Guaranteed Parameter Estimation of Discrete Energy Minimization for 3D Scene Parsing

Mengtian Li

CMU-RI-TR-16-49
July 2016

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Daniel Huber, Advisor
Alexander J. Smola
David Fouhey

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Robotics.

Keywords: 3D scene parsing, energy minimization, complexity, NP-hard, APX, exp-APX, NPO, WCSP, min-sum, MAP MRF, QPBO, planar graph, structural learning, submodularity

For my parents

Abstract

Point clouds data, obtained from RGB-D cameras and laser scanners, or constructed through structural from motion (SfM), are becoming increasingly popular in the field of robotics perception. To allow efficient robot interaction, we require not only the local appearance and geometry, but also a higher level understanding of the scene. Such semantic representation is also necessary for as-built Building Information Model (BIM) creation and infrastructure inspection.

In this work, we present our discrete energy minimization based approach for 3D scene parsing. First, we contribute to the understanding of theoretical hardness of discrete energy minimization problems, which are also known as the MAP inference for MRF/CRFs. This theory explains why a previous scene parsing approach cannot have guaranteed optimality. Second, we propose a max-margin structural learning algorithm with performance guarantee. Finally, we demonstrate the performance and efficiency of our algorithm in the application of semantic labeling.

Acknowledgments

I would like to offer my special thanks to Dr. Daniel Huber, my advisor, for guiding me through various aspects of my research career. It is his patient guidance, enthusiastic encouragement and useful critiques that taught me to become a disciplined researcher.

I would also like to thank Dr. Alexander Shekhovtsov from Graz University of Technology (TU Graz), my coauther, without whose amazing work, our theoretical paper on energy minimization would not have been published.

I want to thank ARIA team members for an active and cooperative lab environments. In particular, I want to thank Zhizhong Li and Meng Song, for their wonderful work on 3D scene parsing, which my work is built upon, and creating the synthetic bridge dataset for the team.

I want give credits to Dr. Alexander J. Smola (MLD), Dr. Martial Hebert (RI), Dr. Venkatesan Guruswami (CSD), Euiwoong Lee (CSD), and Dr. David Fouhey (RI) for their valuable inputs to my papers.

I wish to thank my family, my friends, and my fellow students, for their kind support and for bringing much fun to my 2-year research life at CMU.

This material is based upon work supported by the National Science Foundation under Grant No. IIS-1328930 and by the European Research Council under the Horizon 2020 program, ERC starting grant agreement 640156.

Contents

1 Introduction

With the increased accessibility of 3D sensing, demand is increasing for efficient methods to transform 3D data into higher level, semantically relevant representations. The work of this thesis is part of a larger project with such a goal. This project is named the Aerial Robotic Infrastructure Analyst (ARIA). It aims to rapidly create comprehensive, high-resolution, semantically rich 3D models of infrastructure using unmanned aerial vehicles (UAVs). A semantic representation of the 3D scene is important for reverse engineering the infrastructure. Such semantic representation allows us to interact with the robot in a more natural way. For example, instead of sending precise coordinates to the drone, the operator can now instruct the drone to fly around a particular column and take some pictures of the abutment. Also, semantic representation is necessary for structural analysis, which analyzes the loads among different components of the bridge.

One of the key steps in the ARIA pipeline is 3D scene parsing, which is the focus of this work. Many of the most popular and successful 3D scene parsing algorithms can be reduced to some form of discrete energy minimization (or energy minimization for short) [3, 5, 29, 61, 76, 78, 82, 100]. One of the benefits of energy minimization methods is that they are able to capture contextual information or to encode prior knowledge. These capabilities are particularly important in complex 3D scene parsing, where local cues may be insufficient. For example, in the task of bridge component recognition (Figure 1), attached beams have similar appearance to connecting beams. The difference is that attached beams are usually beneath the deck and on top of connections whereas connecting beams are not. Therefore, to tell these

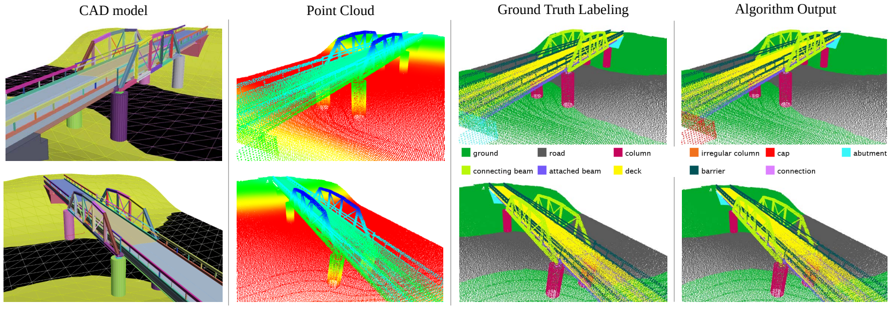


Fig. 1: Semantic labeling of a large-scale outdoor scene. We propose a generic structural learning algorithm with theoretical guarantees. When applied to scene parsing on the Cornell RGB-D dataset [49, 3], it runs three times faster than the competing method while keeping the same level of accuracy. On a larger scale problem of bridge component recognition, our algorithm solves the scene parsing problem intractable to previous methods. The point cloud dataset we created contains 11 domain-specific semantic class and is generated by merging several simulated LiDAR scans taken from multiple locations in the CAD model scene.

two classes apart, the scene parsing algorithms need to incorporate knowledge of how a bridge is typically built, which governs the spatial relationships of the components.

For another example, in 3D indoor scene parsing [100], coplanarity of two planes fitted on point clouds is a strong cue for them to be labeled as "wall." In contrast, the same coplanarity might not be useful if one of them is labeled as clutter. So the existence of certain features on a pair of nodes in the graph encourages certain joint labeling of the two nodes. These relationships can depend on the feature, the label configuration, and the particular edge. In order to encode the interactions, we need a parametrized energy function with a large parameter space [1]. An immediate question with such formulation is how to estimate these parameters autonomously.

Parameter estimation for energy minimization, also called structural learning, fails when the input data becomes large and complex, due to the intractable inference subroutine. Such intractability arises, for example, in 3D scene parsing of complex structures, where a scene can be composed of hundreds or thousands of objects with arbitrary connectivity. For these problems, it might not be possible to solve the inference subroutine exactly or even to approximate to a certain precision. However, the inference subroutine, or the separation oracle to be precise, plays the important role of finding the subgradients of the objective in a structural learning framework. Using unbounded approximation for the separation oracle generates imperfect gradients, causing the learning algorithm to fail, since the quality can be arbitrarily poor [25]. Two immediate questions would be why the separation oracle returns unbounded approximation and whether it would be possible to adopt a separation oracle with an exact solution or bounded approximation in polynomial time.

To answer these questions, we delved into the theory of energy minimization. Unfortunately, our theoretical investigation showed a negative result that the problem is inapproximable in polynomial time. Therefore, the above problem of imperfect gradients cannot be solved simply by adopting better inference subroutines. Such practice is common in structural learning, since the inference subroutine is usually treated as a modular "black box." We work around this intractable formulation issue by exploiting the properties of the joint problem of the overarching training and the inference subroutine. This enables us to propose a theoretically sound structural learning algorithm without the limitation of intractable inference.

We make three contributions in this thesis. First, we show the theoretical hardness of energy minimization. We contribute to this theory by proving that energy minimization, even in the pairwise 2-label case, and planar 3-label energy minimization are in general exp-APX-complete. This implies that these problems are inapproximable. As an auxiliary contribution, we summary existing complexity results of energy minimization in an unified framework. Next, we present our structural learning algorithm. Exploiting the property of the max-margin structural learning framework, we perform a series of binary submodular inferences to learn the weights in the energy for multi-class classification. Lastly, we demonstrate our algorithm's per-

[1] Note that the simple and popular smoothing prior model of energy minimization [17] is unable to capture such sophisticated interactions.

formance on the task of 3D scene parsing. Our algorithm runs much faster than our competing method, with little sacrifice to the accuracy, and is able to solve problems that are intractable to the competing methods.

Part I

2 Energy Minimization and the Complexity Theory

Discrete energy minimization, also known as min-sum labeling [97] or weighted constraint satisfaction (WCSP)[1] [34], is a popular model for many problems in computer vision, machine learning, bioinformatics, and natural language processing. In particular, the problem arises in maximum a posteriori (MAP) inference for Markov (conditional) random fields (MRFs/CRFs) [56]. In the most frequently used pairwise case, the *discrete energy minimization problem* (simply "energy minimization" hereafter) is defined as

$$\min_{x \in \mathcal{L}^{\mathcal{V}}} \sum_{u \in \mathcal{V}} f_u(x_u) + \sum_{(u,v) \in \mathcal{E}} f_{uv}(x_u, x_v), \tag{1}$$

where x_u is the label for node u in a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. When the variables x_u are binary (Boolean): $\mathcal{L} = \mathbb{B} = \{0, 1\}$, the problem can be written as a quadratic polynomial in x [14] and is known as quadratic pseudo-Boolean optimization (QPBO) [14].

In computer vision practice, energy minimization has found its place in semantic segmentation [69], pose estimation [102], scene understanding [76], depth estimation [58], optical flow estimation [101], image in-painting [80], and image denoising [11]. For example, tree-structured models have been used to estimate pictorial structures such as body skeletons or facial landmarks [102], multi-label Potts models have been used to enforce a smoothing prior for semantic segmentation [69], and general pairwise models have been used for optimal flow estimation [101]. However, it may not be appreciated that the energy minimization formulations used to model these vision problems have greatly varied degrees of tractability or *computational complexity*. For the three examples above, the first allows efficient exact inference, the second admits a constant factor approximation, and the third has no quality guarantee on the approximation of the optimum.

The study of complexity of energy minimization is a broad field. Energy minimization problems are often intractable in practice except for special cases. While many researchers analyze the time complexity of their algorithms (e.g., using big O notation), it is beneficial to delve deeper to address the difficulty of the underlying problem. The two most commonly known complexity classes are P (polynomial time) and NP (nondeterministic polynomial time: all decision problems whose solutions can be verified in polynomial time). However, these two complexity classes are only defined for *decision* problems. The analogous complexity classes for *optimization* problems are PO (P optimization) and NPO (NP optimization: all optimization

[1] WCSP is a more general problem, considering a bounded plus operation. It is itself a special case of valued CSP, where the objective takes values in a more general valuation set.

problems whose solution feasibility can be verified in polynomial time). Optimization problems form a superset of decision problems, since any decision problem can be cast as an optimization over the set $\{\text{yes, no}\}$, i.e., $\text{P} \subset \text{PO}$ and $\text{NP} \subset \text{NPO}$. The NP-hardness of an optimization problem means it is at least as hard as (under Turing reduction) the hardest decision problem in the class NP. If a problem is NP-hard, then it is not in PO assuming $\text{P} \neq \text{NP}$.

Although optimal solutions for problems in NPO, but not in PO, are intractable, it is sometimes possible to guarantee that a good solution (i.e., one that is worse than the optimal by no more than a given factor) can be found in polynomial time. These problems can therefore be further classified into class APX (constant factor approximation) and class exp-APX (inapproximable) with increasing complexity (Figure 1). We can arrange energy minimization problems on this more detailed complexity scale, originally established in [7], to provide vision researchers a new viewpoint for complexity classification, with a focus on NP-hard optimization problems.

Here we make three core contributions, as explained in the next three paragraphs. First, we prove the inapproximability result of QPBO and general energy minimization. Second, we show that the same inapproximability result holds when restricting to planar graphs with three or more labels. In the proof, we propose a novel micro-graph structure-based reduction that can be used for algorithmic design as well. Finally, we present a unified framework and an overview of vision-related special cases where the energy minimization problem can be solved in polynomial time or approximated with a constant, logarithmic, or polynomial factor.

Binary and multi-label case (Section 5). It is known that QPBO (2-label case) and the general energy minimization problem (multi-label case) are NP-hard [15], because they generalize such classical NP-hard optimization problems on graphs as vertex packing (maximum independent set) and the minimum and maximum cut problems [37]. In this thesis, we show a stronger conclusion. *We prove that QPBO as well as general energy minimization are complete (being the hardest problems) in the class exp-APX.* Assuming $\text{P} \neq \text{NP}$, this implies that a polynomial time method cannot have a guarantee of finding an approximation within a constant factor of the optimal, and in fact, the only possible factor in polynomial time is exponential in the input size. In practice, this means that a solution may be essentially arbitrarily bad.

Planar three or more label case (Section 6). Planar graphs form the underlying graph structure for many computer vision and image processing tasks. It is known that efficient exact algorithms exist for some special cases of planar 2-label energy minimization problems [74]. In this thesis, we show that for the case of three or more labels, planar energy minimization is exp-APX-complete, which means these problems are as hard as general energy minimization. It is unknown that whether a constant ratio approximation exists for planar 2-label problems in general.

Subclass problems (Section 7). Special cases for some energy minimization algorithms relevant to computer vision are known to be tractable. However, detailed complexity analysis of these algorithms is patchy and spread across numerous papers. In Section 7, we classify the complexity of these subclass problems and il-

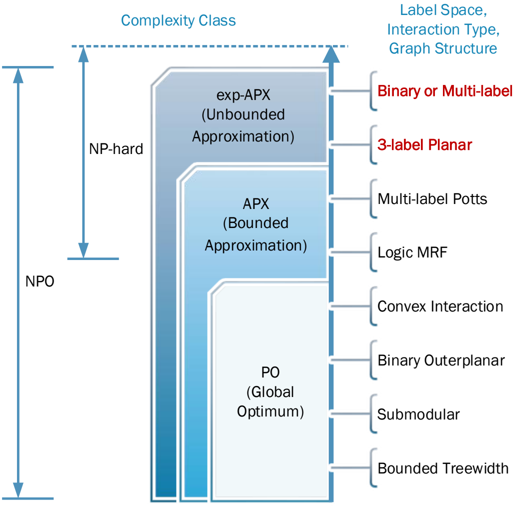


Fig. 1: Discrete energy minimization problems aligned on a complexity axis. Red/boldface indicates new results proven in this thesis. This axis defines a partial ordering, since problems within a complexity class are not ranked. Some problems discussed in this thesis are omitted for simplicity.

lustrate some of their connections. Such an analysis can help computer vision researchers become acquainted with existing complexity results relevant to energy minimization and can aid in selecting an appropriate model for an application or in designing new algorithms.

3 Related Work on the Complexity of Energy Minimization Problems

Much of the work on complexity in computer vision has focused on experimental or empirical comparison of inference methods, including influential studies on choosing the best optimization techniques for specific classes of energy minimization problems [85, 36] and the PASCAL Probabilistic Inference Challenge, which focused on the more general context of inference in graphical models [1]. In contrast, our work focuses on theoretical computational complexity, rather than experimental analysis.

On the theoretical side, the NP-hardness of certain energy minimization problems is well studied. It has been shown that 2-label energy minimization is, in general, NP-hard, but it can be in PO if it is submodular [40] or outerplanar [74]. For multi-label problems, the NP-hardness was proven by reduction from the NP-hard multi-way cut problem [19]. These results, however, say nothing about the complexity of *approximating* the global optimum for the intractable cases. The complexity involving approximation has been studied for classical combinatorial problems, such as MAX-CUT and MAX-2SAT, which are known to be APX-complete [63]. QPBO generalizes such problems and is therefore APX-hard. This leaves a possibility that QPBO may be in APX, i.e., approximable within a constant factor.

Energy minimization is often used to solve MAP inference for undirected graphical models. In contrast to scarce results for energy minimization and undirected graphical models, researchers have more extensively studied the computational complexity of approximating the MAP solution for *Bayesian networks*, also known as *directed graphical models* [54]. Abdelbar and Hedetniemi first proved the NP-hardness for approximating the MAP assignment of directed graphical models in the value of probability, i.e., finding x such that

$$\frac{p(x^*)}{p(x)} \leq r(n) \tag{2}$$

with a constant or polynomial ratio $r(n) \geq 1$ is NP-hard and showing that this problem is poly-APX-hard [2]. The probability approximation ratio is closest to the energy ratio used in our work, but other approximation measures have also been studied. Kwisthout showed the NP-hardness for approximating MAPs with the measure of additive value-, structure-, and rank-approximation [52, 53, 54]. He also investigated the hardness of expectation-approximation of MAP and found that no randomized algorithm can expectation-approximate MAP in polynomial time with a bounded margin of error unless NP $\subseteq$ BPP, an assumption that is highly unlikely to be true [54].

Unfortunately, the complexity results for directed models do not readily transfer to undirected models and vice versa. In directed and undirected models, the graphs represent different conditional independence relations, thus the underlying family of probability distributions encoded by these two models is distinct, as detailed in Appendix B. However, one can ask similar questions on the hardness of undirected models in terms of various approximation measures. In this work, we answer two questions, "How hard is it to approximate the MAP inference in the ratio of energy (log probability) and the ratio of probability?" The complexity of structure-, rank-, and expectation-approximation remain open questions for energy minimization.

4 Definitions and Notation

There are at least two different sets of definitions of what is considered an NP optimization problem [62, 7]. Here, we follow the notation of Ausiello et al [7] and restate the definitions needed for us to state and prove our theorems in Sections 5 and 6 with our explanation of their relevance to our proofs.

Definition 4.1 (Optimization Problem, [7] Def. 1.16)**.** An *optimization problem* $\mathcal{P}$ is characterized by a quadruple $(\mathcal{I}, \mathcal{S}, m, \text{goal})$ where

1. $\mathcal{I}$ is the set of instances of $\mathcal{P}$.

2. $\mathcal{S}$ is a function that associates to any input instance $x \in \mathcal{I}$ the set of *feasible solutions* of x.
3. m is the *measure* function, defined for pairs (x, y) such that $x \in \mathcal{I}$ and $y \in \mathcal{S}(x)$. For every such pair (x, y), $m(x, y)$ provides a positive integer.
4. $\text{goal} \in \{\min, \max\}$.

Notice the assumption that the cost is positive, and, in particular, it cannot be zero.

Definition 4.2 (Class NPO, [7] Def 1.17)**.** An optimization problem $\mathcal{P} = (\mathcal{I}, \mathcal{S}, m, \text{goal})$ belongs to the class of NP optimization (NPO) problems if the following hold:

1. The set of instances $\mathcal{I}$ is recognizable in polynomial time.
2. There exists a polynomial q such that given an instance $x \in \mathcal{I}$, for any $y \in \mathcal{S}(x)$, $|y| < q(x)$ and, besides, for any y such that $|y| < q(x)$, it is decidable in polynomial time whether $y \in \mathcal{S}(x)$.
3. The measure function m is computable in polynomial time.

Definition 4.3 (Class PO, [7] Def 1.18)**.** An optimization problem $\mathcal{P}$ belongs to the class of PO if it is in NPO and there exists a polynomial-time algorithm that, for any instance $x \in \mathcal{I}$, returns an optimal solution $y \in \mathcal{S}^*(x)$, together with its value $m^*(x)$.

For intractable problems, it may be acceptable to seek an approximate solution that is sufficiently close to optimal.

Definition 4.4 (Approximation Algorithm, [7] Def. 3.1)**.** Given an optimization problem $\mathcal{P} = (\mathcal{I}, \mathcal{S}, m, \text{goal})$ an algorithm $\mathcal{A}$ is an *approximation algorithm* for $\mathcal{P}$ if, for any given instance $x \in \mathcal{I}$, it returns an *approximate solution*, that is a feasible solution $\mathcal{A}(x) \in \mathcal{S}(x)$.

Definition 4.5 (Performance Ratio, [7], Def. 3.6)**.** Given an optimization problem $\mathcal{P}$, for any instance x of $\mathcal{P}$ and for any feasible solution $y \in \mathcal{S}(x)$, the *performance ratio*, *approximation ratio* or *approximation factor* of y with respect to x is defined as

$$R(x, y) = \max\left\{\frac{m(x, y)}{m^*(x)}, \frac{m^*(x)}{m(x, y)}\right\}, \tag{3}$$

where $m^*(x)$ is the measure of the optimal solution for the instance x.

Since $m^*(x)$ is a positive integer, the performance ratio is well-defined. It is a rational number in $[1, \infty)$. Notice that from this definition, it follows that if finding a feasible solution, e.g. $y \in \mathcal{S}(x)$, is an NP-hard decision problem, then there exists no polynomial-time approximation algorithm for $\mathcal{P}$, irrespective of the kind of performance evaluation that one could possibly mean.

Definition 4.6 ($r(n)$-approximation, [7], Def. 8.1)**.** Given an optimization problem $\mathcal{P}$ in NPO, an approximation algorithm $\mathcal{A}$ for $\mathcal{P}$, and a function $r\colon \mathbb{N} \to (1, \infty)$, we say that $\mathcal{A}$ is an $r(n)$*-approximate* algorithm for $\mathcal{P}$ if, for any instance x of $\mathcal{P}$ such that $\mathcal{S}(x) \neq \varnothing$, the performance ratio of the feasible solution $\mathcal{A}(x)$ with respect to x verifies the following inequality:

$$R(x, \mathcal{A}(x)) \leq r(|x|). \tag{4}$$

Definition 4.7 (F-APX, [7], Def. 8.2)**.** Given a class of functions F, F-APX is the class of all NPO problems $\mathcal{P}$ such that, for some function $r \in F$, there exists a polynomial-time $r(n)$-approximate algorithm for $\mathcal{P}$.

The class of constant functions for F yields the complexity class APX. Together with logarithmic, polynomial, and exponential functions applied in Definition 4.7, the following *complexity axis* is established:

$$\text{PO} \subseteq \text{APX} \subseteq \text{log-APX} \subseteq \text{poly-APX} \subseteq \text{exp-APX} \subseteq \text{NPO}.$$

Since the measure m needs to be computable in polynomial time for NPO problems, the largest measure and thus the largest performance ratio is an exponential function. But exp-APX is not equal to NPO (assuming P $\neq$ NP) because NPO contains problems whose feasible solutions cannot be found in polynomial time. For an energy minimization problem, any label assignment is a feasible solution, implying that all energy minimization problems are in exp-APX.

The standard approach for proofs in complexity theory is to perform a reduction from a known NP-complete problem. Unfortunately, the most common polynomial-time reductions ignore the quality of the solution in the approximated case. For example, it is shown that any energy minimization problem can be reduced to a factor 2 approximable Potts model [65], however the reduction is not approximation preserving and is unable to show the hardness of general energy minimization in terms of approximation. Therefore, it is necessary to use an approximation preserving (AP) reduction to classify NPO problems that are not in PO, for which only the approximation algorithms are tractable. AP-preserving reductions preserve the approximation ratio in a linear fashion, and thus preserve the membership in these complexity classes. Formally,

Definition 4.8 (AP-reduction, [7] Def. 8.3)**.** Let $\mathcal{P}_1$ and $\mathcal{P}_2$ be two problems in NPO. $\mathcal{P}_1$ is said to be AP-*reducible* to $\mathcal{P}_2$, in symbols $\mathcal{P}_1 \leq_{\text{AP}} \mathcal{P}_2$, if two functions π and σ and a positive constant α exist such that [2]:

1. For any instance $x \in \mathcal{I}_1$, $\pi(x) \in \mathcal{I}_2$.
2. For any instance $x \in \mathcal{I}_1$, if $S_1(x) \neq \varnothing$ then $S_2(\pi(x)) \neq \varnothing$.
3. For any instance $x \in \mathcal{I}_1$ and for any $y \in S_2(\pi(x))$, $\sigma(x, y) \in S_1(x)$.
4. π and σ are computable by algorithms whose running time is polynomial.
5. For any instance $x \in \mathcal{I}_1$, for any rational $r > 1$, and for any $y \in S_2(\pi(x))$,

$$R_2(\pi(x), y) \leq r \quad \text{implies} \tag{5}$$

$$R_1(x, \sigma(x, y)) \leq 1 + \alpha(r - 1). \tag{6}$$

[2] The complete definition contains a rational r for the the two mappings (π and σ) and it is omitted here for simplicity.

AP-reduction is the formal definition of the term ‘as hard as’ used in this thesis unless otherwise specified. It defines a partial order among optimization problems. With respect to this relationship, we can formally define the subclass containing the hardest problems in a complexity class:

Definition 4.9 ($\mathcal{C}$-hard and $\mathcal{C}$-complete, [7] Def. 8.5)**.** Given a class $\mathcal{C}$ of NPO problems, a problem $\mathcal{P}$ is $\mathcal{C}$-hard if, for any $\mathcal{P}' \in \mathcal{C}$, $\mathcal{P}' \leq_{\text{AP}} \mathcal{P}$. A $\mathcal{C}$-hard problem is $\mathcal{C}$-complete if it belongs to $\mathcal{C}$.

Intuitively, a complexity class $\mathcal{C}$ specifies the upper bound on the hardness of the problems within, $\mathcal{C}$-hard specifies the lower bound, and $\mathcal{C}$-complete exactly specifies the hardness.

5 Inapproximability for the General Case

In this section, we show that QPBO and general energy minimization are inapproximable by proving they are exp-APX-complete. As previously mentioned, it is already known that these problems are NP-hard [15], but it was previously unknown whether useful approximation guarantees were possible in the general case. The formal statement of QPBO as an optimization problem is as follows:

Problem 1. **QPBO**

INSTANCE: A pseudo-Boolean function $f\colon \mathbb{B}^{\mathcal{V}} \to \mathbb{N}$:

$$f(x) = \sum_{v \in \mathcal{V}} f_u(x_u) + \sum_{u,v \in \mathcal{V}} f_{uv}(x_u, x_v), \tag{7}$$

given by the collection of unary terms f_u and pairwise terms f_{uv}.

SOLUTION: Assignment of variables $x \in \mathbb{B}^{\mathcal{V}}$.

MEASURE: $\min f(x) > 0$.

Theorem 5.1. QPBO is exp-APX-complete.

Proof Sketch. (Full proof in Appendix A).

1. We observe that W3SAT-triv is known to be exp-APX-complete [7]. W3SAT-triv is a 3-SAT problem with weights on the variables and an artificial, trivial solution.
2. Each 3-clause in the conjunctive normal form can be represented as a polynomial consisting of three binary variables. Together with representing the weights with the unary terms, we arrive at a cubic Boolean minimization problem.
3. We use the method of [32] to transform the cubic Boolean problem into a quadratic one, with polynomially many additional variables, which is an instance of QPBO.
4. Together with an inverse mapping σ that we define, the above transformation defines an AP-reduction from W3SAT-triv to QPBO, i.e. W3SAT-triv $\leq_{\text{AP}}$ QPBO. This proves that QPBO is exp-APX-hard.

5. We observe that all energy minimization problems are in exp-APX and thereby conclude that QPBO is exp-APX-complete.

This inapproximability result can be generalized to more than two labels.

Corollary 5.2. k-label energy minimization is exp-APX-complete for $k \geq 2$.

Proof Sketch. (Full proof in Appendix A). This theorem is proved by showing QPBO $\leq_{\text{AP}}$ k-label energy minimization for $k \geq 2$.

We show in Corollary B.1 the inapproximability in energy (log probability) transfer to probability in Equation (2) as well.

Taken together, this theorem and its corollaries form a very strong inapproximability result for general energy minimization [3]. They imply not only NP-hardness, but also that there is no algorithm that can approximate general energy minimization with two or more labels with an approximation ratio better than some exponential function in the input size. In other words, any approximation algorithm of the general energy minimization problem can perform arbitrarily badly, and it would be pointless to try to prove a bound on the approximation ratio for existing approximation algorithms for the general case. While this conclusion is disappointing, these results serve as a clarification of grounds and guidance for model selection and algorithm design. Instead of counting on an oracle that solves the energy minimization problem, researchers should put efforts into selecting the proper formulation, trading off expressiveness for tractability.

6 Inapproximability for the Planar Case

Efficient algorithms for energy minimization have been found for special cases of 2-label planar graphs. Examples include planar 2-label problems without unary terms and outerplanar 2-label problems (i.e., the graph structure remains planar after connecting to a common node) [74]. Grid structures over image pixels naturally give rise to planar graphs in computer vision. Given their frequency of use in this domain, it is natural to consider the complexity of more general cases involving planar graphs.

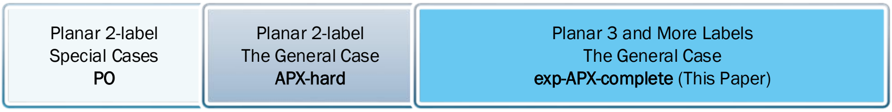


Fig. 2: Complexity for planar energy minimization problems. The "general case" implies no restrictions on the pairwise interaction type. This thesis shows that the third category of problems is not efficiently approximable.

[3] These results automatically generalize to higher order cases as they subsume the pairwise cases discussed here.

Figure 2 visualizes the current state of knowledge of the complexity of energy minimization problems on planar graphs. In this section, we prove that for the case of planar graphs with three or more labels, energy minimization is exp-APX-complete. This result is important because it significantly reduces the space of potentially efficient algorithms on planar graphs. The existence of constant ratio approximation for planar 2-label problems in general remains an open question [4].

Theorem 6.1. Planar 3-label energy minimization is exp-APX-complete.

Proof Sketch. (Full proof in Appendix A).

1. We construct elementary gadgets to reduce any 3-label energy minimization problem to a planar one with polynomially many auxiliary nodes.
2. Together with an inverse mapping σ that we define, the above construction defines an AP-reduction, i.e., 3-label energy minimization $\leq_{\text{AP}}$ planar 3-label energy minimization.
3. Since 3-label energy minimization is exp-APX-complete (Corollary 5.2) and all energy minimization problems are in exp-APX, we thereby conclude that planar 3-label energy minimization is exp-APX-complete.

Corollary 6.2. Planar k-label energy minimization is exp-APX-complete, for $k \geq 3$.

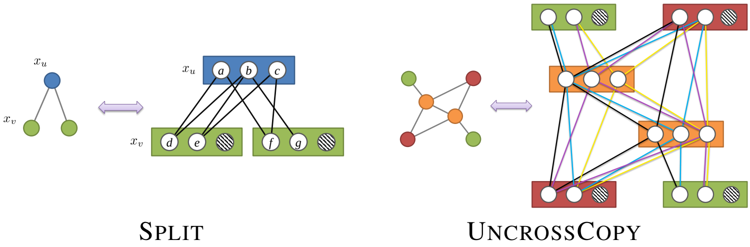
Proof Sketch. (Full proof in Appendix A). This theorem is proved by showing planar 3-label energy minimization $\leq_{\text{AP}}$ planar k-label energy minimization, for $k \geq 3$.

These theorems show that the restricted case of planar graphs with 3 or more labels is as hard as general case for energy minimization problems with the same inapproximable implications discussed in Section 5.

The most novel and useful aspect of the proof of Theorem 6.1 is the planar reduction in Step 1. The reduction creates an equivalent planar representation to any non-planar 3-label graph. That is, the graphs share the same optimal value. The reduction applies elementary constructions or “gadgets” to uncross two intersecting edges. This process is repeated until all intersecting edges are uncrossed. Similar elementary constructions were used to study the complexity of the linear programming formulation of energy minimization problems [66, 65]. Our novel gadgets have three key properties *at the same time*: 1) they are able to uncross intersecting edges, 2) they work on non-relaxed problems, i.e., all indicator variables (or pseudomarginals to be formal) are integral; and 3) they can be applied repeatedly to build an AP-reduction.

The two gadgets used in our reduction are illustrated in Figure 3. A 3-label node can be encoded as a collection of 3 indicator variables with a one-hot constraint. In the figure, a solid colored circle denotes a 3-label node, and a solid colored rectangle denotes the equivalent node expressed with indicator variables (white circles). For example, in Figure 3, $a = 1$ corresponds to the blue node taking the first label value. The pairwise potentials (edges on the left part of the figures) can be viewed as edge costs between the indicator variables (black lines on the right), e.g., $f_{uv}(3, 2)$

[4] The planar 2-label problem in general is APX-hard, since it subsumes the APX problem planar vertex cover [10].



SPLIT

UNCROSSCOPY

Fig. 3: Gadgets to represent a 3-label variable as two 2-label variables (SPLIT) and to copy the values of two diagonal pairs of 2-label variables without edge crossing (UNCROSSCOPY).

is placed onto the edge between indicator c and e and is counted into the overall measure if and only if $c = e = 1$. In our gadgets, drawn edges represent zero cost while omitted edges represent positive infinity[5]. While the set of feasible solutions remains the same, the gadget encourages certain labeling relationships, which, if not satisfied, cause the overall measure to be infinity. Therefore, the encouraged relationships must be satisfied by any optimal solution. The two gadgets serve different purposes:

SPLIT A 3-label node (blue) is split into two 2-label nodes (green). The shaded circle represents a label with a positive infinite unary cost and thus creates a simulated 2-label node. The encouraged relationships are

- $a = 1 \Leftrightarrow d = 1$ and $f = 1$.
- $b = 1 \Leftrightarrow g = 1$.
- $c = 1 \Leftrightarrow e = 1$ and $f = 1$.

Thus (d, f) encodes a, (d, g) and (e, g) both encode b and (e, f) encodes c.

UNCROSSCOPY The values of two 2-label nodes are encouraged to be the same as their diagonal counterparts respectively (red to red, green to green) without crossing with each other. The orange nodes are intermediate nodes that pass on the values. All types of lines represent the same edge cost, which is 0. The color differences visualize the verification for each of the 4 possible states of two 2-label nodes. For example, the cyan lines verify the case where the top-left (green) node takes the values (1, 0) and the top-right (red) node takes the value (0, 1). It is clear that the encouraged solution is for the bottom-left (red) node and the bottom-right (green) node to take the value (0, 1) and (1, 0) respectively.

These two gadgets can be used to uncross the intersecting edges of two pairs of 3-label nodes (Figure 4, left). For a crossing edge (x_u, x_v), first a new 3-label node $x_{v'}$ is introduced preserving the same arbitrary interaction (red line) as before (Figure 4, middle). Then, the crossing edges (enclosed in the dotted circle) are uncrossed

[5] A very large number will also serve the same purpose, e.g., take the sum of the absolute value of all energy terms and add 1. Therefore, we are not expanding the set of allowed energy terms to include ∞.

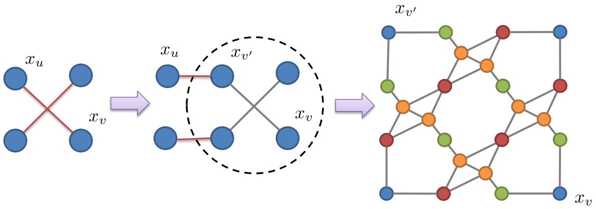


Fig. 4: Planar reduction for 3-label problems

by applying SPLIT and UNCROSSCOPY four times (Figure 4, right). Without loss of generality, we can assume that no more than two edges intersect at a common point except at their endpoints. This process can be applied repeatedly at each edge crossing until there are no edge crossings left in the graph [66].

7 Complexity of Subclass Problems

In this section, we classify some of the special cases of energy minimization according to our complexity axis (Figure 1). This classification can be viewed as a reinterpretation of existing results from the literature into a unified framework.

7.1 Class PO (Global Optimum)

Polynomial time solvability may be achieved by considering two principal restrictions: those restricting the *structure* of the problem, i.e., the graph G, and those restricting the type of allowed interactions, i.e., functions f_{uv}.

Structure Restrictions. When G is a chain, energy minimization reduces to finding a shortest path in the trellis graph, which can be solved using a classical dynamic programming (DP) method known as the Viterbi algorithm [26]. The same DP principle applies to graphs of bounded treewidth. Fixing all variables in a separator set decouples the problem into independent optimization problems. For treewidth 1, the separators are just individual vertices, and the problem is solved by a variant of DP [64, 73]. For larger treewidths, the respective optimization procedure is known as junction tree decomposition [56]. A loop is a simple example of a treewidth 2 problem. However, for a treewidth k problem, the time complexity is exponential in k [56]. When G is an outer-planar graph, the problem can be solved by the method of [74], which reduces it to a planar Ising model, for which efficient algorithms exist [81].

Interaction Restrictions. Submodularity is a restriction closely related to problems solvable by minimum cut. A quadratic pseudo-Boolean function f is *submodular* iff its quadratic terms are non-positive. It is then known to be equivalent with finding a minimum cut in a corresponding network [28]. Another way to state this condition for QPBO is $\forall(u,v) \in \mathcal{E}, f_{uv}(0,1) + f_{uv}(1,0) \geq f_{uv}(0,0) + f_{uv}(1,1)$. However, submodularity is more general. It extends to higher-order and multi-label problems. Submodularity is considered a discrete analog of convexity. Just as convex functions are relatively easy to optimize, general submodular function minimization can be solved in strongly polynomial time [75]. Kolmogorov and Zabin introduced submodularity in computer vision and showed that binary 2$^{\text{nd}}$ order and 3$^{\text{rd}}$ order submodular problems can be always reduced to minimum cut, which is much more efficient than general submodular function minimization [45]. Živný et al. and Ramalingam et al. give more results on functions reducible to minimum cut [96, 67]. For QPBO on an unrestricted graph structure, the following *dichotomy* result has been proven by Cohen et al. [22]: either the problem is submodular and thus in PO or it is NP-hard (i.e., submodular problems are the only ones that are tractable in this case).

For multi-label problems Ishikawa proposed a reduction to minimum cut for problems with convex interactions, i.e., where $f_{uv}(x_u, x_v) = g_{uv}(x_u - x_v)$ and g_{uv} is convex and symmetric [31]. It is worth noting that when the unary terms are convex as well, the problem can be solved even more efficiently [30, 41]. The same reduction [31] remains correct for a more general class of submodular multi-label problems. In modern terminology, component-wise minimum $x \wedge y$ and component-wise maximum $x \vee y$ of complete labelings x, y for all nodes are introduced ($x, y \in \mathcal{L}^{\mathcal{V}}$). These operations depend on the *order of labels* and, in turn, define a lattice on the set of labelings. The function f is called *submodular on the lattice* if $f(x \vee y) + f(x \wedge y) \leq f(x) + f(y)$ for all x, y [92]. In the pairwise case, the condition can be simplified to the form of submodularity common in computer vision [67]: $f_{uv}(i, j+1) + f_{uv}(i+1, j) \geq f_{uv}(i,j) + f_{uv}(i+1, j+1)$. In particular, it is easy to see that a convex f_{uv} satisfies it [31]. Kolmogorov [43] and Arora et al. [6] proposed maxflow-like algorithms for higher order submodular energy minimization. Schlesinger proposed an algorithm to find a reordering in which the problem is submodular if one exists [72]. However, unlike in the binary case, solvable multi-label problems are more diverse. A variety of problems are generalizations of submodularity and are in PO, including symmetric tournament pair, submodularity on arbitrary trees, submodularity on arbitrary lattices, skew bisubmodularity, and bisubmodularity on arbitrary domains (see references in [90]). Thapper and Živný [89] and Kolmogorov [44] characterized these tractable classes and proved a similar dichotomy result: a problem of unrestricted structure is either solvable by LP-relaxation (and thus in PO) or it is NP-hard. It appears that LP relaxation is the most powerful and general solving technique [104].

Mixed Restrictions. In comparison, results with mixed structure and interaction restrictions are rare. One example is a planar Ising model without unary terms [81]. Since there is a restriction on structure (planarity) and unary terms, it does not fall into any of the classes described above. Another example is the restriction to supermodular functions on a bipartite graph, solvable by [72] or by LP relaxation, but not falling under the characterization [90] because of the graph restriction.

Algorithmic Applications. The aforementioned tractable formulations in PO can be used to solve or approximate harder problems. Trees, cycles and planar problems are used in dual decomposition methods [46, 47, 12]. Binary submodular problems are used for finding an optimized crossover of two-candidate multi-label solutions. An example of this technique, the expansion move algorithm, achieves a constant approximation ratio for the Potts model [19]. Extended dynamic programming can be used to solve restricted segmentation problems [24] and as move-making subroutine [95]. LP relaxation also provides approximation guarantees for many problems [8, 21, 38, 48], placing them in the APX or poly-APX class.

7.2 Class APX and Class log-APX (Bounded Approximation)

Problems that have bounded approximation in polynomial time usually have certain restriction on the interaction type. The Potts model may be the simplest and most common way to enforce the smoothness of the labeling. Each pairwise interaction depends on whether the neighboring labellings are the same, i.e. $f_{uv}(x_u, x_v) = c_{uv}\delta(x_u, x_v)$. Boykov et al. showed a reduction to this problem from the NP-hard multiway cut [19], also known to be APX-complete [7, 23]. They also proved that their constructed alpha-expansion algorithm is a 2-approximate algorithm. These results prove that the Potts model is in APX but not in PO. However, their reduction from multiway cut is not an AP-reduction, as it violates the third condition of AP-reducibility. Therefore, it is still an open problem whether the Potts model is APX-complete. Boykov et al. also showed that their algorithm can approximate the more general problem of metric labeling [19]. The energy is called *metric* if, for an arbitrary, finite label space $\mathcal{L}$, the pairwise interaction satisfies a) $f_{uv}(\alpha, \beta) = 0$, b) $f_{uv}(\alpha, \beta) = f_{uv}(\beta, \alpha) \geq 0$, and c) $f_{uv}(\alpha, \beta) \leq f_{uv}(\beta, \gamma) + f_{uv}(\beta, \gamma)$, for any labels α, β, $\gamma \in \mathcal{L}$ and any $uv \in \mathcal{E}$. Although their approximation algorithm has a bound on the performance ratio, the bound depends on the ratio of some pairwise terms, a number that can grow exponentially large. For metric labeling with k labels, Kleinberg et al. proposed an $O(\log k \log\log k)$-approximation algorithm. This bound was further improved to $O(\log k)$ by Chekuri et al. [20], making metric labeling a problem in log-APX [6].

We have seen that a problem with convex pairwise interactions is in PO. An interesting variant is its truncated counterpart, i.e., $f_{uv}(x_u, x_v) = w_{uv} \min\{d(x_u - x_v), M\}$, where w_{uv} is a non-negative weight, d is a convex symmetric function to define the distance between two labels, and M is the truncating constant [94]. This problem is NP-hard [94], but Kumar et al. [51] have proposed an algorithm that yields bounded approximations with a factor of $2 + \sqrt{2}$ for linear distance functions

[6] An $O(\log k)$-approximation implies an $O(\log |x|)$-approximation (see Corollary C.1).

and a factor of $O(\sqrt{M})$ for quadratic distance functions[7]. This bound is analyzed for more general distance functions by Kumar [50].

Another APX problem with implicit restrictions on the interaction type is logic MRF [9]. It is a powerful higher order model able to encode arbitrary logical relations of Boolean variables. It has energy function $f(x) = \sum_i^n w_i C_i$, where each C_i is a disjunctive clause involving a subset of Boolean variables x, and $C_i = 1$ if it is satisfied and 0 otherwise. Each clause C_i is assigned a non-negative weight w_i. The goal is to find an assignment of x to maximize $f(x)$. As disjunctive clauses can be converted into polynomials, this is essentially a pseudo-Boolean optimization problem. However, this is a special case of general 2-label energy minimization, as its polynomial basis spans a subspace of the basis of the latter. Bach et al. [9] proved that logic MRF is in APX by showing that it is a special case of MAX-SAT with non-negative weights.

8 Practical Implications

The algorithmic implications of our inapproximability have been discussed above. Here, we focus on the discussion of practical implications. The existence of an approximation guarantee indicates a practically relevant class of problems where one may expect reasonable performance. In structural learning for example, it is acceptable to have a constant factor approximation for the inference subroutine when efficient exact algorithms are not available. Finley and Joachims proved that this constant factor approximation guarantee yields a multiplicative bound on the learning objective, providing a relative guarantee for the quality of the learned parameters [25]. An optimality guarantee is important, because the inference subroutine is repeatedly called, and even a single poor approximation, which returns a not-so-bad worst violator, will lead to the early termination of the structural learning algorithm.

However, despite having no approximation ratio guarantee, algorithms such as the extended roof duality algorithm for QPBO [71] are still widely used. This gap between theory and application applies not only to our results but to all other complexity results as well. We list several key reasons for the potential lack of correspondence between theoretical complexity guarantees and practical performance.

Complexity results address the worst case scenario. Our inapproximability result guarantees that for any polynomial time algorithm, there exists an input instance for which the algorithm will produce a very poor approximation. However, applications often do not encounter the worst case. Such is the case with the simplex algorithm, whose worst case complexity is exponential, yet it is widely used in practice.

Objective function is not the final evaluation criterion. In many image processing tasks, the final evaluation criterion is the number of pixels correctly labeled.

[7] In these truncated convex problems, the ratio bound is defined for the pairwise part of the energy (1). The approximation ratio in accordance to our definition is obtained assuming the unary terms are non-negative.

The relation between the energy value and the accuracy is implicit. In many cases, a local optimum is good enough to produce a high labeling accuracy and a visually appealing labeling.

Other forms of optimality guarantee or indicator exist. Approximation measures in the distance of solutions or in the expectation of the objective value are likely to be prohibitive for energy minimization, as they are for Bayesian networks [52, 53, 54]. On the other hand, a family of energy minimization algorithms has the property of being *persistent* or *partial optimal*, meaning a subset of nodes have consistent labeling with the global optimal one [13, 14]. Rather than being an optimality guarantee, persistency is an optimality indicator. In the worst case, the set of persistent labelings could be empty, yet the percentage of persistent labelings over the all the nodes gives us a notion of the algorithm's performance on this particular input instance. Persistency is also useful in reducing the size of the search space [39, 79]. Similarly, the per-instance integrality gap of duality based methods is another form of optimality indicator and can be exponentially large for problems in general [48, 84].

Part II

9 An Alternative to the Inapproximable Inference Subroutine

Now we have seen that energy minimization is in general inapproximable. Therefore, without limiting the label set, the potential type or the graph structure, we cannot expect to solve the imperfect gradient problem in structural learning with a better inference algorithm. However, an alternative is to modify the structural learning framework. In the next few sections, we show that considering together the joint problem of the overarching training and the inference subroutine enables us to exploit properties that would not be possible otherwise. First, we propose a theoretically sound structural learning algorithm without the limitation of intractable inference. We review and exploit the properties of the joint problem of training time inference and learning. By modifying the training procedure, we can perform a training time inference corresponding to a binary submodular problem that is much easier than the original one while keeping the testing time inference problem almost the same. This method can be extended to learn higher order potentials as well. Second, while making no assumptions on the structure of the graph or on the potential type, we prove that our algorithm returns a solution within a given absolute error relative to the global optimal within the feasible parameter space. In addition, we demonstrate our algorithm's performance on two 3D scene parsing datasets. On one dataset, our algorithm runs three times faster than the competing method [3] and achieves the same level of accuracy. Our algorithm finds a solution efficiently on the second, more complex problem, which is intractable for competing methods. Also, we show that what is learned by the model captures domain knowledge and is easily interpretable.

10 Related Work on Structural Learning

Most existing literature on structural learning is based on the max-margin formulation proposed by Taskar et al. [70]. Directly minimizing the negative log-likelihood is NP-hard for many problems, and approximation must be used. The max-margin formulation uses a convex surrogate loss, removing the need for computing the partition function. Joachims et al. [35, 93] generalized this max-margin formulation to arbitrary structural outputs, a method known as structural SVM. The concept of max-margin structural learning has been successfully applied to many problems in computer vision. These works usually have limiting assumptions: tree-like or special structure output [59, 76, 102], small structural space [29, 100], or restricted potential type [4, 60, 87, 86]. Under these assumptions, exact inference is possible. However, we don't make these assumptions, yet we can still apply exact inference during training. Other works adopt approximate inference for the separation oracle

[3]. These methods have no guarantee of the solution quality. Notably, a common approximation scheme is convex programming relaxation [33]. Our early experiments show that methods based on this type of relaxation produce results with undesirably low accuracy.

The most similar work to our approach is [25], in which they point out the problem of training structural SVMs when exact inference is not possible and proposed two workarounds. The first one is to assume a constant factor approximation of the inference procedure. However, it was shown in [57] that such an assumption is not reasonable, as the problem cannot be approximated with any meaningful guarantee. The second workaround is to use the persistency property of binary MRFs, yet there is no quality guarantee of the learned parameters. In addition, we find the approach often fails in practice. Many works [55, 68, 77] focus on improving the performance of structural SVM itself, but still they face the problem of an imperfect separation oracle.

Similar to previous works, our algorithm is based on the max-margin formulation [70]. We adopt non-negative constraints to restrict the parameter space [86, 87], but in combination with a different loss and a different separation oracle for tractability.

The separation oracle in structural learning is frequently solved by energy minimization. Here, we highlight energy minimization algorithms used in this work and refer readers to [36] for a complete overview. Boykov and Kolmologrov (BK) [18] solved MAP inference for binary MRFs with a specially optimized max-flow algorithm. Rother et al. [71] proposed the Quadratic Pseudo-Boolean Optimization (QPBO) algorithm for binary problems of arbitrary potentials. They first created a different auxiliary graph, in which each original node corresponds exactly to two non-terminal nodes in the new graph. Then they ran the BK algorithm on this auxiliary graph. Note that some nodes will remain unlabeled if the corresponding non-terminal node pair has conflicting assignments. For multi-class problems of arbitrary potentials, Kolmologrov [42] built a convergent version of the tree-reweighted max-product message passing algorithm (TRW-S). By creating a proper local polytope, an energy minimization problem can be reduced to an integer linear programming (ILP) problem [98], and the integral constraint can be removed to derive an approximation algorithm (LP).

11 Our Structural Learning Algorithm

We propose a max-margin structural learning algorithm for a pairwise model with a linear discriminant function. Our algorithm enables tractable exact training time inference through our submodular formulation, which leads to a guaranteed solution quality. Submodularity cannot be easily enforced because it requires a binary problem and limits the potential type. As adopted in standard machine learning algorithms, multi-class classification can be solved by training a set of 1-vs-all binary classifiers and post-processing the classifier output to make a final one-hot prediction where only a single class is labeled for each example. We adopt a similar idea. During training, we solve a set of binary classification problems but without resolving the conflicts among the binary classifiers. This setup can still learn the desired parameters, since the loss will encourage the parameters to make one-hot predictions. During testing, we enforce one-hot prediction by adding a hard constraint to the inference problem. Because we are enforcing the submoduarity on the transformed binary problems, the potential type of the original energy is not constrained. The rest of this section introduces the desired theoretical properties of the inference procedure and the learning framework before showing our modifications to exploit these properties to build to our structural learning algorithm.

11.1 Problems and Properties

In this subsection, we first review the energy minimization formulation and the submodular property. Then we introduce our testing and training formulation.

Problem 1. **Discrete Energy Minimization**

- Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, define the *energy function*

$$U(\mathbf{y}) = \sum_{u \in \mathcal{V}} U_u(y_u) + \sum_{(u,v) \in \mathcal{E}} U_{uv}(y_u, y_v) \tag{1}$$

  where $U_{uv}(y_u, y_v) = U_{vu}(y_v, y_u)$
- *Energy minimization* assigns to each node a label from a finite label set $\mathcal{L}$ to minimize the energy

$$\mathbf{y}^* = \operatorname*{argmin}_{\mathbf{y} \in \mathcal{L}^{|\mathcal{V}|}} U(\mathbf{y}) \tag{2}$$

Definition 11.1 ([71])**.** A binary (two-class) energy minimization problem is *submodular* if and only if $\forall u, v \in \mathcal{V}$

$$U_{uv}(0,1) + U_{uv}(1,0) \geq U_{uv}(0,0) + U_{uv}(1,1) \tag{3}$$

It is well-known that if the energy is submodular, the global minimum can be found in polynomial time using graph cut. For multi-class problems, submodularity [67] is hard to exploit due to the order dependency and magnitude constraint. The definition of submodularity requires the label set to be a totally ordered set, e.g., a depth value from 0 to 255. This definition also constrains the relative magnitude of potentials on the same edge as in the binary case. These two conditions are not generally applicable.

Another interesting property, which is exploited by [25], is *persistency*, or *partial optimality*. Comparing to submodularity, persistency is an optimality indicator rather than an optimality guarantee. If we run the QPBO algorithm [71] on binary problems with arbitrary potentials, some nodes will be left unlabeled, but labelled nodes are part of the globally optimal solution. Boros et al. [16] showed that in an equivalent linear programming formulation, all variables corresponding to the unlabeled nodes take 0.5 in optimal solution. Let's assume we accept relaxed ([0, 1] instead of $\{0, 1\}$) solutions, then running QPBO and replacing the unlabeled nodes with 0.5 will result in an approximation algorithm, which we denote as QPBO-R.

An immediate question is how good the QPBO-R approximation is. This question is answered from a more general perspective in [57]: assuming P $\neq$ NP, for binary energy minimization in general, there does not exist a constant ratio approximation algorithm or even one with a ratio subexponential in the input size. Unfortunately, the theoretical properties of many structural learning algorithms [25, 55, 77] depend on a separation oracle with at least a constant ratio approximation, and the finding in [57] makes pointless the assumption along with the derived properties for these algorithms when applied to energy minimization in general.

We use full potential structural prediction as our testing time formulation.

Problem 2. **Full Potential Structural Prediction**

- Given a node feature extractor $\delta(\cdot)$ and an edge feature extractor $\delta(\cdot,\cdot), \forall k, l \in \mathcal{L}$ define the *unary* and *pairwise potentials*

$$U_u(y_u = k) := -\mathbf{w}_u^k \cdot \delta(u) \tag{4}$$

$$U_p(y_u = k, y_v = l) := -\mathbf{w}_{uv}^{kl} \cdot \delta(u, v) \tag{5}$$

- Denote the graph $\mathcal{G}$ as $\mathbf{x}$, and define the *linear discriminant function (score function)*

$$f(\mathbf{x}, \mathbf{y}) := -U(\mathbf{y}) = \mathbf{w}^\intercal \Psi(\mathbf{x}, \mathbf{y}) \tag{6}$$

- $\Psi(\mathbf{x}, \mathbf{y})$ is called the *joint feature map*. Using *binary encoding* $y_u^k = \delta(y_u = k)$, $\Psi(\mathbf{x}, \mathbf{y})$ can be decomposed as follows:

$$\Psi(\mathbf{x}, \mathbf{y})_{\mathbf{w}_u^k} = \sum_{u \in \mathcal{V}} y_u^k \delta(u) \tag{7}$$

$$\Psi(\mathbf{x}, \mathbf{y})_{\mathbf{w}_{uv}^{kl}} = \sum_{(u,v) \in \mathcal{E}} y_u^k y_v^l \delta(u, v) \tag{8}$$

- Then the *testing time inference problem* is

$$\hat{\mathbf{y}} = \operatorname*{argmax}_{\mathbf{y} \in \mathcal{L}^{|\mathcal{V}|}} f(\mathbf{x}, \mathbf{y}) = \operatorname*{argmin}_{\mathbf{y} \in \mathcal{L}^{|\mathcal{V}|}} U(\mathbf{y}) \tag{9}$$

- By abuse of notation, let $(\mathbf{x}_i, \mathbf{y}_i)$ be an *example* from a *dataset* $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$.

The potentials depend on both the parameters and the features, so given $\mathbf{w}$, $f = \mathbf{w}^\intercal \Psi(\mathbf{x}_i, \mathbf{y}_i)$ defines an energy function for an example $\mathbf{x}_i$. An ideal set of parameters should put the ground truth at or close to the place of lowest energy/highest score for each example so that the output of testing time inference is at or close to the ground truth. A linear score function makes the parameter estimation easier than non-linear forms. For some structural learning algorithms, kernel tricks can be applied to capture complicated mappings [35].

Full Potential Interaction Notice here we have a full potential matrix $U_p(y_u^k, y_v^l)$ for each edge. This generalizes the well-known Potts model and associative Markov networks [87], where only the diagonal terms are non-zero. The relative magnitude of diagonal terms and off-diagonal terms can be arbitrary. *This implies that the model is more expressive as it can be both attractive (modeling a smoothing prior) or repulsive. Moreover, the potential matrix does not need to be symmetric.* Thus, such a formulation is able to encode directed relationships like relative positions, e.g., a computer monitor is usually placed above desk.

Next, we present the standard learning framework before presenting our modifications.

Problem 3. **Structural SVM** [35]

$$\min_{\mathbf{w},\xi\geq 0} \quad \frac{1}{2}||\mathbf{w}||^{\mathbf{2}} + \mathbf{C}\xi \tag{10}$$

$$\text{s.t.} \quad \forall(\bar{\mathbf{y}}_1, ..., \bar{\mathbf{y}}_n) \in \mathcal{Y}^n :$$

$$\frac{1}{n}\mathbf{w}^\mathsf{T}\sum_{i=1}^{n}\left(\Psi - \bar{\Psi}\right) \geq \frac{1}{n}\sum_{i=1}^{n}\Delta(\mathbf{y}_i, \bar{\mathbf{y}}_i) - \xi, \tag{11}$$

where Ψ and $\bar{\Psi}$ are shorthand for $\Psi(\mathbf{x}_i, \mathbf{y}_i)$ and $\Psi(\mathbf{x}_i, \bar{\mathbf{y}}_i)$.

Structural SVMs are an extension to standard SVMs for structural outputs. A structural SVM finds the optimal set of parameters that creates a large margin relative to the loss for each structural example in the dataset. Here C is the parameter that controls the relative weighting between regularization and risk minimization, and $\Delta(\mathbf{y}_i, \hat{\mathbf{y}})$ is a loss function encoding the penalty for a wrong labeling.

Due to the combinatorial nature of the label space ($\mathcal{Y}^n = \mathcal{L}^{|\mathcal{V}|}$) , its size, i.e., the number of constraints (11) is exponential. Joachims et al. [35, 93] proposed the cutting-plane algorithm, which finds the optimal solution by adding only a polynomial number of constraints, given a separation oracle to compute the subgradients.

Definition 11.2. Given a loss function $\Delta(\mathbf{y}_i, \hat{\mathbf{y}})$, the *loss augmented inference* or *separation oracle* is a procedure that finds

$$\bar{\mathbf{y}}_i = \operatorname*{argmax}_{\hat{\mathbf{y}}\in\mathcal{Y}} \Delta(\mathbf{y}_i, \hat{\mathbf{y}}) + \mathbf{w}^\mathsf{T}\Psi(\mathbf{x}_i, \hat{\mathbf{y}}) \tag{12}$$

The loss augmented inference finds the worst violators of the margin. Instead of bounding in the entire structural space $\hat{\mathbf{y}} \in \mathcal{Y}$, the cutting-plane algorithm bounds the violation of the worst violators. It can be shown that this is equivalent to solving the original problem, but now the algorithm terminates in polynomial time and returns a globally optimal solution.

11.2 The Joint Problem for Parameter Estimation

This subsection describes our modifications to solve the joint problem that is not limited by the intractable separation oracle as in previous approaches. For the loss fuction, we use *Hamming loss* with the goal of labeling each node in the graph correctly:

$$\Delta(\mathbf{y}, \bar{\mathbf{y}}) = \rho \left[1 - \frac{1}{|\mathcal{V}|} \sum_{u \in \mathcal{V}} \delta(y_u = \bar{y}_u) \right] \tag{13}$$

The loss equals to (1 - accuracy) scaled by a factor ρ. The structure of the loss is simple, and the loss can be merged into the unary potentials, making loss augmented problem the same problem as Problem 2.

Multi-class to Binary Transformation For loss augmented inference, we use a binary encoding and remove the sum-up-to-1 constraint ($\sum_{k \in \mathcal{L}} y_u^k = 1$). The loss also needs to be slightly modified to address the removal of the constraint. We adopt the *Hamming loss for binary encoding*:

$$\Delta_b(\mathbf{y}, \bar{\mathbf{y}}) = \frac{\rho}{|\mathcal{V}|} \sum_{u \in \mathcal{V}} \sum_{k \in \mathcal{L}} \delta(y_u^k \neq \bar{y}_u^k). \tag{14}$$

The above modifications are based on the following observations:

- With the sum-up-to-1 constraint, $\Delta(\mathbf{y}, \bar{\mathbf{y}})$ and $\Delta_b(\mathbf{y}, \bar{\mathbf{y}})$ are equivalent;
- Without the sum-up-to-1 constraint, let $\delta(y_u = \bar{y_u}) = \prod_{k \in \mathcal{L}} \delta(y_u^k = \bar{y}_u^k)$, then $\Delta_b(\mathbf{y}, \bar{\mathbf{y}})$ is a tight upper bound of $\Delta(\mathbf{y}, \bar{\mathbf{y}})$ in that $\Delta_b(\mathbf{y}, \bar{\mathbf{y}}) \geq \Delta(\mathbf{y}, \bar{\mathbf{y}})$ and $\Delta_b(\mathbf{y}, \bar{\mathbf{y}}) = 0$ if and only if $\Delta(\mathbf{y}, \bar{\mathbf{y}}) = 0$;

In our approach, the removal of the sum-up-to-1 constraint changes the separation oracle, and the binary labeling might not have a consistent interpretation of the original labeling during training. However, the tightness of the loss function shows that we are effectively learning parameters to minimize the original loss. The sum-up-to-1 constraint is implicitly enforced in a soft manner through the loss minimization during training. Soft labeling ($y_u^k \in [0, 1]$) is adopted in [3, 25]. In this case, the loss is defined by replacing $\delta(y_u^k \neq \bar{y}_u^k)$ with $|y_u^k - \bar{y}_u^k|$ in (14). In contrast to the hard labeling that we use, for soft labeling without the sum-up-to-1 constraint, $\Delta_b(\mathbf{y}, \bar{\mathbf{y}})$ does not have the same property of being a tight upper bound.

Enforcing Submodularity As presented in Section 11.1, without any relaxation, the transformed binary problem puts great challenge to the inference subroutine because the problem is NP-hard and not even possible to approximate with a guarantee. Thus, we need to enforce submodularity to enable tractable exact inference.

The transformed binary problem U^b takes the form

$$U_p^b(y_u^k, y_u^l) = y_u^k y_u^l U_p(y_u = k, y_v = l). \tag{15}$$

Note that it does not have a full potential matrix, and only $U_p^b(1,1)$ can be nonzero. If, for all edges, $U_p^b(1,1)$ is non-positive, the whole energy satisfies (3) and is submodular. Since our algorithm depends on only $U_p^b(1,1)$ being non-zero, the multi-class-to-binary transformation must also be applied to binary classification problems, which is not necessary in the typical 1-vs-all setup.

One way to satisfy the condition of $U_p^b(1,1) \leq 0$ is to have all edge features $\delta(\cdot,\cdot)$ and pairwise parameters $\mathbf{w}_{uv}^{kl}$ be non-negative. It is reasonable to assume pairwise features can be always non-negative, since in many applications, the features are normalized to [0, 1] during a pre-processing step. Therefore, we add additional constraints only on the weights (18). We summarize our formulation as follows:

Problem 4. **Partially Non-negative Structural SVM**

$$\min_{\mathbf{w},\xi\geq 0} \quad \frac{1}{2}||\mathbf{w}||^{\mathbf{2}} + \mathbf{C}\xi \tag{16}$$

$$\text{s.t.} \quad \forall(\bar{\mathbf{y}}_1, ..., \bar{\mathbf{y}}_n) \in \mathcal{Y}^n :$$

$$\frac{1}{n}\mathbf{w}^\intercal \sum_{i=1}^{n} \left(\Psi - \bar{\Psi}\right) \geq \frac{1}{n}\sum_{i=1}^{n} \Delta_b(\mathbf{y}_i, \bar{\mathbf{y}}_i) - \xi \tag{17}$$

$$\forall j \in P, \quad w_j \geq 0 \tag{18}$$

where Ψ and $\bar{\Psi}$ are short for $\Psi(\mathbf{x}_i, \mathbf{y}_i)$ and $\Psi(\mathbf{x}_i, \bar{\mathbf{y}}_i)$. P is the set of indices where the parameter should be non-negative, e.g., the pairwise weights.

To solve this problem, we adopt the standard max-margin formulation. Our complete algorithm is shown in Algorithm 1.

Solving the Modified Quadratic Program Non-negative constraints have been previously employed in structural learning but in a different context. In pose estimation [102, 103], the quadratic spring terms must be non-negative. These works employ a tree-structured model, so exact inference is possible through dynamic programming. It is shown in [68] that for solvers in the primal space, adding non-negative constraints amounts to clipping the parameters during the update step while leaving the rest unchanged. We adopt the dual coordinate descent solver from [68] to solve the minimization problem in Problem 4. In practice, however, we find that a commercial general purpose QP solver, namely Gurobi [27], is several times faster under the same tolerance setting.

Algorithm 1 Submodular Structural SVM for Non-submodular Problems

```
$\mathcal{W} \leftarrow \varnothing$                                  ▷ A working set of worst violators
$\eta \leftarrow \infty$                                  ▷ The new violation in each iteration
$\xi \leftarrow 0$                                  ▷ The violation of the entire working set
while $\eta - \xi > \varepsilon$ do
    $(\mathbf{w}, \xi) \leftarrow \operatorname{argmin}_{\mathbf{w},\xi \geq 0} \quad \frac{1}{2}||\mathbf{w}||^{\mathbf{2}} + \mathbf{C}\xi$
              s.t. $\forall(\bar{\mathbf{y}}_1, ..., \bar{\mathbf{y}}_n) \in \mathcal{W}$,
                   $\frac{1}{n}\mathbf{w}^\intercal \sum_{i=1}^{n} [\Psi(\mathbf{x}_i, \mathbf{y}_i) - \Psi(\mathbf{x}_i, \bar{\mathbf{y}}_i)] \geq \frac{1}{n} \sum_{i=1}^{n} \Delta_b(\mathbf{y}_i, \bar{\mathbf{y}}_i) - \xi$
                   $\forall j \in P, \quad w_j \geq 0$
    for i = 1,...,n do
        $\bar{\mathbf{y}}_i \leftarrow \operatorname{argmax}_{\hat{\mathbf{y}} \in \mathcal{Y}} \Delta_b(\mathbf{y}_i, \hat{\mathbf{y}}) + \mathbf{w}^\intercal \Psi(\mathbf{x}_i, \hat{\mathbf{y}})$     ▷ Exact inference is now possible
    end for
    $\mathcal{W} \leftarrow \mathcal{W} \cup \{(\bar{\mathbf{y}}_1, ..., \bar{\mathbf{y}}_n)\}$
    $\eta \leftarrow \frac{1}{n} \sum_{i=1}^{n} \Delta_b(\mathbf{y}_i, \bar{\mathbf{y}}_i) - \frac{1}{n}\mathbf{w}^\intercal \sum_{i=1}^{n} [\Psi(\mathbf{x}_i, \mathbf{y}_i) - \Psi(\mathbf{x}_i, \bar{\mathbf{y}}_i)]$
end while
return w
```

11.3 Generalization to Higher Order Potentials

Higher order potentials capture more interactions than the pairwise potentials. For example, a column between a pair of abutments is a 3rd order interaction. Our generalization is based on the pairwise reduction from arbitrary high order potentials proposed by Ishikawa et al. [32]. Taking the 3rd order case as an example, the reduction is based on the identity over Boolean variables

$$-xyz = \min_{w \in \{0,1\}} -w(x + y + z - 2) \tag{19}$$

If the 3rd order potential is non-positive, then the constructed pairwise potentials in the reduction are also non-positive and vice versa. This enables us to enforce submodularity on 3rd order energy minimization problems. Likewise, we can apply similar constraints for even higher order problems. Details for general higher order can be found in the supplementary material.

12 Analysis of Our Algorithm

The following theorems prove that our algorithm is both efficient and globally optimal.

Theorem 12.1. Correctness of the algorithm For any training datasets $\mathcal{D}$ and any $\varepsilon > 0$, if $(\mathbf{w}^*, \xi^*)$ is the optimal solution of Problem 4, then Algorithm 1 returns a solution $(\mathbf{w}, \xi)$ that has a better objective value than $(\mathbf{w}^*, \xi^*)$, and for which $(\mathbf{w}, \xi + \varepsilon)$ is feasible in Problem 4.

	Accu	macro P	macro R	Time	Speedup
[3]	81.45	76.79	70.07	4.11h	1.00
Ours	80.72	73.42	69.74	1.34h	**3.06**

Table 1: Performance comparison on the Cornell RGB-D Dataset (office scenes). The second column denotes the overall accuracy. The ’P’ and ’R’ here stand for precision and recall respectively. As defined in [3], the macro P or R equates to class average P or R.

Proof. The original proof presented in [35] holds, since it does not depend on any constraints involving only $\mathbf{w}$, and in our case, all separation oracles during training are exact. □

Theorem 12.2. Convergence of the algorithm Algorithm 1 terminates in polynomial time.

The proof is provided in the supplementary material. Briefly, the separation oracle terminates in polynomial time, and adding negative constraints does not change the nature of the convex optimization in line 5. Note that the actual convergence rate depends on the QP solver used for line 5.

13 Testing Time Inference

While we have a transformed and restricted problem during training, during testing we might still have a full potential matrix for each potential. The only limitation in the expressiveness of the formulation is that all the pairwise potentials are non-positive (in the sense of minimization). We show in our experiments that this restriction has limited effects on the overall accuracy. At testing time, the inference is performed independently on each example, and the error does not accumulate as it does at training time. If the graph is small or sparse, exact inference is possible through ILP. Otherwise, TRW-S [42] provides good approximation in practice [36] for general potentials.

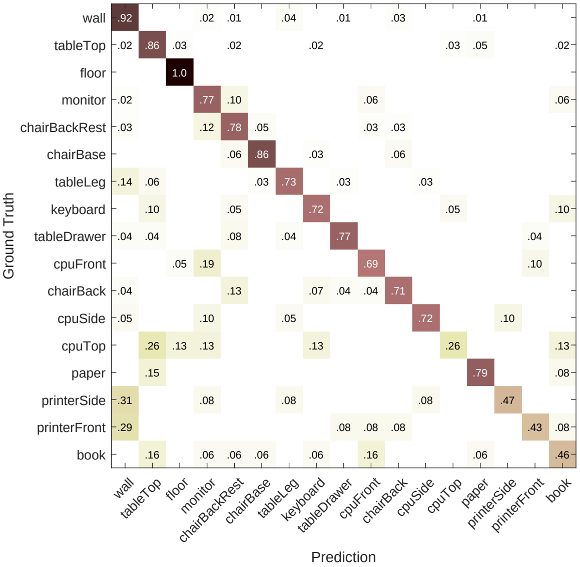
14 Experiments

We demonstrate the performance of our algorithm on the standard Cornell RGB-D dataset and a larger scale bridge dataset, which we created. On Cornell’s dataset, our algorithm runs three times faster while keeping the same level of accuracy as the competing method. On the bridge dataset, the competing methods are unable to solve the scene parsing problem due to the intractable seperation oracle. In contrast, our algorithm is able to solve it efficiently and accurately. In addition, we visualize the weights learned by our algorithm to show that our model captures domain knowledge.

14.1 Cornell RGB-D Dataset: Understanding 3D Scenes

The Cornell RGB-D dataset [49, 3] is an indoor point cloud dataset captured by Microsoft Kinect. The point clouds are obtained through merging multiple RGB-D views using the simultaneous localization and mapping (SLAM) algorithm. The point clouds are clustered into multiple segments. This dataset is suitable for testing structural learning prediction algorithms because it is necessary to take into account the neighborhood interaction for each node in order to label the segments correctly.

We compare our approach with [3] and use the same segmentation and features to ensure a fair comparison. The pairwise features cover visual appearance, local shape and geometry, and geometric context. Their algorithm adopts the persistency based approach in [25] (QPBO-R in Section 11.1). Note this method has no guarantee of optimality and an empirical heuristic needs to be adopted as discussed below. A variant of their algorithm makes use of additional class label information to limit the pairwise interactions to a predefined set of classes. The method assumes some labels are parts of an object, and restricts some potentials to be only among these labels. This information is usually not available on other structural datasets, so we do not include it in our comparison. The 4-fold cross-validation results are summarized in Table 1. The first row is taken from their paper. Our confusion matrix is shown in Figure 1. Notice that even with the additional constraints, our algorithm achieves approximately the same accuracy as [3] in 1/3 the time and with the critical advantage of a theoretical guarantee bounding the error.



	wall	tableTop	floor	monitor	chairBackRest	chairBase	tableLeg	keyboard	tableDrawer	cpuFront	chairBack	cpuSide	cpuTop	paper	printerSide	printerFront	book
wall	.92			.02	.01		.04		.01		.03			.01			
tableTop	.02	.86	.03		.02			.02					.03	.05			.02
floor			1.0														
monitor	.02			.77	.10					.06							.06
chairBackRest	.03			.12	.78	.05				.03	.03						
chairBase					.06	.86		.03			.06						
tableLeg	.14	.06				.03	.73		.03			.03					
keyboard		.10			.05			.72					.05				.10
tableDrawer	.04	.04			.08		.04		.77							.04	
cpuFront			.05	.19						.69						.10	
chairBack	.04				.13			.07	.04	.04	.71						
cpuSide	.05			.10			.05					.72			.10		
cpuTop		.26	.13	.13				.13					.26				.13
paper		.15												.79			.08
printerSide	.31			.08			.08					.08			.47		
printerFront	.29								.08	.08	.08					.43	.08
book		.16		.06	.06	.06		.06		.16				.06			.46

Fig. 1: Confusion matrix of our algorithm on the Cornell RGB-D Dataset (office scenes).

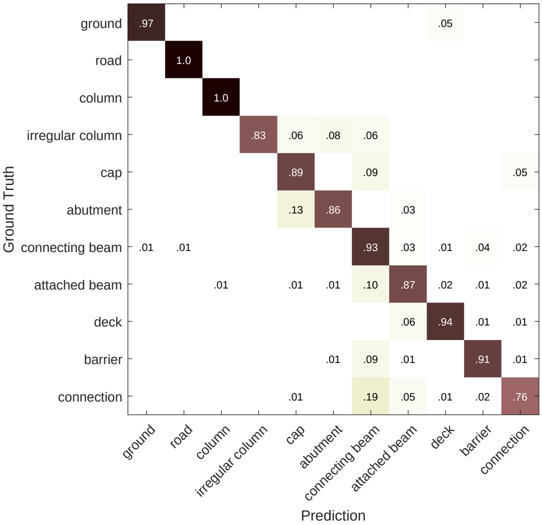


Fig. 2: Confusion matrix of our algorithm on the bridge dataset.

The competing method's implementation uses an undocumented heuristic that is vital for the learning procedure. In our algorithm, there is no need for this heuristic, because no relaxation is involved. Recall the rationale for interpreting an unlabeled node as 0.5 in Section 3. To compute the joint feature map $\Psi(\mathbf{x}, \mathbf{y})$, we need to compute $y_u^k y_v^l$ in (8). If both are unlabeled, then $y_u^k y_v^l$ would be 0.25. In [3], an additional measure is taken when neither side is labeled by QPBO:

- $y_u^k y_v^l$ is interpreted as 0.5, if the coefficient, i.e., $U_p^b(y_u^k, y_u^l)$, is positive;
- $y_u^k y_v^l$ is interpreted as 0, otherwise.

We found that without this rounding heuristic, the learning algorithm in [3] terminates after a dozen or fewer iterations with a newly found violation smaller than the violation of the current working set, which is impossible if the loss augmented inference is exact. Such early termination prevents the structral SVM from learning any meaningful potentials, and the prediction is usually a failure. This effect has been observed using both their implementation and our independent implementation on Cornell's RGB-D Dataset and the bridge dataset in next subsection.

14.2 Bridge Dataset: Scaling up to Complex Structures

For a second experiment, we tested out our algorithm on a domain-specific dataset to evaluate its performance against a large dataset with complex structures. To this aim, we created a synthetic but realistic bridge dataset (Figure 1) modeling complicated building structure. Such a dataset is useful for developing 3D reverse engineering

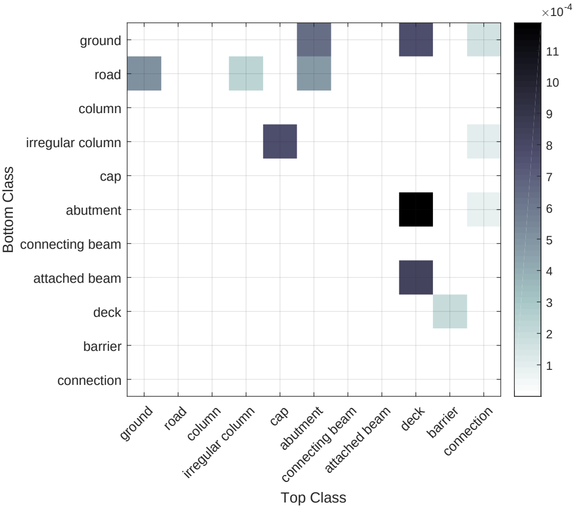


Fig. 3: The pairwise weights for the *on-top-of* feature. These weights capture domain knowledge for bridge architecture.

techniques, which can find their application in as-built Building Information Model (BIM) creation [99] and infrastructure inspection [83]. Unlike color or RGB-D images, full building laser scan datasets are scarce, thus we utilize a realistic synthetic dataset. We constructed CAD models of bridges, and generated the point clouds by placing a virtual laser scanner, complete with a noise model, in the scene as if we are actually conducting actual field scans. Multiple scans are taken per scene and merged into a single point cloud. In total, we have 25 bridge models of five different types. Each model contains 200k to 500k 3D points after down-sampling.

Similar to the Cornell RGB-D dataset, the task is to semantically label the segments, and we define eleven semantic classes for this dataset. We train a random forest classifier on SHOT descriptors [91] to obtain a label class distribution for each point. The descriptor encodes histogram of local surface information. We take the mean class distribution as the node feature for each segment. We use ground truth segmentation for benchmarking the contextual classification algorithms. We build a graph based on the physical adjacency of the segments and use on-top-of, principal direction consistency, and perpendicularity as three edge features. The accuracy is computed at the node level. On average, the bridge scenes contain ten times more segments and nine times more edges than the Cornell RGB-D dataset. We split the dataset into five folds, each containing five bridge models.

The cross-validation result is summarized in Figure 2 and visualized in Figure 4. We obtain 90.07% overall accuracy for semantic labeling the scene with 11 classes. For a single fold, the training takes 1.3 hours, and testing takes 89 seconds for five scenes. We attempted to use [55] and [3] as competing methods. However, the first

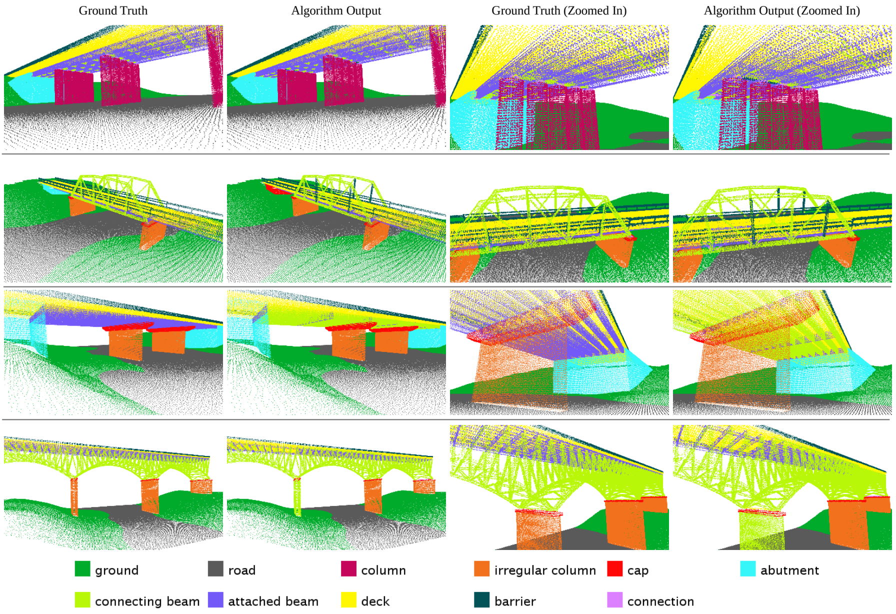


Fig. 4: Output of our algorithm on the bridge dataset. Some errors can be seen by comparing the 3rd and 4th columns.

fails due to the poor separation oracle and the latter could not handle this large scale of data and did not terminate after 7.5 days.

Capturing domain knowledge. Our algorithm is able to encode domain knowledge in the pairwise weights. For instance, we visualize the weights for the *on-top-of* feature in Figure 3. The feature is a binary indicator, and the product of this feature and the corresponding weight adds towards the overall score. The matrix reveals typical structural relationships seen in bridge architecture, e.g., the abutment and attached beam are usually placed beneath the deck.

15 Conclusion

In this thesis, we investigated the problems of identifying the computational complexity of energy minimization and estimating the parameters for energy minimization.

We showed that general energy minimization, even in the 2-label pairwise case, and planar energy minimization with three or more labels are exp-APX-complete. Our finding rules out the existence of any approximation algorithm with a sub-exponential approximation ratio in the input size for these two problems, including constant factor approximations. Moreover, we collected and reviewed the computational complexity of several subclass problems and arranged them on a complexity scale consisting of three major complexity classes – PO, APX, and exp-APX, corresponding to problems that are solvable, approximable, and inapproximable in polynomial time. Problems in the first two complexity classes can serve as alternative tractable formulations to the inapproximable ones. Our work can help vision researchers to select an appropriate model for an application or guide them in designing new algorithms. These altogether set up a new viewpoint for interpreting and classifying the complexity of optimization problems for the computer vision community. In the future, it will be interesting to consider the open questions of the complexity of structure-, rank-, and expectation-approximation for energy minimization.

For the parameter estimation of energy minimization, we proposed a method to overcome the problem caused by using unbounded approximation for the separation oracle in structural learning. Through exploiting the properties of the joint problem of training time inference and learning, we transformed the inapproximable inference problem into a polynomial time solvable one, thereby enabling tractable exact inference while still allowing an arbitrary graph structure and full potential interactions. We were able to retrieve the theoretical guarantees of structural SVMs that were lost when unbounded approximation was used.

Finally, we applied our structural learning algorithm to the 3D scene parsing task. The effectiveness and efficiency of our method was well-demonstrated on the Cornell RGB-D dataset and our bridge dataset.

References

[1] The Probabilistic Inference Challenge (2011), http://www.cs.huji.ac.il/project/PASCAL/

[2] Abdelbar, A., Hedetniemi, S.: Approximating MAPs for belief networks is NP-hard and other theorems. Artificial Intelligence 102(1), 21–38 (6 1998)

[3] Anand, A., Koppula, H.S., Joachims, T., Saxena, A.: Contextually guided semantic labeling and search for three-dimensional point clouds. IJRR p. 0278364912461538 (2012)

[4] Anguelov, D., Taskarf, B., Chatalbashev, V., Koller, D., Gupta, D., Heitz, G., Ng, A.: Discriminative learning of markov random fields for segmentation of 3d scan data. In: CVPR. vol. 2, pp. 169–176. IEEE (2005)

[5] Armeni, I., Sener, O., Zamir, A.R., Jiang, H., Brilakis, I., Fischer, M., Savarese, S.: 3d semantic parsing of large-scale indoor spaces. CVPR (2016)

[6] Arora, C., Banerjee, S., Kalra, P., Maheshwari, S.N.: Generic cuts: An efficient algorithm for optimal inference in higher order MRF-MAP. In: ECCV. pp. 17–30 (2012)

[7] Ausiello, G., Crescenzi, P., Gambosi, G., Kann, V., Marchetti-Spaccamela, A., Protasi, M.: Complexity and approximation: Combinatorial optimization problems and their approximability properties. Springer (1999)

[8] Bach, S.H., Huang, B., Getoor, L.: Unifying local consistency and max sat relaxations for scalable inference with rounding guarantees. In: AISTATS. JMLR Proceedings, vol. 38 (2015)

[9] Bach, S.H., Huang, B., Getoor, L.: Unifying local consistency and max sat relaxations for scalable inference with rounding guarantees. In: AISTATS. pp. 46–55 (2015)

[10] Bar-Yehuda, R., Even, S.: On approximating a vertex cover for planar graphs. In: Proceedings of the Fourteenth Annual ACM Symposium on Theory of Computing. pp. 303–309 (1982)

[11] Barbu, A.: Learning real-time MRF inference for image denoising. In: CVPR. pp. 1574–1581 (2009)

[12] Batra, D., Gallagher, A.C., Parikh, D., Chen, T.: Beyond trees: Mrf inference via outer-planar decomposition. In: CVPR. pp. 2496–2503 (2010)

[13] Boros, E., Hammer, P.L., Sun, X.: Network flows and minimization of quadratic pseudo-Boolean functions. Tech. Rep. RRR 17-1991, RUTCOR (May 1991)

[14] Boros, E., Hammer, P.: Pseudo-Boolean optimization. Tech. rep., RUTCOR (October 2001)

[15] Boros, E., Hammer, P.: Pseudo-Boolean optimization. Discrete Applied Mathematics 1-3(123), 155–225 (2002)

[16] Boros, E., Hammer, P.L.: Pseudo-boolean optimization. Discrete applied mathematics 123(1), 155–225 (2002)

[17] Boykov, Y., Funka-Lea, G.: Graph cuts and efficient nd image segmentation. IJCV 70(2), 109–131 (2006)

[18] Boykov, Y., Kolmogorov, V.: An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. PAMI 26(9), 1124–1137 (2004)

[19] Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. PAMI 23, 1222–1239 (November 2001)

[20] Chekuri, C., Khanna, S., Naor, J., Zosin, L.: A linear programming formulation and approximation algorithms for the metric labeling problem. SIAM Journal on Discrete Mathematics 18(3), 608–625 (2005)

[21] Chekuri, C., Khanna, S., Naor, J., Zosin, L.: Approximation algorithms for the metric labeling problem via a new linear programming formulation. In: In Symposium on Discrete Algorithms. pp. 109–118 (2001)

[22] Cohen, D., Cooper, M., Jeavons, P.: Principles and Practice of Constraint Programming, chap. A Complete Characterization of Complexity for Boolean Constraint Optimization Problems, pp. 212–226 (2004)

[23] Dahlhaus, E., Johnson, D.S., Papadimitriou, C.H., Seymour, P.D., Yannakakis, M.: The complexity of multiterminal cuts. SIAM Journal on Computing 23(4), 864–894 (1994)

[24] Felzenszwalb, P.F., Veksler, O.: Tiered scene labeling with dynamic programming. In: CVPR. pp. 3097–3104 (2010)

[25] Finley, T., Joachims, T.: Training structural svms when exact inference is intractable. In: ICML. pp. 304–311. ACM (2008)

[26] Forney Jr, G.D.: The viterbi algorithm. Proceedings of the IEEE 61(3), 268–278 (1973)

[27] Gurobi Optimization, I.: Gurobi optimizer reference manual (2015), http://www.gurobi.com

[28] Hammer, P.L.: Some network flow problems solved with pseudo-Boolean programming. Operation Research 13, 388–399 (1965)

[29] Hedau, V., Hoiem, D., Forsyth, D.: Recovering the spatial layout of cluttered rooms. In: ICCV. pp. 1849–1856. IEEE (2009)

[30] Hochbaum, D.S.: An efficient algorithm for image segmentation, Markov random fields and related problems. J. ACM 48(4), 686–701 (Jul 2001)

[31] Ishikawa, H.: Exact optimization for Markov random fields with convex priors. PAMI 25(10), 1333–1336 (2003)

[32] Ishikawa, H.: Transformation of general binary MRF minimization to the first-order case. PAMI 33(6), 1234–1249 (2011)

[33] Jancsary, J., Nowozin, S., Rother, C.: Learning convex qp relaxations for structured prediction. In: Proceedings of The 30th International Conference on Machine Learning. pp. 915–923 (2013)

[34] Jeavons, P., Krokhin, A., Živný, S., et al.: The complexity of valued constraint satisfaction. Bulletin of EATCS 2(113) (2014)

[35] Joachims, T., Finley, T., Yu, C.N.J.: Cutting-plane training of structural svms. Machine Learning 77(1), 27–59 (2009)

[36] Kappes, J.H., Andres, B., Hamprecht, F.A., Schnörr, C., Nowozin, S., Batra, D., Kim, S., Kausler, B.X., Kröger, T., Lellmann, J., et al.: A comparative study of modern inference techniques for structured discrete energy minimization problems. IJCV 115, 155–184 (2015)

[37] Karp, R.M.: Reducibility among combinatorial problems. In: Proceedings of a symposium on the Complexity of Computer Computations. pp. 85–103 (1972)

[38] Kleinberg, J., Tardos, E.: Approximation algorithms for classification problems with pairwise relationships: metric labeling and Markov random fields. J. ACM 49(5), 616–639 (2002)

[39] Kohli, P., Shekhovtsov, A., Rother, C., Kolmogorov, V., Torr, P.: On partial optimality in multi-label MRFs. In: ICML. pp. 480–487 (2008)

[40] Kolmogorov, V., Zabih, R.: What energy functions can be minimized via graph cuts? PAMI 26(2), 147–159 (February 2004)

[41] Kolmogorov, V.: Primal-dual algorithm for convex Markov random fields. Tech. Rep. MSR-TR-2005-117, Microsoft Research Cambridge (2005)

[42] Kolmogorov, V.: Convergent tree-reweighted message passing for energy minimization. PAMI 28(10), 1568–1583 (2006)

[43] Kolmogorov, V.: Minimizing a sum of submodular functions. Discrete Applied Mathematics 160(15), 2246 – 2258 (2012)

[44] Kolmogorov, V.: The power of linear programming for finite-valued CSPs: A constructive characterization. In: Automata, Languages, and Programming, Lecture Notes in Computer Science, vol. 7965, pp. 625–636 (2013)

[45] Kolmogorov, V., Zabin, R.: What energy functions can be minimized via graph cuts? PAMI 26(2), 147–159 (2004)

[46] Komodakis, N., Paragios, N., Tziritas, G.: MRF optimization via dual decomposition: Message-passing revisited. In: ICCV. pp. 1–8 (2007)

[47] Komodakis, N., Paragios, N.: Beyond loose LP-relaxations: Optimizing MRFs by repairing cycles. In: ECCV. pp. 806–820 (2008)

[48] Komodakis, N., Tziritas, G.: Approximate labeling via graph cuts based on linear programming. PAMI 29(8), 1436–1453 (2007)

[49] Koppula, H.S., Anand, A., Joachims, T., Saxena, A.: Semantic labeling of 3d point clouds for indoor scenes. In: NIPS. pp. 244–252 (2011)

[50] Kumar, M.P.: Rounding-based moves for metric labeling. In: NIPS. pp. 109–117 (2014)

[51] Kumar, M.P., Veksler, O., Torr, P.H.: Improved moves for truncated convex models. Journal of Machine Learning Research 12, 31–67 (Feb 2011)

[52] Kwisthout, J.: Most probable explanations in Bayesian networks: Complexity and tractability. International Journal of Approximate Reasoning 52(9), 1452 – 1469 (2011)

[53] Kwisthout, J.: Symbolic and Quantitative Approaches to Reasoning with Uncertainty: 12th European Conference, ECSQARU 2013, Utrecht, The Netherlands, July 8-10, 2013. Proceedings, chap. Structure Approximation of Most Probable Explanations in Bayesian Networks, pp. 340–351 (2013)

[54] Kwisthout, J.: Tree-width and the computational complexity of MAP approximations in Bayesian networks. Journal of Artificial Intelligence Research pp. 699–720 (2015)

[55] Lacoste-Julien, S., Jaggi, M., Schmidt, M., Pletscher, P.: Block-coordinate frank-wolfe optimization for structural svms. Machine Learning (2013)

[56] Lauritzen, S.L.: Graphical Models. No. 17 in Oxford Statistical Science Series, Oxford Science Publications (1998)

[57] Li, M., Shekhovtsov, A., Huber, D.: Complexity of discrete energy minimization problems. In: ECCV (2016)

[58] Liu, B., Gould, S., Koller, D.: Single image depth estimation from predicted semantic labels. In: CVPR. pp. 1253–1260 (2010)

[59] Mottaghi, R., Chen, X., Liu, X., Cho, N.G., Lee, S.W., Fidler, S., Urtasun, R., et al.: The role of context for object detection and semantic segmentation in the wild. In: CVPR. pp. 891–898. IEEE (2014)

[60] Munoz, D., Bagnell, J.A., Vandapel, N., Hebert, M.: Contextual classification with functional max-margin markov networks. In: CVPR. pp. 975–982. IEEE (2009)

[61] Nathan Silberman, Derek Hoiem, P.K., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV (2012)

[62] Orponen, P., Mannila, H.: On approximation preserving reductions: complete problems and robust measures. Technical Report (1987)

[63] Papadimitriou, C.H., Yannakakis, M.: Optimization, approximation, and complexity classes. Journal of Computer and System Sciences 43(3), 425 – 440 (1991)

[64] Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers Inc. (1988)

[65] Prusa, D., Werner, T.: How hard is the lp relaxation of the potts min-sum labeling problem. In: Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR). pp. 57–70 (2015)

[66] Prusa, D., Werner, T.: Universality of the local marginal polytope. PAMI 37(4), 898–904 (2015)

[67] Ramalingam, S., Kohli, P., Alahari, K., Torr, P.H.: Exact inference in multi-label crfs with higher order cliques. In: CVPR. pp. 1–8. IEEE (2008)

[68] Ramanan, D.: Dual coordinate solvers for large-scale structural svms. arXiv preprint arXiv:1312.1743 (2014)

[69] Ren, X., Bo, L., Fox, D.: Rgb-(d) scene labeling: Features and algorithms. In: CVPR. pp. 2759–2766. IEEE (2012)

[70] Roller, B.T.C.G.D.: Max-margin markov networks. NIPS 16, 25 (2004)

[71] Rother, C., Kolmogorov, V., Lempitsky, V., Szummer, M.: Optimizing binary MRFs via extended roof duality. In: CVPR. pp. 1–8 (2007)

[72] Schlesinger, D.: Exact solution of permuted submodular minsum problems. In: EMMCVPR. vol. 4679, pp. 28–38. Springer (2007)

[73] Schlesinger, M.I., Hlaváč, V.: Ten lectures on statistical and structural pattern recognition, Computational Imaging and Vision, vol. 24. Kluwer Academic Publishers, Dordrecht, The Netherlands (2002)

[74] Schraudolph, N.: Polynomial-time exact inference in NP-hard binary MRFs via reweighted perfect matching. In: AISTATS. JMLR Proceedings, vol. 9, pp. 717–724 (2010)

[75] Schrijver, A.: A combinatorial algorithm minimizing submodular functions in strongly polynomial time. Journal of Combinatorial Theory Series B(80), 346–355 (2000)

[76] Schwing, A.G., Urtasun, R.: Efficient exact inference for 3d indoor scene understanding. In: ECCV, pp. 299–313. Springer (2012)

[77] Shah, N., Kolmogorov, V., Lampert, C.H.: A multi-plane block-coordinate frank-wolfe algorithm for training structural svms with a costly max-oracle. In: CVPR. IEEE (2015)

[78] Shapovalov, R., Velizhev, A.: Cutting-plane training of non-associative markov network for 3d point cloud segmentation. In: 2011 International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission. pp. 1–8. IEEE (2011)

[79] Shekhovtsov, A., Swoboda, P., Savchynskyy, B.: Maximum persistency via iterative relaxed inference with graphical models. In: CVPR (2015)

[80] Shekhovtsov, A., Kohli, P., Rother, C.: Curvature prior for MRF-based segmentation and shape inpainting. In: DAGM/OAGM. pp. 41–51 (2012)

[81] Shih, W.K., Wu, S., Kuo, Y.S.: Unifying maximum cut and minimum cut of a planar graph. IEEE Transactions on Computers 39(5), 694–697 (May 1990)

[82] Silberman, N., Fergus, R.: Indoor scene segmentation using a structured light sensor. In: ICCV Workshops. pp. 601–608. IEEE (2011)

[83] Song, M., Huber, D.: Automatic recovery of networks of thin structures. 3DV (2015)

[84] Sontag, D., Choe, D.K., Li, Y.: Efficiently searching for frustrated cycles in MAP inference. In: Uncertainty in Artificial Intelligence (UAI). pp. 795–804 (2012)

[85] Szeliski, R., Zabih, R., Scharstein, D., Veksler, O., Kolmogorov, V., Agarwala, A., Tappen, M., Rother, C.: A comparative study of energy minimization methods for Markov random fields with smoothness-based priors. PAMI 30(6), 1068–1080 (2008)

[86] Szummer, M., Kohli, P., Hoiem, D.: Learning crfs using graph cuts. In: ECCV, pp. 582–595. Springer (2008)

[87] Taskar, B., Chatalbashev, V., Koller, D.: Learning associative markov networks. In: ICML. p. 102. ACM (2004)

[88] Teo, C.H., Smola, A., Vishwanathan, S., Le, Q.V.: A scalable modular convex solver for regularized risk minimization. In: SIGKDD. pp. 727–736. ACM (2007)

[89] Thapper, J., Živný, S.: The power of linear programming for valued CSPs. In: Symposium on Foundations of Computer Science (FOCS). pp. 669–678 (2012)

[90] Thapper, J., Živný, S.: The complexity of finite-valued CSPs. In: Symposium on the Theory of Computing (STOC). pp. 695–704 (2013)

[91] Tombari, F., Salti, S., Di Stefano, L.: Unique signatures of histograms for local surface description. In: ECCV, pp. 356–369. Springer (2010)

[92] Topkis, D.M.: Minimizing a submodular function on a lattice. Operations Research 26(2), 305–321 (1978)

[93] Tsochantaridis, I., Hofmann, T., Joachims, T., Altun, Y.: Support vector machine learning for interdependent and structured output spaces. In: ICML. p. 104. ACM (2004)

[94] Veksler, O.: Graph cut based optimization for MRFs with truncated convex priors. In: CVPR. pp. 1–8 (2007)

[95] Vineet, V., Warrell, J., Torr, P.H.S.: A tiered move-making algorithm for general pairwise MRFs. In: CVPR. pp. 1632–1639 (2012)

[96] Živný, S., Cohen, D.A., Jeavons, P.G.: The expressive power of binary submodular functions. Discrete Applied Mathematics 157(15), 3347 – 3358 (2009)

[97] Werner, T.: A linear programming approach to max-sum problem: A review. PAMI 29(7), 1165–1179 (July 2007)

[98] Werner, T.: A linear programming approach to max-sum problem: A review. PAMI 29(7), 1165–1179 (2007)

[99] Xiong, X., Adan, A., Akinci, B., Huber, D.: Automatic creation of semantically rich 3d building models from laser scanner data. Automation in Construction 31, 325–337 (2013)

[100] Xiong, X., Huber, D.: Using context to create semantic 3d models of indoor environments. In: BMVC. pp. 1–11 (2010)

[101] Xu, L., Jia, J., Matsushita, Y.: Motion detail preserving optical flow estimation. PAMI 34(9), 1744–1757 (2012)

[102] Yang, Y., Ramanan, D.: Articulated pose estimation with flexible mixtures-of-parts. In: CVPR. pp. 1385–1392 (2011)

[103] Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: CVPR. pp. 2879–2886. IEEE (2012)

[104] Živný, S., Werner, T., Průša, D.a.: The Power of LP Relaxation for MAP Inference, pp. 19–42. The MIT Press, Cambridge, USA (December 2014)

Appendix

A Formal Proofs

Note for all proofs in this section, we assign integer values to Boolean functions: 0 for False and 1 for True.

A.1 General Case

Theorem 5.1. QPBO is exp-APX-complete.

Proof. We reduce from the following problem.

Problem 1 ([7], Section 8.3.2). **W3SAT-triv**

INSTANCE: Boolean CNF formula F with variables $x_1, \cdots, x_n$ and each clause assuming exactly 3 variables; non-negative integer weights $w_1, \cdots, w_n$.

SOLUTION: Truth assignment τ to the variables that either satisfies F or assigns the trivial, all-true assignment.

MEASURE: $\min \sum_{i=1}^{n} w_i \tau(x_i)$.

W3SAT-triv is known to be exp-APX-complete [7]. We use an AP-reduction from W3SAT-triv to prove the same completeness result for QPBO. The optimal value of W3SAT-triv is upper bounded by $M := \sum_i w_i$ because the all-true assignment is feasible. The objective weight is represented in QPBO as unary terms $f_i(x_i) = w_i x_i$. For every Boolean clause $C(x_i, x_j, x_k) \in F$ we construct a triple-wise term

$$\delta_{ijk}(x_i, x_j, x_k) = M(1 - C(x_i, x_j, x_k)). \tag{1}$$

This term takes the large value M iff C is not satisfied and 0 otherwise. Further, the Boolean clause $C(x_i, x_j, x_k)$ can be represented uniquely as a multi-linear cubic polynomial. For example, a clause $x_1 \vee \bar{x}_2 \vee \bar{x}_3$ can be represented as

$$1 - (1 - x_1)x_2x_3 = x_1x_2x_3 - x_2x_3 + 1. \tag{2}$$

Then we obtain similar representation with a single third order term and a second order multi-linear polynomial for δ_{ijk}:

$$\delta_{ijk} = M(ax_ix_jx_k + \sum_J b_J \prod_{l \in J} x_l), \tag{3}$$

where $J \subseteq \{i, j, k\}, |J| \leq 2$, $\prod_{l \in J} x_l$ is set to 1 if J is empty, $a \in \{-1, 1\}$, and $b_J \in \{-1, 0, 1\}$. We now apply the quadratization techniques [32] to δ_{ijk}. After introducing an auxiliary variable x_w with $w > n$, we observe the following identities:

$$-x_i x_j x_k = \min_{x_w \in \{0,1\}} -x_w(x_i + x_j + x_k - 2) \tag{4}$$

$$x_i x_j x_k = \min_{x_w \in \{0,1\}} \Big((x_w - 1)(x_i + x_j + x_k - 1) + (x_i x_j + x_i x_k + x_j x_k)\Big) \tag{5}$$

In either case, substituting the cubic term $a x_i x_j x_k$ in δ_{ijk} with the expression inside the min operator, we can have a unified quadratic form

$$\psi_{ijk} := M \sum_{J_w} b_{J_w} \prod_{l \in J_w} x_l, \tag{6}$$

where $J_w \subseteq \{i, j, k, w\}, |J_w| \leq 2$ and $\prod_{i \in J_w} x_i$ is set to 1 if J_w is empty. In both cases, the quadratic form takes the same optimal values as its cubic counterpart given the optimal assignment, i.e.,

$$\min_{x_i, x_j, x_k, x_w} \psi_{ijk} = \min_{x_i, x_j, x_k} \delta_{ijk}, \tag{7}$$

but the transformation expands the original range of the cubic term from $\{-1, 0\}$ to $\{-1, 0, 1, 2\}$ and from $\{0, 1\}$ to $\{0, 1, 3\}$ respectively for $a = -1$ and $a = 1$. Therefore, the cost of the constructed instance of QPBO is bounded in the absolute value by $3M$ and the number of added variables is exactly the number of clauses in F. Clearly, this construction can be computed in polynomial time. *Note that when approximation is used, this transformation is no longer exact ($\psi_{ijk} \neq \delta_{ijk}$), as the optimality of the auxiliary variable x_w cannot be guaranteed. However, it can be verified that under all possible assignments (ignoring the min operator) in either case, $\psi_{ijk} \geq 0$, which is the key for the reduction to be an approximation preserving (AP) one.*

The construction above defines a mapping π from any instance of W3SAT-triv ($p_1 \in I_1$) to an instance of QPBO ($p_2 \in I_2$). The mapping σ from feasible solutions of p_2 ($x \in S_2(p_2)$) to that of p_1 is defined as follows: if $f(x) \geq M$, then let the mapped solution $\sigma(p_1, x)$ be the all true assignment, otherwise let the mapped solution $\sigma(p_1, x)$ be $x_i, i \in \{1, ..., n\}$.

Now, we need to show that (π, σ) together with a constant α is an AP-reduction. Let m_1, m_2, m_1^* and m_2^* to be short for $m_1(p_1, \sigma(p_1, x))$, $m_2(p_2)$, $m_1^*(p_1)$, and $m_2^*(\pi(p_2))$ respectively, where $*$ indicates the optimal solution. First, note that $\sigma(p_1, x)$ is always feasible for *W3SAT*-triv: either it satisfies F or $f(x) \geq M$ and therefore $\sigma(p_1, x)$ is the all-true assignment. In the first case, since every quadratic term is non-negative, we have

$$m_1 = \sum_{i=1}^{n} x_i w_i \tag{8}$$

$$\leq \sum_{i=1}^{n} x_i w_i + \sum_{C_{ijk} \in F} \psi_{ijk}(x_i, x_j, x_k) = f(x) = m_2. \tag{9}$$

In the second case, by construction

$$m_1 = M \leq f(x) = m_2. \tag{10}$$

Therefore, no matter which case $m_1 \leq m_2$.

Now for the optimal solution, if F is satisfiable, then by construction $m_1^* = m_2^*$. Recall from Definition 4.5, $R = m/m^*$. For any instance $p_1 \in I_1$, for any rational $r > 1$, and for any $x \in S_2(p_2)$, if

$$R_2(p_2, x) \leq r, \tag{11}$$

then

$$m_1 \leq m_2 \leq r m_2^* = r m_1^* \tag{12}$$

$$R_1(p_1, \sigma(p_1, x)) = \frac{m_1}{m_1^*} \leq r \tag{13}$$

If F is not satisfiable, $m_1^* = M \leq m_2^*$ and $m_2 \geq m_2 * \geq M$. Thus, for any instance $p_1 \in I_1$, for any rational $r > 1$, and for any $x \in S_2(p_2)$,

$$R_1(p_1, \sigma(p_1, x)) = \frac{m_1}{m_1^*} = \frac{M}{M} = 1 \leq r \tag{14}$$

Therefore $(\pi, \sigma, 1)$ is an AP-reduction. Since W3SAT-triv is exp-APX-complete and QPBO is in exp-APX, we prove that QPBO is exp-APX-complete. □

Corollary 5.2. *k-label energy minimization is exp-APX-complete for $k \geq 2$.*

Proof. We create an AP-reduction from QPBO to k-label energy minimization by setting up the unary and pairwise terms to discourage a labeling with the additional $k - 2$ labels.

Denote QPBO as $\mathcal{P}_1 = (\mathcal{I}_1, \mathcal{S}_1, m_1, \min)$ and k-label energy minimization as $\mathcal{P}_2 = (\mathcal{I}_2, \mathcal{S}_2, m_2, \min)$. Given an instance $p_1 = (\mathcal{G} = (\mathcal{V}, \mathcal{E}), \mathcal{L}_1, f) \in \mathcal{I}_1$, let $M(p_1)$ be a large number such that all for all $\mathbf{x}_1 \in \mathcal{S}_1$, $m_1 < M$. For example, we can let

$$M = \sum_{u \in \mathcal{V}} \sum_{x_u \in \mathcal{L}_1} |f_u(x_u)| + \sum_{(u,v) \in \mathcal{E}} \sum_{x_u \in \mathcal{L}_1} \sum_{x_v \in \mathcal{L}_1} |f_{uv}(x_u, x_v)| + 1. \tag{15}$$

We define the forward mapping π from any $p_1 \in I_1$ to $p_2 = (\mathcal{G} = (\mathcal{V}, \mathcal{E}), \mathcal{L}_2, g) \in I_2$ as follows:

- $g_u(a) = f_u(a)$, for $\forall a \in \mathcal{L}_1$, and $\forall u \in \mathcal{V}$;
- $g_u(a) = M$, for $\forall a \notin \mathcal{L}_1$, and $\forall u \in \mathcal{V}$;
- $g_{uv}(a, b) = f_{uv}(a, b)$, for $\forall a, b \in \mathcal{L}_1$, and $\forall (u, v) \in \mathcal{E}$;
- $g_{uv}(a, b) = M$ if either a or $b \notin \mathcal{L}_1$ for $\forall (u, v) \in \mathcal{E}$.

This setup has two properties:

- $m_2 \geq M$ if and only if the labeling $\mathbf{x}_2 \in \mathcal{S}_2$ includes labels that are not in $\mathcal{L}_1$;
- $m_1^* = m_2^*$, for any p_1 and $p_2 = \pi(p_1)$.

Then we define the reverse mapping σ from any $(p_2, \mathbf{x}_2)$ to $\mathbf{x}_1 \in \mathcal{S}_1$ to be

- $\mathbf{x}_1 = \mathbf{x}_2$, if $m_2 < M$;
- $\mathbf{x}_1$ be any fixed feasible solution (e.g., all nodes are labeled as the first label), if $m_2 \geq M$.

Observe that in both cases, $m_1 \leq m_2$. For any instance $p_1 \in I_1$, for any rational $r > 1$, and for any $\mathbf{x}_2 \in S_2$, if

$$R_2(p_2, \mathbf{x}_2) = \frac{m_2}{m_2^*} \leq r, \tag{16}$$

then

$$m_1 \leq m_2 \leq r m_2^* = r m_1^* \tag{17}$$

$$R_1(p_1, \mathbf{x}_1) = \frac{m_1}{m_1^*} \leq r \tag{18}$$

Therefore $(\pi, \sigma, 1)$ is an AP-reduction. As QPBO is exp-APX-complete and all energy minimization problems are in exp-APX, we conclude that k-label energy minimization is exp-APX-complete for $k \geq 2$. □

The above construction also formally shows that the energy minimization problem can only become harder when having a larger labeling space, irrespective of the graph structure and the interaction type.

A.2 Planar Case

Theorem 6.1. Planar 3-label energy minimization is exp-APX-complete.

Proof. We create an AP-reduction from 3-label energy minimization to planar 3-label energy minimization by introducing polynomially many auxiliary nodes and edges.

Denote 3-label energy minimization as $\mathcal{P}_1 = (\mathcal{I}_1, \mathcal{S}_1, m_1, \min)$ and planar 3-label energy minimization as $\mathcal{P}_2 = (\mathcal{I}_2, \mathcal{S}_2, m_2, \min)$. Given an instance $p_1 \in \mathcal{I}_1$, we compute a large number $M(p_1)$ as in Equation (15) in the proof for Corollary 5.2.

The gadget-based reduction presented in Section 6, defines a forward mapping π from any $p_1 = (\mathcal{G}_1 = (\mathcal{V}_1, \mathcal{E}_1), \mathcal{L}, f) \in I_1$ to $p_2 = (\mathcal{G}_2 = (\mathcal{V}_2, \mathcal{E}_2), \mathcal{L}, g) \in I_2$. Let $\mathcal{V}_3$ be the nodes added during the reduction, then $\mathcal{V}_2 = \mathcal{V}_1 \cup \mathcal{V}_3$. The two gadgets SPLIT and UNCROSSCOPY are used 4 times each to replace an edge crossing (point of intersection not at end points) with a planar representation (Figure 4), introducing 22 auxiliary nodes. Since the gadgets can be drawn arbitrarily small so that they are not intersecting with any other edges, we can repeatedly replace all edge crossings in $\mathcal{G}_1$ with this representation. There can be up to $O(|\mathcal{E}_1|^2)$ edge crossings, and we have $|\mathcal{V}_3| = O(|\mathcal{E}_1|^2)$. Given that the reduction adds only a polynomial number of auxiliary nodes, the forward mapping π can be computed by a polynomial time algorithm.

This setup has two properties:

- $m_2 \leq M$ if and only if the labeling $\mathbf{x}_1$ is the same as the partial labeling in $\mathbf{x}_2$ restricting to nodes in $\mathcal{V}_1$ in $\mathcal{G}_2$.
- $m_1^* = m_2^*$, for any p_1 and $p_2 = \pi(p_1)$.

Then we define the reverse mapping σ from any $(p_2, \mathbf{x}_2)$ to $\mathbf{x}_1 \in \mathcal{S}_1$ to be

- $\mathbf{x}_1 = \mathbf{x}_2$, if $m_2 < M$;
- $\mathbf{x}_1$ be any fixed feasible solution (e.g., all nodes are labeled as the first label), if $m_2 \geq M$.

Observe that in both cases, $m_1 \leq m_2$. For any instance $p_1 \in I_1$, for any rational $r > 1$, and for any $\mathbf{x}_2 \in S_2$, if

$$R_2(p_2, \mathbf{x}_2) = \frac{m_2}{m_2^*} \leq r, \tag{19}$$

then

$$m_1 \leq m_2 \leq r m_2^* = r m_1^* \tag{20}$$

$$R_1(p_1, \mathbf{x}_1) = \frac{m_1}{m_1^*} \leq r \tag{21}$$

Therefore $(\pi, \sigma, 1)$ is an AP-reduction. As 3-label energy minimization is exp-APX-complete (Corollary 5.2) and all energy minimization problems are in exp-APX, we conclude that planar 3-label energy minimization is exp-APX-complete.

□

Corollary 6.2. Planar k-label energy minimization is exp-APX-complete, for $k \geq 3$.

Proof. The proof of Corollary 5.2 is graph structure independent. Therefore, the same proof applies here. □

B Relation to Bayesian Networks

There are substantial differences between results for Bayesian networks [2] and our result. Bayesian networks have a probability density function $p(x)$ that factors according to a directed acyclic graph, e.g., as $p(x_1, x_2, x_3) = p(x_1|x_2, x_3)p(x_2)p(x_3)$. Finding the MAP assignment (same as the most probable estimate (MPE)) in a Bayesian network is related to energy minimization (1) by letting $f(x) = -\log(p(x))$. The product is transformed into the sum and so, e.g., factor $p(x_1|x_2, x_3)$ corresponds to term $f_{1,2,3}(x_1, x_2, x_3)$.

The inapproximability result of Abdelbar and Hedetniemi [2] holds even when restricting to binary variables and factors of order three. However, [2, Section 6.1] count incoming edges of the network. For a factor $p(x_1|x_2, x_3)$, there are two, but the total number of variables it couples is three and therefore such a network does not correspond to QPBO. If one restricts to factors of at most two variables, e.g., $p(x_1|x_2)$, in a Bayesian network, then only tree-structured models can be represented, which are easily solvable.

In the other direction, representing pairwise energy (1) as a Bayesian network may require to use factors of order up to $|\mathcal{V}|$ composed of conditional probabilities of the form $p(x_i \mid x_j, x_k, \cdots)$ with the number of variables depending on the vertex degrees. It is seen that while the problems in their most general forms are convertible, fixed-parameter classes (such as order and graph restrictions) differ significantly. In addition, the approximation ratio for probabilities translates to an absolute approximation (an additive bound) for energies. The next corollary of our main result illustrates this point.

Corollary B.1. It is NP-hard to approximate MAP in the value of probability (2) with any exponential ratio $\exp(r(n))$, where r is polynomial.

Proof. Recall that the probability $p(x)$ is given by the exponential map of the energy: $p(x) = \exp(-f(x))$. Assume for contradiction that there is a polynomial time algorithm $\mathcal{A}$ that finds solution x and a polynomial $r(n) \geq 0$ for $n > 0$ such that

$$\frac{p(x^*)}{p(x)} \leq e^{r(n)} \tag{22}$$

for all instances of the problem. Taking the logarithm,

$$-f(x^*) + f(x) \leq r(n). \tag{23}$$

or,

$$f(x) \leq r(n) + f(x^*). \tag{24}$$

Divide by $f(x^*)$, which, by definition of NPO is positive, we obtain

$$\frac{f(x)}{f(x^*)} \leq 1 + \frac{1}{f(x^*)} r(n) \leq 1 + r(n). \tag{25}$$

where we have used that $f(x^*)$ is integer and positive and hence it is greater or equal to 1. Inequality (25) provides a polynomial ratio approximation for energy minimization. Since the latter is exp-APX-complete (Corollary 5.2), this contradicts existence of the polynomial algorithm $\mathcal{A}$, unless P = NP. □

Note, this corollary provides a stronger inapproximability result for probabilities than was proven in [2].

Remark B.2. Abdelbar and Hedetniemi [2] have shown also the following interesting facts. For Bayesian networks, the following problems are also APX-hard (in the value of probability):

- Given the optimal solution, approximate the second best solution;
- Given the optimal solution, approximate the optimal solution conditioned on changing the assignment of one variable.

C Miscellaneous

This result is used in Section 7.2.

Corollary C.1. An $O(\log k)$-approximation implies an $O(\log |x|)$-approximation for k-label energy minimization problems.

Proof. Observe that an instance of the energy minimization problem (1) is completely specified by a set of all unary terms f_u and pairwise terms f_{uv}. This defines a natural encoding scheme to describe an instance of an energy minimization problem with binary alphabet $\{0, 1\}$. Assume each potential is encoded by d digits, the input size

$$|x| = O((k|\mathcal{V}| + k^2|\mathcal{V}|^2)d) = O(k^2|\mathcal{V}|^2). \tag{26}$$

For an $O(\log k)$-approximation algorithm, the performance ratio

$$r = O(\log k) = O(\log k + \log |\mathcal{V}|) = O(\log k|\mathcal{V}|) = O(\log |x|), \tag{27}$$

which implies an $O(\log |x|)$-approximation algorithm.

□

D Proof for Convergence of Algorithm 1

Convergence has been proven in [35, 88] for 1-slack structural SVMs. Here, we show that similar results hold for problems with non-negative constraints. The proof constructs a line search to bound the increase in the objective in each iteration. The non-negative constraints can bring additional increase for the objective when they are activated, resulting in possibly fewer iterations. Symbols used in the proof are summarized in Table 1.

Problem 2. **Primal QP**

Using the new notations, the QP in Algorithm 1, line 5 can be written as

$$\min_{\mathbf{w},\xi\geq 0} \quad \frac{1}{2}||\mathbf{w}||^2 + C\xi \tag{28}$$

$$\text{s.t.} \quad \mathbf{H}^\intercal \mathbf{w} \geq l - \xi\mathbf{1}, \tag{29}$$

$$\mathbf{w}_P \geq \mathbf{0} \tag{30}$$

The Lagrangian is

$$\begin{aligned} L(\mathbf{w}, \xi, \alpha, \beta, \gamma) = \frac{1}{2}||\mathbf{w}||^2 + C\xi \\ - \alpha^\intercal[\mathbf{H}^\intercal\mathbf{w} - l + \xi\mathbf{1}] - \beta^\intercal\mathbf{w} - \gamma\xi \end{aligned} \tag{31}$$

Symbols	Definitions
t	iteration count for Algorithm 1
h_t	$\frac{1}{n}\sum_{i=1}^{n}[\varPsi(\mathbf{x}_i,\mathbf{y}_i)-\varPsi(\mathbf{x}_i,\bar{\mathbf{y}}_i)]$ for all $\bar{\mathbf{y}}_i$ added in the t-th iteration
d_t	$\frac{1}{n}\sum_{i=1}^{n}\varDelta_b(\mathbf{y}_i,\bar{\mathbf{y}}_i)$ for all $\bar{\mathbf{y}}_i$ added in the t-th iteration
$\mathbf{H}$ or $\mathbf{H}_t$	$[h_1\ h_2\ ...\ h_t]$
l or l_t	$[d_1\ d_2\ ...\ d_t]^\intercal$
R	$\max_{\forall i,\bar{\mathbf{y}}}\|\|\varPsi(\mathbf{x}_i,\mathbf{y}_i)-\varPsi(\mathbf{x}_i,\bar{\mathbf{y}}_i)\|\|_2$
$\varDelta$	$\max_{\forall i,\bar{\mathbf{y}}}\varDelta_b(\mathbf{y}_i,\bar{\mathbf{y}})$
α	the dual variables for margin violation
β	the dual variables for non-negativity
$(\mathbf{w}^*,\xi^*)$	the optimal solution of Problem 4.1
(α^*,β^*)	corresponding dual variables for $(\mathbf{w}^*,\xi^*)$
$J_t(\mathbf{w})$	the primal objective value of the QP in Algorithm 1, line 5 at the t-th iteration
$D_t(\alpha,\beta)$	the dual objective value of the QP in Algorithm 1, line 5 at the t-th iteration
δ_t	$D_t(\alpha^*,\beta^*)-D_t(\alpha_t,\beta_t)$

Table 1: List of symbols for the convergence proof. (Section D)

Setting the differential of L with respect to $\mathbf{w}$ to zero yields

$$\mathbf{w}=\mathbf{H}\alpha+\beta \tag{32}$$

Setting the differential of L with respect to ξ to zero yields

$$C-\alpha^\intercal\mathbf{1}=\gamma\geq 0 \tag{33}$$

Note that we define β to be a vector of the same length as $\mathbf{w}$ for simplicity. $(\beta)_j$ is fixed to zero for every coordinate j not required to be non-negative ($j\notin P$).

Problem 3. **Dual QP**

The dual problem is obtained by substituting equations (32) and (33) (KKT-conditions) into the Lagrangian

$$\max_{\alpha\geq\mathbf{0},\beta\geq\mathbf{0}}\quad -\frac{1}{2}\alpha^\intercal\mathbf{H}^\intercal\mathbf{H}\alpha-\beta^\intercal\mathbf{H}\alpha+l^\intercal\alpha-\frac{1}{2}\beta^\intercal\beta \tag{34}$$

$$\text{s.t.}\quad \alpha^\intercal\mathbf{1}\leq C \tag{35}$$

Initially, the working set $\mathcal{W}$ is empty and $J_1=D_1=0$. The trivial solution $\mathbf{w}=\mathbf{0}$ generates an upper bound $C\varDelta$ for the optimality gap δ_t. Next, we show that this gap can be closed through a constant increase in the dual objective in each iteration. The QP is solved by a QP solver in Algorithm 1. However, we cannot bound the change of the objective value. Instead, we resort to a series of line searches. There are two sets of dual variables, α and β. In each iteration, we optimize α, keeping β fixed, and then optimize β, keeping α fixed. The following lemma is introduced to bound the minimal increase in the objective with a line search in α.

Lemma D.1. For any unconstrained quadratic program,

$$f(\mathbf{x}) = -\frac{1}{2}\mathbf{x}^{\intercal}\mathbf{A}\mathbf{x} + \mathbf{b}^{\intercal}\mathbf{x} \tag{36}$$

with positive semi-definite $\mathbf{A}$, a line search starting at $\mathbf{x}$ with maximum step-size s towards a direction $\mathbf{g}$, such that $\nabla f(\mathbf{x})^{\intercal}\mathbf{g} \geq 0$ and $\mathbf{g}^{\intercal}\mathbf{A}\mathbf{g} \neq 0$, increases the objective by at least

$$\begin{aligned}\max_{0\leq\lambda\leq s}&[f(\mathbf{x}{+}\lambda\mathbf{g}) - f(\mathbf{x})]\\ &\geq \frac{1}{2}\min\left\{s\nabla f(\mathbf{x})^{\intercal}\mathbf{g}, \frac{[\nabla f(\mathbf{x})^{\intercal}\mathbf{g}]^2}{\mathbf{g}^{\intercal}\mathbf{A}\mathbf{g}}\right\}\end{aligned} \tag{37}$$

The first case applies when $\frac{\nabla f(\mathbf{x})^{\intercal}\mathbf{g}}{\mathbf{g}^{\intercal}\mathbf{A}\mathbf{g}} > s$, while the latter applies when $\frac{\nabla f(\mathbf{x})^{\intercal}\mathbf{g}}{\mathbf{g}^{\intercal}\mathbf{A}\mathbf{g}} \leq s$.

Proof.

$$f(\mathbf{x} + \lambda\mathbf{g}) - f(\mathbf{x}) = -\frac{1}{2}\mathbf{g}^{\intercal}\mathbf{A}\mathbf{g}\lambda^2 + \nabla f(\mathbf{x})^{\intercal}\mathbf{g}\lambda \tag{38}$$

is a simple quadratic function in λ restricted to $[0, s]$. When $\frac{\nabla f(\mathbf{x})^{\intercal}\mathbf{g}}{\mathbf{g}^{\intercal}\mathbf{A}\mathbf{g}} \leq s$, its optimal value is obtained at $\lambda^* = \frac{\nabla f(\mathbf{x})^{\intercal}\mathbf{g}}{\mathbf{g}^{\intercal}\mathbf{A}\mathbf{g}}$, with value $\frac{[\nabla f(\mathbf{x})^{\intercal}\mathbf{g}]^2}{2\mathbf{g}^{\intercal}\mathbf{A}\mathbf{g}}$; and when $\frac{\nabla f(\mathbf{x})^{\intercal}\mathbf{g}}{\mathbf{g}^{\intercal}\mathbf{A}\mathbf{g}} > s$, its optimal value is obtained at $\lambda^* = s$, with value $\nabla f(\mathbf{x})^{\intercal}\mathbf{g}s - \frac{1}{2}\mathbf{g}^{\intercal}\mathbf{A}\mathbf{g}s^2 \geq \frac{1}{2}s\nabla f(\mathbf{x})^{\intercal}\mathbf{g}$. □

Consider at the beginning of iteration $(t + 1)$, t constraints have been added for the QP. We want to optimize this new QP based on the previous iteration's solution (α, β). Keeping β fixed, the line search in α is constructed as:

$$\tilde{\alpha}(\lambda) := [-\lambda\alpha^{\intercal},\ \lambda C]^{\intercal}, \quad \lambda \in [0, 1] \tag{39}$$

Note the direction ($\tilde{\alpha} = [-\alpha_t^{\intercal},\ C]$) is chosen so that by construction, $\alpha + \tilde{\alpha}(\lambda)$ is always in the feasible region. In order to apply Lemma D.1, we need to bound $\nabla D^{\intercal}\tilde{\alpha}$ and $\tilde{\alpha}^{\intercal}\mathbf{H}^{\intercal}\mathbf{H}\tilde{\alpha}$.

Due to strong duality,

$$\frac{\partial D(\alpha, \beta)}{\partial\alpha} = l - \mathbf{H}^{\intercal}(\mathbf{H}\alpha + \beta) = l - \mathbf{H}^{\intercal}\mathbf{w}, \tag{40}$$

and due to complementary slackness, for each non-zero component i of α,

$$\frac{\partial D(\alpha,\beta))}{\partial(\alpha)_i} = d_i - h_i^{\intercal}\mathbf{w} = \xi \tag{41}$$

For $(\alpha)_t$ corresponding to the newly added constraint and some μ, by construction of Algorithm 1

$$\frac{\partial D(\alpha,\beta))}{\partial \alpha_t} = d_t - h_t^{\intercal}\mathbf{w} = \xi + \mu \geq \xi + \varepsilon \tag{42}$$

Therefore

$$\nabla D^{\intercal}\tilde{\alpha} = -\mathbf{1}^{\intercal}\alpha\xi + C(\xi+\mu) = C\mu \tag{43}$$

On the other hand

$$\begin{aligned}\tilde{\alpha}^{\intercal}\mathbf{H}^{\intercal}\mathbf{H}\tilde{\alpha} &= \tilde{\alpha}^{\intercal}\mathbf{H}_t^{\intercal}\mathbf{H}_t\tilde{\alpha} \\ &= \alpha^{\intercal}\mathbf{H}_{t-1}^{\intercal}\mathbf{H}_{t-1}\alpha - 2C\mathbf{1}^{\intercal}\mathbf{H}_{t-1}^{\intercal}\mathbf{H}_{t-1}\alpha + C^2 h_t^2 \qquad (44) \\ &\leq C^2R^2 + 2C^2R^2 + C^2R^2 \qquad (45) \\ &= 4C^2R^2 \qquad (46)\end{aligned}$$

Applying Lemma D.1, we have

$$\max_{0\leq\lambda\leq 1}\left[D(\alpha+\tilde{\alpha}(\lambda),\beta) - D(\alpha,\beta)\right] \geq \min\left\{\frac{\mu}{2}, \frac{\mu^2}{4C^2R^2}\right\} \tag{47}$$

We update the α using the line search above and then optimize β assuming α fixed. The dual problem 3 is a quadratic function with a diagonal quadratic matrix. Thus there is no interaction between each coordinate of β, and they can be optimized independently.

The optimal solution is

$$\forall j \in P, \quad (\beta^*)_j = \max\left(0, -(\mathbf{H}\alpha)_j\right) \tag{48}$$

with an increase in the objective

$$\begin{aligned}&\frac{1}{2}(\beta)_j^2 + (\mathbf{H}\alpha)_j(\beta)_j, \text{ if } (\beta^*)_j = 0; \qquad (49) \\ &\frac{1}{2}((\beta^*)_j - (\beta)_j)^2, \text{ if } (\beta^*)_j = -(\mathbf{H}\alpha)_j; \qquad (50)\end{aligned}$$

It is important to check that this solution ensures that $\mathbf{w} \geq \mathbf{0}$. In both cases, the component-wise update in β gives the objective a non-negative increase. However, the increase can be zero when $(\beta)_j = 0$ or $(\mathbf{H}\alpha)_j \leq 0$, or equivalently, when the primal constraint $\mathbf{w}_j \geq 0$ is not activated.

In summary, adding the non-negative constraints will not widen the duality gap but will actually decrease the gap, yet the amount of reduction is not guaranteed, as is the case with α.

The remainder of the reasoning is identical to [35]. The reasoning leads to the following theorem:

Theorem D.2. Convergence of Algorithm 1 For any training dataset $\mathcal{D}$ and any $C > 0$, $0 < \varepsilon \leq 4R^2C$, $\rho > 0$, Algorithm 1 terminates after at most

$$\left\lceil \log_2 \frac{\Delta(\rho)}{4R^2C} \right\rceil + \left\lceil \frac{16R^2C}{\varepsilon} \right\rceil \tag{51}$$

iterations.

We have enforced submodularity for the loss augmented inference, thus it can be computed optimally using the BK algorithm [18] with worst case complexity $O(n^2m|\mathcal{C}|)$ or the standard push-relabel based max-flow algorithm [**?**] with worst case complexity $O(n^2\sqrt{m})$ [1]. Here n and m denote the number of nodes and edges in the graph. $|\mathcal{C}|$ is the size of the minimal cut.

In each iteration of Algorithm 1, the loss augmented inference is called exactly n times, with n being the size of the dataset. Putting everything together, we have the proof for Theorem 5.2, i.e., polynomial time termination of Algorithm 1.

E Proof for Generalization to Higher Order Potentials for Enforcing Submodularity

Our algorithm can be generalized to higher order potentials using the reduction described in [32]. Let

$$S_1 = \sum_{i=1}^{d} y_i, \quad S_2 = \sum_{i=1}^{d-1} \sum_{j=i+1}^{d} y_i y_j = \frac{S_1(S_1 - 1)}{2} \tag{52}$$

The two ways of reduction are proposed based on the sign of the coefficient a:

if $a < 0$,

$$a y_1 ... y_d = \min_{z \in \{0,1\}} a z (S_1 - d + 1) \tag{53}$$

if $a > 0$,

$$a y_1 ... y_d = a \min_{z_1,...,z_{n_d} \in \{0,1\}} \sum_{i=1}^{n_d} z_i [c_{i,d}(-S_1 + 2i) - 1] + a S_2 \tag{54}$$

[1] Although the BK algorithm has a worse theoretical complexity, it was shown in [18] to be more efficient for computer vision problems in practice.

where n_d and $c_{i,d}$ are some positive constants.

In our case, $a = -\mathbf{w}_d \cdot \delta(u_1, ..., u_d)$. To enforce submodularity, we want all coefficients of the pairwise terms to be non-positive. It can be verified that if $a < 0$, this condition is satisfied. If $a < 0$, we have, after reduction, the term aS_2, which contains positive coefficients. Thus, we need to impose similar assumptions and restrictions that all high order features are non-negative and the learned higher order potential be non-negative. Applying this reduction, our algorithm is able to learn the parameters for high order potentials exactly in polynomial time.