Duality and the Continuous Graphical Model

Alexander Fix[1] and Sameer Agarwal[2]

[1] Cornell University
[2] Google Inc.

Abstract. Inspired by the Linear Programming based algorithms for discrete MRFs, we show how a corresponding infinite-dimensional dual for continuous-state MRFs can be approximated by a hierarchy of tractable relaxations. This hierarchy of dual programs includes as a special case the methods of Peng et al. [17] and Zach & Kohli [33]. We give approximation bounds for the tightness of our construction, study their relationship to discrete MRFs and give a generic optimization algorithm based on Nesterov's dual-smoothing method [16].

1 Introduction

Consider an optimization problem of the form

$$\min_{\mathbf{x}} \sum_{\alpha} f_\alpha(\mathbf{x}_\alpha). \tag{1}$$

Here, $\mathbf{x}$ is n-dimensional parameter vector. The index α varies over subsets of the variables of $\mathbf{x} = \{x_1, \ldots, x_n\}$, $\mathbf{x}_\alpha$ denotes the the corresponding sub-vector of $\mathbf{x}$, and $f_\alpha(\mathbf{x}_\alpha)$ is a function that only depends on $\mathbf{x}_\alpha$. For problems of interest to us $|\alpha| \ll n$, i.e., the number of parameters that f_α depends on is much smaller than the total number of parameters.

Problems of this form abound in computer vision and machine learning, including image denoising [3], bundle adjustment [26], and stereo matching [25], among many others. One particularly important case is the inference problem in Markov Random Fields(MRFs) [8]. MRFs are probability distributions that can be written in the form

$$p(\mathbf{x}) \propto \prod_{\alpha} e^{-f_\alpha(\mathbf{x}_\alpha)}. \tag{2}$$

Where, α are cliques in the underlying graph and f_α are the associated clique potentials. It is straightforward to see that the MAP inference problem for $p(\mathbf{x})$ is equivalent to solving (1).

In the case where the domain of the f_α (commonly known as the state-space of $\mathbf{x}$) is a finite discrete set, (2) is known as a discrete MRF and finding the optimal solution to (1) is NP-Hard [22]. Despite this, a variety of algorithms have been developed to efficiently compute an approximately optimal solution. These include graph cuts [4], belief propagation [30], and (most relevant to this

work) dual and primal-dual methods [12,31,21]. Most of these methods have been developed for the case of *pairwise discrete* MRFs, i.e., $|\alpha| \leq 2$.

There is however considerable interest in problems where the states are continuous and/or the cliques have size greater than 2, including pose tracking [23,24], structure from motion [27,6], stereo estimation [9,32], and protein folding [17]. Given the success of the methods used for solving discrete pairwise MRFs, it is natural that a number of attempts have been made to extend them to continuous domains and larger clique sizes [10,17,33]. But it is fair to say that their success is limited and the development of these methods is still in its infancy. Current methods for optimizing continuous MRFs include [17], [33] (which we will discuss at greater length below) as well as [1], which can only handle convex hinge-loss functions, but does allow constraints between the variables.

One of the most powerful tools for developing and analyzing discrete optimization algorithms (exact and approximate) is linear programming [14,28]. So it is no surprise that linear programming is at the heart of some of the most successful methods for solving MRFs including [12,31,21] (see [29] for a review). It is straightforward to construct a linear program relaxation of (1), as well as its dual, but both of these end up being abstract infinite dimensional problems that are not amenable to computation.

In this paper we offer a systematic procedure for approximating the infinite dimensional dual using a hierarchy of piecewise polynomial functions. Doing so allows us to handle both the issue of continuous domain as well as larger clique size in a principled manner. It also allows us to unify and generalize the works of Peng et al. [17] and Zach & Kohli [33]. As one would expect, the degree of the polynomial and the granularity of the piecewise construction affect the fidelity of the approximation. We analyze this and provide explicit approximation bounds. We also study the cases where the elements of the hierarchy coincide with a suitably constructed discrete optimization problem thereby enabling the use of existing optimization algorithms. Last but not the least, we propose a dual optimization algorithm applicable to a slice of our hierarchy based on Nesterov's dual-smoothing methods [16], which has recently been used successfully to solve discrete MRFs [11,21].

The rest of the paper is organized as follows. In section 2 we construct the linear programming relaxation to (1), its dual, and make some elementary observations about their structure. In section 3 we present a hierarchy of polynomial approximations to the dual, and in section 4 we generalize to piecewise polynomial approximations to the dual. Section 5 considers the special case of piecewise constant and linear f_α. Section 6 presents optimization methods for solving the dual hierarchy. We conclude with a discussion in Section 7.

1.1 Preliminaries

Without loss, we assume that unary terms exist for each $i \in \{1, \dots n\}$ [3]. Furthermore the vector $\mathbf{x}$ lives in some $\Omega = \Omega_1 \times \cdots \times \Omega_n$. We will assume that the domains Ω_i are 1-dimensional, i.e., $\Omega_i \subseteq \mathbb{R}$ and compact. Compactness ensures that the minimum value is attained. This allows us to re-write (1) as

$$\min_{\mathbf{x} \in \Omega} \sum_i f_i(x_i) + \sum_\alpha f_\alpha(\mathbf{x}_\alpha). \tag{F}$$

Where, $|\alpha| \geq 2$. Let $M = \sum_\alpha |\alpha|$ which is one convenient measure for the size of the input.

We will distinguish between two cases for the domain. If Ω is finite, we will refer to (F) as a discrete problem. If each Ω_i is an interval $[a, b]$ we will refer to (F) as a continuous problem. f_i and f_α will be assumed to be lower semi-continuous (l.s.c.) as functions $\Omega \to \mathbb{R}$. Note that if Ω is discrete then all functions $\Omega \to \mathbb{R}$ are continuous.

For any Ω, $\mathcal{P}[\Omega]$ is the space of all regular[4] probability distributions on Ω. $\langle f, \mu \rangle$ is the expectation of f with respect to the probability distribution μ. $C[\Omega]$ denotes the space of continuous functions $\Omega \to \mathbb{R}$, and $\mathrm{Lip}_L[\Omega] \subseteq C[\Omega]$ are the L-Lipschitz continuous functions, i.e., functions f with $|f(x) - f(y)| \leq L\|x - y\|_1$ for all $x, y \in \Omega$.

The Fenchel Conjugate of a function g is $g^*(\mathbf{y}) = \sup_{\mathbf{x}} \mathbf{y}^\top \mathbf{x} - g(\mathbf{x})$; it is always convex. The double Fenchel Conjugate g^{**} is the convex envelope of g.

For an optimization problem A, $\mathrm{OPT}(A)$ will denote the optimal objective function value.

2 The linear programming relaxation

If $\mu \in \mathcal{P}[\Omega]$, then for every subset α, let $\mu|_\alpha \in \mathcal{P}[\Omega_\alpha]$ be the marginal distribution of μ over the variables in α. Consider the optimization problem

$$\min_{\substack{\mu \in \mathcal{P}[\Omega] \\ \mu_\alpha \in \mathcal{P}[\Omega_\alpha]}} \sum_\alpha \langle f_\alpha, \mu_\alpha \rangle \,, \text{ s.t. } \mu_\alpha = \mu|_\alpha. \tag{3}$$

This is the *Full Marginal Polytope LP*. It has the same optimum as (1). It is however intractable. Even in the discrete case it involves exponentially many variables (it requires specifying a probability for each value of $\mathbf{x} \in \Omega$). To get

[3] This is a notational convenience. If an f_i is not part of the input, set $f_i = 0$, which does not change the optimization problem.

[4] A measure μ is regular if for each $A \subseteq \Omega$ we have $\mu(A) = \inf\{\mu(O) \mid O \supseteq A, O \text{ open}\} = \sup\{\mu(K) \mid K \subseteq A, K \text{ compact}\}$. Many nice distributions, in particular, delta distributions, are regular.

around this, a standard relaxation is [5]:

$$\min_{\substack{\mu_i \in \mathcal{P}[\Omega_i] \\ \mu_\alpha \in \mathcal{P}[\Omega_\alpha]}} \sum_i \langle f_i, \mu_i \rangle + \sum_\alpha \langle f_\alpha, \mu_\alpha \rangle \,, \text{ s.t. } \mu_\alpha|_i = \mu_i. \tag{P-F}$$

This is the *Local Marginal Polytope LP*. The constraints $\mu_\alpha|_i = \mu_i$ (indexed by α, i for all α and $i \in \alpha$.) are equivalent to saying that the signed measures[6] $\mu_\alpha|_i - \mu_i$ are identically 0. By dualizing these constraints, we get the following unconstrained optimization problem [17]:

$$\max_{\boldsymbol{\lambda}} \sum_i \min_{x_i \in \Omega_i} \left[f_i(x_i) - \sum_{\alpha \ni i} \lambda_{\alpha,i}(x_i) \right] + \sum_\alpha \min_{\mathbf{x}_\alpha \in \Omega_\alpha} \left[f_\alpha(\mathbf{x}_\alpha) + \sum_{i \in \alpha} \lambda_{\alpha,i}(x_i) \right] \tag{D-F}$$

From [19], we know that the dual variables $\lambda_{\alpha,i}$ are arbitrary continuous univariate functions defined on Ω_i, i.e., $\lambda_{\alpha,i} \in C[\Omega_i]$. Furthermore, we have strong-duality: OPT(P-F) = OPT(D-F). Here, we have not assumed anything about f_α. In fact, if the f_α satisfy certain smoothness properties, then the optimal dual variables do as well. The proof of the following lemma can be found in Appendix A.

Lemma 1. *If all $f_\alpha \in Lip_L[\Omega_\alpha]$, then there is a dual-optimal $\boldsymbol{\lambda}$ where each $\lambda_{\alpha,i} \in Lip_L[\Omega_i]$.*

Before we go further, let us define some notation. We will use $q(\boldsymbol{\lambda})$ to denote the value of the objective in (D-F) and $q_i(\boldsymbol{\lambda})$ and $q_\alpha(\boldsymbol{\lambda})$ to denote the individual terms of the summation. For convenience, we define $\lambda_i(x_i) = \sum_{\alpha \ni i} \lambda_{\alpha,i}(x_i)$ and $\lambda_\alpha(\mathbf{x}_\alpha) = \sum_{i \in \alpha} \lambda_{\alpha,i}(x_i)$. Given this notation we can write

$$q_i(\boldsymbol{\lambda}) = \min_{x_i} f_i(x_i) - \lambda_i(x_i) \tag{4}$$

$$q_\alpha(\boldsymbol{\lambda}) = \min_{\mathbf{x}_\alpha} f_\alpha(\mathbf{x}_\alpha) + \lambda_\alpha(\mathbf{x}_\alpha) \tag{5}$$

$$q(\boldsymbol{\lambda}) = \sum_i q_i(\boldsymbol{\lambda}) + \sum_\alpha q_\alpha(\boldsymbol{\lambda}). \tag{6}$$

Since the dual variables $\boldsymbol{\lambda}$ are allowed to be arbitrary continuous functions, the dual problem (D-F) is infinite dimensional and hence not computationally tractable. We will instead consider subspaces $\Lambda \subset C[\Omega]$ that lead to computationally tractable duals:

$$\max_{\boldsymbol{\lambda} \in \Lambda} \sum_i q_i(\boldsymbol{\lambda}) + \sum_\alpha q_\alpha(\boldsymbol{\lambda}). \tag{D-Λ}$$

[5] Sometimes, if more is known about the structure of f_α, then more specialized relaxations can be constructed. e.g. if f_α is a polynomial, then problem (3) is the same starting point as that of the Lasserre Heirarchy [13]. The Lasserre hierarchy exploits the fact that expectations of polynomials are linear in the moment variables of a distribution $y_d = \langle x^d, \mu \rangle$. Then, the inverse Moment Problem allows relaxing the original problem to a hierarchy of Semi-Definite Programs.

[6] $\mu_\alpha|_i - \mu_i$ are signed measures on Ω_i, since we may have $\mu_\alpha|_i > \mu_i$ for some events, and $\mu_\alpha|_i < \mu_i$ elsewhere.

Lemma 2. *Let $\Lambda \subset \Lambda' \subseteq C[\Omega]$, and let $\boldsymbol{\lambda}^*$ and $\boldsymbol{\lambda}'^*$ be the solution of the corresponding optimization problems (D-Λ) and (D-Λ'). Then*

$$q(\boldsymbol{\lambda}^*) \leq q(\boldsymbol{\lambda}'^*) \leq q(\boldsymbol{\lambda}^*) + 2M\epsilon \tag{7}$$

where, $M = \sum_\alpha |\alpha|$ and $\epsilon = \max_{\alpha,i} \sup_{x_i} |\lambda^*_{\alpha,i}(x_i) - \lambda'^*_{\alpha,i}(x_i)|$.

Proof. The left inequality holds because (D-Λ') optimizes over a larger set than (D-Λ). For the right inequality, let $\Delta_i = |\{\alpha \ni i\}|$ and $\Delta_\alpha = |\{i \in \alpha\}|$. Then

$$\min_{x_i} f_i(x_i) - \sum_\alpha \lambda^*_{\alpha,i}(x_i) \geq \min_{x_i} f_i(x_i) - \sum_\alpha \lambda'^*_{\alpha,i}(x_i) - \Delta_i\epsilon \tag{8}$$

$$\min_{\mathbf{x}_\alpha} f_\alpha(\mathbf{x}_\alpha) + \sum_i \lambda^*_{\alpha,i}(x_i) \geq \min_{\mathbf{x}_\alpha} f_\alpha(\mathbf{x}_\alpha) + \sum_i \lambda'^*_{\alpha,i}(x_i) - \Delta_\alpha\epsilon \tag{9}$$

hence $q(\boldsymbol{\lambda}^*) \geq \sum_i q_i(\boldsymbol{\lambda}'^*) + \sum_\alpha q_\alpha(\boldsymbol{\lambda}'^*) - (\sum_i \Delta_i + \sum_\alpha \Delta_\alpha)\epsilon = q(\boldsymbol{\lambda}'^*) - 2M\epsilon$.

3 Polynomial dual variables

We begin by considering polynomial dual variables. The subspace $\Lambda^{(d)} \subset C[\Omega]$ will denote dual variables that are polynomials of degree d i.e.,

$$\lambda_{\alpha,i}(x_i) = \lambda^{(0)}_{\alpha,i} + \lambda^{(1)}_{\alpha,i} x_i + \cdots + \lambda^{(d)}_{\alpha,i} x_i^d. \tag{10}$$

Since $\Lambda^{(d-1)} \subset \Lambda^{(d)}$, this forms a hierarchy. Let us look at some special cases.

3.1 Constant dual variables

The simplest subspace of dual variables is $\Lambda^{(0)}$, the space of constant functions, i.e., $\lambda_{\alpha,i}(x_i) = \lambda^{(0)}_{\alpha,i}$. Then we have

Lemma 3. *Let $\boldsymbol{\lambda}^{(0)} \in \Lambda^{(0)}$, then for all $\boldsymbol{\lambda} \in C[\Omega]$, $q(\boldsymbol{\lambda} + \boldsymbol{\lambda}^{(0)}) = q(\boldsymbol{\lambda})$.*

Proof. Observe that $q_i(\boldsymbol{\lambda} + \boldsymbol{\lambda}^{(0)}) = q_i(\boldsymbol{\lambda}) - \lambda^{(0)}_i$ and $q_\alpha(\boldsymbol{\lambda} + \boldsymbol{\lambda}^{(0)}) = q_\alpha(\boldsymbol{\lambda}) + \lambda^{(0)}_\alpha$ and $\sum_i \lambda^{(0)}_i = \sum_{\alpha,i} \lambda^{(0)}_{\alpha,i} = \sum_\alpha \lambda^{(0)}_\alpha$. Therefore,

$$q(\boldsymbol{\lambda} + \boldsymbol{\lambda}^{(0)}) = \sum_i q_i(\boldsymbol{\lambda}) + \sum_\alpha q_\alpha(\boldsymbol{\lambda}) - \sum_i \lambda^{(0)}_i + \sum_\alpha \lambda^{(0)}_\alpha = q(\boldsymbol{\lambda}) \tag{11}$$

In other words, we can ignore constants added to $\lambda_{\alpha,i}$. In addition, this allows us to simplify the optimization problem (D-$\Lambda^{(0)}$).

Corollary 1. $\mathrm{OPT}(\text{D-}\Lambda^{(0)}) = q(\mathbf{0}) = \sum_i \min_{x_i} f_i(x_i) + \sum_\alpha \min_{\mathbf{x}_\alpha} f_\alpha(\mathbf{x}_\alpha)$

Proof. $\mathrm{OPT}(\text{D-}\Lambda^{(0)}) = q(\mathbf{0})$ is a straightforward consequence of Lemma 3. Then

$$q(\mathbf{0}) = \sum_i q_i(\mathbf{0}) + \sum_\alpha q_\alpha(\mathbf{0}) \tag{12}$$

$$= \sum_i \min_{x_i} f_i(x_i) + \sum_\alpha \min_{\mathbf{x}_\alpha} f_\alpha(\mathbf{x}_\alpha) \tag{13}$$

Thus $\mathrm{OPT}(\text{D-}\Lambda^{(0)})$ is the obvious lower bound of f obtained by simply minimizing each term separately.

3.2 Affine dual variables

Next, we consider the space $\Lambda^{(1)}$ of affine dual variables $\lambda_{\alpha,i}(x_i) = \lambda^{(0)}_{\alpha,i} + \lambda^{(1)}_{\alpha,i} x_i$. From Lemma 3 we know that constant offsets cancel out, so we can assume that $\lambda^{(0)}_{\alpha,i} = 0$, i.e., optimizing over affine dual variables is the same as optimizing over linear dual variables. With $\boldsymbol{\lambda} \in \Lambda^{(1)}$ we have:

$$q_i(\boldsymbol{\lambda}) = \min_{x_i} f_i(x_i) - \lambda^{(1)}_i x_i = -f^*_i(\lambda^{(1)}_i) \tag{14}$$

$$q_\alpha(\boldsymbol{\lambda}) = \min_{\mathbf{x}_\alpha} f_\alpha(\mathbf{x}_\alpha) + \boldsymbol{\lambda}^{(1)T}_\alpha \mathbf{x}_\alpha = -f^*_\alpha(-\boldsymbol{\lambda}^{(1)}_\alpha) \tag{15}$$

Here $\boldsymbol{\lambda}^{(j)}_\alpha$ is the vector $(\lambda^{(j)}_{\alpha,i})_{i\in\alpha}$ and $\lambda^{(j)}_i$ is the sum $\sum_\alpha \lambda^{(j)}_{\alpha,i}$. Recall that f^* is the Fenchel conjugate of f. Combining these, D-$\Lambda^{(1)}$ can be simplified to

$$\max_{\boldsymbol{\lambda}} \sum_i -f^*_i\big(\lambda^{(1)}_i\big) + \sum_\alpha -f^*_\alpha(-\boldsymbol{\lambda}^{(1)}_\alpha) \tag{D-$\Lambda^{(1)}$}$$

The Fenchel Conjugate can be explicitly computed for certain analytically defined functions (such as truncated quadratics). In cases where an analytical solution is not possible, there are numerical algorithms that can compute the Fenchel conjugate of a sampled function in linear time (e.g., [15]).

More interestingly, we have

Theorem 1. *Let*

$$\min_{\mathbf{x}} \sum_i f^{**}_i(x_i) + \sum_\alpha f^{**}_\alpha(\mathbf{x}_\alpha) \tag{P-SC}$$

be the convex optimization problem obtained by separately convexifying each term of (F)*, then* OPT(P-SC) = OPT(D-$\Lambda^{(1)}$).

Proof. Introduce copies of the variables $\mathbf{y}_\alpha$ for each clique α in (P-SC) to get the equivalent optimization problem

$$\min_{\mathbf{x},\{\mathbf{y}_\alpha\}} \sum_i f^{**}_i(x_i) + \sum_\alpha f^{**}_\alpha(\mathbf{y}_\alpha),\ \text{s. t. } \mathbf{y}_{\alpha,i} = x_i \tag{16}$$

Dualizing the equality constraints with dual-multipliers $\lambda^{(1)}_{\alpha,i}$ we get (D-SC)

$$\max_{\boldsymbol{\lambda}^{(1)}} \min_{\mathbf{x},\{\mathbf{y}_\alpha\}} \sum_i \Big[f^{**}_i(x_i) - \lambda^{(1)}_i x_i\Big] + \sum_\alpha \Big[f^{**}_\alpha(\mathbf{y}_\alpha) + \boldsymbol{\lambda}^{(1)\top}_\alpha \mathbf{y}_\alpha\Big] \tag{D-SC}$$

$$= \max_{\boldsymbol{\lambda}^{(1)}} \sum_i \min_{x_i} \Big[f^{**}_i(x_i) - \lambda^{(1)}_i x_i\Big] + \sum_\alpha \min_{\mathbf{y}_\alpha} \Big[f^{**}_\alpha(\mathbf{y}_\alpha) + \boldsymbol{\lambda}^{(1)\top}_\alpha \mathbf{y}_\alpha\Big] \tag{17}$$

$$= \max_{\boldsymbol{\lambda}^{(1)}} \sum_i -f^{***}_i(\lambda^{(1)}_i) + \sum_\alpha -f^{***}_\alpha(-\boldsymbol{\lambda}^{(1)}_\alpha) \tag{18}$$

For any f it is the case that $f^{***} = f^*$, i.e., the convex envelope of a convex function is the function itself. Therefore (18) is the same optimization problem as OPT(D-$\Lambda^{(1)}$), and then by strong duality, OPT(P-SC) = OPT(D-$\Lambda^{(1)}$).

3.3 Degree d polynomial dual variables

Let us now consider the subspaces $\Lambda^{(d)}$ when $d > 1$. Recall that $\lambda_{\alpha,i}(x_i) = \lambda^{(1)}_{\alpha,i} x_i + \cdots + \lambda^{(d)}_{\alpha,i} x_i^d$. This gives subproblems of the form

$$q_i(\boldsymbol{\lambda}) = \min_{x_i} f_i(x_i) - \lambda^{(1)}_i x_i - \cdots - \lambda^{(d)}_i x_i^d \tag{19}$$

$$q_\alpha(\boldsymbol{\lambda}) = \min_{\mathbf{x}_\alpha} f_\alpha(\mathbf{x}_\alpha) + \sum_{i \in \alpha} (\lambda^{(1)}_{\alpha,i} x_i + \cdots + \lambda^{(d)}_{\alpha,i} x_i^d) \tag{20}$$

These subproblems look almost like a Fenchel conjugate, with the linear form $\lambda^{(1)}_{\alpha,i} x_i$ replaced with a polynomial. In fact, optimization problems of this form have been studied, under the name of Φ-conjugates [7].

For a function $\Phi : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$, the Φ-conjugate transforms functions $f : \mathbb{R}^n \to \mathbb{R}$ to their conjugate $f^\Phi : \mathbb{R}^m \to \mathbb{R}$, defined by

$$f^\Phi(\mathbf{y}) = \sup_{\mathbf{x}} \Phi(\mathbf{x}, \mathbf{y}) - f(\mathbf{x}) \tag{21}$$

If $\Phi(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \mathbf{y}$, then the Φ-conjugate is just the Fenchel Conjugate. In the case of polynomial dual variables, we can define Φ to be the polynomial evaulation map: $\Phi_{i,d}(x_i, y^{(1)}, \ldots, y^{(d)}) = y^{(1)} x_i + \cdots + y^{(d)} x_i^d$ and $\Phi_{\alpha,d}(\mathbf{x}_\alpha, \mathbf{y}^{(1)}_\alpha, \ldots, \mathbf{y}^{(d)}_\alpha) = \sum_{i \in \alpha} \Phi_{i,d}(x_i, y^{(1)}_i, \ldots, y^{(d)}_i)$. Then, the dual for degree d polynomials becomes

$$\max_{\boldsymbol{\lambda}} \sum_i \left[- f_i^{\Phi_{i,d}}(\lambda^{(1)}_i, \ldots, \lambda^{(d)}_i) \right] + \sum_\alpha \left[- f_\alpha^{\Phi_{\alpha,d}}(-\boldsymbol{\lambda}^{(1)}_\alpha, \ldots, -\boldsymbol{\lambda}^{(d)}_\alpha) \right] \tag{22}$$

In this case, the Φ-conjugate can be computed in terms of the Fenchel conjugate (this is a straightforward generalization of the full quadratic transform, Example 11.65 of [18]).

4 Piecewise defined dual variables

A separate hierarchy, orthogonal to the polynomial hierarchy above, is obtained by considering dual-variables defined piecewise on their domain. That is, each dual variable has some fixed number of pieces and each piece belongs to $\Lambda^{(d)}$ for some fixed degree d, e.g., piecewise constant or piecewise linear functions.

To simplify notation, we assume the domain of x_i is $\Omega_i = [0, K]$ for some integer K and consider dual variables $\lambda_{\alpha,i}$ which are piecewise defined on the subintervals $I_k = [k-1, k)$ for $k = 1, \ldots, K$. We will use superscript notation to denote the pieces of $\boldsymbol{\lambda}$, so that $\lambda_{\alpha,i}(x_i) = \lambda^k_{\alpha,i}(x_i)$ for $x_i \in I_k$. We'll define $\lambda^k_{\alpha,i} = 0$ outside I_k so that $\lambda_{\alpha,i} = \sum_k \lambda^k_{\alpha,i}$.

It will be convenient to correspondingly subdivide the domains of f_i, f_α, so let $f^k_i(x_i) = f_i(x_i)$ for $x_i \in I_k$ and 0 otherwise. For the higher-order functions, we subdivide the cube $[0, K]^{|\alpha|}$ into grid-cells, indexed by $\mathbf{k}_\alpha = (k_i)_{i \in \alpha}$. Then, the grid cells are $I_{\mathbf{k}_\alpha} = \prod_{i \in \alpha} I_{k_i}$, and the pieces of f_α are $f^{\mathbf{k}_\alpha}_\alpha$, where $f^{\mathbf{k}_\alpha}_\alpha(\mathbf{x}_\alpha) =$

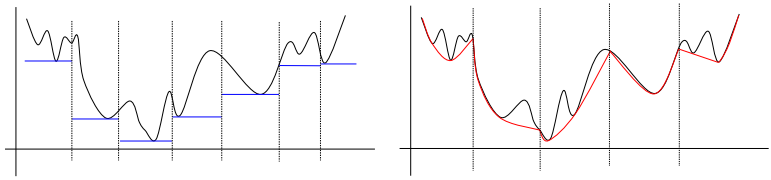

Fig. 1: (left) Piecewise constant dual variables find the minimum value in each "piece". (right) Piecewise linear dual variables convexify each piece separately. Note that the overall function is non-convex.

$f_\alpha(\mathbf{x}_\alpha)$ for $\mathbf{x}_\alpha \in I_{\mathbf{k}_\alpha}$ and 0 otherwise. Finally let us define subproblems for each piece:

$$q_i^k(\boldsymbol{\lambda}) = \min_{x_i \in I_k} f_i^k(x_i) - \lambda_i^k(x_i) \tag{23}$$

$$q_\alpha^{\mathbf{k}_\alpha}(\boldsymbol{\lambda}) = \min_{\mathbf{x}_\alpha \in I_{\mathbf{k}_\alpha}} f_\alpha^{\mathbf{k}_\alpha}(\mathbf{x}_\alpha) + \lambda_\alpha^{\mathbf{k}_\alpha}(x_i) \tag{24}$$

where we have extended our summation-shorthand with $\lambda_i^k = \sum_{\alpha \ni i} \lambda_{\alpha,i}^k$ and $\lambda_\alpha^{\mathbf{k}_\alpha} = \sum_{i \in \alpha} \lambda_{\alpha,i}^{k_i}$.

Lemma 4. *For piecewise defined dual-variables, the dual problem is given by*

$$q(\boldsymbol{\lambda}) = \sum_i \min_k q_i^k(\boldsymbol{\lambda}) + \sum_\alpha \min_{\mathbf{k}_\alpha} q_\alpha^{\mathbf{k}_\alpha}(\boldsymbol{\lambda}). \tag{25}$$

Proof. For piecewise dual variables, we know that $\lambda_{\alpha,i} = \sum_k \lambda_{\alpha,i}^k$, and that $\lambda_{\alpha,i}^k$ and f_i^k (resp. $f_\alpha^{\mathbf{k}_\alpha}$) are 0 for $x_i \notin I_k$ (resp. for $\mathbf{x}_\alpha \notin I_{\mathbf{k}_\alpha}$). Therefore, we have

$$q_i(\boldsymbol{\lambda}) = \min_{x_i} \sum_k [f_i^k(x_i) - \lambda_i^k(x_i)] \tag{26}$$

$$= \min_k \min_{x_i \in I_k} [f_i^k(x_i) - \lambda_i^k(x_i)] = \min_k q_i^k(\boldsymbol{\lambda}) \tag{27}$$

$$q_\alpha(\boldsymbol{\lambda}) = \min_{\mathbf{x}_\alpha} \sum_{\mathbf{k}_\alpha} [f_\alpha^{\mathbf{k}_\alpha}(\mathbf{x}_\alpha) + \lambda_\alpha^{k_\alpha}(\mathbf{x}_\alpha)] \tag{28}$$

$$= \min_{\mathbf{k}_\alpha} \min_{\mathbf{x}_\alpha \in I_{\mathbf{k}_\alpha}} [f_\alpha^{\mathbf{k}_\alpha}(\mathbf{x}_\alpha) + \lambda_\alpha^{k_\alpha}(\mathbf{x}_\alpha)] = \min_{\mathbf{k}_\alpha} q_\alpha^{\mathbf{k}_\alpha}(\boldsymbol{\lambda}) \tag{29}$$

So, if we know the dual subproblems q_i and q_α for a given class of functions, then adding piecewise defined functions just requires a finite minimum over the $K^{|\alpha|}$ subproblems for each piece.

We can combine this approach with the polynomial hierarchy of Sections 3.2 and 3.3 by letting (D-$\Lambda^{(d),K}$) be the dual program with piecewise polynomial dual variables — each piece is a degree d polynomial, and the pieces are K equally-sized intervals of the domain Ω_i. We can bound how close each approximation is to the true dual solution: as we increase the number of pieces of the domain, or the degree of polynomials allowed as dual variables, these bounds converge to 0, as characterized by the following theorem (proof in the Appendix).

Theorem 2. *If each variable has domain $\Omega_i = [0,1]$, and all f_α are L-Lipschitz, then* $OPT(\text{D-}\Lambda^{(d),K}) \geq OPT(D) - O(\frac{ML}{dK})$.

We highlight the duals programs for piecewise constant ($\Lambda^{(0),K}$) and piecewise linear dual variables ($\Lambda^{(1),K}$).

4.1 Piecewise constant dual variables

Substituting the appropriate subproblems $q_i^k, q_\alpha^{\mathbf{k}_\alpha}$ into Lemma 4, we get

$$\max_{\boldsymbol{\lambda}} \left[\sum_i \min_k \left[\min_{x_i} (f_i^k(x_i)) + \lambda_i^{(0),k} \right] + \sum_\alpha \min_{\mathbf{k}_\alpha} \left[\min_{\mathbf{x}_\alpha} (f_\alpha^{\mathbf{k}_\alpha}(\mathbf{x}_\alpha)) + \lambda_\alpha^{(0),\mathbf{k}_\alpha} \right] \right] \tag{D-$\Lambda^{(0),K}$}$$

Recall that the pieces of Ω_i and Ω_α are indexed by $k_i, \mathbf{k}_\alpha$. Define the following discrete functions

$$\overline{f}_i(k_i) = \min_{x_i \in I_k} f_i(x_i) \tag{30}$$

$$\overline{f}_\alpha(\mathbf{k}_\alpha) = \min_{\mathbf{x}_\alpha \in I_{\mathbf{k}_\alpha}} f_\alpha(\mathbf{x}_\alpha) \tag{31}$$

and the discrete optimization problem

$$\min_{\mathbf{k}} \sum_i \overline{f}_i(k_i) + \sum_\alpha \overline{f}(\mathbf{k}_\alpha). \tag{$\overline{\text{F}}$}$$

Note that $\overline{f} = \sum_i \overline{f}_i + \sum_\alpha \overline{f}_\alpha$ is obtained by taking the minimum value in each piece of the domain (see Figure 1 (left) for illustration). The primal LP relaxation of this problem using the Local Marginal Polytope is

$$\min_{\substack{\mu_i \in \mathcal{P}(K) \\ \mu_\alpha \in \mathcal{P}^\alpha(K)}} \sum_i \left\langle \overline{f}_i, \mu_i \right\rangle + \sum_\alpha \left\langle \overline{f}_\alpha, \mu_\alpha \right\rangle, \text{ s.t. } \mu_\alpha|_i = \mu_i. \tag{P-$\overline{\text{F}}$}$$

Where, $\mathcal{P}(K)$ is the space of discrete probability distributions on the integers $\{1, \ldots, K\}$ and $\mathcal{P}^\alpha(K)$ is the space of discrete probability distributions on corresponding $|\alpha|$-dimensional integer grid. Then we have,

Theorem 3. $\text{OPT}(\text{D-}\Lambda^{(0),K}) = \text{OPT}(\text{P-}\overline{\text{F}})$.

Proof. Substituting (30) and (31) into (D-$\Lambda^{(0),K}$), we get

$$\max_{\boldsymbol{\lambda}} \sum_i \min_{k_i} \left[\overline{f}_i(k_i) - \lambda_i^{(0),k} \right] + \sum_\alpha \min_{\mathbf{k}_\alpha} \left[\overline{f}_\alpha(\mathbf{k}_\alpha) + \lambda_\alpha^{(0),\mathbf{k}_\alpha} \right] \tag{32}$$

This is exactly the dual of (P-$\overline{\text{F}}$), and by strong duality our claim holds.

Note that this means we can solve D-$\Lambda^{(0),K}$ by using $\overline{f}$ as input to any Discrete MRF solver which optimizes the dual, such as [31,21].

4.2 Piecewise linear dual variables

Again, substituting the appropriate subproblems $q_i^k, q_\alpha^{\mathbf{k}_\alpha}$ into Lemma 4, we get

$$\max_{\boldsymbol{\lambda}} \left[\sum_i \min_k \left[-(f_i^k)^*(\lambda_i^{(1),k}) + \lambda_i^{(0),k} \right] + \sum_\alpha \min_{\mathbf{k}_\alpha} \left[-(f_\alpha^{\mathbf{k}_\alpha})^*(-\boldsymbol{\lambda}_\alpha^{(1),\mathbf{k}_\alpha}) + \lambda_\alpha^{(0),\mathbf{k}_\alpha} \right] \right] \tag{D-$\Lambda^{(1),K}$}$$

This problem turns out to be closely related to the method proposed by Zach and Kohli [33]. Their method first subdivides the functions f_i, f_α on a grid, and separately convexifies each piece (see Figure 1 (right) for illustration). The resulting problem is solved using a convex program with marginalization constraints. More specifically, their method solves the following convex program[7]:

$$\min_{\mathbf{x},\mathbf{y}} \sum_{i,k} y_i^k (f_i^k)^{**}\left(\frac{x_i^k}{y_i^k}\right) + \sum_{\alpha,\mathbf{k}_\alpha} y_\alpha^{\mathbf{k}_\alpha} (f_\alpha^{\mathbf{k}_\alpha})^{**}\left(\frac{\mathbf{x}_\alpha^{\mathbf{k}_\alpha}}{y_\alpha^{\mathbf{k}_\alpha}}\right) \tag{P-ZK}$$

$$\sum_{k_\alpha : k_i = k} y_\alpha^{\mathbf{k}_\alpha} = y_i^k \tag{33}$$

$$\sum_{\mathbf{k}_\alpha : k_i = k} x_{\alpha,i}^{\mathbf{k}_\alpha} = x_{i,k} \tag{34}$$

$$\langle y_i, 1\rangle = \langle y_\alpha, 1\rangle = 1 \tag{35}$$

$$0 \le x_i^k \le y_i^k \qquad 0 \le x_{\alpha,i}^{\mathbf{k}_\alpha} \le y_\alpha^{\mathbf{k}_\alpha} \tag{36}$$

Theorem 4. $\mathrm{OPT}(\text{D-}\Lambda^{(1),K}) = \mathrm{OPT}(\text{P-ZK})$.

Proof. We can dualize the constraints (33) and (34) with dual variables $\lambda_{\alpha,i}^{(1),k}$ and $\lambda_{\alpha,i}^{(0),k}$ respectively. This gives the program:

$$\max_{\boldsymbol{\lambda}} \min_{\mathbf{x},\mathbf{y}} \sum_{i,k} y_i^k \left[(f_i^k)^{**}\left(\frac{x_i^k}{y_i^k}\right) - \lambda_i^{(1),k} \frac{x_i^k}{y_i^k} - \lambda_i^{(0),k} \right] \tag{37a}$$

$$+ \sum_{\alpha,\mathbf{k}_\alpha} y_\alpha^{\mathbf{k}_\alpha} \left[(f_\alpha^{\mathbf{k}_\alpha})^{**}\left(\frac{\mathbf{x}_\alpha^{\mathbf{k}_\alpha}}{y_\alpha^{\mathbf{k}_\alpha}}\right) + (\boldsymbol{\lambda}_\alpha^{(1),\mathbf{k}_\alpha})^T \frac{\mathbf{x}_\alpha^{\mathbf{k}_\alpha}}{y_\alpha^{\mathbf{k}_\alpha}} + \lambda_\alpha^{(0),\mathbf{k}_\alpha} \right] \tag{37b}$$

$$\langle y_i, 1\rangle = \langle y_\alpha, 1\rangle = 1 \tag{37c}$$

$$0 \le x_i^k \le y_i^k \qquad 0 \le x_{\alpha,i}^{\mathbf{k}_\alpha} \le y_\alpha^{\mathbf{k}_\alpha} \tag{37d}$$

Then, minimizing over $\mathbf{x}$ and then $\mathbf{y}$ (and using the fact that $f^{***} = f^*$), this simplifies to (D-$\Lambda^{(1),K}$). Therefore, (D-$\Lambda^{(1),K}$) is dual to (P-ZK).

[7] (P-ZK) generalizes equation (8) from [33] to higher-order cliques (they consider only pairwise terms), with slight changes in notation to match our own. Additionally, throughout this section we'll define $0/0 = 0$, to make the divisions above always be well-defined.

5 Piecewise defined f_i and f_α

In the last two sections we studied versions of (D-F) where the dual variables are restricted to subspaces of $C[\Omega_i]$ without any restrictions on f_i and f_α. Let us now consider the case where the functions f_i and f_α are also defined piecewise.

Theorem 5. *If f_i, f_α are piecewise constant on $I_k, I_{\mathbf{k}_\alpha}$ respectively, and if $\overline{F}$ is the optimization problem obtained by* discretizing F *(via* (30) *and* (31)*), then*

$$\mathrm{OPT}(\text{P-F}) = \mathrm{OPT}(\text{P-}\overline{\text{F}}) = \mathrm{OPT}(\text{D-}\Lambda^{(0),K})$$

Proof. Since f_i, f_α are all piecewise constant on the grid cells $I_k, I_{\mathbf{k}_\alpha}$, then the construction (30) and (31) reduces to

$$\overline{f}_i(k_i) = \min_{x_i \in I_k} f_i(x_i) = f_i(I_k) \tag{38}$$

$$\overline{f}_\alpha(\mathbf{k}_\alpha) = \min_{\mathbf{x}_\alpha \in I_{\mathbf{k}_\alpha}} f_\alpha(\mathbf{x}_\alpha) = f(I_{\mathbf{k}_\alpha}) \tag{39}$$

Observe that the optimal dual variables $\boldsymbol{\lambda}^*$ for (D) will also be piecewise constant, here is why. Since f_i, f_α are constant on $I_k, I_{\mathbf{k}_\alpha}$, setting $\lambda_{\alpha,i}(x_i)$ to its average value on the interval, $\lambda_{\alpha,i}(x_i) = \frac{1}{|I_k|}\int_{I_k} \lambda_{\alpha,i}(x_i)dx$ for $x_i \in I_k$, does not decrease the objective q. Therefore, we have $\mathrm{OPT}(D) = \mathrm{OPT}(\text{D-}\Lambda^{(0),K})$, which combined with $\mathrm{OPT}(D) = \mathrm{OPT}(\text{P-F})$ and Lemma 3 gives our result.

For piecewise-linear functions, we have a similar result (proof in Appendix).

Theorem 6. *Let the functions f_i, f_α be continuous piecewise-linear on a regular grid*[8] *and let $\widetilde{f}_i$ be f_i restricted to $\{0,\dots,K\}$, and $\widetilde{f}_\alpha$ be f_α restricted to $\{0,\dots,K\}^\alpha$. Then consider the discrete optimization problem:*

$$\min_{\substack{\mu_i \in \mathcal{P}(K) \\ \mu_\alpha \in \mathcal{P}^\alpha(K)}} \sum_i \left\langle \widetilde{f}_i, \mu_i \right\rangle + \sum_\alpha \left\langle \widetilde{f}_\alpha, \mu_\alpha \right\rangle, \ \textit{s.t. } \mu_\alpha|_i = \mu_i \tag{P-$\widetilde{\text{F}}$}$$

$\mathrm{OPT}(\text{P-F}) = \mathrm{OPT}(\text{P-}\widetilde{\text{F}})$.

According to these theorems, for piecewise-constant and piecewise-linear objective functions, the infinite dimensional primal and dual problems have the same value as the finite dimensional problems for the discrete MRF $\widetilde{f}$. This means that the classic discrete MRF optimization methods can be used to solve this class of problems effectively.

[8] f_i is linear on each interval I_k, and that there's a triangulation T of the grid Ω_α such that f_α is linear on each triangle (or simplex) $\tau \in T$.

6 Solving D-$\Lambda^{(d),K}$

The problem D-$\Lambda^{(d),K}$ is a finite dimensional, unconstrained convex non-smooth optimization problem. For non-smooth problems, the best general-purpose optimization algorithms are subgradient methods [2]. However, because subgradient algorithms have slow convergence, requiring $O(\frac{1}{\epsilon^2})$ function evaluations to obtain an ϵ-optimal solution [2]) it is worth considering more specialized methods.

As we noted earlier, as consequence of Theorem 3, (D-$\Lambda^{(0),K}$) can be solved efficiently using discrete MRF solvers that operate on the dual.

Sometimes it is preferable to slightly modify the problem and instead optimize a smooth approximation to the dual and this can lead to a convergence rate of $O(1/\epsilon)$ [16]. For discrete problems, this approach has been used in the Adaptive Diminishing Smoothing method of [21] to obtain state of the art optimization results for discrete MRFs. This can also be applied to (D-$\Lambda^{(1),K}$). There are two sources of discontinuity in dual: the finite minimization from the piecewise part, where $q_i(\boldsymbol{\lambda}) = \min_k q_i^k(\boldsymbol{\lambda})$; and the Fenchel conjugate $q_i^k(\boldsymbol{\lambda}) = -(f_i^k)^*(\lambda_i^{(1),k})$, which may also be non-differentiable. To get a smooth approximation for the finite minimization, we replace min with soft-min[9] as in the work of [21]. For the Fenchel conjugate, we have

Lemma 5. *For $f : [0,1]^n \to \mathbb{R}$ and $t > 0$, let $f_t(\mathbf{x}) = f^{**} + t\|x\|^2$. Then f_t^* is differentiable, and $f^*(\mathbf{y}) \geq f_t^*(\mathbf{y}) \geq f^*(\mathbf{y}) - tn$.*

Proof. We note that (by Lemma 26.3 of [20]) f^* is differentiable if and only if f is strictly-convex, and f^{**} is convex and $t > 0$ so f_t is strictly convex. Then, $f_t \geq f$ for all $\mathbf{x}$, so $f_t^* \leq f^*$ (the Fenchel conjugate is order reversing). Finally,

$$f_t^*(\mathbf{y}) = \sup_{\mathbf{x}} \mathbf{y}^T\mathbf{x} - f^{**}(\mathbf{x}) - t\|x\|^2 \tag{40}$$

$$\geq \sup_{\mathbf{x}} \mathbf{y}^T\mathbf{x} - f^{**}(\mathbf{x}) - tn = f^*(\mathbf{y}) - tn. \tag{41}$$

Combining these two smoothing techniques gives us a differentiable approximation to (D-$\Lambda^{(1),K}$) which can then be efficiently optimized, either using conventional quasi-Newton methods such as L-BFGS, or the special-purpose optimal method of Nesterov [16].

For $d > 1$, the story is less nice. Recall that if there are multiple minimizers for $f(\mathbf{x}) = \min_{i \in I} g_i(\mathbf{x})$, then f is non-differentiable. In particular, for quadratic dual variables, the Φ-conjugate $f^{\Phi}(a_1, a_2) = \sup_x a_1 x + a_2 x^2 - f(x)$ may have multiple minimizers. Ensuring strict convexity does not help this situation: if $f(x) = x^2$, then $f^{\Phi}(0,1) = \sup_x 0 \cdot x + 1 \cdot x^2 - x^2 = \sup_x 0$, which is minimized by every x. So for $d > 1$, the smoothing method of Lemma 5 doesn't work. We do not know a practical way to smooth these subproblems, so we can only propose to use subgradient methods for optimization in this case.

[9] Defined by soft-$\min_{i \in I} g_i(\mathbf{x}) = -t \log \sum_i e^{-g_i(\mathbf{x})/t}$.

7 Discussion

We have given a sequence of dual programs (D-$\Lambda^{(d),K}$), which get increasingly close to the infinite-dimensional dual (D) as d, K increase. To see the tradeoffs in choosing d, K, the bounds from Theorem 2 are $O(\frac{1}{dK})$. If we consider either doubling the number of pieces K, or the degree d, then we get the same improvement in error bound, but both choices use twice as many coefficients $\lambda_{\alpha,i}^{(j),k}$.

However, the computation of the soft-min for the gradient of f_α scales as $K^{|\alpha|}$, whereas if our f_α are analytically defined, our Fenchel conjugate computation may be much cheaper to compute (potentially in constant time). Therefore, depending on the specifics of the problem, increasing the degree d is likely to be a better tradeoff in terms of computational efficiency. Unfortunately, (D-$\Lambda^{(d),K}$) cannot use smooth optimization methods for $d > 1$, which suggests that (D-$\Lambda^{(1),K}$) is the best choice.

A further attractive feature of the dual construction is that it unifies both higher order cliques α, and continuous domains in a single framework. The main complexity of the dual (D) is due to the continuous variables, causing it to be infinite dimensional. The higher-order cliques do cause the number of pieces $I_{\mathbf{k}_\alpha}$ in the piecewise dual variable case to grow exponentially with the clique size, but the same is true for discrete MRFs as well.

Going forward we plan on exploring the practical performance and specialized algorithms for computing the dual for specific f_α of interest in applications. In particular, we will consider the truncated L_1 and L_2 priors, because their Fenchel conjugates can be analytically derived and computed in constant time. We will also investigate the Fast Fenchel Conjugate [15] for handling more general f_α. Using these duals as building blocks, we then plan on building a practical implementation of smoothing based optimization algorithm for (D-$\Lambda^{(1),K}$).

A Proofs

Lemma 1. *If all $f_\alpha \in Lip_L[\Omega_\alpha]$, then there is a dual-optimal $\boldsymbol{\lambda}$ where each $\lambda_{\alpha,i} \in Lip_L[\Omega_i]$.*

Proof. Let $\boldsymbol{\lambda}$ be dual-optimal. We will iterate through α, i updating each $\lambda_{\alpha,i}$ to become L-Lipschitz, without reducing the objective $q(\boldsymbol{\lambda})$. So, let $\lambda'_{\alpha,i}(x_i) = \min_{x'_i} \lambda_{\alpha,i}(x'_i) + L|x_i - x'_i|$. First, note that $\lambda'_{\alpha,i} \leq \lambda_{\alpha,i}$. We also have that $\lambda'_{\alpha,i}$ is L-Lipschitz: for any x there is some z with $\lambda'_{\alpha,i}(x) = \lambda_{\alpha,i}(z) + L|x - z|$ and for all y, $\lambda'_{\alpha,i}(y) \leq \lambda'_{\alpha,i}(z) + L|y - z|$ hence $\lambda'_{\alpha,i}(y) - \lambda'_{\alpha,i}(x) \leq L|y - z| - L|x - z| \leq L|y - x|$. By symmetry, $|\lambda'_{\alpha,i}(y) - \lambda'_{\alpha,i}(x)| \leq L|y - x|$.

Let $\boldsymbol{\lambda}'$ be $\boldsymbol{\lambda}$ where we've updated one $\lambda_{\alpha,i}$ to $\lambda'_{\alpha,i}$. Since $\lambda'_{\alpha,i} \leq \lambda_{\alpha,i}$ we know $q_i(\boldsymbol{\lambda}') \geq q_i(\boldsymbol{\lambda})$. To show $q_\alpha(\boldsymbol{\lambda}') \geq q_\alpha(\boldsymbol{\lambda})$, pick any $\mathbf{x}_\alpha$. There is some x'_i such that $\lambda'_{\alpha,i}(x_i) = \lambda_{\alpha,i}(x'_i) + L|x'_i - x_i|$. Let $\mathbf{x}'_\alpha$ be $\mathbf{x}_\alpha$ with x'_i replacing x_i. Then:

$$f_\alpha(\mathbf{x}_\alpha) + \sum_i \lambda'_{\alpha,i}(x_i) = f_\alpha(\mathbf{x}_\alpha) + \sum_j \lambda_{\alpha,j}(x'_j) + L|x_i - x'_i| \tag{42}$$

$$\geq f_\alpha(\mathbf{x}'_\alpha) + \sum_i \lambda_{\alpha,i}(x'_i) \geq q_\alpha(\boldsymbol{\lambda}) \tag{43}$$

Therefore, $q_\alpha(\boldsymbol{\lambda}') = \min_{\mathbf{x}_\alpha} f_\alpha(\mathbf{x}_\alpha) + \lambda'_\alpha(\mathbf{x}_\alpha) \geq q_\alpha(\boldsymbol{\lambda})$, so $q(\boldsymbol{\lambda}') \geq q(\boldsymbol{\lambda})$.

Theorem 2. *If each variable has domain $\Omega_i = [0,1]$, and all f_α are L-Lipschitz, then $OPT(\text{D-}\mathcal{M}^{(d),K}) \geq OPT(D) - O\left(\frac{ML}{dK}\right)$.*

Proof. We use Jackson's theorem (Corollary 7.5 of [5]): there is a constant C such that if $f \in \text{Lip}_L[0,1]$ then there is a polynomial p_n of degree n with $\|f - p_n\|_\infty \leq O\left(\frac{L}{n}\right)$. This bound may not be tight for small n; however for constant and linear functions, we have $p_0(x) = f(\frac{1}{2})$ and $p_1(x) = f(\frac{1}{4}) + (f(\frac{3}{4}) - f(\frac{1}{4}))x$ with $\|f - p_0\| \leq \frac{L}{2}$ and $\|f - p_1\| \leq \frac{L}{4}$ (the functions $f_0(x) = Lx$ and $f_1(x) = L|x - \frac{1}{2}|$ show that these bounds are tight).

Since the f_α are L-Lipschitz, by lemma 1 there is a dual-optimal $\boldsymbol{\lambda}$ where each $\lambda_{\alpha,i}$ is L-Lipschitz. Then, apply the Jackson-inequality to each piece $[\frac{i}{K}, \frac{i+1}{K}]$ of the domain $[0,1]$ to get a K-piecewise d-degree $\overline{\boldsymbol{\lambda}}$ with $\|\overline{\lambda}_{\alpha,i} - \lambda^*_{\alpha,i}\|_\infty \leq O\left(\frac{L}{dK}\right)$. Consequently, by lemma 2, $\text{OPT}(\text{D-}\mathcal{M}^{(d),K}) \geq q(\overline{\boldsymbol{\lambda}}) \geq \text{OPT}^* - O\left(\frac{ML}{dK}\right)$.

Theorem 6. *If f_i, f_α are continuous piecewise-linear functions on a regular grid, then* $\text{OPT}(\text{P-F}) = \text{OPT}(\text{P-}\widetilde{\text{F}})$.

Proof. Since $\widetilde{f}$ is just a sampled version of f, the discrete LP (P-$\widetilde{F}$) is identical to (P-F) with the restriction that $\widetilde{\mu}_\alpha \in \mathcal{P}^\alpha(K)$. Since $\mathcal{P}^\alpha(K) \subseteq \mathcal{P}[\Omega_\alpha]$ it's clear that the continuous LP is a lower bound on the discrete LP.

For the other direction, take a feasible primal $\mu_\alpha \in \mathcal{P}[\Omega_\alpha]$: we'll construct a feasible $\widetilde{\mu}_\alpha \in \mathcal{P}^\alpha(K)$ with the same objective.

Let T_α be the standard triangulation of the grid $\{0, \ldots, K\}^\alpha$. Each simplex $\tau \in T_\alpha$ has vertices in $\{0, \ldots, K\}^\alpha$ and the projection of τ onto the i-th component is an interval $[j, j+1]$. For each $\widetilde{\mathbf{x}} \in \{0, \ldots, K\}^\alpha$, there is a set of simplices with $\widetilde{\mathbf{x}}$ as a vertex, we will denote this set as $\tau \sim \widetilde{\mathbf{x}}$. Each simplex comes with barycentric coordinates: every point $\mathbf{x} \in \tau$ is a convex combination of the vertices. We'll write these as $\xi_{\tau,\widetilde{\mathbf{x}}}(\mathbf{x})$ which satisfy $\sum_{\widetilde{\mathbf{x}} \sim \tau} \xi_{\tau,\widetilde{\mathbf{x}}}(\mathbf{x})\widetilde{\mathbf{x}} = \mathbf{x}$.

We construct $\widetilde{\mu}_\alpha$ by taking all the mass from μ_α on a simplex τ, and gathering it to each vertex $\widetilde{\mathbf{x}}$, weighted by the barycentric coordinates $\xi_{\tau,\widetilde{\mathbf{x}}}$. More specifically, define $\widetilde{\mu}_\alpha(\widetilde{\mathbf{x}}) := \sum_{\tau \sim \widetilde{\mathbf{x}}} \int_\tau \xi_{\tau,\widetilde{\mathbf{x}}}(\mathbf{x}) d\mu_\alpha$.

The fact that the barycentric coordinates sum to 1 ensures that $\widetilde{\mu}_\alpha$ is a probability distribution on $\widetilde{\Omega_\alpha}$, and since the projections of τ onto the i-th component are intervals $[j, j+1]$ we get that $\{\widetilde{\mu}_\alpha\}$ satisfy the marginalization constraints. Finally, since our objective is linear on each τ, we have $f_\alpha(\mathbf{x}_\alpha) = \sum_{\widetilde{\mathbf{x}} \sim \tau} \xi_{\tau,\widetilde{\mathbf{x}}}(\mathbf{x}_\alpha) f(\widetilde{\mathbf{x}})$. Therefore, we have

$$\langle f_\alpha, \widetilde{\mu_\alpha} \rangle = \sum_{\widetilde{\mathbf{x}}} f_\alpha(\widetilde{\mathbf{x}}) \Big(\sum_{\tau \sim \widetilde{\mathbf{x}}} \int_\tau \xi_{\tau,\widetilde{\mathbf{x}}}(\mathbf{x}_\alpha) d\mu_\alpha \Big) = \sum_\tau \int_\tau \sum_{\widetilde{\mathbf{x}} \sim \tau} f(\widetilde{\mathbf{x}}) \xi_{\tau,\widetilde{\mathbf{x}}}(\mathbf{x}_\alpha) d\mu_\alpha \quad (44)$$

$$= \sum_\tau \int_\tau f(\mathbf{x}_\alpha) d\mu_\alpha = \langle f_\alpha, \mu_\alpha \rangle . \quad (45)$$

References

1. Bach, S.H., Broecheler, M., Getoor, L., O'Leary, D.P.: Scaling MPE inference for constrained continuous markov random fields with consensus optimization. In: Advances in Neural Information Processing Systems. pp. 2663–2671 (2012)
2. Bertsekas, D.: Nonlinear Programming. Athena Scientific (1995)
3. Besag, J., Besag, J.: On the statistical analysis of dirty pictures. Journal of the Royal Statistical Society B pp. 48–259 (1986)
4. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(11), 1222–1239 (2001)
5. Carothers, N.L.: A short course on approximation theory (2009)
6. Crandall, D.J., Owens, A., Snavely, N., Huttenlocher, D.P.: SfM with MRFs: Discrete-continuous optimization for large-scale structure from motion. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(12), 2841–2853 (2013)
7. Dolecki, S., Kurcyusz, S.: On ϕ-convexity in extremal problems. SIAM Journal on Control and Optimization 16(2), 277–300 (1978)
8. Geman, S., Geman, D.: Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence 6(6), 721–741 (Nov 1984)
9. Ihler, A., McAllester, D.: Particle belief propagation. In: Artificial Intelligence and Statistics. pp. 256–263 (2009)
10. Ishikawa, H.: Higher-order gradient descent by fusion-move graph cut. In: IEEE International Conference on Computer Vision. pp. 568–574 (2009)
11. Jojic, V., Gould, S., Koller, D.: Fast and smooth: Accelerated dual decomposition for MAP inference. In: International Conference on Machine Learning (2010)
12. Komodakis, N., Tziritas, G.: Approximate labeling via graph cuts based on linear programming. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(8), 1436–1453 (2007)
13. Lasserre, J.B.: Global optimization with polynomials and the problem of moments. SIAM Journal on Optimization 11, 796–817 (2001)
14. Lee, J.: A first course in combinatorial optimization, vol. 36. Cambridge University Press (2004)
15. Lucet, Y.: Faster than the fast Legendre transform, the linear-time Legendre transform. Numerical Algorithms 16(2), 171–185 (1997)
16. Nesterov, Y.: Smooth minimization of non-smooth functions. Mathematical Programming 103(1), 127–152 (2005)
17. Peng, J., Hazan, T., Mcallester, D., Urtasun, R.: Convex max-product algorithms for continuous MRFs with applications to protein folding. In: International Conference on Machine Learning (2011)
18. Rockafellar, R.T., Wets, R.J.B., Wets, M.: Variational analysis, vol. 317. Springer (1998)
19. Rockafellar, R.: Conjugate Duality and Optimization. Society for Industrial and Applied Mathematics (1974)
20. Rockafellar, R.: Convex Analysis. Convex Analysis, Princeton University Press (1997)
21. Savchynskyy, B., Schmidt, S., Schnrr, C.: Efficient MRF energy minimization via adaptive diminishing smoothing. In: Uncertanity in Artificial Intelligence (2012)

22. Shimony, S.E.: Finding MAPs for belief networks is NP-hard. Artificial Intelligence 68(2), 399–410 (1994)
23. Sigal, L., Bhatia, S., Roth, S., Black, M., Isard, M.: Tracking loose-limbed people. In: IEEE Conference on Computer Vision and Pattern Recognition (2004)
24. Sudderth, E.B., M, M.I., Freeman, W.T., Willsky, A.S.: Distributed occlusion reasoning for tracking with nonparametric belief propagation. In: Advances in Neural Information Processing Systems. pp. 1369–1376 (2004)
25. Sun, J., Zheng, N.N., Shum, H.Y.: Stereo matching using belief propagation. IEEE Transactions on Pattern Analysis and Machine Intelligence 25(7), 787–800 (July 2003)
26. Triggs, B., McLauchlan, P., Hartley, R., Fitzgibbon, A.: Bundle adjustment a modern synthesis. In: Vision Algorithms: Theory and Practice, pp. 298–372 (2000)
27. Trinh, H., McAllester, D.: Particle-based belief propagation for structure from motion and dense stereo vision with unknown camera constraints. In: Robot Vision, vol. 4931, pp. 16–28 (2008)
28. Vazirani, V.V.: Approximation algorithms. Springer (2001)
29. Wainwright, M.J., Jordan, M.I.: Graphical models, exponential families, and variational inference. Found. Trends Mach. Learn. 1(1-2), 1–305 (Jan 2008)
30. Weiss, Y., Freeman, W.T.: On the optimality of solutions of the max-product belief-propagation algorithm in arbitrary graphs. IEEE Trans. Inf. Theor. 47(2), 736–744 (Sep 2006)
31. Werner, T.: A linear programming approach to max-sum problem: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(7), 1165–1179 (July 2007)
32. Yamaguchi, K., Hazan, T., McAllester, D., Urtasun, R.: Continuous Markov random fields for robust stereo estimation. In: European Conference on Computer Vision, pp. 45–58 (2012)
33. Zach, C., Kohli, P.: A convex discrete-continuous approach for Markov random fields. In: ECCV 2012, pp. 386–399 (2012)