Evaluation of a First-Order Primal-Dual Algorithm for MRF Energy Minimization

Stefan Schmidt, Bogdan Savchynskyy, Jörg H. Kappes, and Christoph Schnörr

Heidelberg University, IWR / HCI
Speyerer Str. 6, 69115 Heidelberg, Germany
{schmidt,kappes,schnoerr}@math.uni-heidelberg.de
bogdan.savchynskyy@iwr.uni-heidelberg.de
http://ipa.iwr.uni-heidelberg.de
http://hci.iwr.uni-heidelberg.de

Abstract. We investigate the First-Order Primal-Dual (FPD) algorithm of Chambolle and Pock [1] in connection with MAP inference for general discrete graphical models. We provide a tight analytical upper bound of the stepsize parameter as a function of the underlying graphical structure (number of states, graph connectivity) and thus insight into the dependency of the convergence rate on the problem structure. Furthermore, we provide a method to compute efficiently primal and dual feasible solutions as part of the FPD iteration, which allows to obtain a sound termination criterion based on the primal-dual gap. An experimental comparison with Nesterov's first-order method in connection with dual decomposition shows superiority of the latter one in optimizing the dual problem. However due to the direct optimization of the primal bound, for small-sized (e.g. 20x20 grid graphs) problems with a large number of states, FPD iterations lead to faster improvement of the primal bound and a resulting faster overall convergence.

Keywords: graphical model, MAP inference, LP relaxation, image labeling, sparse convex programming.

1 Introduction

1.1 Overview

Our goal is to compute maximum-a-posteriori (MAP) solutions for discrete Markov random fields (MRF), specified by a hypergraph $\mathcal{G} = (\mathcal{V}, \mathcal{F})$, where the set of hyperedges $\mathcal{F}$ is a subset of the power-set $2^{\mathcal{V}}$ of $\mathcal{V}$, which we will call the *set of factors*.[1] The states of random variables of the MRF belong to finite sets $\mathcal{X}_v,\ v \in \mathcal{V}$. The notation $\mathcal{X}_a,\ a \in \mathcal{F}$ is used the Cartesian product $\otimes_{v \in a} \mathcal{X}_v$ of state sets for variables belonging to the factor $a \in \mathcal{F}$. The associated

[1] Note that we represent factors directly as hyperedges here. In the representation of [2] as a bipartite graph $\tilde{\mathcal{G}} = (\mathcal{V}, \mathcal{F}; \mathcal{E})$, this implies $\mathcal{E} = \{(v, a) \in \mathcal{V} \times \mathcal{F} \mid v \in a\}$.

distribution is $p_{\mathcal{G}}(x;\theta) \propto \exp(-E_{\mathcal{G}}(x;\theta))$, with the energy function

$$E_{\mathcal{G}}(x;\theta) = \sum_{a\in\mathcal{F}} \theta_a(x_a)\,, \tag{1}$$

where θ denotes a collection of potential functions $\theta_a : \mathcal{X}_a \to \mathbb{R}$ associated with each factor $a \in \mathcal{F}$.

1.2 Related Work and Motivation

Computing a MAP solution is equivalent to minimization of energy (1) and is known to be NP-complete in general. Thus we will concentrate on its linear programming (LP) relaxation over the local polytope [3]. The special case when the problem contains only first and second order factors ($\forall a \in \mathcal{F}, |a| \leq 2$), further on referred to as *pairwise* or *second-order* case. Its LP relaxation was originally studied in [4] (see also a modern overview [5]). There is a number of algorithms for solving this LP relaxation. The first group of such algorithms contains DAG and diffusion algorithms by Schlesinger (cf. [5]) and the closely related TRW-S algorithm by Kolmogorov [6]. These algorithms decrease the value of the dual LP monotonically but may not attain its optima in general, since they can be interpreted as (block-)coordinate descent and thus can get stuck, due to the non-smoothness of the dual objective. Indeed, TRW-S is considered as one of the fastest approximate solvers for the problem [7].

As alternatives, sub-gradient schemes for maximizing the dual objective were proposed in [8] and [9]. They are theoretically guaranteed to reach the optimum. However, these schemes are rather slow not only in theory, but also in practice. A recent paper [10] proposed to combine a dual decomposition [11] and Nesterov's first-order optimization scheme [12], which can be considered as a compromise between speed of TRW-S and guarantee of convergence.

Another recent paper [1] proposes a first-order primal-dual iteration scheme and a range of successful applications to variational optimization problems in image processing. Since this method is suited for large-scale non-smooth convex problems, and since MRF based image labeling covers a broad range of applications in computer vision, a competitive assessment of the method is of particular interest.

1.3 Contribution

Our contribution is three-fold:

- We propose a way of applying the first order primal-dual method [1] to the LP relaxation of the general (not obligatory pairwise) MRF energy minimization problem (1) and numerically compare it to Nesterov's optimization scheme [10] and TRW-S [6].
- For the pairwise case we provide a tight bound showing how the step size parameter of the method depends on the model structure and on the number of variable states.

– We generalize a method for computing an approximate primal solution [10] to models of arbitrary order (it was proposed originally only for the pairwise case) and propose a similar approach for constructing an approximate solution of the dual problem. These two approximations result in a sound stopping criterion based on the duality gap.

2 Methods

2.1 LP Relaxation of the MAP problem

We introduce the notation $\mathcal{F}^1 = \{a \in \mathcal{F} \colon |a| = 1\}$ for the set of all unary factors. Without loss of generality, we suppose that the model includes a unary factor for each variable, i.e. $\{\{v\} : v \in \mathcal{V}\} \subseteq \mathcal{F}$, and all non-first order potentials are absorbed into those of the highest order, i.e. if $b \in \mathcal{F}\backslash\mathcal{F}^1$, then $\forall a \in \mathcal{F}$, from $a \subset b$ follows $|a| = 1$.

We start by representing the energy (1) in overcomplete form [3] as

$$E_{\mathcal{G}}(x;\theta) = \sum_{i \in \mathcal{I}(\mathcal{G})} \theta_i \cdot \phi_i(x) = \langle \theta, \phi(x) \rangle \,, \tag{2}$$

where $x \in \otimes_{v\in\mathcal{V}} \mathcal{X}_v$ is *a model configuration* and the potentials $\theta = (\theta_a(x_a), a \in \mathcal{F}, x_a \in \mathcal{X}_a)$ as well as indicator vectors $\phi(x) \in \{0,1\}^{\mathcal{I}(\mathcal{G})}$ are indexed by $\mathcal{I}(\mathcal{G}) = \{(a;x_a) | a \in \mathcal{F},\ x_a \in \mathcal{X}_a\}$. The notation $\langle \cdot, \cdot \rangle$ is used for the standard scalar product.

Relaxing the binary vector ϕ to a vector $\mu = (\mu_a(x_a),\ a \in \mathcal{F},\ x_a \in \mathcal{X}_a)$ with components from the interval $[0;1]$ and imposing consistency constraints between the components leads to the well-known linear programming relaxation of the problem of minimizing (2):

$$\min_{\mu} \quad \langle \theta, \mu \rangle \quad \text{s.t.} \quad L\mu = c, \quad \mu \geq 0 \,. \tag{3}$$

Here L is the matrix of a linear operator and c a vector of corresponding dimension, which we will define next.

Let $\mathbb{1}_a = (\underbrace{1, \ldots, 1}_{|\mathcal{X}_a|})^\top$ denote an $|\mathcal{X}_a|$-dimensional vector of ones. A specific feature of our problem (3) is that the constraint matrix L has a block form. Namely, for $\mu_a \in \mathbb{R}^{\mathcal{X}_a}$, $\mu_b \in \mathbb{R}^{\mathcal{X}_b}$ the problem (3) can be written as

$$\min_{\mu} \quad \langle \theta, \mu \rangle \tag{4}$$

$$\text{s.t.} \quad L_{ab}\mu_b = \mu_a,\ b \in \mathcal{F}\backslash\mathcal{F}^1,\ a \subset b \,, \tag{5}$$

$$\mathbb{1}_a^\top \mu_a = 1, \quad a \in \mathcal{F}^1 \,, \tag{6}$$

$$\mu \geq 0 \,. \tag{7}$$

For the second-order case, problem (4)-(7) reads:

$$\min_{\mu} \quad \langle \theta, \mu \rangle \tag{8}$$

$$\text{s.t.} \quad \sum\nolimits_{x_{a'} \in \mathcal{X}_{a'}} \mu_b(x_a, x_{a'}) = \mu_a(x_a)\ b \in \mathcal{F} \backslash \mathcal{F}^1,\ \forall a \in b, a' \in b \setminus \{a\}\,, \tag{9}$$

$$\sum\nolimits_{x_a \in \mathcal{X}_a} \mu_a(x_a) = 1, \qquad a \in \mathcal{F}^1\,, \tag{10}$$

$$\mu \geq 0\,. \tag{11}$$

The dual to (3),

$$\max_{\nu} \quad \langle c, \nu \rangle \quad \text{s.t.} \quad L^\top \nu \leq \theta\,, \tag{12}$$

plays a significant role in many optimization schemes. We will analyze its structure for the general (non-pairwise) case in Section 2.3.

We cast the pair (3), (12) of optimization problems into a saddle point form via their Lagrangian,

$$\max_{\mu \geq 0} \min_{\nu} \quad \{\langle -c, \nu \rangle + \langle \mu, L^\top \nu \rangle - \langle \theta, \mu \rangle\}\,, \tag{13}$$

which is of the general form to apply the first order primal-dual iteration scheme – Algorithm 1 in [1]. This algorithm will be further on referred to as FPD.

2.2 Primal-Dual Iteration Scheme

Starting from any $\mu^{(1)} \geq 0$, $\nu^{(1)}$, $\zeta^{(1)}$, the FPD algorithm iterates for $t = 2, 3, \ldots$ and step-size $\tau \geq 0$ the updates:

$$\begin{aligned} \mu^{t+1} &\leftarrow \mathit{\Pi}_{\mathbb{R}_+} \left(\mu^t + \tau \left(L^\top \zeta^t - \theta\right)\right) \\ \nu^{t+1} &\leftarrow \nu^t - \tau \left(L \mu^{t+1} - c\right) \\ \zeta^{t+1} &\leftarrow 2\nu^{t+1} - \nu^t\,, \end{aligned} \tag{14}$$

where $\mathit{\Pi}_{\mathbb{R}_+}$ denotes the projection onto the positive orthant $\mathbb{R}_+$.[2]

As shown in [1, Th. 1], the algorithm achieves a $O(1/t)$ convergence rate, where t is the number of iterations.

Note that this algorithm requires only two sparse matrix multiplications (with L and $L^\top$), and a projection $\mathit{\Pi}_{\mathbb{R}_+}$, both of which are simple to implement and easily parallelizable.

Computing the maximal step length τ requires the estimation of the spectral norm of the matrix L. We have the sufficient convergence condition [1, Th. 1]

$$\tau \leq \lambda_{\max}^{-1/2}(L L^\top)\,, \tag{15}$$

where $\lambda_{\max}$ gives the largest eigenvalue of its argument, which depends on the graph structure only, and may be computed *a priori* using power iterations [13].

However, $\lambda_{\max}$ can be estimated also analytically, as is stated by the following theorem.

[2] We consider the case where the step-sizes for primal and dual iterations are set to the same value τ.

Theorem 1 *Let $\mathcal{G}$ be a second-order factor graph and all its variables have an equal number K of possible states. Then for d_{max} denoting the maximal degree (number of adjacent pairwise factors) of any node of $\mathcal{G}$,*

$$\lambda_{max} \leq \frac{1}{2}\left(3K + d_{max} + \sqrt{K^2 + 6d_{max}K + d_{max}^2}\right) . \tag{16}$$

We prove this theorem in the appendix.

Remark 1 *Note that the bound (16) does not depend on the graph size for grids. This bound is exact for regular graphs, i.e. having all nodes of equal degree. A typical example are fully connected graphs ($d_{max} = |\mathcal{V}| - 1$) and infinite grids ($d_{max} = 4$). For finite grids numerical computations show that this bound is quite sharp: for a particular 100×100 grid graph with 5 states the value $\lambda_{max} = 15.8418$ was computed numerically using power iterations [13] and the value 15.8443 is given by (16).*

Remark 2 *We also considered a variant of* (13) *that explicitly enforces the constraint for the unary primal variables μ_a, $a \in \mathcal{F}^1$ to lie in the unit simplex $\Delta(\mathcal{X}_a)$ defined by constraints (6) and (7):*

$$\max_{\substack{\mu \geq 0 \\ a \in \mathcal{F}^1 : \mu_a \in \Delta(\mathcal{X}_a)}} \quad \min_{\nu} \quad \left\{\langle -c, \nu\rangle + \langle \mu, \tilde{L}^\top \nu\rangle - \langle \theta, \mu\rangle\right\}, \tag{17}$$

where $\tilde{L}$ is the matrix, obtained by removing constraints (6) from L. The necessary simplex projections may be computed using e.g. [14], however this requires an inner loop within the first FPD step. It can be shown that for the pairwise case

$$\|\tilde{L}\tilde{L}^\top\| \leq d_{max} + 2K \tag{18}$$

holds, thus the incurred extra computational cost does not necessarily outweigh the larger maximal step-size allowed due to the reduced $\tilde{L}$, which increased only marginally for our problems (compare (18) to (16)).

2.3 Estimating Primal and Dual Bounds

The primal μ^t and dual ν^t iterates are not necessarily feasible in the respective primal (3) and dual (12) problems during the course of the algorithm, therefore obtaining primal and dual bounds to base a stopping criterion on the duality gap is not trivial.

We devise a method for computing sequences of primal and dual feasible points such that primal and dual bounds and the duality gap, respectively, can be estimated as a part of the overall iteration (14). Our method relies on strong duality of the primal (3) and dual (12) pair and (13) which ensures a vanishing duality gap after convergence.

A method to compute feasible points in the local polytope and hence an upper bound for the energy of the relaxed problem was recently proposed in [10]. We generalize this method to problems of arbitrary order, and we show that the

same idea can be applied to obtain feasible points for both primal and dual problems. To simplify understanding of the main idea we will provide explicit formulations for the more common second order problems.

The estimation of feasible primal and dual points is based on the following simple proposition.

Proposition 1 *Let $f\colon \mathbb{R}^N \times \mathbb{R}^M \to \mathbb{R}$ be a proper lower semi-continuous convex function of two vector variables and $(x^*, y^*) = \operatorname{argmin}_{(x,y)\in\mathbb{R}^N\times\mathbb{R}^M} f(x,y)$ be its minimizer. Let $x^t,\ t = 1, 2, \ldots$ be a sequence of points in $\mathbb{R}^N$ converging to x^*. If additionally the function $\varphi(x) = \min_{y\in\mathbb{R}^M} f(x,y)$ is continuous, then $\varphi(x^t) \xrightarrow{t\to\infty} f(x^*, y^*)$.*

The proof of the proposition is straightforward: since $x \xrightarrow{t\to\infty} x^*$ then due to continuity $\varphi(x) \xrightarrow{t\to\infty} \varphi(x^*) = \min_{y\in\mathbb{R}^M} f(x^*, y) = f(x^*, y^*)$.

To make use of the Proposition 1 for the calculation of primal and dual feasible points, we split our set of variables into two parts (x and y according to the notation of the Proposition 1). The subsets of variables should be selected such that the function φ is continuous and easy to compute. Indeed, as observed in [10] for the second-order case, with fixed $\mu_a, a \in \mathcal{F}^1$ satisfying the last two constraints in (8), the primal problem (8) splits into a set of independent small subproblems: one subproblem for each second-order factor. This has a straightforward generalization for problem (4) of arbitrary order, as stated by the following theorem:

Theorem 2 *Let μ^* be any solution of (4)-(7), and let μ^t be a sequence such that $\mu_a^t \xrightarrow{t\to\infty} \mu_a^*,\ \mu_a^t \geq 0,\ a \in \mathcal{F}^1$. Let ${\mu'}^t$ be constructed as follows:*

$$\forall a \in \mathcal{F}^1 \quad {\mu'}^t_a(x_a) = \Pi_{\Delta(\mathcal{X}_a)}(\mu_a^t)\,, \tag{19}$$

where $\Pi_{\Delta(\mathcal{X}_a)}\colon \mathbb{R}^{\mathcal{X}_a} \to \Delta(\mathcal{X}_a)$ denotes a projection operator to the $|\mathcal{X}_a|$-dimensional simplex $\Delta(\mathcal{X}_a)$, and

$$\begin{aligned} \forall b \in \mathcal{F}\backslash\mathcal{F}^1 \quad {\mu'}^t_b = \arg \min_{\mu_b\in\mathbb{R}^{\mathcal{X}_b}} \quad & \langle \theta_b, \mu_b \rangle \\ s.t. \quad & L_{ab}\mu_b = {\mu'}^t_a,\ a \subset b\,, \\ & \mu_b \quad \geq 0\,. \end{aligned} \tag{20}$$

Then

$$\langle \theta, {\mu'}^t \rangle \xrightarrow{t\to\infty} \langle \theta, \mu^* \rangle\,.$$

We prove the theorem in the appendix.

The dual to (20) reads (see its derivation in appendix):

$$\begin{aligned} \max_{\nu} \ & \sum_{a\in\mathcal{F}^1} \nu_a + \sum_{b\in\mathcal{F}\backslash\mathcal{F}^1} \nu_b \\ \text{s.t.} \quad & \theta_a - \sum_{\substack{b\supset a \\ b\in\mathcal{F}\backslash\mathcal{F}^1}} \nu_{ab} \geq \nu_a \cdot \mathbb{1}_a, \quad a \in \mathcal{F}^1\,, \\ & \theta_b + \sum_{\substack{a\subset b \\ a\in\mathcal{F}^1}} L_{ab}^\top \nu_{ab} \geq \nu_b \cdot \mathbb{1}_b,\ b \in \mathcal{F}\backslash\mathcal{F}^1\,. \end{aligned} \tag{21}$$

In the second-order case this formulation has the following well-known (cf. [5]) form:

$$\max_{\nu} \sum_{a\in\mathcal{F}^1} \nu_a + \sum_{b\in\mathcal{F}\setminus\mathcal{F}^1} \nu_b \tag{22}$$

$$\text{s.t. } \theta_a(x_a) - \sum_{\substack{b\supset a \\ b\in\mathcal{F}\setminus\mathcal{F}^1}} \nu_{ab}(x_a) \qquad \geq \nu_a,\ a\in\mathcal{F}^1,\ x_a\in\mathcal{X}_a\,,$$

$$\theta_b(x_a,x_{a'}) + \nu_{ab}(x_a) + \nu_{a'b}(x_{a'}) \geq \nu_b,\ b\in\mathcal{F}\setminus\mathcal{F}^1,\ b=a\cup a',\ (x_a,x_{a'})\in\mathcal{X}_b\,.$$

Since for each $b\in\mathcal{F}\setminus\mathcal{F}^1$ variable x_b is in fact a collection of $x_a,\ a\subset b$, we will use the notation $(x_b)_a$ for such x_a. The dual problem (21) becomes easily solvable with respect to $\nu_a,\ a\in\mathcal{F}^1$ and $\nu_b,\ b\in\mathcal{F}\setminus\mathcal{F}^1$, when ν_{ab} are fixed. This is stated by the following theorem.

Theorem 3 *Let ν^* denote any solution of (21), and ν^t be a sequence such that $\nu^t_{ab}\xrightarrow{t\to\infty}\nu^*_{ab},\ b\in\mathcal{F}\setminus\mathcal{F}^1, a\subset b$. Let ${\nu'}^t$ be constructed as follows:*

$${\nu'}^t_{ab} = \nu^t_{ab},\ b\in\mathcal{F}\setminus\mathcal{F}^1, a\subset b\,, \tag{23}$$

$${\nu'}^t_{a} = \min_{x_a\in\mathcal{X}_a} \theta_a(x_a) - \sum_{\substack{b\supset a \\ b\in\mathcal{F}\setminus\mathcal{F}^1}} \nu'_{ba}(x_a),\ a\in\mathcal{F}^1\,, \tag{24}$$

$${\nu'}^t_{b} = \min_{x_b\in\mathcal{X}_b} \theta_b(x_b) + \sum_{\substack{a\subset b \\ a\in\mathcal{F}^1}} L^\top_{ab}\nu'_{ba}((x_b)_a),\ b\in\mathcal{F}\setminus\mathcal{F}^1\,. \tag{25}$$

Then

$$\sum_{a\in\mathcal{F}^1} {\nu'}^t_{a} + \sum_{b\in\mathcal{F}\setminus\mathcal{F}^1} {\nu'}^t_{b} \xrightarrow{t\to\infty} \sum_{a\in\mathcal{F}^1} \nu^*_a + \sum_{b\in\mathcal{F}\setminus\mathcal{F}^1} \nu^*_b\,.$$

We prove the theorem in the appendix.

3 Experimental results

Test Cases. We compared the FPD approach to other established methods using standard grid-structured models from the Middlebury MRF-Benchmark [7], in particular the well-known Tsukuba stereo problem. We additionally used various synthetic models with a varying number of variable states (range $2,\dots,20$) and grid size (range $2^2,\dots,40^2$). The potential functions θ for nodes and edges were sampled from a uniform distribution.

Furthermore, we tested with a set of specific grids leading to LP-tight relaxations, where the LP problem (3) always has an integer minimizer. These graphical models were constructed as follows: First, starting from graphs with uniformly sampled potentials, for each unary factor ($a\in\mathcal{F}^1$) we chose one state to have the minimum local potential, and also modified the connected pairwise factors to assign the minimum energy to each pair of selected states, such that the

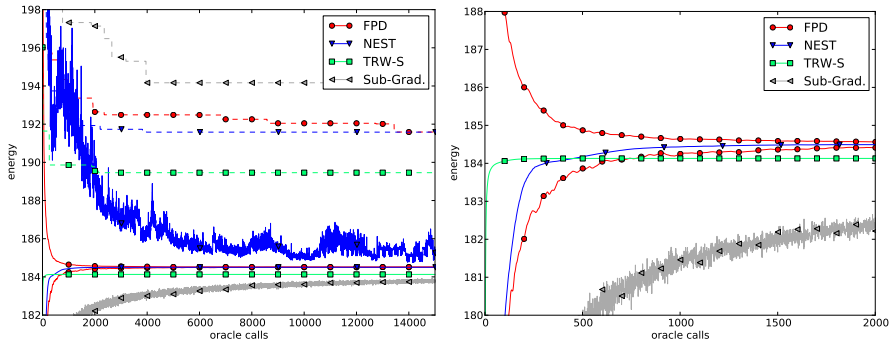

Fig. 1. FPD method for a 20×20 synthetic grid model with 5 states, in comparison to (a) NEST with $\epsilon = 1$, (b) TRW-S and (c) sub-gradient methods. The plots show LP lower and upper bounds (unavailable for TRW-S and subgradients) as well as integer bounds obtained by rounding (dashed). TRW-S is the fastest on this data, but gets stuck in a non-optimal fixed point. FPD gives much better upper bounds than NEST and achieves a low primal-dual gap much earlier, which is attributed to its direct optimization of both primal and dual variables. The right plot displays a close-up, highlighting the superiority w.r.t. subgradients and TRW-S in this case, but also shows that NEST achieves a better lower bound for a given number of iterations.

problem becomes trivial[3]. Second, we randomly sampled $\nu_{ab},\ b \in \mathcal{F}\backslash\mathcal{F}^1,\ a \in b$ (see (22)) associated with the pairwise factors and applied the reparametrization:

$$\theta_a(x_a) \leftarrow \theta_a(x_a) - \sum_{\substack{b \supset a \\ b \in \mathcal{F}\backslash\mathcal{F}^1}} \nu_{ab}(x_a),\ a \in \mathcal{F}^1,\ x_a \in \mathcal{X}_a\,, \tag{26}$$

$$\theta_b(x_a, x_{a'}) \leftarrow \theta_b(x_a, x_{a'}) + \nu_{ab}(x_a) + \nu_{a'b}(x_{a'}),$$
$$b \in \mathcal{F}\backslash\mathcal{F}^1,\ b = a \cup a',\ (x_a, x_{a'}) \in \mathcal{X}_b\,. \tag{27}$$

It is known [4, 5] that this reparametrization does not change the energy of any configuration.

Compared Algorithms. Our comparison includes Nesterov's method (NEST) of [10], a sub-gradient method [8], as well as TRW-S [6], all based on the same dual decomposition to acyclic subgraphs corresponding to the rows and columns of the input grid. For TRW-S and NEST, the authors kindly provided the original implementation of their algorithms. We show the lower (LP-dual due to Th. 3) and upper (LP-primal due to Th. 2 and integer-rounded values of $\mu_a^t,\ a \in \mathcal{F}^1$) bounds on the energy, which the algorithms achieve. To ensure comparability, the algorithm progress is not plotted against time, but instead as a function of the number of oracle calls, i.e. the required number of objective function or gradient evaluations. For FPD, one oracle call corresponds to a single iteration.

[3] For a definition of such problems, please see [5, Sect. III.D] or [9, part I, eq. (7)].

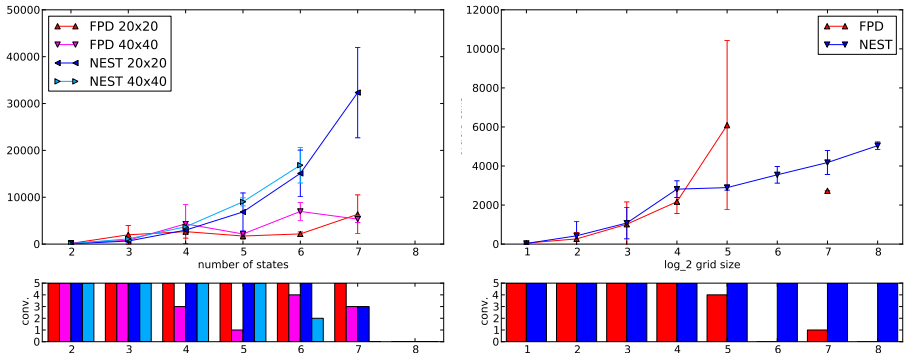

Fig. 2. Convergence behaviour of FPD compared with NEST in terms of the number of iterations (oracle calls) required to reach a 0.1% duality gap, for synthetic grid graph problems. **Left:** varying the number of states; **Right:** varying the grid size. The lower bar-plots show the number of runs (out of 5) which converged within a maximum of 15000 iterations, which were therefore included in the upper plot. The increase with the number of states is more pronounced for NEST, probably because the primal problems becomes more complicated, which only FPD optimizes directly along with the dual. With increasing graph size however, the converse is true, where FPD quickly required more than the maximum number of iterations permitted in the experiment.

Synthetic grid problem. In the first experiment, all four methods were compared using a synthetic grid graph problem (Fig. 1). We first note that TRW-S converges to a suboptimal fixed point. Furthermore, the subgradient scheme is not competitive. The primal bound obtained from our FPD method drops faster than that of the Nesterov-based method, which is attributed to the fact that the latter does not optimize the primal problem directly, while FPD does. Note that NEST employs smoothing, which depends on the required precision ϵ, which also influences the convergence.

In the second group of experiments, we study the dependence of the number of oracle calls required to reach a given precision in the number of variables and the number of states per variable. Here, we compare only FPD against NEST, because none of the other methods provides a primal bound of the relaxed problem (3).

Dependence on the number of states. According to Figs. 2, 3, both FPD and NEST require increasing numbers of iterations with increasing number of states. This increase is more pronounced for NEST, which again probably is due to NEST being a dual optimization method rather than a primal-dual one.

Dependence on the number of variables. When the number of variables in the model increases, the number of iterations increases as Figs. 2, 3 show. As opposed to the dependency on the number of states, here the growth is much more pronounced for the FPD method, quickly exceeding the maximal number of iterations we imposed for the experiment, while for NEST, the increase is

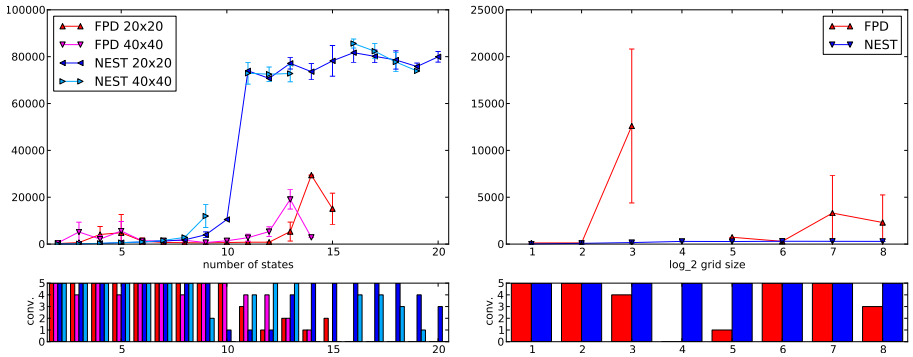

Fig. 3. Convergence behaviour of FPD compared with NEST as in Fig. 2, but for a target precision of a 1% duality gap and a maximum number of 30000 iterations permitted. **Left:** varying the number of states; **Right:** varying the grid size. Again NEST requires more oracle calls than FPD for larger numbers of states, but FPD quickly becomes inferior with increasing numbers of variables.

moderate. This can be explained by dual decomposition into large subgraphs in NEST, that leads to faster propagation of information across the graph.

LP-tight problems. One reason for the differences observed above may be the different optimization strategies of FPD and NEST: While the first optimizes primal and dual simultaneously, the latter only optimizes the dual. Hence we confirmed this conjecture by considering LP-tight graphical models:Fig. 4 shows the corresponding result. The required number of oracle calls for both methods linearly depends on the number of states, and also grows for larger graphs, with NEST achieving the required precision several times faster. The overall number of oracle calls is much lower than for non-LP-tight problems, as the complexity of the primal optimization is absent, and the algorithms mainly optimize the dual, which is apparently much easier.

Tsukuba stereo problem. Fig. 5 finally presents results of a comparison of the energy minimization algorithms for the Tsukuba dataset. Among all methods, FPD shows the slowest convergence. This is indeed to be expected since the problem is quite large for FPD (110592 variables) and contains a relatively large number of states (16 depth states). Other methods (even typically slow subgradient) are more efficient presumably due to use of the dual decomposition to large subgraphs.

Parallelization properties. Besides the moderately-sized dual vector, the algorithm requires to handle the set of primal variables, whose storage requirement in the pairwise case is $O(|\mathcal{F}^1|K + |\mathcal{F} \setminus \mathcal{F}^1|K^2)$, where K is the number of states (assuming an equal number of states per node). Due to the quadratic growth in K, this is a major drawback of the method if the problem has many states.

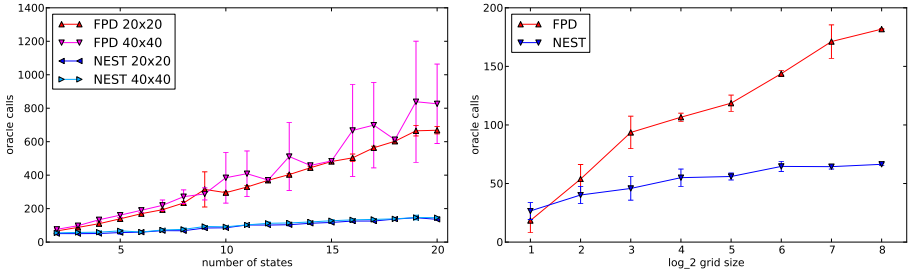

Fig. 4. Convergence behaviour of FPD compared with NEST for LP-tight synthetic grid graph problems, in terms of the number of iterations (oracle calls) required to reach a 0.1% duality gap. **Left:** varying the number of states; **Right:** varying the grid size. The overall number of oracle calls is much less than for non-LP-tight problems, as the complexity of the primal optimization is absent and the algorithms mainly optimize the dual, which is apparently much easier. In that case, NEST is clearly superior to FPD in almost all instances.

However, the method (including the bounds computation) is easily parallelizable, which we exploited to provide a CUDA variant running on GPU hardware[4]. This code allowed practical speedups up to a factor of 160 compared to a sequential variant running on a CPU[5] (both not explicitly optimized for grid graphs).

4 Conclusion

We presented a study of the first order primal-dual algorithmic scheme [1] applied to MAP inference for general discrete graphical models, via the LP relaxation of this problem formulated in a saddle-point form. We supplemented the original scheme by a method for computing upper and lower bounds, which results in a sound stopping condition and thus ensures comparability and reproducibility of results.

Our study shows that the performance of the algorithm rapidly drops as the model size increases. Competitive methods, which use a dual decomposition technique, appear to propagate information across the graph much faster. However due to an explicit optimization of the primal objective, FPD accomplishes faster improvement of the primal bound than the application of Nesterov's scheme [10] to the dual objective, which is not optimizing the primal directly. This effect is clearly visible for small-sized graphical models.

Future work will focus on a combination of efficient optimization and decomposition schemes, both for primal and dual objectives. Good parallelization properties – as given for the presented FPD method – will also play a key role in further improving the efficiency of convergent inference methods.

[4] NVidia GTX 480
[5] Intel Core-i7 860 (using one of 4 cores) at 2.8 GHz

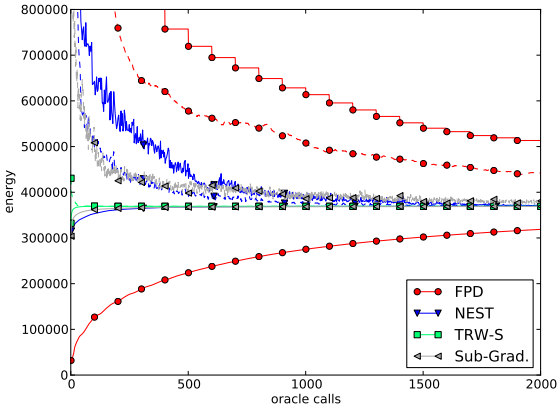

Fig. 5. FPD method for Tsukuba data in comparison to (a) NEST with $\epsilon = 10$, (b) TRW-S and (c) sub-gradient methods. The plot shows LP lower and upper bounds (unavailable for TRW-S and subgradients) as well as integer bounds obtained by rounding (dashed). FPD shows the slowest convergence among all methods, as the problem is relatively large in terms of both the grid size and the number of states (16 depth states). Note however that all other methods use dual decomposition into large subgraphs, and therefore may be able to propagate information faster across the graph.

Acknowledgements The authors gratefully acknowledge support by the German Science Foundation (DFG) within the Excellence Initiative (HCI) and by the joint research project "Spatio/Temporal Graphical Models and Applications in Image Analysis", grant GRK 1653. The authors thank B. Andres, D. Breitenreicher and J. Lellmann for many helpful discussions and sharing software.

References

1. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. J. Math. Imaging and Vision (2010) 1–26
2. Kschischang, F., Frey, B., Loeliger, H.A.: Factor graphs and the sum-product algorithm. IEEE Transactions on Information Theory **47** (2001) 498–519
3. Wainwright, M.J., Jordan, M.I.: Graphical models, exponential families, and variational inference. Found. Trends Mach. Learn. **1** (2008) 1–305
4. Schlesinger, M.: Syntactic analysis of two-dimensional visual signals in the presence of noise. Kibernetika (1976) 113–130
5. Werner, T.: A linear programming approach to max-sum problem: A review. IEEE PAMI **29** (2007)
6. Kolmogorov, V.: Convergent tree-reweighted message passing for energy minimization. IEEE PAMI **28** (2006) 1568–1583
7. 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. IEEE PAMI **30** (2008) 1068–1080

8. Komodakis, N., Paragios, N., Tziritas, G.: MRF optimization via dual decomposition: Message-passing revisited. In: ICCV. (2007)
9. Schlesinger, M., Giginyak, V.: Solution to structural recognition (MAX,+)-problems by their equivalent transformations. In 2 parts. Control Systems and Computers (2007)
10. Savchynskyy, B., Kappes, J., Schmidt, S., Schnörr, C.: A study of Nesterov's scheme for Lagrangian decomposition and MAP labeling. In: CVPR 2011. (2011)
11. Korte, B., Vygen, J.: Combinatorial Optimization. 4th edn. Springer (2008)
12. Nesterov, Y.: Smooth minimization of non-smooth functions. Math. Program. **Ser. A** (2004) 127–152
13. Golub, G., Van Loan, C.: Matrix Computations. 3rd edn. The John Hopkins Univ. Press (1996)
14. Michelot, C.: A finite algorithm for finding the projection of a point onto the canonical simplex of R^n. J. Optim. Theory Appl. **50** (1986) 195–200
15. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, New York, USA (2004)

5 Appendix

Proof of Theorem 1. The matrix L has a block form

$$L = \begin{pmatrix} L_V & 0 \\ L_{VE} & L_E \end{pmatrix} ,$$

where L_V is determined by (10), L_E and L_{VE} respectively correspond to the left hand and right hand sides of (9), and

$$\begin{pmatrix} L_1 & L_2^\top \\ L_2 & L_3 \end{pmatrix} := LL^\top , \qquad L_1 = KI_{|V|} ,$$

where K is the number of possible states. Let $x := \left(\begin{smallmatrix} x_V \\ x_E \end{smallmatrix}\right)$ denote the maximal eigenvector of $LL^\top$ corresponding to $\lambda_{\max}$. From the upper part of the eigenvalue equation $LL^\top x = \lambda_{\max} x$,

$$L_1 x_v + L_2^\top x_E = \lambda_{\max} x_V ,$$

and $L_1 = KI_{|V|}$, we conclude $x_V = \frac{1}{\lambda_{\max} - K} L_2^\top x_E$. Insertion into the lower part of the eigenvalue equation yields

$$\left(\frac{1}{\lambda_{\max} - K} L_2 L_2^\top + L_3 \right) x_E = \lambda_{\max} x_E . \tag{28}$$

In terms of the components of L, this equation reads

$$\left(\frac{\lambda_{\max}}{\lambda_{\max} - K} L_{VE} L_{VE}^\top + L_E L_E^\top \right) x_E = \lambda_{\max} x_E .$$

The maximum eigenvalue of the matrix on the left hand side equals $\lambda_{\max}$ and itself depends on $\lambda_{\max}$. We therefore solve the equation

$$\frac{\lambda}{\lambda - K} d_{\max} + 2K = \lambda \quad \Rightarrow \quad \lambda = \frac{1}{2} \left(3K + d_{\max} + \sqrt{K^2 + 6 d_{\max} K + d_{\max}^2} \right) .$$

where $d_{\max}$ denotes the maximal degree of any node of G. The expression on the left is an upper bound of the ℓ_1-norms of the row vectors of the matrix, which is an upper bound of $\lambda_{\max}$ by Gerschgorin's theorem [13]. Because this function decreases with λ, it follows $\lambda \geq \lambda_{\max}$.

Derivation of the block form of the dual objective (21). The full Lagrangian corresponding to (4) reads:

$$\mathcal{L}(\mu, \tilde{\nu}, \gamma) = \langle \theta, \mu \rangle + \sum_{\substack{b \in F \setminus F^1 \\ a \subset b}} \tilde{\nu}_{ab}(L_{ab}\mu_b - \mu_a) + \sum_{a \in F^1} \tilde{\nu}_a(1 - \mathbb{1}_a^\top \mu_a) - \langle \gamma, \mu \rangle \,. \tag{29}$$

From this follows the dual problem (a detailed presentation of (12)):

$$\max_{\tilde{\nu}} \sum_a \tilde{\nu}_a \tag{30}$$

$$\begin{aligned} \text{s.t.} \quad & \theta_a - \textstyle\sum_{\substack{b \supset a \\ b \in F \setminus F^1}} \tilde{\nu}_{ab} \geq \tilde{\nu}_a \cdot \mathbb{1}_a, \;\; a \in F^1 \\ & \theta_b + \textstyle\sum_{\substack{a \subset b \\ a \in F^1}} L_{ab}^\top \tilde{\nu}_{ab} \geq 0 \cdot \mathbb{1}_b, \;\; b \in F \setminus F^1 \end{aligned} \tag{31}$$

We introduce additional variables $\nu_b \in \mathbb{R},\ b \in F \setminus F^1$ and apply the following change of variables:

$$\nu_{ab} := \tilde{\nu}_{ab} + \nu_b \cdot \mathbb{1}_a \,, \qquad \nu_a := \tilde{\nu}_a + \sum_{b \supset a} \frac{\nu_b}{|b|} \,. \tag{32}$$

The matrix L_{ab} in (4) is of size $|\mathcal{X}_b| \times |\mathcal{X}_a|$, and it possesses the important property

$$\sum_{a \subset b} L_{ab}^\top \mathbb{1}_a = \mathbb{1}_b \,, \tag{33}$$

following from the fact that each column of L_{ab} contains exactly one non-zero entry (equalling 1).

Taking into account (33) and $\sum_{a \in F^1} \sum_{b \supset a} \frac{\nu_b}{|b|} = \sum_{b \in F \setminus F^1} \nu_b$ leads to the equivalent dual problem formulation (21).

Proof of Theorem 2. Due to Proposition 1 and continuity of the projection in (19), it suffices to prove that the objective value of (20) continuously changes with ${\mu'}_a^t$. Our proof is a straightforward generalization of the one given in [10]. Problem (20) satisfies Slater's condition [15] due to affinity of its constraints and it always has at least one feasible point when $\mathbb{1}_a^\top \mu_a = 1$, $a \subset b$. This condition holds due to (19). Since $1 \geq \mu_b \geq 0$ its optimal value is always finite. Thus its Lagrange dual has the same finite optimal value. The Lagrange dual for (20) reads:

$$\max_{\substack{\xi_{ab} \in \mathbb{R}^{\mathcal{X}_a} \\ a \subset b}} \sum_{a \subset b} \xi_{ab} {\mu'}_a^t \qquad \text{s.t.} \quad \theta_b - L_{ab}^\top \xi_{ab} \geq 0 \,. \tag{34}$$

It depends on ${\mu'}_a^t$ only through its objective, which continuously depends on ${\mu'}_a^t$. Since optimal value of (34) is finite, it is attained in one of the vertices of its constraint set, which implies that it changes continuously with ${\mu'}_a^t$.

Proof of Theorem 3. The proof follows from Proposition 1 and continuity of the min-operation in (24)-(25).