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

Master Thesis

Author(s):
Zhang Jian

Publication date:
2015

Permanent link:
https://doi.org/10.3929/ethz-a-010437872

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

M.Sc. Thesis by

Jian Zhang

Supervisor: Prof. Dr. Andreas Krause

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Abstract

Although shown to be a very powerful tool in computer vision, existing higher-order models are mostly restricted to computing MAP configuration for specific energy functions. In this thesis, we propose a multi-class model along with a variational marginal inference formulation for capturing higher-order log-supermodular interactions. Our modeling technique utilizes set functions by incorporating constraints that each variable is assigned to exactly one class. Marginal inference for our model can be done efficiently by either Frank-Wolfe or a soft-move-making algorithm, both of which are easily parallelized. To simutaneously address the associated MAP problem, we extend marginal inference formulation to a parameterized version as smoothed MAP inference. Accompanying the extension, we present a rigorous analysis on the efficiency and accuracy trade-off by varying the smoothing strength.

We evaluate the scalability and the effectiveness of our approach in the task of natural scene image segmentation, demonstrating state-of-the-art performance for both marginal and MAP inference. In addition, we also conduct experiments on the efficiency-accuracy trade-off to verify our theoretical analysis.

Acknowledgements

I would like to express my special appreciation to my supervisor Professor Andreas Krause. He is not only a nice guide in research but also a kind advisor who helps me make important personal decisions. I would also like to say thanks to Josip Djolonga. He helps me explore different ideas with helpful insights and effective discussions. In our collaboration, I have learned a lot both technically and in methodology of logical reasoning.

This work is dedicated to my parents. Without their valuable comprehension and selfless support, I would not be able to face difficulties in various aspects and heading confidently towards my future life.

Contents

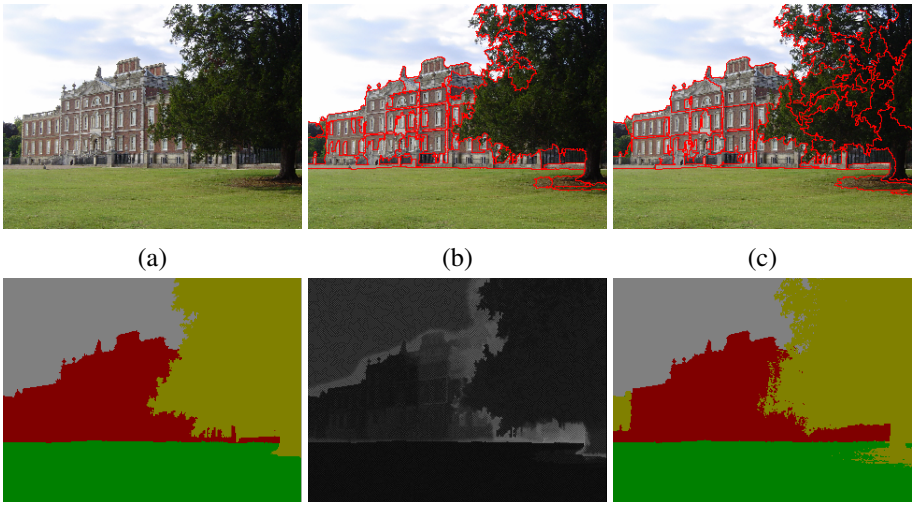
1 Introduction

Probabilistic inference is a powerful mechanism allowing for making decisions under uncertainty. There are many standard inference techniques that have been successfully applied to a plethora of domains (Wainwright & Jordan, 2008; Hazan & Shashua, 2012). However, these techniques are usually limited to low-order interactions as complexity grows *exponentially* in the order of the largest clique. In recent years, higher order modeling has drawn a significant attention in the vision community. It demonstrates superior performance over conventional pairwise MRF in various tasks such as segmentation (Kohli et al., 2009), scene understanding (Zhang et al., 2013) and stereo matching (Woodford et al., 2009). Although exhibiting appealing expressiveness, these higher-order models also exert challenges on the inference techniques.

To do inference over the higher-order models, the alternating approach (Sun et al., 2014; Valgaerts et al., 2010) divides variables into groups. In each iteration, a certain group of variables is fixed and the model is reduced to a lower order one. The algorithms then iteratively alternate the fixed group and do lower order inference. Besides this approach, three main classes of rigorous multi-class inference methods are developed. The first class is based on belief propagation. Zhang et al. (2014) utilize parallelization to achieve constant acceleration for both MAP and marginal inference, but it suffers from the exponential complexity in computing messages. Tarlow et al. (2010) propose a polynomial time message updating approach for special forms of higher order interactions. However, it only works for MAP inference. The second class uses max-flow/min-cut solver as the underlying engine. Representative works like (Kohli et al., 2009) make use of expansion/swap moves for robust Pn model. A recent primal/dual method (Fix et al., 2014) does move-making-like MAP inference for arbitrary higher order potentials while the iteration-wise complexity is still prohibitively exponential. The most recent fashion is based on filtering which serves for both MAP and marginal inference. It is first introduced to image segmentation by (Krähenbühl & Koltun, 2012) and later generalized to solving special forms of higher order CRFs (Vineet et al., 2014).

Submodular functions are a family of set functions with the property of diminishing gains. It is an natural tool modeling utility in different inference and learning settings such as clustering (Narasimhan et al., 2005), structured norm (Bach, 2010b) and variable selection (Krause & Guestrin, 2012). Specifically in the vision literature, techniques based on graph-cut (Boykov et al., 2001) are used for regular (submodular) energy minimization. Recent work (Djolonga & Krause, 2014) initiates the study of submodularity in the Bayesian binary settings. The log-supermodular inference formulation can incorporate higher order binary interactions with polynomial iteration-wise complexity. Djolonga & Krause (2015) connect the binary marginal inference problem to the well-studied min-norm problem which inspires scalable parallel inference algorithms. These inference approaches have been successfully applied to problems such as fore/background segmentation.

In this thesis, we propose a novel multi-class log-supermodular model together



(a) (b) (c)

(d) (e) (f)

Figure 1.1: Applying our inference technique to an image from the MSRC-21 dataset. (a) Original image. (b) and (c) are two layers of mean-shift superpixel generated with different parameters. (d) MAP result from our formulation. (e) Entropy of the marginal from our formulation. (f) Ground truth.

with a marginal inference approach for capturing any log-supermodular higher-order interactions. With minor modifications, our approach also solves approximated MAP inference with guarantees on approximation error and convergence rate. Our framework has polynomial iteration-wise complexity for both approximated marginal and MAP inference.

In Figure 1.1, we demonstrate an example of applying our inference framework to segment an image from MSRC-21 dataset. We generate multiple layer of superpixels from mean-shift algorithm with different parameters. Together with unary potentials, we use the higher-order prior that pixels in a single superpixel tend to share the same label. By running our marginal inference algorithm, we estimate not only marginals but also MAP results from the associated smoothed MAP inference formulation. We show in Figure 1.1e the entropy from the estimated marginals. The high entropy values around semantic boundaries shows MAP estimation is more uncertain where objects interact. In other words, our formulation support Bayesian analysis on uncertainty in addition to point-wise decision-making.

Contributions.

- A new Bayesian modeling framework together with a marginal inference formulation for multi-class log-supermodular distribution.
- Efficient and parallelizable algorithms for approximate marginal inference.
- We propose and analyze an smoothed MAP inference formulation which has deep connections with our marginal inference problem. We also present a rigorous treatment of the efficiency-accuracy trade-off in controlling smoothing strength.
- We demonstrate the scalability and effectiveness of our approach on natural scene segmentation.

2 Preliminaries: submodularity

Submodular functions are set functions with diminishing gains. Given a set function $F : 2^V \to \mathbb{R}$ with finite ground set V, we define $F(i|A) = F(A \cup \{i\}) - F(A)$ as the gain of adding $i \in V$ to A. F is submodular if $F(i|B) \leq F(i|A)$ for $\forall A \subseteq B \subseteq V$ and $i \notin B$. Set function F is supermodular if $-F$ is submodular. One simple example of submodular function is the cut function. Given a graph $G = (V, E)$, a cut function is defined as

$$F(A) = \sum_{(u,v)\in E} w_{i,j} |\mathbf{1}(u \in A) - \mathbf{1}(v \in A)|$$

which is the sum over weights of edges connecting A and $V \setminus A$. Another example is the concave cardinality function. Let $g(x)$ be an arbitrary concave function. With ground set V, for $\forall A \in V$, we define a concave cardinality function as

$$F(A) = f(|A|).$$

It can be shown concave cardinality function is also submodular.

Modular functions are a special class of submodular and supermodular functions simultaneously. By associating real value s_v with $v \in V$, modular function $s : 2^V \to \mathbb{R}$ is defined as $s(A) = \sum_{v\in A} s_v,\ \forall A \subseteq V$. Another essential notion related to submodularity is the base polytope. The base polytope of submodular function F is defined as

$$B(F) = \left\{ \boldsymbol{s} \in \mathbb{R}^{|V|} : s(V) = F(V),\ s(A) \leq F(A),\ \forall A \subset V \right\}.$$

There has been extensive investigation such as (Bach, 2011)[§9.1] and (Jegelka et al., 2013) on convex optimization over the base polytope. One of the frequently referred problem is evaluating Lovász extension for a given submodular function. Formally the Lovász extension of submodular function F is defined as

$$f(\boldsymbol{w}) = \max_{\boldsymbol{s}\in B(F)} \langle \boldsymbol{w},\ \boldsymbol{s} \rangle.$$

Though the feasible set $B(F)$ is defined with exponentially many linear constraints, linear optimization over $B(F)$, i.e. evaluating Lovász extension, can be achieved efficiently in $O(|V| \log |V|)$ time (Edmonds, 1970).

3 Log-supermodular modeling with sets

3.1 Approach

Let $\boldsymbol{X} = \left(X_1, X_2, ..., X_{|\mathcal{I}|}\right)$ be a vector of random variables with index set $\mathcal{I} = \{1, 2, ..., N\}$. Correspondingly $\boldsymbol{x} = \left(x_1, x_2, ..., x_{|\mathcal{I}|}\right)$ is a possible configuration of $\boldsymbol{X}$. For simplicity we assume $x_i \in \mathcal{L} = \{1, 2, ..., L\}$[1] and aim at associating a probability with all the feasible configurations. In the binary setting with $L = 2$, each configuration $\boldsymbol{x}$ can be equivalently represented by a set. More specifically, we define ground set $U = \{u_1, u_2, ..., u_{\mathcal{I}}\}$ with u_i corresponding to x_i, $\boldsymbol{x}$ is equivalently represented by $A_{\boldsymbol{x}} \cup_{i:\, x_i=2} \{u_i\}$. However, this method does not apply to multi-class settings where $|\mathcal{L}| \geq 3$. Let $V_i = \{v_{i,1}, v_{i,2}, ..., v_{i,L}\}$, we instead model with ground set $V = \cup_{i=1}^{N} V_i$ and utilize $A_{\boldsymbol{x}} = \cup_{i=1}^{N} \{v_{i,x_i}\}$ to represent configuration $\boldsymbol{x}$.

To filter subsets corresponding to valid configurations, we define set $\mathcal{M}$ which is the basis of a partition matroid (Welsh, 2010)[§2.1].

$$\mathcal{M} = \{B \subseteq V : \forall i \in \mathcal{I}, |B \cap V_i| = 1\}.$$

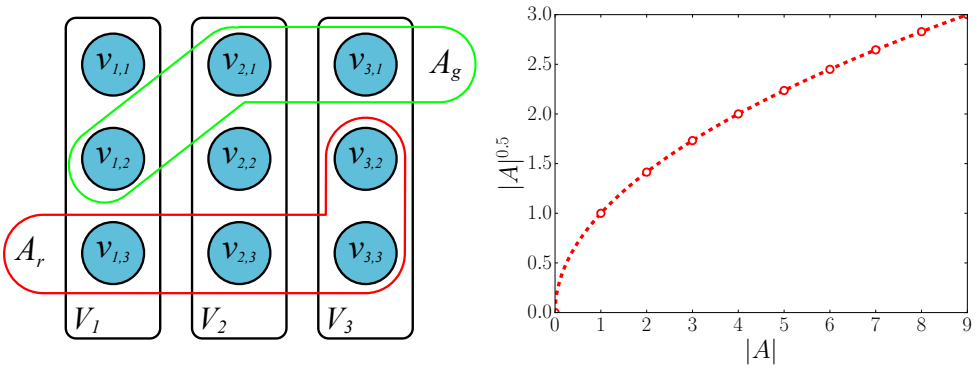
With function $F : 2^V \to \mathbb{R}$, we define the following probability for every $A \subseteq V$.

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

As a concrete example to illustrate the modeling methodology, we consider a pixel-wise labeling problem in Figure 3.1(a). Assume i is the index of a pixel while j is the index for a possible label. $x_i = j$ indicates pixel i is assigned with label j, which is equivalently represented by $v_{i,j} \in A$. A_g, which is represented by the region with green boundaries, assigns exactly one label to every pixel. Thus $A_g \in \mathcal{M}$. However, A_r assigns 2 labels to pixel 3 as shown with the region with green boundaries. It implies $A_r \notin \mathcal{M}$. We can verify set $\mathcal{M}$ includes exactly every feasible subset $A \in V$ which respects the fact that every pixel has exactly one label.

In most applications, the energy function is a sum over functions whose scopes are subsets of V, i.e. $F(A) = \sum_{r=1}^{R} F_r(A \cap V_r)$ with V_r as the scope of F_r. If we do not have any restrictions on F_r, our model essentially reflects a general graphical model represented by a factor graph. The complexity of updates in the corresponding message-passing algorithms is exponential with respect to $|V_r|$, i.e. the order of the interaction modeled by F_r. One possible way to get around the obstacle is to utilize submodularity. More specifically, we assume F_r to be submodular so that the probability $P(A)$ becomes a log-supermodular distribution.

[1] By replacing L with a proper integer function of i, our modeling approach also extends to cases where random variables have different numbers of possible values.



(a) A representation example.

(b) A concave cardinality function.

Figure 3.1: Examples for illustrating modeling approach.

3.2 Examples

In this section, we discuss three examples to illustrate our method in modeling pairwise and higher-order interactions. The first example is modeling Potts pairwise potential with cut function while the other two are about modeling label consistency and elements diversity which are both higher-order interactions.

Potts energy and cut function. Assume a graphical model is represented by graph $G(U, E)$, we define discrete random variable $x_i \in \mathcal{L} = \{1, 2, ..., L\}$, $\forall i \in U$. A pairwise potts energy is defined as

$$\phi(\boldsymbol{x}) = \sum_{(i,j)\in E} \phi_{i,j}(x_i, x_j) = \sum_{(i,j)\in E} \lambda_{i,j}[x_i \neq x_j].$$

Let $V_k = \{v_{k,1}, v_{k,2}, ..., v_{k,L}\}$ for $\forall k \in V$. For each component $\phi_{i,j}(x_i, x_j)$, we define submodular cut function

$$F_{i,j}(A) = \sum_{l\in\mathcal{L}} \frac{1}{2}\lambda_{i,j} |\mathbf{1}(v_{i,l} \in A) - \mathbf{1}(v_{j,l} \in A)|. \tag{3.1}$$

There is a corresponding set representation $A_{i,j}(x_i, x_j) = \{v_{i,x_i}, v_{j,x_j}\}$ for every configuration $\boldsymbol{x}$ with $F_{i,j}(A_{i,j}(x_i, x_j)) = \phi(x_i, x_j)$. We can verify that

$$\forall A_{i,j} \in \mathcal{M}_{i,j} = \{B_{i,j} \in V_i \cup V_j : |B_{i,j} \cap V_i| = 1,\ |B_{i,j} \cap V_j| = 1\},$$

i.e. $F_{i,j}(A_{i,j})$ is an equivalent representation of component $\phi_{i,j}$. As sum of submodular functions is still submodular, we have $F(A) = \sum_{(i,j)\in E} F_{i,j}(A \cap (V_i \cup V_j))$ is submodular. With the constraint $A \in \mathcal{M} = \cup_{(i,j)\in E}\mathcal{M}_{i,j} = \{B : |B \cap V_i| = 1,\ \forall i \in U\}$, we can then conclude submodular function $F(A)$ is equivalent to $\phi(\boldsymbol{x})$.

Higher-order consistency and concave cardinality function. Given a random vector $\boldsymbol{X} = (X_1, X_2, ..., X_\mathcal{I})$ with configuration components $x_i \in \mathcal{L}$, we want to model the prior that components in $\boldsymbol{x}$ tends to have the same label. With $v_{i,j}$ indicating $x_i = j$, we define $V_i = \cup_{l\in\mathcal{L}} \{v_{i,l}\}$ and $V^l = \cup_{i\in\mathcal{I}} \{v_{i,l}\}$. Based on a collection

of concave functions $g_l(x)$, each of which corresponds to a label class, we define the corresponding concave cardinality function as

$$g(A) = \sum_{l \in \mathcal{L}} g_l(|A \cap V^l|).$$

Function $g(A)$ tends to have smaller value when labels concentrate into a single class and larger value when labels scatter in different classes. As a concrete example with $|\mathcal{I}| = 9$ and $|\mathcal{L}| = 3$, we plot $g_l(|A \cap V^l|) = |A \cap V^l|^{0.5}$ in Figure 3.1b. When all the labels concentrate into a single class, we have $g(A) = N^{0.5} = 3$. If values of random variables are uniformly distributed into different classes, we have $g(A) = |\mathcal{L}| \, (N/|\mathcal{L}|)^{0.5} \approx 5.196 > 3$. Thus the concave cardinality function indeed advocates sharing the same label and models our higher-order consistency prior.

Diversity and group coverage function. Let $\{G_1, G_2, ..., G_N\}$ be a set of groups with G_i being subsets of ground set V. To measure the diversity of a set $A \subseteq V$, we define the following group coverage function

$$D(A) = |\{i \colon A \cap G_i \neq \emptyset\}|$$

which is easy shown submodular. It counts the number of groups covered by elements in set A. We can utilize $D(A)$ with our probabilistic framework as a prior to encourage less diverse sets, i.e. sets covering less groups.

4 Marginal inference

4.1 Formulation

Marginal inference reasons about the marginal probability of every random variable in graphical models. As exact marginal inference involves exponentially complex marginalization and is $\#P$-hard in the general setting, approximated inference is the way to get around the obstacle. Our approximated inference formulation for log-supermodular models is a variational one. We approximate submodular the function F with modular function a s and associate with $\boldsymbol{s}$ an approximated probability $\hat{P}(A) = \frac{1}{\hat{\mathcal{Z}}}\exp(-\boldsymbol{s}(A))$. Under the constraint that $\boldsymbol{s}(A) \in B(F)$, the log-partition functions of $P(A)$ and $\hat{P}(A)$ respects

$$\log \mathcal{Z} = \log \sum_{A\in\mathcal{M}} \exp^{-F(A)} \leq \log \sum_{A\in\mathcal{M}} \exp^{-\boldsymbol{s}(A)} = \sum_{i\in\mathcal{I}} \log \sum_{j\in\mathcal{L}} \exp^{-s_{i,j}} = \log \hat{\mathcal{Z}}$$

The equality is a consequence of the fact that $\sum_{A\in\mathcal{M}} \exp^{-\boldsymbol{s}(A)} = \prod_{i\in\mathcal{I}} \sum_{j=1}^{L} \exp^{-s_{i,j}}$. The inequality results from $\boldsymbol{s}(A) \leq F(A)$ because $\boldsymbol{s} \in B(F)$. Our inference formulation utilizes the relation and minimizes a upper bound of the exact log partition function. More specifically, we use the following program

$$\min_{\mathbf{s}\in B(F)} \sum_{i\in\mathcal{I}} \log \sum_{j\in\mathcal{L}} \exp^{-s_{i,j}} \tag{4.1}$$

to find the optimal modular function which minimizes the gap between log partition functions of the exact and approximated probability. With the optimal solution $\boldsymbol{s}^*$, approximated marginal probability can be computed efficiently with

$$\hat{P}(s_{i,j} \in A) = \frac{\exp(-s_{i,j})}{\sum_{j\in\mathcal{L}} \exp(-s_{i,j})}, \forall i \in \mathcal{I}, j \in \mathcal{L}.$$

Fenchel Duality. Interestingly, Equation (4.1) can also be interpreted as entropy maximization which is regularized by Lovasz extension. This fact becomes clear by considering the Fenchel dual of the marginal inference problem. With $\boldsymbol{w}_i \in \mathbb{R}^{|V_i|}$ and $\boldsymbol{w}$ being the concatenation of all $\boldsymbol{w}_i$, we now denote $f(\boldsymbol{w}) = \max_{\boldsymbol{s}\in B(F)} \langle \boldsymbol{w}, \boldsymbol{s} \rangle$ as the Lovasz extension of F, $H_i(\boldsymbol{w}_i)$ as the entropy of a multinomial distribution and $\Delta_i = \left\{ \boldsymbol{w}_i \in \mathbb{R}^{|V_i|} \colon \boldsymbol{w}_i \geq 0, \mathbf{1}^T \boldsymbol{w}_i = 1 \right\}$ as the probabilistic simplex. Using the above notations, the Fenchel dual problem is derived in Claim 1. Assume $(\boldsymbol{s}^\star, \boldsymbol{w}^\star)$ is a optimal primal/dual pair, the dual optimum $\boldsymbol{w}^\star$ is exactly our desired approximated marginal associated with primal optimum $\boldsymbol{s}^\star$. The proof of Claim 1 is delayed as an special case of Claim 6.

Claim 1. *The Fenchel dual of the marginal inference problem in Equation (4.1) is*

$$\begin{aligned} \min_{\boldsymbol{w}\in\mathbb{R}} \quad & f(\boldsymbol{w}) - \sum_{i\in\mathcal{I}} H_i(\boldsymbol{w}_i) \\ s.t. \quad & \boldsymbol{w}_i \in \Delta_i \end{aligned} \tag{4.2}$$

Zero duality gap is achieved at pair $(\boldsymbol{s}^\star, \boldsymbol{w}^\star)$ *if and only if* $\langle \boldsymbol{s}^\star,\ \boldsymbol{w}^\star \rangle = f(\boldsymbol{w}^\star)$ *and* $w^\star_{i,j} = \exp(-s^\star_{i,j}) / \sum_{j \in \mathcal{L}} \exp(-s^\star_{i,j}),\ \forall i \in \mathcal{I}, j \in \mathcal{L}$.

4.2 Algorithms

In the following, we assume the energy function is a sum over functions whose scopes are subsets of the ground set V, i.e. $F(A) = \sum_{r=1}^{R} F_r(A_r)$ with A_r as the scope of F_r. This setting enables us to demonstrate the parallelization of our algorithms.

Inference with Frank-Wolfe. As evaluating Lovász extension, i.e. solving linear programs over a base polytope, requires only $O(|V| \log |V|)$ computation, Frank-Wolfe algorithm is a natural choice for convex optimization over the base polytope. It repeatedly solves linear programs over the base polytope with a $O(1/k)$ convergence rate (Jaggi, 2013). In addition, the base polytope of a sum over submodular functions is the Minkowski sum of the corresponding base polytopes (Fujishige, 2005)[§4.2]. As shown in Claim 2, in each iteration, we can divide the linear programming problem into multiple smaller problems and solve them in parallel. The global optimum can be acquired by simply summing up optima from the smaller problems. These facts give rise to an efficient parallel solver in Algorithm 1 for problems in the sum-of-function setting.

Algorithm 1 Parallel Inference via Frank-Wolfe

1: Input $F = \sum_{r=1}^{R} F_r$, $g(\boldsymbol{s}) = \sum_{i \in \mathcal{I}} \log \sum_{j \in \mathcal{L}} \exp^{-s_{i,j}}$
2: Initialize $\boldsymbol{s} = \boldsymbol{s}_0 \in B(F)$
3: **for** $k = 1 : M$ **do**
4: $\quad \boldsymbol{x}_r = \operatorname{argmin}_{\boldsymbol{y} \in B(F_r)} \langle \nabla g(\boldsymbol{s}), \boldsymbol{y} \rangle$ in parallel for r
5: $\quad \boldsymbol{x} = \sum_{r=1}^{R} \boldsymbol{x}_r$
6: $\quad \boldsymbol{s} = \boldsymbol{s} + \gamma\,(\boldsymbol{x} - \boldsymbol{s})$ with $\gamma = 2/\,(k+2)$
7: **end for**
8: **return** $\hat{P}(s_{i,j} \in A) \propto \exp(-s_{i,j})$

Claim 2. *Let* $\boldsymbol{c} \in \mathbb{R}^{|V|}$ *and* $\hat{\boldsymbol{x}}_r = \operatorname{argmin}_{\boldsymbol{y} \in B(F_r)} \langle \boldsymbol{c}, \boldsymbol{y} \rangle$, $\hat{\boldsymbol{x}} = \sum_{r=1}^{R} \hat{\boldsymbol{x}}_r$ *is an optimum of* $\min_{\boldsymbol{y} \in B(F)} \langle \boldsymbol{c}, \boldsymbol{y} \rangle$

Proof. Let $B(F)$ be the base polytope of $F = \sum_{r=1}^{R} F_r$ and $B(F_r)$ is the base polytope of submodular components F_r. As proven in (Fujishige, 2005)[§4.2], $B(F)$ is the Minkowski sum of the base polytopes $B(F_r)$. Assume $\boldsymbol{x}^\star = \operatorname{argmin}_{\boldsymbol{y} \in B(F)} \langle \boldsymbol{c}, \boldsymbol{y} \rangle$, on the one hand, we know $\hat{\boldsymbol{x}} \in B(F)$ and thus $\langle \boldsymbol{c},\ \hat{\boldsymbol{x}} \rangle \geq \min_{\boldsymbol{y} \in B(F)} \langle \boldsymbol{c}, \boldsymbol{y} \rangle = \langle \boldsymbol{c},\ \boldsymbol{x}^\star \rangle$. On the other hand, $\boldsymbol{x}^\star$ can be decomposed as $\boldsymbol{x}^\star = \sum_{r=1}^{R} \boldsymbol{x}^\star_r$ with $\boldsymbol{x}^\star_r \in B(F)$. As $\langle \boldsymbol{c},\ \boldsymbol{x}^\star_r \rangle \geq \langle \boldsymbol{c},\ \hat{\boldsymbol{x}}_r \rangle$, we have $\langle \boldsymbol{c},\ \boldsymbol{x}^\star \rangle \geq \langle \boldsymbol{c},\ \hat{\boldsymbol{x}} \rangle$. It implies that $\langle \boldsymbol{c},\ \hat{\boldsymbol{x}} \rangle = \langle \boldsymbol{c},\ \boldsymbol{x}^\star \rangle$ and $\hat{\boldsymbol{x}}$ is an optimum of $\min_{\boldsymbol{y} \in B(F)} \langle \boldsymbol{c}, \boldsymbol{y} \rangle$. □

Inference with soft-move-making. Move-making algorithms such as α-β-swap and α-expansion (Boykov et al., 2001) have been extensively applied for MAP inference. α-expansion expands or reduces the region associated with a single label in

each iteration. Our soft-move-making approach is a block coordinate descent algorithm sharing the same intuition with α-expansion. We define $\boldsymbol{s}^l$ as a vector collecting $s_{i,l}, \forall i \in \mathcal{I}$ and $\boldsymbol{s}$ is actually the concatenation of $\boldsymbol{s}^l$. Correspondingly, V^l is defined as $V^l = \cup_{i\in\mathcal{I}} \{v_{i,l}\}$. If F can be decomposed as $F(A) = \sum_{l\in\mathcal{L}} F^l(A \cap V^l)$, we can derive an efficient block coordinate descent algorithm for our inference task. In Algorithm 2, we iteratively descent over $\boldsymbol{s}^l$ for all the possible values of l. Our sub-problem in each iteration is proven to be equivalent to min-norm problem (Djolonga & Krause, 2015) which can be efficiently solved via the Divide-and-Conquer algorithm in (Bach, 2011)[§9.1]. We derived the min-norm form of our block coordinate descent sub-problem in Claim 3. The subproblem solver can be parallelized with standard dual decomposition technique as described in (Jegelka et al., 2013).

Algorithm 2 Parallel Inference via Soft-move-making

1: Input submodular F, set $\mathcal{L}$ of all possible labels
2: Initialize $\boldsymbol{s}^l = 0, \forall l = 1, 2, ..., L$
3: **repeat**
4: **for** $l = 1 : |\mathcal{L}|$ **do**
5: Update $\boldsymbol{t}^l$ with $t^l_i = \log \sum_{j\neq l} \exp^{-s_{i,j}}$
6: $\boldsymbol{s}^l = \underset{\boldsymbol{r}\in B(F^l)}{\operatorname{argmin}} \frac{1}{2}||\boldsymbol{r} - \boldsymbol{t}^l||^2$
7: **end for**
8: **until** Convergence
9: **return** $\hat{P}(s_{i,j} \in A) \propto \exp(-s_{i,j})$

Claim 3. *Let $F^l \colon 2^{V^l} \to \mathbb{R}$ be submodular functions with base polytopes $B(F^l)$. Assume $V^l \cap V^{l'} = \emptyset$ if $l \neq l'$ and $F(A) = \sum_{l\in\mathcal{L}} F^l(A \cap V^l)$, the block coordinate descent sub-problem of the program in Equation (4.1) is equivalent to*

$$\min_{\boldsymbol{r}\in B(F^l)} \frac{1}{2}||\boldsymbol{r} - \boldsymbol{t}^l||^2 \tag{4.3}$$

where $\boldsymbol{t}^l$ collects $t^l_i = \log \sum_{j\neq l} \exp^{-s_{i,j}}, \forall i \in \mathcal{I}$.

Proof. We first prove $B(F)$ is the Cartesian product of $B(F^l)$. For $\forall A \in V$, we define $\hat{F}^l(A) = F^l(A \cap V^l)$ as the extension of F^l from 2^{V^l} to 2^V with $B(\hat{F}^l)$ as the corresponding base polytope. From the definition of base polytopes, $B(\hat{F}^l)$ is a polyhedron in $\mathbb{R}^{|V|}$ while $B(F^l)$ is in $\mathbb{R}^{|V^l|}$. For $\forall \hat{\boldsymbol{s}} \in B(\hat{F}^l)$, $\hat{s}_{i,j} \leq \hat{F}^l(\{v_{i,j}\}]) = F^l(\emptyset) = 0$ when $j \neq l$. On the other hand, $\hat{s}(V \setminus V^l) + \hat{s}(V^l) = \hat{F}^l(V) = F^l(V^l)$. If for any $j \neq l$, $\hat{s}_{i,j} < 0$, we have $\hat{s}(V^l) > \hat{F}^l(V) = F^l(V^l)$, which contradicts the constraint $\hat{s}(V^l) \leq \hat{F}^l(V^l) = F^l(V^l)$ in $B(\hat{F}^l)$. Thus for $\forall \hat{\boldsymbol{s}} \in B(\hat{F}^l)$ and $j \neq l$, $\hat{s}_{i,j} = 0$. In addition, as the base polytope $B(F)$ is the Minkowski sum of base polytopes $B(\hat{F}^l)$, we have $B(F)$ is the Cartesian product of $B(F^l)$, i.e. $B(F) = \sum_{l\in\mathcal{L}} B(\hat{F}^l) = \prod_{l\in\mathcal{L}} B(F^l)$.

As $B(F)$ is the Cartesian product of $B(F^l)$, the block coordinate descent subproblem is simply a constrained problem over $B(F^l)$ when s^l is variable. When updating $\boldsymbol{s}^l$ and fixing other variables, our formulation in Equation (4.1) turns into

$$\min_{\boldsymbol{s}^l\in B(F^l)} \sum_{i\in\mathcal{I}} \log\left(\exp\left(-t^l_i\right) + \exp\left(-s_{i,l}\right)\right) \tag{4.4}$$

where $\boldsymbol{t}^l$ collects $t_i^l = \log \sum_{j \neq l} \exp^{-s_{i,j}}, \forall i \in \mathcal{I}$. From Lemma 3 in (Djolonga & Krause, 2015), we can derive that the problem in Equation (4.3) shares the same optimum with the problem in Equation (4.4). □

5 Smoothed MAP inference

MAP inference gives point estimation on the (approximate) global optimal configuration of random variables in graphical models, i.e. the configuration maximizing the joint probability of all the random variables. In Section 5.1, we present a tractable continuous relaxation of the combinatorial MAP problem. We also discuss about a smoothing technique to improve efficiency, which is tightly connected to our marginal inference formulation. In Section 5.2, we analyze the interaction among convergence rate, approximation error and smoothing strength, presenting an efficiency-accuracy trade-off in the smoothing technique.

5.1 Relaxing and smoothing MAP inference

We first formulate the discrete MAP problem in Equation (5.1). It has an equivalent form stated in Claim 4. In order to design efficient algorithms for MAP inference, we switch to look at a continuous relaxation of the combinatorial problem in Claim 4, i.e. to replace the discrete constraints with $\Delta_i = \left\{\boldsymbol{w}_i \in \mathbb{R}^{|V_i|} \colon \boldsymbol{w}_i \geq 0, \mathbf{1}^T \boldsymbol{w}_i = 1\right\}$. To connect MAP problem with marginal inference formulation, we derive in Claim 5 the Fenchel Dual of the continuous relaxation. The proof of Claim 5 is presented with details in Appendix A.2

$$\begin{aligned} \min_{A \subset V} \quad & F(A) \\ s.t. \quad & |A \cap V_i| = 1, \quad \forall i \in \mathcal{I} \end{aligned} \tag{5.1}$$

Claim 4. *Let $f(\boldsymbol{w})$ be the Lovasz extension of submodular function $F(A)$, the problem in Equation (5.1) is equivalent to $\min_{\boldsymbol{w}_i \in \{0,1\}} f(\boldsymbol{w})$ with constraints $\mathbf{1}^T \boldsymbol{w}_i = 1$. The continuous relaxation of this equivalent problem is*

$$\begin{aligned} \min_{\boldsymbol{w} \in \mathbb{R}} \quad & f(\boldsymbol{w}) \\ s.t. \quad & \boldsymbol{w}_i \in \Delta_i \end{aligned} \tag{5.2}$$

Proof. For all A in constraint $\mathcal{M}$, we define $\boldsymbol{w}_A$ with $(\boldsymbol{w}_A)_{i,j} = 1$ when $v_{i,j} \in A$ and 0 otherwise. From Proposition 3.1 (f) in (Bach, 2011), we have $\forall A \in \mathcal{M},\ F(A) = f(\boldsymbol{w}_A)$. In addition, every feasible set A in Equation (5.1) has a corresponding $\boldsymbol{w}_A$ in Equation (5.3) and vice versa. With the equivalent feasible set and objective function, we can conclude the problem in Equation (5.1) and the one in Equation (5.3) are equivalent.

$$\begin{aligned} \min_{\boldsymbol{w}_i \in \{0,1\}} \quad & f(\boldsymbol{w}) \\ s.t. \quad & \mathbf{1}^T \boldsymbol{w}_i = 1 \quad \forall i \in \mathcal{I} \end{aligned} \tag{5.3}$$

By relaxing $\boldsymbol{w}_i \in \{0,1\}$ to $\boldsymbol{w}_i \in [0,1]$, we can derive the continuous relaxation in Equation (5.2). □

Claim 5. *The Fenchel dual of the continuous relaxation in Equation (5.2) is*

$$\min_{s\in B(F)} \sum_{i\in\mathcal{I}} \max_{j\in\mathcal{L}} (-s_{i,j}) \tag{5.4}$$

By comparing Equation (4.2) with Equation (5.2), we can find the additional entropy term turns the corresponding dual from the non-smooth problem in Equation (5.4) to the smooth one in Equation (4.1). Thus we can alternatively interpret marginal inference as smoothing the dual form of the relaxed MAP problem in Equation (5.4). However, the marginal inference formulation in Equation (4.2) lacks flexibility in controlling the smoothing strength. It naturally inspires the parametric formulation in Equation (5.5) of which the non-smooth relaxed MAP problem is an limit when $\epsilon \to 0$.

$$\begin{aligned} \min_{\boldsymbol{w}\in\mathbb{R}} \quad & f(\boldsymbol{w}) - \epsilon \sum_{i\in\mathcal{I}} H_i(\boldsymbol{w}_i) \\ s.t. \quad & \boldsymbol{w}_i \in \Delta_i \end{aligned} \tag{5.5}$$

To adapt our algorithms from marginal inference to the parametric formulation, we consider in Claim 6 the Fenchel dual of the program in Equation (5.5) when $\epsilon > 0$. With the formulation in Equation (5.6), we can directly apply Algorithm 1 and Algorithm 2 by dividing the input $F(A)$ with ϵ. For $\epsilon = 0$. We also present an algorithm based on Frank-Wolfe in Appendix A.1. As the hard-max function $\max_i s_i$ is a limit of the soft-max function $\epsilon \sum_i \log(s_i/\epsilon)$ when $\epsilon \to 0$, the non-smooth objective in Equation (5.4) is also an extreme case of the formulation in Equation (5.6) when $\epsilon \to 0$. It aligns well with our intuition that the non-smoothed problem should be a limit of the smoothed problem in both primal and dual domains.

Claim 6. *When $\epsilon > 0$, the parametric formulation in Equation (5.5) is the Fenchel dual of*

$$\min_{\boldsymbol{s}\in B(F)} \sum_{i\in\mathcal{I}} \epsilon \log \sum_{j\in\mathcal{L}} \exp^{-\frac{s_{i,j}}{\epsilon}} \tag{5.6}$$

Zero duality is achieved at $(\boldsymbol{s}^\star, \boldsymbol{w}^\star)$ if and only if $w^\star_{i,j} = \frac{\exp(-s^\star_{i,j}/\epsilon)}{\sum_{j\in\mathcal{L}} \exp(-s^\star_{i,j}/\epsilon)}$ and $\langle \boldsymbol{w}^\star, \boldsymbol{s}^\star \rangle = f(\boldsymbol{w}^\star)$.

Proof. From (Boyd & Vandenberghe, 2004)[Ex.3.25], the convex conjugate of $h_i(\boldsymbol{s}_i) = \log \sum_{j\in\mathcal{L}} \exp^{s_{i,j}}$ is

$$h_i^\star(\boldsymbol{p}_i) = -H_i(\boldsymbol{p}_i) = \begin{cases} \sum_{j\in\mathcal{L}} p_{i,j} \log p_{i,j} & \sum_{j\in\mathcal{L}} p_{i,j} = 1, p_{i,j} \geq 0 \\ +\infty & \text{otherwise} \end{cases} \tag{5.7}$$

It implies $\epsilon h_i^\star(-\boldsymbol{p}_i) = -\epsilon H_i(-\boldsymbol{p}_i)$ and $\epsilon h_i(-\frac{\boldsymbol{s}_i}{\epsilon})$ are convex conjugate to each other. For $i \neq j$, h_i and h_j are independent. Thus

$$h^\star(\boldsymbol{p}) = \sum_{i\in\mathcal{I}} \epsilon h_i^\star(-\boldsymbol{p}_i) = -\epsilon \sum_{i\in\mathcal{I}} H_i(-\boldsymbol{p}_i)$$

is the convex conjugate of

$$h(\boldsymbol{s}) = \sum_{i\in\mathcal{I}} \epsilon h_i(-\frac{\boldsymbol{s}_i}{\epsilon}) = \sum_{i\in\mathcal{I}} \epsilon \log \sum_{j\in\mathcal{L}} \exp^{-\frac{s_{i,j}}{\epsilon}}.$$

In addition, the convex conjugate of indicator $\boldsymbol{I}(\boldsymbol{s} \in B(F))$ is the Lovasz extension $f(\boldsymbol{p})$ (Bach, 2010a)[§3]. Thus $g(\boldsymbol{s}) = -\boldsymbol{I}(\boldsymbol{s} \in B(F))$ is the concave conjugate of Lovász extension $g^\star(\boldsymbol{p}) = -f(-\boldsymbol{p})$. From the Fenchel duality theorem in (Nedic et al., 2003)[Prop. 7.2.2], we have

$$\begin{array}{lll} \max\limits_{\boldsymbol{p}\in\mathbb{R}} & g^\star(\boldsymbol{p}) - h^\star(\boldsymbol{p}) & = \max\limits_{\boldsymbol{p}\in\mathbb{R}} \quad -f(-\boldsymbol{p}) + \epsilon \sum\limits_{i\in\mathcal{I}} H_i(-\boldsymbol{p}_i) \\ s.t. & -\boldsymbol{p}_i \in \Delta_i & s.t. \quad -\boldsymbol{p}_i \in \Delta_i \end{array} . \tag{5.8}$$

is the Fenchel dual problem of

$$\min_{\boldsymbol{s}\in\mathbb{R}^{|V|}} h(\boldsymbol{s}) - g(\boldsymbol{s}) \Leftrightarrow \min_{\boldsymbol{s}\in\mathbb{R}^{|V|}} \epsilon \sum_{i\in\mathcal{I}} h_i(-\frac{\boldsymbol{s}_i}{\epsilon}) + \boldsymbol{I}\left(\boldsymbol{s} \in B\left(F\right)\right) \tag{5.9}$$

According to Theorem 1 in (Rockafellar et al., 1966), as $g(\boldsymbol{p})$ is continuous on the whole domain and $\exists \boldsymbol{p}_0$ where $g(\boldsymbol{p}_0)$ and $h(\boldsymbol{p}_0)$ are both finite, strong duality holds for the primal-dual pair at some $(\boldsymbol{s}^\star,\ \boldsymbol{p}^\star)$. As proven in Theorem 2 in (Rockafellar et al., 1966), zero duality gap is achieved if and only if $\exists \boldsymbol{p}^\star \in \partial h(\boldsymbol{s}^\star) \cap \partial\left(-g(\boldsymbol{s}^\star)\right)$. As $h(\boldsymbol{s})$ is differentiable, we have $p^\star_{i,j} = -\exp(-\frac{s^\star_{i,j}}{\epsilon}) / \sum_{j\in\mathcal{L}}^{L} \exp(-\frac{s^\star_{i,j}}{\epsilon})$. To ensure $\boldsymbol{p}^\star \in \partial\left(-g(\boldsymbol{s}^\star)\right)$, we also need $\langle \boldsymbol{p}^\star,\ \boldsymbol{s}^\star \rangle = g(\boldsymbol{s}^\star) + g^\star(\boldsymbol{p}^\star)$, i.e. $\langle \boldsymbol{p}^\star,\ \boldsymbol{s}^\star \rangle = -f(-\boldsymbol{p}^\star)$.

By replacing $-\boldsymbol{p}$ with $\boldsymbol{w}$, we can verify the equivalence of the problem in Equation (5.5) and the one in Equation (5.8). The optimality condition is $\langle \boldsymbol{w}^\star,\ \boldsymbol{s}^\star \rangle = f(\boldsymbol{w}^\star)$ and $w^\star_{i,j} = \exp(-\frac{s^\star_{i,j}}{\epsilon}) / \sum_{j\in\mathcal{L}}^{L} \exp(-\frac{s^\star_{i,j}}{\epsilon})$. □

5.2 Accuracy-efficiency trade-off

For typical smoothing techniques, the stronger the smoothing term is, the faster the underlying solver usually converges. However, when the smoothing term is too strong, the optimum may be significantly different from the non-smooth optimum. E.g. for the program in Equation (5.5), the optimum will be biased to uniform distribution if entropy terms dominate. To rigorously analyze the balance together with the behavior of Frank-Wolfe based algorithm, we present the relation between convergence rate and ϵ in Claim 7 as well as the one between approximation error and ϵ in Theorem 1. On the one hand, if we evaluate the original relaxed MAP objective at the optima from original and smoothed relaxed problem, the difference is linearly bounded by ϵ. On the other hand, the convergence rate of Frank-Wolfe is linearly dependent on the reverse of ϵ. We present the proof of Claim 7 here and delay the one for Theorem 1 to Appendix A.3.

Empirically, we observe dramatic slowing down for Frank-Wolfe algorithm with very small ϵ. Non-differentiable MAP formulation is the case of $\epsilon \to 0$ which also suffers from the slowing down. In our analysis, we can see the the convergency rate $O(1/(\epsilon k))$ is inversely proportional to ϵ, which supports our empirical observations. Thus when using our framework for approximated MAP inference, optimality can be traded for better efficiency by tuning the smoothing strength with ϵ.

Claim 7. *Let ϵ be the parameter in Equation (5.5) and k be the number of iterations, the convergence rate of solving the program in Equation (5.5) with Frank-Wolfe Algorithm is $O\left(\frac{1}{\epsilon k}\right)$.*

Proof. We first consider the negative entropy term $h_\epsilon(\boldsymbol{w}) = -\epsilon \sum_{i\in\mathcal{I}} H_i(\boldsymbol{w}_i) = \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{L}} w_{i,j}\log w_{i,j}$ over domain $D = \{\boldsymbol{w}\colon 0 \le w_{i,j} \le 1,\ \forall i \in \mathcal{I},\ j \in \mathcal{L}\}$. As the hessian is

$$\boldsymbol{H}_\epsilon(\boldsymbol{w}) = \epsilon \begin{bmatrix} \boldsymbol{H}_1 & & & \\ & \boldsymbol{H}_2 & & \\ & & \ddots & \\ & & & \boldsymbol{H}_{|\mathcal{I}|} \end{bmatrix} \text{ with } \boldsymbol{H}_i = \begin{bmatrix} \frac{1}{w_{i,1}} & & & \\ & \frac{1}{w_{i,2}} & & \\ & & \ddots & \\ & & & \frac{1}{w_{i,|\mathcal{L}|}} \end{bmatrix},$$

we know that $\boldsymbol{H}_\epsilon(\boldsymbol{w}) \ge \epsilon \boldsymbol{I}$, i.e. $h_\epsilon(\boldsymbol{w})$ is ϵ-strongly convex in the interior of D. In other words, $\forall \boldsymbol{u}, \boldsymbol{v}$ in the interior of D, we have

$$h_\epsilon(\boldsymbol{u}) \ge h_\epsilon(\boldsymbol{v}) + \langle \nabla h_\epsilon(\boldsymbol{v}), \boldsymbol{u} - \boldsymbol{v}\rangle + \frac{\epsilon}{2}||\boldsymbol{u} - \boldsymbol{v}||_2^2.$$

With $\hat{\Delta}_i = \left\{\boldsymbol{w}_i\colon \mathbf{1}^T\boldsymbol{w} = 1, \boldsymbol{w}_i > 0\right\}$, we have $h_\epsilon(\boldsymbol{w})$ is ϵ-strongly convex over $\prod_{i\in\mathcal{I}}\hat{\Delta}_i$ which is a subset of the interior of D. Let $\hat{h}_\epsilon(\boldsymbol{w}) = h_\epsilon(\boldsymbol{w})$ with domain $\prod_{i\in\mathcal{I}}\hat{\Delta}_i$. From the definition of convex conjugate, we can derive the conjugate of $\hat{h}_\epsilon(\boldsymbol{w})$ is still

$$h_\epsilon^\star(\boldsymbol{s}) = \epsilon \sum_{i\in\mathcal{I}} \log \sum_{j\in\mathcal{L}} \exp^{s_{i,j}/\epsilon}$$

with domain $\mathbb{R}^{|V|}$. From (Borwein & Vanderwerff, 2010)[§5], we know that the conjugate of a ϵ-strongly convex function is $1/\epsilon$-smooth. In our setting, we have $h_\epsilon^\star(\boldsymbol{s})$ is $1/\epsilon$-smooth, i.e.

$$h_\epsilon^\star(\boldsymbol{y}) \le h_\epsilon^\star(\boldsymbol{x}) + \langle \nabla h_\epsilon^\star(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle + \frac{1}{2\epsilon}||\boldsymbol{y} - \boldsymbol{x}||_2^2 \quad \forall \boldsymbol{x},\ \boldsymbol{y} \in \mathbb{R}^{|V|}.$$

Considering the fact that $h_\epsilon^\star(-\boldsymbol{s})$ is still $1/\epsilon$-smooth, we have that the objective function in Equation (5.5) is $1/\epsilon$-smooth with respect to L_2 norm. Lemma 7 in (Jaggi, 2013) implies the curvature parameter for Algorithm 1 is $C_{h_\epsilon^\star} \le \text{diam}_{||\cdot||_2}\left(B(F)\right)^2/\epsilon$ for $h_\epsilon^\star(-\boldsymbol{s})$. Thus the convergence rate is $O(C_{g_\epsilon}/k) = O(\frac{1}{\epsilon k})$. □

Theorem 1. *Let $\boldsymbol{w}^\star(\epsilon)$ be the optimal solution of the program in Equation (5.5) with $\epsilon > 0$, and $\boldsymbol{w}^\star(0)$ is the optimal solution when $\epsilon = 0$, i.e., the optimal solution for the non-smooth relaxed MAP formulation. For arbitrary submodular function F with its Lovász extension being $f(\boldsymbol{w})$, we have*

$$f(\boldsymbol{w}^\star(\epsilon)) - f(\boldsymbol{w}^\star(0)) \le \epsilon|\mathcal{I}|\log|\mathcal{L}|.$$

And given $\forall\epsilon$, $\mathcal{I}$ and $\mathcal{L}$, we can construct a submodular function $\hat{F}$ with Lovász extension $\hat{f}(\boldsymbol{w})$ so that $\hat{f}(\boldsymbol{w}^\star(\epsilon)) - \hat{f}(\boldsymbol{w}^\star(0)) = \frac{1}{2}\epsilon|\mathcal{I}|\log(|\mathcal{L}|)$.

6 Experiments

6.1 Experiment setup

We evaluate our inference approach on MSRC-21 dataset for pixel-wise multi-class image segmentation. The MSRC-21 dataset contains images in portrait format of size 213×320 or landscape format of size 320×213. The image samples range from outdoor views to indoor scenes containing 21 classes of objects such as cow, car, book and human. As shown in Figure 6.1, the original dataset only has coarse-grain annotations with unlabeled ambiguous regions. The annotation typically inflates the foreground objects and does not preserve boundaries properly. In addition to the original dataset, a fine-grain annotation is created as part of the work in (Krähenbühl & Koltun, 2012). The new annotation covers 93 images from MSRC-21 dataset. It preserves the boundaries emerging from complex interactions among objects and produces pixel-wise labeling in high quality. In order to reliably evaluate our higher-order inference technique over the segmentation model, we run experiments on the subset of MSRC-21 with fine-grain annotations. For all our experiments, we use TextonBoost unary features designed in (Krähenbühl & Koltun, 2012). It uses 17-dimensional filter bank suggested by Shotton et al. (Shotton et al., 2009) and is augmented with color, pixel location and HOG features.

Higher-order segmentation model. Oversegmentation methods such as SLIC (Achanta et al., 2010) and mean-shift (Comaniciu & Meer, 2002) generate superpixels which are typically used to construct superpixel-wise model. The superpixel-wise models are more efficient as the scale is dramatically reduced from pixel-wise model. Although superpixel typically can not directly produce high-level semantic segmenta-

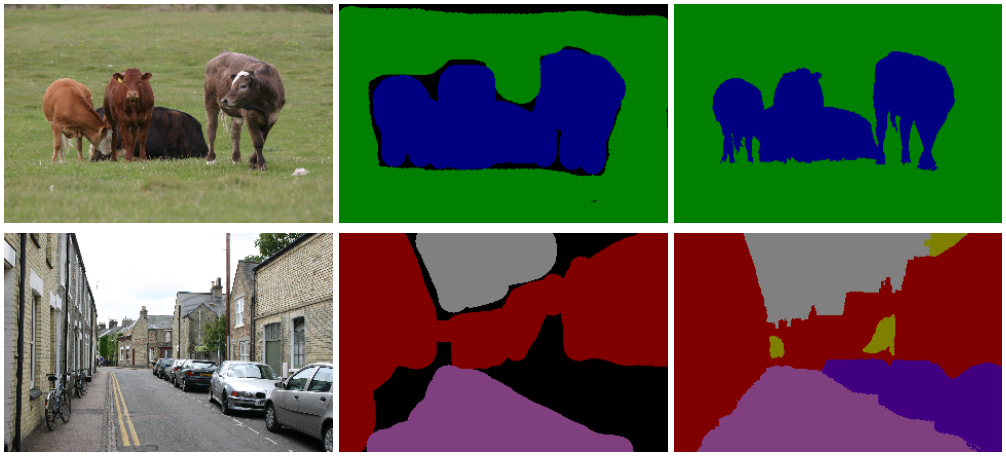


Figure 6.1: Samples from MSRC-21 dataset. Left: image. Middle: coarse-grain annotation. Right: fine-grain annotation.

tion, it generates oversegmentation from low-level image features like colors. As semantic boundaries usually align with boundaries of homogeneous regions, superpixels can be employed to enforce pixel-level label consistency within a single superpixel. This consistency prior usually emerges as interaction of extremely high order because a single superpixel may contains tens to thousands of pixels. Besides modeling conventional Potts pairwise interaction with cut function, we also utilize the higher-order prior in our experiments. More specifically, we generate multiple layers of superpixels with the mean-shift algorithm and softly enforce the label consistency of pixels in each single superpixel with concave cardinality fucntions. Let s_p be the spatial bandwidth parameter, s_r be the range bandwidth parameter and m_r be the minimum size of regions, we consider the following different configurations of our model:

- Submod$_{\text{pair}}$: A grid Potts pairwise model using cut function.

- Submod$_{\text{2-layer}}$: It generates 2 layers of superpixels with $(7, 4, 500)$ and $(7, 10, 100)$ for the parameter (s_p, s_r, m_r). Only higher-order consistency prior is utilized.

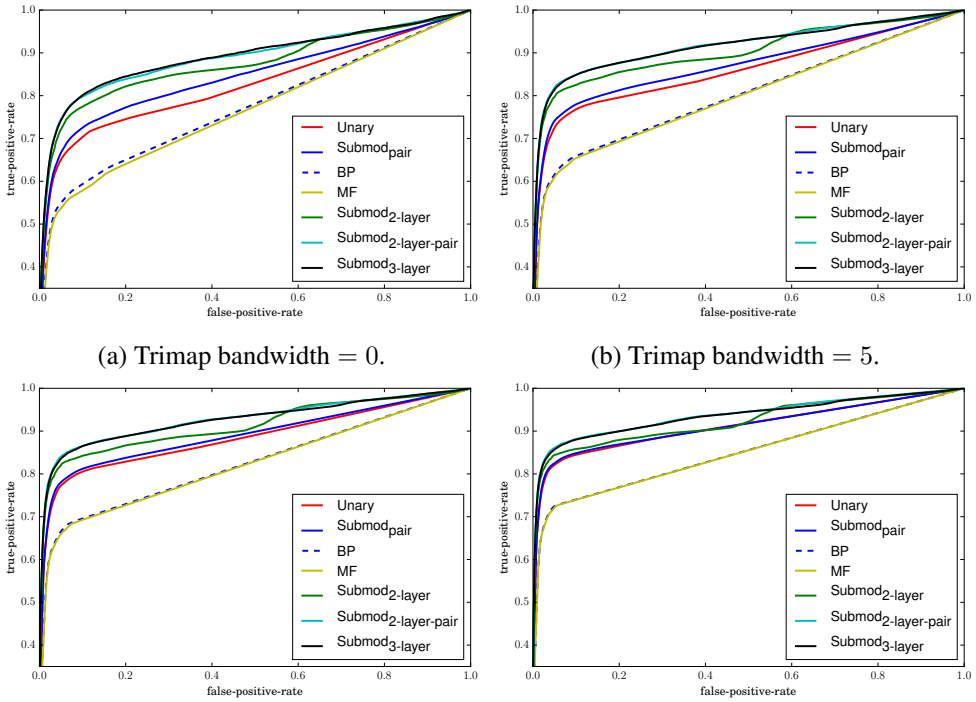
- Submod$_{\text{2-layer-pair}}$: It utilizes the same superpixel as Submod$_{\text{2-layer}}$. Higher-order consistency prior is employed together with pairwise interaction modeled with cut functions.

- Submod$_{\text{3-layer}}$: It generates 3 layers of superpixels with $(7, 4, 500)$, $(7, 7, 300)$ and $7, 10, 100$ as the parameter (s_p, s_r, m_r). Only higher-order consistency prior is utilized.

Along with our models, we do experiments with grid pairwise model with libdai (Mooij, 2010) Belief-Propagation and Mean-Field solver. In addition, we present performance of Robust-Pn model (Kohli et al., 2009) and CRF$_{\text{fully}}$ (Krähenbühl & Koltun, 2012) for MAP inference as reported in (Krähenbühl & Koltun, 2012). Algorithms for log-supermodular inference are parallelized with 4 threads on an Intel Core-i5 quad-core 3.2 GHz processor while those with libdai solvers run on a single core of the same processor. Using the submodular function in Section A.4 and the parameter grid shown in Table A.1, all the experiments are done in 5-fold cross-validation with grid search. The estimation are evaluted using finely annotated dataset of 93 samples. For Submod$_{\text{2-layer-pair}}$ which combines pairwise and higher-order interactions, we fix the pairwise parameter to the one most frequently selected in cross-validation experiment of Submod$_{\text{pair}}$ and then perform cross-validation on the grid of higher-order parameters.

6.2 Marginal inference evaluation

In order to demonstrate the performance in a finer-grain fashion, we adopt the trimap concept from (Kohli et al., 2009). A trimap with bandwidth h is the union of $(2h + 1) \times (2h + 1)$ neighborhoods of all the boundary pixels. As classification error usually happens around the boundaries while most of the pixels are non-boundary pixels, performance on trimap may helps better comparing the ability to distinguish classes. Thus we report results on trimaps with different bandwith.



(a) Trimap bandwidth $= 0$. (b) Trimap bandwidth $= 5$.

(c) Trimap bandwidth $= 10$. (d) Trimap bandwidth $= 20$.

Figure 6.2: Overall ROC curve over trimaps with different bandwidth.

Receiver operating characteristic. Receiver operating characteristic (ROC) is a plot illustrating the performance of binary classifiers. It measures the relationship between false-positive-rate and true-positive-rate. In our multi-class setting, we perform a macro average over ROC curves of each class. Specifically, we first generate ROC curves per class in a 1-vs-all binary setting. The overall ROC curve is then produced by averaging those curves. Together with the overall ROC curve, we use the area under the curve (AUC) to illustrate the quality of marginal probabilities. Higher AUC value typically corresponds to classifiers with higher true-positive-rate at the cost of a certain fixed false-positive-rate. It implies better marginal probability usually produces higher AUC values. As demonstrated in Figure 6.2 and in Table 6.1, both our pairwise and higher-order model improve AUC from Unary model. For different trimap bandwidth, either Submod$_{3\text{-layer}}$ or Submod$_{2\text{-layer-pair}}$, which are both higher-order, achieves the best AUC among all the models. We can observe that all the higher-order models dramatically improve AUC from Unary while pairwise model Submod$_{pair}$ gains little over Unary. We can also notice that MF$_{pair}$ and BP$_{pair}$ degenerate the ROC curve comparing with Unary.

KL divergence. KL-divegence is a conventional measure of the similarity between two probabilities. We also evaluate the pixel-wise average KL divergence $D_{KL}(q||p)$ between the estimated marginal p and the ground truth marginal q. The ground truth marginal is a vector with $0/1$ entries where a single 1 indicates the ground truth label. In Figure 6.3, we can observe all the higher-order models with our log-supermodular inference engine uniformly outperform unary and any pairwise models. Comparing with pairwise model Submod$_{pair}$, our higher-order models achieves substantial de-

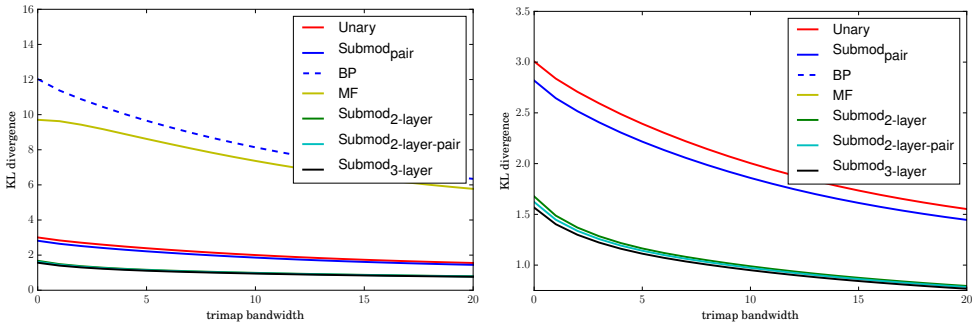
bandwidth	0	5	10	20
Unary	0.8254	0.8602	0.8841	0.9112
$\text{Submod}_{\text{pair}}$	0.8443	0.8708	0.8905	0.9119
BP_{pair}	0.7727	0.8034	0.8245	0.8504
MF_{pair}	0.7663	0.8006	0.8226	0.8499
$\text{Submod}_{\text{2-layer}}$	0.8735	0.9035	0.9132	0.9233
$\text{Submod}_{\text{2-layer-pair}}$	0.8886	**0.9184**	**0.9278**	**0.9371**
$\text{Submod}_{\text{3-layer}}$	**0.8904**	0.9173	0.9264	0.9355

Table 6.1: AUC for macro-average ROC w.r.t. bandwidth of trimap.

bandwidth	0	5	10	20
Unary	3.01	2.39	2.00	1.55
$\text{Submod}_{\text{pair}}$	2.82	2.22	1.86	1.45
BP_{pair}	12.03	9.66	8.14	6.34
MF_{pair}	9.71	8.62	7.37	5.77
$\text{Submod}_{\text{2-layer}}$	1.68	1.16	0.99	0.79
$\text{Submod}_{\text{2-layer-pair}}$	1.62	1.14	0.97	0.78
$\text{Submod}_{\text{3-layer}}$	**1.57**	**1.11**	**0.95**	**0.77**

Table 6.2: Pixel-wise KL divergence w.r.t bandwidth of trimap.

crease of KL divergence from Unary. As the marginals from BP_{pair} and MF_{pair} are substantially more confident than those from other models, an marginal from BP_{pair} or MF_{pair} with its peak in a wrong class will contribute larger KL divergence. Thus we again observe degenerated performance of BP_{pair} and MF_{pair} comparing with Unary in Table 6.2 and Figure 6.3.



(a) Pixel-wise KL divergence between estimated and ground truth marginal of all the models.

(b) Pixel-wise KL divergence between estimated and ground truth marginal of all models based on log-supermodular inference solver.

Figure 6.3: Pixel-wise KL divergence w.r.t. bandwidth of trimaps.

Method	Pixel-wise accuracy	running time(s)
Unary	$83.71\% \pm 1.81\%$	–
$\text{Submod}_{\text{pair}}$	$83.92\% \pm 1.81\%$	12.58
BP_{pair}	$83.91\% \pm 1.81\%$	25.64
MF_{pair}	$83.83\% \pm 1.82\%$	203.53
$\text{Submod}_{\text{2-layer}}$	$88.55\% \pm 1.80\%$	12.53
$\text{Submod}_{\text{2-layer-pair}}$	$88.48\% \pm 1.68\%$	20.10
$\text{Submod}_{\text{3-layer}}$	$\mathbf{88.61\% \pm 1.70\%}$	15.86
$\text{CRF}_{\text{fully}}$[1]	$88.2\% \pm 0.7\%$	0.2
Robust-Pn[2]	$86.5\% \pm 1.0\%$	30

Table 6.3: Average Pixel-wise accuracy and average running times of the 93 samples. Note the error and running times of $\text{CRF}_{\text{fully}}$ and Robust-Pn are reported by Krähenbühl & Koltun (2012) and not from experiments on our machines.

6.3 MAP inference evaluation

Pixel-wise accuracy and running time. In this section, we evaluate our approach as smoothed MAP inference on the finely annotated subset of MSRC-21. We use smoothing strength $\epsilon = 1$ for models using our log-supermodular inference engine. In Table 6.3, we demonstrate the average accuracy with standard derivation in 5-fold cross-validation. Comparing with Unary model and all the grid pairwise ones, models with more complex interactions achieves approximately 3% to 5% improvement in pixel-wise accuracy. Among these complex models, $\text{Submod}_{\text{3-layer}}$ achieves the best result of 88.61%. Robust-Pn is the only higher-order baseline whose underlying max-flow engine is not easy to parallel. $\text{Submod}_{\text{2-layer}}$, $\text{Submod}_{\text{2-layer-pair}}$ and $\text{Submod}_{\text{3-layer}}$, which are easy to parallel, uniformly achieves 2% improvements over Robust-Pn. The running time of our inference approach is in the same magnitude with the one of Robust-Pn. However, our inference engine produces marginal and MAP estimation simultaneously while Robust-Pn only generates MAP results. $\text{CRF}_{\text{fully}}$ achieves achieves the second best results in a fraction of second. As filtering-based inference engine $\text{CRF}_{\text{fully}}$ only deal with gaussian pairwise interaction while log-supermodular engine supports higher-order models, the running time is not directly comparable. We can see that $\text{Submod}_{\text{pair}}$, BP_{pair} and MF_{pair} achieves similar accuracy over the same model. $\text{Submod}_{\text{pair}}$ naively assigns jobs to cores according to the index ordering of pairwise interactions. It can be substantially accelerated with a variant of greedy coloring in (Felzenszwalb & Huttenlocher, 2006), which assigns pairwise interactions to multiple cores and avoid memory race-conditions in parallelization. Note the running time of MF_{pair} varies dramatically with the parameters of pairwise interactions. For those parameters which produce good results, MF_{pair} runs approximately $10\times$ slower than $\text{Submod}_{\text{pair}}$ and BP_{pair}. In Figure 6.4, we demonstrate pixel-wise error with respect to the bandwidth of trimap. we can see error decreases when bandwidth increases and the error happens more frequently near semantic boundaries. It indeed shows improvement from higher-order priors is stronger when we evaluate over a trimap with smaller bandwidth.

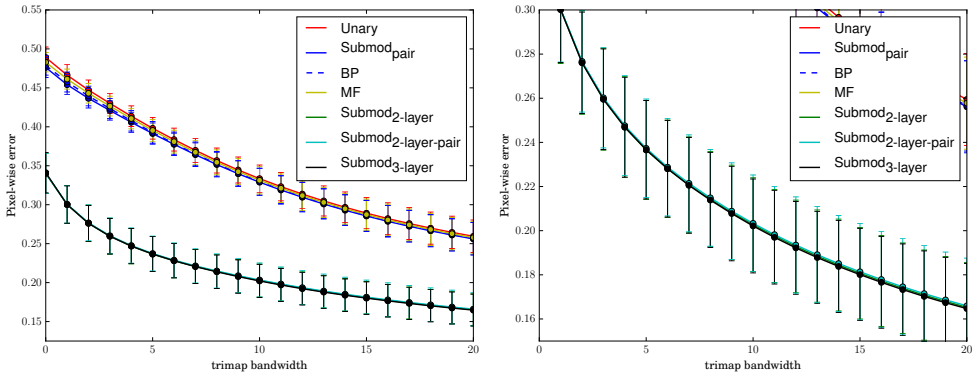
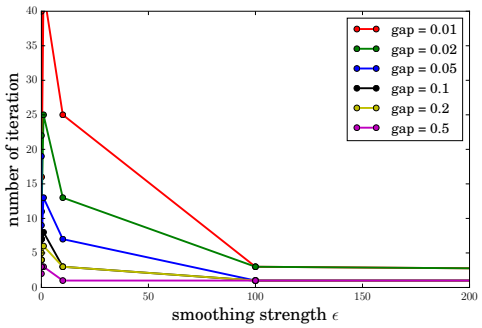
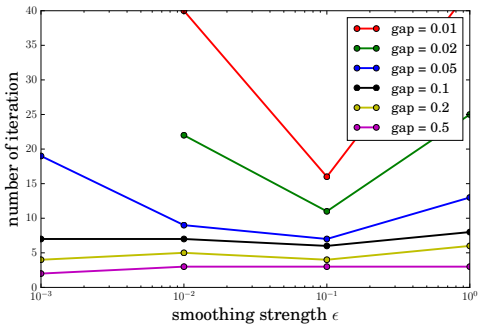


Figure 6.4: Pixel-wise error w.r.t. the bandwidth of trimap. Left: Comparison of all the models. Right: Zoom in for models based on log-supermodular inference engine.

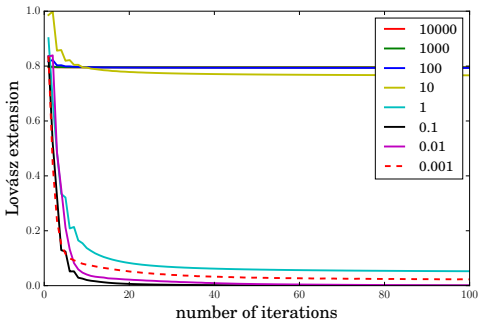
Efficiency-accuracy trade-off. To empirically verify our theoretical analysis on efficiency-accuracy trade-off, we monitor inference process of Submod$_{3\text{-layer}}$ model and do collective analysis over all of the 93 samples. As different samples produce monitored value from different energy, we need to first normalize them before averaging. For each single sample, we monitor its inference process with different ϵ. To normalize the primal gap, we pick the largest primal gap in the union of monitoring data of this sample for all ϵ. The picked largest gap is then rescaled to 1. In order to normalize the value of the Lovász extension, we pick for each sample the largest and smallest value of Lovász extension in the union of monitoring data. The largest value is rescaled to 1 while the smallest value to 0. To get the discrete value, we compute the dual value corresponding to the current primal point in Equation (5.6). As discussed in Claim 6, the dual variable corresponds to a marginal probability. We generate point estimation for every variable with the peak of this marginal probability. We normalize the discrete objective value using the same procedure for Lovász extension. The average curves are produced by averaging the individual normalized curves. Corresponding to Claim 7, we reported the number of iterations needed for the average primal gap curves to achieve a certain primal gap. Note we omit points if a certain primal gap is not yet reached by a curve after 100 iterations. In Figure 6.5a, we can observe the relationship roughly follows the shape of inverse function when ϵ is relatively large. However, in Figure 6.5b, the relationship between ϵ and needed number of iterations fluctuates. Conventional Frank-Wolfe convergence rate analysis treat the feasible set and objective function separately with the diameter of feasible set and curvature parameter (Jaggi, 2013). When smoothing strength becomes smaller, the interaction between objective and feasible set might contribute to curvature parameter. We leave finer-grain analysis as future work. According to Theorem 1, the optimum with smaller ϵ produces smaller Lovász extension value. As shown in Figures 6.5c and 6.5d, for those ϵ producing well converged curves, smaller ϵ value produce curve lower in their ending phase. Exceptions are $\epsilon = 0.001$ and $\epsilon = 0.01$. We believe it is because for these two cases, Frank-Wolfe algorithm is still not close enough to convergence. We also show discrete objective values along with the values of Lovász extension. The variation of discrete objective curves with varying ϵ aligns with the variation tendency of Lovász extension curves.



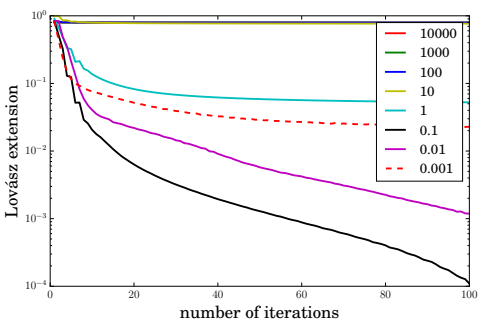
(a) Number of iterations needed to achieve certain primal gaps with different ϵ. Range of ϵ is limited from 0 to 200 for clearness.



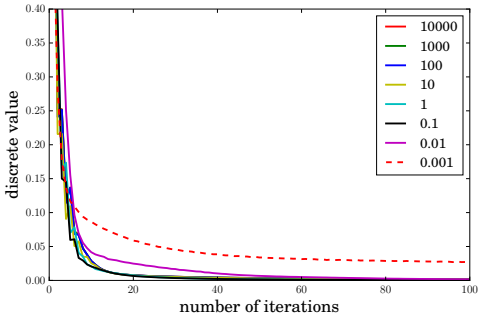
(b) The same curves with Figure 6.5a but only focus on very small ϵ values.



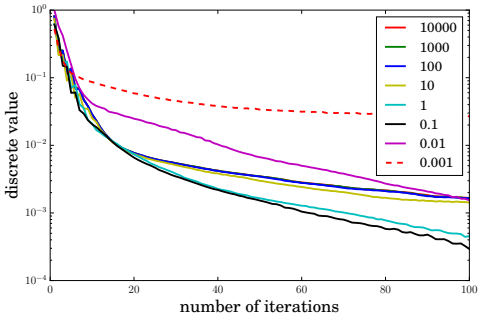
(c) Lovász extension value (linear scale) w.r.t. number of iterations for different ϵ.



(d) Lovász extension value (logarithm scale) w.r.t. number of iterations for different ϵ.



(e) Discrete objective value (linear scale) w.r.t. number of iterations for different ϵ.



(f) Discrete objective value (logarithm scale) w.r.t. number of iterations for different ϵ.

Figure 6.5: Experiment results on the efficiency-accuracy trade-off.

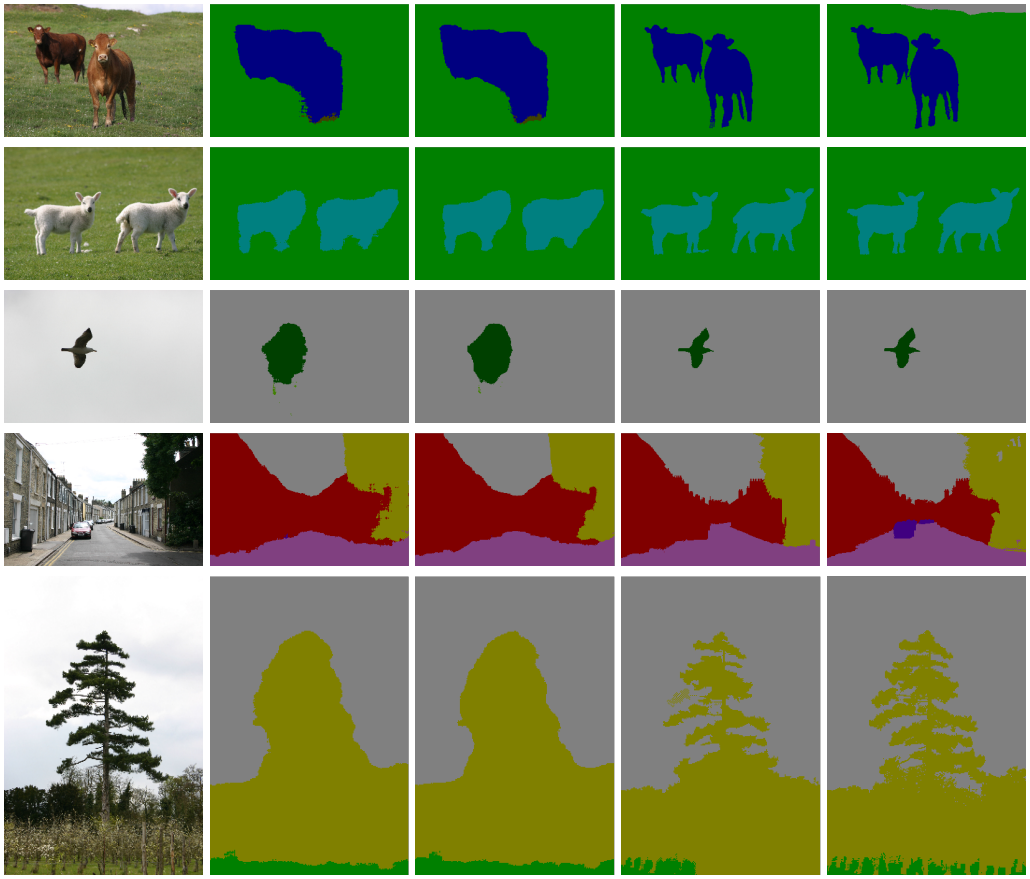


Figure 6.6: Qualitative MAP results.

6.4 Qualitative results

We visualize Qualitative MAP estimation from model Submod$_{\text{3-layer}}$ along with results from pairwise model Submod$_{\text{pair}}$ in Figure 6.6. Pairwise models are able to eliminating small noisy spots but can not smooth out relatively larger regions of noise. E.g. in the second row, the region of noise at the root of tree keeps there even with pairwise smoothing. On the contrary, these relatively larger regions can be eliminated by our higher-order consistency priors. As the oversegmentation from mean-shift usually preserves semantic boundaries in high quality, our higher-order priors is intuitively stronger than the local pairwise smoothing prior. Thus Submod$_{\text{3-layer}}$ can still produce qualitative segmentation even the unary potentials does not align well with object boundaries.

The entropy from marginal estimation of Submod$_{\text{3-layer}}$ as well as those from baselines Submod$_{\text{pair}}$ and BP$_{\text{pair}}$ are visualized in Figure 6.8. Entropy value is linearly rescaled so that the largest possible entropy equals the strongest brightness of a pixel. For Unary, high uncertainty with large entropy value typically lies at the boundary of its MAP estimation. Pairwise model BP$_{\text{pair}}$ produces high uncertainty only in a very narrow band around boundary. It is rather confident about regions other than the narrow band, which aligns with our analysis on the low AUC value for BP$_{\text{pair}}$.

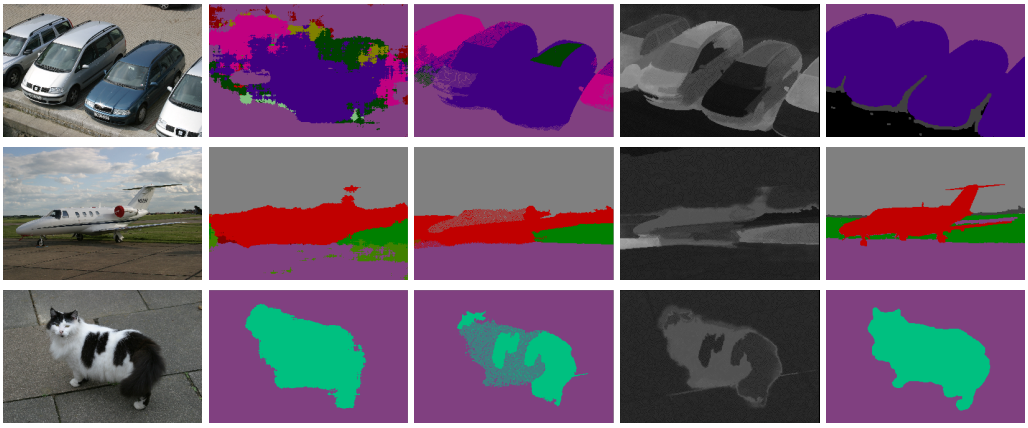


Figure 6.7: Hard examples with estimation results.

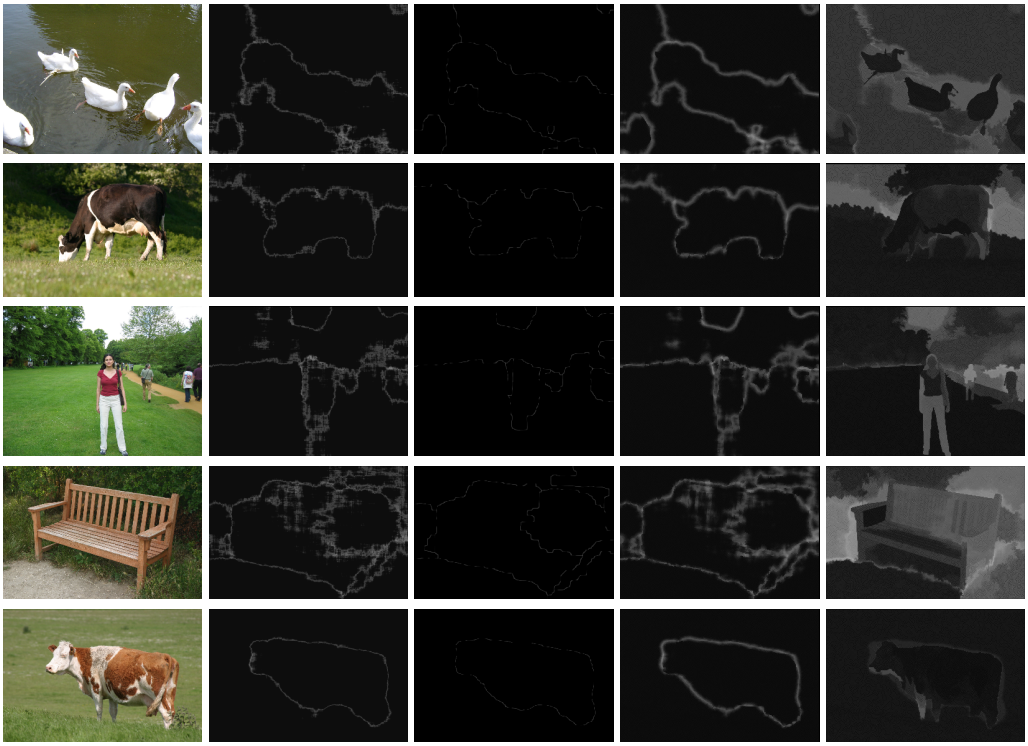


Figure 6.8: Comparison of entropy from Unary, $\text{Submod}_{\text{pair}}$ and $\text{Submod}_{\text{3-layer}}$.

However, $\text{Submod}_{\text{pair}}$ preserves uncertainty in a relatively wide band and eliminate uncertainty in regions corresponding to noise in unary terms. The uncertainty pattern of higher-order $\text{Submod}_{\text{3-layer}}$ comes in a different way. Homogeneous regions of objects usually have similar uncertainty level on pixels within the regions, which may even provide information about the existence of object or object parts.

Failure cases from $\text{Submod}_{\text{3-layer}}$ typically result from two reasons. The first

is strong unary terms which bias MAP to a wrong class such as the first example in Figure 6.7. The other reason is regions on different objects are grouped into a single superpixel due to color similarity. E.g. in the second example, the upper part of the plane and the neighboring sky are actually grouped into a single superpixel by mean-shift. Both of the reasons tend to produce noisy regions in MAP estimation. From the visualization, the upper part of plane can be visually separated as a region with high entropy. It implies the entropy may help to identify regions corresponding to objects or object parts even if MAP estimation fails on the region. One possible application of this object part separation ability lies in active learning setting. For example, machines may ask for label of these objects in crowd-sourcing.

7 Conclusion

In the thesis, we propose the log-supermodular model for multi-class probabilistic modeling with higher-order interactions. Variational Marginal inference in our model is formalized as a convex optimization problem over the base polytope. Associated with the formulation, a Frank-Wolfe-based algorithm and a soft-move-making algorithm are presented. Both of the algorithms are easily parallelized and highly efficient when the energy comes in the form of sum-of-submodular functions.

In order to simutaneously address MAP inference, we extend our marginal inference formulation to a parametric one. The extended version act as smoothed approximate MAP inference with a single parameter controling the smoothing strength. By analyzing the relation among convergence rate, approximation error and smoothing strength, we present an efficiency-accuracy trade-off in solving the smoothed MAP problem.

We evaluate our higher-order modeling approach and inference algorithms with semantic segmentation task over the MSRC-21 dataset. In comparison with multiple pairwise and higher-order baselines, our log-supmodular model achieves state-of-the-art performance in both marginal and MAP inference. We also present empirical analysis on the efficiency-accuracy trade-off in addition to our theoretical proofs. We believe our multi-class log-supermodular framework is one useful step towards statistically modeling complex dependencies in real world problems.

Future work. On the theoretical aspect, we are very interested in discussing the relationship between binary log-supermodular model in (Djolonga & Krause, 2014) and the special case of our model in binary setting, i.e. when $\mathcal{L} = 2$. As the exact MAP configuration of binary log-supermodular model can be recovered by thresholding the marginal, we are also eager to analyze whether we can derive exact MAP results from our multi-class model when $\mathcal{L} = 2$. If we can transform our model with $\mathcal{L} = 2$ to an equivalent binary log-supermodular model, our model may produce exact MAP estimation with similar thresholding techniques. On the empirical aspect, the comparison between Frank-Wolfe-based algorithm and soft-move-making algorithm is an interesting topic to explore. By analyzing the convergence speed and the accuracy of outputs, we might be able to provide practical guidelines in choosing algorithms for specific problems.

A Appendix

A.1 Algorithm for non-smooth relaxed MAP inference

In order to adapt our algorithm based on Frank-Wolfe to the non-smooth MAP problem, we only need to replace the gradient with subgradient which is derived in Claim 8. We can pick any point in the subdifferential as the subgradient utilized in our algorithm based on Frank-Wolfe.

Claim 8. *The subdifferential of* $\sum_{i\in\mathcal{I}} \max_{j\in\mathcal{L}}(-s_{i,j})$ *is a Cartesian product over multiple convex hulls, more specifically*

$$\partial \sum_{i\in\mathcal{I}} \max_{j\in\mathcal{L}}(-s_{i,j}) = \prod_{i\in\mathcal{I}} conv\{-\boldsymbol{e}_{i,j} \in \mathbb{R}^{|V_i|} : j \in \mathcal{M}_i(\boldsymbol{s}_i)\} \tag{A.1}$$

with $j \in \mathcal{M}_i(\boldsymbol{s}_i)$ *if and only if* $-s_{i,j} = \max_{k\in\mathcal{L}}(-s_{i,k})$. $\boldsymbol{e}_{i,j}$ *is the indicator vector with a single 1 in the entry corresponding to element* $v_{i,j}$.

Proof. For a function in the form of $\max_i f_i(\boldsymbol{s})$, the subdifferential can be expressed as

$$conv\{\nabla f_i(\boldsymbol{s}) | i \in \mathcal{M}(\boldsymbol{s})\}$$

with $i \in \mathcal{M}(\boldsymbol{s})$ if and only is $f_i(\boldsymbol{s}) = \max_j f_j(\boldsymbol{s})$. In our case, the subdifferential of each function component is $\partial \max_{j\in\mathcal{L}}(-s_{i,j}) = conv\{-\boldsymbol{e}_{i,j} | j \in \mathcal{M}_i(\boldsymbol{s}_i)\}$. As $\boldsymbol{s}_i$ is independent on $\boldsymbol{s}_j$ if $i \neq j$, the overall subdifferential is the Cartesian product of component-wise subdifferential. □

A.2 Proof of Claim 5

We first prove Lemma 1 to support the argument in proving Claim 5.

Lemma 1. $g_i(\boldsymbol{s}_i) = \max_{j\in\mathcal{L}}(-s_{i,j})$ *is the convex conjugate of the indicator funciton* $I(-\boldsymbol{w}_i \in \Delta_i)$ *where*

$$\Delta_i = \{\boldsymbol{p}_i | \sum_{j\in\mathcal{L}} p_{i,j} = 1, p_{i,j} \geq 0, \forall i \in \mathcal{I},\ j \in \mathcal{L}\} \tag{A.2}$$

Proof. Let $\boldsymbol{I}^\star(\boldsymbol{s})$ be the convex conjugate of $\boldsymbol{I}(-\boldsymbol{w} \in \Delta_i)$, from the definition of conjugate function, we have

$$\boldsymbol{I}^\star(\boldsymbol{s}_i) = \max_{\boldsymbol{w}_i\in\mathbb{R}^{|V_i|}} \langle \boldsymbol{s}_i,\ \boldsymbol{w}_i\rangle - \boldsymbol{I}(-\boldsymbol{w}_i \in \Delta_i) = \max_{\boldsymbol{w}_i\in\mathbb{R}^{|V_i|}} \langle -\boldsymbol{s}_i,\ -\boldsymbol{w}_i\rangle - \boldsymbol{I}(-\boldsymbol{w}_i \in \Delta_i).$$

The maximum can only be achieved when $-w_{i,j}$ is non-negative and sums to 1 because otherwise function $\boldsymbol{I}$ goes to $-\infty$. Assume $k \in \operatorname{argmax}_{j\in\mathcal{L}}(-s_{i,j})$, the maximum can be achieved by setting $w_{i,k}$ to 1 and other entries to 0, which gives $\boldsymbol{I}^\star(\boldsymbol{s}_i) = \max_{j\in\mathcal{L}}(-s_{i,j})$. □

Claim 5. *The continuous relaxation in Equation (5.4) is the Fenchel dual of*

$$\min_{s\in B(F)} \sum_{i\in\mathcal{I}} \max_{j\in\mathcal{L}} (-s_{i,j}) \tag{A.3}$$

Proof. We define $g(\boldsymbol{s}) = \sum_{i\in\mathcal{I}} g_i(\boldsymbol{s}_i)$. As $\forall i \neq j$, $g_i(\boldsymbol{s}_i)$ and $g_j(\boldsymbol{s}_j)$ are independent of each other, the convex conjugate of $g(\boldsymbol{s})$ is

$$g^\star(\boldsymbol{p}) = \sum_{i\in\mathcal{I}} \boldsymbol{I}(-\boldsymbol{p}_i \in \Delta_i) = \boldsymbol{I}(-\boldsymbol{p} \in \prod_{i\in\mathcal{I}} \Delta_i).$$

The problem in Equation (A.3) can be reformulated as unconstrained problem

$$\min_{\boldsymbol{s}\in\mathbb{R}} g(\boldsymbol{s}) - h(\boldsymbol{s})$$

where $h(\boldsymbol{s}) = -\boldsymbol{I}(\boldsymbol{s} \in B(F))$ is the indicator function of the base polytope. In addition, Lovasz extension $f(\boldsymbol{p})$ is the conjugate of indicator function $I\,(\boldsymbol{s} \in B(F))$ (Bach, 2010a)[Prop.8]. It implies the concave conjugate of $h(\boldsymbol{s}) = -\boldsymbol{I}(\boldsymbol{s} \in B(F))$ is $h^\star(\boldsymbol{p}) = -f(-\boldsymbol{p})$. Thus the Fenchel dual of the the problem in Equation (A.3) is

$$\max_{\boldsymbol{p}\in\mathbb{R}} h^\star(\boldsymbol{p}) - g^\star(\boldsymbol{p}) \Leftrightarrow \max_{\boldsymbol{s}\in\mathbb{R}} -f(-\boldsymbol{p}) + I\left(-\boldsymbol{p} \in \prod_{i\in\mathcal{I}} \Delta_i\right) \overset{\boldsymbol{w}=-\boldsymbol{p}}{\Leftrightarrow} \min_{\boldsymbol{w}_i\in\Delta_i} f(\boldsymbol{w})$$

. □

A.3 Proof of Theorem 1

Let $\boldsymbol{w}^\star(\epsilon)$ be the optimal solution of the program in Equation (5.5) with $\epsilon > 0$, and $\boldsymbol{w}^\star(0)$ is the optimal solution when $\epsilon = 0$, i.e. the optimal solution for the non-smooth relaxed MAP formulation. In order to prove Theorem 1, we first construct a resisting oracle for the upper bound of $f(\boldsymbol{w}^\star(\epsilon)) - f(\boldsymbol{w}^\star(0))$ with given ϵ, $|\mathcal{I}|$ and $|\mathcal{L}|$. With the oracle, we know there exists a submodular function such that

$$f(\boldsymbol{w}^\star(\epsilon)) - f(\boldsymbol{w}^\star(0)) = \frac{1}{2}\epsilon|\mathcal{I}|\log\left(|\mathcal{L}| - 1\right).$$

Lemma 2. *Given ϵ, $|\mathcal{I}|$ and $|\mathcal{L}|$, we define modular function $F(A) = s(A)\colon 2^V \to \mathbb{R}$ with $s^\star_{i,1} = -\epsilon\log\left(|\mathcal{L}| - 1\right)$ and $s^\star_{i,j} = 0$, $\forall j \neq 1$. Then $f(\boldsymbol{w}^\star(\epsilon)) - f(\boldsymbol{w}^\star(0)) = \frac{1}{2}\epsilon|\mathcal{I}|\log\left(|\mathcal{L}| - 1\right)$.*

Proof. As submodular function F is modular, we know $\boldsymbol{s}^\star$ is the only feasible point in $B(F)$. Thus it is also the optimum. According to Claim 6, we know strong duality holds and the optimal dual variable in Equation (5.5) is

$$w^\star_{i,j}(\epsilon) = \frac{\exp(-\frac{s^\star_{i,1}}{\epsilon})}{\sum_{k\in\mathcal{L}} \exp(-\frac{s^\star_{i,k}}{\epsilon})} = \begin{cases} \frac{1}{2} & \text{if } j = 1 \\ \frac{1}{2|\mathcal{L}|-2} & \text{otherwise} \end{cases}.$$

From the definition of Lovasz extension, we know $f(\boldsymbol{w}) = \langle \boldsymbol{s}^\star,\ \boldsymbol{w}\rangle$ is a linear function for modular functions F. Thus $f(\boldsymbol{w}^\star(\epsilon)) = \sum_{i\in\mathcal{I}} s^\star_{i,1} w^\star_{i,1}(\epsilon)$. As $w_{i,j} \geq 0$ and

sum to 1, we can also derive $f(\boldsymbol{w}^\star(0)) = \sum_{i\in\mathcal{I}} \min_{j\in\mathcal{L}} s_{i,j} = \sum_{i\in\mathcal{I}} s_{i,1}$, which gives

$$\begin{aligned} & f(\boldsymbol{w}^\star(\epsilon)) - f(\boldsymbol{w}^\star(0)) \\ =& \sum_{i\in\mathcal{I}} s^\star_{i,1} w^\star_{i,1}(\epsilon) - \sum_{i\in\mathcal{I}} s^\star_{i,1} \\ =& \frac{1}{2}\epsilon|\mathcal{I}|\log\left(|\mathcal{L}|-1\right) \end{aligned}$$

□

Theorem 1. *Let $\boldsymbol{w}^\star(\epsilon)$ be the optimal solution of the program in Equation (5.5) with $\epsilon > 0$, and $\boldsymbol{w}^\star(0)$ is the optimal solution when $\epsilon = 0$, i.e., the optimal solution for the non-smooth relaxed MAP formulation. For arbitrary submodular function F with its Lovász extension being $f(\boldsymbol{w})$, we have*

$$f(\boldsymbol{w}^\star(\epsilon)) - f(\boldsymbol{w}^\star(0)) \leq \epsilon|\mathcal{I}|\log|\mathcal{L}|.$$

And given $\forall\epsilon$, $\mathcal{I}$ and $\mathcal{L}$, we can construct a submodular function $\hat{F}$ with its Lovász extension as $\hat{f}(\boldsymbol{w})$ so that $\hat{f}(\boldsymbol{w}^\star(\epsilon)) - \hat{f}(\boldsymbol{w}^\star(0)) = \frac{1}{2}\epsilon|\mathcal{I}|\log\left(|\mathcal{L}|\right)$.

Proof. In the first step, we are to prove $f(\boldsymbol{w}^\star(\epsilon)) - f(\boldsymbol{w}^\star(0)) \leq \epsilon|\mathcal{I}|\log|\mathcal{L}|$. Under the probabilistic simplex constraints, $\boldsymbol{w}^\star(0) \in \operatorname{argmin} f(\boldsymbol{w})$ and $\boldsymbol{w}^\star(\epsilon) \in \operatorname{argmin} f(\boldsymbol{w}) - \epsilon\sum_{i\in\mathcal{I}} H_i(\boldsymbol{w}_i)$. We have

$$\begin{aligned} f(\boldsymbol{w}^\star(0)) \geq& f(\boldsymbol{w}^\star(0)) - \epsilon\sum_{i\in\mathcal{I}} H_i(\boldsymbol{w}^\star_i(0)) \\ \geq& f(\boldsymbol{w}^\star(\epsilon)) - \epsilon\sum_{i\in\mathcal{I}} H_i(\boldsymbol{w}^\star_i(\epsilon)) \\ \geq& f(\boldsymbol{w}^\star(\epsilon)) - \epsilon|\mathcal{I}|\log|\mathcal{L}| \end{aligned}$$

The last inequality is derived by maximizing entropy with uniform distribution, i.e. $\max_{\boldsymbol{w}_i\in\Delta_i} H_i(\boldsymbol{w}_i) = -\sum_{j\in\mathcal{L}} \frac{1}{|\mathcal{L}|}\log\frac{1}{|\mathcal{L}|}$. It implies $\epsilon|\mathcal{I}|\log|\mathcal{L}|$ is a valid upper bound of $f(\boldsymbol{w}^\star(\epsilon)) - f(\boldsymbol{w}^\star(0))$ with given ϵ, $|\mathcal{I}|$ and $|\mathcal{L}|$.

In Lemma 2, we construct a resisting oracle and prove that the exact value of $f(\boldsymbol{w}^\star(\epsilon)) - f(\boldsymbol{w}^\star(0))$ is $\frac{1}{2}\epsilon|\mathcal{I}|\log\left(|\mathcal{L}|-1\right)$, which proves the second statement. □

A.4 Details in experiments

Pairwise interaction modeling. Let $\boldsymbol{I}_i$ and $\boldsymbol{I}_j$ be the color feature of pixel i and j, we define $\lambda_{i,j} = \gamma\exp(\theta||\boldsymbol{I}_i - \boldsymbol{I}_j||^2/255^2)$ for modeling pairwise interaction with the cut fucntion in Equation (3.1).

Higher-order consistency modeling. Let V be the ground set and concave function $h(x) = \beta x^\alpha$. We model higher-order consistency with concave cardinality function

$$g(A) = h(|V\setminus A|) - h(|V|) = \beta|V\setminus A|^\alpha - \beta|V|^\alpha.$$

Parameter grid for grid searching. We use the parameter grid in Table A.1 for grid search in cross-validation

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

Table A.1: Parameter grid for cross-validation based on grid search.

Bibliography

Achanta, Radhakrishna, Shaji, Appu, Smith, Kevin, Lucchi, Aurelien, Fua, Pascal, and Süsstrunk, Sabine. Slic superpixels. Technical report, 2010.

Bach, Francis. Convex analysis and optimization with submodular functions: a tutorial. *arXiv preprint arXiv:1010.4207*, 2010a.

Bach, Francis. Learning with submodular functions: A convex optimization perspective. *arXiv preprint arXiv:1111.6453*, 2011.

Bach, Francis R. Structured sparsity-inducing norms through submodular functions. In *Advances in Neural Information Processing Systems*, pp. 118–126, 2010b.

Borwein, Jonathan M and Vanderwerff, Jon D. *Convex functions: constructions, characterizations and counterexamples*, volume 109. Cambridge University Press Cambridge, 2010.

Boyd, Stephen and Vandenberghe, Lieven. *Convex optimization*. Cambridge university press, 2004.

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

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 *Advances in Neural Information Processing Systems*, pp. 244–252, 2014.

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

Edmonds, Jack. Submodular functions, matroids, and certain polyhedra. *Edited by G. Goos, J. Hartmanis, and J. van Leeuwen*, pp. 11, 1970.

Felzenszwalb, Pedro F and Huttenlocher, Daniel P. Efficient belief propagation for early vision. *International journal of computer vision*, 70(1):41–54, 2006.

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

Fujishige, Satoru. *Submodular functions and optimization*, volume 58. Elsevier, 2005.

Hazan, Tamir and Shashua, Amnon. Convergent message-passing algorithms for inference over general graphs with convex free energies. *arXiv preprint arXiv:1206.3262*, 2012.

Jaggi, Martin. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pp. 427–435, 2013.

Jegelka, Stefanie, Bach, Francis, and Sra, Suvrit. Reflection methods for user-friendly submodular optimization. In *Advances in Neural Information Processing Systems*, pp. 1313–1321, 2013.

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

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

Krause, Andreas and Guestrin, Carlos E. Near-optimal nonmyopic value of information in graphical models. *arXiv preprint arXiv:1207.1394*, 2012.

Mooij, Joris M. 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. URL http://www.jmlr.org/papers/volume11/mooij10a/mooij10a.pdf.

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

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

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

Shotton, Jamie, Winn, John, Rother, Carsten, and Criminisi, Antonio. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *International Journal of Computer Vision*, 81(1):2–23, 2009.

Sun, Deqing, Liu, Ce, and Pfister, Hanspeter. Local layering for joint motion estimation and occlusion detection. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pp. 1098–1105. IEEE, 2014.

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

Valgaerts, Levi, Bruhn, Andrés, Zimmer, Henning, Weickert, Joachim, Stoll, Carsten, and Theobalt, Christian. Joint estimation of motion, structure and geometry from stereo sequences. In *Computer Vision–ECCV 2010*, pp. 568–581. Springer, 2010.

Vineet, Vibhav, Warrell, Jonathan, and Torr, Philip HS. 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.

Wainwright, Martin J and Jordan, Michael I. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2): 1–305, 2008.

Welsh, Dominic James Anthony. *Matroid theory*. Courier Corporation, 2010.

Woodford, Oliver, Torr, Philip, Reid, Ian, and Fitzgibbon, Andrew. Global stereo reconstruction under second-order smoothness priors. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(12):2115–2128, 2009.

Zhang, Jian, Kan, Chen, Schwing, Alexander G, and Urtasun, Raquel. Estimating the 3d layout of indoor scenes and its clutter from depth sensors. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pp. 1273–1280. IEEE, 2013.

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

ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

HIGHER-ORDER INFERENCE FOR MULTI-CLASS LOG-SUPERMODULAR MODELS

Authored by (in block letters):

For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**
ZHANG	JIAN

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date	**Signature(s)**
Zurich Apr. 15. 2015	Jian Zhang

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.

ETH-Bibliothek
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

ETH-Bibliothek
Digitale Bibliothek
ETH E-Collection
Raemistrasse 101
8092 Zurich
Fax +41 44 632 1357
www.ethbib.ethz.ch

Declaration of consent

For the publication of a diploma or master thesis on the ETH E-Collection institutional repository of ETH Zurich

ETH Zurich's institutional repository allows users to access diploma or master theses written at ETH Zurich. This offers academics the opportunity to present their work worldwide. The published theses fulfil the specified formal quality criteria.

Declaration of the author

I hereby declare that I consent to the ETH-Bibliothek making my diploma or master thesis, which I shall submit to the ETH-Bibliothek as a pdf file, available to the public on the institutional repository. The rights of third parties are not infringed by this publication. I consent to any subsequent necessary conversions into other data formats being made.

Author: Jian Zhang
Title of the publication: Higher-order Inference for Multi-class Log-supermodular Models
Department: Computer Science
Publication year: 2015
E-mail/Telephone number: zhangjianthu@gmail.com
Address: Kirchstrasse 5, Dietikon 8953

Zurich, on Apr. 15 2015 Signature Jian Zhang

Recommendation of the responsible ETH professor or research supervisor

I have supervised this diploma or master thesis at ETH Zurich and recommend its publication. In particular, I hereby declare that the publication of this thesis does not infringe the rights of third parties and that any rights to confidentiality are protected.

Zurich, on 15 Apr 2015 Signature Josip Djolonga

Declaration of the ETH-Bibliothek

The ETH-Bibliothek guarantees the long-term availability of the thesis as well as the integrity of its content and its authenticity and will make the document accessible via the Internet. The intellectual property rights remain with the author.