International Journal of Approximate Reasoning 54 (2013) 1279–1299

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Energy distribution view for monotonic dual decomposition

Qiang Cheng, Feng Chen *, Jianwu Dong, Wenli Xu

Tsinghua National Laboratory for Information Science and Technology, Department of Automation, Tsinghua University, Beijing 100084, PR China

ARTICLE INFO

Article history:
Received 19 October 2012
Received in revised form 10 July 2013
Accepted 25 July 2013
Available online 8 August 2013

Keywords:
Graphical models
MAP inference
Dual decomposition
Energy distribution ratios

ABSTRACT

We consider the problem of finding the most probable explanation (also known as the MAP assignment) on probabilistic graphical models. The dual decomposition algorithms based on coordinate descent are efficient approximate techniques for this problem, where the local dual functions are constructed and optimized to monotonically increase the cost of the dual function. In this paper, we present a unifying framework for constructing and optimizing these local dual functions, and introduce an energy distribution view to analyze the convergence rates of these algorithms. To optimize the local dual functions, we first propose a new concept—the energy distribution ratio—to describe the features of the solutions, and then derive an explicit optimal solution, which covers most of the monotonic dual decomposition algorithms. It is shown that the differences of these algorithms lie in both the forms of the local dual functions and the settings of the energy distribution ratios, and the existing algorithms mainly focus on constructing compact and solvable local dual functions. In contrast, we study the impact of the energy distribution ratios and introduce two energy distribution criteria for fast convergence. Moreover, we exploit dynamic energy distribution ratios to optimize the local dual functions, and propose a series of improved algorithms. The experimental results on synthetic and real problems show the improved algorithms outperform the existing ones on the convergence performance.

1. Introduction

Probabilistic graphical models, such as Markov random fields (MRFs) and Bayesian networks, have been widely studied and used as powerful tools in the field of artificial intelligence [1,2], computer vision [3], bioinformatics [4], signal processing [5], and many others. The problems in these applications are often formed as the task of finding the most probable explanation, also known as the maximum a posteriori (MAP) assignment. Unfortunately, the MAP problem is widely known to be NP-hard [6,7].

The linear programming (LP) relaxation is a powerful technique for solving the MAP problem [8]. The LP relaxation is often solved from the dual and reduces to the dual decomposition technique. The dual decomposition framework includes two designing choices: how to decompose the original problem and how to optimize the dual function. For the first designing choice, the original graph can be decomposed into different types of subgraphs over which the MRF energy can be efficiently minimized, such as trees [9–11], outer-planar graphs [12], planar graphs with pure interaction potentials [13], binary planar graphs [14], submodular graphs [15], k-fans [16], triplets or other solvable clusters [17,18], and so on. For the second designing choice, the dual function is a nonsmooth convex function, and can be optimized by nondifferentiable optimization methods, such as the block coordinate descent method [9,10,19], the subgradient method [15,12,20],

* Corresponding author. Tel.: +861062797145; fax: +861062786911.
E-mail address: chenfeng@mail.tsinghua.edu.cn (F. Chen).

0888-613X/$ – see front matter
http://dx.doi.org/10.1016/j.ijar.2013.07.008

the alternating direction method of multipliers (ADMM) [21,22], the bundle method [23], the primal method [24], the convex max-product algorithm [25], and the cutting plane method [26].

The block coordinate descent method is an efficient and successful one for optimizing the nonsmooth dual function, which is shown to converge faster than other methods [21,22]. The basic idea of the block coordinate descent method is first selecting a set of subproblems to construct a local dual function, and then maximizing the local dual function to increase the cost of the dual function. There are many monotonic algorithms based on the block coordinate descent method: the max-sum diffusion (MSD) algorithm [9], the max product linear programming (MPLP) algorithm [10], the tree block coordinate descent (TBCD) algorithm [27,28], the sequential tree-reweighted message passing algorithm (TRW-S) [19], the generalized MPLP (GMPLP) algorithm [10,26], the n-ary max-sum diffusion (GMSD) algorithm [29], the star update (Star) algorithm [8] and the MSD++ algorithm [8]. Meltzer et al. [30] propose a unifying view for convergent message passing algorithms based on the tree-consistency bound optimization (TCBO), where they state that the local dual function can be optimized by enforcing tree consistency between the subproblems. However, they do not give specific solutions on how to enforce tree consistency. Sontag and Jaakkola [27] propose an algorithm to optimize the local dual function defined on tree graphs with uniform energy distribution ratios (introduced later in this paper). However, the solutions for optimizing the local dual function defined on hypergraphs with other *valid* energy distribution ratios are not introduced in their work. Moreover, all these methods focus on constructing different and compact local dual functions, but pay little attention to other factors (such as the energy distribution ratios) that influence their performance.

In this paper, we propose a unifying framework for monotonic dual decomposition algorithms from the perspective of partially ordered set (poset). The poset has been proposed to formulate the general belief propagation (GBP) algorithms on hypergraphs [31–33]. Here, the poset is exploited to formulate the dual decomposition algorithms for MAP inference, and to help construct and optimize the local dual functions. Within our framework, the set of decomposed subgraphs forms a poset, and the local dual function is constructed by selecting some subgraphs from the poset that form a hypertree. To optimize the local dual function, we introduce a new concept—the energy distribution ratio, which describes the major features of the optimal solutions. We then derive an explicit and unifying solution to optimize the local dual function by using the energy distribution ratios, and show that most of the existing monotonic algorithms are covered by our solution.

It is observed that the main differences of the monotonic algorithms lie in the local dual functions and the energy distribution ratios (i.e., the choice of block size and equalization method, which are first proposed in [8, Section 1.5.6]). Previous work mainly focuses on constructing compact and solvable local dual functions. For example, all the above mentioned algorithms (except for MSD++ [8]) differ in the forms of the local dual functions. In contrast to the previous work, we focus on investigating the effect of the energy distribution ratios on these algorithms. Two general energy distribution criteria for fast convergence are introduced, which explain why some algorithms converge faster than others. Moreover, rather than fixing the energy distribution ratios for all local dual functions, we propose to dynamically set the energy distribution ratios, where the energy distribution ratios are set to take different values for different local dual functions. A series of improved algorithms based on dynamic energy distribution ratios are proposed and experimental results on different models show that great improvement can be achieved by the proposed algorithms.

The paper is organized as follows. Section 2 sets up the backgrounds and definitions. In Section 3, we present a unifying framework and an explicit solution based on poset and show that the existing monotonic message passing algorithms are instances of our framework. We introduce two energy distribution criteria to speed up the convergence in Section 4, and propose new algorithms with dynamic energy distribution ratios in Section 5. We show empirical results in Section 6, and conclude this paper in Section 7.

2. Preliminaries

We introduce some necessary concepts in this section. We denote by $G = (\mathcal{V}, \mathcal{E})$ an undirected graph, with nodes set $\mathcal{V} = \{1, 2, \ldots, N\}$ and edges set $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$. Each node $s \in \mathcal{V}$ is associated with a random variable X_s, which takes values in the space $\mathcal{X}_s$. Let $\mathbf{X}$ be the vector of all variables, i.e., $\mathbf{X} = \{X_s, \forall s \in \mathcal{V}\}$, with $\mathbf{x} = \{x_s, \forall s \in \mathcal{V}\}$ corresponding to the assignment of $\mathbf{X}$. Define a unary potential function $\theta_s(\cdot) : \mathcal{X}_s \to \mathbb{R}$ for each node s and a pairwise potential function $\theta_{st}(\cdot,\cdot) : \mathcal{X}_s \times \mathcal{X}_t \to \mathbb{R}$ for each edge st. Then, a pairwise Markov random field (MRF) is defined by the graph G and the potential functions. The task of MAP inference is to minimize the following MRF energy $E(\boldsymbol{\theta}, \mathbf{x})$ w.r.t. $\mathbf{x} \in \mathcal{X}$:

$$E(\boldsymbol{\theta}, \mathbf{x}) = \sum_{s \in \mathcal{V}} \boldsymbol{\theta}_s(x_s) + \sum_{(s,t) \in \mathcal{E}} \boldsymbol{\theta}_{st}(x_s, x_t). \tag{1}$$

The parameter vector of a pairwise MRF can be denoted by $\boldsymbol{\theta} = \{\{\boldsymbol{\theta}_s\}_{s \in \mathcal{V}}, \{\boldsymbol{\theta}_{st}\}_{(s,t) \in \mathcal{E}}\}$, where $\boldsymbol{\theta}_s = \{\theta_s(x_s), \forall x_s \in \mathcal{X}_s\}$ and $\boldsymbol{\theta}_{st} = \{\theta_{st}(x_s, x_t), \forall x_s \in \mathcal{X}_s, x_t \in \mathcal{X}_t\}$.

For tree-structured graphs, the MRF energy $E(\boldsymbol{\theta}, \mathbf{x})$ can be minimized using dynamic programming methods, with computational complexity exponential in the treewidth [19]. For graphs with high treewidth, the energy minimization is often formulated as a linear programming (LP) relaxation problem and then solved from the dual, which reduces to the dual decomposition technique [26].