Efficiently Solving Convex Relaxations for MAP Estimation

M. Pawan Kumar PAWAN@ROBOTS.OX.AC.UK
Department of Engineering Science, University of Oxford

P.H.S. Torr PHILIPTORR@BROOKES.AC.UK
Department of Computing, Oxford Brookes University

Abstract

The problem of obtaining the maximum *a posteriori* (MAP) estimate of a discrete random field is of fundamental importance in many areas of Computer Science. In this work, we build on the tree reweighted message passing (TRW) framework of (Kolmogorov, 2006; Wainwright et al., 2005). TRW iteratively optimizes the Lagrangian dual of a linear programming relaxation for MAP estimation. We show how the dual formulation of TRW can be extended to include cycle inequalities (Barahona & Mahjoub, 1986) and some recently proposed second order cone (SOC) constraints (Kumar et al., 2007). We propose efficient iterative algorithms for solving the resulting duals. Similar to the method described in (Kolmogorov, 2006), these algorithms are guaranteed to converge. We test our approach on a large set of synthetic data, as well as real data. Our experiments show that the additional constraints (i.e. cycle inequalities and SOC constraints) provide better results in cases where the TRW framework fails (namely MAP estimation for non-submodular energy functions).

1. Introduction

The problem of obtaining the maximum *a posteriori* (MAP) estimate of a discrete random field plays a central role in various applications, e.g. stereo reconstruction (Szeliski et al., 2006) and protein side-chain prediction (Sontag & Jaakkola, 2007). Furthermore, it is closely related to many important combinatorial optimization problems such as MAXCUT (Goemans & Williamson, 1995) and 0-extension (Karzanov, 1998). It is therefore not surprising that a number of approximate MAP estimation approaches exist in the literature. One such class of approaches which provides a good approximation, both in theory and in practice, is based on convex relaxations (e.g. see (Kumar et al., 2007) for an overview). In this work, we focus on the issue of solving these relaxations efficiently with the goal of handling a large number of random variables, e.g. variables corresponding to pixels in an image.

A discrete random field is defined over random variables $\mathbf{v} = \{v_0, \cdots, v_{n-1}\}$, each of which can take a label from the set $\mathbf{l} = \{l_0, \cdots, l_{h-1}\}$. Throughout this paper, we will assume a conditional random field (CRF) while noting that all our results are applicable to the Markov random field framework. A CRF describes a neighbourhood relationship $\mathcal{E}$ between the variables such that $(a, b) \in \mathcal{E}$ if, and only if, v_a and v_b are neighbours. A labelling of the CRF is specified by a function $f : \{0, \cdots, n-1\} \longrightarrow \{0, \cdots, h-1\}$ (i.e. variable v_a takes label $l_{f(a)}$). Given data $\mathbf{D}$, the energy of the labelling is given by

$$Q(f; \mathbf{D}, \boldsymbol{\theta}) = \sum_{v_a \in \mathbf{v}} \theta^1_{a;f(a)} + \sum_{(a,b) \in \mathcal{E}} \theta^2_{ab;f(a)f(b)}, \quad (1)$$

where $\theta^1_{a;f(a)}$ and $\theta^2_{ab;f(a)f(b)}$ are the data-dependent unary and pairwise potentials respectively, and $\boldsymbol{\theta}$ denotes the parameter of the CRF. The problem of MAP estimation is to obtain the labelling f^* with the minimum energy (or equivalently the maximum posterior probability), i.e. $f^* = \arg\min_f Q(f; \mathbf{D}, \boldsymbol{\theta})$.

Related Work: We build upon the linear programming (LP) relaxation of (Wainwright et al., 2005), which we call LP-S (since it was first proposed by (Schlesinger, 1976) for the special case of hard constraint pairwise potentials). Although the LP-S relaxation can be solved in polynomial time using Interior Point algorithms, the state of the art softwares can only handle up to a few hundred variables due to their large memory requirements. To overcome this prob-

Appearing in *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008.

lem, two iterative algorithms were proposed by (Wainwright et al., 2005) for solving the dual of the LP-S relaxation. Similar to min-sum belief propagation (BP), these algorithms are not guaranteed to converge. The work of (Kolmogorov, 2006) addressed this problem by proposing a convergent sequential tree-reweighted message passing (TRW-S) algorithm for solving the dual.

Despite its strong theoretical foundation, it was observed that TRW-S yields labellings with very high energies when the energy function contains non-submodular terms (Kolmogorov, 2006). This is not surprising since the LP-S relaxation provides an inaccurate approximation in such cases (e.g. see (Kumar et al., 2007)). In this work, we address this deficiency of TRW-S by appending the LP-S relaxation with some useful constraints.

Our Results: We show how the dual formulation of the LP-S relaxation can be extended to include linear cycle inequalities (Barahona & Mahjoub, 1986) (section 3). Furthermore, we incorporate the recently proposed second order cone (SOC) constraints of (Kumar et al., 2007) within this framework (section 4). Note that although the importance of cycle inequalities and SOC constraints is well-recognized, their use has been limited to a small number of random variables due to the lack of efficient algorithms (Sontag & Jaakkola, 2007). Our results on including these constraints within the TRW formulation allow us to develop efficient convergent algorithms for solving the resulting duals. We successfully apply these algorithms to several synthetic and real problems containing a large number of variables which could not be handled by previous approaches (section 5). Our experiments indicate that incorporating these constraints provides a much better approximation for the MAP estimation problem within reasonable computational times compared to several state of the art algorithms. Additional experimental results and proofs are provided in (Kumar & Torr, 2008).

2. Preliminaries

We begin by introducing some notation which would allow us to describe our results concisely.

Optimal Energy and Min-Marginals: The energy of the optimal labelling and the min-marginals of random variables and neighbouring random variables is given by the following equations respectively:

$$q(\boldsymbol{\theta}) = \min_{f} Q(f : \mathbf{D}), \tag{2}$$

$$q_{a;i}(\boldsymbol{\theta}) = \min_{f, f(a)=i} Q(f; \mathbf{D}, \boldsymbol{\theta}), \tag{3}$$

$$q_{ab;ij}(\boldsymbol{\theta}) = \min_{f, f(a)=i, f(b)=j} Q(f; \mathbf{D}, \boldsymbol{\theta}), \tag{4}$$

where the term $\mathbf{D}$ is dropped from the LHS to make the notation less cluttered.

Reparameterization: A parameter $\overline{\boldsymbol{\theta}}$ is called a reparameterization of the parameter $\boldsymbol{\theta}$ (denoted by $\overline{\boldsymbol{\theta}} \equiv \boldsymbol{\theta}$) if, and only if,

$$Q(f; \mathbf{D}, \overline{\boldsymbol{\theta}}) = Q(f; \mathbf{D}, \boldsymbol{\theta}), \forall f. \tag{5}$$

Over-complete Representations: A labelling f can be represented using an over-complete set of boolean variables $\mathbf{y}$ defined as

$$y_{a;i} = \begin{cases} 1 & \text{if } f(a) = i, \\ 0 & \text{otherwise.} \end{cases}, \quad y_{ab;ij} = y_{a;i} y_{b;j}. \tag{6}$$

We also define variables $(\mathbf{x}, \mathbf{X})$ such that

$$x_{a;i} = 2y_{a;i} - 1, \;\; X_{ab;ij} = 4y_{ab;ij} - 2y_{a;i} - 2y_{b;j} + 1. \tag{7}$$

We will sometimes specify the additional constraints (i.e. cycle inequalities and SOC constraints) using variables $(\mathbf{x}, \mathbf{X})$, since they will allow us to write these constraints concisely.

The LP-S Relaxation: The LP-S relaxation of the MAP estimation problem is given by

$$\mathbf{y}^* = \arg\min_{\mathbf{y} \in LOCAL(\mathbf{v}, \mathcal{E})} \mathbf{y}^\top \boldsymbol{\theta},$$

$$LOCAL(\mathbf{v}, \mathcal{E}) = \begin{cases} y_{a;i} \in [0,1], y_{ab;ij} \in [0,1], \\ \sum_{l_i \in \mathbf{l}} y_{a;i} = 1, \\ \sum_{l_j \in \mathbf{l}} y_{ab;ij} = y_{a;i}. \end{cases} \tag{8}$$

The term $LOCAL(\mathbf{v}, \mathcal{E})$ stands for *local consistency polytope* (Wainwright et al., 2005) and denotes the feasibility region of the LP-S relaxation (specified by the above constraints for all $v_a \in \mathbf{v}, (a,b) \in \mathcal{E}, l_i, l_j \in \mathbf{l}$).

Dual of the LP-S Relaxation: Let $\mathcal{T}$ denote a set of tree-structured CRFs defined over subsets of the given random variables. For a CRF $T \in \mathcal{T}$, we denote its random variables by $\mathbf{v}_T$, its neighbourhood relationship by $\mathcal{E}_T$ and its parameter as $\boldsymbol{\theta}^T$. The parameter $\boldsymbol{\theta}^T$ consists of unary potentials $\theta^{T1}_{a;i}$ and pairwise potentials $\theta^{T2}_{ab;ij}$. Let $\boldsymbol{\rho} = \{\rho(T), T \in \mathcal{T}\}$ be a set of non-negative real numbers which sum to one. Using the above notation, the dual of the LP-S relaxation can be written as follows (Kolmogorov, 2006; Wainwright et al., 2005):

$$\max_{\sum_{T \in \mathcal{T}} \rho(T) \boldsymbol{\theta}^T \equiv \boldsymbol{\theta}} \sum_T \rho(T) q(\boldsymbol{\theta}^T). \tag{9}$$

The TRW-S Algorithm: Table 1 describes the TRW-S algorithm (Kolmogorov, 2006) which attempts to solve the dual of the LP-S relaxation. In other words, it solves for the set of parameters $\boldsymbol{\theta}^T$, $T \in \mathcal{T}$, which maximize the dual (9). There are two main steps: (i)

reparameterization, which involves running one pass of BP on the tree structured CRFs T; and (ii) averaging operation. TRW-S is guaranteed not to decrease the value of the dual (9) at each iteration. Further, it can be shown that it converges to a solution which satisfies the *weak tree agreement* (WTA) (Kolmogorov, 2006).

Initialization

1. For every $\omega \in \mathbf{v} \bigcup \mathcal{E}$, find all trees $\mathcal{T}_\omega \subseteq \mathcal{T}$ which contains ω.
2. Initialize $\boldsymbol{\theta}^T$ such that $\sum_T \rho(T)\boldsymbol{\theta}^T \equiv \boldsymbol{\theta}$.
   Typically, we set $\rho(T) = \frac{1}{|\mathcal{T}|}$ for all $T \in \mathcal{T}$.
   Then we can initialize $\theta^{T1}_{a;i} = \theta^1_{a;i} \frac{|\mathcal{T}|}{|\mathcal{T}_{v_a}|}$ for all $T \in \mathcal{T}_{v_a}$.
   Similarly, $\theta^{T2}_{ab;ij} = \theta^2_{ab;ij} \frac{|\mathcal{T}|}{|\mathcal{T}_{(a,b)}|}$ for all $T \in \mathcal{T}_{(a,b)}$.

Iterative Steps

3. Pick an element $\omega \in \mathbf{v} \bigcup \mathcal{E}$.
4. For all $T \in \mathcal{T}_\omega$, reparameterize $\boldsymbol{\theta}^T$ to $\overline{\boldsymbol{\theta}}^T$ such that
   (i) $\overline{\theta}^{T1}_{a;i} = q_{a;i}(\boldsymbol{\theta}^T)$, if $\omega = v_a \in \mathbf{v}$,
   (ii) $\overline{\theta}^{T1}_{a;i} + \overline{\theta}^{T1}_{b;j} + \overline{\theta}^{T2}_{ab;ij} = q_{ab;ij}(\boldsymbol{\theta}^T)$, if $\omega = (a,b) \in \mathcal{E}$.
   This step involves running one iteration of BP for T.
5. Averaging operation:
   (i) If $\omega = v_a \in \mathbf{v}$,
   (a) Compute $\nu_{a;i} = \frac{1}{\rho_a} \sum_{T \in \mathcal{T}_a} \rho(T) \overline{\theta}^{T1}_{a;i}$.
   (b) Set $\overline{\theta}^{T1}_{a;i} = \nu_{a;i}$, for all $T \in \mathcal{T}_{v_a}$.
   (ii) If $\omega = (a,b) \in \mathcal{T}$,
   (a) Compute $\nu_{ab;ij} = \frac{1}{\rho_{ab}} \sum_{T \in \mathcal{T}_{(a,b)}} \rho(T)(\overline{\theta}^{T1}_{a;i} + \overline{\theta}^{T1}_{b;j} + \overline{\theta}^{T2}_{ab;ij})$.
   (b) Set $\overline{\theta}^{T1}_{a;i} + \overline{\theta}^{T1}_{b;j} + \overline{\theta}^{T2}_{ab;ij} = \nu_{ab;ij}$, for all $T \in \mathcal{T}_{(a,b)}$.
6. Repeat steps 3, 4 and 5 till convergence.

Table 1. The TRW-S algorithm. Recall that $\theta^{T1}_{a;i}$ and $\theta^{T2}_{ab;ij}$ are the unary and pairwise potentials for the parameter $\boldsymbol{\theta}^T$. Similarly, $\overline{\theta}^{T1}_{a;i}$ and $\overline{\theta}^{T2}_{ab;ij}$ are the unary and pairwise potentials defined by the parameter $\overline{\boldsymbol{\theta}}$. The terms $\rho_a = \sum_{T, v_a \in \mathbf{v}_T} \rho(T)$ and $\rho_{ab} = \sum_{T,(a,b) \in \mathcal{E}_T} \rho(T)$ are the variable and edge appearance terms for $v_a \in \mathbf{v}$ and $(a,b) \in \mathcal{E}$ respectively. In step 3, the value of the dual (9) remains unchanged. Step 4, i.e. the averaging operation, ensures that the value of the dual does not decrease. TRW-S converges to a solution which satisfies the WTA condition.

3. Adding Linear Constraints

We now show how the results of (Kolmogorov, 2006; Wainwright et al., 2005) can be extended to include an arbitrary number of linear cycle inequalities (Barahona & Mahjoub, 1986; Kumar et al., 2007). This requires us to incorporate cycle inequalities into the dual (11).

We begin by briefly describing cycle inequalities. Consider a cycle of length c in the graphical model of the given CRF, which is specified over a set of random variables $\mathbf{v}_C = \{v_b, b = a_1, a_2, \cdots, a_c\}$ such that $\mathcal{E}_C = \{(a_1,a_2),(a_2,a_3),\cdots,(a_n,a_1)\} \subseteq \mathcal{E}$. Further, let $\mathcal{E}_F \subseteq \mathcal{E}_C$ such that $|\mathcal{E}_F|$ (i.e. the cardinality of $\mathcal{E}_F$) is odd. Using these sets of edges, a cycle inequality can be specified as

$$\sum_{(a_k,a_m) \in \mathcal{E}_F} X_{a_k a_m; i_k i_m} - \sum_{(a_k,a_m) \in \mathcal{E}_C - \mathcal{E}_F} X_{a_k a_m; i_k i_m} \geq 2 - c, \tag{10}$$

where $l_{i_k}, l_{i_m} \in \mathbf{l}$[1]. The variables $X_{a_k a_m; i_k i_m}$ are defined in equation (7). It can be shown that adding cycle inequalities to LP-S, i.e. problem (8), provides a better relaxation than LP-S alone. Their importance is reflected in their wide use in recent literature such as (Sontag & Jaakkola, 2007; Zwick, 1999).

In general, a set of N_C cycle inequalities defined on a cycle $C = (\mathbf{v}_C, \mathcal{E}_C)$ (using different labels l_{i_k} for variables $v_{a_k} \in \mathbf{v}_C$) can be written as $\mathbf{A}^C \mathbf{y} \geq \mathbf{b}^C$. In other words, for every cycle we can define up to h^c cycle inequalities (where $h = |\mathbf{l}|$), i.e. $N_C \in \{0, 1, \cdots, h^c\}$. Let $\mathcal{C}$ be a set of cycles in the given CRF. Theorem 1 (given below) provides us with the dual of the LP relaxation obtained by appending problem (8) with cycle inequalities (defined over cycles in the set $\mathcal{C}$). We refer to the resulting relaxation as LP-C (where C denotes cycles).

Theorem 1: The following problem is the dual of problem (8) appended with a set of cycle inequalities $\mathbf{A}^C \mathbf{y} \geq \mathbf{b}^C$, for all $C \in \mathcal{C}$ (hereby referred to as the LP-C relaxation):

$$\begin{aligned} \max \quad & \textstyle\sum_T \rho(T) q(\boldsymbol{\theta}^T) + \sum_C \rho'(C) (\mathbf{b}^C)^\top \mathbf{u}^C, \\ \text{s.t.} \quad & \textstyle\sum_T \rho(T) \boldsymbol{\theta}^T + \sum_C \rho'(C) (\mathbf{A}^C)^\top \mathbf{u}^C \equiv \boldsymbol{\theta}, \\ & u^C_k \geq 0, \forall k \in \{1, 2, \cdots, N_C\}, C \in \mathcal{C}. \end{aligned} \tag{11}$$

Here $\boldsymbol{\rho}' = \{\rho'(C), C \in \mathcal{C}\}$ is some (fixed) set of non-negative real numbers which sum to one, and $\mathbf{u}^C = \{u^C_k, k = 1, \cdots, N_C\}$ are some non-negative slack variables.

Similar to the dual (9), the above problem cannot be solved using standard software for a large number of variables $\mathbf{v}$. In order to overcome this deficiency we propose a convergent algorithm (similar to TRW-S) to approximately solve problem (11). We call our approach the TRW-S(LP-C) algorithm. In order to describe TRW-S(LP-C), we need the following definitions.

We say that a tree structured random field $T = (\mathbf{v}_T, \mathcal{E}_T) \in \mathcal{T}$ *belongs to* a cycle $C = (\mathbf{v}_C, \mathcal{E}_C) \in \mathcal{C}$ (denoted by $T \in C$) if, and only if, there exists an edge $(a,b) \in \mathcal{E}_T$ such that $(a,b) \in \mathcal{E}_C$. In other words, $T \in C$ if they share a common pair of neighbouring random variables $(a,b) \in \mathcal{E}$. We also define the following problem:

$$\max \quad \textstyle\sum_{T \in C} \rho(T) q(\boldsymbol{\theta}^T) + \rho'(C) (\mathbf{b}^C)^\top \mathbf{u}^C,$$

[1]Note that using the variable $\mathbf{y}$ would result in a less compact representation of cycle inequalities.

$$\text{s.t.} \quad \sum_{T \in C} \rho(T)\boldsymbol{\theta}^T + \rho'(C)(\mathbf{A}^C)^\top \mathbf{u}^C = \boldsymbol{\theta}^C, \quad u_k^C \geq 0, \forall k \in \{1, 2, \cdots, N_C\}, \tag{12}$$

for some parameter $\boldsymbol{\theta}^C$. The variables of the above problem are restricted to u_k^C, $\theta_{a;i}^{T1}$ and $\theta_{ab;ij}^{T2}$ where $(a, b) \in \mathcal{E}_T \bigcap \mathcal{E}_C$ for some $T \in C$. In other words, problem (12) has fewer variables and constraints than dual (11) and can be solved easily using standard Interior Point algorithms for small cycles C. As will be seen, even using cycles of size 3 or 4 results in much better approximations of the MAP estimation problem for non-submodular energy functions.

Table 2 describes the convergent TRW-S(LP-C) algorithm for approximately solving the dual (11). The algorithm consists of two main steps : (i) solving problem (12) for a cycle; and (ii) running steps 4 and 5 of the TRW-S algorithm. Note that our approach is different from other generalizations of TRW, e.g. (Wiegernick, 2005) which computes marginals. Specifically, we do not cluster random variables but include additional constraints to reduce the feasibility region of the relaxation. Our experiments in section 5 show that, unlike (Wiegernick, 2005), we always outperform BP. The properties of the TRW-S(LP-C) algorithm are summarized below.

Initialization
1. Choose a set of tree structured random fields $\mathcal{T}$. Choose a set of cycles $\mathcal{C}$. For example, if the 4-neighbourhood is employed, $\mathcal{C}$ can be the set of all cycles of size 4.
2. Initialize $\boldsymbol{\theta}^T$ such that $\sum_T \rho(T)\boldsymbol{\theta}^T \equiv \boldsymbol{\theta}$. Initialize $u_k^C = 0$ for all C and k.
Iterative Steps
3. Pick an element $\omega \in \mathbf{v} \bigcup \mathcal{C}$. Find all cycles $\mathcal{C}_\omega \subseteq \mathcal{C}$ which contains ω.
4. For a cycle $C \in \mathcal{C}_\omega$, compute $\boldsymbol{\theta}^C = \sum_{T \in C} \rho(T)\boldsymbol{\theta}^T + \rho'(C)(\mathbf{A}^C)^\top \mathbf{u}^C$ using the values of $\boldsymbol{\theta}^T$ and $\mathbf{u}^C$ obtained in the previous iteration. Solve problem (12) using an Interior Point method. Update the values of $\boldsymbol{\theta}^T$ and $\mathbf{u}^C$.
5. For all trees $T \in \mathcal{T}$ which contain ω, run steps 4 and 5 of the TRW-S algorithm.
6. Repeat steps 3 and 4 for all cycles $C \in \mathcal{C}_\omega$.
7. Repeat steps 3 to 5 for all elements ω till convergence.

Table 2. The TRW-S(LP-C) algorithm.

3.1. Properties of the TRW-S(LP-C) Algorithm.

Property 1: *At each step of the algorithm, the reparameterization constraint is satisfied, i.e.*

$$\sum_T \rho(T)\boldsymbol{\theta}^T + \sum_C \rho'(C)(\mathbf{A}^C)^\top \mathbf{u}^C \equiv \boldsymbol{\theta}. \tag{13}$$

The constraint in problem (12) ensures that parameter vector $\boldsymbol{\theta}^C$ of cycle C remains unchanged. Hence, after step 4 of the TRW-S(LP-C) algorithm, the reparameterization constraint is satisfied. It was also shown that step 5 (i.e. running TRW-S) provides a reparameterization of $\boldsymbol{\theta}$ (see Lemma 3.3 of (Kolmogorov, 2006) for details). This proves Property 1.

Property 2: *At each step of the algorithm, the value of the dual (11) never decreases.* Clearly, step 4 of the TRW-S(LP-C) algorithm does not decrease the value of the dual (11) (since the objective function of problem (12) is part of the objective function of dual (11)). The work of (Kolmogorov, 2006) showed that step 5 (i.e. TRW-S) also does not decrease this value. Note that the LP-C relaxation is guaranteed to be bounded since it dominates the LP-S relaxation (Kumar et al., 2007), which itself is bounded (Kolmogorov, 2006). Therefore, by the Bolzano-Weierstrass theorem (Fitzpatrick, 2006), it follows that TRW-S(LP-C) will converge.

Property 3: *Like* TRW-S*, the necessary condition for convergence of* TRW-S*(*LP-C*) is that the parameter vectors $\boldsymbol{\theta}^T$ of the trees $T \in \mathcal{T}$ satisfy* WTA. This follows from the fact that TRW-S increases the value of the dual in a finite number of steps as long as the set of parameters $\boldsymbol{\theta}^T$, $T \in \mathcal{T}$, do not satisfy WTA (see (Kolmogorov, 2006) for details).

Property 4: *Unlike* TRW-S*,* WTA *is not the sufficient condition for convergence.* One of the main drawbacks of the TRW-S algorithm is that it converges as soon as the WTA condition is satisfied. Experiments in (Kolmogorov, 2006) indicate that this results in high energy solutions for the MAP estimation problem when the energy function is non-submodular. Using a counter-example, it can be shown that WTA is not the sufficient condition for the convergence of TRW-S(LP-C) (Kumar & Torr, 2008).

Obtaining the Labelling: Similar to the TRW-S algorithm, TRW-S(LP-C) solves the dual (11) and not the primal problem. In other words, it does not directly provide a labelling of the random variables. In order to obtain a labelling, we use the same scheme as the one suggested in (Kolmogorov, 2006) for the TRW-S algorithm. Briefly, we assign labels to the variables $\mathbf{v} = \{v_0, v_1, \cdots, v_{n-1}\}$ in increasing order (i.e. we label variable v_0, followed by variable v_1 and so on). Let $\boldsymbol{\theta}^{\mathcal{T}} = \sum_T \rho(T)\boldsymbol{\theta}^T$. At each stage, a variable v_a is assigned the label $l_{f(a)}$ such that

$$f(a) = \arg\min_{i, l_i \in \mathbf{l}} \left(\theta_{a;i}^{\mathcal{T}1} + \sum_{b<a, (a,b) \in \mathcal{E}} \theta_{ab;i,f(b))}^{\mathcal{T}2} \right), \tag{14}$$

where $\theta_{a;i}^{\mathcal{T}1}$ and $\theta_{ab;i,f(b)}^{\mathcal{T}2}$ are the unary and pairwise potentials corresponding to the parameter $\boldsymbol{\theta}^{\mathcal{T}}$ respec-

tively. It can be shown that under certain conditions the above procedure provides the optimal labelling (Meltzer et al., 2005).

4. Adding SOC Constraints

We now show how second order cone (SOC) constraints can be added to the dual (9). Specifically, we consider the two SOC constraints proposed in (Kumar et al., 2007) which result in the SOCP-C and SOCP-Q relaxations described below.

The SOCP-C Relaxation: Consider a set of random variables $\mathbf{v}_C = \{v_b, b = a_1, \cdots, a_c\} \subseteq \mathbf{v}$ such that $\mathcal{E}_C = \{(a_1, a_2), (a_2, a_3), (a_c, a_1)\} \subseteq \mathcal{E}$ (i.e. $\mathbf{v}_C$ forms a cycle of length c). We define a vector $\mathbf{x}_C$ whose k^{th} element is given by $x_{a_k;i_k}$ and a matrix $\mathbf{X}_C$ whose $(k,m)^{th}$ element is given by $X_{a_k a_m;i_k i_m}$ (where $l_{i_k}, l_{i_m} \in \mathbf{l}$). SOCP-C specifies constraints $||\mathbf{U}^\top \mathbf{x}_C||^2 \leq \mathbf{C} \bullet \mathbf{X}_C$ where $\mathbf{C} = \mathbf{D}_c + \lambda_c I = \mathbf{U}\mathbf{U}^\top$ and $(\bullet)$ represents the Frobenius inner product. The $c \times c$ matrix $\mathbf{D}_c$ is given by

$$D_c(i,j) = \begin{cases} (-1)^{c-1} & \text{if} \quad |i-j| = c-1, \\ 1 & \text{if} \quad |i-j| = 1, \\ 0 & \text{otherwise}, \end{cases} \tag{15}$$

and λ_c is the absolute value of the smallest eigenvalue of $\mathbf{D}_c$.

The SOCP-Q Relaxation: Consider a set of random variables $\mathbf{v}_C = \{v_b, b = a_1, \cdots, a_c\} \subseteq \mathbf{v}$ such that $\mathcal{E}_C = \{(a_i, a_j), i, j = 1, \cdots, c\} \subseteq \mathcal{E}$ (i.e. $\mathbf{v}_C$ form a clique of size c). SOCP-Q specifies constraints of the form $||\mathbf{U}^\top \mathbf{x}_C||^2 \leq \mathbf{C} \bullet \mathbf{X}_C$ where $\mathbf{C}$ is a matrix whose elements are all 1.

In general, a set of N_C SOC constraints on a cycle/clique can be defined as

$$||\mathbf{A}_k^C \mathbf{y} + \mathbf{b}_k^C|| \leq \mathbf{y}^\top \mathbf{c}_k^C + d_k^C, k \in \{1, 2, \cdots, N_C\}. \tag{16}$$

Let $\mathcal{C}$ be a set of cycles/cliques in the graphical model of the given random field. The following theorem provides us with the dual of the SOCP relaxation obtained by appending problem (8) with SOC constraints defined over the set $\mathcal{C}$.

Theorem 2: The following problem is the dual of problem (8) appended with a set of SOC constraints $||\mathbf{A}_k^C \mathbf{y} + \mathbf{b}_k^C|| \leq \mathbf{y}^\top \mathbf{c}_k^C + d_k^C$ for $k \in \{1, 2, \cdots, N_C\}$ and $C \in \mathcal{C}$.

$$\begin{aligned} \max \quad & \textstyle\sum_T \rho(T) q(\boldsymbol{\theta}^T) - \sum_C \rho'(C) \sum_k p_k^C, \\ \text{s.t.} \quad & \textstyle\sum_T \rho(T) \boldsymbol{\theta}^T + \sum_C \rho'(C) \sum_k q_k^C \equiv \boldsymbol{\theta}, \\ & ||\mathbf{u}_k^C|| \leq v_k^C, \forall k \in \{1, 2, \cdots, N_C\}, C \in \mathcal{C}. \end{aligned} \tag{17}$$

where

$$p_k^C = (\mathbf{b}_k^C)^\top \mathbf{u}_k^C + d_k^C v_k^C, \tag{18}$$

$$q_k^C = (\mathbf{A}_k^C)^\top \mathbf{u}_k^C + \mathbf{c}_k^C v_k^C. \tag{19}$$

Here $\mathbf{u}_k^C$ and v_k^C are some slack variables.

We can define up to h^c SOC constraints for a cycle/clique, where c is the size of the cycle/clique (i.e. $N_C \in \{0, 1, \cdots, h^c\}$). Before proceeding further, we also define the following problem:

$$\begin{aligned} \max \quad & \textstyle\sum_{T \in C} \rho(T) q(\boldsymbol{\theta}^T) - \rho'(C) \sum_k p_k^C, \\ \text{s.t.} \quad & \textstyle\sum_{T \in C} \rho(T) \boldsymbol{\theta}^T + \rho'(C) \sum_k q_k^C = \boldsymbol{\theta}^C, \\ & ||\mathbf{u}_k^C|| \leq v_k^C, \forall k \in \{1, 2, \cdots, N_C\}, \end{aligned} \tag{20}$$

where $\boldsymbol{\theta}^C$ is some parameter vector. The variables of the above problem are restricted to $\mathbf{u}_k^C$, v_k^C, $\theta_{a;i}^{T1}$ and $\theta_{ab;ij}^{T2}$ where $(a,b) \in \mathcal{E}_T \bigcap \mathcal{E}_C$. Like problem (12), we can solve problem (20) using standard Interior Point algorithms for small cycles/cliques C.

Similar to TRW-S(LP-C), a convergent algorithm can now be described for solving the dual (17). This algorithm differs from TRW-S(LP-C) in only step 4, where it solves problem (20) for a cycle/clique C instead of problem (12). We refer to this algorithm as either TRW-S(SOCP-C) or TRW-S(SOCP-Q) depending upon the SOCP relaxation that we are solving. When using the TRW-S(SOCP-Q) algorithm, we include all slack variables corresponding to the cycle inequalities defined over the cycles in clique C. It can easily be shown that both TRW-S(SOCP-C) and TRW-S(SOCP-Q) satisfy all the properties given in § 3.1. Note that, like TRW-S and TRW-S(LP-C), these algorithms do not directly provide a labelling for the random variables of the CRF. Instead we use the procedure described in § 3.1 to obtain the final solution.

5. Experiments

We tested the approaches described in this paper using both synthetic and real data. For synthetic data experiments, we closely follow the setup of (Kolmogorov, 2006). We show that our algorithms overcome a well-known deficiency of TRW-S, namely that it does not provide good MAP estimates for non-submodular energy functions. Next, we consider the problem of segmentation using real data and show favourable comparison between our methods and several other standard MAP estimation techniques.

5.1. Synthetic Data

Datasets: We conducted two sets of experiments using binary grid CRFs (i.e. $h = |\mathbf{l}| = 2$) of size 30×30. In the first experiment the edges of the graphical model, i.e. $\mathcal{E}$, were defined using a 4-neighbourhood system while the second experiment used an 8-neighbourhood system. Similar to (Kolmogorov, 2006), the unary potentials $\theta_{a;0}^1$ and $\theta_{a;1}^1$ were generated using the normal distribution $\mathcal{N}(0,1)$. The pairwise potentials $\theta_{ab;00}^2$ and

$\theta^2_{ab;11}$ were set to 0 while $\theta^2_{ab;01}$ and $\theta^2_{ab;10}$ were generated using $\mathcal{N}(0, \sigma^2)$. For both experiments, 50 CRFs were generated using the method described above. All the CRFs defined non-submodular energy functions (i.e. there exists an $(a, b) \in \mathcal{E}$ such that $\theta_{ab;01} + \theta_{ab;10} < 0$) which are in general NP-hard to minimize. As noted in (Kolmogorov, 2006), TRW-S performs considerably worse than BP on such examples.

Implementation Details: We tested the LP-C and the SOCP-C relaxations in the first experiment. Constraints were defined on all cycles of size 4. The LP-C and SOCP-Q relaxation were tested in the second experiment. Cycles inequalities were defined on all cycles of size 3. In addition, for SOCP-Q, SOC constraints were defined on all cliques of size 4. In both the experiments, our algorithms were tested using trees defined by individual edges of the graphical model for ease of implementation. In other words, a tree $T = (\mathbf{v}_T, \mathcal{E}_T) \in \mathcal{T}$ such that $\mathbf{v}_T = \{v_a, v_b\}$ and $\mathcal{E}_T = \{(a, b)\} \subseteq \mathcal{E}$. However, we note here that our algorithms are general and can be applied for any choice of trees. Although our current set of trees are quite restrictive, the results show that they outperform several state of the art algorithms. The TRW-S algorithm, as well as other standard approaches, was tested using the publically available code which uses monotonic chains as trees.

The terms $\rho(T)$ and $\rho'(C)$ were set to $1/|\mathcal{T}|$ and $1/|\mathcal{C}|$ respectively for all $T \in \mathcal{T}$ and $C \in \mathcal{C}$. We found it sufficient to define one cycle inequality per cycle C using a set of labels $\{l_{i_1}, l_{i_2}, \cdots, l_{i_c}\}$ which satisfies

$$\sum_{(a_k,a_m)\in\mathcal{E}_F} \theta_{a_k a_m;i_k i_m} - \sum_{(a_k,a_m)\in\mathcal{E}_C-\mathcal{E}_F} \theta^2_{a_k a_m;i_k i_m} \geq \sum_{(a_k,a_m)\in\mathcal{E}_F} \theta_{a_k a_m;j_k j_m} - \sum_{(a_k,a_m)\in\mathcal{E}_C-\mathcal{E}_F} \theta^2_{a_k a_m;j_k j_m},$$

for all sets of labels $\{l_{j_1}, \cdots, l_{j_c}\}$. Here $\mathcal{E}_C = \{(a_1, a_2), \cdots, (a_n, a_1)\}$ and $\mathcal{E}_F \subseteq \mathcal{E}_C$ such that $|\mathcal{E}_F| = 3$. As proposed in (Kumar et al., 2007), we also define only one SOC constraint per cycle/clique when considering the SOCP-C and the SOCP-Q relaxations. At each iteration, problems (12) and (20) were solved using the MOSEK software (available at http://www.mosek.com).

Results: Figure 1 (a) shows the results obtained for the first experiment using $\sigma = \frac{10}{\sqrt{d}}$ (where d is the degree of the variables in the graphical model). Note that since the energy functions are non-submodular, TRW-S provides labellings with higher energies than BP as observed in (Kolmogorov, 2006). However, the additional constraints in the LP-C and SOCP-C algorithm enable us to obtain labelling with lower energies than BP. Further, unlike BP, they also provide us with the value of the dual at each iteration. This value allows us to find out how close we are to the global optimum (since the energy of the optimal labelling cannot be less than the value of the dual). Also note that the value of the LP-C dual is greater than the value of the SOCP-C dual. This provides empirical evidence that LP-C dominates SOCP-C as conjectured in (Kumar et al., 2007).

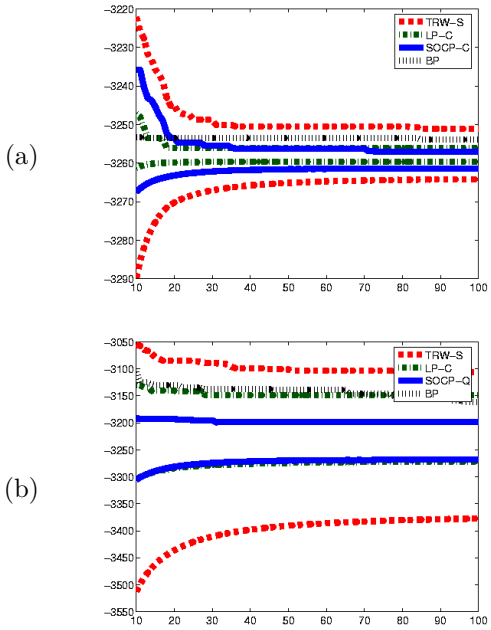

Figure 1. Results of the synthetic data experiment. **(a)** First experiment. The x-axis shows the iteration number. The lower curves show the average value of the dual at each iteration over 50 random CRFs while the upper curves show the average energy of the best labelling found till that iteration. The additional constraints in the LP-C and SOCP-C relaxations enable us to obtain labellings with lower energy compared to TRW-S and BP. Cycle inequalities provide a better approximation than the SOC constraint of the SOCP-C relaxation. **(b)** Second experiment. Note that the value of the dual obtained using SOCP-Q is greater than the value of the dual of the LP-C relaxation.

The results of the second experiment are shown in Figure 1 (b) using $\sigma = \frac{10}{\sqrt{d}}$. Again, BP outperforms TRW-S, while LP-C and SOCP-Q provide better approximations. The SOC constraints defined over cliques in SOCP-Q provide a greater value of the dual compared to the LP-C relaxation. The complexity and timings for all the algorithms are given in tables 3 and 4.

5.2. Real Data - Segmentation

We now present the results of our method on interactive segmentation (Boykov & Jolly, 2001) where, given some seed pixels for all the segments present in an image, we wish to obtain the segmentation of the image.

Problem Formulation: The problem of obtaining the segmentation of an image can be cast within the CRF framework. Specifically, we define a CRF over random variables $\mathbf{v} = \{v_0, \cdots, v_{n-1}\}$, where each variable

Algorithm	No. of Var.	No. of Cons.	Time(sec)
BP	-	-	0.0018
TRW-S	$nh + \|\mathcal{E}\|h^2$	$n + 2\|\mathcal{E}\|h$	0.0018
LP-C	$nh + \|\mathcal{E}\|h^2$	$2n + 2\|\mathcal{E}\|h$	7.5222
SOCP-C	$nh + \|\mathcal{E}\|h^2$	$2n + 2\|\mathcal{E}\|h$	8.9091

Table 3. Complexity and timings of the algorithms for the first synthetic data experiment with a 4-neighbourhood relationship. Recall that $n = |\mathbf{v}|$ is the number of random variables, $h = |\mathbf{l}|$ is the size of the label set and $\mathcal{E}$ is the neighbourhood relationship defined by the CRF. The second and third columns show the number of variables and constraints in the primal problem respectively. The fourth column shows the average time of the each algorithm for one iteration (in seconds). All timings are reported for a Pentium IV 3.3 GHz processor with 2GB RAM.

Algorithm	No. of Var.	No. of Cons.	Time(sec)
BP	-	-	0.0027
TRW-S	$nh + \|\mathcal{E}\|h^2$	$n + 2\|\mathcal{E}\|h$	0.0027
LP-C	$nh + \|\mathcal{E}\|h^2$	$5n + 2\|\mathcal{E}\|h$	7.7778
SOCP-Q	$nh + \|\mathcal{E}\|h^2$	$6n + 2\|\mathcal{E}\|h$	9.1170

Table 4. Complexity and timings for the second synthetic data experiment with an 8-neighbourhood relationship. Note that SOCP-Q includes all the constraints of LP-C.

corresponds to a pixel of the frame. Each label in the set $\mathbf{l} = \{l_0, \cdots, l_{h-1}\}$ corresponds to a segment (where h is the total number of segments). The unary potential of assigning a variable v_a to segment l_i is specified by the negative log-likelihood of the RGB value of pixel a given the seed pixels of the segment l_i. The pairwise potentials encourage continuous segments whose boundaries lie on image edges. For more details, we refer the reader to (Boykov & Jolly, 2001). The problem of obtaining the segmentation of a frame then boils down to that of finding the MAP estimate of the CRF.

Datasets and Implementation Details: We used the well-known 'Garden' sequence to conduct our experiments (with frame size 120×175). The seed pixels were provided using the ground truth segmentation of a keyframe as shown in Fig. 2.

Similar to the synthetic data experiment, we defined the trees as individual edges of the graphical model of the CRF for our algorithms. Other algorithms were tested using publically available code (including TRW-S which uses monotonic chains as trees). We specified one cycle inequality and one SOC constraint for each cycle/clique (as described in the previous section). The terms $\rho(T)$ and $\rho'(C)$ were set to $1/|\mathcal{T}|$ and $1/|\mathcal{C}|$ respectively for all $T \in \mathcal{T}$ and $C \in \mathcal{C}$. Once again, problems (12) and (20) were solved using MOSEK.

Results: For the first set of experiments, we used a 4-neighbourhood system and tested the following algorithms: TRW-S, LP-C, SOCP-C, $\alpha\beta$-swap, α-expansion and BP. Fig. 3 shows the segmentations (of frames

Figure 2. Segmented keyframe of the 'Garden' sequence. The left image shows the keyframe while the right image shows the corresponding segmentation provided by the user. The four different colours indicate pixels belonging to the four segments namely sky, house, garden and tree.

Algorithm	Avg. Time-1 (s)	Avg. Time-2 (s)
BP	0.1400	0.1740
TRW-S	0.1400	0.1740
$\alpha\beta$-swap	0.1052	0.1201
α-expansion	0.1100	0.1240
LP-C	140.3320	142.2226
SOCP-C/SOCP-Q	143.6365	144.9890

Table 5. Average timings of the algorithms (per iteration) for the first experiment on video segmentation with a 4-neighbourhood relationship (column 2) and the second experiment with an 8-neighbourhood relationship (column 3). Again, all timings are reported for a Pentium IV 3.3 GHz processor with 2GB RAM.

other than the keyframe) and the values of the energy function obtained for all algorithms. Note that, by incorporating additional constraints using all cycles of length 4, LP-C and SOCP-C outperform other methods. Further, the cycle inequalities in LP-C provide better results than the SOC constraints of SOCP-C. Table 5 provides the average time for all algorithms.

The second set of experiments used an 8-neighbourhood system and tested the following algorithms: TRW-S, LP-C, SOCP-Q, $\alpha\beta$-swap, α-expansion and BP. For the LP-C algorithm, cycle inequalities were specified for all cycles of size 3. In addition, the SOCP-Q algorithm specifies SOC constraints on all cliques of size 4. Fig. 4 shows the segmentations and energies obtained for all the algorithms. The average timings per iteration are shown in table 5. Note that, similar to the synthetic data examples, SOCP-Q outperforms LP-C by incorporating additional SOC constraints.

6. Discussion

We extended the LP-S relaxation based approach of (Kolmogorov, 2006; Wainwright et al., 2005) for the MAP estimation problem. Specifically, we showed how cycle inequalities and SOC constraints can be incorporated within the TRW framework. We also proposed convergent algorithms for solving the resulting duals. Our experiments indicate that these additional constraints provide a more accurate approximation for MAP estimation when the energy function is non-submodular. Although our algorithm is much faster than Interior Point methods, it is slower than TRW-S and BP. An interesting direction for future re-

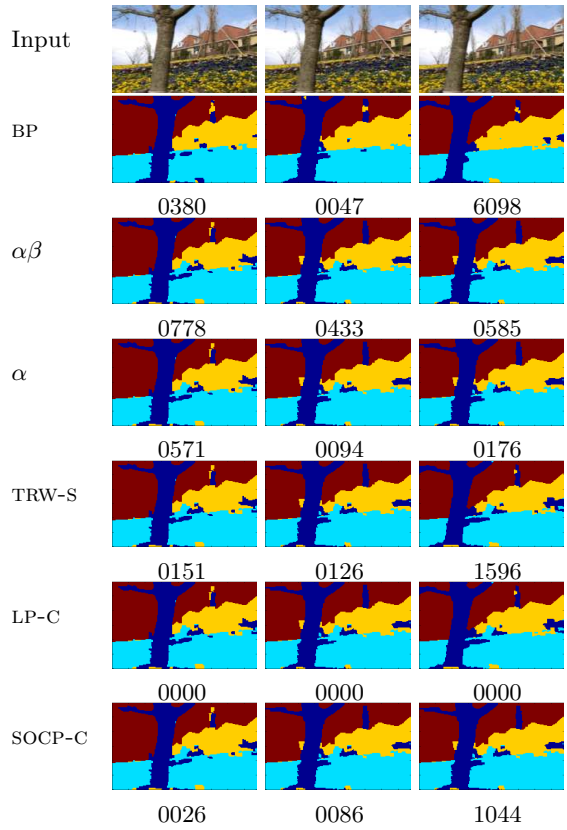

Figure 3. Segmentations obtained for the 'Garden' video sequence using 4-neighbourhood. The corresponding energy values (scaled up to integers for using $\alpha\beta$-swap and α-expansion) of all the algorithms are shown below the segmentation. The following constant terms are subtracted from the energy values of all algorithms for the three frames respectively (to make minimum energy among all algorithms 0): 5139499, 5145234 and 5126941.

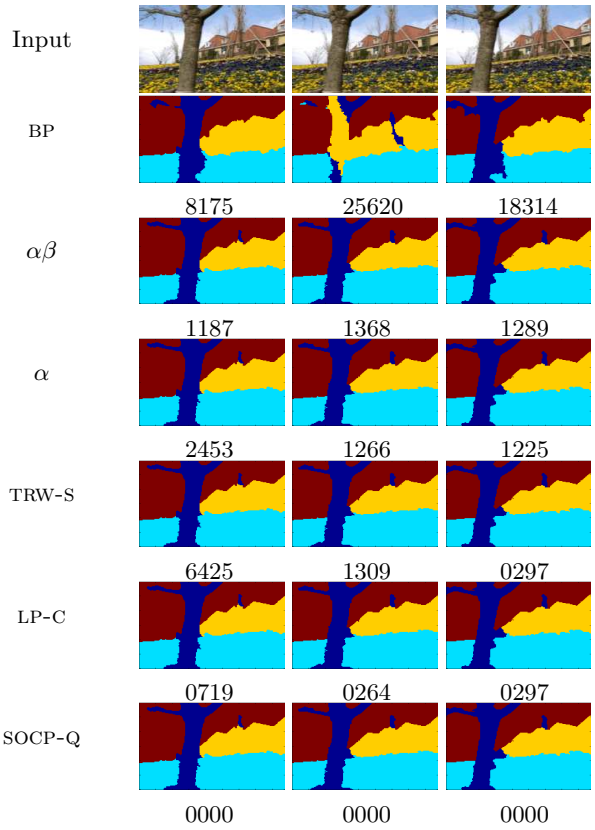

Figure 4. Segmentations obtained for the 'Garden' video sequence using 8-neighbourhood. The corresponding energy values (reduced by 5304466, 5299756 and 5292224 for the three frames respectively) are also shown.

search would be to develop specialized algorithms for solving problems (12) and (20) (which are used in our approach).

References

Barahona, F., & Mahjoub, A. (1986). On the cut polytope. *Mathematical Programming*, *36*, 157–173.

Boykov, Y., & Jolly, M. (2001). Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images. *ICCV* (pp. I: 105–112).

Fitzpatrick, P. (2006). *Advanced calculus*. Thompson Brooks/Cole.

Goemans, M., & Williamson, D. (1995). Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of ACM*, *42*, 1115–1145.

Karzanov, A. (1998). Minimum 0-extension of graph metrics. *European Journal of Combinatorics*, *19*, 71–101.

Kolmogorov, V. (2006). Convergent tree-reweighted message passing for energy minimization. *PAMI*, *28*, 1568–1583.

Kumar, M. P., Kolmogorov, V., & Torr, P. H. S. (2007). An analysis of convex relaxations for MAP estimation. *NIPS*.

Kumar, M. P., & Torr, P. H. S. (2008). *Efficiently solving convex relaxations for MAP estimation* (Technical Report). Oxford Brookes University.

Meltzer, T., Yanover, C., & Weiss, Y. (2005). Globally optimal solutions for energy minimization in stereo vision using reweighted belief propagation. *ICCV*.

Schlesinger, M. (1976). Sintaksicheskiy analiz dvumernykh zritelnikh singnalov v usloviyakh pomekh. *Kibernetika*, *4*, 113–130.

Sontag, D., & Jaakkola (2007). New outer bounds on the marginal polytope. *NIPS*.

Szeliski, R., Zabih, R., Scharstein, D., Veksler, O., Kolmogorov, V., Agarwala, A., Tappen, M., & Rother, C. (2006). A comparative study of energy minimization methods for markov random fields. *ECCV* (pp. II: 16–29).

Wainwright, M., Jaakola, T., & Willsky, A. (2005). MAP estimation via agreement on trees: Message passing and linear programming. *IEEE Trans. on Information Theory*, *51*, 3697–3717.

Wiegernick, W. (2005). Approximations with reweighted generalized belief propagation. *AISTATS*.

Zwick, U. (1999). Outward rotations: A tool for rounding solutions of semidefinite relaxations, with applications to MAX CUT and other problems. *STOC* (pp. 679–687).