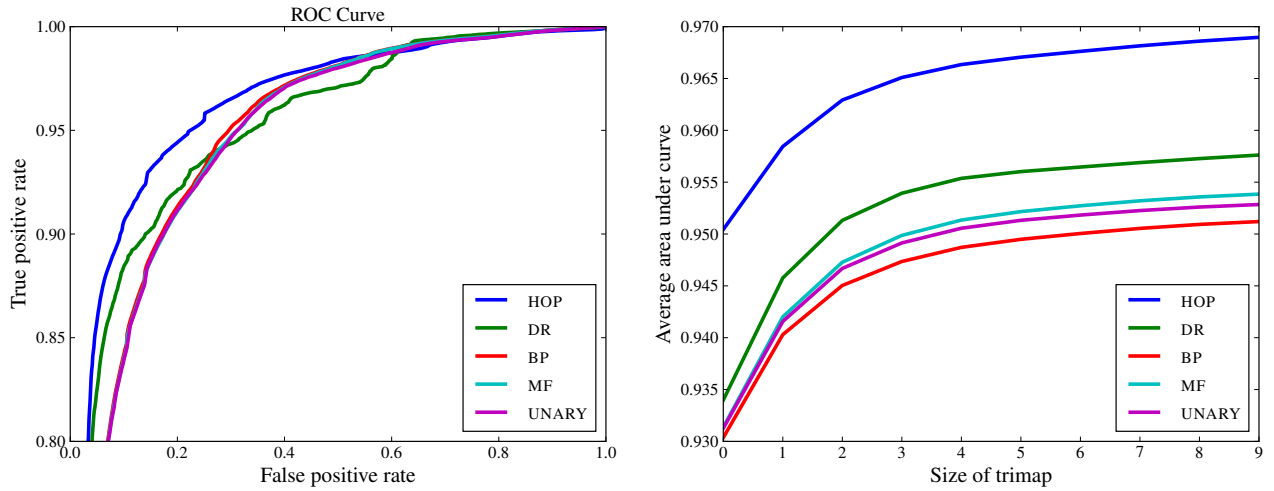
For every method we tested several variants using different combinations for α, β, γ and θ (exact numbers provided in the appendix). Then, we performed a leave-one-out cross-validation for estimating the average AUC. We have also generated a sequence of 10 trimaps by growing the boundary around the true foreground to estimate accuracy over the hardest pixels, namely those at the boundary.

Accuracy. We first wish to quantitatively compare the accuracy of the approximate marginals. We report the aggregate results in Figure 3, and the ROC curves in Figure 4.

Method	Avg. AUC	Std. Dev.	Avg. AUCT	Std. Dev.
HOP	**0.9670**	0.0549	**0.9644**	0.0546
DR	0.9568	0.0663	0.9525	0.0651
BP	0.9500	0.0636	0.9464	0.0734
MF	0.9489	0.0647	0.9487	0.0676
UNARY	0.9484	0.0658	0.9480	0.0681

Figure 3: Average scores of the methods estimated using leave-one-out cross validation. The *Avg. AUC* column is the average area under the ROC curve. The *Avg. AUCT* column reports the average of the mean AUC over the 10 trimaps. The second and the fourth columns are the standard deviation of the preceding columns.

We can clearly see that our approach outperforms the traditional inference methods for both objectives — the AUC over the whole image and over the challenging boundary (trimaps). Sometimes we see very poor behavior of the alternative methods, which can be attributed to either their over-confidence (as verified below), or the fact that they optimize non-convex objectives and can fail to converge within the given number of iterations. Lastly, capturing high-order interactions leads to higher accuracy (in particular around the boundary) than pairwise potentials only.



(a) Average ROC curves for the different methods.

(b) Average AUC for trimaps of increasing size.

Figure 4: Comparison of inference methods in terms of their accuracy. For each method we optimize the parameters via grid-search, and report leave-one-out cross-validation results. (a) ROC curves for classifying pixels as fore- or background by independently thresholding marginals, averaged over the whole image. (b) Results over the trimaps (blurred boundaries around the ground truth segmentation), focusing on "difficult" pixels. For every algorithm and every of the 10 trimap sizes we report the average area under the curve.

Properties of marginals. We would also like to understand the qualitative characteristics of the resulting marginals of our methods when compared with the traditional techniques. From the discussion on the divergence minimization in §5, we would expect the approximate marginals to avoid assigning low probabilities and rather prefer to err conservatively, i.e., on the side of causing false positives. On the other hand, it is known that the results of belief propagation are often over-confident. For this purpose, we provide a visual comparison in Figure 2. Namely, each of the four BP/DR pairs are results using the respective algorithms for the same parameters of the model. We observe exactly what the theory predicts — the distribution obtained via L-FIELD is less concentrated around the object and mass is spread around more. The contrast is starkest on Figures 2 *(b)* and *(h)*, where we use a very strong pairwise prior (high β). On Figures 2 *(e)* and *(k)* we have used a very weak pairwise prior (low β), and as expected the resulting marginals are mainly determined by the unary part and the choice of inference procedure does not make a difference. The results in the last column are from the higher order model, with two different values of γ (the strength of the higher order potential). We can see that the resulting probabilities better preserve the boundaries of the object and the fine details, which is one of the main benefits of using these models.

8. Conclusion

We have addressed the problem of variational inference in log-supermodular distributions. In particular, building on the L-FIELD approach of Djolonga & Krause (2014), we established two natural, important interpretations of their method. First, we showed how L-FIELD can be reduced to solving the well-studied minimum norm point problem, making a wealth of tools from submodular optimization suddenly available for approximate Bayesian inference. Secondly, we showed that the factorized distributions returned by L-FIELD minimize a particular type of information divergence. Both of these theoretical connections are immediately algorithmically useful. In particular, for the common case of decomposable models, both connections lead to efficient message passing algorithms. Exploiting the minimum norm connection, we proved strong convergence rates for a natural parallel approach, with convergence rates dependent on the factor graph structure. Lastly, we demonstrate our approach on a challenging image segmentation task. Our results demonstrate the accuracy of our marginals (in terms of AUC score) compared to those produced by classical techniques like belief propagation, mean field and variants, on models where these can be applied. We also show that performance can be further improved by moving to high-order potentials, leading to models where classical marginal inference techniques become intractable. We believe our results provide an important step towards practical, efficient inference in models with complex, high-order variable interactions.

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

References

Bach, Francis. Structured sparsity-inducing norms through submodular functions. In *NIPS*, 2010.

Bach, Francis. Learning with submodular functions: a convex optimization perspective. *Foundations and Trends® in Machine Learning*, 6(2-3), 2013.

Barbero, Álvaro and Sra, Suvrit. Modular proximal optimization for multidimensional total-variation regularization. 2014.

Barbero, lvaro and Sra, Suvrit. Fast Newton-type methods for total variation regularization. In *ICML*, pp. 313–320, 2011.

Cevher, Volkan and Krause, Andreas. Greedy dictionary selection for sparse representation. *IEEE Journal of Selected Topics in Signal Processing*, 99(5):979–988, September 2011.

Chakrabarty, Deeparnab, Jain, Prateek, and Kothari, Pravesh. Provable submodular minimization using Wolfe's algorithm. In *Advances in Neural Information Processing Systems*, pp. 802–809, 2014.

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

Djolonga, Josip and Krause, Andreas. From MAP to marginals: Variational inference in Bayesian submodular models. In *Neural Information Processing Systems (NIPS)*, 2014.

Edmonds, Jack. Submodular functions, matroids, and certain polyhedra. *Combinatorial structures and their applications*, pp. 69–87, 1970.

Fujishige, Satoru. Lexicographically optimal base of a polymatroid with respect to a weight vector. *Mathematics of Operations Research*, 5(2):186–196, 1980.

Fujishige, Satoru. *Submodular functions and optimization*, volume 58 of *Annals of Discrete Mathematics*. 2005.

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

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

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

Jegelka, Stefanie, Bach, Francis, and Sra, Suvrit. Reflection methods for user-friendly submodular optimization. In *NIPS*, 2013.

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

Kohli, Pushmeet, Ladický, L'ubor, and Torr, Philip H.S. Robust higher order potentials for enforcing label consistency. *International Journal of Computer Vision*, 82 (3):302–324, 2009.

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

Kulesza, A. and Taskar, B. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5(2–3), 2012.

Minka, Tom et al. Divergence measures and message passing. Technical report, Technical report, Microsoft Research, 2005.

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

Nagano, Kiyohito and Aihara, Kazuyuki. Equivalence of convex minimization problems over base polytopes. *Japan journal of industrial and applied mathematics*, 29 (3):519–534, 2012.

Narasimhan, Mukund, Jojic, Nebojsa, and Bilmes, Jeff. Q-clustering. In *NIPS*, volume 5, pp. 5, 2005.

Nishihara, Robert, Jegelka, Stefanie, and Jordan, Michael I. On the convergence rate of decomposable submodular function minimization. In *Advances in Neural Information Processing Systems*, pp. 640–648, 2014.

Orlin, James B. A faster strongly polynomial time algorithm for submodular function minimization. *Mathematical Programming*, 118(2):237–251, 2009.

Pearl, Judea. Fusion, propagation, and structuring in belief networks. *Artificial intelligence*, 29(3):241–288, 1986.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P.,

Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Rényi, Alfréd. On measures of entropy and information. In *Fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pp. 547–561, 1961.

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

Van Erven, Tim and Harremoës, Peter. Rényi divergence and Kullback-Leibler divergence. *arXiv preprint arXiv:1206.2459*, 2012.

Wainwright, Martin J. and Jordan, Michael I. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1(1-2):1–305, 2008.