Dual Decomposition for Marginal Inference

Justin Domke
Rochester Institute of Technology
Rochester, NY 14623

Abstract

We present a dual decomposition approach to the tree-reweighted belief propagation objective. Each tree in the tree-reweighted bound yields one subproblem, which can be solved with the sum-product algorithm. The master problem is a simple differentiable optimization, to which a standard optimization method can be applied. Experimental results on 10x10 Ising models show the dual decomposition approach using L-BFGS is similar in settings where message-passing converges quickly, and one to two orders of magnitude faster in settings where message-passing requires many iterations, specifically high accuracy convergence, and strong interactions.

Introduction

Marginal inference in probabilistic graphical models is used in many applications. In the high-treewidth setting, intractability forces the use of approximate algorithms. Loopy belief propagation(Yedidia, Freeman, and Weiss 2005) often gives good approximations, but can suffer from local minima or non-convergence. In one line of work, several algorithms were proposed guaranteeing convergence (to a local minima) on the loopy belief propagation objective(Teh and Welling 2001; Yuille 2002; Heskes, Albers, and Kappen 2003), though often with a tradeoff in speed.

Another line of work developed convex variants(Globerson and Jaakkola 2007a; Meshi et al. 2009) that do not suffer from local optima. Here, we consider the tree-reweighted (TRW) objective(Wainwright, Jaakkola, and Willsky 2005a). The original TRW message passing algorithm does not always converge, though appropriate "damping" of updates appears to accomplish this in practice. Several algorithms have been proposed that do provably converge on this or similar objectives(Hazan and Shashua 2009; Globerson and Jaakkola 2007b; Meltzer, Globerson, and Weiss 2009). These are not generally claimed to be faster than TRW when it does converge. A recent exception is the marginal inference version of the TRW-S algorithm, which is essentially TRW run with appropriate message update orders on a graph consisting of monotonic chains(Meltzer, Globerson, and Weiss 2009).

A practical limitation of message-passing algorithms is that even when they do converge, they may do so very slowly, particularly on models with strong interactions and when high accuracy is required (Experiments Section).

Dual decomposition is a well-established idea in optimization, whereby an objective function that is a sum of functions over subsets of variables is "decoupled" into independent subproblems by introducing Lagrange multipliers. These multipliers are then adjusted in the "master" problem to assure that the solutions of all subproblems agree on common variables. Dual decomposition has proven useful for MAP inference (Komodakis, Paragios, and Tziritas 2007; Wainwright, Jaakkola, and Willsky 2005b; Kolmogorov 2006), where a linear-programming relaxation of a high-treewidth problem is decoupled into subproblems over trees. Because the master problem is non-differentiable, care must be taken in optimization to assure convergence (e.g. using subgradient methods).

This paper observes that it is relatively straightforward to apply dual decomposition to the TRW objective. This results in a simple algorithm where the traditional sum-product algorithm is called on each tree in the TRW bound as subproblems. Unlike for MAP inference, the master problem is differentiable, and so standard faster-converging optimization methods (e.g. quasi-Newton) can be used.

We present experimental results on 10x10 Ising grids, using L-BFGS to optimize the master problem. Roughly, our results show that when message passing methods converge quickly (e.g. weak coupling) dual decomposition performs similarly, or slightly slower. However, when message-passing methods require many iterations (strong coupling, convergence to high accuracy), dual decomposition is one to two orders of magnitude faster. We also find that the number of iterations required by dual decomposition is more concentrated than for message passing methods.

A limitation of the proposed method is the need to run the sum-product algorithm on each constituent tree in each iteration. Thus, the approach is most attractive when each factor participates in few trees.

Inference

Undirected models can be seen as members of the exponential family

$$p(\mathbf{x};\boldsymbol{\theta}) = \exp\big(\mathbf{f}(\mathbf{x})\cdot\boldsymbol{\theta} - A(\boldsymbol{\theta})\big)$$
$$A(\boldsymbol{\theta}) = \log\sum_{\mathbf{x}}\exp\big(\mathbf{f}(\mathbf{x})\cdot\boldsymbol{\theta}\big),$$

where the vector of sufficient statistics

$$\mathbf{f}(\mathbf{X}=\mathbf{x}) = \{I(\mathbf{X}_\alpha = \mathbf{x}_\alpha)\}\cup\{I(X_i = x_i)\} \quad (1)$$

is the set of all indicator functions on all factors α and variables i. Corresponding to this is a bipartite graph with one node for each factor and variable, and an edge between α and i if and only if $i\in\alpha$.

Marginal inference means recovering the expected value of the sufficient statistics

$$\mathbf{b} = \sum_{\mathbf{x}} p(\mathbf{x};\boldsymbol{\theta})\mathbf{f}(\mathbf{x}), \quad (2)$$

which, when f is as in Eq. 1, is equivalent to finding the marginal distributions for all factors and variables. The sum over all possible vectors $\mathbf{x}$ in Eq. 2 is impractical when $\mathbf{x}$ is not low-dimensional. The sum-product algorithm provides a solution in treelike graphs. In general, approximations are necessary. This motivates the variational characterization(Wainwright and Jordan 2008)

$$A(\boldsymbol{\theta}) = \max_{b\in\mathcal{M}} \boldsymbol{\theta}\cdot\mathbf{b} + H(\mathbf{b}), \quad (3)$$

where the marginal polytope

$$\mathcal{M} = \{\mathbf{b} : \exists\boldsymbol{\theta}, \mathbf{b} = E_{p(\boldsymbol{\theta})}[\mathbf{f}(\mathbf{X})]\}$$

is the set of achievable marginals, and $H(\mathbf{b}) = -\sum_{\mathbf{x}} p(\mathbf{x})\log p(\mathbf{x})$ is the entropy of the distribution p that produces the marginals $\mathbf{b}$.

The tree-reweighted (TRW)(Wainwright, Jaakkola, and Willsky 2005a) bound on A is based on applying two relaxations to Eq. 3, both of which are upper bounds. The first is that H can be upper-bounded by the *projection* onto a tree-graph T. Define H_T to be the entropy corresponding to the tree graph T, and $\mathbf{b}(T)$ to be only those marginals corresponding to T, then we have the tractable bound(Wainwright and Jordan 2008, Prop. 7.1)

$$H(\mathbf{b}) \le H_T(\mathbf{b}(T)).$$

Accordingly, given a distribution ρ_T over a set of trees T, we have

$$H(\mathbf{b}) \le \sum_T \rho_T H(\mathbf{b}(T)).$$

Secondly, the marginal polytope $\mathcal{M}$ is difficult to characterize in general, and so is replaced with the *local polytope*

$$\mathcal{L} = \{\mathbf{b} : \mathbf{b}(T)\in\mathcal{M}_T\ \forall T\},$$

where $\mathcal{M}_T$ is the marginal polytope for tree T. Since $\mathcal{L}\supset\mathcal{M}$, and maximizing over a larger set can only increase the optimum, this also upper bounds A.

Applying both these bounds, we have

$$A(\boldsymbol{\theta}) \le B(\boldsymbol{\theta}) = \max_{\mathbf{b}\in\mathcal{L}} \boldsymbol{\theta}\cdot\mathbf{b} + \sum_T \rho_T H(\mathbf{b}(T)) \quad (4)$$

We are interested in performing the optimization necessary to compute $B(\boldsymbol{\theta})$, as well as the associated marginals.

It is possible to compute $B(\boldsymbol{\theta})$ using a standard constrained optimization approach, such as an interior point algorithm. Done conventionally, this scales poorly (albeit polynomially) to large high-treewidth graphs due to the constraint that the beliefs lie in the local polytope. One can also take the dual of Eq. 4, resulting in an unconstrained problem that can be addressed with methods that scale linearly such as conjugate gradients or L-BFGS. However, this has not been proven faster than message-passing methods in practice(Globerson and Jaakkola 2007b).

Dual Decomposition

The basic idea of dual decomposition is to take an optimization problem of the form

$$\max \sum_i f_i(\mathbf{x})$$

and "decouple" the functions f_i by transforming it into the equivalent constrained problem

$$\max_{\mathbf{x}} \sum_i f_i(\mathbf{x}_i) \text{ s.t. } \forall i,j\ \mathbf{x}_i = \mathbf{x}_j$$

The constraint that the different $\mathbf{x}_i$ are equal can be enforced in various ways. For technical reasons, it is convenient in this paper to enforce that each $\mathbf{x}_i$ is equal to the mean of all $\mathbf{x}_j$, producing the problem

$$\max_{\mathbf{x}} \sum_i f_i(\mathbf{x}_i) \text{ s.t. } \mathbf{x}_i = \frac{1}{N}\sum_j \mathbf{x}_j.$$

This has the Lagrangian

$$\mathbb{L} = \sum_i \big(f_i(\mathbf{x}_i) + \boldsymbol{\lambda}_i\cdot(\mathbf{x}_i - \frac{1}{N}\sum_j \mathbf{x}_j)\big),$$

and so we can solve the optimization via the minimax problem

$$\min_{\{\boldsymbol{\lambda}_i\}}\max_{\{\mathbf{x}_i\}} \mathbb{L}.$$

Notice that for fixed $\{\boldsymbol{\lambda}_i\}$, $\max_{\{\mathbf{x}_i\}}\mathbb{L}$ can be achieved by optimizing over each $\mathbf{x}_i$ *independently*. The dual decomposition strategy is most advantageous when this can be done quickly.

Dual Decomposition of the TRW Objective

This section proves the main result, that the TRW objective can be addressed using dual decomposition. First, recall the TRW optimization problem:

$$B(\boldsymbol{\theta}) = \max_{\mathbf{b}\in\mathcal{L}} \boldsymbol{\theta}\cdot\mathbf{b} + \sum_T \rho_T H_T(\mathbf{b}(T)) \quad (5)$$

Theorem 1. *The TRW objective can be written as*

$$B(\boldsymbol{\theta}) = \min_{\{\boldsymbol{\theta}^T\}}\max_{\{\mathbf{b}^T\in\mathcal{M}_T\}} \sum_T \big(\boldsymbol{\theta}^T\cdot\mathbf{b}^T + \rho_T H_T(\mathbf{b}^T)\big)$$
$$\textit{s.t.} \quad \forall a, \sum_{T:a\in T}\theta^T_a = \theta_a.$$

Algorithm 1 Computing the value and gradient of the master problem M.

Initialize M to 0.
For all T:

1. Set $\bar{\mathbf{b}}^T \leftarrow \arg\max_{\mathbf{b}^T \in \mathcal{M}_T} \boldsymbol{\theta}^T \cdot \mathbf{b}^T + \rho_T H(\mathbf{b}^T)$ by running the sum-product algorithm on the graph T with parameters $\boldsymbol{\theta}^T/\rho_T$.
2. $M \leftarrow M + \boldsymbol{\theta}^T \cdot \mathbf{b}^T + \rho_T H_T(\mathbf{b}^T)$
3. $\frac{dM}{d\boldsymbol{\theta}^T} \leftarrow \bar{\mathbf{b}}^T$

Here, $\mathbf{b}^T$ and $\boldsymbol{\theta}^T$ denote beliefs and parameters for tree T. This theorem shows that the TRW problem can be decomposed into subproblems of the form

$$S_T(\boldsymbol{\theta}^T) = \max_{\mathbf{b}^T \in \mathcal{M}_T} \boldsymbol{\theta}^T \cdot \mathbf{b}^T + \rho_T H_T(\mathbf{b}^T),$$

the optimum of which can be found by running the traditional sum-product algorithm on the graph T with parameters $\boldsymbol{\theta}^T/\rho_T$. Further, by Danskin's theorem, $dS_T/d\boldsymbol{\theta}^T = \bar{\boldsymbol{b}}^T$, where $\bar{\mathbf{b}}^T$ are the maximizing beliefs.

The master problem has the form

$$\min_{\{\boldsymbol{\theta}^T\}} \sum_T S(\boldsymbol{\theta}^T) \quad \text{s.t.} \quad \forall a, \sum_{T:a\in T} \theta_a^T = \theta_a. \tag{6}$$

This is a convex minimization under simple linear constraints, and so can be solved by many standard methods. In particular, notice each constraint in Eq. 6 only affects the block of variables $\{\theta_a^T\}$ for a single a, and so does not "overlap" on variables with other constraints.

Proof of Theorem 1. First, note that we can write Eq. 5 as

$$\begin{aligned} B(\boldsymbol{\theta}) = \quad & \max_{\{\mathbf{b}^T \in \mathcal{M}_T\}} \quad \sum_T \big(\boldsymbol{\phi}^T \cdot \mathbf{b}^T + \rho_T H(\mathbf{b}^T)\big) \\ & \text{s.t.} \qquad b_a^T = \frac{1}{N_a} \sum_{G:a\in G} b_a^G, \end{aligned}$$

where $N_a = |\{G : a \in G\}|$, and the weights $\boldsymbol{\phi}^T$ have been chosen so that

$$\theta_a = \sum_{T:a\in T} \phi_a^T. \tag{7}$$

Taking the Lagrangian, we have

$$\begin{aligned} B(\boldsymbol{\theta}) = \quad & \min_{\{\boldsymbol{\lambda}^T\}} \max_{\{\mathbf{b}^T \in \mathcal{M}_T\}} \quad \sum_T \big(\boldsymbol{\phi}^T \cdot \mathbf{b}^T + \rho_T H(\mathbf{b}^T) \\ & + \textstyle\sum_{a\in T} \lambda_a^T \quad (b_a^T - \frac{1}{N_a} \sum_{G:a\in G} b_a^G)\big). \end{aligned} \tag{8}$$

Now, define $\boldsymbol{\theta}^T$ by

$$\theta_a^T = \phi_a^T + \lambda_a^T - \frac{1}{N_a} \sum_{G:a\in G} \lambda_a^G.$$

We can now transfer the condition on $\boldsymbol{\phi}^T$ from Eq. 7 to $\boldsymbol{\theta}^T$ by observing

$$\begin{aligned} \sum_{T:a\in T} \theta_a^T &= \sum_{T:a\in T} \big(\phi_a^T + \lambda_a^T - \frac{1}{N_a} \sum_{G:a\in G} \lambda_a^G\big) \\ &= \sum_{T:a\in T} \phi_a^T = \theta_a. \end{aligned}$$

Finally, substituting, $\boldsymbol{\theta}^T$ into Eq. 8 gives the result. □

Experiments

These experiments[1] compare the proposed dual decomposition approach to three tree-reweighted message passing algorithms: traditional TRW, TRW with a damping factor of $\frac{1}{2}$ in the log domain (Wainwright and Jordan 2008, p. 174), and the provably convergent variant TRW-S(Meltzer, Globerson, and Weiss 2009).

All experiments are on what has become the most common benchmark for high-treewidth inference algorithms, namely a 10x10 pairwise grid of the form $p(\mathbf{x}) \propto \exp\big(\sum_{ij} \theta(x_i, x_j) + \sum_i \theta(x_i)\big)$, for $x_i \in \{-1, 1\}$. The "field" parameters are of the form $\theta(x_i) = \alpha_F x_i$ where α_F is drawn uniformly from $[-1, 1]$. The "interaction" parameters are of the form $\theta(x_i, x_j) = \alpha_I x_i x_j$, where α_I is chosen from different distributions to represent six different settings: *mixed* potentials of various strengths, $\alpha_I \in [-1, 1], [-3, 3], [-9, 9]$, and *attractive* potentials of various strengths, $\alpha_I \in [0, 1], [0, 3], [0, 9]$.

Here, the dual decomposition objective from Eq. 6 is optimized using the limited-memory BFGS algorithm. We impose the linear equality constraints by reparameterization. Namely, we optimize over the unconstrained parameters $\boldsymbol{\gamma}^T$, setting $\theta_a^T = \frac{1}{N_a}\theta_a + \gamma_a^T - \frac{1}{N_a}\sum_{G:a\in G} \gamma_a^T$, which guarantees that $\sum_{T:a\in T} \theta_a^T = \theta_a$. Given the derivatives of the master problem objective M with respect to $\boldsymbol{\theta}^T$, the derivatives with respect to $\boldsymbol{\gamma}^T$ are also available[2].

All results use uniform edge-appearence probabilities of $\frac{1}{2}$. For dual decomposition, two trees are used: One consisting of all horizontal links, the other of all vertical links.

We compare all algorithms in terms of the number of iterations necessary to reach various levels of convergence. One iteration for TRW denotes a full pass over the grid from top left to bottom right, and then from bottom right to top left. One iteration for dual decomposition denotes one call to the Alg. 1, meaning one call to the sum-product algorithm for each tree. To accurately reflect running-times, this includes calls made during line searches. TRW thus update each message from a factor to a variable twice in one iteration. TRW-S and dual-decomposition however, update each message only once due to the fact that messages are updated along one "direction" and so take roughly half as much time per iteration (The overhead of L-BFGS itself is small).

[1] All algorithms were implemented in Python, with C-extensions for message-passing for efficiency.

[2] $\frac{dM}{d\gamma_a^T} = \sum_{G:a\in G} \frac{dM}{d\theta_a^G}\frac{d\theta_a^G}{d\gamma_a^T} = \frac{dM}{d\theta_a^T} - \frac{1}{N_a}\sum_{G:a\in G} \frac{dM}{d\theta_a^G}$

Here, convergence is measured by first performing inference to a very high degree of accuracy using dual decomposition. Convergence is then measured as $|\mathbf{b}^t - \mathbf{b}^*|_\infty$, where $\mathbf{b}^*$ are the high accuracy marginals, and $\mathbf{b}^t$ are the predicted marginals at iteration t. (For dual decomposition, this is defined by averaging over all trees that contain each element.) It was verified that there is negligible bias introduced by using dual decomposition to generate $\mathbf{b}^*$ by checking that all algorithms converge to an accuracy of better than 10^{-7} if run for sufficiently many iterations (except when TRW fails to converge). Dual decomposition was used simply because it usually converges to high accuracy much faster. Figs. 1 and 2 show scatterplots comparing the number of iterations necessary for dual decomposition to reach each of three convergence levels with each of the message passing algorithms over 500 randomly generated problems, while Fig. 3 shows median statistics.

Discussion

This paper proposes a dual decomposition of the TRW objective. As a standard optimization algorithm can be efficiently applied to the master problem, this method inherits the properties of that algorithm, such as guaranteed convergence, and fast convergence rates. Using L-BFGS, dual decomposition is seen to be one to two orders of magnitude faster on difficult Ising grids.

The reader may question the need to compute approximate marginals to high accuracy. After all, the approximation error in a variational method is probably far larger than 10^{-6}. The need motivating this research was parameter learning. When using approximate inference as part of a parameter fitting procedure, high accuracy is necessary to avoid instability. This issue is particularly pronounced if, for example, computing the loss gradient using finite-difference perturbation(Domke 2010). Secondly, the number of iterations necessary to reach a given level convergence greatly depends on the particular problem. Early termination can be dangerous, since after running a particular number of message-passing iterations, the actual degree of convergence is unknown.

A limitation of the proposed algorithm is the requirement to explicitly process each tree in the TRW bound, rather than just using edge appearance probabilities, as with message passing algorithms. In many cases, this is no problem. Where each edge participates in only one tree (as is typical for grids), this represents no overhead. However, this could be prohibitively expensive if one wanted to make use of a very complicated tree bound. One natural idea would be to instead make use of an arbitrary set of trees *covering* the original graph, rather than the same trees used in the TRW bound. This strategy could also allow the use of other proposed convex entropies(Meshi et al. 2009; Hazan and Shashua 2009; Heskes 2006). This leads to subproblems of the form

$$S_T(\boldsymbol{\theta}^T) = \max_{\mathbf{b}^T \in \mathcal{M}_T} \boldsymbol{\theta}^T \cdot \mathbf{b}^T - \sum_{\alpha \in T} \sum_{\mathbf{x}_\alpha} c_\alpha b(\mathbf{x}_\alpha) \log b(\mathbf{x}_\alpha) - \sum_{i \in T} \sum_{x_i} c_i b(x_i) \log b(x_i).$$

Unfortunately, though these problems can be solved with message-passing(Meshi et al. 2009, Eqs. 9-10), convergence does not occur in a single pass of updates like the sum-product algorithm, even on a tree, meaning a major loss of efficiency in solving subproblems. To the author's knowledge, no algorithm is known that converges in a single pass for this problem.

References

Domke, J. 2010. Implicit differentiataion by perturbation. In *NIPS*.

Globerson, A., and Jaakkola, T. 2007a. Approximate inference using conditional entropy decompositions. In *AISTATS*.

Globerson, A., and Jaakkola, T. 2007b. Convergent propagation algorithms via oriented trees. In *UAI*.

Hazan, T., and Shashua, A. 2009. Norm-prodcut belief propagtion: Primal-dual message-passing for lp-relaxation and approximate-inference. Technical report, Leibniz Center for Research, The Hebrew University.

Heskes, T.; Albers, K.; and Kappen, B. 2003. Approximate inference and constrained optimization. In *UAI*.

Heskes, T. 2006. Convexity arguments for efficient minimization of the bethe and kikuchi free energies. *J. Artif. Intell. Res. (JAIR)* 26:153–190.

Kolmogorov, V. 2006. Convergent tree-reweighted message passing for energy minimization. *IEEE Trans. Pattern Anal. Mach. Intell.* 28(10):1568–1583.

Komodakis, N.; Paragios, N.; and Tziritas, G. 2007. MRF optimization via dual decomposition: Message-passing revisited. In *ICCV*.

Meltzer, T.; Globerson, A.; and Weiss, Y. 2009. Convergent message passing algorithms - a unifying view. In *UAI*.

Meshi, O.; Jaimovich, A.; Globerson, A.; and Friedman, N. 2009. Convexifying the bethe free energy. In *UAI*.

Teh, Y. W., and Welling, M. 2001. The unified propagation and scaling algorithm. In *NIPS*.

Wainwright, M. J., and Jordan, M. I. 2008. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.* 1(1-2):1–305.

Wainwright, M. J.; Jaakkola, T.; and Willsky, A. S. 2005a. A new class of upper bounds on the log partition function. *IEEE Transactions on Information Theory* 51(7):2313–2335.

Wainwright, M. J.; Jaakkola, T. S.; and Willsky, A. S. 2005b. Map estimation via agreement on trees: message-passing and linear programming. *IEEE Trans. on Information Theory* 51(11):3697 – 3717.

Yedidia, J. S.; Freeman, W. T.; and Weiss, Y. 2005. Constructing free energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory* 51:2282–2312.

Yuille, A. L. 2002. CCCP algorithms to minimize the Bethe and Kikuchi free energies: convergent alternatives to belief propagation. *Neural Computation* 14(7):1691–1722.

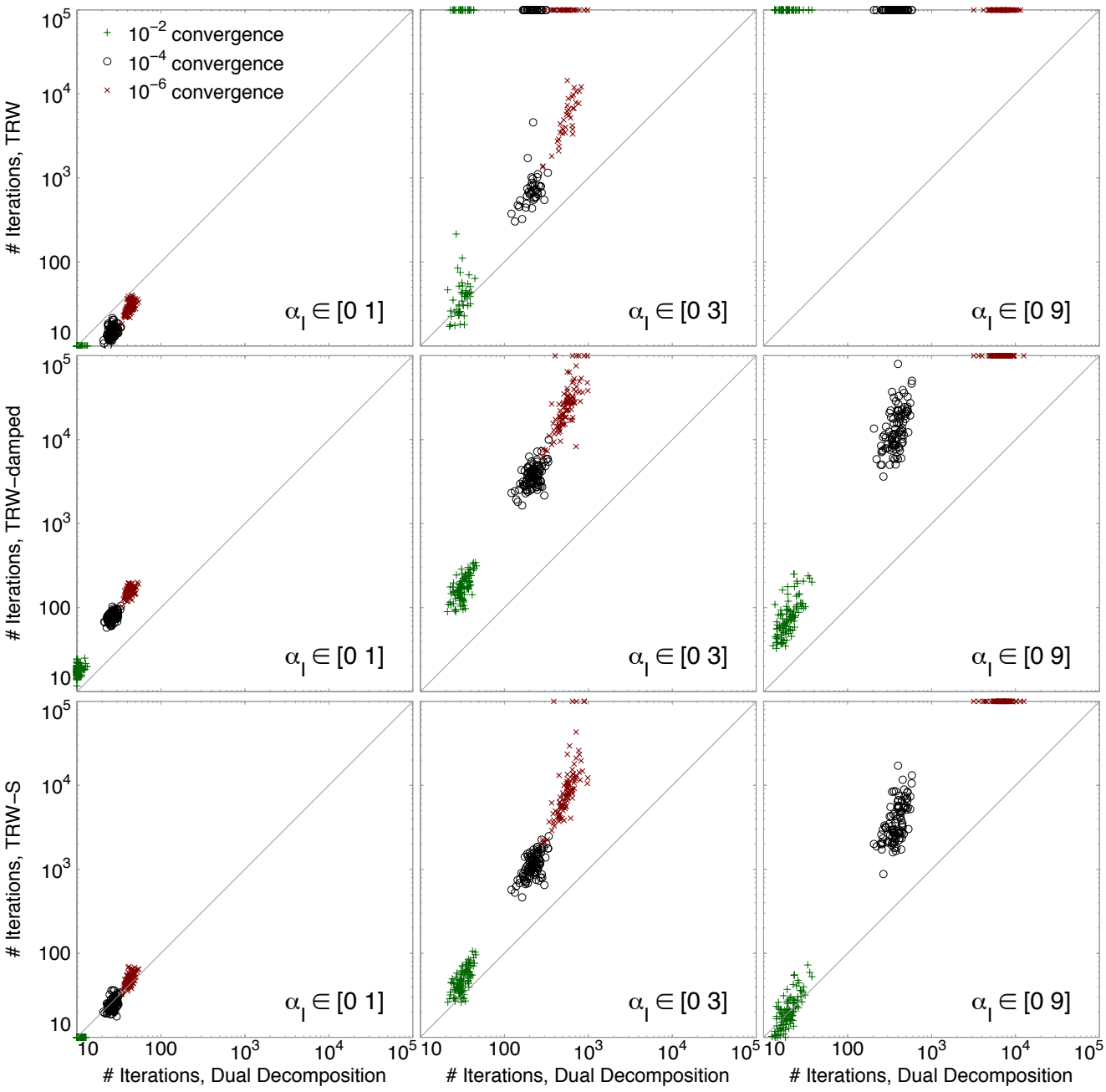

Figure 1: Scatterplots comparing the number of iterations necessary to reach three levels of convergence on 500 random problems with field potentials $\alpha_F \in [-1, 1]$ and interaction potentials α_I of three *attractive* strengths. With weak interactions strengths, the message passing algorithms perform comparably to dual decomposition, while with strong interactions, dual decomposition is faster. The number of iterations required by dual decomposition is more concentrated than for message passing. A point is plotted at 10^5 iterations if convergence was not achieved by then. For low interaction levels and low accuracy convergence, message-passing is somewhat faster, though dual decomposition also performs well. At higher interaction levels and higher levels of convergence dual decomposition is faster.

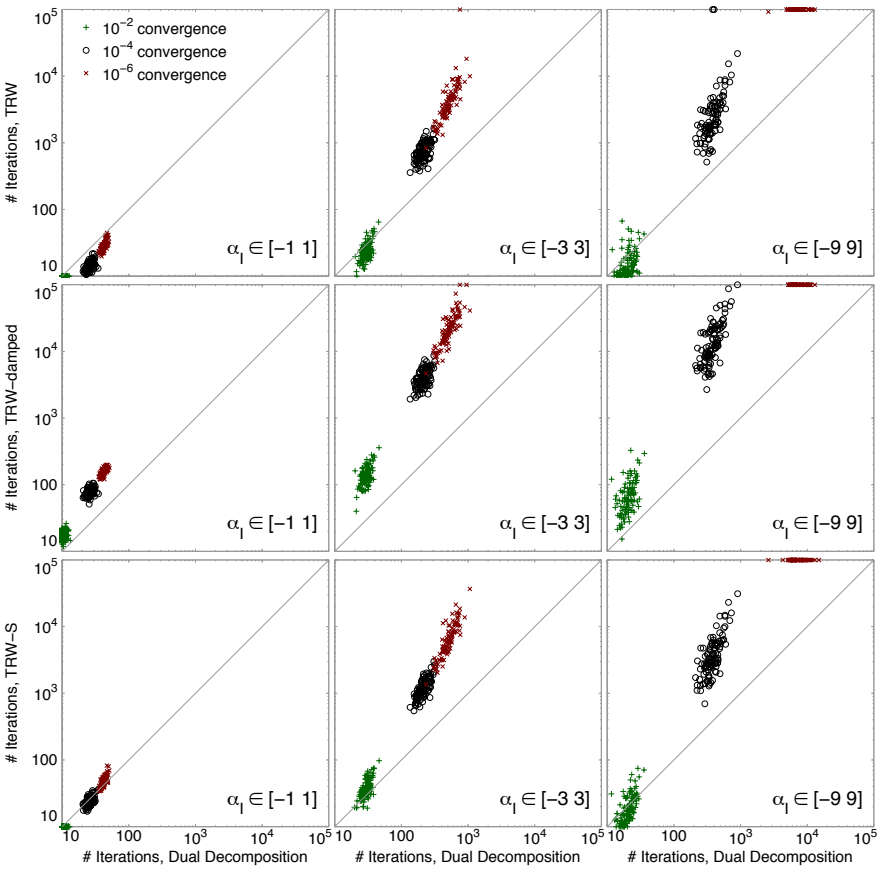

Figure 2: The experiment from Fig. 1 with interaction potentials of three *mixed* strengths. For low interaction levels and low accuracy convergence, message-passing is somewhat faster, though dual decomposition also performs well. At higher interaction levels and higher levels of convergence dual decomposition is faster.

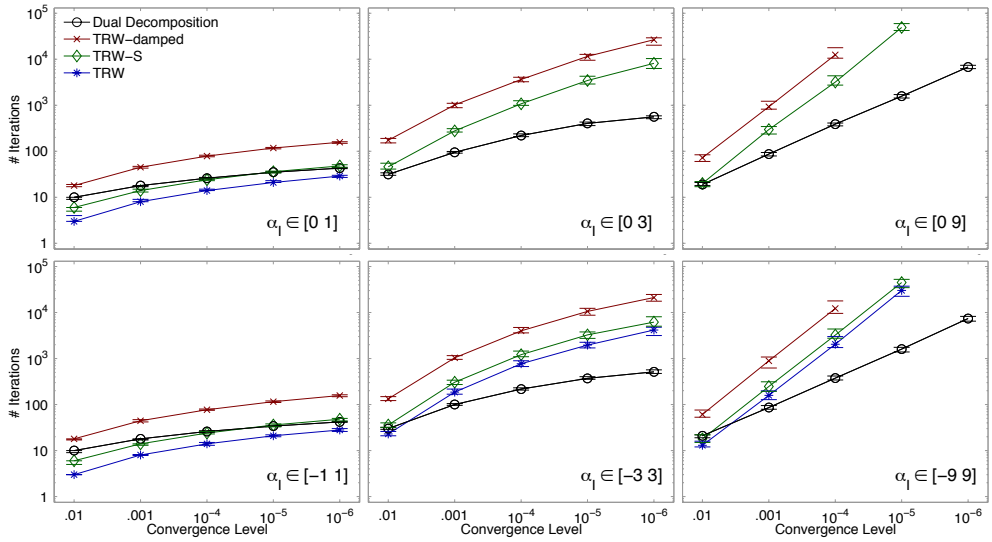

Figure 3: The median number of iterations required for the dual decomposition and message passing methods in various settings, along with 99% confidence intervals, shown with horizontal bars. Top: attractive potentials. Bottom: Mixed potentials. Since TRW does not always converge, it is only shown on a subset of settings. TRW-damped and TRW-S appear to always converge, but sometimes not within 10^5 iterations. These are plotted when enough problems converge within 10^5 iterations to estimate the median and confidence intervals.